Sequence data mining and characterisation of unclassified microbial diversity

Sejal Modha

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

College of Medical, Veterinary and Life Sciences
University of Glasgow

June 2022

Abstract

In the last two decades, sequencing has become increasingly affordable and a routine tool to study the microbial community of a given environment. Metagenomics has revolutionised the way microbes are identified and studied in this age of biological data science because it provides a relatively unbiased view of the composition of microbial communities we interact with every day, which are integral to our ecosystem. These technological advances have led to an exponential growth of raw data repositories that save, distribute and archive these metagenomic datasets. Since metagenomics presents the ultimate opportunity to capture, explore and identify uncultivated microbial genomic sequences, these metagenomic datasets harbour a large proportion of unknown sequences that do not bear any similarity to known sequences readily available in the standard sequence data repositories. The aim of this thesis was to systematically catalogue, quantify and potentially characterise the unknown sequences embedded within the metagenomic datasets. To this end, a comprehensive, portable, modular framework called UnXplore was developed to determine the proportion of unknown sequences included in human microbiome datasets. UnXplore was applied to a range of different human microbiomes and showed that on average 2% of assembled sequences were categorised as unknown meaning that they did not bear any sequence similarity to known sequences. A third of the unknown sequences were shown to contain large open reading frames indicating the coding potential and biological origin of the unknowns. Furthermore, a small proportion of these potentially coding sequences were shown to have functional similarities as they were deemed to contain known protein domain signatures. These results indicated that unknown sequences captured through the UnXplore framework were not artefacts and were indeed of biological origin. To test this formally, supervised k-mer-based machine learning models were devised, tested and validated. These models are currently distributed in a package called TetraPredX that can accurately predict whether a sequence originated from bacteria, archaea, virus or plasmid. TetraPredX models were applied to the unknown sequence dataset and revealed that the majority of unknown sequences are of biological origin. Furthermore, TetraPredX results demonstrated that >70% of all long unknown sequences (i.e. >1kb) are likely to be of virus origin indicating an unexplored diversity of viruses that is yet to be fully characterised and classified. In order to catalogue the diversity of virus sequences in human microbiome samples analysed here, an extensive virus discovery analysis was carried out on the contigs assembled through UnXplore. This helped to characterise a vast diversity of prokaryotic, eukaryotic and unclassified virus sequences captured in a range of human microbiomes. The results obtained here demonstrate the need to systematically interrogate metagenomic datasets to fully comprehend and compile the presence of both known and unknown uncultivated microbes within them. A comprehensive survey of metagenomic datasets carried out in this manner would provide a more complete picture of the known and unknown organisms that surround us.

Contents

List of Tables

List of Figures

Dedicated to
those who are no longer with us but continue to inspire.

Acknowledgements

First, I would like to thank my PhD supervisors, David Robertson, Richard Orton and Joseph Hughes for giving me this opportunity to work on this exciting project that has helped me grow as a researcher in the past few years. Special thanks to Richard and Joseph for being there every step of the way during this memorable journey. I am most grateful for the support I was provided with throughout this PhD, especially during the pandemic. Despite their busy schedules, they have always made time, provided constructive feedback and encouraged stimulating discussion pushing me out of my comfort zone. Without their constant support, encouragement and expertise, this work would not have been possible. I would like to thank MRC Precision Medicine for providing the funding for this project.

I am grateful to my annual review assessors, Andrew Davison, Kathryn Crouch and George Bailey for providing helpful advice along the way. Special thanks to Andrew and Kathryn for providing extensive guidance during this PhD and beyond.

I would like to extend my thanks to all Robertson lab members including Fran, Kieran, Spyros, Vandana and Haiting for their input in my project. Fran and Haiting's input and expertise in the machine learning section of my project were greatly appreciated. A very special thanks to Vandana for introducing me to LaTeX, sharing some cool tips and tricks for using LaTeX, and helping with troubleshooting weird Overleaf/LaTeX errors. I would like to thank Scott for his continuous IT, server and computation requirement-related support and for making sure that computational resources were maintained to the highest standard for me to carry out crucial analyses required for this project.

I would like to thank Donna Macpherson, Fiona Graham, Evelyn McIntosh and Michelle Pearson for their assistance with the administrative aspects of this PhD project at the Centre for Virus Research and the University of Glasgow. Special thanks to both Donna and Fiona for their extended support during the pandemic. I would also like to extend my thanks to Precision Medicine's administrative staff for their prompt and accurate responses to my queries. Thanks to CVR PGR representatives, Joanna and Spyros, for attending to our needs and voicing our opinions to the CVR and MVLS management.

Every professional with whom I have worked has taught me something, and these experiences have been tremendously valuable for my PhD. My thanks are extended to each and every one of them.

On a personal note, I would like to thank my parents Kirti and Bipinchandra for helping me develop the discipline and determination required to succeed in my PhD and beyond. Words cannot express my gratitude to my sister Kruti, and my brother Darshil. They have comforted me and been a constant source of unfailing encouragement, a positive outlook, and humour that has sustained me even in the darkest times. I would like to express my deepest gratitude to my husband, Gautam. This endeavour would not have been possible without him. He encouraged me to pursue my dream and pushed me to make the most of opportunities.

I would also like to acknowledge and thank friends and family in the UK, India and around the world. I'm eternally grateful to them for being supportive in difficult and fun times. Thank you for always making an effort to keep in touch with me and checking in on me. I would like to thank my friends who are in Glasgow; Maha, Quan, Meha and Jeevan for regular meetups, in-person and virtual conversations and much-needed banter that kept me sane.

Many things have changed in the past few years, and I have been challenged in a variety of ways. However, these changes have also brought me a wealth of experience and knowledge.

Declaration

I, Sejal Modha, declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

June 2022

Abbreviations

AF : Alignment fraction

AMG : Auxiliary metabolic genes

ANI : Average nuclotide identity

AUC : Area under the ROC curve

AWS : Amazon web services

BAM : Binary aligment map

BEG : Bioinformatics experts group

BLAST : Basic local aligment search tool

BWA : Burrows-Wheel aligner

CAMI : Critical Assessment of Metagenome Interpretation

CPR : Candidate phyla radiation

CRESS DNA : circular replication(Rep)-encoding single-stranded

CRISPR : Clustered Regularly Interspaced Short Palindromic Repeats

CV : Cross validation

DFAST : DDBJ Fast Annotation And Search Tool

DJR-CP : Double-jelly-roll capsid protein

DNA : Deoxyribonucleotic acid

DPANN : Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, Nanohaloarchaea

DRAM : Distilled and Refined Annotation of Metabolism

dsDNA : Double-tranded DNA

dsRNA : Double-stranded RNA

DTR : Direct Terminal Repeats

EBI : European Bioinformatics Institute

ENA : European Nucleotide Archives

FN : False Negative

FP : False Positive

FPR : False Positive Rate

GCP : Google Cloud Platform

GPD : Gut Phage Database

GVD : Gut Virome Database

HGTs : Horizontal Gene Transfers

HIV : Human Immunodeficiency Virus

HMM : Hidden Markov model

HMP : Human Microbiome Project

HTS : High throughput sequencing

ICEs : Integrative conjugative elements

ICTV : International committee on taxonomy of viruses

INSDC : International Nucleotide Sequence Database Collaboration

ITR : Inverted Terminal Repeats

kb : Kilobases

LCA : Lowest Common Ancestor

lncRNA : Long non-coding RNA

MAG : Metagenome-assembled genome

mb : Megabases

MGE : Mobile Genetic Elements

MGV : Metagenomic Gut Virus

MIUViG : Minimum Information about an Uncultivated Virus Genome

ML : Machine learning

mRNA : Messenger ribonucleic acid

NCBI : National Center for Biotechnology Information

NCLDV : Nucleocytoplasmic Large DNA Viruses

NLP : Natural language proecessing

NR : Non-redundant

NT : Nucleotide

OLC : Overlap Layout Consensus

ORF : Open reading frame

OTU : Operational taxonomic unit

PCA : Principal component analysis

PE : Paired-end

PyPI : Python Package Index

RATT : Rapid annotation transfer tool

RCRE : rolling-circle replication (initiation) endonuclease

RdRp : RNA-directed RNA polymerase

RF : Random forest

RFC : Random forest classifier

RNA : Ribonucleic acid

ROC : Receiver Operating Characteristic

rRNA : Ribosomal RNA

RSCU : Relative synonymous codon usage
RT : Reverse transcriptase
S3H : Superfamily 3 helicase
SAM : Sequence alignment map
SARS-CoV-2 : severe acute respiratory syndrome coronavirus 2
SE : Single-end
SJR-CP : single-jelly-roll capsid protein
SRA : Sequence Read Archives
ssDNA : Single-stranded DNA
ssRNA : Single-stranded RNA
STAT : SRA Taxonomy Anlaysis Tool
SVM : Support Vector Machine
t-SNE : t-distributed stochastic neighbour embedding
TN : True Negative
TNF : Tetranucleotide frequencies
ToL : Tree of Life
TP : True Postive
TPR : True Positive Rate
TRACA : transposon-aided capture
TTMV : Torque teno mini virus
TTV : Torque teno virus
TTVMD : Torque teno midi virus
UCs : Unknown contigs
UMAP : Uniform Manifold Approximation and Projection
UViCs : Unclassified viral contig sequences
UViGs : Uncultivated Virus Genomes
VAPiD : Viral Annotation Pipeline and iDentification
VHGs : Viral hallmark genes
vMAG : viral metagenome-assembled genomes
VMR : Virus Metadata Resource
vOTU : virus operational taxonomic units

Chapter 1

Outline

Microbes, their diversity and their ubiquity in our biosphere has been fully appreciated through the advent of high throughput sequencing (HTS) and metagenomics. Metagenomics has been widely applied to various environments, microbiomes and clinical samples in the last two decades. Public repositories that host raw sequences such as Sequence Read Archives (SRA) and European Nucleotide Archives (ENA) have grown exponentially, housing >18 petabytes of open-access datasets as a result of high throughput sequencing's widespread application. A total of 4,277,855 publicly accessible metagenomic datasets are available via SRA as of 4 September 2022. It has been shown that these metagenomic sequence datasets contain unknown sequences that are often referred to as biological 'dark matter' (Marcy et al., 2007; Krishnamurthy et al., 2017; Thomas et al., 2019). Typically, metagenomic sequence analysis is carried out with a specific research question at hand, and, does not attempt to catalogue the diversity of unknown/dark sequences and they are often excluded from the downstream analyses. These unknown sequences are hypothesised to be originating from uncultivated microbes (for which no isolated representative exists) and their identification and characterisation could help us get a more complete picture of the complex microbial community that surrounds us.

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic that started in 2020 brought global socio-economic interactions to a standstill. This event emphasised and re-iterated the importance of cataloguing the microbial world that surrounds us. In this era of biological data science and the unstoppable growth of sequence data repositories, it is essential to ensure that each dataset is carefully examined to look for signs of microbes that we regularly interact with as a variety of microbes have a significant role to play in human health, environment, and ecology. My research project aims to identify, catalogue and classify the biological unknowns present in the known sequence space. I intend to achieve this by employing tools, methods, resources and expertise in bioinformatics and computational virology. My endeavours to expand the current scientific knowledge and advance the frontiers of microbial dark matter research are described in detail in this thesis. A brief narrative of the research content included in each chapter is described below.

Chapter 2 of my thesis aims to introduce the main topics and research avenues that form the basis of this dissertation. A thorough background is provided on viruses, metagenomics, and the shift in virus discovery in this era of sequencing and metaviromics. In addition, it introduces the computational approaches utilised to identify virus sequences captured using metagenomics. I want to emphasise the importance of analysing metagenomic datasets to discover unknown sequences that are often described as biological ‘dark matter’. The final section discusses briefly a number of studies that provided the foundation and motivation for investigating the biological unknowns inherent in the microbiome datasets. From a detailed literature review described in this chapter, it was hypothesised that the biological ‘dark matter’ is likely to be mainly of microbial and/or viral origin.

Chapter 3 focuses on the characterisations and quantification of unknowns in the human microbiome dataset. To address this, I developed a comprehensive, modular and portable analysis framework called UnXplore which is described in detail. UnXplore was applied to quantify the unknown sequences (‘dark matter’) in human metagenomic datasets. To determine whether the dark matter catalogued here originated from uncultivated (micro)organisms and to understand their distribution, a detailed comparison of the unknown sequences between samples, studies and microbiomes was carried out. Furthermore, the unknown contigs obtained in this analysis were compared to currently known sequences in publicly available resources such as GenBank over the period of the study to determine the rate at which these unknown contig sequences are being taxonomically classified. This chapter 3 has been published as a research article in mSystems.

Unknown contigs obtained in chapter 3 did not bear any significant sequence similarity to known sequences available in general-purpose nucleotide and protein databases. A third of them were shown to contain open reading frames that were at least 100 amino acid residues long, and a small proportion of them was shown to contain known protein domains. These results supported our initial hypothesis that unknown sequences are likely to be originating from uncultivated microorganisms, specifically viruses. To test this further, alignment-free, supervised machine learning models were explored, developed and tested. A tetranucleotide frequency-based machine learning prediction model embedded within a package called TetraPredX was designed, which is described in detail in Chapter 4. The machine learning prediction models developed in TetraPredX were applied to the unknown sequence dataset and showed that >70% of unknown sequences were of viral origin supporting our hypothesis. TetraPredX is published on Python Package Index (PyPI) and is available on https://pypi.org/project/TetraPredX/.

A comprehensive analysis of the contigs assembled through UnXplore was carried out with a focus on cataloguing human virome and virus discovery in Chapter 5. Over seven million contigs that were greater than 1kb were initially tested to predict those originating from viruses and these predictions were further validated using gold-standard alignment-based approaches. This in-depth survey led to the discovery of thousands of potentially novel prokaryotic and eukaryotic virus sequences found across various microbiomes, samples and BioProjects. Moreover, a large proportion of virus sequences that could not be associated with a known virus family, order or realm were also catalogued. This reiterates the importance of systematic exploration of already ‘analysed’ public datasets.

Finally, Chapter 6 discusses major findings from this research dissertation with the aim to provide an outlook for future analysis. In addition to that, a critical appraisal of the methodology and the overall approach is discussed in detail. Lastly, future avenues of unknown sequence explorations, virus discovery and implementation of effective computation approaches to address the ever-increasing big data challenges in biology are contextualised.

The work documented in this thesis - conceptualisation, data curation, methodology, formal analysis, project management, research, resource and software development, validation, visualisation, and original writing - has been conducted by me. My supervisors contributed to the conceptualisation of the project, the development of the methodology, and the editing and revision of the original text.

Chapter 2

Introduction

The concept of the Tree of Life (ToL) was initially realised by Charles Darwin and the first evidence of it is found in his notebook, The Origin of Species, in 1837 (Darwin, 1859; Harris et al., 2021). This formulates the basis of evolutionary biology as it provided the basic notions of evolution and introduced the concepts of relatedness between living forms and their universal common origin. The ToL encompasses all living organisms that are grouped into three major domains: bacteria, archaea and eukaryotes (Woese et al., 1990). These organisms are grouped into two major categories based on their fundamental cell structure. Prokaryotes (bacteria and archaea) are unicellular and lack membrane-bound structures such as a nucleus, whereas eukaryotes can be either single or multi-cellular organisms with a clearly defined nucleus as well as other cell organelles like mitochondria and golgi apparatus (Vellai et al., 1999). The eukaryotic branch of the ToL on the other hand includes all other living things including animals, plants, fungi, protists and algae.

Bacteria are considered to be the simplest living forms that typically have a single loop of Deoxyribonucleotic acid (DNA) as their genetic material. Some bacterial cells also possess small circular genetic material known as plasmids. The plasmids typically contain genetic material that would give bacteria some advantage over other bacteria. Bacteria replicate by the mechanism of binary fission whereby a parent bacterial cell is divided into two daughter cells with the exact same genetic makeup (Cossart, 2018). Archaea were initially thought to be bacteria and were grouped together with bacteria until the 1970s when this domain of life was added to the ToL (Woese et al., 1977; Fox et al., 1977; McInerney et al., 2008; Woese et al., 1990; DasSarma et al., 2009). Archaea have been found in some of the most extreme and hostile environments (DasSarma et al., 2009; Rampelotto, 2013). Archaea also harbour characteristics found in both bacteria and eukaryotes. Despite their prokaryotic cellular makeup, archaea share some metabolic genes, pathways and enzymes with eukaryotes (DasSarma et al., 2009). Though they are microscopic, prokaryotes make up the majority of known organisms in our biosphere.

2.1 Viruses

In 1898, Martinus Willem Beijerinck initially used the term “contagium vivum fluidum” (Latin for “contagious living liquid”) to describe the tobacco mosaic virus that retained its infectious nature after dilution (Beijerinck, 1898). The term ‘virus’ was then being used to describe anything from toxins to infection agents, eventually, it started getting associated with this new type of pathogen. This pioneered the new branch of biology that focused on studying viruses - Virology (Bos, 2000). Despite the identification of viruses, the structure of the ToL did not change. Viruses are not included in the ToL as they are not considered living things because they don’t have universal genes that are present across all viruses. Moreover, a virus does not possess the required machinery to multiply without a host. In fact, due to the lack of replication mechanism and inability to carry out metabolic processes required to qualify them as living things, they are considered an inert organic matter in absence of a suitable living host cell (Moreira et al., 2009). Viruses can have either DNA or ribonucleic acid (RNA) as their genetic material that is typically wrapped in a protective protein overlay called a capsid. Viruses can also have an additional protein layer (coded by the virus) called an envelope which can help them to evade the host immune system (Rowlands, 2021).

2.1.1 Viruses and the tree of life

Viruses are fundamentally different to living organisms as they cannot independently replicate or produce energy like other cellular lifeforms. Viruses are passive agents that cannot do anything in absence of a suitable host cell. Upon successful entry into a suitable host cell, viruses “hijack” the cell’s replication machinery to propagate and assemble virus particles that are then released to infect and invade more cells accessible in a given environment (Cann, 2021). In their simplest forms, viruses have been deemed the most abundant biological complexes that can infect all cellular life (Harris et al., 2021). However, without any generally acceptable consensus, it is undecided whether viruses are considered “alive” as the core definition of being alive in itself is a debatable topic (Koonin et al., 2016). Due to the polyphyletic nature of virus origin that lacks a shared common ancestor, and a single gene that is shared by all viruses, it has been argued that viruses cannot be placed in the tree of life. Recent studies based on the wealth of virus genome data derived from metagenomics have revealed that viruses may be more interrelated than previously thought (Iranzo et al., 2016; Bin Jang et al., 2019). Viruses have been shown to harbour viral hallmark genes (VHGs) and these genes are shared among different groups of viruses (Koonin et al., 2020a). It has been proposed that viruses should be analysed with a network-based perspective as opposed to tree-based hierarchical approaches to fully comprehend the viral diversity, their relationship with one another and their complex interactions (Bin Jang et al., 2019). Viruses are ubiquitous entities that can infect all cellular life, evolve as biological entities, influence host evolutionary mechanisms and have co-evolved with cellular life. These unique properties harboured exclusively by viruses have been argued as indicative of viruses being strongly linked to cellular organisms and potentially influencing cellular evolution, and hence, as some experts conclude, qualify to be included in the network of tree of life (Harris et al., 2021; Forterre et al., 2021). Viruses are part of the continuum of life, whether they are alive or not is just semantics. Arguably, they are more akin to spores/gametes i.e. the dispersal part of the "virocell" (Forterre, 2011). The "are viruses alive?" debate is considered futile as it misses the point that they are part of life as dependent replicators (Koonin et al., 2016).

2.1.2 Virus Diversity

Regardless of their status, viruses are deemed the most compact, fast-evolving biological entities that are currently known to be present in our biosphere (Roux et al., 2021a). Viruses can vary in shape, size and nucleic acid composition. Overall, viruses are classified into four groups based on their shapes: enveloped, filamentous (long and cylindrical), spherical (isometric or icosahedral) and head and tail (Sevvana et al., 2021). The first identified virus, the tobacco mosaic virus is a filamentous virus (Klug, 1999). In fact, a large proportion of plant viruses are filamentous. Spherical viruses are indeed icosahedral when looked at closely under a microscope (Rux et al., 1998). They consist of equilateral triangles that are fused together to form a spherical shape. Some examples of viruses in this group are herpesvirus, adenovirus, rhinovirus, and poliovirus. Enveloped viruses have an additional outer layer or membrane as indicated in the name. This envelope could be typically derived from the host cell membrane and also contain some glycoproteins that are coded by the virus (Sevvana et al., 2021). Animal viruses such as Human immunodeficiency virus (HIV), influenza, and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) are enveloped viruses. Head and tail viruses typically include those that infect prokaryotes i.e. bacteria and archaea. These viruses have a head that is similar to icosahedral viruses and a tail shape like filamentous viruses. These groups of viruses that typically infect bacteria are referred to as bacteriophages (Sevvana et al., 2021). Although most viruses have a defined shape, some of them such as influenza have been observed to exist in both spherical and filamentous shapes (Bouvier et al., 2008). A number of viruses use a form of glycoprotein to attach to their host cells via molecules on the cell called viral receptors. These surface glycoprotein attachments are used as tunnels to penetrate virus' genetic material into the cell and subsequently replicate inside the cell (Dimitrov, 2004; Casasnovas et al., 2021). Because of the range of shapes, sizes, infection mechanisms, and nucleic acid types, viruses can be classified based on one or more of these characteristics.

Unlike cellular life that uses double-stranded DNA to store genetic material, viruses can use DNA or RNA as their genetic material, which can be either double or single-stranded (Sanjuán et al., 2021). Cellular organisms use their double-stranded DNA to store, replicate and cypher its genetic information. The DNA is transcribed into messenger RNA (mRNA) by DNA polymerase, and mRNA is translated into functional units i.e. proteins with the help of ribosomes. This process commonly referred to as the 'central dogma' of molecular biology is deemed the most important flow of genetic information within a biological system (Crick, 1970). As viruses have different genetic makeup, their version of this information flow can be different. For example, RNA viruses can have either positive or negative sense genomes. A negative-sense genome implies that the viral nucleic acid (typically found in RNA viruses) cannot be readily translated into protein and it requires synthesis of the complementary strand first that can be translated into proteins later. Depending on the nucleic acid types, viruses require different replication mechanisms, transcription strategies and polymerase interactions to complete the genome synthesis (Baltimore, 1971; Cann, 2021). For example, viruses that infect and stay in the cell cytoplasm have to encode their own polymerase as they do not have access to host polymerase. On the contrary, viruses that localise in the cell nucleus rely completely on the host cell machinery for complete genome synthesis (Choi, 2012).

Although the virus genome diversity, relatedness among different virus groups, host range and other factors that allow or restrict viruses to replicate in a given environment is well understood, there is no consensus among virology experts about the origin of viruses (Forterre et al., 2021). Currently, there are three major hypotheses (Forterre, 2006): (1) Virus-first hypothesis (2) Reduction hypothesis (regressive) (3) Escape hypothesis (progressive). The virus-first hypothesis states that viruses predate the origin of cellular life and it potentially played an important role in shaping the first cellular life. Support for this hypothesis is that viruses have RNA as their genetic material and that scientists generally understand that RNA predates DNA as a replicative material (Poole et al., 2000; Forterre, 2001; Forterre, 2002; Wolf et al., 2007; Durzyńska et al., 2015). The reduction hypothesis states that viruses originated from small primordial cells that lost a range of cellular functional components but retained the most important functional elements required for replication. The escape hypothesis states that viruses are derived from mobile genetic elements such as plasmids and transposons (Koonin et al., 2006). To understand the origin of viruses, it may be necessary to explore more than one explanation. Regardless of the origin of viruses, they are capable of infecting all cellular lifeforms in the tree of life from bacteria, fungi, and plants to vertebrates (Harris et al., 2021). Moreover, although viruses have been associated with infectious diseases and have a reputation as disease-causing agents, a large proportion of viruses are in fact an integral part of the natural world and ecology. Naturally occurring viruses have been found in most parts of ecosystems and microbiomes, suggesting a reasonably underappreciated relationship between viruses and their hosts potentially as symbionts that shaped the living world as we know it today (Roossinck, 2011; Roossinck et al., 2017; Koonin et al., 2021a).

Depending on the nucleic acid type of the viral genome, and the mechanism and strategies needed to synthesise mRNA, viruses are grouped into seven different classes. This is referred to as the Baltimore classification system (Baltimore, 1971) of viruses and includes the following seven classes:

- Baltimore Class I: Double-stranded DNA (dsDNA) viruses

- Baltimore Class II: Single-stranded DNA (ssDNA) viruses
- Baltimore Class III: Double-stranded RNA (dsRNA) viruses
- Baltimore Class IV: Positive-sense single-stranded RNA (ssRNA) viruses
- Baltimore Class V: Negative-sense single-stranded RNA viruses
- Baltimore Class VI: Single-stranded RNA-RT viruses with positive-sense RNA with DNA intermediates produced by reverse transcription of the viral genome
- Baltimore Class VII: Double-stranded DNA viruses with an RNA intermediate in their life-cycle

Largely, Baltimore classes are referred to as the informal highest ranks representing virus diversity as each Baltimore class BC was assumed to share a common ancestor (i.e. being monophyletic). However, recent studies and comprehensive phylogenetic analyses by Koonin et al. (2020a) and Wolf et al. (2018) have challenged the idea of Baltimore classes being monophyletic as they observed significant overlap and gene exchange among them suggesting that they may not be suitable to be used as top-level ranks for virus groupings.

Virus genome lengths range from <2 kilobases (kb) to several thousand kilobases. CRESS (circular replication-associated protein (Rep)-encoding single-stranded) DNA viruses that are classified in the family *Circoviridae* have the smallest genome of around 1.7-2.1kb (Breitbart et al., 2017). Compared to these small single-stranded DNA viruses, double-stranded DNA viruses included in phylum *Nucleocytoviricota* have the largest genomes observed in the viral world so far. These viruses are often referred to as giant viruses that depict their large physical and genome sizes, and, are included in a group called nucleocytoplasmic large DNA viruses (NCLDV). The largest NCLDV known to date is *Megavirus chilensis* and it has a genome length of around 1.26Mb. These giant viruses that infect amoeba are included in the family *Mimiviridae* (Legendre et al., 2012). In contrast to DNA viruses, RNA viruses can have segmented genomes meaning that their genomes can be split into more than one fragment, each coding for specific proteins required for virus infection, replication and host immune system invasion. Among RNA viruses, the largest non-segmented genomes are found in viruses included in the order *Nidovirales*. Perhaps the most studied viruses within this group are included in the family *Coronaviridae* which contain single-stranded positive-sense RNA viruses that can infect mammals, birds and fish. Their genome sizes can vary between 26-32kb in length (ICTV, 2012). Double-stranded RNA viruses included in the family *Reoviridae* genomes are composed of the largest number of segments. They have 18–29 kbp of segmented (9–12) linear dsRNA genomes with the segments ranging from 0.7–5.8 kbp (https://talk.ictvonline.org/ictv-reports/ictv_online_report/dsrna-viruses/w/reoviridae). These viruses infect a range of cellular life from algae, fungi, plants, invertebrates, aquatic animals, birds to mammals. Recently, a new group of negative-sense single-stranded RNA (ssRNA) viruses have been discovered that can have segmented and/or nonsegmented genomes that may be also circular. These viruses are included in order *Jingchuvirales* and have been predominantly found to infect invertebrates (Di Paola et al., 2022). The smallest RNA viruses are those that are included in the family *Kolmioviridae* and have negative-sense single-stranded RNA genomes that are around 1.7kb long. These viruses were initially found in Hepatitis-B infected humans but have recently been identified in the transcriptomes of a range of invertebrates and vertebrates (Chang et al., 2019; Bergner et al., 2021).

2.1.3 Virus taxonomy

Virus taxonomy is a very important field of science that specialises in the grouping of viruses into artificial categories called taxa. It also develops, executes and utilises systematic naming conventions to group viruses into different taxa. It is formed of expert virologists that specialise in grouping viruses according to their properties (Fauquet, 2008; King et al., 2021). Viruses - at species level and above - are formally classified by the International Committee of Virus Taxonomy (ICTV) which was initially formed in 1966 at the International Congress for Microbiology in Moscow. At the time, it was known as the International Committee on Nomenclature of Viruses (ICVN), which became today's ICTV in 1974. The first ICTV report was published in 1971 (King et al., 2021) and it contained 290 virus species grouped into two virus families and 43 genera. Today, ICTV executive committee has established 100 international study groups covering all virus taxa. Each study group consists of world-leading virology experts and researchers who voluntarily contribute to streamlining virus taxonomy (https://talk.ictvonline.org/taxonomy/w/ictv-taxonomy). They play a critical role in rigorously assessing and reviewing taxonomic proposals submitted to the ICTV (King et al., 2021).

Virus taxonomy ratified and compiled by the ICTV contains viruses, viroids and satellite viruses. Initially, only viruses that were isolated using the traditional culture-based methods were recognised and incorporated into the taxonomy framework. Due to this method of virus identification relevant external properties including morphology, pathogenicity, host range etc have been traditionally used to classify viruses. These characteristics are often specific and customised to a group of viruses. These properties of virus classification have been adapted and extended to include other features that focus on genomic composition and evolutionary relatedness to keep up with the virus discovery in this new era of DNA sequencing and metagenomics. Virus taxonomy structures and organises individual viruses into the tree of life-like structure that has 7 mandatory and 7 optional taxonomic ranks that has 'realm' as the highest taxonomic rank and 'species' as the lowest (King et al., 2021). Although viruses are grouped based on artificial and often arbitrary demarcation criteria, the taxonomic framework is deemed crucial to studying viruses. It compiles and connects all viruses into a systematic framework that facilitates the research community to understand the global organisation of known viruses, as well as enabling them to rapidly explore their biological, clinical and evolutionary relationship with one another and their corresponding host(s) (Davison et al., 2020).

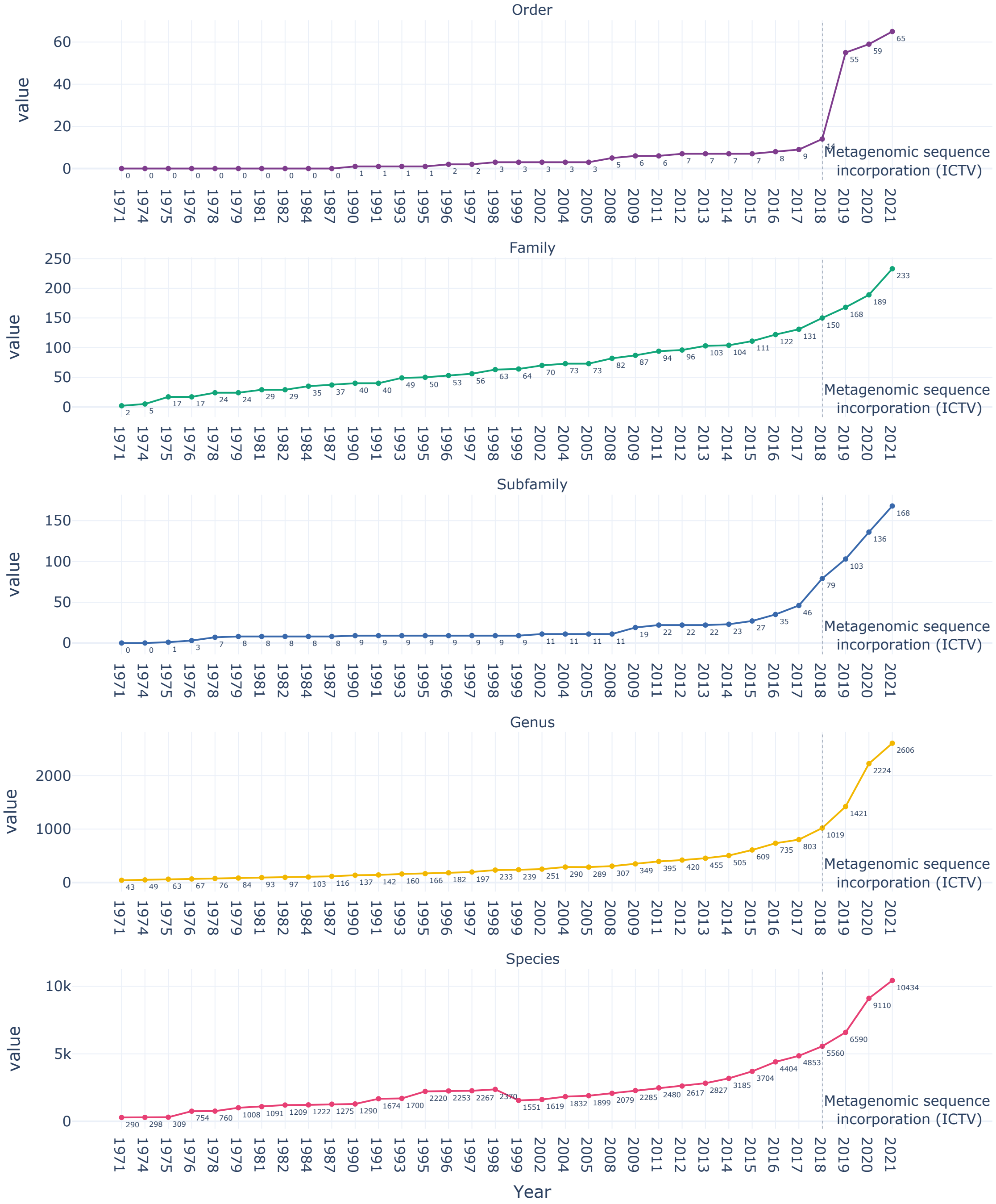


Figure 2.1: An overview of ICTV approved viruses in the context of taxonomic classification and its growth in the last 50 years. The graph shows the number of taxa approved by ICTV for each taxonomic rank. The Y-axis shows the count and the X-axis shows the year. Plots are separated by taxonomic ranks order, family, subfamily, genera and species.

A shift was observed in the field of virology with the advancements in DNA sequencing technology and its application in metagenomics (Described in detail in section 2.2). Metagenomics enabled a massively parallel sequencing and exploration of microbial communities targeting the uncultivated viral diversity that simply could not be accessed through standard cultivation-based approaches. A detailed outlook of how metagenomics led to the discovery of previously inaccessible virome (total collection of viruses found in an environment, microbiome or sample) and viral communities are described in detail in section 2.2.4. ICTV realised and responded to the scientific community's need to classify the diversity of viruses that were identified by the means of metagenomics. In 2017, ICTV officially announced the incorporation of metagenomically assembled virus genomes into its framework (Simmonds et al., 2017a) which led to a massive increase in the number of species added to the taxonomic framework (figure 2.1). Several large-scale metagenomics and metatranscriptomic studies have also influenced virus taxonomy, adding a number of new orders, families, and genera in recent years (https://talk.ictvonline.org/taxonomy/p/taxonomy_releases). As of 2021, there are 10,434 ICTV recognised virus species divided into 2,606 genera that are grouped into 233 virus families. Although a number of different taxa ranks have been added, it is worth noting that a range of virus families are yet to be associated with higher taxonomy ranks such as order and/or realms. All RNA viruses on the other hand have been classified at the top-level rank realm called *Riboviria* (King et al., 2021). Despite this progress, it has been noted that the majority of uncultivated viruses remain yet to be included in the taxonomy framework. Due to the sheer number of viruses that have been identified using uncultivated approaches, virus taxonomy will probably remain in flux for some time, requiring the taxonomy framework to be dynamic and adaptable in response to the large diversity of novel viruses discovered on a daily basis. One suggestion from the Bioinformatics Expert Group (BEG) that is a part of ICTV, is to explore computational and automated approaches to cope with the high demand for the classification of metagenomically discovered uncultivated viruses (Bas E Dutilh et al., 2021). For example, an extensive uncultivated virus database, IMG/VR, contains 868,178 viral operational taxonomy units (vOTUs) as of 29 May 2022 (https://img.jgi.doe.gov/cgi-bin/vr/main.cgi) and these sequences are currently not included in the ICTV taxonomy framework. Indeed, computational virology-based approaches could also be employed to fill in the knowledge gap and systematically place these uncultivated virus sequences into the existing virus taxonomy. Major challenges associated with metagenomically derived virus genome classification such as virus-host association, quality and completeness assessment of uncultivated virus genomes could be tackled with the amalgamation of existing knowledge of virology and computer science to rapidly place uncultivated viruses into the existing taxonomy framework. This advancement would be critical and can potentially transform virus taxonomy and its broader application in understanding virus diversity.

2.2 High throughput sequencing and metagenomics

Microbes are ubiquitous and are an integral part of our living world. They are important to study in health settings because they have been linked to a plethora of diseases and health. It is essential to understand the microbial world in order to obtain a complete picture of the living world. It is suggested that humans should be seen as 'holobionts' due to the co-dependence and intricate relationships between us and microbes (Bordenstein et al., 2015; Guchte et al., 2018). Traditionally, microbes were identified using cell culture techniques. However, there are many challenges to isolating, growing and identifying the microbial communities living around us solely based on these techniques, as it is not always possible to grow them in a laboratory environment.

The development of high throughput sequencing (HTS) technologies has made it possible to survey the microbial communities that surround us. HTS technologies such as 16S ribosomal RNA (rRNA) metabarcoding (metataxonomics), metagenomics and metatranscriptomics have increased the rate at which novel microbes can be discovered. Metabarcoding uses a single marker gene-based identification method such as 16S and/or 18S rRNA for classification and is often misleadingly referred to as metagenomics. 16S rRNA based classification methods are only capturing one genomic marker of a sample and therefore cannot detect organisms that do not possess these specific target genes. A large variety of bacterial species has been discovered using this technique, specifically those that are difficult to cultivate in standard laboratory conditions.

Metagenomics is defined more systematically as an area of research comprising a range of methods that targets the entire microbial sequences at an aggregate level with unbiased sequencing approaches. This term was first used by (Schloss et al., 2003) and, as they predicted in their paper, metagenomics has become a critical tool to enhance research in the field of environmental and microbial genomics. The authors of the book 'The New Science of Metagenomics: Revealing the secrets of our microbial planet' (*The New Science of Metagenomics* 2007), go as far as defining metagenomics as the science of discovery, modelling, understanding and ultimately managing the molecular level dynamic relationships between the molecules that define the living communities and their biosphere; this definition derives mainly from Hood's definition of systems biology (Hood, 2003). It is clear that metagenomics is an area of scientific research that provides powerful tools to study the interactions within and between different communities. Metatranscriptomics is a similar branch of research that focuses on the expressed or active part of the communities instead of their genomic profiling. The transcribed mRNA content of a given sample is sequenced through HTS technologies. When applied together, metagenomics and metatranscriptomics present immense opportunities to gain insights into the microbial communities that are present as well as active around us.

Conventional cultivation approaches cannot capture the complete view of a given community as it is almost impossible to mimic the ecological or health environment on a petri dish, thus limiting the study and characterisation of uncultivated microbes. Moreover, these cultivation methods cannot be used to quantify the proportion of microbial abundance in a biological niche due to the biases associated with the culture-based approaches. On contrary, metagenomic and metatranscriptomic approaches can help to overcome this cultivation-dependent issue faced by the classical microbial study techniques, and they provide a relatively unbiased view of the microbial community directly sampled and sequenced from its natural environment.

With the advent of HTS, the field of bioinformatics/computational biology has also been advancing at a fast pace to match the demand for analysing the large amount of sequence data that originated from a range of sequencing technologies. Though the field of bioinformatics predates HTS technologies, widespread applications of HTS in all branches of biology have accelerated the growth and the maturation of bioinformatics as an area of research that combines the knowledge of biology and uses cutting-edge computational approaches to address complex biological questions. A rapid expansion of biological sequence data analysis algorithms, techniques and computational resources has led to the development of a plethora of tools, databases and workflows. These computational resources are needing to be updated, customised and maintained continually to keep up with the demands of this area of research.

2.2.1 Metagenomic data analysis

Metagenomics, metabarcoding and/or metatranscriptomics have a wide range of applications. They can be applied to study the microbial makeup of a given environment such as soil, aquatic or other biogeographical and ecological locations. These methods can also be employed to study the host-associated microbiomes in both natural and clinical settings. Though all three techniques serve the purpose of investigating the microbial composition of relevant samples broadly, their specific applications are distinct. Metataxonomics or metabarcoding involves sequencing of specific marker genes such as 16S/18S rRNA of prokaryotes or internal transcribed spacer regions of fungal ribosomes (Breitwieser et al., 2018). As this technique takes the advantage of marker gene-based identifications and does not capture the complete content of a given environment, they are not considered shotgun metagenomics (Quince et al., 2017). Despite this, metabarcoding is the most cost-effective way to profile targeted microbial communities and can be used to study bacteria and microbial eukaryotes. The major disadvantage of metataxonomics is that it cannot be used to study viruses as they lack marker genes and it provides a limited resolution of microbial genomes as they only capture the diversity within specific genes. Moreover, metabarcoding sequencing involves a target gene amplification step that can lead to biases if the research project aims to capture the quantitative aspects of the microbes present in the samples. To overcome this, a read-based normalisation step is required to draw abundance-based correlations.

In contrast to metabarcoding, both metagenomics and metatranscriptomics methods capture the total nucleic acid content of a given biological sample. Shotgun metagenomics which involves the sequencing of random untargeted DNA or RNA present in the sample is regarded as more powerful due to its robustness and ability to capture all forms of life present in the

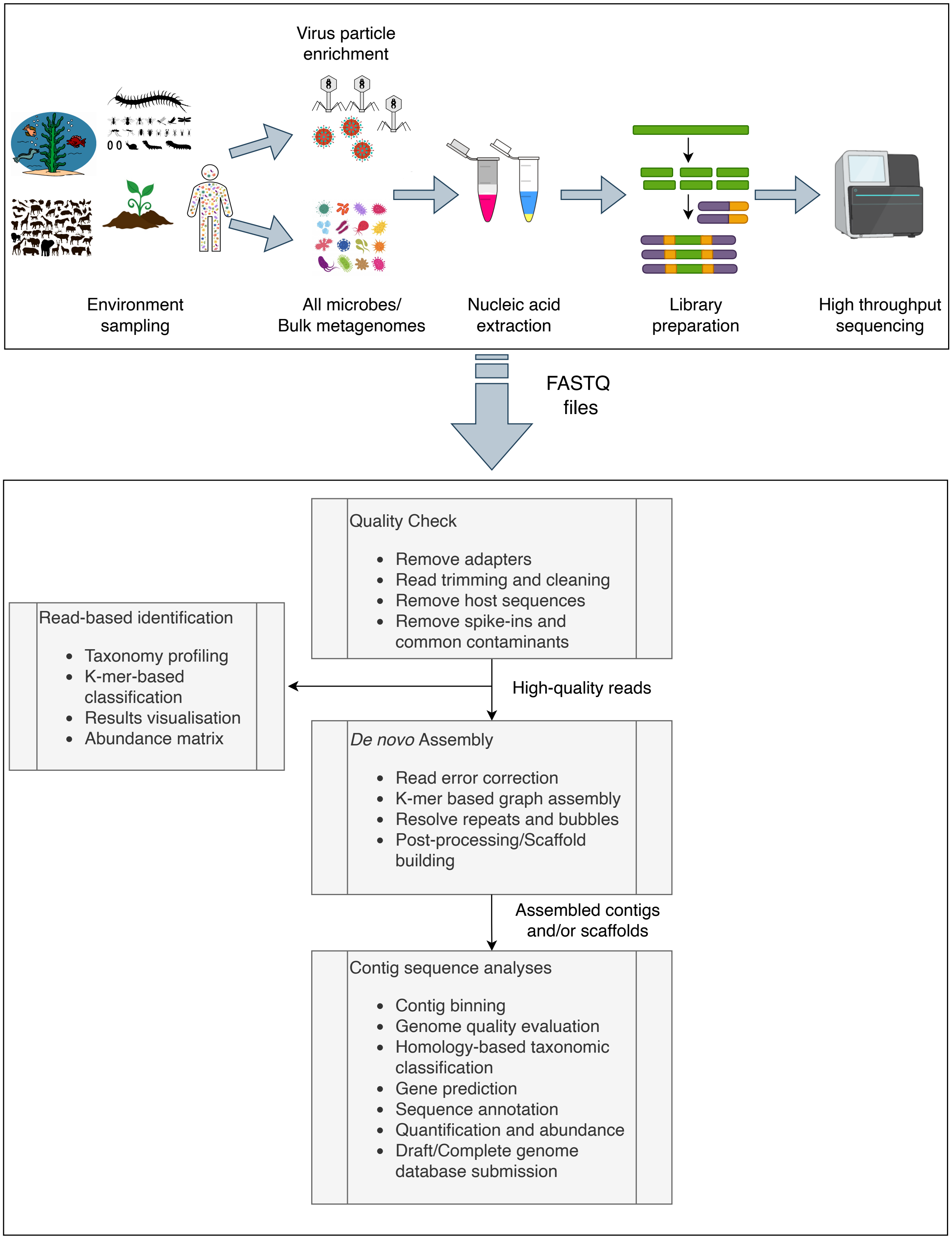

Virus particle enrichment
Environment sampling
All microbes/ Bulk metagenomes
Nucleic acid extraction
Library preparation
High throughput sequencing
FASTQ files
Quality Check
• Remove adapters
• Read trimming and cleaning
• Remove host sequences
• Remove spike-ins and common contaminants
High-quality reads
Read-based identification
• Taxonomy profiling
• K-mer-based classification
• Results visualisation
• Abundance matrix
De novo Assembly
• Read error correction
• K-mer based graph assembly
• Resolve repeats and bubbles
• Post-processing/Scaffold building
Assembled contigs and/or scaffolds
Contig sequence analyses
• Contig binning
• Genome quality evaluation
• Homology-based taxonomic classification
• Gene prediction
• Sequence annotation
• Quantification and abundance
• Draft/Complete genome database submission

Figure 2.2: An overview of metagenomic sample preparations, sequencing and data analysis. A typical HTS metagenomic analysis is outlined in the following stages. Briefly, it is categorised into Quality Check, *De novo* assembly and/or Read-based identification and Contig sequence analysis.

captured nucleic acid. A brief and abstract overview of the metagenomic and metatranscriptomic sequencing and analysis workflow is described in the figure 2.2. A typical pipeline includes DNA or RNA extraction from the target biogeographical environment or clinical samples. The nucleic acid is chopped into short segments that are ligated with adapters that enable sequencing library generation. These libraries are then sequenced using short or long-read sequencing platforms such as Illumina, Oxford Nanopore or Pacbio. As the field of metagenomics has matured, a wide range of laboratory kits and reagents are made available that help with the sequencing library preparation specific to various environments. Metagenomics is the sequencing of the DNA or RNA whereas metatranscriptomics is the sequencing of mRNA (messenger RNA). Shotgun metagenomics can be used to profile the genomic content originating from all domains of life including bacteria, viruses, archaea and eukaryotes. It also enables the *de novo* assembly of genomes present in the samples and can enable functional genome analyses. Moreover, unlike metabarcoding, this unbiased approach provides a more complete picture of what is present in the sample and can be utilised to identify completely novel organisms and pathogens. Metagenomics enables researchers to address 'What is present' in the sample and metatranscriptomics enables the explorations of 'What is active' in a given sample as it targets the transcribed part of the genomes. Both techniques require a high-depth sequencing approach to capture the microbial community present at sufficient resolution. An alternative approach of metagenomics or metatranscriptomics that specifically focuses on the study of the virome content of a given sample involves an additional step in the sequencing process. This virus particle enrichment process is a filtering step that is applied to filter out particles of specific size and can help to capture small virus particles (figure 2.2). This filtering process is not unique to viruses (e.g. it is used to separate bacteria from eukaryotes), but it is helpful in separating small virus particles from other microbial material. Further, depending on the aims of the research project, the nucleic acid samples are treated with DNAase or RNAase to degrade the nucleic acid sequences that are not of interest. These virome-specific meta-omic methods are termed metaviromics and are used to study the virome content of a sample.

Metagenomic sequencing data analysis generally requires an extensive pipeline that is tailored to address the specific research aims. A general overview of the metagenomic data analysis pipeline is shown in figure 2.2. Although a wide range of sequencing platforms have been used and are being tested for metagenomic sequencing, the Illumina sequencing platform is the *de facto* default technology used in most metagenomic sequencing (Breitwieser et al., 2018). The sequencing reads produced from Illumina are around 150-300bp long. The first step to analysing sequence data is to assess the quality of these short sequencing reads. This quality assessment step is also important as it includes read trimming, sequencing adapter removal and low-quality read filtering. A number of tools such as Cutadapt (Martin, 2011), TrimGalore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), Trimmomatic (Bolger et al., 2014), PRINSEQ (Schmieder et al., 2011), BBDuk (Bushnell B., 2015a), FastQC (https://www.

bioinformatics.babraham.ac.uk/projects/fastqc/), and MultiQC (Ewels et al., 2016) are available to assess the read quality and clean the reads to retain only high-quality sequencing reads. The QC step also includes the removal of common lab contaminants and/or known spike-ins which can have an impact on the results. QC tools such as BBDuk, Trimmomatic, and TrimGalore have a range of built-in settings and known sequence sets that can be readily removed if found from the input read set. In the case of the host-associated microbiome, the QC step involves mapping a cleaned set of reads to the associated host and extracting reads that are unmapped. This step reduces the number of reads down as well as helping to speed up the *de novo* assembly process. General-purpose read alignment programs such as Burrows-Wheel aligner(BWA) (Heng Li et al., 2009), bowtie2 (Langmead et al., 2012) or BBMap (Bushnell B., 2019) can be used to map short reads to the known host genomes. In the case of long reads, corresponding long read mapping programs such as minimap2 (Heng Li, 2018) can be used. This step generates output in sequence alignment map (SAM) or binary alignment map (BAM) formats. Both formats consist of the exact same data but the BAM file format is machine-readable binary data that cannot be read by humans but the SAM file contains the mapping information in the standard human-readable format. To obtain the unmapped reads from these files, samtools (H. Li et al., 2009) or BBTools (Bushnell B., 2019) packages can be used.

It is known that complex microbial communities are non-uniformly distributed and the proportion of each species' genomic content captured in a typical metagenomic sample varies widely. In this scenario, it is likely that the process of *de novo* assembly could potentially lead to the assembly of the species that are most abundantly present in the sample. To overcome this, an optional step of read normalisation can be utilised. This step reduces the redundancy of the reads by removing the duplicate reads and by down-sampling the reads it can achieve a reasonably uniform sampling coverage that in turn helps with the *de novo* assembly process. Reducing the number of duplicate reads also helps to accelerate the assembly process and can help to reduce the computational resources required for the assembly part of the pipeline. Tools such as BBNorm (Bushnell B., 2015b), and Diginorm (from the Khmer package) (Crusoe et al., 2015) can be used to normalise the reads prior to the assembly.

The *de novo* assembly process is arguably the most computationally intensive part of the metagenomic data analyses workflow. Genome assembly in itself is regarded as one of the most challenging steps of HTS data analyses, the sequence assembly of metagenomic samples is more challenging due to the presence of many different organisms/species in varied abundance. There are two ways in which the short reads can be assembled. A single sample assembly entails assembly of the reads from a metagenomic sample and a co-assembly which treats all samples as a uniform big sample and combines all reads from a range of samples that are typically originating from the same experiments and/or microbiome or ecological sites. The co-assembly of samples is more likely to lead to fragmented assembly and generally requires a substantially extensive computational framework to assemble short reads into contigs. Moreover, the co-assembly step is more likely to yield fragmented assemblies due to the complexity of microbial communities and other factors including the uneven sample coverage. On the contrary, single sample assembly is a relatively straightforward process and can be carried out on smaller computational resources compared to co-assembly. *De novo* assembly obtained from a single sample is shown to be less likely to be fragmented and can result in higher quality genomes (Olm et al., 2017).

A number of *de novo* assembly tools that implement various algorithms can be used to achieve a good quality assembly. The two most established approaches are Overlap Layout Consensus (OLC) and *De Bruijin* Graph (DBG) which have been implemented in a wide range of tools. Briefly, OLC approaches work by finding overlap among sequence reads. A contiguous sequence is obtained by stitching overlapping reads together into longer sequences. The amount of overlap can be varied from a short k-mer to the length of the read. However, the OLC approaches struggle to find the best continuous sequences as most of them only used the far ends of the reads to search for the overlaps. The DBG approaches work by chopping the short reads into shorter k-mers generated using a sliding window across the length of the read, hence utilising all information and bases captured in the reads. These k-mers are organised in a graph layout such that each node represents a k-mer and the edges represent k-1 overlap among them. These complex large graph paths are joined together to generate longer sequences that are termed contigs (Wenyu Zhang et al., 2011). General-purpose microbial assembly tools such as Abyss (Simpson et al., 2009), SOAPdenovo (Luo et al., 2012), MIRA (Chevreux et al., 2004), and Velvet (Zerbino et al., 2008) have been used to assemble microbial genomes from metagenomic samples. However, these tools often struggle with the high complexity and uneven coverage that is unique to the metagenomic datasets and often lead to short contigs with low N50 (a measure to assess assembly quality). The quality of assembly is highly dependent on the size of choice of k and it can be very tricky to determine the best k-mer size that would help to recover the largest genome fragments from a sample. To overcome this, iterative DBG approaches have been implemented in the *de novo* assembly tools that are specifically designed for metagenomic datasets. These tools including IDBA-UD (Peng et al., 2012), metaSPAdes (Nurk et al., 2017), and Megahit (D. Li et al., 2015) have been more successful in retrieving complex microbial genomic structures embedded within the metagenomic datasets as they carry out the assembly using a range of k-mers and use the contigs generated at each step of the assembly for the next iteration. Most of these sophisticated methods of assembly also offer a built-in assembly pipeline that also takes care of pre-assembly read error correction, low complexity contig filtering, contig coverage and depth calculation and scaffold building steps included in the pipeline.

In a standard metagenomic sequencing project, millions of reads are condensed down to tens/hundreds of thousands of contigs that need to be taxonomically labelled to identify their biological origins. To assess the quality of the metagenomic assembly such as to measure the proportion of chimeric contigs, identify misassemblies as well as compare the contigs to known sequence databases, MetaQUAST (Mikheenko et al., 2016) can be used. Moreover, a general-purpose quality check such as aligning reads back to a contig and/or bin can also be carried out to gather statistics related to the quality of the contigs. To perform this, standard assembly mapping tools such as BWA, bowtie2 or BBMap can be used and the resulting SAM/BAM files can be analysed using samtools to gather the assembly statistics. As short-read assemblies can often be fragmented due to the uneven coverage and the presence of multiple strains, metagenomic binning can be utilised to cluster these highly similar sequences that potentially originate from the same species into metagenomic bins. The representative sequence obtained through binning can be used to denote a single species or more generally an operational taxonomic unit (OTU). Metagenomic binning tools cluster sequences based on one or more features. These features could be obtained from previously known reference databases e.g. phylogenetic marker gene presence (implemented in MyCC (Lin et al., 2016)) or are calculated dynamically from the dataset in question. These features include read coverage and linkage, differential coverage, tetranucleotide frequencies, and/or multi-sample coverage. Genome binning tools widely used for this analysis include MetaBAT (D. Kang et al., 2019), MyCC (Lin et al., 2016), CONCOCT (Alneberg et al., 2014), GroopM (Imelfort et al., 2014) and MetaWatt (Strous et al., 2012). However, as with any other analyses, genome binning can often lead to different results based on the distinct methods applied to bin the contig dataset and the parameters used to perform the binning. To overcome this, DAS Tool (Sieber et al., 2018) was developed as it can compile the binning results, remove redundancies from them and consolidate the bins into better assemblies reflecting more complete genomes. Moreover, to check the quality of binning, tools such as CheckM (Parks et al., 2015) can be used for prokaryotic datasets and CheckV (Nayfach et al., 2020a) can be used for viruses and viral OTUs.

Once good quality contigs or contig bins are generated, the next analysis requires the taxonomic identification of the sequences to be carried out. Notably, it is also possible to carry out read-based taxonomic profiling prior to assembling contigs. Typically the read-based taxonomy profiling tools utilise a short exact k-mer matching approach to taxonomically label the reads, this is due to the sheer amount of reads that need to be processed and standard approaches such as Basic Local Alignment Search tool (BLAST) are deemed too slow for carrying out similarity searches on millions of reads as they were designed to work on longer sequences. Though a number of k-mer-based taxonomy assignment tools such as Kraken (Wood et al., 2014), Kaiju (Menzel et al., 2016), CLARK (Ounit et al., 2015), Centrifuge (Kim et al., 2016), DisCVR (Maabar et al., 2019) exist now, Kraken was one of the first to implement this approach. In Kraken, short k-mers of k=31 were used and implemented as part of the exact string matching algorithm that created k-mers of length 31 from reads and compared them against a reference database of k=31 from known genome sequences. Typically a lowest common ancestor (LCA) containing the specific k-mers is inferred from the taxonomic databases and sequences are "classified" at a specific taxonomy level using the LCA information. A complete sequence-based taxonomic profiling output can be generated by these k-mer profiling tools and the results can be visualised using the hierarchical taxonomic profile visualisation tools such as KronaTools (Brian D Ondov et al., 2011b; Brian D. Ondov et al., 2011a). It is worth noting that k-mer-based profiling can be efficient in some cases but has disadvantages as long k-mers are likely to be too stringent and would not be effective candidates for exact matches due to issues such as sequencing errors, and short k-mers can lead to false positives. Striking a balance between the best precision and recall is a very challenging task and the choice of k size would often be highly dependent on the research question, the size of the database and the type of data being investigated. Similar to k-mer-based classification, another alternative method for metagenomic profiling is using the overlapping MinHash signatures. Two popular tools Mash (Brian D. Ondov et al., 2016) and sourmash (Pierce et al., 2019) have implemented this approach that allows users to run quick similarity estimates of datasets on a laptop as they require much smaller datasets and less computation power. Though most of the k-mer and hash-based taxonomic profiling tools were designed for reads, they can be used with contigs. Other tools such as taxator-tk (Dröge et al., 2015) enable users to perform binning, taxonomic assignment and microbial community profiling using a single package and work well with prokaryotic metagenomic datasets.

In general, a standard homology-based sequence identification is carried out on contigs to infer their evolutionary relatedness to the existing sequences. Conventional sequence similarity tools such as BLAST+ suite are considered the gold standard due to their widespread usage and applications. Nucleotide similarity searches can be carried out against extensive known sequence databases such as nt or against a specific set of marker genes using tools like MetaPhlAn (Segata et al., 2012; Truong et al., 2015). If International Nucleotide Sequence Database Collaboration (INSDC) databases are used, they are often too big to host on a small personal computer and are hosted either on a database server or searched using National Center for Biotechnology Information (NCBI)'s remote BLAST options. However, nucleotide searches against extensive databases such as nt can be computationally expensive and may not be feasible. In such cases, other taxonomic profiling and classification tools such as Kraken + Bracken (for classification) (J. Lu et al., 2017), CLARK, mOTU (Sunagawa et al., 2013), MetaPhlAn2 (Truong et al., 2015), Kallisto (pseudo alignment algorithm) (Bray et al., 2016), LAST or Centrifuge can be used. However, these short sequence-based taxonomy profiling tools can often be less sensitive and may not be effective in some specific use cases. Nucleotide sequences vary largely even between highly similar species or taxonomic groups, but protein sequences are more conserved due to evolutionary convergence. To exploit this, and enable the identification of novel and/or phylogenetically distantly related sequences, protein alignments can be utilised. BLASTX can be used to translate a nucleotide query into the amino acid sequence in all six frames and then search these sequences in a protein sequence database. This is a very powerful approach and can be particularly effective for the classification of virus genome sequences. Although BLASTX works very well, it is very slow at processing a large number of contigs and alternatives such as DIAMOND (Buchfink et al., 2014; Buchfink et al., 2021), and MMSeqs2 (Steinegger et al., 2017) have replaced the traditional BLASTP and BLASTX tools as they are able to achieve BLAST equivalent precision and recall performance in a fraction of time.

Post metagenomic analyses involve a number of different post-processing steps before the genome sequences are ready to be submitted to the relevant database. One of the most popular and advisable steps is to carry out gene prediction on the genome to identify the open reading frames (ORF). Several standard ORF prediction tools e.g. getorf (Rice et al., 2000) or translate can be used to extract ORF information from the contig sequences. Other genes and ORF prediction tools such as prodigal that has a specialised mode for the metagenomic dataset which uses previously trained models for gene prediction are also widely used (Hyatt et al., 2010). These predicted protein sequences can also be used to further carry out translated sequence analysis and/or phylogenetic analysis where applicable. In cases where no significant sequence similarity is observed, predicted ORFs can be used to identify the presence of protein domain signatures. This can be achieved by using protein profile analyses and/or domain analysis tools such as HMMER (requires a protein profile databases e.g. Pfam) (Finn et al., 2011) or multi-purpose protein homology analysis tools such as HHPred (Söding et al., 2005) or InterProScan (Mitchell et al., 2018a). ORF-based analyses are largely suitable for the discovery of new proteins in microbes and work efficiently for novel virus sequences that may be significantly diverse from known virus sequences available in the databases. This analysis cannot identify and annotate long non-coding RNAs (lncRNAs) that play regulatory roles, and to identify such lncRNAs from metagenomic datasets, specialised pipelines and tools such as DRAGoM (Liu et al., 2021) could be used.

All metadata and features obtained through various stages of the metagenomic analysis could be gathered to formulate a complete picture of the novel genome. If a close relative of the newly identified sequence exists in the database, the newly discovered microbial genome can be annotated with relevant features. The feature annotation step may include but is not limited to gene start and end positions, corresponding translated ORF sequences, domains that were found to be associated with the ORFs, untranslated regions, terminal repeat sequences and any other features that may be unique and relevant to the organism being interrogated. A number of tools and pipelines such as Prokka (Seemann, 2014), Distilled and Refined Annotation of Metabolism (DRAM) (Shaffer et al., 2020), Rapid Annotation Transfer Tool (RATT) (Otto et al., 2011), DDBJ Fast Annotation And Search Tool (DFAST) (Tanizawa et al., 2018), Viral Annotation Pipeline and iDentification (VAPiD) (Shean et al., 2019) are developed that can automate this process and can be used to either create annotations or transfer annotations from a currently known genome sequence. The final output from this pipeline is prepared in an INSDC compatible data format e.g. GenBank or EMBL that can be directly submitted to the databases to share the newly identified organism draft/complete genome with the research community.

It has to be appreciated that metagenomic sequence analysis is an involved and complex process that requires a lot of attention to detail and customisation to address specific biological questions. Despite the complexity, there have been efforts by the researchers to encapsulate this process into a meaningful, modular and adaptable workflow that could be applied readily to answer specific questions. Moreover, there have also been efforts by the research community to benchmark the tools and resources available for a specific analysis step to devise standardised ways to carry out the analyses. Community-led efforts such as the Critical Assessment of Metagenome Interpretation (CAMI) challenge have also helped researchers in making the right choice of tools to use for metagenomic analyses by performing extensive benchmarks of an ever-increasing catalogue of assembly, binning, classification and annotation tools against comprehensive and diverse real and simulated datasets. Furthermore, there have also been efforts to develop modular metagenomic workflows that incorporate several tools for each step of the analysis and can be tailored to be used for specific datasets. MetAMOS (Treangen et al., 2013), Anvi'o (Eren et al., 2015), MetaWRAP (Uritskiy et al., 2018), MGnify (Mitchell et al., 2018b) and many more such pipelines are available that can be used for any metagenomic datasets. Moreover, a range of experiment-specific pipelines such as MetaViC (Modha et al., 2019), VIP (Y. Li et al., 2016), Taxonomer (Flygare et al., 2016) etc is available to choose from. Most of the pipelines and software mentioned here are either open source or available for free under academic licenses. Other commercial software tools such as CLC Workbench, One Codex, and CosmosID are also used widely in metagenomic sequence analysis. Overall, a range of different bioinformatic tools and pipelines are available to perform these analyses (Breitwieser et al., 2018). A recent review focusing on the viral metagenomics methods and data analysis pipeline discussed 49 currently available tools and provided a decision tree for the different pipelines dependent on application such as clinical diagnosis for viral discovery (Nooij et al., 2018).

It is evident that a large number of metagenomic sequence analysis steps rely heavily on reference databases for microbial species characterisation and annotation. In order to classify as many contigs as possible, it is essential to use general-purpose large-scale databases. There is a notable trade-off between using the general-purpose databases that contain all types of sequences as they are often not as well-curated as other more specialised databases that encompass organism-specific sequences. This leads to the sensitivity versus specificity challenges that need to be considered with respect to the aims of the metagenomic projects. Moreover, as more and more environments are readily sequenced using metagenomic and/or metatranscriptomic approaches, the databases that harbour these sequences are also expanding exponentially in size. This poses a computational and resource-oriented issue as local searches, i.e. databases hosted locally on a server within the same network, are much quicker to perform than those required to be done via the internet.

2.2.2 Application of metagenomics in microbe discovery

In the last two decades, HTS metagenomic methods have been applied to a range of environments and bodily sites and have led to the identification of numerous novel microorganisms and their interactions. For instance, a study by Brown et al. (2015) applied HTS and cultivation-independent approaches to identify novel bacterial phyla referred to as candidate phyla radiation (CPR), defined as new phyla with no isolated representative. A more recent study has demonstrated how the diversity of the tree of life has been expanded due to these CPRs as they capture around 25% of bacterial diversity and their interactions with archaea (Castelle et al., 2018). These CPRs have been shown to be genetically distinct and often lack genes or metabolic pathways that were identified to be universal for bacterial species. Their unique genomic organisation and other features including self-splicing introns, split genes and downsizing genomes to gain a competitive edge to their parasitic hosts suggest that these organisms tend to be highly dependent on their host microbes, and cannot be grown in laboratory culture and could only be identified with metagenomic techniques (Koonin, 2018). Other studies that strengthen this host dependency have studied novel archaeal DPANN (Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, Nanohaloarchaea) superphylum and the Asgard phylum, and their strikingly parallel evolution with CPR bacteria. The Asgard archaea have also been shown to possess many genes that are considered to be eukaryotic signature genes and have been phylogenetically shown to be a sister clade of the eukaryotes (Castelle et al., 2015; Spang et al., 2015; Hug et al., 2016; Spang et al., 2017; Zaremba-Niedzwiedzka et al., 2017; Castelle et al., 2018; Koonin, 2018). Metagenomics has also led to the identification of Nitrospira bacterial species that encode all the enzymes required to carry out nitrification of ammonia (Daims et al., 2015). A study by Kessel et al. (2015) also mined public sequence databases (such as NR; non-redundant protein database) for the signature sequences of amoA genes and identified amoA sequences that were misclassified in these databases as methane monooxygenases. This finding highlights two important points: firstly, unbiased metagenomic sequencing led to the discovery of the first-ever bacterial species that was shown to be capable of carrying out the complete cycle of nitrification of ammonia and secondly, the data mining exercise undertaken by this project helped to correctly classify the existing sequences in the databases that were labelled "unusual" methane monooxygenase as amoA gene signatures.

2.2.3 Uncultivated microbial diversity in the human microbiome

Metagenomics has been applied to samples from a diverse range of environments such as soil and oceans, as well as samples from a variety of hosts including humans. The Human Microbiome Project (HMP) consortium proposed and led by the NIH in 2007 has taken great advantage of metagenomics to study the microorganisms that live both inside and on humans. These microorganisms are known as human microbiota and they are estimated to be present in humans with a 1 to 1 ratio of human cells, making up to 1-3 percent of human body mass (*NIH Human Microbiome Project - About the Human Microbiome* 2020). The HMP has generated over 7 Tb of sequence data that has been submitted to the Sequence Read Archives (SRA). The HMP was designed to identify the “core” microbiome of humans. It has served as a resource that has led to the identification of 2200 microbial reference genomes (*NIH Human Microbiome Project - About the Human Microbiome* 2020), 700 metagenomes and 60 million predicted genes from healthy adult microbiomes. Results derived from the HMP indicate that the microbiomes between individuals can differ significantly (Consortium et al., 2012; Gevers et al., 2012). The second phase of the HMP was established in 2014 and is titled the 'The Integrative Human Microbiome Project' (iHMP). The iHMP's aim was to understand the impact of the microbiome on human health and disease. This involves studying microbes and their interactions with their hosts in disease-specific cohorts by employing a multi-omics data collection and integration approach.

Another similar study that focused solely on the Metagenomics of the Human Intestinal Tract (MetaHIT) has identified between 1,000-1,150 prevalent bacterial species from 124 individuals (Qin et al., 2010). A recent study by Pasolli et al. (2019), recovered 154,723 microbial genomes by mining nearly 10,000 human metagenomes. Two other gut microbiome dataset mining analyses led by Almeida et al. (2019) and Nayfach et al. (2019) also expanded the uncultivated bacterial diversity with the identification of thousands of novel Metagenome Assembled Genomes (MAGs). Subsequently, Almeida et al. (2020) consolidated the MAGs from the above three gut microbiome surveys and created a unified catalogue of 204,938 reference genomes from the human gut microbiome. Although the gut microbiome has been a major focus for a number of years due to its crucial role in human metabolic health and diseases, other human microbiomes e.g. skin (Kashaf et al., 2022), oral (Dewhirst et al., 2010), and blood (Whittle et al., 2019) have also been explored to characterise their microbial makeup. Overall, these studies have shown that the microbes around us have a large effect on an individual's health status. Microbiome studies of various conditions and environments such as disease versus healthy individuals, and of various human body sites such as the skin, gut, blood, oral and respiratory microbiome, have demonstrated that human microbiota varies among individuals and within body sites. These variations play a role in the unique interactions each individual has with other microbial organisms in a given environmental context. For example, Donia et al. (2014) identified an antibiotic bacterial gene that was found to be widespread in the gut bacteria of the HMP cohort metagenomic samples, whilst Norman et al. (2015) and Gevers et al. (2014) identified the important role played by changes in viral and bacterial microbiomes in Crohn's disease. The human microbiome has been considered our second genome and with a realisation, that imbalance or dysbiosis of human microbial communities can have implications in health status. The human microbiome is a network of complex interactions between microbes and humans are merely a host that facilitates these interactions. In turn, microbial communities that are found to be associated with humans have been shown to be playing an important role in shaping and playing a critical role in human health (Malla et al., 2019). Problematically, most 'microbiome' studies have focused on the bacterial component of the microbiome, ignoring bacteriophage and RNA viruses, and therefore the study of the human virome has been identified as a key priority (Zou et al., 2016).

2.2.4 Viral metagenomics

Traditionally, viruses were discovered using cell culture methods and most viruses known to date were identified using this technique. However, viruses can be difficult to grow in cell culture, which hinders the process of virus identification, replication and further functional studies. A lot of viruses need specific hosts in order to replicate. Most viruses often require particular host cells and suitable functional environments that can be very challenging to imitate in a standard laboratory setting. In recent years, metagenomics and metatranscriptomics have been the most effective techniques to identify and characterise viruses in samples. When these techniques are applied to viruses, it is termed “metaviromics”, and has led to the characterisation of the global virome or “virosphere” (Mizuno et al., 2013; Zablocki et al., 2014; Koonin et al., 2018). Metagenomic and metaviromic techniques are excellent tools to enumerate the virosphere specifically in the context of viruses that infect uncultivated prokaryotic and eukaryotic microbial species.

The first viruses to be discovered using metaviromics were in the marine environment (Breitbart et al., 2002). Since then a range of studies has found viruses to be abundant in a diverse range of environments. This includes a range of projects such as the Earth’s virome project which discovered over 125,000 partial DNA viral genomes and the largest phage identified (Paez-Espino et al., 2016). In the case of the marine environment, a study by Mizuno et al. (2013) identified over 200 novel marine phage genomes. Additionally, other marine metagenomics projects such as the Tara Ocean Virome and the Pacific Ocean Virome projects have shown these ecosystems to be rich in viral communities. These consortiums have identified novel viruses that bear little or no similarity to previously known viruses (Hurwitz et al., 2013; Jennifer R Brum et al., 2015). A range of studies have shown that humans harbour diverse viruses that can play a role in the healthy or diseased status of an individual, this could be either due to direct interactions between viruses and humans or through the virus interactions with other microbiota. It has also been shown that the viruses that regulate bacteria in humans have an impact on antiviral immune response and viral infectivity (Honda et al., 2012; Duerkop et al., 2013; Lecuit et al., 2013; Popgeorgiev et al., 2013; Rascovan et al., 2016; Zárate et al., 2017). Moreover, HTS has also been applied to identify novel viral pathogens (Briese et al., 2009; Zaki et al., 2012; F. Wu et al., 2020; Jerome et al., 2019) and has been instrumental in virus outbreak tracking (T. Li et al., 2019; Gardy et al., 2017; Luk et al., 2015), and metaviromic projects have also contributed to advancing other related fields such as human and animal pathogen identification (Hoffmann et al., 2012) and clinical diagnostics (Nakamura et al., 2009; Stremlau et al., 2015; Moustafa et al., 2017; Thorburn et al., 2015). Recently a comprehensive consolidated catalogue of the human virome has been created by Liang et al. (2021) that summarised the family-level overview of eukaryotic and prokaryotic viruses found in various human body sites.

Initially, metagenomic advances predominantly led to the discovery of DNA viruses that infect prokaryotes. This was mainly due to the fact that a lot of studies sequenced the DNA captured and isolated from the given environments and bodily sites. The first human gut virome study by Breitbart et al. (2003) showed that a large majority of virus sequences identified from the human gut were not matching to anything and that gut virome contained around 1200 viral genotypes. Since then, many large studies have identified hundreds of thousands of DNA viruses associated with human bodily sites using metagenomics (Tisza et al., 2020; Nayfach et al., 2019; Paez-Espino et al., 2016; Benler et al., 2021). A large proportion of uncultured virus databases such as IMG/VR, Gut Phage Database (GPD), and Gut Virome Database(GVD) are made up of DNA viruses that were catalogued through the means of metagenomics (Roux et al., 2021c; Luis F. Camarillo-Guerrero et al., 2021; Gregory et al., 2020). Although a large of a proportion of prokaryotic viruses have been catalogued, due to the unbiased nature of metagenomics, they cannot be readily linked to their hosts. Virus-host association of uncultivated viruses has been deemed one of the major challenges in metaviromics (Roux et al., 2021b).

In 2005, a study by Breitbart et al. (2005) isolated and sequenced ssDNA viruses using the shotgun metagenomics from blood. Since then, many metaviromics studies have led to the discovery and acknowledgement of small circular virus diversity in human and non-human samples (Abbas et al., 2019; Ng et al., 2015; Tisza et al., 2020). Small circular viruses such as anelloviruses have been found to be highly diverse and ubiquitously present in human blood (Tisza et al., 2021b; Arze et al., 2021). These ssDNA viruses are found in different human microbiomes including urine, fecal and saliva microbiomes (Kaczorowska et al., 2020), and their genomic diversity was shown to be driven by the mechanism of recombination (Worobey, 2000; Arze et al., 2021). The first anellovirus, then called TT virus, was identified from the blood sample of a hepatitis patient (Nishizawa et al., 1997). Since then, anelloviruses have been found in several mammals including primates (chimpanzee, macaque, tamarin and douroucouli), cows, dogs, cats, pigs, rodents and bats as well as seals (Varsani et al., 2021; Souza et al., 2018; Kaczorowska et al., 2020). Though these viruses have been hypothesised to be associated with a range of diseases and conditions in humans, this opinion has been contested with an alternative hypothesis that they have co-evolved with their hosts (Koonin et al., 2021b) and they may be a part of the commensal human virome (Freer et al., 2018). The virus experts have theorised that anelloviruses may have a symbiotic relationship with their hosts due to their omnipresence in samples obtained from both human and other mammalian species (Souza et al., 2018; Kaczorowska et al., 2020). It is also notable that a complete clearance from anellovirus infection is considered impossible and is instead driven by the host immune system with a higher viral load evident in those with compromised immune responses (Freer et al., 2018; Webb et al., 2020; Koonin et al., 2021b). The authors of Arze et al. (2021) coined a new term “anellome" describing the diversity of anellovirus genomes in human microbiome datasets.

All RNA viruses have RNA-directed RNA polymerase (RdRp) - a hallmark gene that can be used to check the presence of RNA viruses in metatranscriptomic samples. Due to the robustness of RdRp-based virus identification of RNA viruses, it has recently become a popular strategy to search for RNA viruses in various environments. Metatranscriptomics has been instrumental in discovering the highly divergent RNA viruses. Due to their high mutation rates and lack of proofreading mechanisms, RNA viruses are the most challenging viruses to isolate, study and understand (Drake et al., 1999; Duffy, 2018). A recent study by Neri et al. (2022) led to a five fold expansion of RNA phages with the discovery of 5,150 publicly available in aquatic, terrestrial, host-associated and engineered metatranscriptomes. A separate study by Zayed et al. (2022) identified thousands of novel RNA viruses from the samples collected from Tara Ocean expeditions. They analysed 28 terabases of Global Ocean RNA sequences and discovered a globally distributed phylum termed *Taraviricota* that may potentially provide the missing link for the evolutionary origins of RNA viruses in connection with retroelements present in both eukaryotes and prokaryotes. RNA virome analysis of aquatic sampling from China's Yangtze River led to the discovery of 4,500 distinct RNA viruses expanding the previously known RNA virus diversity twofold (Wolf et al., 2020). Similarly, metatranscriptomic sequencing of 220 invertebrate species identified over 1400 novel RNA viruses and these viruses were found to be significantly divergent from already known species (M. Shi et al., 2016a). Another study focusing on the virome of the vertebrates including reptiles, amphibians and a number of fish identified 214 novel vertebrate-associated RNA viruses using metatranscriptomic analysis (M. Shi et al., 2018a). Other studies have found a range of novel viral species belonging to existing and proposed viral families in other arthropods such as mosquitoes, honey bees and ticks (Coffey et al., 2014; Lara Pinto et al., 2017; Pettersson et al., 2017; Remnant et al., 2017; M. Shi et al., 2017). A novel computational framework called Serratus was developed by Edgar et al. (2022) that exploits the RdRp landscape, for mining RNA viruses in publicly available data repositories. Serratus was applied to >5 million SRA datasets and led to the discovery of >100,000 RNA viruses embedded within them.

Although viruses are not monophyletic, recent sequence-led computational analyses have highlighted a number of virus hallmark genes (VHGs) that are shared between different groups of viruses. A study by Iranzo et al. (2016) utilised a hierarchical gene sharing network approach to characterise the VHGs among the dsDNA viruses. This network-based approach identified 19 modules that were representative of dsDNA viruses which formed five major and three minor supermodules. They discovered 14 VHGs including terminase, integrase, helicase, DNA primase, DNA polymerase protease, which highlighted intermodule connections. These hallmark genes included essential viral structural proteins and those involved in virus replication. As network hubs for the two largest supermodules, two major capsid proteins (double jelly roll and HK97-like) were observed. The HK97-like were found in order *Caudovirales* (an order that comprise of bacteriophages) and order *Herpesvirales* (the order that include herpesviruses). The double jelly roll was shared among the putative order Megavirales and smaller viruses, as well as polintons, which are large DNA transposons (Iranzo et al., 2016). These VHGs were subsequently utilised to identify DNA virus sequences from metagenomic datasets and these features were implemented in various virus identification tools including VirSorter (Roux et al., 2015a), CheckV (Nayfach et al., 2020b) and Cenote-Taker 2 (Tisza et al., 2021a). A more recent study by Koonin et al. (2020a) further refined these VHGs and organised them into superviral hallmark genes which were observed to be present across different Baltimore classes. These super hallmark genes included double-jelly-roll capsid protein (DJR-CP; spans BC I and BC II), rolling-circle replication (initiation) endonuclease (RCRE; spans BC I and BC II), RNA-directed RNA polymerase (RdRp; spans BC III, BC IV and BC V), reverse transcriptase (RT; spans BC VI and BC VII), superfamily 3 helicase (S3H; spans BC I, BC II and BC IV) and single-jelly-roll capsid protein (SJR-CP; spans BC I, BC II, BC III and BC IV). This phylogenomics-led network analysis showed that viral super hallmark genes span multiple Baltimore classes, suggesting a network-based taxonomy approach may be more suitable to explain and capture the diversity encompassed within the virus world (Koonin et al., 2020a).

It is important to note that without the powerful metagenomics and metatranscriptomics techniques, the current knowledge of microbial diversity would be very limited. The discovery of new uncultivated microbes through metagenomics can enable us to reveal previously unseen diversity of sequences that can contribute iteratively to increasing the efficiency of current approaches. Incorporating novel microbial sequences identified through metagenomics could help to refine computational methods, which will help us answer some of the most pressing questions in biology and relating to the integrated microbial community surrounding us.

2.3 ‘Unknown’ sequence matter embedded in metagenomic datasets

With the advents in applications of HTS in microbial research and, the advances in the field of metagenomics, the public repositories that hold these raw sequencing data such as the Sequence Read Archive (SRA) and European Nucleotide Archive (ENA) have also grown rapidly in the last decade (Katz et al., 2022). As of 4 September 2022, over 4.2 million open access metagenomic datasets are available on the SRA. This is due to the importance of data sharing for reproducible results: a requirement of funding bodies and scientific journals that sequence data should be published along with the results highlighted in the research papers. This has led to an expansion of sequence databases such as GenBank, that store nucleotide and protein sequence data from various organisms. Although the raw sequences generated as part of metagenomic experiments are made publicly available through SRA or ENA repositories, the assembled contigs data are rarely submitted to the relevant databases. Although it is possible to submit assembled contigs to repositories such as ENA (Hunter et al., 2014) (https://ena-docs.readthedocs.io/en/latest/faq/metagenomes.html), typically, only the contigs that can be classified using metagenomic pipelines and are of interest to the scientific study are submitted to sequence databases. However, in a typical metagenomic dataset, a range of assembled sequences cannot be functionally classified, a large proportion of which, even after excluding spurious contigs, bear no functional or sequence similarity to known sequences and is referred to as biological ‘dark sequence matter’ (Marcy et al., 2007; Bernard et al., 2018).

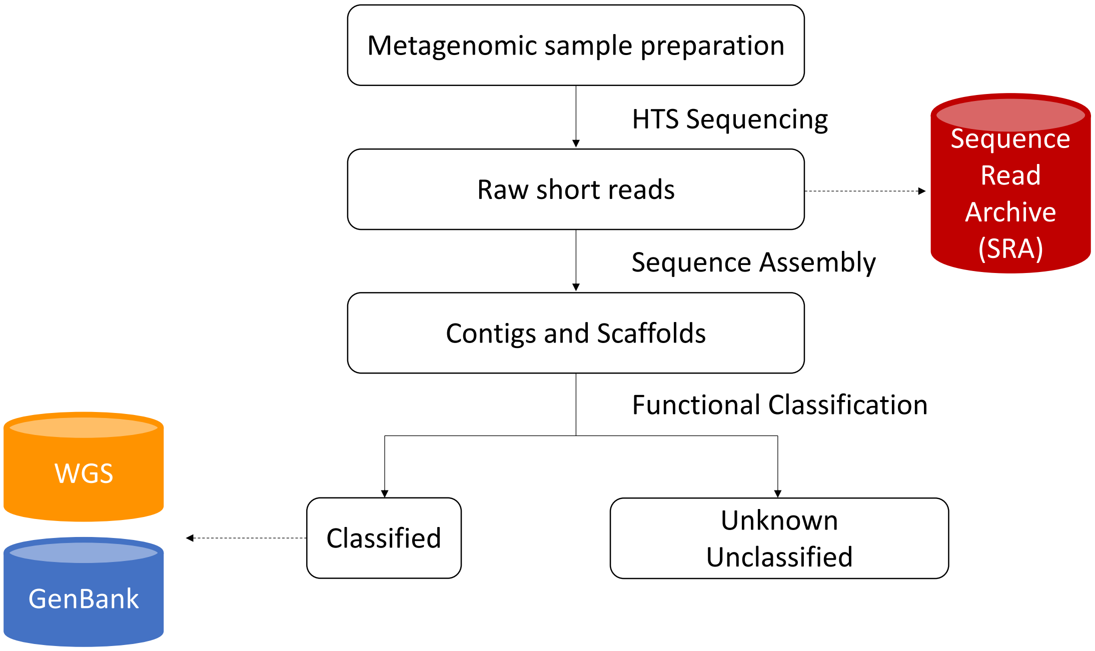


Figure 2.3: This diagram illustrates a typical metagenomic data sharing workflow. The raw HTS sequence reads are generally submitted to short-read databases such as ENA and SRA. The assembled sequences that can be functionally classified are annotated and submitted to relevant INSDC databases such as GenBank. The sequences that cannot be characterised and/or annotated are excluded and would not be shared.

2.3.1 Microbial dark matter

So-called dark sequence matter is defined as any genetic sequence that originates from the biosphere and cannot be assigned to a taxonomic lineage and functional category using the known reference nucleotide and/or protein sequence data (Youle et al., 2012; Krishnamurthy et al., 2017; Bernard et al., 2018). Although the definition of biological dark matter is strictly defined as sequence matter belonging to unknown unknowns, it is important to note that novel sequences identified that are substantially distinct from the known lineage and/or functions are frequently categorised as ‘grey matter’. This refers to sequences that may be distantly related to currently classified functional units and bear very little similarity to them. A schematic representation of this is illustrated in Figure 2. Depending on the analyses, such viral dark matter could constitute approximately 40-90% of all unidentifiable sequence matter (Youle et al., 2012; Hurwitz et al., 2013; Minot et al., 2013; Jennifer R Brum et al., 2015; Fawaz et al., 2016; Krishnamurthy et al., 2017). There are many microbial genome discovery pipelines that can identify novel bacteria, archaea or viruses in metagenomic datasets (Mitchell et al., 2018b; Nooij et al., 2018), however, there is not a streamlined functional analytical pipeline that focuses on the identification, clustering or classification of sequence dark matter. The advent of cultivation-independent approaches like metagenomics and metatranscriptomics has made the microbial dark matter, which is thought to be composed of bacteria, archaea, and viruses, widely accessible. In addition to enabling large-scale exploration and identification of unknown microbial diversity, metagenomics has ultimately led to the exponential growth of a wide range of data repositories. In order to fully understand the uncultivated microbial diversity embedded within these data repositories, a large proportion of their sequences must be systematically analysed. It is possible to look for novel sequences or genomes from microbes that are currently uncultivated by mining publicly available datasets.

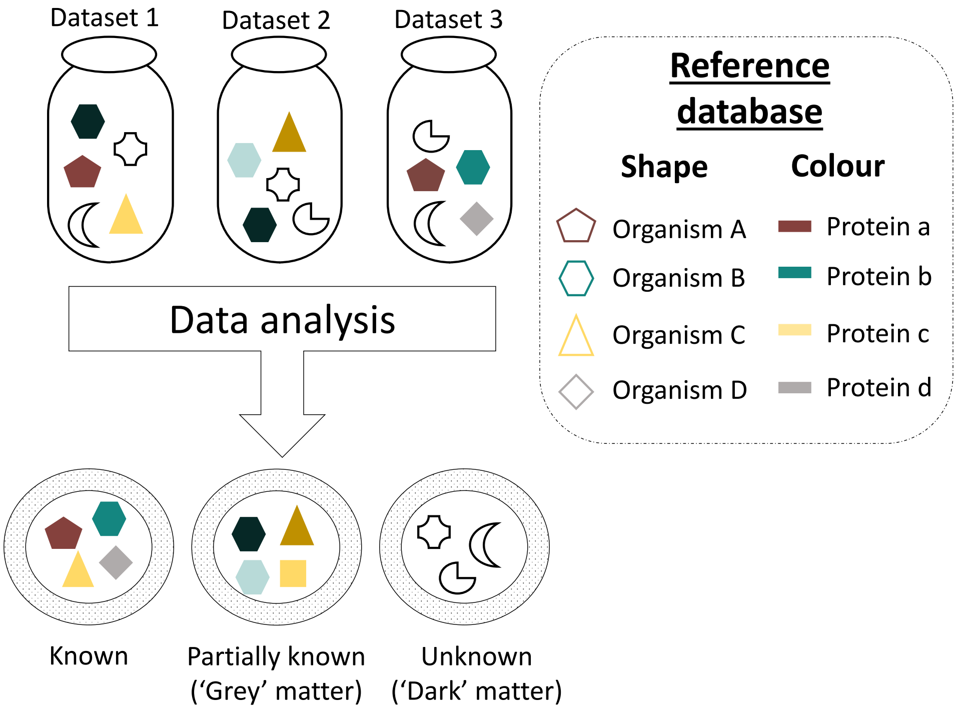


Figure 2.4: A schematic representation of known, partially known and unknown sequence matter in the metagenomic datasets.

Viruses are the most abundant entities on the planet with an estimated 10^{31} particles with the ability to infect microbial populations (Youle et al., 2012). Statistical methods have estimated that each mammalian species harbours around 58 different viruses and extrapolated that 320,000 viruses yet to be discovered that could infect mammalian species (Anthony et al., 2013). Although this study focuses on nine virus families, it provides an estimate of the viral unknown that is yet to be discovered. If this analysis was applied to the 1,740,330 known species of vertebrates, invertebrates, plants, lichens, mushrooms and brown algae then the number would increase to 100,939,140 viruses that are yet to be discovered (*How many viruses on Earth?* 2019). This is equally true for humans, for a given human microbiome e.g. human gut, it is estimated at least the same number of bacteriophages are likely to be present as the number of bacteria and other microbes (Shkoporov et al., 2019b; Sausset et al., 2020). The viruses that can jump from animal species to humans are considered zoonotic viruses (Rahman et al., 2020). An estimate based on the known zoonotic viruses and data extrapolation by the Global Virome Project indicates that around 1.67 million novel viruses are yet to be discovered in mammals and birds. Among these, it was estimated that between 631,000 and 827,00 novel viruses may have zoonotic potential. Though it is very difficult to determine the exact number of viruses that could jump the species barriers and pose potential threat to humans, this approximation highlights the importance of identifying such biological dark matter as well as its scope.

In the early 2000s when metagenomics was first applied to various environments and microbiomes, the concept of dark matter was introduced. This term was typically used to describe the genomic sequences that were so diverse that they did not match any known sequences in the databases (Krishnamurthy et al., 2017; Roux et al., 2021b). In the last two decades, a number of studies have tried to dive deep into microbial dark matter to make sense of these unknowns. A 2013 study by Rinke et al. (2013) that employed single-cell genomics and sequenced nine diverse habitats identified 201 uncultivated microbes belonging to 29 previously uncharted branches of the tree of life. In 2016, Hug et al. (2016) used over 1000 uncultivated and little-known organism genomes from IMG/M (Markowitz et al., 2012), combined with public genomic data, to infer the tree of life and defined a hyper-diverse group of microbial dark matter, called the Candidate Phyla Radiation (CPR), which subdivides the domain Bacteria (Hug et al., 2016). Similarly, Parks et al. (2017) mined 1,550 metagenomes downloaded from the SRA datasets and identified 7,903 novel bacterial and archaeal genomes spanning 17 bacterial and three archaeal candidate phyla. Their study also led to the discovery of 245 genomes from CPR and showed that the relative diversity of this group differs significantly with different protein marker sets (Parks et al., 2017). A recent review focusing on the importance and challenges in studying microbial dark matter highlighted the importance of microbial dark matter mining and reiterated that a number of major bacteria and archaea lineages have been recovered solely through metagenomic sequence mining (Jiao et al., 2021).

2.3.2 Viral dark matter

In 2015, Roux et al. (2015b) mined 14,977 publicly available bacterial and archaeal genomes and identified 12,498 high-quality viral genomes. They utilised VirSorter (Roux et al., 2015a) - a computational method to identify virus-specific signals from bacterial and archaeal datasets leading to the accurate prediction of prophages embedded within the host genomes. However, these novel virus sequences were not recognised as novel viruses in the official virus classification framework. A solely computationally identified crAssphage was shown to be omnipresent in the human fecal microbiome and made up 1.7% of all fecal metagenomic sequences (Bas E. Dutilh et al., 2014). Subsequent studies identified bacterial members from the genus Bacteroides as the natural hosts of crAssphage and also led to the revelations that crAssphages are a group of viruses that reside in the human gut (Robert A. Edwards et al., 2019; Koonin et al., 2020b; Yutin et al., 2018; Guerin et al., 2018). According to ICTV Virus Metadata Resource Master Species List 37 (ratified in March 2022), crAssphages are included in order *Crassvirales* in class *Caudoviricetes* and realm *Duplodnaviria*. The order *Crassvirales* contains 73 species included in 4 families and 42 genera (https://talk.ictvonline.org/taxonomy/vmr/m/vmr-file-repository/13426). Jennifer R. Brum et al. (2016) applied metagenomics and metaproteomics to marine samples and identified 1,875 novel virion-associated proteins specific to dsDNA viruses. The viral dark matter that consists of unknown viruses was demonstrated to play an important role in inflammatory bowel disease as the composition of human virome was observed to be altered compared to healthy human gut virome (Clooney et al., 2019).

Recently a study (Nayfach et al., 2021) analysed 11,810 publicly accessible human gut microbiome samples and generated a comprehensive Metagenomic Gut Virus catalogue that comprises 189,680 viral genomes. Another study, conducted the same year, utilised systematic data mining to identify a range of novel virus sequences in human gut metagenomes. Luis F Camarillo-Guerrero et al. (2020) recovered 142,809 non-redundant gut phage genomes from 28,060 metagenomes and isolate genomes from the human gut. A relevant viral metagenomic study by Gregory et al. (2020) explored the human gut microbiome of 1,986 individuals representing 16 countries and identified >33,000 novel gut virus sequences. This study also explored that human gut virome patterns are age and health status-dependent (Gregory et al., 2020). These studies highlight the importance of identifying and cataloguing viral dark matter. In this day and age where metagenomics has become a routine tool to study the microbial composition of a given sample, dark matter analyses often become synonymous with the discovery of novel viruses.

2.3.3 Mobile genetic elements (MGEs)

The mobilome is defined as any mobile genetic element (MGE) that spread horizontally and within a microbial community (Siefert, 2009). Bulk metagenomic samples often contain plasmids, bacteriophages, mobilisable genetic elements, integrative conjugative elements (ICEs or conjugative transposons), insertion sequences, integrons, and gene cassettes (Carr et al., 2021). MGEs are important in studying the functional elements of the microbiome that affects microbial community composition, antimicrobial resistance genes and virulence factors. The mobilome is the agent of change that facilitates the process of horizontal gene transfers (HGTs). The community-level microbiome datasets that are sequenced through metagenomics, provide a unique opportunity to understand the roles that MGEs play in shaping the microbial evolution. It is also anticipated that mobilome may play an instrumental role in determining how selection pressure impacts microbial communities and their impact on host organisms or tissues (Hall et al., 2022). Moreover, aside from transporting resistance determinants, they also transmit virulence factors and antimicrobial resistance determinants between bacteria (Partridge et al., 2018) which is highly relevant in clinical settings such as antimicrobial resistance.

The metagenomic dataset provides an opportunity to gain further insights into the microbial mobilome. To address the different parts of the mobilome embedded within the microbiome dataset, customised metagenomic data analysis protocols are required as the mobilome elements vary greatly in size and have diverse mechanisms of movement (Carr et al., 2021). For example, plasmid sequences can vary from <1 kilobases (kb) to several megabases (mb) whereas ICEs are at least 18kb long (Siefert, 2009). The bacteriophages can have diverse genome lengths and can integrate their genetic material into the host genomes (prophages). Furthermore, the microbial mobilome is often composed of a mixture of highly heterogeneous elements; moreover, certain elements are difficult to distinguish from one another (Carr et al., 2021). These unique features entailed by MGEs hinder their identification using standard metagenomic data analyses approaches. To capture specific types of MGE, tailored isolation and sequencing approaches need to be employed. For example, a high-throughput transposon-aided capture (TRACA) method is used to isolate circular plasmids from metagenomic DNA. They are then transformed into *Escherichia coli* for cloning, before being sequenced using shotgun approaches or PCR to fill in gaps in sequences (B. V. Jones et al., 2006; Smalla et al., 2015). Other approaches such as size filtering for phage/virus isolation that is described in detail in section 2.2.1 can also be applied to target specific parts of the mobilome. Though these targeted approaches help tackle the resolution issue, they could also lead to a slightly biased representation of MGE abundance and MGE load in a given sample or environments.

Current metagenomic sequencing experiments are heavily reliant on the short reads technology that works well in the case of assembling microbial genomes from scratch but provides limited resolution for MGEs. As fragmented assemblies derived from short read sequences often lack the required resolution to fully reconstruct the mobilome from metagenomic datasets, this challenge is more comprehensive for MGEs. To this end, MGE-specific tools and pipelines have been developed and applied to the microbiome datasets to identify the specific parts of the mobilome, for example there are a large number of phage assembly and identification pipelines developed exclusively to interrogate the bacteriophages present in a given microbiome sample (Fung et al., 2022). There are also plasmid-specific tools available that can be used to identify these specific categories of MGEs from metagenomic data (Carr et al., 2021). However, it is clear that all part of the mobilome are yet to be fully characterised and their genomic signatures are likely to be captured in the current microbiome datasets but exist as biological dark matter or unknowns. It is anticipated that a hybrid sequencing approach and the development of additional specialised tools targeting the mobilome embedded within microbiomes would provide further insights into these mobile genetic elements that are currently considered biological unknowns.

In the era of HTS, researchers are faced with the problem that an increasing amount of sequence data exists that does not match with the currently known genetic sequences using the standard sequence similarity-based approaches such as BLAST, thus it is important to employ other non-similarity-based computational approaches such as sequence prediction and identification using machine learning (Ren et al., 2017; Ren et al., 2020; Barrientos-Somarribas et al., 2018; Maarala et al., 2018). Although dark matter expeditions predominantly focus on sequence identification, the addition of further important information including sequence annotation and gene identification and prediction are deemed equally as important to add value to the newly assembled sequences. A recent framework called Agnostos was developed (Vanni et al., 2022) to categorise the genes with known and unknown functions. It was applied to 400 million microbial genes predicted from 1,749 metagenomes and 28,941 bacterial archaeal genomes. The results showed that around 30% of these genes were deemed of unknown functions. This percentage was smaller than the previous estimates of around 60% in ocean datasets (Salazar et al., 2019) and 40% in human datasets (Thomas et al., 2019). Their analysis also showed that these genes of unknown functions are highly diverse. Moreover, by combining targeted hypothesis testing and an experimental approach, they were able to identify a novel gene that could be involved in antibiotic resistance (Vanni et al., 2022). This gene-level dark matter analysis approach can provide the basis for further research in the field of dark matter that encompasses multiple levels of unknowns.

It is notable that the proportion of unknowns embedded within the microbiome and metagenomic samples changes over time. In the early 2000s, a large proportion of sequences identified from metagenomics could not be attributed to known organisms and the corresponding sequences available in those databases. Hence, the initial quantification of unknown sequences identified from early metagenomics studies was expected to be up to 90% for certain environments and/or microbiomes and it was anticipated to be environment-dependent (Rinke et al., 2013; Krishnamurthy et al., 2017; Solden et al., 2016). However, as metagenomic and single-cell sequencing became popular, they were routinely employed to study various environments and microbiomes, and as a result, a number of previously unknown hidden species were discovered. Moreover, with new taxa included in the set of reference genomes, microbiomes can be analysed more comprehensively since a higher proportion of reads generated from shotgun sequencing experiments match a catalogued microbial genome, which increases the mappability of the metagenome. As an example, in recent years the mappability of the human gut microbiome has increased to an average of 85 percent (Pasolli et al., 2019; Thomas et al., 2019), suggesting that a more comprehensive picture is emerging of the microbial community contained within it. Despite these advances, some of the most studied environments such as human microbiomes may still contain about 20% of unknown sequences that cannot be related to any known sequence (Thomas et al., 2019) emphasising the importance of cataloguing and characterising the genetics of these unknowns that are captured, present yet hidden in sequence datasets.

2.4 Machine learning and its applications in microbial sequence analysis

Exploration and identification of uncultivated microbial diversity have benefited hugely from computational and algorithmic advances including machine learning (ML). Due to their widespread accessibility and application across the field of microbiology, ML and Deep Learning have become regular resource in Bioinformatician/Computational Biologists' tool kits. These methods help researchers understand the mechanisms underpinning uncultivated microbial diversity, and their interactions with each other and can help unravel the diverse role they play in nature and how it affects their hosts including humans (Ching et al., 2018; Qu et al., 2019; Ghannam et al., 2021).

ML provides opportunities to determine and learn patterns from the big datasets such as those generated from HTS technologies and can help to understand complex biological systems of uncultivated microbes. ML models that are disseminated for wider research applications often entail one of the following: a) predictive modelling (supervised learning) and b) description or inferring relationships from the data (unsupervised learning). A typical supervised learning workflow entails the extraction of features from a given dataset and then training various models by sub-setting proportions of these observations captured in the dataset. Data is typically split into two categories: a training set and a testing set. The training set is used to train the models, these trained models are then used to interrogate the test dataset to predict the outcome. The model performance is evaluated by comparing the expected outcome with the model output. Unsupervised ML approaches are utilised in cases where the target outcome of the data is unavailable. Briefly, unsupervised ML models work by analysing unlabelled, unclassified data and attempting to detect hidden patterns in it. Clustering is one of the most popular examples of unsupervised learning (Sarker, 2021).

In recent years, ML has become an increasingly mainstream method in bioinformatics and computational biology, and, has been applied to all types of biological dataset ranging from genomics to protein structure predictions. For example, natural language processing (NLP) methods have been widely applied to sequence datasets whereby shorter sequences (k-mers) derived from genome sequences are converted into embeddings or vectors. These biological sequence vectors can then be used to estimate function and structure, or to feed into other probabilistic models (Yandell et al., 2002; Iuchi et al., 2021). In recent years, ML has been applied to address outstanding biological questions that led to significant advances of the field. Accurate protein structure prediction models implemented in AlphaFold (Jumper et al., 2021) and RoseTTAFold (Baek et al., 2021) utilise neural networks and deep learning algorithms combined with sequence alignment and known protein structures to build and predict structures of novel proteins from sequence data. AlphaFold predicts the local regions of the protein structure that are derived from the protein sequence and structural homology first, and then stitches them together to obtain a complete structure of the input protein sequence. AlphaFold predictions have been shown to achieve similar accuracy as conventional protein structure derivation techniques such as x-ray crystallography (Jumper et al., 2021). These highly accurate protein structure models have been subsequently made publicly accessible through European Bioinformatics Institute (EBI)'s AlphaFold DB platform. This resource currently contains more than 200 million protein structure predictions, allowing researchers to uncover complex mechanisms underlying protein function and interactions (Varadi et al., 2022).

ML can help untangle the complex microbial signals in microbial research. For example, the classification and predictive models can help to detect the presence of a specific microbial species and/or other taxonomic groups in a given sample. A range of tools and packages have utilised this approach to develop models that can accurately predict the taxonomic group of a given sequence. For example, a popular microbiome species prediction tool that can predict the microbial class based on marker genes e.g. 16S and other rRNA signatures, IDTAXA developed by Murali et al. (2018) was shown to outperform more traditional homology search based 16S classification tools such as BLAST, RDP and QIIME. IDTAXA tool has been extended further and a more recent version of this tool that incorporated amino acid signature-based prediction is shown to outperform traditional protein assignment tools such as HMMER and BLAST by accurately linking sequences to KEGG ortholog groups (Cooley et al., 2021). In the case of virus predictions, tools such as MARVEL (Amgarten et al., 2018) and VirSorter (Roux et al., 2015a) have been extremely popular and successful in predicting bacteriophage sequences from microbiomes datasets using virus hallmark genes. Other methods that implemented entirely k-mer signature-based tools like VirFinder (Ren et al., 2017) and DeepVirFinder (Ren et al., 2020) have also been efficiently used to predict both DNA and RNA viruses from metagenomic datasets. An updated version of the virus prediction tool VirSorter, VirSorter2 that combines genomic signatures with other features such as hallmark genes and protein domains can accurately predict DNA and RNA virus sequences from microbiome data (Guo et al., 2021a). Random forest and artificial neural network models that implement simple features such as relative synonymous codon usage (RSCU) have also been shown to perform well in discovering viruses from metagenomic datasets (Bzhalava et al., 2018).

Identification of virus-host associations has been argued as one of the most challenging areas of research that has been even more prominent due to the recent emergence of SARS-CoV-2 (Cobbin et al., 2021; Coclet et al., 2021; Holmes, 2022). Though metagenomics has unravelled previously unseen virus sequence diversity, linking viruses to their host(s) remains a major challenge (Roux et al., 2021b). The host information associated with metagenomically derived viruses is often not readily available through standard metagenomic analyses, hence, the crucial virus-host linkage remains largely unknown (Roux et al., 2021c; Roux et al., 2019; Coclet et al., 2021). As there are no well-established high-throughput experimental method, researchers rely on bioinformatic predictions to link uncultivated phages with their potential hosts since. These predictions are typically based on molecular signals (features) of coevolution and/or an arms race between phages and their hosts, such as identical matches to reference host genomes or Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) spacers, and comparisons of sequence compositions. Computational tools for host predictions typically require a phage genome sequence or genome-derived features that are searched against host features and/or host genome databases to perform host prediction analyses (reviewed in (Coclet et al., 2021)). Broadly this type of analyses can be grouped into three major categories. The alignment-dependent approach typically relies on a host database-led similarity searches that are carried out by BLASTN or equivalent alignment search tools. Sequence similarities are often reflected in integrated proviruses, host-encoded CRISPR spacers, auxiliary metabolic genes (AMG), and/or shared tRNAs, all reflecting different arms races and/or coevolutions between phages and their hosts. There are several tools and databases such as SpacePHARER (R. Zhang et al., 2021), CrisprOpenDB (Dion et al., 2021) that implement CRISPR spacer-led identification approach that would yield highly sensitive and accurate virus-host predictions as spacers represent the previous interactions between phages and their host(s). The similarity between the query phage and a CRISPR spacer (typically 20-70 nucleotide (nt) long) reflects a successful defence of bacteria against a closely related phage. Due to the ongoing arms race between phages and hosts, phage genomes and CRISPR spacers are highly similar in sequence (Horvath et al., 2010; Stern et al., 2012; Dion et al., 2021) leading to accurate host predictions. However, these approaches can have low recalls as they are highly reliant on host genome databases (Coclet et al., 2021; Dion et al., 2021). Other alignment-based approaches that utilise phage marker genes to predict virus-host associations also suffers the same shortcoming as they rely on existing databases and similarity searches. An alternative to alignment-based approach is alignment-free approach whereby short nucleotide sequences (k-mers) and/or protein composition features are used to predict virus host(s). Genome sequence similarities between phages and hosts can be attributed to the adaptation of the phage genome to host replication, transcription, and translation machinery (Roux et al., 2015a). These methods allow host prediction for a broader range of phages than alignment-based approaches because they do not require the presence of closely related phages or hosts in the database. However, they tend to be less accurate than alignment-based approaches. A number of these alignment-free prediction tools implement machine learning models to carry out predictions. Due to the high level of uncultivated virus diversity being identified from microbiome and metagenomic datasets, the race to accurately link these viruses to their host has been deemed a welcome challenge by ML researchers. A range of virus-host prediction strategies such as prophage detection (Roux et al., 2015b), CRISPR spacers signatures (Dion et al., 2021), and virus-host genomic composition analyses (Babayan et al., 2018; Young et al., 2020) has been translated into ML models that can accurately assign a host taxa to uncultivated virus species. A study led by Young et al. (2020) showed that Support Vector Machines (SVM) models trained on short nucleotide and protein *k*-mers, protein domains as well as physio-chemical properties of amino acid sequences were accurately able to predict the host of both prokaryotic and eukaryotic viruses. Young et al. (2020) also recommended an integrative approach of combining a range of these features can lead to improved virus-host predictions. Another study by Babayan et al. (2018) was able to implement gradient boost ML models to accurately predict reservoir hosts and vectors of RNA viruses based on their genomic signatures. VIDHOP - a deep neural network approach devised by Mock et al. (2021) that can accurately predict the host based on short 100-400 bases signatures from viruses without needing the models to be trained on the host genomes, provides promising advancement to link uncultivated viruses to their host(s), especially in the context of pandemics and zoonosis. Beyond these widely popular research avenues, ML has also been used to predict disease based on the microbial communities in microbiome samples (J. Y. Shi et al., 2018), identifying interactions and associations between microorganisms (Leite et al., 2018) as well as exploration of microbiome-disease associations (X. Chen et al., 2017; Yan et al., 2020; Yan et al., 2021). The third and final category integrates both alignment-led and alignment-free approaches. These integrated approaches maximise both the recall and accuracy of phage–host predictions as they integrate multiple approaches and address specific challenges and limitations of each method. Integrative approaches implemented in VirHostMatcher-Net (W. Wang et al., 2020) and PHISDetector (F. Zhang et al., 2020) use machine-learning models and score the overall probability of individual phage-host pairs using a combination of alignment-free and alignment-based features. A combination of alignment-free and alignment-based features is incorporated into both tools based on (i) k-mer frequencies based similarities between phages and hosts; (ii) CRISPR spacers shared by both phages and hosts; and (iii) alignment-based matches between phages and hosts. To achieve best recall and accuracy of phage–host pair detection, the use of multiple host prediction tools appears to be a reliable strategy.

ML and Deep Learning methods present a significant opportunity to apply these extensively used data analysis techniques and tools to explore microbial as well as dark sequence matter. As microbial dark sequences are likely to possess significantly different genomic signatures to currently known sequences, the ML approaches can be customised, trained and adapted to comprehend the current knowledge of microbial genomics, and apply it to explore the unknown microbial diversity embedded within sequence data repositories.

2.5 Motivation

The study of biological dark sequence matter is a very dynamic field of microbial research as new organisms and environments are sequenced at a fast pace, and the use of sequencing technologies has been increasing the data output. Identification of sequences of unknown origin should be considered an iterative process. As more microenvironments in the biosphere are sequenced, more novel organisms and their genomic contents will be catalogued. The sequences that may have been the biological dark matter a few years ago before their discovery, such as crAssphage and Chuviruses, have now become mainstream identifiable and classifiable sequence matter.

These discoveries are important to understanding the interactions between humans and other microorganisms present in the biosphere. By mining the metagenomic data for such unknown signatures, I aim to explore, catalogue and potentially classify the microbes that may be present in the different organisms and/or environments. Identification of novel microbial species and their corresponding genomes provide the first insight into their sophisticated ecosphere. This would lead to a better understanding of the microbial world around us. This project aims to identify, quantify and potentially classify the unknown microbial matter in human metagenomic and vector metagenomic datasets. These unknown microbes could be potential disease-causing agents. A better understanding of the microbial biosphere around us could also help us understand, predict and control disease outbreaks. Cataloguing and characterising the dark sequence matter is important in the context of emerging viruses. It will provide a means of determining whether a novel virus has already been sequenced and in which studies it has been sequenced, lead to a better understanding of virus-host interactions and can also help to detect and catalogue common viral contaminants in various datasets. In order to reduce observational bias or the street light effect, it is important and inevitable to move away from looking where it is most obvious to look to discover truly new organisms.

Chapter 3

Identification and quantification of ‘unknown’ biological sequences in human microbiomes

Unknown explorations

The UnXplore framework and unknown sequence analyses described in this chapter are published in mSystems. A copy of the accepted manuscript is included in Appendix D.

Modha S, Robertson DL, Hughes J, Orton RJ. Quantifying and Cataloguing Unknown Sequences within Human Microbiomes. mSystems. 2022 Apr 26;7(2):e0146821. DOI: 10.1128/msystems.01468-21.

3.1 Abstract

Advances in high throughput sequencing technologies and cheaper sequencing costs have led to the rapid growth of the data repositories that hold these data. With these advances, metagenomics and metatranscriptomics have become popular tools to study and acquire a snapshot of the microbial communities in various environments. However, due to the limitations of the various databases used in the microbial identification analysis, there are a large number of unknown sequences that are embedded within these repositories that often remain unidentified. To this end, a portable and extendable framework was developed to systematically quantify the amount of unknown biological matter in publicly available metagenomic repositories. A survey of this data suggests less extensively explored microbiomes such as skin and oral microbiomes contain a large amount of unknown biological sequences on average. These unknown sequences are found in most microbiomes and could potentially belong to uncultured and unidentified novel microbes that surround us and that we interact with on a daily basis.

3.2 Introduction

Metagenomics has become an increasingly mainstream tool to catalogue the microbial makeup of any given habitat (Aguiar-Pulido et al., 2016; Koonin, 2018; Quince et al., 2017; Thomas et al., 2019). It has been applied to a diverse range of environments from human body sites (Foulongne et al., 2012; Gevers et al., 2012; Consortium et al., 2012; Qin et al., 2010) to the depth of vast oceans (Breitbart et al., 2002; Hurwitz et al., 2013; Mizuno et al., 2013). Metagenomics provides a relatively unbiased approach compared to culture-based methods; to observe, measure and understand the interactions of the microbes within communities as well as with their hosts, including humans (Quince et al., 2017). Underpinned by powerful insights and relatively cheaper sequencing costs, metagenomics has become a routine technique to study the microbial content of any environment (Koonin, 2018).

These advances in sequencing technologies have led to the rapid expansion of publicly available sequence repositories. This is due to the importance of data sharing for reproducible results: a requirement of funding bodies and scientific journals is that sequence data should be published along with the results highlighted in the research papers. This has led to the growth of sequence databases such as GenBank, that store nucleotide and protein sequence data from various organisms (Cochrane et al., 2015; Karsch-Mizrachi et al., 2017). However, although the raw sequences generated as part of metagenomic experiments are made publicly available through Sequence Read Archive (SRA) or European Nucleotide Archive (ENA) repositories, the assembled contigs data are rarely submitted to the relevant databases (Connor et al., 2019). The reason for the absence of these types of data can be associated with the requirement for sequences to be annotated before their submission to annotated databases such as GenBank, which is not possible when the organism the sequence came from is unknown. INSDC databases such as ENA allow scientists to submit assembled and unannotated contigs, but this practice is not always followed. Moreover, all of the contigs generated as part of this analysis may not be relevant and/or of interest for a specific research goal. Furthermore, the unidentified contigs are often discarded and excluded from downstream analysis.

The raw data in public databases are typically analysed using metagenomic protocols designed to address specific project aims. There is a range of different tools and pipelines available for metagenomic sequence analysis. There is a limited comparison of these pipelines as they are usually developed to address a specific research question. For example, there are approximately 50 workflows available for virus metagenomic analysis that have been used in different publications with primarily different aims (Nooij et al., 2018). As part of the routine metagenomic analysis, only the contigs that can be classified using a specific workflow and that are of interest to the scientific study are submitted to sequence repositories such as GenBank. However, in a typical metagenomic dataset, a range of assembled contigs cannot be functionally or taxonomically classified, a large proportion of which, even after excluding spurious contigs, bear no functional or sequence similarity to known sequences and are often referred to as biological ‘dark’ or uncharacterised sequence matter (Youle et al., 2012; Krishnamurthy et al., 2017; Bernard et al., 2018). Although the terminology itself has been controversial (Murat A., 2020), it typically refers to the sequences of unidentified taxonomic or functional origin.

Generally, these contigs are often excluded from downstream analyses. However, a number of recent studies have highlighted the importance of identification and categorisation of such unknown sequences: A study led by Almeida et al. (2019) have mined over 11,850 human gut microbiome datasets and has identified nearly 2000 novel uncultured bacterial species from 92,143 genomes assembled from metagenomic datasets. Similarly, another focusing on multiple human microbiomes assembled 150,000 microbial genomes from 9,428 metagenomic datasets (Pasolli et al., 2019).

Characterisation of Metagenomically Assembled Genomes (MAGs) as microbial origin has strengthened the hypothesis that the uncharacterised biological sequence matter is highly likely to belong to the uncultured bacteria, archaea and viruses that surround us (Rinke et al., 2013; Bernard et al., 2018; Thomas et al., 2019; Woyke et al., 2019). A range of different 'dark' matter studies has led to the identification of novel microbes, including the identification of novel bacterial and archaeal phyla and superphyla (Rinke et al., 2013; Saw et al., 2015). Previous studies have shown that dark sequences of unknown lineage and unknown functions tend to be of viral origin (Youle et al., 2012). For example, a novel identified phage species crAssphage has been shown to constitute approximately 1.7% of all fecal metagenomic sequences (Bas E. Dutilh et al., 2014). A more recent study by (Yutin et al., 2018), predicted that this phage is likely to belong to a crAss-like family of viruses that are associated with diverse bacteria from the phylum Bacteriodetes. A study by Roux et al. (2015b) mined 14,977 publicly available bacterial and archaeal genomes and identified 12,498 viral genomes linked to their hosts. This is applicable to human datasets too, a study mined human metagenomic data and identified 32 novel predicted putative gene families of which one family is shown to be related to the Torque Teno virus and has led to the identification of a novel bacteriophage called bacteriophage HFM (Barrientos-Somarribas et al., 2018). A study led by Kowarsky et al. (2017), found that 1% of cell-free DNA sequences appear to be of non-human origin in human blood samples and only a small fraction of them can be mapped to currently known microbial sequences. Despite this, multiple levels of unknowns remain an ongoing challenge in microbiome research (Thomas et al., 2019) and the identification of viruses in 'dark' matter remains an even greater challenge due to the absence of a universal gene signature and the high diversity among virus genome content (D. Wang, 2020).

There has been a community-wide effort to address the above challenges by mining the sequences present in the short-read archives (Connor et al., 2019; Mitchell et al., 2018b) to compile a complete list of assembled sequences and then, annotate them to identify the diversity that is harboured within these unexplored sequences. A range of different tools and pipelines have been developed to forward this field of research (Pasolli et al., 2017; Sczyrba et al., 2017; Mitchell et al., 2018b; Von Meijenfeldt et al., 2019; Paez-Espino et al., 2019; Tisza et al., 2020;

Galloway-Peña et al., 2020). However, a comprehensive computational framework and associated database that provide details about the presence of uncharacterised biological matter in different metagenomic samples are still to be designed.

In this project, a framework was developed that can be applied to metagenomic datasets and can enable the detection of sequences of unknown taxonomic origins. This framework was employed to mine human microbiome datasets to quantify the extent of unknown sequences embedded within them. Additionally, the UnXplore framework could be expanded to incorporate potential classification and comparison of biological unknown sequence matter between different datasets.

3.3 Methods

Metagenomic sequence analysis is a computationally intensive task and requires an appropriate computational environment to analyse the large volumes of short-read data. The European Bioinformatics Institute (EBI) has developed a 'standard' metagenomic analysis pipeline - MGnify (Mitchell et al., 2018b; Mitchell et al., 2019) that has been made available to all researchers. This online platform allows users to submit their data to the ENA and offers standard metagenomic analysis facilities. Although the workflow of this pipeline is not tailored to serve a specific project, it helps researchers to get an overview of the microbial communities present in their samples. Briefly, MGnify includes a *de novo* assembly step that generates contigs from the studies and all contigs generated with the pipeline. These contigs are submitted to downstream analysis that includes ribosomal profiling, open reading frame predictions, domain identification and functional annotations (Mitchell et al., 2019). However, this general-purpose framework cannot quantify the proportion of unknown sequences in samples analysed using this pipeline.

In this study, datasets available within MGnify resources were included. All human microbiomes submitted to ENA which were included in the MGnify databases were downloaded with corresponding metadata on 19 April 2019. This included a set of 351 unique studies comprising a range of different microbiome datasets from human hosts. In order to obtain further metadata, each study was linked to the corresponding SRA repository using NCBI eutilities (Sayers, 2018). One study with ENA accession MGYS00000314 could not be linked to SRA databases. As this project focuses on metagenomic datasets, studies targeting metabarcoding-based sequencing methods such as 16S and/or amplicon sequencing were excluded with studies that solely focused on third-party annotation i.e. analysis of previously published data that lack primary data were also excluded (n=190). In order to reduce sequencing technology-related bias, studies that utilised sequencing platforms other than Illumina were excluded (n=49). Any Illumina platform samples with amplicon library preparation `LibraryStrategy == AMPLICON` were also excluded as they typically do not represent metagenomic sequencing (n=51). A number of studies utilised multiple sequencing platforms (n=5), and, a few studies included multiple types of library preparations (n=3), in these cases, samples were sequenced on the Illumina platform along with non-amplicon samples were included (n=3). To keep the pilot datasets to a reasonably manageable size, studies with more than 100 samples were also limited to a random sampling of 100 samples (n=19: fecal=12, human=4, intestine=2, oral=1). Overall, the generated curated set included in the pilot study comprised 44 distinct studies with 1130 samples.

Initially, all samples were downloaded using the `DownloadSRAReads.py` (https://github.com/sejmodha/UnXplore) script that used parallel-fastq-dump (Valieris R., 2020) to get a local copy of the published short reads archives data files in fastq format. These samples were divided into two major categories according to the sample's library layout: single-end and paired-end. This categorisation also helped to design a customised analytical pipeline suitable for the corresponding reads layout. The pilot dataset included 790 paired-end (PE) and 340 single-end (SE) samples. However, a range of sample-specific raw data files were missing from the SRA. 8 samples from BioProject PRJEB19188 and 1 sample from BioProject PRJEB14383 were not publicly available on the SRA, and, were excluded from the analysis. In total, 1121 samples (789 PE, 332 SE) from 43 distinct studies were successfully downloaded and submitted to the metagenomic analysis pipeline.

3.4 Results

3.4.1 Metagenomic analysis

A comprehensive metagenomic workflow was designed to analyse samples included in this study. An overview of the analytical approach is shown in figure 3.1

Quality assessment

Out of 1121 samples, 158 could not be assembled due to insufficient reads and were excluded from downstream analysis. A majority of these samples were from BioProjects PRJEB14782 and PRJEB15057. Further details about these samples are shown in the appendix tables A.1 and A.3.

In summary, 963 samples from 40 distinct studies were included in the pilot study and were processed using the complete metagenomic pipeline developed as part of this project. All results described in this chapter are summarised based on these 963 (784 PE, 179 SE) samples. A brief overview of each study, description and number of samples is listed in the appendix table A.2

In order to assess the quality of the samples and remove sequencing adaptors, all samples were submitted to 'bbduk' from the BBTools package (Bushnell B., 2015a; Bushnell B., 2019). The word Duk in bbduk stands for Decontamination Using Kmers. BBDuk is capable of efficiently trimming, cleaning and filtering sequences based on kmer matches. The recommended kmer threshold of `mink=11` and `k=23` were used. All samples included in this study were sequenced using the Illumina sequencing platform, however, the sequencing methods varied between samples. It is often not possible to trace the exact sequencing adapters used during sequencing, a fasta file containing a range of adapters was used to trim the adapter sequences from the reads for all samples. BBDuk auto-detected the presence of the relevant adapter sequences from the input fastq file specified and trimmed them. Additionally, commonly known sequencing contamination and spike-in sequences including PhiX adapters, PhiX Illumina sequences and other miscellaneous sequences were also trimmed and removed as part of the quality trimming step. In order to retain the best quality reads, all reads below the average PHRED quality score of 20 were trimmed after kmer-based filtering. BBDuk recommended parameters for paired-end samples `tbo` and `tpe` were also applied to trim adapters where pair overlap was detected (`tpe`), and, to ensure that both reads were trimmed to the same length if the adapter sequence was only detected in one of the pairs (`tbo`) (Bushnell B., 2015a). All reads that pass the rigorous trimming and quality filters were retained and submitted to the next step of the pipeline.

The quality trimmed reads were mapped to the human genome sequence build GRCh38 using the Burrows-Wheeler Aligner (BWA); a tool for mapping short reads to its corresponding reference genome (H. Li et al., 2009). BWA is one of the fastest and most accurate tools for mapping reads back to large reference genomes such as the human genome (Hatem et al., 2013). These alignments are stored in Binary Alignment/Map (BAM) files. These files were processed to obtain the unmapped reads that were extracted using SAMTools (Heng Li et al., 2009).

Read normalisation

Biological samples processed using metagenomic protocols typically contain short sequencing reads distributed unevenly across sequenced genetic material. If these reads are assembled using *de novo* assembly tools, they often result in the assembly of the most abundant species in the samples, thus missing the low-abundant species. In order to reduce sequence assembly bias, the best practice is to normalise the reads prior to the *de novo* assembly step (Howe et al., 2014). In this pipeline, BBNorm (Bushnell B., 2015b) was used to normalise reads based on the kmer coverage composition. This step also enabled the acceleration of the assembly process as only a subset of reads were used to build the *de novo* assembly and resulting in better assembly quality overall (Crusoe et al., 2015). BBNorm employs a kmer-based coverage algorithm whereby the user can define the minimum number of kmers coverage cut-off to be used for the read normalisation. Typically, sequencing depth under 2x is understood to be sequencing errors (Bushnell B., 2015b), therefore, a kmer threshold of 3 (`mindepth=3`) was implemented in the pipeline and any kmers below that threshold are deleted. This step also accounts for sequencing errors and helps to remove reads with very low-frequency kmers that usually occur due to sequencing errors, and, retains the critical reads that account for the real diversity in the samples.

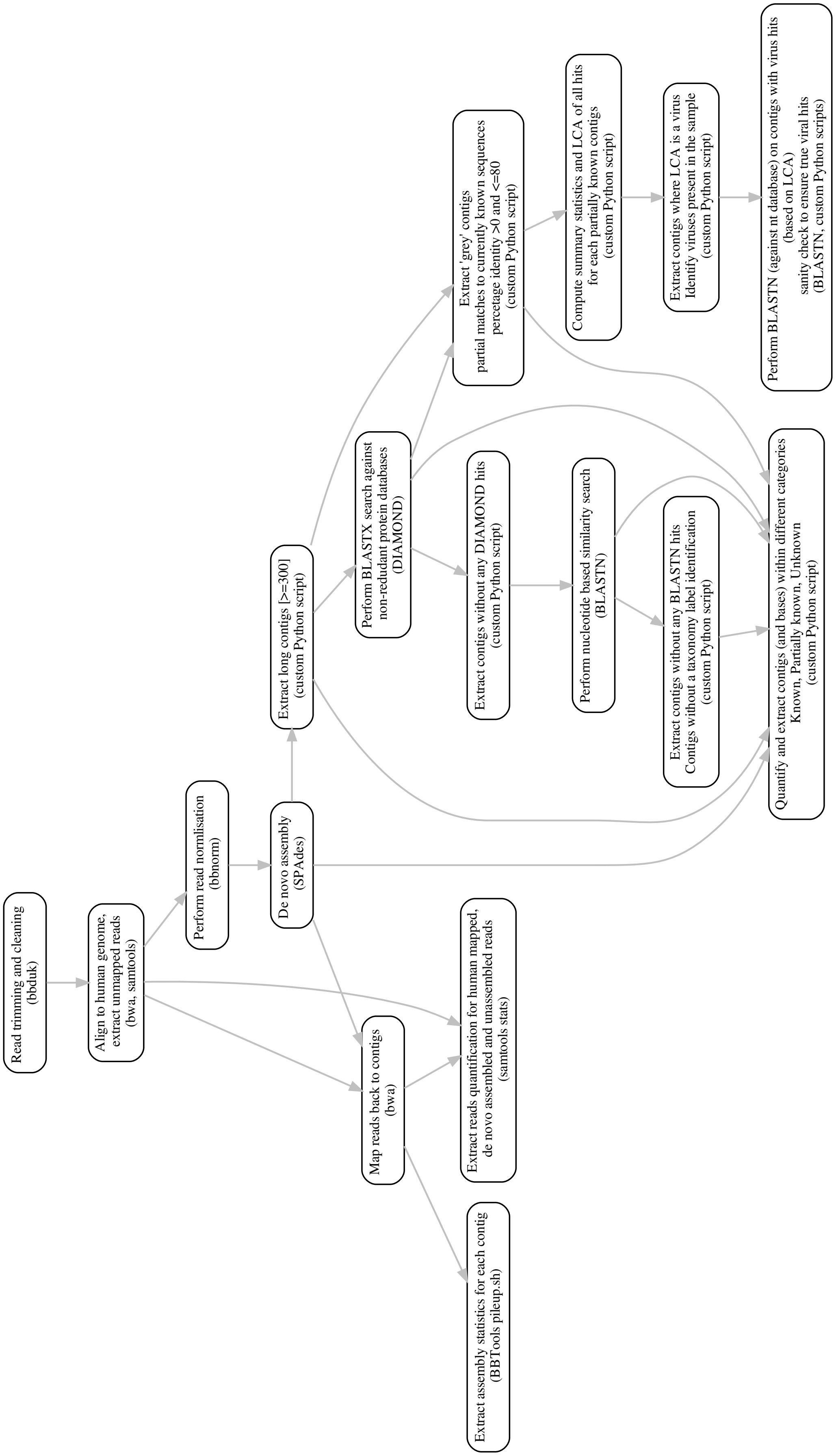

Read trimming and cleaning
(bbduk)
Align to human genome,
extract unmapped reads
(bwa, samtools)
Perform read normlisation
(bbnorm)
De novo assembly
(SPAdes)
Map reads back to contigs
(bwa)
Extract reads quantification for human mapped,
de novo assembled and unassembled reads
(samtools stats)
Extract assembly statistics for each contig
(BBTools pileup.sh)
Extract long contigs [>=300]
(custom Python script)
Perform BLASTX search against
non-redudant protein databases
(DIAMOND)
Extract contigs without any DIAMOND hits
(custom Python script)
Perform nucleotide based similarity search
(BLASTN)
Extract contigs without any BLASTN hits
Contigs without a taxonomy label identification
(custom Python script)
Extract 'grey' contigs
partial matches to currently known sequences
percetage identity >0 and <=80
(custom Python script)
Compute summary statistics and LCA of all hits
for each partially known contigs
(custom Python script)
Extract contigs where LCA is a virus
Identify viruses present in the sample
(custom Python script)
Perform BLASTN (against nt database) on contigs with virus hits
(based on LCA)
sanity check to ensure true viral hits
(BLASTN, custom Python scripts)
Quantify and extract contigs (and bases) within different categories
Known, Partially known, Unknown
(custom Python script)

Figure 3.1: UnXplore: A metagenomic analysis and unknown sequence identification pipeline

De novo assembly

The normalised reads were used for *de novo* assembly using the SPAdes assembly pipeline, part of the SPAdes package (Nurk et al., 2013). metaSPAdes is deemed to be the best tool for assembling microbial genomes (Nurk et al., 2017). It employs a *de bruijin* graph-based approach to assemble the reads into longer stretches of sequences - labelled as contigs. metaSPAdes assembly pipeline is not available for single-end reads and the samples with single-end reads were assembled using the standard SPAdes pipeline using the default parameters.

The `FilterFasta.py` (https://github.com/sejmodha/UnXplore/) script was developed to extract contigs that were longer than 300 nucleotides. *De bruijin* graph base approaches break short sequence reads into even shorter kmers, hence this approach often leads to misassemblies. A threshold of length 300 (that is the length of two reads in a pair or twice the length of a read) was applied to filter out short contigs as they often represent noise and misassemblies generated using the short kmer-based approach implemented in *de brujin* graph-based assemblies (Bergner et al., 2020). Such short contigs do not contain adequate information and were excluded from downstream analysis as a precautionary measure, and the remaining long contigs were used in the subsequent steps.

The normalised subset of reads was used to generate assemblies, however, these reads cannot be used to assess the assembly quality as they represent a small subset of the actual reads. To assess the assembly quality, the complete set of reads that did not map to the human genome was mapped onto the *de novo* assembled contigs with BWA using the default parameters.

Taxonomic annotation

The 'long' contigs were searched against the non-redundant (nr) protein databases using the BLASTX algorithm implemented in DIAMOND (Buchfink et al., 2014). It has been demonstrated that this algorithm is 10,000 times faster than the stand-alone BLASTX (Altschul et al., 1990) and has a similar level of sensitivity. It carries out the six-frame translation of the nucleotide sequences and then searches those translated sequences against the nr protein databases. This step is very important and enables the identification of distantly related homologues of the queried sequences due to the 6 frame translation from nucleotide into protein sequences. The top 25 hits for each contig were extracted and analysed downstream (`--unal 1 --evalue 0.001`).

The contigs that did not have any protein matches were extracted and searched against the comprehensive nucleotide database (nt) using BLASTN (`--evalue=0.001`). This step helped to identify and remove non-coding sequences such as ribosomal RNA and untranslated regions of currently sequenced organisms included in the databases.

To determine the most appropriate organism that the contigs were matching to, the lowest common ancestor (LCA) was obtained from the top 25 hits. Python scripts `ExtractLCA.py` and `ExtractLCABLASTM6.py` were used to determine the taxonomic LCA from DIAMOND and BLASTN tabular output (https://github.com/sejmodha/UnXplore/).

Validation and statistics

In order to validate the quality of the contigs, all reads that did not map to the human genome were mapped back to the contigs. The assembly quality statistics such as coverage, length, and the number of mapped reads were generated for each contig using `pileup.sh` - a script included in the BBTools package (Bushnell B., 2019). Additionally, `samtools stats` was used to gather metrics about the reads mapped to the human genome, reads that were submitted to *de novo* assembly and the reads that could not be assembled.

To analyse partial matches in more detail, `ExtractPartiallyKnownSeq.py` script (https://github.com/sejmodha/UnXplore) was used to filter BLASTX results to identify contigs with <=80 percent identity at the protein level. All relevant hits for each of those contigs were grouped together, and, the Lowest Common Ancestors (LCA) species were identified using the `ExtractLCA.py` script (https://github.com/sejmodha/UnXplore). The contigs that match exclusively to viruses i.e. the LCA was deemed to be a virus taxonomic group, were investigated further to identify the virus species using `ExtractViralHits.py` (https://github.com/sejmodha/UnXplore). The contigs that match exclusively to virus proteins were searched against the nucleotide database once again to carry out a final sanity check on the assembled contig sequences that could potentially originate from viruses. This step also worked as a quality assessment to ensure that the contigs matched to viruses and were not spurious hits.

The pilot study set included a range of different sample types as described in figure 3.2(a). It is important to note that this set is highly skewed towards the human gut microbiome that is normally sampled through fecal material. This skewness highlights the current bias towards the gut microbiome studies over other human microbiomes. The second most common microbiome included in the study was the oral microbiome. Although other microbiomes were under-represented in the pilot study, it is clear that the initial pilot project covered a wide range of samples from various human bodily sites and fluids. An ambiguous microbiome ‘Human’ was included in this dataset that represents 3 distinct studies including PRJEB14301 (CSF, n=1), PRJEB21827 (A/B testing for colon model, n=12) and PRJEB6045 (metagenomics of medieval human remains from Sardinia, n=1).

The microbiomes originated from different countries around the globe as shown in the figure 3.2(b). This figure shows the global distribution of 861 samples analysed for which geographical location was available. Most studies were from western Europe and the geographic distribution of the samples included in the pilot study is skewed toward western countries in the world. The location data was extracted from the SRA metadata resources using `pysradb` (Choudhary, 2019) for each study. The location information could not be found for PRJEB11554 (n=1), PRJEB12998 (n=1), PRJEB21827 (n=12), PRJEB5761 (n=81), PRJNA264728 (n=8) and PRJNA43253 (n=7) and those data points were excluded from the figure 3.2(b). A complete list of study locations is shown in the appendix table A.4. These samples were sequenced in various sequencing facilities across the world, and the complete distribution of the sequencing centre is shown in the appendix

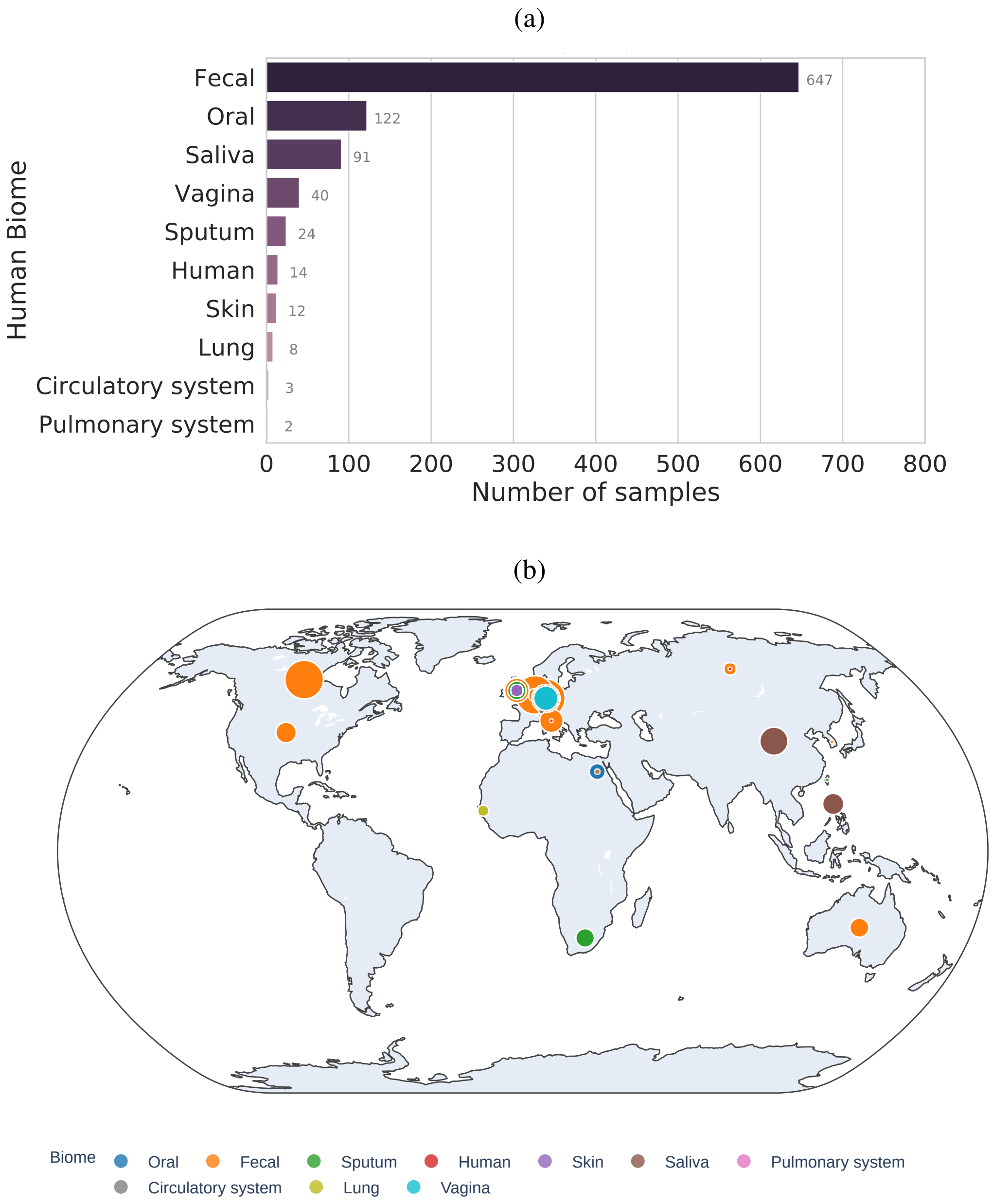


Figure 3.2: Overview of human microbiome samples included in this analysis and their geographic distribution. (a) The number of samples included in this study per microbiome (n=963). (b) Overview of the geographical distribution of the samples included in the pilot study (n=861). Circles are coloured according to the different microbiomes and the size of the circle corresponds to the number of samples. Geographical locations were not available for 102 samples.

figure A.1.

Table 3.1: A comparison between sequence lengths of raw reads downloaded from the SRA and the cleaned reads used for *de novo* assembly.

	SRA Reads	**Clean Reads**
count	963	963
mean	129	108
std	58	40
min	36	30
max	301	264

To assess the quality of data, extensive metrics related to the sequence data were generated. The read length distributions were calculated for each sample as the read lengths varied among different studies, samples and microbiomes. Read lengths ranged from 36 to 301 bases for raw sequence data downloaded from the SRA. After trimming and QC cleaning, assembled reads were shorter due to adapter trimming and low-quality base cleaning. The average length of the cleaned reads was 108 bases as shown in table 3.1. As the read lengths varied widely between the samples and the studies, it was not possible to compare the quality metrics using the read length measure as it could be misleading. To enable this comparison, quality assessment metrics were carried out on a number of bases.

The QC step of the assembly led to the loss of bases that were trimmed due to the rigorous quality trimming and clipping criteria mentioned in the section 3.4.1 in the methods. Overall, 11.87% of bases were lost during the QC step (figure 3.3(a)) compared to the raw data. This is expected as bases from adapter sequences, spike-ins and those that were of low quality were trimmed as part of the QC step. A more detailed overview of the data is shown in table 3.2. On average, 2-28% of bases were lost. The mean proportion of bases lost due to trimming and cleaning was 13% (standard deviation: 14.27%). These values varied largely between different microbiomes. For example, in the case of oral microbiome studies PRJEB12831 and PRJEB15334, 94.6% and 83.3%, bases were lost respectively compared to vaginal samples where less than 2% of bases were trimmed off on average.

Table 3.2: Percentage of bases lost after QC grouped by microbiome. The count represents the number of samples included in each microbiome group.

	Circulatory system	**Fecal**	**Human**	**Lung**	**Oral**	**Pulmonary system**	**Saliva**	**Skin**	**Sputum**	**Vagina**
count	3	647	14	8	122	2	91	12	24	40
mean	5.94	10.48	20.11	23.45	25.41	21.01	27.66	15.26	7.72	1.95

	Circulatory system	**Fecal**	**Human**	**Lung**	**Oral**	**Pulmonary system**	**Saliva**	**Skin**	**Sputum**	**Vagina**
std	8.48	9.49	5.34	6.17	17.36	20.78	23.37	12.54	2.24	0.23
min	0.86	0.03	15.1	13.99	2.78	6.32	0.8	5.19	3.02	1.64
max	15.73	63.94	34.8	33.44	94.63	35.7	77.03	50.83	13.26	2.55

3.4.2 Reads classification and assembly analysis

All QC passed reads were subjected to reference mapping to the human genome. This is an important step to remove unwanted and known human sequences from the samples as the aim was to explore the microbial makeup and unknown sequences of these samples. Overall, 30.75% of all bases were mapped to the human genome (figure 3.3(a)).

On average, 14.5% of bases were mapped to the human genome. This proportion also varied largely between all microbiomes. Microbiomes including skin, sputum, vagina and lung had between 60-70% of all cleaned bases mapped to the human genome. In contrast, saliva and fecal microbiomes contained <8% of bases mapped to the human genome on average. Any bases that were lost when extracting unmapped reads from the BAM files were also calculated. This category would include bases that were lost due to one of the PE read mapping to the human genome and the other not and these bases were discarded and were not analysed further. This category contains the smallest proportion of bases lost and on average 0.03% of bases were lost in this filtering step.

All remaining reads that did not map to the human genome were extracted, normalised and submitted to the *de novo* assembly step. In total, 12,038,529,159 (>12 billion) reads representing nearly 1,195,949,958,509 (>1.1 trillion) bases were assembled in this study. Pink bars represent these assembled bases in figure 3.3(b). It is important to check that reads map back to the contigs as *de Bruijn* graph-based assemblers break reads down into kmers, and can assemble spurious contigs that may not represent the original sequence data. On average, 61% of bases were able to map back to the *de novo* assembled contigs. The saliva microbiome had the highest proportion of assembled contigs with an average of 81.7% of bases mapping back to the contigs. In contrast, the skin microbiome only had around 10%. However, all microbiomes had a proportion of bases that could not be assembled into contigs. QC passed reads were divided into three major categories: (i) mapped to the human genome, (ii) assembled and mapped to contigs, and (iii) reads that could not be assembled. Figure 3.3 provides an overview of the average proportion of bases in each of these three categories for each microbiome.

Unassembled sequences

‘Unassembled’ bases are defined as those that did not map to the human genomes and could not be assembled into contigs. These sequences could not be classified as part of this project but

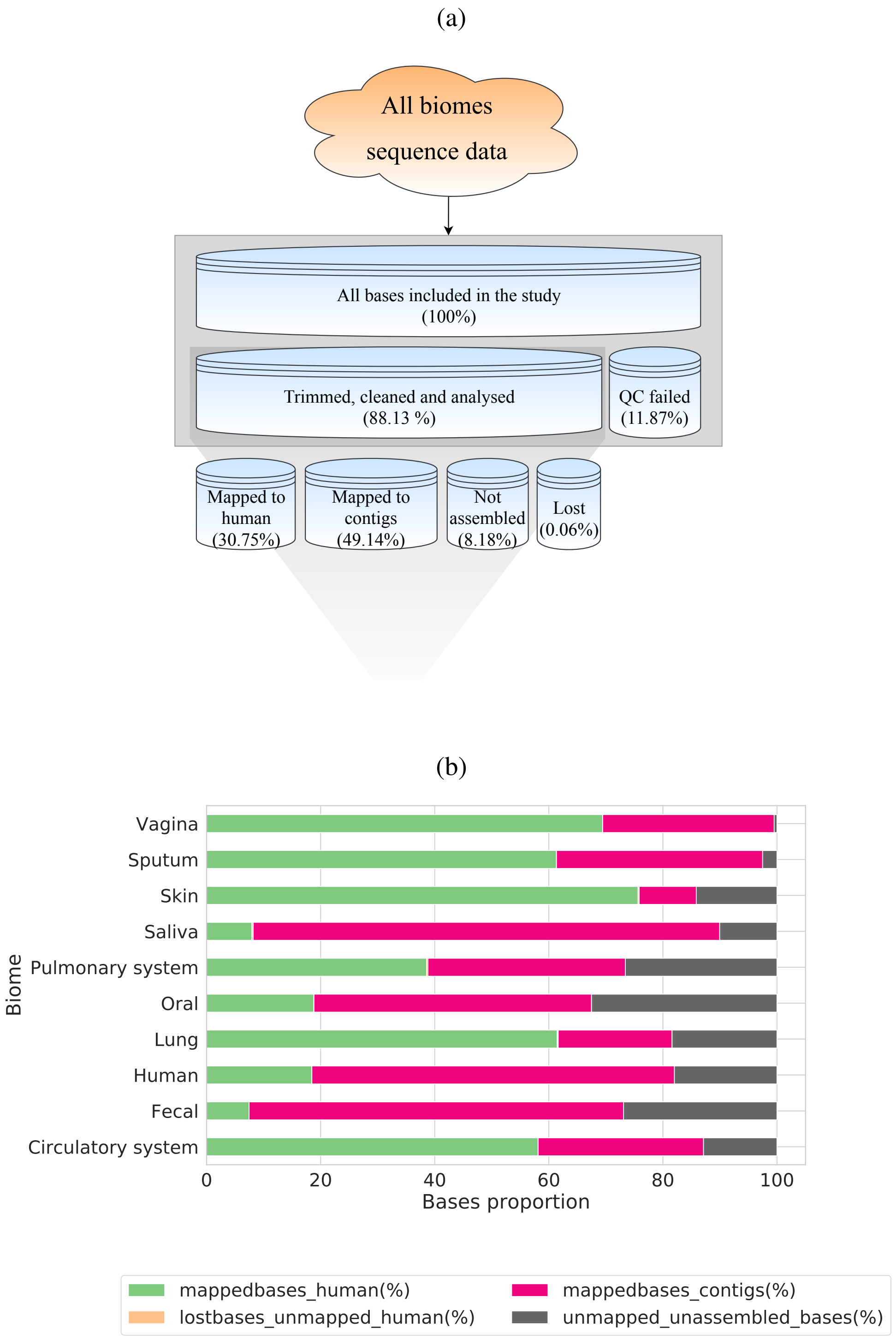


Figure 3.3: An overview of all bases analysed and categorised in this study. (a) Overall categorisation of all bases included in the study. (b) The proportion of bases mapped to humans, assembled contigs and bases that were not assembled in different microbiomes

were quantified as shown in figure 3.3(b) - grey bars. Our quantification suggests that almost all microbiomes have some proportion of such unassembled sequences which ranges from 0.03-98.83% depending on the sample with an average of 23.9% (std: 26.59%). Overall, 8.18% of all data fell into this category as described in figure 3.3(a).

This measure can potentially help to define how easy it may be to assemble a certain microbiome. It can also provide a measure of the quality of the sequenced nucleic acid. Oral, fecal and pulmonary system microbiomes had the highest proportion of such unassembled bases ranging from 26.5 to 32.5%. 68 samples from the BioProject PRJEB17784 contained 87.7-98.8% of such unassembled bases. This BioProject contains the samples of fecal microbiome in L-DOPA naive Parkinson's disease. Ancient dental calculus from skeletons from the Radcliff hospital burial ground samples also contain a large proportion of unassembled bases for a range of samples. This oral microbiome is represented with BioProject PRJEB15334. The mean value for such unknown was 51% for this study with unassembled ranging from 9-97%. At least 50% of samples from this study contained 50% bases that could not be assembled. Due to the type of samples included in this project, it is very likely that these samples contained deteriorated or damaged DNA which could lead to poor quality sequences that could not be assembled.

3.4.3 Defining known, partially known and unknown matter

A total of 44,238,374 contigs were generated and 28,505,777 of them were longer than 300 nucleotides. These contigs were submitted to downstream analysis for classification. In order to bin the contigs into these three major categories, sequence similarity thresholds were used. The identity threshold that defines the highest percent identity for a set of aligned segments to the same subject sequence was used to categorise the known sequences.

All BLASTX hits were grouped and the contigs with >80% protein sequence identity were classified as contigs with 'known' taxonomic origin. In total, 25,148,829 (88.22%) contigs were classified as known contigs in this study. A threshold of 80% was selected based on the prior knowledge of protein homology as it is anticipated that when two proteins have conserved active sites, and share and more than 80% similarity, they have similar functions (Pearson, 2013). In contrast, it is more difficult to make such an argument at a much greater evolutionary distance suggesting that this threshold should distinguish between functionally known sequences and unknown sequences. BLAST/DIAMOND hits were not filtered for the query coverage and all hits were categorised solely based on the percent identity criteria.

Partially known sequences were categorised based on the protein sequence similarity cut-off of >0 and <=80%. 2,517,700 contigs that matched these criteria were classified as partially known. 2,517,700 (8.83%) of all analysed contigs were grouped into this category.

This study systematically measured the proportion of biological sequences that cannot be labelled taxonomically for all microbiomes as they did not have any sequence similarity to known sequences. Overall, 651,529 (2.29%) contigs could not be mapped to a known taxonomic group of organisms using our approach and were categorised as unknown.

Table 3.3: An overall summary of assembled contigs categories defined in this analysis. Total number of contigs in each category is shown in the table below.

Contig category	**Number of contigs**
All	44,238,374
Analysed (>=300 nucleotide)	28,505,777
Known (protein identity >=80)	25,148,829 (88.22%)
Partially known (protein identity <80)	2,517,700 (8.83%)
Unknown (no similarity to any existing sequence)	651,529 (2.29%)
Contigs with BLASTN hits (no BLASTX hits)	187,671 (0.66%)
LCA taxon could not be determined despite DIAMOND/BLASTN hits	75

In total, 25,148,829 (88.22%) contigs were classified as known contigs whilst 2,517,700 (8.83%) of all analysed contigs were classified as partially known. The remaining sequences, referred to as unknown contigs (UCs), are sequences that did not bear significant similarity to known sequences in the databases. Overall, 651,529 (2.29%) of contigs did not match any currently known sequences using our approach and were categorised as UCs. On average 1.3% of assembled bases per sample were found to be unknown. The proportion of unknown varied significantly between different assembled metagenomes as shown in figure 3.5(a). Samples from some microbiomes such as the circulatory system did not contain any unknown sequences compared to the skin microbiome where this proportion was up to 25.85% for some samples.

Box 3.4.3: The definition and categorisation of reads and contigs

- **Unmapped unassembled reads**: Read that did not map to the human genome and *de novo* assembled contigs
- **Unknown contigs (UCs)**: An assembled sequence that could not be labelled taxonomically and/or functionally
- **Partially known sequence**: Assembled sequence with DIAMOND BLASTX hits with <80% sequence similarity
- **Partially known virus sequence**: A partially known sequence that matches exclusively to virus protein sequences
- **Known sequence**: Assembled sequence with DIAMOND BLASTX hits with 80% or higher sequence similarity
- **Known virus sequence**: A known sequence that matches exclusively to virus protein sequences

3.4.4 Quantification of unknown

The UCs varied largely in length and most of the UCs were 300-1000 nucleotides long (figure 3.5(b)). 95.36% (n=621,302) of all UCs were shorter than 1kb and 4.59% (n=29,879) UCs were between 1-5kb long. A set of 320 UCs fell within the 5-10kb length category and 28 UCs were >10kb long. The largest UCs were 42.3kb long and the second largest UCs were 21.3kb long. A complete distribution of UCs across different microbiomes is shown in the figure 3.4 that shows that the largest UCs were assembled from fecal, oral and saliva microbiomes.

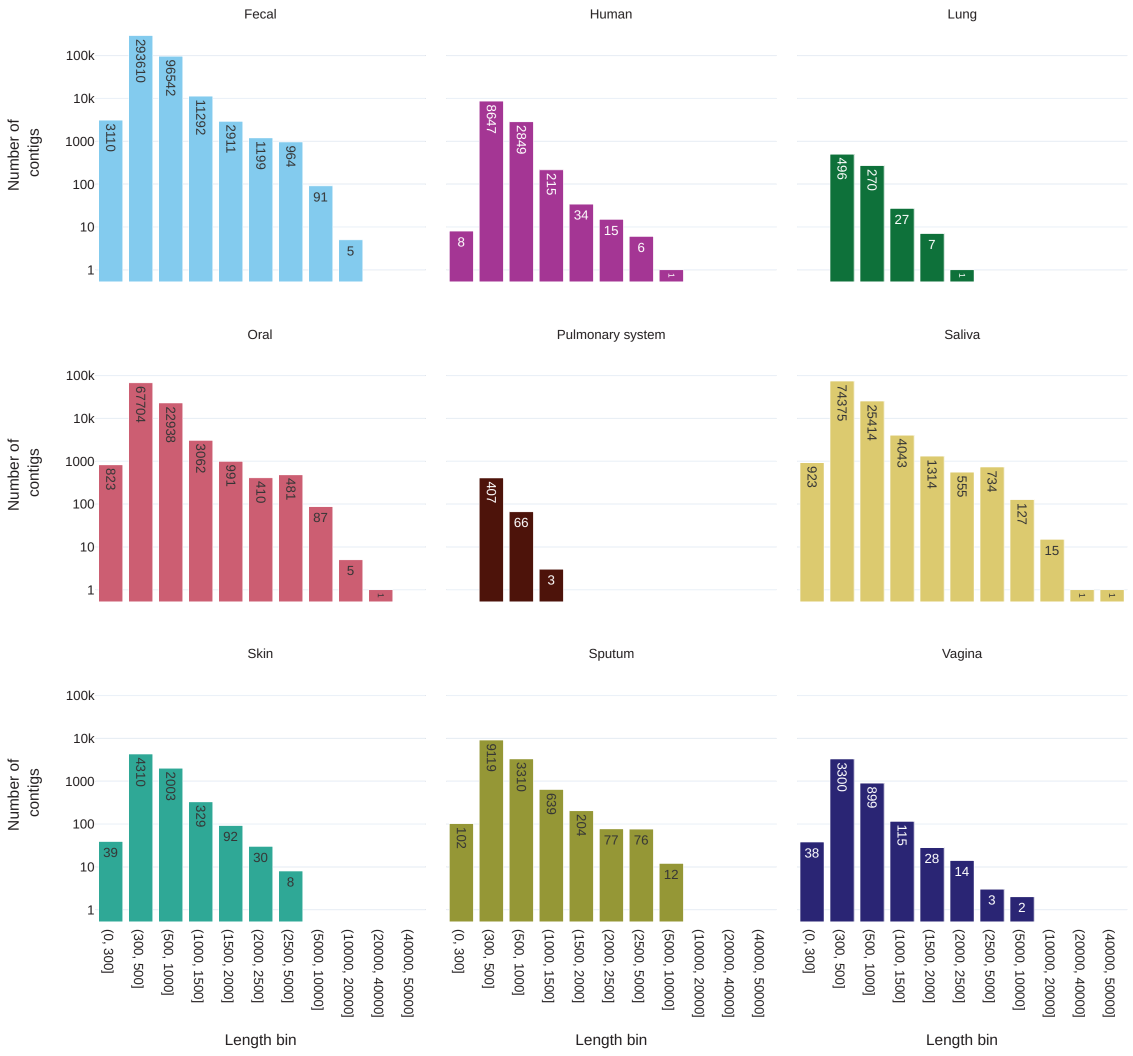


Figure 3.4: A detailed distribution of unknown contigs across all microbiomes where each microbiome is represented by a subplot in the faceted plot. The X-axis shows the interval for the length bin and the Y-axis shows the number of contigs in each interval category. Each bar is annotated with the total number of contigs corresponding to the interval on the X-axis. An ambiguous microbiome ‘Human’ was included in this dataset that represents 3 distinct studies including PRJEB14301 (CSF, n=1), PRJEB21827 (A/B testing for colon model, n=12) and PRJEB6045 (metagenomics of medieval human remains from Sardinia, n=1).

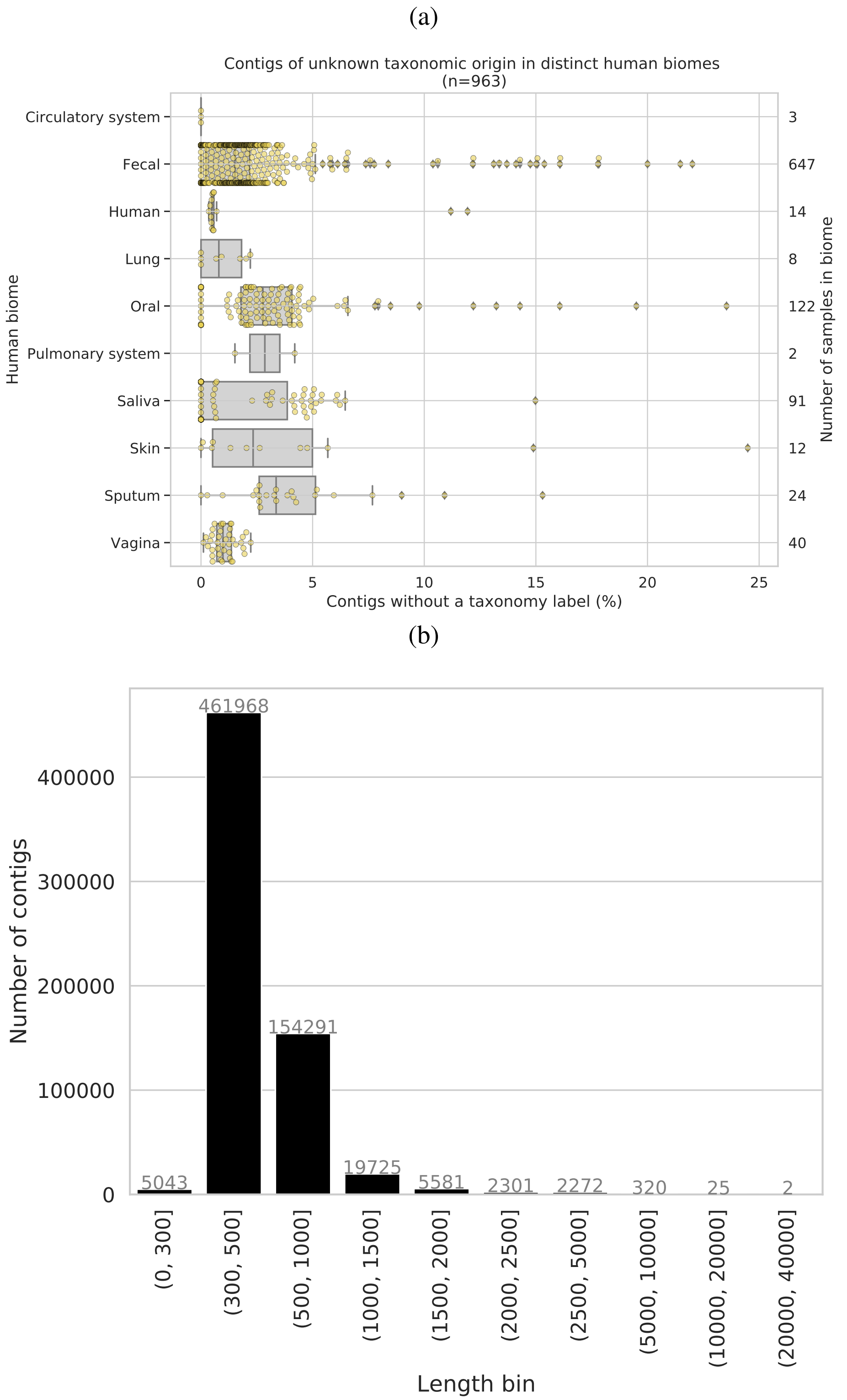


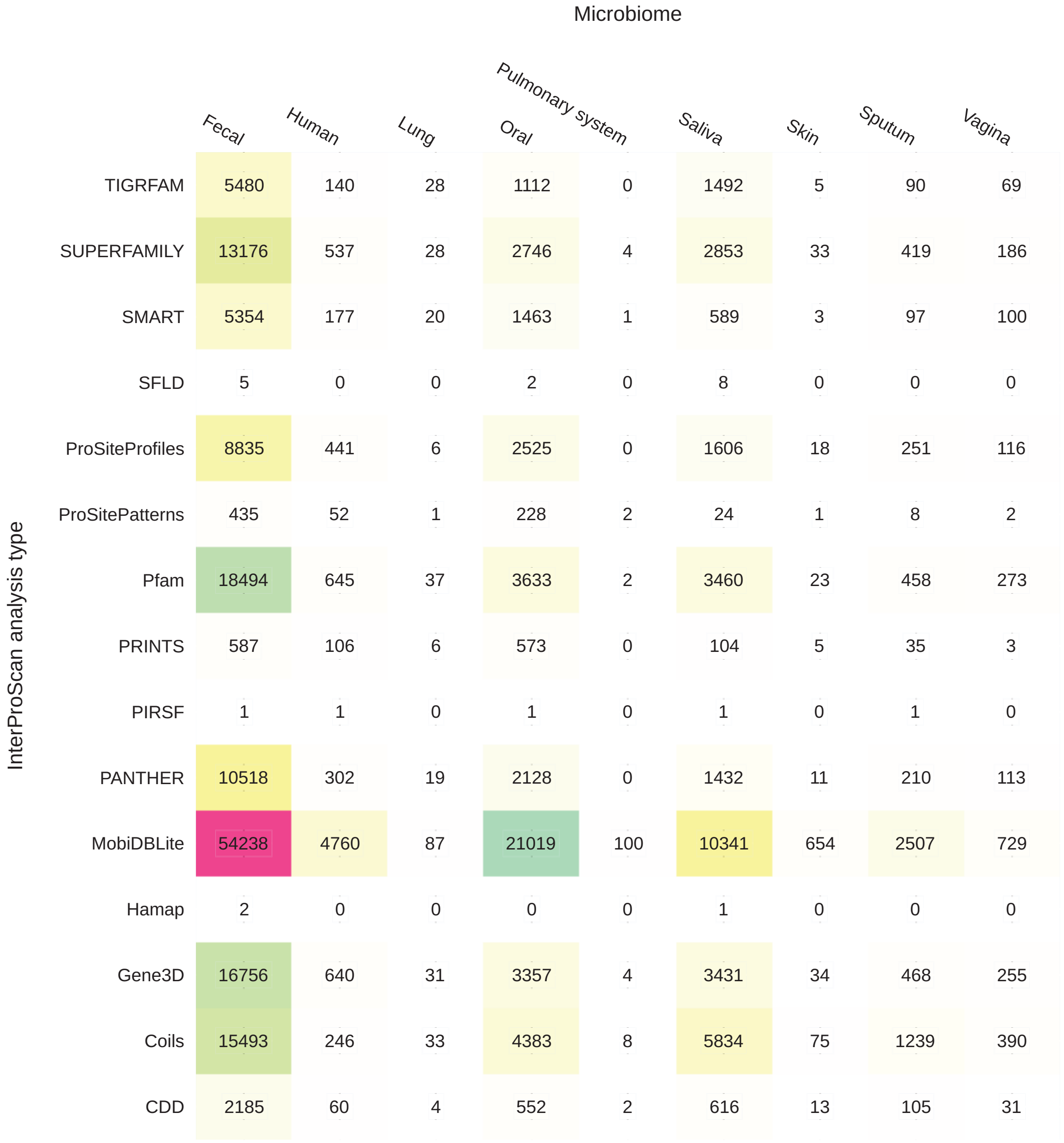
Figure 3.5: Quantification of unknown contigs (UCs) in different human microbiomes. (a) The proportion of UCs in different human microbiomes. The distribution is shown on the X-axis with each microbiome represented on the left-hand side Y-axis. Y-axis on the right-hand side shows the number of samples in each microbiome and corresponds to the number of dots on the plot for the given microbiome. (b) Distribution of contig lengths for all UCs. The X-axis shows length intervals and the number of contigs is shown on the Y-axis and is annotated on top of the bar.

3.4.5 Coding potential of the unknown

To understand the coding potential of the unknown sequences, open reading frames (ORFs) were predicted. 273,590 ORFs that were at least 100 amino acids in length were generated using the standard genetic code. A threshold of 100 AA was selected, this is similar to that used in the taxonomic classification tool GRAViTy which demonstrated only a 5-10% gene loss at this cutoff for viral sequences (Aiewsakun et al., 2018). These ORFs originated from 215,985 distinct UC, showing that 33.15% of all UCs contained large ORFs. On average, ORFs were 157 amino acid (AA) long with a standard deviation of 87 AA residues. The longest ORF was 6,898 AA long (figure 3.10(a)). This set also included 2,713 ORFs with lengths of at least 500 AA and 256 that were at least 1000 AA long.

A detailed protein domain analysis for these ORFs was carried out using the InterProScan (Mitchell et al., 2018a) protein analysis software. As a database, InterPro combines information about proteins' function from several databases, giving an overview of which families proteins belong to, and what domains and sites they contain. The InterProScan software package allows users to run scanning algorithms directly from the InterPro database on novel nucleotide or protein sequences. InterProScan searches the domain and functional signature of amino acid sequences against a range of distinct domain databases including Pfam (El-Gebali et al., 2018), CDD (S. Lu et al., 2019) and SUPERFAMILY (Gough et al., 2001). 36,354 ORFs originating from 35,760 UCs could be functionally annotated using the InterProScan analyses, this number excludes hits to MobiDBLite and Coils databases as they predict disordered regions and coils structure of predicted ORFs as opposed to the domain signatures. An overview of the number of hits found to various InterProScan databases for each microbiome is shown in the figure 3.6.

The highest number of hits were found in the MobiDBlite (Necci et al., 2017) - a database that can predict the intrinsic disorder regions in the proteins. Overall, 5.49% of UCs (n=35,760) contained ORFs (n=36,354) with at least one identifiable domain. The functional classification of the ORFs was prominently centred around the Pfam database resource (El-Gebali et al., 2018). Pfam databases facilitate the domain-based searches against the set of protein sequences using profile hidden Markov models (HMMs). These types of searches can identify distantly related protein sequences. Individual Pfam hits were treated as independent entities and overlapping hits were not consolidated based on the subject/Pfam entries. 16,839 ORFs originating from 16,705 UCs were found to match at least one Pfam entry and in total, 27,025 Pfam hits were derived (figure 3.6). All Pfam entries were collapsed down to their corresponding protein clans (grouping of related protein families) by mapping the Pfam IDs back to their clan membership. Figure 3.7 shows a heatmap of the top 50 Pfam clans with hits to UCs ORFs predicted in different metagenomes. The most abundant hits were identified to clans tetratrico peptide repeat superfamily and leucine-rich repeats. The largest number of hits was found in the fecal microbiome due to the high number of fecal microbiomes included in this study. Additionally, a range of other protein clans including those that represent Helix-turn-helix, beta-strands,



	Fecal	Human	Lung	Oral	Pulmonary system	Saliva	Skin	Sputum	Vagina
TIGRFAM	5480	140	28	1112	0	1492	5	90	69
SUPERFAMILY	13176	537	28	2746	4	2853	33	419	186
SMART	5354	177	20	1463	1	589	3	97	100
SFLD	5	0	0	2	0	8	0	0	0
ProSiteProfiles	8835	441	6	2525	0	1606	18	251	116
ProSitePatterns	435	52	1	228	2	24	1	8	2
Pfam	18494	645	37	3633	2	3460	23	458	273
PRINTS	587	106	6	573	0	104	5	35	3
PIRSF	1	1	0	1	0	1	0	1	0
PANTHER	10518	302	19	2128	0	1432	11	210	113
MobiDBLite	54238	4760	87	21019	100	10341	654	2507	729
Hamap	2	0	0	0	0	1	0	0	0
Gene3D	16756	640	31	3357	4	3431	34	468	255
Coils	15493	246	33	4383	8	5834	75	1239	390
CDD	2185	60	4	552	2	616	13	105	31

Figure 3.6: Functional homologues were identified in the InterProScan analyses for UCs generated for each microbiome. Darker colours represent the higher number of hits to specific Pfam clans.

polymerase and nuclease proteins were also found in this set. These results illustrate that the UCs sequences have known protein domains suggesting that these unknown sequences are functional and belong to organisms that are not yet fully sequenced or taxonomically classified. This step of the analysis looks at the UCs at protein resolution. These results are intriguing and, provide more insights into these taxonomically uncharacterised sequences. This also adds a new layer of information associated with the unknown sequences and sheds light on the dark sequences from a functional classification perspective.

3.4.6 Unknown sequence clustering

To investigate the extent of sequence diversity and to identify UCs sequences present in multiple samples and microbiomes, sequence clustering was performed. Clustering analysis was carried out using MMSeqs2 (Steinegger et al., 2017; Steinegger et al., 2018) to group the unknown sequences based on the sequence similarity and coverage. All sequences with at least 90% sequence identity and 80% overlap were clustered using the MMSeqs2's linclust algorithm (Steinegger et al., 2018). A unidirectional clustering was performed with respect to target sequence coverage using parameter `cov_mode 1`. In brief, linclust is a shared k-mer alignment-based approach where only sequences that share a minimum number of k-mers are aligned, and the longest sequence is set as the cluster representative. Consequently, sequences with shared k-mers are aligned to the cluster representative and sequences that pass the specified clustering criteria are clustered together. This approach means that if the sequences are missed (i.e. false negative), then too many clusters are generated as a result.

MMSeqs2 (Steinegger et al., 2018) generated 464,181 clusters of which 377,855 were singletons i.e. did not cluster with any other sequences. These singletons were excluded from the cluster analysis described below. 86,326 clusters comprised two or more sequences with a mean cluster size of 5.7 contigs and a standard deviation of 8.1. Cluster representatives were extracted from MMSeq's clustering output which are the longest sequences in the cluster. The largest cluster contained 153 sequences which originated from the fecal microbiome from 8 distinct BioProjects (figure 3.10(c)(c)). A cluster size distribution across different microbiomes is shown in figure 3.8 and a detailed cluster size distribution with cluster representative length is shown in the figure 3.9). 89.42% of 273,674 UCs (n=244,730) were clustered into single microbiome clusters, 10.58% UCs (n=28,944) were found in clusters that contained sequences from two or more microbiomes. To compare that with specific studies, 39.4% UCs were clustered into BioProject-specific clusters and the remaining 60.6% UCs (n=165,851) were grouped into clusters originating from two or more BioProjects. 78,139 (90.52%) clusters contained sequences from a single microbiome and 7,645 (8.86%) clusters included sequences from two microbiomes. Only a few clusters were comprised of members from 3 (n=512) or 4 (n=30) microbiomes. The largest multi-microbiome cluster contained 57 sequences (304-9,080 bases long) from 4 distinct microbiomes and BioProjects and contigs assembled from 12 samples. The largest single

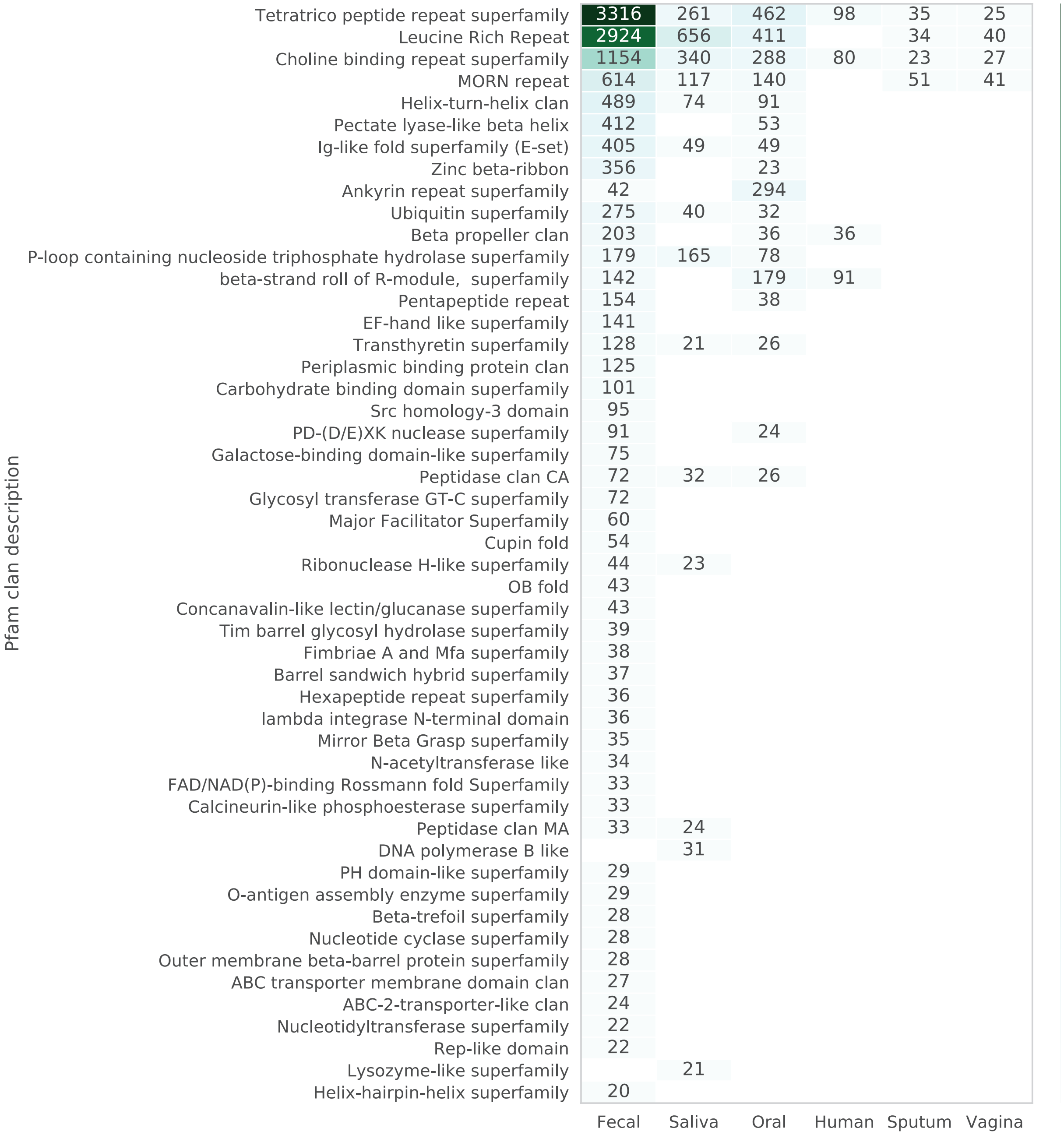


Figure 3.7: The heatmap shows the UCs mapped to various Pfam clans found in the different microbiomes. The darker shades represent the larger number of UCs and lighter shades of the colour represent a smaller number of UCs that are also annotated in the boxes of the heatmap plot here. The protein clans are shown on the Y-axis, human microbiomes are shown on the X-axis and the number of UCs in corresponding categories are annotated on the heatmap.

microbiome cluster contained 153 sequences (6,640-300 bases long) from fecal microbiomes with contigs assembled from 46 distinct samples covering 8 different studies. Overall, this clustering method produced very small, study-specific clusters. A set of 464,181 UCs was obtained by combining the cluster representative sequences with the unclustered singleton UCs and used to determine the rate at which UCs are classified.

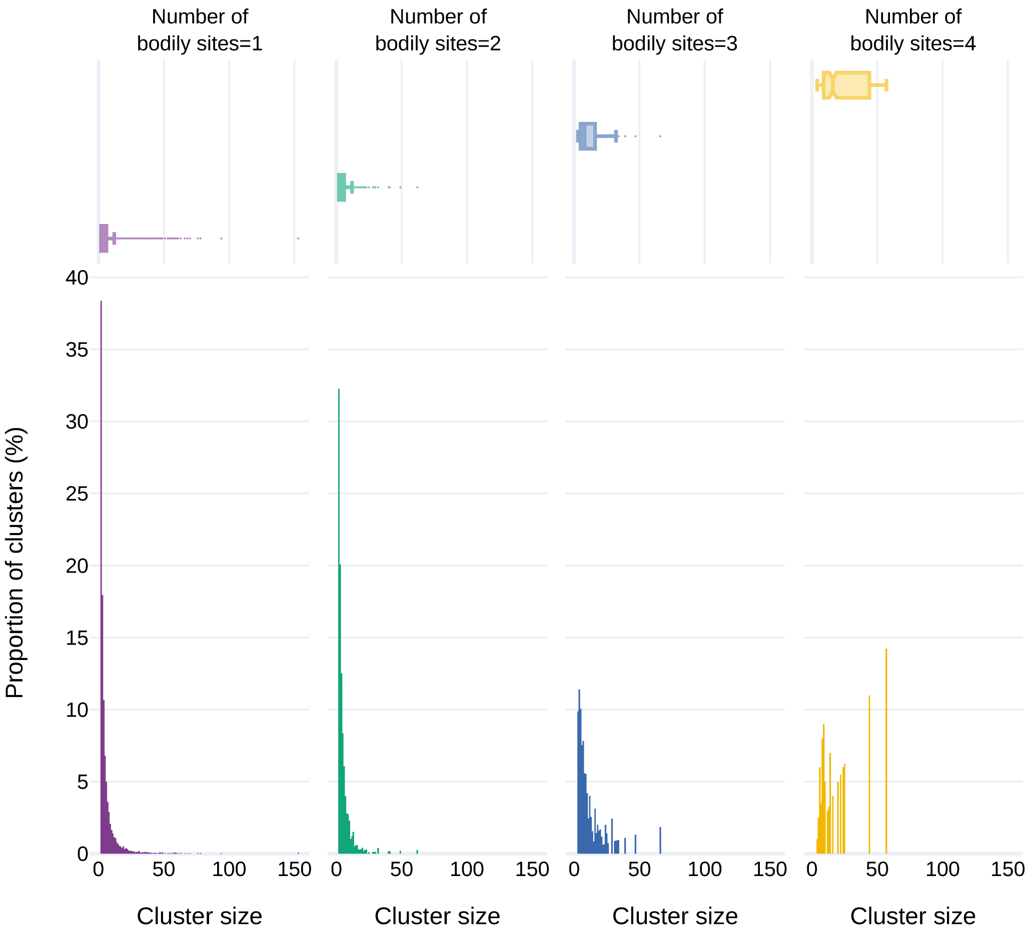


Figure 3.8: MMSeq cluster analysis results and the distribution of UC clusters identified in the different microbiomes. Distribution of cluster sizes on the X-axis and their proportion on the Y-axis. The marginal box plot shows the distribution of cluster sizes for each category. The plots are grouped and coloured according to the number of distinct bodily sites the clusters are found in; e.g. Number of bodily sites = 2 in green, means that members of each cluster are found in data sets from two distinct bodily sites (e.g. gut, skin, fecal, oral), all clusters from this plot come from 2 distinct bodily sites, but may (or may not) come from different bodily sites compared to other clusters within the plot, with one cluster coming from gut and skin, for example, and another from the skin and fecal etc.

Given the novelty of the unknown sequences, it is very difficult to determine the actual number of clusters present in this set. As the unknown dataset contain very large numbers of sequences of variable length, a lot of standard clustering tools such as cd-hit and uclust that employ a greedy algorithm for clustering take too long and are deemed inadequate for this task.

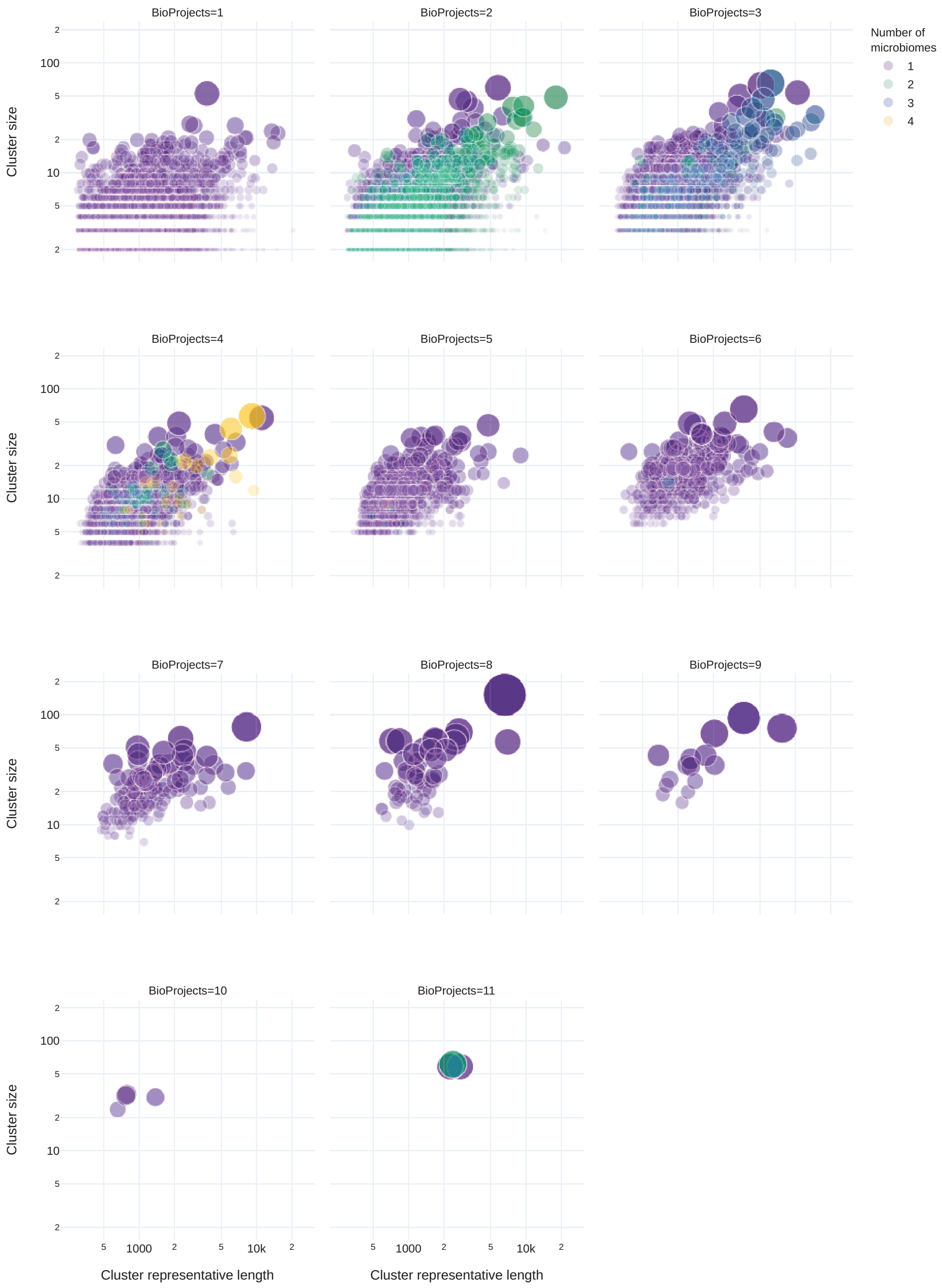


Figure 3.9: Overview of clusters found in the unknown sequence dataset. A number of microbiomes included in distinct clusters are represented by the columns and the number of BioProjects is represented by rows in the facetted plot. For each subplot, the X-axis represent the length of the cluster representative sequence and the Y-axis represent the cluster size for all cluster of size >=2. The size of the bubbles corresponds to the cluster sizes.

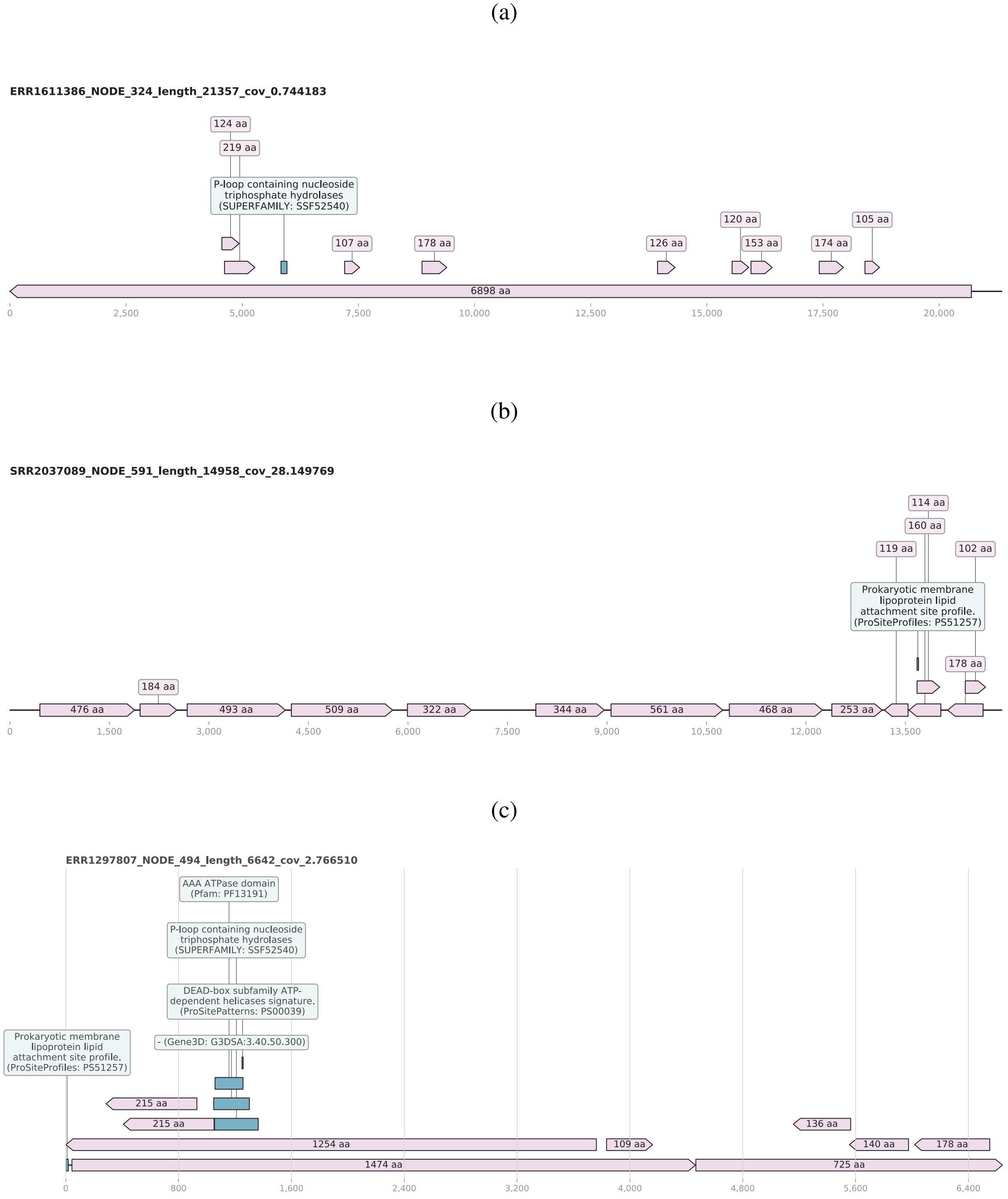

(a)
ERR1611386_NODE_324_length_21357_cov_0.744183
124 aa
219 aa
P-loop containing nucleoside triphosphate hydrolases (SUPERFAMILY: SSF52540)
107 aa
178 aa
126 aa
120 aa
153 aa
174 aa
105 aa
6898 aa
0
2,500
5,000
7,500
10,000
12,500
15,000
17,500
20,000
(b)
SRR2037089_NODE_591_length_14958_cov_28.149769
114 aa
160 aa
119 aa
102 aa
Prokaryotic membrane lipoprotein lipid attachment site profile. (ProSiteProfiles: PS51257)
178 aa
184 aa
476 aa
493 aa
509 aa
322 aa
344 aa
561 aa
468 aa
253 aa
0
1,500
3,000
4,500
6,000
7,500
9,000
10,500
12,000
13,500
(c)
ERR1297807_NODE_494_length_6642_cov_2.766510
AAA ATPase domain (Pfam: PF13191)
P-loop containing nucleoside triphosphate hydrolases (SUPERFAMILY: SSF52540)
DEAD-box subfamily ATP-dependent helicases signature. (ProSitePatterns: PS00039)
- (Gene3D: G3DSA:3.40.50.300)
Prokaryotic membrane lipoprotein lipid attachment site profile. (ProSiteProfiles: PS51257)
215 aa
215 aa
136 aa
1254 aa
109 aa
140 aa
178 aa
1474 aa
725 aa
0
800
1,600
2,400
3,200
4,000
4,800
5,600
6,400

Figure 3.10: The genome diagrams of large unknown contigs (UCs) show the open reading frames (ORFs) in the light pink shade with the ORFs lengths as their corresponding labels and the green boxes illustrating the InterProScan predicted presence of domain signature. (a) Cluster representative of the largest cluster comprising UCs across multiple microbiomes. (b) UC with the largest predicted ORF (6,898 AA). (c) Cluster representative of the largest single microbiome cluster derived from fecal microbiome. This cluster comprised of 153 unknown contigs assembled from 46 samples across 8 different studies.

3.4.7 Classification of the unknown over time

In this framework, the unknown sequence identification is dependent on the publicly available nucleotide or protein sequence databases. These data repositories are updated regularly with new sequence data being deposited from around the world. However, typically, the sequence searches are carried out against static versions of the databases. Our analysis conducted against the databases downloaded on 18 April 2019 identified 651,529 UCs that collapsed down to a set of 464,181 UCs following the cluster analysis. Subsequent analyses on 31 October 2019 and 5 March 2020 produced a set of 613,726 and 558,711 UCs respectively. The final number of sequences that still lacked a taxonomy label was down to 459,147 after the most recent analysis carried out against the databases downloaded on 14 October 2020. 29.5% (n=192,382) of the sequences compared to the initial set of unknowns matched at least one sequence from the updated databases in the BLASTX and the BLASTN steps of the analysis. Similarly, 27.6% (n=128,288) of the representative set sequences could be labelled taxonomically with the updated databases. A rate of taxonomic characterisation of 1.64% of unknown sequences being characterised per month was calculated from the complete set. This rate was estimated to be 1.54% for the representative set. Moreover, as shown in the figure 3.11, a range of long UCs still remained unknown even after the similarity sequence-based analysis carried out on 14 Oct 2020.

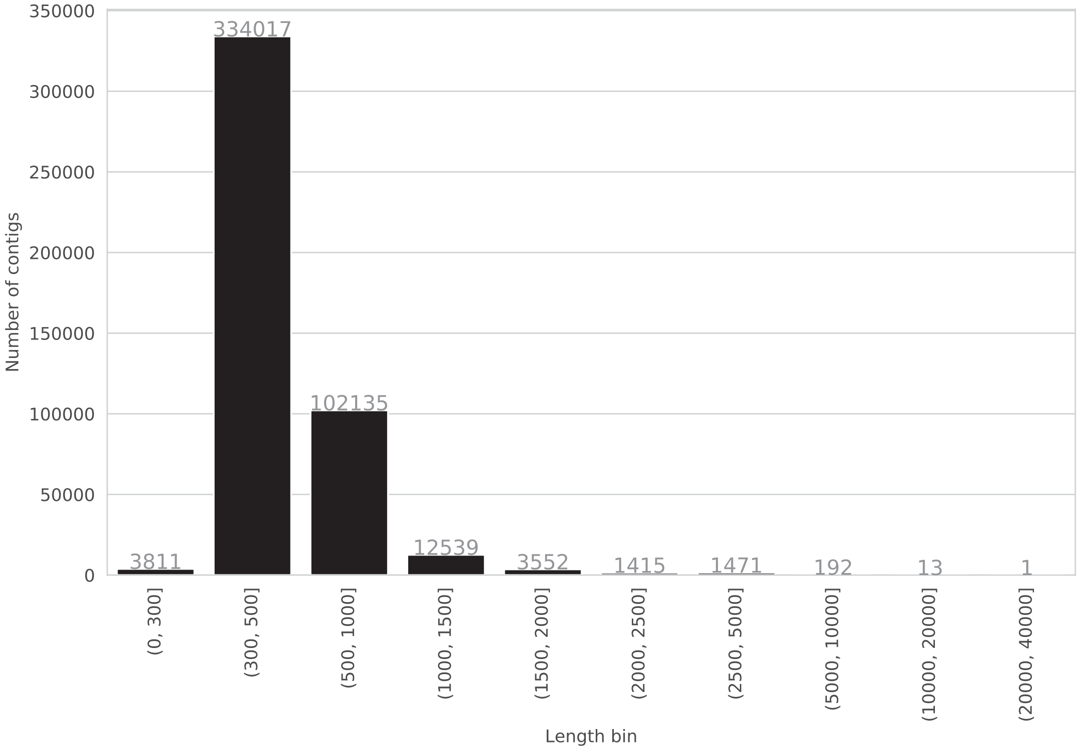


Figure 3.11: Distribution of contig lengths for all unknown contigs after the final time point (14 Oct 2020).

From a set of 192,382 contigs that were labelled taxonomically after the most recent analyses carried out on 14 Oct 2020, 167,864 were identified using BLASTX and 24,518 were identified using BLASTN. 106,739 UCs from the BLASTX classified set were categorised as known and 61,125 contigs were categorised as partially known. A large majority of these contigs (97.11%, n=162,987) were also deemed to be bacterial. The remaining contigs were divided between cellular organisms (n=2,104), archaea (n=930), viruses (n=858), root (n=827) and Eukaryota (n=140). 76.55% of all BLASTN hits were matching to bacteria (n=18,768), 17.88% matched to viruses (n=4,383), 1.99% matched to Eukaryota (n=487) and 0.03% archaea (n=7). The hits that could not be mapped to a superkingdom were divided between unidentified plasmid (n=544), root (n=294), cellular organisms (n=20), and uncultured organisms (n=14) and synthetic construct (n=1). These results reiterate our initial hypothesis that the majority of UCs represent currently unknown microbial genomes.

3.4.8 Viral domain signature identification

195 UCs were shown to contain a virus-specific functional domain which was parsed using the term ‘virus’ or ‘viral’ in the InterProScan analysis signature description column. Results with the term ‘phage’ were not included in this subset as a range of phage domains are also present in the host bacterial genomes. These domains were predominantly identified using the Pfam (n=125) analysis. The most abundant virus-specific domain was Vaccinia Virus protein VP39 and it was found in 53 UCs derived from fecal (n=23), saliva (n=14), oral (n=12), sputum (n=1) and human (ambiguous; n=3) microbiomes and it was identified by Gene3D analysis. The largest UCs containing this domain were 3,661 bases long and were found in sample ERR1474567. Another frequently found domain in the UCs was the podovirus DNA encapsidation protein Gp16 domain. It was found in 25 UC, out of this set 23 UCs were assembled from fecal microbiome. The largest UCs containing this virus-specific domain was 9kb long contig shown in figure 3.15(b), assembled from PRJEB18265. These UCs were clustered with 24 other sequences (See section 3.4.6) that were assembled from 11 samples representing 5 distinct fecal microbiome studies. These results indicate that these UCs represent a completely novel genome of a virus that is likely related to currently known podoviruses.

The largest UCs containing a viral RNA dependent RNA polymerase (Pfam: PF00680) domain was found in the sputum microbiome sample ERR1022511. This UC was 5,894 bases long and contained seven ORFs that were at least 100 AA long (figure 3.12). A 269 AA long ORF contained ATPase P4 of dsRNA bacteriophage phi-12 (Pfam: PF11602) domain suggesting that these UCs represent the large segment of a novel double-stranded RNA phage which is usually categorised in the virus family *Cystoviridae*. The genomes of these phages are composed of three linear dsRNA segments with a total genome length of 12.7–15kb and all segments code for various proteins (Poranen et al., 2017). Although several other UCs were found in the same sample, none of them displayed any sequence or functional similarity to the other two segments, i.e. small and medium segments of Cystoviruses. However, UCs that could potentially belong to novel cystovirus-like genomes were extracted based on the sequence length, GC content and sequencing depth criteria. Moreover, this UC representing a potentially novel relative cystoviruses did not match any known protein or nucleotide sequences even in the most recent analyses confirming the discovery of a novel virus.

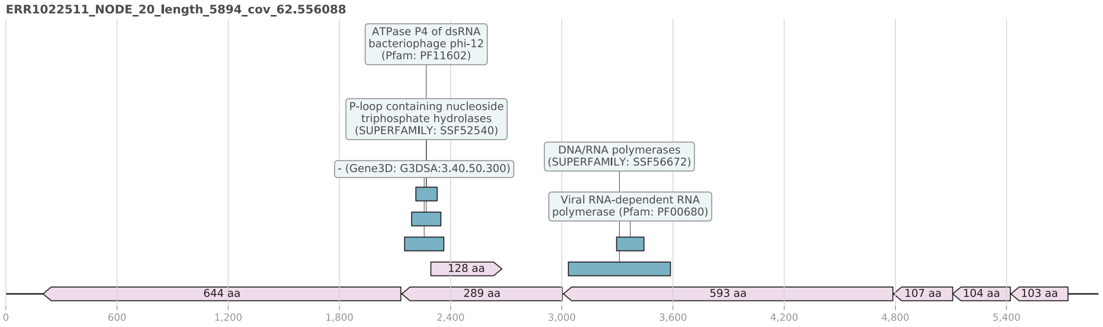

ERR1022511_NODE_20_length_5894_cov_62.556088
ATPase P4 of dsRNA bacteriophage phi-12 (Pfam: PF11602)
P-loop containing nucleoside triphosphate hydrolases (SUPERFAMILY: SSF52540)
- (Gene3D: G3DSA:3.40.50.300)
DNA/RNA polymerases (SUPERFAMILY: SSF56672)
Viral RNA-dependent RNA polymerase (Pfam: PF00680)
128 aa
644 aa
289 aa
593 aa
107 aa
104 aa
103 aa
0
600
1,200
1,800
2,400
3,000
3,600
4,200
4,800
5,400

Figure 3.12: The genome diagrams of a potentially novel dsRNA phage segment found among the UC set that is hypothesised to be related to currently known Cystoviruses. The open reading frames (ORFs) are highlighted in the light pink shade with the ORF lengths as their corresponding labels and the green boxes illustrating the InterProScan computed presence of domain signature.

3.4.9 Virus prediction and comparisons to uncultured virus databases

From the complete set of the UCs, 323,395 (49.64%) UCs were predicted as viruses by DeepVirFinder (see figure 3.13). This set included 300,271 UCs that were under 1kb long which represents 48.33% of UCs identified in this length category. A number of larger contigs were also predicted as viruses: 76.27% (n=22,788) of UCs in the 1-5kb length category and 96.55% (n=336) of UCs in the 5-50kb category. These results strongly support our hypothesis that the large majority of the UCs are of virus origin, albeit a large proportion of short UCs is likely to be fragments of unknown viruses. These results are discussed in detail in section 4.4.7 of Chapter 4.

These predicted virus sequences (n=323,395) were clustered with other known and partially known sequences using MMSeqs with 90% sequence similarity across 80% of the sequence. 50.18% (162,271) of UCs were either singletons or were clustered with other UCs, whilst the remaining 49.82% (161,124) of UCs were clustered with known and partially known. However, a large proportion (n=152,295; 94.52%) of the UCs that clustered with these were shorter than 1kb. 8,829 UCs (out of 22,788; 38.74%) were at least 1kb long among which 1,402 UCs (out of 4,419; 31.73%) were at least 2kb long, 75 UCs (out of 336; 22.32%) were at least 5kb long and 5 UCs (out of 28; 17.86%) were at least 10kb long. Moreover, 47.52% of sequences that match the UCs were deemed partially known (i.e. had a protein sequence hit with <80% sequence similarity) in this analysis suggesting that these known and partially known sequences are still significantly

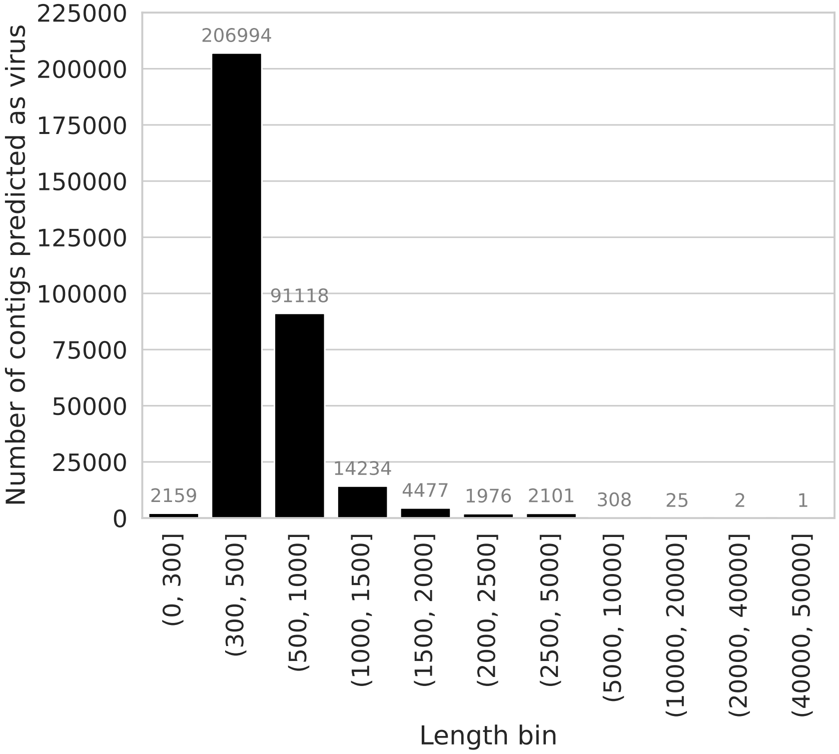


Figure 3.13: Contig length distribution of unknown contigs predicted as virus using DeepVirFinder with qvalue<0.05. 48.33% of UCs that were <1kb long, 76.27% of UCs between 1-5kb long, and 96.55% of UCs that were at least 5kb long were predicted as viruses.

divergent from those present in the databases.

To identify the “known unknowns” i.e. uncultured viruses categorised as UCs in this study and also observed in previous meta-analyses, the IMG/VR databases were used as a reference and the UCs were searched against the nucleotide and protein repositories. 182,293 (27.98% of all UCs) UCs had at least one hit to uncultivated viral genomes (UViGs) included in the IMG/VR using BLASTN and 175,372 (26.92%) UCs were found to match at least one UViGs using the BLASTX approach (figure 3.14). Out of the 273,590 predicted ORF set, 85,852 ORFs were found to match protein sequences included in IMG/VR. 64,779 (9.94%) of UCs were found to match the uncultured viruses in IMG/VR using all three approaches.

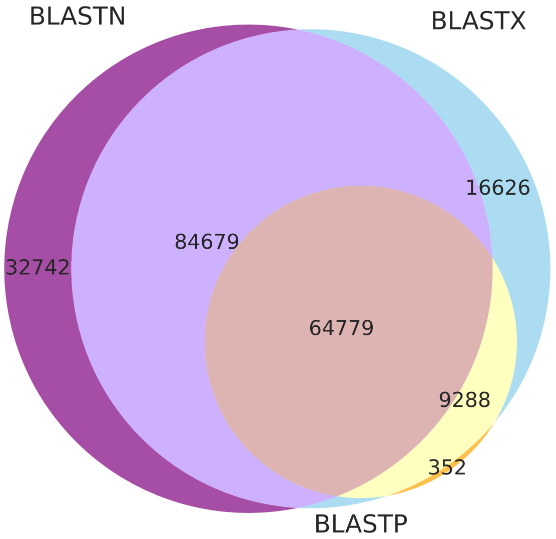


Figure 3.14: A Venn diagram comparing UCs to IMG/VR databases. Out of the complete set of 651,529 UCs, 442,970 UCs did not have a hit to protein and nucleotide sequences included in IMG/VR. 208,559 UCs were matching to IMG/VR sequences using one or more of three different approaches; BLASTN (n=182,293), BLASTP (n=74,512) and BLASTX (n=175,372). Overlapping UCs shared between all different approaches are shown in the Venn diagram here. 64,779 UCs were found to have an IMG/VR hit for BLASTN, BLASTP and BLASTX methods. The largest overlapping UC set was between BLASTN and BLASTX which shared 149,458 UCs whereas the smallest overlap was found between BLASTN and BLASTP which was 64,872 UCs. BLASTP and BLASTX methods shared 74,067 UCs.

3.4.10 The large unknown contigs

All UCs described in this section were predicted to be viruses by DeepVirFinder and did not cluster with known and partially known sequences. The largest UCs were assembled from the saliva sample ERR1474583 and were 42,357 bases long. This contig did not cluster with any other contigs and has 23 ORFs that were over 100 AA long. One of the ORFs that is 434 AA long comprised of the cysteine proteinases domain (SUPERFAMILY: SSF54001) according to the InterProScan analysis. This contig still remained unknown after searches against the most recent version of the databases suggesting that the organism this genomic sequence belongs to is still to be identified and fully sequenced. A snapshot of the ORFs and domain is shown in figure 3.15(a), highlighting the presence of coding regions across the entire length of the UCs sequence. Based on the results we have obtained here, we predict that this UCs sequence is likely to be of microbial origin as it lacks a non-coding region. CheckV analysis predicted it to be a viral genome fragment with the presence of two identifiable viral genes albeit with low quality as per the Minimum Information about an Uncultivated Virus Genome (MIUViG) (Roux et al., 2019) standards due to the lack of similarity to any known sequences. This strongly suggests that this UC can potentially be a representative or partial genome sequence of a currently unknown and completely novel virus.

A 20,309 nucleotide long contig from saliva sample ERR1474612 clustered with two very short contigs from the same study. As shown in figure 3.15(c), long ORFs were predicted

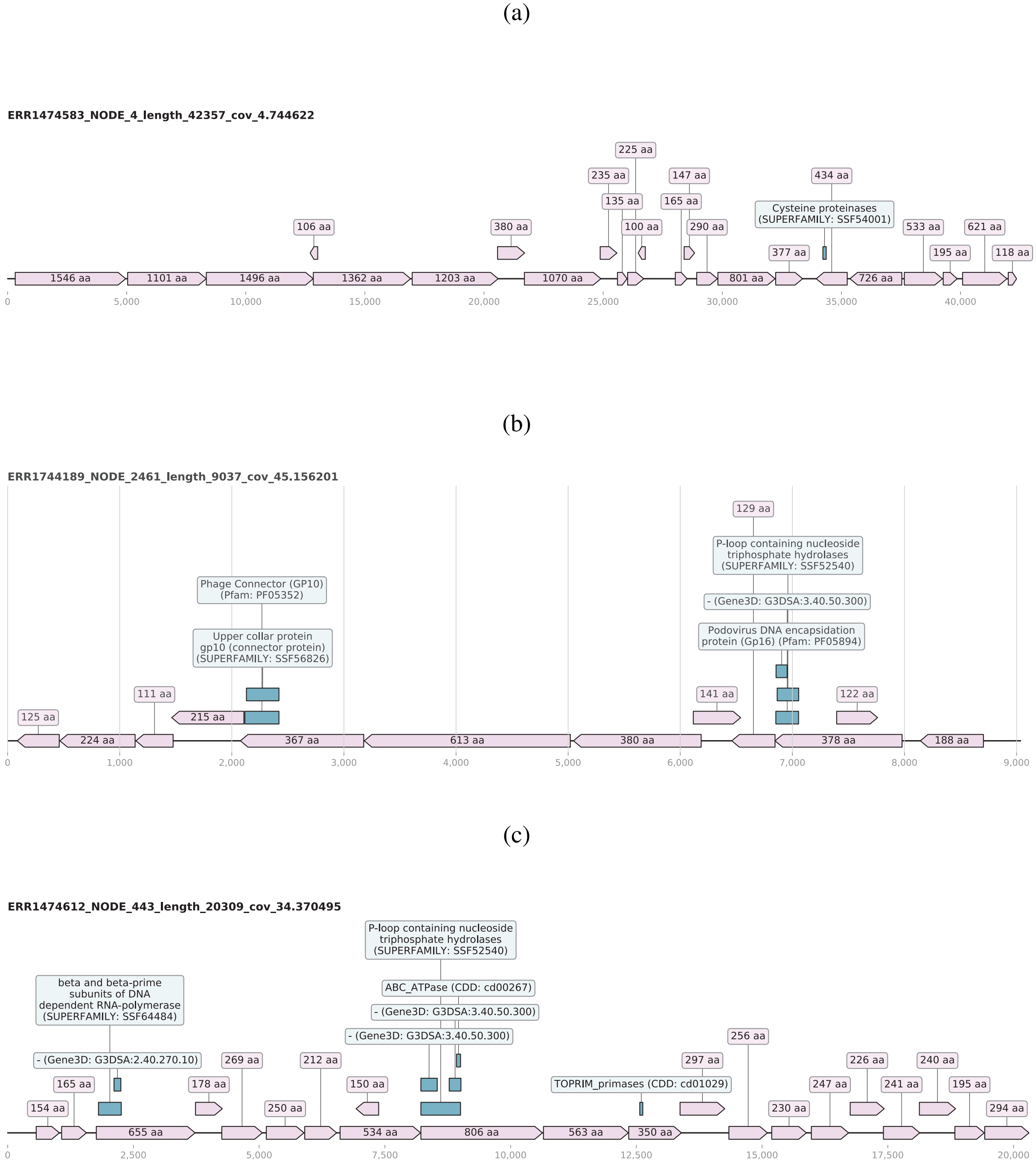


Figure 3.15: The genome diagrams of large unknown contigs show the open reading frames (ORFs) in the light pink shade with the ORFs lengths as their corresponding labels and the green boxes illustrating the InterProScan computed presence of domain signature. (a) The largest unknown contig assembled in the set is categorised as unknown even after the most recent similarity-based search on 14 Oct 2020. (b) The largest contig with podovirus DNA encapsidation protein Gp16 domain. (c) An unknown contig of length 20,309 bases was described to contain a range of domains including a potential virus-specific RNA polymerase domain.

across the whole sequence. Some of the predicted ORFs were found to have interesting domain signatures (figure 3.15(c)) such as enzymes for nucleic acid replication e.g. polymerases. An ORF that is 655 AA long shows the presence of DNA dependent RNA polymerase domain (SUPERFAMILY: SSF64484). A CheckV (Nayfach et al., 2020b) analysis of the contig also predicted it to be of viral genomic origin, however, it was predicted to be an incomplete genome. This UC was shown to have a very low identity (<30% sequence identity with 2% of query coverage) to a hypothetical protein of Firmicutes bacterium (HAB66316.1) and AAA family ATPase from Sharpea azabuensis (23% sequence similarity). When the e-value threshold was removed, a total of 8 BLAST hits were obtained and 3 out of 8 hits were to a range of phages including Bacillus phage vB_BpuM-BpSp, Vibrio phage 2 TSL-2019 and Ralstonia phage RP12. These hits range from hypothetical and putative proteins. All these matches were localised to a short region between 8,217-8,915 which was shown to contain ATPase and P-loop containing nucleotide triphosphate hydrolases domains (figure 3.15(c)). Notably, no nucleotide sequence hits were identified for this UC. Although these results have bacterial hits, it is likely that this UC represents a complete or partial genome of a novel phage that infects the host bacteria e.g. firmicutes.

3.4.11 Short circular contigs

A range of circular contigs with direct terminal repeat (DTR) and inverted terminal repeat (ITR) signatures were identified using CheckV in the UCs data set. A total of 1,839 containing repeat signatures were predicted of which 1,771 contained DTR signatures and 68 contained ITR signatures. 94 of these UCs were at least 1kb long suggesting circular genomes and 48 of them contained a range of 55 bases long terminal repeats. A cluster of 8 sequences from 2 different microbiomes and studies were identified to contain similar sequences (71-100% similarity) assembled from different samples (table 3.4). Four cluster members were 2,110 bases long, one sequence was 1,983 nucleotides long and the cluster representative was 3,165 nucleotides long. The cluster representative sequence contained multiple copies of the same ORFs suggesting the presence of multiple genome copies, sequencing error or mis-assembly. Most of these sequences contained a 50 bp long DTR sequence signature 'GTGCATTTTTTTTGTGCACTTTTTCAAAAAAACCGTGAAAAAAATTCATT'. These contigs contained two distinct ORFs, which were 125 AA and 144 AA long. Similarly another 50 bases long DTR signature 'AATGAATTTTTTTCACGGTTTTTTTGAAAAAGTGCACAAAAAAAATGCAC' was observed in another cluster that had 7 member sequences ranging in similarity from 31 to 100 percent and assembled from 7 distinct samples. All but one member were 1,770-1,771 bases long. These contigs also contained two ORFs that were 102 AA and 106 AA long. These ORFs did not match any existing protein sequences in the databases. These circular contigs were assembled from a range of oral microbiome samples from study PRJNA230363. Similarly, a range of contigs (n=9) that contained Inverted Terminal Repeats (ITR) was also identified in this data set. A cluster of 5 distinct circular contigs assembled from distinct samples from the fecal microbiome (PRJEB7949). Four out of five of these circular contigs contained the ITR sequence 'CGAAACGATTGCCCAGAGAGATGACTGTCAATCCGCCCGATTATTGGGCGCTTAC'. They also contained a 138 AA long ORF. These short circular UCs did not bear any sequence or functional similarity to known sequences or domains so their biological origin is difficult to predict. However, based on their genome organisation and size distribution, it was hypothesised that they are likely to represent either novel circular replication-associated protein (Rep)-encoding single-stranded (CRESS) DNA virus groups or novel satellite virus-like groups. 16 out of 20 UCs described in the table 3.4 were predicted to be viruses by DeepVirFinder (see: Virus prediction and uncultured virus databases).

Table 3.4: Circular contig clusters with direct and inverted terminal repeats

Study ID(s)	**Cluster size**	**Typical contig length for contigs in this cluster**	**Repeat type**	**Sample type**	**Sequence similarity (min-max)**
PRJEB14383; PRJNA230363	8	2110	DTR	Saliva; Oral	71-100
PRJNA230363	7	1771	DTR	Oral	31-100
PRJEB7949	5	1337	ITR	Fecal	67-100

3.4.12 Quantification of partially known contig sequences

All contigs that matched other proteins with less than 80% sequence similarity were clustered into the category of partially known sequences. These partially known sequences make up around 8.83% of all classified contigs. An overview of the proportion of the partially known sequences is shown in the figure 3.16. It is interesting to note the proportion of partially known bases is very high in exposed microbiomes including oral and saliva. Oral microbiomes can harbour up to 80% of such partially known sequences with an average of 25.63% compared to an overall average of 9.75% among all microbiomes. The partially known category represents a very interesting set of sequences in this study as it represents sequences that are distantly related to the currently known sequences available in the databases.

The average length of partially known contigs was around 986 bases with a standard deviation of 3054 bases. The largest contig included in this category was 823,704 long. Among the partially known contigs, 511,977 were at least 1kb long and 45,022 contigs that were >=5kb. This category had the longest contigs with 16,341 contigs being at least 10kb or longer.

Partially known contigs were unevenly distributed among different microbiomes and BioProjects. 13% of all partially known contigs belonged to a study that sequenced the fecal microbiome of patients suffering from IBD and compared this microbiome with the counterparts of control individuals (BioProject PRJEB7949). A second fecal microbiome study (PRJEB12357), looking at the impact of fecal microbiota transplantation on the intestinal microbiome in metabolic syndrome patients, contained 12.81% of all partially known contigs. Two other human fecal microbiome studies, PRJEB8094 and PRJEB6092, comprised over 20% of all partially known contigs. Additionally, another study that compared the fecal metagenome of sickle cell disease patients and the healthy controls comprised 5.13% of partially known contigs. Collectively, these 5 fecal metagenomes contained nearly 50% of all partially known contigs. This result could be reflective of the initial bias in the sampling i.e. high number of fecal microbiome samples included in the data. Two other microbiomes represented by BioProjects PRJEB14383 and PRJEB12831 were the only non-fecal metagenome studies that contained >5% of partially known contigs identified in this analysis. This might be due to the approach employed here. General-purpose, extensive nucleotide and protein databases used here are likely to contain a large number of diverse sequences and the probability of the known and partially known contigs matching to any of these sequences from the databases is very high. Moreover, the first step of the sequence similarity-based search was carried out at the protein level in the UnXplore framework and a lot of new microbial species could bear high similarity at the protein level (often higher than 80% criteria used here) despite originating from different species and genera. This could have skewed some of the partially known contigs results obtained here.

Table 3.5 provides a very brief overview of the partially known contigs classification based on the LCA and taxonomic superkingdom determined for each contig. Bacteria were found to be the dominant superkingdom among the partially known sequences with 2,323,916 contigs exclusively matching bacterial proteins. The average % identity for these hits was 63% with 24% query coverage and a mean alignment length of around 169. The second-largest number of contigs were mapped to the cellular organisms with 94,355 partial hits. The third most prominent group included in this classification was root comprising 69,529 contigs. This group covered the hits that could originate from two or more superkingdoms and the LCA which was determined to be the root of the taxonomic tree. Typically this group could include bacteriophage contigs that often match to both bacteria and virus sequences. 6,576 partially known contigs matched to viruses. These results are summarised in detail in the following section.

Table 3.5: Distribution of partially known sequences across different taxonomy groups

Taxonomy group	**Count**	**%Identity (mean)**	**%Query coverage (mean)**	**Alignment length (mean)**
Archaea	6250	60.72	22.82	152.0
Bacteria	2323916	63.08	23.99	169.17
Eukaryota	16885	59.38	18.46	139.15
Plasmid pTD1	1	70.0	10.22	60.0

Taxonomy group	**Count**	**%Identity (mean)**	**%Query coverage (mean)**	**Alignment length (mean)**
Plasmid pVT736-1	1	71.2	12.52	66.0
Viruses	6571	45.69	23.37	163.89
cellular organisms	94355	61.54	22.17	151.74
root	69529	58.61	20.35	177.71
unclassified Iapetusvirus	5	34.48	25.25	158.2
uncultured marine microorganism HF4000_APKG8K5	2	48.0	30.67	194.0
uncultured microorganism	1	45.0	11.05	60.0
uncultured organism	5	60.8	14.59	68.0
uncultured organism HF70_19B12	1	32.1	30.35	112.0
uncultured prokaryote	164	55.32	17.64	95.3

3.4.13 Partially known contigs matching to viruses in different microbiomes

In total, 6,576 contigs had exclusively virus hits. This category of contigs was identified as those that had DIAMOND BLASTX homologues exclusive to viruses. In order to validate these contigs, they were searched against the entire nucleotide (nt) database using BLASTN. This step should have identified any untranslated sequences from other organisms that match viral proteins by chance. This search showed that these partially known contigs matching exclusively to viruses did not have a hit to a genomic sequence confirming that sequences included in this category were likely to be of virus origin.

Figure 3.17 provides an overview of the length distribution of contigs in this category. Approximately 45% i.e. 2,999 contigs out of 6,576 were 300-500 nucleotide long. These short contigs were most prominent in all microbiomes as shown in figure 3.17(a) and 3.17(b). A subset of contigs that were longer than 1kb and the microbiomes they were found in is shown in the figure 3.17(c). Some of the longest contigs were found predominantly in saliva and oral microbiomes. The fecal microbiome also contained >400 contigs that were longer than 1kb. 31 of these were at least 10kb long and the five longest contigs were >54kb and were found in oral (PRJNA230363) and saliva (PRJEB14383) microbiomes.

The genomic composition of target sequences found to be matching the contigs is shown in the figure 3.18(a). The contigs matching the dsDNA viruses were found in all except the pulmonary system microbiome. The highest numbers of these contigs were in saliva, fecal and oral microbiomes. The second most prominent group of viruses was negative ssDNA viruses

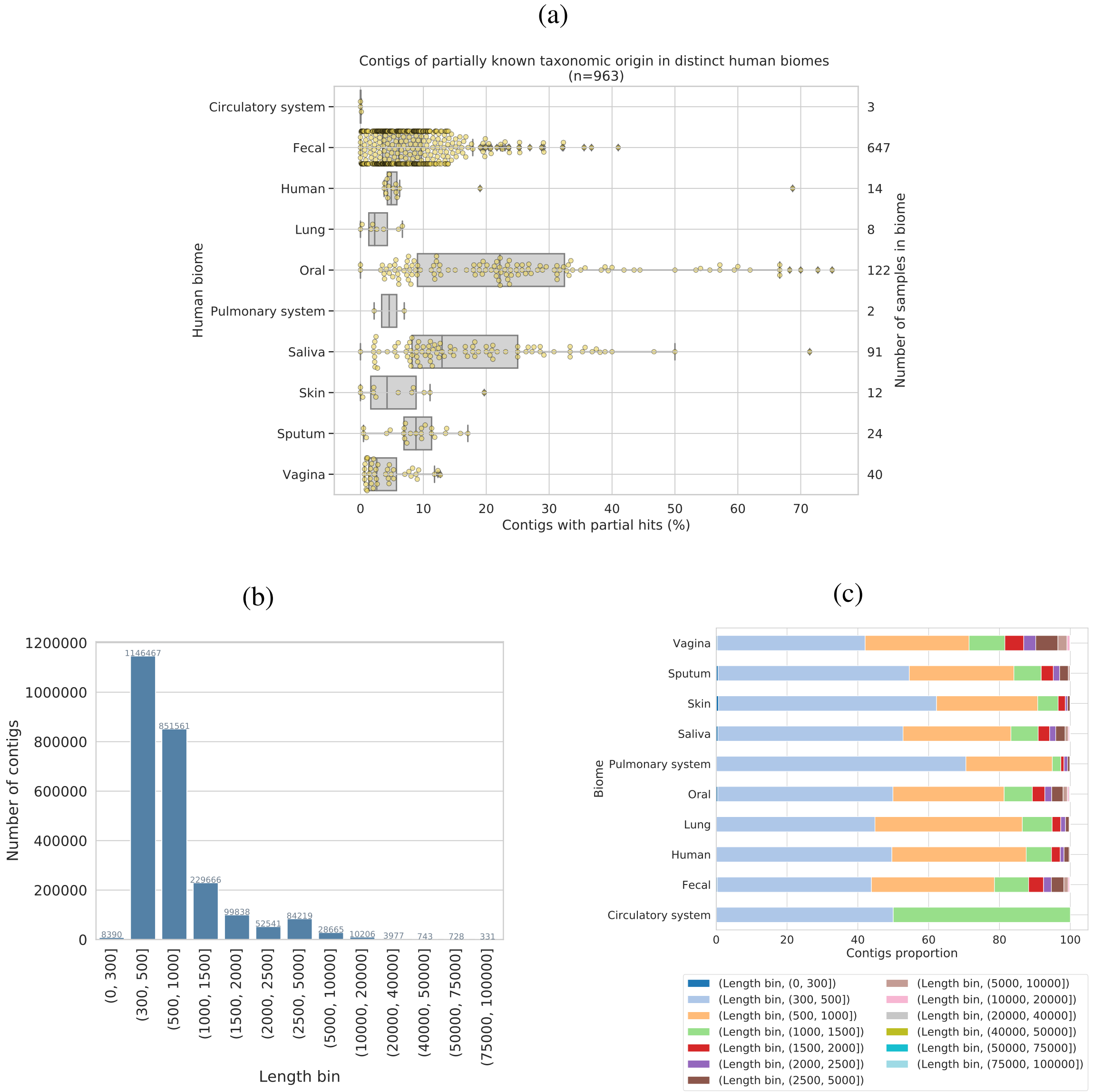


Figure 3.16: Quantification of partially known sequences in different microbiomes. (a) The percentage of partially known contigs in different human microbiomes. The X-axis shows the percentage of partially known contigs for corresponding human microbiomes shown on the Y-axis. The boxplot represents the distribution of partially known contigs and each sample in the study is denoted using a yellow circle. The total number of samples in each category is shown on the secondary Y-axis on the right-hand side. (b) The Length distribution of partially known contigs is shown using a bar plot. The length intervals are shown on the X-axis and the total number of contigs in each length interval is shown on the Y-axis. The actual number of contigs present in each length category are annotated on the top of the bar. (c) A proportion plot is used to visualise the proportion of partially known contigs in different human microbiomes with different length bins. Each interval bin is coloured according to the colour key shown below the plot.

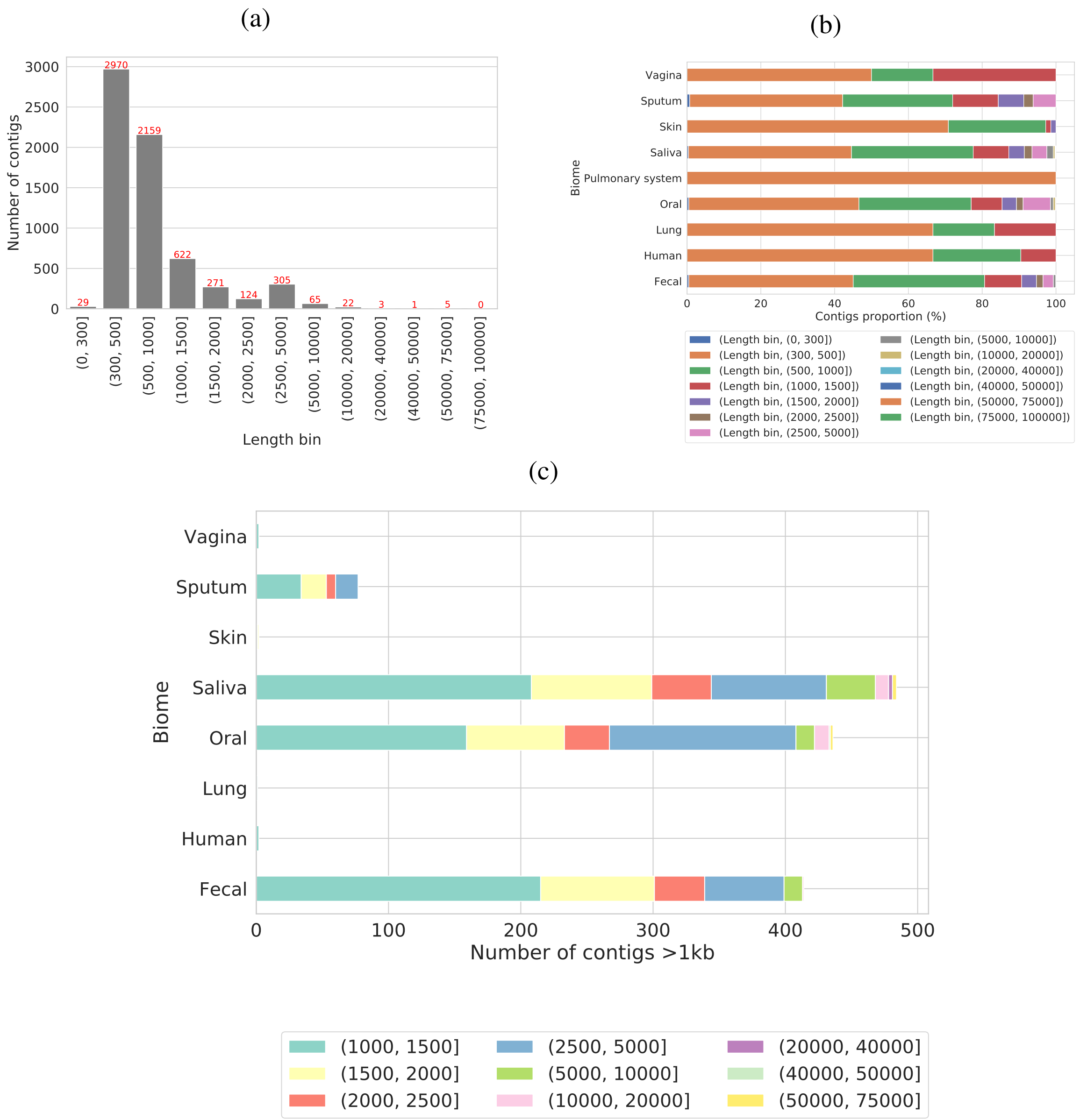


Figure 3.17: An overview of partially known viral contig lengths in different microbiomes. (a) The distribution of contig lengths of partially known contigs with virus hits. The X-axis displays distinct bins of non-overlapping length. Y-axis shows the number of contigs within each category. (b) A proportion plot is used to visualise the proportion of partially known viral contigs in different human microbiomes with different length bins. Each interval bin is coloured according to the colour key shown below the plot. (c) A stacked bar plot illustrating the distribution of partially known viral contigs larger than 1kb in different microbiomes. The X-axis shows the total number of partially known viral contigs and the Y-axis shows different human microbiomes. The colours of the stacked bar plot depict the corresponding length interval.

that were mostly concentrated in the oral microbiome. The third dominant group of viruses was negative ssRNA viruses. They were not found in lung, pulmonary system and vaginal microbiomes. They were mostly concentrated in the fecal microbiome. Interestingly, oral, saliva and sputum microbiomes included the broadest range of viruses matching partially known viral contigs. On the contrary, only one or two distinct genomic groups of viruses were identified in pulmonary, lung and vaginal microbiomes. This distribution could be explained by the low number of samples associated with specific microbiomes and the number of samples analysed for that microbiomes. Moreover, another factor could be the different types of sequencing library preparation steps i.e. metagenomic vs metatranscriptomic that would lead to the capture and sequencing of different types of genomic materials from individual samples.

In order to get further insights into the different virus families represented in the dataset, a heatmap showing different virus families, their genomic composition and the distribution across different microbiomes was generated shown in the figure 3.18(b). The dsDNA families, *Myoviridae*, *Siphoviridae* and *Podoviridae* contain viruses that most commonly infect bacteria and were the most dominant across all microbiomes. Other dsDNA virus families including *Ackermannviridae*, *Ascoviridae*, *Herelleviridae*, *Iridoviridae*, *Marseilleviridae*, *Mimiviridae*, *Nudiviridae*, *Phycodnaviridae* and *Tectiviridae* were also found in different microbiomes in different proportions (fig 3.18(b)). The contigs matching the negative ssDNA viruses were all found to be from the family *Anelloviridae*, predominantly in the oral microbiome. Anelloviruses are thought to be omnipresent in various human microbiomes and have not been linked to any specific diseases or health conditions yet (Kaczorowska et al., 2020).

Other double and single-stranded DNA virus families that have partially known contigs were matching with *Caulimoviridae*, *Genomoviridae*, *Inoviridae*, *Microviridae* and *Circoviridae*. 14 different contigs from 8 different BioProject (PRJEB12357, PRJEB14383, PRJEB8094, PRJEB23207, PRJEB18265, PRJNA230363, PRJEB7949, PRJEB19367) and 3 different microbiomes (fecal, oral and saliva) also matched a group of unclassified DNA viruses named pithoviruses, that are also known as giant viruses.

Single and double-stranded RNA viruses were found in a very small proportion across all microbiomes. However, 5 different RNA virus families: *Partitiviridae*, *Picobirnaviridae*, *Leviviridae*, *Narnaviridae* and *Retroviridae*, were represented in this dataset. A subset of negative ssRNA viruses is excluded from the figure 3.18(b). These viruses were classified at the order level instead of the family level. 641 of these contigs were matching to order *Bunyavirales*.

It is notable that this chapter focuses on the unknown sequences and hence, individual contigs matching to various known viruses were not analysed in detail here. However, a thorough virus metagenomic sequence analysis described in Chapter 5 addressed the individual contigs and virus groups of interest.

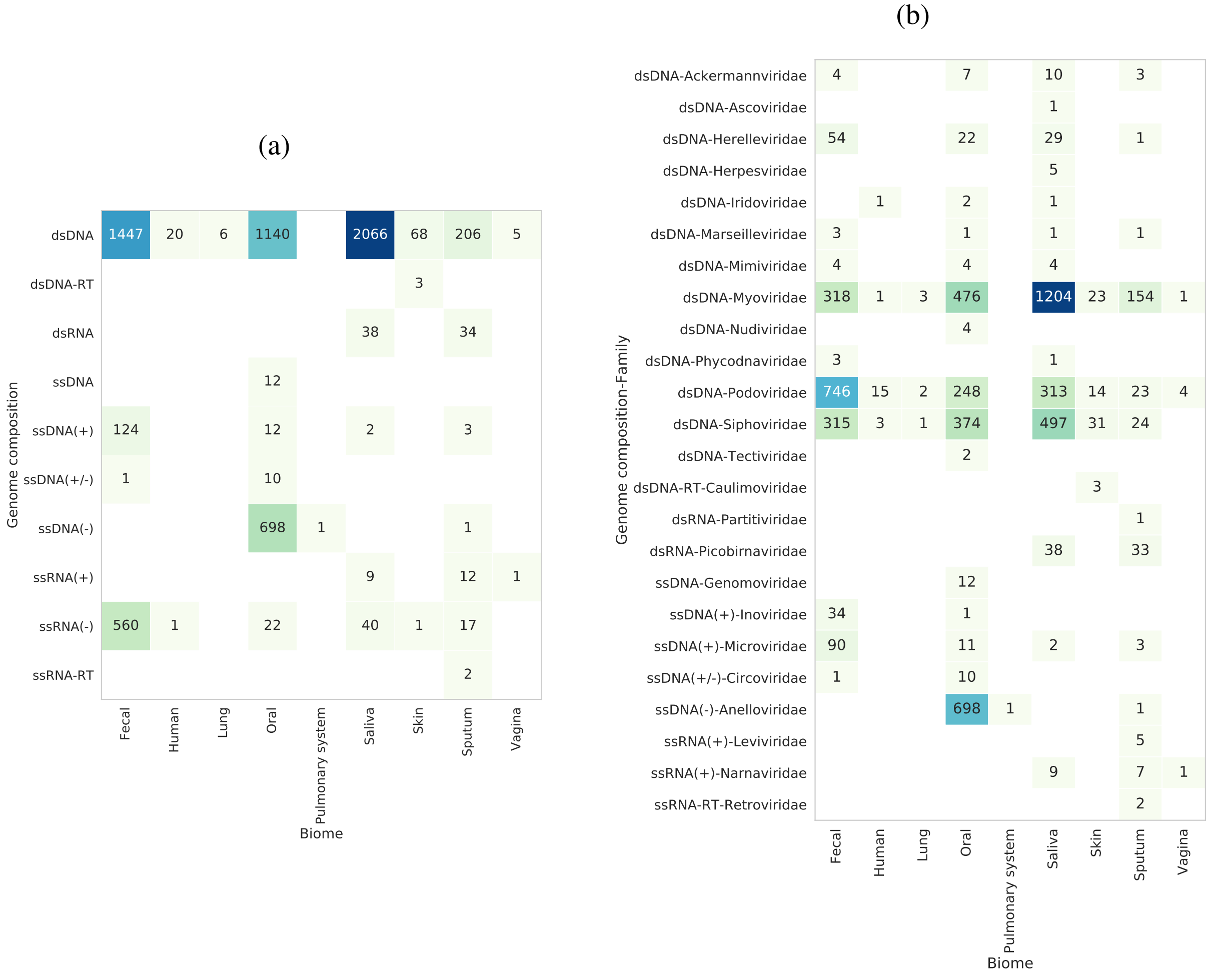


Figure 3.18: The taxonomic grouping of partially known viral contigs in different microbiomes. The virus family and genome composition were derived from the lowest common ancestor (LCA) of the DIAMOND hits. Heat maps depict partially known viral contigs in different microbiomes. The X-axis shows different human microbiomes and Y-axis shows the virus groups: (a) Partially known virus hits are grouped according to the virus genome composition indicated on the Y-axis and the number of contigs in the corresponding categories are annotated on the plot. (b) Partially known virus hits are grouped according to the virus family and genome composition indicated on the Y-axis and the number of contigs in the corresponding categories is annotated on the plot.

3.4.14 Known sequence classification

This category of classification contained the largest number of sequences. A total of 88.22% of all contigs assembled fell into this category. An overview of a brief superkingdom level classification of these contigs is shown in a heatmap in the appendix figure A.4. Each row of the heatmap represents a BioProject label with the corresponding microbiome represented using a coloured box. As can be observed in the appendix figure A.4, bacteria were the most dominant superkingdom in most microbiomes as expected with metagenomic samples. However, the saliva (PRJNA306560) and the oral (PRJEB12998) microbiomes had Eukaryota as the most dominant superkingdom for all known contigs.

It is noticeable that the known contigs matching the viruses were present in a very small proportion across all microbiomes and BioProjects. In total, 3,484 known contigs with exclusive hits to viruses were identified and analysed further. The length distribution of these contigs is shown in the figure 3.19(a). The overall pattern of the distribution of the lengths is very similar to partially known contigs matching to viruses with a higher proportion of contigs in 300-1kb categories in all microbiomes (figure 3.19(b)).

A comprehensive family-level categorisation of the viruses matching to known contigs is shown in the figure 3.19(c). The dsDNA viruses were found to be the most common ones in this classification and were present in all but the pulmonary system and vagina microbiomes. The families representing the viruses of bacteria i.e. phages were most abundant among this set of contigs. Virus families *Herpesviridae* were over-represented in the salivary microbiome. Viruses from the negative single-stranded DNA virus family *Anelloviridae* were abundant in the oral microbiome set. In contrast to the viruses matching the partially known contigs, the RNA virus families *Picornaviridae*, *Virgaviridae* and *Paramyxoviridae* were only found in the known set.

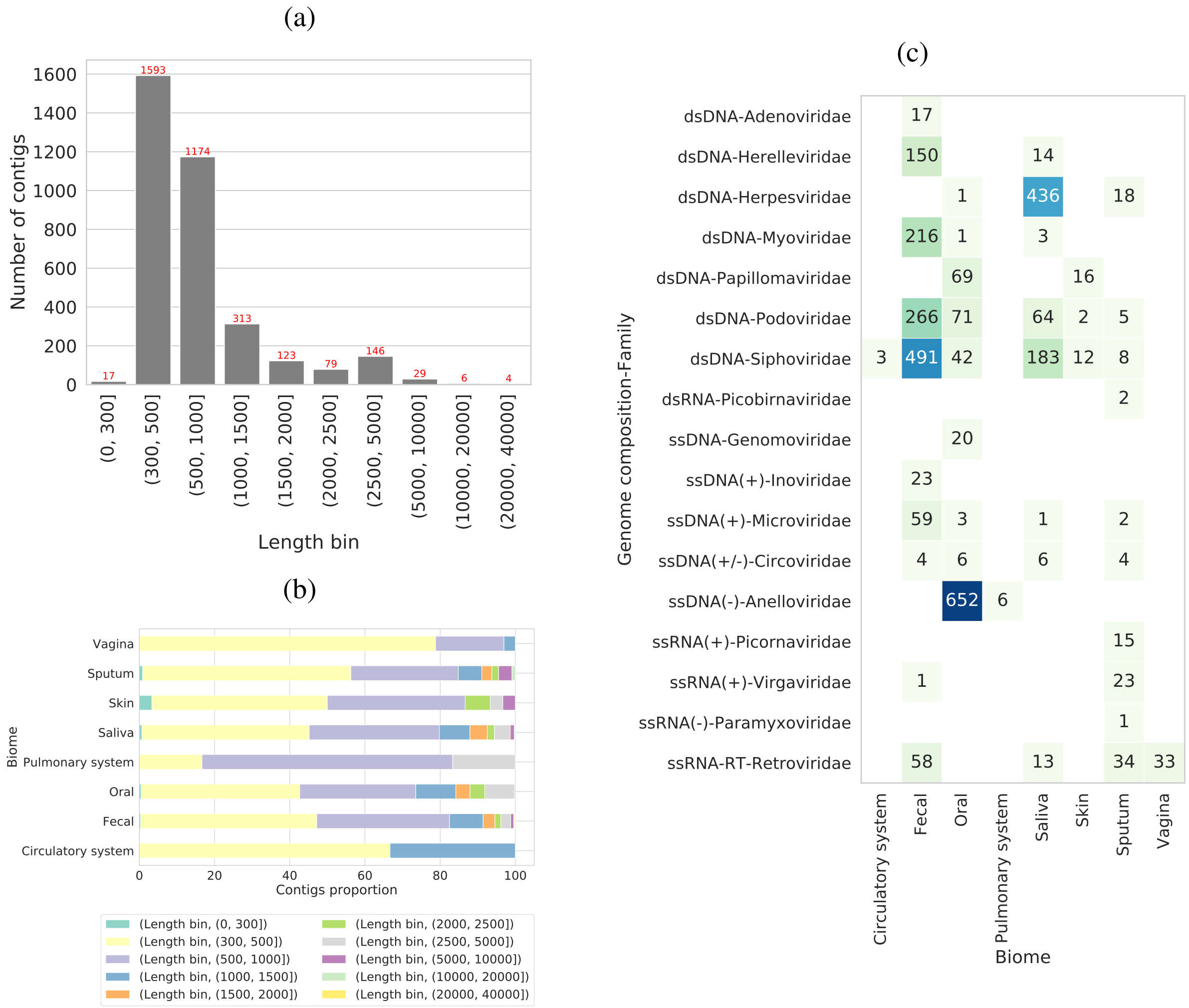


Figure 3.19: Distribution of known viral contigs. (a) The distribution of contig lengths of partially known contigs with virus hits. The X-axis displays distinct bins of non-overlapping length. Y-axis shows the number of contigs within each category. (b) A proportion plot is used to visualise the proportion of known viral contigs in different human microbiomes with different length bins. Each interval bin is coloured according to the colour key shown below the plot. (c) Known virus hits are grouped according to the virus family and genome composition indicated on the Y-axis and the number of contigs in the corresponding categories is annotated on the plot. The virus family and genome composition were derived from the lowest common ancestor (LCA) of the DIAMOND hits.

3.5 Discussion

In this study, we have developed an automated framework that can systematically quantify the proportion of unknown contigs (UCs) in meta-omics samples. Whilst the presence of UCs is well recognised, this is the first study that addresses the question of UCs comprehensively and quantifies it across different human microbiomes. Our approach utilises sequence similarity-based taxonomic categorisation to identify the sequences that cannot be categorised. We define these UCs as the sequences that do not match known sequences in the databases with a predefined sequence similarity threshold of evalue 0.001 which is a very lenient threshold, anything with evalue higher than this is unlikely to truly be related to the database sequence hit. We show that on average 2.29% of assembled contigs are categorised as unknown in different human microbiome studies. Moreover, a subset of those with unknown sequences could be translated and contained protein domains, thus we were able to find functional similarity to 5.49% of taxonomically unknown contigs. We have generated a comprehensive catalogue of 651,529 UCs that do not bear any sequence similarity to sequences present in the widely used GenBank protein and nucleotide databases. Although sequence similarity-based approaches are dependent on the databases, the protein sequence-based approach implemented here is highly effective in fishing out distantly related homologues of known sequences available in the databases (Altschul et al., 1990) and thus provides better resolution for sequence classification compared to those solely based on the genomic signature-based binning (L. X. Chen et al., 2020). This study highlights the importance of avoiding the “street light" effect i.e. observational bias arising from classifying metagenomic sequences on the basis of related sequences that already exist in the databases. Here, we have aimed to eliminate such observational bias by performing a comprehensive data mining of the human microbiome data and cataloguing the UCs, their frequency in different human microbiomes and their overlap between different samples.

This study has enabled the identification of a range of genomic sequences that are hypothesised to belong to currently uncharacterised organisms that are often found in similar samples and/or microbiomes. A range of large UCs with and without known protein domains are presented here. However, the complete set includes a large number of UCs that still remain unknown and can be mined further to study their biological origin. A third of all UCs (n=215,985) contained large predicted open reading frames (at least 100 amino acids long) that were predicted using the standard genetic code. Using alternative genetic codes may expand this set further by revealing novel, potentially different open reading frames generated from the UCs. A small proportion of these open reading frames contained domain signatures confirming the presence of currently unidentified organisms. Moreover, a comprehensive clustering analysis has led to the identification of UCs that were present across different human microbiomes (as well as from different samples/studies investigating the same human microbiome) indicating that we have discovered potentially widespread and as yet unclassified novel biological organisms within the human microbiome. The multi-microbiome clustering approach applied here provides an interesting way to understand the diversity and the distribution of the UCs across different microbiomes and geographical sites. For example, this approach led to the identification of 30 clusters that spanned 4 distinct microbiomes. The largest multi-microbiome cluster comprised 57 UCs recovered from saliva, sputum, oral and lung microbiomes and was assembled from 12 different samples. Although it is impossible to identify the true clusters present in the data due to the novelty of the UCs, the clustering approach helps to identify obvious patterns of sequence similarity between microbiomes and studies. This approach provides an additional dimension by capturing unknown sequences that are shared between different projects or human microbiomes.

Virus predictions carried out by DeepVirFinder - a machine learning-based virus prediction tool for identifying viruses from metagenomic datasets - have shown that approximately 50% of all UCs are likely to be of virus origin. Additionally, nearly 30% of all UCs identified in this study have an overlap with uncultivated viral genomes currently catalogued in IMG/VR databases. As with most similarity-based approaches, we used an arbitrary threshold for determining a match to the IMG/VR database and thus a match does not mean they are closely related. Interestingly, this study provides an added dimension to these matching uncultivated viral genomes (UViGs) by providing information on the type of microbiome they have been found in. It is anticipated that UCs catalogued in this study may have some overlap with other viral genome databases such as Gut Phage Database (Luis F. Camarillo-Guerrero et al., 2021) and Gut Virome Database (Gregory et al., 2020). Short contigs i.e. those less than 1-5kb are often ignored in most data mining and exploration research typically in studies that employ a contig binning step as binning has been shown to be less sensitive for short contigs (Breitwieser et al., 2018; L. X. Chen et al., 2020; Mallawaarachchi et al., 2020). The clustering and time point analyses carried out on short UCs have shown that these short UCs are originating from biological entities and predominantly represent the novel microbial sequences that are currently uncatalogued. This has been demonstrated with the example of short circular sequences with terminal repeats. Short contigs, which are typically excluded from large microbiome mining studies employing the metagenomic binning approach, were studied in detail here. These short UCs are found across multiple human microbiomes and samples, we speculate that these are of viral origin and could potentially represent novel CRESS DNA or satellite viruses, although the ORFs originating from these genomes do not bear any sequence of functional similarity to the typical rep and cap genes. Moreover, a number of large contigs were found to contain various functional ORFs and domains often originating from viruses or phages indicating that a proportion of UCs is very likely to be novel viruses that infect currently uncharacterised microbes. In our approach, we have implemented a protein sequence similarity-based identification that enables the identification of distantly related sequence homologues (Altschul et al., 1990). This approach can potentially 'classify' contigs of viruses or phages as their corresponding host with very low sequence similarity. Indeed, viruses are well known to mimic their host genomic signatures by incorporating genomic sequences from their host into their genome. We anticipate that the virus diversity described in this manuscript is reasonably underestimated due to this specific characteristic of viruses and speculate that a range of assembled contigs classified as bacterial with very low sequence similarity across a short genomic coverage is likely to be of virus origin. This hypothesis will need to be tested further by mining the ‘known’ and ‘partially known’ contigs systematically. In order to explore virus sequences assembled with UnXplore, a comprehensive analysis of all contigs was carried out, and this process is described in detail in Chapter 5. We note that a range of UCs matching known and partially known sequences could be taxonomically uncharacterised in GenBank databases such as unclassified viruses. Assembled contigs matching these sequences are categorised as known (protein sequence similarity >80%) or partially known (protein sequence similarity <80%) in this study. Those contigs would need to be investigated further to identify potentially novel and divergent sequences assembled in this study. The HMP control sample analyses resulted in only a few UCs validating the UC identification approach implemented in our framework. The results generated from this study can be extended to identify the organisms that co-occur in different microbiomes, which in turn can help to inform the interactions between these organisms and how it affects their hosts - humans. Despite having sequenced human microbiomes extensively, our understanding of how these microbes interact with humans remains limited. These large-scale explorations can help to understand the human holobionts and the interactions of macro- and microorganisms. Based on these results, we do not know whether the microbes identified in different studies are consistently associated with humans or they are just passing associations captured at the time of sampling, the latter would make it even harder to make comparisons between samples and microbiomes.

The UCs landscape changes over time as more sequences get characterised and added to the ever-expanding sequence repositories. This was demonstrated by comparing the UCs to different GenBank databases over the course of 18 months. We have estimated that 1.64% of the UCs identified in this study are getting characterised each month. However, this number would be highly dependent on the types of data deposited in the International Nucleotide Sequence Database Collaboration (INSDC) resources. This study provides a strong foundation for preliminary estimation of this rate and UCs would need to be analysed at multiple future time points to determine how the rate at which the UCs are being classified, changes over time. Additionally, the time-point analysis also provides strong evidence of the real biological entities being assembled and characterised in our study. Indeed, a proportion of the UCs was taxonomically classified during the period of the study. This delineation of the UCs demonstrates that the unknown matter that surrounds us largely belongs to currently uncultured, unidentified microbes that we interact with on a daily basis. The technological advances have accelerated the speed at which genomic sequences belonging to novel uncultured organisms are being deposited in INSDC databases. This sharp increase of metagenomically assembled microbial genomes has led to the scientific community driving the development of genomic data and metadata standards such as MIMAG (for bacteria and archaea) (Bowers et al., 2017) and MIUViG (for viruses) (Roux et al., 2019) for consistency and comparison purposes. The taxonomic classification landscape has also faced a tectonic shift whereby it is moving from the phenotype-based classification to a more holistic sequence-centric phylogenetic classification, e.g. GTDB (bacteria and archaea) (Parks et al., 2018) and ICTV (viruses) (Simmonds et al., 2017a). These changes enable the incorporation of the uncultured sequence diversity into the microbial taxonomy and will provide a more comprehensive understanding of the complex phylogenetic relationships and interactions between different microbes.

The metagenomics analysis framework developed here works as a proof of concept for overcoming the challenge of the quantification of the unknown in already 'analysed' data sets. The pipeline developed here is flexible and can be applied to any microbiome. To get a cross-section of different human microbiomes and geographical locations whilst keeping the overall data set size manageable large studies involving $>$100 samples were down-sampled. This framework can readily be applied to routine metagenomic exploration, which can help to gain further understanding of the landscape of sequences of unknown origins. The framework applied here is easily portable to metatranscriptomics data. In fact, a couple of the BioProjects (PRJEB10919 and PRJEB21446) analysed in this study were indeed from a metatranscriptomic study. It is important to note that, unlike other studies that often focus on the cross assembly of different samples, each sample was assembled individually here. This is regarded as best practice when a cocktail of samples from unrelated studies is analysed in bulk. The co-assembly would often lead to fragmented assembly as the complexity of sequences originating from multiple samples would be much higher compared to a single sample (Olm et al., 2017). On the contrary, independent assembly is expected to capture better diversity across each sample with high-quality genomes assembled from each sample (Olm et al., 2017). Typically the sequence similarity-based approach is less reliable for unrelated sequences as the similarity search tools heavily rely on the databases used in the analysis. Like most other pipelines, this framework classifies the sequences with respect to a static version of the reference sequence databases. The search results are as good as the data in the ever-expanding repositories that are often too large to be hosted on a local computer. Furthermore, a number of these gold-standard repositories have been shown to contain erroneous and contaminated sequences (Steinegger et al., 2020) which could potentially impact the sequence similarity-based approach implemented here. In order to improve this, an alignment-free approach could be explored. The development of a general-purpose alignment-free prediction method that can categorise the sequences based on the genomic composition would be suitable for the downstream analysis of the UCs. The UCs classification is highly dependent on the methods employed to identify and quantify the unknown. Moving away from the sequence similarity-based methods would help to categorise and classify the currently unknown sequences better. Machine learning-based approaches might be deemed suitable in certain circumstances to overcome the similarity threshold-based approaches. In the case of completely novel sequences that bear no similarity to currently known sequences, significantly rigorous training sets and features would need to be identified and be built into the models in order to make accurate predictions as machine learning approaches are highly reliant on the training data the models have been developed with. Moreover, a recent study by Krishnamurthy and Wand (Krishnamurthy et al., 2018) made predictions for picobirnaviruses to be bacteriophages rather than eukaryotic viruses based on the presence of bacterial ribosome-binding sites in front of the coding sequences. This approach could potentially be applied to check whether viral UCs are bacteriophages.

This study demonstrates that there is a large diversity of unknown sequences embedded within various human meta-omic samples available in public repositories. It is clear that the unknown sequence landscape observed in this study is likely to be the tip of the iceberg, and, as we scan more microbiomes and extend this to less-studied environments e.g. insect metagenomes, we are likely to gather a better understanding of the unknown sequence space. As more species and environments are sequenced more readily, the rate at which the unknown sequences become known would also change. Our results of novel viruses indicate that the unknown microbes and their genomic signatures are likely to be more divergent to those currently present in widely used sequence databases; however, it should be noted that many of the short contigs found in our study are likely to represent fragments of larger viral genomes rather than being short but complete viral genomes. Our study also shows that at least some of these unknown microorganisms are prevalent in nature. To overcome this, more comprehensive resources including searchable databases such as those enabled using BIGSI (Bradley et al., 2019) and federated indexes (Martí-Carreras et al., 2020) could be created for the unknown sequence data and metadata. This would allow researchers to explore the human metagenomic sequence space in a more holistic manner and in turn, provide a better understanding of microbial diversity interacting with and within human hosts. It would enable researchers to search, link and explore the unknown sequences present in different microbiomes, studies and samples. Such resources could help in speeding up the pace at which unknown sequences can be ‘classified’ and make it easier for researchers to determine the functional and/or ecological importance of the organisms the sequence comes from. A concerted effort could help to pin down human-microbial interactions in a broader context such as linking unknown microbes to human diseases and disorders of unknown aetiologies.

In conclusion, more raw sequence data should be mined in this holistic manner to discover novel, uncultivated species of microorganisms. The unknown sequence classification is highly dependent on the methods employed to identify and quantify the unknown. Moving away from the sequence similarity-based methods would help to categorise and classify the currently unknown sequences better. A novel machine learning-based prediction method developed and described in Chapter 4 was explored and applied to “classify” all unknown sequences without a taxonomy label into the higher-order ‘classification’ such as bacteria, archaea or viruses. The partially known and known sequences that were found to be matching to viruses were systematically analysed to catalogue the virosphere. The results of a more standard metagenomic analysis are described in detail in Chapter 5.

Chapter 4

Predicting the biological origin of unknown sequences using machine learning

Shedding light on the unknown.

4.1 Abstract

There are unknown sequences embedded within metagenomic and metatranscriptomic datasets which cannot be classified taxonomically or functionally. They represent the genetic signatures of entirely new microbes that could be interacting with known microbes and their hosts on a regular basis. Previous studies have found tetranucleotide signatures are unique to microbial species and contain phylogenetic information. The main objective of this study was the development of simple genome-composition-based machine learning models that could be used to classify archaea, bacteria, plasmids, and viruses with high levels of accuracy and precision based solely on their k-mer composition. These models were packaged and made available to the scientific community through a PyPI package - TetraPredX. TetraPredX was applied to unknown sequences catalogued from human microbiomes and more than 70% of the unknown sequences were accurately determined to be viruses. These results support our initial hypothesis that the vast majority of unknown sequences are likely to originate from viruses. An analysis of TetraPredX shows that in the absence of alignment-based sequence similarity, it can assist in identifying novel microbial sequences embedded within unknown sequence matter efficiently and with high accuracy. A comparison of TetraPredX's models against DeepVirFinder showed TetraPredX was able to identify most DeepVirFinder predicted sequences with equivalent accuracy. TetraPredX includes a set of models that can be applied to any metagenomic or metatranscriptomic dataset to assist in the identification of unknown sequences.

4.2 Introduction

A number of studies have shown that a significant proportion of metagenomic and metatranscriptomic samples contain unknown sequences, that is, sequences that do not match those in public databases, such as GenBank (Tisza et al., 2021b; Bernard et al., 2018; Krishnamurthy et al., 2017; Shkoporov et al., 2019a; Zamkovaya et al., 2020; Aevarsson et al., 2021). New microbiome databases have been created as a result of the identification of novel uncultivated microbes, primarily viruses. These include Gut Virome Database (GVD), Metagenomic Gut Virus (MGV) and Gut Phage Database (GPD) (Gregory et al., 2020; Luis F. Camarillo-Guerrero et al., 2021; Benler et al., 2021; Nayfach et al., 2021). As discussed in Chapter 3, the UnXplore framework was developed to systematically identify and quantify unknown sequences from 40 distinct microbiome studies spanning 963 samples and produced a comprehensive set of 651,529 unknown contigs (UCs). These UCs were predicted to belong to microbes that have not yet been identified based on functional characteristics such as protein domains. Despite this, most of the UCs remain novel and unknown since they bear no sequence or domain similarity to currently known sequences available in NCBI databases.

Sequence similarity approaches such as BLAST are not able to categorise these UCs and thus alternative approaches are required to ascertain the origin of unknown sequences. Machine learning algorithms are computational approaches combined with statistics implemented in a programming language that can potentially identify complex hidden patterns embedded in large datasets. These methods, also referred to as pattern recognition or data mining algorithms, can decode signals specific to different data points, classify them into different categories, and can be trained to predict the outcome of previously unobserved data points (Tarca et al., 2007; Chicco, 2017). Machine learning (ML) has become an increasingly mainstream bioinformatics tool applied to address a range of biological problems in this genomics-led data science era (Larrañaga et al., 2006; D. T. Jones, 2019; Libbrecht et al., 2015). ML methods have been employed to tackle various complex biological questions from drug-target identification to gene prediction, virus-host predictions and image analysis (Zitnik et al., 2019; Babayan et al., 2018; Nami et al., 2021).

k-mer frequencies and composition have also been used in prokaryotic classification. Tetra-nucleotide frequencies (TNF) have been shown to contain classification signals similar to phylogenetic signals (Pride et al., 2003; Teeling et al., 2004). TNFs combined with abundance can provide up to species-level classification in bacteria. Metagenomic binning tools e.g. CONCOCT (Alneberg et al., 2014), MaxBin2 (Y. W. Wu et al., 2016), MetaBAT (D. D. Kang et al., 2015), MetaBAT2 (D. Kang et al., 2019) make use of TNF (often combined with other sequencing metrics such as coverage and abundance) to provide up to strain-level resolution and clustering in metagenomic datasets. TNF can also aid virus-host interactions and accurate host predictions for novel viruses (Pride et al., 2006; Roux et al., 2015b).

Due to the diversity encompassed by viruses, prediction tools often combine multiple methods to accurately predict novel viruses. Popular methods such as VirSorter and the more recently published VirSorter2 combine multiple methods such as Hidden Markov Models (HMM), homology to known viruses, hallmark virus gene identification, and genomic composition metrics incorporated into machine learning algorithms that can enable accurate identification of novel viruses in unexplored environments. VirFinder was the first reference-free k-mer-based predictions tool (Ren et al., 2017). An updated version of the same tool, DeepVirFinder (Ren et al., 2020) has been deemed the most efficient for the identification of bacteriophages in metaviromic datasets in an extensive benchmark carried out by Fung et al. (2022). These k-mer composition-based prediction methods have been shown to outperform other more extensive database-based methods in predicting viral genomes and segments accurately from assembled contigs (Fung et al., 2022).

TNF signals combined with machine learning methods were used as motivation to investigate UCs further by developing and implementing simple TNF-based ML algorithms to predict the origin of the UCs discovered in this study. As the samples analysed in Chapter 3 originated from the EBI MGnify subset and contained a very limited set of blood microbiomes, we complement our previous UCs catalogue with additional human blood microbiome samples processed through UnXplore.

4.3 Methods

4.3.1 Human blood microbiome dataset

Blood is the liquid channel that transports and preserves life's most fundamental, but vital, ingredients. Traditionally, human blood is considered a relatively sterile environment compared to other bodily sites such as the gastrointestinal tract or oral cavity. However, a number of studies have observed the presence of various microbial and virus sequences in human blood which has led to a debate surrounding its 'sterile' status (Païssé et al., 2016; Castillo et al., 2019; Moustafa et al., 2017; Cebriá-Mendoza et al., 2021). It is important to characterise and catalogue the microbial content of the blood as it is relevant in the context of epidemiological surveillance and more importantly for transfusion safety (Sauvage et al., 2016; Païssé et al., 2016; Wen Zhang et al., 2016). To quantify and identify the proportion of unknown contigs present in the human blood microbiome datasets, the UnXplore framework was applied to human blood microbiome samples. NCBI Entrez utilities were used to search SRA repositories using query 'human blood metagenome AND "platform Illumina"[Properties]' as well as -query 'txid1504969[Organism:noexp]'. Illumina sequence data were combined from these two sets and 16S/amplicon samples were excluded. The resulting set included 21 distinct BioProjects covering 3,312 samples of which 2,625 samples were downloaded using 'parallel-fastq-dump' (Valieris R., 2020).

All samples were analysed using the UnXplore framework (described in detail in Chapter *Identification and quantification of 'unknown' biological sequences in human microbiomes*) with the nucleotide and protein reference databases downloaded in February 2020. To update this analysis and carry out similarity searches against a more recent version of databases, all contigs classified as unknown, as well as those UCs previously identified in Chapter 3 were searched against databases downloaded in February 2021. A final set of UCs that were at least 1 kb long was created (n=20,552).

4.3.2 Machine learning models

UCs do not bear any alignment-based sequence similarity to known sequences in the databases and hence it is difficult to categorise them. There have been a variety of applications of machine learning techniques to biological data in recent years. ML methods can be applied as a useful tool that can help to predict the origin of the microbial UCs catalogued here. Nucleotide composition-based ML methods were explored in this study to obtain potential predictions of UCs based on their sequence composition. An overview of the various models, datasets and optimisation steps is shown in figure 4.1.

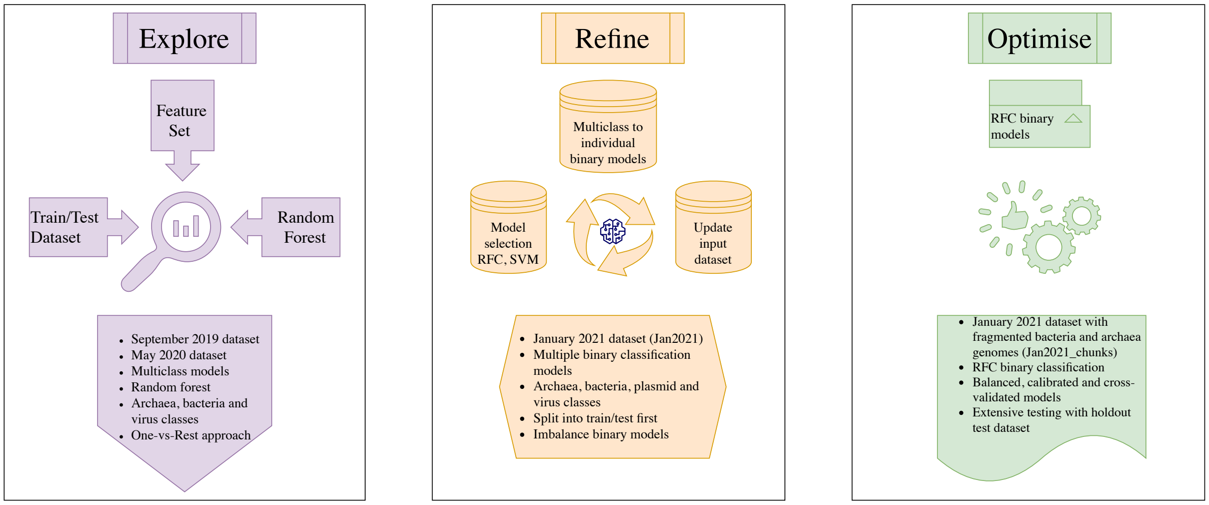

Explore
Feature Set
Train/Test Dataset
Random Forest
September 2019 dataset
May 2020 dataset
Multiclass models
Random forest
Archaea, bacteria and virus classes
One-vs-Rest approach
Refine
Multiclass to individual binary models
Model selection RFC, SVM
Update input dataset
January 2021 dataset (Jan2021)
Multiple binary classification models
Archaea, bacteria, plasmid and virus classes
Split into train/test first
Imbalance binary models
Optimise
RFC binary models
January 2021 dataset with fragmented bacteria and archaea genomes (Jan2021_chunks)
RFC binary classification
Balanced, calibrated and cross-validated models
Extensive testing with holdout test dataset

Figure 4.1: An overview of different stages of model development with descriptions of algorithms, datasets, and optimisation procedures. Briefly, the 'Explore' component entailed determining the suitable feature set, algorithms and classes for supervised learning. Initially, multi-class multilabel prediction models were designed. These models were improved as shown in the 'Refine' block. The multi-class models were transformed into binary classification models designed with updated datasets. The final binary classification models were validated, calibrated and packaged into TetraPredX as shown in the 'Optimise' block.

4.3.3 Machine learning datasets

The complete genome sequences used to train and test the ML models were downloaded at three different time points: September 2019, May 2020 and January 2021, and are labelled accordingly (Table 4.1 and Appendix table B.1). September 2019 dataset was generated by downloading complete genomes for archaea, and bacteria using the assembly summary file available on https://ftp.ncbi.nlm.nih.gov/genomes/. For viruses, reference genomes were downloaded using NCBI eutilities. May 2020 dataset contained archaea references and complete genomes downloaded from GenBank nucleotide databases downloaded on 19/05/2020. It contained bacteria reference and representative genomes downloaded on 13/05/2020 and ICTV species exemplar for virus genomes downloaded on 13/05/2020. January 2021 dataset contained archaea genomes downloaded from GenBank nucleotide on 14/01/2021, bacteria reference and representative genomes downloaded on 15/01/2021 and ICTV species exemplar genomes downloaded on 14/01/2021. Additionally, this set also contained reference plasmid genomes downloaded from https://doi.org/10.15146/R33X2J. Plasmid sequences from bacterial genomes were also separated using the 'plasmid' string in the header and were added to the plasmid set.

All multi-class multilabel exploration models described below were trained and tested using the September 2019 and May 2020 datasets. All binary models described below were trained and tested using the January 2021 dataset as well as a curated plasmid dataset published in Brooks et al. (2019). A variant of the January 2021 dataset was used to optimise binary models whereby bacteria and archaea genomes from this set were fragmented into non-overlapping chunks to increase the number of observations for the models without introducing any artificial bias in the data. This strategy has been shown to be efficient to predict the microbial class of assembled sequences identified from the metagenomic datasets (Ren et al., 2020).

Table 4.1: An overview of the various datasets used for building and developing machine learning models.

Dataset	Archaea	Bacteria	Virus	Plasmid	Model
September 2019	54,896	28,666	12,148		Multiclass Random Forest Classifier (RFC)
May 2020	1,268	5,441	7,143		PCA, t-SNE, One-vs-rest
January 2021	10,319	9,814	7,953	6,642	Binary RFC, Support Vector Classifier

Feature and data selection

Genomic composition-based ML models were trained using existing bacteria, archaea, viruses and plasmid sequences downloaded from GenBank (Table B.1). In the case of prokaryotes, reference and representative genomes were downloaded. Briefly, RefSeq reference genomes are NCBI curated sequences that are high-quality genomes and identified as being important, and clade-specific representative genome sequences are provided by the NCBI in absence of reference genomes (https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/faq/). Initial models were developed using archaea, bacteria and virus genomes and plasmid sequences were added to models at a later stage. A python script ‘DownloadReferenceGenomes2020.py’ was developed to download reference and representative genome sequences in FASTA format for bacteria, archaea and viruses from NCBI databases. This set included reference and representative complete genomes of bacteria, ICTV virus species exemplars (ICTV, 2021b) identified from NCBI virus RefSeq (Brister et al., 2015) and all archaea sequences from NCBI nuccore were downloaded. A range of k-mer features described below was extracted from these sequences and their taxonomy was used as labels for supervised learning models explained in detail below.

Briefly, k-mers are overlapping words of size k generated from a sequence. The possible number of unique k-mers is calculated with n^k where n represents the number of unique monomers. For example, in the case of DNA bases these values would be $n = 4$, (i.e. A, C, G and T), and k of size 2 would yield $4^2 = 16$ k-mers. For a given sequence of length L, the total number of k-mers for size k can be calculated from $N = L - k + 1$. In supervised machine learning, such measures (e.g. k-mer frequencies) are often termed ‘features’. These features are used to train ML models to predict the corresponding class outcome i.e. ‘label’. Features and labels are generic terminologies used in ML. To normalise k-mer frequencies for the length of the sequence, the frequency value for each k-mer was divided by the total number of k-mers generated for each sequence. These normalised frequencies were calculated for each sequence in the model training and test dataset.

The Python programming library scikit-learn (sklearn) provides an extensive toolkit for predictive data analysis (Pedregosa et al., 2011). All models and methods used for modelling described below were used from the pre-implemented sklearn library. k-mer length of 4 was selected for the feature set which led to 256 unique k-mers used in this analysis. As described in the Introduction, TNF i.e. k=4 features have been shown to contain classification signals similar to phylogenetic signals (Pride et al., 2003; Pride et al., 2006; Teeling et al., 2004).

To count the word frequencies ($k = 4$) for each sequence, the sklearn feature extraction method TfidfVectorizer was applied to the FASTA sequences. A tabular output with each sequence record identifier such as FASTA header or NCBI accession, k-mer frequencies for each feature and sequence label (e.g. bacteria, archaea or virus) was generated. This dataset was subsequently used for modelling.

Plasmids are naturally occurring extra-chromosomal circular sequences that are often present in bacterial genomes. Initial models were developed where plasmid sequences were not separated from the bacterial genomic sequences. In later versions of models, the plasmid class was separated from the bacterial sequences and it was compiled with a curated plasmid database that was published by Brooks et al. (2019).

In the final dataset, four distinct classes of records were included: archaea (n=10,319), bacteria (n=9,814), plasmid (n=6,642) and viruses (n=7,953).

Algorithm selection

Initially, a multilabel modelling approach whereby sequences were classified into one of three classes was assessed and decision tree-based ensemble methods were explored. A random forest classification model for multilabel prediction was deemed suitable as it supports multi-class classification and performs well with an imbalanced dataset (i.e. different number of observations in each class) which was the problem at hand in our case (Dittman et al., 2015). Briefly, Random Forest (RF) is a supervised machine learning algorithm. It is an ensemble of decision trees that are typically trained with bagging and feature randomness (figure 4.2(a)). The bagging method applies a combination of learning models/trees to cast votes for the most popular class for the given input and this process increases the overall accuracy of the result (Breiman, 2001). Each class in the input data was split into training and testing sets with 70% of observations being used to train the model and the remaining 30% being used to test the model performance. Although promising results were obtained using this approach (see results section: 4.4.4), forcing input into one of the predefined classes (3 classes) included in the model was deemed a major drawback of this multi-class model. To overcome the issue of forcing the classification of the input to one of the predefined classes included in the model, multiple binary classification models were explored. In simple terms, in binary classification, each sequence is independently predicted as to whether it is bacteria (yes/no), virus (yes/no) etc for class with a probability. Binary classification alleviated the forced predictions of unknown contigs into one of the defined classes. They are deemed more suitable for this use case as unknown contigs that are of non-microbial origin will not be forced into one of the pre-defined classes.

Initially, one of the most popular binary classification algorithms, Support Vector Machines (SVMs), was considered along with converting the existing RF model into multiple binary classification models. SVMs are simple models that work by finding the decision boundaries between two classes (Cortes et al., 1995). The SVM algorithm implementation in sklearn can help to identify the best shape of this classification boundary as it may not always be a straight line. Datasets were split into positive and negative sets for each class (virus, bacteria, archaea, and plasmid) and independent individual models were trained for each binary classification. Similarly, independent individual models were also developed using the RF algorithm. The train/test ratio of 70:30 was used for all models. The binary models were developed using two different data sampling strategies. The first strategy involved splitting the data into train and test sets first and then developing the binary classifier model for each class. Notably there were unequal numbers of observations for each class in the training set; 7,267 archaea, 6,857 bacteria, 4,614 plasmid and 5,571 viruses. The result is an imbalance class problem because the negative observations (i.e. not bacteria, not viruses, etc) typically include all remaining sequences in the training class. For example, for archaea models, the negative set included 17,042 observations (bacteria, plasmids,

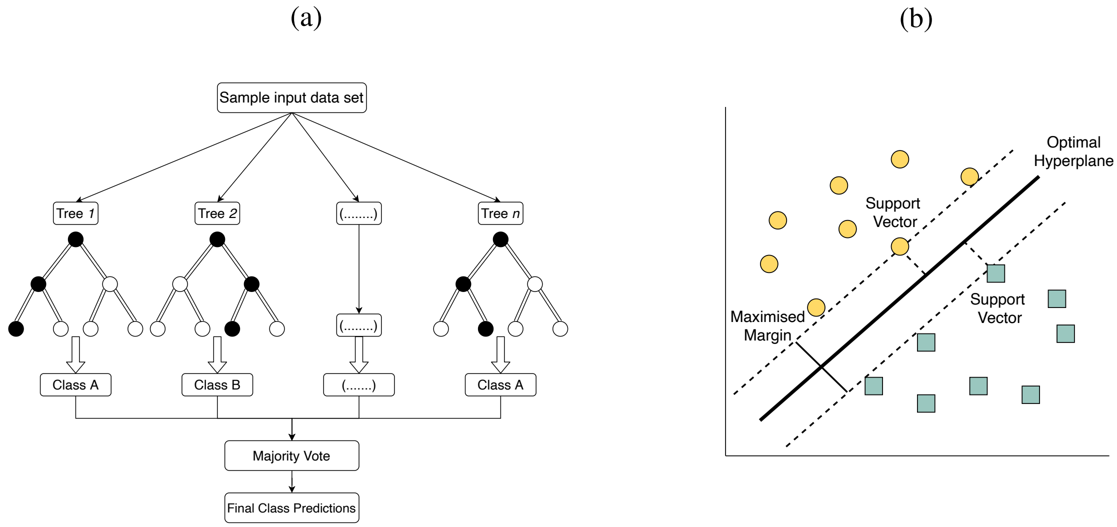


Figure 4.2: Illustration of two different machine learning algorithms used to predict sequence classes derived from microbial genomic compositions. (a) Ensemble decision tree-based classification algorithm implemented in random forest classifiers. (b) Standard binary classification method implemented in Support Vector Machines.

and viruses). This demonstrates that the number of observations in the positive and negative classes is not balanced i.e. more negative observations compared to positive observations are taken into account by each model. An alternative to this is to train each model such that for each class, an equal number of negative observations are drawn to match the number of positive observations which leads to a balanced dataset.

Model calibration and Holdout data

ML models that output prediction probabilities or scores often need to be calibrated to ensure that the model is not predicting outcomes in favour of the majority class (in the case of imbalanced data). The predicted probabilities can be over or underestimated in the case of a non-calibrated classifier. sklearn CalibratedClassifierCV was used to calibrate the final model to overcome this issue.

To assess model performance and validate the model predictions, 4 different additional holdout datasets were designed for bacteria and virus classes. Technically, splitting datasets into training and testing sets would automatically treat the test set as a holdout set but in order to further assess the model predictions on previously unseen data, additional data was used which is termed as a holdout dataset in this context. The contig and scaffold level assembled sequences (>=1kb) from a completely novel species of bacteria *Paraburkholderia madseniana* (n=385, NCBI taxonomy ID: 2599607) were used to assess the predictions made by the final calibrated RF models. In the case of viruses, 3 virus families; *Geminiviridae* (n=578), *Chuviridae*(n=32) and *Siphoviridae* (n=785) were chosen as holdout sets. The three virus families chosen here are representative of the distinct genomic compositions of viruses. *Geminiviridae* are ssDNA viruses, *Siphoviridae* are dsDNA viruses and *Chuviridae* are ssRNA viruses that can be segmented or unsegmented. In order to assess the quality of the predictions and test whether a completely novel family of viruses could be identified using the TNF-based ML models, each set of virus family-specific sequences were removed from the training set and those models were used to predict the class for the holdout data representing the individual virus families.

All models were rigorously tested using a range of datasets shown in tables B.1 and B.3.

Virus-specific models

To predict whether tetranucleotide frequencies could also be utilised to perform virus-specific predictions e.g. genome type, realm, and segmentation, a range of other ML models for these properties was developed and tested. Virus metadata related to these labels were extracted from the ICTV Virus Metadata Resource (VMR) version 010820 MSL35 (ICTV, 2021a). Negative and positive datasets for each of these models were generated from virus data included in the previous models.

Unknown data predictions

The final binary RF models were used to predict the class of the UCs. Additionally, two other widely used virus prediction tools, VirSorter2 and DeepVirFinder were also used to predict the proportion of UCs that may be of virus origin. Both tools were run with the default parameters. The results for DeepVirFinder were filtered for score >=0.5 was used as it is comparable to probability >0.5 in our models and the qvalue threshold of <0.05 was applied 95% confidence threshold.

4.4 Results

4.4.1 Human blood microbiome mining

Initially, 21 blood microbiome studies comprising 3,312 samples were shortlisted, of which 2,625 were successfully downloaded. A total of 2,596 samples from 18 BioProjects could be assembled and analysed further. 22 samples could not be assembled and were excluded. The geographic location of 2,430 samples was extracted from SRA metadata and is shown in figure 4.3(a). Approximately 70% of samples were collected from the USA. The second-largest number of samples were collected from Sweden. In total, 2,999,668 contigs were assembled and 331,079 were at least 300 bases long. These ‘long’ sequences were submitted to the downstream analysis of the UnXplore pipeline. Following the approach defined in Chapter 3, 161,730 contigs (48.85%) were classified as known, and 86,873 contigs were classified as partially known (26.24%). 12,673 contigs (3.83%) were shown to have at least one BLASTN hit despite not having any BLASTX hits (3.83%). 69,803 (21.08%) long contigs were classified as unknown (UCs). The proportion of these unknown contigs (UCs) for each study is shown in figure 4.3(b). On average, 24.75% of unknown contigs were found in these samples with a standard deviation of 21.29%. 14 samples originating from 6 different studies contained 100% unknown bases, however, these samples contained less than 20 long contigs (i.e. >=300 bases long) each with most of them containing only 1 UC.

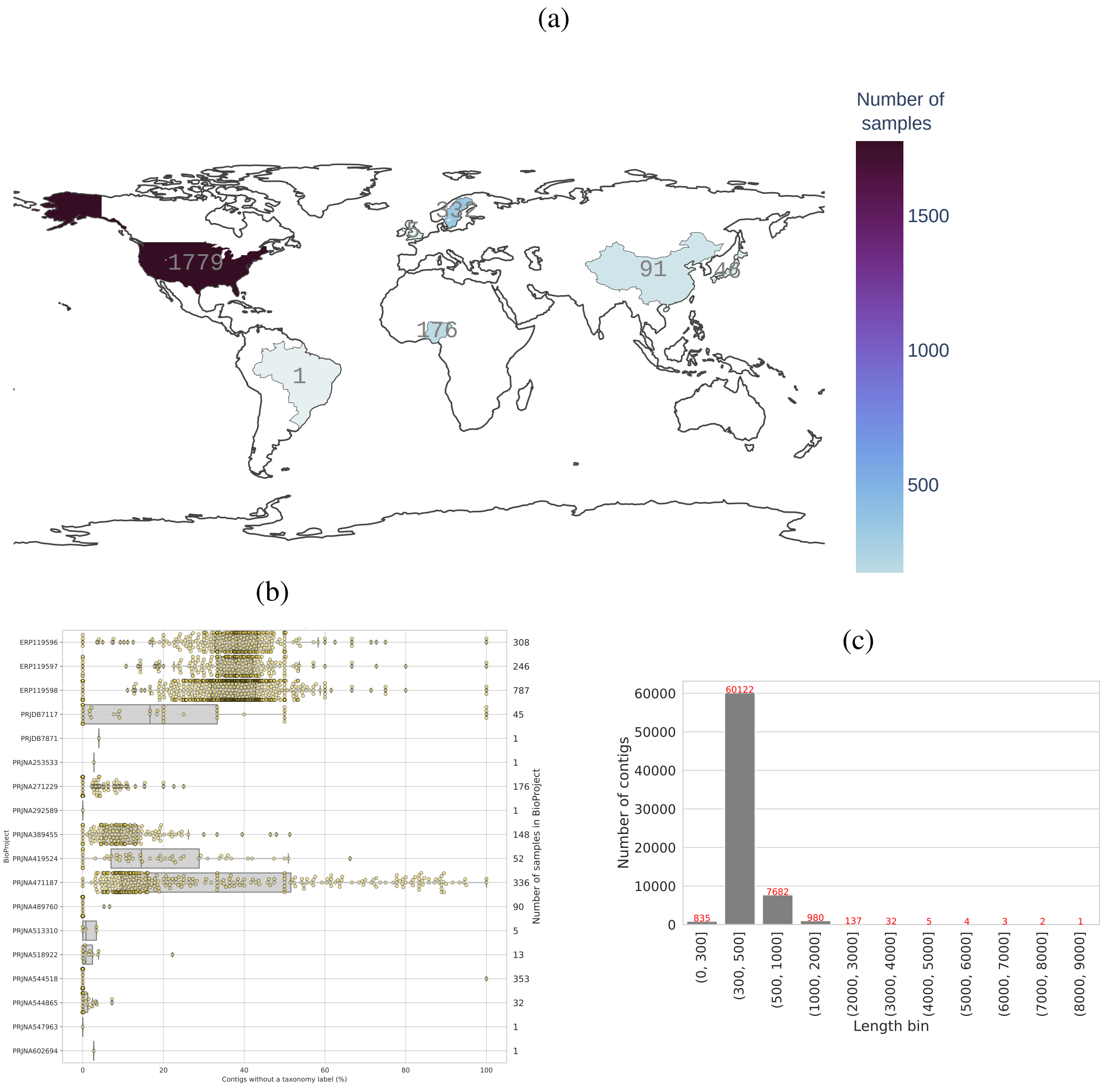


Figure 4.3: (a) Geographical distribution of the human blood microbiome samples included in this study (n=2,430). Countries are coloured according to the number of samples with darker shades representing the higher number of samples analysed. (b) The proportion of unknown contigs in human blood microbiome samples grouped by BioProject. The distribution is shown on the X-axis with each BioProject represented on the left-hand side Y-axis. Y-axis on the right-hand side shows the number of samples in each BioProject and corresponds to the number of dots on the plot for the given BioProject. (c) Distribution of unknown contigs length across different length bins. The X-axis shows length intervals and the number of contigs is shown on the Y-axis and is annotated on top of the bar.

Figure 4.3 shows that a majority of UCs were <1kb long (n=68,640). 1,169 UCs were at least 1 kb long, from this set, 184 were found to be at least 2 kb long and 10 UCs were 5kb or longer. The largest UC found in the human blood microbiome set was 8,557 bases long and was assembled from SRR7167036. UCs from the human blood microbiome set were significantly shorter than those from the previous study described in Chapter 3 . This length distribution of UCs was compared to all contigs generated from this set to identify whether short contigs were produced due to assembly anomalies in human blood microbiome datasets. However, this was not the case and the largest known contig assembled from sample SRR8862013 was 482,734 bases long suggesting that the assembly strategy was effective.

4.4.2 Feature visualisation

To understand and visualise how features are represented across these three major classes (archaea, bacteria and viruses), dimensionality reduction techniques including Principal Component Analysis (PCA), t-distributed Stochastic Neighbour Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) were explored. Briefly, dimensionality reduction is the process of reducing the number of features to the most relevant ones (Sivarajah, 2021). These methods can compress multidimensional data into a simpler 2 or 3 dimension representation that can be used to visualise the data in a 2D or 3D plot. Although such dimensionality reduction can lead to the loss of some information, these easily interpretable 2D/3D plots can provide meaningful and important insights into the underlying patterns embedded within the feature set and how they correlate with the data points or classes e.g. identification of clusters within the data.

This analysis was carried out using the May 2020 dataset that comprised archaea (n=1,268), bacteria (n=5,441) and virus (n=7,143) reference sequences (downloaded in May 2020). The September 2019 dataset was not deemed suitable for this analysis as it was deemed highly skewed and contained a very large number of archaea and bacteria observations (Table 4.1). PCA projects data onto linear hyperplane/directions that explain the most amount of variance as it is a variance maximiser method. Principal components one and two are plotted against each other in figure 4.4(a) that demonstrates reasonably tight clustering between all three classes which suggests that it is not possible to explain the variance embedded within the dataset using the first 2 principal components. Although it is recommended to keep the principal components to 2 or 3 to visualise the data, in this case, explained variation per principal component was 0.56, 0.08 and 0.03 respectively for the first three components. In such a case, a cumulative explained variance is explored. Cumulative explained variation for 50 principal components added up to 0.93 suggesting that up to 93% variance could be explained by 50 features identified by the PCA. t-SNE is deemed suitable for high dimension datasets as it tries to cluster data points by keeping similar data points together and dissimilar data points apart (Maaten, 2021). This approach can provide meaningful clusters as the underlying relationship between data points is preserved within the embeddings. Unlike PCA which relies on linear relationships, t-SNE

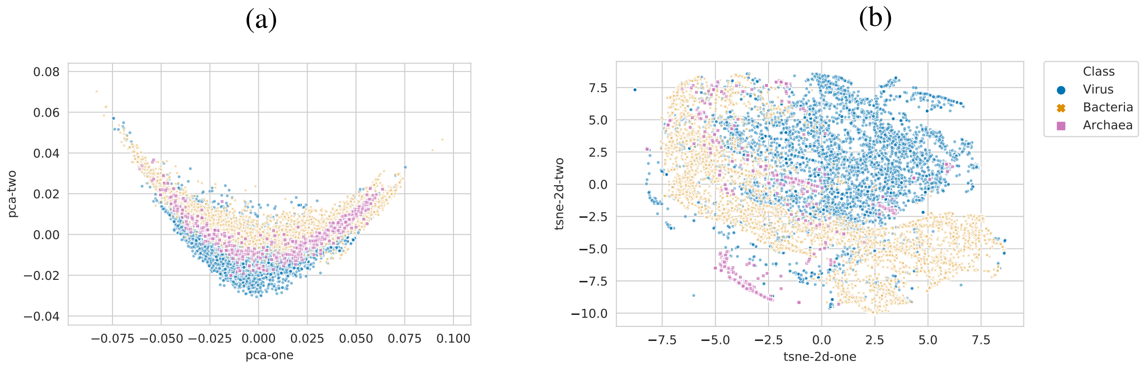


Figure 4.4: Two popular dimensionality reduction and feature data visualisation plots are shown here that were generated for May 2020 dataset and contained 3 classes; archaea, bacteria and virus. Different colours and shapes are used to denote each class. Archaea is represented with pink squares, bacteria are shown as orange crosses and viruses are shown as blue circles. (a) Principal Component Analysis (PCA) plot showing first and second principal components derived from May 2020 dataset. Plotting the PCA of multiple classes is a simple way to view their overall relatedness. A Principal Component (PC) is a weighted set of probes used to identify the strongest signals in the data and separate them into Principal Components (PCs). 56% of variance was explained by PC1 and 8% variance was explained by PC2. (b) t-distributed Stochastic Neighbour Embedding (t-SNE) plot generated from May 2020 dataset. As a dimensionality reduction technique, t-SNE plots are a visual way to simplify very large datasets.

can capture non-linear relationships. A t-SNE plot generated from this data is shown in 4.4(b). It shows some clustering within viruses and bacteria whereas archaea sequences did not show similar patterns of clustering. It also highlighted that viruses have the largest variance among the three classes included here. This is expected as viruses are known to mimic host genome signatures (Pride et al., 2006; Babayan et al., 2018). Overlapping points in t-SNE built using all features could potentially indicate the low classification potential between the three classes. However, these results were preliminary and should not be over-interpreted as they are highly sensitive to parameter optimisation and tuning.

4.4.3 Evaluation metrics

There are various metrics that can be applied to evaluate the performance of the predictions made by a model. Some of the most popular metrics are defined in Box 4.4.3.

Box 4.4.3: Model evaluation metrics

- **True positive (TP)**: Correctly identified positive class
- **True negative (TN)**: Correctly identified negative class
- **False positive (FP)**: Negative class incorrectly predicted as positive

- **False negative (FN)**: Negative class missed
- **Precision**:
$$Precision = \frac{TP}{TP+FP}$$
- **Recall (Sensitivity)**:
$$Recall = \frac{TP}{TP+FN}$$
- **True negative rate (Specificity)**:
$$Specificity = \frac{TN}{TN+FP}$$
- **Accuracy**: Proportion of correct predictions made by the model.
$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$
- **Confusion matrix**: A table comparing the known outcome to the predicted outcome generated from a model/classifier.
- **F1-score**: Harmonic mean of precision and recall.
$$F1-score = 2 * \frac{Precision * Recall}{Precision + Recall}$$
- **Support**: The support is the actual number of instances of the class in the dataset.
- **Micro average**: Micro average value is computed by taking the unweighted mean of all the per-class scores. All classes are treated equally regardless of their support values.
- **Weighted average**: Weighted-average scores are calculated by taking the mean of all per-class scores while taking into account each class's support. Weight refers to the proportion of each class's support in relation to the total.
- **K-fold cross-validation**: A technique to evaluate ML model performance by training/testing several ML models on the data that is divided into k subsets.
- **Receiver operating characteristic (ROC) curve**: Receiver operating characteristic curve that is plotted with TPR (true positive rate) against the FPR (false positive rate) where TPR ($Recall$) is on the y-axis and FPR ($1-Specificity$) is on the x-axis.

- **Area Under Curve (AUC)**: AUC represents the degree of separability represented by the ROC curve. It indicates how well the model can distinguish between classes. Higher AUCs result in more accurate class separation.

4.4.4 Multilabel multiclass modelling

Initially, a multilabel multiclass model was developed using archaea, bacteria and virus datasets downloaded in September 2019. This multiclass model attempted to place each sequence into one of the three classes; archaea, bacteria or virus. As shown in the table 4.1, September 2019 dataset used in these models was imbalanced meaning all classes did not have an equal number of observations. As this was a multiclass imbalance learn the problem, random forest models were deemed suitable to be applied to this data. As shown in table 4.2, this model performs well despite the skewed dataset. It is notable that the recall for the virus class was the lowest due to the nature of the underlying data i.e. lowest number of data points for class virus.

To address the imbalance class issue, two different data manipulation methods were explored; undersampling major classes and oversampling of minor classes to match the number of records and recreate a more balanced dataset. Undersampling methods achieve a balanced dataset by removing additional data points from majority classes whereas oversampling methods achieve a balanced set by generating multiple copies of records from the minority class. Undersampling and oversampling methods were applied to this dataset to generate an artificially "balanced" dataset. As shown in figure 4.5(a), the September 2019 dataset is highly skewed with around 54,890 observations representing archaea class and 12,148 observations included for virus class. These classes were balanced using undersampling and oversampling techniques. Figures 4.5(b) show the observations in each class reduced to match the smallest class i.e. viruses by undersampling and 4.5(c) shows the number of observations in each class increased (oversampled) by duplicating them to match the number of observations in the largest class. Model performance after oversampling and undersampling datasets are shown in table 4.2. These results were shown to improve accuracy and recall for minority classes without impacting the classification of the majority class. Although these methods are powerful they have limitations. Undersampling can lead to loss of information as valuable observations relevant to the majority classes are discarded whereas oversampling can lead to overfitting of the model. Moreover, with oversampling approach, it is better to split the data into train/test split before copying the observations in order to ensure that model is not learning and testing on the same data points (overfitting).

Table 4.2: Classification report for multiclass multilabel RFC models.

	precision	**recall**	**F1-score**	**support**	**Model**
archaea	0.96	1.0	0.98	16455.0	Sept2019_RFC
bacteria	0.99	0.97	0.98	8555.0	Sept2019_RFC
virus	0.96	0.84	0.9	3703.0	Sept2019_RFC
accuracy	0.97	0.97	0.97	0.97	Sept2019_RFC

macro avg	0.97	0.94	0.95	28713.0	Sept2019_RFC
weighted avg	0.97	0.97	0.97	28713.0	Sept2019_RFC
archaea	0.96	0.95	0.95	3680.0	Sept2019_RFC_Undersample
bacteria	0.97	0.96	0.97	3628.0	Sept2019_RFC_Undersample
virus	0.93	0.96	0.94	3626.0	Sept2019_RFC_Undersample
accuracy	0.95	0.95	0.95	0.95	Sept2019_RFC_Undersample
macro avg	0.95	0.95	0.95	10934.0	Sept2019_RFC_Undersample
weighted avg	0.95	0.95	0.95	10934.0	Sept2019_RFC_Undersample
archaea	0.96	0.95	0.96	3680.0	Sept2019_RFC_Oversample
bacteria	0.98	0.96	0.97	3680.0	Sept2019_RFC_Oversample
virus	0.93	0.96	0.94	3680.0	Sept2019_RFC_Oversample
accuracy	0.96	0.96	0.96	0.96	Sept2019_RFC_Oversample
macro avg	0.96	0.96	0.96	11040.0	Sept2019_RFC_Oversample
weighted avg	0.96	0.96	0.96	11040.0	Sept2019_RFC_Oversample

4.4.5 Data partitioning and model optimisation

An overview of the datasets used in the model predictions is shown in the figure 4.1 and supplementary table B.1. Two binary classification approaches namely SVM and random forest classifier (RFC) were applied to the January 2021 dataset (Jan2021). For SVM models, a standard scaling step was applied to the data before it was submitted to the SVM models. In the first instance of this model's development, the datasets were split into the train/test category first with a ratio of 70:30. From the training set, all observations of a class were used as positive labels and the other observations from the other classes were used as negative labels. Results from these models shown in the figures 4.7-4.10 demonstrate that these binary classification models were based on imbalanced data. Despite the imbalance dataset, these models yielded good predictions for each class.

For each class and classification model, raw and normalised confusion matrices were generated. For class archaea, 877 positive and 12,285 negative observations were included. Prediction accuracy, area under the curve (AUC) and average Precision (AP) for this set were 0.98-1.00 for SVC and RFC models. Recall for predicting the archaea as a positive label was 0.94 for RFC and 0.93 for SVC as shown in confusion matrices in figures 4.7. 10-fold cross-validation yielded the accuracy of 99.07% and 99.14% for RFC and SVC respectively for class archaea. It is worth noting that the F1-score for the same was around 0.189 for both SVC and RFC. Such contrast between accuracy and F1-score indicates that the model was accurately predicting the negative class i.e. not archaea efficiently but was struggling to predict the positive observation. This is due to the lower number of positive observations. When the dataset is split into 10 different chunks for cross-validation, some chunks could have very few or no observations of smaller (i.e. archaea) classes included in them which would in turn lead to an inaccurate representation of the prediction signal if only accuracy values are compared. This also highlights the importance of using a more appropriate model performance metric such as the

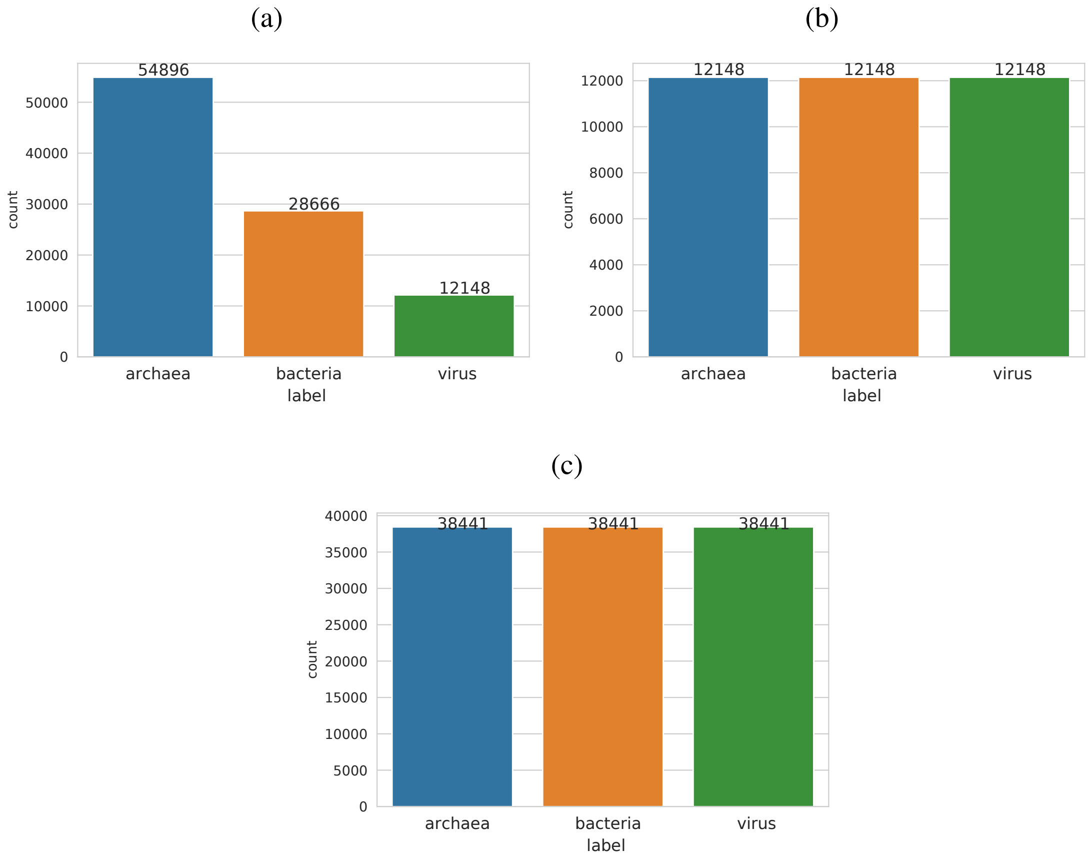


Figure 4.5: Number of observations included in each class is shown in the bar chart. The X-axis shows the class and Y-axis shows the total number of observations in each class. (a) A bar chart showing the raw number of observations for each class in the September 2019 dataset. (b) A bar chart showing the number of observations reduced to match the smallest class with 12,148 observations after undersampling techniques were applied to the September 2019 dataset. (c) A bar chart showing the number of observations duplicated (oversampled) to match the total number of observations in the train split of the largest class (n=38,441).

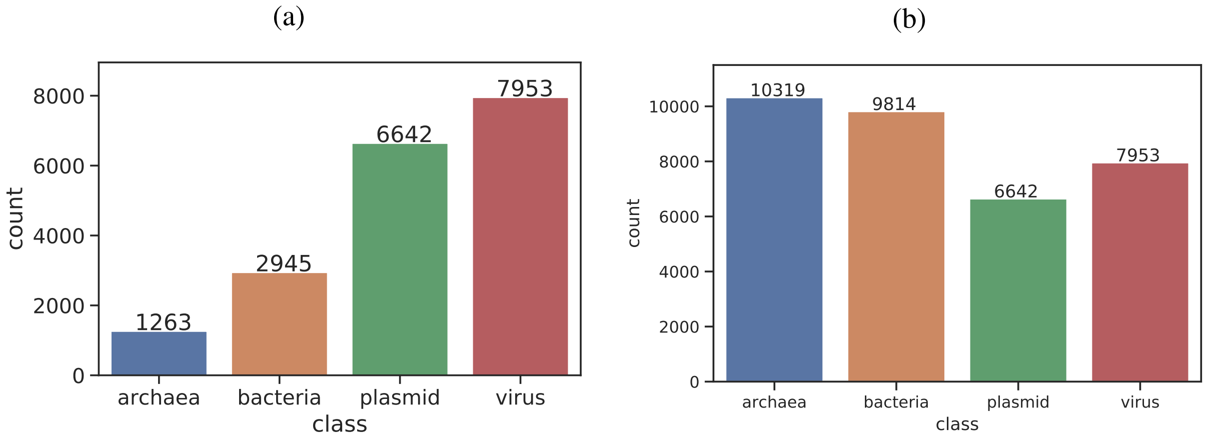


Figure 4.6: Number of observations included in each class is shown in the bar chart. The X-axis shows the class and Y-axis shows the total number of observations in each class. (a) A bar chart showing the raw number of sequence observations included in each class in the January 2021 dataset was used to develop binary classification models. (b) A bar chart showing the January 2021 dataset with chopped bacteria and archaea genome sequences. Briefly to achieve a relatively similar number of observations in each class, sequences from classes with a smaller number of observations i.e. bacteria and archaea - were split into non-overlapping fragments to increase the number of observations without introducing bias in the dataset.

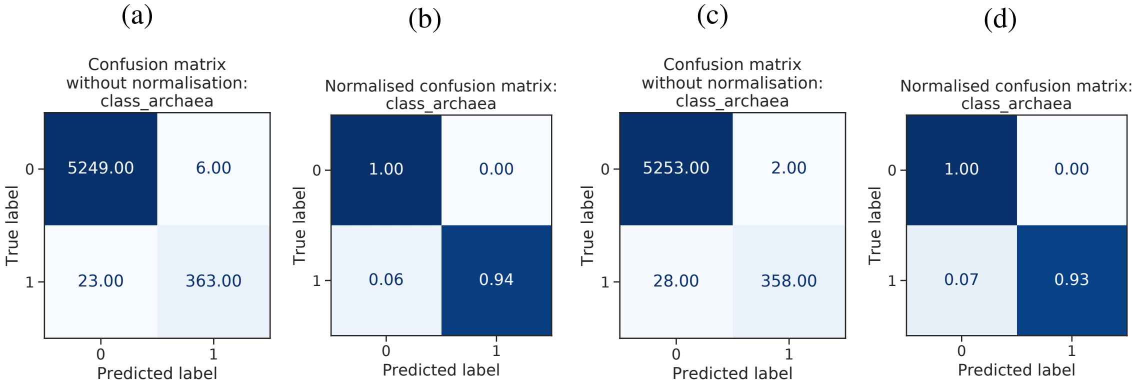


Figure 4.7: Binary model predictions results were obtained for class archaea using January 2021 dataset. A confusion matrix for SVC performance is shown with (a) raw data and (b) normalised data. Similarly, the confusion matrix showing the prediction results for RFC is shown in (c) raw data and (d) normalised data. Note: As matrix values are rounded up to two floating points, it may lead to minor round errors if the numbers are too small. The normalised data represents the raw counts that are transformed into proportions.

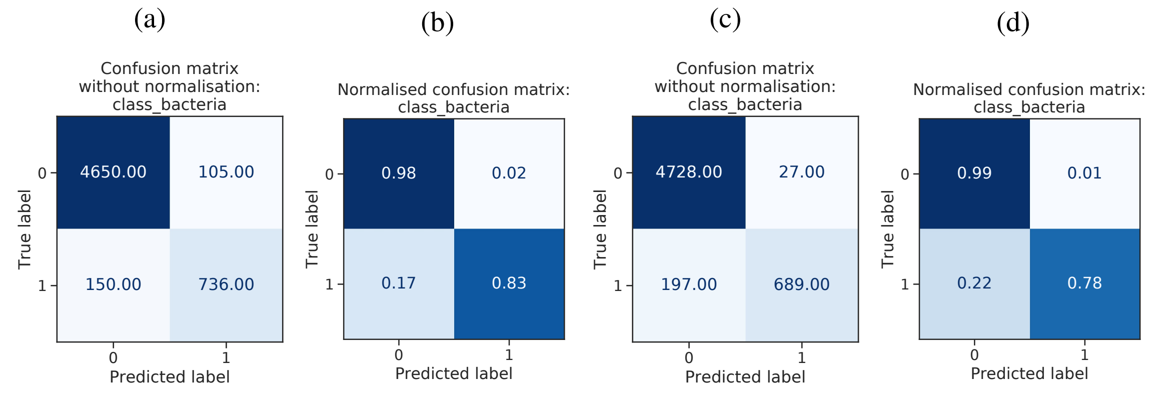


Figure 4.8: Binary model predictions results obtained for class bacteria using January 2021 dataset. A confusion matrix for SVC performance is shown with (a) raw data and (b) normalised data. Similarly, the confusion matrix showing the prediction results for RFC is shown in (c) raw data and (d) normalised data. Note: As matrix values are rounded up to two floating points, it may lead to minor round errors if the numbers are too small. The normalised data represents the raw counts that are transformed into proportions.

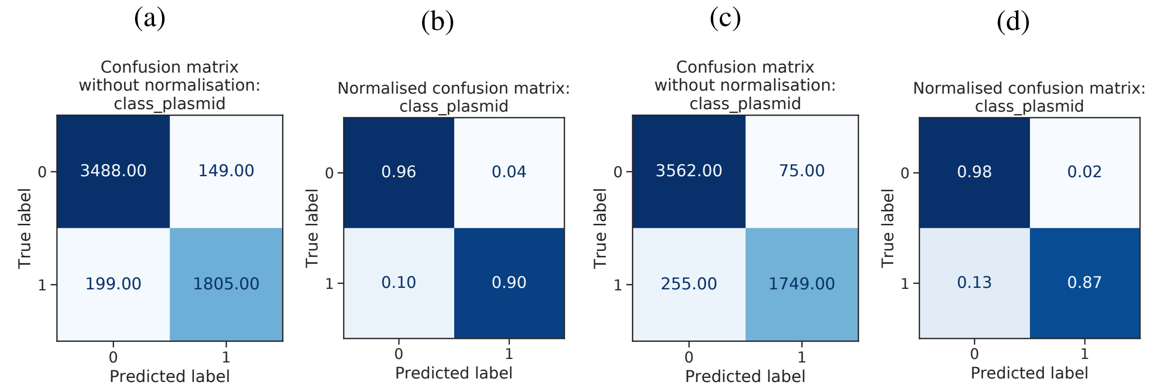


Figure 4.9: Binary model predictions results obtained for class bacteria using January 2021 dataset. A confusion matrix for SVC performance is shown with (a) raw data and (b) normalised data. Similarly, the confusion matrix showing the prediction results for RFC is shown in (c) raw data and (d) normalised data. Note: As matrix values are rounded up to two floating points, it may lead to minor round errors if the numbers are too small. The normalised data represents the raw counts that are transformed into proportions.

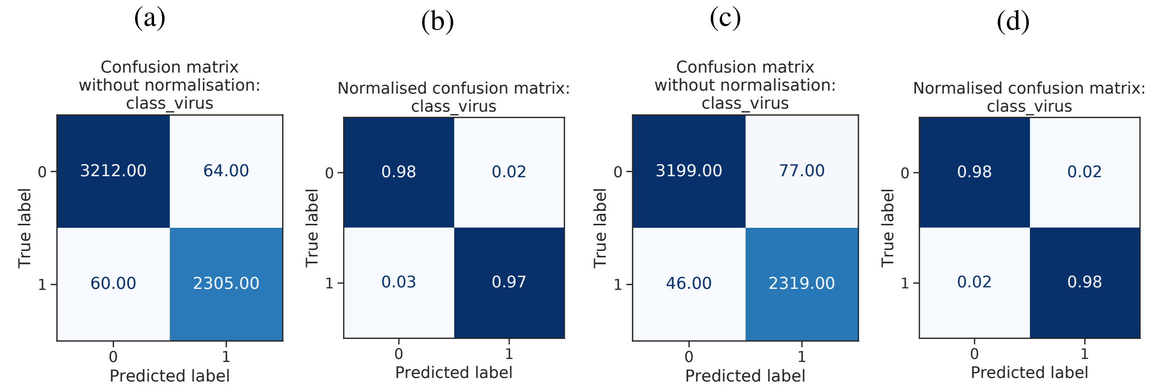


Figure 4.10: Binary model predictions results obtained for class bacteria using January 2021 dataset. A confusion matrix for SVC performance is shown with (a) raw data and (b) normalised data. Similarly, the confusion matrix showing the prediction results for RFC is shown in (c) raw data and (d) normalised data. Note: As matrix values are rounded up to two floating points, it may lead to minor round errors if the numbers are too small. The normalised data represents the raw counts that are transformed into proportions.

F1-score that provides further insights into the precision and recall for binary classification models.

In the case of class bacteria, 2,059 positive and 11,103 negative data points were included. Although these models have high accuracy of 0.95 for SVC and 0.96 for RFC, recall for positive class bacteria was 0.83 and 0.78 respectively for SVC and RFC (figure 4.8). Lower recall values implied that true bacterial sequences were being missed by these models. The average accuracy for RFC for bacteria after 10-fold cross-validation was 92.88% for RFC and 92.30% for SVC. The mean F1-score was 0.152 for RFC and 0.163 for SVC. For class plasmid, 4,638 positive and 8,524 negative data points were used. In the case of the plasmid class, an accuracy of 0.94 was obtained for both classifiers whereas the sensitivity was 0.90 for SVC and 0.87 for RFC (figure 4.9). The average accuracy for 10-fold cross-validation was around 89.54% and the average F1-score was around 0.356 for RFC models whereas for SVC models mean accuracy was 90.54% and the mean F1-score was 0.373. The virus class contained 5,588 positive and 7,574 negative observations. Accuracy, AUC and AP were between 0.98-1.00 for both classifiers (figure 4.10). The average accuracy of 95.95% with an average F1-score of 0.486 was obtained for RFC after cross-validation. Similarly, an average accuracy of 96.47% and an average F1-score of 0.486 was obtained for SVC for viruses.

These results demonstrate that the features implemented in these models are able to distinguish all four classes with high precision and recall. The cross-validation results obtained for RFC highlight the problems with imbalanced data being used to train the models. These models lack robustness for classes with a smaller number of observations i.e. archaea and bacteria. To overcome this issue, further efforts were made that are discussed in detail below.

Notably, both classification method predictions and performance were comparable for all classes. This suggested that either method would be suitable for this specific classification problem. The wall time for training and testing models using SVC was 19 minutes 24 seconds compared to RFC which took 17.7 seconds. This is due to the ability of RFC-based models to be run in parallel using the `n_jobs` parameter in sklearn. Moreover, RFC-based models were chosen for further analysis as they are natively probabilistic which means that it is easier to extract probabilities for each classification in addition to the final classification outcome. Probability calibration methods such as Platt scaling can be used to convert class outcomes to probabilities in deterministic models like SVC.

An alternative data partitioning approach was explored whereby genome sequences from smaller classes, namely bacteria and archaea, were split into non-overlapping chunks to inflate the number of sequences without introducing any bias in the dataset. This dataset is referred to as Jan2021_chunks from hereon. A number of k-mer classification tools like VirFinder and (Ren et al., 2017), DeepVirFinder (Ren et al., 2020) have successfully implemented this approach in their models. As features are ultimately normalised for the length, these models are assumed to be not sensitive to the length of the sequence. Hence, splitting the larger genomes into shorter fragments can help to increase the number of observations in the class without introducing bias that is typically introduced through other techniques such as over/undersampling. To achieve this, all larger genomes (namely bacteria and archaea) were split into non-overlapping chunks of a given size. The chunk size was selected to roughly match up the number of observations from different classes. A chunk size of 1mb (for bacteria) and 200kb (for archaea) were arbitrarily selected to roughly match up the number of records in each class. All k-mer frequencies were normalised for the length of the sequence such that sequence length would not bias the features and the observed frequencies of the feature set. This generated a less imbalanced class compared to the previous dataset (figure 4.6).

Additionally, calibration methods were also explored to calibrate the prediction probabilities of random forest-based classifiers. The predicted probabilities of a well-calibrated classifier can be directly interpreted as confidence levels. For example, a well-calibrated binary classifier should classify the samples such that among the samples to which it gave a predict_proba value close to 0.9, approximately 90% actually belong to the positive class (Sklearn, 2021). Brier score, which compares the actual probability with the predicted probabilities was used to measure the accuracy of the predictions. A Brier score of 0 indicates complete/total accuracy and a value of 1 indicates completely inaccurate predictions. In current models, the train/test split is always performed on the fly and before separating the classes which in turn results in a dynamic set of training and testing data points. In general, these minor changes should not make a huge difference to the actual results of the classifiers. This was demonstrated and tested using 10-fold cross-validation. Additionally, a predefined random seed was used that can ensure the reproducibility of the results.

Overall results for the Jan2021_chunks dataset were comparable to those described for the Jan2021 dataset. 10-fold cross-validation carried out using these datasets indicated average accuracy of 99.11% for archaea (average F1-score: 0.472), 91.33% for bacteria (average F1-score: 0.264), 90.26% for plasmid (average F1-score: 0.217) and 95.27% for virus (average F1-score: 0.276) classification using RFC. Furthermore, the average F1-score for archaea and bacteria was improved suggesting better prediction ability with more data points for the minority class. However, these additional archaea and bacteria observations negatively impacted the F1-score of plasmid and virus classes suggesting that a further improvement was required.

Brier's score of calibration for archaea was 0.005 before calibration and was improved slightly to 0.003 after calibration albeit the initial model was well-calibrated. For bacteria, the Brier score was 0.033 for the uncalibrated classifier and could be improved to 0.024 after sigmoid calibration was applied. In the case of the plasmid class, the sigmoid calibrated classifier's Brier score was 0.032 compared to the uncalibrated classifier with a Brier score of 0.041. Finally, for the virus class, the Brier score of the sigmoid calibrated classifier was 0.012 improved from 0.022 to its uncalibrated counterpart.

These binary models work well in the given scenario, however, they could be slightly biased towards predicting the negative label. This is simply due to the fact that the model has "seen" more negative labels compared to the positive ones. To improve these models further, alternative balanced data strategies were explored. This involved a more traditional binary classification approach whereby for each class positive labels are separated and an equal number of negative data points were drawn resulting in a balanced dataset.

Balanced binary classification, cross-validation and calibration

Balanced binary RFC models that include the same number of positive and negative observations for each class were trained, tested and validated on Jan2021 and Jan2021_chunks datasets. Overall, the results obtained for both datasets were similar with slightly better results for the Jan2021_chunks datasets simply because of a higher number of observations included in the set for bacteria and archaea classes and that positive and negative observation were balanced for each class. All RFC binary models were calibrated using a sigmoid calibration and 10-fold cross-validation. Performance metrics of the final balanced, calibrated and cross-validated models are shown in the table 4.3. Average accuracy from 10-fold cross-validation (CV) was obtained for each class. The highest accuracy of 99.68% and AUC of 1.00 were achieved for class archaea. The Brier score for the archaea was the lowest among all binary classifiers. For class bacteria and plasmid, a significant improvement was observed from the previous unbalanced classification. For bacterial classification, the overall accuracy of 96.48% and F1-score of 0.96 were achieved. In the case of the plasmid class, an accuracy of 94.22% was achievable with F1-score of 0.93. These results show significant improvement from the previous models where though the accuracy was very high the F1-score denoting the precision and recall for each class was not very high. Balancing the number of observations used for training and cross-validation steps demonstrated greater prediction precision and recall for each class. Finally, for the virus class, an average accuracy of 97.74% and F1-score of 0.97 were achieved. These final classifiers were tested further using the holdout datasets. These models are implemented as final models and are shared through TetraPredX.

Table 4.3: Balanced, binary calibrated and cross-validated model performance metrics

Class	**N**	**AUC**	**F1-score**	**Accuracy**	**Brier score§**
Archaea	10319	1.00	0.99	99.68%	0.005
Bacteria	9814	0.99	0.96	96.48%	0.029
Plasmid	6642	0.98	0.93	94.22%	0.050
Virus	7953	1.00	0.97	97.74%	0.020

Classification of holdout dataset

To test and validate the predictions made by the final calibrated binary classifiers, four distinct holdout datasets were designed. These are as follows: 1) *Geminiviridae* 2) *Chuviridae* 3) *Siphoviridae* and 4) contigs and scaffold level assembly of *Paraburkholderia madseniana* strain RP11. Three out of four datasets were specific to virus families and one was for a novel bacteria strain. To perform these experiments, each group of sequences were excluded (held out) from the RFC training set and the resulting models were tested on the holdout set. This was to evaluate if a previously unknown set of sequences could be identified using the classifiers. Probability-based classification results were divided into two categories: 1) maximum/highest probability-based classification and 2) signal-based classification. Maximum probability-based classification considers the highest probability of the given sequence being classified in one or more microbial classes whereas the signal-based classification would simply show the classes for which the predicted probability is >0.5. For example, if a contig was predicted to belong to the class virus with a probability score of 0.7 and class plasmid with a probability score of 0.65 then it will be classified as a virus in the highest probability-based method and will be classified as a virus/plasmid using the signal-based method. This is due to the fact that there are independent signals (i.e. probability score >0.5) that this contig could belong to either class with different probability scores. Any sequences that were predicted with probability <=0.5 were categorised as "undetermined" indicating that there wasn't sufficient signal to predict their class using our models with >50% confidence.

Prediction results for *Geminiviridae* and *Chuviridae* are shown in the figure 4.11. It is notable that most sequences for both holdout sets were predicted to belong to the virus class. The sequences from the family *Geminiviridae* were predicted as viruses with a mean probability of 0.991 (SD:0.008). Moreover, the mean probability for these sequences to be any other class was between 0.004-0.016. For the family *Chuviridae*, the mean probability for the virus class was 0.991 (SD:0.022) and the mean probability for all other classes was between 0.007-0.016. These results suggest that the virus-specific signals are strong enough to be picked up by the model even though the models were trained using the sequences that were left out of the training set. Compared to these two virus families, predictions made for the family *Siphoviridae* were different as shown in the figure 4.11(c). Out of the complete set of 785 sequences, 298 could not be predicted to belong to any class, 401 sequences were predicted to belong to the virus class, 55 were predicted to belong to the plasmid class, 29 were predicted to belong to the class bacteria and 2 were predicted to be archaea. However, the classification described above was based on the absolute highest probability where a class probability was >0.5. If the "signal" (i.e. where predicted probabilities for the class was >0.5) was considered then the classification was as follows: 387 were viruses, 46 were plasmid, 27 were bacteria and 2 were classified as archaea, 20 were predicted to be virus/plasmid, 3 were predicted to virus/bacteria and 2 were predicted to be bacteria/plasmid. These mixed signals can be observed in figure 4.11(c) where the predictions are plotted for each class with the highest probabilities. Notably, these models struggled with the classification of siphoviruses. Bacteria and archaea serve as natural hosts for siphoviruses, and viruses often mimic their host's genomic signature. Tetramer frequencies have been shown to be predictive of the host for prokaryotic viruses (Young et al., 2020). The removal of siphovirus sequences would have resulted in the loss of key distinctive signals used by models to distinguish viruses from other classes, which could severely influence the prediction abilities of models.

Predicted probabilities and model-based classification for *Paraburkholderia madseniana* strain RP11 sequences are shown in the figure 4.12. Out of the set of 284 sequences, 21 could not be predicted to belong to any of the classes using the calibrated RFC models developed here i.e. did not have a predicted probability >0.5 (figure 4.12). Sequence classification based on the maximum predicted probabilities was as follows: 133 as bacteria, 70 as plasmid and 60 as a virus. This is shown in the figure 4.12. However, this set of predictions also contained "mixed" classification signals whereby more than one class had >0.5 probabilities for sequence classification. Based on that "signal", 120 sequences were predicted to be bacteria, 42 were predicted to be plasmid and 54 were predicted to be of virus class. In the remaining set, 38 sequences had >0.5 probabilities for bacteria as well as plasmid class, 8 sequences had probabilities >0.5 for plasmid/virus classes and 1 sequence was in bacteria/virus classes. These results imply that some of the incomplete bacterial genome sequences are difficult to predict accurately using the TNF-based models developed here.

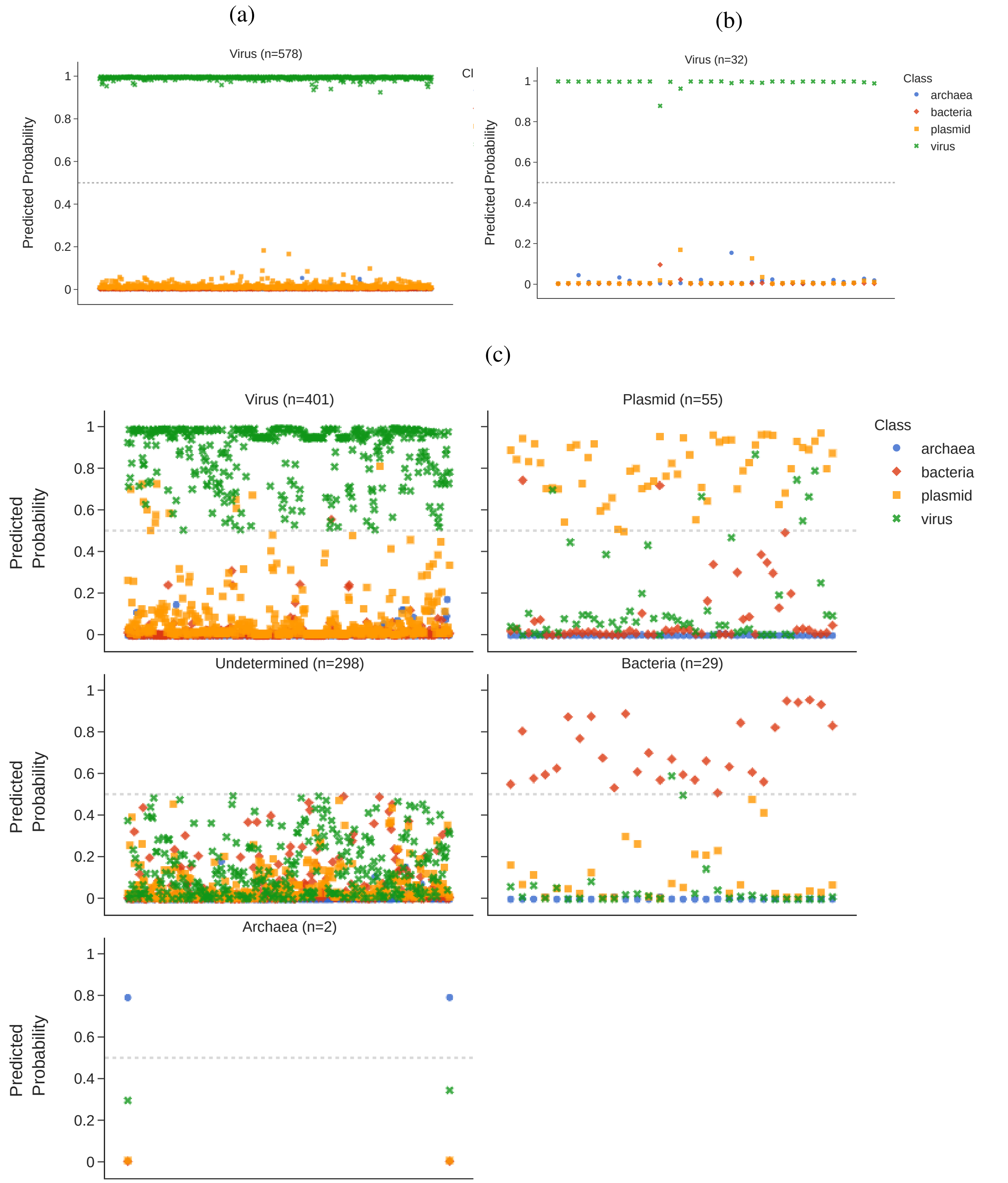


Figure 4.11: Virus holdout dataset prediction results obtained using Jan2021_chunk models are shown in various scatter plots. Predicted probabilities were plotted for each instance of for holdout dataset. The X-axis represents the query sequence, Y-axis represents the probability for each model and the dotted grey line represents the probability threshold of 0.5. The colours and shapes denote individual classes. Class archaea is represented with blue circles, class bacteria is shown in red diamonds, yellow squares represent the class plasmid and green crosses represent the class virus. (a) Prediction results are shown for the holdout set for the virus family *Geminiviridae* (n=578). These results show that the Jan2021_chunk model predicted all sequences in this dataset as viruses with very high probabilities. (b) Prediction results for virus family *Chuviridae* (n=32). All 32 sequences were predicted to belong to the class virus using Jan2021_chunk models. (c) Prediction results for virus family *Siphoviridae* (n=784). Each panel shows the class and the corresponding number of sequences predicted to belong to that class based on the maximum prediction probability. 298 sequences could not be classified accurately (probability <=0.5) using Jan2021_chunk models.

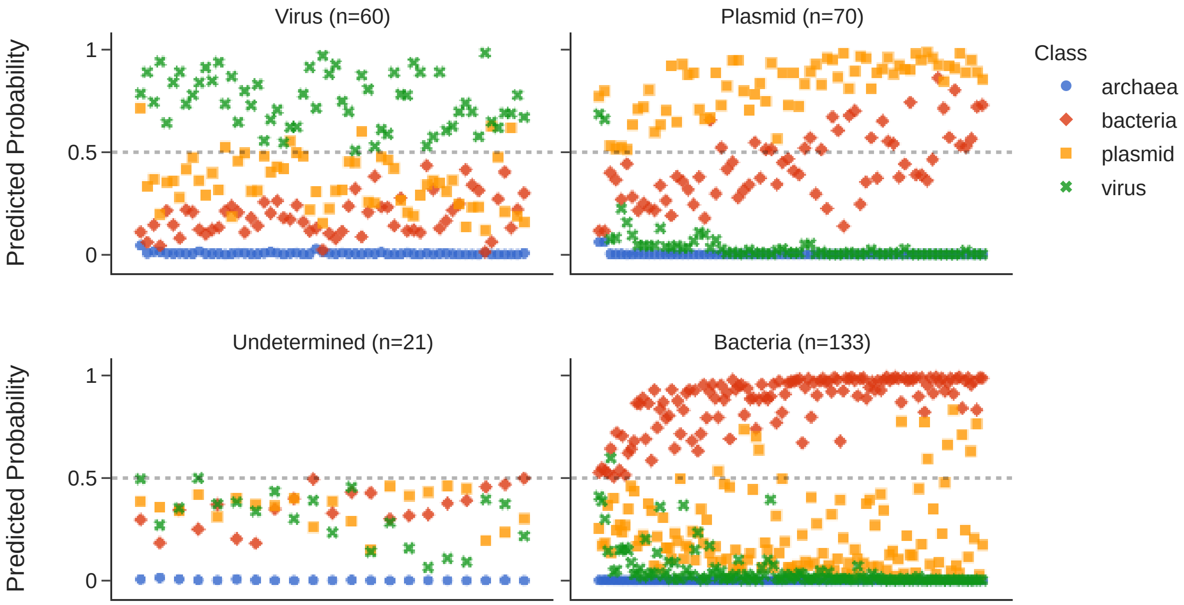


Figure 4.12: Prediction results for holdout set containing sequences from bacteria *Paraburkholderia madseniana* strain RP11 predictions based on the highest prediction probability. This set contained 284 contig/scaffold sequences. Each panel shows a class and the corresponding number of sequences predicted to belong to that class based on the maximum prediction probability.

Reverse complement features

Typical metagenomic sequencing involves fragmentation of the genomic content and these fragments are then sequenced using the high throughput sequencing approaches. The *de novo* assembly process often leads to the contig assembly in the opposite orientation. To efficiently predict the origin of such contigs, it is important to gather the k-mer frequencies in both the forward and reverse strands of the contig sequences. To achieve this, a minor modification to the ML models was applied whereby the tetranucleotide frequencies were calculated from both strands of the sequences for all sequences in the subsequent analyses.

These model prediction performances were not greatly affected by this tweak as shown in the supplementary table B.4. The holdout dataset results were very slightly affected by the implementation of this approach. For example, 2 of the *Geminiviridae* sequences and 1 chuvirus sequence could not be predicted accurately as shown in the figure 4.13. In the case of siphoviruses, TetraPredX models could not determine the class of the majority of sequences (n=469). However, from those that could be predicted, 195 sequences were predicted as viruses, 74 were predicted as plasmids and 43 were predicted to belong to class bacteria. Using reverse complement features helped to improve the predictions for holdout set *Paraburkholderia madseniana* with 152 sequences correctly predicted as bacteria (figure 4.14). The results were interesting since they illuminated the importance of feature selection and how k-mer frequency calculations can affect the model's performance. Here, no changes were made in the dataset but the k-mer frequencies in the feature set had been impacted due to the addition of reverse complement k-mers.

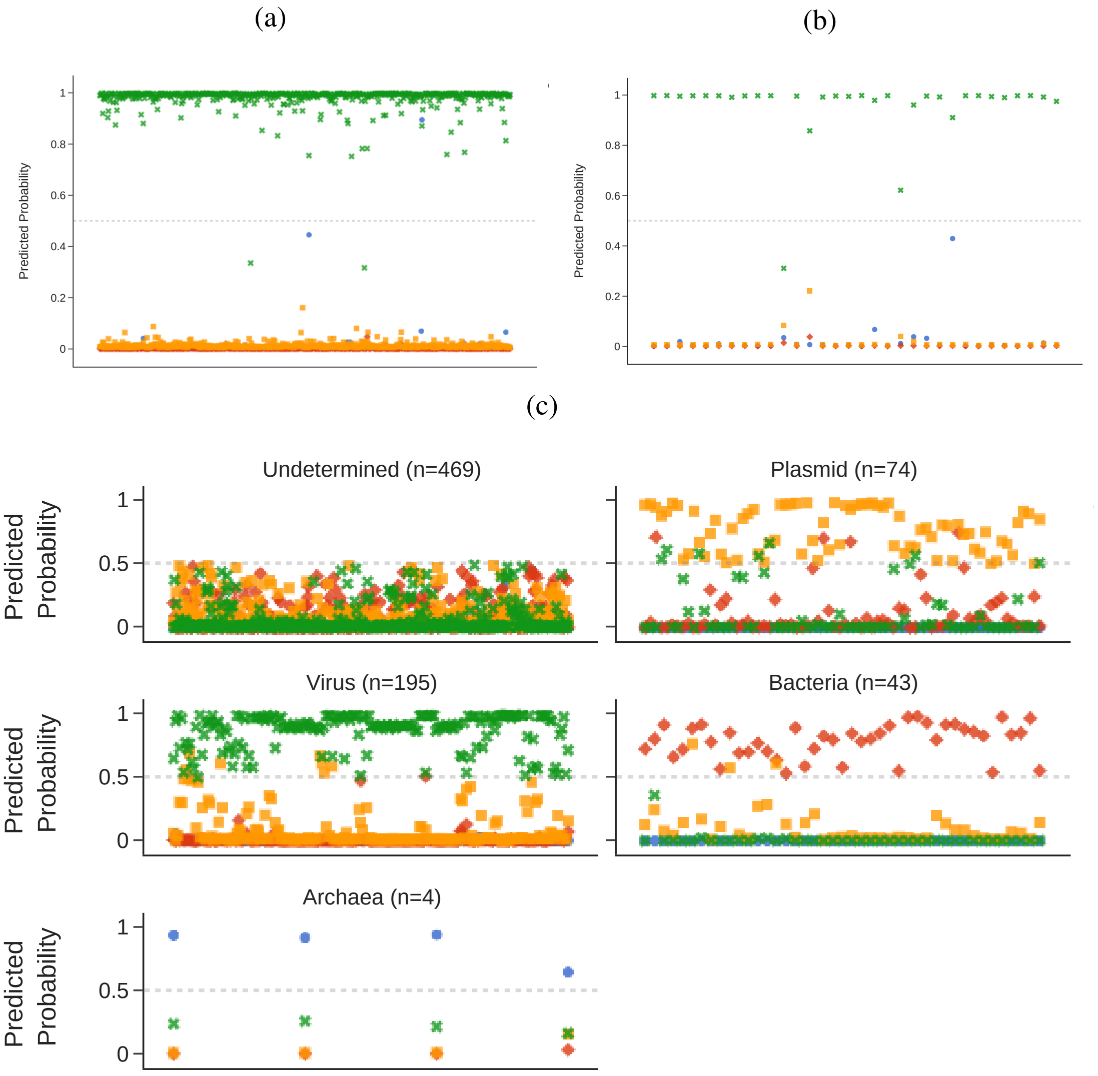


Figure 4.13: Holdout virus dataset prediction results after the introduction of reverse complement features to Jan2021_chunk models. The X-axis represents the query sequence, Y-axis represent the probability for each model and the dotted grey line represents the probability threshold of 0.5. The colours and shapes denote individual classes. Class archaea are represented with blue circles, class bacteria is shown in red diamonds, yellow squares represent the class plasmid and green crosses represent the class virus. (a) Prediction results are shown for the holdout set for the virus family *Geminiviridae* (n=578). (b) Prediction results are shown for the holdout set for the virus family *Chuviridae* (n=32). (c) Prediction results shown for the holdout set for virus family *Siphoviridae* (n=784). Each panel shows a class and the corresponding number of sequences predicted to belong to that class based on the maximum prediction probability.

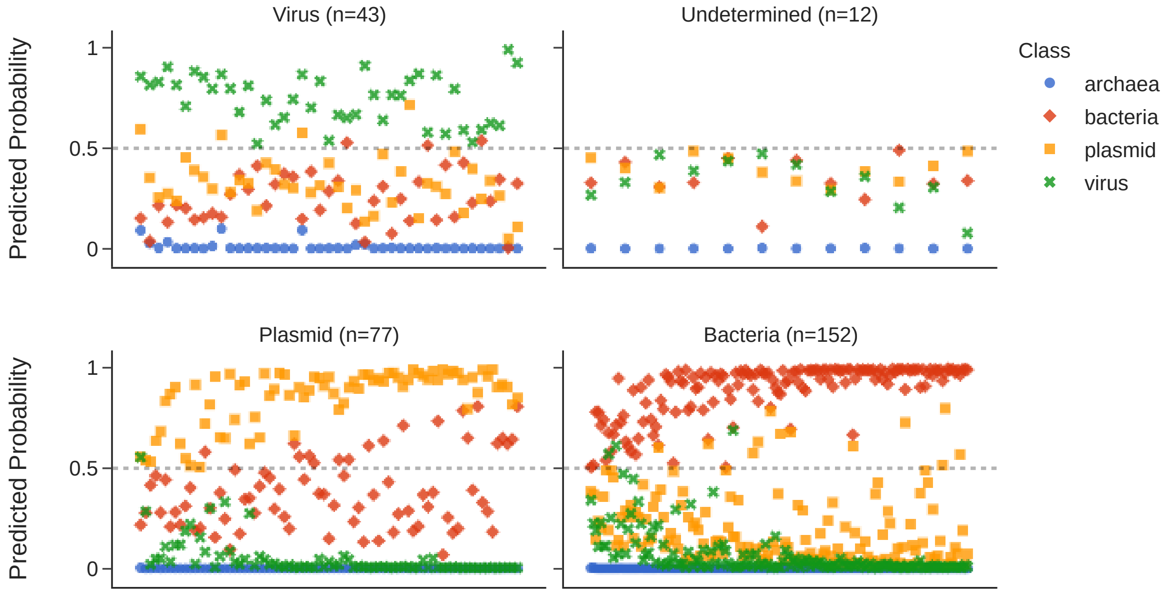


Figure 4.14: Prediction results for holdout set for bacteria *Paraburkholderia madseniana* strain RP11 predictions (n=284) after incorporation of reverse complement features in Jan2021_chunk models. Each panel in this figure shows a class and the corresponding number of sequences predicted to belong to that class based on the maximum prediction probability. The X-axis represents the query sequence, Y-axis represent the probability for each model and the dotted grey line represents the probability threshold of 0.5. The colours and shapes denote individual classes. Class archaea is represented with blue circles, class bacteria is shown in red diamonds, yellow squares represent the class plasmid and green crosses represent the class virus.

4.4.6 Virus-specific modelling

To identify if the existing models and feature set can provide further resolution on the type of virus a given sequence is, virus-specific classification modelling was undertaken. Three different taxonomic and genotypic classification properties; realm, genome type and segmentation were selected for the virus-specific classification models. Random forest binary classifiers were trained for three distinct virus properties that were extracted from ICTV VMR 010820 MSL35. This included segmentation, genome type and virus realms. For each category, individual binary prediction models were trained and tested. Results and performance metrics for balanced binary models that were calibrated and cross-validated as described in the previous sections are summarised in table 4.4.

For all four realms (*Duplodnaviria, Monodnaviria, Riboviria* and *Varidnaviria*), the average 10-fold CV accuracy was over 85% and AUC was >=0.93. All models were reasonably well-calibrated with a Brier score between 0.036 and 0.105. The highest average accuracy of 96.10% and F1-score of 0.95 were achieved for realm *Duplodnaviria*, followed by realm *Riboviria* with an F1-score 0.94 and accuracy of 94.47%. Calibrated RFC models for *Varidnaviria* performed the worst out of the set with an average accuracy of 87.85%, F1-score of 0.85 and Brier score of 0.105. This could be either due to the small number of observation sequences in this set (n=231) included in this model or that these specific model methods and/or features are not suitable for accurately predicting virus realm(s).

The genome type property was derived from the genome composition details included in the VMR. As shown in the figure 4.15, certain genome composition groups e.g. ssDNA(+), ssRNA, and dsDNA-RT contained a very low number of records thus it would not be feasible to develop models for each of these categories. To simplify this dataset, genome composition was collapsed into four distinct genome types; dsDNA (n=3,066), dsRNA (n=951), ssDNA (n=1,189) and ssRNA (n=2,695). Calibrated binary classification models were able to distinguish among these groups with an average accuracy of 94.13% for dsDNA, 93.27% for ssDNA, 89.64% for dsRNA and 92.41% for ssRNA (table 4.4). In addition to this, a balanced binary model was also developed to predict the segmentation i.e. whether a virus can be segmented or unsegmented. A total of 2,909 records of segmented viruses were used as a positive set against an equal number of unsegmented viruses to train this model which demonstrated an average accuracy of 89% with an F1-score of 0.88 to predict that the given virus record could be segmented.

These results showed that all three virus-specific properties could be predicted well using the TNF features. All models were able to achieve >85% accuracy with >0.85% F1-score suggesting that these models can be applied to virus sequences to potentially characterise their realm, genome type and segmentation properties.

To test these models further, two separate test datasets were designed. In the first test set, all virus sequences used to train these models were chopped into non-overlapping 2kb fragments and the models were used to predict realm, genome type and segmentation properties for each

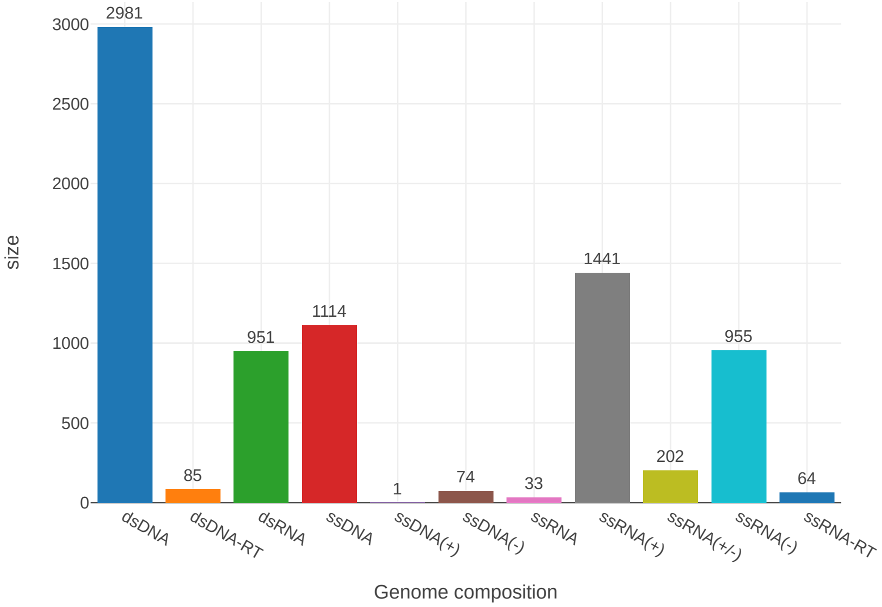


Figure 4.15: Number of viruses grouped by the genome composition derived from VMR MSL35. The X-axis shows different genome compositions and Y-axis shows the number of viruses in the corresponding genome composition category. Colours represent distinct genome composition categories.

Table 4.4: Virus taxonomy/properties-based model performance metrics

Realm	**N**	**AUC**	**F1-score (positive label)**	**Accuracy**	**Brier score**
Duplodnaviria	2051	0.99	0.95	96.10%	0.036
Monodnaviria	1233	0.98	0.93	93.27%	0.056
Riboviria	3698 (3515)	0.98	0.94	94.47%	0.045
Varidnaviria	231	0.93	0.85	87.85%	0.105
Genome type					
dsDNA	3066	0.98	0.94	94.13%	0.046
dsRNA	951	0.96	0.89	89.64%	0.078
ssDNA	1189	0.97	0.90	93.27%	0.063
ssRNA	2695	0.97	0.92	92.41%	0.062
Segmentation					
Segmented	2909	0.95	0.88	89.00%	0.089

of them. This is to test whether incomplete or fragmented virus genome sequences can be accurately predicted using these virus-specific models. 107,368 virus sequence fragments were generated. The comparison between actual, correctly predicted and incorrectly predicted label are shown in figure 4.16(a) and 4.16(b). Notably, realm *Duplodnaviria* and genome type dsDNA were significantly under-predicted for these short fragmented sequences. On the contrary, realm

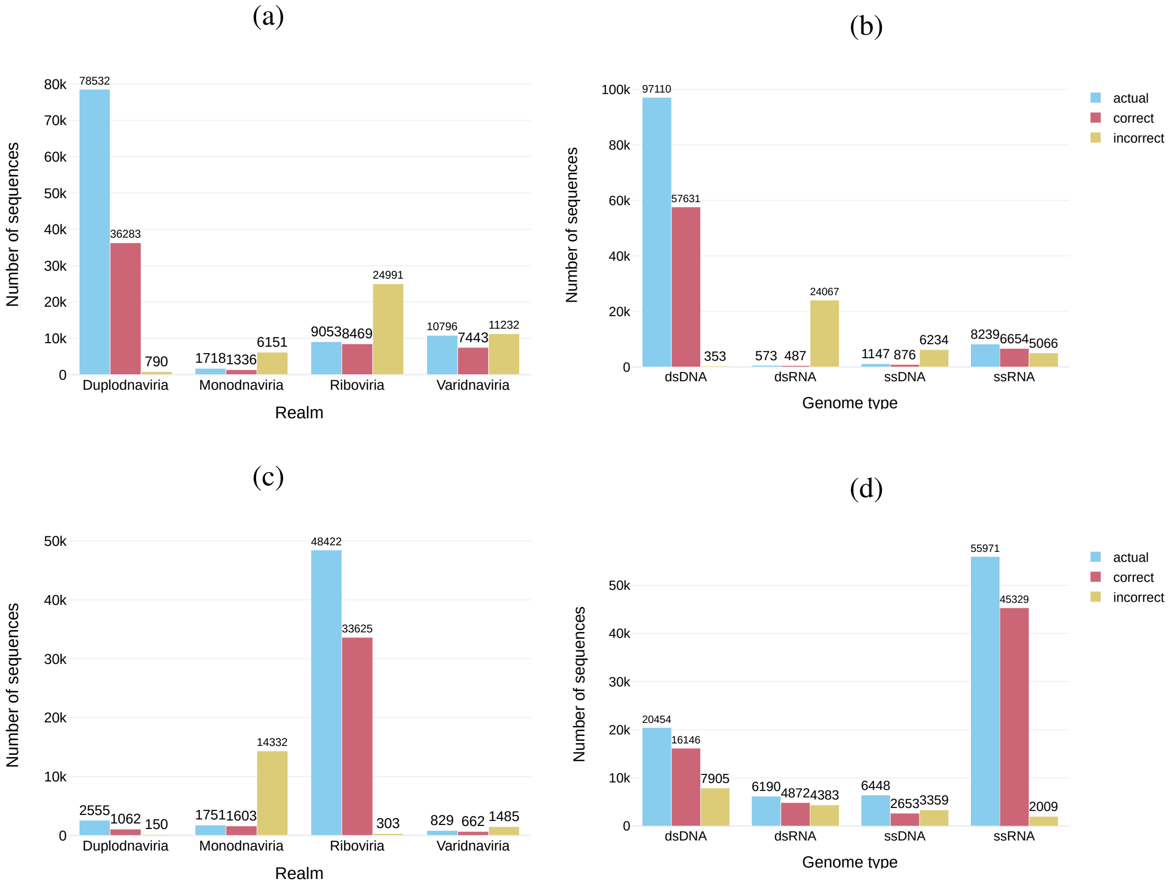


Figure 4.16: Virus-specific prediction results using TNF models developed to predict virus realm and genome type for two test datasets. (a) and (b) shows the comparisons between actual realm and genome type properties of fragmented RefSeq virus sequences; (c) Realm and (d) genome type comparisons for representative GenBank sequence subset that were not included in the initial model training and/or testing. It is notable that virus-specific TNF models are sensitive to the sequence length. In case of the RefSeq set, a large number sequences that belong to realm *Duplodnaviria* were incorrectly predicted as realms *Monodnaviria, Ribodnaviria* and *Varidnaviria*. This was also reflected in the genome type models as shown in (b). In contrast, fragmented sequences derived from GenBank were comparatively better predicted for both genome type and realm models, except for realm *Monodnaviria*. Overall, models that were developed with smaller number of observations for each of the properties e.g. *Monodnaviria, Varidnaviria*, dsRNA were shown to lack robustness with these categories of predictions.

Monodnaviria, *Riboviria* and *Varidnaviria* models were able to recover actual fragments better albeit with more false-positive predictions. Similar patterns were observed for genome type dsRNA and ssDNA.

For the second test set, all GenBank nucleotide sequences for the virus (txid10239) were downloaded from the NCBI on 10 June 2021 using NCBI e-utilities. This set comprised 4,192,727 sequences which were filtered for the length of 1kb and shorter sequences were discarded. To dereplicate these sequences and remove duplicates, these sequences were clustered using MMSeqs2 for 90% sequence similarity with at least 80% target sequence coverage. A cluster representative set with 108,045 was generated which was then filtered for any RefSeq virus genome sequences previously used to train and/or test the models. The final set comprised 102,221 sequences for which feature extraction and model predictions were carried out. The test dataset was annotated with relevant taxonomic attributes using the metadata derived from NCBI Virus Resource and ICTV VMR MSL35. 101,924 records from the test set could be mapped to relevant metadata downloaded from the NCBI Virus resource. The results for this subset for realm and genome type predictions are shown in figure 4.16(c) and 4.16(d). Realm and genome type prediction for this test set was comparable to the original classification labels. Notably, far fewer false-positive predictions were obtained for all realms and genome types. High numbers of false positives were identified in the fragmented test set predictions highlighting a major limitation of the genome composition-based model predictions. Shorter genomic fragments and incomplete sequences often lack adequate specific signals and can lead to misclassification which is a known limitation of k-mer-based classification (Ren et al., 2017).

4.4.7 Unknown sequence predictions

To predict the class of the unknown contigs (UCs) generated from Chapter 3 and the blood microbiome dataset analysed in this chapter, TetraPredX models were employed. Genome composition-based prediction models have been shown to be sensitive to contig sequence length and often lead to false positives when applied to short contigs as they lack enough genomic information required for accurate predictions (Ren et al., 2020; Guo et al., 2021a). To avoid this, UC set was filtered to include contigs that were at least 1kb long. The contigs that were still categorised as UCs after the latest analysis carried out on 14 October 2020 (see section 3.4.7) were combined with the UCs obtained from the blood microbiome dataset and subsequently filtered to exclude any UCs that were <1kb long. Though a large number (n=9,900) of UCs included here was assembled from fecal samples, these UCs were originating from ten different microbiomes analysed in this study. Overall, the final set of 20,552 UCs was spanning 40 BioProjects and represented 866 samples. To use the TetraPredX models, feature extraction was carried out from the FASTA sequences and all four calibrated models (archaea, bacteria, plasmid and virus) were applied to the consolidated UC set of 20,552 sequences. From the UCs set, 18,236 sequences (88.7%) were predicted as viruses and the second most predicted class for these UCs was plasmid which comprised a set of 1,239 sequences. 163 UCs were predicted to be bacteria, 11 UC was predicted as archaea and the remaining 903 UCs did not have probabilities >0.5 for any classes. When the signal-based classification was taken into account, 16,601 (80.7%) UCs were predicted to belong to the virus class only while 717 UCs were classified as a plasmid. 2,106 UCs had probabilities >0.5 for both virus and plasmid classes, 15 UCs were classified into virus/archaea classes and 21 UC was in bacteria/plasmid classes (table 4.5).

Table 4.5: Predicted microbial class for unknown sequences

Unpredicted	903 (probability <=0.5)
	Predicted class based on maximum probability
Archaea	11
Bacteria	163
Plasmid	1239
Virus	18236
	Predicted class based on signal
Archaea	8
Bacteria	141
Bacteria/Plasmid	21
Plasmid	717
Virus	16601
Virus/Archaea	15
Virus/Bacteria	39
Virus/Bacteria/Plasmid	1
Virus/Plasmid	2106

Notably, >80% UCs were predicted as likely originating from virus class using the models developed here. To compare these predictions with other virus prediction methods, two widely used virus prediction tools, VirSorter2 and DeepVirFinder, were applied to the UC set. Briefly, DeepVirFinder is a widely used genomic sequence composition-based machine learning prediction tool whereas VirSorter2 uses some genomic composition metrics e.g. GC content combined with hallmark genes to predict the likelihood of a sequence being of virus origin. Although DeepVirFinder employs more sophisticated deep learning algorithms to carry out predictions, the modelling methods i.e. k-mer based predictions are more comparable to the models described in this section.

Overall, 11,617 and 2,084 UCs were predicted as viruses using DeepVirFinder and VirSorter2 respectively. These predicted viral UCs were compared between all three methods described here and the overlapping UCs found with these three methods are shown in the Venn diagram shown in figure 4.17. These results demonstrate that there was variation in UCs predicted to be viral in origin. 1,463 UCs were predicted to be of virus origin by all three methods. Among the predictions made by DeepVirFinder and TetraPredX, there was significant overlap. The tools predicted 10,221 UCs as viruses. There were 1,463 contigs that were common to all three tools. This could be due to similar genome-composition-based prediction approaches employed by both

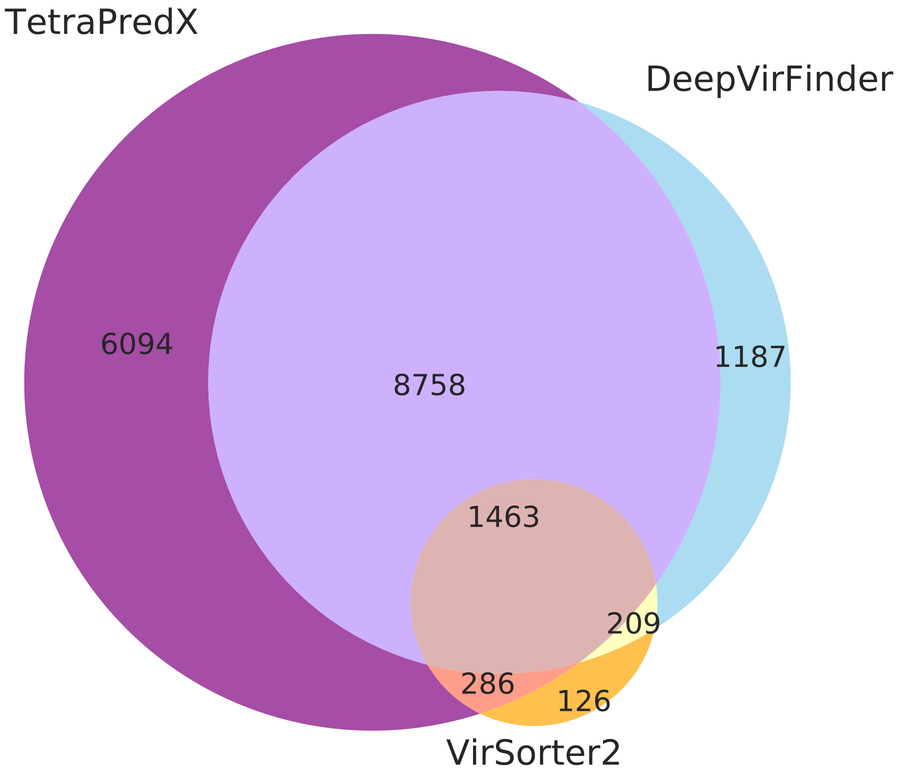


Figure 4.17: Virus predictions results for unknown contigs showing a comparison between TetraPredX, VirSorter2 and DeepVirFinder output. Venn diagram shows the overlap among predicted virus sequences from the unknown contigs dataset. Two popular virus prediction tools VirSorter2 and DeepVirFinder were applied to the same dataset and their results are compared with TetraPredX output.

tools. Among the predicted set, 6,094 UCs were unique to TetraPredX. These prediction results were slightly different if the signal-based predictions were taken into account for TetraPredX (figure B.2). 18,762 UCs were predicted to belong to the class virus with a predicted probability >0.5. 1,603 UCs were shared and predicted as viruses using all three methods.

The set of 10,221 UCs that were predicted to be viruses using TetraPredX and DeepVirFinder were analysed further using anicalc/aniclust approach to dereplicate the sequences into virus operational taxonomic units (OTU). `anicalc.py` and `aniclust.py` scripts available within CheckV (Nayfach et al., 2020b) are shown to be the most efficient to create virus OTUs (Tisza et al., 2021b). Virus OTUs were generated using aniclust.py with `--min_ani 95`, `--min_qcov 0` and `--min_tcov 85` parameters. These analyses yielded 9,118 unique virus OTUs suggesting the presence of 9,118 distinct viruses in this set.

To assign further virus-specific attributes to the predicted viruses, 14,830 UCs that were predicted as viruses were extracted and virus-specific models were used to predict the realm, genome type and segmentation properties of these UCs. Although these results are preliminary, they are described below. Based on the highest probability of classification, 8,881 UCs were predicted to belong to realm *Riboviria*, 2,572 were assigned to realm *Monodnaviria*, 1,671 were assigned to realm *Varidnaviria* and 498 were assigned to realm *Duplodnaviria*. A realm could not be determined for the remaining 1,208 UCs. Predicted genome types based on the maximum signals were as follows: 7,195 dsRNA, 2,392 ssDNA, 2,296 dsDNA, and 1,941 ssRNA with the remaining 1,006 UCs that could not be predicted to belong to any of these classes. Signal-based predictions for each of these virus-specific properties were often contradictory as shown in table B.5 highlighting the potential limitations of these models.

4.5 Discussion

Unknown sequences exist in most shotgun metagenomic and metatranscriptomic datasets. A range of new studies that aim to classify this unknown sequence matter embedded within metagenomes have found novel viruses and phages (Gregory et al., 2020; Luis F. Camarillo-Guerrero et al., 2021; Benler et al., 2021; Tisza et al., 2021b; Nayfach et al., 2021). In this study, I have expanded the catalogue of unknown contigs (UCs) by analysing 2,625 human blood microbiome samples using the UnXplore framework discussed in *Identification and quantification of 'unknown' biological sequences in human microbiomes* chapter. We found that on average 22% of assembled sequences were deemed to be of unknown taxonomic origin. These UCs were compiled with UCs discovered in Chapter 3 resulting in a set of 20,552 sequences that cannot be classified using traditional sequence similarity-based methods such as BLAST. Hence, the classification of these UCs remains a major challenge. The k-mer composition-based machine learning (ML) models we developed and optimised here aimed to solve this issue and classify the UCs surveyed here into microbial sequences.

Two separate approaches namely Support Vector Classifier (SVC) and Random Forest Classifier (RFC) were implemented to build prediction models that can classify the UCs into one of the four categories. The balanced binary models that included the reverse complement k-mer features were deemed most suitable for this task. The overall performance obtained for both methods was comparable and RFC was chosen over SVC as they were faster to run due to their inherited multithread processing attribute available in sklearn. Moreover, the issue of overfitting i.e. fitting exactly against training data in a statistical model and unable to make accurate predictions on previously unseen data - is less prominent to RFC as these models consist of several weak classifiers which are all trained independently on different subsets of training data.

Simple tetranucleotide frequency-based prediction models were developed to predict the origin of sequences. These models were trained and tested on known microbial sequences, refined to suit UC derived from microbiome datasets and polished to predict the microbial class of the unknown sequences identified in this as well as the previous chapter. These models were packaged into a python tool that can be widely applied to metagenomic and metatranscriptomic datasets to "classify" the sequences that do not have alignment-based sequence similarity to known sequences. An extensive comparison between different models was carried out and random forest models were deemed suitable for this task as they are less likely to be biased towards observed data. A wide range of microbial genome datasets were assessed to determine their suitability for prediction models. RefSeq genomic sequences combined with non-overlapping fragments of bacteria and archaea sequences proved to be the most effective for metagenomic sequence classification. Final calibrated random forest classification models with fragmented bacterial and archaeal sequences, ICTV species exemplar viruses and reference plasmid sequences were shown to be able to predict all four microbial sequence classes with very high accuracy and precision.

Our TNF-based models were applied to the UCs catalogued in this Chapter as well as in Chapter 3. Predictions made by these models clearly show that >70% of all UCs can be confidently assigned to virus class. These results support our initial hypothesis that a large proportion of unknowns embedded in public repositories could be originating from uncultivated virus genomes. These sequences represent completely novel virus genomic signatures that are currently missing from the public databases. Although our models are much simpler compared to the more sophisticated deep learning algorithms implemented in DeepVirFinder, the prediction results obtained using these models are largely overlapping with those generated using DeepVirFinder. This is promising as DeepVirFinder is shown to perform the best among all virus/phage prediction tools (Fung et al., 2022; Ren et al., 2020).

TNF-based prediction models were tested using a range of test datasets and were deemed very sensitive toward the identification of each class. However, some bacterial sequences could potentially be predicted as plasmids or viruses due to the nature of these mobile genetic elements to acquire bacterial genomic sequences. There is a known trade-off between k-mer length and specificity. Shorter k-mers could lead to more false positives whereas longer k-mers could be overly specific to known sequences. As the exact k-mer matching approach is implemented in TetraPredX, the implementation of longer k-mers could be extremely sensitive to pitfalls associated with exact k-mer matching. Popular prediction tools such as IDTAXA for bacteria (Murali et al., 2018) and VirFinder/DeepVirFinder (Ren et al., 2017; Ren et al., 2020) for viruses have implemented their models with k-mers of length 8-10. Shorter k-mers implemented in TetraPredX could be less specific compared to other prediction tools that use longer k-mers. However, it could be argued that short signatures such as that implemented in TetraPredX could be more effective at predicting high diversity sequences e.g. RNA viruses (Ul et al., 2020). Moreover, our results show that TNF models implemented in the random forest are in fact able to deconvolute the signals associated with different classes tested here with very high precision and recall.

Numerous virus-host prediction tools have demonstrated that short k-mer signatures including di-, tri- and tetra-nucleotides are effective at predicting the host specifically for prokaryotic viruses (Tang et al., 2015; Villarroel et al., 2016; Ahlgren et al., 2017; Babayan et al., 2018; Young et al., 2020). This virus-host signature sharing arising from viruses' properties of genome mimicry could potentially be attributed to false-positive predictions. However, this phenomenon would be more prominent in the case of previously unseen host (i.e. bacteria and archaea) sequences as viruses possess a broader genomic diversity compared to their hosts. As previously noted, TNF models' performance could be poorer if applied to short contigs (typically less than 1kb) as these sequences lack enough observations required for the models to make accurate predictions. Due to such complex interactions between viruses and their hosts, further explorations to identify the virus-host prediction boundaries and feature optimisation would be required. One solution to avoid this could be to implement suitable longer k-mers in predicting various classes, however, exact k-mer matching with longer k-mer could in turn negatively impact the prediction accuracy. By expanding the feature set and incorporating other relevant features such as nucleotide-based features such as GC content, amino acid k-mers, and protein domain signatures that are implemented in other prediction tools such as VirSorter2, these challenges could be overcome (Guo et al., 2021a). Identification and implementation of unique and specific features for each class included in TetraPredX would be a substantial undertaking and could be extremely challenging, however, it would be a significant upgrade from the existing models. All models implemented in TetraPredX were trained and tested using the gold-standard NCBI RefSeq datasets. To encompass a wider diversity of sequences, alternative data repositories that hold a much larger number of sequences such as GenBank or specialised databases such as GTDB (Parks et al., 2022) or NCBI Virus (Brister et al., 2015), IMG/VR (Roux et al., 2021c) can be explored in order to provide a more comprehensive dataset.

ML models for virus properties including realm, genome type and segmentation were explored but the predictions made using these models would need to be optimised further, specifically in the case of realm *Varidnaviria* and RNA viruses. Due to the poor resolution and signals embedded in shorter sequences, virus-specific model predictions and performance has plenty of room for improvement. For example, more sophisticated algorithms e.g deep learning method implemented in DeepVirFinder and longer k-mers may be more suitable for modelling these properties. It is worth noting that taxonomic classification is an arbitrary method of grouping viral sequences into a variety of classes often based on the genomic (e.g. dsDNA viruses are included in realm *Duplodnaviria*) as well phenotypic properties. Genomic composition-based signals alone may not provide sufficient resolution to achieve this classification accurately. Virus taxonomic classification at various levels of taxonomy is currently in flux e.g. realms *Adnaviria* and *Riboviria* have been added in the newer version of the master species list released in May 2021. Moreover, a range of virus families (e.g. *Baculoviridae*, *Anelloviridae*) still remain to be classified into one of the realms. Given the fluid nature of virus taxonomy and the complexity presented by virus genomes/segments modelling virus properties at this level remains a fascinating yet challenging task.

It is notable that models developed and distributed through TetraPredX are highly reliant on the reference databases that often suffer from shortcomings related to misannotation and mislabelling of reference sequences. These errors can potentially impact the model performance and could potentially introduce bias. Due to the fluid nature of taxonomy and delays in incorporation of taxonomic changes in corresponding INSDC databases such as GenBank, it is anticipated that some results obtained here may be difficult to replicate. Moreover, as mobile genetic elements such as phages and plasmids often share certain genomic sequences with the host genomes, some unknown sequences would not be classified using TetraPredX models and further research and development of targeted approaches would be required to classify them.

Models developed here are aimed at identifying microbial sequences, and if eukaryotic sequences are submitted to these models, they are likely to be predicted as viruses since eukaryotic sequences were not included in the models. Viruses are known to imitate their host sequences and often have host nucleic acid embedded within their genomes which can lead to misclassification which is a known limitation of these k-mer-based classification models (Ponsero et al., 2019). UCs that were analysed here were free from eukaryotic contamination as human host sequences as well any known sequences with reasonable sequence similarity were removed as part of the UnXplore framework. UnXplore framework has been made available to the wider scientific research community via GitHub (https://github.com/sejmodha/UnXplore). These models are wrapped up in a Python package called TetraPredX which is currently under alpha testing and is also available to download via a simple Python package installation (PyPI) interface via https://pypi.org/project/TetraPredX/. This package can automatically extract features from a given input FASTA file and use calibrated, optimised models developed here to predict the microbial origin unknown sequences assembled from the metagenomic datasets.

Chapter 5

Exploring the diversity of virus genome sequences embedded within the human microbiome data

Cataloguing virosphere.

5.1 Abstract

Metagenomics has enabled researchers to obtain greater insights into the uncultivated microbial diversity in many types of samples. With the advent of sequencing technologies and cheaper costs of sequencing, it has become a standard laboratory technique that has been applied to a range of different environments and samples to study their microbial content. Metagenomics poses great opportunities in the field of virus discovery and has enabled the identification, characterisation and study of a large number of viruses over the last decade. A systematic metagenomic analysis was carried out on *de novo* assemblies of >3000 human microbiome samples to identify known and novel virus genomes in the contigs generated from the UnXplore framework. Extensive analyses focusing on individual samples interrogating their viral content led to the identification of hundreds of novel virus genomes including >300 prokaryotic viruses, >200 novel anellovirus species and >30 RNA viruses. These results have revealed that viruses are present in a large number of already “analysed" microbiome datasets emphasising that public data repositories contain a wealth of novel viruses that are yet to be catalogued.

5.2 Introduction

Viruses are universal to all ecosystems of the planet earth. They are found in all environments from the human gut to marine water. The virus particles are thought to be one of the most abundant entities in our living communities. Traditionally, viruses have been identified using laboratory-based cell culture techniques but the advances in the high-throughput sequencing technologies (HTS) and their application in the field of metagenomics and metatranscriptomics has opened a vast avenue of discovery of uncultivated virus diversity. These methods serve as very powerful tools to study and identify currently uncultured viruses, and their interactions with each other and with their hosts, and can help us get insights into the key role they play in our ecosystem.

This field of study as a whole is often referred to as metaviromics - a term first coined by Zablocki et al. (2014) and refers to the study of viruses made accessible and possible through the advent of HTS technologies. In the last two decades, a range of novel viruses, viral families and unclassified virus diversity has been discovered and catalogued by researchers through metaviromics (Koonin et al., 2018; Dance, 2021). Metagenomics and metatranscriptomics have enabled the identification of diverse microbes that cannot be cultured in the lab including viruses. However, in striking contrast to other microbes such as bacteria, one of the major challenges to discovering viruses is the lack of universally present conserved genes (such as ribosomal RNA for bacteria) across the entire viral genomic landscape (Koonin et al., 2018). Nonetheless, significant progress has been made through the use of certain clade-specific genes such as the RNA-dependent RNA polymerase (RdRp) gene in RNA viruses (M. Shi et al., 2016a; M. Shi et al., 2016b; Edgar et al., 2022) and host-specific signatures embedded within the prokaryotic viruses that are often referred to as phages.

A large catalogue of studies has captured the virosphere in a variety of natural environments including the Earth virome project (Paez-Espino et al., 2016), and marine virome projects such as Tara Ocean Virome and the Pacific Ocean Virome projects (Mizuno et al., 2013; Hurwitz et al., 2013; Jennifer R Brum et al., 2015). Viruses have undoubtedly played a major role in shaping the human ecosystem as humans are surrounded by and interact with a diverse range of viruses and their hosts on a daily basis. Studies have shown that the viruses that regulate microbial components in humans, e.g. bacteriophages, have an impact on antiviral immune response and viral infectivity (Honda et al., 2012; Duerkop et al., 2013; Lecuit et al., 2013; Popgeorgiev et al., 2013; Rascovan et al., 2016; Zárate et al., 2017; Tisza et al., 2021b). In 2014, the identification of crAssphage led by Bas E. Dutilh et al. (2014) highlighted the importance of shedding light on viral dark matter embedded within the human microbiome. Subsequently, an array of phages related to crAssphage were discovered in publicly available datasets. These phages that infect bacteria included in the phylum *Bacteroidetes* were found to constitute up to 90% of the human gut virome (Yutin et al., 2018). More recently, studies led by Luis F. Camarillo-Guerrero et al. (2021) and Gregory et al. (2020) have catalogued and consolidated community-level human gut virome into the Gut Phage Database (GPD) and Gut Virome Database (GVD) respectively. Moreover, another study led by Benler et al. (2021) has identified three novel candidate families of order *Caudovirales* by mining community-level human gut microbiome samples for circular phage contigs. This analysis has also identified diverse putative mechanisms underlying phage-host interactions in the human gut (Benler et al., 2021). Similarly, the Skin Microbial Genome Collection (SMGC) found thousands of novel prokaryotic and eukaryotic virus sequences that were absent from uncultivated virome-specific repositories such as IMG/VR (Paez-Espino et al., 2019; Roux et al., 2021c) and GPD (Kashaf et al., 2022).

The application of metaviromics has been crucial in human and animal pathogen identification (Jerome et al., 2019; Briese et al., 2009), and virus outbreak tracking (such as sewage monitoring for SARS-CoV-2 (Crits-Christoph et al., 2021)) as well as in clinical diagnostics (Nakamura et al., 2009; Stremlau et al., 2015; Moustafa et al., 2017; Thorburn et al., 2015; Jerome et al., 2019). Since the Coronavirus pandemic, there has been a newfound interest in virus discovery including their identification and presence in reservoir species (Wahba et al., 2020). Many community-led projects have been undertaken and massively parallelised the process of virus discovery by mining the existing datasets available in the public repositories (Connor et al., 2019; Martí-Carreras et al., 2020; Edgar et al., 2022). This has been possible through the employment of novel computation algorithms and resources; for example, a peta-base scale data mining has led to the discovery of 131,957 novel RNA viruses from publicly accessible raw sequence data repositories (Edgar et al., 2022).

A large number of uncultivated virus genomes have been identified through metaviromics compared to the traditional culture-based protocols in the last ten years (Roux et al., 2021c). Viruses require hosts to grow and replicate but due to the limitations of laboratory cultivation techniques, a lot of viruses and their hosts cannot be cultivated in controlled laboratory environments. As a result, in 2017, the International Committee on Taxonomy of Viruses (ICTV) formally started accepting proposals to classify uncultured viruses identified through metaviromics in their framework given that they were adequately analysed and met the required quality standards (Simmonds et al., 2017a). In light of this, researchers have also devised appropriate standards to catalogue the diversity of uncultivated viruses in a more systematic way. Minimum Information about an Uncultivated Virus Genome (MIUViG) standards chiefly aims to capture relevant information about the uncultivated viruses such as their origin, genome quality, genome annotation, taxonomic classification, biogeographic distribution and *in silico* host prediction and fittingly provide the best practices for metaviromics research (Roux et al., 2019). ICTV Bioinformatics Expert Group has also highlighted the importance of including the uncultivated virus diversity in the taxonomic framework and indicated that this has added valuable insights into virus taxa, viruses and virus evolution (Bas E Dutilh et al., 2021).

Typically, a large diversity of DNA viruses is expected to be discovered in the bulk metagenomic samples owing to the sample preparation approaches that target the total DNA present in the samples. Bulk metagenomic samples that include viruses, as well as other microbes, often do not have a specific virus enrichment step that is crucial to identifying small DNA viruses and RNA viruses. It is notable that such metagenome explorations often lead to taxonomically unclassified virus genomes and a variety of terminologies are found in the literature to represent the assembled virus sequences from metagenomes. However, MIUViG recommendations suggest the use of the term "viral operational taxonomic units" or vOTUs that represents species-level ranking equivalent for uncultivated viruses and it is conventionally derived after clustering of assembled viral sequences is performed. Moreover, other terminologies such as metagenome-assembled genomes i.e MAGs, single amplified genomes (SAGs - often obtained by sorting individual virus particles using methods such as flow cytometry), Uncultivated Virus Genomes (UViGs) are used to refer to contigs/genomes that originate from metagenomic datasets.

To meet the demands of the ever-expanding field, a range of different virus prediction tools and pipelines have been developed and have been used widely by the community to perform sophisticated metaviromic analysis leading to the successful identification of various novel uncultured viruses from different environments (Nooij et al., 2018). Some of the most popular tools that initially focused on prokaryotic viruses discoveries such as VirSorter (Roux et al., 2015a) have been updated with the advances in sequence mining techniques such as machine learning. A more recent version of this successful user-friendly tool - VirSorter2; combines various different nucleotide and protein level features including the presence of hallmark genes, profile HMMs (derived from Pfam), GC content, and short sequence motifs to successfully predict viruses from a given set of contigs originating from any environment (Guo et al., 2021a). Other tools such as DeepVirFinder (Ren et al., 2020) and TetraPredX (described in Chapter *Predicting the biological origin of unknown sequences using machine learning*; https://github.com/sejmodha/TetraPredX) employ solely sequence feature-based machine learning models to successfully predict viruses from any given set of assembled contigs. To further assess and refine the quality of these predictions, including the completeness (based on the MIUViG criteria) of the viral genomes, more targeted pipelines such as CheckV (Nayfach et al., 2020a) have been developed and have been widely applied to virus discovery projects. Moreover, the final step of the analysis, virus annotation, can also be automated with virus annotation pipelines such as RATT (Otto et al., 2011), VAPiD (Shean et al., 2019), and DRAM-v (Shaffer et al., 2020). A general-purpose metagenomic analysis pipeline developed by the EBI called MGnify is currently being adapted to accommodate virus discovery. EBI's new VIRify (Rangel-Pineros et al., n.d.)(https://github.com/EBI-Metagenomics/emg-viral-pipeline) pipeline that is currently being developed can detect, annotate, and taxonomically classify metaviromic assemblies and it makes use of a range of different tools mentioned above to provide a one-stop solution to researchers to analyse their metaviromic datasets. Similarly, IMG/VR provides an integrated data management and analysis platform to store and adequately analyse metaviromic datasets (Roux et al., 2021c). Moreover, environment-specific virus community resources such as Gut Virome Database (GVD) and Gut Phage Database (GPD) have also been developed to accommodate and facilitate data management and integration of uncultivated virus sequences arising from large virus discovery projects.

In light of the large-scale identification of uncultivated viruses, a major challenge is to determine the host association for the uncultivated viruses as it cannot be readily made available through metagenomics. Host linkage could be argued as one of the most important undertakings due to its importance in understanding virus ecology, interactions, potential role in shaping human (and animal) health and the disease state as well microbial evolution (Roux et al., 2021b). For example, >95% of all UViGs in IMG/VR databases lack host linkage information (Roux et al., 2021c). In absence of host information, computational approaches have been used to predict and determine the potential hosts of uncultivated viruses (Robert A Edwards et al., 2016). These approaches leverage currently known virus-host genome interactions such as the presence of prophage sequences (Roux et al., 2015b; Roux et al., 2015a), Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) spacers (Dion et al., 2021), or short nucleotide and protein signatures (k-mers) to predict the potential host of the uncultivated viruses (Villarroel et al., 2016; Babayan et al., 2018; Young et al., 2020). Although a range of tools and models are available that can link uncultured viruses to their hosts, this area of research is deemed to be one of the most challenging tasks by the experts (Roux et al., 2021b; Coclet et al., 2021) and computational approaches are being adapted to accurately predict the hosts of newly discovered viruses.

UnXplore provided an extensive metagenomic analysis opportunity as high quality *de novo* assembly was performed using the framework, as described in Chapter 3. To identify and characterise the human virome and catalogue virus genomic sequences embedded within human microbiome datasets assembled through UnXplore, a comprehensive virus discovery analysis was conducted.

5.3 Method

3,559 human microbiome datasets surveyed and assembled using the UnXplore framework (described in Chapters *Identification and quantification of 'unknown' biological sequences in human microbiomes* and *Predicting the biological origin of unknown sequences using machine learning*) were systematically interrogated to identify and quantify known and novel virus genome sequences present in them. These samples were distributed across 58 BioProjects. A list of these BioProjects along with the metadata including the location of sampling, microbiome type as well as the primary publication associated with the dataset are shown in the appendix table C.1. 24 BioProjects were linked to a primary publication and the remaining 34 could not be linked to a primary publication and/or published data analyses. 51 BioProjects included here were sequenced with metagenomic (i.e. DNA sequencing) protocols. On contrary, the following BioProjects were sequenced with metatranscriptomic protocols which included RNA sequencing: Sputum samples from PRJEB14539 and PRJEB10919, Vagina samples from PRJEB21446, Blood samples from PRJNA513310 and PRJNA271229, and Saliva samples from PRJNA264728.

From the complete set of assembled contigs generated using UnXplore, assembled contigs

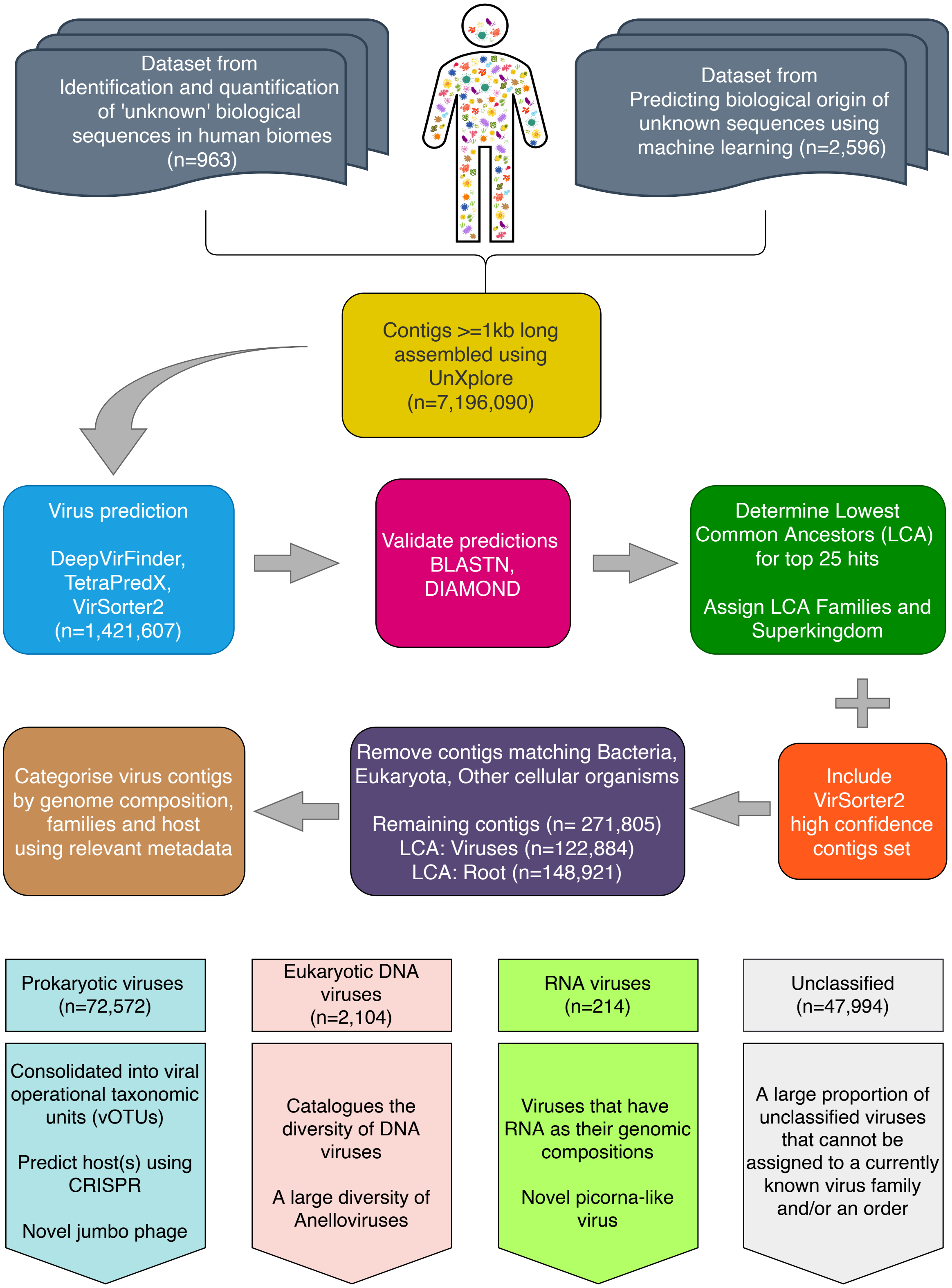

Dataset from Identification and quantification of 'unknown' biological sequences in human biomes (n=963)
Dataset from Predicting biological origin of unknown sequences using machine learning (n=2,596)
Contigs >=1kb long assembled using UnXplore (n=7,196,090)
Virus prediction
DeepVirFinder, TetraPredX, VirSorter2 (n=1,421,607)
Validate predictions BLASTN, DIAMOND
Determine Lowest Common Ancestors (LCA) for top 25 hits
Assign LCA Families and Superkingdom
Include VirSorter2 high confidence contigs set
Remove contigs matching Bacteria, Eukaryota, Other cellular organisms
Remaining contigs (n= 271,805)
LCA: Viruses (n=122,884)
LCA: Root (n=148,921)
Categorise virus contigs by genome composition, families and host using relevant metadata
Prokaryotic viruses (n=72,572)
Consolidated into viral operational taxonomic units (vOTUs)
Predict host(s) using CRISPR
Novel jumbo phage
Eukaryotic DNA viruses (n=2,104)
Catalogues the diversity of DNA viruses
A large diversity of Anelloviruses
RNA viruses (n=214)
Viruses that have RNA as their genomic compositions
Novel picorna-like virus
Unclassified (n=47,994)
A large proportion of unclassified viruses that cannot be assigned to a currently known virus family and/or an order

Figure 5.1: A schematic representation of the virus prediction and discovery process is described in this chapter.

that were at least 1kb long were retained to carry out virus discovery analyses. An overview of the virus discovery process highlighting the major findings is shown in the figure 5.1. Various steps of the analyses are described in detail below.

5.3.1 Virus sequence prediction

Two widely-used virus prediction tools; VirSorter2 (Guo et al., 2021a) and DeepVirFinder (Ren et al., 2020) were used to predict viral contigs from UnXplore filtered set. TetraPredX (described in Chapter 4; https://github.com/sejmodha/TetraPredX) was also applied to the same contig dataset for virus predictions. Briefly, both TetraPredX and DeepVirFinder are machine learning-based tools that predict the sequence class solely based on the k-mer composition whereas VirSorter2 uses 27 distinct sequence features such as viral HMMs, gene density, average gene size along with other sequence-derived features to predict virus sequences (Guo et al., 2021a). VirSorter2 predictions were run with `--include-groups dsDNAphage, NCLDV, RNA, ssDNA, lavidaviridae` `--provirus-off` and `--min-score 0.5` parameters; all predicted contigs included in "final-viral-combined.fa" were considered.

In order to capture as many predicted virus sequences as possible three virus sequence prediction tools that employ different prediction algorithms were used. Three prediction tools were used to make predictions for the 7,196,090 contigs at least 1kb in length assembled from the UnXplore analysis in Chapter 3. To retain high confidence predictions these results were filtered using the following criteria: 1) DeepVirFinder, contigs with a score >=0.9 and pvalue <0.05 were selected. 2) VirSorter2: all predicted contigs with a minimum score of 0.5 were selected. 3) TetraPredX: all contigs with viral signal probability >=0.95 and all other class probability <=0.5 were selected. These prediction thresholds were similar to those that were used for prediction accuracy measurement in the VirSorter2 paper (Guo et al., 2021a). As DeepVirFinder and TetraPredX prediction values are based on probabilities, applying a high probability threshold of 95% should exclude most false-positive sequence predictions. The score threshold of 0.9 was applied to DeepVirFinder contigs as the higher the score indicates the more likely a sequence is from viral genomes. To validate these prediction results further, all sequences that were predicted to be viral by any tool after filtering were searched against nucleotide database nt (downloaded on 27 October 2021) using BLASTN with evalue 0.0001 and the top 25 hits for each contig were extracted. The results were generated in the standard BLAST tabular output format with additional columns staxids, stitle, qcovs and qcovus. The lowest common ancestor (LCA) was computed for each contig from its corresponding hits using the Python script `ExtractLCABLASTM6.py` (https://github.com/sejmodha/UnXplore) and the superkingdom of the LCA was computed using Python script `GetSuperkingdomFromLCA.py` (https://github.com/sejmodha/UnXplore). The proportion of virus-specific hits for each contig was included in the output generated using these scripts and was subsequently used to confirm the results of the prediction along with taxonomic superkingdom level classification. This percentage value was derived by dividing the total number of virus hits by the total number of hits i.e. 25. For a contig with all 25 hits to viruses, the percentage of virus hits would be 100%, however, if a contig had hits to bacteria and viruses then the LCA would be determined as root and the percentage of virus hits would correspond to the number of hits specific to viruses. This value was only used in cases where the LCA would be assigned as "root" suggesting that the contig had hits from more than one superkingdom. It is noteworthy that with this approach a range of phages where even one hit is from bacteria, the LCA would be determined to root as opposed to viruses. Additionally, all contigs were also searched against the comprehensive RefSeq protein database (downloaded on 9 November 2021) using DIAMOND BLASTX run with evalue 0.001 -b5 -c1 and standard tabular output with additional columns qframe, stitle, staxids, qcovhsp and scovhsp was generated. The LCA and the superkingdom for each contig were also obtained for these BLASTX searches using custom Python script `ExtractLCA.py` (https://github.com/sejmodha/UnXplore). Although protein level analyses were carried out for all predicted contigs, DIAMOND results were only examined when BLASTN hits were absent.

5.3.2 Contig quality assessment

All contigs (n=7,196,090) were scanned for their quality and completeness using CheckV (Nayfach et al., 2020a). CheckV is an automated pipeline that can determine the quality of viral contigs using a range of different strategies including the presence of direct and inverted terminal repeats, virus hallmark genes and identification of viral protein domain signatures derived from large complete viral genome sequences including a comprehensive set of 76,262 virus sequences identified from publicly available metagenomic datasets (Nayfach et al., 2020a).

5.3.3 Consolidated viral contigs set

To obtain a set of high confidence contigs that are most likely to be viruses, a consolidated set of viral contigs was generated. As VirSorter2 has been demonstrated to recover more viral contigs compared to other tools, VirSorter2 predicted sequences that satisfied the criteria described in (2) were also included.

1. Any predicted viral contig where LCA was assigned to viruses (n=122,884).

2. VirSorter2 specific criteria: VirSorter2 predicted contigs that were filtered with criteria adapted from the protocol described in ‘Viral sequence identification SOP with VirSorter2 V.3’ (Guo et al., 2021b).

   Briefly, this set included:

   (a) contigs with more viral genes than host genes (Set 1)

(b) contigs with 0 viral and 0 host genes with max score value >=0.95 and hallmark >2 (Set 2)

(c) contigs that weren't included in Set 1 and Set 2 that had 0 viral genes, 1 host gene and were at least 10kb long (Set 3). This set would include novel phage sequences with integrated host genes.

Filtering based on (a), (b) and (c) generated a set of 366,883 contigs. These 366,883 contigs were filtered further by incorporating LCA information. The contigs where LCA was bacteria (n=128,829), archaea (n=231), eukaryota (n=200), cellular organisms (n=1,880) or undetermined (e.g. plasmid, synthetic constructs; n=1,022) were filtered out leaving a set of 234,721 contigs. From this set of 234,721 contigs, those with LCA Viruses (n=85,800) and root (n=148,921) were retained.

These contigs obtained from (1) and (2) were consolidated to remove any duplicates and the final set of 271,805 high confidence contigs was analysed further. This set included 122,884 contigs that had "Viruses" as a superkingdom and 148,921 contigs where superkingdom was assigned as "root".

For all contigs where superkingdom LCA was Viruses, the LCA was determined from the top 25 hits extracted from DIAMOND and/or BLAST results and where possible potential virus family was identified based on the LCA of contig hits. However, in some cases, a virus family could not be assigned. This could happen in cases where the LCA taxon was determined to be an unclassified virus sequence (typically refers to sequences that belong to NCBI taxonomy ID: 12429), a sequence that was not associated with a family, a virus sequence extracted from an environmental sample, or had BLAST hits at higher taxonomic ranks such as order. These metadata derived from similarity sequence results were used to apply various filters to the dataset. It was combined with the Virus Metadata Resource (VMR; https://talk.ictvonline.org/taxonomy/vmr/) and was used to determine the potential genomic composition, potential hosts as well as other taxonomic ranks such as order and family of the predicted viral contigs.

A final consolidated set of 271,805 contigs was collated for further analysis - this set included contigs where LCA was either Viruses or root (figure 5.1). These contigs were annotated with a range of different features including the LCA, virus family, CheckV quality and other relevant metadata. This metadata was used to categorise contigs of interest into three broad categories: prokaryotic viruses, non-prokaryotic viruses (i.e. DNA viruses excluding phages) and RNA viruses. Additionally, a range of unclassified viruses where LCA was determined to be either viruses or root was also identified but were excluded from the above categories as they could not be 'classified' into one of the above categories based on the prediction and similarity analysis carried out here. 122,884 contigs from this set had the superkingdom "viruses" as the top-level LCA.

5.3.4 Viral contigs categorisation

The confirmed viral contigs, i.e. those with LCA superkingdom as "Viruses" (n=122,884) were further categorised into different virus groups based on the LCA obtained from BLAST hits in the first instance. If a contig did not have any hits to nucleotide sequences, then LCA was determined using protein sequences obtained through DIAMOND results. These four categories were as follows:

1. Prokaryotic viruses (contigs with hits to virus sequences that infect bacteria and archaea)
2. Eukaryotic DNA viruses (contigs with hits to DNA virus sequences are not known to infect prokaryotic organisms such as bacteria and archaea)
3. RNA viruses (contigs with hits to RNA virus sequences - RNA viruses have RNA as their genetic material)
4. Unclassified viruses (contigs with hits to virus sequences that could not be mapped to a known virus family, realm and/or order e.g. those originating from environmental samples or hits spanning more than one order or realms in virus taxonomy)

Table 5.1: Summary of different sets of contigs used for virus discovery

Input contigs	Type of analysis	Note
28,837,029	Length filtering (>=1kb)	Remaining contigs: 7,196,090
7,196,090	Virus prediction	Predicted contigs: 1,421,607
1,421,607	Validate predictions	Label contigs' origin using validation output
122,884		Confirmed virus sequences validated using BLAST/DIAMOND
271,805	Prokaryotic virus discovery	Confirmed virus sequences combined with those where LCA = root to identify potentially new phages
122,884	Geographic distribution of viruses	

In order to minimise false positives, all sections of the diversity exploration except for the prokaryotic section used sequence data comprising 122,884 contigs with BLASTN/DIAMOND hits match exclusively to viruses derived from the 1,421,607 contigs obtained during virus prediction analysis (see table 5.1). The prokaryotic virus discovery section included an additional 149,943 sequences scanned for prokaryotic virus features described in section 5.4.4 (Table 5.1).

5.3.5 Prokaryotic viruses

To identify prokaryotic viruses i.e. viruses that infect bacteria and archaea, viral contigs were filtered from the consolidated set of 271,805 contigs. This more comprehensive set with contigs where the LCA was "root" as well as "Viruses" was included for prokaryotic virus sequence hunting to ensure that phages that mimic their hosts are not excluded from this analysis. To aid that, Virus Metadata Resource (VMR) version 200721 with the Master Species List version MSL36 was downloaded from https://talk.ictvonline.org/taxonomy/vmr/m/vmr-file-repository/13175. VMR data was joined with the virus data using the shared columns virus family. VMR also provided additional metadata associated with virus families including genome composition and host(s). The following criteria were then applied to filter contigs that belong to prokaryotic viruses:

- Contig length >=10000; This would lead to exclusion of sequences that belong to families *Microviridae* and *Inoviridae* but helps to exclude false-positive sequences.
  AND
  - Host is bacteria or archaea
    OR
  - LCA TaxonName contains string ‘caudovirales’, ‘phage’ or ‘caudo’ where LCA family could not be determined
    OR
  - Contigs where:
    - host genes < 5 AND (based on VirSorter2 prediction) AND
    - max_score_group == "dsDNAphage" (based on VirSorter2 prediction) AND
    - Percentage of viral contigs >=50 (based on BLAST results)

vOTU identifications and comparison to other phage genome sequences

Complete genome sequences for prokaryotic virus orders *Ligamenvirales* (n=59), *Primavirales* (n=5), *Caudovirales* (n=14,016), *Tubulavirales* (n=248), *Petitvirales* (n=2,979), *Haloruvirales* (n=32), *Mindivirales* (n=83), *Norzivirales* (n=427), *Timlovirales* (n=447), *Belfryvirales* (n=4), *Kalamavirales* (n=33), *Vinavirales* (n=17), *Halopanivirales* (n=16) were downloaded from NCBI databases in FASTA format.

Nucleotide sequences of the 142,809 viral clusters (95% nucleotide identity) belonging to the Gut Phage Database (GPD) were downloaded from http://ftp.ebi.ac.uk/pub/databases/metagenomics/genome_sets/gut_phage_database/. GPD sequences were clustered with prokaryotic virus sequences obtained in this study using the rapid genome clustering based on pairwise Average Nucleotide Identity (ANI).

Rapid genome clustering based on pairwise Average Nucleotide Identity (ANI)

To obtain a sequence level clustering and a representative sequence set, pairwise Average nucleotide identity (ANI) can be utilised. ANI combined with Alignment fractions (AF) which represent sequence similarity between two sequences with sequence coverage provides a quick but efficient clustering of sequences. `anicalc.py` and `aniclust.py` scripts included in CheckV package were used to determine Virus Operating Taxonomic Units (vOTUs) with 95% ANI and 85% Alignment Fraction (AF) with parameters `--min_ani 95`, `--min_tcov 85` and `--min_qcov 0`.

Virus-host predictions

To predict potential hosts of these prokaryotic viruses, CrisprOpenDB was used on prokaryotic viral contigs. CrisprOpenDB requires PhageHostFinder databases that were downloaded from http://crispr.genome.ulaval.ca/dash/PhageHostIdentifier_DBfiles.zip. Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) spacer are short nucleotide sequences can be used to predict hosts of unknown phages, as spacers represent biological records of past phage–bacteria interactions. DNA spacers are short segments (26-72bp long) that are homologous to phages or plasmids.

Prokaryotic vOTUs compiled from results described in the "Prokaryotic viruses" sections (see 5.4) were searched against the comprehensive phage-host database using CrisprOpenDB to determine potential hosts with default parameters that allow up to two mismatches in spacer sequences.

5.3.6 Anellovirus diversity exploration

A number of contigs matching to anelloviruses were found (see Results section 5.4.5) which were investigated further. A comprehensive phylogenetic analysis was carried out to study and compare the diversity of the anelloviruses. To obtain a set of complete sequences, sequences shorter than 2kb and larger than 4kb were excluded. ORFs were predicted using getorf (Rice et al., 2000) `-find 1` and `-minsize 600` parameters and the largest ORF (typically refers to ORF1 in the family *Anelloviridae*) were extracted. The largest ORF from each anellovirus contig was assumed to be the ORF1 sequences. The ORFs were not subjected to any other quality assessment steps. 1,275 anellovirus contigs matched the above ORF prediction criteria and were included in the downstream analysis.

Anelloviridae is a family of small circular ssDNA viruses that are known to infect both primates and non-primate mammals (Varsani et al., 2021; Souza et al., 2018; Kaczorowska et al., 2020). Anellovirus genome is not segmented and contains a single molecule of negative-sense single-stranded circular DNA that is 2000-4000 nucleotides long. Human anelloviruses are classified into four genera: *Alphatorquevirus*, *Betatorquevirus*, *Gammatorquevirus* and *Hetorquevirus* (Varsani et al., 2021). To obtain an up-to-date set of human anelloviruses references, representative sequences for each human anellovirus species (n=72) were selected from a recent study led by Varsani et al. (2021). Each species was searched against NCBI GenBank to fetch a RefSeq genome sequence, ORFs and other metadata available in GenBank. In cases where a RefSeq for a species could not be found, any sequence used for classification for that specific species from Varsani et al. (2021) was used. The largest ORF amino acid sequences for each species were extracted.

A complete set of 1,347 sequences combining the reference set (n=72) and the contig set (n=1,275) was generated. These sequences were analysed further to study the phylogenetic relatedness with respect to currently available species and reference sequences. A phylogenetic analysis protocol to obtain a genera level phylogeny as outlined (Varsani et al., 2021) was used. Briefly, amino acid sequences for the largest ORF (i.e. ORF1) were aligned using MAFFT `--auto` mode (Katoh et al., 2013). The resulting multiple sequence alignment was filtered to remove gaps using the TrimAL `--gappyout` option. The resulting filtered alignment was submitted to IQTree with `-m TEST` to automatically test and select the appropriate model for maximum-likelihood-based phylogenetic inference with 1000 bootstraps.

Sequence Demarcation Tool (SDT) (Muhire et al., 2014) was used to calculate ORF1 nucleotide level all-vs-all pairwise sequence similarity between reference set and anelloviruses assembled in this study. This set of sequences was run through SDT with MAFFT as the alignment program. The resulting matrix of similarities was analysed using sklearn implementation of hierarchical clustering for novel species identification with predetermined species demarcation criteria.

5.3.7 RNA viruses

To identify RNA viruses, all contigs where the LCA family was an RNA virus family were extracted. For contigs where the family could not be determined, the complete lineage of the LCA was obtained using ete3 (Huerta-Cepas et al., 2016) implementation in Python. The lineage was searched for the term ‘Riboviria’ which is the highest taxonomic level (Realm) that encompasses all RNA viruses. These contigs with hits exclusive to viruses included in realm *Riboviria* were also considered and labelled as potential RNA viral contigs.

Phylogenetic analysis

To study the relatedness of novel RNA viral contigs to existing viruses, a comprehensive phylogenetic analysis was carried out. First, contigs that match to a specific group of RNA viruses were extracted. For the specific group of interest, relevant RNA virus protein sequence hits were extracted and consolidated to remove duplicate protein entries retrieved from the RefSeq protein database. As all RNA virus genomes code for RNA-dependent RNA polymerase (RdRp), the ORF containing RNA-dependent RNA polymerase (RdRp) were extracted from the matching virus sequences from the protein sequence hits and/or relevant databases used for searches. The RdRp sequences were used as RdRp reference set as they spanned all database entries matching to contig set. Similarly, RdRp containing ORFs were also extracted from the contigs using getorf and a confirmatory BLASTP analysis was carried out against the reference RdRp sequence set to identify the ORF containing RdRp signatures. A protein sequence-based multiple sequence alignment was generated using MAFFT (Katoh et al., 2013) in `--auto` mode. The alignments were filtered to remove columns that contained 20% or more gaps. The final alignment was used to generate a maximum-likelihood-based phylogenetic tree using IQTree (Minh et al., 2020) with model testing and 1000 bootstraps. The resulting phylogeny was visualised and annotated in FigTree (http://tree.bio.ed.ac.uk/software/figtree/) and the annotations obtained from the VMR and NCBI were overlayed on the phylogeny.

5.4 Results

5.4.1 Virus predictions and validation

Out of the complete set of 7,196,090 contigs, 1,421,607 unique contigs were predicted as viral using any of the three virus prediction methods. 625,476 were predicted as viruses using TetraPredX, 495,425 were predicted as viruses using DeepVirFinder and 597,812 were predicted as viral using VirSorter2. A total of 23,769 contigs were predicted as viruses using all three prediction tools as shown in the figure 5.2(a). Out of these 23,769 contigs predicted by all three virus prediction tools, 21,141 were found to match at least one sequence in the nt database using BLASTN and 10,509 (50%) of them were shown to have viruses as their LCA superkingdom. Moreover, as shown in figure 5.2(b), 6,515 of these 21,141 contigs were found to match at least one virus sequence using BLASTN despite their LCA being defined as root which suggests that these contigs were matching sequences from one or more taxonomic superkingdoms. 42% of all contigs that were predicted as viral using both DeepVirFinder and VirSorter2 could be validated using BLASTN (figure 5.2(c)) and this percentage was around 18% for contigs that were shared between VirSorter2 and TetraPredX (figure 5.2(d)). A range of predicted viral contigs was shown to have “root” as LCA as their BLASTN hits were composed of sequences that originated from more than one superkingdom. In general, between 25-45% of any predicted viral contigs were categorised as matching the root (figure 5.2). However, among this set, >80% of them were shown to have a BLAST hit to a known virus sequence in the database indicating their viral origin. The k-mer-based prediction tools such as DeepVirFinder and TetraPredX as well as the random forest classifier implemented in VirSorter2, rely on these short signatures to differentiate between viral and non-viral sequences. Due to their ability to efficiently distinguish virus-specific signatures present in contig sequences, all three prediction tools are able to predict a number of contigs as viruses even if they may have a variety of matches to other organisms. As viruses are known to mimic their host sequences, other sequences that predicted viral contigs are matching could be associated with the viruses e.g. a host. This analysis and virus prediction tool comparison highlights that none of the three virus prediction tools was 100% accurate.

Overall, there was a substantial difference in the contigs predicted as viral between the three different prediction tools. A number of predicted viral contigs were unique to each prediction software (figure 5.2(a)). To validate these predictions further, BLASTN searches were carried out for all predicted contigs (shared as well unique contigs for each software) and these results are shown in the sunburst charts figure5.2(e), 5.2(h) and 5.2(f). Confirmation of whether a contig is originating from a virus sequence was obtained with the LCA superkingdom derived from the BLASTN results. A conservative approach of labelling a contig to be of virus origin only if the LCA superkingdom was assigned as "Viruses" was utilised. In cases where the LCA was assigned to be "root" - which suggested that the contig was matching to sequences from more than one superkingdom, a number of virus hits for each contig were checked. Notably, the proportions of sequences with hits exclusive to virus sequences in the databases, varied between 2-11% of the total number of predicted contigs for each tool. For contigs where LCA was determined to be "root", a large proportion of them were found to be matching at least one virus sequence from the databases. It was hypothesised that although assigned to root, these sequences could still originate from a virus genome as they showed some similarity to currently known virus sequences included in the databases. It is appreciated that despite their similarity to known virus sequences in INSDC databases, these sequences could also originate from other mobile genetic elements (MGE), such as plasmids. The current analytical approaches are unable to differentiate these MGEs, and when combined with other factors such as erroneous entries in INSDC databases, some of these ambiguous sequences that were hypothesised to be viruses could be misclassified.

BLASTN-based validation showed that 11% of predicted viral contigs unique to VirSorter2 were matching exclusively to virus genome sequences in the databases whereas this proportion was 7% for DeepVirFinder and 2% TetraPredX. This could be explained by the different approaches implemented in these virus prediction tools. VirSorter2 uses a wide array of sequence features including nucleotide and protein level information as well as virus-specific domain signatures for predictions whereas DeepVirFinder and TetraPredX rely solely on short k-mers of 10 and 4 respectively. In turn, it is much faster to run DeepVirFinder and TetraPredX prediction analysis and VirSorter2 was deemed much slower in comparison and required much greater computational resources such as RAM and CPU.

If a contig was predicted to be viral (by any tool) and whose BLASTN LCA was determined to be viruses, the viral family was derived from BLASTN (nucleotide). In cases where a BLASTN hit was not found but a protein hit was obtained using DIAMOND, the viral family was derived using DIAMOND output. This family level classification is summarised in figure 5.3 which represents a total of 124,493 contigs. A range of DNA virus families whilst prokaryotic virus families *Siphoviridae, Myoviridae* and *Podoviridae* were dominant among them. This was expected as most of the datasets analysed in this study originated from metagenomic DNA library preparations. RNA viruses spanning 9 distinct families *Arenaviridae, Flaviviridae,*

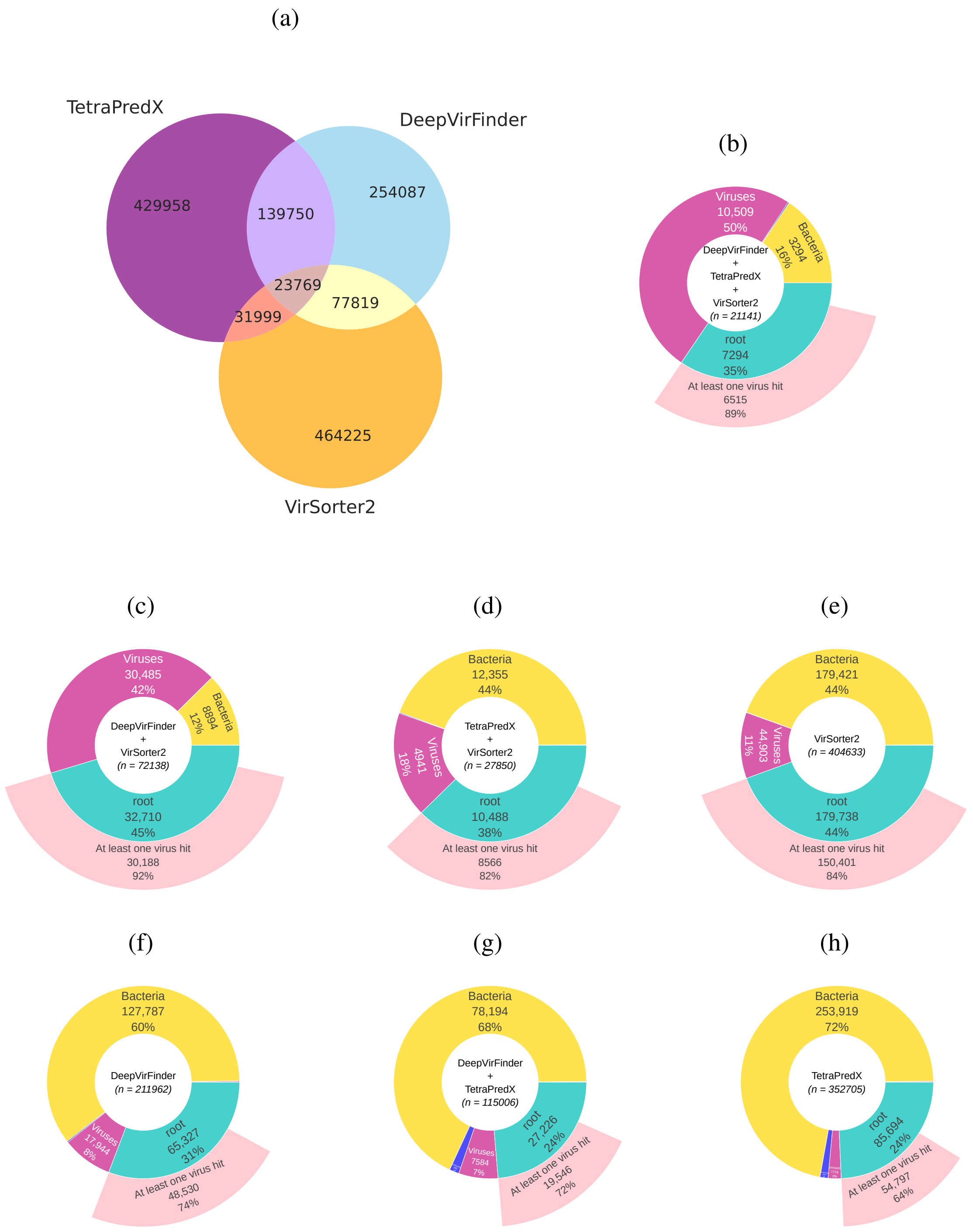
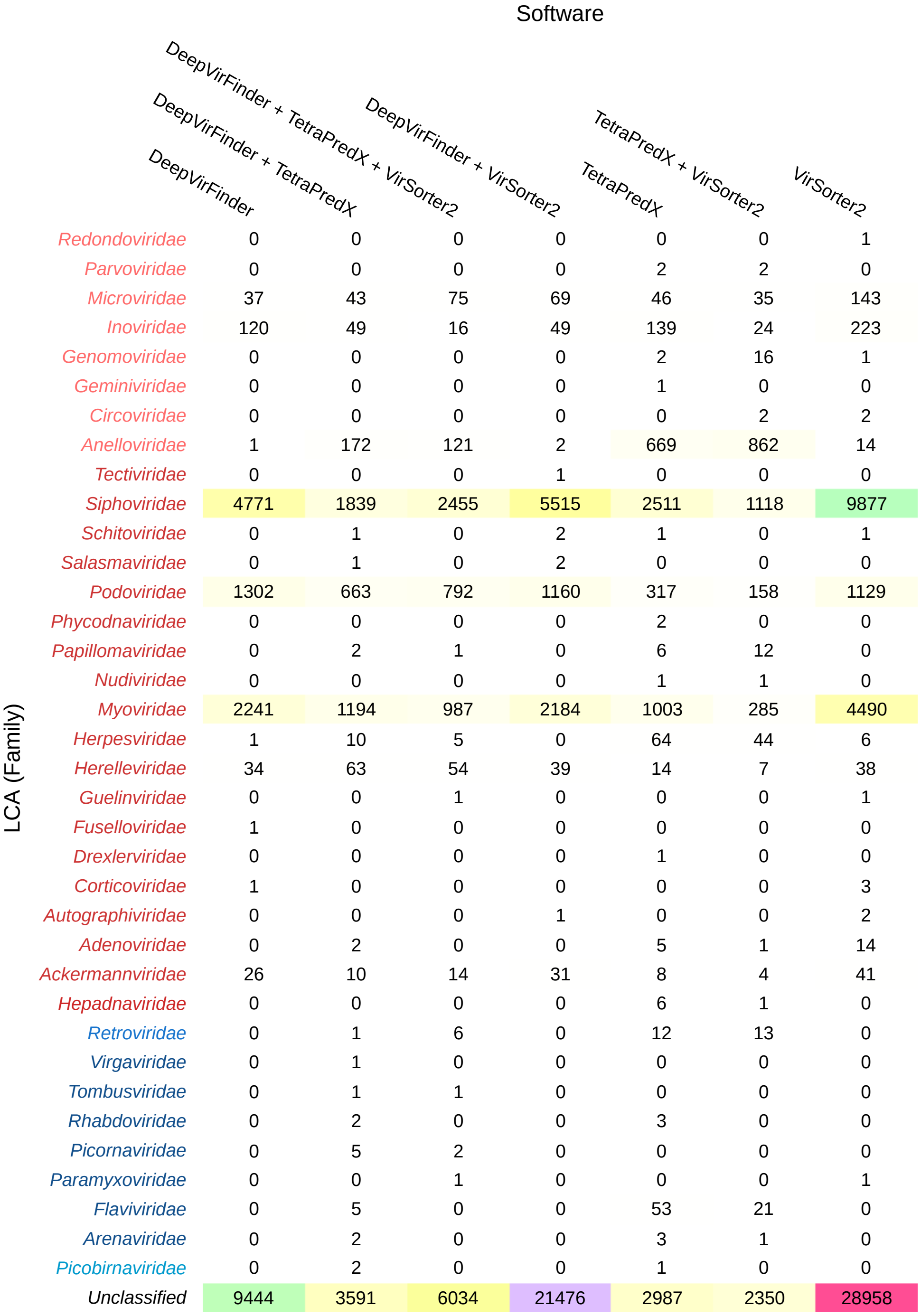


Figure 5.2: Virus prediction results for all three prediction software: DeepVirFinder, TetraPredX and VirSorter2 (a) A Venn diagram showing the overlap between contigs predicted as viral using all three tools. (b-h) Individual sunburst plots showing the number and proportion of sequences classified into different taxonomy superkingdoms using BLAST against the nt database for all sets included in the Venn diagram. Purple bands represent sequences that match exclusively to virus sequences in the databases, and yellow bands represent the sequences that match exclusively to bacterial sequences. Green bands represent the sequences that have matches to more than one superkingdoms (hence assigned to "root") and the light pink band represents the proportion of these sequences with at least one virus hit. Blue bands represent sequences that match eukaryotic sequences.



Software

LCA (Family)	DeepVirFinder	DeepVirFinder + TetraPredX	DeepVirFinder + TetraPredX + VirSorter2	DeepVirFinder + VirSorter2	TetraPredX	TetraPredX + VirSorter2	VirSorter2
Redondoviridae	0	0	0	0	0	0	1
Parvoviridae	0	0	0	0	2	2	0
Microviridae	37	43	75	69	46	35	143
Inoviridae	120	49	16	49	139	24	223
Genomoviridae	0	0	0	0	2	16	1
Geminiviridae	0	0	0	0	1	0	0
Circoviridae	0	0	0	0	0	2	2
Anelloviridae	1	172	121	2	669	862	14
Tectiviridae	0	0	0	1	0	0	0
Siphoviridae	4771	1839	2455	5515	2511	1118	9877
Schitoviridae	0	1	0	2	1	0	1
Salasmaviridae	0	1	0	2	0	0	0
Podoviridae	1302	663	792	1160	317	158	1129
Phycodnaviridae	0	0	0	0	2	0	0
Papillomaviridae	0	2	1	0	6	12	0
Nudiviridae	0	0	0	0	1	1	0
Myoviridae	2241	1194	987	2184	1003	285	4490
Herpesviridae	1	10	5	0	64	44	6
Herelleviridae	34	63	54	39	14	7	38
Guelinviridae	0	0	1	0	0	0	1
Fuselloviridae	1	0	0	0	0	0	0
Drexlerviridae	0	0	0	0	1	0	0
Corticoviridae	1	0	0	0	0	0	3
Autographiviridae	0	0	0	1	0	0	2
Adenoviridae	0	2	0	0	5	1	14
Ackermannviridae	26	10	14	31	8	4	41
Hepadnaviridae	0	0	0	0	6	1	0
Retroviridae	0	1	6	0	12	13	0
Virgaviridae	0	1	0	0	0	0	0
Tombusviridae	0	1	1	0	0	0	0
Rhabdoviridae	0	2	0	0	3	0	0
Picornaviridae	0	5	2	0	0	0	0
Paramyxoviridae	0	0	1	0	0	0	1
Flaviviridae	0	5	0	0	53	21	0
Arenaviridae	0	2	0	0	3	1	0
Picobirnaviridae	0	2	0	0	1	0	0
Unclassified	9444	3591	6034	21476	2987	2350	28958

Figure 5.3: A heatmap showing the number of predicted viral contigs using any prediction tool for each family (derived from the LCA). Darker shades of colours represent the higher number of contigs in the heatmap. The LCA family is coloured according to the genomic composition. Shades of red colour represent DNA virus families, shades of blue represent RNA virus families and black represents unclassified viruses.

Paramyxoviridae, Picornaviridae, Rhabdoviridae, Tombusviridae, Virgaviridae, Retroviridae and *Picobirnaviridae* were also identified. However, the majority of viral contig sequences identified were determined to be unclassified viruses (figure 5.3). A large proportion of these unclassified virus sequences were shown to be unclassified phages as shown in the appendix figure C.1. TetraPredX identified more RNA viruses belonging to the families *Retroviridae* and *Flaviviridae* (figure 5.3) as well as unclassified viruses belonging to the LCA realm *Riboviria* (n=42, figure C.1).

MIUViG quality was also assessed for viral contigs and the results are shown in figure 5.4. 122,993 (98.8%) of viral contigs were categorised as genome fragments and the remaining 1,499 viral contigs were deemed high quality meaning that they were at least 90% complete genome sequences. The highest number of complete viral genome sequences was predicted by VirSorter2 (n=1,417) followed by TetraPredX (n=683) and the least number of complete viral genome sequences were included in the DeepVirFinder (n=647) set. 186 high-quality virus genome sequences were identified by all three prediction tools, 447 were predicted by DeepvirFinder and VirSorter2, and 419 sequences were identified by VirSorter2 and TetraPredX. Additionally, 365, 68, and 4 were unique to VirSorter2, TetraPredX and DeepVirFinder respectively. It is worth noting that this set contains 1,609 provirus contigs that were identified using CheckV analysis hence the total number of contigs analysed in this section was 124,493 instead of 122,884. The majority of the provirus contigs had MIUVIG quality assigned as Genome-fragment (n=1,532). These provirus contigs were subsequently removed from the downstream analysis.

5.4.2 Bio-sample distribution of viruses

A top-level overview of various viral taxonomic groups across all 58 BioProjects sampled and analysed in this study is shown in figures 5.5 and 5.6. This analysis was based on 122,884 contig sequences that were deemed to be originating from virus genome sequences based on the virus prediction and validation analyses.

This analysis of human microbiome datasets captured samples and studies from around the world as shown in figure 5.5. From 58 BioProjects, the top three countries were the USA (n=9), the UK (n=8), and China (n=4) and the remaining samples were from a range of other countries spanning all continents of the world. 15 BioProjects out of 58 did not have any confirmed virus hits consolidated in the final set of viral contigs. Realm *Duplodnaviria* was present across most BioProjects (figure 5.6). The major prokaryotic virus families such as *Inoviridae*, *Microviridae*, *Myoviridae*, *Podoviridae* and *Siphoviridae* were also present across all studies reflecting their corresponding realm. RNA virus realm *Riboviria* was found in ten BioProjects and they were associated with nine distinct RNA virus families (figure 5.5). Notably, unclassified viruses i.e. that that could not be associated with a virus family, were found in 37 studies. This suggests that contig sequences matching novel/unclassified virus sequences were found across many BioProjects and geolocations.

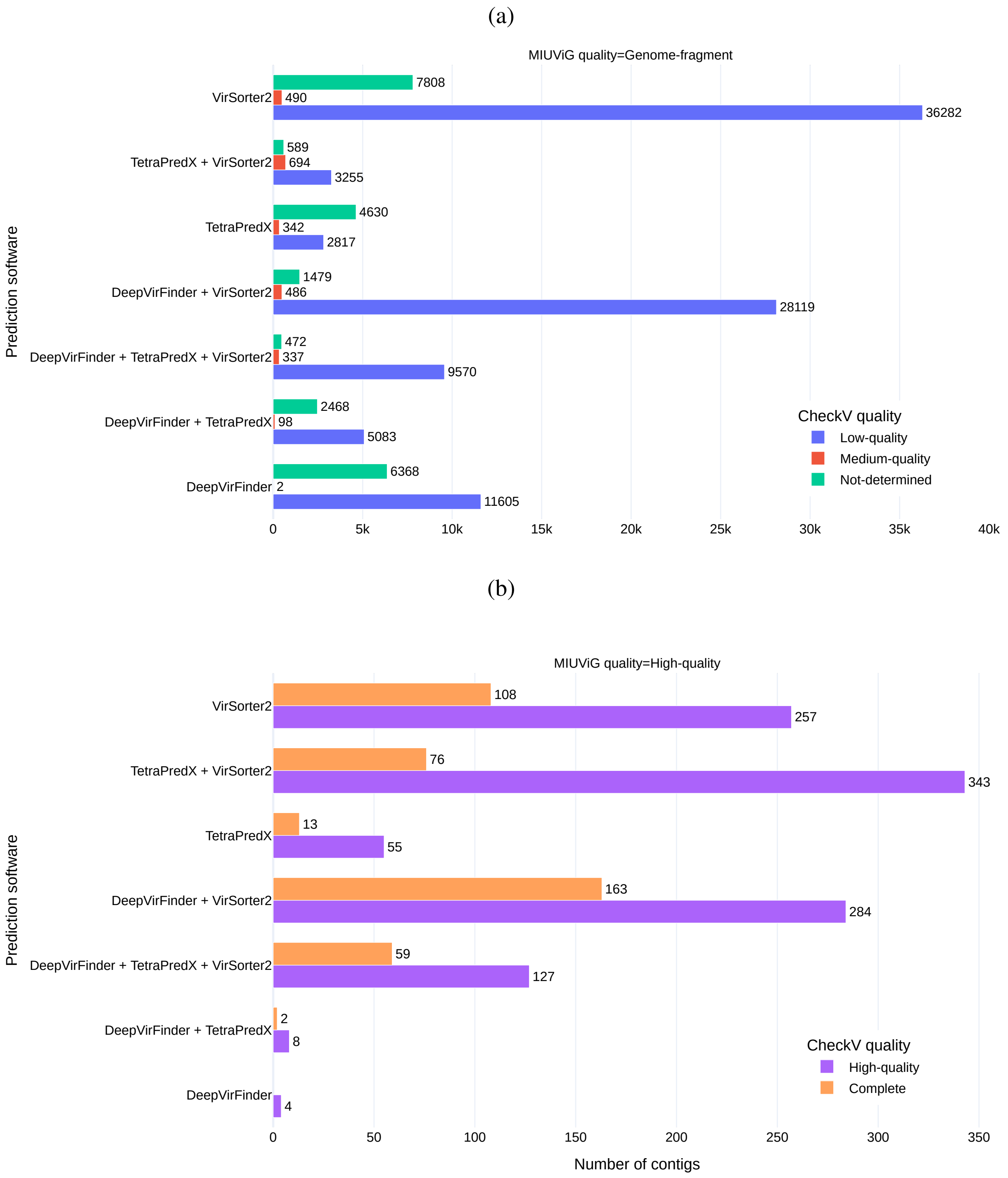


Figure 5.4: Grouped bar plots showing the quality of predicted viral contigs where similarity search derived LCA was “Viruses". The colours of the bar represent the CheckV quality categories for the prediction tool(s) specified on the Y-axis and the corresponding number of contigs are shown on the X-axis. (a) The MIUViG quality category Genome-fragment includes the following three CheckV quality criteria: Low-quality, Medium-quality and Not-determined. This is shown in the top grouped bar plot. (b) The MIUViG quality category High-quality includes the following two CheckV quality criteria: High-quality and Complete. The contigs for this category is shown in the bottom grouped bar plot.

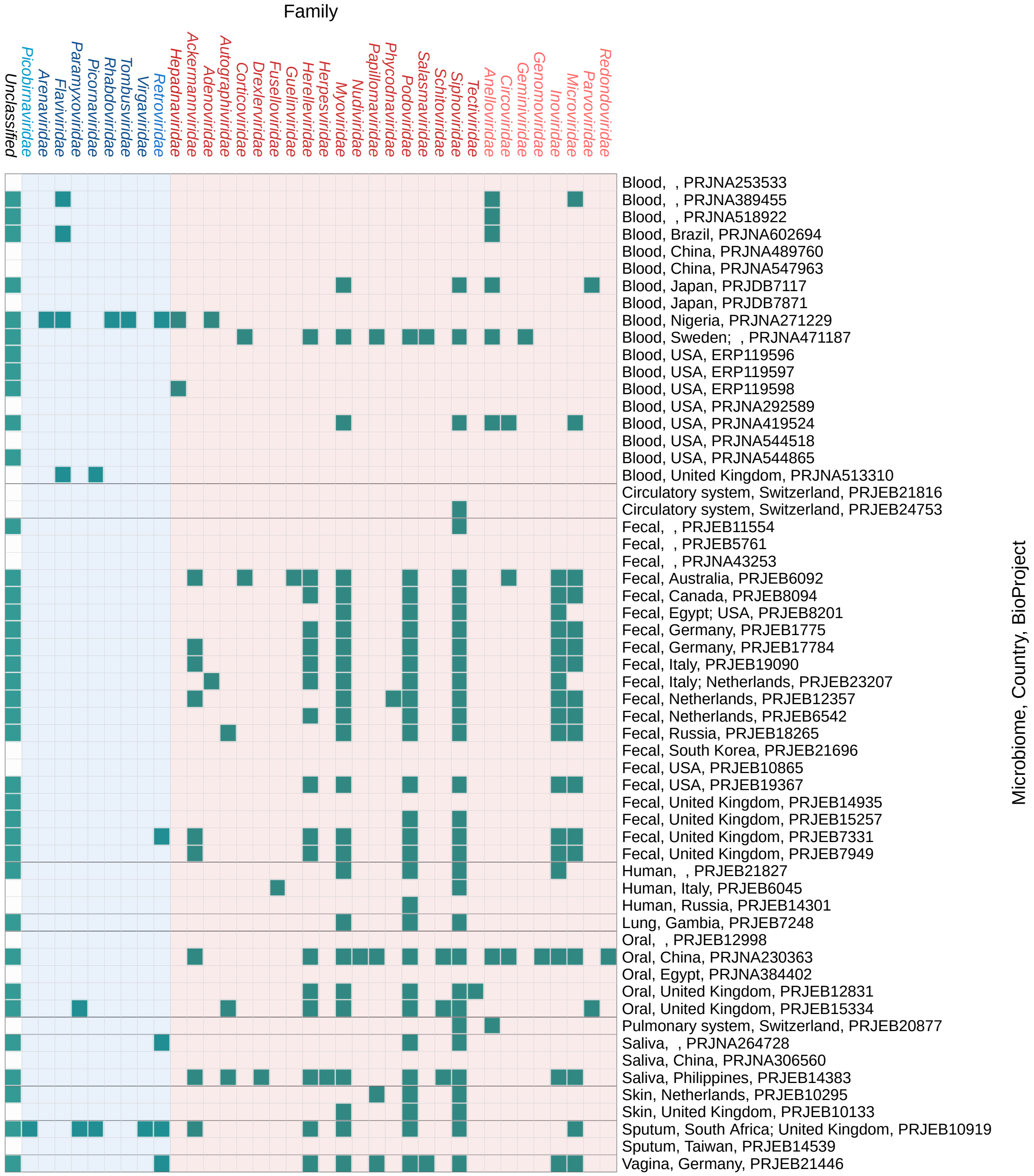


Figure 5.5: Distribution of 36 virus families across 58 BioProjects analysed in this study. The BioProjects on the Y-axis are grouped according to the microbiome (sample type) and sample locations. The virus families are shown on the X-axis. DNA virus families are labelled in shades of red, RNA virus families are labelled in shades of blue and unclassified viral sequences are labelled in black. Green squares represent the presence of the virus family in the corresponding BioProject specified on the Y-axis and non-green squares represent an absence of the specific group of viral contigs in the corresponding BioProject(s). Samples from 9 studies could not be associated with a geographic location and their country locations are shown as blank.

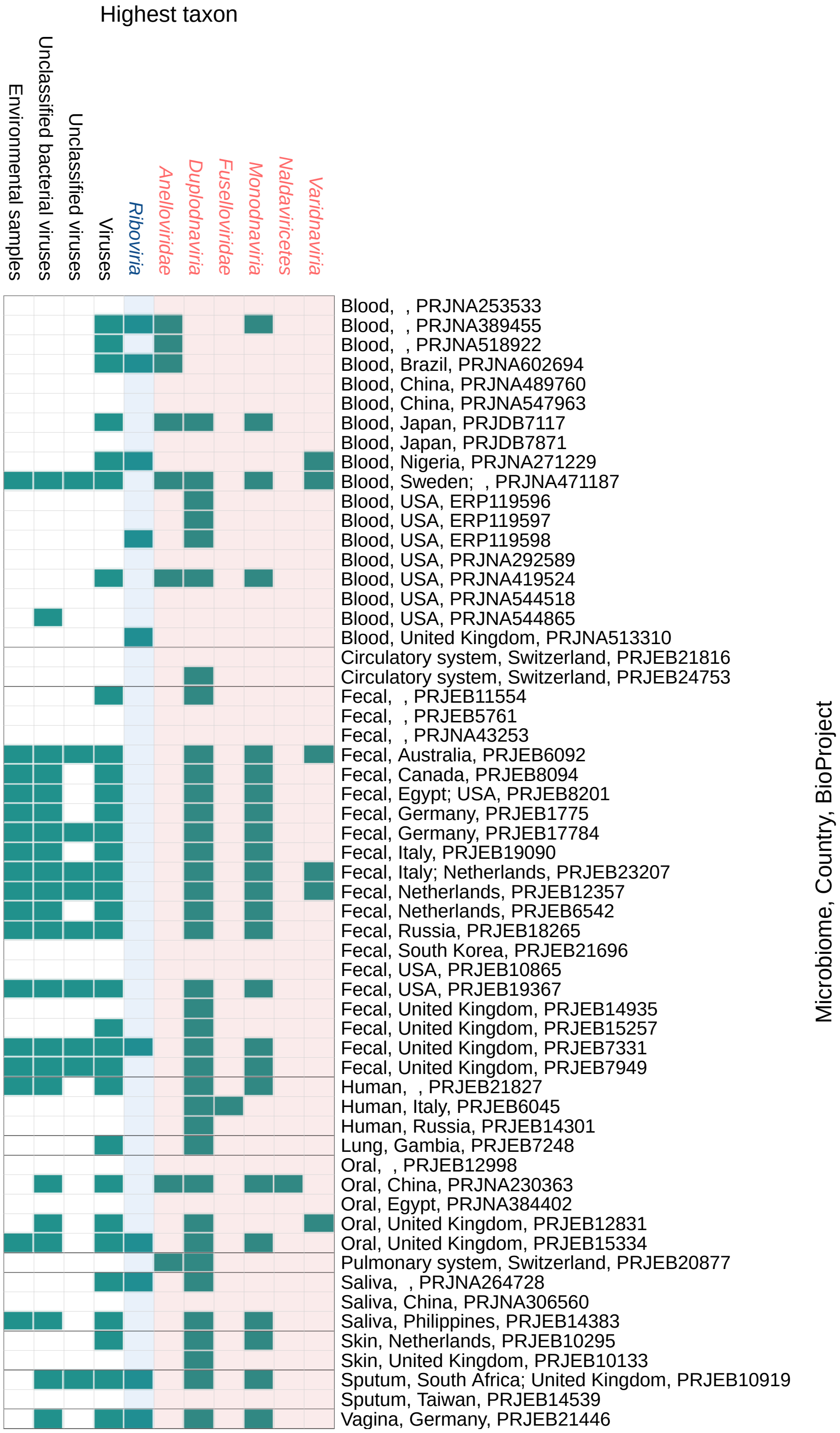


Figure 5.6: Distribution of higher-level taxon groups (i.e. realms or unclassified sequence groups) across 58 BioProjects analysed in this study. The BioProjects on the Y-axis are grouped according to the microbiome (sample type) and sample locations. The virus taxonomic groups are shown on the X-axis. DNA virus groups are labelled in red, RNA virus groups are highlighted in blue and unclassified viral sequences in black. Green squares represent the presence of the virus taxonomic group in the corresponding BioProject specified on the Y-axis and the non-green squares represent an absence of the virus taxonomic group. Samples from 9 studies could not be associated with a geographic location and their country locations are shown as empty boxes.

Although this overview comparing all 58 BioProjects provide some insights into the viruses that were found in relevant BioProjects, sample size and sample selection are highly skewed in this study. For example, as described in the Chapter 3, a range of studies that contained >100 samples were arbitrarily limited to 100 samples and no such filter was applied to the blood microbiome set. The uneven sampling of distinct microbiomes here can provide some idea of viruses that are present in these samples, but we are unable to rule out those that are absent. This is because viruses that were not found in these samples could be present at a low level and they are not detectable by the virus discovery analysis approaches employed here.

5.4.3 Co-occurrence analysis

The co-occurrence among viruses can be measured in microbiome datasets such as those analysed and described in this chapter. This analysis can provide insights into interactions of various virus groups and it can be used to interpret these interactions further in the context of other relevant metadata obtained from SRA databases such as geolocation and/or microbiome. Co-occurrence analysis could also indicate correlated species/family distributions across different microbiome, BioProjects and/or geolocations. To explore the co-occurrence patterns of virus families, co-occurrence analysis was carried out using the R package cooccur (Griffith et al., 2016). The results obtained from this analysis are categorised as either random, positive or negative co-occurrences. Random co-occurrence simply indicates that two taxa are distributed independently of each other. Positive co-occurrence indicates that a pair of taxonomic groups are more likely to be observed together and a negative co-occurrence stipulates the opposite meaning they are less likely to be observed together. It is notable that co-occurrence does not indicate direct interactions between taxa but simply shows that certain taxa are more or less likely to be found together in a given environment based on the presence/absence data.

All observed pairwise interactions are compared against the expected probabilities in probabilistic models implemented in this package and are described in detail (Veech, 2013). Briefly, if the observed probabilities of co-occurrence are significantly greater than the expected probabilities then the model deemed those interactions as positively correlated. In contrast, if the observed probabilities of co-occurrence are significantly less than the expected probabilities then those interactions are deemed negatively correlated. Finally, if there is no significant difference between the observed and expected probabilities then the interactions are considered random (Veech, 2013; Griffith et al., 2016).

To facilitate this analysis, all samples i.e. individual SRA IDs were treated as independent entities. A sample versus virus family matrix was generated indicating the presence/absence of the 36 virus families identified. This dataset was generated from the confirmed viruses set (n=122,884) where the family of a viral contig hits could be determined. To gather a top-level overview of virus family-level interactions regardless of other metadata (e.g. microbiome, geolocation, DNA/RNA sampling) all interactions were analysed as a whole. A total of 630 family

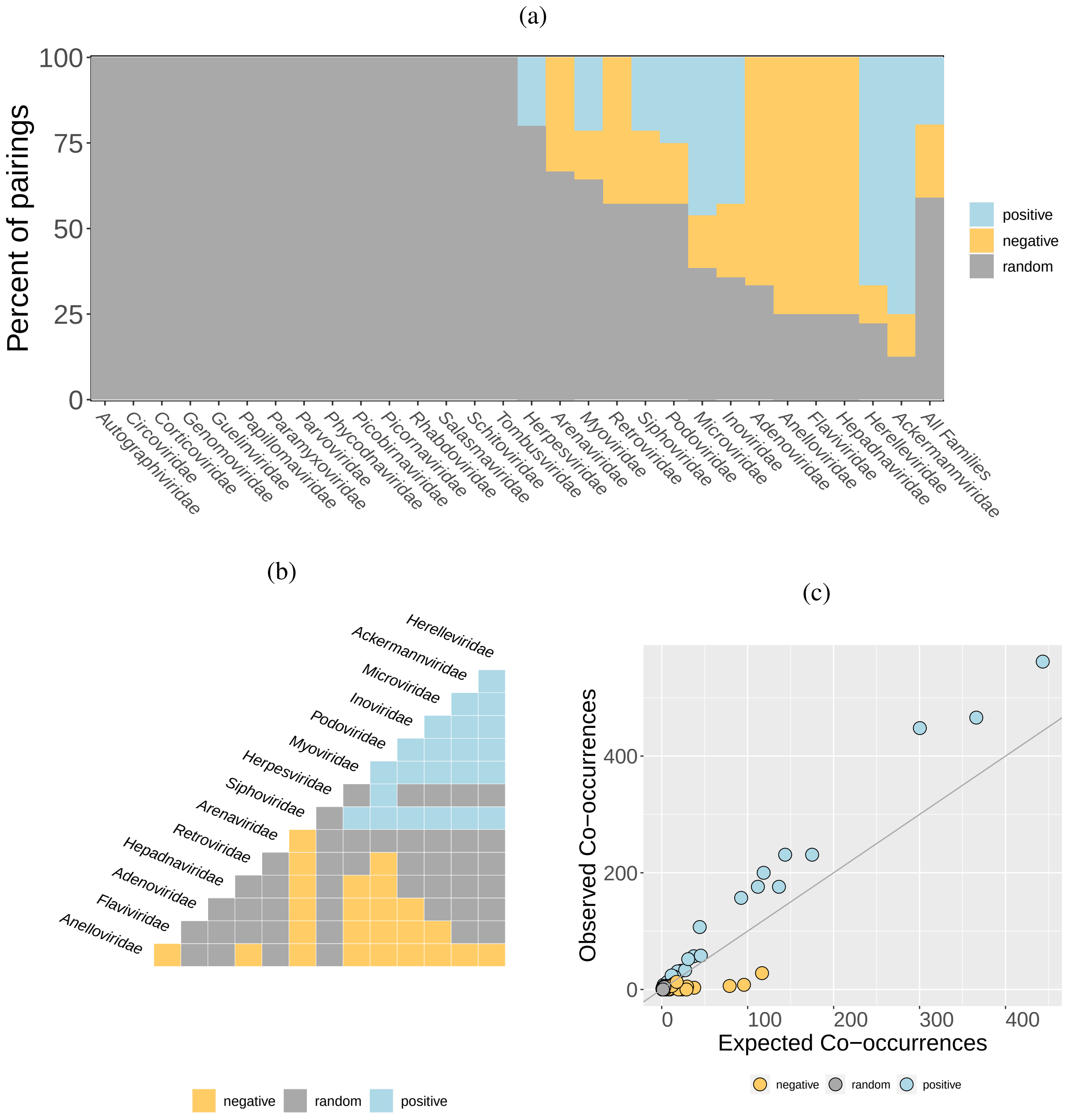


Figure 5.7: Co-occurrence analysis of virus families found in all samples. Blue represents a positive association; yellow represents a negative association and grey represents the random association in all plots. (a) Family association plot quantifying the percentage of positive, negative and random associations that each virus family has with all other viral families identified. (b) A heatmap describing pairwise associations between families, only families with at least one positive or negative association is shown. (c) Observed vs expected co-occurrences plotted for each pair of associations. The grey line represents the random co-occurrence.

pair combinations were analysed for 36 unique families (number of unique pairs = $n(n-1)/2$ where $n = 36$). 518 pairs (82.22 %) were removed from the analysis because the expected co-occurrence was < 1 and 112 pairs were analysed further. This step only retained virus family pairs that were observed together in a given sample. The results obtained for all families across all samples are shown in figure 5.7. Most virus families were randomly co-occurring with other families leading to insignificant interactions (figure 5.7(a)). This is also highlighted in the cumulative interactions as overall 60% of all pairwise associations between virus families were deemed to be random (figure 5.7(a) ‘All Families’) suggesting that the presence/absences of most virus families were distributed independently. Negative co-occurrences were observed between DNA and RNA virus families in figure 5.7(b) that is likely to be representative of library preparation and sequencing techniques. DNA viruses such as those included in the family *Anelloviridae* are typically captured through metagenomic approaches that sequence the DNA present in a sample whereas RNA families such as *Flaviviridae* and *Hepadnaviridae* are sampled through metatranscriptomic techniques that target and sequence the total RNA present in a sample. Unless a sample was prepared in a manner that both DNA and RNA contents were sequenced homogeneously, it is expected that either DNA viruses or RNA of viruses will be captured. Moreover, negative associations between DNA virus families *Anelloviridae* and a range of phage families e.g. *Siphoviridae* observed here are likely to represent the type of microbiomes (e.g. fecal, blood, skin etc) these viruses are typically found in. For example, as shown in figure 5.5 anelloviruses are predominantly found in blood samples whereas bacteriophages are common in fecal samples. This is because phages rely on their bacterial hosts to replicate and these bacteria are more likely to be commensal in fecal microbiome samples compared to blood. Hence, the negative co-occurrences observed here are most likely to be a shortfall in grouping different types of samples together to perform a high-level analysis. In contrast, positive co-occurrences were common between bacteriophage families as they are found in similar microbiome types. A positive co-occurrence between families *Herpesviridae* and *Podoviridae* was observed. Further analysis indicated that these two families were found together in 6 saliva samples from the Philippines included in the BioProject PRJEB14383 (figure 5.5). It is worth pointing out that these two virus families have been known to coexist in human virome and have been observed to be present together in the oral cavity in previous studies (Liang et al., 2021).

To explore these correlations further, and test whether some of the co-occurrence patterns observed were indeed due to artificial grouping of different microbiome samples, a similar co-occurrence analysis was carried out for a subset of microbiome specific. This dataset included to the following microbiomes: fecal, blood and oral. Only a subset of microbiomes could be analysed in this manner as enough samples and observations were required to carry out an analysis that tests the statistical significance of observing two taxonomic group together in a given sample. Other microbiomes such as skin, sputum, and pulmonary system contained too few samples to quantify statistically significant interactions using `cooccur` package.

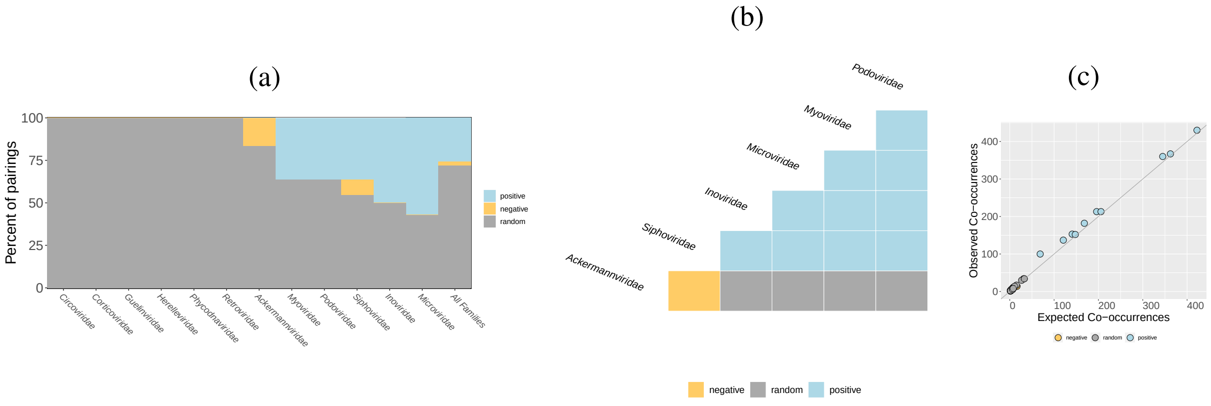


Figure 5.8: Co-occurrence analysis of virus families found in fecal samples. Blue colour indicates positive association; yellow indicates negative association and grey represents random associations in all plots. (a) Family association plot quantifying the positive, negative and random associations for each virus family. (b) A heatmap describing pairwise associations between families. (c) Observed vs expected co-occurrences plotted for each pair of associations.

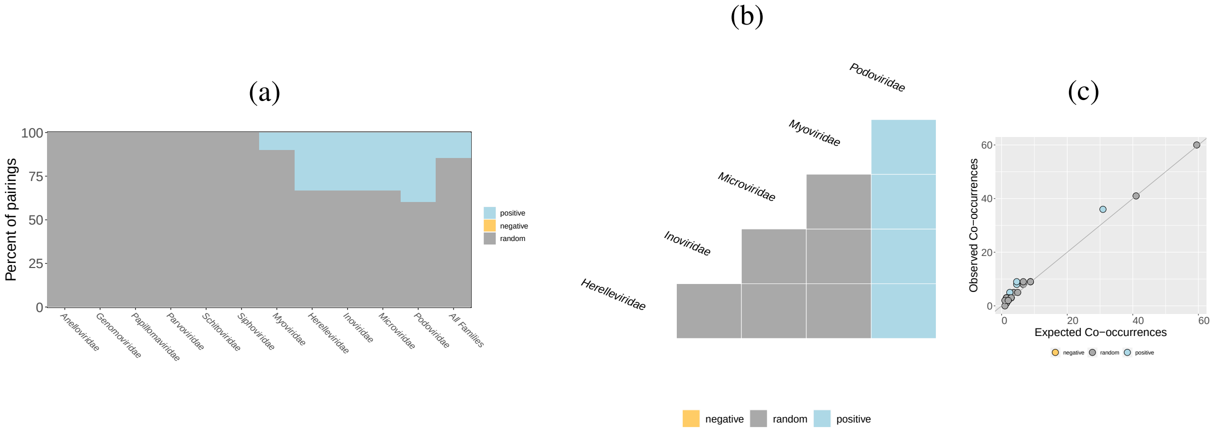


Figure 5.9: Co-occurrence analysis of virus families found in oral samples. Blue colour indicates positive association; yellow indicates negative association and grey represents random associations in all plots. (a) Family association plot quantifying the positive, negative and random associations for each virus family. (b) A heatmap describing pairwise associations between families. (c) Observed vs expected co-occurrences plotted for each pair of associations.

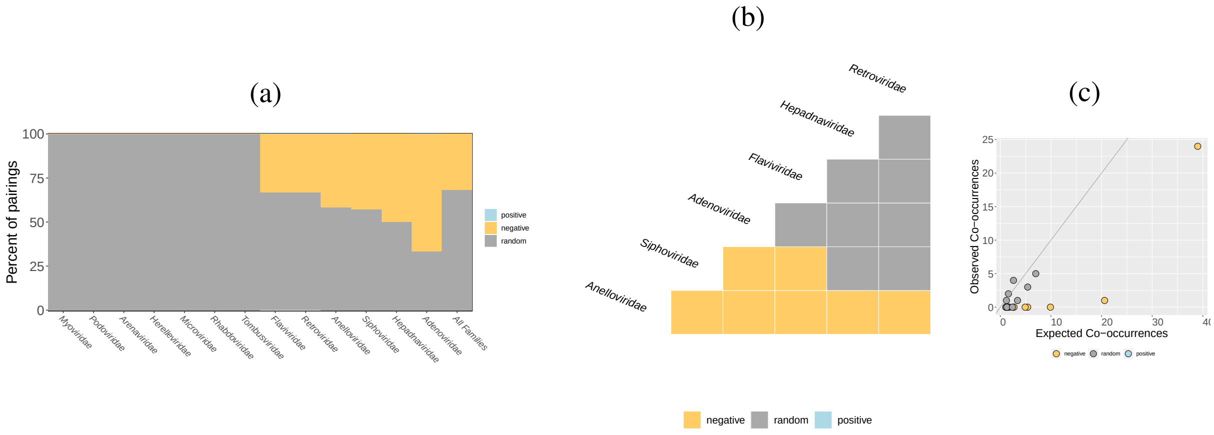


Figure 5.10: Co-occurrence analysis of virus families found in blood samples. Blue colour indicates positive association; yellow indicates negative association and grey represents random associations in all plots. (a) Family association plot quantifying the positive, negative and random associations for each virus family. (b) A heatmap describing pairwise associations between families. (c) Observed vs expected co-occurrences plotted for each pair of associations.

Out of 630 unique pairwise combinations, 39 associations were analysed for fecal microbiome, 27 pairs were analysed for oral microbiome and 22 pairs were analysed for blood microbiome. These microbiome-specific co-occurrence results are shown in figures 5.8, 5.9 and 5.10. Overall, 70% of all associations were determined to be random for all three microbiomes. The negative associations observed at all sample levels between ssDNA virus family *Anelloviridae* and bacteriophage families were absent in the fecal microbiome-specific analysis suggesting that these two types of viruses were unlikely to be present together in the fecal samples analysed here. On the contrary, negative associations between anelloviruses and siphoviruses were observed in the blood microbiome-specific analysis (figure 5.10). No positive co-occurrences were observed in blood microbiome specific set and no negative correlations were found between virus families found in the oral microbiome specific set.

5.4.4 Exploring the prokaryotic virus diversity embedded within human microbiomes

9,218 potentially prokaryotic virus sequences were extracted from the set of 271,805 final contigs set derived here and were analysed along with GPD (n=142,809) and NCBI (n=18,366) genomes datasets. This generated a final set of 170,939 sequences that were analysed further. The contigs obtained from our analysis are labelled as SM set for ease of understanding. A distribution of LCA for each of these contigs is shown in the bar plot shown in the figure 5.11 coloured and faceted by the contig quality. 1,141 contigs were deemed of High-quality suggesting that they are nearly complete or complete genomes and the remaining 8,077 contigs were categorised as Genome-fragment as per the MIUViG quality assessment criteria. Moreover, as shown in figure 5.11, a large number of contigs were classified at higher taxonomy levels such as the superkingdom level ‘Viruses’ or the order level *Caudovirales* suggesting that virus sequences assembled in this study are matching to a range of different sequences included in nt databases as opposed to matching to a specific species or genus of phage classification. High-quality contig sequences were found in fecal, saliva, oral and lung microbiomes whereas those categorised as Genome-fragments were found in all microbiomes.

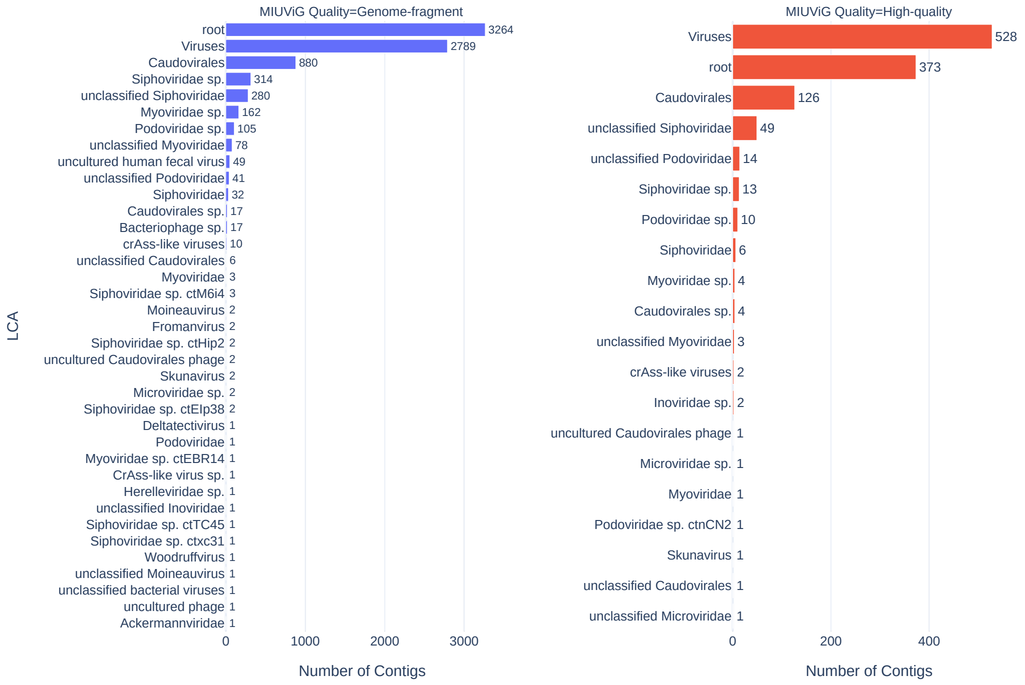


Figure 5.11: Bar chart showing the number of contigs (X-axis) and their corresponding lowest common ancestors (LCA) on the Y-axis. Blue bars (left-hand plot) represent contigs with MIUViG quality = Genome-fragments whereas red bars (right-hand plot) represent those with MIUViG quality = High-quality.

USEARCH-like clustering carried out using the 95% average nucleotide identity with 85% coverage of 170,393 sequences yielded 97,956 clusters. A total of 78,252 sequences were singletons and 2,243 of them were from the SM dataset. Out of these 2,243 singletons, 215 were High-quality as per MIUViG quality criteria indicating that they were near-complete or complete genomes. The remaining 2,028 were categorised as Genome-fragments of which 402 were medium-quality as per the CheckV quality assessment criteria suggesting that they were at least 50% complete genome sequences.

In total, 19,704 clusters with at least 2 sequences were generated, comprising 92,141 sequences. For each cluster with at least 2 sequences, a sequence is defined as a cluster representative which is typically the longest sequence in the cluster. Henceforth, the term ‘representative’ is used to describe a cluster representative sequence in this context. These representative sequences are often labelled as viral operating taxonomic units (vOTUs) as recommended by MIUViG (Roux et al., 2019). As we have utilised the standard MIUViG criteria of 95% average nucleotide identity (ANI) over 85% alignment fraction (AF) relative to the shorter sequence - to cluster sequences in NCBI, GPD and SM set, the cluster representatives, as well as the singletons (sequences that do not cluster with anything else), can also be referred to as vOTUs. vOTUs typically describe the species level, and virus groups (Roux et al., 2019). As GPD sequences made up the largest proportion of sequences in the overall dataset, as expected the largest number of clusters i.e. 16,328 were represented by a sequence assembled and catalogued in GPD. 14,392 (88.14%) of these clusters were exclusive to the GPD dataset meaning they only contained sequences from the GPD set whereas 1,936 (11.86%) clusters comprised sequences from all three datasets (GPD, NCBI, SM). 82 clusters comprised sequences from both GPD and NCBI set where the cluster representative was from GPD set, and similarly, there were 1,832 clusters from GPD and SM set with cluster representative from the GPD set. 2,728 clusters were represented by vOTUs originating from the NCBI set and among them, 94.21% (n=2,570) were represented by genome sequences exclusive to the NCBI set only, whereas 5.79% (n=158) were comprised of sequences from all three datasets. 7 clusters contained sequences that originated only from NCBI and SM datasets where a sequence from the NCBI dataset was cluster representative whereas 142 clusters contained sequences from both NCBI and GPD datasets where the cluster representative was from the NCBI set. Finally, 648 clusters were represented by the vOTUs belonging to the SM dataset. 72.99% (n=473) of these clusters were exclusive to the SM dataset and 27.01% (n=175) of clusters included sequences from all three datasets. 175 clusters that had SM dataset sequence as a cluster representative contained sequences from both SM and GPD sets. No clusters exclusive to SM and NCBI set were observed where a cluster representative was from the SM set. Overall, the largest cluster contained 896 sequences of which 809 were from the GPD set, 77 sequences were from the SM dataset and the remaining 10 sequences were from the NCBI dataset. The vOTU for this cluster was found to be similar to crAssphage which is known to be the most dominant phage identified in the gut microbiome samples (Bas E. Dutilh et al., 2014) and it was also found in fecal samples originating from 7 different BioProjects in the SM set.

The clustering analysis yielded 2,891 (2,243 singletons and 648 cluster representatives) viral operational taxonomy units (vOTUs) where the cluster representative vOTU was a sequence originating from the SM dataset. 396 (13.7%) vOTU sequences from this set were complete or nearly complete genomes (MIUViG quality = High-quality) and the remaining 2,614 sequences (90.42%) were classified as Genome-fragment as per the MIUViG quality criteria. 180 of these High-quality vOTUs were cluster representatives and 106 of these clusters were exclusive to the SM dataset. vOTU distributions were biased towards the fecal microbiome, reflecting both sampling biases for the microbiome as well as sampling biases caused by prokaryotic viruses present in fecal samples. 1,416 vOTUs originated from the fecal microbiome followed by 947 vOTUs from saliva and 482 from oral microbiomes. The remaining vOTUs were as follows: 12

each were from blood and vagina microbiomes, 11 from sputum, 5 from misc label human, 3 from the pulmonary system and 2 from lung microbiomes.

The largest SM dataset-specific cluster was obtained from the fecal microbiome BioProject PRJEB8094 which contained 30 sequences from 26 samples from the same study. The cluster representative sequence was deemed Genome-fragment suggesting that this was a partial sequence. Another cluster from the same study with 28 cluster members was a complete genome and was represented by vOTU ERR719882_NODE_74. This sequence matched to bacteria and viruses using BLASTN analysis against nt, was predicted as dsDNA phage using VirSorter2 (score=0.993) and was deemed complete genome with the presence of 55 bases direct terminal repeat sequence 'CCGCCTTGTAAATGCCTGACCTTTTATTCGTTTACCGTTTTTCATAAAAATATAT'. A schematic representation of this virus genome is shown in the figure 5.12 with ORFs identified using the DRAM-v annotations pipeline shown along with the genome and GC content of the sequence shown in the plot below.

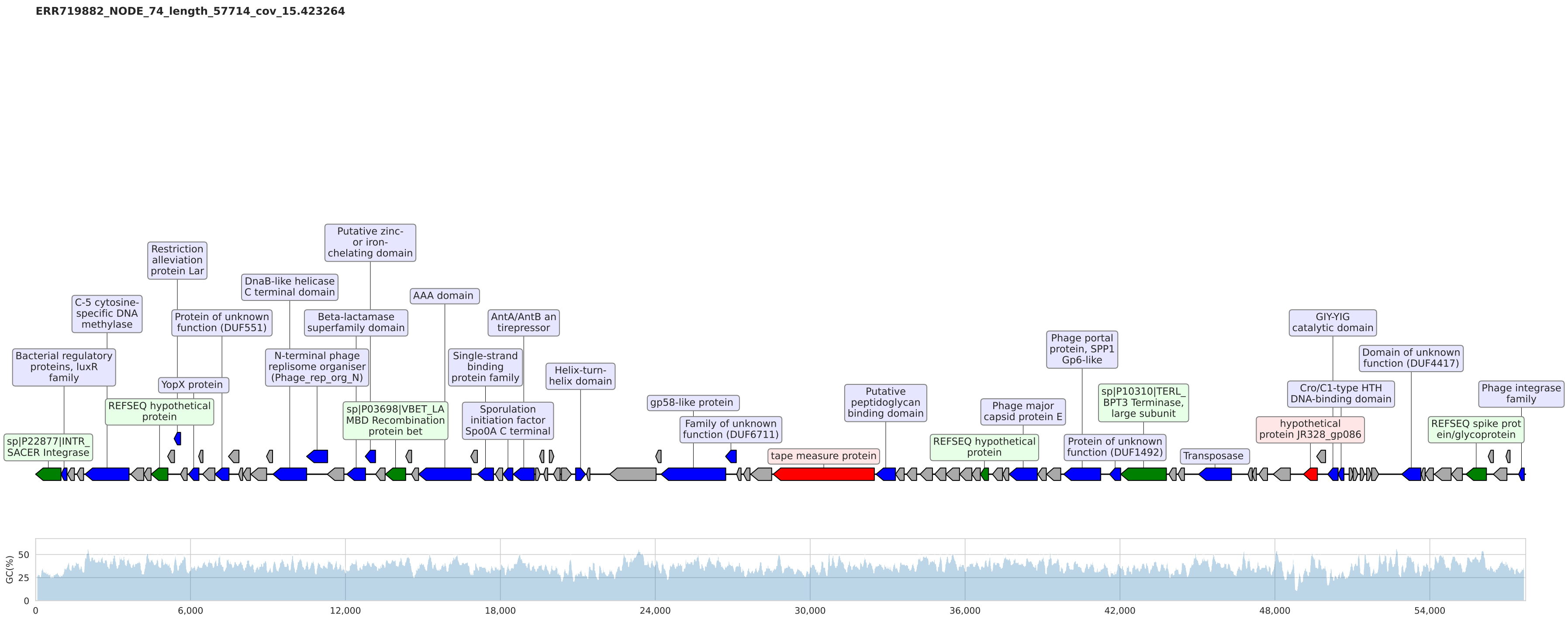

ERR719882_NODE_74_length_57714_cov_15.423264
sp|P22877|INTR_SACER Integrase
Bacterial regulatory proteins, luxR family
C-5 cytosine-specific DNA methylase
REFSEQ hypothetical protein
Restriction alleviation protein Lar
YopX protein
Protein of unknown function (DUF551)
DnaB-like helicase C terminal domain
N-terminal phage replisome organiser (Phage_rep_org_N)
Beta-lactamase superfamily domain
Putative zinc- or iron-chelating domain
sp|P03698|VBET_LAMBD Recombination protein bet
AAA domain
Single-strand binding protein family
Sporulation initiation factor Spo0A C terminal
AntA/AntB antirepressor
Helix-turn-helix domain
gp58-like protein
Family of unknown function (DUF6711)
tape measure protein
Putative peptidoglycan binding domain
REFSEQ hypothetical protein
Phage major capsid protein E
Phage portal protein, SPP1 Gp6-like
Protein of unknown function (DUF1492)
sp|P10310|TERL_BPT3 Terminase, large subunit
Transposase
hypothetical protein JR328_gp086
Cro/C1-type HTH DNA-binding domain
GIY-YIG catalytic domain
Domain of unknown function (DUF4417)
REFSEQ spike protein/glycoprotein
Phage integrase family
GC(%)
50
25
0
0
6,000
12,000
18,000
24,000
30,000
36,000
42,000
48,000
54,000

Figure 5.12: The cluster representative of the largest SM dataset-specific cluster. This cluster contained 28 sequence originating from 25 samples. The representative vOTU is 57,714 bases long and this sequence was annotated using DRAM-v (Shaffer et al., 2020) to identify open reading frames (ORFs) and domain signatures. A schematic representation of the genome created using DNA feature viewer (Zulkower et al., 2020). The GC-content across the genome is shown in the subplot below. The ORFs are coloured according to the database hits and their code is as follows: Pfam (blue), KEGG (pink), VOGDB (green), NCBI virus sequence database (red). The grey ORFs show that no known viral domains were found in them.

Twenty vOTU clusters identified exclusively in the SM dataset contained sequences from more than one microbiome. The most commonly clustered microbiomes were either oral and saliva or saliva and sputum microbiomes which covered 19 out of 20 multi-microbiome clusters. The remaining multi-microbiome cluster contained two sequences; one from fecal and one from vagina samples. Six of these 20 multi-microbiome clusters were represented by vOTUs that were deemed High-quality indicating that they were complete genomes. All six of these vOTUs were predicted to be virus sequences using all three prediction tools. Four of them had viruses as LCA and two had root as LCA. These vOTUs sequences were between 33-98kb long and could not be linked to a known virus family. The VirSorter2 predictions score for all of them was 1 and they were predicted to be dsDNA phages. This suggests that these vOTUs are highly likely to be novel phages and their genomic makeup is likely to be significantly different compared to known phage sequences. The largest multi-microbiome cluster contained 9 sequences. This cluster had sequences from oral and saliva microbiomes from China (PRJNA230363; n=7) and the Philippines (PRJEB14383; n=2). It was represented by SRR2037090_NODE_11 - a partial genome sequence from the sample SRR2037090 that was most likely to be a member of the family *Myoviridae* based on the LCA analysis.

A total of 322 vOTUs that were unique to the SM dataset were deemed high-quality suggesting that they were either complete or nearly complete genomes. 84 of these vOTUs were predicted as virus genomes using all three prediction tools included in section 5.4.1, 111 were predicted as viral genomes using DeepVirFinder and VirSorter2 and 12 were predicted as viral genomes using both TetraPredX and VirSorter2. 321 of these contigs were annotated using the DRAM-v (Shaffer et al., 2020) pipeline using `annotate` and `distill` workflows. DRAM-v works very well with VirSorter2 workflow as VirSorter2 results can be manipulated to generate DRAM-v compatible input files. Although viral contigs that were not predicted using VirSorter2 can be annotated using DRAM-v, the `distill` part of the pipeline that consolidates the results and generated viral metagenome-assembled genomes (vMAGs) level summaries, cannot be executed without VirSorter2 prediction output. Out of 322 vOTUs, 321 were predicted using VirSorter2 (as well as other predictions tools) and 1 vOTU was exclusively predicted using DeepVirFinder and subsequently could not be efficiently annotated using DRAM-v. A summary of the resultant 321 vMAGs with metadata including cluster size, predicted genes, and a number of viral genes with the CRISPR host prediction are detailed in table C.2. 312 of these vMAGs were predicted to be in VirSorter (original) category 1 predictions indicating that they were high confidence predictions. The category 1 predictions indicate the presence of viral hallmark genes and viral-like genes. 9 vMAGs were categorised into VirSorter category 2 predictions indicating that these vMAGs were likely to be viral genomes. Category 2 predicted vMAGs tend to have genomic regions that have either enrichment in viral-like or non-Caudovirales genes, or a viral hallmark gene detected. These vMAGs have regions that are associated with at least one other metric defined by the VirSorter tool which includes: depletion (reduction) in Pfam affiliated genes, enrichment in uncharacterised genes, enrichment in short genes, and depletion in strand switch (Roux et al., 2015a). These metrics are also included in VirSorter2 and the categorisation was obtained from VirSorter2 and DRAM-v output.

Out of 321 vMAGs, 122 were assembled from the saliva microbiome, 117 were assembled from fecal microbiomes, 81 were from the oral microbiome and 1 was from the lung microbiome. All 122 vMAGs originating from the saliva microbiome were from BioProject PRJEB14383. This study originally included metagenomic sequencing of saliva/oral samples from individuals with hunter-gatherer or agriculturalist lifestyles from locations in the Philippines. The second-largest number of vMAGs (n=75) were assembled from BioProject PRJNA230363 that including oral samples from China. A range of vMAGs were associated with a number of fecal microbiome BioProjects. One vMAG was assembled from lung microbiome BioProject PRJEB7248.

These SM dataset-specific vMAGs contained 63.09 predicted genes on average with a standard deviation of 37.21. The largest number of genes were predicted for jumbo viruses with genome size >200kb. These are described in detail in section 5.4.4. These 321 vMAGs contained 38.23 viral hypothetical genes on average with a very high standard deviation of 33.66. Moreover, on average each vMAG contained 8.47 genes of unknown function (standard deviation of 12.56). vMAGs also contained 2.86 viral structural genes (standard deviation: 4.48) and 0.75 viral replication genes (standard deviation: 1.22). These results alluded that vMAGs identified and catalogued here, are likely to contain novel genes and proteins that are yet to be catalogued. These novel genes and proteins may be significantly different to those that are currently captured in the standard nucleotide and protein databases.

Jumbo phages

Prokaryotic virus genomes that are larger than 200kb are termed ‘jumbo phages’. In our set, 8 jumbo phage vOTUs were identified. Six of these vOTUs were singletons whereas the remaining two were clustered with 6 and 5 other contigs respectively. 5 of the vOTUs were predicted as viruses using all three prediction tools, 2 were predicted as viruses using DeepVirFinder and VirSorter2 and one was predicted as dsDNA phage by VirSorter2. Five of these jumbo phage vOTUs were found in four saliva samples from the Philippines in BioProject PRJEB14383 which also contained a large proportion of vOTUs as described in table C.2. The remaining were found in 2 oral samples from PRJNA230363 (China) and 1 oral sample from PRJEB15334 (UK). Overall, all jumbo phage vOTUs were associated with oral and/or saliva samples. vOTU ERR1474612_NODE_7 represented the largest cluster with 6 other sequences. This cluster contained 3 members from the GPD dataset and 3 other members from the SM dataset. The second cluster was represented by vOTU SRR2037090_NODE_1 and 4 cluster members were from the same BioProject - PRJNA230363. 7 out of 8 vOTUs were deemed High-quality suggesting that they were complete genomes. The largest jumbo phage was predicted as a dsDNA phage using VirSorter2 with a score of 0.993. This vOTU ERR1611403_NODE_2 was shown to have hits to sequences belonging to the virus family *Myoviridae* using BLAST analysis, however, it did not have significant sequence similarity to known viruses. The closest virus hit for this vOTU was found to Myoviridae sp. isolate ctjeh30 (accession: BK042204.1), however, the matches were sparse with the largest hit covering 5% of the vOTU sequence. Additionally, the DRAM-v annotations found a range of ORFs and domains within this sequence (figure 5.13) suggesting that this novel bacteriophage genome sequence was significantly different and potentially distantly related to existing and known myoviruses. This vOTU was shown to contain 127 base direct termini repeat sequences TGATATAATTACTGCAAAAAAATAAGGAAGGGCTCAATGCCCTTCCAATTCTTTTTCATTTTATGCAGATAATATTGTAGGTTTTCTACATTCTATAAAATTATCAAATGATTGTCTTAATCTATTACTACATTCTATAAAATTATCAAATGATTGTCTTAATCTATTA.

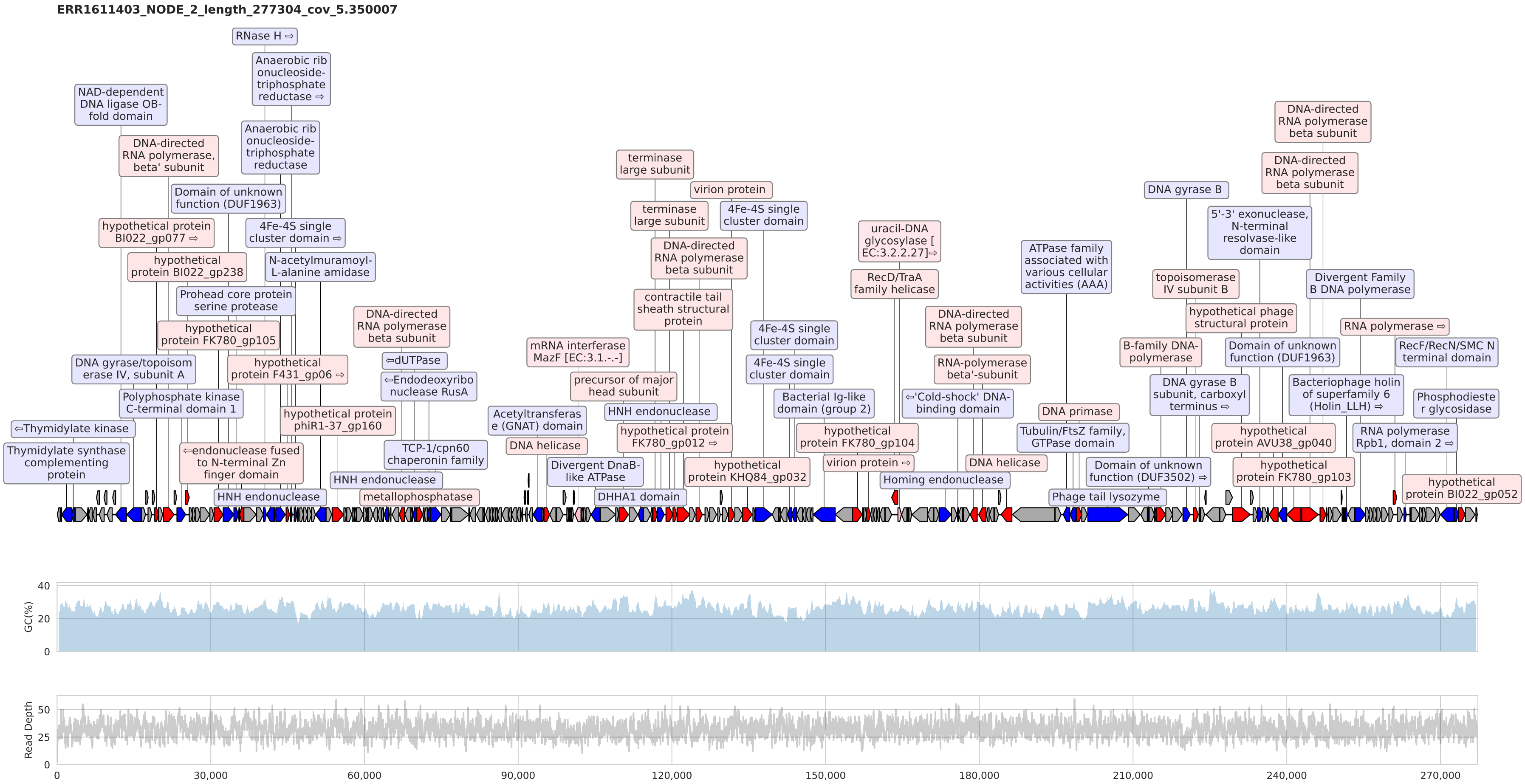

ERR1611403_NODE_2_length_277304_cov_5.350007
RNase H ⇨
Anaerobic rib onucleoside-triphosphate reductase ⇨
NAD-dependent DNA ligase OB-fold domain
Anaerobic rib onucleoside-triphosphate reductase
DNA-directed RNA polymerase, beta' subunit
Domain of unknown function (DUF1963)
hypothetical protein BI022_gp077 ⇨
4Fe-4S single cluster domain ⇨
hypothetical protein BI022_gp238
N-acetylmuramoyl-L-alanine amidase
Prohead core protein serine protease
hypothetical protein FK780_gp105
DNA gyrase/topoisom erase IV, subunit A
hypothetical protein F431_gp06 ⇨
Polyphosphate kinase C-terminal domain 1
hypothetical protein phiR1-37_gp160
⇦Thymidylate kinase
Thymidylate synthase complementing protein
⇦endonuclease fused to N-terminal Zn finger domain
HNH endonuclease
DNA-directed RNA polymerase beta subunit
⇦dUTPase
⇦Endodeoxyribo nuclease RusA
TCP-1/cpn60 chaperonin family
HNH endonuclease
metallophosphatase
mRNA interferase MazF [EC:3.1.-.-]
Acetyltransferas e (GNAT) domain
DNA helicase
Divergent DnaB-like ATPase
precursor of major head subunit
HNH endonuclease
hypothetical protein FK780_gp012 ⇨
DHHA1 domain
terminase large subunit
terminase large subunit
virion protein
4Fe-4S single cluster domain
DNA-directed RNA polymerase beta subunit
contractile tail sheath structural protein
hypothetical protein KHQ84_gp032
4Fe-4S single cluster domain
4Fe-4S single cluster domain
Bacterial Ig-like domain (group 2)
uracil-DNA glycosylase [EC:3.2.2.27]⇨
RecD/TraA family helicase
hypothetical protein FK780_gp104
virion protein ⇨
DNA-directed RNA polymerase beta subunit
RNA-polymerase beta'-subunit
⇦'Cold-shock' DNA-binding domain
DNA helicase
Homing endonuclease
ATPase family associated with various cellular activities (AAA)
DNA primase
Tubulin/FtsZ family, GTPase domain
Phage tail lysozyme
Domain of unknown function (DUF3502) ⇨
DNA gyrase B
topoisomerase IV subunit B
B-family DNA-polymerase
DNA gyrase B subunit, carboxyl terminus ⇨
5'-3' exonuclease, N-terminal resolvase-like domain
hypothetical phage structural protein
Domain of unknown function (DUF1963)
hypothetical protein AVU38_gp040
hypothetical protein FK780_gp103
DNA-directed RNA polymerase beta subunit
DNA-directed RNA polymerase beta subunit
Divergent Family B DNA polymerase
Bacteriophage holin of superfamily 6 (Holin_LLH) ⇨
RNA polymerase ⇨
RNA polymerase Rpb1, domain 2 ⇨
RecF/RecN/SMC N terminal domain
Phosphodieste r glycosidase
hypothetical protein BI022_gp052
GC(%)
40
20
0
Read Depth
50
25
0
0
30,000
60,000
90,000
120,000
150,000
180,000
210,000
240,000
270,000

Figure 5.13: A novel Jumbo phage genome assembled from SRA run ID ERR1611403. (a) A snapshot of the BLAST webpage shows that this novel virus genome does not bear significant sequence similarity to any known virus sequence in the database. (b) A genome diagram of the novel phage showing the different ORFs and domains across the jumbo phage genome that was annotated using DRAM-v. The GC-content across the genome is shown in the plot below the genome diagram that was calculated with a window size of 1000 bases. The read depth across the genome was calculated using samtools depth and is shown as grey track.

Virus host prediction

To predict the bacteriophage hosts, the CrisprOpenDB package was utilised. It includes a comprehensive database of more than 11 million sequences of spacers that can be searched extensively using dedicated software included in the package that can execute host predictions on large viral datasets. The spacers can provide accurate associations between phages and their bacterial hosts since they are derived from past interactions between phages and their hosts. From the set of 2,891 prokaryotic vOTUs obtained in section 5.4.4, a host could be predicted for 1,314 (45.45%) using CrisprOpenDB. Four different criteria are used for predictions made by the CrisprOpenDB tool as described in Dion et al. (2021). Level 1 predictions indicate that only one host genus is identified and it is assigned as the predicted host. Level 2 predictions include those that match more than one genera and in this case host targeting the highest number of regions in the phage genome will be assigned as the predicted host. Level 3 predictions are the same as level 2 except that if two or more genera have an equal number of matches then the host with spacers closest to the 5' end of the CRISPR array is selected. Finally, level 4 predictions were applied when all three aforementioned criteria failed to match a single host genus. In level 4 predictions, the last common ancestor of the remaining bacterial genera was calculated and was predicted to be the phage host (Dion et al., 2021).

Out of the 1,314 predicted hosts, 88.96% (n=1,169) were level 1 predictions, 7.99% (n=105) were level 2 predictions, 2.89% (n=38) were level 3 predictions and the remaining 2 sequences' host were predicted using level 4 criteria. The distribution of predicted hosts with >5 vOTUs is shown in figure 5.14(a). The most commonly predicted host genus was *Streptococcus* (n=176), followed by *Prevotella* (n=115), *Veillonella* (n=85) and *Bacteroides* (n=83). Host predictions included 254 vOTUs that were High-quality and 1,060 that were Genome-fragment as per the MIUViG quality assessment. As shown in figure 5.14(b), the most High-quality phage vOTUs were from fecal, oral and saliva microbiomes whereas the Genome-fragment quality sequences were from almost all microbiomes analysed. A virus-host interaction network based on the above result is shown in the figure 5.14(c).

It is noteworthy that a large proportion of phage vOTUs (n=1,072) was unclassified i.e. could not be linked to a known phage family and were potentially novel as represented by triangles in figure 5.14(c). These phage vOTUs can be associated with their corresponding hosts suggesting that these could be completely novel phages. The virus-host network also provides further insights into the predicted hosts, for example, it is apparent that phages that infect the genus *Streptococcus* are predominantly found in fecal, saliva and oral microbiomes whereas those phages that are specific to the genus *Bacteroides* are exclusively found in the fecal samples. *Bacteroides* is a genus of bacteria that is dominant and commonly found in the gut microbiota of humans. Bacteriophages that infect the genus *Neisseria* - a genus of bacteria commonly found in mucosal surfaces of many animals including humans were specific to oral and saliva microbiomes. Phages that infect the genus *Cutibacterium* were found in the blood microbiomes. The members of the

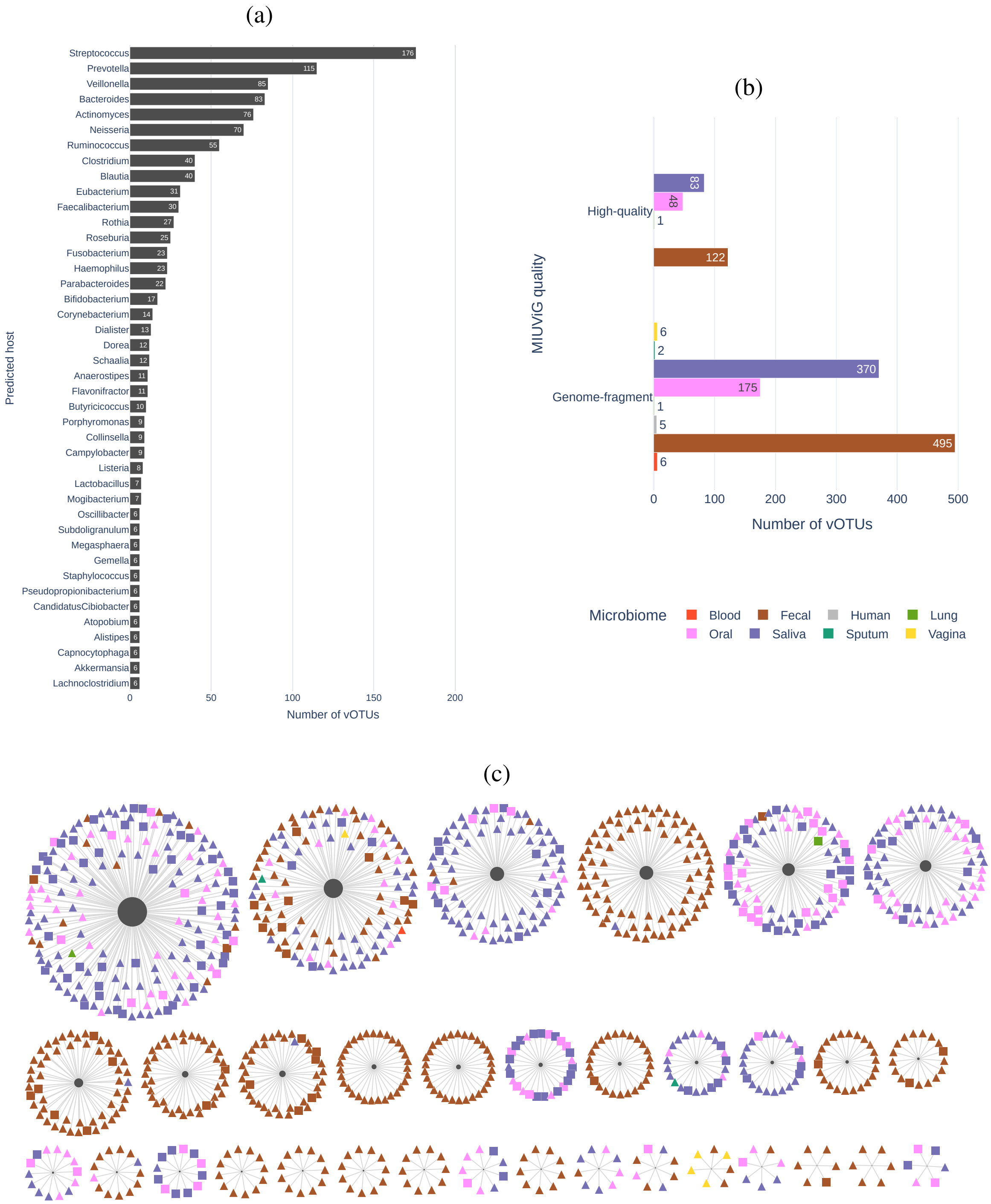


Figure 5.14: Host predictions and analysis for prokaryotic virus OTUs. (a) A bar chart showing the number of vOTUs (X-axis) and their predicted hosts (Y-axis) by the CrisprOpenDB tool. (b) A grouped bar chart showing the phage vOTUs and their completeness measured using the MIUViG quality. Each bar is coloured according to the number of vOTUs identified from different human microbiomes. (c) Predicted hosts of 1,314 bacteriophage vOTUs. Colours represent different microbiomes: The same colour scheme as shown in (b). Shapes represent LCA family rectangles = family known (Inoviridae, Microviridae, Siphoviridae, Myoviridae and Podoviridae) and triangles represent vOTUs that are unclassified. The size of the dark grey circle represents the number of vOTUs where the corresponding bacterial taxonomic rank was predicted as host. In each network, the host node is represented by the dark grey circle in the middle with the circle size relative to the number of phage vOTUs predicted for the host. All hosts are displayed in the same order as figure 5.14(a). The phage nodes are represented by the square or triangle shapes where squares indicate that the LCA of the BLASTN hits for the corresponding phage vOTU was a known prokaryotic virus family e.g. *Inoviridae, Microviridae, Myoviridae, Podoviridae and Siphoviridae* and triangles represent the contrary e.g. LCA taxa were unclassified phages or were higher taxonomic ranks such as an order or realm. The colours of the phage nodes represent the microbiomes that they were assembled from and the colour scheme is the same as that shown in (b).

genus *Cutibacterium* are commensal bacteria of human skin that are also known as common contaminants of blood and body fluid cultures.

Overall, CRISPR signature-based host identification found the interconnections between phages and their known bacterial hosts that are commensal in different microbiomes meaning that they are part of the normal human microbiota. It is appreciated that the prokaryotic virus-host interactions are typically measured at a much finer scale. Host prediction analysis carried out here provided a much higher-level overview of the virus-host relationships with a limited number of sequences available in the CRISPR spacer database utilised in this analysis. Moreover, though virus-host relationships highlighted here are preliminary data-driven, it is worth noting that a large proportion of contigs that are likely to be matching to a range of bacteria were not analysed in detail here limiting the resolution of virus-host interaction analyses. Additional analyses with a focus on the presence/absence of specific virus-host pairs could provide further insights into the microbiome-specific virus-host interactions. The CrisprOpenDB package utilised here has much lower recall which was also indicated in these results as large proportion of phage contigs could not be associated with the corresponding hosts. Furthermore, the current version of CrisprOpenDB package only includes bacteria sequences and completely lacks archaea-specific spacer signatures. It is hypothesised that the large number of phages that could not be associated with a host owing to the limitations of CrisprOpenDB, could potentially be linked to their hosts using other virus-host predictions approaches.

5.4.5 Diversity of eukaryotic DNA viruses

This section of results focuses on viral contigs that match known DNA virus groups (family, order, genera) excluding prokaryotic virus-specific groups. 2,104 DNA viral contigs were identified to be originating from the genomes of eukaryotic DNA viruses. Among this set, 130 contigs were predicted as viruses using all three virus prediction tools (DeepVirFinder, TetraPredX and VirSorter2), 189 were predicted using DeepVirFinder and TetraPredX, 946 were predicted using VirSorter2 and TetraPredX and 19 were predicted between DeepVirFinder and VirSorter2. There were 11 predicted contigs that were unique to DeepVirFinder, 758 were unique to TetraPredX and 51 were uniquely predicted by VirSorter2. These 2,104 DNA viral contigs were originating from 21 distinct BioProjects as shown in the contig distribution across different studies in the bar chart shown in figure 5.15(c).

969 contigs were found in the blood microbiome studies and they originated from a range of BioProjects. As noted in figure 5.15(c), the highest number of DNA viral contigs (n=946) was found in the oral microbiome BioProject PRJNA230363 which contained samples from China. This BioProject samples were obtained from dental plaques, which are likely to contain blood. Thus, anelloviruses, which are usually found in human blood, are associated with the oral microbiome in this study. Other microbiomes such as saliva and fecal were also shown to contain contigs originating from DNA viruses. To characterise the viruses taxonomically, LCA

was obtained for each contig hit and these results are described in two separate heatmaps shown in figure 5.15(a) (family-level classification) and figure 5.15(b) (the LCA could not be classified at a family level). These contigs were classified into 10 distinct DNA virus families, and, were found across 7 different microbiomes (figure 5.15(a)). Notably, the *Anelloviridae* family was found to be dominant in this classification and anellovirus contigs were predominantly found in oral (n=897, 48.72%) and blood (n=940; 51.06%) microbiomes. The second most common viral family found in this set was *Herpesviridae* and contigs matching herpesviruses were only found in the saliva microbiome. 58 contigs were unclassified at the family level and were found in 3 distinct microbiomes and a large majority of them had "Monodnaviria" as their LCA indicating that they were matching a range of different single-stranded DNA viruses from different families (figure 5.15(b)).

Among the 2,044 contigs where the LCA could be classified at a family level, 394 (19.28%) contigs were deemed as nearly complete genomes based on the MIUViG quality assessment and 1,650 (80.72%) were characterised as Genome-fragments (figure 5.15(d)). A majority of complete genomes were classified to be anelloviruses (n=373), whilst a small proportion was classified as genomoviruses, circoviruses, parvoviruses and papillomaviruses. MIUViG high-quality sequences indicating that they were nearly complete genomes were classified as Arfiviricetes, CRESS virus sp., Circular genetic element sp., *Cressdnaviricota*, *Monodnaviria*, Viruses, unclassified viruses, and uncultured human fecal virus.

To explore this dataset further, the contigs matching the most dominant family found in this dataset i.e. *Anelloviridae* were investigated in detail.

Cataloguing the diversity of Anelloviruses

Anelloviruses were the most abundant DNA viruses found in this study. They are small, circular ssDNA viruses with negative-sense genomes and belong to the family *Anelloviridae*. Anellovirus genome sizes range typically between 2-3.9kb. There is a total of 155 viral species classified into 31 genera in the family *Anelloviridae* according to ICTV MSL 36 (ratified in March 2021).

Initially, 1,841 contigs with BLASTN hits exclusive to various anelloviruses were identified. This set was expanded with an addition of 164 contigs that had "Viruses" as LCA but were predominately similar to a range of anelloviruses. These 164 contigs had hits to at least one of three NCBI sequences (JX157237.1, JX157238.1, JX157239.1) which are classified under environmental samples under viruses resulting in the LCA being determined to be Viruses instead of family *Anelloviridae*. All three of these NCBI sequences contained Torque teno virus (TTV)-like ORFs. These additional 164 contigs were combined with the original set of 1,841 contigs resulting in a final set of 2,005 contigs matching anelloviruses. These anellovirus contigs were between 1-8kb long. The mean contig length was 2356 bases and 75% of sequences were under 3kb long (figure 5.16(b)). The contigs that were larger than expected genome sizes are likely to be results of either chimaeric sequences or misassemblies. The *de novo* assembly tools often

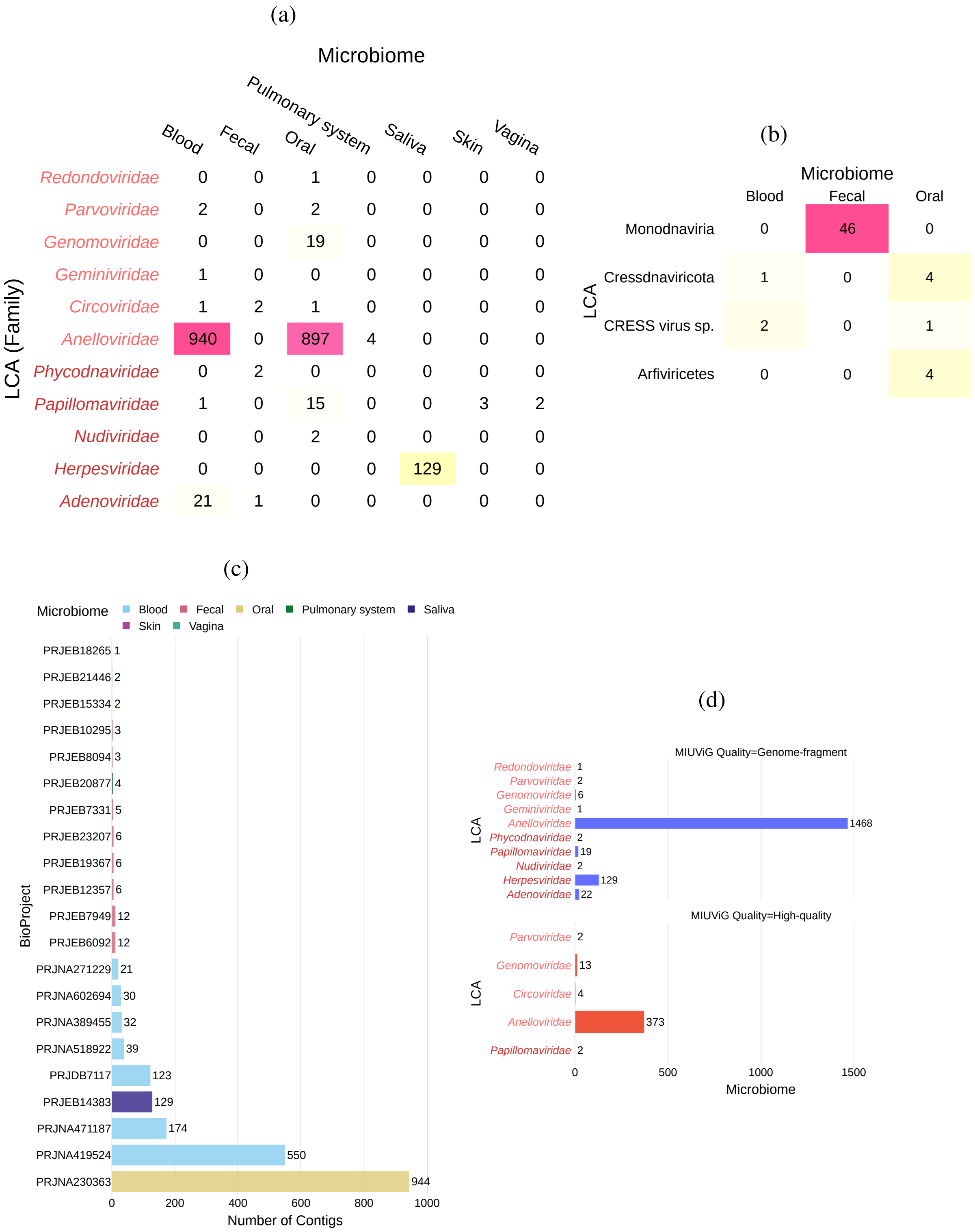


Figure 5.15: An overview of contigs matching to DNA viruses excluding phages. (a) A heatmap showing the number of contigs found in different microbiomes (X-axis) grouped by eukaryotic DNA virus families (Y-axis). (b) A heatmap showing unclassified DNA viruses (Y-axis) contigs found across different microbiomes (X-axis). This category included viruses that could not be classified in a known virus family. (c) Distribution of the number of contigs matching to DNA viruses found in different BioProjects. (d) MIUViG quality assessment of DNA viral contigs that were classified into a known DNA virus family based on the LCA. The plot is split into two categories. Blue bars in the top plot show contigs in MIUViG category Genome-fragments, and, the red bar shows the MIUViG category High-quality.

join two sequences with overlapping regions together and this could cause trouble in assembling circular viruses (Hunt et al., 2015). Out of 2,005 contigs, 19 contigs were >4kb long (i.e. larger than the expected genome size), 709 contigs were <2kb long (smaller than the expected genome size) and the remaining 1,277 contigs were between 2-4kb long which was the expected genome size of this virus family.

To initially assess their similarity to existing anellovirus sequences in the database, the top hit with the highest sequence similarity was extracted. The sequence similarity was plotted against the query coverage as shown in figure 5.16(a). As it can be seen in the figure 5.16(a), a large majority of contigs were found to fall within <80% query coverage range suggesting that these sequences are relatively different to those that were currently catalogued. A total of 1,006 anelloviruses were found in oral microbiome, 995 were found in the blood microbiome and 4 were found in the pulmonary system microbiome. These microbiomes spanned 8 distinct BioProjects (figure 5.16(c)) and 162 distinct samples originating from these studies. The largest number of anellovirus contigs (n=1,006) were found in the oral microbiome samples from China included in the BioProject PRJNA230363. The original study that sequenced these oral samples was led by J. Wang et al. (2016) and aimed to analyse the phage-host interaction network in the human oral microbiome. The highest number of anellovirus contigs from the blood microbiome were from BioProject PRJNA419524 (n=573). This study included blood samples of patients with organ transplants and was carried out in three different sites (Pennsylvania, Wisconsin and New York) in the USA. The CheckV analysis indicated that 395 (19.7%) of all anellovirus contigs were High-quality (complete or nearly complete genomes) and 1,610 (80.3%) were partial genome sequences (figure 5.16(c)).

Typically, the ORF1 coding sequence is used to investigate the phylogenetic relationships in anelloviruses and it is the largest ORF shared among all anelloviruses and hypothesised to code for virus replication-associated and capsid proteins. Thus, the ORF1 amino acid sequence-based phylogenetic tree built with IQTree model LG+F+I+G4 selected via the model selection using IQTree, and 1000 bootstrap is shown in figure 5.17(a). Four distinct clades representing all four anellovirus genera are well separated with strong bootstrap support. The anellovirus contigs were assigned a genus based on their respective clade membership. A breakdown of distinct genera found in different BioProjects is shown in figure 5.17(b).

From the set of 1,275 anellovirus contigs, 693 were assigned to genus *Alphatorquevirus*, 196 were assigned to genus *Betatorquevirus*, 354 were assigned to genus *Gammatorquevirus* and 1 was assigned to genus *Hetorquevirus*. 17 contigs did not fall within any of these clades and are shown in grey in the figure 5.17(b). These are also shown in a separate cluster in the figure 5.17(a). The members of genus *Alphatorquevirus* are referred to as Torque teno viruses (TTVs), those in genus *Betatorquevirus* are referred to as Torque teno mini viruses (TTMVs) and those in genus *Gammatorquevirus* are referred to as Torque teno midi viruses (TTVMDs). The largest number of TTVs were found in the blood microbiomes of BioProject

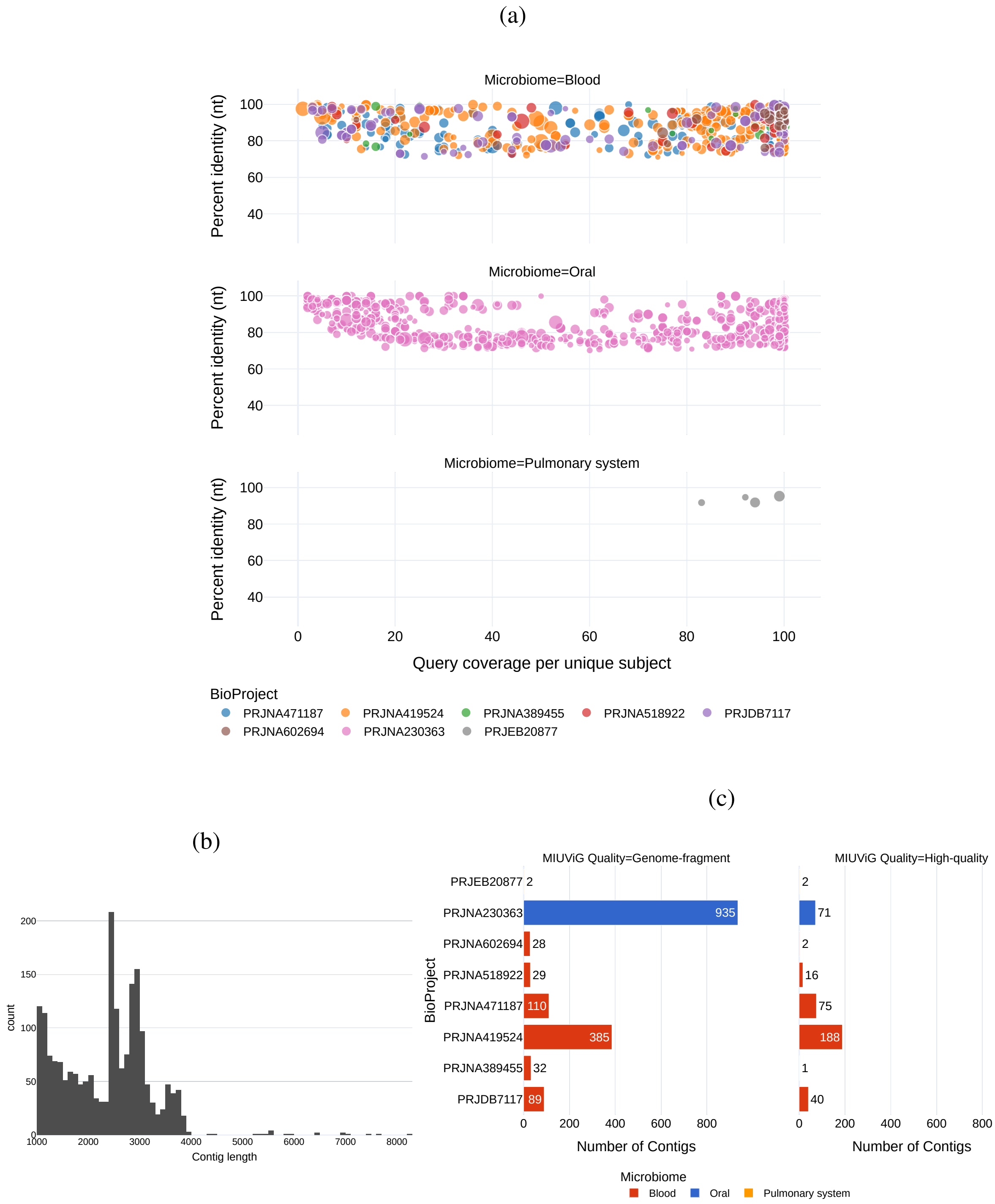


Figure 5.16: Overview of the contigs matching to anelloviruses. (a) A bubble chart showing the sequence identity and query coverage with respect to the top BLASTN hit for each anellovirus contigs. The size of the bubble represents the contig size and the colours represent the BioProjects for each microbiome category (subplots). (b) The distribution of contig lengths of 2,005 anellovirus contigs was identified in this study. (c) MIUViG quality assessment of anellovirus contigs grouped by study/BioProject. This figure is split into two MIUViG categories. The bar plot on the left shows the contigs that were categorised as genome-fragments and the right plot shows the contigs that were categorised as high-quality. Each BioProject is shown on the Y-axis and the number of anellovirus contigs is shown on the X-axis. The colour of the bar is indicative of the microbiome type described in the legend.

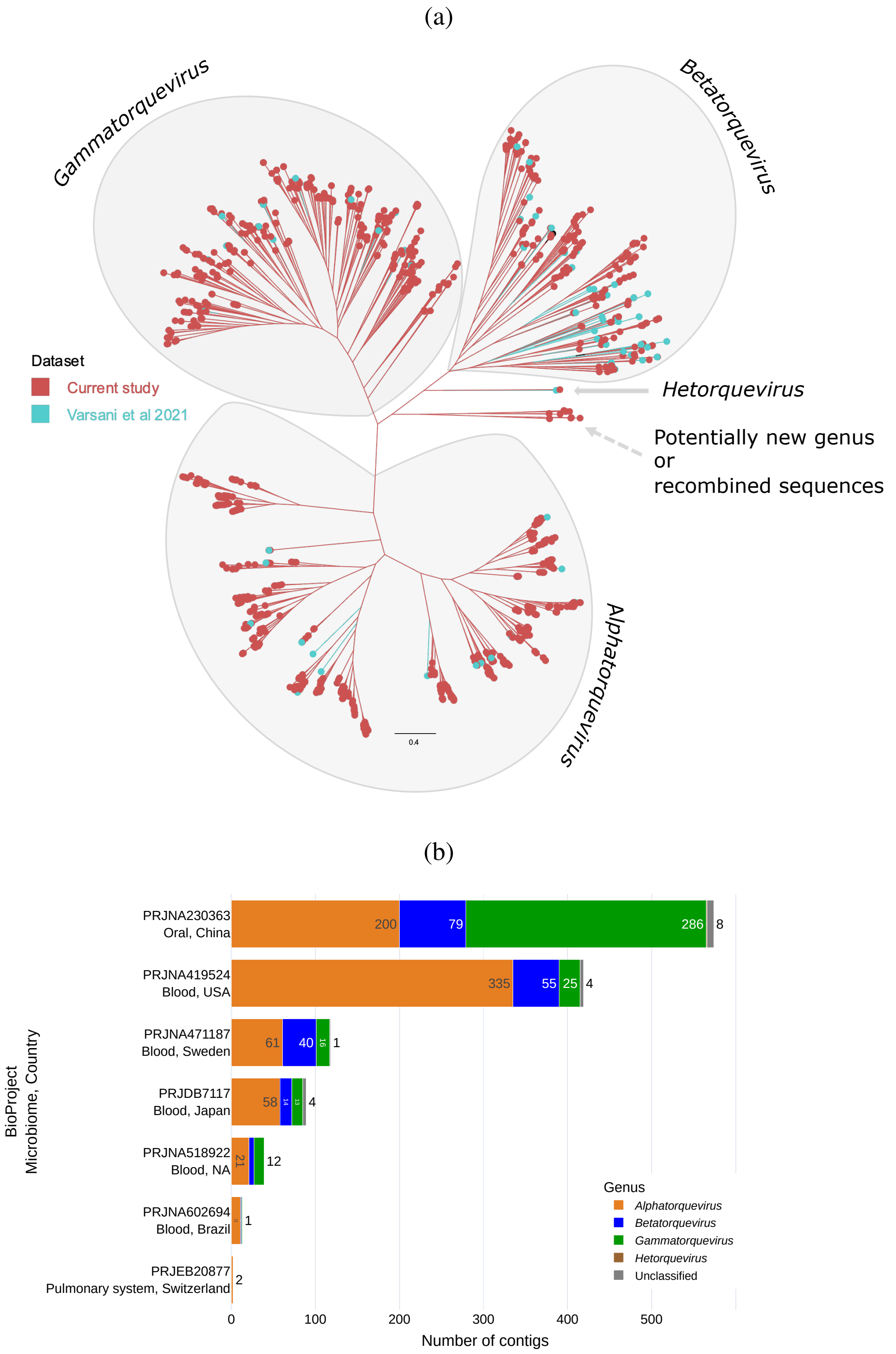


Figure 5.17: Phylogenetic analysis and classification of anellovirus contigs. (a) A maximum-likelihood tree is inferred from ORF1 amino acid sequences of anelloviruses. Circles represent nodes which are coloured according to the study; reference sequences retrieved from Varsani et al. (2021) are shown in cyan and anellovirus contigs identified in this study are shown in burnt sienna. (b) Human anellovirus genera assignment is inferred from the phylogenetic tree and its distribution across different BioProjects and microbiomes. Overall, all genera were found in most BioProjects. Anelloviruses from the genus *Gammatorquevirus* were predominantly found in oral microbiome samples from China.

PRJNA419524 whereas the highest number of TTVMDs were found in oral microbiomes of BioProject PRJNA230363. A large diversity of TTMDV present in the oral samples from China suggest either a microbiome or geolocation-specific link. However, this hypothesis was not formally tested here to explore this further. The unclassified sequences were found in 4 different BioProjects including PRJDB7117 (n=4), PRJNA230363 (n=8), PRJNA419524 (n=4) and PRJNA471187 (n=1). A number of sequences were shown to form a separate cluster from known human anellovirus genera. These sequences were initially thought to belong to the genus *Omegatorquevirus*. However, an updated *Anelloviridae* tree with the entire *Anelloviridae* showing all genera (see appendix figure C.2) phylogeny built with sequences from Varsani et al. (2021) showed that these unclassified sequences form a distinct clade from the genus *Omegatorquevirus*. This indicates that either these sequences may form a potential new human anellovirus genus or that they are likely to be recombinant sequences meaning that they may be composed of sequences that may have originated from two or more genera. Anellovirus sequence diversity has been evidently driven by recombination (Worobey, 2000; Arze et al., 2021). These recombination events have been observed within different clades and/or genera of anelloviruses and are not limited to closely related anellovirus species (Arze et al., 2021). The sequences that fall within this separate new clade were isolated from blood and oral microbiomes. These BioProjects were not specific to a geographic location and they originated from various parts of the world. For example, the blood microbiome BioProject PRJDB7117 was from Japan, PRJNA471187 was from Sweden, PRJNA419524 was from the USA and the oral microbiome BioProject PRJNA230363 contained sequences from China.

Novel anellovirus species and genera are classified based on the ORF1 coding nucleotide sequences. Based on the recently updated species classification criteria defined by Varsani et al. (2021), any sequences that bear <69% sequence similarity to currently known anellovirus species are identified as novel species. Varsani et al. (2021) also recommend using SDT to determine the species demarcation based on pairwise sequence identity. The similarity matrix was extracted from SDT and was converted to distances. Clusters were computed from this set using the Agglomerative Clustering (Hierarchical clustering) implemented in the Scikit-learn package with distance_threshold = 0.31, linkage = complete, affinity = precomputed, compute_full_tree = True and compute_distances = False parameters for a total of 1,419 sequences (1,275 anellovirus contigs plus 144 reference sequences from Varsani 2021 set).

In total 371 clusters were generated using the above demarcation clustering method. 143 clusters contained at least one known anellovirus species, these clusters comprised 443 sequences including 143 known anelloviruses species from the Varsani et al. (2021) set as well as 299 anellovirus contigs from the SM set; the remaining 228 clusters contained anellovirus contigs that were assembled from the SM set. 299 human anellovirus contigs were found to cluster within 46 clusters representing existing human anellovirus species. These species' demarcation results align well with the phylogenetic analysis described above. As shown by the phylogenetic clade membership in figure 5.17(a) all 46 species that clustered with anellovirus contigs were members of human anellovirus genera *Alphatorquevirus*, *Betatorquevirus* and *Gammatorquevirus*. One cluster with *Gammatorquevirus* species Torque teno midi virus 15 that was isolated from Pan troglodytes was shown to include two human contigs sequences SRR7166951_NODE_76 and SRR6316221_NODE_11 assembled in this analysis. The remaining 976 sequences were grouped into 228 clusters based on the species demarcation criteria and were uniquely found in the dataset assembled and catalogued from this analysis. Based on these clustering results, it is hypothesised that 228 potentially novel anellovirus species in humans have been discovered (table C.3).

A complete list of these novel species of human anelloviruses discovered from 3 distinct microbiomes (1 from the Pulmonary system, 72 from the Blood and 155 from the Oral) is shown in the table C.3. These new species originated from 50 different SRA samples included in 8 distinct BioProjects. 155 novel species were identified from the oral microbiome from China (PRJNA230363). 72 novel species were identified from a range of different blood microbiome BioProjects. One novel species was found in the pulmonary system BioProject PRJEB20877 from Switzerland. Novel species were identified in all human anellovirus genera. To reiterate, the genera were assigned based on the phylogenetic clade membership. 46 new species were from the genus *Alphatorquevirus*. 70 novel species were from the genus *Betatorquevirus* and 106 novel species were from genus *Gammatorquevirus*. One novel species of the genus *Hetorquevirus* was also identified. Moreover, 5 new species were identified that could not be associated with a known human anellvirus genus.

5.4.6 RNA viruses

Out of the four major categories of viruses described in the figure 5.1, RNA virus categories contained the smallest number of contigs. This is due to the fewer metranscriptomic BioProject included in the previous analyses carried out in Chapters 3 and 4. In total, 269 contigs matched a range of RNA viruses. Among this set 54 contigs were deemed High-quality and 215 were partial sequences. However, it is worth noting that MIUViG criteria do not always provide the best measure of the completeness of RNA virus genomes due to RNA virus-specific features such as segmentation. The distribution of these RNA viral contigs across different BioProjects and microbiomes is shown in figure 5.18. 125 of these contigs' LCA could not be determined at a family level and the remaining contigs were categorised into RNA virus families as follows: *Arenaviridae* (n=6), *Flaviviridae*(n=79), *Hepadnaviridae* (n=7), *Paramixoviridae* (n=2), *Picobirnaviridae* (n=3), *Retroviridae* (n=32), *Rhabdoviridae* (n=5), *Tombusviridae* (n=2) and *Virgaviridae* (n=1). Genus *Pegivirus* which comprises the family *Flaviviridae* contained the most number of hits (n=69) and it was found in 4 distinct BioProjects. Pegivirus name is derived from **pe**rsistant, and **g** in historical reference to GB virus and hepatitis G virus names. Pegiviruses have a broad host range and can infect humans, non-human primates, pigs, horses and a range of rodent and bat species (Simmonds et al., 2017b).

The contigs that could not be classified at a family level were predominantly classified at higher taxonomy levels including the realm *Riboviria* (n=52), phylum *Lenarviricota* (n=6) and Viruses (n=55). Additionally, these contigs were also matching to a range of currently unclassified RNA viruses such as unclassified RNA viruses ShiM-2016 (n=6), Hubei narna-like virus 22 (n=4), le Maire virus (n=1) and Beihai picorna-like virus 64 (n=1).

A number of RNA viral contigs were found in the human blood microbiome. This is chiefly because BioProject PRJNA271229, which was comprised of human blood samples of patients with an unknown fever collected in Nigeria in 2011, contained the most RNA viral contigs (n=177) found in this study. The majority of RNA viral contigs identified in this BioProject could not be linked back to a known RNA virus family and their LCA was assigned to the realm *Riboviria* (n=47) or Viruses (n=47). The metatranscriptomic study that sequenced total RNA from sputum samples (BioProject: PRJEB10919) of patients with active tuberculosis contained 37 contigs matching RNA viruses. This BioProject contained the most diverse range of unclassified RNA viruses including Chicken picobirnavirus (n=1), Coxsackievirus A21 (n=1), Enterovirus (n=2), Hubei narna-like virus 22 (n=3), Human immunodeficiency virus 1 (n=9), Human respirovirus 3 (n=1), *Lenarviricota* (n=6), Lentivirus (n=1), Otarine picobirnavirus (n=1), Picobirnavirus (n=1), Rhinovirus A (n=1), Rhinovirus B (n=2), *Riboviria* (n=3), Tobamovirus (n=1), Viruses (n=2), le Maire virus (n=1) and unclassified RNA viruses ShiM-2016 (n=1).

A general-purpose clustering was performed using the anicalc and aniclust script available within the CheckV package to identify sequence similarity among contigs matching RNA viruses. A minimum sequence identity threshold of 95% and a target coverage of 85% were applied. From the set of 269 contigs, 141 clusters were obtained. 103 cluster representatives had a nucleotide hit to nt databases (BLASTN) whereas 38 cluster representatives did not have nucleotide hits and only had protein hits against viral proteins. Two of the largest clusters were generated from two separate contigs of the same length of 7,940 nucleotides representing a potentially novel picorna-like virus. Both contigs were assembled from blood microbiome sample SRR1748182. These results are discussed in detail in the section 5.4.6 below.

Unclassified RNA viruses

Due to their short genomes, error-prone replication mechanism and high sequence diversity, RNA viruses are often trickier to identify in metagenomic samples and require special wet-lab and dry-lab protocols to successfully capture their signature sequences from metagenomic/metatranscriptomic studies (Holmes, 2009; Greninger, 2018; M. Shi et al., 2018b). Moreover, novel RNA viruses often do not bear any nucleotide-level sequence similarity to known viruses at the nucleotide level. This is a distinct feature of completely novel RNA viruses that leads to no hits when searched against comprehensive nucleotide databases such as nt, however, they may bear some similarity to their distant relatives at the protein level. RNA viral contig representatives (n=38) that did not bear any nucleotide level similarity to known

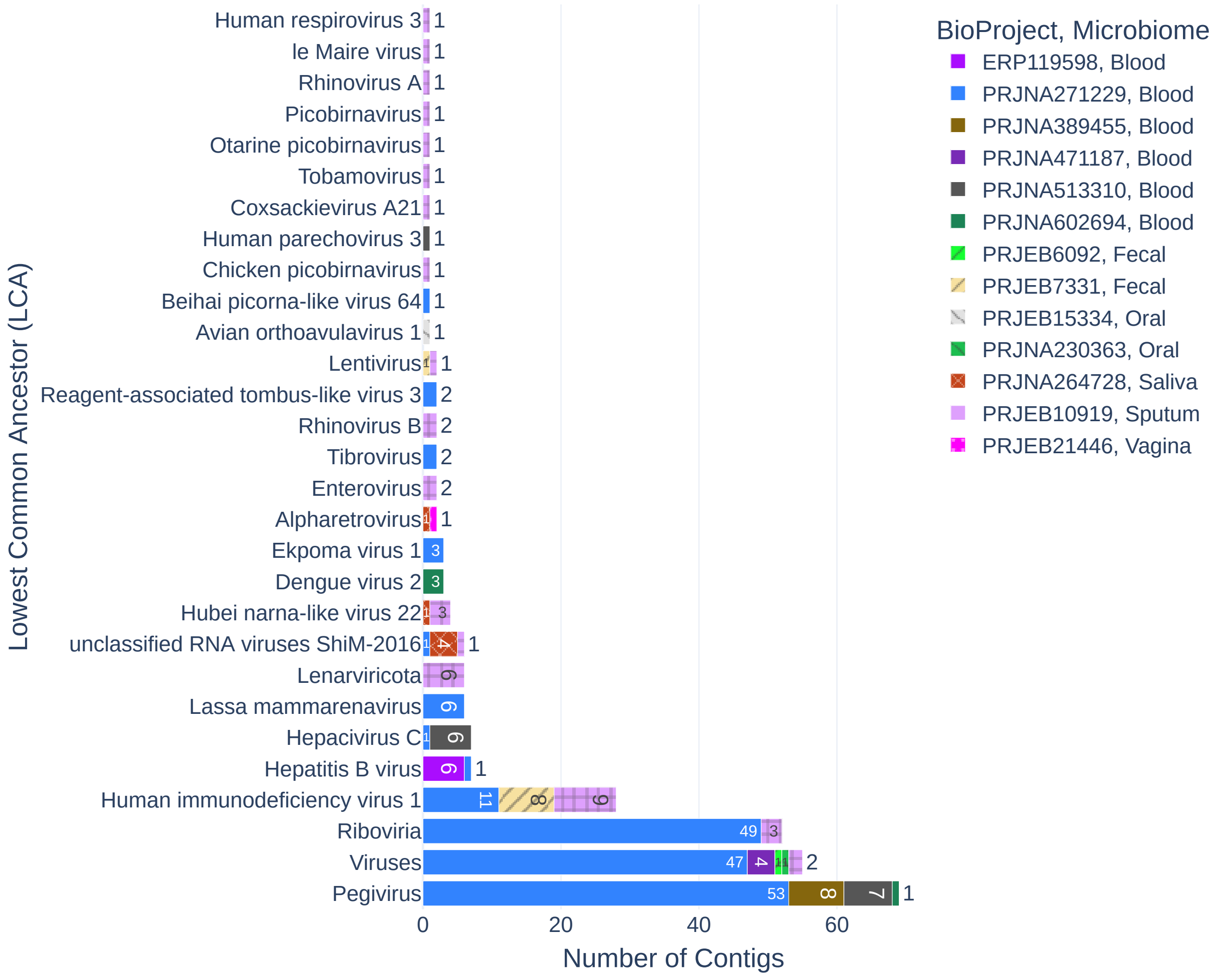


Figure 5.18: Overview of contigs matching to RNA viruses. The lowest common ancestor (LCA) of BLAST/DIAMOND hits is shown on the Y-axis with the number of contigs identified in that category shown on the X-axis. The bars are coloured according to the BioProject the contigs are found in, and the patterns represent the corresponding microbiomes.

sequences in the databases but displayed some protein similarity to a range of unclassified RNA virus protein sequences were explored further. The top hit for these contigs was extracted to obtain query coverage and percent identity and these two measures were plotted against each other as shown in figure 5.19. All these sequences have very low sequence similarity to their closest match even at the protein sequence level compared to their BLASTN counterpart shown in the appendix figure C.3. Moreover, a large number of these contigs have low query coverage of <80% (X-axis). These low sequence similarity and low sequence coverage suggest that these contigs represent completely novel RNA virus sequences that are not yet catalogued in the databases.

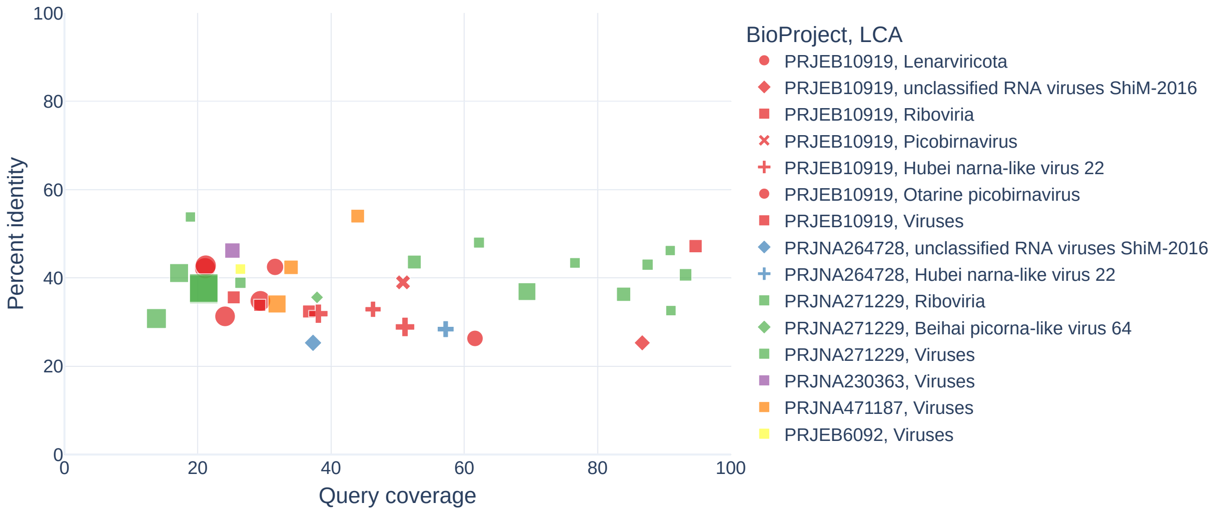


Figure 5.19: A scatter plot showing query coverage (X-axis) and corresponding percent identity (Y-axis) to its protein level best hit for each contig. The shapes of the markers represent distinct studies, the colours of the markers represent the LCA taxa and the size of the marker is relative to the contig length.

These potentially novel unclassified viruses were identified in 6 distinct studies; PRJNA271229 (n=16; blood), PRJEB10919 (n=15; sputum), PRJNA471187 (n=3; blood), PRJNA264728 (n=2; saliva), PRJNA230363 (n=1; oral) and PRJEB609 (n=1; fecal). These cluster representative sequences (or vOTUs) had hits to viral proteins originating from the following LCAs: *Riboviria* (n=13), Viruses (n=11), *Lenarviricota* (n=5), Hubei narna-like virus 22 (n=4), unclassified RNA viruses ShiM-2016 (n=2), Picobirnavirus (n=1), Otarine picobirnavirus (n=1), and Beihai picorna-like virus 64 (n=1). These results are captured in detail in table C.4. To catalogue completely novel RNA viruses that were classified at the higher taxonomy levels, an in-depth analysis was carried out on specific contigs of interest.

Novel picorna-like virus

17 contigs of interest were found in the blood microbiome study PRJNA271229 that were at least 5kb long. This study sequenced human serum samples derived from healthy individuals

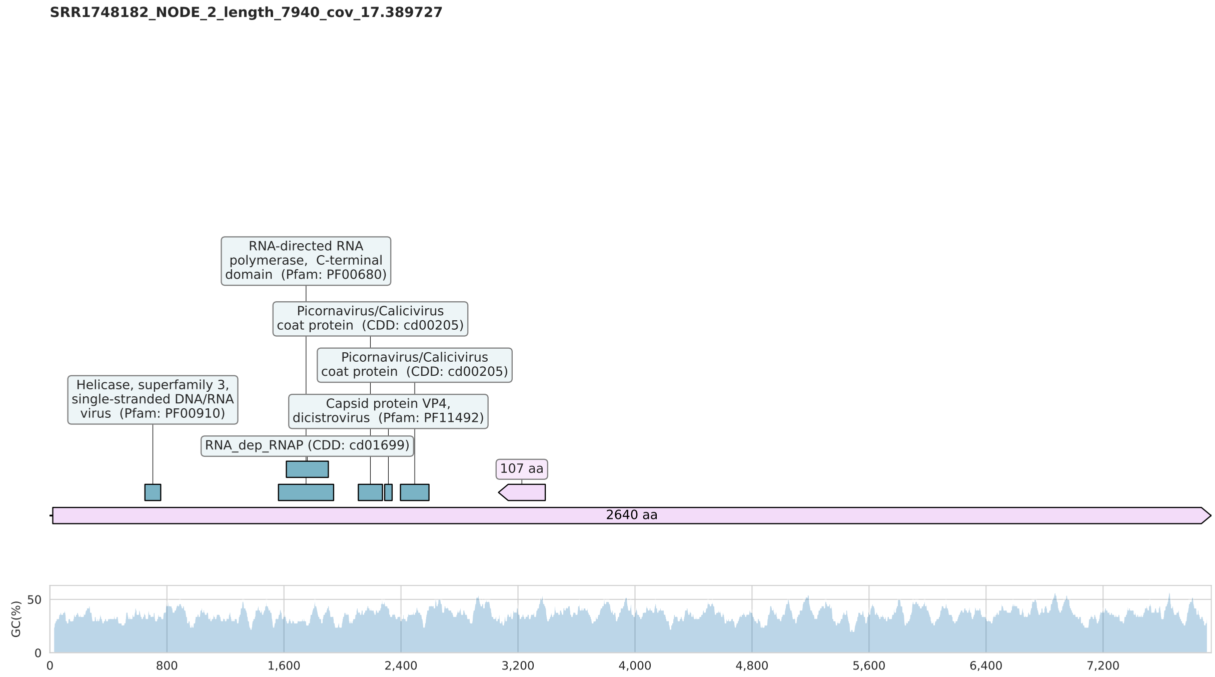
and patients with fevers of unknown origin from Nigeria in 2011 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA271229). All contigs >=5kb were extracted and their corresponding LCA identification showed that all of them were matching to a range of unclassified picorna-like viruses with very low sequence similarity at the protein level and no significant matches were found at the nucleotides sequence level. These contigs were 5,888-7,940 nucleotide long. The figure 5.20 shows the sequence similarity of the largest contig (SRR1748182_NODE_2) against the NCBI RefSeq protein databases on BLASTX run on the web on 10 April 2022. Notably, this sequence matched an unclassified picorna-like virus called Bat dicibavirus with 27% similarity at the protein level. These results are shown in figures 5.20(b) and 5.20(c). Moreover, these contigs were also shown to contain a large polyprotein ORF which contained RdRp and VP4 domains identified by BLAST domain signature analysis (figure 5.20(a)) .

The presence of the polyprotein ORF, RNA virus-specific domains and protein sequence divergence indicated that these contigs were likely to be originated from a novel picornavirus or picorna-like virus. Picornaviruses have a monopartite or bipartite positive-strand RNA genome that ranges from 7-12kb nucleotides in length. Monopartite genomes code for a single ORF that encodes a single large polyprotein (King et al., 2012). A range of picorna-like viruses has been identified predominantly from marine environment (Culley et al., 2003; Lang et al., 2009), however, since the widespread application of metagenomics, picorna-like viruses are identified in faeces and organs of a variety of organisms (M. Shi et al., 2016a; M. Shi et al., 2018a; Zell et al., 2022). This indicates both a wide distribution of picorna-like viruses among organismic kingdoms as well as an accumulation of unspecific viruses without subsequent infection (Zell et al., 2022). To investigate this hypothesis further, all protein sequences that these 17 contigs were matching to were extracted from the DIAMOND results. In total, 37 unique protein sequences matching these contigs were identified that belong to a range of picorna-like viruses discovered in freshwater arthropods and molluscs as well as other hosts including octopus and algae. In total 54 sequences were used to perform the analysis and this dataset is referred to as PL01. A vertebrate picornavirus, rabbit hemorrhagic diseases virus was used as an outgroup for the phylogeny shown in figure 5.21 which was generated using the steps described in the Method section 5.3.7.

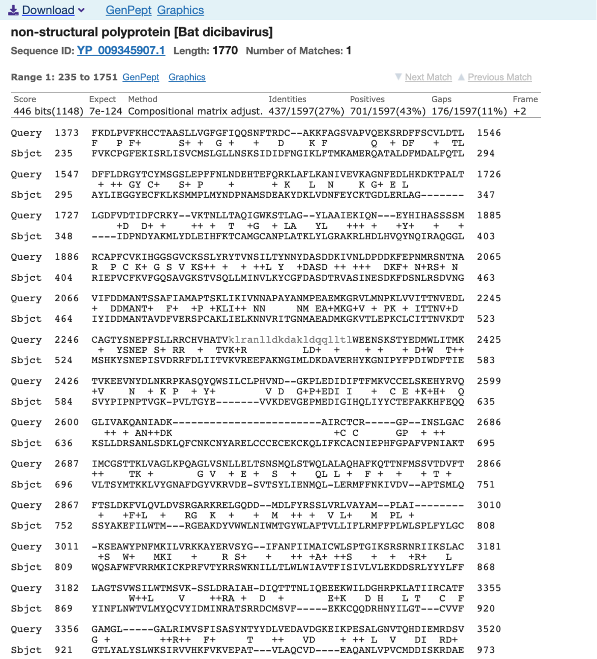
The novel picorna-like virus sequences showed very high sequence similarity (>94-100% at the nucleotide level) among each other suggesting that they may be originated from the same virus (figure 5.21). However, they form a cluster separate from their closest relative Shahe picorna-like virus 1 which was extracted from freshwater arthropod by Shi et al (M. Shi et al., 2016a). Although most of the viruses included in the PL01 phylogeny shown here are unclassified sequences, 7 sequences from this set were classified to the family *Marnaviridae*. The family *Marnaviridae* represents sequences that typically infect marine protists, and algae and also have unclassified members identified using metagenomic methods that were found in marine and freshwater environments (Lang et al., 2021).

To study this novel picorna-like virus further, all classified sequences of order *Picornavirales*

(a)


SRR1748182_NODE_2_length_7940_cov_17.389727
RNA-directed RNA polymerase, C-terminal domain (Pfam: PF00680)
Picornavirus/Calicivirus coat protein (CDD: cd00205)
Picornavirus/Calicivirus coat protein (CDD: cd00205)
Helicase, superfamily 3, single-stranded DNA/RNA virus (Pfam: PF00910)
Capsid protein VP4, dicistrovirus (Pfam: PF11492)
RNA_dep_RNAP (CDD: cd01699)
107 aa
2640 aa
GC(%)
50
0
0
800
1,600
2,400
3,200
4,000
4,800
5,600
6,400
7,200

(b)



non-structural polyprotein [Bat dicibavirus]

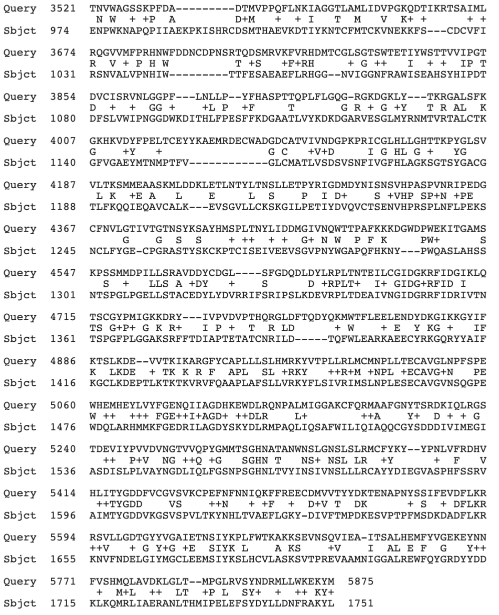
Sequence ID: YP_009345907.1 Length: 1770 Number of Matches: 1

Range 1: 235 to 1751 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
446 bits(1148)	7e-124	Compositional matrix adjust.	437/1597(27%)	701/1597(43%)	176/1597(11%)	+2

```
Query  1373  FKDLPVFKHCCTAASLLVGFGFIQQSNFTRDC--AKKFAGSVAPVQEKSRDFFSCVLDTL  1546
             F   P F+      S+ +  G +   +   D    K F         Q  + DF   +  TL
Sbjct  235   FVKCPGFEKISRLISVCMSLGLLNSKSIDIDFNGIKLFTMKAMERQATALDFMDALFQTL  294

Query  1547  DFFLDRGYTCYMSGSLEPFFNLNDEHTEFQRKLAFLKANIVEVKAGNFEDLHKDKTPALT  1726
              + ++ GY C+   S+ P     +      + K   L  N    K G+ E L
Sbjct  295   AYLIEGGYECFKLKSMMPLMYNDPNAMSDEAKYDKLVDNFEYCKTGDLERLAG-------  347

Query  1727  LGDFVDTIDFCRKY--VKTNLLTAQIGWKSTLAG--YLAAIEKIQN---EYHIHASSSSM  1885
                 +D  D+ +    ++ +  T  +G  + LA   YL   +++ +   +Y+    +   +
Sbjct  348   ----IDPNDYAKMLYDLEIHFKTCAMGCANPLATKLYLGRAKRLHDLHVQYNQIRAQGGL  403

Query  1886  RCAPFCVKIHGGSGVCKSSLYRYTVNSILTYNNYDASDDKIVNLDPDDKFEPNMRSNTNA  2065
             R  P C K+ G S V KS++ +  + ++L Y  +DASD ++ +++  DKF+ N+RS+ N
Sbjct  404   RIEPVCFKVFGQSAVGKSTVSQLLMINVLKYCGFDASDTRVASINESDKFDSNLRSDVNG  463

Query  2066  VIFDDMANTSSAFIAMAPTSKLIKIVNNAPAYANMPEAEMKGRVLMNPKLVVITTNVEDL  2245
             +  DDMANT+  F+  +P +KLI++ NN     NM EA+MKG+V + PK + ITTNV+D
Sbjct  464   IYIDDMANTAVDFVERSPCAKLIELKNNVRITGNMAEADMKGKVTLEPKCLCITTNVKDT  523

Query  2246  CAGTYSNEPFSLLRRCHVHATVklranlldkdakldqqlltlWEENSKSTYEDMWLITMK  2425
              +  YSNEP S+ RR  +  TVK+R        LD+  +    + +   + D+W  T++
Sbjct  524   MSHKYSNEPISVDRRFDLIITVKVREEFAKNGIMLDKDAVERHYKGNIPYFPDIWDFTIE  583

Query  2426  TVKEEVNYDLNKRPKASQYQWSILCLPHVND--GKPLEDIDIFTFMKVCCELSKEHYRVQ  2599
             +V   N  + K P  + Y+       V D  G+P+EDI I   +  C E +K+H+  Q
Sbjct  584   SVYPIPNPTVGK-PVLTGYE-------VVKDEVGEPMEDIGIHQLIYYCTEFAKKHFEQQ  635

Query  2600  GLIVAKQANIADK------------------------AIRCTCR-----GP--INSLGAC  2686
               ++ + AN++DK                          +C C       GP  + ++
Sbjct  636   KSLLDRSANLSDKLQFCNKCNYARELCCCECEKCKQLIFKCACNIEPHFGPAFVPNIAKT  695

Query  2687  IMCGSTTKLVAGLKPQAGLVSNLLELTSNSMQLSTWQLALAQHAFKQTTNFMSSVTDVFT  2866
             ++    TK +      G V  + E +  S  +  QL L +  F +  +   + T +
Sbjct  696   VLTSYMTKKLVYGNAFDGYVKRVDE-SVTSYLIENMQL-LERMFFNKIVDV--APTSMLQ  751

Query  2867  FTSLDKFVLQVLDVSRGARKRELGQDD--MDLFYRSSLVRLVAYAM--PLAI--------  3010
              +   +F+L  +   RG  K  +   +  M  ++ +  V L+    M  PL +
Sbjct  752   SSYAKEFILWTM---RGEAKDYVWWLNIWMTGYWLAFTVLLIFLRMFFPLWLSPLFYLGC  808

Query  3011  -KSEAWYPNFMKILVRKKAYERVSYG--IFANFIIMAICWLSPTGIKSRSRNRIIKSLAC  3181
              +S  W+   MKI   +    R S+   +  ++ +A+ ++S   +  +  +R+   L
Sbjct  809   WQSAFWFVRRMKICKPRFVTYRRSWKNILLTLWLWIAVTFISIVLVLEKDDSRLYYYLFF  868

Query  3182  LAGTSVWSILWTMSVK-SSLDRAIAH-DIQTTTNLIQEEEKWILDGHRPKLATIIRCATF  3355
                   W++L   V   ++RA + D  +      E+K  D H  L T  C  F
Sbjct  869   YINFLNWTVLMYQCVYIDMINRATSRRDCMSVF-----EKKCQQDRHNYILGT---CVVF  920

Query  3356  GAMGL-----GALRIMVSFISASYNTYYDLVEDAVDGKEIKPESALGNVTQHDIEMRDSV  3520
             G +       ++R++  F+     T   ++   VD    + ++ L  V   DI  RD+
Sbjct  921   GTLYALYSLWKSIRVVHKFVKVEPAT---VLAQCVD----EAQANLVPVCMDDISKRDAE  973
```

(c)



```
Query  3521  TNVWAGSSKPFDA---------DTMVPPQFLNKIAGGTLAMLIDVPGKQDTIKRTSAIML  3673
              N W  + +P  A          D+M   +  + I  T  M   V  K+ +       + +
Sbjct  974   ENPWKNAPQPIIAEKPKISHRCDSMTHAEVKDTIYKNTCFMTCKVNEKKFS---CDCVFI  1030

Query  3674  RQGVVMFPRHNWFDDNCDPNSRTQDSMRVKFVRHDMTCGLSGTSWTETIYWSTTVVIPGT  3853
             R  V + P H W          T +S   +F+RH      + G ++   I  + +  IP T
Sbjct  1031  RSNVALVPNHIW---------TTFESAEAEFLRHGG--NVIGGNFRAWISEAHSYHIPDT  1079

Query  3854  DVCISRVNLGGPF---LNLLP--YFHASPTTQPLFLGQG-RGKDGKLY---TKRGALSFK
             D  +  +  GG +    +L P  +F     T         G R + G +Y   T R AL  K
Sbjct  1080  DFSLVWIPNGGDWKDITHLFPESFFKDGAATLVYKDKDGARVESGLMYRNMTVRTALCTK  1139

Query  4007  GKHKVDYFPELTCEYYKAEMRDECWADGDCATVIVNDGPKPRICGLHLLGHTTKPYGLSV  4186
             G    +Y   +                  G C   +V+D       I G HL G +   YG
Sbjct  1140  GFVGAEYMTNMPTFV------------GLCMATLVSDSVSNFIVGFHLAGKSGTSYGACG  1187

Query  4187  VLTKSMMEAASKMLDDKLETLNTYLTNSLLETPYRIGDMDYNISNSVHPASPVNRIPEDG  4366
              L K  +E A   L    E     L  S    P  I D+     S +VHP SP+N +PE
Sbjct  1188  TLFKQQIEQAVCALK---EVSGVLLCKSKGILPETIYDVQVCTSENVHPRSPLNFLPEKS  1244

Query  4367  CFNVLGTIVTGTNSYKSAYHMSPLTNYLIDDMGIVNQWTTPAFKKKDGWDPWEKITGAMS  4546
                   G   G  S  S     + ++  + +  G+ N W  P F K      PW+      S
Sbjct  1245  NCLFYGE-CPGRASTYSKCKPTCISEIVEEVSGVPNYWGAPQFHKNY---PWQASLAHSS  1300

Query  4547  KPSSMMDPILLSRAVDDYCDGL----SFGDQDLDYLRPLTNTEILCGIDGKRFIDGIKLQ  4714
               S  +  LLS A +DY   +    S      D +RPLT+  I+ GIDG+RFID I
Sbjct  1301  NTSPGLPGELLSTACEDYLYDVRRIFSRIPSLKDEVRPLTDEAIVNGIDGRRFIDRIVTN  1360

Query  4715  TSCGYPMIGKKDRY---IVPVDVPTHQRGLDFTQDYQKMWTFLEELENDYDKGIKKGYIF  4885
             TS G+P+ G K R+   I P +  T R LD       + W   + E  Y KG +    IF
Sbjct  1361  TSPGFPLGGAKSRFFTDIAPTETATCNRILD-----TQFWLEARKAEECYRKGQRYYAIF  1415

Query  4886  KTSLKDE--VVTKIKARGFYCAPLLLSLHMRKYVTPLLRLMCMNPLLTECAVGLNPFSPE  5059
             K  LKDE  + TK K R F  APL  SL +RKY   ++R+M +NPL +ECAVG+N   PE
Sbjct  1416  KGCLKDEPTLKTKTKVRVFQAAPLAFSLLVRKYFLSIVRIMSLNPLESECAVGVNSQGPE  1475

Query  5060  WHEMHEYLVYFGENQIIAGDHKEWDLRQNPALMIGGAKCFQRMAAFGNYTSRDKIQLRGS  5239
             W ++  +++ FGE++I+AGD+ ++DLR    L+          ++A   Y+  D + + G
Sbjct  1476  WDQLARHMMKFGEDRILAGDYSKYDLRMPAQLIQSAFWILIQIAQQCGYSDDDIVIMEGI  1535

Query  5240  TDEVIYPVVDVNGTVVQPYGMMTSGHNATANWNSLGNSLSLRMCFYKY--YPNLVFRDHV  5413
               ++  P+V  NG ++Q +G   SGHN T   NS+ NSL LR  +Y       +  F  V
Sbjct  1536  ASDISLPLVAYNGDLIQLFGSNPSGHNLTVYINSIVNSLLLRCAYYDIEGVASPHFSSRV  1595

Query  5414  HLITYGDDFVCGVSVKCPEFNFNNIQKFFREECDMVVTYYDKTENAPNYSSIFEVDFLKR  5593
              ++TYGDD    VS    ++N   + +F  +  D+V T  DK        + S  + DFLKR
Sbjct  1596  AIMTYGDDVKGSVSPVLTKYNHLTVAEFLGKY-DIVFTMPDKESVPTPFMSDKDADFLKR  1654

Query  5594  RSVLLGDTGYYVGAIETNSIYKPLFWTKAKKSEVNSQVIEA-ITSALHEMFYVGEKEYNN  5770
             ++V   + G Y+G +E  SIYK L    A KS    +V    I  AL E F  G    Y++
Sbjct  1655  KNVFNDELGIYMGCLEEMSIYKSLHCVLASKSVTPREVAAMNIGGALREWFQYGRDYYDD  1714

Query  5771  FVSHMQLAVDKLGLT--MPGLRVSYNDRMLLWKEKYM  5875
              +  M+L  ++  LT  +P L  SY+  +  ++ KY+
Sbjct  1715  KLKQMRLIAERANLTHMIPELEFSYDYLLDNFRAKYL  1751
```

Figure 5.20: Protein level sequence similarity of novel picorna-like viral contig to the existing sequences in RefSeq protein databases. (a) Schematic representation of the novel picorna-like virus. Open Reading Frames (ORFs) are highlighted with pink boxes and the Pfam domains are highlighted in blue. The GC-content of the sequence was obtained using window size 50 and it is shown in the plot below. (b) and (c) Protein sequence alignments for novel picorna-like virus against its closest relative found in the NCBI RefSeq protein databases. The sparse protein-level similarity indicates the diversity of the novel sequence being distantly related to its closest relative found in the BLAST databases.

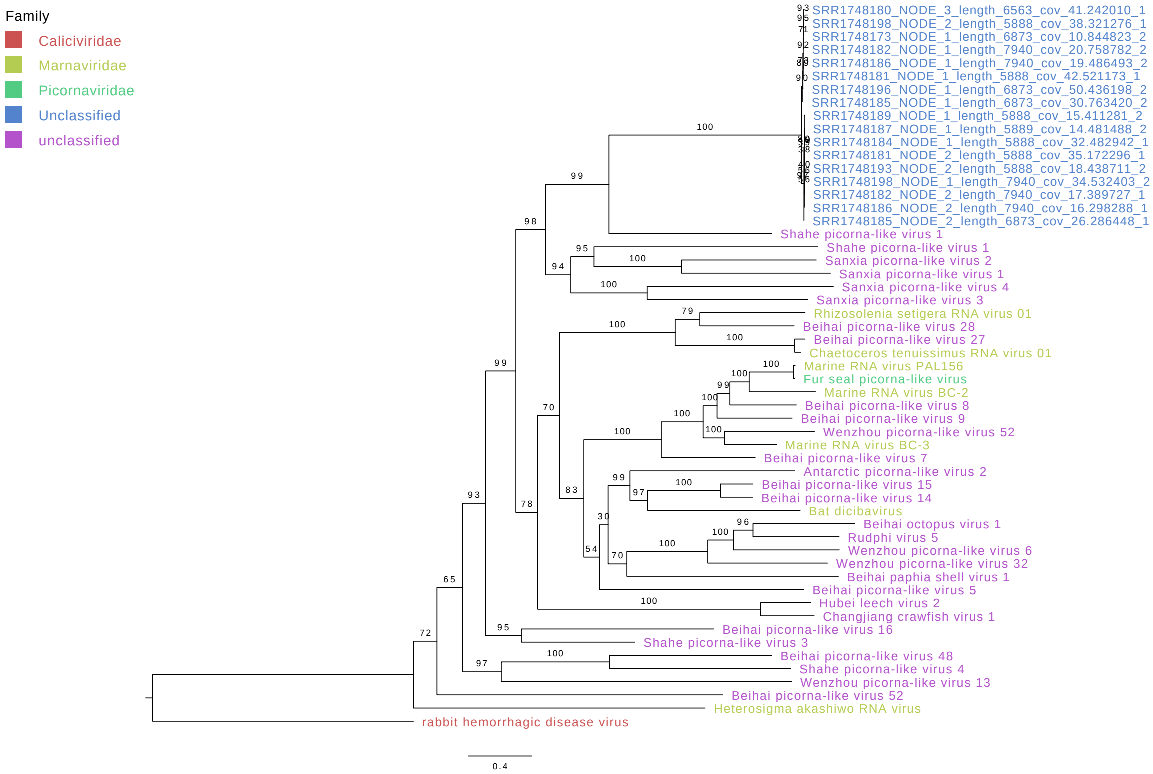

Family
🟥 Caliciviridae
🟩 Marnaviridae
🟩 Picornaviridae
🟦 Unclassified
🟪 unclassified
SRR1748180_NODE_3_length_6563_cov_41.242010_1
SRR1748198_NODE_2_length_5888_cov_38.321276_1
SRR1748173_NODE_1_length_6873_cov_10.844823_2
SRR1748182_NODE_1_length_7940_cov_20.758782_2
SRR1748186_NODE_1_length_7940_cov_19.486493_2
SRR1748181_NODE_1_length_5888_cov_42.521173_1
SRR1748196_NODE_1_length_6873_cov_50.436198_2
SRR1748185_NODE_1_length_6873_cov_30.763420_2
SRR1748189_NODE_1_length_5888_cov_15.411281_2
SRR1748187_NODE_1_length_5889_cov_14.481488_2
SRR1748184_NODE_1_length_5888_cov_32.482942_1
SRR1748181_NODE_2_length_5888_cov_35.172296_1
SRR1748193_NODE_2_length_5888_cov_18.438711_2
SRR1748198_NODE_1_length_7940_cov_34.532403_2
SRR1748182_NODE_2_length_7940_cov_17.389727_1
SRR1748186_NODE_2_length_7940_cov_16.298288_1
SRR1748185_NODE_2_length_6873_cov_26.286448_1
Shahe picorna-like virus 1
Shahe picorna-like virus 1
Sanxia picorna-like virus 2
Sanxia picorna-like virus 1
Sanxia picorna-like virus 4
Sanxia picorna-like virus 3
Rhizosolenia setigera RNA virus 01
Beihai picorna-like virus 28
Beihai picorna-like virus 27
Chaetoceros tenuissimus RNA virus 01
Marine RNA virus PAL156
Fur seal picorna-like virus
Marine RNA virus BC-2
Beihai picorna-like virus 8
Beihai picorna-like virus 9
Wenzhou picorna-like virus 52
Marine RNA virus BC-3
Beihai picorna-like virus 7
Antarctic picorna-like virus 2
Beihai picorna-like virus 15
Beihai picorna-like virus 14
Bat dicibavirus
Beihai octopus virus 1
Rudphi virus 5
Wenzhou picorna-like virus 6
Wenzhou picorna-like virus 32
Beihai paphia shell virus 1
Beihai picorna-like virus 5
Hubei leech virus 2
Changjiang crawfish virus 1
Beihai picorna-like virus 16
Shahe picorna-like virus 3
Beihai picorna-like virus 48
Shahe picorna-like virus 4
Wenzhou picorna-like virus 13
Beihai picorna-like virus 52
Heterosigma akashiwo RNA virus
rabbit hemorrhagic disease virus
93
95
71
92
73
89
90
100
38
40
56
97
56
99
98
95
100
94
100
79
100
100
100
100
99
99
100
70
100
100
100
100
83
99
100
78
93
97
30
96
100
100
54
70
100
65
95
72
100
97
0.4

Figure 5.21: A maximum likelihood phylogeny derived based on the polyprotein ORF (largest ORF) amino acid sequences showing the clustering of the novel picrona-like virus with other picorna-like viruses. The phylogenetic tree was generated using 1000 bootstrap in IQTree showing the relationship between the novel picorna-like contig to its BLASTX hits in RefSeq protein databases. The virus names are coloured according to the corresponding virus family.

were extracted from VMR 200721 version with MSL 36. This set was filtered (n=59 removed) further to include all species exemplar with a RefSeq sequence which yielded 272 sequences. The largest ORFs that coded for polyproteins were extracted for each of these sequences and then this set was merged with sequences from PL01. All sequences were sanity checked to remove any duplicate sequence identifiers and a final set with 319 sequences was created which is referred to as P01 hereafter.

P01 set contained sequences representing 8 distinct families; *Caliciviridae* (n=15), *Dicistroviridae* (n=15), *Iflaviridae* (n=15), *Marnaviridae* (n=20), *Picornaviridae* (n=137), *Polycipiviridae* (n=8), *Secoviridae* (n=62), *Solinviviridae* (n=2) of the order *Picornavirales* as well as unclassified picorna-like sequences from PL01 along with picorna-like virus sequences identified from PRJNA271229. A polyprotein amino acid sequence-based phylogeny of the P01 set was generated and it is shown in figure 5.22. The phylogenetic tree of the order *Picornavirales* shows clear clustering of the families based on the polyprotein ORF. Most of the sequences from each family are clustered with other sequences of the same virus family with high bootstrap support. Unclassified sequences that are termed picorna-like by M. Shi et al. (2016a) are shown to cluster with existing members of the family *Marnaviridae* often with very high bootstrap support suggesting that these picorna-like viruses could potentially be classified with respect to their sequence similarity with existing marnaviruses. The sequences assembled from PRJNA271229 do not cluster with any other sequences, confirming that they represent a currently unidentified novel picorna-like virus.

It has been noted by the ICTV that most of the families (except *Iflaviridae*) in the order *Picornavirales* form monophyletic branches in a maximum-likelihood-based phylogenetic tree derived using the amino acid sequence extracted from the protein-polymerase region, and each family has a defined host range (https://talk.ictvonline.org/ictv-reports/ictv_9th_report/positive-sense-rna-viruses-2011/w/posrna_viruses/227/picornavirales). For example, members of the family *Secoviridae* have a bipartite genome and infect plants, members of the family *Picornaviridae* contain monopartite genomes and infect vertebrates and families *Dicistroviridae* and *Iflaviridae* include viruses with monopartite genomes that infect arthropods (King et al., 2012). As this novel picorna-like virus that was found in the human blood samples from multiple individuals from Nigeria, clusters with unclassified viruses and members of the family *Marnaviridae* that infect unicellular organisms including marine algae and other marine invertebrates, it is difficult to determine its host species.

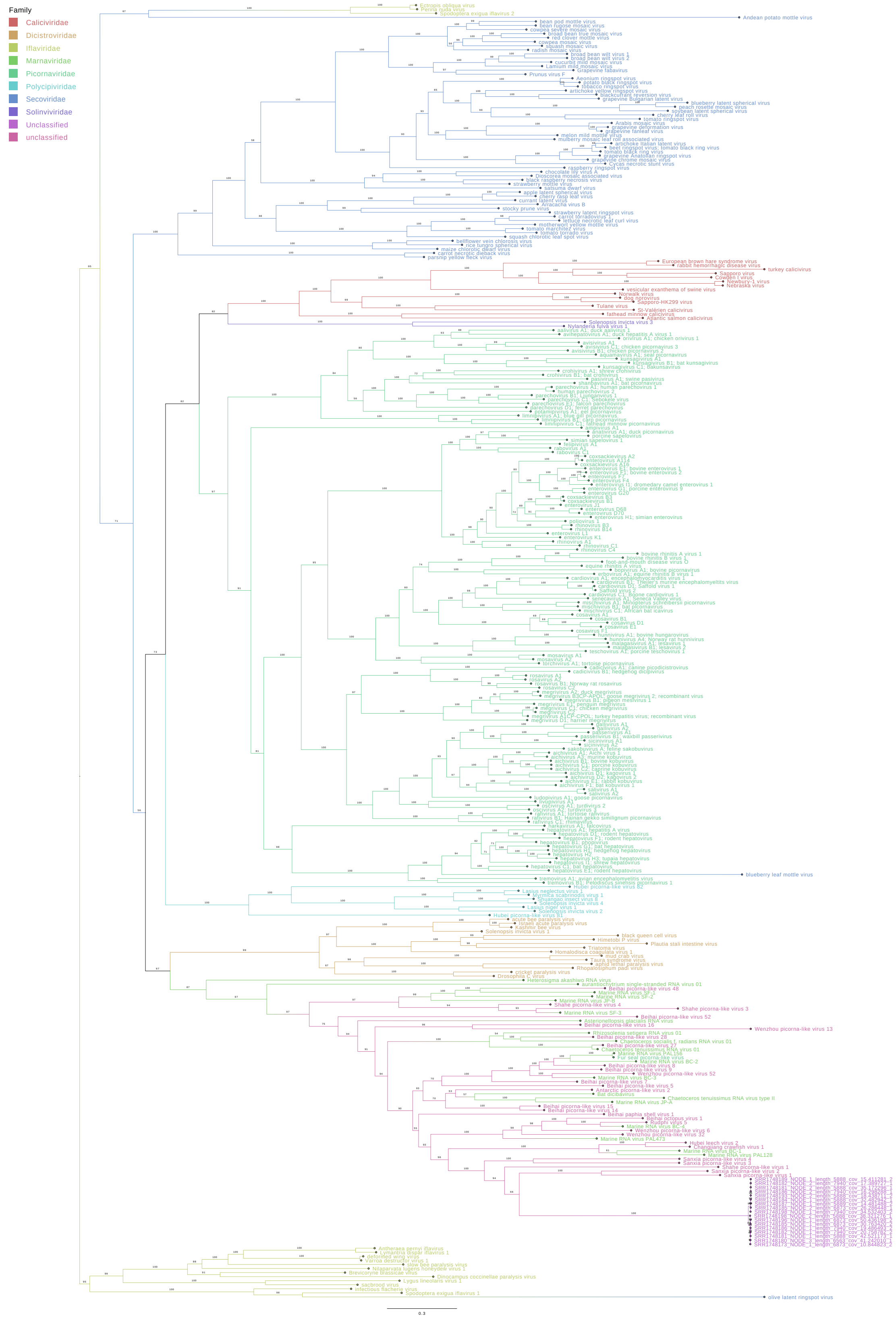


Figure 5.22: A midpoint rooted maximum likelihood-based on polyprotein ORF amino acid sequences for all representative species in order *Picornavirales*. This phylogenetic tree was generated using 1000 bootstrap in IQTree and it was visualised in FigTree with midpoint root settings. All species are coloured according to their corresponding family membership in the order *Picornavirales*. The contigs assembled in this study are shown in purple with complete contig names used as their labels.

5.4.7 Unclassified viral contig sequences (UViCs)

Unclassified viral contig sequences (UViCs) represented by the grey box in the figure 5.1 included confirmed viral contig sequences that could not be confidently associated with a virus family and/or order. A total of 47,994 contigs (out of 122,884) representing more than a third (39.06%) of all confirmed viral contigs were included in this category. 47,341 UViCs were deemed to be Genome-fragments and 653 were High-quality according to the MIUViG criteria. Among this set of 653 High-quality contigs, 557 were predicted to belong to the dsDNA phage group according to the VirSorter2 categorisation suggesting they are likely to be novel phage sequences. Out of 47,994 UViCs, 4,451 were predicted as being viral using all three prediction software. 14,375 UViCs were predicted using DeepVirFinder and VirSorter2; 2,437 were predicted by DeepVirFinder and TetraPredX, and 1,677 were predicted to be originating from virus genomes using TetraPredX and VirSorter2. A range of software-specific predicted viral contigs was also identified. 17,553 UViCs were predicted exclusively by VirSorter2 whereas 5,819 and 1,682 UViCs were predicted exclusively using DeepVirFinder and TetraPredX respectively. Among all three prediction tools, VirSorter2 can provide a virus group that the viral contig may be originating from. The UViCs were categorised as follows based on their VirSorter2 `max_score_group` categorisation: 1,890 NCLDV, 55 RNA, 19,389 dsDNAphage, 1,653 lavidaviridae, and 13,838 ssDNA. Max score group could not be determined for the remaining 11,169 UViCs.

Table 5.2: Distribution of Unclassified viral contig sequences (UViCs) across different microbiomes and BioProjects.

BioProject	Microbiome	Country	Number of Contigs	Number of SRA Run ID
PRJNA471187	Blood		1	1
PRJNA602694	Blood	Brazil	2	1
PRJNA389455	Blood		3	3
PRJNA518922	Blood		10	3
PRJDB7117	Blood	Japan	17	7
PRJNA271229	Blood	Nigeria	47	20
PRJNA471187	Blood	Sweden	64	50
PRJNA419524	Blood	USA	91	28
PRJEB11554	Fecal		1	1
PRJEB15257	Fecal	United Kingdom	65	6
PRJEB23207	Fecal	Italy	195	7
PRJEB8201	Fecal	Egypt	208	1
PRJEB23207	Fecal	Netherlands	260	4
PRJEB8201	Fecal	USA	261	1
PRJEB17784	Fecal	Germany	672	52
PRJEB1775	Fecal	Germany	930	38
PRJEB19090	Fecal	Italy	1268	37
PRJEB18265	Fecal	Russia	1523	9
PRJEB19367	Fecal	USA	2031	28
PRJEB7331	Fecal	United Kingdom	2397	24
PRJEB6542	Fecal	Netherlands	2714	8
PRJEB7949	Fecal	United Kingdom	5240	40
PRJEB6092	Fecal	Australia	5479	24
PRJEB8094	Fecal	Canada	6661	100
PRJEB12357	Fecal	Netherlands	9835	100
PRJEB21827	Human		184	12
PRJEB7248	Lung	Gambia	23	3
PRJEB12831	Oral	United Kingdom	27	11
PRJEB15334	Oral	United Kingdom	164	23
PRJNA230363	Oral	China	2584	28
PRJNA264728	Saliva		81	8
PRJEB14383	Saliva	Philippines	4599	30

BioProject	Microbiome	Country	Number of Contigs	Number of SRA Run ID
PRJEB10295	Skin	Netherlands	2	1
PRJEB10919	Sputum	South Africa	343	17
PRJEB21446	Vagina	Germany	12	8

UViCs were not specific to a study and/or microbiomes and were found in all major microbiomes represented by 35 BioProjects shown in table 5.2. However, they were predominantly found in fecal microbiome datasets. A total of 39,740 UViCs were found in fecal microbiome datasets which represented 82.8% of all UViCs identified here. The largest number of UViCs were found in fecal microbiome BioProjects PRJEB12357 (n=9,835; Netherlands), PRJEB8094 (n=6,661; Canada), PRJEB6092 (n=5,479; Australia) and PRJEB7949 (n=5,240; United Kingdom). 4,680 UViCs were found in saliva microbiome datasets. 4,599 of these saliva UViCs were found in BioProject PRJEB14383 (Philippines) and 81 were identified from BioProject PRJNA264728. 2,775 UViCs were assembled from oral microbiomes and a large proportion of these (n=2,584) were assembled from BioProject PRJNA230363 from China. 343 UViCs were identified from sputum samples from BioProject PRJEB10919 from South Africa. 235 UViCs were assembled from blood microbiome and these UViCs originated from BioProjects PRJNA419524 (n=91; USA), PRJNA471187 (n=65; Sweden), PRJNA271229 (n=47; Nigeria), PRJDB7117 (n=17; Japan), PRJNA518922 (n=10), PRJNA389455 (n=3) and PRJNA602694 (n=2). Miscellaneous BioProject PRJEB21827 with microbiome specified as ‘Human’ and no geolocation metadata comprised of 184 UViCs. Remaining UViCs were assembled from lung (n=23), vagina (n=12) and skin (n=2) microbiomes. These results highlight the importance of cataloguing the virosphere as it represents the viral sequences that are yet to be classified and incorporated into the formal virus classification framework such as taxonomy. It also emphasises the significance of carrying out systematic virus sequence prediction and identification analyses such as that applied here as it can lead to the expansion of viral genome sequence space.

A superficial high-level clustering of UViCs carried out using the standard protocol specified in 5.3.5 generated 20,734 vOTUs from 47,994 UViCs. 5,479 clusters (with at least 2 sequences) were generated and 15,255 UViCs were deemed to be singletons (i.e. UViCs that did not cluster with any other sequences). This suggests that a third of UViCs bear some sequence similarity to one another. These UViCs could be explored further to investigate their potential virus origin as currently they represent the “viral unknowns” and cannot be grouped with known virus families, order or realm in the existing virus taxonomy framework.

5.5 Discussion

In this study, nearly 7.2 million contigs originating from 3,559 samples covering 11 distinct human microbiomes from 58 studies (BioProjects) were systematically interrogated for the purpose of virus discovery. This analysis enabled the identification of 122,884 confirmed viral contigs that could be confidently assigned to the taxonomic rank ’Viruses’ as superkingdom. This has led to the creation of a rich dataset with 49,247 (40.02%) contigs that could be confidently associated with virus families and the remaining 59.92% (n=73,637) that could only be classified at above family ranks, sometimes only at the top level e.g. viruses. 1,422 (1.16%) of these viruses were deemed High-quality and were complete genomes whereas 121,462 (98.84%) were likely to be genome segments or partial viral sequences according to the MIUViG quality criteria. However, it is worth noting that the MIUViG criteria can provide limited information in cases of ssDNA as well as RNA viruses due to their shorter genome lengths. Moreover, the resolutions required for RNA virus identifications are often limited as RNA viruses can possess segmented genomes.

Three distinct virus prediction tools were applied to the assembled contigs set and the prediction results were validated using gold-standard sequence similarity methods including BLAST and DIAMOND. These results also show that k-mer-based prediction tools often lead to more false-positive compared to the more traditional sequence similarity methods and a combination of these methods such as the ones implemented in VirSorter2. VirSorter2 approach performs extremely well and can help to reduce false positives in virus discoveries. However, it is worth noting that a conservative and gold-standard BLAST approach for validating the prediction results has been applied here which has its own limitations. Moreover, it can be argued that VirSorter2’s algorithm that uses tools such as Prodigal for ORF prediction and HMMER3 for HMM profile identification for viral genes is similar to the sequence similarity approach implemented in BLAST. This integrated approach that utilises existing databases, and known and uncultivated virus sequence space is certainly better at virus sequence prediction compared to other tools that solely rely on short signature k-mers. Nevertheless, it should be noted that all prediction methods were susceptible to false-positive predictions albeit to a varying degree. The validation results show that TetraPredX was able to predict more RNA viruses compared to DeepVirFinder and VirSorter2. Although it is likely that those RNA viral contigs were simply missed because of the various prediction threshold applied to the output of prediction tools, it is worth noting that TetraPreX was able to predict those RNA viral contigs with very high confidence suggesting that our simple k-mer based prediction method is able to deconvolute the microbial genomic signals embedded within tetra-mer frequencies. VirSorter2 gave null confident scores for a number of RNA viral contigs, which may be because these contigs had only one complete gene and possessed a recognisable hallmark gene (personal communication with Jiarong Guo at EVBC 2022, https://github.com/jiarong/VirSorter2/issues/68). Moreover, as a more comprehensive prediction analysis is carried out by VirSorter2, it was the slowest among the three prediction tools used here. In reality, it is impossible to completely measure the exact number of microbial species present in a metagenomic or metaviromic sample except those originating from pre-designed mock communities. Thus, it should be appreciated that sequences that belong to root and match more than one superkingdom can potentially be originating from virus genomes but they cannot be confidently assigned to a group of viruses due to the limitation of our validation approach.

Despite the algorithmic advances in sequence classification, taxonomic classification and characterisation are deemed to be challenging tasks. Metagenomic data analysis is overly subjective and it is especially more difficult for virus classification. The reason for this is partly due to the sheer amount of genomic diversity embedded in viral genomes, as well as the specialised approach required to categorise different types of viruses, e.g. phages vs RNA viruses. In order to capture various different types of viruses, customised analytical approaches were developed that are described in the results section. To verify the viral origin of predicted viral contigs, a general-purpose, but much slower BLAST-based identification method was employed here. Arbitrary criteria of evalue 0.0001 for nucleotide level classification and 0.001 for protein level classifications without a query coverage filter were applied to obtain the top 25 hits for each predicted viral contig. This BLAST-based validation method could be considered more suitable for viruses in contrast to the short k-mer-based classification method as viruses have been shown to mimic their host genomic signatures to evade host immunity.

In order to strengthen the taxonomic assignment of viral contigs, I have also utilised the approach whereby the top 25 hits for a sequence are extracted and then an LCA of these hits is obtained to determine the final taxonomic “class” of the contig. This LCA-based classification strategy is deemed more effective and works in most cases (Mande et al., 2012; Huson et al., 2007; McIntyre et al., 2017) but fails if the sequences present in the databases being interrogated are misclassified. For example, it was noted that due to misannotation of GenBank entries JX157239 and JX157238, a range of anellovirus contigs was assigned to root despite the majority of BLAST homologues for those contigs being anelloviruses. Despite this limitation, LCA taxonomy classification approach works better than other alternatives such as top hit which can be misleading and may not provide the complete picture as the parameter that is used to restrict the output to "top hit" (`max_target_seqs` = 1) is not designed to provide the best match (Shah et al., 2019). According to the BLAST+ manual, at least 5 hits for each match should be retrieved (https://www.ncbi.nlm.nih.gov/books/NBK279690/). Alternatively, a majority voting approach could be applied whereby a taxonomic group could be decided for each taxonomic level of the BLAST hits and the final taxonomy level could be decided based on the majority voting. This majority voting approach would overcome the specific issue discussed above however, it would become ineffective in cases where the majority votes are too close to call (Watson, 2021).

These results also confirm an overall expected pattern of virus presence in various studies and geolocations. As anticipated, a range of omnipresent viruses including bacteriophages and anelloviruses were found in most samples. Interestingly, unclassified viruses and those that cannot be linked back to a known virus family were more prominent in all types of microbiomes and studies sampled across different countries and locations. This suggests that though there have been advances in cataloguing the virosphere around us, a large proportion of viruses still remain unclassified, uncultured and uncharacterised. There were some relevant patterns to be explored further e.g. the diversity of RNA viruses in unexplored and undersampled regions such as countries in Africa; these patterns are more likely to be consistent with sample preparations meaning that we are able to capture more RNA viruses through metatranscriptomic studies compared to their metagenomic counterparts. It is worth noting that the geolocation analysis described in this study aims to provide a cursory overview of patterns of viruses we observed in this study and cannot be used to draw conclusions about the presence or more importantly absence of specific viruses in certain geographic regions.

The co-occurrence analysis carried out at various microbiome levels suggest that a large proportion of interactions between virus families that were observed here are likely to be random. Although some positive, as well as negative associations, were determined between virus families, these results were largely based on a very small number of pairs of samples. It is important not to over-analyse these associations due to the nature of the dataset used here. It was assumed that all samples were originating from individual human subjects, however, that is unlikely to be true as a range of studies included samples pooled from more than one individual. Further validation of the most interesting associations should be carried out using more suitable microbiome-specific datasets.

To predict the hosts for bacteriophages, a conservative but effective approach that predicts the host based on the CRISPR spacer interactions and looks for the overlap of these short sequences between virus and host genomes was applied here. It is important to note other tools and methods such as those that utilise short nucleotide and amino acid k-mers shared between virus and hosts (Ahlgren et al., 2017; Villarroel et al., 2016; Babayan et al., 2018; Young et al., 2020), CpG composition (Simmonds et al., 2013), and the presence of a prophage genome (Roux et al., 2015b; Roux et al., 2015a) have also been widely applied to efficiently predict hosts for both prokaryotic and eukaryotic viruses. Identification of novel bacteriophages can help to provide further insights into their interactions with their hosts as well as any implications that it can have on human health. Furthermore, the specificity of such virus-host interactions can be explored further and can be contextualised in treatments such as phage therapy.

A range of recent studies has identified a previously unexplored diversity of anelloviruses, collectively known as the anellome (Moustafa et al., 2017; Tisza et al., 2020; Arze et al., 2021). The results described here extend the diversity of human anelloviruses found in human blood and oral microbiomes. Only 23.45% (299 out of 1,275 genomes) of anelloviruses discovered in this study belong to a currently known human anellovirus species. A thorough analysis carried out here indicates that the remaining sequences can be represented with 228 novel human anellovirus species which expands the diversity of known human anellovirus species threefold. The largest catalogue of previously unknown Torque teno midi viruses belonging to the genus *Gammatorquevirus* that was found in oral samples from BioProject PRJNA230363 from China was discovered. Anellovirus phylogenetic analysis also found a novel clade of anelloviruses that is separate from currently known human anellovirus genera. Our analysis also shows that this new clade is also separate from the genus *Omegatorquevirus* which currently has only one species namely Torque teno hominid virus 1 that was isolated from Gorilla. Due to their compact genomes, anelloviruses are known to recombine within the same host (Arze et al., 2021). It would be interesting to explore the recombination patterns in the species we have assembled in this study. Moreover, a further detailed recombination analysis could also shed light on whether the genomes included in the novel clade identified here are results of recombination. There are 72 currently known species of anelloviruses and our analysis has greatly expanded the known anellovirus sequence species diversity by identifying three times as many novel species in this study. A further analysis comparing these novel anellovirus species with those identified in other studies such as Tisza et al. (2020) and Arze et al. (2021) could help reveal the distribution of these novel species across different samples, studies and geolocations, as well as help, get insights into the world of these omnipresent human viruses that have been co-evolving with their hosts. Notably, although they have been termed the "friendly" human viruses, their role in shaping human immunity specifically in immune-compromised individuals remains unknown. The advances made through metagenomics have helped to understand the extent of their diversity and further confirmatory studies can help to determine their evolutionary history, host range as well as a critical role in the wider area of virology research.

Although our study focuses on the metagenomic samples, the virus discovery approach devised here can also be applied to metatranscriptomic samples. Only a small proportion of metatranscriptomics samples were included here that were investigated for the presence of RNA viruses. A range of known RNA viruses was found in various samples originating from blood, fecal and saliva microbiomes from different locations including Brazil, Nigeria, the USA and the UK. A novel picorna-like virus was found in the blood sample from Nigeria that was shown to have a very diverse genomic composition compared to any known members of the order *Picornavirales*. The phylogenetic analysis showed that this novel picorna-like virus is likely to be very divergent from any known viruses as it shows sparse homology even at the protein sequence level to its nearest relatives in the phylogeny. Although it was not possible to determine its potential host, this virus sequence was shown to cluster with other viruses that infect freshwater arthropods albeit with very low sequence similarity. A range of other contigs likely to represent potentially novel RNA viruses were identified. These contigs were predominantly matched to unclassified viruses and displayed very low sequence identity to known unclassified RNA viruses at the protein level, suggesting that they too are likely to be distantly related to any known RNA viruses.

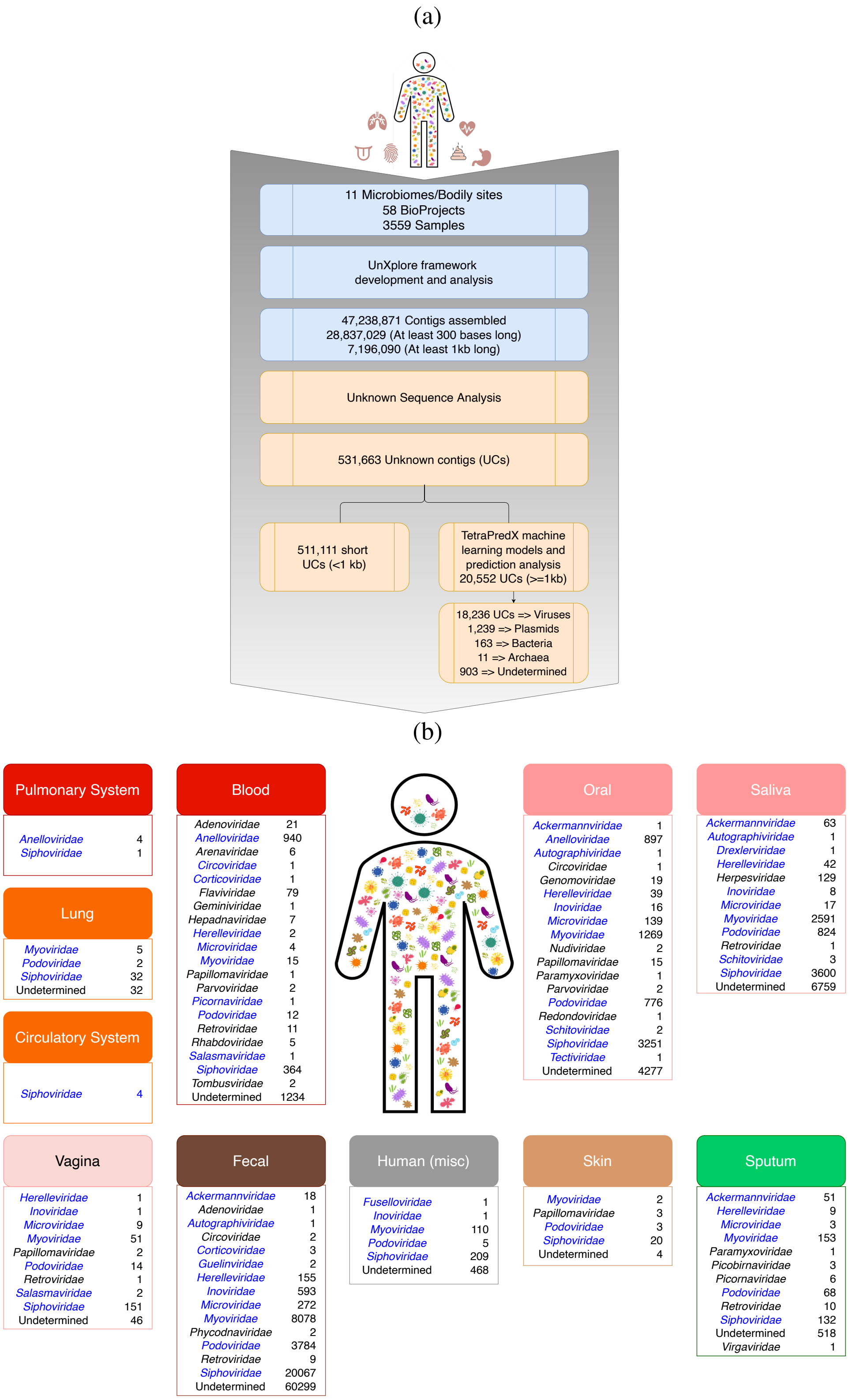
A large diversity of unclassified virus sequences was also discovered embedded within different human microbiome datasets. Here, the unclassified virus sequences were defined as those whereby the LCA of the nucleotide or protein hits could not be mapped to a known virus family and/or order suggesting that viral contig was matching to sequences from multiple virus families and/or orders. These unclassified sequences are likely to represent uncultivated and novel families, orders or realms of viruses. An overlap among these unclassified virus sequences and uncultivated virus sequence databases such as GPD, GVD and IMG/VR is anticipated. It is noteworthy that, unlike the RefSeq virus sequence database, the uncultivated virus sequence databases are currently being developed and managed in silos. There is a need for a federated or curated non-redundant database that can link these repositories. Further endeavours could be made by the virus discovery community to centralise and unify the uncultivated virus sequences such that they are automatically updated, integrated and can be queried in a more systematic fashion. To enable users to access such databases efficiently and programmatically, an application programming interface (API) could also be made available through standard database services such as ENA, NCBI or IMG/VR.

Chapter 6

Discussion

Advances in metagenomic and metatranscriptomic techniques have opened up a completely new horizon of microbial research. These techniques have inevitably revolutionised the way we study microbes. A plethora of microbiome and environmental datasets get submitted to the International Nucleotide Sequence Database Collaboration (INSDC) repositories. These raw sequence datasets have been shown to contain a large of number uncultivated and unclassified microbial sequences that are often referred to as microbial dark matter or unknown sequence matter. In the inquest of cataloguing the unknown sequences, an extensive, modular, portable analyses framework - UnXplore was designed and developed. UnXplore is a modular pipeline written and implemented in SnakeMake that can automatically assemble, analyse and quantify the presence of unknown sequences in the short read sequence datasets. It is a portable framework that can be installed and run on any computer and requires minimal setup to enable users to analyse their own datasets of interest. As it is contained in preconfigured Conda environment minimal user intervention is required. The UnXplore framework simply requires the download of the relevant databases of interest and specifying these locations in a simple configuration file along with the location of the input fastq files. UnXplore can QC the reads, remove adapters, remove human host sequences, assemble the non-human reads into high-quality contigs and search the contigs against existing nucleotide and protein data repositories in a systematic and automated fashion. Once the database searches are completed, a range of Python scripts included in UnXplore can quantify the proportion of known, partially known and unknown sequences in them. UnXplore can also be used as an effective metagenomic analysis pipeline as it carries out extensive searches against the relevant INSDC databases and provides a high-quality per sequence quantification of known and partially known sequences along with their best hits and the lowest common ancestor (LCA) of the top 25 hits. Moreover, UnXplore also produces contig metadata including the number of reads aligning to contigs, depth and breadth of reads coverage for each contig assembled using this pipeline.

UnXplore framework which is now published in mSystems (Modha et al., 2022) was applied to 3,559 human microbiome datasets from 58 BioProjects that were sequenced on the Illumina



Pulmonary System	
Anelloviridae	4
Siphoviridae	1

Lung	
Myoviridae	5
Podoviridae	2
Siphoviridae	32
Undetermined	32

Circulatory System	
Siphoviridae	4

Blood	
Adenoviridae	21
Anelloviridae	940
Arenaviridae	6
Circoviridae	1
Corticoviridae	1
Flaviviridae	79
Geminiviridae	1
Hepadnaviridae	7
Herelleviridae	2
Microviridae	4
Myoviridae	15
Papillomaviridae	1
Parvoviridae	2
Picornaviridae	1
Podoviridae	12
Retroviridae	11
Rhabdoviridae	5
Salasmaviridae	1
Siphoviridae	364
Tombusviridae	2
Undetermined	1234

Oral	
Ackermannviridae	1
Anelloviridae	897
Autographiviridae	1
Circoviridae	1
Genomoviridae	19
Herelleviridae	39
Inoviridae	16
Microviridae	139
Myoviridae	1269
Nudiviridae	2
Papillomaviridae	15
Paramyxoviridae	1
Parvoviridae	2
Podoviridae	776
Redondoviridae	1
Schitoviridae	2
Siphoviridae	3251
Tectiviridae	1
Undetermined	4277

Saliva	
Ackermannviridae	63
Autographiviridae	1
Drexlerviridae	1
Herelleviridae	42
Herpesviridae	129
Inoviridae	8
Microviridae	17
Myoviridae	2591
Podoviridae	824
Retroviridae	1
Schitoviridae	3
Siphoviridae	3600
Undetermined	6759

Vagina	
Herelleviridae	1
Inoviridae	1
Microviridae	9
Myoviridae	51
Papillomaviridae	2
Podoviridae	14
Retroviridae	1
Salasmaviridae	2
Siphoviridae	151
Undetermined	46

Fecal	
Ackermannviridae	18
Adenoviridae	1
Autographiviridae	1
Circoviridae	2
Corticoviridae	3
Guelinviridae	2
Herelleviridae	155
Inoviridae	593
Microviridae	272
Myoviridae	8078
Phycodnaviridae	2
Podoviridae	3784
Retroviridae	9
Siphoviridae	20067
Undetermined	60299

Human (misc)	
Fuselloviridae	1
Inoviridae	1
Myoviridae	110
Podoviridae	5
Siphoviridae	209
Undetermined	468

Skin	
Myoviridae	2
Papillomaviridae	3
Podoviridae	3
Siphoviridae	20
Undetermined	4

Sputum	
Ackermannviridae	51
Herelleviridae	9
Microviridae	3
Myoviridae	153
Paramyxoviridae	1
Picobirnaviridae	3
Picornaviridae	6
Podoviridae	68
Retroviridae	10
Siphoviridae	132
Undetermined	518
Virgaviridae	1

Figure 6.1: A schematic illustrating the main findings described in this thesis. (a) A summary of unknown sequence analysis and results, stating the number of samples, microbiomes, and BioProjects analysed and described in Chapters 3 and 4. (b) An overview of microbiome-associated virus families provides a snapshot of the human-associated virome characterised and described in detail in Chapter 5. Virus families associated with various microbiomes are shown in the corresponding microbiome boxes along with the number of contigs associated with each family. The blue shading indicates the contigs associated with specific virus families that were analysed in depth.

sequencing platform (figure 6.1(a)). By utilising this methodology it was shown that on average 2% of assembled sequences were shown to be categorised as unknown contigs (UCs) meaning that they did not have significant sequence similarity to any known sequence at the nucleotide or protein level. Despite the lack of nucleotide and protein level sequence similarity to any known sequences, a third of the UCs were shown to contain open reading frames (ORFs) that were at least 100 AA long. Moreover, a large proportion of these predicted ORFs was also shown to have protein domain signatures. Overall, 5.49% of UCs that could not be labelled taxonomically, could be annotated functionally based on the domain signatures. New sequences are added to INSDC repositories on a daily basis. To compute the rate at which the UCs become taxonomically classified - i.e. have a match to a known sequence in the INSDC databases, the sequence similarity-based analysis was carried out at four distinct time points over the course of 18 months. As new sequences get analysed, catalogued and added to the INSDC repositories, the probability of finding a match/hit to the UCs would increase over time. This was demonstrated using the database searches carried out at different time points, as the proportions of UCs were shown to decrease over time as the databases are updated. Overall, an estimated 1.64% of the UCs were identified per month, however, it was anticipated that this rate would plateau as more time points would be considered. The presence of ORFs, domains and the UCs matching to known sequences as time passes, strongly supported the hypothesis that UCs are of biological origin. A number of large and small unknown contigs were analysed further. These contigs were shown to contain large ORFs and some of those ORFs also had protein domains embedded within them further confirming the biological origin of the unknown sequences.

The proportion of UCs varied greatly between different types of microbiomes. For example, less studied microbiomes such as skin and oral that are often more exposed to the outside environment were shown to harbour a higher proportion of UCs. On the contrary, the blood microbiome contained a much smaller number of UCs. One explanation for this can be that some human microbiomes such as fecal, skin or oral microbiomes tend to contain a range of commensal and non-commensal microbes, however, compared to these microbiomes, blood is considered a sterile environment. This was observed when despite analysing a large number of blood microbiomes using UnXplore a very small number of large UCs (>=1kb) were identified. However, it is still very important to analyse human blood microbiome datasets in this context as diseases of unknown aetiology could potentially be associated with the presence of blood-borne microbes, specifically viruses.

It was hypothesised that a large number of UCs could be originating from genomes of uncultivated microbes, specifically viruses. To test this, extensive supervised machine learning models were designed, developed, tested and validated. It was noted from the literature that short nucleotide frequencies can be unique to each microbe and have been shown to encompass robust phylogenetic signals (Pride et al., 2006). To build on this, TetraPredX - a Python package that can learn and predict the microbial origin of a UC based on its tetranucleotide k-mer (k=4) frequency was developed as part of this project. TetraPredX is a supervised machine learning method that implements k-mer-based prediction using the Random Forest (RF) algorithm. The models were trained, tested and cross-validated using the gold standard genomes and sequences originating from four different microbial classes namely archaea, bacteria, plasmid and viruses. TetraPredX models were shown to demonstrate very high precision and recall for all four classes and were effectively able to separate the genomic signatures embedded within the sequences to predict the class of a given input sequence with very high accuracy.

Sequence predictions carried out using TetraPredX, DeepVirFinder and VirSorter2 indicated that a large proportion of UCs was likely to be originating from viruses (figure 6.1(a)). These results obtained using a method developed here, as well as using the widely used virus predictions software supported the initial hypothesis that UCs are indeed representatives of uncultured virus diversity that is yet to be catalogued in the standard INSDC databases. This finding emphasises the importance of scanning the microbiome and environmental datasets available in public repositories that are often analysed with a specific research question in mind, and can often contain a completely unknown and uncharacterised diversity of microbes that are yet to be identified.

It is appreciated that TetraPredX is not the first implementation of k-mer frequency-based taxonomy predictions and many such tools e.g. VirFinder, DeepVirFinder, IDTAXA exist that can predict a class of the microbial sequences based on the short nucleotide frequencies. However, TetraPredX is the first tool to incorporate models that can predict the microbial class across multiple superkingdoms as well as plasmids. TetraPredX provides an opportunity to think of microbial communities as a whole and look for all microbial and mobile element-related signals present in them and then use them efficiently to predict the potential class of the UCs. Though TetraPredX was designed to predict the class of UCs, it can be easily used for standard metagenomic contigs originating from microbiome datasets as it is designed to work off any given fasta input.

Despite its very high precision and recall, TetraPredX models suffer the same shortcomings as those that are known for use of short k-mer frequencies. ML models developed based solely on short k-mer frequencies are sensitive to false positives. This feature is more prominent in the case of viruses as viruses are known to mimic their hosts' genomic signatures in order to escape the immune responses associated with their hosts. All k-mer-based prediction methods including those that are widely accepted and used by the research communities are challenged with this limitation and are continuously being improved to tackle the complexity posed by such biological phenomenons. Moreover, k-mer frequency-based predictions have been shown to work very well in the case of short sequences as well as prediction of RNA viruses as they are able to capture the compound signals obtained from gold standard datasets and can overcome other sequence similarities related limitations such as finding things that are largely similar to those that are currently included in the nucleotide and protein databases. In absence of significant sequence similarity, ML methods provide a valuable layer of information albeit with a probability associated with it which can be analysed further to identify the sequence features and their potential biological origin. It is becoming more and more apparent that ML methods combined with other approaches such as domain information, GC content, presence/absence of hallmark genes and protein signatures can often yield better predictions and can help to reduce false positives in metagenomic sequence analyses (Guo et al., 2021a).

Machine learning models serve as a powerful tool to explore the relationship of novel sequences by learning and applying the patterns observed from the real-world known datasets. Nevertheless, with the ever-expanding INSDC data repositories with new biological sequences being added to them on a daily basis, the application of these models can often be temporary. The ML models are as good as the data that was used to build and train these models, hence to retain the robustness of these models, they need to be revised with updated information and new data periodically. This could be a challenging task as scientific projects are commonly funded through short cycles of aid and financial support. The huge demand of keeping the data, models and relevant metadata up-to-date is a very demanding job that can be potentially challenging with researchers who develop and devise these models tending to also move on and away from the initial project due to the high turnaround of people and research needs. Moreover, all ML models suffer from sampling limitations meaning that if predictions are being made on substantially novel and previously unseen observations, the models tend to perform worse compared to the data that was previously "seen" by these models. Hence, the models would need to be updated perpetually to keep up with the high influx of novel data points. To address this limitation, an automated framework that can update, train and test models with newer datasets could be explored. Alternatively, instead of continuously updating and retraining the models, comprehensive standard test datasets could also be devised that are updated regularly with new sequences as model testing is often considered less burdensome than completely retraining the models. These datasets could be used to monitor the model performance over a period of time and when the model performance fails the desired standard, they would be expanded and/or trained, tested and validated with newer datasets.

Following the potential classification of UCs as viruses, the apparent next step was deemed to look for known and novel virus sequence diversity embedded within the SRA repositories. A number of recent studies have shown that microbiome and environmental datasets harbour a large number of uncultivated virus genome sequences. Systematic meta-omic analyses have led to the expansion of the virosphere by cataloguing viruses that simply could not have been identified due to the limitation of culture-based approaches. The culture-independent approaches are immensely useful in virus discovery as viruses are extremely tricky to grow in labs as they often require specific host(s) cells and precise environmental conditions are very hard to simulate in traditional laboratory settings.

UnXplore framework provided an important foundation to carry out virus discovery analyses on 3,559 samples that were already assembled into high-quality contigs. Over 7 million contigs were analysed using three distinct virus prediction tools TetraPredX, DeepVirFinder and VirSorter2. Each prediction method generated a large number of predicted contigs that were validated by nucleotide and protein level sequence similarity searches carried out using comprehensive nt and RefSeq protein databases. This analysis helped to characterise human-associated virome (figure 6.1(b)). A core set of 122,884 confirmed virus sequences was generated which was then analysed further to identify known and novel viruses. These explorations led to the identification of 321 novel bacteriophage species, 228 novel anellovirus species and 214 novel RNA virus contigs representing potentially novel and unclassified RNA viruses. Moreover, approximately 48 thousand unclassified virus contigs were also identified that could not be linked to a known virus family and/or order indicating the vast diversity of potentially novel virus genomes embedded within this dataset. Geographic and co-occurrence patterns were explored for the classified - i.e. where a virus contig could be linked to a known virus family, which indicated that most of the family-level interactions observed in these datasets were likely to be random. However, there were some interesting positive association patterns to be found between bacteriophage families in specific human microbiomes e.g. fecal. These patterns would be required to be validated with much larger, microbiome-specific datasets. As the initial datasets analysed here were not set out to be aimed at co-occurrence analysis and exploration of a positive and negative association between viruses, the co-occurrence analysis results should be carefully interpreted in the context of the presence and/or absence of specific virus families in specific microbiome samples.

The process of making sense of viral dark matter, including the sequencing and analyses, has matured significantly in the last decade with arguably a vast number of novel viruses identified through metagenomic and/or metatranscriptomic approaches compared to the traditional culture-based approach (Roux et al., 2021c). The virus-related knowledge derived from these datasets can then be fed back into making informed decisions about specific research questions and/or navigating hypotheses-driven science. One such striking finding was the identification of virus hallmark genes. Although viruses lack a universal signature gene such as 16S for bacteria, sequence-led analyses by Koonin et al. (2020a) provided a megataxonomy view of the viruses which helped to define and identify a range of virus hallmark genes that were shown to be conserved through the virus realm or subrealm level (Koonin et al., 2020a). These findings not only helped to gather and organise the global viral diversity but can also help to design and implement suitable algorithms and software that can efficiently translate these complex biological findings into resources and tools that can be readily applied to new and existing datasets. A real-world example of such a cyclic approach can be demonstrated with the example of RdRp which helped recover tends of thousands of novel RNA viruses from a wide range of environments (M. Shi et al., 2016a; M. Shi et al., 2018b; Wolf et al., 2020; Neri et al., 2022) as well as the specific RdRp palm domain (Venkataraman et al., 2018; H. Jia et al., 2019) and how it was further scrutinised to mine petabase scale SRA resources and led to the discovery of around 130,000 novel RNA virus sequences embedded within them (Babaian et al., 2021; Edgar et al., 2022). Similarly, virus prediction tools e.g. VirSorter2 (Guo et al., 2021a) heavily rely on detecting the presence of hallmark genes to predict the probability of a sequence originating from a virus genome.

With the help of advances made in computational virology, the process of virus identification from microbiome, environmental and clinical datasets has become an increasingly substantial topic of conversation in the virus research community. A community-level endeavour that enables quick and easy processing of metagenomic and metaviromic datasets has been undertaken by organisations such as JGI, NCBI and EBI. NCBI has developed new resources to increase the usability of NCBI viral genomes (Brister et al., 2015) and virus genome variation data (Hatcher et al., 2017). JGI has made a nontargeted virus genome detection pipeline available through their platform that users can apply to detect uncultivated viruses present in their sample. This pipeline formed the basis of one of the largest uncultivated virus data repositories IMG/VR (Paez-Espino et al., 2017; Paez-Espino et al., 2016). Virify - a pipeline derived from EBI's MGnify is currently being developed that enables users to systematically and automatically detect, annotate, and taxonomic classify viral contigs in metagenomic and metatranscriptomic assemblies (Rangel-Pineros et al., n.d.) (https://github.com/EBI-Metagenomics/emg-viral-pipeline). The inclusion of these analytical protocols as part of routine sequence data analyses enables non-expert users to understand their data to answer virus-centric research questions.

In recent years, several hundred thousand uncultivated viruses have been identified solely through computational virological approaches. As a result of the Coronavirus pandemic that began in 2020 and brought the world to a complete standstill, it has led us to realise that we must better understand the viruses that surround us. These viruses that exist as inert entities in environments that we regularly interact with, when provided with the suitable conditions can pose a very serious threat to humankind. Hence, it is very important to develop our understanding of the viral world and the first step in this process is to catalogue, characterise and potentially classify the diverse viruses that surround us. Typically viruses are classified in a hierarchical taxonomy framework to organise the virosphere and access their biological properties readily. However, in this era of metagenomics, classification of all uncultivated virus sequences into a taxonomy framework is considered a very ambitious task, albeit an important one as uncultured virus sequences represent a significant amount of biological insights that simply cannot be overlooked. Efforts are being made by the taxonomy communities to keep up with the pace and computational virology researchers are called upon to help with devising novel tools and algorithms that can capture the majority of uncultured virus diversity into the ICTV framework (Bas E Dutilh et al., 2021). It is equally important to realise that the taxonomy framework is likely to be in flux for some time due to the sheer amount of data that needs to be incorporated into the existing framework. Though alternative frameworks such as purely sequence-based taxonomy is suggested, a critical coherent assessment of such proposals must be carried out before completely breaking away from the norms and starting something completely new.

In addition to the virus sequence taxonomic classification, the databases that store and share the uncultivated virus sequence information are also confronted with challenges related to the exponential increase of these uncultured virus sequences. A number of resources including IMG/VR, GPD, GVD, unclassified and environmental assembled sequences in INSDC e.g. GenBank, nt or NCBI virus resource are all accessible to the researchers but they are developed and maintained independently by various organisations. The uncultured and unclassified virus data is dispersed across multiple independent platforms that do not communicate with one another which inevitably leads to information fragmentation. Due to the nature of uncultivated sequences included in them, these data repositories often contain a high number of false-positive sequences and require manual curation to identify such anomalies and where appropriate remove them from confirmed virus sequences. A scientific community-wide collaborative effort is required to standardise uncultivated virus data sharing. Some progress in this field has been made since the introduction of MIUViG criteria, specifically in the case of DNA viruses, however, certain measures such as genome completeness remain difficult to determine computationally for viruses that have multipartite genomes; e.g. RNA viruses have very short genomes and many DNA viruses have short genomes that are <10kb long. A more collaborative approach between different data repositories could help achieve the common goal of cataloguing virosphere and can potentially aim to unify the virus sequences representing the virus diversity scattered across multiple silos. As the unknown sequence data analysis stems from the metagenomic datasets, it is also highly dependent on the databases the contig sequences are searched against. For example, in additional analysis carried out for Modha et al. (2022), it was observed that though sequences were classified as “unknown" with respect to widely used nt and nr databases, approximately 28% of UCs were found to match at least one known uncultivated virus sequence in IMG/VR suggesting that around a third of unknowns are “known-unknowns”. This further emphasises the importance of the unification of uncultivated sequence resources that can potentially help find unknown sequences present in various data repositories that are similar to one another. Moreover, this result also provides further confirmation to the initial hypothesis that unknown sequence matter is likely to be associated with uncultivated viruses and represents the genome sequences originating from these viruses.

This research project initially endeavoured to discover, categorise and potentially classify unknowns and has highlighted the importance of looking for things that are not related to currently known sequences. By enabling the identification and categorisation of unknowns that are embedded in the publicly available datasets, this project has led to the identification of a myriad of sequences that could potentially be originating from the unknown uncultured organisms that surround us, interact with us and are yet to be discovered. The unbiased approach implemented to gain further insights into the biological unknown matter has effectively shown that our knowledge about the uncultured virus diversity is limited. With the advent of the field of computational biology research, truly no sequence can be labelled unknown. Unknown sequences can be assigned a denomination by the means of various taxonomic and/or functional features associated with them albeit with the help of the right types of models and algorithms.

Although the raw sequence data added to repositories such as ENA and SRA are growing at an exponential rate, it has to be appreciated that the data deposited in these repositories are often imperfect. For example, in the case of the 58 studies analysed in detail here, it was noted that the SRA metadata associated with these studies was inconsistent. As these repositories and data submission into these repositories rules are lenient to serve a broad scientific research community, the metadata associated with the raw data can often be incomplete and/or inconsistent. In some cases such as BioProjects PRJEB11554 and PRJEB14301, a limited amount of information was provided describing the study itself. Datasets submitted to the SRA and ENA repositories often contain customised data fields that can be redundant, uninformative and potentially misleading. It was notable that around 190 unique column names were captured among the 58 BioProjects analysed here. The mislabelling of datasets can also cause problems if the datasets are not handled with meticulous detail and utmost care. Although a range of new pseudo taxonomy groups has been created in NCBI to cope with and categorise various metagenomic studies and samples, the data and metadata quality submission responsibilities lie with the research community to ensure that raw data is submitted with appropriate relevant metadata which enables the reproducibility in science.

The sequence data examined here merely scratches the surface. There are more than 18 petabytes of publicly accessible raw sequence data available in the SRA. Analysing the whole of SRA with an aim of cataloguing the unknowns would be a very challenging but extremely helpful task. It would require relevant search techniques that are adaptable to the level of multi-terabytes of data. This could be achieved by taking advantage of SRA Taxonomy Analysis Tool (STAT) - a new MinHash k-mer-based method that calculates the taxonomic distribution of SRA reads from HTS runs submitted to SRA repositories (Katz et al., 2022). The STAT reports for each SRA run can be mined further to identify the SRA runs with a high proportion of unknown sequence reads to target the BioProjects with a large proportion of unknown reads embedded within them. This could potentially be a good starting point to scale this analysis to terabyte-level data. It is worth noting that SRA data repositories are being moved to cloud platforms such as Amazon Web Services (AWS) and Google Cloud Platform (GCP) for easier access for processing in the cloud. However, this also means that users who wish to download data on a local server available on their own premise will need to pay egress charges (https://www.ncbi.nlm.nih.gov/sra/docs/sra-cloud-access-costs/). It may be an expensive task both financially as well as computationally to have local copies of these large repositories. The current version of the UnXplore analysis pipeline requires the data to be locally accessible and has not been tested in the cloud compute environment. To scale up, an alternative cloud-based analysis model such as that developed by Serratus authors (Edgar et al., 2022) might need to be explored as a more viable option.

This thesis targeted the characterisation and cataloguing of genetic unknowns but it is appreciated that there may be other types of unknowns embedded within the datasets analysed here that were not pursued here. A systematic quantitative measure of genetic unknowns indicated that the integration of new taxa into a set of reference genomes in the last decade resulted in a significant reduction of genetic unknowns in human microbiomes. However, the percentage of the unknown is highly dependent on experimental factors such as bodily sites/microbiomes, sampled populations, as well as other analysis factors such as resources and databases used to compare and classify the assembled sequences. Although the average proportion of unknowns identified in human microbiome samples analysed in this thesis was around 2% which is lower than previous measures (Krishnamurthy et al., 2017; Thomas et al., 2019), this was shown to be highly divergent between microbiomes, BioProjects and samples. Furthermore, the definition of unknowns applied in this thesis could be considered more conservative compared to those used in similar studies. To label something as genetically unknown, the sequences were required to be phylogenetically significantly different to any known protein and nucleotide sequences available in the general-purpose databases. A criticism of this approach would be that it excluded other types of unknowns, such as novel strains, or species-level taxa since sequences originating from them are likely to have correlations with known sequences present in the databases, although at lower degrees of similarity. To measure such hidden unknowns, results obtained from the similarity sequence analysis part of the UnXplore framework could be investigated further. An educated hypothesis would be that such analysis could likely result in the characterisation of a number of novel microbial strains and species that are missed by the current analysis carried out here.

Human-associated virome catalogued here provides a comprehensive picture of specific virus families identified in a range of human microbiomes. The virome described here is largely similar to that included in a recent review (Liang et al., 2021) that compiled a list of viruses identified from healthy humans sampled in various virome surveys. Considering that our analysis did not exclusively use healthy human samples, it is likely that the discrepancy between our results and those reported in Liang et al. (2021) can be attributed to sampling type and human subject health status. Though this catalogue is not a complete snapshot of the human virome, as the human gut alone is believed to contain a staggering amount of viruses, it can nonetheless be utilised as a powerful resource and could be used to determine virus-host relationships. Moreover, the human virome has also been shown to be unique to individuals and is considered dynamic. As a result, it is difficult to distinguish between cases of vertical or horizontal virus transmission, or differences in composition, of the virus in individuals from different geographical locations, those consuming different diets, or those ageing or younger. Our current understanding of human viruses is highly skewed toward western populations. Though efforts are being made to achieve a more inclusive catalogue of the human virome, it can be argued that a complete picture of the human virome is yet to be painted. There need to be more concerted efforts on the part of the scientific community to decode the viral dark matter surrounding us, so that it can be used for better therapeutics and precision medicine in the future. Furthermore, a comprehensive survey of viruses beyond human populations is also as important as viruses with zoonotic potential can emerge, evolve and jump from other species to humans. It was reaffirmed to the greatest extent by the Coronavirus pandemic. An extensive survey of viruses could help predict, control, and possibly completely avoid future pandemics.

While many viruses are being identified in this era of metagenomics using computational virology, these probable viruses are yet to be grown in cell cultures. Viral taxa have historically been determined by virus isolation and cell culture studies, but today viral lineages are often referred to as sequences before successful culture can occur. As a result, at present, we do not know how many members of the virome are replicated. Most viruses found through metagenomics do not have host associations, so further research is needed to determine host associations. Changes in virome are increasingly associated with disease states, but the molecular basis and causality of many cases remain unclear. Human virology is a vast field, and we are beginning to understand it, laying the foundation for future research.

In this thesis, the sole focus was on the genetic unknowns, and the functional unknown landscape was not explored. Although some open reading frames were analysed in the context of unknown contigs of interests, a comprehensive functional unknown analysis was largely excluded. To overcome this, UnXplore can be applied in conjunction with a specialised unknown functional analysis framework such as AGNOSTOS (Vanni et al., 2022) to investigate the functional unknowns embedded within this dataset. To pursue this, results obtained in Chapter 5 could serve as a good starting point because a number of virus sequences have been shown to have protein-coding ORFs of unknown function when annotated using DRAM-v. However, it is worth noting that functional unknowns would not be limited to viruses/virome sequences and other microbial gene prediction and ORF analysis would also need to be considered.

The results obtained from the current UnXplore analysis could be extended to characterise and survey a broader microbial co-occurrence network. Such network analysis could in turn be used to investigate within and between microbiome interactions of the microbial community that surrounds us. Furthermore, to strengthen these findings, virus-host interactions could also be explored as viruses often mimic their host genetic properties, those robust signals could be harnessed to identify important links in the co-occurrence network. This could help better our understanding of the microbial interactions and it can be utilised and contribute to other fields of biology such as antimicrobial resistance and the application of phage therapy. In order to obtain the complete picture of microbial interactions taking place within and around humans, integrated approaches that incorporate metagenomics with meta-transcriptomics, metabolomics and metaproteomics could be employed and explored.

All datasets included in this analysis originated from Illumina short-read sequencing technology. The largest overhead associated with the short-read sequences is the genome reconstructions and assemblies. These challenges are being tackled with long-read sequencing. Although they are not as widely used as short-read technologies at present, long-read technologies have been applied to study and explore microbial communities in clinical, ecological, and epidemiological settings (Leggett et al., 2019; Van Goethem et al., 2021; Warwick-Dugdale et al., 2019; X. Jia et al., 2021; Moss et al., 2020; Warwick-Dugdale et al., 2019; Yahara et al., 2021; Zablocki et al., 2021; X. Deng et al., 2020). However, due to the higher sequencing error rates associated with the long-read sequencing, its applications in metagenomics has been limited. As long-read sequencing technologies improve, they would be deemed increasingly suitable for sequence analysis of environmental microbial communities. Rather than having to assemble microbial genomes from short reads, which can yield inaccurate and/or incomplete sequences, a complete genome can be sequenced either as a single read or from relatively fewer but much longer reads. Through constant acceleration and rapid optimisation, the long-read sequencing technique could easily become the next big thing that can provide unprecedented insights into microbes, particularly viruses. Because their genomes are relatively short, viruses can be sequenced in a single read per genome with the advent of long-read sequencing. Long-read metagenomic technology promises to transform viral metagenomics, and, because it enables whole-nucleotide phasing of polymorphisms, it could not only solve the assembly difficulties associated with viral genomes but could also provide valuable understanding into the evolutionary process of viruses. Hybrid approaches, which combine the strengths of both short and long-read technologies, can also help gather a more comprehensive view of the microbial community around us. Furthermore, enhanced long-read and hybrid sequencing technologies can also shed light on microbial dark matter that cannot be detected due to current technological limitations.

Due to the lack of standardised dark matter identification and annotation protocols, the proportion of microbial dark matter reported in literature varied greatly between studies. Here, I have tried to establish a general-purpose unknown sequence analysis framework, UnXplore, that can systematically quantify and characterise dark matter embedded within publicly accessible datasets. To classify the unknown sequences identified here, accurate microbial sequence prediction models were developed using a supervised machine learning approach. These models are implemented in a portable sequence prediction framework called TetraPredX. TetraPredX was applied to the unknown contigs found here and revealed that a majority of unknown contigs were viral in origin. Virus genomes are being discovered in record numbers with metagenomic sequencing. To mine the metagenomic samples analysed using UnXplore, customised and comprehensive virus discovery analyses were carried out. This analysis highlighted a large number of known and novel virus sequences present in human microbiome samples.

Through the use of computational approaches to search and analyse publicly available data, I have revealed various biological insights previously hidden from the scientific research community; by doing this, I have made a contribution to the field of unknown sequence research in an attempt to shed light on this microbial dark matter.

Appendix A

UnXplore Resources

A.1 Supplementary data

A.1.1 Supplementary tables

Table A.1: List of samples excluded from the metagenomic analysis.

BioProject	**Total number of samples**	**Samples excluded**	**Reason for exclusion**
PRJEB14383	31	1	raw sequence data unavailable (PE)
PRJEB14782	93	93	not assembled (SE)
PRJEB15057	56	56	not assembled (SE)
PRJEB15334	48	3	not assembled (SE)
PRJEB19090	38	1	not assembled (PE)
PRJEB19188	8	8	raw sequence data unavailable (PE)
PRJEB20595	4	4	not assembled (PE)
PRJEB21696	2	1	not assembled (SE)

Table A.2: Brief description of BioProjects included in the study.

Biome	BioProject	Description	Library Layout	Samples	Reference
Circulatory system	PRJEB21816	Aortite	PAIRED	1	Foulex et al., 2019
Circulatory system	PRJEB24753	Endocarditis due to Neisseria meningitidis	PAIRED	2	Choutko et al., 2019
Fecal	PRJEB10865	Study of the abundance of bacteria from human samples	PAIRED	2	
Fecal	PRJEB11554	NeoM	PAIRED	1	
Fecal	PRJEB12357	Impact of faecal microbiota transplantation on the intestinal microbiome in metabolic syndrome patients	PAIRED	100	S. S. Li et al., 2016
Fecal	PRJEB14935	Term and preterm shotgun samples	PAIRED	3	Alcon-Giner et al., 2017
Fecal	PRJEB15257	The antibiotic resistance potential of the preterm infant gut microbiome measured using shotgun metagenomics.	PAIRED	15	Rose et al., 2017
Fecal	PRJEB1775	Diagnostic Metagenomics: A Culture-Independent Approach to the Investigation of Bacterial Infections	PAIRED	53	Loman et al., 2013
Fecal	PRJEB17784	The fecal microbiota in L-DOPA naive PD patients	SINGLE	100	
Fecal	PRJEB18265	Estimation of variability in the gut microbiota resistome of the Russian citizens aimed at identification of pathways for transmission and spread of antibiotic resistance.	PAIRED	10	Olekhnovich et al., 2019
Fecal	PRJEB19090	Potential and active functions in the gut microbiota of a healthy human cohort	PAIRED	37	Tanca et al., 2017

Biome	BioProject	Description	Library Layout	Samples	Reference
Fecal	PRJEB19367	Analysis of stool samples from sickle cell disease patients and healthy controls	PAIRED	28	
Fecal	PRJEB21696	Metagenomics 1st 5 data	SINGLE	1	
Fecal	PRJEB23207	Metagenomic characterization of the human intestinal microbiota in faecal samples from STEC-infected patients	PAIRED	11	
Fecal	PRJEB5761	Gut microbiota in chronic kidney disease	PAIRED	81	
Fecal	PRJEB6092	Metagenome fecal microbiota- Illumina seq reads of 12 individuals at 2 timepoints	PAIRED	24	
Fecal	PRJEB6542	Gut microbial metabolism shifts towards a more toxic profile with supplementary iron in a kinetic model of the human large intestine	PAIRED	8	
Fecal	PRJEB7331	metagenomic analysis of human gut microbiome	PAIRED	24	
Fecal	PRJEB7949	The fecal microbiome was studied in a group of IBD suffers and compared to a control group's fecal microbiome	PAIRED	40	
Fecal	PRJEB8094	The initial state of the human gut microbiome determines its reshaping by antibiotics	PAIRED	100	Raymond et al., 2016
Fecal	PRJEB8201	Comparison of distal gut microbiota structure and function in US and Egyptian children	SINGLE	2	
Fecal	PRJNA43253	Human fecal microbiome	SINGLE	7	Turnbaugh et al., 2010
Human	PRJEB14301	CSF	SINGLE	1	
Human	PRJEB21827	A/B testing for colon model	PAIRED	12	

Biome	**BioProject**	**Description**	**Library Layout**	**Samples**	**Reference**
Human	PRJEB6045	metagenomics of medieval human remains from Sardinaia	PAIRED	1	
Lung	PRJEB7248	Metagenomics of TB-associated sputum	PAIRED	8	
Oral	PRJEB12831	A plaque on both your houses. Exploring the history of urbanisation and infectious diseases through the study of archaeological dental tartar	PAIRED	31	
Oral	PRJEB12998	Test file for Oralfungi project	PAIRED	1	
Oral	PRJEB15334	Radcliffe dental calculus	SINGLE	45	
Oral	PRJNA230363	Oral Microbiome	PAIRED	28	
Oral	PRJNA384402	oral metagenome Metagenome	PAIRED	17	
Pulmonary system	PRJEB20877	Detection of bacterial pathogens from broncho-alveolar lavage by next-generation sequencing	PAIRED	2	Leo et al., 2017
Saliva	PRJEB14383	Oral microbiome samples from the Philippines	PAIRED	30	Lassalle et al., 2018
Saliva	PRJNA264728	Gene expression anlayses of saliva-derived in vitro biofilms during carbohydrate fermentation and pH stress	PAIRED	8	Edlund et al., 2015
Saliva	PRJNA306560	Human oral saliva Metagenome	PAIRED	53	
Skin	PRJEB10133	These samples are selections from a larger cohort that were selected for the participation in the EBI metagenomics training in Sept. 2015	PAIRED	10	
Skin	PRJEB10295	Whole genome sequencing of metagenomes extracted from palms of two individuals	PAIRED	2	
Sputum	PRJEB10919	Total RNA-Seq on sputum samples from patients with active tuberculosis	SINGLE	23	Schnettger et al., 2017

Biome	BioProject	Description	Library Layout	Samples	Reference
Sputum	PRJEB14539	Metagenome and metatranscriptome profiling of moderate and severe COPD sputum in Taiwanese Han males	PAIRED	1	
Vagina	PRJEB21446	Metatranscriptome reveals the function shifts of the vaginal microbiome during the treatment of bacterial vaginosis	PAIRED	40	Z.-L. Deng et al., 2018

Table A.3: List of sample accession and the corresponding BioProject identifiers of samples excluded from the metagenomic analysis.

BioProject	**Run**
PRJEB14782	ERR1529686
PRJEB14782	ERR1529669
PRJEB14782	ERR1529668
PRJEB14782	ERR1529667
PRJEB14782	ERR1529666
PRJEB14782	ERR1529665
PRJEB14782	ERR1529664
PRJEB14782	ERR1529670
PRJEB14782	ERR1529663
PRJEB14782	ERR1529661
PRJEB14782	ERR1529660
PRJEB14782	ERR1529659
PRJEB14782	ERR1529658
PRJEB14782	ERR1529657
PRJEB14782	ERR1529656
PRJEB14782	ERR1529662
PRJEB14782	ERR1529671
PRJEB14782	ERR1529672
PRJEB14782	ERR1529673
PRJEB14782	ERR1529688
PRJEB14782	ERR1529687
PRJEB14782	ERR1529615
PRJEB14782	ERR1529685
PRJEB14782	ERR1529684
PRJEB14782	ERR1529683
PRJEB14782	ERR1529682
PRJEB14782	ERR1529681
PRJEB14782	ERR1529680
PRJEB14782	ERR1529679
PRJEB14782	ERR1529678
PRJEB14782	ERR1529677
PRJEB14782	ERR1529676
PRJEB14782	ERR1529675
PRJEB14782	ERR1529674
PRJEB14782	ERR1529655
PRJEB14782	ERR1529654

BioProject	Run
PRJEB14782	ERR1529653
PRJEB14782	ERR1529652
PRJEB14782	ERR1529632
PRJEB14782	ERR1529631
PRJEB14782	ERR1529630
PRJEB14782	ERR1529629
PRJEB14782	ERR1529628
PRJEB14782	ERR1529627
PRJEB14782	ERR1529626
PRJEB14782	ERR1529625
PRJEB14782	ERR1529624
PRJEB14782	ERR1529623
PRJEB14782	ERR1529622
PRJEB14782	ERR1529621
PRJEB14782	ERR1529620
PRJEB14782	ERR1529619
PRJEB14782	ERR1529618
PRJEB14782	ERR1529633
PRJEB14782	ERR1529689
PRJEB14782	ERR1529634
PRJEB14782	ERR1529636
PRJEB14782	ERR1529651
PRJEB14782	ERR1529650
PRJEB14782	ERR1529649
PRJEB14782	ERR1529648
PRJEB14782	ERR1529647
PRJEB14782	ERR1529646
PRJEB14782	ERR1529645
PRJEB14782	ERR1529644
PRJEB14782	ERR1529643
PRJEB14782	ERR1529642
PRJEB14782	ERR1529641
PRJEB14782	ERR1529640
PRJEB14782	ERR1529639
PRJEB14782	ERR1529638
PRJEB14782	ERR1529637
PRJEB14782	ERR1529635
PRJEB14782	ERR1529617
PRJEB14782	ERR1529690

BioProject	Run
PRJEB14782	ERR1529692
PRJEB14782	ERR1529613
PRJEB14782	ERR1529614
PRJEB14782	ERR1529691
PRJEB14782	ERR1529705
PRJEB14782	ERR1529704
PRJEB14782	ERR1529703
PRJEB14782	ERR1529702
PRJEB14782	ERR1529701
PRJEB14782	ERR1529616
PRJEB14782	ERR1529699
PRJEB14782	ERR1529698
PRJEB14782	ERR1529697
PRJEB14782	ERR1529696
PRJEB14782	ERR1529693
PRJEB14782	ERR1529694
PRJEB14782	ERR1529700
PRJEB14782	ERR1529695
PRJEB15057	ERR1558907
PRJEB15057	ERR1558906
PRJEB15057	ERR1558905
PRJEB15057	ERR1558904
PRJEB15057	ERR1558901
PRJEB15057	ERR1558902
PRJEB15057	ERR1558908
PRJEB15057	ERR1558900
PRJEB15057	ERR1558899
PRJEB15057	ERR1558903
PRJEB15057	ERR1558909
PRJEB15057	ERR1558912
PRJEB15057	ERR1558911
PRJEB15057	ERR1558898
PRJEB15057	ERR1558913
PRJEB15057	ERR1558914
PRJEB15057	ERR1558915
PRJEB15057	ERR1558916
PRJEB15057	ERR1558917
PRJEB15057	ERR1558918
PRJEB15057	ERR1558919

BioProject	**Run**
PRJEB15057	ERR1558920
PRJEB15057	ERR1558921
PRJEB15057	ERR1558922
PRJEB15057	ERR1558910
PRJEB15057	ERR1558897
PRJEB15057	ERR1558894
PRJEB15057	ERR1558895
PRJEB15057	ERR1558867
PRJEB15057	ERR1558868
PRJEB15057	ERR1558869
PRJEB15057	ERR1558870
PRJEB15057	ERR1558871
PRJEB15057	ERR1558872
PRJEB15057	ERR1558873
PRJEB15057	ERR1558874
PRJEB15057	ERR1558875
PRJEB15057	ERR1558876
PRJEB15057	ERR1558877
PRJEB15057	ERR1558878
PRJEB15057	ERR1558896
PRJEB15057	ERR1558879
PRJEB15057	ERR1558881
PRJEB15057	ERR1558882
PRJEB15057	ERR1558883
PRJEB15057	ERR1558884
PRJEB15057	ERR1558885
PRJEB15057	ERR1558886
PRJEB15057	ERR1558887
PRJEB15057	ERR1558888
PRJEB15057	ERR1558890
PRJEB15057	ERR1558891
PRJEB15057	ERR1558892
PRJEB15057	ERR1558893
PRJEB15057	ERR1558880
PRJEB15057	ERR1558889
PRJEB15334	ERR1611424
PRJEB15334	ERR1611401
PRJEB15334	ERR1611418
PRJEB19090	ERR1809127

BioProject	Run
PRJEB20595	ERR1951441
PRJEB20595	ERR1951440
PRJEB20595	ERR1951439
PRJEB20595	ERR1951438
PRJEB21696	ERR2028014

Table A.4: List of analysed BioProjects with associated metadata including the location, microbiome and number of samples included in each BioProject.

Country	BioProject	Biome	Count
Australia	PRJEB6092	Fecal	24
Canada	PRJEB8094	Fecal	100
China	PRJNA230363	Oral	28
China: Sichuan	PRJNA306560	Saliva	53
Egypt	PRJEB8201	Fecal	1
Egypt	PRJNA384402	Oral	17
Gambia	PRJEB7248	Lung	8
Germany	PRJEB1775	Fecal	53
Germany	PRJEB17784	Fecal	100
Germany	PRJEB21446	Vagina	40
Italy	PRJEB19090	Fecal	37
Italy	PRJEB23207	Fecal	7
Italy: Sardinia	PRJEB6045	Human	1
Netherlands	PRJEB10295	Skin	2
Netherlands	PRJEB12357	Fecal	100
Netherlands	PRJEB23207	Fecal	4
Netherlands	PRJEB6542	Fecal	8
Philippines	PRJEB14383	Saliva	30
Russia	PRJEB14301	Human	1
Russia	PRJEB18265	Fecal	10
South Africa	PRJEB10919	Sputum	19
South Korea	PRJEB21696	Fecal	1
Switzerland	PRJEB20877	Pulmonary system	2
Switzerland	PRJEB21816	Circulatory system	1
Switzerland	PRJEB24753	Circulatory system	2
Taiwan	PRJEB14539	Sputum	1
USA	PRJEB10865	Fecal	2
USA	PRJEB19367	Fecal	28

Country	BioProject	Biome	Count
USA	PRJEB8201	Fecal	1
United Kingdom	PRJEB10133	Skin	10
United Kingdom	PRJEB10919	Sputum	4
United Kingdom	PRJEB12831	Oral	31
United Kingdom	PRJEB14935	Fecal	3
United Kingdom	PRJEB15257	Fecal	15
United Kingdom	PRJEB15334	Oral	45
United Kingdom	PRJEB7331	Fecal	24
United Kingdom	PRJEB7949	Fecal	40
nan	PRJEB11554	Fecal	1
nan	PRJEB12998	Oral	1
nan	PRJEB21827	Human	12
nan	PRJEB5761	Fecal	81
nan	PRJNA43253	Fecal	7
not applicable	PRJNA264728	Saliva	8

Table A.5: An overview of open reading frames generated from unknown contigs that were at least >=1.5kb long.

Genetic code	Number of ORFs
0	26,280
1	34,234
2	25,282
3	53,621
4	68,092
5	65,949
6	69,533
9	49,234
10	41,608
11	39,440
12	31,435
13	41,608
14	79,610
15	33,285
16	33,285
21	49,234
22	18,960
23	19,948

A.1.2 Supplementary figures

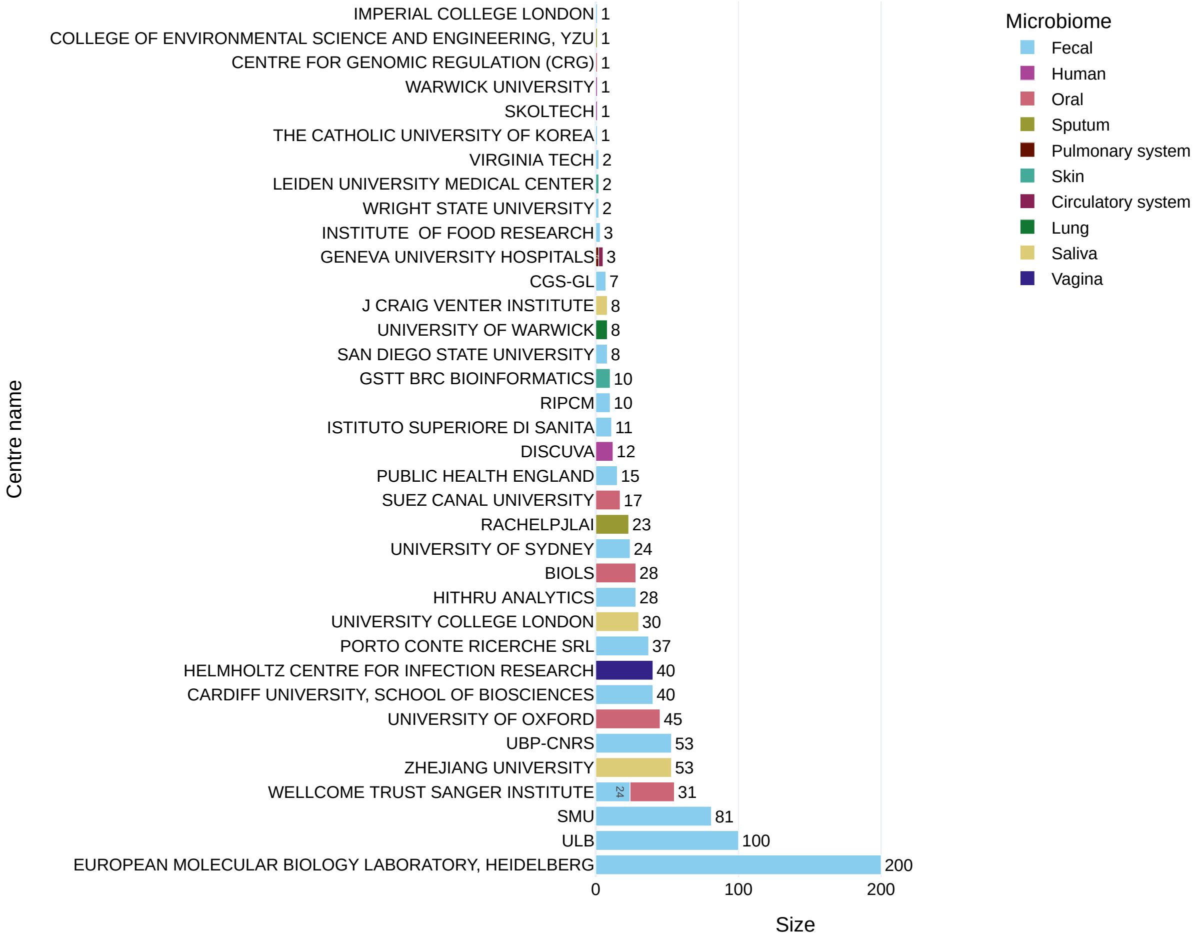


Figure A.1: Detailed information on the research centres associated with the BioProjects (n=963). A bar chart shows the research centre/organisation on the Y-axis and corresponding number of samples on the X-axis.

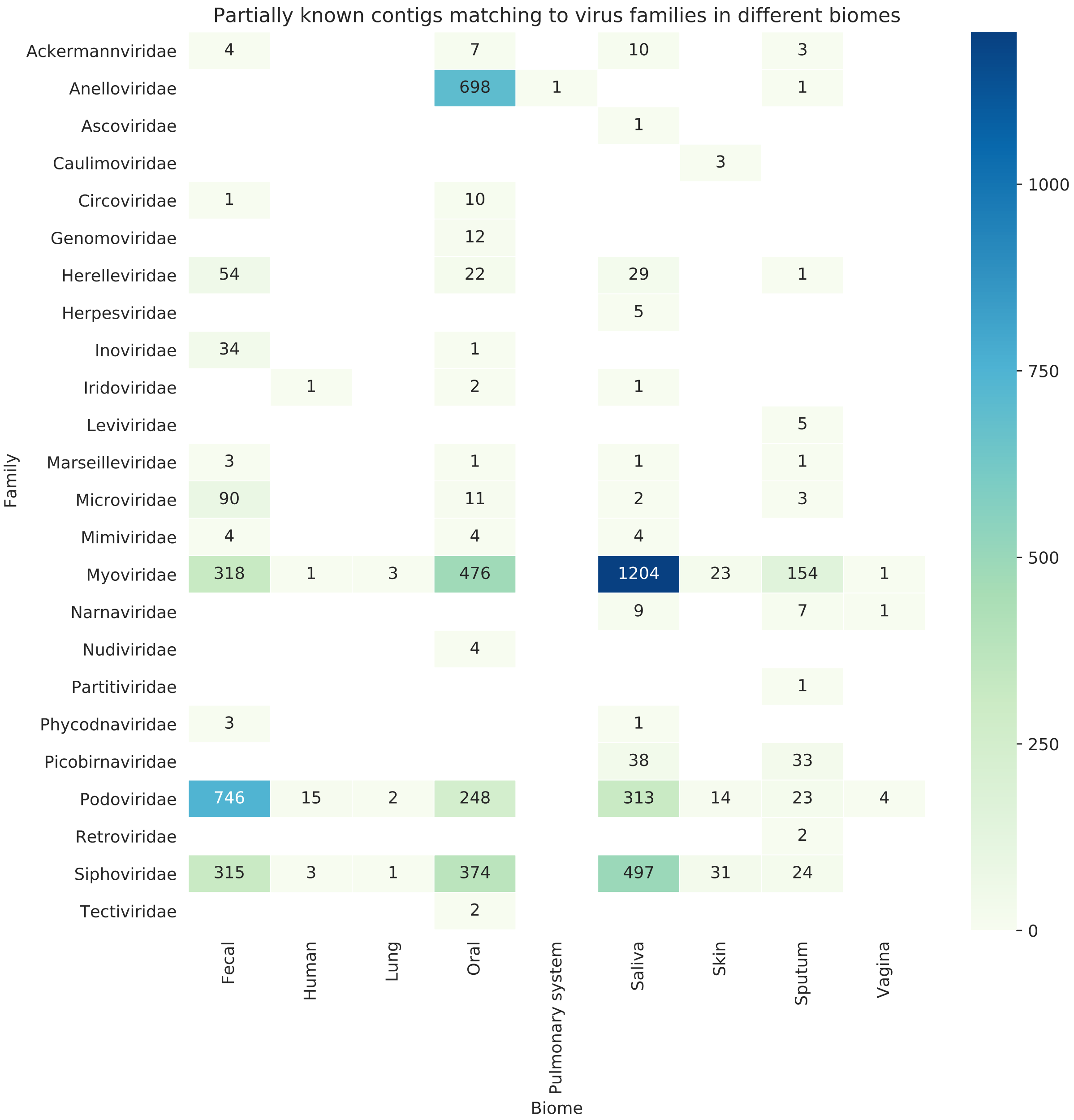


Figure A.2: A heatmap showing partially known viral contigs grouped according to the corresponding virus family (Y-axis) determined from the LCA hits and the microbiome they originate from which is shown on the X-axis. The heatmap is annotated with the number of contigs for each category.

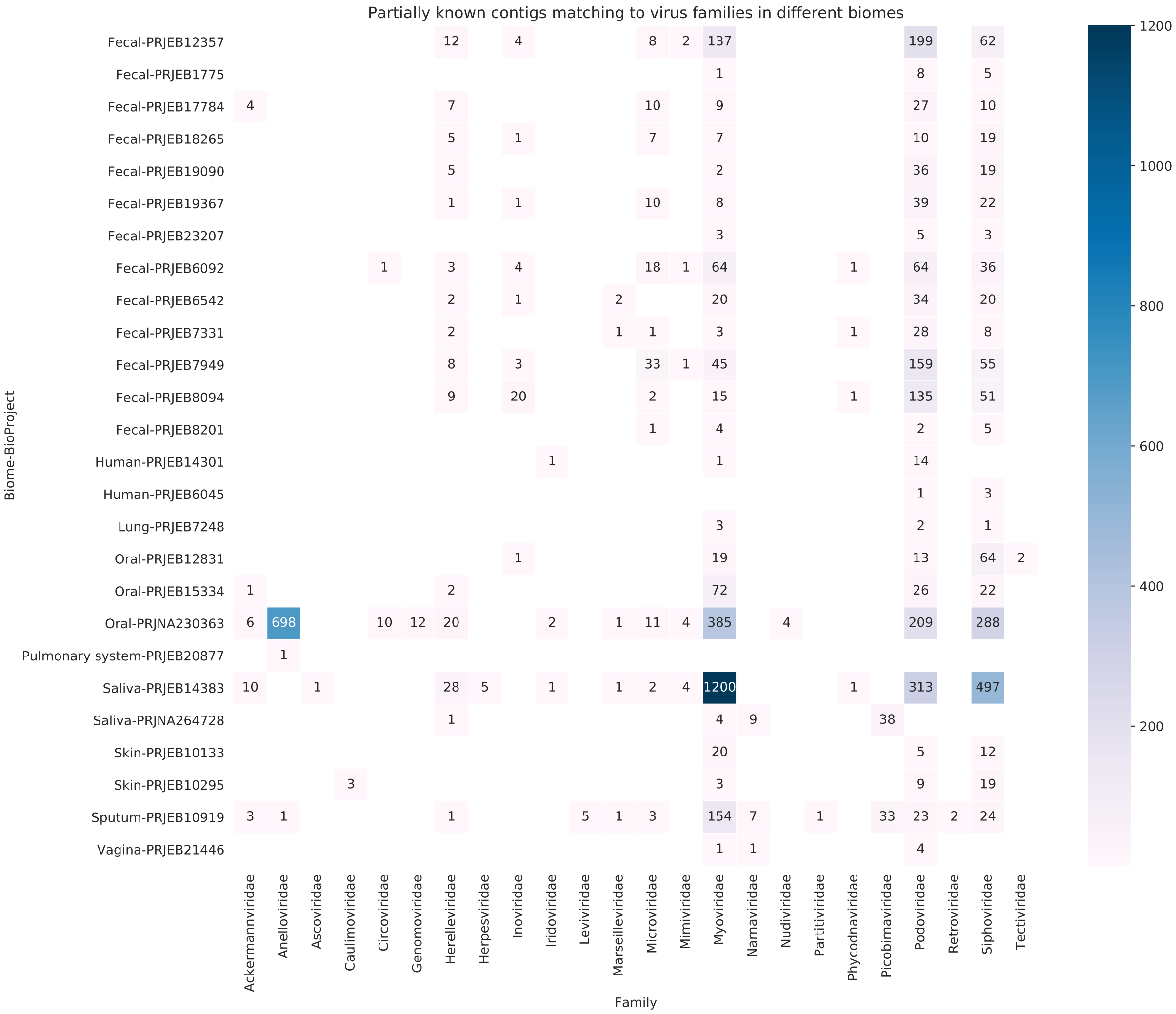


Figure A.3: A heatmap showing partially known viral contigs grouped according to the corresponding virus family (X-axis) determined from the LCA hits and the BioProject and microbiome they originate from which is shown on the Y-axis. The heatmap is annotated with the number of contigs for each category.

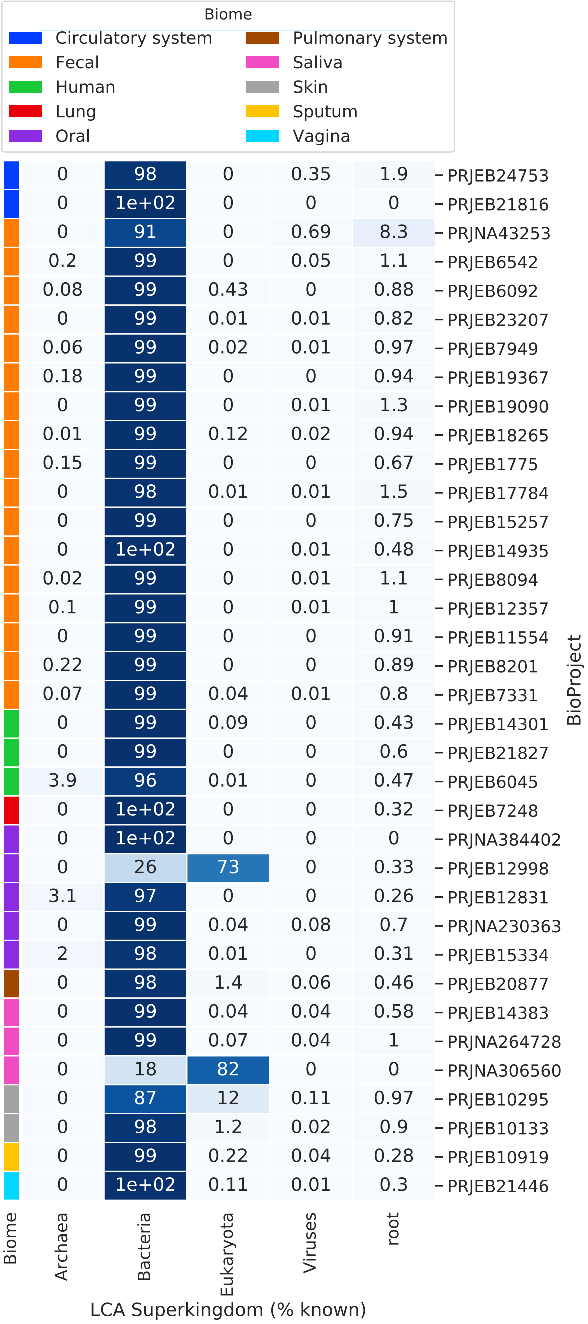


Figure A.4: Distribution of all known contigs across superkingdoms. The percentage of known contigs in each superkingdom is shown in the heatmap wtih superkingdoms specified on the X-axis and BioProjects shown on the Y-axis. BioProjects are grouped according to microbiomes are denoted in different colours specified in the colour legend at the top of the plot.

A.2 Tools and databases

Table A.6: Software and algorithms currently implemented in the metagenomic analysis pipeline

Software	Version	Reference	Source
BBDuk	38.22	https://sourceforge.net/projects/bbmap/	https://sourceforge.net/projects/bbmap/
BBNorm	38.22	https://sourceforge.net/projects/bbmap/	https://sourceforge.net/projects/bbmap/
BioPython	1.77		
BLASTN	2.9.0	Altschul et al., 1990	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
BWA	0.7.17-r1188	H. Li et al., 2009	http://bio-bwa.sourceforge.net/
DIAMOND	0.9.21.122	Buchfink et al., 2014	https://github.com/bbuchfink/diamond
ete3 (Python)			
InterProScan	5.38-76.0	Mitchell et al., 2018a	https://github.com/ebi-pf-team/interproscan
Parallel-fastq-dump	0.6.6		https://github.com/rvalieris/parallel-fastq-dump/
pysradb			
Python	3.6.7	https://www.python.org/download/releases/3.0/	https://www.python.org/downloads/
SAMTools	1.7		
Snakemake	5.4.5	Köster et al., 2012	https://bitbucket.org/snakemake/
SPAdes	3.11.1	Nurk et al., 2017	http://cab.spbu.ru/software/spades/

Table A.7: Reference genomes and databases used in the current analysis pipeline

Database/Genome	Version	Download Link	Notes
Non-redundant protein (nr)	114	https://ftp.ncbi.nlm.nih.gov/blast/db/v5/	nr databases downloaded in FASTA format and DIAMOND databasese were generated from that set
Nucleotide databases (nt)	71	https://ftp.ncbi.nlm.nih.gov/blast/db/v5/	BLAST v5 databases were used in this analysis
Human Genome	GRCh38	https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.38/	Path release: p12 Assembly identifier: GCF_000001405.38

Appendix B

TetraPredX Resources

B.1 Supplementary data

B.2 One-vs-Rest modelling approach

An alternative approach whereby each class is tested against the remaining classes was also explored. This One-vs-Rest (OVR) RFC model performance report is described in table B.2. The overall accuracy of this model was 0.97. The average F1-score was 0.96, 0.96 and 0.97 for archaea, bacteria and virus classes respectively. Notably, the recall for bacteria and archaea class was slightly lower compared to the virus class, which could be due to the lower number of observations and imbalanced datasets. Similar results were obtained for the gradient boost model. An OVR model was also trained with SVM classification which predicted each of the three classes with similar accuracy and F1-score. The predictions made by these models were very encouraging. However, a major limitation of these multiclass models is that a certain probability is assigned to each class and the sum of probabilities is always 1. This means that when a new dataset is interrogated with these models, each data point would be classified as either archaea, bacteria or virus. In the case of an unknown sequence prediction, these models would force each UC into one of these categories which can lead to an artificial classification.

Although this model performs really well with respect to the training and test data, it will force the classification of each contig into one of the classes included in the multiclass model, therefore independent binary prediction models were explored. To achieve this, multiple binary classification models were developed that predicted the given class as a positive label and if the model could not predict the positive class then the contig would be assigned a negative label. This breakdown of multiple classes into multiple binary classification problems was deemed more intuitive and suitable for the UC dataset.

B.3 Supplementary figures

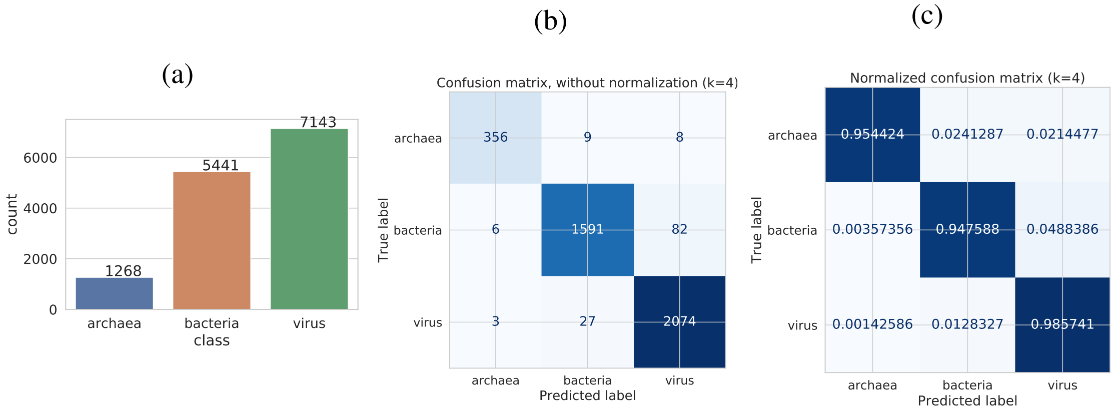


Figure B.1: (a) A bar chart showing May 2020 dataset. (b) Random forest one-vs-all (RFC OVR) performance confusion matrix whereby actual labels are plotted against the predicted labels with the number of observations in the test dataset. (c) A normalised confusion matrix of RFC OVR for all three classes.

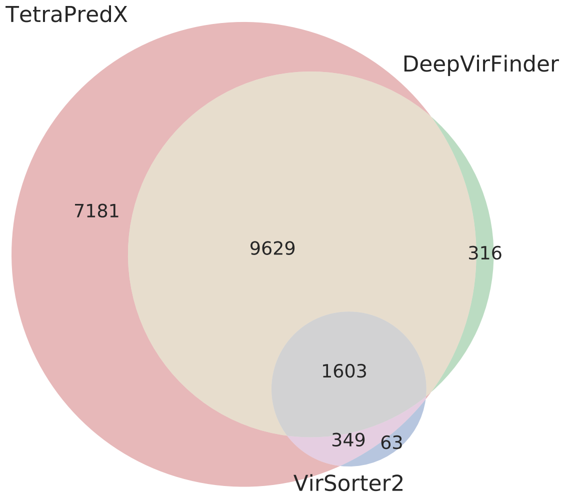


Figure B.2: Venn diagram showing the overlap among predicted virus sequences i.e. those with probability >0.5 for virus class from the unknown dataset using TetraPredX, VirSorter2 and DeepVirFinder.

B.3.1 Supplementary tables

Table B.1: Overview of machine learning model datasets

dataset	Version	Models	Notes
September 2019	Sept2019	Multiclass RFC	All RefSeq complete genomes • 3 majors classes • Bacteria and virus genomes downloaded on 26/09/2019 • Archaea genomes downloaded on 01/10/2019
May 2020	May2020	PCA, t-SNE, RFC OVR	RefSeq representative genomes • 3 major classes • Archaea reference genomes downloaded on 13/05/2020. Archaea genomes from nuccore downloaded on 19/05/2020 • Bacteria reference and representative genomes downloaded on 13/05/2020 • ICTV species exemplar virus genomes downloaded on 13/05/2020
January 2021	Jan2021	Binary RFC, SVM	RefSeq representative genomes • 4 major classes (bacteria, archaea, virus, plasmid) • Archaea genomes from nuccore downloaded on 14/01/2021 • Bacteria reference and representative genomes downloaded on 15/01/2021 • ICTV species exemplar genomes downloaded on 14/01/2021 • Reference plasmid genomes downloaded from https://doi.org/10.15146/R33X2J Plasmid sequences from bacterial genomes were also separated using ‘plasmid’ string in header

Table B.2: Classification report for multiclass multilabel RFC OVR model.

	precision	recall	F1-score	support
archaea	0.98	0.95	0.96	373.0
bacteria	0.98	0.95	0.96	1679.0
virus	0.96	0.99	0.97	2104.0
accuracy			0.97	4156.0
macro avg	0.97	0.96	0.97	4156.0
weighted avg	0.97	0.97	0.97	4156.0

Table B.3: Overview of additional holdout datasets

Dataset	Alias	Notes
Paraburkholderia madseniana strain RP11		284 contigs and scaffold sequences >=1kb
Geminiviridae		Holdout set containing 578 complete genome sequences from Jan2021 RefSeq virus set
Chuviridae		Holdout set containing 32 complete genome sequences/segments from Jan2021 RefSeq virus set
Siphoviridae		Holdout set containing 784 complete genome sequences from Jan2021 RefSeq virus set
Virus RefSeq Jan2021 chopped		RefSeq virus sequences chopped into 2kb non-overlapping fragments
Virus GenBank June 2021		All nucleotide sequences for viruses (txid10239). These sequences were filtered (>=1kb), clustered (MMSeqs2) and viruses from Jan2021 were removed from cluster representatives

Table B.4: Balanced, binary calibrated and cross-validated model performance metrics for models with reverse complement features

Class	N	AUC	F1-score	Accuracy	Brier score
Archaea	10319	1.00	1.00	99.78%	0.003
Bacteria	9814	0.99	0.97	96.97%	0.025
Plasmid	6642	0.99	0.94	94.60%	0.046
Virus	7953	1.00	0.97	97.64%	0.024

Table B.5: Virus specific model signal-based predictions (probability >0.5) for UCs that were predicted to be viruses (n=14,830).

Predicted class	**Number of UCs**
Realm	
Unpredicted	1208
Duplodnaviria	459
Duplodnaviria/Monodnaviria	41
Duplodnaviria/Riboviria	18
Duplodnaviria/Varidnaviria	30
Duplodnaviria/Varidnaviria/Monodnaviria	11
Monodnaviria	1496
Riboviria	6533
Riboviria/Monodnaviria	1391
Riboviria/Varidnaviria	2180
Riboviria/Varidnaviria/Monodnaviria	243
Varidnaviria	554
Varidnaviria/Monodnaviria	666
Genome type	
Unpredicted	1006
dsDNA	1471
dsDNA/dsRNA	1085
dsDNA/dsRNA/ssDNA	394
dsDNA/dsRNA/ssRNA	4
dsDNA/ssDNA	447
dsDNA/ssDNA/ssRNA	1
dsDNA/ssRNA	150
dsRNA	4270
dsRNA/ssDNA	2292
dsRNA/ssDNA/ssRNA	75
dsRNA/ssRNA	829
ssDNA	1222
ssDNA/ssRNA	137
ssRNA	1447
Segmentation	
Unpredicted	2370
Segmented	12460

Appendix C

Virus Discovery Resources

C.1 Supplementary tables

Table C.1: Metadata and publications associated with BioProjects analysed (n=58).

Microbiome	BioProject	Description	Country	Reference
Blood	ERP119596		USA	Poore et al., 2020; “Microbiome analyses of blood and tissues suggest cancer diagnostic approach”
Blood	ERP119597		USA	Poore et al., 2020; “Microbiome analyses of blood and tissues suggest cancer diagnostic approach”
Blood	ERP119598		USA	Poore et al., 2020; “Microbiome analyses of blood and tissues suggest cancer diagnostic approach”
Blood	PRJDB7117	Metagenomic analysis of biological samples collected from patients with Kawasaki disease.	Japan	
Blood	PRJDB7871	Whole genome analysis of Mycoplasma haemohominis identified from a patient with pyrexia	Japan	Hattori et al., 2020; “Candidatus Mycoplasma haemohominis in Human, Japan”
Blood	PRJNA253533	Human blood metagenome Genome sequencing - Mutation screening in retinitis pigmentosa		
Blood	PRJNA271229	Illumina sequencing directly from human serum samples derived from healthy individuals and patients with fevers of unknown origin. All samples are sourced from Nigeria in 2011.	Nigeria	Stremlau et al., 2015; “Discovery of Novel Rhabdoviruses in the Blood of Healthy Individuals from West Africa”
Blood	PRJNA292589	Human blood sample Raw sequence reads	USA	
Blood	PRJNA389455	Human blood metagenome of patients with acute liver failure of unknown etiology. Sequence analysis was performed to detect viral infections missed by conventional clinical testing. Blood from patients with known viral infections and non-infectious ALF were also sequenced as controls.		
Blood	PRJNA419524	Lung Transplant Recipient Lung and Blood Viral Microbiome	USA	
Blood	PRJNA471187	Viremia preceding multiple sclerosis - Metagenomic sequencing of serum from patients preceding multiple sclerosis diagnosis	Sweden	
Blood	PRJNA513310	Efficient and unbiased metagenomic recovery of RNA virus genomes from human plasma samples	United Kingdom	
Blood	PRJNA518922	Metagenomic sequencing of blood from patients suffering from a variety of vector-borne diseases including anaplasmosis babesiosis and mansonelliasis.		Vijayvargiya et al., 2019; “Application of metagenomic shotgun sequencing to detect vector-borne pathogens in clinical blood samples”
Blood	PRJNA544518	Bacterial microbiome of serum samples from leukemic and allogeneic stem cell transplant patients	USA	
Blood	PRJNA544865	Investigating Transfusion-Related Sepsis with Metagenomics	USA	Crawford et al., 2020; “Investigating Transfusion-related Sepsis Using Culture-Independent Metagenomic Sequencing”
Blood	PRJNA547963	Microbial 16S rRNA Gene in the Serum of Patients With Gastric Cancer	China	
Blood	PRJNA602694	Viral metagenomics in deferred blood donations with post-donation information from Brazil	Brazil	dos Santos Bezerra et al., 2021; “Viral metagenomics in blood donations with post-donation illness reports from Brazil”
Circulatory system	PRJEB21816	This project describe the first case of Listeria monocytogenes abdominal periaortitis associated to a vascular graft. Cultures of intraoperative samples were positive for Listeria monocytogenes. Results were further confirmed by a broad-range PCR and next-generation sequencing	Switzerland	Foulex et al., 2019; “Listeria monocytogenes infectious periaortitis: A case report from the infectious disease standpoint”
Circulatory system	PRJEB24753	A metagenomic analysis of a heart valve community acquired from a patient with blood culture negative infective endocarditis due to Neisseria meningitidis.	Switzerland	Choutko et al., 2019; “Rare Case of Community-Acquired Endocarditis Caused by Neisseria meningitidis Assessed by Clinical Metagenomics”

Microbiome	BioProject	Description	Country	Reference
Fecal	PRJEB10865	This study contains 12 samples for the analysis of abundance of bacterial communities on human residues farm derived samples	USA	
Fecal	PRJEB11554	NeoM		
Fecal	PRJEB12357	Impact of faecal microbiota transplantation on the intestinal microbiome in metabolic syndrome patients	Netherlands	S. S. Li et al., 2016; “Durable coexistence of donor and recipient strains after fecal microbiota transplantation”
Fecal	PRJEB14935	Term and preterm shotgun samples	United Kingdom	Alcon-Giner et al., 2017; “Optimisation of 16S rRNA gut microbiota profiling of extremely low birth weight infants”
Fecal	PRJEB15257	The antibiotic resistance potential of the preterm infant gut microbiome measured using shotgun metagenomics.	United Kingdom	Rose et al., 2017; “Antibiotic resistance potential of the healthy preterm infant gut microbiome”
Fecal	PRJEB1775	Diagnostic Metagenomics: A Culture-Independent Approach to the Investigation of Bacterial Infections	Germany	Loman et al., 2013; “A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of shiga-toxigenic Escherichia coli O104:H4”
Fecal	PRJEB17784	The fecal microbiota in L-DOPA naive PD patients	Germany	
Fecal	PRJEB18265	Estimation of variability in the gut microbiota resistome of the Russian citizens aimed at identification of pathways for transmission and spread of antibiotic resistance.	Russia	Olekhnovich et al., 2019; “Shifts in the human gut microbiota structure caused by quadruple *helicobacter pylori* eradication therapy”
Fecal	PRJEB19090	Potential and active functions in the gut microbiota of a healthy human cohort	Italy	Tanca et al., 2017; “Potential and active functions in the gut microbiota of a healthy human cohort”
Fecal	PRJEB19367	Analysis of stool samples from sickle cell disease patients and healthy controls	USA	
Fecal	PRJEB21696	Metagenomics 1st 5 data	South Korea	
Fecal	PRJEB23207	Metagenomic characterization of the human intestinal microbiota in faecal samples from STEC-infected patients	Italy; Netherlands	
Fecal	PRJEB5761	Gut microbiota in chronic kidney disease		
Fecal	PRJEB6092	Metagenome fecal microbiota- Illumina seq reads of 12 individuals at 2 timepoints	Australia	
Fecal	PRJEB6542	Gut microbial metabolism shifts towards a more toxic profile with supplementary iron in a kinetic model of the human large intestine	Netherlands	
Fecal	PRJEB7331	Metagenomic analysis of human gut microbiome	United Kingdom	
Fecal	PRJEB7949	The fecal microbiome was studied in a group of IBD suffers and compared to a control group’s fecal microbiome	United Kingdom	
Fecal	PRJEB8094	The initial state of the human gut microbiome determines its reshaping by antibiotics	Canada	Raymond et al., 2016; “The initial state of the human gut microbiome determines its reshaping by antibiotics”
Fecal	PRJEB8201	Comparison of distal gut microbiota structure and function in US and Egyptian children	Egypt; USA	
Human	PRJEB14301	CSF	Russia	
Human	PRJEB21827	A/B testing for colon model		
Human	PRJEB6045	Metagenomics of medieval human remains from Sardinaia	Italy	
Lung	PRJEB7248	Metagenomics of TB-associated sputum	Gambia	
Oral	PRJEB12831	A plaque on both your houses. Exploring the history of urbanisation and infectious diseases through the study of archaeological dental tartar	United Kingdom	
Oral	PRJEB12998	Test file for Oralfungi project		
Oral	PRJEB15334	Ancient dental calculus from skeletons from the Radcliffe Hospital burial ground	United Kingdom	
Oral	PRJNA230363	Oral Microbiome	China	J. Wang et al., 2016; “Phage-bacteria interaction network in human oral microbiome”

Microbiome	BioProject	Description	Country	Reference
Oral	PRJNA384402	Oral metagenome. Utilizing culture-independent molecular techniques to extend our knowledge on the breadth of bacterial diversity in the healthy human oral cavity in Egyptian individuals.	Egypt	
Pulmonary system	PRJEB20877	Detection of bacterial pathogens from broncho-alveolar lavage by next-generation sequencing	Switzerland	Leo et al., 2017; "Detection of Bacterial Pathogens from Broncho-Alveolar Lavage by Next-Generation Sequencing"
Saliva	PRJEB14383	Oral microbiome samples from the Philippines	Philippines	Lassalle et al., 2018; "Oral microbiomes from hunter-gatherers and traditional farmers reveal shifts in commensal balance and pathogen load linked to diet"
Saliva	PRJNA264728	Gene expression anlayses of saliva-derived in vitro biofilms during carbohydrate fermentation and pH stress		Edlund et al., 2015; "Meta-omics uncover temporal regulation of pathways across oral microbiome genera during in vitro sugar metabolism"
Saliva	PRJNA306560	Human oral saliva Metagenome	China	
Skin	PRJEB10133	These samples are selections from a larger cohort that were selected for the participation in the EBI metagenomics training in Sept. 2015	United Kingdom	
Skin	PRJEB10295	Whole genome sequencing of metagenomes extracted from palms of two individuals	Netherlands	
Sputum	PRJEB10919	Total RNA-Seq on sputum samples from patients with active tuberculosis	South Africa; United Kingdom	Schnettger et al., 2017; "A Rab20-Dependent Membrane Trafficking Pathway Controls M. tuberculosis Replication by Regulating Phagosome Spaciousness and Integrity"
Sputum	PRJEB14539	Metagenome and metatranscriptome profiling of moderate and severe COPD sputum in Taiwanese Han males	Taiwan	
Vagina	PRJEB21446	Metatranscriptome reveals the function shifts of the vaginal microbiome during the treatment of bacterial vaginosis	Germany	Z.-L. Deng et al., 2018; "Metatranscriptome Analysis of the Vaginal Microbiota Reveals Potential Mechanisms for Protection against Metronidazole in Bacterial Vaginosis"

Table C.2: Novel prokaryotic virus (complete genomes only) specific to SM dataset (n=321).

Viral Operating Taxonomy Unit (vOTU)	vOTU BioProject	vOTU Microbiome	Cluster size	Gene count	Potential AMG count	Viral hypothetical genes	Viral genes with viral benefits	Viral structure genes	Viral genes with host benefits	Viral replication genes	Viral genes with unknown function	Predicted host (genus)
ERR719882_NODE_74_length_57714_cov_15.423264	PRJEB8094	Fecal	28	94.0	1.0	77.0	10.0	5.0	1.0	0.0	1.0	Roseburia
ERR719902_NODE_14_length_43888_cov_6.237584	PRJEB8094	Fecal	17	70.0	0.0	51.0	12.0	1.0	2.0	1.0	3.0	
ERR719395_NODE_56_length_41380_cov_7.455293	PRJEB8094	Fecal	14	71.0	0.0	55.0	1.0	2.0	1.0	1.0	11.0	Faecalibacterium
ERR719387_NODE_54_length_43833_cov_6.971264	PRJEB8094	Fecal	12	55.0	0.0	38.0	11.0	0.0	0.0	0.0	6.0	
ERR1297780_NODE_36_length_43908_cov_4.344264	PRJEB12357	Fecal	10	46.0	0.0	37.0	1.0	0.0	2.0	1.0	5.0	Prevotella
ERR1744189_NODE_291_length_46440_cov_31.239905	PRJEB18265	Fecal	9	59.0	0.0	5.0	53.0	0.0	0.0	0.0	1.0	
ERR695616_NODE_121_length_44307_cov_13.069669	PRJEB7949	Fecal	8	80.0	1.0	68.0	1.0	1.0	2.0	0.0	8.0	
ERR719447_NODE_14_length_42699_cov_6.470875	PRJEB8094	Fecal	8	78.0	0.0	61.0	11.0	2.0	1.0	1.0	2.0	Butyricicoccus
SRR2037087_NODE_5_length_18753_cov_8.338111	PRJNA230363	Oral	8	28.0	0.0	9.0	6.0	1.0	1.0	0.0	11.0	Streptococcus
ERR695626_NODE_346_length_32111_cov_5.711380	PRJEB7949	Fecal	8	40.0	0.0	28.0	10.0	2.0	0.0	0.0	0.0	
ERR537006_NODE_447_length_40653_cov_88.953126	PRJEB6542	Fecal	7	48.0	0.0	5.0	28.0	1.0	6.0	1.0	7.0	Dialister
ERR719445_NODE_25_length_34762_cov_3.629037	PRJEB8094	Fecal	7	57.0	0.0	42.0	2.0	2.0	2.0	1.0	8.0	Oscillibacter
ERR1297769_NODE_62_length_67608_cov_4.175581	PRJEB12357	Fecal	6	92.0	0.0	6.0	5.0	17.0	1.0	0.0	63.0	Ruminococcus
ERR1297782_NODE_385_length_14391_cov_4.051897	PRJEB12357	Fecal	6	21.0	0.0	3.0	17.0	1.0	0.0	0.0	0.0	
SRR2037083_NODE_33_length_38830_cov_9.525106	PRJNA230363	Oral	6	49.0	1.0	38.0	1.0	2.0	0.0	0.0	8.0	
ERR537009_NODE_183_length_58007_cov_7.061810	PRJEB6542	Fecal	6	85.0	1.0	59.0	1.0	0.0	2.0	0.0	23.0	
ERR719934_NODE_18_length_85666_cov_20.951081	PRJEB8094	Fecal	6	110.0	1.0	90.0	8.0	6.0	0.0	3.0	3.0	
ERR537010_NODE_502_length_34437_cov_12.681287	PRJEB6542	Fecal	6	47.0	0.0	4.0	32.0	2.0	1.0	0.0	8.0	Dialister
SRR2037083_NODE_1_length_187518_cov_15.996853	PRJNA230363	Oral	6	204.0	1.0	179.0	9.0	1.0	0.0	2.0	13.0	Neisseria
SRR2037090_NODE_27_length_54970_cov_12.317509	PRJNA230363	Oral	6	84.0	0.0	32.0	2.0	43.0	1.0	2.0	4.0	Streptococcus
SRR2037084_NODE_2_length_98062_cov_15.335711	PRJNA230363	Oral	6	147.0	1.0	111.0	31.0	4.0	0.0	1.0	0.0	Prevotella
ERR695610_NODE_275_length_35404_cov_15.196271	PRJEB7949	Fecal	6	36.0	0.0	5.0	29.0	1.0	0.0	0.0	1.0	Faecalibacterium
ERR695601_NODE_42_length_146390_cov_19.134336	PRJEB7949	Fecal	6	175.0	0.0	136.0	29.0	8.0	0.0	0.0	2.0	Parabacteroides
SRR2037089_NODE_66_length_40530_cov_15.205287	PRJNA230363	Oral	6	64.0	0.0	36.0	12.0	6.0	2.0	0.0	8.0	Lachnoanaerobaculum
ERR537007_NODE_1990_length_11624_cov_48.453799	PRJEB6542	Fecal	6	16.0	0.0	13.0	0.0	0.0	0.0	0.0	3.0	
ERR537009_NODE_362_length_42241_cov_20.142109	PRJEB6542	Fecal	5	50.0	0.0	25.0	1.0	0.0	1.0	0.0	23.0	Bacteroides
ERR1297847_NODE_138_length_44773_cov_11.152064	PRJEB12357	Fecal	5	56.0	0.0	23.0	2.0	6.0	3.0	2.0	20.0	Ruminococcus
SRR2037088_NODE_6_length_34938_cov_46.321733	PRJNA230363	Oral	5	47.0	1.0	33.0	8.0	2.0	0.0	3.0	1.0	Actinomyces
ERR695636_NODE_151_length_57452_cov_8.836786	PRJEB7949	Fecal	5	66.0	0.0	36.0	16.0	3.0	4.0	2.0	5.0	Anaerostipes
SRR2037090_NODE_1_length_235026_cov_7.662026	PRJNA230363	Oral	5	214.0	2.0	173.0	29.0	1.0	0.0	0.0	11.0	Veillonella
SRR2037083_NODE_28_length_40314_cov_57.845078	PRJNA230363	Oral	5	57.0	0.0	26.0	25.0	1.0	3.0	0.0	2.0	Streptococcus
SRR2037083_NODE_13_length_49609_cov_68.045869	PRJNA230363	Oral	5	67.0	0.0	61.0	1.0	0.0	0.0	1.0	4.0	
SRR2037084_NODE_4_length_57284_cov_35.885285	PRJNA230363	Oral	5	95.0	0.0	36.0	49.0	2.0	3.0	2.0	3.0	Streptococcus
ERR719909_NODE_4_length_57910_cov_64.527975	PRJEB8094	Fecal	5	85.0	0.0	61.0	2.0	1.0	1.0	0.0	20.0	
SRR2037089_NODE_123_length_32627_cov_8.511421	PRJNA230363	Oral	5	44.0	0.0	31.0	7.0	1.0	2.0	2.0	1.0	
ERR537005_NODE_1464_length_18293_cov_19.692894	PRJEB6542	Fecal	4	22.0	0.0	9.0	5.0	1.0	0.0	0.0	7.0	Bifidobacterium
SRR2037085_NODE_24_length_36643_cov_11.529163	PRJNA230363	Oral	4	51.0	0.0	22.0	2.0	1.0	3.0	0.0	23.0	Streptococcus
ERR695637_NODE_80_length_86316_cov_9.261300	PRJEB7949	Fecal	4	101.0	2.0	81.0	12.0	2.0	3.0	0.0	3.0	Prevotella
ERR1474585_NODE_116_length_33529_cov_9.028798	PRJEB14383	Saliva	4	38.0	0.0	2.0	12.0	1.0	2.0	1.0	20.0	Corynebacterium
ERR537007_NODE_359_length_38471_cov_12.448277	PRJEB6542	Fecal	4	47.0	0.0	32.0	2.0	2.0	2.0	0.0	9.0	
ERR695607_NODE_1034_length_11961_cov_8.249874	PRJEB7949	Fecal	4	15.0	0.0	11.0	1.0	1.0	0.0	0.0	2.0	

Viral Operating Taxonomy Unit (vOTU)	vOTU BioProject	vOTU Microbiome	Cluster size	Gene count	Potential AMG count	Viral hypothetical genes	Viral genes with viral benefits	Viral structure genes	Viral genes with host benefits	Viral replication genes	Viral genes with unknown function	Predicted host (genus)
ERR537005_NODE_2347_length_12467_cov_55.597029	PRJEB6542	Fecal	4	17.0	0.0	13.0	0.0	2.0	0.0	0.0	2.0	Blautia
SRR2037089_NODE_76_length_38983_cov_5.180590	PRJNA230363	Oral	4	48.0	1.0	39.0	4.0	3.0	1.0	0.0	1.0	
ERR537010_NODE_420_length_37789_cov_21.032994	PRJEB6542	Fecal	4	49.0	0.0	44.0	2.0	0.0	2.0	0.0	1.0	
SRR2037083_NODE_30_length_39488_cov_31.367864	PRJNA230363	Oral	3	55.0	1.0	0.0	35.0	1.0	4.0	0.0	15.0	Rothia
ERR1474570_NODE_21_length_87482_cov_13.315612	PRJEB14383	Saliva	3	79.0	1.0	33.0	3.0	0.0	0.0	0.0	43.0	Prevotella
ERR1474565_NODE_2_length_90476_cov_16.924796	PRJEB14383	Saliva	3	91.0	0.0	70.0	20.0	0.0	0.0	0.0	1.0	
SRR2037085_NODE_33_length_32021_cov_20.060001	PRJNA230363	Oral	3	44.0	0.0	3.0	35.0	0.0	1.0	1.0	4.0	
SRR2037084_NODE_17_length_38779_cov_18.420566	PRJNA230363	Oral	3	58.0	1.0	5.0	44.0	3.0	1.0	0.0	5.0	
ERR634991_NODE_225_length_43775_cov_45.406747	PRJEB7331	Fecal	3	54.0	0.0	19.0	1.0	9.0	1.0	2.0	22.0	Clostridium
ERR695626_NODE_172_length_50725_cov_7.046300	PRJEB7949	Fecal	3	65.0	1.0	41.0	12.0	0.0	1.0	8.0	3.0	
SRR2037084_NODE_34_length_27844_cov_68.277880	PRJNA230363	Oral	3	29.0	0.0	22.0	2.0	2.0	2.0	1.0	0.0	
ERR695606_NODE_117_length_50784_cov_9.879950	PRJEB7949	Fecal	3	57.0	2.0	29.0	16.0	2.0	5.0	1.0	4.0	
SRR2037089_NODE_387_length_19098_cov_12.273539	PRJNA230363	Oral	3	29.0	0.0	20.0	6.0	2.0	0.0	1.0	0.0	Actinomyces
SRR2037084_NODE_30_length_29636_cov_54.770292	PRJNA230363	Oral	3	42.0	1.0	7.0	28.0	1.0	3.0	1.0	2.0	
ERR695631_NODE_210_length_45659_cov_11.465003	PRJEB7949	Fecal	3	62.0	0.0	7.0	52.0	0.0	0.0	0.0	3.0	Bacteroides
ERR695636_NODE_200_length_47760_cov_11.282172	PRJEB7949	Fecal	3	72.0	0.0	40.0	1.0	18.0	2.0	1.0	10.0	Megamonas
SRR2037083_NODE_51_length_32298_cov_37.294203	PRJNA230363	Oral	3	56.0	0.0	41.0	1.0	10.0	2.0	0.0	2.0	
SRR2037083_NODE_129_length_17229_cov_70.770001	PRJNA230363	Oral	3	24.0	0.0	15.0	2.0	1.0	0.0	0.0	6.0	
SRR2037089_NODE_69_length_39733_cov_14.279450	PRJNA230363	Oral	3	60.0	1.0	41.0	11.0	2.0	2.0	2.0	2.0	Sanguibacter
SRR2037085_NODE_2_length_60127_cov_66.623668	PRJNA230363	Oral	3	78.0	0.0	64.0	3.0	0.0	0.0	0.0	11.0	
ERR505084_NODE_817_length_36637_cov_48.700399	PRJEB6092	Fecal	3	58.0	1.0	50.0	1.0	6.0	0.0	0.0	1.0	
SRR2037085_NODE_3_length_58075_cov_57.380748	PRJNA230363	Oral	3	77.0	0.0	21.0	0.0	0.0	1.0	1.0	54.0	Streptococcus
ERR1474612_NODE_96_length_63046_cov_68.919401	PRJEB14383	Saliva	3	82.0	3.0	68.0	4.0	0.0	0.0	0.0	10.0	Streptococcus
ERR1474570_NODE_70_length_45433_cov_30.952246	PRJEB14383	Saliva	3	73.0	0.0	62.0	4.0	1.0	1.0	2.0	3.0	
ERR634990_NODE_402_length_50762_cov_15.636283	PRJEB7331	Fecal	2	81.0	1.0	52.0	16.0	1.0	3.0	6.0	3.0	Faecalibacterium
ERR537010_NODE_219_length_51339_cov_6.648896	PRJEB6542	Fecal	2	60.0	3.0	40.0	3.0	4.0	1.0	0.0	12.0	
SRR2037090_NODE_104_length_35541_cov_6.746435	PRJNA230363	Oral	2	36.0	1.0	28.0	4.0	2.0	1.0	0.0	1.0	Prevotella
ERR719931_NODE_83_length_50911_cov_5.815990	PRJEB8094	Fecal	2	79.0	0.0	64.0	10.0	2.0	2.0	0.0	1.0	Bacteroides
ERR695600_NODE_427_length_35744_cov_9.072431	PRJEB7949	Fecal	2	49.0	0.0	33.0	7.0	2.0	2.0	3.0	2.0	Ruminococcus
ERR1297808_NODE_94_length_42739_cov_2.836449	PRJEB12357	Fecal	2	51.0	0.0	26.0	5.0	1.0	2.0	1.0	16.0	Anaerostipes
SRR2037083_NODE_43_length_34719_cov_19.506289	PRJNA230363	Oral	2	45.0	0.0	1.0	37.0	0.0	1.0	1.0	5.0	Prevotella
ERR1823593_NODE_269_length_44245_cov_19.461643	PRJEB19367	Fecal	2	61.0	0.0	45.0	8.0	2.0	1.0	2.0	3.0	
ERR1474567_NODE_10_length_40996_cov_23.251826	PRJEB14383	Saliva	2	58.0	0.0	35.0	14.0	5.0	0.0	3.0	1.0	Neisseria
SRR2037084_NODE_22_length_34613_cov_8.349702	PRJNA230363	Oral	2	45.0	0.0	30.0	2.0	0.0	0.0	2.0	11.0	Streptococcus
ERR537005_NODE_454_length_41243_cov_8.178207	PRJEB6542	Fecal	2	66.0	0.0	34.0	23.0	1.0	2.0	3.0	3.0	Streptococcus
ERR1297756_NODE_110_length_42288_cov_18.610755	PRJEB12357	Fecal	2	56.0	0.0	34.0	2.0	3.0	0.0	0.0	17.0	Bacteroides
ERR1474580_NODE_61_length_35356_cov_7.205178	PRJEB14383	Saliva	2	46.0	0.0	42.0	2.0	0.0	0.0	0.0	2.0	
SRR2037083_NODE_21_length_43840_cov_16.988375	PRJNA230363	Oral	2	69.0	0.0	40.0	22.0	1.0	2.0	0.0	4.0	Neisseria
SRR2037083_NODE_26_length_41820_cov_10.871088	PRJNA230363	Oral	2	42.0	1.0	28.0	1.0	8.0	2.0	0.0	3.0	Veillonella
SRR2037083_NODE_71_length_25833_cov_59.198192	PRJNA230363	Oral	2	40.0	0.0	32.0	3.0	0.0	1.0	0.0	4.0	
ERR1474612_NODE_124_length_51800_cov_67.659677	PRJEB14383	Saliva	2	69.0	0.0	46.0	20.0	0.0	1.0	1.0	1.0	
ERR537009_NODE_190_length_57073_cov_35.054071	PRJEB6542	Fecal	2	89.0	0.0	71.0	2.0	11.0	1.0	0.0	4.0	Blautia
ERR1297770_NODE_755_length_12482_cov_7.405201	PRJEB12357	Fecal	2	15.0	0.0	10.0	1.0	2.0	1.0	1.0	0.0	Ruminococcus
ERR634977_NODE_2426_length_12979_cov_9.731662	PRJEB7331	Fecal	2	19.0	0.0	16.0	0.0	0.0	0.0	0.0	3.0	

Viral Operating Taxonomy Unit (vOTU)	vOTU BioProject	vOTU Microbiome	Cluster size	Gene count	Potential AMG count	Viral hypothetical genes	Viral genes with viral benefits	Viral structure genes	Viral genes with host benefits	Viral replication genes	Viral genes with unknown function	Predicted host (genus)
ERR1744189_NODE_1514_length_13281_cov_15.208453	PRJEB18265	Fecal	2	16.0	0.0	14.0	1.0	0.0	0.0	0.0	1.0	
ERR1474585_NODE_12_length_83028_cov_61.401444	PRJEB14383	Saliva	2	96.0	0.0	69.0	17.0	1.0	0.0	0.0	9.0	
ERR537007_NODE_375_length_37409_cov_29.732184	PRJEB6542	Fecal	2	52.0	0.0	34.0	3.0	10.0	1.0	0.0	4.0	Dialister
SRR2037083_NODE_32_length_38904_cov_38.451286	PRJNA230363	Oral	2	55.0	0.0	18.0	31.0	3.0	2.0	0.0	1.0	Streptococcus
ERR634978_NODE_338_length_36966_cov_65.320609	PRJEB7331	Fecal	2	52.0	0.0	44.0	1.0	7.0	0.0	0.0	0.0	
ERR1297754_NODE_82_length_64645_cov_4.072767	PRJEB12357	Fecal	2	83.0	0.0	72.0	8.0	0.0	0.0	0.0	3.0	Subdoligranulum
ERR1474586_NODE_37_length_64507_cov_11.234593	PRJEB14383	Saliva	2	92.0	0.0	77.0	13.0	1.0	1.0	0.0	0.0	Streptococcus
SRR2037090_NODE_80_length_39552_cov_8.694331	PRJNA230363	Oral	2	58.0	0.0	16.0	25.0	2.0	0.0	1.0	14.0	Neisseria
ERR1744186_NODE_74_length_63368_cov_9.474168	PRJEB18265	Fecal	2	84.0	1.0	59.0	15.0	3.0	0.0	1.0	6.0	Subdoligranulum
ERR634990_NODE_299_length_63341_cov_25.882075	PRJEB7331	Fecal	2	91.0	0.0	29.0	1.0	0.0	1.0	1.0	59.0	
SRR2037084_NODE_64_length_18367_cov_20.005898	PRJNA230363	Oral	2	25.0	0.0	1.0	2.0	3.0	0.0	3.0	16.0	Actinomyces
ERR634979_NODE_224_length_36326_cov_64.753991	PRJEB7331	Fecal	2	40.0	0.0	35.0	3.0	0.0	1.0	0.0	1.0	
SRR2037083_NODE_119_length_18088_cov_23.624355	PRJNA230363	Oral	2	24.0	0.0	15.0	3.0	5.0	1.0	0.0	0.0	
SRR2037083_NODE_65_length_27347_cov_72.796277	PRJNA230363	Oral	2	49.0	0.0	0.0	38.0	0.0	2.0	0.0	9.0	
ERR1297783_NODE_88_length_28318_cov_46.164101	PRJEB12357	Fecal	2	49.0	0.0	6.0	30.0	7.0	4.0	1.0	1.0	
ERR695628_NODE_235_length_42764_cov_51.350465	PRJEB7949	Fecal	2	64.0	0.0	41.0	8.0	0.0	0.0	0.0	15.0	Intestinibacter
ERR1474612_NODE_183_length_40470_cov_58.148979	PRJEB14383	Saliva	2	51.0	1.0	23.0	2.0	18.0	4.0	1.0	3.0	Listeria
SRR2037083_NODE_58_length_29321_cov_35.452812	PRJNA230363	Oral	2	40.0	0.0	2.0	5.0	2.0	0.0	1.0	30.0	
ERR1297794_NODE_85_length_55672_cov_5.451049	PRJEB12357	Fecal	2	60.0	4.0	31.0	16.0	5.0	5.0	2.0	1.0	
ERR695604_NODE_474_length_27960_cov_27.059631	PRJEB7949	Fecal	2	43.0	0.0	33.0	1.0	5.0	1.0	1.0	2.0	Bifidobacterium
SRR2034639_NODE_3_length_40558_cov_15.866306	PRJNA230363	Oral	2	55.0	0.0	40.0	7.0	3.0	0.0	3.0	2.0	
ERR1474577_NODE_8_length_36981_cov_8.327520	PRJEB14383	Saliva	1	51.0	0.0	18.0	23.0	2.0	1.0	2.0	5.0	Streptococcus
ERR1823606_NODE_274_length_37163_cov_4.706775	PRJEB19367	Fecal	1	55.0	0.0	42.0	5.0	1.0	0.0	4.0	3.0	Dialister
SRR2037083_NODE_37_length_37213_cov_10.510792	PRJNA230363	Oral	1	44.0	1.0	23.0	1.0	4.0	5.0	2.0	9.0	Dialister
SRR2034638_NODE_4_length_35370_cov_11.332210	PRJNA230363	Oral	1	59.0	1.0	50.0	2.0	0.0	3.0	0.0	4.0	
ERR695628_NODE_290_length_36945_cov_9.479019	PRJEB7949	Fecal	1	56.0	0.0	44.0	0.0	9.0	1.0	0.0	2.0	Bifidobacterium
ERR634980_NODE_527_length_35857_cov_14.551142	PRJEB7331	Fecal	1	52.0	1.0	38.0	7.0	5.0	1.0	0.0	1.0	Collinsella
SRR2037083_NODE_35_length_37959_cov_11.282609	PRJNA230363	Oral	1	66.0	0.0	5.0	26.0	1.0	1.0	0.0	33.0	Streptococcus
ERR2270961_NODE_98_length_38106_cov_6.405482	PRJEB23207	Fecal	1	53.0	0.0	19.0	3.0	4.0	0.0	2.0	25.0	Streptococcus
SRR2037085_NODE_29_length_34552_cov_21.255790	PRJNA230363	Oral	1	48.0	1.0	35.0	1.0	2.0	1.0	0.0	9.0	Veillonella
ERR1474608_NODE_155_length_35532_cov_22.818361	PRJEB14383	Saliva	1	51.0	1.0	16.0	23.0	4.0	1.0	1.0	6.0	Streptococcus
ERR505084_NODE_810_length_36899_cov_8.814434	PRJEB6092	Fecal	1	44.0	0.0	25.0	2.0	2.0	7.0	2.0	6.0	
ERR1474586_NODE_153_length_34726_cov_17.247873	PRJEB14383	Saliva	1	46.0	0.0	0.0	28.0	2.0	1.0	0.0	15.0	Actinomyces
ERR1474568_NODE_211_length_35968_cov_45.072787	PRJEB14383	Saliva	1	49.0	0.0	29.0	16.0	1.0	1.0	2.0	0.0	Schaalia
ERR1823595_NODE_199_length_36798_cov_16.885448	PRJEB19367	Fecal	1	53.0	0.0	37.0	4.0	1.0	0.0	0.0	11.0	Collinsella
ERR1474570_NODE_114_length_35473_cov_20.433367	PRJEB14383	Saliva	1	50.0	0.0	37.0	8.0	2.0	1.0	0.0	2.0	Schaalia
ERR695617_NODE_181_length_35459_cov_6.860976	PRJEB7949	Fecal	1	36.0	0.0	30.0	4.0	0.0	1.0	1.0	0.0	
ERR1474584_NODE_6_length_35435_cov_7.967326	PRJEB14383	Saliva	1	56.0	1.0	33.0	4.0	4.0	2.0	0.0	13.0	Veillonella
ERR2271043_NODE_212_length_35861_cov_31.127493	PRJEB23207	Fecal	1	48.0	0.0	27.0	15.0	2.0	1.0	0.0	3.0	Ruminococcus
ERR505106_NODE_348_length_35060_cov_11.196258	PRJEB6092	Fecal	1	57.0	0.0	47.0	1.0	0.0	0.0	0.0	9.0	Coprobacter
ERR1474571_NODE_11_length_35554_cov_8.241134	PRJEB14383	Saliva	1	47.0	0.0	21.0	3.0	17.0	2.0	3.0	1.0	Streptococcus
ERR1474564_NODE_13_length_35078_cov_6.308483	PRJEB14383	Saliva	1	54.0	0.0	15.0	27.0	3.0	1.0	3.0	5.0	Streptococcus
ERR1823609_NODE_130_length_36584_cov_26.298092	PRJEB19367	Fecal	1	57.0	0.0	39.0	10.0	0.0	2.0	0.0	6.0	Eubacterium
ERR1474570_NODE_106_length_36430_cov_12.143038	PRJEB14383	Saliva	1	56.0	0.0	43.0	1.0	10.0	1.0	0.0	1.0	Actinomyces

Viral Operating Taxonomy Unit (vOTU)	vOTU BioProject	vOTU Microbiome	Cluster size	Gene count	Potential AMG count	Viral hypothetical genes	Viral genes with viral benefits	Viral structure genes	Viral genes with host benefits	Viral replication genes	Viral genes with unknown function	Predicted host (genus)
ERR505085_NODE_838_length_36058_cov_24.284282	PRJEB6092	Fecal	1	59.0	0.0	6.0	52.0	1.0	0.0	0.0	0.0	Coprobacter
ERR1611403_NODE_2_length_277304_cov_5.350007	PRJEB15334	Oral	1	243.0	4.0	196.0	30.0	16.0	0.0	0.0	1.0	
ERR1474612_NODE_263_length_31734_cov_68.192904	PRJEB14383	Saliva	1	49.0	1.0	40.0	6.0	1.0	0.0	0.0	2.0	
ERR1474587_NODE_123_length_34469_cov_9.197129	PRJEB14383	Saliva	1	56.0	0.0	9.0	4.0	0.0	7.0	7.0	29.0	Rodentibacter
ERR1474612_NODE_329_length_26380_cov_28.810864	PRJEB14383	Saliva	1	41.0	0.0	34.0	3.0	0.0	1.0	0.0	3.0	
ERR1474608_NODE_324_length_22548_cov_54.161206	PRJEB14383	Saliva	1	26.0	0.0	22.0	0.0	0.0	2.0	0.0	2.0	
ERR1474568_NODE_586_length_19991_cov_16.738012	PRJEB14383	Saliva	1	23.0	0.0	20.0	1.0	1.0	0.0	0.0	1.0	
ERR1474570_NODE_379_length_19199_cov_11.287161	PRJEB14383	Saliva	1	21.0	0.0	12.0	2.0	1.0	0.0	0.0	6.0	Streptococcus
SRR2034639_NODE_22_length_19028_cov_6.316186	PRJNA230363	Oral	1	26.0	0.0	16.0	2.0	1.0	0.0	0.0	7.0	
SRR1044017_NODE_1_length_18999_cov_12.173195	PRJNA230363	Oral	1	24.0	0.0	14.0	2.0	5.0	1.0	0.0	2.0	
ERR1474582_NODE_60_length_18936_cov_4.408824	PRJEB14383	Saliva	1	27.0	0.0	9.0	1.0	5.0	2.0	0.0	10.0	Streptococcus
ERR1297780_NODE_132_length_18814_cov_2.260195	PRJEB12357	Fecal	1	29.0	1.0	0.0	27.0	0.0	0.0	0.0	2.0	Roseburia
ERR1474564_NODE_67_length_18607_cov_3.966563	PRJEB14383	Saliva	1	28.0	0.0	0.0	10.0	4.0	3.0	1.0	10.0	Streptococcus
ERR1823597_NODE_713_length_17854_cov_6.232316	PRJEB19367	Fecal	1	29.0	0.0	24.0	0.0	0.0	1.0	4.0	0.0	
SRR2037083_NODE_122_length_17846_cov_11.608229	PRJNA230363	Oral	1	24.0	0.0	6.0	10.0	1.0	0.0	1.0	6.0	Streptococcus
SRR1044006_NODE_225_length_17792_cov_21.447877	PRJNA230363	Oral	1	25.0	0.0	15.0	3.0	3.0	4.0	0.0	0.0	Actinomyces
ERR1474585_NODE_301_length_17779_cov_4.424171	PRJEB14383	Saliva	1	24.0	0.0	6.0	10.0	3.0	4.0	0.0	1.0	Streptococcus
ERR1474570_NODE_467_length_17454_cov_47.630956	PRJEB14383	Saliva	1	20.0	0.0	11.0	6.0	0.0	1.0	2.0	0.0	
ERR1474568_NODE_736_length_17208_cov_6.286597	PRJEB14383	Saliva	1	23.0	0.0	14.0	3.0	2.0	1.0	0.0	3.0	Actinomyces
ERR1474570_NODE_508_length_16703_cov_5.885632	PRJEB14383	Saliva	1	25.0	0.0	16.0	6.0	1.0	0.0	0.0	2.0	Mycobacteroides
ERR1474584_NODE_63_length_16069_cov_4.631447	PRJEB14383	Saliva	1	17.0	0.0	9.0	2.0	0.0	0.0	4.0	2.0	
ERR1744184_NODE_1262_length_15526_cov_17.757352	PRJEB18265	Fecal	1	22.0	0.0	16.0	0.0	3.0	0.0	0.0	3.0	
SRR1044034_NODE_2_length_15465_cov_9.481311	PRJNA230363	Oral	1	22.0	0.0	13.0	0.0	8.0	0.0	0.0	1.0	Corynebacterium
SRR1044035_NODE_48_length_14873_cov_34.197328	PRJNA230363	Oral	1	23.0	0.0	14.0	8.0	0.0	0.0	0.0	1.0	Corynebacterium
ERR1474568_NODE_915_length_14583_cov_14.688670	PRJEB14383	Saliva	1	21.0	0.0	11.0	9.0	0.0	0.0	0.0	1.0	Corynebacterium
SRR2037084_NODE_94_length_14182_cov_47.491046	PRJNA230363	Oral	1	31.0	0.0	29.0	2.0	0.0	0.0	0.0	0.0	Galactobacillus
ERR1744189_NODE_1487_length_13442_cov_24.399268	PRJEB18265	Fecal	1	19.0	0.0	0.0	15.0	1.0	0.0	0.0	3.0	CandidatusCibiobacter
SRR2037085_NODE_292_length_13087_cov_57.397406	PRJNA230363	Oral	1	6.0	0.0	4.0	0.0	0.0	1.0	0.0	1.0	
SRR2037084_NODE_112_length_12752_cov_11.140506	PRJNA230363	Oral	1	11.0	0.0	7.0	0.0	2.0	0.0	0.0	2.0	Prevotella
ERR1823587_NODE_1682_length_11265_cov_88.731757	PRJEB19367	Fecal	1	16.0	0.0	1.0	0.0	1.0	0.0	0.0	14.0	
ERR1823600_NODE_489_length_11170_cov_95.394922	PRJEB19367	Fecal	1	13.0	0.0	9.0	1.0	3.0	0.0	0.0	0.0	Bacteroides
SRR2037085_NODE_469_length_10511_cov_69.315800	PRJNA230363	Oral	1	7.0	0.0	0.0	4.0	0.0	0.0	1.0	2.0	
SRR1045095_NODE_35_length_26171_cov_32.536338	PRJNA230363	Oral	1	25.0	0.0	18.0	1.0	4.0	1.0	0.0	1.0	
ERR598794_NODE_19_length_26955_cov_11.350632	PRJEB7248	Lung	1	39.0	0.0	29.0	4.0	0.0	0.0	0.0	6.0	Streptococcus
ERR1474587_NODE_124_length_34278_cov_6.914473	PRJEB14383	Saliva	1	49.0	1.0	35.0	9.0	1.0	2.0	2.0	0.0	Veillonella
ERR1474564_NODE_22_length_29067_cov_11.276472	PRJEB14383	Saliva	1	30.0	0.0	24.0	2.0	1.0	0.0	0.0	3.0	
ERR1474571_NODE_15_length_34010_cov_8.087145	PRJEB14383	Saliva	1	42.0	0.0	3.0	23.0	1.0	0.0	0.0	15.0	Actinomyces
ERR2270961_NODE_111_length_33948_cov_10.064999	PRJEB23207	Fecal	1	46.0	1.0	33.0	10.0	0.0	1.0	1.0	1.0	Veillonella
ERR1474584_NODE_8_length_33932_cov_18.216578	PRJEB14383	Saliva	1	45.0	2.0	2.0	0.0	10.0	1.0	2.0	30.0	Veillonella
ERR1474580_NODE_67_length_33493_cov_8.884832	PRJEB14383	Saliva	1	53.0	1.0	38.0	8.0	2.0	1.0	0.0	4.0	
ERR1474580_NODE_68_length_33235_cov_7.793460	PRJEB14383	Saliva	1	49.0	1.0	40.0	4.0	1.0	0.0	1.0	3.0	
ERR2270941_NODE_170_length_33133_cov_7.603059	PRJEB23207	Fecal	1	53.0	0.0	2.0	40.0	7.0	1.0	2.0	1.0	Intestinimonas
ERR634986_NODE_720_length_33047_cov_20.139276	PRJEB7331	Fecal	1	51.0	0.0	2.0	34.0	2.0	0.0	0.0	13.0	Dorea
ERR1474585_NODE_120_length_33031_cov_13.579755	PRJEB14383	Saliva	1	43.0	0.0	35.0	5.0	0.0	3.0	0.0	0.0	

Viral Operating Taxonomy Unit (vOTU)	vOTU BioProject	vOTU Microbiome	Cluster size	Gene count	Potential AMG count	Viral hypothetical genes	Viral genes with viral benefits	Viral structure genes	Viral genes with host benefits	Viral replication genes	Viral genes with unknown function	Predicted host (genus)
ERR634974_NODE_881_length_32973_cov_3.537578	PRJEB7331	Fecal	1	50.0	0.0	39.0	0.0	7.0	1.0	0.0	3.0	Bifidobacterium
ERR1474612_NODE_251_length_32605_cov_67.573456	PRJEB14383	Saliva	1	44.0	0.0	39.0	4.0	1.0	0.0	0.0	0.0	
ERR1823601_NODE_280_length_32593_cov_20.240888	PRJEB19367	Fecal	1	41.0	0.0	29.0	6.0	2.0	1.0	1.0	2.0	Dialister
ERR1474586_NODE_174_length_32584_cov_20.300624	PRJEB14383	Saliva	1	51.0	0.0	22.0	3.0	1.0	1.0	1.0	23.0	Streptococcus
ERR634988_NODE_423_length_32524_cov_10.308048	PRJEB7331	Fecal	1	36.0	0.0	21.0	11.0	1.0	0.0	0.0	3.0	
ERR1474570_NODE_146_length_32053_cov_5.335927	PRJEB14383	Saliva	1	51.0	0.0	32.0	1.0	10.0	5.0	2.0	1.0	
SRR1045097_NODE_34_length_32053_cov_4.297394	PRJNA230363	Oral	1	37.0	0.0	32.0	3.0	0.0	2.0	0.0	0.0	
ERR634984_NODE_132_length_38147_cov_12.443951	PRJEB7331	Fecal	1	58.0	0.0	40.0	10.0	0.0	4.0	1.0	3.0	Clostridium
ERR1474587_NODE_138_length_31612_cov_6.892100	PRJEB14383	Saliva	1	53.0	0.0	34.0	12.0	2.0	2.0	0.0	3.0	Veillonella
ERR1474570_NODE_157_length_31429_cov_5.253554	PRJEB14383	Saliva	1	44.0	0.0	4.0	36.0	3.0	0.0	0.0	1.0	
ERR1474568_NODE_275_length_31404_cov_5.641424	PRJEB14383	Saliva	1	31.0	0.0	27.0	1.0	2.0	0.0	1.0	0.0	
ERR1474610_NODE_4_length_31034_cov_10.984893	PRJEB14383	Saliva	1	43.0	1.0	29.0	7.0	1.0	1.0	4.0	1.0	Pseudopropionibacterium
SRR2034640_NODE_14_length_30992_cov_41.188641	PRJNA230363	Oral	1	34.0	0.0	29.0	0.0	2.0	0.0	1.0	2.0	
SRR2037085_NODE_36_length_30693_cov_58.187839	PRJNA230363	Oral	1	42.0	0.0	33.0	2.0	1.0	1.0	0.0	5.0	Rothia
ERR1474612_NODE_275_length_30287_cov_23.678883	PRJEB14383	Saliva	1	44.0	0.0	36.0	3.0	0.0	3.0	1.0	1.0	
SRR1045100_NODE_1_length_30271_cov_4.744175	PRJNA230363	Oral	1	49.0	0.0	13.0	4.0	0.0	3.0	0.0	29.0	Streptococcus
ERR1474612_NODE_277_length_30173_cov_30.179760	PRJEB14383	Saliva	1	47.0	1.0	32.0	0.0	10.0	1.0	2.0	2.0	
ERR1474586_NODE_201_length_29630_cov_7.805342	PRJEB14383	Saliva	1	32.0	0.0	27.0	0.0	1.0	3.0	0.0	1.0	
ERR1297845_NODE_278_length_29159_cov_4.725914	PRJEB12357	Fecal	1	41.0	0.0	5.0	27.0	1.0	3.0	2.0	3.0	Streptococcus
ERR1823587_NODE_333_length_38136_cov_9.295817	PRJEB19367	Fecal	1	51.0	0.0	35.0	1.0	0.0	5.0	0.0	10.0	
ERR1744187_NODE_393_length_40628_cov_6.104552	PRJEB18265	Fecal	1	66.0	0.0	43.0	13.0	3.0	6.0	0.0	1.0	Oscillibacter
ERR537008_NODE_348_length_38231_cov_8.966654	PRJEB6542	Fecal	1	58.0	0.0	16.0	34.0	4.0	1.0	0.0	3.0	Lactobacillus
SRR1045096_NODE_3_length_55588_cov_6.711343	PRJNA230363	Oral	1	62.0	2.0	8.0	50.0	0.0	0.0	0.0	4.0	Cardiobacterium
ERR1474571_NODE_3_length_58990_cov_33.425774	PRJEB14383	Saliva	1	80.0	0.0	57.0	21.0	1.0	0.0	0.0	1.0	Streptococcus
ERR2271236_NODE_76_length_58762_cov_9.813566	PRJEB23207	Fecal	1	103.0	0.0	84.0	13.0	4.0	0.0	0.0	2.0	Faecalibacterium
ERR1474585_NODE_32_length_58414_cov_16.084871	PRJEB14383	Saliva	1	78.0	0.0	3.0	56.0	1.0	0.0	0.0	18.0	Streptococcus
ERR1474581_NODE_1_length_58306_cov_5.061716	PRJEB14383	Saliva	1	73.0	1.0	43.0	18.0	2.0	1.0	0.0	9.0	Neisseria
ERR1474607_NODE_1_length_57875_cov_10.728952	PRJEB14383	Saliva	1	79.0	0.0	59.0	18.0	1.0	0.0	0.0	1.0	Streptococcus
ERR1823610_NODE_90_length_57793_cov_23.986612	PRJEB19367	Fecal	1	97.0	0.0	79.0	5.0	0.0	1.0	1.0	11.0	Faecalibacterium
SRR2037085_NODE_4_length_57601_cov_19.184183	PRJNA230363	Oral	1	79.0	0.0	63.0	4.0	0.0	0.0	0.0	12.0	
ERR1474567_NODE_7_length_57425_cov_3.921980	PRJEB14383	Saliva	1	72.0	0.0	50.0	1.0	0.0	1.0	0.0	20.0	Streptococcus
ERR1474566_NODE_15_length_56787_cov_6.830219	PRJEB14383	Saliva	1	79.0	0.0	58.0	0.0	19.0	2.0	0.0	0.0	Streptococcus
ERR1474612_NODE_109_length_56772_cov_64.132377	PRJEB14383	Saliva	1	62.0	1.0	39.0	2.0	2.0	1.0	1.0	17.0	Actinomyces
ERR1611388_NODE_1_length_56197_cov_3.904905	PRJEB15334	Oral	1	78.0	0.0	20.0	56.0	1.0	0.0	0.0	1.0	Streptococcus
ERR634977_NODE_416_length_56066_cov_5.665780	PRJEB7331	Fecal	1	82.0	0.0	65.0	3.0	1.0	0.0	0.0	13.0	Odoribacter
ERR1474612_NODE_114_length_55600_cov_53.913314	PRJEB14383	Saliva	1	60.0	1.0	40.0	15.0	0.0	4.0	0.0	1.0	Rothia
ERR1823604_NODE_194_length_55034_cov_9.185744	PRJEB19367	Fecal	1	77.0	0.0	43.0	16.0	13.0	3.0	2.0	0.0	
ERR1474585_NODE_27_length_60893_cov_24.373582	PRJEB14383	Saliva	1	74.0	0.0	21.0	49.0	1.0	0.0	0.0	3.0	Streptococcus
SRR2037089_NODE_22_length_54860_cov_7.358015	PRJNA230363	Oral	1	66.0	1.0	2.0	1.0	10.0	1.0	0.0	52.0	Alloprevotella
ERR1474608_NODE_71_length_54564_cov_69.317544	PRJEB14383	Saliva	1	96.0	0.0	4.0	14.0	0.0	0.0	0.0	78.0	
ERR1474570_NODE_47_length_53857_cov_12.533344	PRJEB14383	Saliva	1	72.0	4.0	35.0	22.0	1.0	1.0	0.0	13.0	Streptococcus
ERR1474580_NODE_35_length_52873_cov_9.443315	PRJEB14383	Saliva	1	75.0	2.0	50.0	16.0	4.0	2.0	1.0	2.0	
ERR1474570_NODE_50_length_52215_cov_65.877301	PRJEB14383	Saliva	1	87.0	2.0	13.0	72.0	0.0	0.0	0.0	2.0	Actinomyces
ERR1474580_NODE_37_length_52122_cov_19.040813	PRJEB14383	Saliva	1	69.0	0.0	61.0	0.0	4.0	0.0	3.0	1.0	

Viral Operating Taxonomy Unit (vOTU)	vOTU BioProject	vOTU Microbiome	Cluster size	Gene count	Potential AMG count	Viral hypothetical genes	Viral genes with viral benefits	Viral structure genes	Viral genes with host benefits	Viral replication genes	Viral genes with unknown function	Predicted host (genus)
SRR1045093_NODE_12_length_51462_cov_60.859202	PRJNA230363	Oral	1	84.0	1.0	67.0	14.0	2.0	0.0	1.0	0.0	Actinomyces
ERR1474567_NODE_8_length_51400_cov_5.252839	PRJEB14383	Saliva	1	69.0	0.0	46.0	20.0	1.0	1.0	1.0	0.0	
ERR1474565_NODE_7_length_51085_cov_8.146639	PRJEB14383	Saliva	1	67.0	1.0	52.0	10.0	0.0	2.0	0.0	3.0	Rothia
ERR1474568_NODE_97_length_50983_cov_4.584708	PRJEB14383	Saliva	1	78.0	0.0	0.0	22.0	4.0	1.0	6.0	45.0	
ERR1823612_NODE_130_length_50247_cov_13.191684	PRJEB19367	Fecal	1	68.0	1.0	36.0	17.0	1.0	6.0	1.0	7.0	
ERR634973_NODE_591_length_50101_cov_12.843784	PRJEB7331	Fecal	1	71.0	2.0	45.0	4.0	0.0	0.0	0.0	22.0	Parabacteroides
ERR1474587_NODE_65_length_49718_cov_21.435193	PRJEB14383	Saliva	1	83.0	1.0	67.0	3.0	0.0	0.0	0.0	13.0	Actinomyces
ERR1823609_NODE_74_length_59404_cov_88.581307	PRJEB19367	Fecal	1	84.0	1.0	57.0	17.0	3.0	3.0	3.0	1.0	Bacteroides
ERR1744188_NODE_133_length_61366_cov_9.902399	PRJEB18265	Fecal	1	122.0	3.0	91.0	5.0	20.0	2.0	1.0	3.0	Phascolarctobacterium
ERR505092_NODE_140_length_49522_cov_67.845594	PRJEB6092	Fecal	1	67.0	1.0	55.0	0.0	8.0	2.0	1.0	1.0	Succinatimonas
ERR1474567_NODE_4_length_79009_cov_8.879335	PRJEB14383	Saliva	1	120.0	5.0	96.0	8.0	13.0	0.0	0.0	3.0	Bacteroides
ERR1474586_NODE_1_length_260795_cov_9.820488	PRJEB14383	Saliva	1	237.0	2.0	196.0	28.0	0.0	0.0	0.0	13.0	Veillonella
SRR1044035_NODE_1_length_254250_cov_6.379114	PRJNA230363	Oral	1	231.0	0.0	172.0	11.0	0.0	2.0	0.0	46.0	
ERR1474584_NODE_1_length_239549_cov_43.101861	PRJEB14383	Saliva	1	225.0	3.0	147.0	63.0	6.0	0.0	0.0	9.0	Neisseria
ERR1474587_NODE_1_length_236676_cov_7.196208	PRJEB14383	Saliva	1	225.0	1.0	180.0	31.0	5.0	1.0	1.0	7.0	Veillonella
ERR1474585_NODE_3_length_153209_cov_46.169352	PRJEB14383	Saliva	1	199.0	5.0	155.0	11.0	28.0	2.0	0.0	3.0	
ERR1611403_NODE_21_length_133800_cov_4.738040	PRJEB15334	Oral	1	166.0	2.0	126.0	30.0	6.0	0.0	0.0	4.0	Listeria
ERR1297849_NODE_18_length_101710_cov_4.406935	PRJEB12357	Fecal	1	113.0	0.0	84.0	12.0	1.0	1.0	1.0	14.0	Parabacteroides
ERR1474586_NODE_17_length_98573_cov_30.651191	PRJEB14383	Saliva	1	167.0	0.0	128.0	34.0	0.0	1.0	0.0	4.0	Prevotella
ERR1474612_NODE_46_length_95109_cov_61.483904	PRJEB14383	Saliva	1	82.0	2.0	6.0	27.0	0.0	1.0	0.0	48.0	Prevotella
ERR1474571_NODE_1_length_94733_cov_18.727265	PRJEB14383	Saliva	1	100.0	1.0	60.0	3.0	0.0	1.0	0.0	36.0	
ERR1474575_NODE_1_length_84450_cov_11.011103	PRJEB14383	Saliva	1	81.0	0.0	36.0	2.0	1.0	1.0	1.0	40.0	
ERR1474612_NODE_59_length_81144_cov_67.834527	PRJEB14383	Saliva	1	88.0	2.0	2.0	6.0	1.0	2.0	1.0	76.0	Prevotella
ERR1823591_NODE_46_length_80574_cov_10.589848	PRJEB19367	Fecal	1	129.0	2.0	7.0	112.0	7.0	0.0	2.0	1.0	
ERR1474565_NODE_3_length_78971_cov_6.562814	PRJEB14383	Saliva	1	85.0	0.0	58.0	10.0	0.0	1.0	1.0	15.0	Aggregatibacter
ERR1823610_NODE_82_length_61509_cov_21.510528	PRJEB19367	Fecal	1	88.0	0.0	63.0	18.0	2.0	1.0	0.0	4.0	Bacteroides
ERR1611403_NODE_47_length_78480_cov_5.218473	PRJEB15334	Oral	1	109.0	1.0	90.0	12.0	5.0	0.0	0.0	2.0	Streptococcus
ERR2270960_NODE_71_length_73896_cov_8.426755	PRJEB23207	Fecal	1	111.0	1.0	33.0	7.0	0.0	0.0	2.0	69.0	
ERR1474568_NODE_48_length_69644_cov_12.655190	PRJEB14383	Saliva	1	107.0	2.0	91.0	6.0	5.0	2.0	0.0	3.0	Veillonella
ERR1474612_NODE_83_length_69558_cov_69.476656	PRJEB14383	Saliva	1	82.0	0.0	50.0	9.0	0.0	0.0	1.0	22.0	
ERR1474570_NODE_32_length_67740_cov_44.524651	PRJEB14383	Saliva	1	98.0	1.0	85.0	8.0	0.0	0.0	0.0	5.0	Veillonella
SRR1045095_NODE_4_length_67184_cov_12.706490	PRJNA230363	Oral	1	80.0	1.0	13.0	60.0	2.0	0.0	1.0	4.0	
ERR1474608_NODE_54_length_66673_cov_67.068540	PRJEB14383	Saliva	1	87.0	1.0	76.0	5.0	4.0	0.0	1.0	1.0	Veillonella
ERR1474580_NODE_20_length_66417_cov_38.834032	PRJEB14383	Saliva	1	95.0	1.0	80.0	5.0	1.0	1.0	0.0	8.0	
ERR1474577_NODE_3_length_65452_cov_9.690919	PRJEB14383	Saliva	1	89.0	1.0	78.0	5.0	4.0	2.0	0.0	0.0	
ERR2271237_NODE_18_length_64169_cov_14.516190	PRJEB23207	Fecal	1	89.0	0.0	59.0	27.0	1.0	0.0	0.0	2.0	
SRR2037085_NODE_1_length_63862_cov_65.534487	PRJNA230363	Oral	1	77.0	0.0	63.0	0.0	11.0	0.0	0.0	3.0	
ERR1474582_NODE_10_length_63844_cov_9.812334	PRJEB14383	Saliva	1	81.0	0.0	54.0	2.0	2.0	1.0	0.0	22.0	Streptococcus
ERR1474571_NODE_2_length_61943_cov_7.462335	PRJEB14383	Saliva	1	77.0	0.0	54.0	3.0	0.0	1.0	0.0	19.0	Streptococcus
SRR2034637_NODE_1_length_49647_cov_11.096770	PRJNA230363	Oral	1	61.0	1.0	48.0	8.0	1.0	0.0	2.0	2.0	Rothia
SRR1045096_NODE_6_length_48978_cov_6.404166	PRJNA230363	Oral	1	49.0	0.0	33.0	6.0	1.0	6.0	0.0	3.0	
ERR1474570_NODE_95_length_38465_cov_8.902109	PRJEB14383	Saliva	1	59.0	0.0	2.0	55.0	0.0	0.0	0.0	2.0	
ERR505088_NODE_451_length_39833_cov_24.038011	PRJEB6092	Fecal	1	76.0	0.0	60.0	9.0	3.0	1.0	1.0	2.0	Clostridium
SRR2037085_NODE_17_length_41192_cov_7.797700	PRJNA230363	Oral	1	42.0	0.0	26.0	4.0	11.0	1.0	0.0	0.0	Neisseria

Viral Operating Taxonomy Unit (vOTU)	vOTU BioProject	vOTU Microbiome	Cluster size	Gene count	Potential AMG count	Viral hypothetical genes	Viral genes with viral benefits	Viral structure genes	Viral genes with host benefits	Viral replication genes	Viral genes with unknown function	Predicted host (genus)
ERR1474608_NODE_126_length_41132_cov_33.187453	PRJEB14383	Saliva	1	58.0	0.0	49.0	3.0	5.0	1.0	0.0	0.0	
ERR1474607_NODE_2_length_41127_cov_7.708974	PRJEB14383	Saliva	1	55.0	1.0	38.0	4.0	2.0	0.0	0.0	11.0	Veillonella
ERR1549321_NODE_133_length_40982_cov_56.574389	PRJEB14935	Fecal	1	46.0	0.0	27.0	8.0	1.0	1.0	1.0	8.0	Veillonella
ERR1823613_NODE_236_length_40969_cov_22.644669	PRJEB19367	Fecal	1	49.0	0.0	28.0	4.0	0.0	2.0	4.0	11.0	
SRR2034638_NODE_2_length_40950_cov_19.270131	PRJNA230363	Oral	1	54.0	0.0	29.0	2.0	3.0	0.0	0.0	20.0	
ERR1474587_NODE_94_length_40747_cov_30.327583	PRJEB14383	Saliva	1	60.0	0.0	27.0	18.0	6.0	1.0	2.0	6.0	Haemophilus
SRR2037083_NODE_27_length_40667_cov_11.813109	PRJNA230363	Oral	1	57.0	0.0	48.0	4.0	2.0	0.0	0.0	3.0	
ERR1474612_NODE_3_length_268263_cov_22.763814	PRJEB14383	Saliva	1	249.0	2.0	201.0	11.0	1.0	1.0	2.0	33.0	
ERR1474585_NODE_86_length_40484_cov_8.694378	PRJEB14383	Saliva	1	56.0	0.0	52.0	3.0	1.0	0.0	0.0	0.0	
ERR1681524_NODE_4_length_40221_cov_75.572823	PRJEB12831	Oral	1	66.0	0.0	55.0	7.0	2.0	2.0	0.0	0.0	
ERR1474608_NODE_131_length_39931_cov_36.269686	PRJEB14383	Saliva	1	59.0	0.0	2.0	27.0	4.0	2.0	2.0	22.0	Streptococcus
ERR1474585_NODE_88_length_39888_cov_55.501017	PRJEB14383	Saliva	1	56.0	1.0	19.0	28.0	4.0	2.0	0.0	3.0	Streptococcus
SRR1045097_NODE_26_length_39789_cov_41.662732	PRJNA230363	Oral	1	52.0	0.0	37.0	7.0	4.0	1.0	0.0	3.0	
ERR1474608_NODE_125_length_41403_cov_23.946890	PRJEB14383	Saliva	1	54.0	1.0	37.0	12.0	1.0	2.0	0.0	2.0	Veillonella
ERR1611403_NODE_163_length_39662_cov_2.826812	PRJEB15334	Oral	1	54.0	0.0	41.0	6.0	1.0	4.0	0.0	2.0	Actinomyces
ERR1474612_NODE_188_length_39572_cov_61.384493	PRJEB14383	Saliva	1	70.0	0.0	50.0	1.0	1.0	3.0	3.0	12.0	Actinomyces
ERR1474612_NODE_192_length_39134_cov_20.179994	PRJEB14383	Saliva	1	49.0	1.0	40.0	2.0	5.0	0.0	2.0	0.0	Ruminococcus
ERR1474586_NODE_122_length_39113_cov_18.669415	PRJEB14383	Saliva	1	60.0	0.0	54.0	0.0	5.0	0.0	0.0	1.0	
ERR1474568_NODE_177_length_39060_cov_32.231919	PRJEB14383	Saliva	1	54.0	0.0	17.0	4.0	1.0	2.0	1.0	29.0	Streptococcus
ERR726369_NODE_203_length_39045_cov_3.839534	PRJEB8201	Fecal	1	53.0	0.0	39.0	0.0	9.0	3.0	0.0	2.0	Ruminococcus
ERR1474585_NODE_94_length_39012_cov_9.980132	PRJEB14383	Saliva	1	51.0	0.0	21.0	2.0	3.0	2.0	0.0	23.0	Streptococcus
ERR1474568_NODE_178_length_38992_cov_6.434651	PRJEB14383	Saliva	1	45.0	0.0	10.0	3.0	2.0	2.0	3.0	25.0	Veillonella
ERR1474571_NODE_9_length_38939_cov_18.541688	PRJEB14383	Saliva	1	61.0	0.0	46.0	7.0	2.0	0.0	3.0	3.0	
ERR634981_NODE_505_length_38785_cov_8.443919	PRJEB7331	Fecal	1	53.0	0.0	31.0	2.0	2.0	2.0	1.0	15.0	
SRR2037083_NODE_34_length_38779_cov_9.707442	PRJNA230363	Oral	1	53.0	0.0	42.0	5.0	0.0	1.0	0.0	5.0	Oribacterium
ERR1474570_NODE_93_length_38658_cov_18.401005	PRJEB14383	Saliva	1	54.0	0.0	35.0	11.0	2.0	2.0	1.0	3.0	Granulicatella
ERR1474608_NODE_136_length_38532_cov_16.387556	PRJEB14383	Saliva	1	53.0	0.0	17.0	27.0	1.0	2.0	2.0	4.0	Streptococcus
ERR1474587_NODE_90_length_41279_cov_9.237701	PRJEB14383	Saliva	1	56.0	1.0	38.0	13.0	2.0	1.0	0.0	2.0	Veillonella
ERR1474568_NODE_150_length_41544_cov_7.671817	PRJEB14383	Saliva	1	57.0	1.0	34.0	0.0	15.0	1.0	5.0	2.0	Veillonella
SRR1045093_NODE_16_length_48977_cov_51.339397	PRJNA230363	Oral	1	78.0	0.0	54.0	17.0	5.0	0.0	1.0	1.0	Cardiobacterium
SRR1045099_NODE_4_length_43943_cov_9.131334	PRJNA230363	Oral	1	62.0	0.0	46.0	7.0	1.0	1.0	0.0	7.0	Actinomyces
ERR1474586_NODE_71_length_48851_cov_11.651734	PRJEB14383	Saliva	1	63.0	1.0	2.0	48.0	1.0	1.0	1.0	10.0	Rothia
ERR634994_NODE_288_length_48834_cov_18.722114	PRJEB7331	Fecal	1	74.0	0.0	50.0	10.0	1.0	3.0	0.0	10.0	
ERR1474571_NODE_4_length_48245_cov_6.625794	PRJEB14383	Saliva	1	66.0	1.0	50.0	1.0	1.0	0.0	0.0	14.0	
ERR537010_NODE_268_length_46963_cov_25.067579	PRJEB6542	Fecal	1	54.0	0.0	29.0	18.0	4.0	1.0	0.0	2.0	Clostridium
ERR634976_NODE_387_length_46081_cov_4.970321	PRJEB7331	Fecal	1	78.0	0.0	64.0	2.0	0.0	2.0	1.0	9.0	Ruminococcus
ERR1823589_NODE_204_length_46019_cov_7.035245	PRJEB19367	Fecal	1	59.0	0.0	3.0	7.0	1.0	0.0	0.0	48.0	Bacteroides
ERR1474570_NODE_69_length_45450_cov_6.435268	PRJEB14383	Saliva	1	67.0	0.0	43.0	17.0	3.0	0.0	0.0	4.0	Haemophilus
SRR2037083_NODE_19_length_45085_cov_9.699778	PRJNA230363	Oral	1	66.0	0.0	54.0	2.0	1.0	0.0	1.0	8.0	Porphyromonas
ERR695625_NODE_396_length_44706_cov_10.101789	PRJEB7949	Fecal	1	79.0	1.0	63.0	9.0	3.0	1.0	2.0	1.0	Flavonifractor
ERR1744185_NODE_242_length_44654_cov_13.017982	PRJEB18265	Fecal	1	61.0	0.0	31.0	2.0	0.0	2.0	0.0	26.0	Bacteroides
ERR1474586_NODE_92_length_44393_cov_17.669426	PRJEB14383	Saliva	1	63.0	1.0	35.0	5.0	0.0	0.0	0.0	23.0	Prevotella
ERR634983_NODE_477_length_44263_cov_15.667979	PRJEB7331	Fecal	1	54.0	2.0	13.0	31.0	2.0	3.0	0.0	5.0	
ERR1474587_NODE_80_length_44122_cov_6.975378	PRJEB14383	Saliva	1	65.0	1.0	53.0	3.0	1.0	0.0	1.0	7.0	Stomatobaculum

Viral Operating Taxonomy Unit (vOTU)	**vOTU BioProject**	**vOTU Microbiome**	**Cluster size**	**Gene count**	**Potential AMG count**	**Viral hypothetical genes**	**Viral genes with viral benefits**	**Viral structure genes**	**Viral genes with host benefits**	**Viral replication genes**	**Viral genes with unknown function**	**Predicted host (genus)**
ERR505097_NODE_193_length_43789_cov_51.006700	PRJEB6092	Fecal	1	54.0	0.0	45.0	7.0	0.0	0.0	0.0	2.0	Prevotella
ERR1474587_NODE_89_length_41594_cov_10.382701	PRJEB14383	Saliva	1	55.0	1.0	32.0	15.0	1.0	5.0	0.0	2.0	Veillonella
ERR634971_NODE_467_length_43591_cov_6.102949	PRJEB7331	Fecal	1	61.0	1.0	54.0	2.0	0.0	0.0	1.0	4.0	
ERR1474570_NODE_77_length_43559_cov_67.178122	PRJEB14383	Saliva	1	61.0	0.0	33.0	3.0	2.0	0.0	0.0	23.0	Prevotella
ERR1823591_NODE_175_length_43133_cov_10.082432	PRJEB19367	Fecal	1	77.0	1.0	66.0	2.0	6.0	1.0	1.0	1.0	
ERR1474568_NODE_135_length_43128_cov_12.956748	PRJEB14383	Saliva	1	57.0	1.0	43.0	12.0	0.0	1.0	0.0	1.0	Kineosphaera
ERR1474570_NODE_79_length_43103_cov_9.389705	PRJEB14383	Saliva	1	60.0	0.0	4.0	18.0	1.0	3.0	0.0	34.0	Streptococcus
SRR1045094_NODE_1_length_42825_cov_6.415548	PRJNA230363	Oral	1	67.0	2.0	53.0	10.0	0.0	3.0	0.0	1.0	
ERR1474612_NODE_171_length_42773_cov_36.439698	PRJEB14383	Saliva	1	59.0	1.0	10.0	46.0	2.0	0.0	0.0	1.0	
ERR634991_NODE_230_length_42743_cov_14.215400	PRJEB7331	Fecal	1	56.0	0.0	37.0	11.0	4.0	2.0	0.0	2.0	Lachnoclostridium
ERR1474608_NODE_117_length_42436_cov_22.329676	PRJEB14383	Saliva	1	66.0	0.0	16.0	41.0	0.0	1.0	4.0	4.0	Streptococcus
ERR1823590_NODE_254_length_42335_cov_84.180535	PRJEB19367	Fecal	1	55.0	0.0	27.0	0.0	0.0	4.0	0.0	24.0	Bacteroides
ERR1474564_NODE_5_length_41963_cov_13.255417	PRJEB14383	Saliva	1	66.0	3.0	5.0	6.0	12.0	0.0	4.0	39.0	Veillonella
ERR1474586_NODE_107_length_41650_cov_5.490684	PRJEB14383	Saliva	1	51.0	2.0	32.0	0.0	13.0	1.0	2.0	3.0	
ERR1474582_NODE_21_length_41618_cov_20.990953	PRJEB14383	Saliva	1	44.0	0.0	28.0	0.0	4.0	1.0	0.0	11.0	Neisseria
ERR719854_NODE_363_length_10137_cov_3.307677	PRJEB8094	Fecal	1	12.0	0.0	4.0	2.0	1.0	1.0	2.0	2.0	Flavonifractor

Table C.3: Contigs predicted to be novel anellovirus *Anelloviridae* species (n=228) assembled and analysed in this study.

ContigID	SRARunID	BioProject	Microbiome	Country	Genus
ERR1989828_NODE_605_length_3247_cov_15.901629	ERR1989828	PRJEB20877	Pulmonary system	Switzerland	Alphatorquevirus
SRR2037083_NODE_1414_length_2801_cov_8.129643	SRR2037083	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037083_NODE_1216_length_3083_cov_71.353699	SRR2037083	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037083_NODE_1192_length_3146_cov_93.224199	SRR2037083	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037083_NODE_1207_length_3105_cov_177.622623	SRR2037083	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037083_NODE_1250_length_3025_cov_80.289562	SRR2037083	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037083_NODE_1261_length_3010_cov_84.541117	SRR2037083	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037083_NODE_1287_length_2975_cov_86.436644	SRR2037083	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037083_NODE_1351_length_2892_cov_19.944307	SRR2037083	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037083_NODE_1225_length_3068_cov_111.693993	SRR2037083	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037083_NODE_1614_length_2530_cov_127.395556	SRR2037083	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037083_NODE_1226_length_3067_cov_70.074037	SRR2037083	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037083_NODE_1247_length_3026_cov_31.711208	SRR2037083	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037083_NODE_1307_length_2949_cov_72.204561	SRR2037083	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037083_NODE_1243_length_3036_cov_66.863133	SRR2037083	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037083_NODE_1298_length_2956_cov_80.619097	SRR2037083	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037083_NODE_1336_length_2912_cov_102.922296	SRR2037083	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037083_NODE_1418_length_2795_cov_59.138321	SRR2037083	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037083_NODE_1526_length_2636_cov_7.888803	SRR2037083	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037083_NODE_1450_length_2739_cov_5.714232	SRR2037083	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037083_NODE_1055_length_3506_cov_100.455810	SRR2037083	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037083_NODE_1236_length_3045_cov_75.244816	SRR2037083	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037083_NODE_1316_length_2937_cov_74.704025	SRR2037083	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037083_NODE_1373_length_2845_cov_5.214337	SRR2037083	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037083_NODE_1283_length_2979_cov_23.466826	SRR2037083	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037084_NODE_664_length_3035_cov_99.495638	SRR2037084	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037084_NODE_720_length_2896_cov_67.324182	SRR2037084	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037084_NODE_641_length_3105_cov_93.142295	SRR2037084	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037085_NODE_5249_length_2843_cov_18.285868	SRR2037085	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037085_NODE_6587_length_2480_cov_6.510515	SRR2037085	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037085_NODE_5468_length_2777_cov_115.985305	SRR2037085	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037085_NODE_5103_length_2890_cov_6.426455	SRR2037085	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037085_NODE_5740_length_2697_cov_3.995836	SRR2037085	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037085_NODE_5287_length_2832_cov_109.884768	SRR2037085	PRJNA230363	Oral	China	Alphatorquevirus
SRR2037085_NODE_5358_length_2813_cov_5.227339	SRR2037085	PRJNA230363	Oral	China	Alphatorquevirus
SRR6316209_NODE_21_length_3898_cov_133.309654	SRR6316209	PRJNA419524	Blood	USA	Alphatorquevirus
SRR6316221_NODE_16_length_2904_cov_63.788698	SRR6316221	PRJNA419524	Blood	USA	Alphatorquevirus
SRR6316270_NODE_6_length_3044_cov_21.549682	SRR6316270	PRJNA419524	Blood	USA	Alphatorquevirus
SRR6316286_NODE_1_length_3840_cov_124.134478	SRR6316286	PRJNA419524	Blood	USA	Alphatorquevirus
SRR6316314_NODE_10_length_2592_cov_4.729602	SRR6316314	PRJNA419524	Blood	USA	Alphatorquevirus
SRR7166762_NODE_22_length_3522_cov_35.503317	SRR7166762	PRJNA471187	Blood	Sweden	Alphatorquevirus
SRR7166826_NODE_23_length_2307_cov_7.345027	SRR7166826	PRJNA471187	Blood	Sweden	Alphatorquevirus
SRR7166877_NODE_24_length_3132_cov_5.864153	SRR7166877	PRJNA471187	Blood	Sweden	Alphatorquevirus
SRR7166877_NODE_44_length_2009_cov_15.385363	SRR7166877	PRJNA471187	Blood	Sweden	Alphatorquevirus
SRR7166943_NODE_1_length_3711_cov_81.121718	SRR7166943	PRJNA471187	Blood	Sweden	Alphatorquevirus
SRR7167030_NODE_11_length_3840_cov_73.389696	SRR7167030	PRJNA471187	Blood	Sweden	Alphatorquevirus
DRR140164_NODE_13_length_2885_cov_7.748410	DRR140164	PRJDB7117	Blood	Japan	Betatorquevirus
DRR140165_NODE_22_length_3550_cov_2.654077	DRR140165	PRJDB7117	Blood	Japan	Betatorquevirus
DRR140166_NODE_60_length_2632_cov_8.199069	DRR140166	PRJDB7117	Blood	Japan	Betatorquevirus
DRR140166_NODE_54_length_2692_cov_11.335988	DRR140166	PRJDB7117	Blood	Japan	Betatorquevirus
DRR140173_NODE_16_length_2430_cov_3.621053	DRR140173	PRJDB7117	Blood	Japan	Betatorquevirus
DRR140173_NODE_17_length_2149_cov_2.801815	DRR140173	PRJDB7117	Blood	Japan	Betatorquevirus
DRR140174_NODE_16_length_2050_cov_2.710777	DRR140174	PRJDB7117	Blood	Japan	Betatorquevirus
DRR140178_NODE_3_length_2980_cov_15.155556	DRR140178	PRJDB7117	Blood	Japan	Betatorquevirus
SRR10951765_NODE_1409_length_2824_cov_3.586854	SRR10951765	PRJNA602694	Blood	Brazil	Betatorquevirus
SRR2037083_NODE_1440_length_2757_cov_3.851221	SRR2037083	PRJNA230363	Oral	China	Betatorquevirus
SRR2037083_NODE_1407_length_2810_cov_71.443194	SRR2037083	PRJNA230363	Oral	China	Betatorquevirus
SRR2037083_NODE_1525_length_2637_cov_6.265686	SRR2037083	PRJNA230363	Oral	China	Betatorquevirus
SRR2037083_NODE_1730_length_2449_cov_3.651211	SRR2037083	PRJNA230363	Oral	China	Betatorquevirus
SRR2037083_NODE_1433_length_2769_cov_39.274134	SRR2037083	PRJNA230363	Oral	China	Betatorquevirus
SRR2037083_NODE_1273_length_2990_cov_48.493015	SRR2037083	PRJNA230363	Oral	China	Betatorquevirus
SRR2037084_NODE_760_length_2822_cov_75.136249	SRR2037084	PRJNA230363	Oral	China	Betatorquevirus
SRR2037085_NODE_5387_length_2804_cov_5.351401	SRR2037085	PRJNA230363	Oral	China	Betatorquevirus
SRR2037085_NODE_6414_length_2511_cov_8.802932	SRR2037085	PRJNA230363	Oral	China	Betatorquevirus
SRR2037085_NODE_5107_length_2889_cov_9.442131	SRR2037085	PRJNA230363	Oral	China	Betatorquevirus
SRR2037085_NODE_6450_length_2505_cov_5.317143	SRR2037085	PRJNA230363	Oral	China	Betatorquevirus
SRR2037085_NODE_6018_length_2618_cov_4.453765	SRR2037085	PRJNA230363	Oral	China	Betatorquevirus
SRR2037085_NODE_5390_length_2801_cov_51.262564	SRR2037085	PRJNA230363	Oral	China	Betatorquevirus

ContigID	SRARunID	BioProject	Microbiome	Country	Genus
SRR2037085_NODE_5400_length_2799_cov_8.270773	SRR2037085	PRJNA230363	Oral	China	Betatorquevirus
SRR2037085_NODE_6297_length_2542_cov_3.851628	SRR2037085	PRJNA230363	Oral	China	Betatorquevirus
SRR2037085_NODE_7975_length_2196_cov_3.409155	SRR2037085	PRJNA230363	Oral	China	Betatorquevirus
SRR2037085_NODE_6385_length_2519_cov_9.820211	SRR2037085	PRJNA230363	Oral	China	Betatorquevirus
SRR2037085_NODE_5919_length_2647_cov_3.398534	SRR2037085	PRJNA230363	Oral	China	Betatorquevirus
SRR2037085_NODE_5501_length_2767_cov_9.144174	SRR2037085	PRJNA230363	Oral	China	Betatorquevirus
SRR2037085_NODE_6391_length_2518_cov_9.021112	SRR2037085	PRJNA230363	Oral	China	Betatorquevirus
SRR2037085_NODE_5670_length_2718_cov_4.489673	SRR2037085	PRJNA230363	Oral	China	Betatorquevirus
SRR2037085_NODE_8719_length_2068_cov_3.507700	SRR2037085	PRJNA230363	Oral	China	Betatorquevirus
SRR2037085_NODE_5404_length_2798_cov_9.382793	SRR2037085	PRJNA230363	Oral	China	Betatorquevirus
SRR2037085_NODE_5339_length_2819_cov_15.157019	SRR2037085	PRJNA230363	Oral	China	Betatorquevirus
SRR2037085_NODE_5981_length_2629_cov_6.550117	SRR2037085	PRJNA230363	Oral	China	Betatorquevirus
SRR2037085_NODE_6009_length_2623_cov_5.651480	SRR2037085	PRJNA230363	Oral	China	Betatorquevirus
SRR2037085_NODE_5966_length_2634_cov_5.871656	SRR2037085	PRJNA230363	Oral	China	Betatorquevirus
SRR5681769_NODE_1_length_2523_cov_21.696648	SRR5681769	PRJNA389455	Blood	NA	Betatorquevirus
SRR6316202_NODE_13_length_2856_cov_4.798286	SRR6316202	PRJNA419524	Blood	USA	Betatorquevirus
SRR6316202_NODE_17_length_2692_cov_7.566932	SRR6316202	PRJNA419524	Blood	USA	Betatorquevirus
SRR6316203_NODE_18_length_2827_cov_7.550505	SRR6316203	PRJNA419524	Blood	USA	Betatorquevirus
SRR6316221_NODE_20_length_2788_cov_5.750823	SRR6316221	PRJNA419524	Blood	USA	Betatorquevirus
SRR6316221_NODE_17_length_2896_cov_17.302006	SRR6316221	PRJNA419524	Blood	USA	Betatorquevirus
SRR6316233_NODE_27_length_2807_cov_15.835756	SRR6316233	PRJNA419524	Blood	USA	Betatorquevirus
SRR6316233_NODE_24_length_2862_cov_16.978269	SRR6316233	PRJNA419524	Blood	USA	Betatorquevirus
SRR6316233_NODE_23_length_2869_cov_50.130775	SRR6316233	PRJNA419524	Blood	USA	Betatorquevirus
SRR6316233_NODE_16_length_2950_cov_88.579620	SRR6316233	PRJNA419524	Blood	USA	Betatorquevirus
SRR6316236_NODE_3_length_2976_cov_8.485108	SRR6316236	PRJNA419524	Blood	USA	Betatorquevirus
SRR6316247_NODE_19_length_2023_cov_4.168191	SRR6316247	PRJNA419524	Blood	USA	Betatorquevirus
SRR6316257_NODE_19_length_2573_cov_7.963463	SRR6316257	PRJNA419524	Blood	USA	Betatorquevirus
SRR6316270_NODE_11_length_2700_cov_4.344423	SRR6316270	PRJNA419524	Blood	USA	Betatorquevirus
SRR6316284_NODE_11_length_2885_cov_6.932862	SRR6316284	PRJNA419524	Blood	USA	Betatorquevirus
SRR6316307_NODE_8_length_2837_cov_77.074766	SRR6316307	PRJNA419524	Blood	USA	Betatorquevirus
SRR7166769_NODE_118_length_2722_cov_7.362955	SRR7166769	PRJNA471187	Blood	Sweden	Betatorquevirus
SRR7166809_NODE_1_length_2005_cov_3.933333	SRR7166809	PRJNA471187	Blood	Sweden	Betatorquevirus
SRR7166826_NODE_20_length_2511_cov_2.880700	SRR7166826	PRJNA471187	Blood	Sweden	Betatorquevirus
SRR7166829_NODE_7_length_2983_cov_66.900615	SRR7166829	PRJNA471187	Blood	Sweden	Betatorquevirus
SRR7166860_NODE_1_length_2940_cov_63.382322	SRR7166860	PRJNA471187	Blood	Sweden	Betatorquevirus
SRR7166883_NODE_14_length_2147_cov_76.689293	SRR7166883	PRJNA471187	Blood	Sweden	Betatorquevirus
SRR7166917_NODE_46_length_2958_cov_33.011712	SRR7166917	PRJNA471187	Blood	Sweden	Betatorquevirus
SRR7166946_NODE_3_length_2838_cov_6.518146	SRR7166946	PRJNA471187	Blood	Sweden	Betatorquevirus
SRR7166965_NODE_6_length_2904_cov_14.952615	SRR7166965	PRJNA471187	Blood	Sweden	Betatorquevirus
SRR7167022_NODE_102_length_2985_cov_44.358020	SRR7167022	PRJNA471187	Blood	Sweden	Betatorquevirus
SRR7167078_NODE_119_length_2844_cov_9.391180	SRR7167078	PRJNA471187	Blood	Sweden	Betatorquevirus
SRR7167078_NODE_138_length_2635_cov_4.973643	SRR7167078	PRJNA471187	Blood	Sweden	Betatorquevirus
SRR7167078_NODE_107_length_2931_cov_22.772601	SRR7167078	PRJNA471187	Blood	Sweden	Betatorquevirus
SRR7167078_NODE_133_length_2666_cov_33.724627	SRR7167078	PRJNA471187	Blood	Sweden	Betatorquevirus
SRR7167078_NODE_113_length_2881_cov_12.172328	SRR7167078	PRJNA471187	Blood	Sweden	Betatorquevirus
SRR7167078_NODE_139_length_2615_cov_32.800391	SRR7167078	PRJNA471187	Blood	Sweden	Betatorquevirus
SRR7167078_NODE_129_length_2717_cov_11.590158	SRR7167078	PRJNA471187	Blood	Sweden	Betatorquevirus
SRR8862005_NODE_7_length_2998_cov_82.991845	SRR8862005	PRJNA518922	Blood	NA	Betatorquevirus
DRR140165_NODE_49_length_2854_cov_18.827081	DRR140165	PRJDB7117	Blood	Japan	Gammatorquevirus
DRR140166_NODE_99_length_2031_cov_6.637652	DRR140166	PRJDB7117	Blood	Japan	Gammatorquevirus
DRR140173_NODE_14_length_2503_cov_7.160131	DRR140173	PRJDB7117	Blood	Japan	Gammatorquevirus
SRR2037083_NODE_1617_length_2528_cov_69.701981	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1692_length_2476_cov_65.699711	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1695_length_2474_cov_15.520050	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1490_length_2690_cov_40.173435	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1531_length_2630_cov_65.916117	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1702_length_2470_cov_5.663768	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1579_length_2569_cov_24.115354	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1639_length_2514_cov_80.590484	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1713_length_2463_cov_31.497093	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1672_length_2488_cov_17.247842	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1685_length_2480_cov_9.928660	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1635_length_2516_cov_11.924015	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1715_length_2460_cov_11.387942	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1701_length_2470_cov_13.613665	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1746_length_2438_cov_28.350399	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1698_length_2470_cov_75.913043	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1664_length_2494_cov_65.056991	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1591_length_2556_cov_24.097961	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1718_length_2457_cov_9.904246	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1676_length_2484_cov_34.350350	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1680_length_2482_cov_23.366296	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1487_length_2693_cov_76.602350	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus

ContigID	SRARunID	BioProject	Microbiome	Country	Genus
SRR2037083_NODE_1666_length_2493_cov_99.588597	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1596_length_2553_cov_72.678143	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1622_length_2523_cov_41.655592	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1683_length_2481_cov_3.532564	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1699_length_2470_cov_61.400000	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1726_length_2450_cov_54.724426	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1697_length_2471_cov_6.011589	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037083_NODE_1649_length_2504_cov_76.756635	SRR2037083	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037084_NODE_910_length_2464_cov_11.052304	SRR2037084	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037084_NODE_892_length_2489_cov_13.798685	SRR2037084	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037084_NODE_944_length_2415_cov_5.583475	SRR2037084	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037084_NODE_824_length_2624_cov_11.172441	SRR2037084	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037084_NODE_842_length_2570_cov_17.083897	SRR2037084	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037084_NODE_899_length_2478_cov_9.038795	SRR2037084	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_5200_length_2857_cov_38.860457	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6643_length_2471_cov_22.883278	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6695_length_2463_cov_13.318522	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_7023_length_2387_cov_12.205403	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6607_length_2477_cov_24.983485	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6544_length_2489_cov_4.679951	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6550_length_2487_cov_36.356908	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_5475_length_2775_cov_31.245588	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6491_length_2496_cov_30.374027	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6508_length_2493_cov_4.894586	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_7436_length_2297_cov_3.397413	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_8039_length_2184_cov_5.918741	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6955_length_2405_cov_7.004255	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6654_length_2469_cov_10.145816	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6789_length_2442_cov_29.776288	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6552_length_2487_cov_5.763158	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6479_length_2498_cov_67.277118	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6307_length_2539_cov_20.448068	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6559_length_2485_cov_53.960082	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6829_length_2434_cov_6.817150	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_7098_length_2368_cov_4.104194	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_8055_length_2181_cov_5.339605	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6480_length_2498_cov_16.431437	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6631_length_2473_cov_21.275848	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6722_length_2456_cov_42.191587	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6706_length_2461_cov_80.141313	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6704_length_2462_cov_4.831325	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6798_length_2441_cov_5.411567	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6487_length_2497_cov_20.082310	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6439_length_2507_cov_53.873165	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6580_length_2481_cov_5.784831	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6689_length_2463_cov_93.828073	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6716_length_2459_cov_6.049085	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6565_length_2484_cov_27.446274	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6619_length_2475_cov_59.180579	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6710_length_2460_cov_31.163410	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6632_length_2473_cov_18.833333	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6604_length_2478_cov_13.554684	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6531_length_2490_cov_103.070637	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6473_length_2500_cov_23.070348	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6586_length_2480_cov_6.567423	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_8723_length_2067_cov_4.818091	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6781_length_2443_cov_15.309464	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_7456_length_2292_cov_5.023692	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6602_length_2478_cov_46.224928	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6693_length_2463_cov_25.809801	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6533_length_2490_cov_12.678439	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6827_length_2434_cov_20.583438	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6645_length_2470_cov_68.151967	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6570_length_2482_cov_73.878039	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6765_length_2446_cov_69.466750	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6800_length_2440_cov_66.693082	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6426_length_2509_cov_86.914833	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR2037085_NODE_6615_length_2476_cov_27.154895	SRR2037085	PRJNA230363	Oral	China	Gammatorquevirus
SRR5681814_NODE_1_length_2643_cov_22.722136	SRR5681814	PRJNA389455	Blood	NA	Gammatorquevirus
SRR6316199_NODE_1_length_3219_cov_11.170670	SRR6316199	PRJNA419524	Blood	USA	Gammatorquevirus
SRR6316202_NODE_18_length_2448_cov_4.269954	SRR6316202	PRJNA419524	Blood	USA	Gammatorquevirus
SRR6316203_NODE_9_length_3018_cov_80.558218	SRR6316203	PRJNA419524	Blood	USA	Gammatorquevirus
SRR6316247_NODE_16_length_2461_cov_9.033250	SRR6316247	PRJNA419524	Blood	USA	Gammatorquevirus

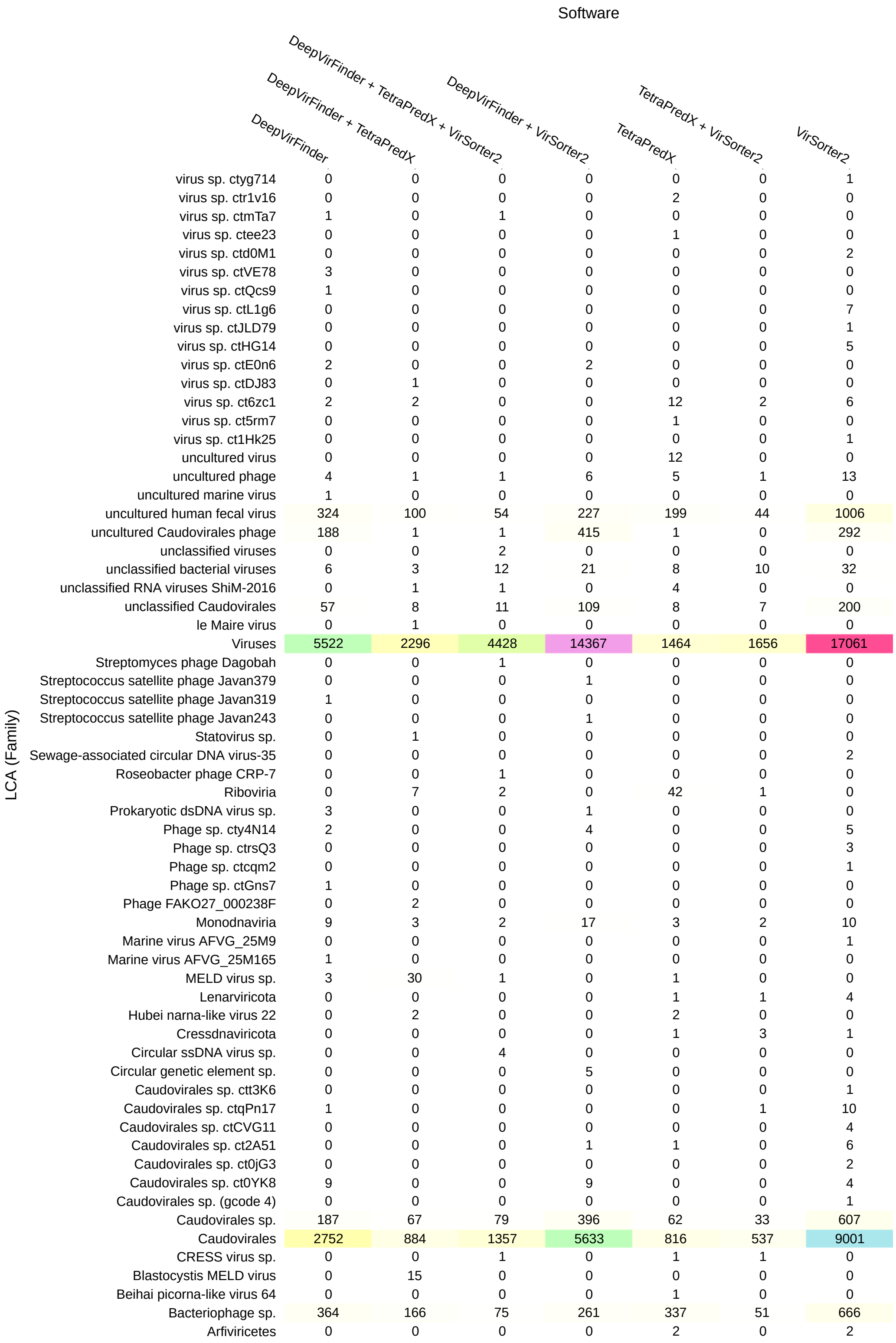
ContigID	SRARunID	BioProject	Microbiome	Country	Genus
SRR6316272_NODE_5_length_2482_cov_4.211372	SRR6316272	PRJNA419524	Blood	USA	Gammatorquevirus
SRR6316298_NODE_19_length_2905_cov_76.388070	SRR6316298	PRJNA419524	Blood	USA	Gammatorquevirus
SRR7166769_NODE_99_length_3238_cov_63.412190	SRR7166769	PRJNA471187	Blood	Sweden	Gammatorquevirus
SRR7166772_NODE_78_length_2820_cov_9.428571	SRR7166772	PRJNA471187	Blood	Sweden	Gammatorquevirus
SRR7166791_NODE_91_length_2982_cov_23.078579	SRR7166791	PRJNA471187	Blood	Sweden	Gammatorquevirus
SRR7166883_NODE_12_length_2402_cov_65.199403	SRR7166883	PRJNA471187	Blood	Sweden	Gammatorquevirus
SRR7167078_NODE_150_length_2498_cov_79.879247	SRR7167078	PRJNA471187	Blood	Sweden	Gammatorquevirus
SRR8862010_NODE_54_length_2554_cov_9.011605	SRR8862010	PRJNA518922	Blood	NA	Gammatorquevirus
SRR2037083_NODE_2038_length_2150_cov_5.638186	SRR2037083	PRJNA230363	Oral	China	Hetorquevirus
SRR2037083_NODE_1583_length_2564_cov_57.915903	SRR2037083	PRJNA230363	Oral	China	Unclassified
SRR2037084_NODE_746_length_2845_cov_23.757706	SRR2037084	PRJNA230363	Oral	China	Unclassified
SRR2037085_NODE_5950_length_2639_cov_9.782508	SRR2037085	PRJNA230363	Oral	China	Unclassified
SRR6316308_NODE_9_length_3148_cov_65.976398	SRR6316308	PRJNA419524	Blood	USA	Unclassified
SRR7166865_NODE_11_length_2799_cov_13.776968	SRR7166865	PRJNA471187	Blood	Sweden	Unclassified

Table C.4: Potentially novel unclassified RNA virus contigs (n=38).

Contig ID	SRA Run ID	BioProject	Microbiome	% Identity	Top hit	Location	Contig length	LCA
ERR1022500_NODE_96_length_4384_cov_27.812197	ERR1022500	PRJEB10919	Sputum	42.2	YP_009640127.1 hypothetical protein MS2g4 [Escherichia phage MS2]	South Africa	4384	Lenarviricota
ERR1022515_NODE_284_length_1721_cov_58.716086	ERR1022515	PRJEB10919	Sputum	47.2	YP_009351841.1 putative RNA-dependent RNA polymerase [Otarine picobirnavirus]	South Africa	1721	Riboviria
ERR1022514_NODE_1939_length_1415_cov_13.957353	ERR1022514	PRJEB10919	Sputum	33.8	YP_009508065.1 RNA-dependent RNA polymerase [Cryptosporidium parvum virus 1]	South Africa	1415	Viruses
ERR1022514_NODE_1221_length_1788_cov_19.901327	ERR1022514	PRJEB10919	Sputum	32.9	YP_009336749.1 RNA-dependent RNA polymerase [Hubei narna-like virus 22]	South Africa	1788	Hubei narna-like virus 22
ERR1022500_NODE_1723_length_1596_cov_30.505516	ERR1022500	PRJEB10919	Sputum	25.3	YP_009333352.1 RdRp [Beihai picobirna-like virus 7]	South Africa	1596	unclassified RNA viruses ShiM-2016
ERR1022500_NODE_449_length_2653_cov_51.551963	ERR1022500	PRJEB10919	Sputum	26.3	YP_009351840.1 capsid protein [Otarine picobirnavirus]	South Africa	2653	Otarine picobirnavirus
ERR1022501_NODE_255_length_2573_cov_9.942017	ERR1022501	PRJEB10919	Sputum	28.9	YP_009336749.1 RNA-dependent RNA polymerase [Hubei narna-like virus 22]	South Africa	2573	Hubei narna-like virus 22
ERR1022502_NODE_2423_length_1658_cov_6.269495	ERR1022502	PRJEB10919	Sputum	32.4	YP_009272899.1 RNA-dependent RNA polymerase [Fusarium poae mitovirus 2]	South Africa	1658	Riboviria
ERR1022501_NODE_270_length_2525_cov_39.081781	ERR1022501	PRJEB10919	Sputum	31.9	YP_009336749.1 RNA-dependent RNA polymerase [Hubei narna-like virus 22]	South Africa	2525	Hubei narna-like virus 22
ERR1022515_NODE_328_length_1631_cov_47.637056	ERR1022515	PRJEB10919	Sputum	35.6	YP_009508065.1 RNA-dependent RNA polymerase [Cryptosporidium parvum virus 1]	South Africa	1631	Viruses
ERR1022512_NODE_6_length_4167_cov_63.667072	ERR1022512	PRJEB10919	Sputum	34.8	NP_085473.1 replicase [Acinetobacter phage AP205]	South Africa	4167	Lenarviricota
ERR1022507_NODE_109_length_4151_cov_63.085205	ERR1022507	PRJEB10919	Sputum	31.3	NP_085473.1 replicase [Acinetobacter phage AP205]	South Africa	4151	Lenarviricota
ERR1022506_NODE_40_length_2936_cov_16.501215	ERR1022506	PRJEB10919	Sputum	42.5	YP_009640127.1 hypothetical protein MS2g4 [Escherichia phage MS2]	South Africa	2936	Lenarviricota
ERR1022505_NODE_26_length_4377_cov_40.463674	ERR1022505	PRJEB10919	Sputum	42.8	YP_009640127.1 hypothetical protein MS2g4 [Escherichia phage MS2]	South Africa	4377	Lenarviricota
ERR1022502_NODE_2009_length_1831_cov_8.564752	ERR1022502	PRJEB10919	Sputum	39.0	YP_009241385.1 capsid protein [Porcine picobirnavirus]	South Africa	1831	Picobirnavirus
ERR505106_NODE_35191_length_1081_cov_2.953216	ERR505106	PRJEB6092	Fecal	42.0	YP_009163924.1 putative replication initiation protein [Mytilus sp. clam associated circular virus]	Australia	1081	Viruses
SRR2037085_NODE_7401_length_2302_cov_3.786827	SRR2037085	PRJNA230363	Oral	46.2	YP_009109677.1 replication-associated protein [Circoviridae 16 LDMD-2013]	China	2302	Viruses
SRR1635854_NODE_966_length_1859_cov_7.512749	SRR1635854	PRJNA264728	Saliva	25.3	YP_009329892.1 RdRp [Hubei diptera virus 18]		1859	unclassified RNA viruses ShiM-2016
SRR1636507_NODE_1462_length_1936_cov_5.306220	SRR1636507	PRJNA264728	Saliva	28.4	YP_009336749.1 RNA-dependent RNA polymerase [Hubei narna-like virus 22]		1936	Hubei narna-like virus 22

Contig ID	SRA Run ID	BioProject	Microbiome	% Identity	Top hit	Location	Contig length	LCA
SRR1748196_NODE_6_length_3877_cov_4.734432	SRR1748196	PRJNA271229	Blood	30.8	YP_009333603.1 hypothetical protein 1 [Beihai picorna-like virus 82]	Nigeria	3877	Viruses
SRR1748182_NODE_2_length_7940_cov_17.389727	SRR1748182	PRJNA271229	Blood	37.7	YP_009336912.1 hypothetical protein 2 [Shahe picorna-like virus 1]	Nigeria	7940	Riboviria
SRR1748196_NODE_8_length_3490_cov_5.342649	SRR1748196	PRJNA271229	Blood	41.1	YP_009336781.1 hypothetical protein 1 [Changjiang picorna-like virus 13]	Nigeria	3490	Viruses
SRR1748185_NODE_16_length_1536_cov_3.962188	SRR1748185	PRJNA271229	Blood	40.7	YP_009667033.1 RNA dependent RNA polymerase [Magnaporthe oryzae ourmia-like virus]	Nigeria	1536	Riboviria
SRR1748182_NODE_1_length_7940_cov_20.758782	SRR1748182	PRJNA271229	Blood	37.4	YP_009336912.1 hypothetical protein 2 [Shahe picorna-like virus 1]	Nigeria	7940	Riboviria
SRR1748196_NODE_55_length_1239_cov_2.301520	SRR1748196	PRJNA271229	Blood	38.9	YP_009336489.1 hypothetical protein 2 [Hubei tombus-like virus 4]	Nigeria	1239	Riboviria
SRR1748181_NODE_15_length_1181_cov_3.009769	SRR1748181	PRJNA271229	Blood	48.0	YP_009333390.1 hypothetical protein 1 [Beihai picorna-like virus 47]	Nigeria	1181	Riboviria
SRR1748184_NODE_7_length_1988_cov_4.106053	SRR1748184	PRJNA271229	Blood	36.3	YP_009336781.1 hypothetical protein 1 [Changjiang picorna-like virus 13]	Nigeria	1988	Riboviria
SRR1748184_NODE_22_length_1033_cov_2.209611	SRR1748184	PRJNA271229	Blood	46.2	YP_052925.1 87 kDa replicase protein [Pelargonium chlorotic ring pattern virus]	Nigeria	1033	Riboviria
SRR1748188_NODE_6_length_1055_cov_3.328000	SRR1748188	PRJNA271229	Blood	32.6	YP_009337699.1 hypothetical protein 2 [Sanxia picorna-like virus 1]	Nigeria	1055	Viruses
SRR1748196_NODE_54_length_1241_cov_1.935076	SRR1748196	PRJNA271229	Blood	43.0	YP_009333391.1 hypothetical protein 2 [Beihai picorna-like virus 47]	Nigeria	1241	Riboviria
SRR1748196_NODE_82_length_1014_cov_2.471324	SRR1748196	PRJNA271229	Blood	35.6	YP_009333388.1 hypothetical protein 1 [Beihai picorna-like virus 64]	Nigeria	1014	Beihai picorna-like virus 64
SRR1748181_NODE_20_length_1033_cov_6.087935	SRR1748181	PRJNA271229	Blood	53.8	YP_009337432.1 hypothetical protein 2 [Changjiang tombus-like virus 5]	Nigeria	1033	Riboviria
SRR1748196_NODE_17_length_1870_cov_2.704132	SRR1748196	PRJNA271229	Blood	43.6	YP_459960.2 p86 [Angelonia flower break virus]	Nigeria	1870	Riboviria
SRR1748181_NODE_17_length_1104_cov_3.046711	SRR1748181	PRJNA271229	Blood	43.4	YP_009270620.1 putative RNA dependent RNA polymerase [Gompholobium virus A]	Nigeria	1104	Riboviria
SRR1748295_NODE_2_length_3088_cov_4.287504	SRR1748295	PRJNA271229	Blood	36.9	YP_009336912.1 hypothetical protein 2 [Shahe picorna-like virus 1]	Nigeria	3088	Viruses
SRR7166913_NODE_99_length_3175_cov_15.859615	SRR7166913	PRJNA471187	Blood	34.1	NP_042772.1 pB354L [African swine fever virus] >YP_009702326.1 pB354L [African swine fever virus]	Sweden	3175	Viruses
SRR7166841_NODE_4_length_2046_cov_3.834756	SRR7166841	PRJNA471187	Blood	42.4	YP_009329369.1 Hypothetical protein BQ3484_429 [Cedratvirus A11]	Sweden	2046	Viruses
SRR7166795_NODE_12_length_1842_cov_2.996642	SRR7166795	PRJNA471187	Blood	54.0	YP_009117082.1 replication-associated protein [Sewage-associated circular DNA virus-35]	Sweden	1842	Viruses

C.1.1 Supplementary figures



LCA (Family)	DeepVirFinder	DeepVirFinder + TetraPredX	DeepVirFinder + TetraPredX + VirSorter2	DeepVirFinder + VirSorter2	TetraPredX	TetraPredX + VirSorter2	VirSorter2
virus sp. ctyg714	0	0	0	0	0	0	1
virus sp. ctr1v16	0	0	0	0	2	0	0
virus sp. ctmTa7	1	0	1	0	0	0	0
virus sp. ctee23	0	0	0	0	1	0	0
virus sp. ctd0M1	0	0	0	0	0	0	2
virus sp. ctVE78	3	0	0	0	0	0	0
virus sp. ctQcs9	1	0	0	0	0	0	0
virus sp. ctL1g6	0	0	0	0	0	0	7
virus sp. ctJLD79	0	0	0	0	0	0	1
virus sp. ctHG14	0	0	0	0	0	0	5
virus sp. ctE0n6	2	0	0	2	0	0	0
virus sp. ctDJ83	0	1	0	0	0	0	0
virus sp. ct6zc1	2	2	0	0	12	2	6
virus sp. ct5rm7	0	0	0	0	1	0	0
virus sp. ct1Hk25	0	0	0	0	0	0	1
uncultured virus	0	0	0	0	12	0	0
uncultured phage	4	1	1	6	5	1	13
uncultured marine virus	1	0	0	0	0	0	0
uncultured human fecal virus	324	100	54	227	199	44	1006
uncultured Caudovirales phage	188	1	1	415	1	0	292
unclassified viruses	0	0	2	0	0	0	0
unclassified bacterial viruses	6	3	12	21	8	10	32
unclassified RNA viruses ShiM-2016	0	1	1	0	4	0	0
unclassified Caudovirales	57	8	11	109	8	7	200
le Maire virus	0	1	0	0	0	0	0
Viruses	5522	2296	4428	14367	1464	1656	17061
Streptomyces phage Dagobah	0	0	1	0	0	0	0
Streptococcus satellite phage Javan379	0	0	0	1	0	0	0
Streptococcus satellite phage Javan319	1	0	0	0	0	0	0
Streptococcus satellite phage Javan243	0	0	0	1	0	0	0
Statovirus sp.	0	1	0	0	0	0	0
Sewage-associated circular DNA virus-35	0	0	0	0	0	0	2
Roseobacter phage CRP-7	0	0	1	0	0	0	0
Riboviria	0	7	2	0	42	1	0
Prokaryotic dsDNA virus sp.	3	0	0	1	0	0	0
Phage sp. cty4N14	2	0	0	4	0	0	5
Phage sp. ctrsQ3	0	0	0	0	0	0	3
Phage sp. ctcqm2	0	0	0	0	0	0	1
Phage sp. ctGns7	1	0	0	0	0	0	0
Phage FAKO27_000238F	0	2	0	0	0	0	0
Monodnaviria	9	3	2	17	3	2	10
Marine virus AFVG_25M9	0	0	0	0	0	0	1
Marine virus AFVG_25M165	1	0	0	0	0	0	0
MELD virus sp.	3	30	1	0	1	0	0
Lenarviricota	0	0	0	0	1	1	4
Hubei narna-like virus 22	0	2	0	0	2	0	0
Cressdnaviricota	0	0	0	0	1	3	1
Circular ssDNA virus sp.	0	0	4	0	0	0	0
Circular genetic element sp.	0	0	0	5	0	0	0
Caudovirales sp. ctt3K6	0	0	0	0	0	0	1
Caudovirales sp. ctqPn17	1	0	0	0	0	1	10
Caudovirales sp. ctCVG11	0	0	0	0	0	0	4
Caudovirales sp. ct2A51	0	0	0	1	1	0	6
Caudovirales sp. ct0jG3	0	0	0	0	0	0	2
Caudovirales sp. ct0YK8	9	0	0	9	0	0	4
Caudovirales sp. (gcode 4)	0	0	0	0	0	0	1
Caudovirales sp.	187	67	79	396	62	33	607
Caudovirales	2752	884	1357	5633	816	537	9001
CRESS virus sp.	0	0	1	0	1	1	0
Blastocystis MELD virus	0	15	0	0	0	0	0
Beihai picorna-like virus 64	0	0	0	0	1	0	0
Bacteriophage sp.	364	166	75	261	337	51	666
Arfiviricetes	0	0	0	0	2	0	2

Figure C.1: A heatmap showing the number of predicted virus contigs using any prediction tool for unclassified virus group (derived from the LCA). Darker shades of colours represent the higher number of contigs in the heatmap. Shades of red colour represent DNA virus families, shades of blue represent RNA virus families and black represents unclassified viruses.

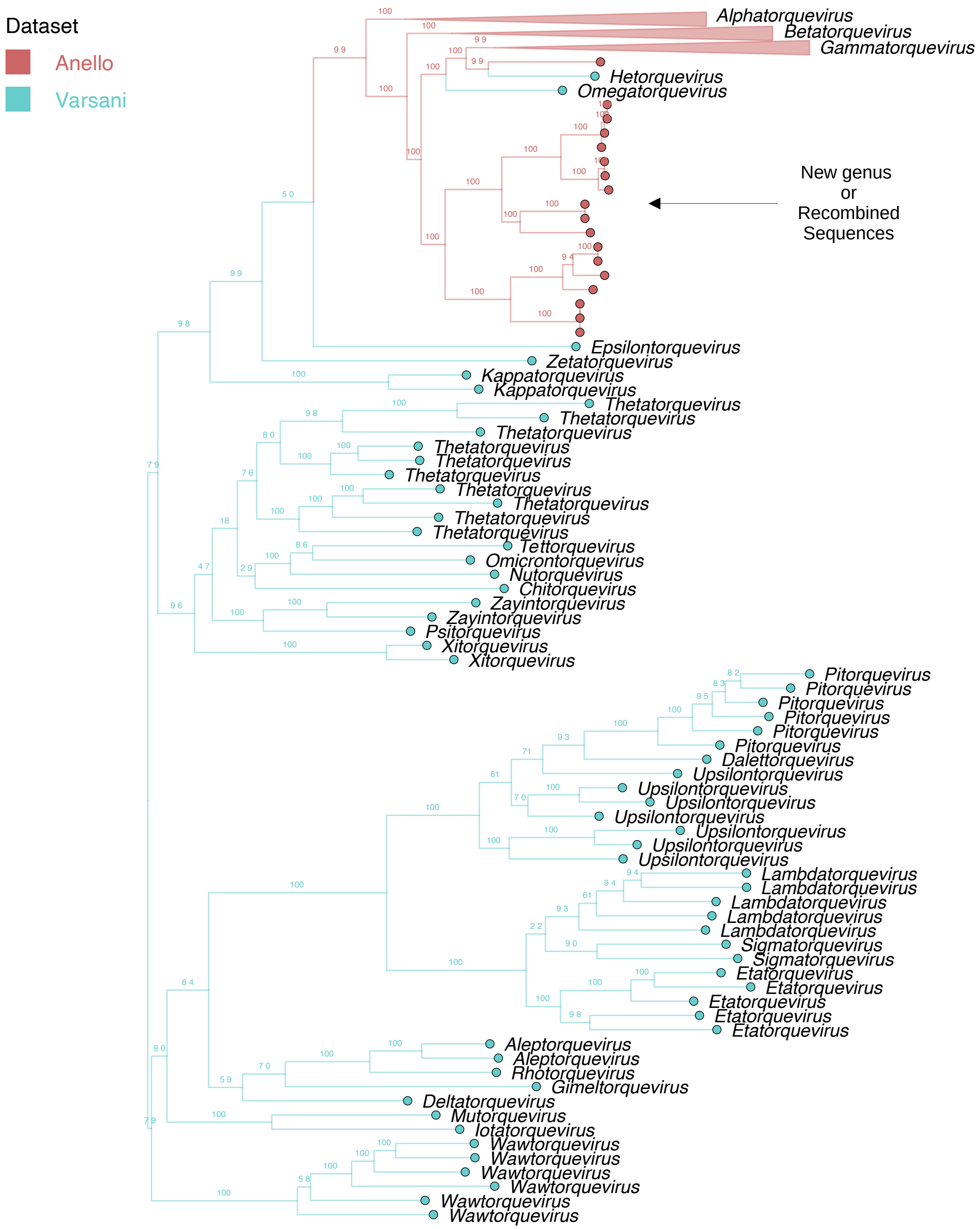


Figure C.2: Phylogenetic analysis and classification of anellovirus contigs. A maximum-likelihood tree inferred from ORF1 amino acid sequences of all anelloviruses. Nodes are coloured according to the study; reference anellovirus sequences retrieved from Varsani et al. (2021) are shown in cyan and anellovirus contigs identified in this study are shown in burnt sienna. The tip labels indicate the corresponding genera of anellovirus sequences used to build the phylogentic tree.

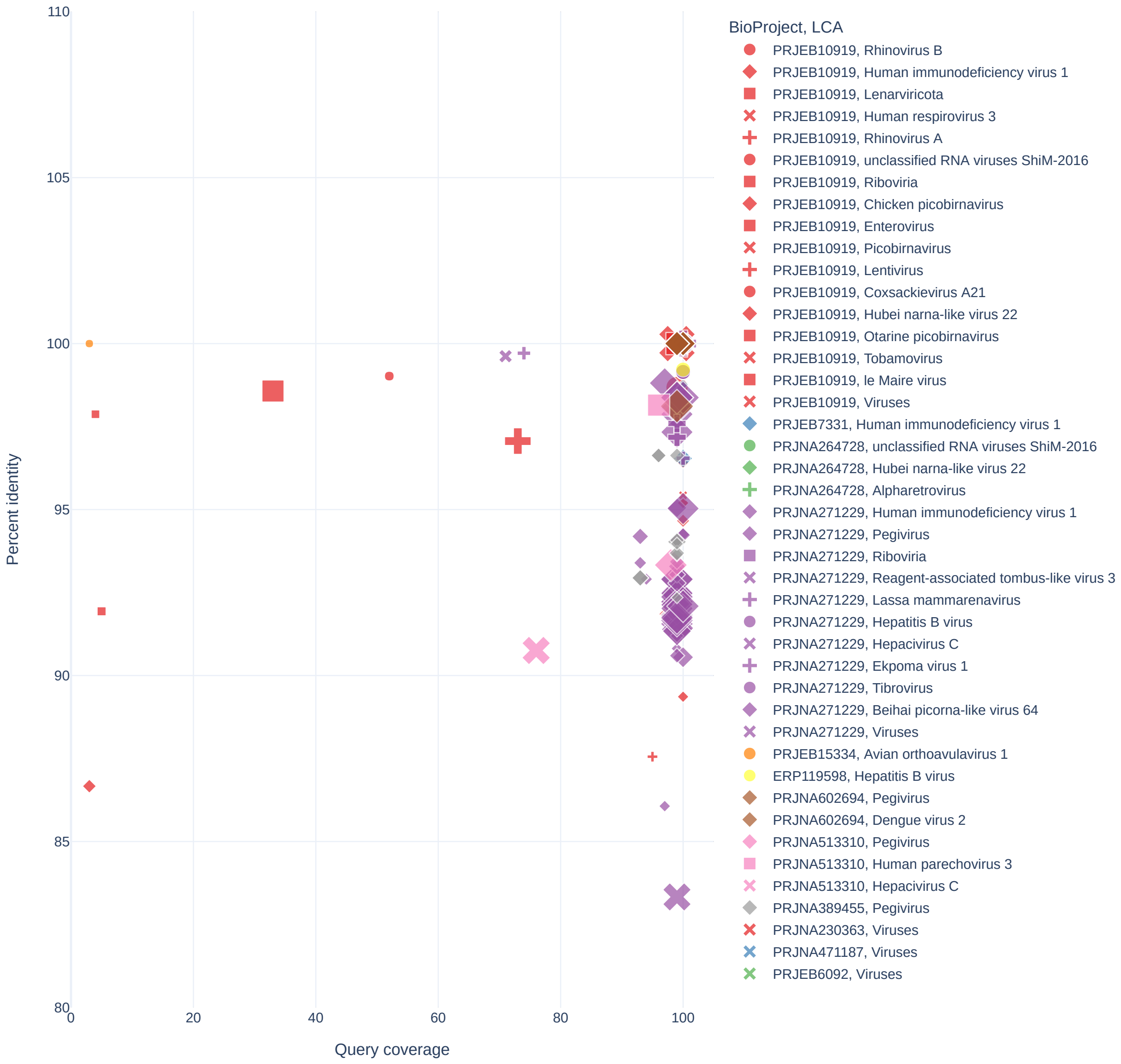


Figure C.3: A scatter plot showing query coverage (X-axis) and corresponding percent identity (Y-axis) to its nucleotide level best hit for each contig. The shapes of the markers represent distinct studies, the colours of the markers represent the LCA taxa and the size of the marker is relative to the contig length.

Appendix D

Publication

Quantifying and cataloguing unknown sequences within human microbiomes

Sejal Modha[1*], David L. Robertson[1], Joseph Hughes[1#], and Richard J. Orton[1#]

[1]MRC University of Glasgow Centre for Virus Research, 464 Bearsden road,
Garscube campus, Glasgow, G61 1QH

Email: s.modha.1@research.gla.ac.uk, david.l.robertson@glasgow.ac.uk,
joseph.hughes@glasgow.ac.uk, richard.orton@glasgow.ac.uk

[*]Corresponding author

[#]*Indicates last author*

Wednesday 9th February, 2022

Abstract

Background

Advances in genome sequencing technologies and lower costs have enabled the exploration of a multitude of known and novel environments and microbiomes. This has led to an exponential growth in the raw sequence data that is deposited in online repositories. Metagenomic and metatranscriptomic data sets are typically analysed with regards to a specific biological question. However, it is widely acknowledged that these data sets are comprised of a proportion of sequences that bear no similarity to any currently known biological sequence, and this so-called 'dark matter' is often excluded from downstream analyses.

Results

In this study, a systematic framework was developed to assemble, identify, and measure the proportion of unknown sequences present in distinct human microbiomes. This framework was applied to forty distinct studies, comprising 963 samples, and covering ten different human microbiomes including fecal, oral, lung, skin and circulatory system microbiomes. We found that whilst the human microbiome is one of the most extensively studied, on average 2% of assembled sequences have not yet been taxonomically defined. However, this proportion varied extensively among different microbiomes and was as high as 25% for skin and oral microbiomes that have more interactions with the environment. A rate of taxonomic characterisation of 1.64% of unknown sequences being characterised per month was calculated from these taxonomically unknown sequences discovered in this study. A cross-study comparison led to the identification of similar unknown sequences in different samples and/or microbiomes. Both our computational framework and the novel unknown sequences produced are publicly available for future cross-referencing.

Conclusions

Our approach led to the discovery of several novel viral genomes that bear no similarity to sequences in the public databases. Some of these are widespread as they have been found in different microbiomes and in studies. Hence, our study illustrates how the systematic characterisation of unknown sequences can help the discovery of novel microbes and we call on the research community to systematically collate and share the unknown sequences from metagenomic studies to speed up the rate at which the unknown sequence space can be classified.

Background

Metagenomics has become an increasingly mainstream tool to catalogue the microbial makeup of any given habitat [1–4]. It has been applied to a diverse range of environments from human body sites [5–8] to the depths of vast oceans [9–11]. Metagenomics, compared to culture-based methods, provides a relatively unbiased approach to observe, measure and understand the interactions of the microbes within communities as well as with their hosts [3]. Underpinned by relatively cheap sequencing costs and providing powerful insights, metagenomic has become a routine technique to study the microbial content of any environment [2].

These advances in sequencing technologies and the importance of data sharing for reproducible research have led to the rapid expansion of publicly available sequence data. This has led to rapid growth in online sequence databases such as GenBank, that store nucleotide and protein sequence data from various organisms [12, 13]. However, although the raw sequences generated as part of metagenomic experiments are made publicly available through the Short Read Archive (SRA) or European Nucleotide Archive (ENA) repositories, the complete set of assembled contigs from a study are rarely submitted to online databases [14]. The reason for the absence of this type of data can be associated with the sheer number of contigs generated and the requirement for sequences to be annotated before their submission, which is difficult when the organism the sequence came from is unknown, and when the number of contigs is large. Additionally, taxonomically unidentifiable contigs are typically discarded and excluded from downstream analyses (Figure 1(a)), but such sequences represent novel, and potentially widespread biological entities and cataloguing their sequences and where they are found will aid taxonomic classification and our understanding of their biological nature in the future.

The raw data in public databases are typically analysed using metagenomic protocols designed to address specific biological questions. There is a range of different tools and pipelines available for metagenomic sequence analysis, but there are limited comparisons of these pipelines as they are usually developed to address specific research questions. For example, there are approximately 50 workflows available for virus metagenomic analysis that were used in different publications with primarily different aims [15]. As part of the routine metagenomic analysis, only the contigs that can be classified using a specific workflow and that are of interest to the scientific study are typically submitted to sequence repositories such as GenBank. The current approaches used for metagenomics extensively rely on similarity searches to known organisms and proteins, thus, suffers from the street light effect i.e. observational bias which occurs when people only search for something where it is easier to look. However, in a typical metagenomic data set, a range of assembled contigs cannot be functionally or taxonomically classified, a large proportion of which, even after excluding spurious contigs, bear no functional or sequence similarity to any known sequences and are often referred to as unknown or ‘dark’ sequence matter [16–19]. Although the terminology itself has been controversial [19, 20], it typically refers to the sequences of unidentified taxonomic and/or functional origin (figure 1(b)). Generally, these unknown contigs

(UCs) are excluded from downstream analyses. However, a number of recent studies have highlighted the importance of identification and categorisation of such unknown sequences [17, 19, 21, 22].

Characterisation of metagenomically assembled genomes (MAGs) as microbial origin has strengthened the hypothesis that uncharacterised biological sequence matter is highly likely to belong to uncultured or unculturable bacteria, archaea and viruses present in the microbiome sampled [4, 17, 19, 23]. A study by Almeida *et al.* [24] mined over 11,850 human gut microbiome data sets and identified nearly 2000 novel uncultured bacterial species from 92,143 genomes assembled from metagenomics data sets. Similarly, another focusing on multiple human biomes assembled 150,000 microbial genomes from 9428 metagenomic data sets [25]. The MAGs generated from these studies were consolidated to create a unified catalogue of 204,938 gut microbiome reference genomes [26]. A range of different data mining studies have led to the identification of novel microbes, including the identification of novel bacterial and archaeal phyla and superphyla [17, 27].

Previous studies have shown that sequences of unknown lineage and unknown functions tend to be of viral origin [16]. For example, a computationally identified phage crAssphage has been shown to constitute approximately 1.7% of all fecal metagenomic sequences [28]. A study by Roux et al. [21] mined 14,977 publicly available bacterial and archaeal genomes and identified 12,498 completely novel viral genomes linked to their hosts. Kowarsky et al. [29] found that 1% of cell-free DNA sequences appear to be of non-human origin in human blood samples and only a small fraction of them can be mapped to currently known microbial sequences. Despite this, the characterisation of unknown sequences in publicly available data repositories remains an ongoing challenge in microbiome research [4] and the identification of viruses in such UCs remain an even greater challenge due to the absence of a universal gene signatures and the high diversity in virus genome content [30]. Overall, this highlights the widespread existence of potentially novel viruses and bacteria in the currently available sequence data sets, and that a systematic method to identify and catalogue them, especially in human data sets, would be extremely useful. The European Bioinformatic Institute (EBI) has developed MGnify that enables researchers to analyse their data using a standard metagenomic workflow [31, 32]. Similarly, there have been other community initiatives developed to forward this field of research [31, 33–38]. Here, we have focused on the development of a robust, portable and reproducible analyses framework that aims to identify and quantify the UCs in different microbiome samples.

In this study, 1) we develop a framework to quantify the unknown sequence matter in human metagenomic data sets; 2) we compare the unknown sequences between samples, studies and microbiomes to determine whether these sequences are likely to be of biological origin and whether they are broadly distributed and 3) we compare the unknown contigs to currently known sequences in GenBank over the period of the study to determine the rate at which these unknown contig sequences are being taxonomically classified. All unknown sequences and associated metadata have been made publicly available for the research community and the original submitter.

Methods

This study includes the data sets available within the EBI MGnify resource. All human microbiome studies submitted to ENA which were included in the MGnify databases were downloaded with the corresponding metadata on 19 April 2019. In order to obtain detailed metadata, each study was linked to the corresponding SRA repository using NCBI e-utilities [39]. As the focus was on shotgun metagenomic data sets, studies targeting metabarcoding-based sequencing methods such as 16S and amplicon sequencing were excluded as well as studies that solely focused on third party annotation i.e. analysis of previously published data and lack primary data were also excluded. In order to reduce sequencing technology-related bias, the studies that utilised sequencing platforms other than Illumina were excluded. Very large studies involving >100 samples were discounted in order to get a cross-section of different human microbiomes and geographical locations whilst keeping the overall data set size manageable. The filtered set initially comprised 44 distinct studies with 1130 samples of which 1121 were available to download. A script that uses parallel-fastq-dump [40] was developed to download reads in fastq format. In total, 1121 samples (789 paired-end [PE], 332 single-end [SE]) from 43 distinct studies were successfully downloaded and submitted to the pipeline. Out of 1121 samples, 158 could not be assembled due to insufficient reads and were excluded from downstream analysis (see supplementary method). In summary, 963 (784 PE, 179 SE) samples from 40 distinct studies were included and were processed using the complete metagenomic analyses pipeline described below (figure 2).

This study set included a range of different sample types as described in the figure 3. It is important to note that this set is highly skewed towards the human gut metagenome that is normally sampled through fecal material and the oral microbiome was the second most common sample type included in the study. Although other metagenomes were under-represented, our study covered a wide range of samples from various human bodily sites and fluids. A miscellaneous metagenome labelled only as ‘Human’ was included in this data set that represents 3 distinct studies including PRJEB14301 (CSF, n=1), PRJEB21827 (A/B testing for colon model, n=12) and PRJEB6045 (metagenomics of medieval human remains from Sardinia, n=1).

In order to assess the quality of the samples and remove sequencing adaptors, all samples were processed through BBDuk from BBTools package [41]. BBDuk auto-detected the presence of the relevant adapter sequences from the input files specified and trimmed them. Additionally, commonly known sequencing contamination and spike-in sequences were also removed as part of this QC step. All reads that pass QC were retained and mapped to the human genome sequence build GRCh38 using the Burrows-Wheeler Aligner (BWA) [42], and unmapped reads were subsequently extracted using SAMTools [43]. BBNorm [41] was used to normalise reads based on the kmer coverage composition with a kmer threshold of 3 (`mindepth=3`). This step also enabled the acceleration of the assembly process as only a subset of reads were used to build the *de novo* assembly and resulted in better assembly quality overall [44]. The read lengths varied widely between the samples and the studies, thus it was not possible to compare the quality metrics using the read-based measures as it would be misleading. To enable a comparison, quality assessment metrics were carried out for a number of bases.

De novo assembly and taxonomy label assignment

The normalised reads were *de novo* assembled using the SPAdes [45] assembly pipeline, with the default parameters. A script was developed to extract contigs that were longer than 300 bases as short contigs do not contain a lot of information and they were excluded from downstream analysis as a precautionary measure. Although the normalised subset of reads was used to generate assemblies, these reads cannot be used to assess the assembly quality as they represent a small subset of the actual reads. To assess the assembly quality, the complete set of reads that did not map to the human genome were mapped onto the *de novo* assembled contigs with BWA [42] using the default parameters. The assembly quality statistics such as coverage, length, number of mapped reads were generated for each contig using `pileup.sh` from BBTools package [41].

Contigs were searched against the GenBank non-redundant (nr) protein databases using the BLASTX algorithm implemented in DIAMOND [46]. It carries out a six-frame translation of the nucleotide sequences and then searches those translated sequences against the nr protein databases. This step enables the identification of distantly related homologues of the currently known sequences. The default DIAMOND tabular output format with additional columns ‘qframe, staxids stitle’ was generated for aligned sequences. The top 25 hits for each contig were extracted and analysed downstream (`--evalue=0.001`). The Lowest Common Ancestors (LCA) of these hits was computed and superkingdom was assigned based on the LCA using the Python ete3 package [47]. The contigs that did not have any protein match were extracted and searched against the GenBank comprehensive nucleotide database (nt) using BLASTN (`--evalue=0.001`); BLAST output format 7 with additional columns ‘qframe, staxids, stitle’ was generated. This step helped to identify and remove non-coding sequences such as ribosomal RNA and untranslated regions of currently sequenced organisms included in the databases.

To identify the geographical distribution of the raw data, location data was mined from the SRA metadata resources using pysradb [48] for each study. Geo-location information was available for 861 samples as shown in figure 4 . A complete list of study location is shown in the supplementary table S1. These samples were sequenced in various sequencing facilities across the world, and the complete distribution of the sequencing centre is shown in the supplementary figure S6(a).

Unassembled sequences

'Unassembled' bases are defined as bases from reads that did not map to the human genome and could not be assembled into contigs. These were calculated from reads that did not map to assembled contigs. These bases/reads could not be classified as part of this project but were quantified as shown in the figure S6(c) - grey bars. Our quantification suggests that almost all microbiome samples have a proportion of unassembled sequences and on the sample, the average value for this is around 23.91% (std: 26.59%). This unassembled sequences proportion was very high for samples originated from PRJEB15334 (mean: 51.17%, max: 97.67%, std: 24.30%) and PRJEB17784 (mean: 82.59%, max: 98.83%, std: 17.84%). Overall, 8.18% of all data fell into this category as described in figure S6(b). A range of possibilities from degraded nucleic acid to sequencing protocols could lead to poor quality data that cannot be used for *de novo* assembly.

Control samples

The Human Microbiome Project mock community samples(n=9) were downloaded for study PRJNA298489 and were analysed using the metagenomic framework described above for quality control and workflow assessment. This would also allow us to validate the metagenomic analyses pipeline for this study.

Post metagenomic analysis

All unknown contigs (UCs) were analysed further to get insights into the coding potential of those sequences. `getorf` tool from EMBOSS [49] suite was used to generate open reading frames (ORFs) from contigs (`-find 1, -minsize 300`) using the standard genetic code. These ORFs were searched against a range of different domains and functional identification databases included in the InterProScan.

To explore the sequence similarity between samples and the diversity of the unknown sequences, a nucleotide-based sequence similarity clustering which also used coverage was carried out using MMSeqs2 [50, 51]. All sequences with at least 90% sequence identity and at least 80% overlap were clustered using the MMSeqs2 `easy-cluster` pipeline [51]. All UCs were processed through CheckV [52] pipeline to identify the UCs that were likely to belong to viruses.

The most widely applied sequence similarity-based approaches rely on static versions of the databases to carry out the classification step of the analysis. In this study, the sequence databases utilised were downloaded on the 18th of April 2019. All results included in the study are based on the searches against this static version of the databases. However, the sequence database is ever-expanding with new sequences being added to the databases each day. With newer sequences being added to these databases, it is very likely that unknown sequences transition into the "known sequence space" over time. In order to identify the proportion of the unknown sequences classified over the period of the study, 4 distinct time points were considered. Static versions of the databases were downloaded on 31 October 2019, 5 March 2020 and 14 October 2020.

To predict the proportion of UCs that are likely to be viruses, the virus prediction tool DeepVirFinder was used. DeepVirFinder has been demonstrated to accurately predict viruses from metagenomic datasets and has been shown to work well even with short contigs [53]. It was deemed suitable for UCs as a large proportion of UCs identified in this study are under 1kb long. DeepVirFinder was run on all UCs with default parameters and q-values (false discovery rate) were computed for the predictions using the R library q-value as recommended in the DeepVirFinder tutorial. The q-value output was rounded to 3 decimal points and a cut-off of q-values <0.05 was applied.

In order to identify if the UCs captured in this study have any overlap with other uncultured virus databases such as IMG/VR [36], initial nucleotide (BLASTN) and protein sequences-based (BLASTX, BLASTP in DIAMOND) searches were carried out against nucleotide and protein sequence data downloaded for the latest IMG/VR version 2020-10-12_5.1. BLASTN searches were carried out with default parameters except for the evalue which was set to 0.0001 and the output was generated in standard tabular format. For BLASTP searches, predicted ORFs were used.

Results

To quantify the presence of unknown sequences in human metagenomes, datasets included in the EBI MGnify were filtered to select for metagenomic data sets sequenced on the Illumina platform (see Methods). A set of 963 samples from forty studies covering ten different microbiomes were downloaded from SRA repositories and analysed using the framework described in the Methods in order to characterise and quantify the unknown sequences in these samples. The studies included a total of 2.08×10^{12} bases of raw sequence data that was derived from a range of human microbiome studies including the following microbiomes (figure 3(a)): (1) circulatory system (n=2) (2) fecal (n=20) (3) lung (n=1) (4) oral (n=5) (5) pulmonary system (n=1) (6) saliva (n=3) (7) skin (n=2) (8) sputum (n=2) (9) vagina (n=1) and (10) human (n=3; miscellaneous). Geo-location information available for 861 of these samples shows that the data sets are globally distributed, but skewed towards western Europe (figure 4 and figure 3(b)). All samples were individually processed through the metagenomic analysis framework designed in this study (see Methods). The framework included an individual sample-based *de novo* assembly step resulting in a total of 44,238,374 *de novo* assembled contigs, 28,505,777 of them were longer than 300 nucleotide. Out of this set, 7,155,624 contigs were at least 1kb long, 970,507 were at least 5kb and 415,719 were at least 10kb long. The largest assembled contig was 1,380,230 bases long and was found in the human gut microbiome sample ERR505090. These contigs were then systematically processed by our metagenomic framework for BLASTX sequence similarity classification against the GenBank non-redundant protein database. Sequence similarity thresholds were used to sort the contigs into three classes: known (>80% similarity to a known protein sequence), partially known (>0 and <80% similarity to a known protein sequence), and unknown (no similarity to any existing sequence).

In total, 25,148,829 (88.22%) contigs were classified as known contigs whilst 2,517,700 (8.83%) of all analysed contigs were classified as partially known. The remaining sequences, referred to as unknown contigs (UC), are sequences that did not bear significant similarity to known sequences in the databases. Overall, 651,529 (2.29%) of contigs did not match any currently known sequences using our approach and were categorised as UCs. On average 1.3% of assembled bases per sample were found to be unknown. The proportion of unknown varied significantly between different assembled metagenomes as shown in the figure 5(a). Samples from some microbiomes such as the circulatory system did not contain any unknown sequences compared to the skin microbiome where this proportion was up to 25.85% for some samples.

The UCs varied largely in length and most of the UCs were 300-1000 nucleotides long (figure 5(b)). 95.36% (n=621,302) of all UCs were shorter than 1kb and 4.59% (n=29,879) UCs were between 1-5kb long. A set of 320 UCs fell within the 5-10kb length category and 28 UCs were >10kb long. The largest UCs was 42.3kb long and the second largest UCs was 21.3kb long. A complete distribution of UCs across different microbiomes is shown in the figure S1 that shows that the largest UCs were assembled from fecal, oral and saliva microbiome.

To understand the coding potential of the unknown sequences, open reading frames (ORFs) were predicted. 273,590 ORFs that were at least 100 amino acid in length were generated using the standard genetic code. A threshold of 100 AA was selected, this is similar to that used in the taxonomic classification tool GRAViTy which demonstrated only a 5-10% gene loss at this cutoff for viral sequences [54]. These ORFs originated from 215,985 distinct UC, showing that 33.15% of all UCs contained large ORFs. On average, ORFs were 157 amino acid (AA) long with a standard deviation of 87 AA residues. The longest ORF was 6,898 AA long. This set also included 2,713 ORFs with length of at least 500 AA and 256 that were at least 1000 AA long.

A detailed protein domain analysis for these ORFs was carried out using the InterProScan [55] protein analysis software. This tool searches the domain and functional signature of amino acid sequences against a range of distinct domain databases including Pfam [56], CDD [57] and SUPERFAMILY [58]. 36,354 ORFs originating from 35,760 UCs could be functionally annotated using the InterProScan analyses, this number excludes hits to MobiDBLite and Coils databases as they predict disordered regions and coils structure of predicted ORFs as opposed to the domain signatures. An overview of the number of hits found to various InterProScan databases for each microbiome is shown in the figure S2(a). The most number of hits were found in the MobiDBlite [59] - a database that can predict the intrinsic disorder regions in the proteins. Overall, 5.49% of UCs (n=35,760) contained ORFs (n=36,354) with at least one identifiable domain. The functional classification of the ORFs was prominently centred around the Pfam database resource [56]. Pfam databases facilitate the domain-based searches against the set of protein sequences using Hidden Markov Model profiles. These types of searches can identify distantly related protein sequences. 16,839 ORFs originating from 16,705 UCs were found to match at least one Pfam entry and in total, 27,025 Pfam hits were derived (figure S2(a)) All Pfam entries were collapsed down to their corresponding protein clans (grouping of related protein families) by mapping the Pfam IDs back to their clan membership. Figure S2(b) shows a heatmap of top 50 Pfam clans with hits to UCs ORFs predicted in different metagenomes. The most abundant hits were identified to clans tetratrico peptide repeat superfamily and leucin rich repeats. The largest number of hits was found in the fecal microbiome due to the high number of fecal microbiomes included in this study. Additionally, a range of other protein clans including those that represent Helix-turn-helix, beta-strands, polymerase and nuclease proteins were also found in this set. These results illustrate that the UCs sequences have known protein domains suggesting that these unknown sequences are functional and belong to organisms that are not yet fully sequenced or taxonomically classified.

Unknown sequence clustering

To investigate the extent of sequence diversity and to identify UCs sequences present in multiple samples and microbiomes, sequence clustering was performed. MMSeqs2 [51] generated 464,181 clusters of which 377,855 were singletons i.e. did not cluster with any other sequences. These singletons were excluded from the cluster analysis described below. 86,326 clusters comprised two or more sequences with a mean cluster size of 5.7 contigs and a standard deviation of 8.1. Cluster representatives were extracted from MMSeq's clustering output which are the longest sequences in the cluster. The largest cluster contained 153 sequences which originated from the fecal microbiome from 8 distinct BioProjects (figure S5(c)). A cluster size distribution across different microbiomes is shown in figure 6 and a detailed cluster size distribution with cluster representative length is shown in the figure S3). 89.42% of 273,674 UCs (n=244,730) were clustered into single microbiome clusters, 10.58% UCs (n=28,944) were found in clusters that contained sequences from two or more microbiomes. To compare that with specific studies, 39.4% UCs were clustered into BioProject specific clusters and the remaining 60.6% UCs (n=165,851) were grouped into clusters originating from two or more BioProjects. 78,139 (90.52%) clusters contained sequences from a single microbiome and 7,645 (8.86%) clusters included sequences from two microbiomes. Only a few clusters were comprised of members from 3 (n=512) or 4 (n=30) microbiomes. The largest multi-microbiome cluster contained 57 sequences (304-9,080 bases long) from 4 distinct microbiomes and BioProjects and contigs assembled from 12 samples. The largest single microbiome cluster contained 153 sequences (6,640-300 bases long) from fecal microbiomes with contigs assembled from 46 distinct samples covering 8 different studies. Overall, this clustering method produced very small, study-specific clusters. A set of 464,181 UCs was obtained by combining the cluster representative sequences with the unclustered singleton

UCs and used to determine the rate at which UCs are classified.

Classification of unknown over time

In this framework, the unknown sequence identification is dependent on the publicly available nucleotide or protein sequence databases. These data repositories are updated regularly with new sequence data being deposited from around the world. However, typically, the sequence searches are carried out against static versions of the databases. Our analysis conducted against the databases downloaded on 18 April 2019 identified 651,529 UCs that were collapsed down to a set of 464,181 UCs following the cluster analysis. Subsequent analyses on 31 October 2019 and 5 March 2020 produced a set of 613,726 and 558,711 UCs respectively. The final number of sequences that still lacked a taxonomy label was down to 459,147 after the most recent analysis carried out against the databases downloaded on 14 October 2020. 29.5% (n=192,382) of the sequences compared to the initial set of unknown matched to at least one sequence from the updated databases in the BLASTX and the BLASTN steps of the analysis. Similarly, 27.6% (n=128,288) of the representative set sequences could be labelled taxonomically with the updated databases. A rate of taxonomic characterisation of 1.64% of unknown sequences being characterised per month was calculated from the complete set. This rate was estimated to be 1.54% for the representative set. Moreover, as shown in the supplementary figure S4, a range of long UCs still remained unknown even after the similarity sequence-based analysis carried out on 14 Oct 2020.

From a set of 192,382 contigs that were labelled taxonomically after the most recent analyses carried out on 14 Oct 2020, 167,864 were identified using BLASTX and 24,518 were identified using BLASTN. 106,739 UCs from the BLASTX classified set were categorised as known and 61,125 contigs were categorised as partially known. A large majority of these contigs (97.11%, n=162,987) were also deemed to be bacterial. The remaining contigs were divided between cellular organisms (n=2,104), archaea (n=930), viruses (n=858), root (n=827) and Eukaryota (n=140). 76.55% of all BLASTN hits were matching to bacteria (n=18,768), 17.88% matched to viruses (n=4,383), 1.99% matched to Eukaryota (n=487) and 0.03% archaea (n=7). The hits that could not be mapped to a superkingdom and were divided between unidentified plasmid (n=544), root (n=294), cellular organisms (n=20), uncultured organisms (n=14) and synthetic construct (n=1). These results reiterate our initial hypothesis that the majority of UCs represent currently unknown microbial genomes.

Viral domain signature identification

195 UCs were shown to contain a virus-specific functional domain which was parsed using the term ‘virus’ or ‘viral’ in the InterProScan analysis signature description column. Results with the term ‘phage’ were not included in this subset as a range of phage domains are also

present in the host bacterial genomes. These domains were predominantly identified using the Pfam (n=125) analysis. The most abundant virus-specific domain was Vaccinia Virus protein VP39 and it was found in 53 UCs derived from fecal (n=23), saliva (n=14), oral (n=12), sputum (n=1) and human (n=3) microbiomes and it was identified by Gene3D analysis. The largest UCs containing this domain was 3,661 bases long and was found in sample ERR1474567. Another frequently found domain in the UCs was podovirus DNA encapsidation protein Gp16 domain. It was found in 25 UC, out of this set 23 UCs were assembled from fecal microbiome. The largest UCs containing this virus-specific domain was 9kb long contig shown in figure 8(a), assembled from PRJEB18265. These UCs was clustered with 24 other sequences (See Unknown sequence clustering) that were assembled from 11 samples representing 5 distinct fecal microbiome studies. These results indicate that these UCs represents a completely novel genome of a virus that is likely related to currently known podoviruses.

The largest UCs containing a viral RNA dependent RNA polymerase (Pfam: PF00680) domain was found in the sputum microbiome sample ERR1022511. This UC was 5,894 bases long and contained seven ORFs that were at least 100 AA long (figure 7). A 269 AA long ORF contained ATPase P4 of dsRNA bacteriophage phi-12 (Pfam: PF11602) domain suggesting that this UCs represents the large segment of a novel double-stranded RNA phage which are usually categorised in the virus family *Cystoviridae*. The genomes of these phages are composed of three linear dsRNA segments with a total genome length of 12.7–15kb and all segments code for various proteins [60]. Although several other UCs were found in the same sample, none of them displayed any sequence or functional similarity to the other two segments i.e. small and medium segments of Cystoviruses. However, UCs that could potentially belong to novel cystovirus-like genomes were extracted based on the sequence length, GC content and sequencing depth criteria. Moreover, this UC representing a potentially novel relative cystoviruses did not match to any known protein or nucleotide sequences even in the most recent analyses confirming the discovery of a novel virus.

Virus prediction and comparisons to uncultured virus databases

From the complete set of the UCs, 323,395 (49.64%) UCs were predicted as viruses by DeepVirFinder (see figure S7(a)). This set included 300,271 UCs that were under 1kb long which represents 48.33% of UCs identified in this length category. A number of larger contigs were also predicted as viruses: 76.27% (n=22,788) of UCs in the 1-5kb length category and 96.55% (n=336) of UCs in the 5-50kb category. These results strongly support our hypothesis that the large majority of the UCs are of virus origin, albeit a large proportion short UCs are likely to be fragments of unknown viruses.

These predicted virus sequences (n=323,395) were clustered with other known and partially known sequences using MMSeqs with 90% sequence similarity across 80% of the sequence. 50.18% (162,271) of UCs were either singletons or were clustered with other UCs, whilst the remaining 49.82% (161,124) of UCs were clustered with known and partially known. However, a large proportion (n=152,295; 94.52%) of the UCs that clustered with these were shorter than 1kb. 8,829 UCs (out of 22,788; 38.74%) were at least 1kb long among which 1,402 UCs (out of 4,419; 31.73%) were at least 2kb long, 75 UCs (out of 336; 22.32%) were at least 5kb long and 5 UCs (out of 28; 17.86%) were at least 10kb long. Moreover, 47.52% of sequences that match the UCs were deemed partially known (i.e. had a protein sequence hit with <80% sequence similarity) in this analysis suggesting that these known and partially known sequences are still significantly divergent from those present in the databases.

To identify the "known unknowns" i.e. uncultured viruses categorised as UCs in this study and also observed in previous meta-analyses, the IMG/VR databases were used as a reference and the UCs were searched against the nucleotide and protein repositories. 182,293 (27.98% of all UCs) UCs had at least one hit to uncultivated viral genomes (UViGs) included in the IMG/VR using BLASTN and 175,372 (26.92%) UCs were found to match at least one UViGs using the BLASTX approach (figure S7(b)). Out of the 273,590 predicted ORF set, 85,852 ORFs were found to match protein sequences included in IMG/VR. 64,779 (9.94%) of UCs were found to match the uncultured viruses in IMG/VR using all three approaches.

The large unknown contigs

All UCs described in this section were predicted to be viruses by DeepVirFinder and did not cluster with known and partially known sequences. The largest UCs was assembled from the saliva sample ERR1474583 and was 42,357 bases long. This contig did not cluster with any other contigs and has 23 ORFs that were over 100 AA long. One of the ORFs that is 434 AA long comprised of the cysteine proteinases domain (SUPERFAMILY: SSF54001) according to the InterProScan analysis. This contig still remained unknown after searches against the most recent version of the databases suggesting that the organism this genomic sequence belongs to is still to be identified and fully sequenced. A snapshot of the ORFs and domain is shown in figure 8(b), highlighting the presence of coding regions across the entire length of the UCs sequence. Based on the results we have obtained here, we predict that this UCs sequence is likely to be of microbial origin as it lacks a non-coding region. CheckV analysis predicted it to be a viral genome fragment with the presence of two identifiable viral genes albeit with low quality as per the MIUVIG [61] standards due to the lack of similarity to any known sequences. This strongly suggests that this UC can potentially be a representative or partial genome sequence of a currently unknown and completely novel virus.

A 20,309 nucleotide long contig from saliva sample ERR1474612 clustered with two very short contigs from the same study. As shown in figure 8(c), long ORFs were predicted across the whole sequence. Some of the predicted ORFs were found to have interesting domain signatures (figure 8(c)) such as enzymes for nucleic acid replication e.g. polymerases. An ORF that is 655 AA long shows the presence of DNA dependent RNA polymerase domain (SUPERFAMILY:

SSF64484). A CheckV [52] analysis of the contig also predicted it to be of viral genomic origin, however, it was predicted to be an incomplete genome. This UC was shown to have a very low identity ($<30\%$ sequence identity with 2% of query coverage) to a hypothetical protein of Firmicutes bacterium (HAB66316.1) and AAA family ATPase from Sharpea azabuensis (23% sequence similarity). When the e-value threshold was removed, a total of 8 BLAST hits were obtained and 3 out of 8 hits were to a range of phages including Bacillus phage vB_BpuM-BpSp, Vibrio phage 2 TSL-2019 and Ralstonia phage RP12. These hits range from hypothetical and putative proteins. All these matches were localised to a short region between 8,217-8,915 which was shown to contain ATPase and P-loop containing nucleotide triphosphate hydrolases domains (figure 8(c)). Notably, no nucleotide sequence hits were identified for this UC. Although these results have bacterial hits, it is likely that this UC represents a complete or partial genome of a novel phage that infects the host bacteria e.g. firmicutes.

Short circular contigs

A range of circular contigs with direct terminal repeat (DTR) and inverted terminal repeat (ITR) signatures were identified using CheckV in the UCs data set. A total of 1,839 containing repeat signatures were predicted of which 1,771 contained DTR signatures and 68 contained ITR signatures. 94 of these UCs were at least 1kb long suggesting circular genomes and 48 of them contained a range of 55 bases long terminal repeats. A cluster of 8 sequences from 2 different microbiomes and studies were identified to contain similar sequences (71-100% similarity) assembled from different samples (table 1). Four cluster members were 2,110 bases long, one sequence was 1,983 nucleotides long and the cluster representative was 3,165 nucleotides long. The cluster representative sequence contained multiple copies of the same ORFs suggesting the presence of multiple genome copies, sequencing error or miss-assembly. Most of these sequences contained a 50 bp long DTR sequence signature ‘GTGCATTTTTTTGTGCACTTTTTCAAAAAAAC-CGTGAAAAAAATTCATT’. These contigs contained two distinct ORFs, which were 125 AA and 144 AA long. Similarly another 50 bases long DTR signature ‘AATGAATTTTTTTCACG-GTTTTTTTGAAAAAGTGCACAAAAAAAATGCAC’ was observed in another cluster that had 7 member sequences ranging in similarity from 31 to 100 percent and assembled from 7 distinct samples. All but one member were 1,770-1,771 bases long. These contigs also contained two ORFs that were 102 AA and 106 AA long. These ORFs did not match any existing protein sequences in the databases. These circular contigs were assembled from a range of oral microbiome samples from study PRJNA230363. Similarly, a range of contigs (n=9) that contained Inverted Terminal Repeats (ITR) were also identified in this data set. A cluster of 5 distinct circular contigs assembled from distinct samples from the fecal microbiome (PRJEB7949). Four out of five of these circular contigs contained the ITR sequence ‘CGAAACGATTGCCCAGA-GAGATGACTGTCAATCCGCCCGATTATTGGGCGCTTAC’. They also contained a 138 AA long ORF. These short circular UCs did not bear any sequence or functional similarity to known sequences or domains so their biological origin is difficult to predict. However, based on their genome organisation and size distribution, we predict that they are likely to represent either novel circular replication-associated protein (Rep)-encoding single stranded (CRESS) DNA virus groups or novel satellite virus-like groups. 16 out of 20 UCs described in table 1 were predicted to be viruses by DeepVirFinder (see: Virus prediction and uncultured virus databases).

Control samples

The HMP mock community samples (n=9) were downloaded for study PRJNA298489 and were analysed for quality control and workflow assessment. These are control samples that are not expected to yield UCs, but if they do, those UCs could be due to sequencing/assembly error or common lab contaminants. Out of the complete set, four HMP samples did not contain any UCs as expected whereas SRR2726666, SRR2726669 and SRR2726672 contained one UC each but their lengths were short varying from 323 to 449 bases. The remaining two samples SRR2726670 and SRR2726671 contained 28 and 18 UCs each. The largest UC assembled in the mock sample was 3,965 bases long and was found in SRR2726670, only 3 UCs were >= 1kb. These UCs were searched against the most recent version of the databases downloaded on 14 Oct 2020 and only 8 short contigs; 4 from SRR2726670, 2 from SRR2726671 and one each from SRR2726666 and SRR2726669 remained in the UCs category. These remaining UCs were only 330-513 bases long. These results validate the UC analysis framework developed here and highlight that even in control samples, there are a very minor number of short UCs to be found. New sequence data gets uploaded to public repositories daily and these updated databases contain a greater diversity of sequences most of which are taxonomically classified. Therefore, UCs identified in the initial analysis of these mock samples were subsequently found to match to a known sequence in the updated version of the database as more sequence data was available and classified.

Resources

We have developed a modular metagenomic and unknown sequence analysis framework using the sophisticated pipeline management tool Snakemake. Our analysis pipeline takes advantage of portability and flexibility offered by Python, BioPython and Snakemake tools which allow reproducible analysis of large meta-omic data on any processing servers and clusters. The framework developed here is capable of utilising multiple cores enabling users to analyse large data sets in a parallel fashion. Results and code generated in this study is available on https://github.com/sejmodha/UnXplore. All assembled unknown sequences generated here are submitted to ENA as third party annotations and are accessible under BioProject PRJEB41812. This allows the UCs data to be properly linked to its original samples and studies. A consistent data labelling scheme is utilised across all studies and samples. For traceability, all UCs fasta identifiers start with SRA sample identifier. All ORFs contain the exact same naming scheme with a suffix '_' and ORF number starting with 1. A complete metadata table is provided to link any new sequence data to its corresponding BioProject and sample. Functional domain predictions and clustering results are annotated with relevant metadata and provided in a tabular format.

Discussion

In this study, we have developed an automated framework that can systematically quantify the proportion of unknown contigs (UCs) in meta-omics samples. Whilst the presence of UCs is well recognised, this is the first study that addresses the question of UCs comprehensively and quantifies it across different human microbiomes. Our approach utilises sequence similarity-based taxonomic categorisation to identify the sequences that cannot be categorised. We define these UCs as the sequences that do not match known sequences in the databases with a predefined sequence similarity threshold of evalue 0.001 which is a very lenient threshold, anything with evalue higher than this is unlikely to truly be related to the database sequence hit. We show that on average 2.29% of assembled contigs are categorised as unknown in different human microbiome studies. Moreover, a subset of those with unknown sequences could be translated and contained protein domains, thus we were able to find functional similarity to 5.49% of taxonomically unknown contigs. We have generated a comprehensive catalogue of 651,529 UCs that do not bear any sequence similarity to sequences present in the widely used GenBank protein and nucleotide databases. Although sequence similarity-based approaches are dependent on the databases, the protein sequence-based approached implemented here is highly effective in fishing out distantly related homologues of known sequences available in the databases [62] and thus provides better resolution for sequence classification compared to those solely based on the genomic signature-based binning [63]. This study highlights the importance of avoiding the “street light" effect i.e. observational bias arising from classifying metagenomic sequences on the basis of related sequences that already exist in the databases. Here, we have aimed to eliminate such observational bias by performing a comprehensive data mining of the human microbiome data and cataloguing the UCs, their frequency in different human microbiomes and their overlap between different samples.

This study has enabled the identification of a range of genomic sequences that are hypothesised to belong to currently uncharacterised organisms that are often found in similar samples and/or microbiomes. A range of large UCs with and without known protein domains are presented here. However, the complete set includes a large number of UCs that still remain unknown and can be mined further to study their biological origin. A third of all UCs (n=215,985) contained large predicted open reading frames (at least 100 amino acid long) that were predicted using the standard genetic code. Using alternative genetic codes may expand this set further by revealing novel, potentially different open reading frames generates from the UCs. A small proportion of these open reading frames contained domain signatures confirming the presence of currently unidentified organisms. Moreover, a comprehensive clustering analysis has led to the identification of UCs that were present across different human microbiomes (as well as from different samples/studies investigating the same human microbiome) indicating that we have discovered potentially widespread and as yet unclassified novel biological organisms within the human microbiome. The multi-microbiome clustering approach applied here provides an interesting way to understand the diversity and the distribution of the UCs across different microbiomes and geographical sites. For example, this approach led to the identification of 30 clusters that spanned 4 distinct microbiomes. The largest multi-microbiome cluster comprised of 57 UCs recovered from saliva, sputum, oral and lung microbiomes and were assembled from 12 different samples. Although it is impossible to identify the true clusters present in the data due to the novelty of the UCs, the clustering approach helps to identify obvious patterns of sequences similarity between microbiomes and studies. This approach provides an additional dimension by capturing unknown sequences that are shared between different projects or human microbiomes.

Virus predictions carried out by DeepVirFinder - a machine learning-based virus prediction tool for identifying viruses from metagenomic datasets - have shown that approximately 50% of all UCs are likely to be of virus origin. Additionally, nearly 30% of all UCs identified in this study have an overlap with uncultivated viral genomes currently catalogued in IMG/VR databases. As with most similarity-based approaches, we used an arbitrary threshold for determining a match to the IMG/VR database and thus a match does not mean they are closely related. Interestingly, this study provides an added dimension to these matching uncultivated viral genomes (UViGs) by providing information on the type of microbiome they have been found in. It is anticipated that UCs catalogued in this study may have some overlap with other viral genome databases such as Gut Phage Database [64] and Gut Virome Database [65]. Short contigs i.e. those less than 1-5kb are often ignored in most data mining and exploration research typically in studies that employ a contig binning step as binning has been shown to be less sensitive for short contigs [63, 66, 67]. The clustering and time point analyses carried out on short UCs has shown that these short UCs are originating from biological entities and predominantly represent the novel microbial sequences that are currently uncatalogued. This has been demonstrated with the example of short circular sequences with terminal repeats. Short contigs that are typically excluded from large microbiome mining studies employing the metagenomic binning approach but were studied in detail here. These short UCs are found across multiple human microbiomes and samples, we speculate that these are of viral origin and could potentially represent novel CRESS DNA or satellite viruses, although the ORFs originating from these genomes do not bear any sequence of functional similarity to the typical rep and cap genes. Moreover, a number of large contigs were found to contain various functional ORFs and domains often originating from virus or phages indicating that a proportion of UCs are very likely to be novel viruses that infect currently uncharacterised microbes. In our approach, we have implemented a protein sequence similarity-based identification that enable the identification of distantly related sequence homologues [62]. This approach can potentially 'classify' contigs of viruses or phages as their corresponding host with very low sequence similarity. Indeed, viruses are well known to mimic their host genomic signatures by incorporating genomic sequences from their host into their genome. We anticipate that the virus diversity described in this manuscript is reasonably underestimated due to this specific characteristic of viruses and speculate that a range of assembled contigs classified as bacterial with very low sequence similarity across a short genomic coverage are likely to be of virus origin. This hypothesis will need to be tested further by mining the 'known' and 'partially known' contigs systematically. We note that a range of UCs matching to known and partially known sequences could be taxonomically uncharacterised in GenBank databases such as unclassified viruses. Assembled contigs matching to these sequences are categorised as known (protein sequence similarity >80%) or partially known (protein sequence similarity <80%) in this study. Those contigs would need to be investigated further to identify potentially novel and divergent sequences assembled in this study. The HMP control sample analyses resulted in only a few UCs validating the UC identification approach implemented in our framework. The results generated from this study can be extended to identify the organisms that co-occur in different microbiomes, which in turn can help to inform the interactions between these organisms and how it affects their hosts - humans. Despite having sequenced human microbiomes extensively, our understanding about how these microbes interact with humans remains limited. These large scale explorations can help to understand the human holobionts and the interactions of macro- and microorganisms. Based on these results, we do not know whether the microbes identified in different studies are consistently associated with human or they are just passing association captured at the time of sampling, the latter would make it even harder to make comparisons between samples and microbiomes.

The UCs landscape changes over time as more sequences get characterised and added to the ever expanding sequence repositories. This was demonstrated by comparing the UCs to different GenBank databases over the course of 18 months. We have estimated that 1.64% of the UCs identified in this study are getting characterised each month. However, this number would be highly dependent on the types of data deposited in the International Nucleotide Sequence Database Collaboration (INSDC) resources. This study provides a strong foundation of preliminary estimation of this rate and UCs would need to be analysed at multiple future time-points to determine how the rate at which the UCs are being classified, changes over time. Additionally, the time-point analysis also provides strong evidence of the real biological entities being assembled and characterised in our study. Indeed, a proportion of the UCs were taxonomically classified during the period of the study. This delineation of the UCs demonstrates that the unknown matter that surrounds us largely belongs to currently uncultured, unidentified microbes that we interact with on a daily basis. The technological advances have accelerated the speed at which genomic sequences belonging to novel uncultured organisms are being deposited in INSDC databases.

This sharp increase of metagenomically assembled microbial genomes has led to the scientific community driving the development of genomic data and metadata standards such as MIMAG (for bacteria and archaea) [68] and MIUVIG (for viruses) [61] for consistency and comparison purposes. The taxonomic classification landscape has also faced a tectonic shift whereby it is moving from the phenotype-based classification to more holistic sequence-centric phylogenetic classification, e.g. GTDB (bacteria and archaea) [69] and ICTV (viruses) [70]. These changes enable the incorporation of the uncultured sequence diversity into the microbial taxonomy and will provide a more comprehensive understanding of the complex phylogenetic relationships and interactions between different microbes.

The metagenomics analysis framework developed here works as a proof of concept for overcoming the challenge of the quantification of the unknown in already 'analysed' data sets. The pipeline developed here is flexible and can be applied to any microbiome. To get a cross-section of different human microbiomes and geographical locations whilst keeping the overall data set size manageable large studies involving $>$100 samples were discounted. This framework can readily be applied to routine metagenomic exploration, which can help to gain further understanding of the landscape of sequences of unknown origins. However, the framework applied here is easily portable to metatranscriptomics data. In fact, a couple of the BioProjects (PRJEB10919 and PRJEB21446) analysed in this study were indeed from a metatranscriptomic study. It is important to note that, unlike other studies that often focus on the cross assembly of different samples, each sample was assembled individually here. This is regarded as best practice when a cocktail of samples from unrelated studies are analysed in bulk. The co-assembly would often lead to fragmented assembly as the complexity of sequences originating from multiple samples would be much higher compared to a single sample [71]. On the contrary, independent assembly is expected to capture better diversity across each sample with high-quality genomes assembled from each sample [71]. Typically the sequence similarity-based approach is less reliable for unrelated sequences as the similarity search tools heavily rely on the databases used in the analysis. Like most other pipelines, this framework classifies the sequences with respect to a static version of the reference sequence databases. The search results are as good as the data in the ever-expanding repositories that are often too large to be hosted on a local computer. In order to improve this, an alignment-free approach could be explored. The development of a general purpose alignment-free prediction method that can categorise the sequences based on the genomic composition would be suitable for the downstream analysis of the UCs. The UCs classification is highly dependent on the methods employed to identify and quantify the unknown. Moving away from the sequence similarity-based methods would help to categorise and classify the currently unknown sequences better. Machine learning-based approaches might be deemed suitable in certain circumstances to overcome the similarity threshold-based approaches. In case of completely novel sequences that bear no similarity to currently known sequences, significantly rigorous training sets and features would need to be identified and be built into the models in order to make accurate predictions as machine learning approaches are highly reliant on the training data the models have been developed with. Moreover, a recent study by Krishnamurthy and Wand Krishnamurthy and Wang [72] made predictions for picobirnaviruses to be bacteriophages rather than eukaryotic viruses based on the presence of bacterial ribosome-binding sites in front of the coding sequences. This approach could potentially be applied to check whether viral UCs are bacteriophages.

Conclusion

This study demonstrates that there is a large diversity of unknown sequences embedded within various human meta-omic samples available in public repositories. It is clear that the unknown sequence landscape observed in this study is likely to be the tip of the iceberg, and, as we scan more microbiomes and extend this to less-studied environments e.g. insect metagenomes, we are likely to gather a better understanding of the unknown sequence space. As more species and environments are sequenced more readily, the rate at which the unknown sequences become known would also change. Our results of novel viruses indicate that the unknown microbes and their genomic signatures are likely to be more divergent to those currently present in widely used sequence databases; however, it should be noted that many of the short contigs found in our study are likely to represent fragments of larger viral genomes rather than being short but complete viral genomes. Our study also shows that at least some of these unknown microorganisms are prevalent in nature. To overcome this, more comprehensive resources including searchable databases such as those enabled using BIGSI [73] and federated indexes [74] could be created for the unknown sequence data and metadata. This would allow researchers to explore the human metagenomic sequence space in a more holistic manner and in turn, provide a better understanding of microbial diversity interacting with and within human hosts. It would enable researchers to search, link and explore the unknown sequences present in different microbiomes, studies and samples. Such resources could help in speeding up the pace at which unknown sequences can be 'classified' and make it easier for researchers to determine the functional and/or ecological importance of the organisms the sequence comes from. A concerted effort could help to pin down human-microbial interactions in a broader context such as linking unknown microbes to human diseases and disorders of unknown aetiologies.

Acknowledgements

All authors would like to thank Dr Quan Gu, Maha Maabar, Dr Sreenu Vattipally, Dr Josh Singer, Hilary Fawcett and Prof Andrew Davison for contributing to an internal hackathon organised in 2016 where the idea of mining metagenomic data sets for identification of 'dark matter' was developed.

Funding

SM is funded by a MRC Precision Medicine PhD studentship (MR/S502479/1). DLR, JH and RJO are funded by the MRC (MC_UU_1201412).

Availability of data and materials

All assembled unknown sequences generated here are submitted to ENA as third party annotations and are accessible through BioProject PRJEB41812. Results and code generated in this study are available on Zenodo: https://zenodo.org/record/5907223 and GitHub: https://github.com/sejmodha/UnXplore.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Conceptualisation: SM, JH, RJO. Data curation, Formal Analysis, Project administration, Investigation, Resources, Software, Validation, Visualisation and Writing - original draft SM. Funding acquisition SM, DLR, JH, RJO. Methodology SM, JH, RJO. Supervision DLR, JH, RJO. Writing - review & editing SM, JH, RJO.

References

[1] Vanessa Aguiar-Pulido, Wenrui Huang, Victoria Suarez-Ulloa, Trevor Cickovski, Kalai Mathee, and Giri Narasimhan. “Metagenomics, Metatranscriptomics, and Metabolomics Approaches for Microbiome Analysis.” In: *Evolutionary bioinformatics online* 12.Suppl 1 (2016), pp. 5–16. DOI: 10.4137/EBO.S36436.

[2] Eugene V. Koonin. “Environmental microbiology and metagenomics: the Brave New World is here, what’s next?” In: *Environmental Microbiology* 20.12 (2018), pp. 4210–4212. DOI: 10.1111/1462-2920.14403.

[3] Christopher Quince, Alan W Walker, Jared T Simpson, Nicholas J Loman, and Nicola Segata. “Shotgun metagenomics, from sampling to analysis”. In: *Nature Biotechnology* 35.9 (2017), pp. 833–844. DOI: 10.1038/nbt.3935.

[4] Andrew Maltez Thomas and Nicola Segata. “Multiple levels of the unknown in microbiome research”. In: *BMC Biology* 17.1 (2019), p. 48. DOI: 10.1186/s12915-019-0667-z.

[5] Vincent Foulongne, Virginie Sauvage, Charles Hebert, Olivier Dereure, Justine Cheval, Meriadeg Ar Gouilh, Kevin Pariente, Michel Segondy, Ana Burguière, Jean-Claude Manuguerra, et al. “Human Skin Microbiota: High Diversity of DNA Viruses Identified on the Human Skin by High Throughput Sequencing”. In: *PLoS ONE* 7.6 (2012). Ed. by Amanda Ewart Toland, e38499. DOI: 10.1371/journal.pone.0038499.

[6] Dirk Gevers, Rob Knight, Joseph F. Petrosino, Katherine Huang, Amy L. McGuire, Bruce W. Birren, Karen E. Nelson, Owen White, Barbara A. Methé, and Curtis Huttenhower. “The Human Microbiome Project: A Community Resource for the Healthy Human Microbiome”. In: *PLoS Biology* 10.8 (2012), e1001377. DOI: 10.1371/journal.pbio.1001377.

[7] The Human Microbiome Project Consortium, Curtis Huttenhower, Dirk Gevers, Rob Knight, Sahar Abubucker, Jonathan H. Badger, Asif T. Chinwalla, Heather H. Creasy, Ashlee M. Earl, Michael G. FitzGerald, et al. “Structure, function and diversity of the healthy human microbiome”. In: *Nature* 486.7402 (2012), pp. 207–214. DOI: 10.1038/nature11234.

[8] Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, et al. “A human gut microbial gene catalogue established by metagenomic sequencing”. In: *Nature* 464.7285 (2010), pp. 59–65. DOI: 10.1038/nature08821.

[9] Mya Breitbart, Peter Salamon, Bjarne Andresen, Joseph M Mahaffy, Anca M Segall, David Mead, Farooq Azam, and Forest Rohwer. “Genomic analysis of uncultured marine viral communities.” In: *Proceedings of the National Academy of Sciences of the United States of America* 99.22 (2002), pp. 14250–5. DOI: 10.1073/pnas.202488399.

[10] Bonnie L. Hurwitz and Matthew B. Sullivan. "The Pacific Ocean Virome (POV): A Marine Viral Metagenomic Dataset and Associated Protein Clusters for Quantitative Viral Ecology". In: *PLoS ONE* 8.2 (2013). Ed. by Fabiano Thompson, e57355. DOI: 10.1371/journal.pone.0057355.

[11] Carolina Megumi Mizuno, Francisco Rodriguez-Valera, Nikole E. Kimes, and Rohit Ghai. "Expanding the Marine Virosphere Using Metagenomics". In: *PLoS Genetics* 9.12 (2013). Ed. by Eduardo P. C. Rocha, e1003987. DOI: 10.1371/journal.pgen.1003987.

[12] Guy Cochrane, Ilene Karsch-Mizrachi, Toshihisa Takagi, and International Nucleotide Sequence Database Collaboration. "The International Nucleotide Sequence Database Collaboration". In: *Nucleic Acids Research* 44.D1 (2015), pp. D48–D50. DOI: 10.1093/nar/gkv1323.

[13] Ilene Karsch-Mizrachi, Toshihisa Takagi, Guy Cochrane, and on behalf of the International Nucleotide Sequence Database Collaboration. "The international nucleotide sequence database collaboration". In: *Nucleic Acids Research* 46.D1 (2017), pp. D48–D51. DOI: 10.1093/nar/gkx1097.

[14] Ryan Connor, Rodney Brister, Jan Buchmann, Ward Deboutte, Rob Edwards, Joan Martí-Carreras, Mike Tisza, Vadim Zalunin, Juan Andrade-Martínez, Adrian Cantu, et al. "NCBI's Virus Discovery Hackathon: Engaging Research Communities to Identify Cloud Infrastructure Requirements". In: *Genes* 10.9 (2019), p. 714. DOI: 10.3390/genes10090714.

[15] Sam Nooij, Dennis Schmitz, Harry Vennema, Annelies Kroneman, and Marion P G Koopmans. "Overview of Virus Metagenomic Classification Methods and Their Biological Applications." In: *Frontiers in microbiology* 9 (2018), p. 749. DOI: 10.3389/fmicb.2018.00749.

[16] Merry Youle, Matthew Haynes, and Forest Rohwer. "Scratching the Surface of Biology's Dark Matter". In: *Viruses: Essential Agents of Life*. Dordrecht: Springer Netherlands, 2012, pp. 61–81. DOI: 10.1007/978-94-007-4899-6_4.

[17] Christian Rinke, Patrick Schwientek, Alexander Sczyrba, Natalia N. Ivanova, Iain J. Anderson, Jan-Fang Cheng, Aaron Darling, Stephanie Malfatti, Brandon K. Swan, Esther A. Gies, et al. "Insights into the phylogeny and coding potential of microbial dark matter". In: *Nature* 499.7459 (2013), pp. 431–437. DOI: 10.1038/nature12352.

[18] Siddharth R. Krishnamurthy and David Wang. "Origins and challenges of viral dark matter". In: *Virus Research* 239 (2017), pp. 136–142. DOI: 10.1016/J.VIRUSRES.2017.02.002.

[19] Guillaume Bernard, Jananan S Pathmanathan, Romain Lannes, Philippe Lopez, and Eric Bapteste. “Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery”. In: *Genome Biology and Evolution* 10.3 (2018), pp. 707–715. DOI: 10.1093/gbe/evy031.

[20] Murat A. *Microbial Dark Matter: The mullet of microbial ecology – Meren Lab*. URL: http://merenlab.org/2017/06/22/microbial-dark-matter/ (visited on 06/19/2020).

[21] Simon Roux, Steven J Hallam, Tanja Woyke, and Matthew B Sullivan. “Viral dark matter and virus–host interactions resolved from publicly available microbial genomes”. In: *eLife* 4 (2015). DOI: 10.7554/eLife.08490.

[22] Lindsey Solden, Karen Lloyd, and Kelly Wrighton. “The bright side of microbial dark matter: lessons learned from the uncultivated majority”. In: *Current Opinion in Microbiology* 31 (2016), pp. 217–226. DOI: 10.1016/J.MIB.2016.04.020.

[23] Tanja Woyke, Devin F.R. Doud, and Emiley A. Eloe-Fadrosh. “Genomes from uncultivated microorganisms”. In: *Encyclopedia of Microbiology*. Elsevier, 2019, pp. 437–442. DOI: 10.1016/B978-0-12-809633-8.90682-4.

[24] Alexandre Almeida, Alex L. Mitchell, Miguel Boland, Samuel C. Forster, Gregory B. Gloor, Aleksandra Tarkowska, Trevor D. Lawley, and Robert D. Finn. “A new genomic blueprint of the human gut microbiota”. In: *Nature* 568.7753 (2019), p. 1. DOI: 10.1038/s41586-019-0965-1.

[25] Edoardo Pasolli, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, Paolo Manghi, Adrian Tett, Paolo Ghensi, et al. “Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle”. In: *Cell* 176.0 (2019), 649–662.e20. DOI: 10.1016/j.cell.2019.01.001.

[26] Alexandre Almeida, Stephen Nayfach, Miguel Boland, Francesco Strozzi, Martin Beracochea, Zhou Jason Shi, Katherine S. Pollard, Ekaterina Sakharova, Donovan H. Parks, Philip Hugenholtz, et al. “A unified catalog of 204,938 reference genomes from the human gut microbiome”. In: *Nature Biotechnology* (2020), pp. 1–10. DOI: 10.1038/s41587-020-0603-3.

[27] Jimmy H. Saw, Anja Spang, Katarzyna Zaremba-Niedzwiedzka, Lina Juzokaite, Jeremy A. Dodsworth, Senthil K. Murugapiran, Dan R. Colman, Cristina Takacs-Vesbach, Brian P. Hedlund, Lionel Guy, et al. “Exploring microbial dark matter to resolve the deep archaeal ancestry of eukaryotes”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1678 (2015). DOI: 10.1098/rstb.2014.0328.

[28] Bas E. Dutilh, Noriko Cassman, Katelyn McNair, Savannah E. Sanchez, Genivaldo G. Z. Silva, Lance Boling, Jeremy J. Barr, Daan R. Speth, Victor Seguritan, Ramy K. Aziz, et al. “A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes”. In: *Nature Communications* 5.1 (2014), p. 4498. DOI: 10.1038/ncomms5498.

[29] Mark Kowarsky, Joan Camunas-Soler, Michael Kertesz, Iwijn De Vlaminck, Winston Koh, Wenying Pan, Lance Martin, Norma F Neff, Jennifer Okamoto, Ronald J Wong, et al. “Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA.” In: *Proceedings of the National Academy of Sciences of the United States of America* 114.36 (2017), pp. 9623–9628. DOI: 10.1073/pnas.1707009114.

[30] David Wang. “5 challenges in understanding the role of the virome in health and disease”. In: *PLOS Pathogens* 16.3 (2020). Ed. by Katherine R. Spindler, e1008318. DOI: 10.1371/journal.ppat.1008318.

[31] Alex L Mitchell, Maxim Scheremetjew, Hubert Denise, Simon Potter, Aleksandra Tarkowska, Matloob Qureshi, Gustavo A Salazar, Sebastien Pesseat, Miguel A Boland, Fiona M I Hunter, et al. “EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies”. In: *Nucleic Acids Research* 46.D1 (2018), pp. D726–D735. DOI: 10.1093/nar/gkx967.

[32] Alex L Mitchell, Alexandre Almeida, Martin Beracochea, Miguel Boland, Josephine Burgin, Guy Cochrane, Michael R Crusoe, Varsha Kale, Simon C Potter, Lorna J Richardson, et al. “MGnify: the microbiome analysis resource in 2020”. In: *Nucleic Acids Research* 48.D1 (2019), pp. D570–D578. DOI: 10.1093/nar/gkz1035.

[33] Edoardo Pasolli, Lucas Schiffer, Paolo Manghi, Audrey Renson, Valerie Obenchain, Duy Tin Truong, Francesco Beghini, Faizan Malik, Marcel Ramos, Jennifer B Dowd, et al. “Accessible, curated metagenomic data through ExperimentHub”. In: *Nature Methods* 14.11 (2017), pp. 1023–1024. DOI: 10.1038/nmeth.4468.

[34] Alexander Sczyrba, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, Stephan Majda, Jessika Fiedler, Eik Dahms, et al. “Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software”. In: *Nature Methods* 14.11 (2017), pp. 1063–1071. DOI: 10.1038/nmeth.4458.

[35] F. A.Bastiaan Von Meijenfeldt, Ksenia Arkhipova, Diego D. Cambuy, Felipe H. Coutinho, and Bas E. Dutilh. “Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT”. In: *Genome Biology* 20.1 (2019), p. 217. DOI: 10.1186/s13059-019-1817-x.

[36] David Paez-Espino, Simon Roux, I-Min A Chen, Krishna Palaniappan, Anna Ratner, Ken Chu, Marcel Huntemann, T B K Reddy, Joan Carles Pons, Mercè Llabrés, et al. "IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes". In: *Nucleic Acids Research* 47.D1 (2019), pp. D678–D686. DOI: 10.1093/nar/gky1127.

[37] Michael J. Tisza, Diana V. Pastrana, Nicole L. Welch, Brittany Stewart, Alberto Peretti, Gabriel J. Starrett, Yuk-Ying S Ying S. Pang, Siddharth R. Krishnamurthy, Patricia A. Pesavento, David H. Mcdermott, et al. "Discovery of several thousand highly diverse circular DNA viruses". In: *eLife* 9 (2020). DOI: 10.7554/eLife.51971.

[38] Jessica Galloway-Peña and Blake Hanson. *Tools for Analysis of the Microbiome*. 2020. DOI: 10.1007/s10620-020-06091-y.

[39] Eric Sayers. *E-utilities Quick Start*. National Center for Biotechnology Information (US), 2018.

[40] Valieris R. *parallel fastq-dump wrapper*. URL: https://github.com/rvalieris/parallel-fastq-dump (visited on 06/19/2020).

[41] Bushnell B. *BBMap download | SourceForge.net*.

[42] H. Li and R. Durbin. "Fast and accurate short read alignment with Burrows-Wheeler transform". In: *Bioinformatics* 25.14 (2009), pp. 1754–1760. DOI: 10.1093/bioinformatics/btp324.

[43] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing 1000 Genome Project Data Processing Subgroup. "The Sequence Alignment/Map format and SAMtools." In: *Bioinformatics (Oxford, England)* 25.16 (2009), pp. 2078–9. DOI: 10.1093/bioinformatics/btp352.

[44] Michael R Crusoe, Hussien F Alameldin, Sherine Awad, Elmar Boucher, Adam Caldwell, Reed Cartwright, Amanda Charbonneau, Bede Constantinides, Greg Edvenson, Scott Fay, et al. "The khmer software package: enabling efficient nucleotide sequence analysis". In: *F1000Research* 4 (2015). DOI: 10.12688/f1000research.6924.1.

[45] Sergey Nurk, Anton Bankevich, Dmitry Antipov, Alexey Gurevich, Anton Korobeynikov, Alla Lapidus, Andrey Prjibelsky, Alexey Pyshkin, Alexander Sirotkin, Yakov Sirotkin, et al. "Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads". In: Springer Berlin Heidelberg, 2013, pp. 158–170. DOI: 10.1007/978-3-642-37195-0_13.

[46] Benjamin Buchfink, Chao Xie, and Daniel H Huson. "Fast and sensitive protein alignment using DIAMOND". In: *Nature Methods* 12.1 (2014), pp. 59–60. DOI: 10.1038/nmeth.3176.

[47] Jaime Huerta-Cepas, François Serra, and Peer Bork. "ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data". In: *Molecular Biology and Evolution* 33.6 (2016), pp. 1635–1638. DOI: 10.1093/MOLBEV/MSW046.

[48] Saket Choudhary. "Pysradb: A Python package to query next-generation sequencing metadata and data from NCBI sequence read archive". In: *F1000Research* 8 (2019), p. 532. DOI: 10.12688/f1000research.18676.1.

[49] Peter Rice, Ian Longden, Alan Bleasby, Lan Longden, and Alan Bleasby. "EMBOSS: The European Molecular Biology Open Software Suite". In: *Trends in Genetics* 16.6 (2000), pp. 276–277. DOI: 10.1016/S0168-9525(00)02024-2.

[50] Martin Steinegger and Johannes Söding. *MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets*. 2017. DOI: 10.1038/nbt.3988.

[51] Martin Steinegger and Johannes Söding. "Clustering huge protein sequence sets in linear time". In: *Nature Communications* 9.1 (2018), pp. 1–8. DOI: 10.1038/s41467-018-04964-5.

[52] Stephen Nayfach, Antonio Pedro Camargo, Frederik Schulz, Emiley Eloe-Fadrosh, Simon Roux, and Nikos C. Kyrpides. "CheckV assesses the quality and completeness of metagenome-assembled viral genomes". In: *Nature Biotechnology 2020 39:5* 39.5 (2020), pp. 578–585. DOI: 10.1038/s41587-020-00774-7.

[53] Jie Ren, Kai Song, Chao Deng, Nathan A Ahlgren, Jed A Fuhrman, Yi Li, Xiaohui Xie, Ryan Poplin, and Fengzhu Sun. *Identifying viruses from metagenomic data using deep learning*. Tech. rep. 2020, pp. 1–14.

[54] Pakorn Aiewsakun and Peter Simmonds. "The genomic underpinnings of eukaryotic virus taxonomy: Creating a sequence-based framework for family-level virus classification". In: *Microbiome* 6.1 (2018), pp. 1–24. DOI: 10.1186/S40168-018-0422-7/FIGURES/9.

[55] Alex L Mitchell, Teresa K Attwood, Patricia C Babbitt, Matthias Blum, Alan Bridge, Shoshana D Brown, Hsin-Yu Chang, Sara El-Gebali, Matthew I Fraser, Julian Gough, et al. "InterPro in 2019: improving coverage, classification and access to protein sequence annotations". In: *Nucleic Acids Research* 47 (2018). DOI: 10.1093/nar/gky1100.

[56] Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Matloob Qureshi, Lorna J Richardson, Gustavo A Salazar, Alfredo Smart, et al. "The Pfam protein families database in 2019". In: *Nucleic Acids Research* 47 (2018), pp. 427–432. DOI: 10.1093/nar/gky995.

[57] Shennan Lu, Jiyao Wang, Farideh Chitsaz, Myra K Derbyshire, Renata C Geer, Noreen R Gonzales, Marc Gwadz, David I Hurwitz, Gabriele H Marchler, James S Song, et al. "CDD/SPARCLE: the conserved domain database in 2020". In: *Nucleic Acids Research* 48.D1 (2019), pp. D265–D268. DOI: 10.1093/nar/gkz991.

[58] Julian Gough, Kevin Karplus, Richard Hughey, and Cyrus Chothia. “Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure”. In: *Journal of Molecular Biology* 313.4 (2001), pp. 903–919. DOI: 10.1006/jmbi.2001.5080.

[59] Marco Necci, Damiano Piovesan, Zsuzsanna Doszt Anyi, Silvio C E Tosatto, Zsuzsanna Dosztányi, and Silvio C E Tosatto. “MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins”. In: *Bioinformatics* 33.9 (2017), pp. 1402–1404. DOI: 10.1093/bioinformatics/btx015.

[60] Minna M. Poranen and Sari Mäntynen. “ICTV virus taxonomy profile: Cystoviridae”. In: *Journal of General Virology* 98.10 (2017), pp. 2423–2424. DOI: 10.1099/jgv.0.000928.

[61] Simon Roux, Evelien M. Adriaenssens, Bas E. Dutilh, Eugene V. Koonin, Andrew M. Kropinski, Mart Krupovic, Jens H. Kuhn, Rob Lavigne, J. Rodney Brister, Arvind Varsani, et al. “Minimum information about an uncultivated virus genome (MIUVIG)”. In: *Nature Biotechnology* 37.1 (2019), pp. 29–37. DOI: 10.1038/nbt.4306.

[62] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. “Basic local alignment search tool”. In: *Journal of Molecular Biology* 215.3 (1990), pp. 403–410. DOI: 10.1016/S0022-2836(05)80360-2.

[63] Lin Xing Chen, Karthik Anantharaman, Alon Shaiber, A. Murat Eren, and Jillian F. Banfield. “Accurate and complete genomes from metagenomes”. In: 30.3 (2020), pp. 315–333. DOI: 10.1101/gr.258640.119.

[64] Luis F. Camarillo-Guerrero, Alexandre Almeida, Guillermo Rangel-Pineros, Robert D. Finn, and Trevor D. Lawley. “Massive expansion of human gut bacteriophage diversity”. In: *Cell* 184.4 (2021), 1098–1109.e9. DOI: 10.1016/j.cell.2021.01.029.

[65] Ann C. Gregory, Olivier Zablocki, Ahmed A. Zayed, Allison Howell, Benjamin Bolduc, and Matthew B. Sullivan. “The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut”. In: *Cell Host and Microbe* 28.5 (2020), 724–740.e8. DOI: 10.1016/j.chom.2020.08.003.

[66] Florian P. Breitwieser, Jennifer Lu, and Steven L. Salzberg. “A review of methods and databases for metagenomic classification and assembly”. In: *Briefings in Bioinformatics* 20.4 (2018), pp. 1125–1139. DOI: 10.1093/bib/bbx120.

[67] Vijini Mallawaarachchi, Anuradha Wickramarachchi, and Yu Lin. “GraphBin: refined binning of metagenomic contigs using assembly graphs”. In: *Bioinformatics (Oxford, England)* 36.11 (2020), pp. 3307–3313. DOI: 10.1093/bioinformatics/btaa180.

[68] Robert M. Bowers, Nikos C. Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, T. B.K. Reddy, Frederik Schulz, Jessica Jarett, Adam R. Rivers, Emiley A. Eloe-Fadrosh, et al. “Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea”. In: *Nature Biotechnology* 35.8 (2017), pp. 725–731. DOI: 10.1038/nbt.3893.

[69] Donovan H. Parks, Maria Chuvochina, David W. Waite, Christian Rinke, Adam Skarshewski, Pierre Alain Chaumeil, and Philip Hugenholtz. “A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life”. In: *Nature Biotechnology* 36.10 (2018), p. 996. DOI: 10.1038/nbt.4229.

[70] Peter Simmonds, Mike J. Adams, Mária Benk, Mya Breitbart, J. Rodney Brister, Eric B. Carstens, Andrew J. Davison, Eric Delwart, Alexander E. Gorbalenya, Balázs Harrach, et al. “Consensus statement: Virus taxonomy in the age of metagenomics”. In: *Nature Reviews Microbiology* 15.3 (2017), pp. 161–168. DOI: 10.1038/nrmicro.2016.177.

[71] Matthew R. Olm, Christopher T. Brown, Brandon Brooks, and Jillian F. Banfield. “DRep: A tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication”. In: *ISME Journal* 11.12 (2017), pp. 2864–2868. DOI: 10.1038/ismej.2017.126.

[72] Siddharth R. Krishnamurthy and David Wang. “Extensive conservation of prokaryotic ribosomal binding sites in known and novel picobirnaviruses”. In: *Virology* 516 (2018), pp. 108–114. DOI: 10.1016/J.VIROL.2018.01.006.

[73] Phelim Bradley, Henk C. den Bakker, Eduardo P.C. C. Rocha, Gil McVean, and Zamin Iqbal. “Ultrafast search of all deposited bacterial and viral genomic data”. In: *Nature Biotechnology* 37.2 (2019), pp. 152–159. DOI: 10.1038/s41587-018-0010-1.

[74] Joan Martí-Carreras, Alejandro Rafael Gener, Sierra D. Miller, Anderson F. Brito, Christiam E. Camacho, Ryan Connor, Ward Deboutte, Cody Glickman, David M. Kristensen, Wynn K. Meyer, et al. “NCBI’s Virus Discovery Codeathon: Building “FIVE” —The Federated Index of Viral Experiments API Index”. In: *Viruses* 12.12 (2020), p. 1424. DOI: 10.3390/v12121424.

Tables

Table 1: Circular contig clusters with direct and inverted terminal repeats

Study ID(s)	**Cluster size**	**Typical contig length**	**Repeat type**	**Sample type**	**Sequence similarity (min-max)**
PRJEB14383; PRJNA230363	8	2110	DTR	Saliva; Oral	71-100
PRJNA230363	7	1771	DTR	Oral	31-100
PRJEB7949	5	1337	ITR	Fecal	67-100

Figures

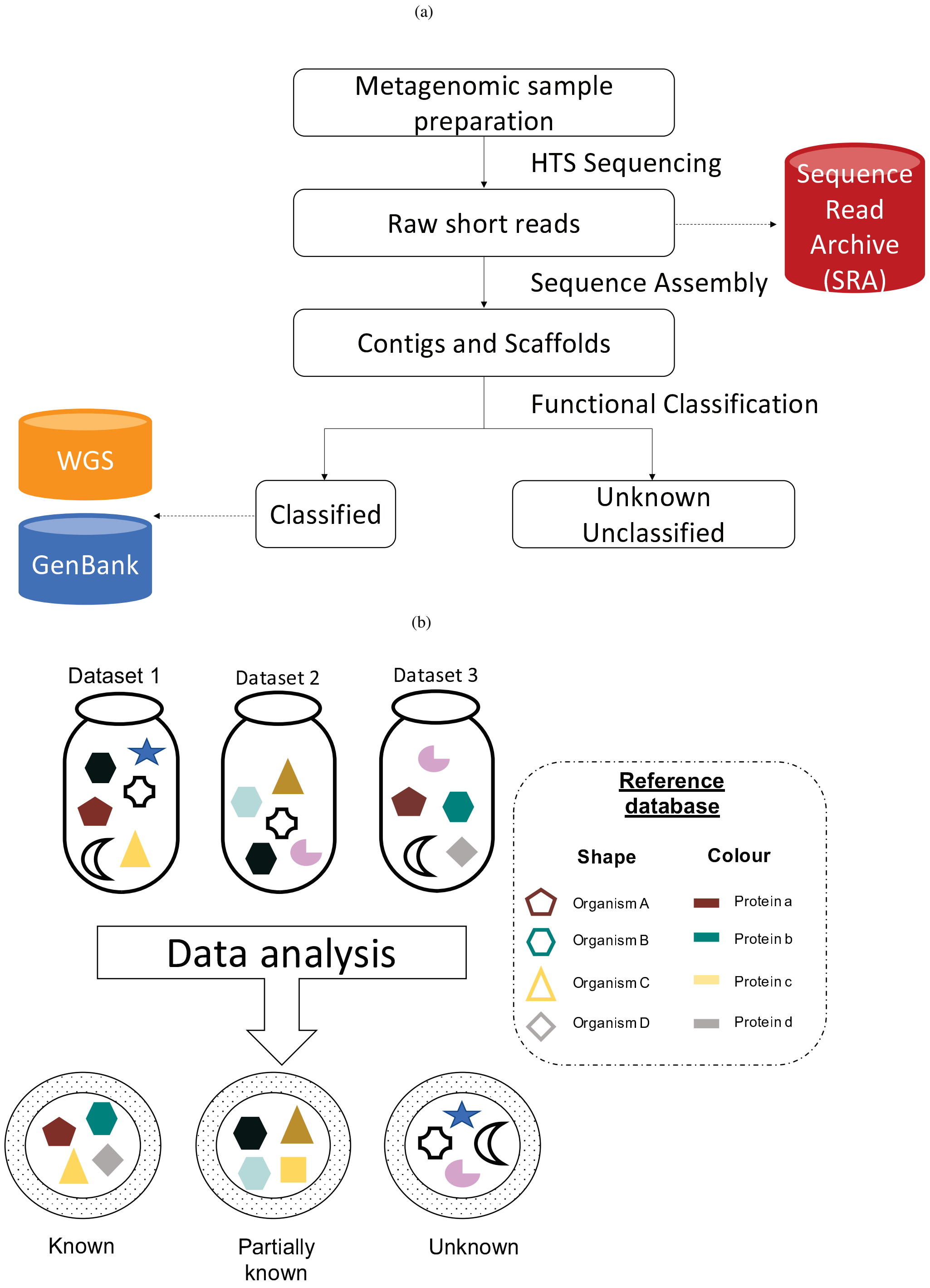


Figure 1: **Typical metagenomic analysis and data submission to public repositories.** Overview of existing metagenomic analytical workflow and the definition of unknown sequence matter. (a) Typical metagenomic analytical workflow with data submission steps. (b) A schematic representation of known, partially known and unknown sequence matter in the metagenomic data sets.

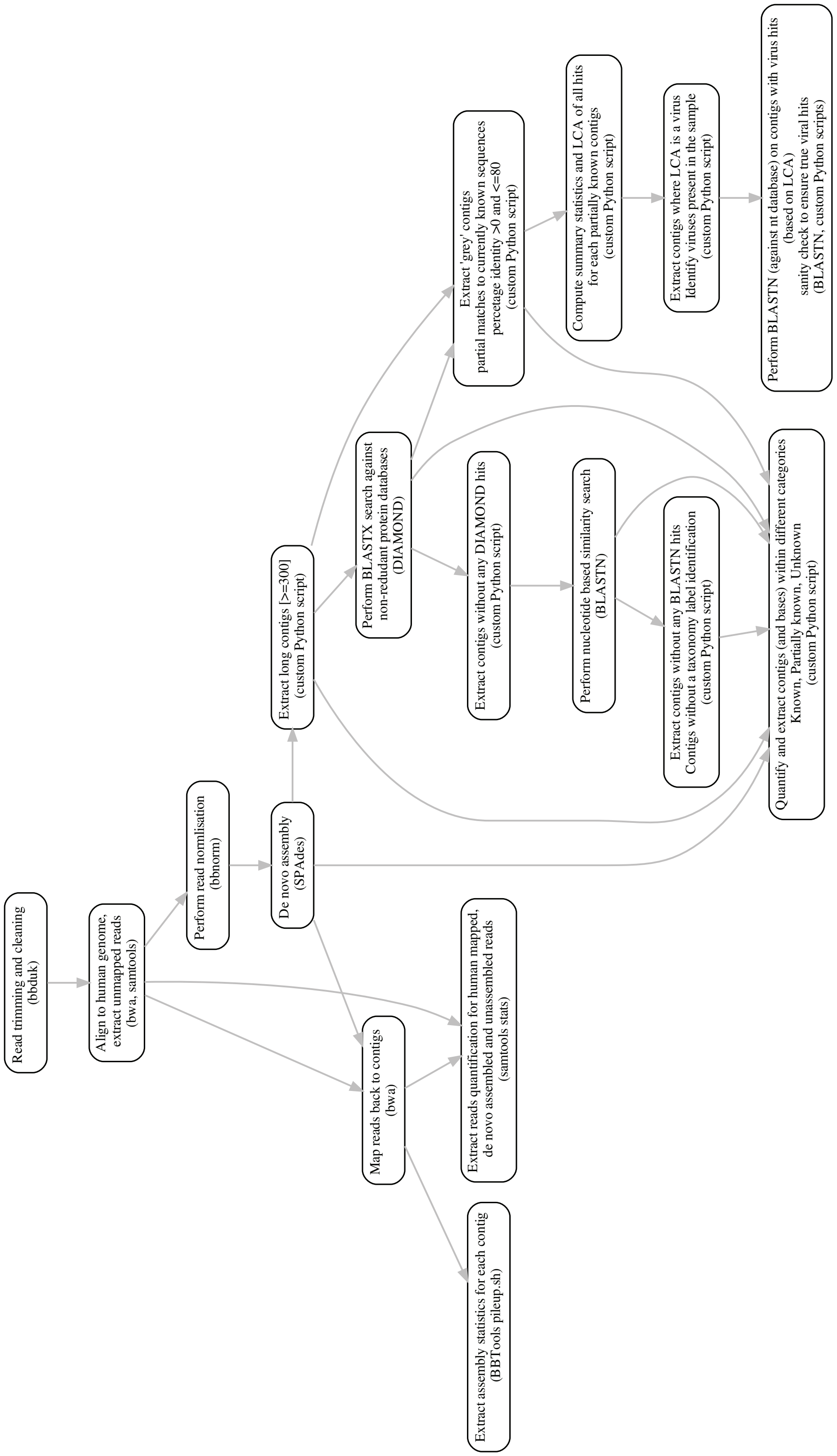

Read trimming and cleaning (bbduk)
Align to human genome, extract unmapped reads (bwa, samtools)
Perform read normlisation (bbnorm)
De novo assembly (SPAdes)
Map reads back to contigs (bwa)
Extract reads quantification for human mapped, de novo assembled and unassembled reads (samtools stats)
Extract assembly statistics for each contig (BBTools pileup.sh)
Extract long contigs [>=300] (custom Python script)
Perform BLASTX search against non-redudant protein databases (DIAMOND)
Extract contigs without any DIAMOND hits (custom Python script)
Perform nucleotide based similarity search (BLASTN)
Extract contigs without any BLASTN hits Contigs without a taxonomy label identification (custom Python script)
Quantify and extract contigs (and bases) within different categories Known, Partially known, Unknown (custom Python script)
Extract 'grey' contigs partial matches to currently known sequences percetage identity >0 and <=80 (custom Python script)
Compute summary statistics and LCA of all hits for each partially known contigs (custom Python script)
Extract contigs where LCA is a virus Identify viruses present in the sample (custom Python script)
Perform BLASTN (against nt database) on contigs with virus hits (based on LCA) sanity check to ensure true viral hits (BLASTN, custom Python scripts)

Figure 2: **UnXplore workflow designed to identify unknown sequences in metagenomic datasets.** Detailed workflow of the metagenomic analysis and unknown sequence identification pipeline.

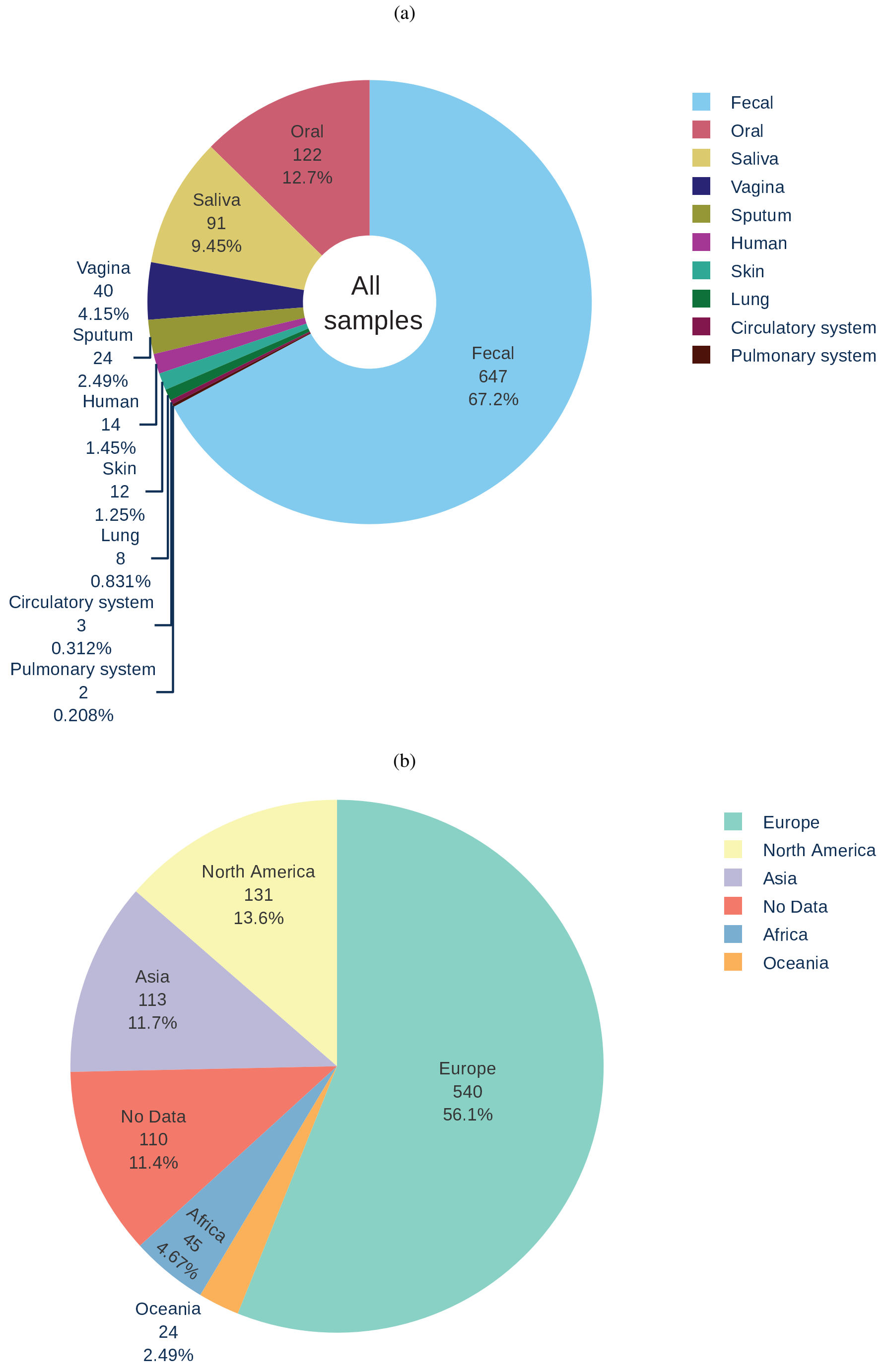


Figure 3: **An overview of the human microbiome data set included in this study.** (a) Distribution of samples included in this study for each microbiome (n=963). (b) Overview of the geographical distribution of the samples included in the study (n=861) coloured according to the distinct microbiome. The size of the slice represents the number and the proportion of samples.

Footnote: As Russia spans two continents; Asia and Europe, samples from Russia were included in Europe to simplify the illustration in this figure.

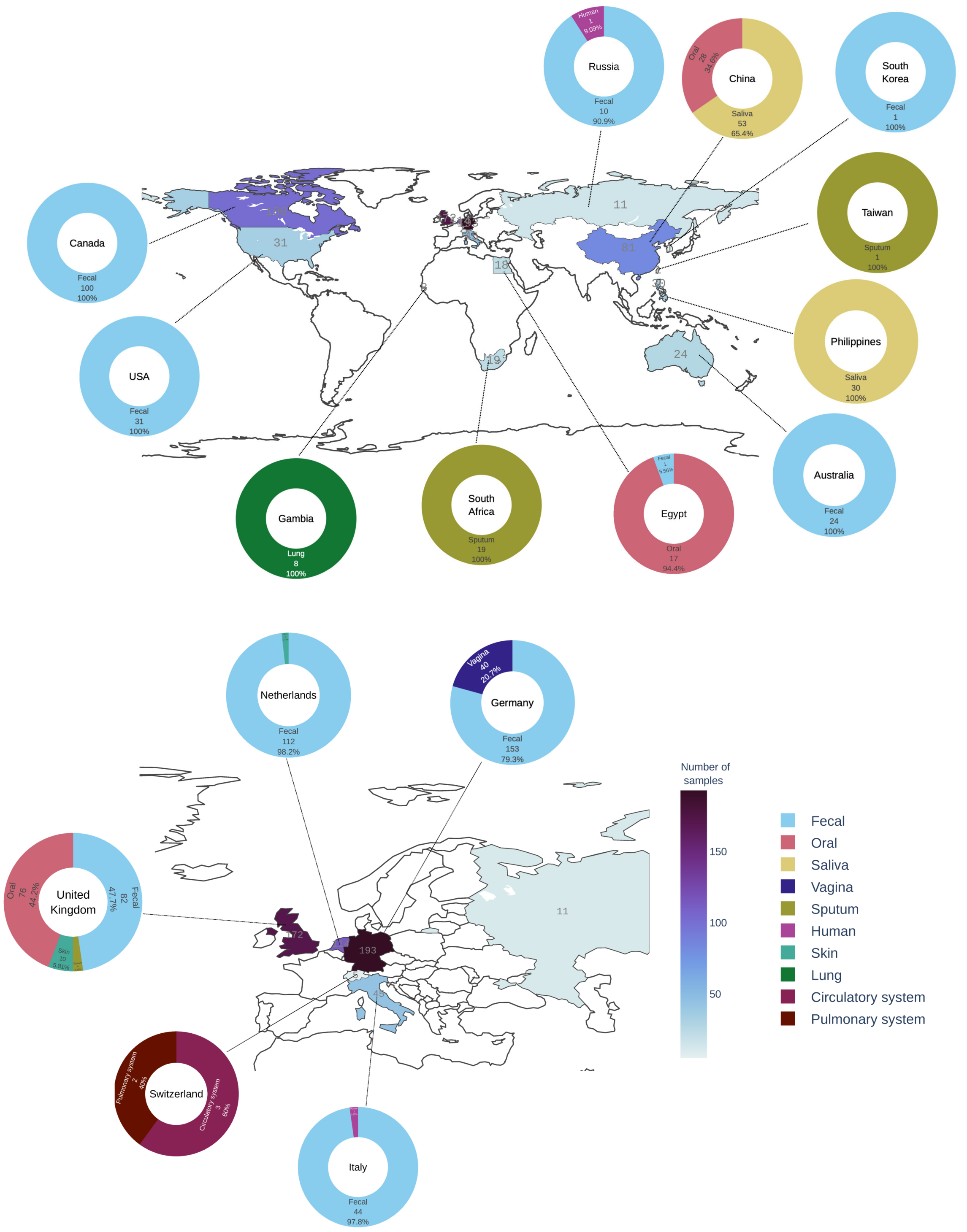


Figure 4: **The geographical distribution of human microbiome samples included in this study.** Geographic locations are coloured according to the number of samples (n=861) with darker shades representing the higher number of samples analysed. Samples originating from each location are represented by a doughnut chart. Each doughnut is coloured according to the microbiome and its proportion is represented by the slice of the doughnut.

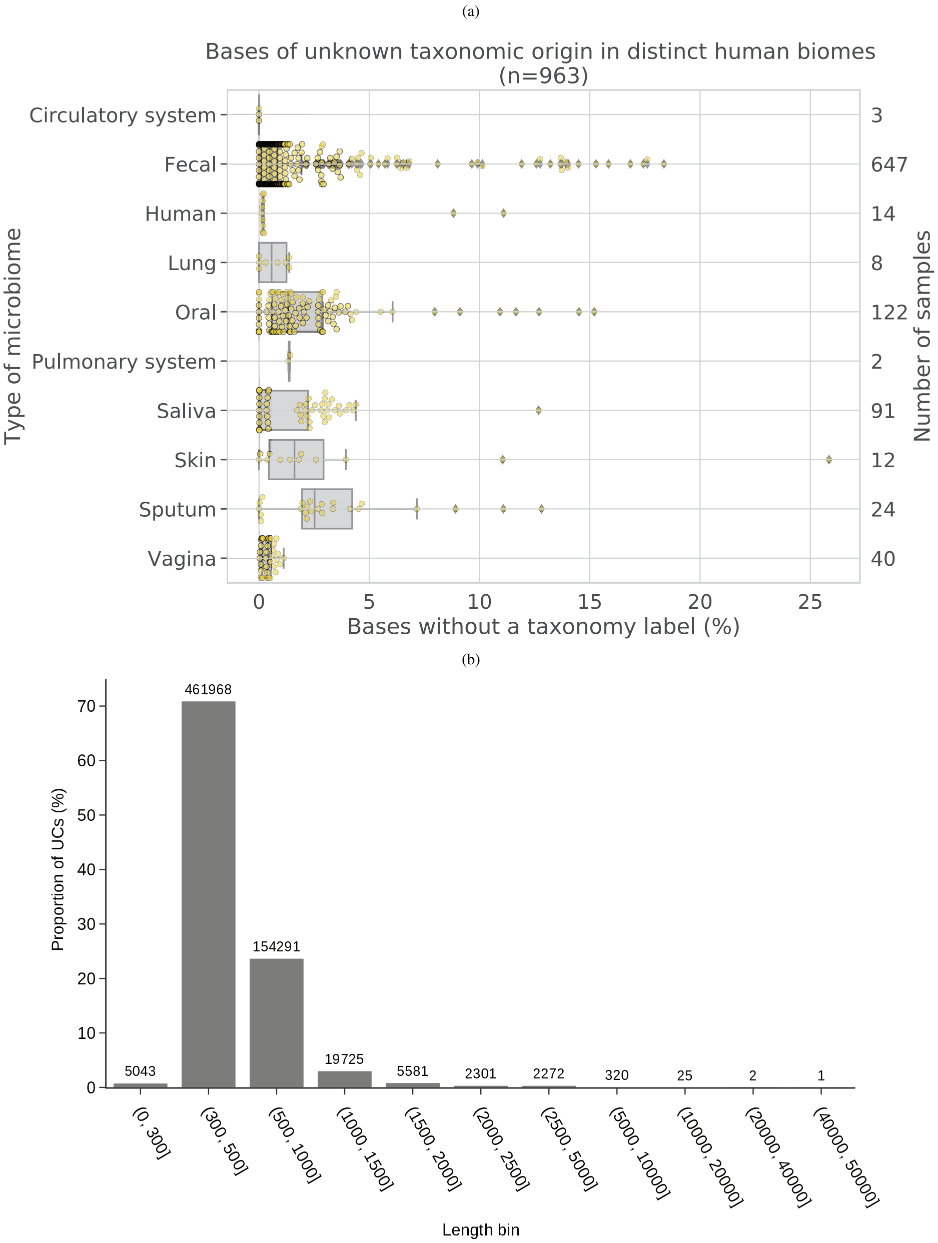


Figure 5: **Quantification of unknown sequences in different human microbiomes.** (a) The proportion of unknown bases in different human microbiomes. The proportion of unknown bases was calculated from the unknown contigs for each microbiome. The secondary Y-axis shows the number of samples analysed in each category. Each individual sample is overlayed on the boxplot and is represented by small yellow circles. (b) The distribution of all unknown contigs in ten distinct length categories. Each bar represents the proportion of UCs on the Y-axis with the number of contigs in the given category annotated at the top of the bar. Bin sizes are shown in the interval format, which means that sizes are exclusive on start values and inclusive on end values.

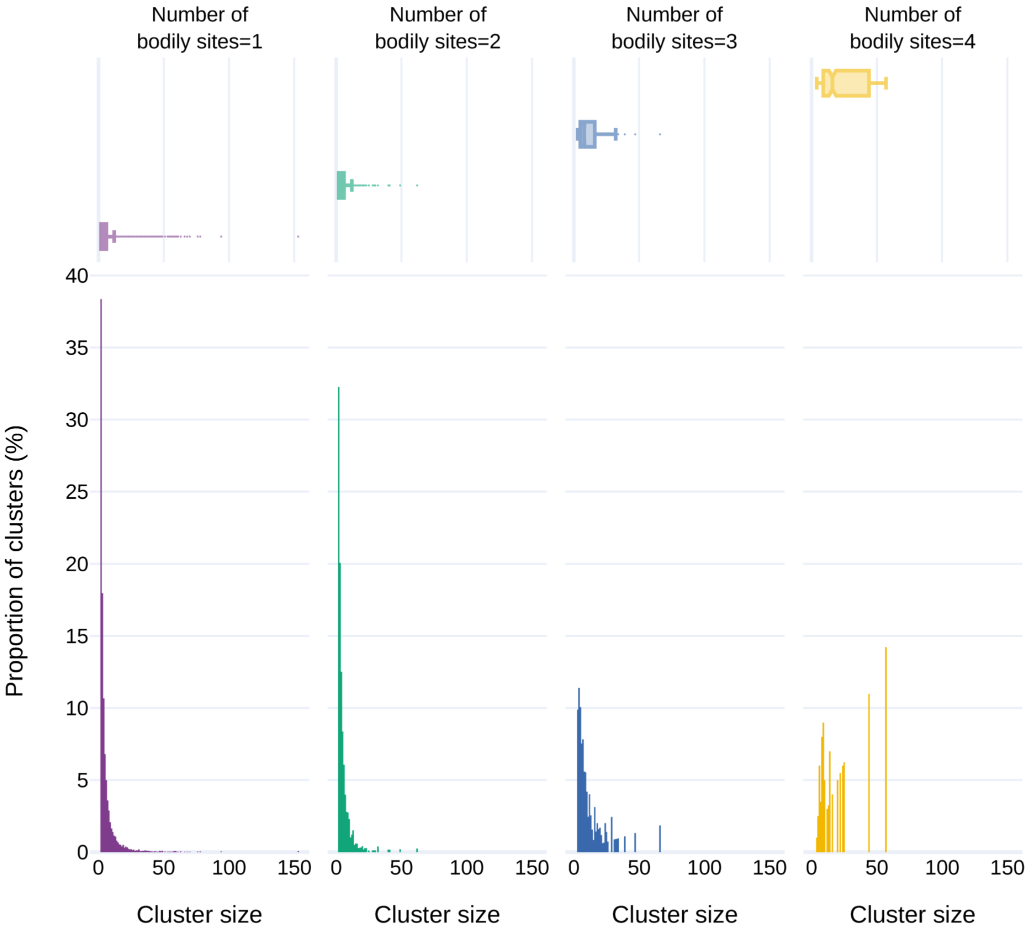


Figure 6: Distribution of cluster sizes on the X-axis and their proportion on the Y-axis. The marginal box plot shows the distribution of cluster sizes for each category. The plots are grouped and coloured according to the number of distinct bodily sites the clusters are found in; e.g. Number of bodily sites = 2 in green, means that members of each cluster are found in data sets from two distinct bodily sites (e.g. gut, skin, fecal, oral), all clusters from this plot come from 2 distinct bodily sites, but may (or may not) come from different bodily sites compared to other clusters within the plot, with one cluster coming from gut and skin, for example, and another from skin and fecal etc.

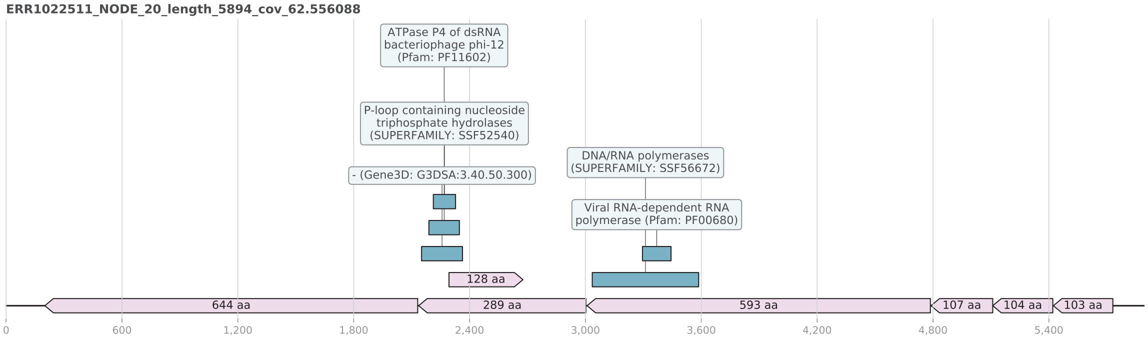
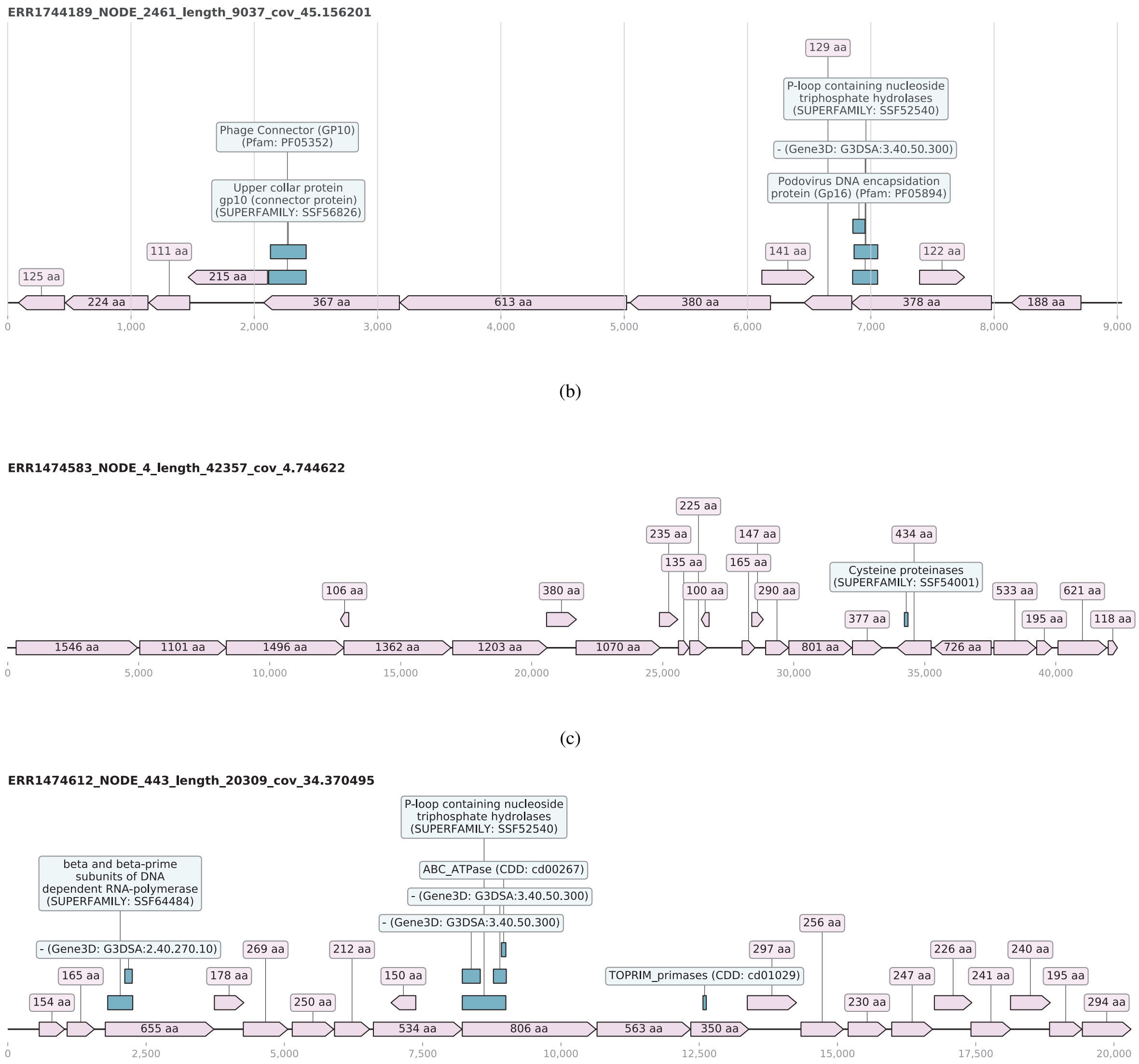

ERR1022511_NODE_20_length_5894_cov_62.556088
ATPase P4 of dsRNA bacteriophage phi-12 (Pfam: PF11602)
P-loop containing nucleoside triphosphate hydrolases (SUPERFAMILY: SSF52540)
- (Gene3D: G3DSA:3.40.50.300)
DNA/RNA polymerases (SUPERFAMILY: SSF56672)
Viral RNA-dependent RNA polymerase (Pfam: PF00680)
128 aa
644 aa
289 aa
593 aa
107 aa
104 aa
103 aa
0
600
1,200
1,800
2,400
3,000
3,600
4,200
4,800
5,400

Figure 7: The genome diagrams of a potentially novel dsRNA phage segment found among the UC set that is hypothesised to be related to currently known Cystoviruses. The open reading frames (ORFs) are highlighted in the light pink shade with the ORF lengths as their corresponding labels and the green boxes illustrating the InterProScan computed presence of domain signature.

(a)



(b)

(c)

ERR1744189_NODE_2461_length_9037_cov_45.156201
129 aa
P-loop containing nucleoside triphosphate hydrolases (SUPERFAMILY: SSF52540)
Phage Connector (GP10) (Pfam: PF05352)
- (Gene3D: G3DSA:3.40.50.300)
Upper collar protein gp10 (connector protein) (SUPERFAMILY: SSF56826)
Podovirus DNA encapsidation protein (Gp16) (Pfam: PF05894)
111 aa
141 aa
122 aa
125 aa
215 aa
224 aa
367 aa
613 aa
380 aa
378 aa
188 aa
0
1,000
2,000
3,000
4,000
5,000
6,000
7,000
8,000
9,000
ERR1474583_NODE_4_length_42357_cov_4.744622
225 aa
235 aa
147 aa
434 aa
135 aa
165 aa
Cysteine proteinases (SUPERFAMILY: SSF54001)
106 aa
380 aa
100 aa
290 aa
533 aa
621 aa
377 aa
195 aa
118 aa
1546 aa
1101 aa
1496 aa
1362 aa
1203 aa
1070 aa
801 aa
726 aa
0
5,000
10,000
15,000
20,000
25,000
30,000
35,000
40,000
ERR1474612_NODE_443_length_20309_cov_34.370495
P-loop containing nucleoside triphosphate hydrolases (SUPERFAMILY: SSF52540)
beta and beta-prime subunits of DNA dependent RNA-polymerase (SUPERFAMILY: SSF64484)
ABC_ATPase (CDD: cd00267)
- (Gene3D: G3DSA:3.40.50.300)
- (Gene3D: G3DSA:3.40.50.300)
256 aa
- (Gene3D: G3DSA:2.40.270.10)
269 aa
212 aa
297 aa
226 aa
240 aa
165 aa
178 aa
150 aa
TOPRIM_primases (CDD: cd01029)
247 aa
241 aa
195 aa
154 aa
250 aa
230 aa
294 aa
655 aa
534 aa
806 aa
563 aa
350 aa
0
2,500
5,000
7,500
10,000
12,500
15,000
17,500
20,000

Figure 8: The genome diagrams of large unknown contigs show the open reading frames (ORFs) in the light pink shade with the ORFs lengths as their corresponding labels and the green boxes illustrating the InterProScan computed presence of domain signature. (a) The largest contig with podovirus DNA encapsidation protein Gp16 domain. (b) The largest unknown contig assembled in the set is categorised as unknown even after the most recent similarity-based search on 14 Oct 2020. (c) An unknown contig of length 20,309 bases was described to contain a range of domains including a potential virus-specific RNA polymerase domain.

Supplemental Material

Supplementary figures

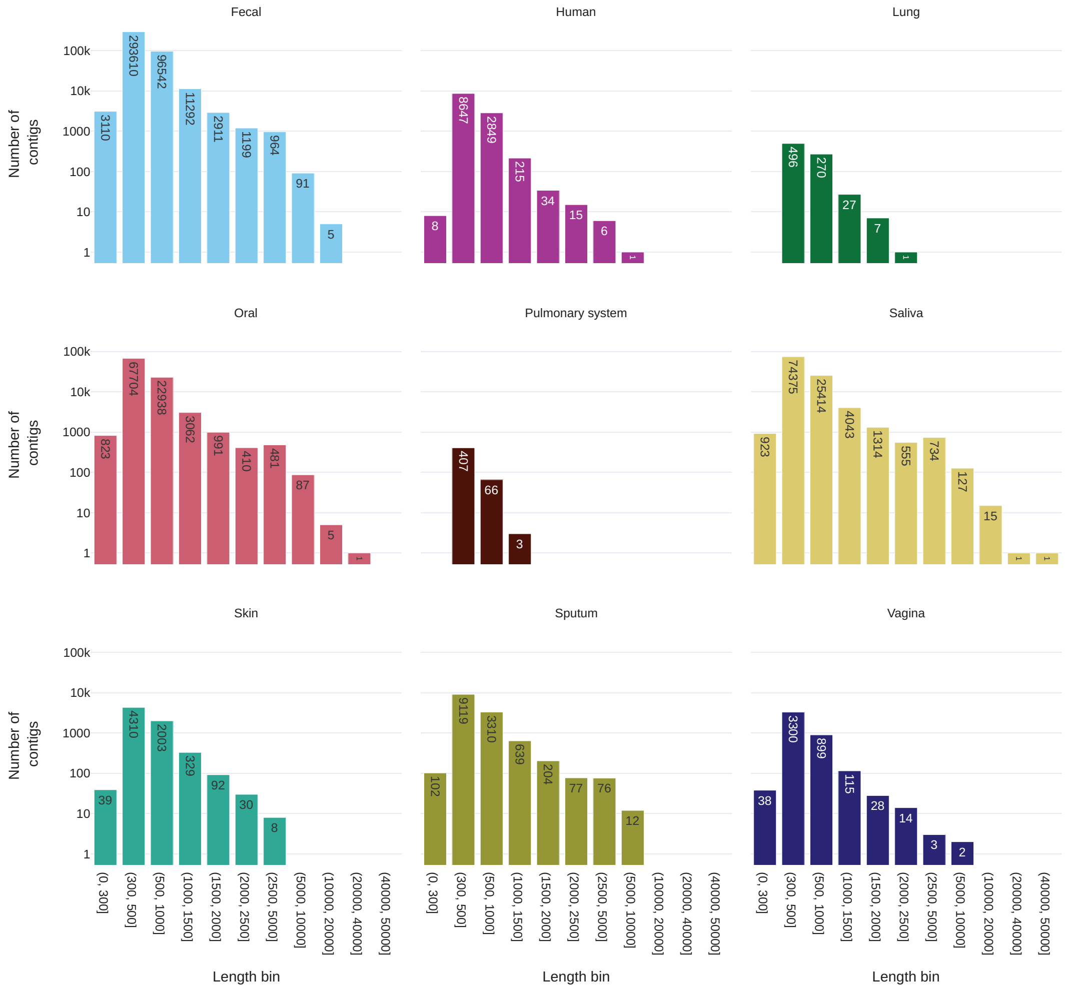
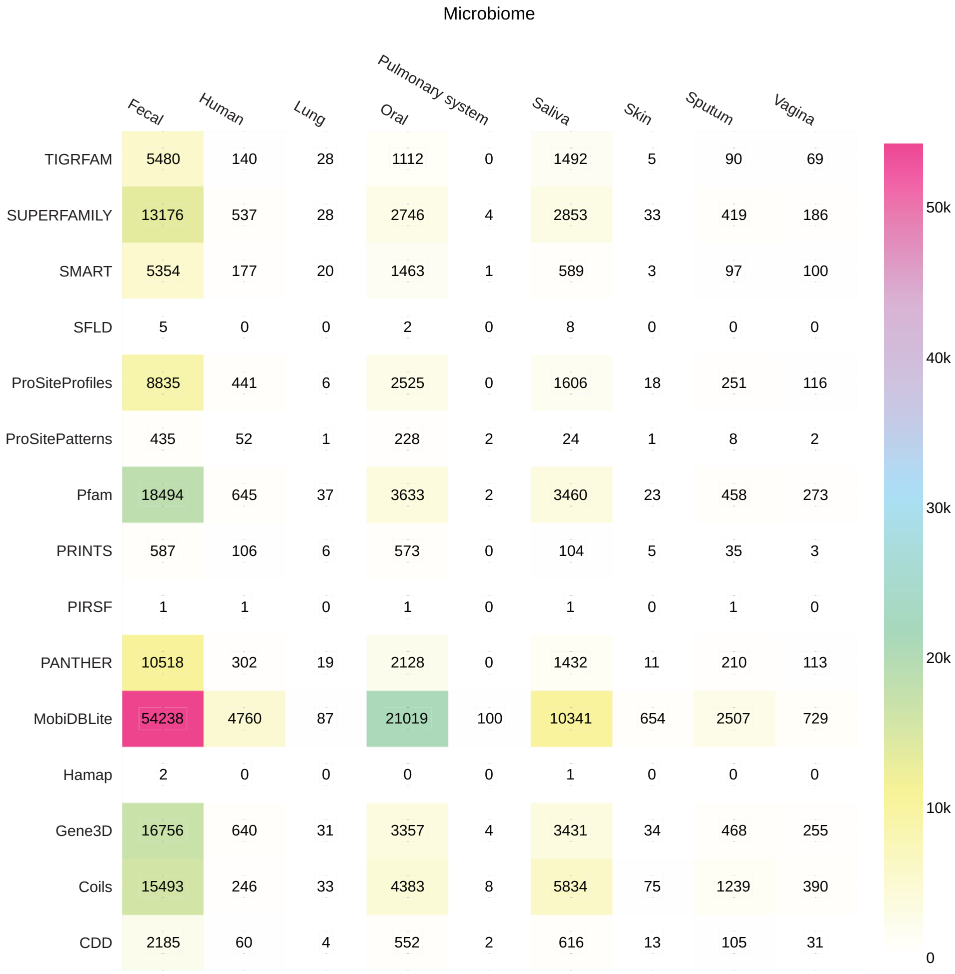


Figure S1: A detailed distribution of unknown contigs across all microbiomes where each microbiome is represented by a subplot in the faceted plot.



	Fecal	Human	Lung	Oral	Pulmonary system	Saliva	Skin	Sputum	Vagina
TIGRFAM	5480	140	28	1112	0	1492	5	90	69
SUPERFAMILY	13176	537	28	2746	4	2853	33	419	186
SMART	5354	177	20	1463	1	589	3	97	100
SFLD	5	0	0	2	0	8	0	0	0
ProSiteProfiles	8835	441	6	2525	0	1606	18	251	116
ProSitePatterns	435	52	1	228	2	24	1	8	2
Pfam	18494	645	37	3633	2	3460	23	458	273
PRINTS	587	106	6	573	0	104	5	35	3
PIRSF	1	1	0	1	0	1	0	1	0
PANTHER	10518	302	19	2128	0	1432	11	210	113
MobiDBLite	54238	4760	87	21019	100	10341	654	2507	729
Hamap	2	0	0	0	0	1	0	0	0
Gene3D	16756	640	31	3357	4	3431	34	468	255
Coils	15493	246	33	4383	8	5834	75	1239	390
CDD	2185	60	4	552	2	616	13	105	31

Figure S2: (a) Functional homologues identified in the InterProScan analyses for UCs generated for each microbiome. Darker colours represent the higher number of hits to specific Pfam clans. (b) UCs matching to various Pfam clans found in different microbiomes. The darker shades represent the larger number of UCs and lighter shades of the colour represent a smaller number of UCs that are also annotated in the boxes of the heatmap plot here.

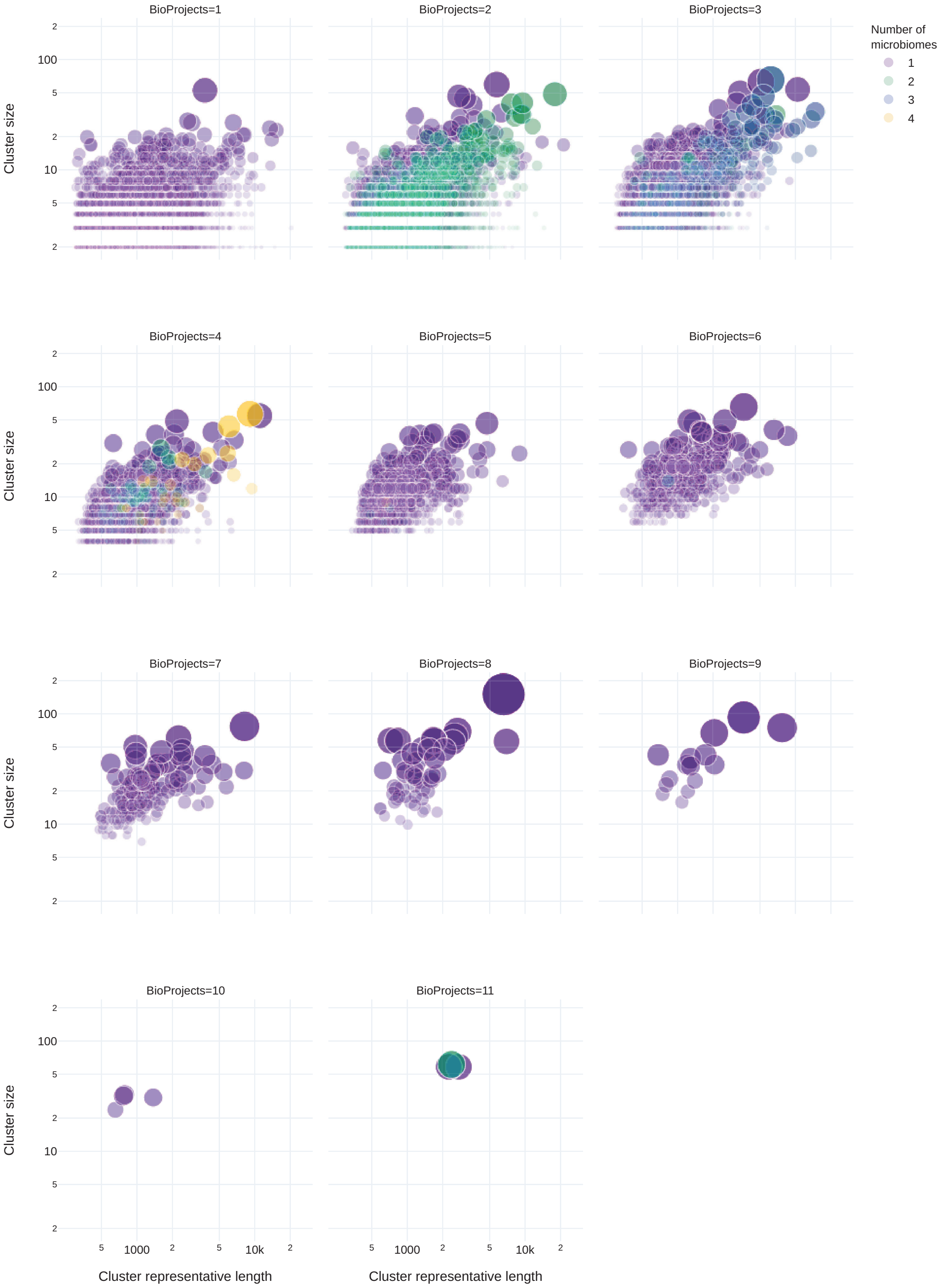


Figure S3: Overview of clusters found in the unknown sequence dataset. A number of microbiomes included in distinct clusters are represented by the columns and the number of BioProjects are represented by rows in the facetted plot. For each subplot, the X-axis represent the length of the cluster representative sequence and the Y-axis represent the cluster size for all cluster of size >=2. The size of the bubbles corresponds to the cluster sizes.

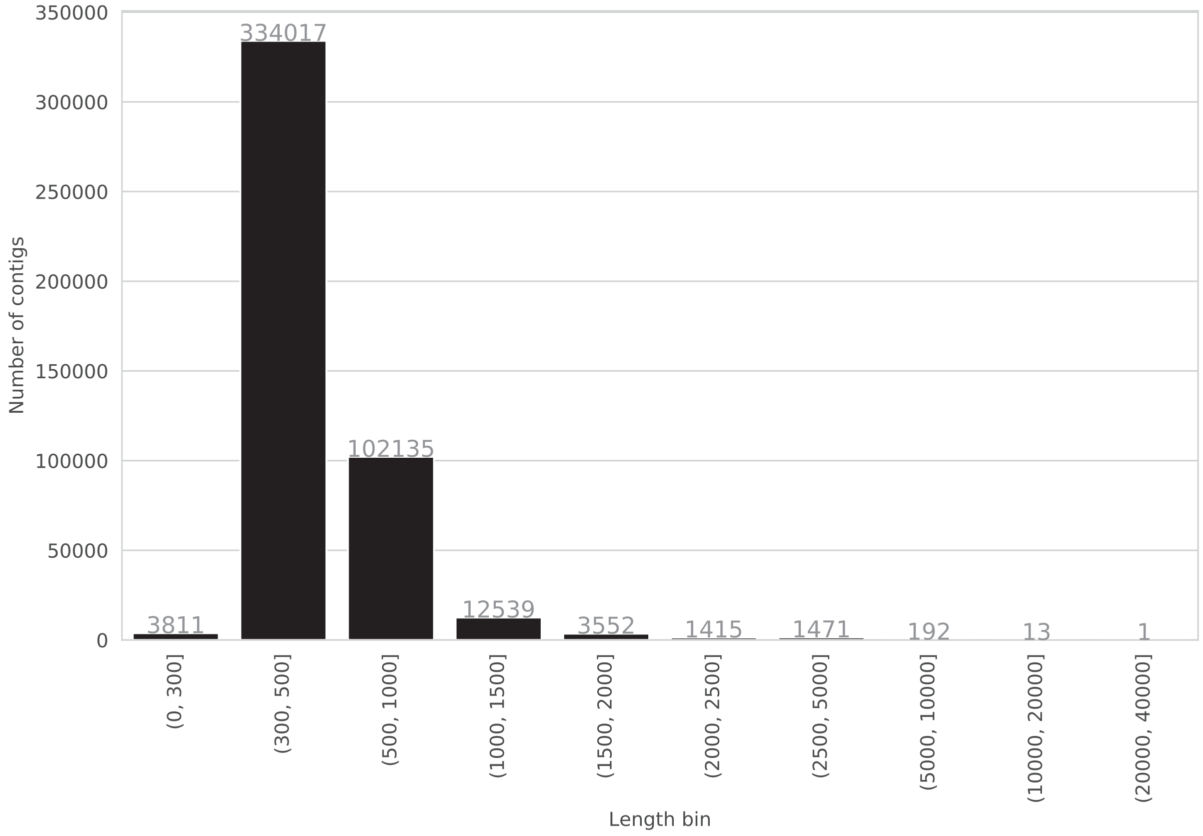
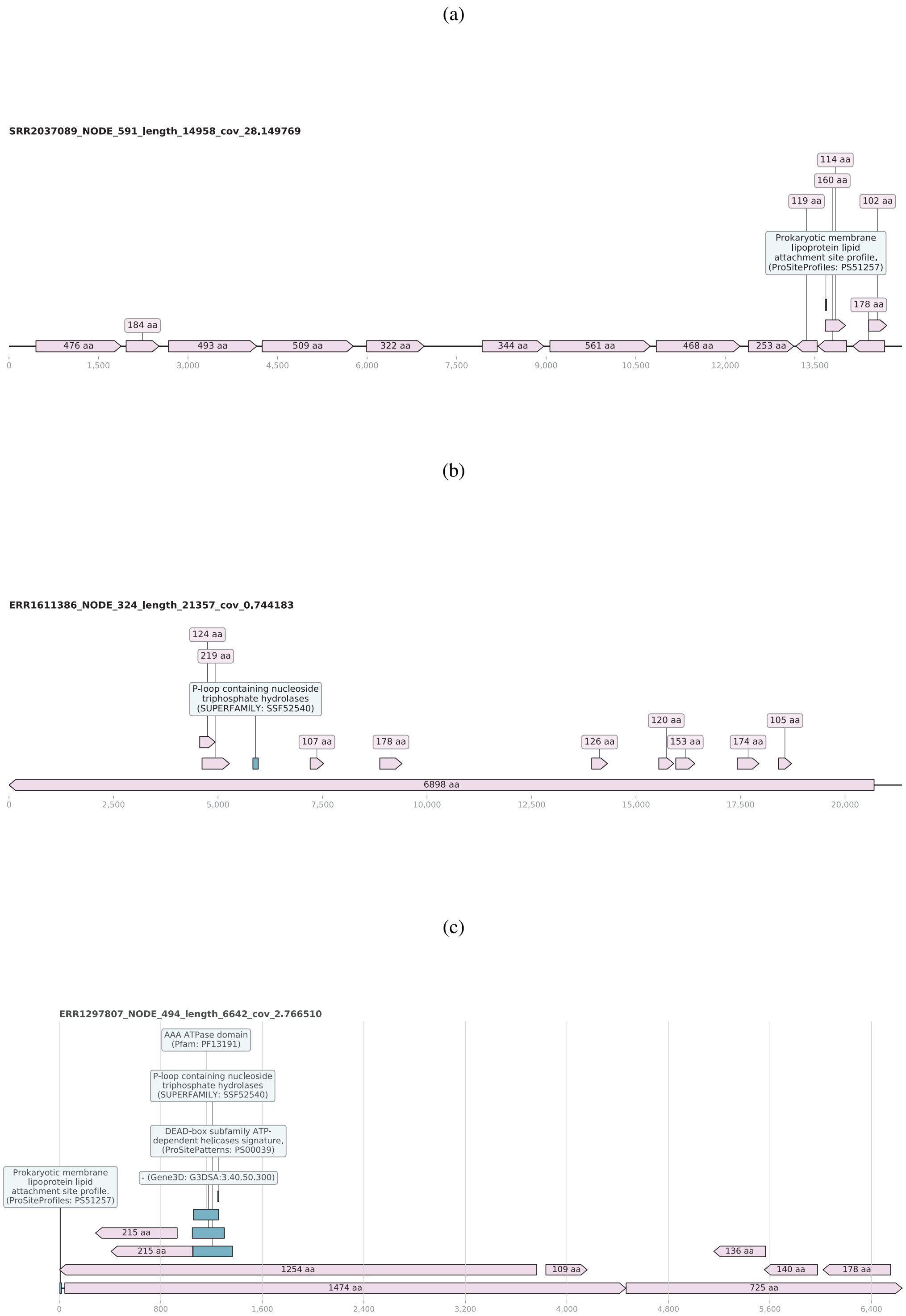


Figure S4: Distribution of contig lengths for all unknown contigs after the final time point (14 Oct 2020)



(a)
SRR2037089_NODE_591_length_14958_cov_28.149769
114 aa
160 aa
119 aa
102 aa
Prokaryotic membrane lipoprotein lipid attachment site profile. (ProSiteProfiles: PS51257)
178 aa
184 aa
476 aa
493 aa
509 aa
322 aa
344 aa
561 aa
468 aa
253 aa
0
1,500
3,000
4,500
6,000
7,500
9,000
10,500
12,000
13,500
(b)
ERR1611386_NODE_324_length_21357_cov_0.744183
124 aa
219 aa
P-loop containing nucleoside triphosphate hydrolases (SUPERFAMILY: SSF52540)
107 aa
178 aa
126 aa
120 aa
153 aa
174 aa
105 aa
6898 aa
0
2,500
5,000
7,500
10,000
12,500
15,000
17,500
20,000
(c)
ERR1297807_NODE_494_length_6642_cov_2.766510
AAA ATPase domain (Pfam: PF13191)
P-loop containing nucleoside triphosphate hydrolases (SUPERFAMILY: SSF52540)
DEAD-box subfamily ATP-dependent helicases signature. (ProSitePatterns: PS00039)
- (Gene3D: G3DSA:3.40.50.300)
Prokaryotic membrane lipoprotein lipid attachment site profile. (ProSiteProfiles: PS51257)
215 aa
215 aa
136 aa
1254 aa
109 aa
140 aa
178 aa
1474 aa
725 aa
0
800
1,600
2,400
3,200
4,000
4,800
5,600
6,400

Figure S5: The genome diagrams of large unknown contigs (UCs) show the open reading frames (ORFs) in the light pink shade with the ORFs lengths as their corresponding labels and the green boxes illustrating the InterProScan predicted presence of domain signature. (a) Cluster representative of the largest cluster comprising UCs across multiple microbiomes. (b) UC with the largest predicted ORF (6,898 AA). (c) Cluster representative of the largest single microbiome cluster derived from fecal microbiome. This cluster comprised of 153 unknown contigs assembled from 46 samples across 8 different studies.

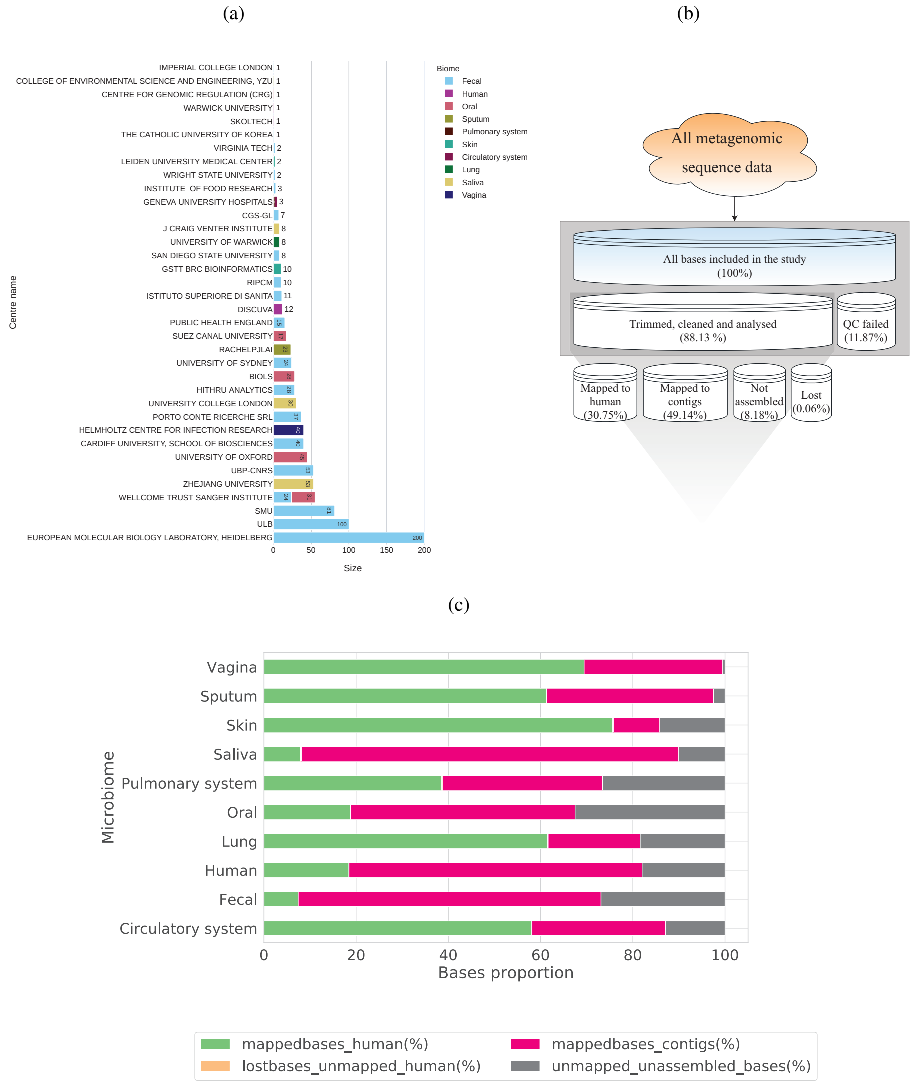
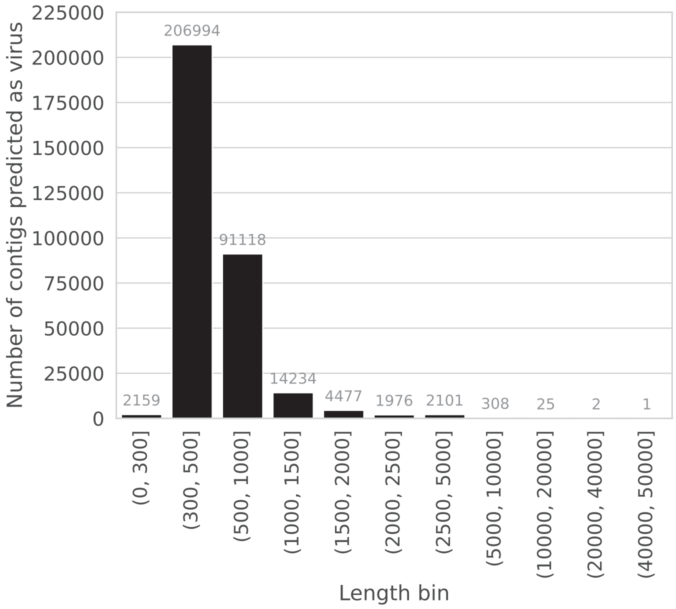


Figure S6: (a) Research centres around the world coloured according to microbiome data sets distributed among them (n=963) An overview of all bases analysed and categorised in this study. (b) Overall categorisation of all bases included in the study. (c) The proportion of bases mapped to the human genome, assembled contigs and bases that were not assembled in the different microbiomes.



(b)

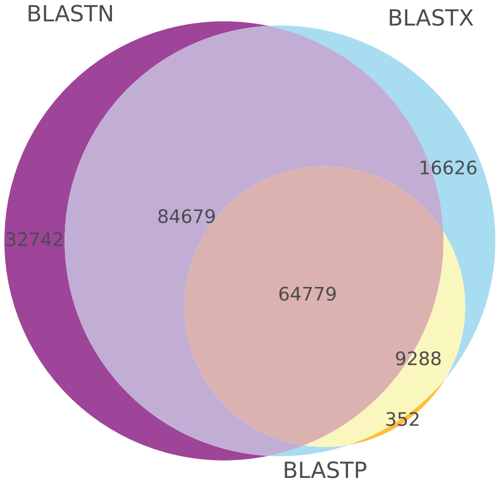


Figure S7: (a) Contig length distribution of unknown contigs predicted as virus using DeepVirFinder with qvalue<0.05. 48.33% of UCs that were <1kb long, 76.27% of UCs between 1-5kb long, and 96.55% of UCs that were at least 5kb long were predicted as viruses. (b) A Venn diagram comparing UCs to IMG/VR databases. Out of the complete set of 651,529 UCs, 442,970 UCs did not have a hit to protein and nucleotide sequences included in IMG/VR. 208,559 UCs were matching to IMG/VR sequences using one or more of three different approaches; BLASTN (n=182,293), BLASTP (n=74,512) and BLASTX (n=175,372). Overlapping UCs shared between all different approaches are shown in the Venn diagram here. 64,779 UCs were found to have an IMG/VR hit for BLASTN, BLASTP and BLASTX methods. The largest overlapping UC set was between BLASTN and BLASTX that shared 149,458 UCs whereas the smallest overlap was found between BLASTN and BLASTP which was 64,872 UCs. BLASTP and BLASTX methods shared 74,067 UCs.

Supplementary tables

Table S1: List of BioProjects analysed in this study with corresponding metadata.

Microbiome	BioProject	Description	Library Layout	Samples	Country	PMID
Circulatory system	PRJEB21816	Aortite	PAIRED	1	Switzerland	30991963
Circulatory system	PRJEB24753	Endocarditis due to Neisseria meningitidis	PAIRED	2	Switzerland	31448292
Fecal	PRJEB10865	Study of the abundance of bacteria from human samples	PAIRED	2	USA	
Fecal	PRJEB11554	NeoM	PAIRED	1		
Fecal	PRJEB12357	Impact of faecal microbiota transplantation on the intestinal microbiome in metabolic syndrome patients	PAIRED	100	Netherlands	27126044
Fecal	PRJEB14935	Term and preterm shotgun samples	PAIRED	3	United Kingdom	29096601
Fecal	PRJEB15257	The antibiotic resistance potential of the preterm infant gut microbiome measured using shotgun metagenomics.	PAIRED	15	United Kingdom	28149696
Fecal	PRJEB1775	Diagnostic Metagenomics: A Culture-Independent Approach to the Investigation of Bacterial Infections	PAIRED	53	Germany	23571589
Fecal	PRJEB17784	The fecal microbiota in L-DOPA naive PD patients	SINGLE	100	Germany	
Fecal	PRJEB18265	Estimation of variability in the gut microbiota resistome of the Russian citizens aimed at identification of pathways for transmission and spread of antibiotic resistance.	PAIRED	10	Russia	31507546
Fecal	PRJEB19090	Potential and active functions in the gut microbiota of a healthy human cohort	PAIRED	37	Italy	28709472

Microbiome	BioProject	Description	Library Layout	Samples	Country	PMID
Fecal	PRJEB19367	Analysis of stool samples from sickle cell disease patients and healthy controls	PAIRED	28	USA	
Fecal	PRJEB21696	Metagenomics 1st 5 data	SINGLE	1	South Korea	
Fecal	PRJEB23207	Metagenomic characterization of the human intestinal microbiota in faecal samples from STEC-infected patients	PAIRED	11	Italy	
Fecal	PRJEB23207	Metagenomic characterization of the human intestinal microbiota in faecal samples from STEC-infected patients	PAIRED	11	Netherlands	
Fecal	PRJEB5761	Gut microbiota in chronic kidney disease	PAIRED	81		
Fecal	PRJEB6092	Metagenome fecal microbiota- Illumina seq reads of 12 individuals at 2 timepoints	PAIRED	24	Australia	
Fecal	PRJEB6542	Gut microbial metabolism shifts towards a more toxic profile with supplementary iron in a kinetic model of the human large intestine	PAIRED	8	Netherlands	
Fecal	PRJEB7331	metagenomic analysis of human gut microbiome	PAIRED	24	United Kingdom	
Fecal	PRJEB7949	The fecal microbiome was studied in a group of IBD suffers and compared to a control group's fecal microbiome	PAIRED	40	United Kingdom	
Fecal	PRJEB8094	The initial state of the human gut microbiome determines its reshaping by antibiotics	PAIRED	100	Canada	26359913
Fecal	PRJEB8201	Comparison of distal gut microbiota structure and function in US and Egyptian children	SINGLE	2	Egypt	

Microbiome	BioProject	Description	Library Layout	Samples	Country	PMID
Fecal	PRJEB8201	Comparison of distal gut microbiota structure and function in US and Egyptian children	SINGLE	2	USA	
Fecal	PRJNA43253	Human fecal microbiome	SINGLE	7		20363958
Human	PRJEB14301	CSF	SINGLE	1	Russia	
Human	PRJEB21827	A/B testing for colon model	PAIRED	12		
Human	PRJEB6045	metagenomics of medieval human remains from Sardinaia	PAIRED	1	Italy: Sardinia	
Lung	PRJEB7248	Metagenomics of TB-associated sputum	PAIRED	8	Gambia	
Oral	PRJEB12831	A plaque on both your houses. Exploring the history of urbanisation and infectious diseases through the study of archaeological dental tartar	PAIRED	31	United Kingdom	
Oral	PRJEB12998	Test file for Oralfungi project	PAIRED	1		
Oral	PRJEB15334	Radcliffe dental calculus	SINGLE	45	United Kingdom	
Oral	PRJNA230363	Oral Microbiome	PAIRED	28	China	
Oral	PRJNA384402	oral metagenome Metagenome	PAIRED	17	Egypt	
Pulmonary system	PRJEB20877	Detection of bacterial pathogens from broncho-alveolar lavage by next-generation sequencing	PAIRED	2	Switzerland	28930150
Saliva	PRJEB14383	Oral microbiome samples from the Philippines	PAIRED	30	Philippines	29165844
Saliva	PRJNA264728	Gene expression anlayses of saliva-derived in vitro biofilms during carbohydrate fermentation and pH stress	PAIRED	8	not applicable	26023872
Saliva	PRJNA306560	Human oral saliva Metagenome	PAIRED	53	China: Sichuan	

Microbiome	BioProject	Description	Library Layout	Samples	Country	PMID
Skin	PRJEB10133	These samples are selections from a larger cohort that were selected for the participation in the EBI metagenomics training in Sept. 2015	PAIRED	10	United Kingdom	
Skin	PRJEB10295	Whole genome sequencing of metagenomes extracted from palms of two individuals	PAIRED	2	Netherlands	
Sputum	PRJEB10919	Total RNA-Seq on sputum samples from patients with active tuberculosis	SINGLE	23	South Africa	28494243
Sputum	PRJEB10919	Total RNA-Seq on sputum samples from patients with active tuberculosis	SINGLE	23	United Kingdom	28494243
Sputum	PRJEB14539	Metagenome and metatranscriptome profiling of moderate and severe COPD sputum in Taiwanese Han males	PAIRED	1	Taiwan	
Vagina	PRJEB21446	Metatranscriptome reveals the function shifts of the vaginal microbiome during the treatment of bacterial vaginosis	PAIRED	40	Germany	29875146

Table S2: Tools and their versions used for UnXplore analyses described in this study.

Software	Version	Source
BBDuk	38.22	https://sourceforge.net/projects/bbmap/
BBNorm	38.22	https://sourceforge.net/projects/bbmap/
BLASTN	2.9.0	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+ /LATEST/
BWA	0.7.17-r1188	http://bio-bwa.sourceforge.net/
DIAMOND	0.9.21.122	https://github.com/bbuchfink/diamond
ETE3	3.1.2	http://etetoolkit.org/
InterProScan	5.38-76.0	https://github.com/ebi-pf-team/interproscan
Parallel-fastq-dump	0.6.6	https://github.com/rvalieris/parallel-fastq-dump/
Python	3.6.7	https://www.python.org/downloads/
SAMTools	1.7	http://www.htslib.org/
Snakemake	5.4.5	https://bitbucket.org/snakemake/
SPAdes	3.11.1	http://cab.spbu.ru/software/spades/

Supplementary text

CheckV predicted viral contigs

All UCs were processed through the CheckV pipeline [52] to identify the UCs that were likely to belong to viruses. A total of 47,702 UCs were predicted to be of viral origin and 7,696 of them were at least 1kb long. A set of 11,121 of these UCs were predicted to have at least one viral gene and 1,712 UCs of this subset were at least 1kb long. These 1,712 UCs were mapped against the results of the most recently carried out BLASTX analysis for validation. 529 of the predicted viral contigs matched to bacterial protein sequences with low sequence identity with a mean percent identity of 48.43. However, these results are based on short protein sequence hits on bacterial proteins indicating that these UCs are likely to be phage genomic signatures that matched bacterial protein in absence of a phage sequence in the database specifically as protein-based similarity searches are able to identify distantly related homologues of query sequences. These results can help to stipulate that the actual diversity of virus sequences present in the UCs set is largely underestimated. It is highly likely that a range of contigs that match distantly related protein sequences of bacterial origin are in fact derived from unknown and uncultured novel viruses, such as phages, that infect bacteria.

The large unknown contigs

The largest multi metagenome UCs (figure S5(a)), was assembled from SRR2037089 from the oral metagenome and was 14,958 bases long. It was clustered with 33 other contig sequence assembled from 12 distinct samples from oral (n=8; PRJNA230363), sputum (n=3; PRJEB10919) and saliva (n=23; PRJEB14383) microbiomes. These three distinct studies contained samples from distinct geographic locations: PRJNA230363 from China, PRJEB14383 from the Philippines and PRJEB10919 from South Africa suggesting that this unknown organism is broadly distributed in its association with humans. The second-largest member of this cluster was 9,791 bases long and was assembled from a separate sample (SRR2037087) from the same study. This large contig was deemed to be identical to the cluster representative. The largest cluster member from the saliva microbiome was 1.5kb long and was assembled from ERR1474566. On the contrary, the contigs assembled from the sputum microbiomes were significantly smaller with lengths ranging between 479-533 bases, indicating the fragmented assembly and the presence of partial sequences.

A large contig of length 21,357 was identified in the oral microbiome shown in figure S5(b). This contig was assembled from run ERR1611386 and was clustered with 16 other sequences from BioProjects PRJEB12831 and PRJEB15334. Other members of the clusters originated from 5 distinct samples and were between 306-6,109 bases long. This contig contained the largest predicted ORF that was 6,898 residues long. 14 out of the 16 other contigs within the cluster contained partial sequences belonging to this ORF. This contig also did not have a taxonomic homologue identified in any of the most recent similarity sequence-based searches. Additionally, the largest ORF was predicted to contain P-loop containing nucleoside triphosphate hydrolases (SUPERFAMILY: SSF52540) signatures.

The largest cluster contained 153 sequences (figure S5(c)) had a cluster representative that was 6,642 bases long assembled from sample ERR1297807 from PRJEB12357. This cluster representative was predicted to contain 9 distinct ORFs. The cluster contained 35 other sequences that were at least 1kb long. Additionally, other contigs (6,015 bases long from ERR537012 and 5,344 bases long from ERR537011) from as a separate study (PRJEB6542) were found in this large cluster.

viruses

MDPI

Article

NCBI's Virus Discovery Codeathon: Building "FIVE" —The Federated Index of Viral Experiments API Index

Joan Martí-Carreras [1,*], Alejandro Rafael Gener [2,3,4,5,*], Sierra D. Miller [6], Anderson F. Brito [7], Christiam E. Camacho [8], Ryan Connor [8,*], Ward Deboutte [1], Cody Glickman [9], David M. Kristensen [10], Wynn K. Meyer [11], Sejal Modha [12], Alexis L. Norris [13], Surya Saha [14,15], Anna K. Belford [16], Evan Biederstedt [17], James Rodney Brister [8], Jan P. Buchmann [18], Nicholas P. Cooley [19], Robert A. Edwards [20], Kiran Javkar [21,22], Michael Muchow [23], Harihara Subrahmaniam Muralidharan [24,25], Charles Pepe-Ranney [26], Nidhi Shah [21], Migun Shakya [27], Michael J. Tisza [16], Benjamin J. Tully [28], Bert Vanmechelen [1], Valerie C. Virta [29], Jake L. Weissman [30], Vadim Zalunin [8], Alexandre Efremov [8] and Ben Busby [8,31,*]

[1] Laboratory of Clinical and Epidemiological Virology, KU Leuven Department of Microbiology, Immunology and Transplantation, Rega Institute, Leuven BE3000, Belgium; ward.deboutte@kuleuven.be (W.D.); bert.vanmechelen@kuleuven.be (B.V.)

[2] Integrative Molecular and Biomedical Sciences Program, Baylor College of Medicine, Houston, TX 77030, USA

[3] Margaret M. and Albert B. Alkek Department of Medicine, Nephrology, Baylor College of Medicine, Houston, TX 77030, USA

[4] Department of Genetics, MD Anderson Cancer Center, Houston, TX 77030, USA

[5] School of Medicine, Universidad Central del Caribe, Bayamón, Puerto Rico 00960, USA

[6] Genetics & Molecular Biology, Millersville University, 40 Dilworth Rd, Millersville, PA 17551; sierradesireemiller@gmail.com

[7] Department of Epidemiology of Microbial Diseases, Yale School of Public Health (YSPH), 60 College Street, New Haven, CT 06510, USA; anderson.brito@yale.edu

[8] National Center for Biotechnology Information, U.S. National Library of Medicine, National Institutes of Health, 9000 Rockville Pike, Bethesda, MD 20894, USA; camacho@ncbi.nlm.nih.gov (C.E.C); jamesbr@ncbi.nlm.nih.gov (J.R.B.)zaluninvv@ncbi.nlm.nih.gov (V.Z.); alexandre.efremov@nih.gov (A.E.)

[9] Laboratory of Clinical and Epidemiological Virology, KU Leuven Department of Microbiology, Immunology and Transplantation, Rega Institute, Leuven BE3000, Belgium; cody.glickman@cuanshutz.edu

[10] Computational Bioscience Program, University of Colorado Anschutz, Aurora, CO 80045, USA; dk131363@gmail.com

[11] AAAS Science and Technology Policy Fellow, American Association for the Advancement of Science, 1200 New York Ave NW, Washington, DC 20005, USA; wynn.meyer@gmail.com

[12] MRC-University of Glasgow Centre for Virus Research, G61 1QH Glasgow, UK; s.modha.1@research.gla.ac.uk

[13] Biotechnology Graduate Program, University of Maryland Global Campus, 1616 McCormick Drive, Largo, MD 20774, USA; alexisleighnorris@gmail.com

[14] Boyce Thompson Institute, Ithaca, NY 14850, USA; ss2489@cornell.edu

[15] School of Animal and Comparative Biomedical Sciences, The University of Arizona, Tucson, AZ 85721, USA

[16] Laboratory of Cellular Oncology, National Cancer Institute, 37 Convent Dr., Bethesda, MD 20894, USA; belfordak@nih.gov

[17] Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA; evan.biederstedt@gmail.com

[18] School of Life and Environmental Sciences and School of Medical Sciences, Marie Bashir Institute for Infectious Diseases and Biosecurity, The University of Sydney, Sydney, Australia; jan.buchmann@sydney.edu.au

[19] Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15260, USA; npc19@pitt.edu
[20] College of Science and Engineering, Flinders University, Bedford Park, SA 5042, Australia; robert.edwards@flinders.edu.au
[21] Department of Computer Science, University of Maryland, College Park, MD 20740, USA; kjavkar@cs.umd.edu
[22] Joint Institute for Food Safety and Applied Nutrition, College Park, University of Maryland, MD 20740, USA
[23] Novel Microdevices, Nucleic Acids, Baltimore, MD 21202, USA; michael@novelmicrodevices.com
[24] Department of Computer Science, University of Maryland, College Park, MD 20740, USA; hsmurali@cs.umd.edu
[25] Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20740, USA
[26] AgBiome, 104 TW Alexander, Research Triangle, NC 27709, USA; cpeperanney@agbiome.com
[27] Bioscience Division, Bikini Atoll Road, Los Alamos National Laboratory, Los Alamos, NM 87545, USA; migun@lanl.gov
[28] Center for Dark Energy Biosphere Investigations, University of Southern California, Los Angeles, CA 90089, USA; tully.bj@gmail.com
[29] AAAS Science & Technology Policy Fellow, National Institutes of Health, Center for Information Technology, 6555 Rock Spring Drive, Bethesda, MD 20817, USA; valerie.virta@nih.gov
[30] Department of Marine and Environmental Biology, University of Southern California, Los Angeles, CA 90089, USA; jakeweis@usc.edu
[31] DNANexus, 1975 W El Camino Real #204, Mountain View, CA 94040, USA; ben.busby@gmail.com; ORCiD: 0000-0001-5267-4988
* Correspondence: joan.marti@kuleuven.be (J.M.C); gener@bcm.edu (A.R.G.); ryan.connor@nih.gov (R.C.); ben.busby@gmail.com (B.B.)

Academic Editor: Manja Marz, Bashar Ibrahim, Franziska Hufsky, Ronald Dijkman, Alban Ramette and Jenna Kelly

Received: 16 October 2020; Accepted: 2 December 2020; Published: 8 December 2020

Abstract: Viruses represent important test cases for data federation due to their genome size and the rapid increase in sequence data in publicly available databases. However, some consequences of previously decentralized (unfederated) data are lack of consensus or comparisons between feature annotations. Unifying or displaying alternative annotations should be a priority both for communities with robust entry representation and for nascent communities with burgeoning data sources. To this end, during this three-day continuation of the Virus Hunting Toolkit codeathon series (VHT-2), a new integrated and federated viral index was elaborated. This Federated Index of Viral Experiments (FIVE) integrates pre-existing and novel functional and taxonomy annotations and virus–host pairings. Variability in the context of viral genomic diversity is often overlooked in virus databases. As a proof-of-concept, FIVE was the first attempt to include viral genome variation for HIV, the most well-studied human pathogen, through viral genome diversity graphs. As per the publication of this manuscript, FIVE is the first implementation of a virus-specific federated index of such scope. FIVE is coded in BigQuery for optimal access of large quantities of data and is publicly accessible. Many projects of database or index federation fail to provide easier alternatives to access or query information. To this end, a Python API query system was developed to enhance the accessibility of FIVE.

Keywords: Data Federation, CRISPR, Protein Domain, Metagenomics, Virus, Genome Graphs, HIV-1

1. Introduction

While the sharp reduction in the cost of sequencing over the past 15 years [1] is leading to the progressive democratization of biomolecule sequencing and experimental data production, the resulting data influx represents a nightmare of data storage, management, accessibility, and analysis. As of November 2019, the Sequence Read Archive (SRA) contained almost 13 petabases of open information, with close to 20 more petabases in the queue [2] (data growth is periodically updated at [3]). Tackling the complexity and depth of this information is daunting. Despite the existence of stable general repositories (Genbank, ENA, DDBJ), a growing number of specialized databases are leaving the results of biological experiments (mostly sequence data) disconnected, sparse, disorganized, and often inaccessible. Data accessibility and data federation, the virtualization of sparse databases into a common platform, represent the most important assets to the wider scientific community. There have been previous attempts to federate such databases and to make them open access, specifically for viral sequences [4].

The National Institutes of Health (NIH) continues to promote such efforts through the Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative, which "provides cost-effective access to industry-leading partners to help advance biomedical research" ([5]). These partnerships enable access to rich datasets and advanced computational infrastructure, tools, and services. The STRIDES Initiative is one of many NIH-wide efforts to implement the NIH Strategic Plan for Data Science, which provides a roadmap for modernizing the NIH-funded biomedical data science ecosystem ([6]). The National Center for Biotechnology Information (NCBI) [7] leverages their participation in the STRIDES Initiative in part by organizing and supporting a series of events, such as hackathons and codeathons, to engage researchers and general users of NIH resources to improve NIH resources ([8]). These events usually span a three-day period and are geared towards addressing a specific research problem or topic. Through these events, the NIH receives active feedback from their user community on existing resources that helps improve the quality and output of future NCBI products. Typically, milestones or minor objectives are brainstormed during an online organizational meeting before the event. These objectives form the core of the working groups during the event. Over the course of the three days, each working group produces a solution to a specific task, often collaborating and integrating their solution with the products from the other working groups. The event concludes with working groups presenting their solutions.

NCBI also leveraged their participation in the STRIDES Initiative by moving SRA to the cloud. SRA is the largest source of open, publicly available next-generation sequencing (NGS) data from diverse biological sources. SRA serves as an umbrella for a variety of sequencing experiments (e.g., amplicons, whole genome sequencing, and environmental metagenomics) from different platforms (namely IonTorrent, Illumina, Oxford Nanopore, and PacBio) and applications. The first collaborative attempt to annotate and index SRA datasets in bulk was conducted as part of the inaugural of this series of events, the Virus Hunting Toolkit (VHT) [4]. This first event (VHT-1) challenged users to harness the power of the Google Cloud environment to test and develop bioinformatics pipelines to identify all viruses, including previously characterized and novel, in existing publicly available SRA datasets. The working groups were organized by specific tasks to emphasize the exhaustive separation of known viral diversity from the bulk data, and the identification of possible new viral sequences: data selection, taxonomic and cluster identification, and annotation of domains and genes. From that first codeathon, a set of 5.5×10^7 reassembled contigs, from 2953 SRA entries, were produced (see Table 1 on accessing these resources) [4]. All contigs were classified and annotated, and this information was integrated into a complete dataset [9]. Despite the efforts made, some areas were not covered in the final annotation files, such as virus–host pairing predictions. An additional impending limitation was developing concise strategies to store and visualize sequencing data from related biological entities. In our second series of Virus Hunting in the Cloud 2.0 (VHT-2), we present here the first Federated Index of Viral (Sequencing) Experiments (FIVE).

In contrast to traditional sequence data databases, federated indices do not store raw sequence data. Instead, federated indices store the results from different sequence analysis. For example, to identify putative novel viruses in SRA contigs, a federated index would store and link results from several analyses on these contigs, allowing researchers to quickly identify SRA contigs of interest without the need to perform the same extensive underlying analysis. Therefore, the output of FIVE is a collation of analysis results from different methods indicating if and where a sequence contains viral or virus-like signals, and not a single similarity score to a known virus sequence or structure.

This novel resource indexes sequences, metadata, and hyperdata from the VHT-1 hackathon, sparse databases, and online resources. More than 2953 SRA entries assembled during VHT-1 [4] were reanalyzed and their annotations included in FIVE. Additionally, existing databases (i.e., taxonomy, CRISPR, etc.) were federated into our index, expanded with novel annotations. Finally, FIVE includes the first attempt to condense viral genome variation (nucleotide diversity along the genome) in an indexable genome graph ([10]), using HIV-1 as a proof-of-concept. The emerging field of genome graphs can provide an efficient method to summarize and index sequence diversity data for a single species [11]. Unlike previous efforts, FIVE is a publicly accessible index where several methods were built to easily query and retrieve information from the index. The index accession methods were wrapped into an easy-to-use Application Programming Interface (API) written in Python to further improve its accessibility. The FIVE index links SRA with accurately assembled contigs, viral and host taxonomy annotation, protein/functional annotation (assisting taxonomy identification), and virus–host pairing predictions. FIVE was generated by (i) federating or mining existing datasets, (ii) implementing novel methods for annotation, and (iii) indexing and improving data access. Several teams were formed, focusing on different aspects to generate FIVE. During and after the VHT-2, four core aspects of FIVE can be distinguished: (i) protein domain recognition and computation scalability, (ii) virus–host pairing prediction, (iii) viral genome variability indexing in genome graphs, and (iv) index structure and accessibility.

New functional and taxonomic annotation methods were developed, covering similarity search by alignment, probabilistic models (hidden Markov models [HMMs]), and *k*-mer distances. These methods were added to the extensive information in our previously published SRA annotations [4]. Known virus–host pairings were federated from existing databases for both eukaryotic and prokaryotic viruses and expanded through novel phage annotation and CRISPR profiling. Besides taxonomy or function, genome variability is another desirable layer of information to consider for virus diversity. Mutations in viral genomes can help to trace geographical patterns, distinguish closely related viruses, or to predict protein functionality and therefore infective properties. Genome graphs are an emerging tool to compress genomic information (e.g., variants) from closely related genomes into more compact formats compared to multiple linear reference genomes or multiple sequence alignments. Despite their compactness, if many variants exist, such graphs can become considerably large. Such diversity is of interest to have indexed together with other types of data. With FIVE we present a proof-of-concept of viral genome graphs indexing and systematization. We refined a new and compact approximate *k*-mer graph creation tool, SWIft Genomes in a Graph (SWIGG), which was used to model the genome variation of full-length HIV-1 reference genomes. Finally, FIVE is intended to be a free, public, usable database by a broader audience, therefore besides the publication of FIVE as a public BigQuery index, we designed an easy-to-use Python application programming interface (API), and some methods, to query FIVE. This index is the first attempt to centralize and federate viral (meta)data and annotations directly from SRA. As such, FIVE is distinct from, yet complementary to, other resources such as ViPR [12] or NCBI Viral Resources.

2. Materials and Methods

2.1. Protein Domain Recognition and Computation Scalability

To improve the contig annotation index in FIVE, we evaluated 2953 datasets from VHT-1 [4,13] against 2082 viral-specific protein domains selected from the CDD database (see [14]). We designed two pipelines: (i) Reverse Position Specific tBLASTn (RPS-tBLASTn) [15] and (ii) Mash pipelines [16] to identify known viral protein domains within VHT-1-assembled contigs (see Table 1). RPS-tBLASTn is a robust method for domain detection, although its computational demands render it challenging for large-scale data annotation, such as in VHT-2. In this context, a distance estimator, based on protein sequence sketches, can be used to identify protein domains with less computational demand. Mash, a metagenome distance estimation tool (based on MinHash dimensionality reduction), [16] was used with amino acid *k*-mer size = 6. A subset of the 728 datasets was used to compare RPS-tBLASTN and Mash ([17]) to assess the performance of the distance estimation. Recall percentages for the Mash pipeline were calculated per dataset by dividing the 'true positive' viral CDDs by the sum of 'true positive' and 'false negative' viral CDDs. Precision percentages for the Mash pipeline were calculated per dataset by dividing the 'true positive' viral CDDs by the sum of 'true positive' and 'false positive' viral CDDs. A CDD hit was considered either a true positive if it was retrieved in the same dataset, a false positive if it was retrieved by the Mash pipeline but not by the RPS-tBLASTn pipeline, and a false negative if it was retrieved by the RPS-tBLASTn pipeline but not by the Mash pipeline. Clustering was performed on the Canberra distance matrices derived from the domain counts matrices using base R function *hclust* (stats package v3.5.3) [18]. Correlation between both matrices was calculated with the Mantel test implemented in the *ade4* R package (v1.7.-15) [19]. Normalized Robinson–Foulds metrics were calculated with the *RF.dist* function in the *Phangorn* R package (v2.5.5) [20], and entanglement values and tanglegram were calculated and plotted using the *dendextend* package (v1.13.4) in R (v3.6.2) [21]. Additionally, HMMER (v3.1) [22] was explored as a short read taxonomic annotator, providing an alternative to a computationally intensive de novo assembly step.

A schematic of both pipelines can be seen in Figure 1, each of which two sets of inputs: (i) a set of 2953 datasets containing assembled contigs constructed during VHT-1 [4], and (ii) a selected set of 2082 virus-associated Conserved Domains Database (CDD) entries (personal communication from J. Rodney Brister, NCBI RefSeq [23]; [24] Supplementary File S1).

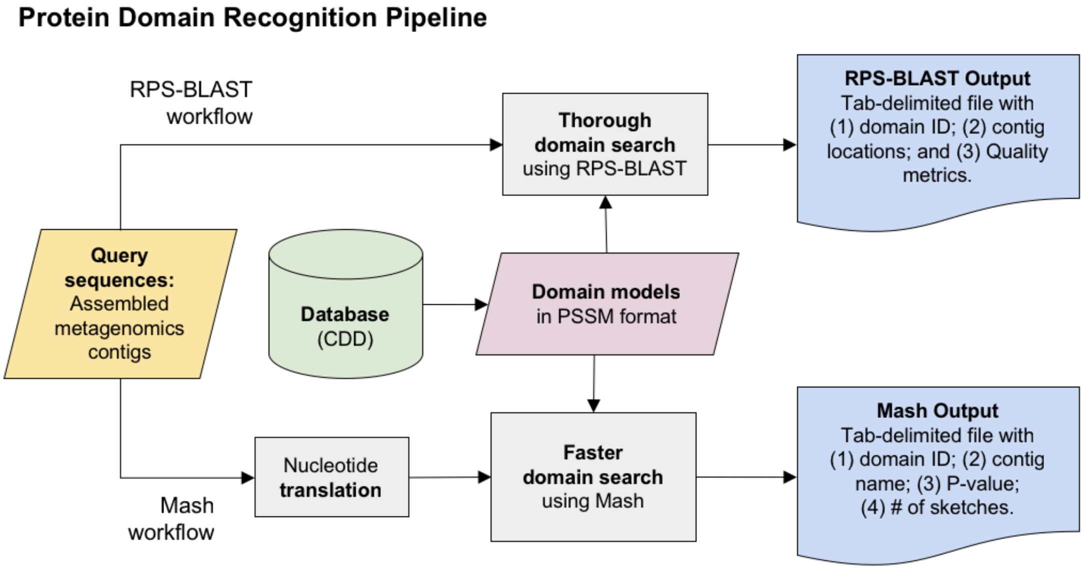

Protein Domain Recognition Pipeline
RPS-BLAST workflow
Thorough domain search using RPS-BLAST
RPS-BLAST Output
Tab-delimited file with (1) domain ID; (2) contig locations; and (3) Quality metrics.
Query sequences: Assembled metagenomics contigs
Database (CDD)
Domain models in PSSM format
Nucleotide translation
Faster domain search using Mash
Mash Output
Tab-delimited file with (1) domain ID; (2) contig name; (3) P-value; (4) # of sketches.
Mash workflow

Figure 1. Protein Domain Recognition Pipeline. Using 2082 entries from CDD (Conserved Domains Database) domain models in PSSM (Position-Specific Scoring Matrix) format, we tested two pipelines: RPS-BLAST and Mash. RPS-BLAST, with known domain models matched against assembled contigs, is accurate but computationally expensive. Mash pipeline was tested, which is significantly faster, and can be applied directly on unassembled reads.

We matched CDD entries against the dataset using RPS-BLAST, with 6-frame translation (RPS-tBLASTn) ([25]). We filtered them to include only higher-quality output hits with e-value ≤ $1x10^{-10}$ and coverage threshold length ≥ 50 nucleotides.

In addition to RPS-BLAST and Mash, we tested the feasibility of domain detection using HMMER [22] directly against short sequence reads. We generated a simulated dataset of 1000 reads (each 150 bases long) by randomly extracting simulated reads from a complete genome of Human Herpesvirus 1 (HHV-1, GenBank accession JN555585). The DNA sequences were translated into the six possible reading frames using Biopython ([26]), yielding 6000 short "peptides" (~50 amino acids). This simulated dataset enabled us to evaluate HMMER (*hmmscan*) for searching for domains in short reads.

2.2. Virus–Host pairing Prediction

FIVE included an index for virus–host pairing (which viruses can infect which organisms). Initially, we federated a reference dataset of experimentally confirmed viral–host pairings, generating 22,896 known virus–host pairings from the NCBI Virus Variation Resource database [27] and supplemented with records from PhagesDB [28]. Missing information from PhagesDB, including host and virus taxonomic identification numbers and taxonomic lineages, was retrieved using the NCBI Taxonomy Browser [29] and incorporated into the index (see custom scripts in [30]). In several instances, phage genomes without a specified host in the database possessed a bacterial genome in the phage name that allowed for relationship inference.

The federated index was expanded upon using CRISPR spacer connections. A large-scale CRISPR spacer merged from the initial federated index and four datasets generated using distinct sources: The first dataset of CRISPR spacers was CRISPRCasdb ([31]; accessed November 6, 2019), where spacers were identified using the tool CRISPRCasFinder [32] from "reference" and "representative" microbial genomes available in RefSeq [23]. The second dataset of CRISPR spacers [33] was identified from all prokaryotic assemblies in RefSeq [23] (December 2017) using CRISPRDetect [34]. The third dataset of CRISPR spacers was identified from 24,345 metagenome-assembled genomes (MAGs) of the human microbiome [35] using MinCED ([36]), based on the CRISPR Recognition Tool [37] (parameters: -spacers -gffFull). The fourth dataset of CRISPR spacers was identified from the 24,706 species-representative sequences in Genome Taxonomy Database (GTDB) [38] using MinCED. All genomes/MAGs were provided using standardized taxonomy, based on the GTDB taxonomy. The resulting CRISPR spacer database was searched against the initially federated index of known virus–host pairings using BLASTn, with parameters set to account for the short size of the spacer regions (parameters: -task blastn-short, -evalue 0.01, -outfmt 6, -gapopen 10, -gapextend 2, -penalty "-1", -word_size 7, -dust no).

2.3. Viral Genome Diversity Indexing in Genome Graphs

As introduced earlier, we used genome graphs as proof-of-concept to index viral genome variability for a given set of closely related viruses. We used SWIft Genomes in a Graph (SWIGG) (commit 48c4661), a nascent genome graph builder, as a back-bone for FIVE's own implementation ([39]). In short, SWIGG creates genome graphs using *k*-mers from input genome sequences. The *k*-mer length can be set by the user, and *k*-mers used for genome graphs can be excluded by fine tuning the maximum and/or minimum *k*-mer counts within or across analyzed sequences. We added the functionality to parse metadata from sequence headers and incorporate both into the individual nodes of the genome graph. We used human Immunodeficiency Virus 1 (HIV-1) as test case for the integration of viral genome diversity into FIVE (sequences available at [40]). HIV-1 was selected because at the time of submission, it was the most well-studied human pathogen, having the most high-quality full- or near full-length genomes available ([41]), with robust feature annotations including structural features at the proviral DNA, viral RNA, and viral protein levels.

To balance sequence diversity with representative HIV-1 sequences, we used 170 HIV-1 reference genome sequences from the Los Alamos National Laboratory's HIV Sequence Database ([42]; accessed on November 4th, 2019). To retrieve these sequences, curated alignments were accessed with the following parameters: from "Alignments", "Curated alignments" was selected; Alignment type = "Subtype reference"; Pre-defined region of the genome = "GENOME"; subtype = "ALL"; DNA/protein = "DNA"; year = "2010". This number was narrowed down from 170 to 167 after removing "cpz" or SIV sequences. We also used a subset of these yielding 39 sequences by changing subtype = "M group without recombinants (A–K)". The implementations require Python v3.6 or higher and the Python package *NetworkX* (v2.4) [43].

2.4. Index Structure and Accessibility

We indexed data generated by working groups during the second codeathon ("Virus Hunting in the Cloud 2.0") in a relational database format on Google Cloud's BigQuery, called Federated Index of Viral Experiments (FIVE). This index can be visualized as distinct silos of the data generated by each of the analysis pipelines presented in this manuscript. Data was parsed ([44]), loaded into BigQuery using google-cloud-sdk (v288.0.0) tools, and SQL-like manipulation and subsetting of the tables was performed (scripts available at [45]). Data loaded into FIVE is made accessible to the public (see Table 1).

Table 1. List of repositories used in the generation of FIVE and its accessory information (contigs and FIVE link) hosted in GitHub and the first release of each repository frozen in ZENODO. VHT—Virus Hunting Toolkit.

Relevant Repositories	GitHub Project	Dataset Citation
Connor *et al.*, 2019 VHT [4]	https://github.com/NCBI-Hackathons/VirusDiscoveryProject	10.3390/genes10090714
VHT contig list	https://github.com/NCBI-Hackathons/VirusDiscoveryProject/blob/master/contigs_readme.md	10.17605/osf.io/g9w8r
VHT contig repository	https://storage.googleapis.com/experimental-sra-metagenome-contigs	10.17605/osf.io/g9w8r
Protein domain recognition and computation scalability	https://github.com/NCBI-Codeathons/Domain_HMM_Boundaries	10.5281/zenodo.4027168
Virus–Host pairing prediction	https://github.com/NCBI-Codeathons/Host_Phage_Interactions	10.5281/zenodo.4027172
Viral genome diversity indexing in genome graphs	https://github.com/NCBI-Codeathons/Virus_Graphs	10.5281/zenodo.4027629
Index structure and accessibility	https://github.com/NCBI-Codeathons/The_Virus_Index	10.5281/zenodo.4027617
FIVE	https://console.cloud.google.com/bigquery?p=virus-hunting-2-codeathon&d=viasq&page=dataset	-

One of the most important aspects of an index is accessibility. BigQuery is a relatively new framework and hence it was decided to generate a series of queryable actions, answering the most frequent research questions. To facilitate this research, a Python-based API was developed. This API, called viral-index (v0.0.3) (Figure 2, is freely available to download from PyPI ([46]). Installation of the viral-index module requires Python 3.7 and the Python packages pip (v20.0.2) and virtualenv (v20.0.18). The viral-index module relies on google-cloud-bigquery (v1.27), google-auth (v1.21.1), and twine (v3.2.0) Python3 modules, which should be downloaded as part of the viral-index module dependencies. After installation, the user must add the path to their google credentials, as system variable "GOOGLE_APPLICATION_CREDENTIALS". These credentials allow the user to access the BigQuery databases (detailed instructions at [47]) and query the federated indexes using a range of functionalities implemented in the viral-index module, which we describe in this manuscript. It is important to note that the viral-index API module supports data retrieval but not data manipulation. The viral-index module returns the data as standard list() or dict() objects that can be easily manipulated in order to carry out further analysis. Additionally, this framework enables incorporation of new or updated datasets when they become available.

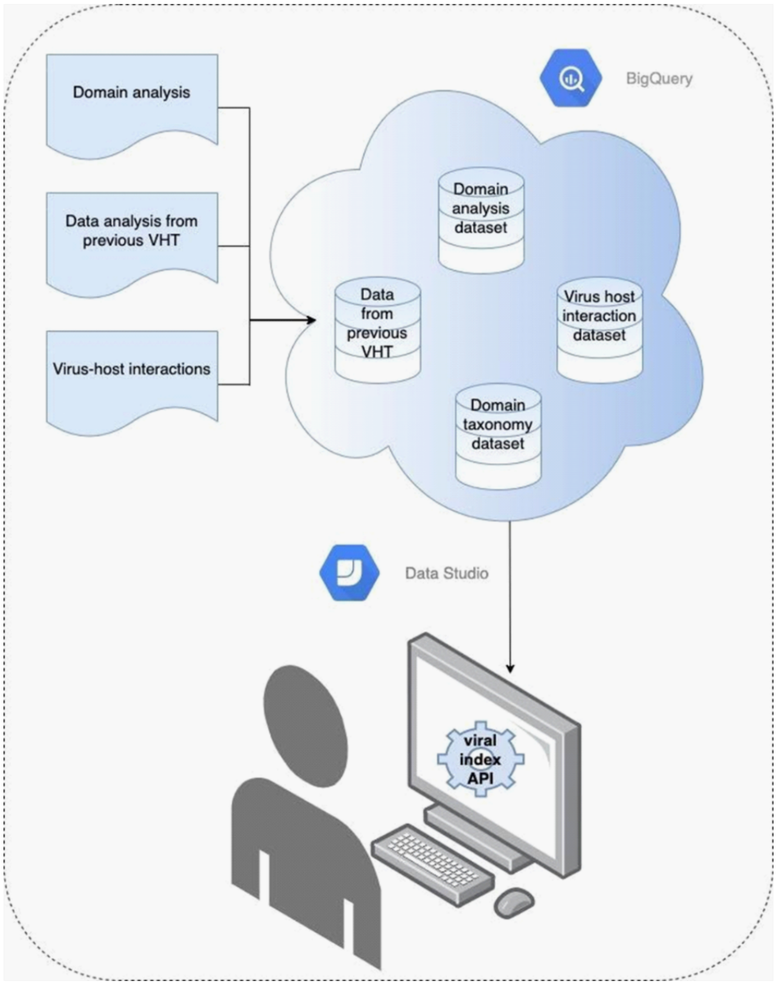


Figure 2. A schematic representation of Federated Index of Viral Experiments (FIVE) implementation, and interactions with users, enabled through the viral-index Application Programming Interface (API). Viral information generated in both codeathons is indexed in BigQuery on FIVE, accessible from Google Cloud, which can be easily queried using the viral-index API ([48]). This API enables users to perform a range of flexible searches on the FIVE databases with minimum code.

2.5. Data and Software Availability

A broad and detailed explanation of each method section can be found in the corresponding GitHub projects (Table 1). Each project contains complete instructions to fully reproduce the data generated and to reproduce FIVE. Links to the VHT contigs and FIVE are also made available (Table 1).

3. Results and Discussion

3.1. Protein Domain Recognition and Computation Scalability

We ran RPS-tBLASTn pipeline on contigs derived from 2953 public sequencing datasets assembled in VHT-1 [4] (see Table 1 for access to the contig list and contig repository from GoogleCloud), but we found it to be too computationally expensive to run with the complete CDD database over three days. Instead we used a subset of 2082 viral CDD models.

Since RPS-tBLASTn does not allow for query parallelization natively, we tried several parallelization strategies for RPS-tBLASTn in order to scale up its performance. Subsequently, we attempted to parallelize the RPS-tBLASTn search using a procedure whereby we divided the database into 60 segments and later combined the results for each segment. Splitting the database and later rejoining results affected the search space, and therefore the e-value of our results. While appropriate for testing purposes, for production runs we strongly recommended to use the *dbsize* parameter to account for the changed search space as a correction factor.

Among the 55,503,968 contigs that we searched against the viral specific CDDs ([49]), 10% of the contigs (5,606,754 from 2745 SRA datasets) had at least one CDD hit with e-value ≤ 1×10^{-3}. Using a more stringent e-value ≤ 1×10^{-10}, the number of contigs having at least one viral CDD was reduced to 0.5% (278,725; from 2534 entries). Hit distribution varied enormously, some contigs had multiple CDD hits, peaking at 22,560 hits (Contig ID = NC_003663.2:1.224499, Cowpox virus), but the majority of contigs (77.3%) had one unique CDD hit. The most common CDD was CDD:222853 (a transposase specific to the *Caudovirales* lineage).

In parallel, we tested the (meta)genomic distance estimation tool Mash (MinHash dimensionality reduction) [16,50] on predicted proteins (obtained through Prodigal protein prediction) from the contigs derived from 728 datasets ([51]), a subset of the 2953 datasets used in the previous RPS-tBLASTn analysis ([52]). Mash's default amino acid *k*-mer length of $k = 21$ (for both of its input sketches) retrieved almost no hits, a length of $k = 6$ was chosen arbitrarily. We could not assess other *k* values (which may have yielded better estimations) due to time restrictions.

A representative subset of 728 datasets was used to compare both the Mash and the RPS-tBLASTn pipelines. Out of 728 datasets, the Mash pipeline had an order of magnitude fewer hits (133,452 hits) than with RPS-tBLASTn (2,574,452 hits). The Mash pipeline was found to be substantially less sensitive than the RPS-tBLASTn analysis, with an average recall of 15.3% and a precision of 37.0%. Despite the low recall value, Canberra distance matrices retrieved from the Mash datasets and RPS-tBLASTn datasets were strongly correlated (Mantel test, p-value = 0.0009). Additionally, hierarchical clustering implied that, despite the loss of global structure in the dataset (as shown by a Robinson–Foulds distance of 0.91 and an entanglement of 0.2), the Mash pipeline can be used as a fast tool to quickly identify datasets containing roughly similar viral domains (see Figure 3). While the Mash pipeline performed significantly faster than RPS-tBLASTn, the prodigal translation step still represented a significant bottleneck for scaling. Faster translation algorithms would increase the feasibility of the Mash pipeline. While the fast Mash pipeline method proved to be promising, as an alternative to more computationally heavy methods, such as RPS-tBLASTn, a lack of recall and precision deemed this method unsuitable for exhaustive research.

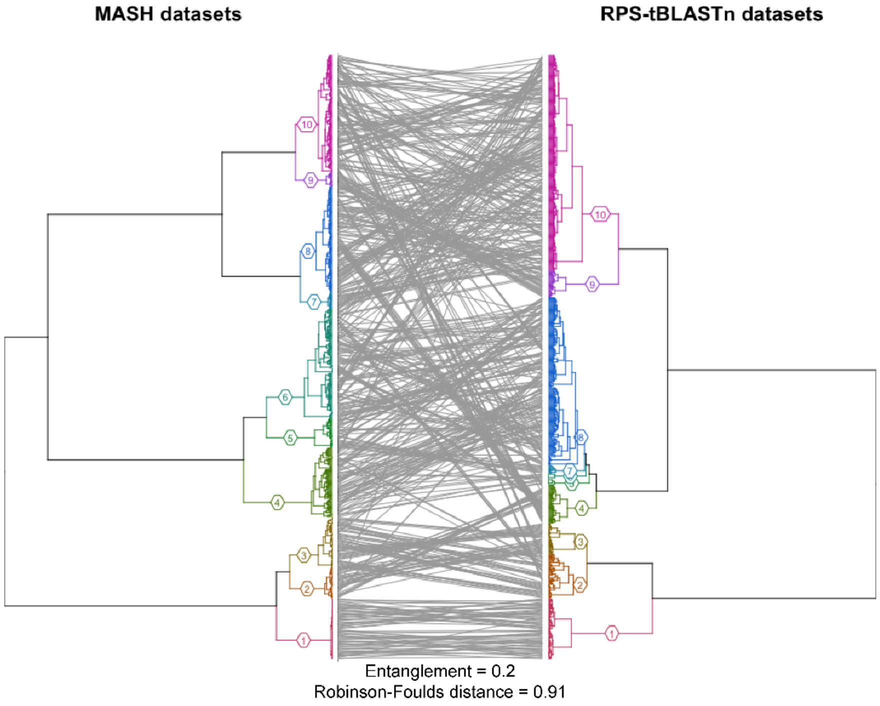


Figure 3. Tanglegram depicting hierarchical clustering performed on the Canberra distance matrices derived from the domain counts matrices of both Mash and RPS-tBLASTn pipelines. Both dendrograms are colored by their cluster id with $k = 10$. Base R function hclust was used to generate the clustering [18]. Correlation between both matrices was calculated with the Mantel test implemented in the ade4 R package [19]. The entanglement value and plot were generated with the Entanglement and Tanglegram functions implemented in the dendextend package [21]. Robinson–Foulds distance was calculated using the RF.dist function implemented in the Phangorm package [20].

The HMMER pipeline (hmmscan against Pfam-A v33.0) was applied to a simulated viral dataset of short reads. A total of 75 herpesvirus-specific domains were detected, with e-values $\leq 1 \times 10^{-3}$. On a node with 64 2.30GHz cores, a total of 136 peptide sequences per minute were scanned for domain detection. In order to improve comparability, the same procedure was applied to search domains in a real, non-simulated sequence dataset (Illumina paired-end data; SRA: ERR1137115). A total of 1000 randomly selected reads were extracted, translated into the six possible frames, and scanned for domains using the HMMER pipeline. For this dataset, 125 peptides were searched per minute, detecting 109 domains. Despite its ability to detect viral domains in short sequences, the HMMER pipeline performance did not scale well enough for datasets containing millions of reads over the course of three codeathon days. In order to deploy the HMMER pipeline for such large datasets, we propose the following: (i) collapsing identical or near-identical sequences to reduce redundancy; (ii) splitting translated frames yielding truncated peptides into distinct peptides, using stop codons as peptide boundaries; and, *(iii)* filtering amino acid sequences by length (≥ 50 amino acids) to decide whether to accept them as queries for domain detection. Applying these premises might still not be enough to yield reliable resources over the course of an event like a codeathon but might be sufficient to yield results under reasonable research time (e.g., over a week).

3.2. Virus–Host Pairing Prediction

To establish a baseline for known virus–host pairings, we federated several resources to act as references for experimentally confirmed and inferred interactions for the FIVE. In addition, we included queryable datasets designed to detect putative viral elements and linkable to a putative host. These putative pairings included phages and prophages mainly, which may help to understand prophage variability in well-categorized host systems. We federated existing databases providing information on hosts (including Bacteria, Archaea, and Eukarya) and the identity of confirmed viral pairings from PhagesDB and NCBI Virus Variation Resource database. The aforementioned resources were combined, expanded, and standardized to produce a comprehensive virus–host pairing index, containing 44,975 virus–host pairs (29,847 unique viruses and 7974 unique hosts) that can be queried from FIVE.

Identified CRISPR spacers from four datasets (CRISPRCasdb, RefSeq, 24,345 human microbiome MAGs and 24,706 GTDB species-representative sequences) were curated and compiled into a comprehensive CRISPR spacer database, with 1 million unique spacer sequences linked to a formalized host taxonomy. CRISPR spacers were compared against the 29,847 unique viruses with known hosts identified in the virus–host pairing index. In addition, the CRISPR spacers are compared against 2953 raw datasets selected from NCBI's SRA used in VHT-1 [4].

3.3. Viral Genome Diversity Indexing in Genome Graphs

To demonstrate the ability of FIVE to index a wide variety of data types, we made genome graphs with HIV-1 reference genomes. Genome graphs facilitate the analysis of the diversity of a set of closely related sequences by counting and connecting *k*-mers from multiple sources like full viral genomes or virus segments. Examples of our HIV-1 genome graphs are shown in Figure 4. The FASTA header for each sequence analyzed to create the graph was extended using brackets as key value pairs, i.e., ">Accession [key=value] [key=value]". Specifically, we indexed the viral diversity using the metadata of individual *k*-mer graph nodes. This metadata was later included into the individual graph nodes (not shown). The sequences were assembled into a graph using the Python3 package *NetworkX* and stored as a GraphML file that can be visualized and further analyzed using free open-source software such as Gephi ([53]) or Cytoscape ([54]) [55].

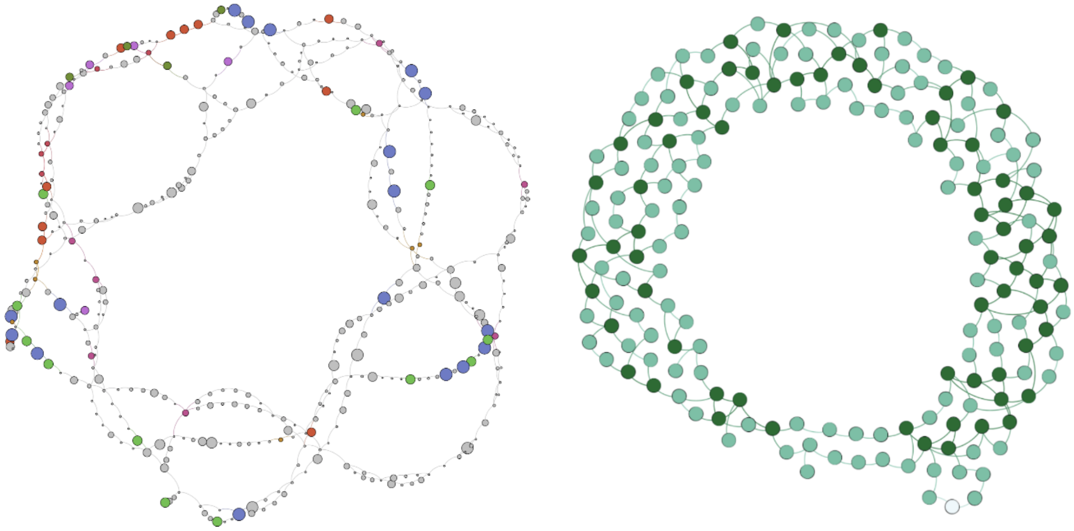


Figure 4. (Left) HIV-1 reference genome graphs generated with SWIft Genomes in a Graph (SWIGG) with annotated *k*-mers/nodes. Number of input sequences (*n*) = 167. Node color corresponds to taxonomic distribution of *k*-mer. Size of nodes is proportional to occurrence of taxonomic category. (Right) HIV-1 subtypes A–J (*n* = 39), *k*-mer size = 41, threshold ≥ 2. Note that both example graphs are circular, which may represent the fact that common nodes occur within long terminal repeats (LTRs). Most of the HIV references used in this work were modeled after the proviral sequence, which includes 5′ and 3′ LTRs.

3.4. Index Structure and Accessibility

Contig annotations and graph data derived from protein domain recognition, viral-specific HMM, virus–host pairing, and HIV-1 genome variability were produced in tabular format (as detailed at the end of each previous sections). Each table was loaded into the FIVE BigQuery GoogleCloud index ([56]) to be queryable. FIVE consists of seven interconnected tables (accession2species [3,174,289 entries], combined_known_interactions [46,979 entries], cdd_data [2,765,472 entries], spacer_db [18,521,874 entries], domains_viral_cds_tblastn [26,902,443 entries], and hiv_a_jrefs_k41_t2 [247 entries]), as seen in Figure 5, and can be freely accessed at [57] (link provided in Table 1).

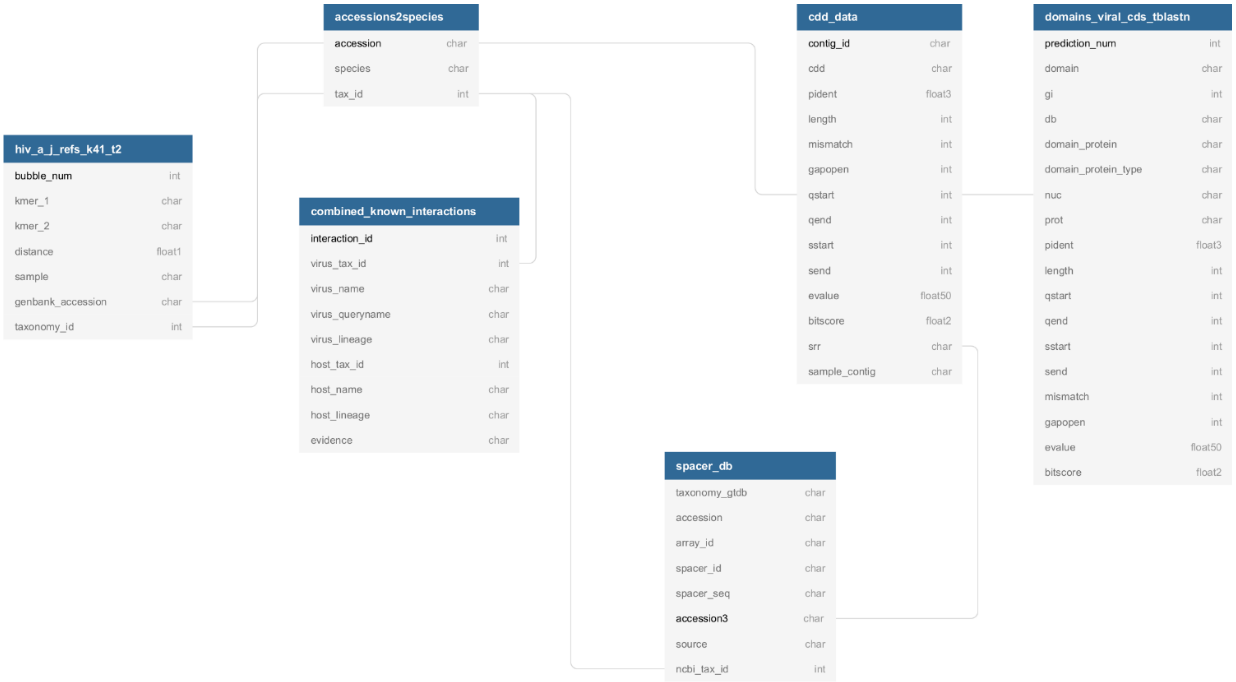


Figure 5. FIVE index schema. Each table (boxes) represents the output from the different annotation efforts towards FIVE. For each table, the title of the table is white in a blue rectangle (*accession2species*, *combined_known_interactions*, *cdd_data*, *spacer_db*, *domains_viral_cds_tblastn*, and *hiv_a_jrefs_k41_t2*), immediately followed by the field names or categories for that given table. Each line corresponds to a field, in which the first column gives the abbreviation name for the content of the field and the second column the format of the content (int for integers, char for strings of characters, float and decimals). Primary keys for each table are found in bold. It is possible to both access each one of the tables independently and to link primary keys from one table to fields from another table, generating a link (in grey).

A range of different functions are implemented in the viral-index API module that enable easy access to FIVE. The viral-index can search the federated indexes by SRA run ID, virus and host taxonomy ID, and CRISPR spacer sequences (as seen in Table 2).

Table 2. Summary and description of primary viral-index API query functions.

Function	Description
get_viruses_for_host_taxonomy	Retrieve host(s) for a given virus taxonomy ID
get_host_from_virus_taxonomy	Retrieve virus(es) that can infect a given host
get_potential_hosts_for_virus_domain	Get all potential host(s) given a domain that is found in viruses
get_virus_host_interactions_from_confidence_level	Get all virus–host interactions for specified confidence level
get_SRAs_where_CDD_is_found	Get Sequence Read Archive (SRA) accessions of studies wherein a viral protein domain is found
get_domains	Find all domains present in a virus

Virus–host pairs can be searched using the *get_host_for_virus_taxonomy* function that takes a NCBI virus taxonomy ID as input and returns all hosts that the given virus could infect. In order to perform the inverse search, *get_viruses_for_host_taxonomy* can be applied, and it may allow users to search for viruses that could infect a given host taxonomy ID; this search can be expanded to incorporate the protein domain-based information. The *get_potential_hosts_for_virus_domain* function integrates the data generated for the protein domains and the virus–host interactions. Thus, it allows searching potential virus hosts that viruses, with a specific domain, could infect by searching the federated data using a CDD domain ID. Other domain-based functions include (i) *get_SRAs_where_CDD_is_found* and (ii) *get_domains,* whereby users can retrieve specific SRA studies where (i) a virus-specific domain is present and (ii) the viruses that may contain a domain, respectively. The former function can be used to get a snapshot of virus domains in several SRA studies analyzed in VHT-1 [4]. Two additional functions are implemented to retrieve CRISPR signature-based spacer indexes. The *get_spacer_seqs* enables the users to fetch all spacer sequences present in the spacer datasets for a given taxonomy ID. The *get_metadata_from_spacer_seq* retrieves spacer ID, spacer sequence, GenBank accession, and taxonomy identification of organisms where the given spacer sequence is present. An example, showing how to use the *viral-index* to retrieve all viruses that infect pigs, is provided at [58].

It is important to note that data integration is becoming the norm; powerful analysis can be performed when it is possible to interlink data generated and enriched with multiple layers of known and novel information. The *viral-index* API enables researchers to interrogate increasingly sophisticated biological questions from FIVE through the multi-layer information available in this federated database indexes.

4. Conclusions and Future Directions

During this three-day continuation of the VHT codeathon series (VHT-2), a new integrated and federated viral index was elaborated. This Federated Index of Viral Experiments—FIVE—integrated new functional and taxonomy annotations, novel virus–host pairings, and for the first time, introduced virus genome diversity as genome graphs. Additionally, FIVE contains a federation of annotations and pairings from pre-existing sources. As per the publication of this manuscript, FIVE is the first implementation of a virus-specific federated index of such scope.

Several metagenomic annotation pipelines were developed and tested, building on top of the foundations laid out in previous editions. Three pipelines for annotation of viral contigs through protein domains were proposed: (i) RPS-tBLASTn, (ii) Mash, and (iii*)* HMMER. Results showed the differences in recall, accuracy, and speed between RPS-tBLASTn and Mash. RPS-tBLASTn may have been more computationally expensive than Mash, but it had better recall and overall accuracy. Additionally, as evidenced from VHT-1 [4], HMMER searches could not be fully scaled in a cloud environment, representing a bottleneck in protein domain classification using HMMs. Despite the ability of RPS-tBLASTn to be pseudo-parallelized, the main bottleneck for high-throughput cloud computing was scalability. Based on the current results, the RPS-tBLASTn pipeline was the best-performing implementation out of the three and the one we recommend for other large-scale cloud computing initiatives.

We made an additional effort to expand and federate not only the annotation tools for viral datasets, but also its taxonomical pairing with a given host. This original work expanded the number of known viral–host taxonomical pairings by 129% over VirHostNet 2.0 (release 1/2019) [59], by integrating a federated high-confidence dataset and a novel dataset based on de novo assignations. The high-confidence dataset is based on a federation of the NCBI Virus Variation Resource and PhagesDB databases. The novel dataset is built with predicted past pairings using CRISPR spacers. An expanded CRISPR dataset was created with 1M unique spacers to identify previously unknown relationships between complete viral genomes with taxonomy and the CRISPR spacer isolation source.

One of the last challenges during the codeathon was to start developing a pipeline and indexing strategy for virus genome diversity. It is known that genome graphs can be used to efficiently summarize known virus genome diversity, thus as a proof-of-concept we decided to build an HIV-1 genome diversity graph to index the variability into a federated index, such as FIVE. The two main challenges were (i) finding appropriate *k*-mer settings and (ii) adding multiple metadata values to virus genome diversity graphs. Proper attachment of metadata is crucial for indexing datasets and ensuring that data is findable, accessible, interoperable, and reusable (FAIR) [60]. Metadata can also be overlaid onto genome diversity graphs to improve functional interpretation. Metadata was added from the analyzed sequences but was inadequate when evaluating features within graphs. A limitation of available reference genomes included nonuniform feature annotation formatting. Follow-up work will be needed to analyze the influence of SWIGG parameters on subsets of HIV-1. Linking individual *k*-mer nodes to component sequence annotations will further enhance the possibility to mine the structural information represented in graphs and to connect it to biological function. In the current state, creating genome diversity graphs is a convoluted process involving iterative testing of multiple parameters and visual inspection of their resulting graphs. Visualizing multiple metadata layers simultaneously becomes challenging and, ultimately, the manual analysis of multiple and complex graphs becomes an unfeasible task. It will be important to develop automated assessments of virus diversity graphs to adjust construction parameters, evolve visualization methods for multiple metadata values, and create methods to automate graph analysis. Input from viral genomics is needed in order to standardize a genome diversity graph format.

Annotations and metadata from the different projects were integrated into the FIVE BigQuery index and later made queryable, making it the first implementation of a virus-specific federated index that is easily accessible and queryable. As per the development of a Python-based API, the FIVE can be *de facto* used by a larger part of the research community (possessing basic scripting abilities). Accessibility and ease of implementation are often the limiting factors for the broad use of public resources. A graphical user interface (GUI) is under discussion to further broaden accessibility. The final aim is to link FIVE to other widely used viral and host resources, such as those supported NCBI, centralizing the resource and improving its connectivity to other services. Efforts are underway to maintain FIVE through annual updates and continuous federation.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, File S1: Hoja1.

Author Contributions: All authors contributed equally to the conceptualization, methodology, investigation performed, and resources produced during this codeathon. The following authors contributed to software design: J.M.C., A.R.G., A.F.B., C.E.C., S.M., S.S., J.P.B., and K.J. The following authors contributed to validation: J.M.C., A.R.G., C.E.C., and S.M. The following authors contributed to formal analyses: J.M.C., A.R.G., A.F.B., and S.M. The following authors contributed to data curation: J.M.C., A.R.G., S.D.M., A.F.M., C.E.C., W.D., C.G., S.M., A.K.B., E.B., N.P.C., B.J.T., and J.L.W.. The following authors contributed to writing—original draft preparation: J.M.C., A.R.G., A.F.B., C.E.C., R.C., W.D., C.G., W.K.M., S.M., A.L.N., S.S., E.B., J.P.B., B.J.T., V.C.V., and B.B. The following authors contributed to writing—review and editing: J.M.C., A.R.G., A.F.B., C.E.C., R.C., W.D., C.G., S.M., W.K.M, A.L.N., S.S., E.B., J.P.B., N.P.C., B.J.T., B.V., V.C.V., and B.B. The following authors contributed to visualization: A.R.G., A.F.B., S.M., V.Z., and A.E.. Working group leads supervised their group members' activity during the codeathon. These leads and team members included D.M.K. for *domain search* with members M.S., W.K.M., W.D., A.F.B.; A.K.B. and V.Z. for *taxonomy-domain integration* with members B.V., N.S., N.C., K.J., R.C.; C.G. for *viral-host interactions* with members R.A.E., C.P.R., M.M., S.D.M., B.J.T., J.L.W.; E.B. and M.J.T *virus graphs* with members A.R.G., A.L.N., H.S.M., J.P.B., S.S., V.C.V; and C.E.C. and A.E. for *virus indexing* with members J.M.C., S.M. The following authors contributed to project administration: J.M.C., A.R.G., A.F.B., J.P.B., B.J.T., and B.B. The following authors contributed to funding acquisition: B.B., J.R.B., and R.C. All authors have read and agreed to the published version of the manuscript.

Funding: This codeathon was funded by the Office of Data Science Strategy and supported by the National Library of Medicine, National Center for Biotechnology Information, as well as the STRIDES Initiative at the National Institutes of Health. J.M.C. was supported by a doctoral grant from HONOURs Marie-Sklodowska-Curie training network (721367). A.R.G. is funded in part by institutional support from Baylor College of Medicine; private funding by East Coast Oils, Inc., Jacksonville, Florida, and A.R.G.'s own private funding, including Student Genomics (manuscripts in prep). A.R.G. also received the PFLAG of Jacksonville scholarship for multiple years. S.M. is funded by the UK Medical Research Council Precision Medicine PhD studentship grant number MR/S502479/1. S.S. is funded by USDA NIFA grant number 2015-70016-23028. N.P.C. was supported by NIH Training Grant T15 LM007059. R.A.E. is funded by NIH NIDDK Grant 1RC2DK116713 and DTRA Grant 13081-32220. M.S. is funded by Laboratory Directed Laboratory Directed Research and Development - Exploratory Research (LDRD-ER #20200252ER). B.J.T was supported by the National Science Foundation Science and Technology Center, the Center for Dark Energy Biosphere Investigations (C-DEBI) [OCE- 0939654]. B.V. was supported by the "Fonds Wetenschappelijk Onderzoek"/Research foundation—Flanders, Brussels, Belgium (1S28617N). The work of R.C., R.B., B.B., C.E.C., and E.H. was funded by the Intramural Research Program of the National Library of Medicine. S.G. holds a scholarship from the Graduate School of Medical Sciences, University of Groningen.

Acknowledgments: The authors would like to acknowledge the role of Carl Leubsdorf, Mihai Pop, Rob Patro, and Tom Ventsias for their scientific and logistical assistance that made this second codeathon possible. Additionally, we would like to especially acknowledge the contribution of Sanzhima Garmaeva (S.G.) for her on-line collaboration during the codeathon and the revision of the manuscript.

Conflicts of Interest: J.M.C and A.R.G have received travel awards and bursaries from Oxford Nanopore Technologies, Oxford, UK. This material should not be interpreted as representing the viewpoint of the U.S. Department of Health and Human Services, the National Institutes of Health, Food and Drug Administration, National Library of Medicine, National Center for Biotechnology Information, Center for Information Technology, or Office of Data Science Strategy. No other competing interests to disclose.

References

1. Mardis, E.R. A decade's perspective on DNA sequencing technology. *Nature* **2011**, *470*, 198–203.
2. Kodama, Y.; Shumway, M.; Leinonen, R. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.* **2012**, *40*, D54–D56.
3. SRA Database Growth. Available online: https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/ (accessed on 3 December 2020).
4. Connor, R.; Brister, R.; Buchmann, J.; Deboutte, W.; Edwards, R.; Martí-Carreras, J.; Tisza, M.; Zalunin, V.; Andrade-Martínez, J.; Cantu, A.; et al. NCBI's Virus Discovery Hackathon: Engaging Research Communities to Identify Cloud Infrastructure Requirements. *Genes (Basel).* **2019**, *10*, 714.
5. STRIDES Initiative. Available online: https://datascience.nih.gov/strides (accessed on 3 December 2020).
6. NIH Strategic Plan for Data Science. Available online: https://datascience.nih.gov/strategicplan (accessed on 3 December 2020).
7. Sayers, E.W.; Agarwala, R.; Bolton, E.E.; Brister, J.R.; Canese, K.; Clark, K.; Connor, R.; Fiorini, N.; Funk, K.; Hefferon, T.; et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2019**, *47*, D23–D28.
8. NCBI Codeathons. Available online: https://ncbi-codeathons.github.io/ (accessed on 3 December 2020).
9. Busby, B.; Saha, S.; Martí-Carreras, J. Virus Discovery Project 2019. Available online: https://osf.io/g9w8r/ (accseesed on 4 December 2019).
10. NCBI-Codeathons/Virus_Graphs. Available online: https://github.com/NCBI-Codeathons/Virus_Graphs/tree/master/data (accessed on 3 December 2020).
11. Paten, B.; Novak, A.M.; Eizenga, J.M.; Garrison, E. Genome graphs and the evolution of genome inference. *Genome Res.* **2017**, *27*, 665–676.
12. Pickett, B.E.; Sadat, E.L.; Zhang, Y.; Noronha, J.M.; Squires, R.B.; Hunt, V.; Liu, M.; Kumar, S.; Zaremba, S.; Gu, Z.; et al. ViPR: An Open Bioinformatics Database and Analysis Resource for Virology Research. *Nucleic Acids Res.* **2012** 40, D593–D598.
13. NCBI-Hackathons/VirusDiscoveryProject. Available online: https://github.com/NCBI-Hackathons/VirusDiscoveryProject (accessed on 3 December 2020).
14. NCBI-Codeathons/Domain_HMM_Boundaries. Available online: https://github.com/NCBI-Codeathons/Domain_HMM_Boundaries/tree/master/viral-cdd-models (accessed on 3 December 2020).
15. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST plus: architecture and applications. *BMC Bioinformatics* **2009**, *10*, 1.
16. Ondov, B.D.; Treangen, T.J.; Melsted, P.; Mallonee, A.B.; Bergman, N.H.; Koren, S.; Phillippy, A.M. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **2016**, *17*, 132.
17. NCBI-Codeathons/Domain_HMM_Boundaries. Available online: https://github.com/NCBI-Codeathons/Domain_HMM_Boundaries/blob/master/dataset_accessions/Mash_accessions.txt (accessed on 3 December 2020).
18. R Core Team R: A Language and Environment for Statistical Computing 2019. Available Online: https://www.r-project.org/.
19. Bougeard, S.; Dray, S. Supervised Multiblock Analysis in R with the ade4 Package. *J. Stat. Softw.* **2018**, *86*.
20. Schliep, K.P. phangorn: phylogenetic analysis in R. *Bioinformatics* **2011**, *27*, 592–593.
21. Galili, T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* **2015**, *31*, 3718–3720.
22. Eddy, S.R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **2011**, *7*, 10.

23. O'Leary, N.A.; Wright, M.W.; Brister, J.R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **2016**, *44*, D733–D745.
24. Domain_HMM_Boundaries/viral-cdd-models/virus_models.txt. Available online: https://github.com/NCBI-Codeathons/Domain_HMM_Boundaries/blob/master/viral-cdd-models/virus_models.txt (accessed on 3 December 2020).
25. Conserved Domains and Protein Classification Help. Available online: https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd_help.shtml (accessed on 3 December 2020).
26. Domain_HMM_Boundaries/scripts/tReads.py. Available Online: https://github.com/NCBI-Codeathons/Domain_HMM_Boundaries/tree/master/scripts/tReads.py (accessed on 3 December 2020).
27. Hatcher, E.L.; Zhdanov, S.A.; Bao, Y.; Blinkova, O.; Nawrocki, E.P.; Ostapchuck, Y.; Schäffer, A.A.; Brister, J.R. Virus Variation Resource – improved response to emergent viral outbreaks. *Nucleic Acids Res.* **2017**, *45*, D482–D490.
28. Russell, D.A.; Hatfull, G.F. PhagesDB: the actinobacteriophage database. *Bioinformatics* **2017**, *33*, 784–786.
29. Sayers, E.W.; Barrett, T.; Benson, D.A.; Bryant, S.H.; Canese, K.; Chetvernin, V.; Church, D.M.; DiCuccio, M.; Edgar, R.; Federhen, S.; et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2009**, *37*, D5–D15.
30. NCBI-Codeathons/Host_Phage_Interactions. Available Online: https://github.com/NCBI-Codeathons/Host_Phage_Interactions/tree/development/src (accessed on 3 December 2020).
31. CRISPR-Cas++. Available Online: https://crisprcas.i2bc.paris-saclay.fr/Home/Download (accessed on 3 December 2020).
32. Couvin, D.; Bernheim, A.; Toffano-Nioche, C.; Touchon, M.; Michalik, J.; Néron, B.; Rocha, E.P.C.; Vergnaud, G.; Gautheret, D.; Pourcel, C. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res.* **2018**, *46*, W246–W251.
33. Weissman, J.L.; Fagan, W.F.; Johnson, P.L.F. Selective Maintenance of Multiple CRISPR Arrays Across Prokaryotes. *Cris. J.* **2018**, *1*, 405–413.
34. Biswas, A.; Staals, R.H.J.; Morales, S.E.; Fineran, P.C.; Brown, C.M. CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics* **2016**, *17*, 356.
35. Nayfach, S.; Shi, Z.J.; Seshadri, R.; Pollard, K.S.; Kyrpides, N.C. New insights from uncultivated genomes of the global human gut microbiome. *Nature* **2019**, *568*, 505–510.
36. ctSkennerton/minced. Available Online: https://github.com/ctSkennerton/minced (accessed on 3 December 2020).
37. Bland, C.; Ramsey, T.L.; Sabree, F.; Lowe, M.; Brown, K.; Kyrpides, N.C.; Hugenholtz, P. CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **2007**, *8*, 209.
38. Parks D.H..; Chuvochina, M.; Chaumeil, PA.; Rinke, C..; Mussig, AJ.; P, Hugenholtz, P. Selection of representative genomes for 24,706 bacterial and archaeal species clusters provide a complete genome-based taxonomy. *bioRxiv* **2019, 771964**.
39. NCBI-Codeathons/Virus_Graphs. Available Online: https://github.com/NCBI-Codeathons/Virus_Graphs (accessed on 3 December 2020).
40. Virus_Graphs/Reference_Seq.fasta. Available Online: https://github.com/NCBI-Codeathons/Virus_Graphs/blob/master/Reference_Seq.fasta (accessed on 3 December 2020).
41. HIV Databases. Available Online: http://www.hiv.lanl.gov/ (accessed on 3 December 2020).
42. HIV Sequence Database. Available Online: https://www.hiv.lanl.gov/content/sequence/HIV/mainpage.html (accessed on 3 December 2020).
43. Hagberg, A.A.; Schult, D.A.; Swart, P.J. Exploring Network Structure, Dynamics, and Function using NetworkX. In Proceedings of the Proceedings of the 7th Python in Science Conference; Varoquaux, G., Vaught, T., Millman, J., Eds.; Pasadena, CA USA, 2008; pp. 11–15.
44. NCBI-Codeathons/The_Virus_Index. Available Online: https://github.com/NCBI-Codeathons/The_Virus_Index/tree/master/python (accessed on 3 December 2020).
45. The_Virus_Index/schema. Available Online: https://github.com/NCBI-Codeathons/The_Virus_Index/tree/master/schema (accessed on 3 December 2020).
46. viral-index 0.0.3. Available Online: https://test.pypi.org/project/viral-index/ (accessed on 3 December 2020).

47. NCBI-Codeathons/The_Virus_Index. Available Online: https://github.com/NCBI-Codeathons/The_Virus_Index (accessed on 3 December 2020).
48. Viral-index API. Available Online: https://github.com/NCBI-Codeathons/The_Virus_Index#api (accessed on 3 December 2020).
49. Domain_HMM_Boundaries/viral-cdd-models/virus_models.txt. Available Online: https://github.com/NCBI-Codeathons/Domain_HMM_Boundaries/blob/master/viral-cdd-models/virus_models.txt (accessed on 3 December 2020).
50. Broder, A.Z. On the Resemblance and Containment of Documents. In Proceedings of the In Compression and Complexity of Sequences (SEQUENCES'97; IEEE Computer Society, NW Washington, DC, USA 1997; pp. 21–29.
51. Domain_HMM_Boundaries/dataset_accessions/Mash_accessions.txt. Available Online: https://github.com/NCBI-Codeathons/Domain_HMM_Boundaries/blob/master/dataset_accessions/Mash_accessions.txt (accessed on 3 December 2020).
52. Domain_HMM_Boundaries/dataset_accessions/RPStbln_accessions.txt. Available Online: https://github.com/NCBI-Codeathons/Domain_HMM_Boundaries/blob/master/dataset_accessions/RPStbln_accessions.txt (accessed on 3 December 2020).
53. The Open Graph Viz Platform. Available Online: https://gephi.org (accessed on 3 December 2020).
54. Cytoscape. Available Online: https://cytoscape.org/ (accessed on 3 December 2020).
55. Shannon, P. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13*, 2498–2504.
56. The_Virus_Index/schema/. Available Online: https://github.com/NCBI-Codeathons/The_Virus_Index/tree/master/schema (accessed on 3 December 2020).
57. Google cloud platform. Available Online: https://console.cloud.google.com/bigquery?p=virus-hunting-2-codeathon&d=viasq&page=dataset (accessed on 3 December 2020).
58. Sample code. Available Online: https://github.com/NCBI-Codeathons/The_Virus_Index#sample-code (accessed on 3 December 2020).
59. Guirimand, T.; Delmotte, S.; Navratil, V. VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. *Nucleic Acids Res.* **2015**, *43*, D583–D587.
60. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, Ij.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018.

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Bibliography

Abbas, Arwa A., Louis J. Taylor, Marisol I. Dothard, Jacob S. Leiby, Ayannah S. Fitzgerald, et al. (2019). " *Redondoviridae*, a Family of Small, Circular DNA Viruses of the Human Oro-Respiratory Tract Associated with Periodontitis and Critical Illness". In: *Cell Host and Microbe* 25.5, 719–729.e4.

Aevarsson, Arnthór, Anna-Karina Kaczorowska, Björn Thor Adalsteinsson, Josefin Ahlqvist, Salam Al-Karadaghi, et al. (2021). "Going to extremes – a metagenomic journey into the dark matter of life". In: *FEMS Microbiology Letters*.

Aguiar-Pulido, Vanessa, Wenrui Huang, Victoria Suarez-Ulloa, Trevor Cickovski, Kalai Mathee, et al. (2016). "Metagenomics, Metatranscriptomics, and Metabolomics Approaches for Microbiome Analysis." In: *Evolutionary bioinformatics online* 12.Suppl 1, pp. 5–16.

Ahlgren, Nathan A., Jie Ren, Yang Young Lu, Jed A. Fuhrman, and Fengzhu Sun (2017). "Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences". In: *Nucleic Acids Research* 45.1, pp. 39–53.

Aiewsakun, Pakorn and Peter Simmonds (2018). "The genomic underpinnings of eukaryotic virus taxonomy: Creating a sequence-based framework for family-level virus classification". In: *Microbiome* 6.1, pp. 1–24.

Alcon-Giner, Cristina, Shabhonam Caim, Suparna Mitra, Jennifer Ketskemety, Udo Wegmann, et al. (2017). "Optimisation of 16S rRNA gut microbiota profiling of extremely low birth weight infants". In: *BMC Genomics* 18.1, p. 841.

Almeida, Alexandre, Alex L. Mitchell, Miguel Boland, Samuel C. Forster, Gregory B. Gloor, et al. (2019). "A new genomic blueprint of the human gut microbiota". In: *Nature* 568.7753, p. 1.

Almeida, Alexandre, Stephen Nayfach, Miguel Boland, Francesco Strozzi, Martin Beracochea, et al. (2020). "A unified catalog of 204,938 reference genomes from the human gut microbiome". In: *Nature Biotechnology*, pp. 1–10.

Alneberg, Johannes, Brynjar Smári Bjarnason, Ino De Bruijn, Melanie Schirmer, Joshua Quick, et al. (2014). "Binning metagenomic contigs by coverage and composition". In: *Nature Methods* 11.11, pp. 1144–1146.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman (1990). “Basic local alignment search tool”. In: *Journal of Molecular Biology* 215.3, pp. 403–410.

Amgarten, Deyvid, Lucas P.P. Braga, Aline M. da Silva, and João C. Setubal (2018). “MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins”. In: *Frontiers in Genetics* 9.AUG, p. 304.

Anthony, Simon J, Jonathan H Epstein, Kris A Murray, Isamara Navarrete-Macias, Carlos M Zambrana-Torrelio, et al. (2013). “A strategy to estimate unknown viral diversity in mammals.” In: *mBio* 4.5, e00598–13.

Arze, Cesar A., Simeon Springer, Gytis Dudas, Sneha Patel, Agamoni Bhattacharyya, et al. (2021). “Global genome analysis reveals a vast and dynamic anellovirus landscape within the human virome”. In: *Cell Host & Microbe*.

Babaian, Artem and Robert C. Edgar (2021). “Ribovirus classification by a polymerase barcode sequence”. In: *bioRxiv*, p. 2021.03.02.433648.

Babayan, Simon A., Richard J. Orton, and Daniel G. Streicker (2018). “Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes”. In: *Science* 362.6414, pp. 577–580.

Baek, Minkyung, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, et al. (2021). “Accurate prediction of protein structures and interactions using a three-track neural network”. In: *Science* 373.6557, pp. 871–876.

Baltimore, D (1971). “Expression of animal virus genomes”. In: *Bacteriological Reviews* 35.3, pp. 235–241.

Barrientos-Somarribas, Mauricio, David N. Messina, Christian Pou, Fredrik Lysholm, Annelie Bjerkner, et al. (2018). “Discovering viral genomes in human metagenomic data by predicting unknown protein families”. In: *Scientific Reports* 8.1, p. 28.

Beijerinck, M.W. (1898). “Over een contagium vivum fluidum als oorzaak van de vlekzickte der tabaksbladen.” In: *Versle. Gewone Vergad. Wis-Natuurkd. Afd. K. Akad. Wet. Amsterdam* 7, pp. 229–235.

Benler, Sean, Natalya Yutin, Mikhail Raykov, Sergey Shmakov, Ayal B. Gussow, et al. (2021). “Thousands of previously unknown phages discovered in whole-community human gut metagenomes”. In: *Microbiome* 9.1, pp. 1–17.

Bergner, Laura M., Richard J. Orton, Julio A. Benavides, Daniel J. Becker, Carlos Tello, et al. (2020). “Demographic and environmental drivers of metagenomic viral diversity in vampire bats”. In: *Molecular Ecology* 29.1, pp. 26–39.

Bergner, Laura M., Richard J. Orton, Alice Broos, Carlos Tello, Daniel J. Becker, et al. (2021). “Diversification of mammalian deltaviruses by host shifting”. In: *Proceedings of the National Academy of Sciences of the United States of America* 118.3, p. 2019907118.

Bernard, Guillaume, Jananan S Pathmanathan, Romain Lannes, Philippe Lopez, and Eric Bapteste (2018). “Microbial Dark Matter Investigations: How Microbial Studies Transform Biological

Knowledge and Empirically Sketch a Logic of Scientific Discovery". In: *Genome Biology and Evolution* 10.3, pp. 707–715.

Bin Jang, Ho, Benjamin Bolduc, Olivier Zablocki, Jens H. Kuhn, Simon Roux, et al. (2019). "Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks". In: *Nature Biotechnology 2019 37:6* 37.6, pp. 632–639.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data". In: *Bioinformatics* 30.15, p. 2114.

Bordenstein, Seth R. and Kevin R. Theis (2015). "Host Biology in Light of the Microbiome: Ten Principles of Holobionts and Hologenomes". In: *PLOS Biology* 13.8. Ed. by Matthew K. Waldor, e1002226.

Bos, Lute (2000). "100 years of virology: from vitalism via molecular biology to genetic engineering". In: *Trends in Microbiology* 8.2, pp. 82–87.

Bouvier, Nicole M. and Peter Palese (2008). "THE BIOLOGY OF INFLUENZA VIRUSES". In: *Vaccine* 26.Suppl 4, p. D49.

Bowers, Robert M, Nikos C Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, et al. (2017). "Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea". In: *Nature biotechnology* 35.8, p. 725.

Bradley, Phelim, Henk C. den Bakker, Eduardo P.C. C. Rocha, Gil McVean, and Zamin Iqbal (2019). "Ultrafast search of all deposited bacterial and viral genomic data". In: *Nature Biotechnology* 37.2, pp. 152–159.

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter (2016). "Near-optimal probabilistic RNA-seq quantification". In: *Nature Biotechnology 2016 34:5* 34.5, pp. 525–527.

Breiman, Leo (2001). "Random forests". In: *Machine Learning* 45.1, pp. 5–32.

Breitbart, Mya, Eric Delwart, Karyna Rosario, Joaquim Segalés, and Arvind Varsani (2017). "ICTV virus taxonomy profile: *Circoviridae*". In: *Journal of General Virology* 98.8, pp. 1997–1998.

Breitbart, Mya, Ian Hewson, Ben Felts, Joseph M. Mahaffy, James Nulton, et al. (2003). "Metagenomic Analyses of an Uncultured Viral Community from Human Feces". In: *Journal of Bacteriology* 185.20, p. 6220.

Breitbart, Mya and Forest Rohwer (2005). "Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing". In: *BioTechniques* 39.5, pp. 729–736.

Breitbart, Mya, Peter Salamon, Bjarne Andresen, Joseph M Mahaffy, Anca M Segall, et al. (2002). "Genomic analysis of uncultured marine viral communities." In: *Proceedings of the National Academy of Sciences of the United States of America* 99.22, pp. 14250–5.

Breitwieser, Florian P., Jennifer Lu, and Steven L. Salzberg (2018). “A review of methods and databases for metagenomic classification and assembly”. In: *Briefings in Bioinformatics* 20.4, pp. 1125–1139.

Briese, Thomas, Janusz T. Paweska, Laura K. McMullan, Stephen K. Hutchison, Craig Street, et al. (2009). “Genetic Detection and Characterization of Lujo Virus, a New Hemorrhagic Fever–Associated Arenavirus from Southern Africa”. In: *PLoS Pathogens* 5.5.

Brister, J. Rodney, Danso Ako-Adjei, Yiming Bao, and Olga Blinkova (2015). “NCBI viral Genomes resource”. In: *Nucleic Acids Research* 43.D1, pp. D571–D577.

Brooks, Lauren, Mo Kaze, and Mark Sistrom (2019). “A Curated, Comprehensive Database of Plasmid Sequences”. In: *Microbiology Resource Announcements* 8.1.

Brown, Christopher T., Laura A. Hug, Brian C. Thomas, Itai Sharon, Cindy J. Castelle, et al. (2015). “Unusual biology across a group comprising more than 15% of domain Bacteria”. In: *Nature* 523.7559, pp. 208–211.

Brum, Jennifer R., J. Cesar Ignacio-Espinoza, Eun-Hae Kim, Gareth Trubl, Robert M. Jones, et al. (2016). “Illuminating structural proteins in viral “dark matter” with metaproteomics”. In: *Proceedings of the National Academy of Sciences* 113.9, pp. 2436–2441.

Brum, Jennifer R, J Cesar Ignacio-Espinoza, Simon Roux, Guilhem Doulcier, Silvia G Acinas, et al. (2015). “Ocean plankton. Patterns and ecological drivers of ocean viral communities.” In: *Science (New York, N.Y.)* 348.6237, p. 1261498.

Buchfink, Benjamin, Klaus Reuter, and Hajk Georg Drost (2021). “Sensitive protein alignments at tree-of-life scale using DIAMOND”. In: *Nature Methods* 18.4, pp. 366–368.

Buchfink, Benjamin, Chao Xie, and Daniel H Huson (2014). “Fast and sensitive protein alignment using DIAMOND”. In: *Nature Methods* 12.1, pp. 59–60.

Bushnell B. (2015a). *Introducing BBDuk: Adapter/Quality Trimming and Filtering*. URL: http://seqanswers.com/forums/showthread.php?t=42776 (visited on 06/19/2020).

— (2015b). *Introducing BBNorm, a read normalization and error-correction tool*. URL: http://seqanswers.com/forums/showthread.php?t=49763 (visited on 06/19/2020).

— (2019). *BBMap download | SourceForge.net*. URL: https://sourceforge.net/projects/bbmap/ (visited on 06/14/2019).

Bzhalava, Zurab, Emilie Hultin, and Joakim Dillner (2018). “Extension of the viral ecology in humans using viral profile hidden Markov models”. In: *PLOS ONE* 13.1. Ed. by Ulrich Melcher, e0190938.

Camarillo-Guerrero, Luis F., Alexandre Almeida, Guillermo Rangel-Pineros, Robert D. Finn, and Trevor D. Lawley (2021). “Massive expansion of human gut bacteriophage diversity”. In: *Cell* 184.4, 1098–1109.e9.

Camarillo-Guerrero, Luis F, Alexandre Almeida, Guillermo Rangel-Pineros, and Trevor D Lawley (2020). “Massive expansion of human gut bacteriophage diversity 1 2”. In: *bioRxiv*, p. 2020.09.03.280214.

Cann, Alan J. (2021). “Viral Replication Cycle”. In: *Encyclopedia of Virology (Fourth Edition)*. Ed. by Dennis H. Bamford and Mark Zuckerman. Fourth Edition. Oxford: Academic Press, pp. 382–387.

Carr, Victoria R., Andrey Shkoporov, Colin Hill, Peter Mullany, and David L. Moyes (2021). “Probing the Mobilome: Discoveries in the Dynamic Microbiome”. In: *Trends in Microbiology* 29.2, pp. 158–170.

Casasnovas, José M. and Thilo Stehle (2021). “Viral Receptors”. In: *Encyclopedia of Virology (Fourth Edition)*. Ed. by Dennis H. Bamford and Mark Zuckerman. Fourth Edition. Oxford: Academic Press, pp. 388–401.

Castelle, Cindy J. and Jillian F. Banfield (2018). “Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life”. In: *Cell* 172.6, pp. 1181–1197.

Castelle, Cindy J., Kelly C. Wrighton, Brian C. Thomas, Laura A. Hug, Christopher T. Brown, et al. (2015). “Genomic Expansion of Domain Archaea Highlights Roles for Organisms from New Phyla in Anaerobic Carbon Cycling”. In: *Current Biology* 25.6, pp. 690–701.

Castillo, Diego J., Riaan F. Rifkin, Don A. Cowan, and Marnie Potgieter (2019). “The Healthy Human Blood Microbiome: Fact or Fiction?” In: *Frontiers in Cellular and Infection Microbiology* 9.MAY, p. 148.

Cebriá-Mendoza, María, Cristina Arbona, Luís Larrea, Wladimiro Díaz, Vicente Arnau, et al. (2021). “Deep viral blood metagenomics reveals extensive anellovirus diversity in healthy humans”. In: *Scientific Reports 2021 11:1* 11.1, pp. 1–11.

Chang, Wei Shan, John H.O. Pettersson, Callum Le Lay, Mang Shi, Nathan Lo, et al. (2019). “Novel hepatitis D-like agents in vertebrates and invertebrates”. In: *Virus Evolution* 5.2.

Chen, Lin Xing, Karthik Anantharaman, Alon Shaiber, A. Murat Eren, and Jillian F. Banfield (2020). “Accurate and complete genomes from metagenomes”. In: *Genome Research* 30.3, pp. 315–333.

Chen, Xing, Yu An Huang, Zhu Hong You, Gui Ying Yan, and Xue Song Wang (2017). “A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases”. In: *Bioinformatics* 33.5, pp. 733–739.

Chevreux, Bastien, Thomas Pfisterer, Bernd Drescher, Albert J. Driesel, Werner E.G. Müller, et al. (2004). “Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs”. In: *Genome Research* 14.6, p. 1147.

Chicco, Davide (2017). “Ten quick tips for machine learning in computational biology”. In: 10.1, p. 35.

Ching, Travers, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, Brian T. Do, et al. (2018). “Opportunities And Obstacles For Deep Learning In Biology And Medicine”. In: *bioRxiv*, p. 142760.

Choi, Kyung H. (2012). “Viral Polymerases”. In: *Viral Molecular Machines*. Ed. by Michael G. Rossmann and Venigalla B. Rao. Boston, MA: Springer US, pp. 267–304.

Choudhary, Saket (2019). “Pysradb: A Python package to query next-generation sequencing metadata and data from NCBI sequence read archive”. In: *F1000Research* 8, p. 532.

Choutko, Vassili, Vladimir Lazarevic, Nadia Gaïa, Myriam Girard, Gesuele Renzi, et al. (2019). “Rare Case of Community-Acquired Endocarditis Caused by Neisseria meningitidis Assessed by Clinical Metagenomics”. In: *Frontiers in Cardiovascular Medicine* 6, p. 112.

Clooney, Adam G., Thomas D.S. Sutton, Andrey N. Shkoporov, Ross K. Holohan, Karen M. Daly, et al. (2019). “Whole-Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease”. In: *Cell Host & Microbe* 26.6, 764–778.e5.

Cobbin, Joanna CA, Justine Charon, Erin Harvey, Edward C. Holmes, and Jackie E. Mahar (2021). “Current challenges to virus discovery by meta-transcriptomics”. In: *Current Opinion in Virology* 51, pp. 48–55.

Cochrane, Guy, Ilene Karsch-Mizrachi, Toshihisa Takagi, and International Nucleotide Sequence Database Collaboration (2015). “The International Nucleotide Sequence Database Collaboration”. In: *Nucleic Acids Research* 44.D1, pp. D48–D50.

Coclet, Clément and Simon Roux (2021). “Global overview and major challenges of host prediction methods for uncultivated phages”. In: *Current Opinion in Virology* 49, pp. 117–126.

Coffey, Lark L., Brady L. Page, Alexander L. Greninger, Belinda L. Herring, Richard C. Russell, et al. (2014). “Enhanced arbovirus surveillance with deep sequencing: Identification of novel rhabdoviruses and bunyaviruses in Australian mosquitoes”. In: *Virology* 448, pp. 146–158.

Connor, Ryan, Rodney Brister, Jan Buchmann, Ward Deboutte, Rob Edwards, et al. (2019). “NCBI’s Virus Discovery Hackathon: Engaging Research Communities to Identify Cloud Infrastructure Requirements”. In: *Genes* 10.9, p. 714.

Consortium, The Human Microbiome Project, Curtis Huttenhower, Dirk Gevers, Rob Knight, Sahar Abubucker, et al. (2012). “Structure, function and diversity of the healthy human microbiome”. In: *Nature* 486.7402, pp. 207–214.

Cooley, Nicholas P. and Erik S. Wright (2021). “Accurate annotation of protein coding sequences with IDTAXA”. In: *NAR Genomics and Bioinformatics* 3.3.

Cortes, Corinna and Vladimir Vapnik (1995). “Support-vector networks”. In: *Machine Learning* 20.3, pp. 273–297.

Cossart, Pascale (2018). *The New Microbiology*. ASM Press.

Crawford, Emily, Jack Kamm, Steve Miller, Lucy M. Li, Saharai Caldera, et al. (2020). “Investigating Transfusion-related Sepsis Using Culture-Independent Metagenomic Sequencing”. In: *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 71.5, p. 1179.

Crick, Francis (1970). “Central Dogma of Molecular Biology”. In: *Nature 1970 227:5258* 227.5258, pp. 561–563.

Crits-Christoph, Alexander, Rose S. Kantor, Matthew R. Olm, Oscar N. Whitney, Basem Al-Shayeb, et al. (2021). “Genome sequencing of sewage detects regionally prevalent SARS-CoV-2 variants”. In: *mBio* 12.1, pp. 1–9.

Crusoe, Michael R, Hussien F Alameldin, Sherine Awad, Elmar Boucher, Adam Caldwell, et al. (2015). “The khmer software package: enabling efficient nucleotide sequence analysis”. In: *F1000Research* 4.

Culley, Alexander I., Andrew S. Lang, and Curtis A. Suttle (2003). “High diversity of unknown picorna-like viruses in the sea”. In: *Nature* 424.6952, pp. 1054–1057.

Daims, Holger, Elena V. Lebedeva, Petra Pjevac, Ping Han, Craig Herbold, et al. (2015). “Complete nitrification by Nitrospira bacteria”. In: *Nature* 528.7583, pp. 504–509.

Dance, Amber (2021). “Beyond coronavirus: the virus discoveries transforming biology”. In: *Nature* 595.7865, pp. 22–25.

Darwin, Charles (1859). *On the Origin of Species by Means of Natural Selection, or Preservation of Favoured Races in the Struggle for Life*. London: John Murray.

DasSarma, S., J. A. Coker, and P. DasSarma (2009). “Archaea (overview)”. In: *Encyclopedia of Microbiology*, pp. 1–23.

Davison, Andrew and Stuart Siddell (2020). *What’s the point of virus taxonomy? - International Science Council*. URL: https://council.science/current/blog/whats-the-point-of-virus-taxonomy/ (visited on 05/28/2022).

Deng, Xianding, Asmeeta Achari, Scot Federman, Guixia Yu, Sneha Somasekar, et al. (2020). “Metagenomic sequencing with spiked primer enrichment for viral diagnostics and genomic surveillance”. In: *Nature Microbiology 2020 5:3* 5.3, pp. 443–454.

Deng, Zhi-Luo, Cornelia Gottschick, Sabin Bhuju, Clarissa Masur, Christoph Abels, et al. (2018). “Metatranscriptome Analysis of the Vaginal Microbiota Reveals Potential Mechanisms for Protection against Metronidazole in Bacterial Vaginosis”. In: *mSphere* 3.3.

Dewhirst, Floyd E., Tuste Chen, Jacques Izard, Bruce J. Paster, Anne C.R. Tanner, et al. (2010). “The human oral microbiome”. In: *Journal of Bacteriology* 192.19, pp. 5002–5017.

Di Paola, Nicholas, Nolwenn M. Dheilly, Sandra Junglen, Sofia Paraskevopoulou, Thomas S. Postler, et al. (2022). “ Jingchuvirales : a New Taxonomical Framework for a Rapidly Expanding Order of Unusual Monjiviricete Viruses Broadly Distributed among Arthropod Subphyla ”. In: *Applied and Environmental Microbiology* 88.6.

Dimitrov, Dimiter S. (2004). “Virus entry: molecular mechanisms and biomedical applications”. In: *Nature Reviews Microbiology 2004 2:2* 2.2, pp. 109–122.

Dion, Moïra B., Pier Luc Plante, Edwige Zufferey, Shiraz A. Shah, Jacques Corbeil, et al. (2021). “Streamlining CRISPR spacer-based bacterial host predictions to decipher the viral dark matter”. In: *Nucleic Acids Research* 49.6, pp. 3127–3138.

Dittman, David J., Taghi M. Khoshgoftaar, and Amri Napolitano (2015). “The Effect of Data Sampling When Using Random Forest on Imbalanced Bioinformatics Data”. In: *2015 IEEE*

International Conference on Information Reuse and Integration. Institute of Electrical and Electronics Engineers Inc., pp. 457–463.

Donia, Mohamed S., Peter Cimermancic, Christopher J. Schulze, Laura C. Wieland Brown, John Martin, et al. (2014). “A Systematic Analysis of Biosynthetic Gene Clusters in the Human Microbiome Reveals a Common Family of Antibiotics”. In: *Cell* 158.6, pp. 1402–1414.

dos Santos Bezerra, Rafael, Leonardo Scalon de Oliveira, Edson L. Moretto, Eugênia M. Amorim Ubiali, Roberta Maraninchi Silveira, et al. (2021). “Viral metagenomics in blood donations with post-donation illness reports from Brazil”. In: *Blood Transfusion* 19.2, p. 93.

Drake, J W and J J Holland (1999). “Mutation rates among RNA viruses.” In: *Proceedings of the National Academy of Sciences of the United States of America* 96.24, pp. 13910–3.

Dröge, J., I. Gregor, and A. C. McHardy (2015). “Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods”. In: *Bioinformatics* 31.6, pp. 817–824.

Duerkop, Breck A and Lora V Hooper (2013). “Resident viruses and their interactions with the immune system”. In: *Nature Immunology* 14.7, pp. 654–659.

Duffy, Siobain (2018). “Why are RNA virus mutation rates so damn high?” In: *PLOS Biology* 16.8, e3000003.

Durzyńska, Julia and Anna Goździcka-Józefiak (2015). “Viruses and cells intertwined since the dawn of evolution Emerging viruses”. In: *Virology Journal* 12.1, pp. 1–10.

Dutilh, Bas E., Noriko Cassman, Katelyn McNair, Savannah E. Sanchez, Genivaldo G. Z. Silva, et al. (2014). “A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes”. In: *Nature Communications* 5.1, p. 4498.

Dutilh, Bas E, Arvind Varsani, Yigang Tong, Peter Simmonds, Sead Sabanadzovic, et al. (2021). “Perspective on taxonomic classification of uncultivated viruses”. In: *Current Opinion in Virology* 51, pp. 207–215.

Edgar, Robert C., Jeff Taylor, Victor Lin, Tomer Altman, Pierre Barbera, et al. (2022). “Petabase-scale sequence alignment catalyses viral discovery”. In: *Nature 2022*, pp. 1–6.

Edlund, Anna, Youngik Yang, Shibu Yooseph, Adam P. Hall, Don D. Nguyen, et al. (2015). “Meta-omics uncover temporal regulation of pathways across oral microbiome genera during in vitro sugar metabolism”. In: *ISME Journal* 9.12, pp. 2605–2619.

Edwards, Robert A., Alejandro A. Vega, Holly M. Norman, Maria Ohaeri, Kyle Levi, et al. (2019). “Global phylogeography and ancient evolution of the widespread human gut virus crAssphage”. In: *Nature Microbiology*, p. 1.

Edwards, Robert A, Katelyn Mcnair, Karoline Faust, Jeroen Raes, and Bas E Dutilh (2016). “Computational approaches to predict bacteriophage-host relationships”. In: *FEMS Microbiology Reviews* 048, pp. 258–272.

Eren, A. Murat, Ozcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, et al. (2015). “Anvi’o: An advanced analysis and visualization platformfor ’omics data”. In: *PeerJ* 2015.10, e1319.

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller (2016). “MultiQC: summarize analysis results for multiple tools and samples in a single report”. In: *Bioinformatics* 32.19, pp. 3047–3048.

Fauquet, C. M. (2008). “Taxonomy, Classification and Nomenclature of Viruses”. In: *Encyclopedia of Virology*, p. 9.

Fawaz, Mohammed, Periyasamy Vijayakumar, Anamika Mishra, Pradeep N. Gandhale, Rupam Dutta, et al. (2016). “Duck gut viral metagenome analysis captures snapshot of viral diversity”. In: *Gut Pathogens* 8.1, p. 30.

Finn, Robert D., Jody Clements, and Sean R. Eddy (2011). “HMMER web server: interactive sequence similarity searching”. In: *Nucleic Acids Research* 39.Web Server issue, W29.

Flygare, Steven, Keith Simmon, Chase Miller, Yi Qiao, Brett Kennedy, et al. (2016). “Taxonomer: An interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling”. In: *Genome Biology* 17.1, pp. 1–18.

Forterre, Patrick (2001). “Genomics and early cellular evolution. The origin of the DNA world”. In: *Comptes Rendus de l’Académie des Sciences - Series III - Sciences de la Vie* 324.12, pp. 1067–1076.

— (2002). “The origin of DNA genomes and DNA replication proteins”. In: *Current Opinion in Microbiology* 5.5, pp. 525–532.

— (2006). “The origin of viruses and their possible roles in major evolutionary transitions”. In: *Virus Research* 117.1, pp. 5–16.

— (2011). “Manipulation of cellular syntheses and the nature of viruses: The virocell concept”. In: *Comptes Rendus Chimie* 14.4, pp. 392–399.

Forterre, Patrick and Morgan Gaïa (2021). “The Origin of Viruses”. In: *Encyclopedia of Virology (Fourth Edition)*. Ed. by Dennis H. Bamford and Mark Zuckerman. Fourth Edition. Oxford: Academic Press, pp. 14–22.

Foulex, Aurélie, Matteo Coen, Abdessalam Cherkaoui, Vladimir Lazarevic, Nadia Gaïa, et al. (2019). “Listeria monocytogenes infectious periaortitis: A case report from the infectious disease standpoint”. In: *BMC Infectious Diseases* 19.1.

Foulongne, Vincent, Virginie Sauvage, Charles Hebert, Olivier Dereure, Justine Cheval, et al. (2012). “Human Skin Microbiota: High Diversity of DNA Viruses Identified on the Human Skin by High Throughput Sequencing”. In: *PLoS ONE* 7.6. Ed. by Amanda Ewart Toland, e38499.

Fox, G. E., K. R. Pechman, and C. R. Woese (1977). “Comparative cataloging of 16S ribosomal ribonucleic acid: molecular approach to procaryotic systematics”. In: *International Journal of Systematic Bacteriology* 27.1, pp. 44–57.

Freer, Giulia, Fabrizio Maggi, Massimo Pifferi, Maria E. Di Cicco, Diego G. Peroni, et al. (2018). "The virome and its major component, Anellovirus, a convoluted system molding human immune defenses and possibly affecting the development of asthma and respiratory diseases in childhood". In: *Frontiers in Microbiology* 9.APR, p. 686.

Fung, Siu, Stanley Ho, Nicole Wheeler, Andrew D Millard, and Willem Van Schaik (2022). "Gauge your phage: Benchmarking of bacteriophage identification tools in metagenomic sequencing data". In: *bioRxiv*, p. 2021.04.12.438782.

Galloway-Peña, Jessica and Blake Hanson (2020). "Tools for Analysis of the Microbiome". In: *Digestive Diseases and Sciences* 65.3, pp. 674–685.

Gardy, Jennifer L. and Nicholas J. Loman (2017). "Towards a genomics-informed, real-time, global pathogen surveillance system". In: *Nature Reviews Genetics 2017 19:1* 19.1, pp. 9–20.

El-Gebali, Sara, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, et al. (2018). "The Pfam protein families database in 2019". In: *Nucleic Acids Research* 47, pp. 427–432.

Gevers, Dirk, Rob Knight, Joseph F. Petrosino, Katherine Huang, Amy L. McGuire, et al. (2012). "The Human Microbiome Project: A Community Resource for the Healthy Human Microbiome". In: *PLoS Biology* 10.8, e1001377.

Gevers, Dirk, Subra Kugathasan, Lee A Denson, Yoshiki Vázquez-Baeza, Will Van Treuren, et al. (2014). "The treatment-naive microbiome in new-onset Crohn's disease." In: *Cell host & microbe* 15.3, pp. 382–392.

Ghannam, Ryan B. and Stephen M. Techtmann (2021). "Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring". In: *Computational and Structural Biotechnology Journal* 19, pp. 1092–1107.

Gough, Julian, Kevin Karplus, Richard Hughey, and Cyrus Chothia (2001). "Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure". In: *Journal of Molecular Biology* 313.4, pp. 903–919.

Gregory, Ann C., Olivier Zablocki, Ahmed A. Zayed, Allison Howell, Benjamin Bolduc, et al. (2020). "The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut". In: *Cell Host and Microbe* 28.5, 724–740.e8.

Greninger, Alexander L. (2018). "A decade of RNA virus metagenomics is (not) enough". In: *Virus Research* 244, pp. 218–229.

Griffith, Daniel M., Joseph A. Veech, and Charles J. Marsh (2016). "cooccur: Probabilistic Species Co-Occurrence Analysis in R". In: *Journal of Statistical Software* 69.1, pp. 1–17.

Guchte, Maarten van de, Hervé M. Blottière, and Joël Doré (2018). "Humans as holobionts: implications for prevention and therapy". In: *Microbiome* 6.1, p. 81.

Guerin, Emma, Andrey Shkoporov, Stephen R. Stockdale, Adam G. Clooney, Feargal J. Ryan, et al. (2018). "Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut". In: *Cell Host & Microbe* 24.5, 653–664.e6.

Guo, Jiarong, Ben Bolduc, Ahmed A. Zayed, Arvind Varsani, Guillermo Dominguez-Huerta, et al. (2021a). “VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses”. In: *Microbiome* 9.1, p. 37.

Guo, Jiarong, Dean Vik, Akbar Adjie Pratama, Simon Roux, and Matthew Sullivan (2021b). *Viral sequence identification SOP with VirSorter2*.

Hall, James P.J., Ellie Harrison, and David A. Baltrus (2022). “Introduction: the secret lives of microbial mobile genetic elements”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 377 (1842).

Harris, Hugh M.B. and Colin Hill (2021). “A Place for Viruses on the Tree of Life”. In: *Frontiers in Microbiology* 11, p. 3449.

Hatcher, Eneida L., Sergey A. Zhdanov, Yiming Bao, Olga Blinkova, Eric P. Nawrocki, et al. (2017). “Virus Variation Resource - improved response to emergent viral outbreaks”. In: *Nucleic acids research* 45.D1, pp. D482–D490.

Hatem, Ayat, Doruk Bozdağ, Amanda E Toland, and Ümit V Çatalyürek (2013). “Benchmarking short sequence mapping tools.” En. In: *BMC bioinformatics* 14.1, p. 184.

Hattori, Norimichi, Makoto Kuroda, Harutaka Katano, Takahiro Takuma, Takayoshi Ito, et al. (2020). “Candidatus Mycoplasma haemohominis in Human, Japan”. In: *Emerging Infectious Diseases* 26.1, p. 11.

Hoffmann, Bernd, Matthias Scheuch, Dirk Höper, Ralf Jungblut, Mark Holsteg, et al. (2012). “Novel Orthobunyavirus in Cattle, Europe, 2011”. In: *Emerging Infectious Diseases* 18.3, pp. 469–472.

Holmes, Edward C. (2009). “The Evolutionary Genetics of Emerging Viruses”. In: *http://dx.doi.org/10.1146/annurev.ecolsys.110308.120248* 40, pp. 353–372.

— (2022). “COVID-19—lessons for zoonotic disease”. In: *Science* 375.6585, pp. 1114–1115.

Honda, Kenya and Dan R. Littman (2012). “The Microbiome in Infectious Disease and Inflammation”. In: *Annual Review of Immunology* 30.1, pp. 759–795.

Hood, Leroy (2003). “Systems biology: integrating technology, biology, and computation”. In: *Mechanisms of Ageing and Development* 124.1, pp. 9–16.

Horvath, Philippe and Rodolphe Barrangou (2010). “CRISPR/Cas, the immune system of Bacteria and Archaea”. In: *Science* 327.5962, pp. 167–170.

How many viruses on Earth? (2019). URL: http://www.virology.ws/2013/09/06/how-many-viruses-on-earth/ (visited on 01/29/2019).

Howe, Adina Chuang, Janet K Jansson, Stephanie A Malfatti, Susannah G Tringe, James M Tiedje, et al. (2014). “Tackling soil diversity with the assembly of large, complex metagenomes.” In: *Proceedings of the National Academy of Sciences of the United States of America* 111.13, pp. 4904–9.

Huerta-Cepas, Jaime, François Serra, and Peer Bork (2016). "ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data". In: *Molecular Biology and Evolution* 33.6, pp. 1635–1638.

Hug, Laura A., Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, et al. (2016). "A new view of the tree of life". In: *Nature Microbiology* 1.5, p. 16048.

Hunt, Martin, Nishadi De Silva, Thomas D. Otto, Julian Parkhill, Jacqueline A. Keane, et al. (2015). "Circlator: Automated circularization of genome assemblies using long sequencing reads". In: *Genome Biology* 16.1, pp. 1–10.

Hunter, Sarah, Matthew Corbett, Hubert Denise, Matthew Fraser, Alejandra Gonzalez-Beltran, et al. (2014). "EBI metagenomics - A new resource for the analysis and archiving of metagenomic data". In: *Nucleic Acids Research* 42.D1, pp. 600–606.

Hurwitz, Bonnie L. and Matthew B. Sullivan (2013). "The Pacific Ocean Virome (POV): A Marine Viral Metagenomic Dataset and Associated Protein Clusters for Quantitative Viral Ecology". In: *PLoS ONE* 8.2. Ed. by Fabiano Thompson, e57355.

Huson, Daniel H., Alexander F. Auch, Ji Qi, and Stephan C. Schuster (2007). "MEGAN analysis of metagenomic data". In: *Genome Research* 17.3, pp. 377–386.

Hyatt, Doug, Gwo Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, et al. (2010). "Prodigal: Prokaryotic gene recognition and translation initiation site identification". In: *BMC Bioinformatics* 11.1, pp. 1–11.

ICTV (2012). "Nidovirales". In: *Virus Taxonomy*. Ed. by Andrew M.Q. King, Michael J. Adams, Eric B. Carstens, and Elliot J. Lefkowitz. San Diego: Elsevier, p. i.

— (2021a). *Virus Metadata Repository: version August 1, 2020; MSL35*. URL: https://talk.ictvonline.org/taxonomy/vmr/m/vmr-file-repository/10312 (visited on 06/24/2021).

— (2021b). *VMR: Virus Metadata Resource*. URL: https://talk.ictvonline.org/taxonomy/vmr/ (visited on 06/24/2021).

Imelfort, Michael, Donovan Parks, Ben J. Woodcroft, Paul Dennis, Philip Hugenholtz, et al. (2014). "GroopM: An automated tool for the recovery of population genomes from related metagenomes". In: *PeerJ* 2014.1.

Iranzo, Jaime, Mart Krupovic, and Eugene V. Koonin (2016). "The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing". In: *mBio* 7.4.

Iuchi, Hitoshi, Taro Matsutani, Keisuke Yamada, Natsuki Iwano, Shunsuke Sumi, et al. (2021). "Representation learning applications in biological sequence analysis". In: *Computational and Structural Biotechnology Journal* 19, pp. 3198–3208.

Jerome, Hanna, Callum Taylor, Vattipally B Sreenu, Tanya Klymenko, Ana Da, et al. (2019). "Metagenomic next-generation sequencing aids the diagnosis of viral infections in febrile returning travellers". In: *Journal of Infection* 79, pp. 383–388.

Jia, Hengxia and Peng Gong (2019). “A structure-function diversity survey of the rna-dependent rna polymerases from the positive-strand rna viruses”. In: *Frontiers in Microbiology* 10.AUG, p. 1945.

Jia, Xiaofang, Lvyin Hu, Min Wu, Yun Ling, Wei Wang, et al. (2021). “A streamlined clinical metagenomic sequencing protocol for rapid pathogen identification”. In: *Scientific Reports 2021 11:1* 11.1, pp. 1–10.

Jiao, Jian Yu, Lan Liu, Zheng Shuang Hua, Bao Zhu Fang, En Min Zhou, et al. (2021). “Microbial dark matter coming to light: challenges and opportunities”. In: *National Science Review* 8.3, p. 2021.

Jones, Brian V. and Julian R. Marchesi (2006). “Transposon-aided capture (TRACA) of plasmids resident in the human gut mobile metagenome”. In: *Nature Methods 2007 4:1* 4.1, pp. 55–61.

Jones, David T. (2019). “Setting the standards for machine learning in biology”. In: *Nature Reviews Molecular Cell Biology* 20.11, pp. 659–660.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, et al. (2021). “Highly accurate protein structure prediction with AlphaFold”. In: *Nature 2021 596:7873* 596.7873, pp. 583–589.

Kaczorowska, Joanna and Lia van der Hoek (2020). “Human anelloviruses: diverse, omnipresent and commensal members of the virome”. In: *FEMS Microbiology Reviews*.

Kang, Dongwan D., Jeff Froula, Rob Egan, and Zhong Wang (2015). “MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities”. In: *PeerJ* 3, e1165.

Kang, Dongwan, Feng Li, Edward S Kirton, Ashleigh Thomas, Rob S Egan, et al. (2019). “MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies”. In:

Karsch-Mizrachi, Ilene, Toshihisa Takagi, Guy Cochrane, and on behalf of the International Nucleotide Sequence Database Collaboration (2017). “The international nucleotide sequence database collaboration”. In: *Nucleic Acids Research* 46.D1, pp. D48–D51.

Kashaf, Sara Saheb, Diana M Proctor, Clay Deming, Paul Saary, Martin Hölzer, et al. (2022). “Integrating cultivation and metagenomics for a multi-kingdom view of skin microbiome diversity and functions”. In: *Nature Microbiology* 7, pp. 169–179.

Katoh, K. and D. M. Standley (2013). “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability”. In: *Molecular Biology and Evolution* 30.4, pp. 772–780.

Katz, Kenneth, Oleg Shutov, Richard Lapoint, Michael Kimelman, J. Rodney Brister, et al. (2022). “The Sequence Read Archive: a decade more of explosive growth”. In: *Nucleic Acids Research* 50.D1, pp. D387–D390.

Kessel, Maartje A. H. J. van, Daan R. Speth, Mads Albertsen, Per H. Nielsen, Huub J. M. Op den Camp, et al. (2015). “Complete nitrification by a single microorganism”. In: *Nature* 528.7583, pp. 555–559.

Kim, Daehwan, Li Song, Florian P. Breitwieser, and Steven L. Salzberg (2016). “Centrifuge: Rapid and sensitive classification of metagenomic sequences”. In: *Genome Research* 26.12, pp. 1721–1729.

“Order - Picornavirales” (2012). In: *Virus Taxonomy*. Ed. by Andrew M.Q. King, Michael J. Adams, Eric B. Carstens, and Elliot J. Lefkowitz. San Diego: Elsevier, pp. 835–839.

King, Andrew M.Q., Elliot Lefkowitz, Michael J. Adams, and Eric B. Carstens (2021). “Virus Taxonomy”. In: *Virus Taxonomy*, pp. 28–37.

Klug, A. (1999). “The tobacco mosaic virus particle: structure and assembly.” In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 354.1383, p. 531.

Koonin, Eugene V. (2018). “Environmental microbiology and metagenomics: the Brave New World is here, what’s next?” In: *Environmental Microbiology* 20.12, pp. 4210–4212.

Koonin, Eugene V. and Valerian V. Dolja (2018). “Metaviromics: a tectonic shift in understanding virus evolution”. In: *Virus Research* 246, A1–A3.

— (2021a). “The Greater Virus World and Its Evolution”. In: *Encyclopedia of Virology (Fourth Edition)*. Ed. by Dennis H. Bamford and Mark Zuckerman. Fourth Edition. Oxford: Academic Press, pp. 38–46.

Koonin, Eugene V., Valerian V. Dolja, and Mart Krupovic (2021b). “The healthy human virome: from virus–host symbiosis to disease”. In: *Current Opinion in Virology* 47, pp. 86–94.

Koonin, Eugene V., Valerian V. Dolja, Mart Krupovic, Arvind Varsani, Yuri I. Wolf, et al. (2020a). “Global Organization and Proposed Megataxonomy of the Virus World”. In: *Microbiology and Molecular Biology Reviews* 84.2.

Koonin, Eugene V., Tatiana G. Senkevich, and Valerian V. Dolja (2006). “The ancient virus world and evolution of cells”. In: *Biology Direct* 1.1, pp. 1–27.

Koonin, Eugene V. and Petro Starokadomskyy (2016). “Are viruses alive? The replicator paradigm sheds decisive light on an old but misguided question”. In: *Studies in history and philosophy of biological and biomedical sciences* 59, p. 125.

Koonin, Eugene V. and Natalya Yutin (2020b). “The crAss-like Phage Group: How Metagenomics Reshaped the Human Virome”. In: *Trends in Microbiology* 28.5, pp. 349–359.

Köster, Johannes and Sven Rahmann (2012). “Snakemake—a scalable bioinformatics workflow engine”. In: *Bioinformatics* 28.19, pp. 2520–2522.

Kowarsky, Mark, Joan Camunas-Soler, Michael Kertesz, Iwijn De Vlaminck, Winston Koh, et al. (2017). “Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA.” In: *Proceedings of the National Academy of Sciences of the United States of America* 114.36, pp. 9623–9628.

Krishnamurthy, Siddharth R. and David Wang (2017). “Origins and challenges of viral dark matter”. In: *Virus Research* 239, pp. 136–142.

— (2018). “Extensive conservation of prokaryotic ribosomal binding sites in known and novel picobirnaviruses”. In: *Virology* 516, pp. 108–114.

Lang, Andrew S., Matthew L. Rise, Alexander I. Culley, and Grieg F. Steward (2009). “RNA viruses in the sea”. In: *FEMS Microbiology Reviews* 33.2, pp. 295–323.

Lang, Andrew S., Marli Vlok, Alexander I. Culley, Curtis A. Suttle, Yoshitake Takao, et al. (2021). “ICTV virus taxonomy profile: *Marnaviridae* 2021”. In: *Journal of General Virology* 102.8, p. 001633.

Langmead, Ben and Steven L Salzberg (2012). “Fast gapped-read alignment with Bowtie 2”. In: *Nature Methods* 9.4, pp. 357–359.

Lara Pinto, Andressa Zelenski de, Michellen Santos de Carvalho, Fernando Lucas de Melo, Ana Lúcia Maria Ribeiro, Bergmann Morais Ribeiro, et al. (2017). “Novel viruses in salivary glands of mosquitoes from sylvatic Cerrado, Midwestern Brazil”. In: *PLOS ONE* 12.11. Ed. by Bradley S. Schneider, e0187429.

Larrañaga, Pedro, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, et al. (2006). “Machine learning in bioinformatics”. In: *Briefings in Bioinformatics* 7.1, pp. 86–112.

Lassalle, Florent, Matteo Spagnoletti, Matteo Fumagalli, Liam Shaw, Mark Dyble, et al. (2018). “Oral microbiomes from hunter-gatherers and traditional farmers reveal shifts in commensal balance and pathogen load linked to diet”. In: *Molecular Ecology* 27.1, pp. 182–195.

Lecuit, Marc and Marc Eloit (2013). “The human virome: new tools and concepts”. In: *Trends in Microbiology* 21.10, pp. 510–515.

Legendre, Matthieu, Defne Arslan, Chantal Abergel, and Jean-Michel Claverie (2012). “Genomics of Megavirus and the elusive fourth domain of Life”. In: *Communicative & Integrative Biology* 5.1, p. 102.

Leggett, Richard M., Cristina Alcon-Giner, Darren Heavens, Shabhonam Caim, Thomas C. Brook, et al. (2019). “Rapid MinION profiling of preterm microbiota and antimicrobial-resistant pathogens”. In: *Nature Microbiology 2019 5:3* 5.3, pp. 430–442.

Leite, Diogo Manuel Carvalho, Xavier Brochet, Grégory Resch, Yok Ai Que, Aitana Neves, et al. (2018). “Computational prediction of inter-species relationships through omics data analysis and machine learning”. In: *BMC Bioinformatics* 19.14, pp. 151–159.

Leo, Stefano, Nadia Gaïa, Etienne Ruppé, Stephane Emonet, Myriam Girard, et al. (2017). “Detection of Bacterial Pathogens from Broncho-Alveolar Lavage by Next-Generation Sequencing”. In: *International journal of molecular sciences* 18.9.

Li, Dinghua, Chi Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak Wah Lam (2015). “MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph”. In: *Bioinformatics (Oxford, England)* 31.10, pp. 1674–1676.

Li, H. and R. Durbin (2009). “Fast and accurate short read alignment with Burrows-Wheeler transform”. In: *Bioinformatics* 25.14, pp. 1754–1760.

Li, Heng (2018). “Minimap2: pairwise alignment for nucleotide sequences”. In: *Bioinformatics* 34.18, pp. 3094–3100.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, et al. (2009). “The Sequence Alignment/Map format and SAMtools.” In: *Bioinformatics (Oxford, England)* 25.16, pp. 2078–9.

Li, Simone S., Ana Zhu, Vladimir Benes, Paul I. Costea, Rajna Hercog, et al. (2016). “Durable coexistence of donor and recipient strains after fecal microbiota transplantation”. In: *Science* 352.6285, pp. 586–589.

Li, Tony, Placide Mbala-Kingebeni, Samia N. Naccache, Julien Thézé, Jerome Bouquet, et al. (2019). “Metagenomic Next-Generation Sequencing of the 2014 Ebola Virus Disease Outbreak in the Democratic Republic of the Congo”. In: *Journal of Clinical Microbiology* 57.9.

Li, Yang, Hao Wang, Kai Nie, Chen Zhang, Yi Zhang, et al. (2016). “VIP: an integrated pipeline for metagenomics of virus identification and discovery”. In: *Scientific Reports 2016 6:1* 6.1, pp. 1–10.

Liang, Guanxiang and Frederic D. Bushman (2021). “The human virome: assembly, composition and host interactions”. In: *Nature Reviews Microbiology* 19.8, pp. 514–527.

Libbrecht, Maxwell W. and William Stafford Noble (2015). “Machine learning applications in genetics and genomics”. In: *Nature Reviews Genetics* 16.6, pp. 321–332.

Lin, Hsin Hung and Yu Chieh Liao (2016). “Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes”. In: *Scientific Reports 2016 6:1* 6.1, pp. 1–8.

Liu, Ben, Sirisha Thippabhotla, Jun Zhang, and Cuncong Zhong (2021). “DRAGoM: Classification and Quantification of Noncoding RNA in Metagenomic Data”. In: *Frontiers in Genetics* 12, p. 669495.

Loman, Nicholas J., Chrystala Constantinidou, Martin Christner, Jacqueline Z.-M. Z.M. Chan, Joshua Quick, et al. (2013). “A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of shiga-toxigenic Escherichia coli O104:H4”. In: *JAMA - Journal of the American Medical Association* 309.14, pp. 1502–1510.

Lu, Jennifer, Florian P. Breitwieser, Peter Thielen, and Steven L. Salzberg (2017). “Bracken: Estimating species abundance in metagenomics data”. In: *PeerJ Computer Science* 2017.1, e104.

Lu, Shennan, Jiyao Wang, Farideh Chitsaz, Myra K Derbyshire, Renata C Geer, et al. (2019). “CDD/SPARCLE: the conserved domain database in 2020”. In: *Nucleic Acids Research* 48.D1, pp. D265–D268.

Luk, Ka Cheung, Michael G. Berg, Samia N. Naccache, Beniwende Kabre, Scot Federman, et al. (2015). “Utility of Metagenomic Next-Generation Sequencing for Characterization of HIV and Human Pegivirus Diversity”. In: *PLOS ONE* 10.11, e0141723.

Luo, Ruibang, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, et al. (2012). “SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler”. In: *GigaScience* 1.1, p. 18.

Maabar, Maha, Andrew J. Davison, Matej Vučak, Fiona Thorburn, Pablo R. Murcia, et al. (2019). “DisCVR: Rapid viral diagnosis from high-throughput sequencing data”. In: *Virus Evolution* 5.2.

Maarala, Altti Ilari, Zurab Bzhalava, Joakim Dillner, Keijo Heljanko, and Davit Bzhalava (2018). “ViraPipe: scalable parallel pipeline for viral metagenome analysis from next generation sequencing reads”. In: *Bioinformatics* 34.6. Ed. by Bonnie Berger, pp. 928–935.

Maaten, Laurens van der (2021). *t-Distributed Stochastic Neighbor Embedding (t-SNE)*. URL: https://lvdmaaten.github.io/tsne/ (visited on 06/20/2021).

Malla, Muneer Ahmad, Anamika Dubey, Ashwani Kumar, Shweta Yadav, Abeer Hashem, et al. (2019). “Exploring the human microbiome: The potential future role of next-generation sequencing in disease diagnosis and treatment”. In: *Frontiers in Immunology* 10.JAN, p. 2868.

Mallawaarachchi, Vijini, Anuradha Wickramarachchi, and Yu Lin (2020). “GraphBin: refined binning of metagenomic contigs using assembly graphs”. In: *Bioinformatics (Oxford, England)* 36.11, pp. 3307–3313.

Mande, Sharmila S., Monzoorul Haque Mohammed, and Tarini Shankar Ghosh (2012). “Classification of metagenomic sequences: methods and challenges”. In: *Briefings in Bioinformatics* 13.6, pp. 669–681.

Marcy, Yann, Cleber Ouverney, Elisabeth M Bik, Tina Lösekann, Natalia Ivanova, et al. (2007). “Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth.” In: *Proceedings of the National Academy of Sciences of the United States of America* 104.29, pp. 11889–94.

Markowitz, Victor M., I. Min A. Chen, Ken Chu, Ernest Szeto, Krishna Palaniappan, et al. (2012). “IMG/M: the integrated metagenome data management and comparative analysis system”. In: *Nucleic Acids Research* 40.Database issue, p. D123.

Martí-Carreras, Joan, Alejandro Rafael Gener, Sierra D. Miller, Anderson F. Brito, Christiam E. Camacho, et al. (2020). “NCBI’s Virus Discovery Codeathon: Building “FIVE” —The Federated Index of Viral Experiments API Index”. In: *Viruses* 12.12, p. 1424.

Martin, Marcel (2011). “Cutadapt removes adapter sequences from high-throughput sequencing reads”. In: *EMBnet.journal* 17.1, pp. 10–12.

McInerney, James O., James A. Cotton, and Davide Pisani (2008). “The prokaryotic tree of life: past, present... and future?” In: *Trends in ecology & evolution* 23.5, pp. 276–281.

McIntyre, Alexa B.R., Rachid Ounit, Ebrahim Afshinnekoo, Robert J. Prill, Elizabeth Hénaff, et al. (2017). “Comprehensive benchmarking and ensemble approaches for metagenomic classifiers”. In: *Genome Biology* 18.1, pp. 1–19.

Menzel, Peter, Kim Lee Ng, and Anders Krogh (2016). “Fast and sensitive taxonomic classification for metagenomics with Kaiju”. In: *Nature Communications 2016 7:1* 7.1, pp. 1–9.

Mikheenko, Alla, Vladislav Saveliev, and Alexey Gurevich (2016). “MetaQUAST: evaluation of metagenome assemblies”. In: *Bioinformatics* 32.7, pp. 1088–1090.

Minh, Bui Quang, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D. Woodhams, et al. (2020). “IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era”. In: *Molecular Biology and Evolution* 37.5, pp. 1530–1534.

Minot, Samuel, Alexandra Bryson, Christel Chehoud, Gary D Wu, James D Lewis, et al. (2013). “Rapid evolution of the human gut virome.” In: *Proceedings of the National Academy of Sciences of the United States of America* 110.30, pp. 12450–5.

Mitchell, Alex L, Alexandre Almeida, Martin Beracochea, Miguel Boland, Josephine Burgin, et al. (2019). “MGnify: the microbiome analysis resource in 2020”. In: *Nucleic Acids Research* 48.D1, pp. D570–D578.

Mitchell, Alex L, Teresa K Attwood, Patricia C Babbitt, Matthias Blum, Alan Bridge, et al. (2018a). “InterPro in 2019: improving coverage, classification and access to protein sequence annotations”. In: *Nucleic Acids Research* 47.

Mitchell, Alex L, Maxim Scheremetjew, Hubert Denise, Simon Potter, Aleksandra Tarkowska, et al. (2018b). “EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies”. In: *Nucleic Acids Research* 46.D1, pp. D726–D735.

Mizuno, Carolina Megumi, Francisco Rodriguez-Valera, Nikole E. Kimes, and Rohit Ghai (2013). “Expanding the Marine Virosphere Using Metagenomics”. In: *PLoS Genetics* 9.12. Ed. by Eduardo P. C. Rocha, e1003987.

Mock, Florian, Adrian Viehweger, Emanuel Barth, and Manja Marz (2021). “VIDHOP, viral host prediction with deep learning”. In: *Bioinformatics* 37.3, pp. 318–325.

Modha, Sejal, Joseph Hughes, Giovanni Bianco, Heather M. H.M. Ferguson, Barbara Helm, et al. (2019). “Metaviromics Reveals Unknown Viral Diversity in the Biting Midge *Culicoides impunctatus*”. In: *Viruses* 11.9, p. 865.

Modha, Sejal, David L. Robertson, Joseph Hughes, and Richard J. Orton (2022). “Quantifying and Cataloguing Unknown Sequences within Human Microbiomes”. In: *mSystems* 7.2.

Moreira, David and Purificación López-García (2009). “Ten reasons to exclude viruses from the tree of life”. In: *Nature Reviews Microbiology 2009 7:4* 7.4, pp. 306–311.

Moss, Eli L., Dylan G. Maghini, and Ami S. Bhatt (2020). "Complete, closed bacterial genomes from microbiomes using nanopore sequencing". In: *Nature Biotechnology 2020 38:6* 38.6, pp. 701–707.

Moustafa, Ahmed, Chao Xie, Ewen Kirkness, William Biggs, Emily Wong, et al. (2017). "The blood DNA virome in 8,000 humans". In: *PLOS Pathogens* 13.3. Ed. by Robert Belshaw, e1006292.

Muhire, Brejnev Muhizi, Arvind Varsani, and Darren Patrick Martin (2014). "SDT: A Virus Classification Tool Based on Pairwise Sequence Alignment and Identity Calculation". In: *PLOS ONE* 9.9, e108277.

Murali, Adithya, Aniruddha Bhargava, and Erik S. Wright (2018). "IDTAXA: A novel approach for accurate taxonomic classification of microbiome sequences". In: *Microbiome* 6.1, pp. 1–14.

Murat A. (2020). *Microbial Dark Matter: The mullet of microbial ecology – Meren Lab*. URL: http://merenlab.org/2017/06/22/microbial-dark-matter/ (visited on 06/19/2020).

Nakamura, Shota, Cheng-Song Yang, Naomi Sakon, Mayo Ueda, Takahiro Tougan, et al. (2009). "Direct Metagenomic Detection of Viral Pathogens in Nasal and Fecal Specimens Using an Unbiased High-Throughput Sequencing Approach". In: *PLoS ONE* 4.1. Ed. by Peter Sommer, e4219.

Nami, Yousef, Nazila Imeni, and Bahman Panahi (2021). "Application of machine learning in bacteriophage research". In: *BMC Microbiology* 21.1, p. 193.

Nayfach, Stephen, Antonio Pedro Camargo, Frederik Schulz, Emiley Eloe-Fadrosh, Simon Roux, et al. (2020a). "CheckV assesses the quality and completeness of metagenome-assembled viral genomes". In: *Nature Biotechnology 2020 39:5* 39.5, pp. 578–585.

Nayfach, Stephen, David Páez-Espino, Lee Call, Soo Jen Low, Hila Sberro, et al. (2021). "Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome". In: *Nature Microbiology*, pp. 1–11.

Nayfach, Stephen, Antonio Pedro Camargo, Emiley Eloe-Fadrosh, and Simon Roux (2020b). "CheckV: assessing the quality of metagenome-assembled viral genomes". In: *bioRxiv*, p. 2020.05.06.081778.

Nayfach, Stephen, Zhou Jason Shi, Rekha Seshadri, Katherine S. Pollard, and Nikos C. Kyrpides (2019). "New insights from uncultivated genomes of the global human gut microbiome". In: *Nature 2019 568:7753* 568.7753, pp. 505–510.

Necci, Marco, Damiano Piovesan, Zsuzsanna Doszt Anyi, Silvio C E Tosatto, Zsuzsanna Dosztányi, et al. (2017). "MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins". In: *Bioinformatics* 33.9, pp. 1402–1404.

Neri, Uri, Yuri I. Wolf, Simon Roux, Antonio Pedro Camargo, Benjamin D. Lee, et al. (2022). "A five-fold expansion of the global RNA virome reveals multiple new clades of RNA bacteriophages". In: *bioRxiv*, p. 2022.02.15.480533.

Ng, Terry Fei Fan, Wen Zhang, Jana Sachsenrö Der, Nikola O. Kondov, Antonio Charlys Da Costa, et al. (2015). "A diverse group of small circular ssDNA viral genomes in human and non-human primate stools". In: *Virus Evolution* 1.1, p. 17.

NIH Human Microbiome Project - About the Human Microbiome (2020). URL: https://www.hmpdacc.org/hmp/overview/ (visited on 06/19/2020).

Nishizawa, Tsutomu, Hiroaki Okamoto, Keiko Konishi, Hiroshi Yoshizawa, Yuzo Miyakawa, et al. (1997). "A Novel DNA Virus (TTV) Associated with Elevated Transaminase Levels in Posttransfusion Hepatitis of Unknown Etiology". In: *Biochemical and Biophysical Research Communications* 241.1, pp. 92–97.

Nooij, Sam, Dennis Schmitz, Harry Vennema, Annelies Kroneman, and Marion P G Koopmans (2018). "Overview of Virus Metagenomic Classification Methods and Their Biological Applications." In: *Frontiers in microbiology* 9, p. 749.

Norman, Jason M., Scott A. Handley, Megan T. Baldridge, Lindsay Droit, Catherine Y. Liu, et al. (2015). "Disease-Specific Alterations in the Enteric Virome in Inflammatory Bowel Disease". In: *Cell* 160.3, pp. 447–460.

Nurk, Sergey, Anton Bankevich, Dmitry Antipov, Alexey Gurevich, Anton Korobeynikov, et al. (2013). "Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads". In: Springer Berlin Heidelberg, pp. 158–170.

Nurk, Sergey, Dmitry Meleshko, Anton Korobeynikov, and Pavel A Pevzner (2017). "metaSPAdes: a new versatile metagenomic assembler." In: *Genome research* 27.5, pp. 824–834.

Olekhnovich, Evgenii I., Alexander I. Manolov, Andrey E. Samoilov, Nikita A. Prianichnikov, Maja V. Malakhova, et al. (2019). "Shifts in the human gut microbiota structure caused by quadruple *helicobacter pylori* eradication therapy". In: *Frontiers in Microbiology* 10.AUG.

Olm, Matthew R., Christopher T. Brown, Brandon Brooks, and Jillian F. Banfield (2017). "DRep: A tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication". In: *ISME Journal* 11.12, pp. 2864–2868.

Ondov, Brian D., Nicholas H. Bergman, and Adam M. Phillippy (2011a). "Interactive metagenomic visualization in a Web browser". In: *BMC Bioinformatics* 12.1, pp. 1–10.

Ondov, Brian D., Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, et al. (2016). "Mash: Fast genome and metagenome distance estimation using MinHash". In: *Genome Biology* 17.1, pp. 1–14.

Ondov, Brian D, Nicholas H Bergman, Adam M Phillippy, DH Huson, AF Auch, et al. (2011b). "Interactive metagenomic visualization in a Web browser". In: *BMC Bioinformatics* 12.1, p. 385.

Otto, Thomas D., Gary P. Dillon, Wim S. Degrave, and Matthew Berriman (2011). "RATT: Rapid Annotation Transfer Tool". In: *Nucleic Acids Research* 39.9, e57–e57.

Ounit, Rachid, Steve Wanamaker, Timothy J. Close, and Stefano Lonardi (2015). “CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers”. In: *BMC Genomics* 16.1, pp. 1–13.

Paez-Espino, David, Emiley A. Eloe-Fadrosh, Georgios A. Pavlopoulos, Alex D. Thomas, Marcel Huntemann, et al. (2016). “Uncovering Earth’s virome”. In: *Nature* 536.7617, pp. 425–430.

Paez-Espino, David, Georgios A Pavlopoulos, Natalia N Ivanova, and Nikos C Kyrpides (2017). “Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data”. In: *Nature Protocols* 12.8, pp. 1673–1682.

Paez-Espino, David, Simon Roux, I-Min A Chen, Krishna Palaniappan, Anna Ratner, et al. (2019). “IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes”. In: *Nucleic Acids Research* 47.D1, pp. D678–D686.

Païssé, Sandrine, Carine Valle, Florence Servant, Michael Courtney, Rémy Burcelin, et al. (2016). “Comprehensive description of blood microbiome from healthy donors assessed by 16S targeted metagenomic sequencing”. In: *Transfusion* 56.5, pp. 1138–1147.

Parks, Donovan H., Maria Chuvochina, Christian Rinke, Aaron J. Mussig, Pierre Alain Chaumeil, et al. (2022). “GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy”. In: *Nucleic Acids Research* 50.D1, pp. D785–D794.

Parks, Donovan H., Maria Chuvochina, David W. Waite, Christian Rinke, Adam Skarshewski, et al. (2018). “A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life”. In: *Nature Biotechnology* 36.10, p. 996.

Parks, Donovan H., Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson (2015). “CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes”. In: *Genome Research* 25.7, p. 1043.

Parks, Donovan H., Christian Rinke, Maria Chuvochina, Pierre Alain Chaumeil, Ben J. Woodcroft, et al. (2017). “Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life”. In: *Nature Microbiology 2017 2:11* 2 (11), pp. 1533–1542.

Partridge, Sally R., Stephen M. Kwong, Neville Firth, and Slade O. Jensen (2018). “Mobile Genetic Elements Associated with Antimicrobial Resistance”. In: *Clinical Microbiology Reviews* 31 (4).

Pasolli, Edoardo, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, et al. (2019). “Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle”. In: *Cell* 176.0, 649–662.e20.

Pasolli, Edoardo, Lucas Schiffer, Paolo Manghi, Audrey Renson, Valerie Obenchain, et al. (2017). “Accessible, curated metagenomic data through ExperimentHub”. In: *Nature Methods* 14.11, pp. 1023–1024.

Pearson, William R. (2013). “An introduction to sequence similarity ("homology") searching”. In: *Current Protocols in Bioinformatics* 0 3.SUPPL.42.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, et al. (2011). “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Peng, Yu, Henry C M Leung, S M Yiu, and Francis Y L Chin (2012). “IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth”. In: *BIOINFORMATICS ORIGINAL PAPER* 28.11, pp. 1420–142810.

Pettersson, John H.O., Mang Shi, Jon Bohlin, Vegard Eldholm, Ola B. Brynildsrud, et al. (2017). “Characterizing the virome of Ixodes ricinus ticks from northern Europe”. In: *Scientific Reports* 7.1.

Pierce, N. Tessa, Luiz Irber, Taylor Reiter, Phillip Brooks, and C. Titus Brown (2019). “Large-scale sequence comparisons with sourmash”. In: *F1000Research* 8.

Ponsero, Alise J. and Bonnie L. Hurwitz (2019). “The Promises and Pitfalls of Machine Learning for Detecting Viruses in Aquatic Metagenomes”. In: *Frontiers in Microbiology* 10, p. 806.

Poole, Anthony, David Penny, and Britt Marie Sjöberg (2000). “Methyl-RNA: an evolutionary bridge between RNA and DNA?” In: *Chemistry & Biology* 7.12, R207–R216.

Poore, Gregory D., Evguenia Kopylova, Qiyun Zhu, Carolina Carpenter, Serena Fraraccio, et al. (2020). “Microbiome analyses of blood and tissues suggest cancer diagnostic approach”. In: *Nature* 579.7800, p. 567.

Popgeorgiev, Nikolay, Sarah Temmam, Didier Raoult, and Christelle Desnues (2013). “Describing the silent human virome with an emphasis on giant viruses.” In: *Intervirology* 56.6, pp. 395–412.

Poranen, Minna M. and Sari Mäntynen (2017). “ICTV virus taxonomy profile: *Cystoviridae*”. In: *Journal of General Virology* 98.10, pp. 2423–2424.

Pride, David T., Richard J. Meinersmann, Trudy M. Wassenaar, and Martin J. Blaser (2003). “Evolutionary Implications of Microbial Genome Tetranucleotide Frequency Biases”. In: *Genome Research* 13.2, pp. 145–158.

Pride, David T., Trudy M. Wassenaar, Chandrabali Ghose, and Martin J. Blaser (2006). “Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses”. In: *BMC Genomics* 7.1, p. 8.

Qin, Junjie, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, et al. (2010). “A human gut microbial gene catalogue established by metagenomic sequencing”. In: *Nature* 464.7285, pp. 59–65.

Qu, Kaiyang, Fei Guo, Xiangrong Liu, Yuan Lin, and Quan Zou (2019). “Application of machine learning in microbiology”. In: *Frontiers in Microbiology* 10.APR, p. 827.

Quince, Christopher, Alan W Walker, Jared T Simpson, Nicholas J Loman, and Nicola Segata (2017). “Shotgun metagenomics, from sampling to analysis”. In: *Nature Biotechnology* 35.9, pp. 833–844.

Rahman, Md Tanvir, Md Abdus Sobur, Md Saiful Islam, Samina Ievy, Md Jannat Hossain, et al. (2020). “Zoonotic diseases: Etiology, impact, and control”. In: *Microorganisms* 8.9, pp. 1–34.

Rampelotto, Pabulo Henrique (2013). “Extremophiles and Extreme Environments”. In: *Life : Open Access Journal* 3.3, p. 482.

Rangel-Pineros, Guillermo, Alexandre Almeida, Martin Beracochea, Ekaterina Sakharova, Manja Marz, et al. (n.d.). “VIRify: an integrated detection, annotation and taxonomic classification pipeline using virus-specific protein profile hidden Markov models”. In: ().

Rascovan, Nicolás, Raja Duraisamy, and Christelle Desnues (2016). “Metagenomics and the Human Virome in Asymptomatic Individuals”. In: *Annual Review of Microbiology* 70.1, pp. 125–141.

Raymond, Frédéric, Amin A. Ouameur, Maxime Déraspe, Naeem Iqbal, Hélène Gingras, et al. (2016). “The initial state of the human gut microbiome determines its reshaping by antibiotics”. In: *ISME Journal* 10.3, pp. 707–720.

Remnant, Emily J, Mang Shi, Gabriele Buchmann, Tjeerd Blacquière, Edward C Holmes, et al. (2017). “A Diverse Range of Novel RNA Viruses in Geographically Distinct Honey Bee Populations.” In: *Journal of virology* 91.16.

Ren, Jie, Nathan A. Ahlgren, Yang Young Lu, Jed A. Fuhrman, and Fengzhu Sun (2017). “VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data”. In: *Microbiome* 5.1, p. 69.

Ren, Jie, Kai Song, Chao Deng, Nathan A. Ahlgren, Jed A. Fuhrman, et al. (2020). “Identifying viruses from metagenomic data using deep learning”. In: *Quantitative Biology* 8.1, pp. 64–77.

Rice, Peter, Ian Longden, Alan Bleasby, Lan Longden, and Alan Bleasby (2000). “EMBOSS: The European Molecular Biology Open Software Suite”. In: *Trends in Genetics* 16.6, pp. 276–277.

Rinke, Christian, Patrick Schwientek, Alexander Sczyrba, Natalia N. Ivanova, Iain J. Anderson, et al. (2013). “Insights into the phylogeny and coding potential of microbial dark matter”. In: *Nature* 499.7459, pp. 431–437.

Roossinck, Marilyn J. (2011). “The good viruses: viral mutualistic symbioses”. In: *Nature Reviews Microbiology 2011 9:2* 9.2, pp. 99–108.

Roossinck, Marilyn J. and Edelio R. Bazán (2017). “Symbiosis: Viruses as Intimate Partners”. In: *Annual review of virology* 4.1, pp. 123–139.

Rose, Graham, Alexander G. Shaw, Kathleen Sim, David J. Wooldridge, Ming Shi Li, et al. (2017). “Antibiotic resistance potential of the healthy preterm infant gut microbiome”. In: *PeerJ* 2017.1.

Roux, Simon, Evelien M. Adriaenssens, Bas E. Dutilh, Eugene V. Koonin, Andrew M. Kropinski, et al. (2019). “Minimum information about an uncultivated virus genome (MIUVIG)”. In: *Nature Biotechnology* 37.1, pp. 29–37.

Roux, Simon, Francois Enault, Bonnie L. Hurwitz, and Matthew B. Sullivan (2015a). “VirSorter: mining viral signal from microbial genomic data”. In: *PeerJ* 3, e985.

Roux, Simon, Steven J Hallam, Tanja Woyke, and Matthew B Sullivan (2015b). “Viral dark matter and virus–host interactions resolved from publicly available microbial genomes”. In: *eLife* 4.

Roux, Simon, Jelle Matthijnssens, and Bas E. Dutilh (2021a). “Metagenomics in Virology”. In: *Encyclopedia of Virology (Fourth Edition)*. Ed. by Dennis H. Bamford and Mark Zuckerman. Fourth Edition. Oxford: Academic Press, pp. 133–140.

— (2021b). “Metagenomics in Virology”. In: *Encyclopedia of Virology*, p. 133.

Roux, Simon, David Páez-Espino, I. Min A. Chen, Krishna Palaniappan, Anna Ratner, et al. (2021c). “IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses”. In: *Nucleic Acids Research* 49.D1, pp. D764–D775.

Rowlands, David J. (2021). “A Brief History of Virology”. In: *Encyclopedia of Virology (Fourth Edition)*. Ed. by Dennis H. Bamford and Mark Zuckerman. Fourth Edition. Oxford: Academic Press, pp. 3–13.

Rux, John J. and Roger M. Burnett (1998). “Spherical viruses”. In: *Current Opinion in Structural Biology* 8.2, pp. 142–149.

Salazar, Guillem, Lucas Paoli, Adriana Alberti, Jaime Huerta-Cepas, Hans Joachim Ruscheweyh, et al. (2019). “Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome”. In: *Cell* 179.5, p. 1068.

Sanjuán, Rafael and Pilar Domingo-Calap (2021). “Genetic Diversity and Evolution of Viral Populations”. In: *Encyclopedia of Virology (Fourth Edition)*. Ed. by Dennis H. Bamford and Mark Zuckerman. Fourth Edition. Oxford: Academic Press, pp. 53–61.

Sarker, Iqbal H. (2021). “Machine Learning: Algorithms, Real-World Applications and Research Directions”. In: *SN Computer Science 2021 2:3* 2.3, pp. 1–21.

Sausset, R., M. A. Petit, V. Gaboriau-Routhiau, and M. De Paepe (2020). “New insights into intestinal phages”. In: *Mucosal Immunology 2020 13:2* 13.2, pp. 205–215.

Sauvage, V. and M. Eloit (2016). “Viral metagenomics and blood safety”. In: *Transfusion Clinique et Biologique* 23.1, p. 28.

Saw, Jimmy H., Anja Spang, Katarzyna Zaremba-Niedzwiedzka, Lina Juzokaite, Jeremy A. Dodsworth, et al. (2015). “Exploring microbial dark matter to resolve the deep archaeal ancestry of eukaryotes”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1678.

Sayers, Eric (2018). *E-utilities Quick Start*. National Center for Biotechnology Information (US).

Schloss, Patrick D and Jo Handelsman (2003). “Biotechnological prospects from metagenomics”. In: *Current Opinion in Biotechnology* 14.3, pp. 303–310.

Schmieder, Robert, Robert Edwards, and Alex Bateman (2011). “Quality control and preprocessing of metagenomic datasets”. In: *BIOINFORMATICS APPLICATIONS NOTE* 27.6, pp. 863–86410.

Schnettger, Laura, Angela Rodgers, Urska Repnik, Rachel P. Lai, Gang Pei, et al. (2017). “A Rab20-Dependent Membrane Trafficking Pathway Controls M. tuberculosis Replication by Regulating Phagosome Spaciousness and Integrity”. In: *Cell Host and Microbe* 21.5, 619–628.e5.

Sczyrba, Alexander, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, et al. (2017). “Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software”. In: *Nature Methods* 14.11, pp. 1063–1071.

Seemann, Torsten (2014). “Prokka: rapid prokaryotic genome annotation”. In: *Bioinformatics* 30.14, pp. 2068–2069.

Segata, Nicola, Levi Waldron, Annalisa Ballarini, Vagheesh Narasimhan, Olivier Jousson, et al. (2012). “Metagenomic microbial community profiling using unique clade-specific marker genes”. In: *Nature methods* 9.8, p. 811.

Sevvana, Madhumati, Thomas Klose, and Michael G. Rossmann (2021). “Principles of Virus Structure”. In: *Encyclopedia of Virology (Fourth Edition)*. Ed. by Dennis H. Bamford and Mark Zuckerman. Fourth Edition. Oxford: Academic Press, pp. 257–277.

Shaffer, Michael, Mikayla A. Borton, Bridget B. McGivern, Ahmed A. Zayed, Sabina Leanti0000 0003 3527 8101 La Rosa, et al. (2020). “DRAM for distilling microbial metabolism to automate the curation of microbiome function”. In: *Nucleic Acids Research* 48.16, pp. 8883–8900.

Shah, Nidhi, Michael G. Nute, Tandy Warnow, and Mihai Pop (2019). “Misunderstood parameter of NCBI BLAST impacts the correctness of bioinformatics workflows”. In: *Bioinformatics* 35.9, pp. 1613–1614.

Shean, Ryan C., Negar Makhsous, Graham D. Stoddard, Michelle J. Lin, and Alexander L. Greninger (2019). “VAPiD: A lightweight cross-platform viral annotation pipeline and identification tool to facilitate virus genome submissions to NCBI GenBank”. In: *BMC Bioinformatics* 20.1, pp. 1–8.

Shi, Jian Yu, Hua Huang, Yan Ning Zhang, Jiang Bo Cao, and Siu Ming Yiu (2018). “BMCMDA: A novel model for predicting human microbe-disease associations via binary matrix completion”. In: *BMC Bioinformatics* 19.9, pp. 85–92.

Shi, Mang, Xian Dan Lin, Xiao Chen, Jun Hua Tian, Liang Jun Chen, et al. (2018a). “The evolutionary history of vertebrate RNA viruses”. In: *Nature 2018 556:7700* 556.7700, pp. 197–202.

Shi, Mang, Xian-Dan Lin, Jun-Hua Tian, Liang-Jun Chen, Xiao Chen, et al. (2016a). “Redefining the invertebrate RNA virosphere”. In: *Nature* 540.7634, pp. 539–543.

Shi, Mang, Xian-Dan Lin, Nikos Vasilakis, Jun-Hua Tian, Ci-Xiu Li, et al. (2016b). “Divergent Viruses Discovered in Arthropods and Vertebrates Revise the Evolutionary History of the *Flaviviridae* and Related Viruses”. In: *Journal of Virology* 90.2, pp. 659–669.

Shi, Mang, Peter Neville, Jay Nicholson, John-Sebastian Eden, Allison Imrie, et al. (2017). “High-Resolution Metatranscriptomics Reveals the Ecological Dynamics of Mosquito-Associated RNA Viruses in Western Australia.” In: *Journal of virology* 91.17, e00680–17.

Shi, Mang, Yong Zhen Zhang, and Edward C. Holmes (2018b). “Meta-transcriptomics and the evolutionary biology of RNA viruses”. In: *Virus Research* 243, pp. 83–90.

Shkoporov, Andrey N., Adam G. Clooney, Thomas D.S. Sutton, Feargal J. Ryan, Karen M. Daly, et al. (2019a). “The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific”. In: *Cell Host & Microbe* 26.4, 527–541.e5.

Shkoporov, Andrey N. and Colin Hill (2019b). “Bacteriophages of the Human Gut: The "Known Unknown" of the Microbiome”. In: *Cell host & microbe* 25.2, pp. 195–209.

Sieber, Christian M. K., Alexander J. Probst, Allison Sharrar, Brian C. Thomas, Matthias Hess, et al. (2018). “Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy”. In: *Nature Microbiology* 3.7, pp. 836–843.

Siefert, Janet L (2009). “Defining the Mobilome Horizontal Gene Transfer”. In: *Methods in molecular biology (Clifton, N.J.)* Methods in Molecular Biology 532. Ed. by Maria B Gogarten, Johann P Gogarten, and Lorraine C Olendzenski, pp. 13–27.

Simmonds, Peter, Mike J. Adams, Mária Benk, Mya Breitbart, J. Rodney Brister, et al. (2017a). “Consensus statement: Virus taxonomy in the age of metagenomics”. In: *Nature Reviews Microbiology* 15.3, pp. 161–168.

Simmonds, Peter, Paul Becher, Jens Bukh, Ernest A. Gould, Gregor Meyers, et al. (2017b). “ICTV Virus Taxonomy Profile: *Flaviviridae*”. In: *The Journal of general virology* 98.1, pp. 2–3.

Simmonds, Peter, Wenjun Xia, J. K. Baillie, and Ken McKinnon (2013). “Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla -selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses”. In: *BMC Genomics* 14.1, pp. 1–16.

Simpson, Jared T., Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven J.M. Jones, et al. (2009). “ABySS: A parallel assembler for short read sequence data”. In: *Genome Research* 19.6, p. 1117.

Sivarajah, Sivakar (2021). *Dimensionality Reduction for Data Visualization: PCA vs TSNE vs UMAP vs LDA*. URL: https://towardsdatascience.com/dimensionality-reduction-for-data-visualization-pca-vs-tsne-vs-umap-be4aa7b1cb29 (visited on 06/20/2021).

Sklearn (2021). *1.16. Probability calibration — scikit-learn 0.24.2 documentation*. URL: https://scikit-learn.org/stable/modules/calibration.html (visited on 06/20/2021).

Smalla, Kornelia, Sven Jechalke, and Eva M. Top (2015). “Plasmid Detection, Characterization, and Ecology”. In: *Microbiology Spectrum* 3.1.

Söding, Johannes, Andreas Biegert, and Andrei N. Lupas (2005). “The HHpred interactive server for protein homology detection and structure prediction”. In: *Nucleic Acids Research* 33.Web Server issue, W244.

Solden, Lindsey, Karen Lloyd, and Kelly Wrighton (2016). “The bright side of microbial dark matter: lessons learned from the uncultivated majority”. In: *Current Opinion in Microbiology* 31, pp. 217–226.

Souza, William Marciel de, Marcílio Jorge Marc\'\ilio Jorge Fumagalli, Jansen de Araujo, Gilberto Sabino-Santos, Felipe Gonçalves Motta Maia, et al. (2018). “Discovery of novel anelloviruses in small mammals expands the host range and diversity of the *Anelloviridae*”. In: *Virology* 514.July 2017, pp. 9–17.

Spang, Anja, Eva F. Caceres, and Thijs J. G. Ettema (2017). “Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life”. In: *Science* 357.6351, eaaf3883.

Spang, Anja, Jimmy H. Saw, Steffen L. Jørgensen, Katarzyna Zaremba-Niedzwiedzka, Joran Martijn, et al. (2015). “Complex archaea that bridge the gap between prokaryotes and eukaryotes”. In: *Nature* 521.7551, pp. 173–179.

Steinegger, Martin and Steven L. Salzberg (2020). “Terminating contamination: Large-scale search identifies more than 2,000,000 contaminated entries in GenBank”. In: *Genome Biology* 21.1, p. 115.

Steinegger, Martin and Johannes Söding (2017). *MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets.*

— (2018). “Clustering huge protein sequence sets in linear time”. In: *Nature Communications* 9.1, pp. 1–8.

Stern, Adi, Eran Mick, Itay Tirosh, Or Sagy, and Rotem Sorek (2012). “CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome”. In: *Genome Research* 22.10, p. 1985.

Stremlau, Matthew H., Kristian G. Andersen, Onikepe A. Folarin, Jessica N. Grove, Ikponmwonsa Odia, et al. (2015). “Discovery of Novel Rhabdoviruses in the Blood of Healthy Individuals from West Africa”. In: *PLOS Neglected Tropical Diseases* 9.3. Ed. by Charles E Rupprecht, e0003631.

Strous, Marc, Beate Kraft, Regina Bisdorf, and Halina E. Tegetmeyer (2012). “The binning of metagenomic contigs for microbial physiology of mixed cultures”. In: *Frontiers in Microbiology* 3.DEC.

Sunagawa, Shinichi, Daniel R. Mende, Georg Zeller, Fernando Izquierdo-Carrasco, Simon A. Berger, et al. (2013). “Metagenomic species profiling using universal phylogenetic marker genes”. In: *Nature methods* 10.12, pp. 1196–1199.

Tanca, Alessandro, Marcello Abbondio, Antonio Palomba, Cristina Fraumene, Valeria Manghina, et al. (2017). “Potential and active functions in the gut microbiota of a healthy human cohort”. In: *Microbiome* 5.1, pp. 1–15.

Tang, Qin, Yulong Song, Mijuan Shi, Yingyin Cheng, Wanting Zhang, et al. (2015). “Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition”. In: *Scientific Reports 2015 5:1* 5.1, pp. 1–8.

Tanizawa, Yasuhiro, Takatomo Fujisawa, and Yasukazu Nakamura (2018). “DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication”. In: *Bioinformatics* 34.6, pp. 1037–1039.

Tarca, Adi L., Vincent J. Carey, Xue wen Chen, Roberto Romero, and Sorin Drăghici (2007). “Machine learning and its applications to biology.” In: *PLoS computational biology* 3.6, e116.

Teeling, Hanno, Anke Meyerdierks, Margarete Bauer, Rudolf Amann, and Frank Oliver Glöckner (2004). “Application of tetranucleotide frequencies for the assignment of genomic fragments”. In: *Environmental Microbiology* 6.9, pp. 938–947.

The New Science of Metagenomics (2007). Washington, D.C.: National Academies Press.

Thomas, Andrew Maltez and Nicola Segata (2019). “Multiple levels of the unknown in microbiome research”. In: *BMC Biology* 17.1, p. 48.

Thorburn, Fiona, Susan Bennett, Sejal Modha, David Murdoch, Rory Gunson, et al. (2015). “The use of next generation sequencing in the diagnosis and typing of respiratory infections”. eng. In: *Journal of Clinical Virology* 69, pp. 96–100.

Tisza, Michael J., Anna K. Belford, Guillermo Dominguez-Huerta, Benjamin Bolduc, and Christopher B. Buck (2021a). “Cenote-Taker 2 democratizes virus discovery and sequence annotation”. In: *Virus Evolution* 7.1.

Tisza, Michael J., Christopher B. Buck, and Yuan Chang (2021b). “A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases”. In: *Proceedings of the National Academy of Sciences* 118.23, e2023202118.

Tisza, Michael J., Diana V. Pastrana, Nicole L. Welch, Brittany Stewart, Alberto Peretti, et al. (2020). “Discovery of several thousand highly diverse circular DNA viruses”. In: *eLife* 9.

Treangen, Todd J, Sergey Koren, Daniel D Sommer, Bo Liu, Irina Astrovskaya, et al. (2013). “MetAMOS: a modular and open source metagenomic assembly and analysis pipeline.” En. In: *Genome biology* 14.1, R2.

Truong, Duy Tin, Eric A. Franzosa, Timothy L. Tickle, Matthias Scholz, George Weingart, et al. (2015). “MetaPhlAn2 for enhanced metagenomic taxonomic profiling”. In: *Nature Methods 2015 12:10* 12.10, pp. 902–903.

Turnbaugh, Peter J., Christopher Quince, Jeremiah J. Faith, Alice C. McHardy, Tanya Yatsunenko, et al. (2010). “Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins”. In: *Proceedings of the National Academy of Sciences of the United States of America* 107.16, pp. 7503–7508.

Ul, Md Nafis, Alam Id, Umar Faruq, and Chowdhury Id (2020). “Short k-mer abundance profiles yield robust machine learning features and accurate classifiers for RNA viruses”. In: *PLOS ONE* 15.9, e0239381.

Uritskiy, Gherman V., Jocelyne Diruggiero, and James Taylor (2018). “MetaWRAP - A flexible pipeline for genome-resolved metagenomic data analysis 08 Information and Computing

Sciences 0803 Computer Software 08 Information and Computing Sciences 0806 Information Systems”. In: *Microbiome* 6.1, pp. 1–13.

Valieris R. (2020). *parallel fastq-dump wrapper*. URL: https://github.com/rvalieris/parallel-fastq-dump (visited on 06/19/2020).

Van Goethem, Marc W., Andrew R. Osborn, Benjamin P. Bowen, Peter F. Andeer, Tami L. Swenson, et al. (2021). “Long-read metagenomics of soil communities reveals phylum-specific secondary metabolite dynamics”. In: *Communications Biology 2021 4:1* 4.1, pp. 1–10.

Vanni, Chiara, Matthew S Schechter, Silvia G Acinas, Albert Barberán, Pier Luigi Buttigieg, et al. (2022). “Unifying the known and unknown microbial coding sequence space”. In: *eLife* 11.

Varadi, Mihaly, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, et al. (2022). “AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models”. In: *Nucleic Acids Research* 50.D1, pp. D439–D444.

Varsani, Arvind, Tanja Opriessnig, Vladimir Celer, Fabrizio Maggi, Hiroaki Okamoto, et al. (2021). “Taxonomic update for mammalian anelloviruses (family *Anelloviridae*)”. In: *Archives of Virology* 166.10, pp. 2943–2953.

Veech, Joseph A. (2013). “A probabilistic model for analysing species co-occurrence”. In: *Global Ecology and Biogeography* 22.2, pp. 252–260.

Vellai, Tibor and Gábor Vida (1999). “The origin of eukaryotes: the difference between prokaryotic and eukaryotic cells.” In: *Proceedings of the Royal Society B: Biological Sciences* 266.1428, p. 1571.

Venkataraman, Sangita, Burra V.L.S. Prasad, and Ramasamy Selvarajan (2018). “RNA Dependent RNA Polymerases: Insights from Structure, Function and Evolution”. In: *Viruses* 10.2.

Vijayvargiya, Prakhar, Patricio R. Jeraldo, Matthew J. Thoendel, Kerryl E. Greenwood-Quaintance, Zerelda Esquer Garrigos, et al. (2019). “Application of metagenomic shotgun sequencing to detect vector-borne pathogens in clinical blood samples”. In: *PLoS ONE* 14.10.

Villarroel, Julia, Kortine Annina Kleinheinz, Vanessa Isabell Jurtz, Henrike Zschach, Ole Lund, et al. (2016). “HostPhinder: A Phage Host Prediction Tool”. In: *Viruses* 8.5.

Von Meijenfeldt, F. A.Bastiaan, Ksenia Arkhipova, Diego D. Cambuy, Felipe H. Coutinho, and Bas E. Dutilh (2019). “Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT”. In: *Genome Biology* 20.1, p. 217.

Wahba, Lamia, Nimit Jain, Andrew Z Fire, Massa J Shoura, Karen L Artiles, et al. (2020). “An Extensive Meta-Metagenomic Search Identifies SARS-CoV-2-Homologous Sequences in Pangolin Lung Viromes”. In:

Wang, David (2020). “5 challenges in understanding the role of the virome in health and disease”. In: *PLOS Pathogens* 16.3. Ed. by Katherine R. Spindler, e1008318.

Wang, Jinfeng, Yuan Gao, and Fangqing Zhao (2016). “Phage-bacteria interaction network in human oral microbiome”. In: *Environmental microbiology* 18.7, pp. 2143–2158.

Wang, Weili, Jie Ren, Kujin Tang, Emily Dart, Julio Cesar Ignacio-Espinoza, et al. (2020). “A network-based integrated framework for predicting virus–prokaryote interactions”. In: *NAR Genomics and Bioinformatics* 2.2.

Warwick-Dugdale, Joanna, Natalie Solonenko, Karen Moore, Lauren Chittick, Ann C. Gregory, et al. (2019). “Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands”. In: *PeerJ* 2019.4, e6800.

Watson, Mick (2021). *Assigning a taxonomy to a sequence or genome*. URL: https://twitter.com/biomickwatson/status/1421014627111120900 (visited on 07/30/2021).

Webb, Brett, A. G.M. Rakibuzzaman, and Sheela Ramamoorthy (2020). “Torque teno viruses in health and disease”. In: *Virus Research* 285, p. 198013.

Whittle, Emma, Martin O. Leonard, Rebecca Harrison, Timothy W. Gant, and Daniel Paul Tonge (2019). “Multi-method characterization of the human circulating microbiome”. In: *Frontiers in Microbiology* 10.JAN, p. 3266.

Woese, C. R. and G. E. Fox (1977). “Phylogenetic structure of the prokaryotic domain: The primary kingdoms”. In: *Proceedings of the National Academy of Sciences of the United States of America* 74.11, pp. 5088–5090.

Woese, C. R., O. Kandler, and M. L. Wheelis (1990). “Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya”. In: *Proceedings of the National Academy of Sciences of the United States of America* 87.12, pp. 4576–4579.

Wolf, Yuri I., Darius Kazlauskas, Jaime Iranzo, Adriana Lucía-Sanz, Jens H. Kuhn, et al. (2018). “Origins and Evolution of the Global RNA Virome”. In: *mBio* 9.6, e02329–18.

Wolf, Yuri I. and Eugene V. Koonin (2007). “On the origin of the translation system and the genetic code in the RNA world by means of natural selection, exaptation, and subfunctionalization”. In: *Biology Direct* 2.1, pp. 1–25.

Wolf, Yuri I., Sukrit Silas, Yongjie Wang, Shuang Wu, Michael Bocek, et al. (2020). “Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome”. In: *Nature Microbiology 2020 5:10* 5.10, pp. 1262–1270.

Wood, Derrick E. and Steven L. Salzberg (2014). “Kraken: Ultrafast metagenomic sequence classification using exact alignments”. In: *Genome Biology* 15.3, pp. 1–12.

Worobey, Michael (2000). “Extensive Homologous Recombination among Widely Divergent TT Viruses”. In: *Journal of Virology* 74.16, pp. 7666–7670.

Woyke, Tanja, Devin F.R. Doud, and Emiley A. Eloe-Fadrosh (2019). “Genomes from uncultivated microorganisms”. In: *Encyclopedia of Microbiology*. Elsevier, pp. 437–442.

Wu, Fan, Su Zhao, Bin Yu, Yan Mei Chen, Wen Wang, et al. (2020). “A new coronavirus associated with human respiratory disease in China”. In: *Nature 2020 579:7798* 579.7798, pp. 265–269.

Wu, Yu Wei, Blake A. Simmons, and Steven W. Singer (2016). “MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets”. In: *Bioinformatics* 32.4, pp. 605–607.

Yahara, Koji, Masato Suzuki, Aki Hirabayashi, Wataru Suda, Masahira Hattori, et al. (2021). “Long-read metagenomics using PromethION uncovers oral bacteriophages and their interaction with host bacteria”. In: *Nature Communications 2021 12:1* 12.1, pp. 1–12.

Yan, Cheng, Guihua Duan, Fang Xiang Wu, Yi Pan, and Jianxin Wang (2020). “BRWMDA:Predicting Microbe-Disease Associations Based on Similarities and Bi-Random Walk on Disease and Microbe Networks”. In: *IEEE/ACM transactions on computational biology and bioinformatics* 17.5, pp. 1595–1604.

— (2021). “MCHMDA:Predicting Microbe-Disease Associations Based on Similarities and Low-Rank Matrix Completion”. In: *IEEE/ACM transactions on computational biology and bioinformatics* 18.2, pp. 611–620.

Yandell, Mark D. and William H. Majoros (2002). “Genomics and natural language processing”. In: *Nature Reviews Genetics 2002 3:8* 3.8, pp. 601–610.

Youle, Merry, Matthew Haynes, and Forest Rohwer (2012). “Scratching the Surface of Biology’s Dark Matter”. In: *Viruses: Essential Agents of Life*. Dordrecht: Springer Netherlands, pp. 61–81.

Young, Francesca, Simon Rogers, and David L. Robertson (2020). “Predicting host taxonomic information from viral genomes: A comparison of feature representations”. In: *PLoS Computational Biology* 16.5, e1007894.

Yutin, Natalya, Kira S. Makarova, Ayal B. Gussow, Mart Krupovic, Anca Segall, et al. (2018). “Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut”. In: *Nature Microbiology* 3.1, pp. 38–46.

Zablocki, Olivier, Michelle Michelsen, Marie Burris, Natalie Solonenko, Joanna Warwick-Dugdale, et al. (2021). “VirION2: A short and long-read sequencing and informatics workflow to study the genomic diversity of viruses in nature”. In: *PeerJ* 9, e11088.

Zablocki, Olivier, Lonnie van Zyl, Evelien M Adriaenssens, Enrico Rubagotti, Marla Tuffin, et al. (2014). “Niche-dependent genetic diversity in Antarctic metaviromes”. In: *Bacteriophage* 4.4, e980125.

Zaki, Ali M., Sander van Boheemen, Theo M. Bestebroer, Albert D.M.E. Osterhaus, and Ron A.M. Fouchier (2012). “Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia”. In: *New England Journal of Medicine* 367.19, pp. 1814–1820.

Zamkovaya, Tatyana, Jamie S. Foster, Valérie de Crécy-Lagard, and Ana Conesa (2020). “A network approach to elucidate and prioritize microbial dark matter in microbial communities”. In: *The ISME Journal*, pp. 1–17.

Zárate, Selene, Blanca Taboada, Martha Yocupicio-Monroy, and Carlos F. Arias (2017). “Human Virome”. In: *Archives of Medical Research* 48.8, pp. 701–716.

Zaremba-Niedzwiedzka, Katarzyna, Eva F. Caceres, Jimmy H. Saw, Disa Bäckström, Lina Juzokaite, et al. (2017). “Asgard archaea illuminate the origin of eukaryotic cellular complexity”. In: *Nature* 541.7637, pp. 353–358.

Zayed, Ahmed A., James M. Wainaina, Guillermo Dominguez-Huerta, Eric Pelletier, Jiarong Guo, et al. (2022). “Cryptic and abundant marine viruses at the evolutionary origins of Earth’s RNA virome”. In: *Science* 376.6589, pp. 156–162.

Zell, Roland, Marco Groth, Lukas Selinka, and Hans Christoph Selinka (2022). “Picorna-Like Viruses of the Havel River, Germany”. In: *Frontiers in Microbiology* 13, p. 865287.

Zerbino, Daniel R. and Ewan Birney (2008). “Velvet: Algorithms for de novo short read assembly using de Bruijn graphs”. In: *Genome Research* 18.5, p. 821.

Zhang, Fan, Fengxia Zhou, Rui Gan, Chunyan Ren, Yuqiang Jia, et al. (2020). “PHISDetector: a tool to detect diverse in silico phage-host interaction signals for virome studies”. In: *bioRxiv*, p. 661074.

Zhang, Ruoshi, Milot Mirdita, Eli Levy Karin, Clovis Norroy, Clovis Galiez, et al. (2021). “SpacePHARER: sensitive identification of phages from CRISPR spacers in prokaryotic hosts”. In: *Bioinformatics* 37.19, pp. 3364–3366.

Zhang, Wen, Linlin Li, Xutao Deng, Johannes Blümel, C. Micha Nübling, et al. (2016). “Viral nucleic acids in human plasma pools”. In: *Transfusion* 56.9, pp. 2248–2255.

Zhang, Wenyu, Jiajia Chen, Yang Yang, Yifei Tang, Jing Shang, et al. (2011). “A Practical Comparison of De Novo Genome Assembly Software Tools for Next-Generation Sequencing Technologies”. In: *PLOS ONE* 6.3, e17915.

Zitnik, Marinka, Francis Nguyen, Bo Wang, Jure Leskovec, Anna Goldenberg, et al. (2019). “Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities”. In: *Information Fusion* 50, pp. 71–91.

Zou, Shimian, Lis Caler, Sandra Colombini-Hatch, Simone Glynn, and Pothur Srinivas (2016). “Research on the human virome: where are we and what is next”. In: *Microbiome* 4.1, p. 32.

Zulkower, Valentin and Susan Rosser (2020). “DNA Features Viewer: a sequence annotation formatting and plotting library for Python”. In: *Bioinformatics* 36.15, pp. 4350–4352.