The Use of Next Generation Sequencing in the Diagnosis and Typing of Viral Infections

Fiona Thorburn

M.B.Ch.B., M.R.C.P. (UK), D.T.M&H

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

MRC – University of Glasgow Centre for Virus Research

College of Medical, Veterinary and Life Sciences

University of Glasgow

May 2016

Abstract

Viral respiratory infections are associated with substantial mortality, morbidity, and a vast economic and healthcare burden. The diagnosis of such infections has been revolutionised by the introduction of molecular methods such as RT-PCR. This has resulted in high levels of sensitivity and specificity along with a rapid turnaround time in comparison to previous methods. As a product of this success, the diagnosis of respiratory infections makes up a large proportion of the workload in most diagnostic laboratories.

The development of next generation sequencing (NGS) may be the next revolution in the field of virus diagnostics. This allows a metagenomic approach to specimen processing whereby target independent sequencing of all genetic material is carried out.

The research presented in this thesis initially sought to examine if NGS would be feasible in the field of respiratory virus diagnostics. The aim was to apply NGS to clinical specimens in parallel with the current diagnostic RT-PCR assays employed by the West of Scotland Specialist Virology Centre (WoSSVC) to determine if NGS could give the same level of results as RT-PCR and whether the sequence information generated in the process could be used in further characterising the detected pathogens. Further to this, the NGS method was then applied to the detection of norovirus from faecal specimens to demonstrate the utility in other areas of viral diagnostics.

The results show that multiple viral pathogens can be detected from clinical specimens without specific virus targeting. The method was less sensitive than RT-PCR but sequence data generated during the process was utilised viral detection, subtyping and phylogenetic analysis. We also demonstrated that a single workflow could be applied to multiple specimen types in the detection of RNA viral pathogens.

Publications

Thorburn, F., Bennett, S., Modha, S., Murdoch, D., Gunson, R., and Murcia, P. R. (2015) The use of next generation sequencing in the diagnosis and typing of respiratory infections. Journal of Clinical Virology, 69, pp. 96-100. (doi:10.1016/j.jcv.2015.06.082)

Nickbakhsh, S., Thorburn, F., Von Wissmann, B., McMenamin, J., Gunson, R. N., and Murcia, P. R. (2016) Extensive multiplex PCR diagnostics reveals new insights into the epidemiology of viral respiratory infections. Epidemiology and Infection, (doi:10.1017/S0950268816000339)

Abbreviations

ARTI	Acute respiratory tract infection
CPE	Cytopathic effect
Ct	Cycle threshold
DIF	Direct immunofluorescence
DNA	Deoxyribonucleic acid
ELISA	Enzyme-linked Immunosorbent assay
EV	Enterovirus
HA	Haemagglutinin
hBoV	Human bocavirus
hCoV	Human coronavirus
hMPV	Human metapneumovirus
hRV	Human rhinovirus
LRTI	Lower respiratory tract infection
MERS	Middle East respiratory syndrome
NA	Neuraminidase
NGS	Next generation sequencing
PCR	Polymerase chain reaction
RNA	Ribonucleic acid
RSV	Respiratory syncytial virus
RT-PCR	Reverse transcription polymerase chain reaction
SARS	Severe acute respiratory syndrome
URTI	Upper respiratory tract infection
VTM	Viral transport medium
WoSSVC	West of Scotland Specialist Virology Centre

Table of contents

List of Tables

List of Figures

Acknowledgements

There are many people to whom I am very thankful for supporting me during these last few years. Were it not for the support of Dr Pablo Murcia and all the members of the Murcia group, past and present, it would not have been possible to complete this project. A special thank you goes to Dr Gaelle Gonzalez who was so helpful and cheerful every day.

I would also like to thank Dr Rory Gunson for his invaluable guidance and support during this “emotional roller-coaster” that we have all been riding. Many thanks are also due to the staff at the West of Scotland Specialist Virology Centre, especially Susan Bennett, who kindly carried out the multiple PCRs on top of an already hectic work schedule.

A great deal of thanks is due for the various staff at the Centre for Virus Research who have assisted during the course of this project, Dr Joseph Hughes and Sejal Modha for their tireless efforts it the Bioinformatics Department and Dr Gavin Wilkie for ensuring I did not break any sequencing equipment. I would also like to thank Prof Ruth Jarrett and her group for their support and providing laboratory space when I encountered a sudden ceiling malfunction.

To Mary and Stuart (Mum and Dad) thank you for believing that I could do this. Bob, my brilliant husband, you have the patience of a saint and for that I am truly grateful. Thank you for your never faltering support, encouragement and belief in me. I promise I will do my fair share of the housework from now on.

Authors Declaration

I declare that, except where explicit reference is made to the contribution of others, that this research 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.

Introduction

Diagnostic virology and the role of the laboratory has evolved and expanded in recent decades with the implementation of genome detection methods being a significant driving force behind this (Storch 2000). Perhaps one of the areas which have seen the greatest change is that of viral respiratory infections.

Until recently only a small number of respiratory pathogens were recognised and tested for. The methods used in their diagnosis required high levels of expertise, lacked sensitivity and resulted in long turnaround times therefore the outcome would not impact on patient management. The implementation of molecular genome detection methods in a diagnostic setting has brought about a reduction in test turnaround times along with an increase in test sensitivity. Generating results in a timely fashion impacts on the management of individual patients through to population level, through initiation of appropriate therapy, institution of infection control measures and accurate epidemiological surveillance. This ability to generate rapid and accurate results has driven demand of the service, as demonstrated by the progressive increase in the number of sample requests received in Greater Glasgow and Clyde. These are now supported by powerful molecular epidemiological methods which together have had a significant impact on disease surveillance, vaccine development and monitoring.

The technologies used in viral diagnostics continue to evolve. The advent of point-of-care testing may see the next major change in the role of the diagnostic service. Such test can be carried out on a ward, general practice or even community allowing clinical decisions and diagnoses to be made in real time.

This chapter outlines the burden of viral respiratory illness in developing and developed countries. Following this the viral causes of respiratory illness are discussed with reference to their virology epidemiology, clinical presentation, treatment and prevention. I will then discuss the diagnostic methods used to detect these pathogens and in the process show how technology has evolved. The method of next generation sequencing will be outlined and the potential application to the diagnosis of viral respiratory illness discussed. Finally the chapter will conclude with an overview of the aims of this thesis.

At a patient level, the development of antiviral therapeutic agents has increased the requirement for laboratory confirmed diagnoses. The substantial cost of many of these drugs, the duration of treatment required and potential toxicity associated with such compounds necessitates a laboratory confirmed case in order to proceed with treatment.

Diagnosis and the role of patient placement are imperative in the control of disease spread. Though the spread of infection from person to person in a healthcare setting was recognised since the nineteenth century, it took many decades for infection control procedures to be formally introduced in the majority of hospitals. Hospitalised patients can be a vulnerable group with co-morbidities which make them more susceptible to, or more likely to develop severe infections. The recognition of infections with the potential to be spread among patients allows for heightened procedures to be put in place to prevent such spread. Examples include patient cohorting of respiratory syncytial virus positive infants or isolation of suspected influenza or norovirus positive patients. In these cases rapid diagnoses are critical to prevent avoidable patient exposure and unnecessary isolation.

At a population level virus detection has public health implications, such as the identification of notifiable conditions or pathogens that require enhanced surveillance, contact tracing or follow up. An example of this would be the Scottish Enhanced Respiratory Virus Infection Surveillance (SERVIS) system. As part of this surveillance system, primary care practitioners submit samples for viral diagnostic testing in patients who present with an influenza-like illness. As not all of these will be caused by influenza viruses this information is important to gather to determine the start and end of an influenza season. This subsequently impacts on vaccination programs and resource planning. The diagnostic laboratory can also play a part in epidemiological surveillance programmes, such as vaccine effectiveness monitoring.

The common methods employed in the detection of viral pathogens will be discussed in detail later in the chapter, but in brief, the move from classical techniques relying on virus culture towards the use of molecular tests in the diagnostic service has dramatically reduced the turn-around time to generate

results, thus improving the effectiveness of the system and therefore increasing reliance and demand of clinicians for rapid and accurate results.

The greatest burden on a routine viral diagnostic service is that of respiratory infections. This will be in part due to the high incidence in a population but also the large number of pathogens capable of causing such infections. For this reason the main focus of this research will be on the detection of respiratory viruses. The common viral pathogens found in respiratory infections will be discussed in detail later in the chapter.

1.1 Importance of Acute Respiratory Tract Infections

Acute respiratory tract infections (ARTI) represent a significant global health issue. The World Health Organization determined that they are the third leading cause of death in all age groups and the leading cause of death in children under five years old (ARIA 2010). This occurrence is particularly marked in developing countries with 70% of all childhood deaths due to ARTI occurring in Africa and South East Asia (Williams, Gouws et al. 2002). Pneumonia, the most severe form of ARTI, is estimated to affect 35 million children under five years old in Africa alone and resulting in over 750,000 deaths annually (Robert F Breiman 2015). It should be noted that the true burden of disease in developing areas may be underestimated as diagnoses, documentation and information dissemination are often impaired.

The majority of mortality associated with ARTI occurs in developing countries but there are multiple points to consider when determining the true burden of respiratory infections. The common cold is often thought of as a benign infection in the developed world and while it is true that the vast majority of cases are mild and self-limiting, there are substantial economic losses associated with the illness. In the US alone there are an estimated 214 million lost work days per year due to absenteeism and loss of productivity associated with ARTI and a further 2 million lost work days when care giving is taken into account (Bramley, Lerner et al. 2002). The indirect financial cost associated with such losses in the US has been estimated to approach $22.5 billion (Fendrick, Monto et al. 2003).

Another aspect to consider when calculating the burden of ARTI is that of healthcare attendance and treatment costs. Respiratory diseases are the commonest reason for general practice consultations in the UK (Ashworth, Charlton et al. 2005); in fact it is estimated that around 20% of the population present to their GP with a respiratory infection each year (Fleming, Smith et al. 2002). Many of these consultations result in a prescription and indeed the vast majority of antibiotics prescribed in the UK originate from general practices. The most commonly cited reason for antibiotic prescription is that of respiratory infection, despite the assumption that many of these cases are of viral aetiology (Lindbaek 2006). The efficacy of antibiotics in such cases is overestimated as the decrease in the mortality and morbidity associated with respiratory infections is multifactorial; general improvements in socioeconomic factors such as housing and sanitation have likely improved the general prognosis of illness and the introduction of vaccination programs such as those for influenza and pneumococcal disease will have an impact on the number of severe cases.

The individual viral causes of respiratory illness will be discussed in detail later in this chapter, but in brief, viruses such as influenza and respiratory syncytial virus are well documented to occur in seasonal epidemics in the UK and are major contributors to both hospitalisation rates and excess winter mortality. There is now increasing evidence that viral respiratory infections themselves may be self limiting in childhood but correlates with the development of asthma later in life (Kusel, de Klerk et al. 2007). It is not yet clear if these infections are causal in the development of asthma or merely highlight already at risk individuals.

1.2 Clinical Presentation

Acute respiratory tract infection (ARTI) is a broad term which encompasses a multitude of clinical syndromes affecting the full length of the respiratory tract. These range from the benign common cold to life threatening conditions including bronchiolitis and pneumonia. Upper respiratory tract infection (URTI) refers to any infection involving the middle ear, nose, paranasal sinuses, pharynx and larynx whereas lower respiratory tract infections (LRTI) affect the trachea, bronchi, bronchioles and alveoli. The specific symptoms associated with infection will depend on the site affected. Involvement of the upper respiratory

tract will often produce inflammation, increased secretions, cough, with or without systemic signs of fever and loss of appetite. Involvement of the lower airways can produce more severe symptoms including wheeze, tachypnoea, breathlessness and signs of respiratory distress, along with signs of a systemic inflammatory response including fever and haemodynamic compromise manifesting as hypotension and tachycardia. Specific respiratory clinical syndromes that are frequently attributed to viruses include croup, inflammation of the larynx resulting in narrowing of the airway, and bronchiolitis, inflammation of the small airways. Many viral pathogens are associated with each presentation; an overview is presented in Table 1-1.

Virus	Colds	Pharyngitis	Tracheobronchitis	Croup	Bronchiolitis	Pneumonia		
						Children	Adults	Immunocompromised
Influenza								
Type A	+	++	+++	++	+	++	++++	+
Type B	+	++	++	+	+	+	++	+
Parainfluenza								
Type1	+	++	+	++++	+			+
Type 2	+	++	+	++	+			+
Type 3	+	++	+	++	++	+++	+	+
Respiratory syncytial virus	++	+	+	++	++++	++++	+	++
Rhinovirus	++++	++	+	+	+	+		
Coronavirus	++	+						
Adenovirus		++	+	++	++	++	++	++

Table 1-1. Frequency of viral respiratory syndromes caused by specific viral pathogens.

Adapted from Clinical Virology, Douglas D. Richman et al, 3rd Edition.

1.3 Aetiology of ARTI

ARTI are the commonest infections encountered by humans with the average adult suffering up to four episodes and children up to 11 episodes per year (Gruber, Keil et al. 2008).

Vast arrays of pathogens are capable of producing ARTIs including bacteria and fungi but the majority of episodes are attributed to viruses (Clark, Medina et al. 2014). The diagnosis of a respiratory tract infection is clinical but establishing the precise aetiology can be a challenging task. It is not possible to determine the causative agent on clinical symptoms alone due to the vast overlap between pathogens. Despite recent advances in diagnostic techniques, which will be discussed in more detail later in the chapter, a large proportion of diagnostic samples return negative results (Murdoch, Slow et al. 2012)

Environment factors affecting the aetiology of infection include temperature and humidity. Aetiology of infection is also influenced by multiple host factors such as age, comorbidities, immune status and possibly sex (Nickbakhsh, Thorburn et al. 2016), thus the cause of infection will vary temporally, geographically and within a population. It is challenging to determine the relative burden of each causative agent as many population based studies which have shaped our understanding of the condition were carried out prior to the discovery of a number of common pathogens and using less sensitive diagnostic methods than those employed today.

The common viral causes of respiratory illness will vary with age group, season and locations. There are some rarer pathogens in the general public that will have a higher incidence depending on patient risk factors such as immunocompromise.

The upper respiratory tract is colonised by multiple bacterial species, many of which are also potentially pathogenic such as *Streptococcus pneumonia*, *Staphyloccoccus aureus* and *Haemophilus influenzae*. Fungal infections are rare in immunocompetent individuals but both *Pneumocystis* and *Aspergillus* are associated with high mortality levels in certain populations. *Pneumocystis*

jirovecii remains one of the commonest opportunistic infections among those with HIV and despite the introduction of anti-retroviral drugs this still carries a mortality rate of 30 – 60 % (Lee, Park et al. 2015). The incidence of *Pneumocystis* in non-HIV cases is increasing though many are associated with immunocompromise relating to haematological malignancies or immunosuppressive therapies. *Aspergillus sp.* are common within the environment but also associated with a spectrum of clinical conditions (Kousha, Tadi et al. 2011). Infection is frequently associated with chronic lung conditions such as cystic fibrosis or asthma or immunocompromise following solid organ transplantation (Doligalski, Benedict et al. 2014; Munoz, Ceron et al. 2014). It should also be noted that repeated exposure can induce severe asthma and eventually result in chronic fibrotic changes in the lungs.

1.3.1 Viral Causes of ARTI

Of the viruses known to cause upper respiratory infections rhinoviruses are consistently found to be the most common (Makela, Puhakka et al. 1998; Ruohola, Waris et al. 2009; Zhang, Hu et al. 2012), followed by respiratory syncytial virus and coronaviruses. In the last 20 years many new pathogens have been described, including human metapneumovirus, human bocavirus and novel coronaviruses (van den Hoogen, de Jong et al. 2001; Falsey and Walsh 2003; Allander, Tammi et al. 2005) along with the pandemic influenza A, emerging in 2009. It is likely that some of these viruses have circulated for many years where as those such as MERS coronavirus occurred as a zoonotic disease, adapting to a new host and the novel influenza A resulted from reassortment of other viral strains. These are discussed in more detail below.

There are multiple viruses established as causes of respiratory infections in humans with many more capable of producing respiratory symptoms, even if this is not the primary infection syndrome. A number of novel viruses and subtypes have emerged in recent years with many causing primary respiratory infections and others proposed as respiratory pathogens but lacking in definitive evidence. The common causes of viral respiratory infections are outlined below.

1.3.2 Influenza Viruses

There are three genera within the *Orthomyxoviridae* family which are recognised as human pathogens, influenza A, B and C. These viruses are enveloped with a negative sense single stranded RNA genome comprising of seven or eight segments (Figure 1-1). The genera are determined by the antigenic structure of the nucleoprotein (NP) and Matrix (M) antigens and in the case of influenza A may be further subdivided based on the structure of the surface proteins.

1.3.2.1 Influenza A

The natural reservoir of influenza A is in wild birds but multiple species are known to be susceptible including horses, dogs and poultry. This wide host range contributes to the diversity of viruses. There are 18 known HA proteins but only 1, 2 and 3 are widely associated with human infection, with H5, H7 and H9 causing human infections in specific circumstances.

The genus of influenza A is subtyped based on the morphology of the surface proteins, haemagglutinin (HA) and neuraminidase (NA). The commonly circulating seasonal subtypes in humans are H3N2 and H1N1 with multiple strains within theses subtypes. The commonly accepted nomenclature of influenza A virus strains is natural host (which is omitted if human), geographical location of isolation, strain number and year of isolation, with the addition of HA and NA subtypes in parentheses e.g. A/Puerto Rico/8/34 (H1N1) (WHO 1980).

The virus attaches to the host cell by the HA binding to sialic acid on the cell surface. The virus then enters the cell through clathrin-mediated endocytosis. Upon cell entry the virus is contained within an acidified endosome. The acidic environment leads to a conformational change in the HA structure allowing it to associate with the endosomal membrane. The M2 ion channel allows influx of H^+ into the virus particle and both these processes mediate the release of the virus particle contents into the cell cytoplasm. The particle content consists of viral RNA bound to the nucleoprotein and the polymerase complex (PB2, PB1 and PA). The RNP and polymerase complex are then transported into the nucleus, where transcription can take place. Viral transcription is initiated through a process known as cap-snatching: following transcription to positive sense RNA the viral

NS1 binds to cellular mRNA transcripts, PA then cleaves a short portion from the host RNA to be added to the viral positive sense RNA. This process allows the viral RNA to be recognised and translated by host ribosomes.

NEP/NS2 and M1 mediate the export of newly synthesised viral RNP complexes from the nucleus to the cytoplasm. The virus particles are assembled and bud from the apical surface of the infected cells. The enzymatic activity of NA is essential in cleaving the sialic acid from the budding virus to release it from the cell surface.

The genome of influenza A also encodes several non-structural proteins by alternative splicing, including NS1, PB1-F2 and PA-X. NS1 inhibits the interferon response of the host, PB1-F2 is a pro-apoptotic protein and PA-X contributes to the pathogenicity of the virus.

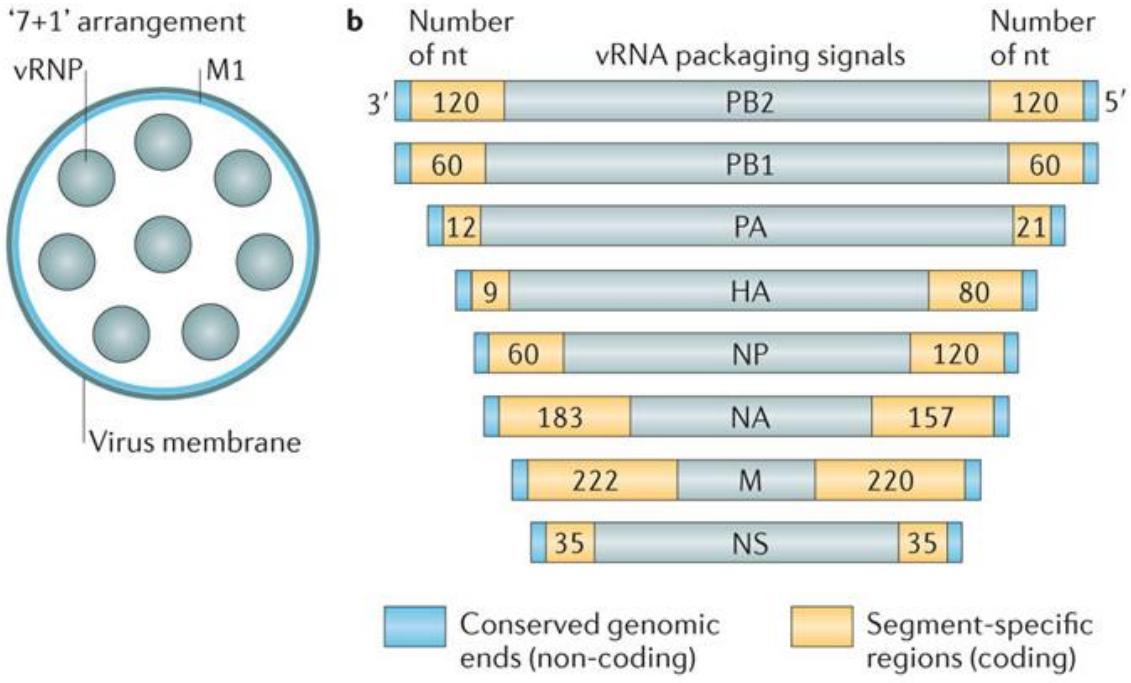


Figure 1-1. Influenza A virus genome.
Reproduced with permission (Eisfeld, Neumann et al. 2015)

1.3.2.2 Influenza B

The structure of Influenza B is essentially synonymous with that of influenza A and their appearances are indistinguishable on electron microscopy. The

nomenclature also follows similar rules to that of influenza A with the exception of host specification and HA/NA subtype.

Influenza B has a much narrower host range, infecting only humans and seals. This is thought to be the reason behind the lack of virus recombination and diversity; indeed there are only two influenza B lineages which have been in co-circulation since the 1980s.

These lineages, B/Yamagata and B/Victoria contain a number of subgroups. The majority of the B/Yamagata in circulation belong to Group 2 B/Brisbane/2/2007 and group 3 B/Bangladesh/3333/2007 and B/Victoria to B/Brisbane/60/2008.

1.3.2.3 Influenza C

The influenza C genome possesses only seven RNA segments, encoding nine proteins and differs from other influenza viruses in that it has a single surface protein, haemagglutinin-esterase fusion (HEF). Influenza C uses 9-O-acetyl-N-acetylneuraminic acid as a receptor rather than sialic acid. This is relevant to the treatment of influenza which will be discussed below.

Influenza C tends to result in a mild or asymptomatic infection, with many acquiring antibodies in childhood (Dykes, Cherry et al. 1980). It was accepted that only one subtype of the virus circulated and there was no animal reservoir, however a novel influenza C virus was recently isolated from pigs and cattle, showing only a 50% homology with those found in humans (Hause, Collin et al. 2014).

1.3.3 Influenza Epidemiology

Influenza is by far the most frequently studied of the respiratory viruses. This is likely a result of the wide host range and devastating pandemics caused by the virus. Influenza is present worldwide, occurring sporadically throughout the year in tropical regions whereas circulation is more predictable in temperate regions with infections peaking in winter months. Influenza is responsible for up to 5 million cases and 250,000 – 500,000 deaths per annum worldwide and is a major

contributor to excess winter mortality rates in the UK (Mann, Mangtani et al. 2013).

The host range of the viruses is determined by the affinity of the receptor to sialic acids. Most human influenza infections target the 2-6-linked sialic acid receptor which is mainly found on the epithelial cells of the upper airway. The 2-3-linked sialic acids, found in the lower airways of humans, are the preferred receptor of avian influenza viruses and proposed to be the reason why human infection with avian viruses results in a more severe infection involving the lower airways.

1.3.4 Clinical Presentation

Influenza viruses are spread by droplet transmission with an incubation period in the region of one to four days. Symptoms of an uncomplicated infection usually manifest as fever, myalgia, headache and photophobia, with respiratory symptoms of dry cough and sore throat. The infection usually resolves after seven days but may be associated with feelings of lethargy following the acute infection. Influenza C differs in that it associated with a mild upper respiratory illness in childhood.

1.3.4.1 Complicated Infection

The initial infection may become complicated by either a viral pneumonia or secondary bacterial pneumonia. Superimposed infection with other pathogens, such as *Staphylococcus aureus*, *Haemophilus influenzae* and *Aspergillus spp* are well documented (Adalja, Sappington et al. 2011).

The RNA dependent RNA polymerase has poor proof reading capabilities and as a result mutations occur frequently in virus progeny. These mutations accumulate over time, eventually affecting epitopes associated with the host immune response. When this occurs to the point where the host immune system can no longer recognise the virus then it will be seen as a novel infection. This process is referred to as antigenic drift. Alternatively a new infection can arise when a human virus undergoes reassortment with a virus associated with infection of another species, gaining a new segment of RNA. This is referred to as antigenic

shift and results in an abrupt change in the virus and as a result there is little residual immunity in the population. This can result in a pandemic outbreak of infection as a large number of individuals will be susceptible hosts.

1.3.5 Paramyxoviridae

The *Paramyxoviridae* are enveloped viruses with a single stranded negative sense RNA genome which is non-segmented and 13 – 15 kb in length. There are three major *Paramyxoviridae* groups associated with human respiratory tract infections, respiratory syncytial virus, the parainfluenzae viruses and human metapneumovirus, which are outlined below.

1.3.5.1 Respiratory Syncytial Virus

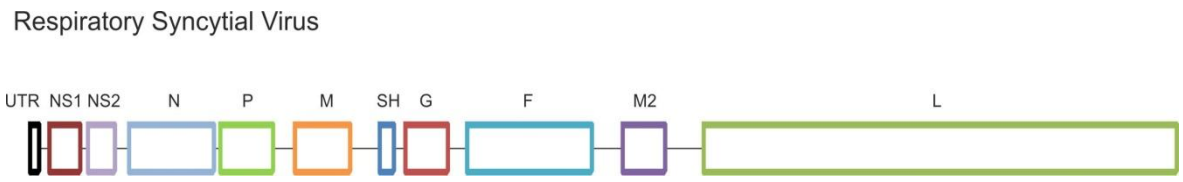


Figure 1-2. The respiratory syncytial virus genome.

1.3.5.1.1 Structural Proteins

A schematic of the RSV genome is presented in Figure 1-2. The major surface proteins on the RSV virion are the fusion (F), attachment glycoprotein (G) and small hydrophobic (SH) protein. The G protein mediates attachment of the virus to the cell receptor nucleolin (Tayyari, Marchant et al. 2011) whereas the F protein is responsible for fusion with the host cell membrane and cell entry. Expression of the F protein on the surface of the infected host cell initiates the formation of syncytia, fusion of surrounding cells to form giant multinucleated cells, from which the name is derived. The majority of paramyxoviruses require both F and G proteins to enter a host cell, though the F protein of RSV can do this independently of G, suggesting that F is the major contributor to cell entry. Both of these surface proteins elicit a neutralising antibody response indicative that both have a significant role in the development of infection.

The role of the SH protein is unclear. Infection is not attenuated in cell culture using recombinant RSV lacking SH but is attenuated in mice and chimpanzees. There is evidence that it may act as an ion channel and inhibit the interferon alpha signalling (Fuentes, Tran et al. 2007; Gan, Tan et al. 2012).

The virus core contains ribonucleoprotein complex, consisting of the viral RNA, nucleoprotein (N), phosphoprotein (P) and L. Binding of the N protein with RNA is essential for replication and transcription as it is required for recognition by the viral polymerase. Binding of P to the RNA-bound N protein opens the structure to allow the L protein to interact with the RNA template. Together with the P, N and M2 protein the L protein forms the viral RNA dependent RNA polymerase complex.

The M protein associates with the RNP complex and lipid envelope during the formation of viral particles and also acts as an inhibitor of cellular transcription in early infection and viral transcription prior to viral assembly (Ghildyal, Baulch-Brown et al. 2003).

The non-structural proteins NS1 and NS2 counteract the host immune response through the inhibition of the type 1 interferon pathway (Bossert and Conzelmann 2002). This in turn inhibits cell apoptosis, thus increasing the survival time of the infected cell and in turn increasing the yield of virus progeny.

1.3.5.1.2 Classification

RSV can be subdivided into two types, A and B, based on the serological differences. There is cross reactivity between the F antibodies of both subtypes A and B and despite the G protein being markedly more variable with only 53% homology between the two subgroups (Johnson, Spriggs et al. 1987) there are conserved regions between the two subtypes, therefore a degree of antibody cross reactivity. There are multiple genotypes described within the two groups, determined on the genetic sequence encoding the second hypervariable region of the G protein.

1.3.5.1.3 Clinical Presentation

RSV was first isolated in 1957 and has long been recognised as an important cause of respiratory illness in infants with nearly all children being infected with RSV by age 2. In the UK RSV accounts for around 20,000 hospital admissions per year and is the leading cause of hospitalisation in children with respiratory infections (Handforth, Friedland et al. 2000). Around 3% of infants under the age of 1 will require hospitalisation due to an RSV related illness (Deshpande and Northern 2003).

RSV will elicit both a cell mediated and humoral immune response during a primary infection but this is incomplete and can wane with time, leaving individuals at risk of future infections, even with the same strain. The exact mechanism behind this is uncertain. Although further infections may be less severe the elderly, immunocompromised and those with chronic lung conditions remain at risk of severe infection. It is now appreciated that RSV is a significant cause of mortality and morbidity amongst immunocompromised individuals and the elderly (Falsey, Hennessey et al. 2005).

1.3.5.2 Human Parainfluenza Viruses

Human parainfluenza viruses are divided into five species which are classified into the genera Rubulavirus (HPIV-2, HPIV-4A and HPIV-4B) and Respirovirus (HPIV-1 and HPIV-3) based on molecular and serological differences.

Figure 1-3. The Parainfluenza virus genome.
The genome structure of the parainfluenza viruses are similar however the nucleotide length of the protein coding regions varies between the different genera.

1.3.5.2.1 Structural Proteins and Non-Structural Proteins

The haemagglutinin neuraminidase (HN) surface glycoprotein has two major functions at the beginning and end of the infection cycle. HN binds to sialic acid on the surface of the host cell to initiate cell entry and the neuraminidase activity will release newly formed virus particles from the surface of the infected cells. Following HN binding with sialic acid the F protein aids fusion of between the virus and host cell. The matrix protein M forms a protective layer beneath the lipid envelope and is essential in the formation of the virus envelope and budding from an infected cell (Lawrence, Borg et al. 2004).

The nucleocapsid (N), phosphoprotein (P) and large (L) protein form the nucleocapsid core along with viral RNA.

A non-structural protein, V, is detected in the rubulaviruses (HPIV-2 and 4), coded by a +1 frameshift in the P reading frame. It has been shown to degrade STAT2, thus inhibiting the cellular interferon response (Nishio, Ohtsuka et al. 2008).

1.3.5.2.2 Clinical Presentation

The parainfluenzae viruses, despite their similarities, circulate with distinct temporal patterns. HPIV-1 is a common cause of croup and is also the most commonly reported of the parainfluenza group. HPIV-2 often affects young children, aged one to two years, and again can cause croup. It is noted to occur with biennial peaks in odd numbered years. HPIV-3 occurs annually in late spring or summer and is the commonest cause of hospitalisation amongst the parainfluenzae viruses. HPIV-4 is more commonly seen in autumn and winter but is rarely associated with severe disease.

1.3.5.3 Human Metapneumovirus

Human Metapneumovirus (hMPV) is classified in the genus Metapneumovirus and sub-family Pneumovirinae. The genome is ordered in a similar manner to that of the other paramyxoviruses but it lacks any non-structural proteins. The genome codes for nine proteins (Figure 1-4).

Figure 1-4. The human metapneumovirus genome.

The F, G and SH proteins are found on the surface. The M protein forms a protective layer beneath the lipid envelope and also coordinated virion assembly. The virion core contains the ribonucleoprotein complex, containing the viral genome and replication proteins N, P and L. M2-1 is a transcription factor which interacts with the RNP complex. M2-2 is involved in genome replication.

1.3.5.3.1 Classification

hMPV comprises of two main subtypes, each with minor subgroups, A1, A2, B1 and B2 based on serology and the genetic sequence of the surface protein F (Huck, Scharf et al. 2006). A further group, A3, has been proposed but further study is needed to confirm this (Escobar, Luchsinger et al. 2009).

1.3.5.3.2 Clinical Presentation

hMPV was first isolated from children in the Netherlands with upper respiratory infections in 2001, but retrospective serological studies suggest it has circulated for over 50 years and was probably of zoonotic origin from birds (Njenga, Lwamba et al. 2003). There is a worldwide distribution of hMPV and infections peak in winter months in temperate regions.

The incubation period of hMPV is about four to five days. Infection is initiated in the nasopharynx and can quickly spread to the lower respiratory tract and therefore hMPV is a common cause of lower airway infections such as bronchiolitis and pneumonia. Despite there being no described viraemic stage of infection there are documented cases of encephalitis attributed to hMPV (Schildgen, Glatzel et al. 2005).

Most children are infected by age five nevertheless immunity is transient, thus reinfections can occur at any age. As is often the case those with underlying lung disease, immunocompromised and the elderly are at risk of developing severe disease, with mortality rates of up to 50% in outbreaks in elderly care home facilities (Liao, Appelgate et al. 2012).

1.3.6 Picornaviridae

There are four species within the genus enterovirus, of the *Picornaviridae* family which are known to cause respiratory tract infections in humans, rhinovirus A, B, C and enterovirus D. The enteroviruses have a non-segmented single-stranded genome of positive sense which is approximately 7.3kb in length (Figure 1-5).

1.3.6.1 Human Rhinovirus

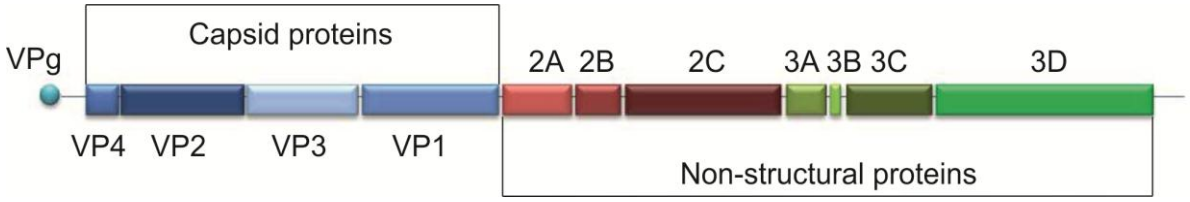


Figure 1-5. The rhinovirus genome.

1.3.6.1.1 Structural Proteins

Three distinct groups of HRV exist, A, B and C which are determined by the genetic sequence of the structural proteins VP1-4 which form the capsid of the virus particle (Jacobs, Lamson et al. 2013). VP1, 2 and 3 make up the outer capsid of the virus and are therefore responsible for the antigenic diversity. VP4 anchors the viral RNA in the core of the virus to the capsid. Multiple host cell receptors have been described for HRV, including ICAM-1, LDLR and heparin sulphate. The method of entry to the cell is dependent on the receptor used but once inside the cell a drop in the pH of the endosome results in viral uncoating and release of the RNA. RNA is then exported across the endosomal membrane into the cytosol where host ribosomes translate the positive sense RNA into a polyprotein.

The genome codes for two proteases, an RNA-dependent RNA polymerase (RdRp) and the VPg protein, all of which are involved in virus replication. The VPg protein is flanked to the 5' UTR and serves as a primer for genome replication (Palmenberg, Rathe et al. 2010). The RdRp replicates RNA genome and the proteases cleave translated polyproteins. The RdRp lacks any proof reading capabilities and therefore mutations occur frequently during HRV replication.

1.3.6.1.2 Classification

Multiple serotypes exist within each group and over 100 serotypes of HRV have been described. There is extensive genome variability between serotypes and this is most marked in the VP1 protein with up to 25% nucleotide variation between serotypes. The 5' UTR is relatively conserved between serotypes and groups and is the common target for molecular diagnostic tests, allowing detection of the virus but not differentiation between groups.

1.3.6.1.3 Clinical Presentation

The human rhinoviruses are often thought of as causing only a mild illness and indeed they are amongst the commonest causes of the common cold. They are now also acknowledged to be involved in severe infections of the elderly and immunocompromised and exacerbation of chronic lung diseases.

The optimal viral replication of hRV is 33ºC and replication of some serotypes is markedly reduced at higher temperatures. For this reason hRVs were thought to be a pathogen of the upper respiratory tract only, however further studies of various serotypes demonstrated only a minimal difference in replication at 37 (Papadopoulos, Sanderson et al. 1999) and infection of the lower airways has also been demonstrated in healthy volunteers by experimental inoculation (Gern, Galagan et al. 1997).

There are high rates of hRV found in patients with acute exacerbations of chronic obstructive pulmonary disease (COPD) in comparison with stable COPD patients, and the development of common cold symptoms was associated with a prolonged recovery in such patients (Seemungal, Donaldson et al. 2000; Seemungal, Harper-Owen et al. 2001). The most recently described group, HRV-

C, appears to cause more severe illness in comparison to HRV-A and B and is also found more frequently in exacerbations of asthma requiring hospitalisation (Miller, Edwards et al. 2009).

1.3.7 Adenoviridae

Adenoviruses belong to the family *Adenoviridae* and genus Mastadenovirus. They comprise of a double stranded DNA genome of 32kb in length (Figure 1-6). They are divided into subgroups A to G and are further classified into at least 52 serotypes. The first adenovirus was isolated from the adenoid tissue of individuals with respiratory infections though they are now known to cause a myriad of clinical infections in humans depending on the subgroup. Subgroups C and E are commonly associated with infections of the respiratory tract.

Multiple receptors have been proposed in the initial attachment of the AdV to the host cell, including the coxsackie-adenovirus receptor (CAR), sialic acid, heparin sulphate and CD46. Following attachment the virus enters the cell through endocytosis, mediated by cell surface integrins. Lysis of the endosome then releases the virions into the cytoplasm where the capsid is dismantled and the DNA is transferred to the nucleus.

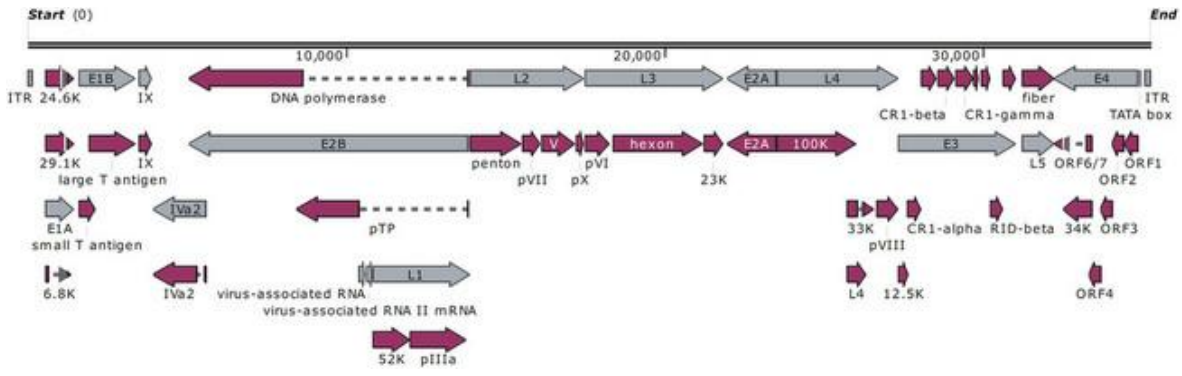


Figure 1-6. The adenovirus genome.

Adapted from Zhao, Wan et al. 2014.

Adenoviruses code for at least 38 genes from 17 transcriptional units. The infection can be divided into early (E) and late (L) phase depending on the genes expressed. Following infection the early transcription factors are activated which modify the host immune response and cell cycle to maximise viral

replication. Following initiation of viral DNA replication, the activation of the major late promoter (MLP) results in expression of late phase genes, including those coding for the structural proteins. Virions are then assembled in the nucleus of the infected cell and released by lysis (Lenaerts, De Clercq et al. 2008).

The clinical symptoms of an adenovirus infection of the respiratory tract include cough, coryza and nasal congestion along with systemic symptoms but it may also manifest as exudative tonsillitis. AdV infection can result in lower respiratory tract infections such as pneumonia with certain serotypes being associated with severe disease, notably Ad3, Ad4 and Ad7, outbreaks of which are associated with high rates of morbidity and rarely death.

AdV infection has long been noted as a major cause of morbidity amongst military personnel with outbreaks in training camps reported worldwide, leading to the development of a live vaccine against the commonly detected serotypes AdV 4 and 7. This vaccine was mandatory in US military training camps, with good effect, between 1971 and 1999 however due to manufacturing issues the vaccine was withdrawn. Subsequent surveillance suggested 15,000 cases of vaccine preventable cases annually in the US (Hoke and Snyder 2013).

In an immunocompromised individual the infection can persist, with detectable virus shed for months to years following infection. This is a particular problem in both bone marrow and solid organ transplant patients, as prior infection can reactivate causing clinical disease (Ison 2006).

1.3.8 Coronaviridae

Coronaviruses are of the family Coronaviridae and are enveloped viruses with a single-strand positive sense RNA genome of 27-33 kb (Figure 1-7), the longest of the RNA viruses. The name is derived from the large spike (S) surface proteins giving a crown appearance when visualised on EM.

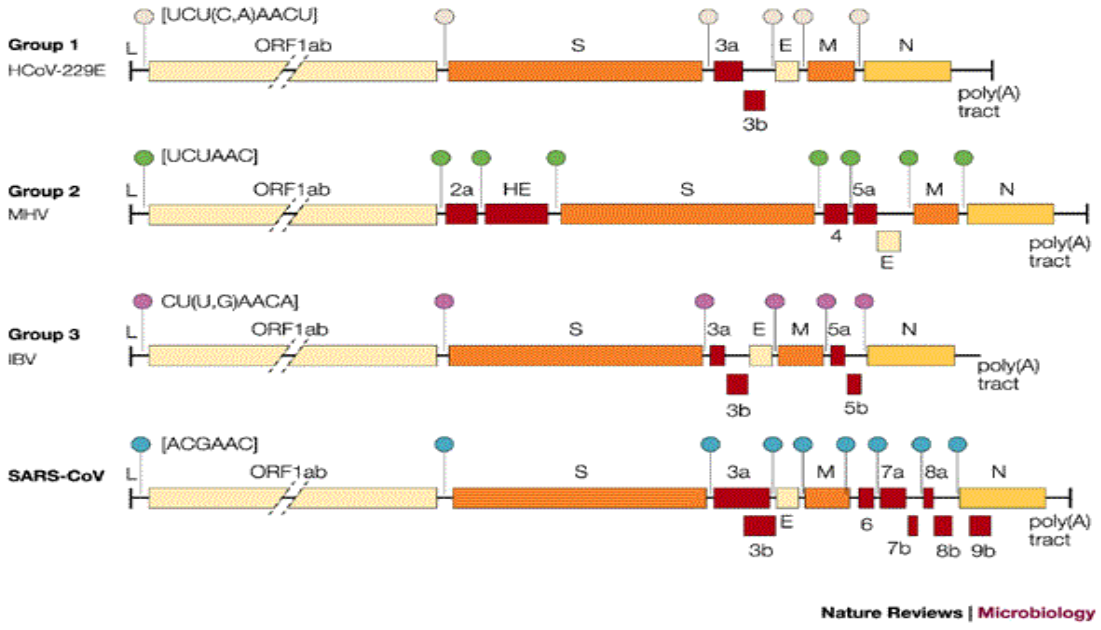


Figure 1-7. Coronavirus genomes.

Adapted, with permission (Stadler, Masignani et al. 2003).

The virus enters the host cell either by direct fusion with the membrane or receptor mediated endocytosis. Multiple receptors have been identified, including CD13, ACE-2 and sialic acid (Gierer, Bertram et al. 2013). This process is mediated by the spike surface protein (S) and haemagglutinin-esterase (HE). The viruses have two additional structural proteins, the membrane protein M which associates with the nucleocapsid and is responsible for the virus shape, and the envelope protein E which associates with the lipid envelope and together with the M protein is essential in the budding and release of virus progeny from the infected cell.

Upon cell entry the viral S protein is cleaved by endosomal acid proteases activating fusion with the endosome membrane and release of RNA into the cell cytoplasm. The positive sense RNA acts as mRNA to be translated into the replicase proteins. The replicase gene comprises of a 22kb portion of the genome with two open reading frames, 1a and 1b, coding the transcription-replication complex. Negative sense RNA is then synthesised as the template for progeny virus by RNA dependent RNA polymerase.

The number of accessory proteins differs between coronavirus species, with the greatest number being found in HCoV-SARS. The function of some of these

proteins remains unclear and whilst not essential for virus replication, there is attenuation of infection up on their deletion.

1.3.8.1 Classification

The coronaviruses are divided into four genera containing at least 29 species. The genera are based on multiple criteria including genetic sequence, natural host and serology. Two genera, alpha and betacoronaviruses, and six species cause human infections, namely HCoV 229E, HKU1, OC43, NL63, SARS along with the recently described MERS.

1.3.8.2 Clinical Presentation

The first human coronaviruses, HCoC-229E and HCoV-OC43, were isolated in the 1960s and were associated with a common-cold like illness. The coronaviruses were thought to only cause either a mild respiratory or gastrointestinal illnesses until the emergence of the SARS coronavirus (HCoV-SARS) in 2003 (Fehr and Perlman 2015). This completely novel coronavirus was of zoonotic origin from bats with the Himalayan palm civet as an intermediate host. HCoV-SARS quickly spread worldwide with 8098 cases reported and 774 deaths. The mortality rate varied within the population and was as high as 55% in some groups with the elderly being at greatest risk of severe disease (Alghamdi, Hussain et al. 2014). Interestingly children were not severely affected by the infection which differs significantly from the burden associated with many other viral respiratory infections though the mechanisms responsible for this pattern are not well understood. This virus no longer circulates in humans.

At least three other coronaviruses have been described since the HCoV-SARS outbreak, namely HKU-1, NL63 and MERS. The origin of MERS is again likely zoonotic, originating in bats with camels having been proposed as the intermediate host (Badawi and Ryoo 2016). In 2012 the Middle East Respiratory Syndrome coronavirus (MERS-CoV) was isolated from a patient in Saudi Arabia. Within months cases had been reported in multiple countries although all had a history of travel to or contact with the region. A significant healthcare associated outbreak occurred in South Korea, 2015, resulting in 186 cases (Ki

2015). This episode highlights the need for effective detection and isolation of cases to reduce spread within healthcare settings and the community. The latest WHO update in May 2016 confirmed 1,733 notified cases and at least 628 deaths from 27 countries (WHO 2015). This virus continues to circulate in the Middle East.

In comparison with these highly pathogenic viruses, the remaining coronaviruses are associated with a range of clinical illnesses including both upper and lower respiratory infections but are generally considered non-severe. There are of course exceptions and severe infections may occur, generally at the extremes of age or in those with underlying immunocompromise.

The HCoV-HKU1, NL63 and OC43 continue to circulate seasonally in humans, with a peak incidence in winter months in temperate regions. HCoV-229E has a less distinct seasonal pattern and can be detected sporadically throughout the year (Gaunt, Hardie et al. 2010). There is a suggestion of alternating dominance of strains each year although a longer duration of study would be required to confirm this.

1.3.9 Human Bocavirus

The human bocavirus is a member of the Parvoviridae. It is non-enveloped and has a single stranded DNA genome of 5.3kb in length. HBoV was first identified in 2005 (Allander, Tammi et al. 2005) but the role in clinical disease has since been debated. Seroepidemiological studies suggest up to 20% of the population have been exposed (Li, He et al. 2015). And while bocavirus is frequently identified alongside a further respiratory pathogen (Sloots, McErlean et al. 2006) the virus has been detected as a single agent from symptomatic children (Kahn 2008).

1.3.10 Polyomaviruses

The WUV and KIPyV are both newly discovered polyomaviruses identified from human respiratory samples using high throughput screening methods. Their pathogenicity is also still debated. Their presence has been confirmed in many countries, with both serological and DNA evidence. However, the carriage in asymptomatic individuals has not been fully explored. In the United Kingdom a

small study reported the carriage rates of WUV and KIPyV to be similar in a symptomatic and asymptomatic control group (Norja, Ubillos et al. 2007). It should also be noted that in a large proportion of symptomatic individuals a further virus was detected, most commonly rhinovirus (Gaynor, Nissen et al. 2007).

1.4 Management of Viral Acute Respiratory Infections

A small number of anti-viral drugs are currently available for use in respiratory infections but, in general, specific drugs are limited to influenza infections. The treatment of any other virus has little in the way of clinical evidence to support the use. This small number of drugs is dwarfed by the relative abundance of antibacterial agents licensed for use in bacterial respiratory infections.

There are many challenges which have hampered the development of specific treatment and prevention strategies for viral infections. Vaccine development has thus far proved unsuccessful, likely due to the vast number of aetiological viruses compounded by their extensive antigenic variability (Simancas-Racines, Guerra et al. 2013).

Trials in the prevention of the common cold using compounds such as Vitamin D, Vitamin C or Zinc have also yielded little success (Murdoch, Slow et al. 2012; Hemila and Chalker 2013; Singh and Das 2013).

Presently management options of these infections are limited to symptomatic relief. Antibiotics are not recommended in the treatment as there has been no clinical benefit demonstrated, nevertheless over 30% of all antibiotics prescribed by ambulatory care physicians in the United States are associated with common colds, upper respiratory tract infections and bronchitis and similar findings have been documented worldwide (Gonzales, Steiner et al. 1997; Higashi and Fukuhara 2009).

This is a major contributor to the global issue of inappropriate antibiotic use. This is a cause of great concern, both due to the economic costs – antibiotics prescribed in the US for upper respiratory tract infections cost $227 million each

year (Bertino 2002) – and the association with the rise in antibiotic resistant bacteria (Gonzales, Bartlett et al. 2001). For example, *Streptococcus pneumoniae* is a member of the commensal flora but is also a leading cause of bacterial meningitis, community acquired pneumonia and otitis media. Prior to 1967 all strains of the bacteria were fully sensitive to penicillin but penicillin-resistant as well as multidrug-resistant strains have emerged which is thought to be linked to antibiotic exposure. Many studies have demonstrated that resistance patterns vary geographically and positively correlate with the level of antibiotic consumption; conversely, rationalising antibiotic prescribing is associated with a reduction in drug-resistant strains (Linares, Ardanuy et al. 2010).

This pattern has heightened the realisation that we have two options, develop new methods of combating infections (Riley, Robinson et al. 2012) or rationalise antibiotic prescribing. An essential component of rationalising antibiotic prescribing is to improve diagnostic testing of respiratory tract infections in order to target the use of therapies currently available (Pavia 2011). Characterising the viruses which are present during these episodes will be an important step in this process. Epidemiological studies to determine the major viruses contributing to the burden of respiratory infection will aid in focusing vaccine development as well as specific drug therapies.

Current management strategies, treatments and proposed treatments are discussed below.

1.4.1 Anti-viral Drugs

1.4.1.1 Adamantane derivatives

Some of the first commonly used anti-viral drugs for respiratory infections were the adamantane derivatives amantadine and rimantidine. Their mechanism of action is to block the influenza M2 ion channel, inhibiting viral uncoating and release of RNA into the cell. The M2 proteins of influenza A and B show very little homology and as a result influenza B is inherently resistant to adamantanes. These drugs were used for many years in both the prevention and treatment of

influenza infections with very few cases of resistance noted. There was however a spike in 2000 in the circulation of resistant strains, both seasonal H3N2 and avian influenza A H5N1 in Asia and later in North America. Resistance was conferred by a single amino acid substitution in the M2 protein, S31N however a further five mutations that confer resistance have been described. High rates of resistance are now documented in H1N1 and H3N2 isolates worldwide (Dong, Peng et al. 2015). For this reason these drugs are no longer recommended in treatment or prophylaxis.

1.4.1.2 Neuraminidase Inhibitors

Drugs of the class neuraminidase inhibitors act on influenza A and B by blocking the function of the viral neuraminidase thus inhibiting viral budding and release from an infected cell. Resistance has also been noted to this class of drugs; the H275Y mutation in the NA segment is widely associated with oseltamivir resistance. Initially this was not concerning as the mutation also negatively affected virulence and infectivity of the virus (Carr, Ives et al. 2002) but a resistant strain has emerged, A/Brisbane/59/07(H1N1), which retained virulence and the ability to transmit from person to person.

The benefit of these therapies has been a subject of debate in recent years as a Cochrane review found that the duration of symptoms is reduced by less than half a day and there was no evidence to support reduction in complications or hospitalisation associated with influenza (Jefferson, Jones et al. 2014).

1.4.1.3 Nucleoside/nucleotide analogues

Ribavarin is a nucleoside analogue which has a mutagenic effect when incorporated into the viral genome. It is licensed for use in humans and has been shown to inhibit viral replication of coronaviruses, influenza, RSV and adenoviruses. Ribavirin has not been used routinely in the treatment of coronaviruses however with the emergence of highly pathogenic viruses such as SARS and MERS there are ongoing trials into the use of interferon-α2b in combination with ribavirin (Falzarano, de Wit et al. 2013). The results in animal studies are promising but there are only a small number of case reports in

humans (Khalid, Al Rabiah et al. 2015). Toxicity is limiting factor as the drug can cause both leukocytopenia and haemolysis.

Cidofovir is a monophosphate nucleotide analogue which has a broad spectrum of activity against DNA viruses. The diphosphate metabolite of cidfovir is competitively incorporated into DNA in place of the nucleotide cytosine resulting in chain termination. Cidofovir is recommended as a first line therapy in the treatment of adenovirus infections in immunocompromised individuals such as those undergoing stem cell or solid organ transplantations and a recent study suggests clinical benefit in immunocompetent individuals with pneumonia (Kim, Kim et al. 2015). The drug is only available as an intravenous preparation and it is associated with potentially significant side effects including nephrotoxicity. Brincidofovir, a prodrug of cidofovir, is conjugated to a lipid and therefore associated with higher intracellular levels of drug. This results in lower plasma levels of the drug which in turn reduces nephrotoxicity.

1.4.2 Symptomatic Management

Where no specific prevention measure or drug treatments are available individuals and clinicians are limited to recommending symptomatic management. There are many options available as part of this multi-billion pound economy. Pharmacological agents such as anti-inflammatories, mucolytics and decongestants are readily available over the counter but the small number of varied studies carried out on such measures makes it difficult to corroborate any beneficial role. In the vast majority of cases it is unlikely that these measures are associated with harm (Smith, Schroeder et al. 2014). The exception to this is centrally acting antitussives where the active ingredient is derived from opiates, associated with addiction and respiratory suppression. Additional simple measures such as vapour and maintaining hydration are seen as helpful but again there is little evidence available to corroborate any positive role in disease outcome (Singh 2013).

1.4.3 Vaccines and Immunisations

An important strategy in the management of viral respiratory infections is that of prevention. There is a great need for vaccines against viral respiratory pathogens particularly in groups at risk of severe disease. However there are multiple barriers to the development of effective viral vaccinations. As discussed previously there are multiple virus groups which cause acute respiratory infections and within each group there are several, up to hundreds of serotypes. For a vaccine to be effective it must target a conserved yet neutralising epitope of the pathogen.

Some viruses, for example RSV, present the greatest risk to young infants who still possess maternal antibodies, interfering with infants' own ability to mount a vaccine response. Antibodies generated at this age can be transient with a lower affinity compared with those generated in later life (Siegrist 2000). At the other extreme of age, the elderly have a marked reduction in the immune response to infection and altered reaction to vaccination. Aside from these host factors, the number of viruses and array of serotypes within each group of viruses responsible for respiratory illnesses makes it difficult to develop an effective vaccination. The large number of serotypes is often a representation of the viral mutations which occur over time as an effective escape mechanism from the host immune system.

As a result of these barriers, few viral vaccines exist with the exception of an influenza vaccine. Multiple vaccine preparations are available with the most commonly used vaccine in the United Kingdom being the trivalent injection. The 2015 season vaccine contained two IFA strains (A/Switzerland/9715293/2013 (H3N2) and A/California/7/2009 (H1N1)pdm09) and one IFB strain (B/Phuket/3073/2013). The trivalent injection preparations contain inactivated viruses but a newer intranasal Quardravalent preparation with live attenuated viruses is licensed and contained the same strains as the trivalent vaccine along with an additional IFB strain (B/Brisbane/60/2008) (WHO 2016). In the United Kingdom the influenza vaccine is recommended to all people over the age of 65 years and under 5 years, pregnant females and those with chronic medical conditions such as diabetes, cardiovascular disease or chronic lung disease, as

these individuals may be more severely affected by the infection. The vaccine contains two strains of influenza A and one strain of influenza B (influenza C rarely circulates and is associated with mild illness and is therefore not included). The vaccine is modified regularly to contain the strains most likely to circulate during the standard influenza epidemic. This choice is based on surveillance information from previous seasons along with modelling data and must be modified frequently as the viruses are prone to antigenic shift and drift which will impact on vaccine effectiveness. No vaccine is associated with 100% efficacy and therefore not all cases will be prevented but on average around 50% of cases are prevented through vaccination programs.

The group at greatest risk from severe RSV infections is infants, due to their degree of immunosuppression related to circulating maternal antibodies and immature immune systems. A passive immunisation against RSV is licensed for use, the humanised monoclonal antibody palivizumab. Its use remains limited to a small and specific population of at risk individuals, in part due to the cost of the preparation and the short half-life, requiring frequent dosing during an RSV outbreak. It is administered to premature infants but is neither practical nor cost effective to be used in the general population.

A live vaccine is available against adenovirus types 4 and 7 which have been responsible for outbreaks associated with high levels of morbidity in specific areas, particularly those with close living quarters such as military recruits, holiday camps and healthcare settings. The vaccine was routinely administered to military recruits during basic training in the United States which was associated with a reduction in the number of cases. It has been hypothesised that the reduction may in part be due to falling numbers of recruits as similar patterns had been noted in countries not administering the vaccine. Due to manufacturing issues and a lack of supplies the vaccination programme ceased in 1999. Surveillance of US military training camps suggested there were around 15,000 vaccine preventable cases of AdV infection per year, therefore the program was reinstated in 2011 with a novel vaccine (Hoke and Snyder 2013).

1.5 Diagnosis of Infection

The idea of an external cause of illness dates back many centuries, however our understanding of infectious disease has evolved with the ability to magnify and visualise the minute. Hippocrates documented outbreaks of disease in great detail in the fifth century BC but he was limited to describing what could be observed with the naked eye such as location, habit and appearance of individuals.

"Whoever wishes to investigate medicine properly, should proceed thus: in the first place to consider the seasons of the year, and what effects each of them produces for they are not at all alike, but differ much from themselves in regard to their changes. Then the winds, the hot and the cold, especially such as are common to all countries, and then such as are peculiar to each locality. We must also consider the qualities of the waters, for as they differ from one another in taste and weight, so also do they differ much in their qualities."

On Airs, Waters And Places
Hippocrates (translated by Francis Adams)

While Hippocrates is undoubtedly the father of epidemiology it is only with the technological advances of recent centuries that the aetiology of many diseases could be explored.

1.5.1 Historical aspect of viral diagnostics

The methods used to diagnose infections have run in parallel with the ability to visualise microorganisms and their effects. Robert Hooke was the first to publish a description of microorganisms in his 1665 seminal work Micrographia where he used light microscopy to give a detailed account of microfungi. Within a few years Antonie van Leeuwenhoek had used similar methods to describe the appearance of bacteria and protozoa. It would be many years later, in 1876, that

Robert Koch would demonstrate the link between such organisms and disease through his work with *Bacillus anthracis* as the cause of anthrax. Following on from this initial discovery many bacteria were identified and characterised as pathogens but it was acknowledged that there was a yet unidentified entity capable of causing disease. The work of Adolf Mayer and Dmitri Ivanovsky demonstrated a non-bacterial pathogen which was able to replicate and transmit disease between tobacco plants. The experiment was reproduce by Martinus Beijerinck, who coined the term “virus”, derived from the Latin meaning “slimy, poisonous liquid” in reference to the filtered plant sap which transmitted disease. Due to the small size of virus particles they remained invisible until the 1930s when they were unveiled by electron microscopy; however the effects of viral infections could be demonstrated much earlier than this using experimental animals and later cultured cells. As this can be used as a diagnostic tool, the method will be discussed in more detail in the following section.

1.5.2 Diagnosis of Viral Infections

1.5.2.1 Case definitions

Perhaps the most basic way of diagnosing an infection is based on the clinical syndrome with which the patient presents. This is a particularly useful tool in resource poor settings where laboratory tests may not be accessible or affordable. A standardised definition of symptoms used across many countries will also allow a comprehensive geographical and chronological comparison. This method is often employed in surveillance programmes where the goal is not to diagnose an individual, rather to look for trends in illness presentation on a larger scale. The commonest utilisation of case definitions is in the context of influenza-like illness (ILI) and severe acute respiratory infection (SARI) surveillance. There are no pathognomonic features of infection which distinguish influenza from other respiratory pathogens therefore often a subset of patients will also have diagnostic tests carried out to confirm aetiology of infection. Case definitions can also be used to triage those at greatest risk of having a disease. This method was used during the recent outbreak of the ebola virus in West Africa. Using the clinical symptoms and the history of presentation, patients can be identified and, where appropriate, isolated while awaiting laboratory

confirmation. This use of triage and patient cohorting aims to reduce the spread of infection (WHO 2010).

1.5.2.2 Culture and Virus Isolation

It has been possible to maintain living tissues and cells in a laboratory setting for many decades but advances made in the 1940s and 50s brought the ability to grow viruses in cultured cells. Henry Eagle described the nutrients required to maintain living cells for an extended period (Eagle 1955) thus allowing the propagation of virus. One of the first uses for this technique was growing purified polio virus to be used in the mass production of vaccines.

For many years this technique was the workhorse of virus diagnostic laboratories. The presence of a viral pathogen can be confirmed by directly inoculating a healthy monolayer of cells with a clinical sample from a symptomatic individual. If the virus is present in the sample and an appropriate susceptible and permissive cell line is used, the virus will replicate. The cells can then be observed for signs of damage due to infection – cytopathic effect (CPE). This process may take days to weeks depending on the virus in question, if indeed it develops at all. Although it can inform an experienced individual of the presence of an infectious agent, one must rely on previously documented patterns of cell damage. This is of course not possible when dealing with novel pathogens, though in such cases the presence of CPE in the absence of known pathogens can indicate a yet unidentified pathogen. Today this technique is used infrequently in routine diagnostics but its value must not be forgotten; it was the process used to initially identify the novel SARS corona virus during the outbreak in 2003.

1.5.2.3 Direct Detection Assays

1.5.2.3.1 Direct and Indirect Immunofluorescence

Using direct immunofluorescence (DIF), also known as direct fluorescence antibody testing (DFA), antigens can be detected in clinical samples by using antisera to a specific protein or epitope. Using the direct method, antisera containing antibodies coupled with a fluorescent dye which will bind to the

antigen if present, are combined with a clinical sample. The reaction is thoroughly washed to remove any unbound antibody and mounted on a slide. If antigen bound antibody remains it will fluoresce when viewed with a fluorescent microscope (Figure 1-8).

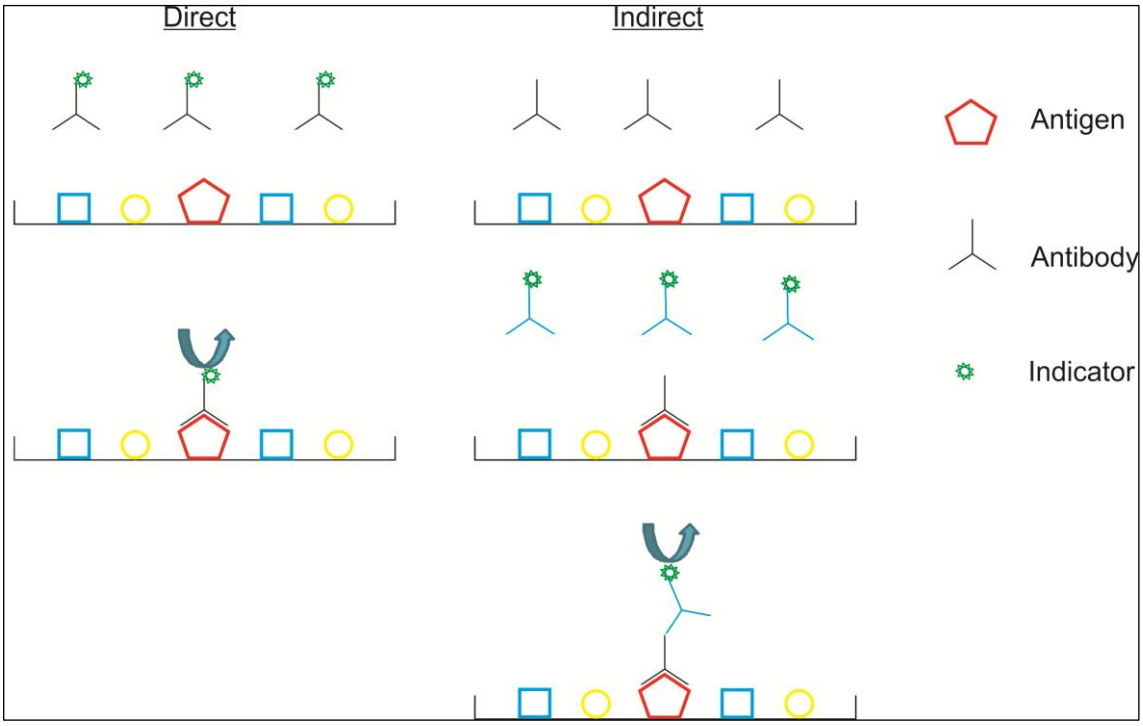


Figure 1-8. Direct and indirect immunofluorescence.
This shows a schematic representation of the direct and indirect immunofluorescence methods. The arrows represent a fluorescent indicator which can be detected using microscopy.

The indirect method is very similar to that outlined above, with the exception of the virus specific antibody is not coupled with the fluorescent dye, rather a secondary antibody directed at the virus specific primary antibody is labelled. This method is more time consuming as two incubation periods are required but only a single labelled antiserum is required to detect multiple antigens, in comparison to the direct method where a labelled antiserum is required for each antigen.

Regardless of which method is used, the reactions must first be optimised to find the antibody concentration which results in a sensitive test whilst reducing the background fluorescence to an acceptable level (Odell and Cook 2013).

1.5.3 Serological Assays

In these methods the study of patient serum can detect current infection (antigen) or inform of an immune response mounted to recent or previous infection (antibody).

1.5.3.1 Complement Fixation and Haemagglutination

In this process the naturally occurring complement proteins are removed from the patient serum, usually through heat treatment which will degrade the complement proteins but preserve antibodies. The serum is then supplemented with a standardised quantity of complement proteins, red blood cells and the antigen under study. If antibodies to the antigen are present the antibody-antigen complexes will consume the complement proteins but if no antibodies are present the complement will be free to lyse the red blood cells, turning the reaction pink. If antibodies are detected the levels can be quantified through serial dilutions of patient serum, to determine an antibody titre. This can be used to determine if an immune response has been mounted, in that the antibody titre will increase over the time of the infection.

Agglutination of red blood cells can be exploited in a similar fashion if the virus is capable of producing haemagglutination to determine virus and antibody quantity. If a sample contains virus, when it is added to red blood cells (typically chicken or turkey) the virus and red blood cells will agglutinate and form a diffuse lattice structure. If the virus is not present the red blood cells will settle in a dot or clump at the bottom of the reaction well. Dilutions of the initial sample are used to determine a quantity. Inhibition of red blood cell agglutination in the presence of virus (influenza) and patient serum confirms the presence of neutralising antibodies in the serum. Again serial dilutions of this process can be used to quantify antibody levels.

1.5.3.2 Enzyme-linked Immunosorbent Assay

The Enzyme-Linked Immunosorbent Assay (ELISA) has similar applications to indirect immunofluorescence. Antibody is bound and immobilised on a surface. The clinical specimen is then added and if antigen is present within the

specimen it will bind to the antibodies. A second antibody is then added to the reaction. This may be labelled with an enzyme (direct), typically horseradish peroxidise (HRP) or alkaline phosphatise (AP), or not labelled (indirect). Any unbound antibody is washed from the reaction and in the direct reaction a substrate is then washed over the reaction. In the presence of labelled antibody the substrate reacts with the enzyme resulting in a colour change which can be read by photometry. As with indirect immunofluorescence, in the indirect reaction an unlabelled antibody is washed over the reaction followed by a labelled antibody directed against the previous. Again a substrate is added and colour change detected (Figure 1-9).

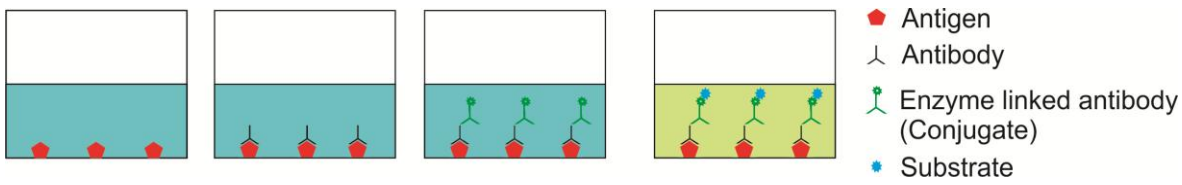


Figure 1-9. Enzyme-linked Immunosorbent assay.
A graphical representation of an enzyme-link immunosorbent assay. The presence of antibody to the target under study results in a reaction colour change.

The detection of antibodies against a pathogen in the serum of a symptomatic individual can inform if they have mounted an immune response to a pathogen. A rise in antibody titres in paired serum samples taken three to four weeks apart (as this process takes time to develop) indicates recent infection. The production of antibodies is not only a process which takes times but also one which can be affected by many inherited and acquired conditions such as primary immune deficiencies, haematological malignancies or infections e.g. HIV (Lazzarotto, Dal Monte et al. 1992). In people affected by these conditions serology results may be unreliable or difficult to interpret. Autoimmune conditions such as rheumatoid arthritis or systemic lupus erythematosus have been associated with false positive results due to non-specific antibody cross-reactivity (Salonen, Vaheri et al. 1980).

Many of the viral pathogens involved in respiratory tract infections are not commonly associated with a viraemic stage therefore serological testing is limited to providing a retrospective diagnosis as antigen is unlikely to be present. In general serological assays inform of past URTIs and would not guide

treatment options as infections will have resolved by the time the second convalescent sample is collected. Serology does provide important epidemiological data which can influence patient management, a good example being strain selection for influenza vaccines.

1.5.4 Molecular Diagnostic Methods

1.5.4.1 Nucleic Acid Amplification Testing

The first account of deoxyribonucleic acid, DNA, was in the 1860s, when Friedrich Miescher identified a new substance which behaved like no known protein or carbohydrate. He called this “nuclein” which Oskar Hertwig determined was responsible for hereditary characteristics and would later be known as DNA. Many decades later Phoebus Levene went on to describe the structure of individual bases which make up DNA with the overall double helix structure being resolved in 1953 by Watson and Crick, the names most frequently associated with DNA and the recipients of the Nobel prize for their work (Watson and Crick 1953).

The recognition that DNA, the building blocks of life, is present in all living organisms revealed a further target in the detection of pathogens.

The introduction of culture-independent molecular methods to the diagnostic laboratory revolutionised routine diagnostics. Not only has the turn-around time been reduced dramatically but as the viral genome is the target the tests are both sensitive and specific. For these reasons molecular testing constitutes the majority of work in many diagnostic laboratories. There is however a need for *a priori* knowledge of the pathogen genome. Diagnosing a viral infection with these methods requires the pathogen to be present in the clinical specimen, so appropriate and effective sampling must be carried out within the period of detectable viral shedding.

1.5.4.1.1 Polymerase Chain Reaction

Nucleic acids must be extracted from the cells or particles within a given sample. Oligonucleotide primers which complement the target genome, one

against each strand, are added to the extracted nucleic acids along with a thermostable DNA polymerase. Heating of the mix melts any secondary structures of the DNA to expose the area complementary to the oligonucleotide primer. Primer binding acts as a starting point for the DNA polymerase enzyme which can then extend a complementary sequence in the 5' and 3' direction. Multiple cycles of this process are carried out, resulting in an exponential increase in target genome copies. These can then be visualised by gel electrophoresis. An additional reverse transcription step to generate cDNA from RNA allows amplification of RNA genomes.

1.5.4.1.2 Nested Polymerase Chain Reaction

Nested PCR is essentially the same process as two standard PCRs. The primers used in the first reaction are designed against a target and the product of this is used as the template in a subsequent reaction. The primers used in the second reaction are designed against a target within the first target. The specificity of PCR amplification is dependent on primer binding and it is possible for a primer set to bind at more than one locus on a template genome. To increase the assay specificity this second set of primers which target the product of the first reaction are added, thus if the wrong locus is amplified during the first reaction it is extremely unlikely that this unwanted product will be further amplified in the second reaction.

1.5.4.1.3 Real-Time Polymerase Chain Reaction

The PCR process can be modified to allow real-time detection and quantification of the target. The addition of a fluorescent dye which only binds to double stranded DNA, i.e. the target, means as the target increases the fluorescence intensity will in turn increase. Alternatively a third oligonucleotide probe can be added to the reaction. This probe is specific to the target and labelled with a fluorophore and quencher. When the fluorophore and quencher are in close proximity the fluorescence cannot be detected. However when the probe binds to the target the exonuclease activity of the polymerase degrades the probe, releasing the quencher and thus exposing the fluorescence which can be detected following excitation by laser.

Using both PCR and real-time PCR it is possible to detect several targets in a single reaction by adding multiple primer sets. This is referred to as a multiplex PCR. It must be considered that there is a limit to the number of primer sets which can be added, usually up to four, before the sensitivity of the assay is affected in comparison with the individual assay.

Testing samples with standardised quantities of DNA in parallel with the unknown samples can be used to determine the quantity of the unknown samples. A standard curve can be generated from results of samples with known DNA quantities; this is used as a reference by which the DNA quantity in unknown samples can be extrapolated.

1.5.4.2 Microarrays

Microarrays (also known as DNA chips or biochips) utilise a collection of oligonucleotide probes about 70 bases in length, immobilised on a solid surface. The probes are complementary portions of DNA or RNA, designed against conserved regions of a genome or gene and thus if the target is present within a given sample it will bind to the corresponding probe. Probe binding is then quantified using a fluorophore or chemilluminescent reaction. Multiple probes can be attached to a single surface, allowing for screening of a large number of pathogens in a single reaction. As probes are targeted against conserved regions of the pathogen genome they may also detect related but novel pathogens.

1.5.5 Point-of-Care Testing

All of the above methods will be carried out in a specialist laboratory but point-of-care tests (POCT) are being introduced, so called as they can be performed in an area of patient care such as an emergency department or intensive care unit, in community or even in the home. The introduction of the Cepheid GeneXpert® has brought PCR into the clinical area. This portable, self-contained device utilises pre-packed cartridges of reagents therefore does not require extensive training or additional equipment. As a result, tests can be offered in the clinical areas, community or even in the field (Jenson, Dize et al. 2013).

Many POCT are available using the methods described above e.g. ELISA and haemagglutination however the majority of the POCT currently available are based on lateral flow immunochromatography (LFI) (Gubbins, Klepser et al. 2014). Using this method, a conjugate e.g. antibody, is bound to a porous surface. The liquid sample migrates across the porous surface and releases the conjugate. This mix migrates further to a capture molecule which will bind the analyte-conjugate complex, often resulting in a colour change to indicate a positive result. A second non-specific capture molecule is often included as a positive control to indicate sample migration was adequate (Figure 1-10).

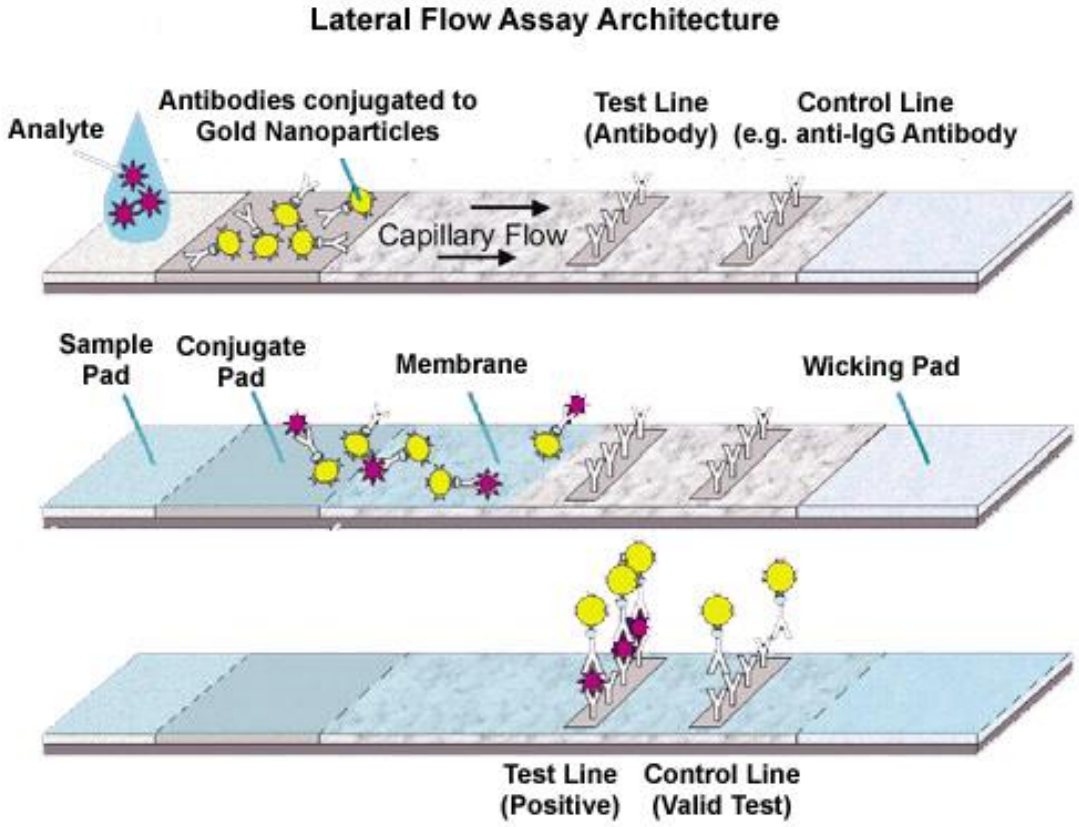


Figure 1-10. Lateral flow immunochromatography.
(Adapted from http://www.cytodiagnostics.com/store/pc/Lateral-Flow-Immunoassays-d6.htm)

A positive result with be indicated by a colour change which can be visualised by the user, although this can vary depending on the user. These rapid diagnostic assays will often provide results that will facilitate patient management, be that initiation of appropriate treatment or cohort care based on positive results i.e.

those with the same condition are treated in the same space to limit further spread of infection. Such tests could also help avoid unnecessary hospital admission and treatment. In the case of febrile illness in malaria endemic regions, such tests result in more accurate diagnosis, therefore a reduction in inappropriate anti-malarial treatment use (Ameyaw, Nguah et al. 2014).

Access to such tests could aid in detection of diseases in populations where it can be difficult to engage with medical services such as developing countries or rural areas far from a central diagnostic laboratory. Results are available rapidly and do not necessarily require healthcare trained staff providing training and on-going performance evaluation is carried out (Drancourt, Michel-Lepage et al. 2016).

1.5.6 DNA Sequencing

Following the discovery of DNA and its gross structure it then seems a natural progression to determine the finer details and patterns of the structure. This information can provide great insight into the pathogens responsible for infection. This was successfully achieved independently by two groups using very different methods as outlined below.

DNA sequencing is not routinely used as a first line diagnostic test but is used to further characterise a pathogen or determine virulence/resistance patterns. The commonly used sequencing methods are discussed below followed by an overview of the Next Generation Sequencing techniques which are currently available.

1.5.6.1 Maxam-Gilbert Sequencing

This method uses four chemical cleavage reactions to split DNA at specific bases or pairs of bases (Maxam and Gilbert 1977) (Figure 1-11). The DNA under study is cleaved using a restriction enzyme and a radioactive label is placed at the 5’ end of the fragment. The labelled DNA fragment is separated using gel electrophoresis and divided into four separate chemical reactions. Each reaction cleaves DNA at a specific base or pair of bases (e.g. G, G+A, C and C+T). The fragments produced are electrophoresed in parallel and the DNA sequence can

be inferred from the pattern and placement of the resulting bands as the rate of travel trough the gel will be related to the size of fragment (Figure 1-11). This method is rarely used due to the complexity of the process and exposure to hazardous and radioactive materials.

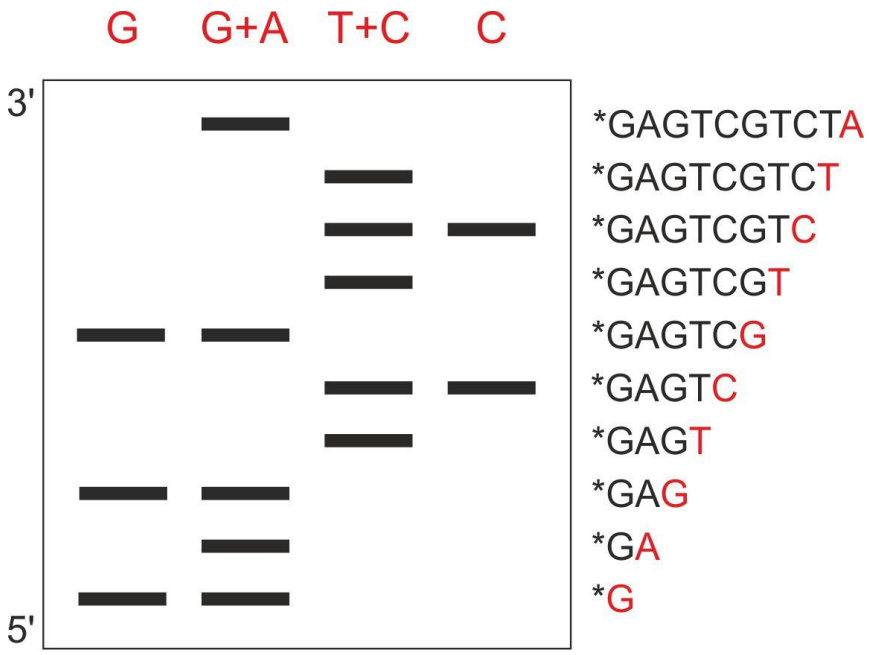


Figure 1-11. Maxam-Gilbert sequencing.
An example of gel electrophoresis output and interpretation following the Maxam-Gilbert sequencing process.

1.5.6.2 Sanger Sequencing

The names most frequently associated with genetic sequencing are Sanger and Coulson who successfully determined the sequence of an E. Coli bacteriophage (Sanger and Coulson 1975) using chain-termination. This method involves purification and denaturation of DNA followed by amplification through cloning or polymerase chain reaction (PCR). The resulting DNA is then divided into four tubes. In each tube one of the four dideoxynucleotides ddATP, ddCTP, ddGTP or ddTTP is added in the presence of DNA polymerase along with the four normal nucleotides. The ddNTP act as chain terminators, therefore if they are incorporated into a DNA strand extension will terminate at that point. The ddNTPs will be incorporated over the normal nucleotide by chance, resulting in fragments of varying lengths. As only one ddNTP is added to each tube the final nucleotide of each chain is known. The resulting double stranded DNA is

denatured and gel electrophoresis performed with one lane used for each ddNTP and from this the DNA sequence can be inferred. The use of fluorescent labels for each individual nucleotide allowed this four tube reaction to be combined in one reaction and the individual ddNTPs can be read by a laser as each emits a different signal.

This new found ability to determine the genetic make-up of living organisms led to the establishment of Human Genome Project in 1986, aiming to map the human genome within 15 years. The Sanger method was used in this project and whilst it took slightly longer than anticipated, the 3 billion base pair sequence of the human genome was completed in 2003 at a cost of $2.7 billion.

The cost of Sanger sequencing on such a massive scale remains prohibitive to most laboratories but is perhaps one of the driving forces behind the search for alternative sequencing methods.

1.5.6.3 Next Generation Sequencing

Next generation sequencing (NGS), also referred to as massively parallel, high throughput or deep sequencing, has vastly increased the throughput of genetic sequencing in comparison with previous methods. There are multiple platforms available, utilising different methods, which are continually evolving and improving. Below is an overview of the currently available platforms but we are in a period of exponential change therefore it is likely that this will soon be out of date.

1.5.6.3.1 Illumina

Illumina (previously known as Solexa) utilise a sequencing-by-synthesis (SBS) method, where DNA fragments are immobilised onto a flow cell with primers where PCR is carried out to produce multiple copies of the target in clusters. The clusters are washed with four types of fluorescently labelled reversible terminating bases in the presence of DNA polymerase. The bases compete for binding sites and if incorporated a laser is used to excite the labelled dyes and photograph the location on the flow cell. The terminating dye is removed and the process is then repeated. All resulting sequenced fragments will be the same

length as length is determined by the number of cycles carried out. With the reagents currently available 600 cycles can be carried out to generate pair-end fragments of 300 bases each. In the paired-end process the DNA target fragment is sequenced in the forward direction and then folded over and sequenced in the reverse direction. This generates two high quality reads from a single target fragment (Illumina 2016).

1.5.6.3.2 Ion PGM Sequencing

In this method nucleotides are sequentially washed over microwells containing target DNA. As a nucleotide is incorporated into the sequence by the DNA polymerase a hydrogen ion is released, resulting in a change in pH in the microwell. It is this change in pH which is detected and used for base calling (Bragg, Stone et al. 2013). As with 454 pyrosequencing it is possible that more than one nucleotide will be incorporated when homopolymers are encountered but the hydrogen ions released and thus the change in pH is proportional to the number of bases added.

1.5.6.3.3 454 Pyrosequencing

Pyrosequencing also uses a SBS approach and relies upon detection of phosphate release as a nucleotide is incorporated. Fragmented DNA is attached to a bead and enclosed within a water droplet where PCR amplification coats the bead with multiple copies of the fragment. Beads are then immobilised on a plate and bases washed over the plate sequentially. Pyrophosphate is released as a base is incorporated, providing the substrate in a luciferase reaction. Visible light is emitted by this reaction which is then detected by a camera. The quantity of light emitted is proportional to the quantity of pyrophosphate released and therefore if multiple bases are incorporated in a reaction this can be detected. The immobilisation of the beads allows the camera to maintain the position of each bead and thus multiple sequences are simultaneously created.

1.5.6.3.4 SOLiD Sequencing

Sequencing by Oligonucleotide Ligation and Detection (SOLiD) employs sequencing by ligation. The template is hybridised to a bead where clonal

populations of the template are prepared with emulsion PCR. All the components of this PCR microreaction are contained in a droplet, suspended in oil. The beads are then immobilised on a glass slide where fluorescently labelled probes ligate to the target releasing a fluorescent signal. The ligation process is repeated up to five times, starting in the n-1 position on subsequent cycles. As a result nearly every base is covered by ligation probes multiple times, hence increasing accuracy.

1.5.6.4 Third Generation Sequencing

1.5.6.4.1 Pacific Biosciences

Pacific Biosciences utilise a Single Molecule Real-Time (SMRT) sequencing method. Essentially this is a sequence-by-synthesis method where a single DNA polymerase compound is immobilised at the bottom of a pore in a solid surface, a zero-mode waveguide (ZMW). A single-stranded DNA molecule attaches and runs through the polymerase compound. Each of the four bases is labelled with a different fluorescent dye at the terminal phosphate and can therefore be added simultaneously. When the polymerase incorporates a nucleotide into the DNA sequence fluorescent dye is clipped off and, upon excitation by a laser, a light signal is emitted which can be detected in real-time. In this method all reagents are incorporated at the beginning of the process with no wash steps required during the process therefore it is rapid in comparison to the alternative (Roberts, Carneiro et al. 2013).

1.6 Importance of diagnostics

The specific therapeutic options for viral respiratory infections may be limited but this does not take away from the importance of accurate diagnosis. Correctly identifying a virus as the cause of illness can still impact on the clinical management of patients through the prevention of inappropriate therapies.

Antibacterial resistance poses a real threat to human health and should be considered a major public health issue. No new classes of antibiotics have been developed in over 20 years and the circulation of multidrug resistant organisms is increasing at an alarming rate. The vast majority of antibiotics are prescribed

in primary care and most of these are for respiratory infections. Antibiotic prescribing rates vary throughout Europe with the UK being amongst the highest (Butler, Hood et al. 2009). This is despite the knowledge that a large proportion of infections, particularly respiratory infections are viral in origin. This approach stems from the historical observations that a simple upper respiratory infection could quickly progress and become complicated by pneumonia (NICE 2008). As a result of the general improvements in healthcare, sanitation and the introduction of vaccinations this is now a much rarer occurrence. There is also a perceived expectation that patients feel antibiotics should be used but these ideas are being challenged.

The use of antibiotics in simple respiratory infections results in no clinical benefit to patients and exposes them to potential drug reactions and side effects (Butler, Hood et al. 2009) as well as being at great expense to healthcare systems. The increasing use of antibiotics correlates with a rise in the rates of antibiotic resistant organisms and rationalising prescribing has been shown to reduce the circulation of resistant bacteria. One of the commonest reasons for antibiotic prescribing is that of viral respiratory infections (van Buul, Veenhuizen et al. 2014). It is now well documented that the presence of a viral infection does not preclude a bacterial infection, in fact co-infections and secondary bacterial infections are seen in up to 15% of childhood pneumonia cases (Jain, Williams et al. 2015). But this same study demonstrated the evidence of a pathogen in 81% of childhood pneumonia cases, 66% of which were viral alone and only 7% were bacteria-viral co-infections. Investigations into the aetiology of acute respiratory infections frequently find viruses as the most common cause of illness.

Penicillin was first introduced to clinical practice in 1940 and within months resistant organisms were detected (Abraham and Chain 1940). Of course at this time the methods of resistance development were not known and it is now understood that multiple mechanisms are responsible. There was an early realisation that antibiotic use could be associated with drug resistance, indeed erythromycin was removed from clinical practice for *S. aureus* treatment as over 25% of cases were found to be erythromycin resistant (Dagan and Bar-David 1992).

Perhaps one of the first cases to highlight this issue as a major public health concern was the emergence of *methicillin-resistant Staphylococcus aureus* (MRSA). MRSA is frequently associated with colonisation; however it poses a great risk to hospitalised individuals as it can result in severe pneumonia, infections of surgical wounds and prosthetic equipment e.g. venous or urinary catheters.

1.7 Potential gains from the utilisation of NGS in a diagnostic setting

As discussed above, the diagnosis of respiratory infection aetiology is an important aspect of clinical management, despite the lack of specific treatment measures. The process of NGS generates vast quantities of data which could be of important use to a health service. The target independent nature of the process could allow for the detection of multiple pathogens using a single sample processing method.

The sequence information generated could be used in molecular epidemiological studies into the presence, spread and evolution of viral diseases. Projects aiming to develop such a system have highlighted issues with regard to incomplete clinical data accompanying sequence data though developing a database in conjunction with clinical testing could improve this situation (Araujo, Souza-Brito et al. 2012).

The long term benefit of such information could be a greater understanding of the viruses which cause disease in humans. This could be used to support vaccine development and clinical research into treatment options through long-term study of viral pathogens and therefore allow the targeting of viral weaknesses which could be exploited to generate further treatment or prevention options.

In a shorter term period such information could allow the real-time study of outbreaks of viral disease. Again, the knowledge of such cases could be essential in development of mechanisms to prevent such instances in the future.

1.8 Research Aims

The aims of the research presented in this thesis were to use next generation sequencing methods and bioinformatic analyses to detect viral pathogens in clinical respiratory samples and compare the outcomes with the current molecular diagnostic assays.

Specific aims:

1) Develop a reliable and reproducible protocol for the processing of clinical samples for next generation sequencing.
2) Develop a data analysis pipeline for the detection of viral genetic sequences from the resulting data sets.
3) Use sequenced information to characterise the viral respiratory pathogens detected in clinical samples.
4) Compare the sequencing outcomes with the current standard respiratory diagnostic molecular assays.
5) Test the developed protocol with further specimen types to determine if the methods are transferable to other clinical areas of viral diagnostics.

Materials and Methods

2.1 Materials

2.1.1 Kits

Kit	Source
NEBNext mRNA Second Strand Synthesis Module	New England BioLabs Inc.
Advantage 2 PCR Kit	Clontech
Nextera XT DNA Sample Preparation Kit	Illumina
KAPA SYBR FAST qPCR Complete kit	Kapa Biosystemes
MagJET Viral DNA and RNA Purification Kit	Thermo Fisher Scientific
Tapestation D1000 ScreenTape System	Agilent
Bioanalyzer DNA 7500 Kit	Agilent

Table 2-1. Kits used in sample processing.

2.1.2 Enzymes

Enzyme Name	Supplier
TURBO DNase	Life Technologies
DNA Polymerase I, Large (Klenow) Fragment	New England BioLabs Inc.
RNase A	Life Technologies

Table 2-2. Enzymes used in sample processing

2.1.3 Primers

Primer Name	Sequence
FR26RV-N	5' GCC GGA GCT CTG CAG ATA TCN NNN NN 3'
FR20RV	5' GCC GGA GCT CTG CAG ATA TC 3’

Table 2-3 Primers used in PCR

2.1.4 Chemicals

Chemical	Abbreviation	Supplier
Agarose	-	Sigma-Aldrich
Ethanol	EtOH	Sigma-Aldrich
Nuclease free water	NFH2O	Thermo Fisher Scientific
Sodium hydroxide	NaOH	Sigma-Aldrich
Ethidium Bromide	EtBr	Sigma-Aldrich
Isopropanol	IPA	Sigma-Aldrich
Tris-acetate-EDTA	TAE	In-house
Polyethylene Glycol	PEG	Sigma-Aldrich
Phosphate buffered saline	PBS	Sigma-Aldrich

Table 2-4. List of chemicals

2.1.5 Illumina reagents and abbreviations

Reagent	Abbreviation
Tagment DNA Buffer	TD
Amplicon Tagment Mix	ATM
Neutralize Tagment Buffer	NT
Nextera PCR Master Mix	NPM
Hybridization buffer	HT1

Table 2-5. Illumina sequencing reagents

2.2 Methods

2.2.1 Nasopharyngeal swab sample preparation

After obtaining a diagnostic swab from a patient, it was placed in viral transport media (VTM) and submitted to the diagnostic laboratory. Upon arrival in the laboratory the VTM tube was opened, the swab pressed against the wall of the tube to remove any cells or fluid and from this point the VTM was treated as the specimen. When samples were not in use they were stored at -20ºC or -80ºC for long term storage.

The NPS used in the main body of the project were obtained from the University of Otago, New Zealand. These samples were collected as part of a large cohort study (Murdoch, Slow et al. 2012). Following completion of this project the residual samples were obtained. Frozen samples were thawed in a water bath at 37ºC then transferred to a sterile 1.5 ml tube and centrifuged at 1500 x g to remove any debris which may hinder nucleic acid extraction. The supernatant was retained and transferred to a fresh tube.

2.2.2 Nucleic acid extraction

Initially DNA and RNA were extracted from clinical samples manually. This method was not thought to be appropriate for high throughput of samples and the number of steps requiring hands-on time and transferring samples between tubes was a risk for potential contamination. Nucleic acids for next generation sequencing were therefore extracted from samples using the MagJET Viral DNA and RNA Kit on the KingFisher™ Flex Purification System (ThermoFisher Scientific). This process utilises guanidium thiocyanate to lyse the cells and particles within samples releasing nucleic acids which bind to beads in the presence of a chaotropic agent before washing and elution from the beads.

A sample volume of 200 µl was transferred to a fresh 1.5 ml tube. To this 200 µl of lysis buffer and 50 µl of proteinase K was added and mixed well. This mixture was transferred to a microtitre deep well 96 plate (plate 1). Plates 2 - 4 were loaded with wash buffers as per manufacturer's instructions and plate 5 loaded with 100 µl of nuclease free water for elution. The Viral_NA_Flex protocol was

initiated on the machine, which pauses following the lysis step to add 25 µl MagJET Magnetic beads and 450 µl 100% isopropanol to each sample well. The protocol was the recommenced and upon completion the eluted nucleic acids were removed from the 96-well plate, transferred to individual 0.2 ml tubes, to be stored at -80^{o}C until processed.

2.2.3 DNase Treatment

DNA was removed by DNase treatment. Reactions comprised of 10 µl of 10x Turbo DNase buffer, 2 µl of 2U/µl DNase (Turbo DNase, Ambion Life Technologies), 10 µl of nucleic acids and 78 µl of nuclease-free water. Reactions were incubated for 30 min at 37^{o}C and mixed every 10 min during this time.

2.2.4 Reverse Transcription

Reverse transcription of purified RNA was carried out using Maxima H Minus (Life Technologies). Reactions comprised of 13 µl purified RNA, 1 µl 10mM dNTPs and 1 µl FR26RV-N primer which was heated to 65^{o}C for 5 min. Then, on ice, 4 µl of 5x Maxima H Minus buffer and 1 µl Maxima H Minus enzyme mix was added. This was incubated at 25^{o}C for 10 min, 55^{o}C for 60 min and finally 85^{o}C for 5 min to terminate the reaction.

2.2.5 RNA Purification

RNA was purified from the DNase reaction using RNAClean XP beads (Agencourt). The beads were vortexed to resuspend and added to the DNase treated solution at a ratio of 1:1.8 and mixed well. The mixture was incubated at room temperature for 5 min then placed on a magnetic rack until the beads were pelleted on the tube wall and the supernatant clear. The supernatant was removed and the bead pellet washed carefully with 200 µl of 70% ethanol. The wash step was repeated and all ethanol removed. The pellet was left to air dry for 15 min and then resuspended in 15 µl of nuclease free water. (This procedure has been shown to effectively purify miRNA with average lengths of 22 nts therefore larger RNA fragments should also be retained (Beckman Coulter Life Sciences)).

2.2.6 Second Strand Synthesis

Second strand DNA was synthesised using NEBNext mRNA Second Strand Synthesis Module (New England BioLabs). DNA was heated to 95°C for 3 min, then on ice 36 µl of nuclease-free water, 6 µl of 10x NEBNext mRNA Second Strand Synthesis Module buffer and 3 µl of NEBNext mRNA Second Strand Module enzyme mix was added. This was mixed gently and incubated at 16°C for 2 hours and 30 min on a heat block.

2.2.7 DNA Purification

DNA was purified using AMPure XP beads (Agencourt, Beckman Coulter). The beads were vortexed to resuspend and allowed to equilibrate to room temperature and then added to the second strand cDNA mix at a ratio of 1:1.8 and incubated at room temperature for 5 min. The solution was then placed on a magnetic rack until the beads pelleted on the wall of the tube and the supernatant was clear. The supernatant was removed and the pellet washed with 200 µl of 70% ethanol. The ethanol was removed and the wash step repeated. All ethanol was then removed and the pellet allowed to air dry for 15 min at room temperature. The dried pellet was resuspended in 20 µl of nuclease-free water.

2.2.8 Sequence Independent Single Primer Amplification (SISPA) Polymerase Chain Reaction

PCR reactions were set up in a total volume of 50 µl. A template volume of 10 µl was added to a master mix containing 1µl Advantage 2 Polymerase, 5 µl 10x buffer, 1 µl 10nM dNTPs, 0.2 µl primer FR26RV (10mM), 1 µl primer FR20RV (10mM) (Froussard 1992) and nuclease-free water 31.8 µl. The PCR was carried out on an ABI 2720 (Applied Biosystems) with the following cycling parameters: 2 min 95ºC, followed by 35 cycles of 95ºC for 30 seconds, 63.5ºC for 1 minute and 72ºC for 3 min, then a final extension time of 72ºC for 10 min.

2.2.9 Agarose Gel Electrophoresis

The presence of PCR products were confirmed using 1% agarose gels, prepared using 1x TAE buffer and a final ethidium bromide concentration of 0.4 µg/ml.

The gels were immersed in 1xTAE and samples loaded with a 3:1 volumes loading dye in parallel with a 1kbp ladder to determine fragment size. The gels were loaded and run at 100V until separation was achieved then fragments were visualised and imaged using a UV transilluminator (GeneFlash, Syngene).

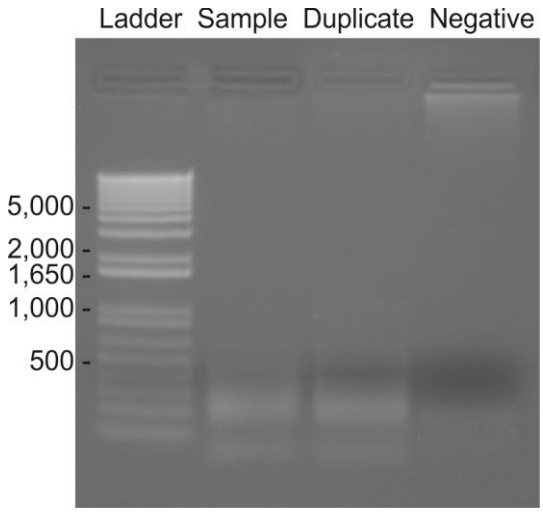


Figure 2-1. An example of the gel electrophoresis output following a SISPA reaction.
The lanes containing samples show a smear rather than distinct band as the PCR does not have a specific target, rather aims to non-specifically bind to and amplify all DNA within the sample.

2.2.10 DNA Quantification

DNA was quantified prior to library preparation using the Qubit dsDNA HS Assay Kit for the Qubit Fluorometric Quantification system (Life Technologies). The Qubit working solution was prepared by diluting the Qubit dsDNA HS reagent 1:200 in Qubit dsDNA HS buffer in a nuclease free tube, with 200 µl in total being prepared for each sample and standard. Prior to each set of samples the fluorometer was calibrated using the supplied standards. 190 µl of working solution was added to 10 µl of standard 1 and 2 and 198 µl of working solution was added to 2 µl of each cDNA sample in a fresh thin-wall, clear 0.5mL PCR tube. All mixes were allowed to incubate for 2 min at room temperature prior to reading. Standards 1 and 2 were read by the fluorometer followed by the samples. The stock concentration was calculated as Qubit value x (200/X) where X is the initial volume of sample used and converted to ng/ul.

2.2.11 Nextera XT Library Preparation

Following Qubit quantification 1ng of DNA was diluted to a total volume of 5 µl with nuclease-free water. Sequencing libraries were prepared using the Nextera XT DNA Library Preparation Kit (Illumina). The kit was removed from -20^{o}C and allowed to thaw on ice. Before use each reagent tube was inverted 5 times to mix thoroughly. Reactions were set up in nuclease-free 0.2 ml PCR tubes.

2.2.11.1 Tagmentation

To each reaction 10 µl of TD, 1 ng of DNA and 5 µl ATM was added, gently pipetted up and down 5 times to mix and spun down briefly before incubating at 55^{o}C for 5 min. NT 5 µl was added to the mixture, mixed and spun down briefly to halt the reaction.

2.2.11.2 Library Indexing

Barcode tags were added to each sample through PCR amplification. 15 µl of NPM was added to each tube containing a library then 5 µl of the corresponding index 1 and 2 primers. The original caps of each primer were discarded and replaced with new caps to prevent cross-contamination. The mix was gently pipetted and briefly spun down before placing on a thermocycler (ABI 2720, Applied Biosystems) using the following programme; 72^{o}C for 3 min, 95^{o}C for 30 seconds, 12 cycles of 95^{o}C for 10 seconds, 55 ^{o}C for 30 seconds, 72^{o}C for 30 seconds, then a final extension step of 72^{o}C for 5 min. The reactions were purified using AMPure XP beads as described for cDNA purification.

2.2.11.3 Library Quantification

Libraries were quantified using the KAPA SYBR FAST qPCR Complete kit. Each library was diluted to 1:1000, 1:2000 and 1:10000 with nuclease-free water and 4 µl of each dilution was transferred to a 96-well plate along with duplicate wells for each of the 6 supplied standards. To each well containing a library or standard 4 µl of PCR grade water and 12 µl of KAPA SYBR FAST qPCR Master Mix was added. The KAPA SYBR FAST qPCR Master contained primers Primer P1 (5'-AAT GAT ACG GCG ACC ACC GA-3') and Primer P2 (5'-CAA GCA GAA GAC GGC ATA

CGA-3') targeting the Illumina adapters used in barcoding during library preparation. The following cycling parameters were used; 95°C for 5 min then 35 cycles of 95°C for 30 seconds and 60°C for 45 seconds (ABI 7500, Applied Biosystems). The concentration was corrected for the dilution and the mean of the three measurements taken as the final concentration.

2.2.11.4 Library Insert Size Calculation

The average library insert size was calculated using the High Sensitivity D1000 ScreenTape System (Agilent). A 2 µl aliquot of the prepared library was transferred to a fresh 0.2 ml tube, to which 2 µl of High Sensitivity D1000 Sample Buffer was added. For each set of samples 1 lane contained High Sensitivity D1000 Ladder in place of a library to allow fragment size quantification. The tubes were loaded into the 2200 Tapestation and sample names entered into the controller software before commencing the run.

The corrected library concentration was calculated by:

$$\frac{\text{Quantity (ng/µl)}}{\text{Mean insert size (bp)}} \times 452 = \text{Concentration (ng/µl)}$$

Figure 2-2. Formula used to determine library concentration.

2.2.11.5 Library Pooling and Denaturing

A reagent cartridge was removed from storage at -20 °C and placed in a water bath at room temperature for 60 min. The cartridge was inverted several times to ensure adequate mixing of reagents and to check for any precipitation which would indicate that the cartridge was not usable.

Figure 2-3. This single use reagent cartridge is pre-filled with sequencing reagents.
The prepared sample libraries are added to the highlighted aperture while the other apertures allow access to the separate regents during the process. The cartridge is then loaded onto the MiSeq, shown on the right. *Image obtained from Illumina Inc.*

The addition of barcode indexes during the library preparation process allowed multiple samples to be sequenced during a single run of the sequencer. In order to ensure equal coverage between samples the concentration of each individual library was diluted to 4nM with nuclease-free water and a 5 µl aliquot of each diluted library was pooled in a single tube. The concentration of the pooled library was calculated by a combination of the concentration measured by Qubit and the average insert size by Tapestation using the formula below.

$$\frac{\text{Concentration (ng/µl)}}{\text{Molecular weight dsDNA (660 Da)} \times \text{Insert size (Base pairs)}} \times 1\text{x}10^{6} = \text{Molar Concentration (nM)}$$

Figure 2-4. Calculation of library molar concentration.

The pooled library was denatured into single stranded DNA using freshly prepared sodium hydroxide 0.2 M and left to stand at room temperature for five min. The denatured library was then diluted to 12.5 pM using HT1 buffer (Illumina). An internal control, Phi X (Illumina), was added to concentration of 1%. This not only acts as an internal control but also helps to increase sample diversity which improves base-calling. The final concentration of the libraries is important as it determines the density of clustering on the sequencing flow cell

and can vary between machines. This concentration was optimised for the in-house machine following installation. The diluted library was loaded on to the thawed MiSeq reagent cartridge which was placed on the MiSeq. The time taken to complete the run will depend on the number of cycles being carried out i.e. the length of reads generated. At the time of optimisation the longest available length of reads was 2 x 150 bases.

2.3 Data Analysis

The steps taken and tools used during data analysis are outlined below (Figure 2-5).

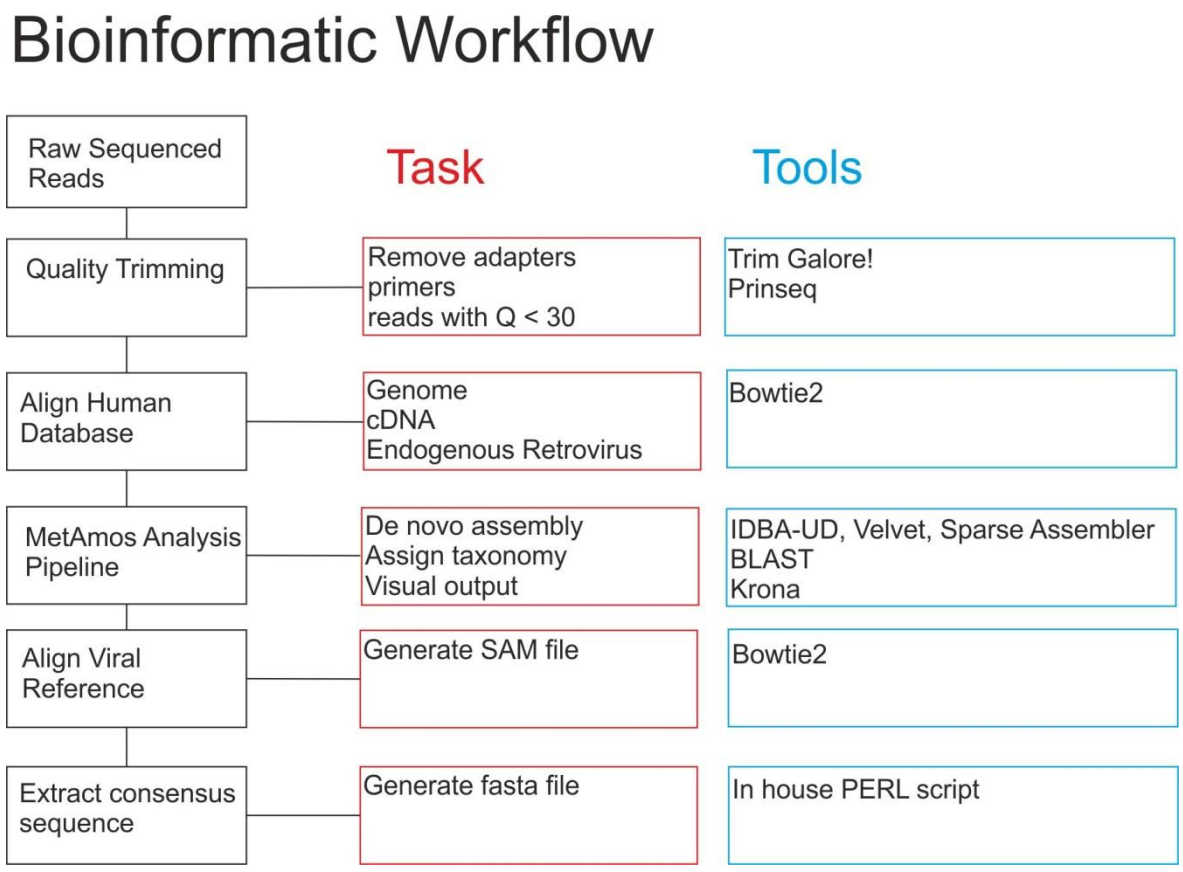


Figure 2-5. Graphical overview of steps taken and tools used during analysis of sequenced reads.

2.3.1 Quality Trimming of Sequenced Reads

During cDNA and library preparation primers and adapters were introduced into the DNA fragments to be sequenced. These artificial sequences have the potential to interfere with subsequent analyses and were therefore removed from the dataset. This was carried out using a combination of the command line

tools Trim Galore! (Lindgreen 2012) and PrinSeq (Schmieder and Edwards 2011). The read quality was visualised with FastQC (Andrews) before and after trimming and any dataset with an average quality score persistently below Q30 was deemed to have failed quality control and was not used in downstream analyses. Example outputs are shown in Figure 2-6.

Figure 2-6. FastQC visualisation of Phred quality scores.
The quality scores of sequenced reads visualised with FastQC before and after quality trimming. The red line is the median value and blue line the mean, the yellow boxes show the interquartile range (25-75%) and the whiskers represent the 10 and 90% points.

2.3.2 De Novo Assembly

Sequenced reads were assembled into longer contiguous sequences i.e. contigs as part of the MetAmos pipeline (Treangen, Koren et al. 2013). Multiple assemblers were used during this step, Velvet, IDBA-ud and SparseAssembler, with the default parameters. The contigs from the most effective assembly, as determined by the LAP score (log average probability), were used for downstream analysis. As part of this pipeline the assembled contigs were assigned taxonomy using a BLASTp search. The output of this process was saved to a text file as well a visual representation in a Krona chart, an example of which is shown below in Figure 2-7.

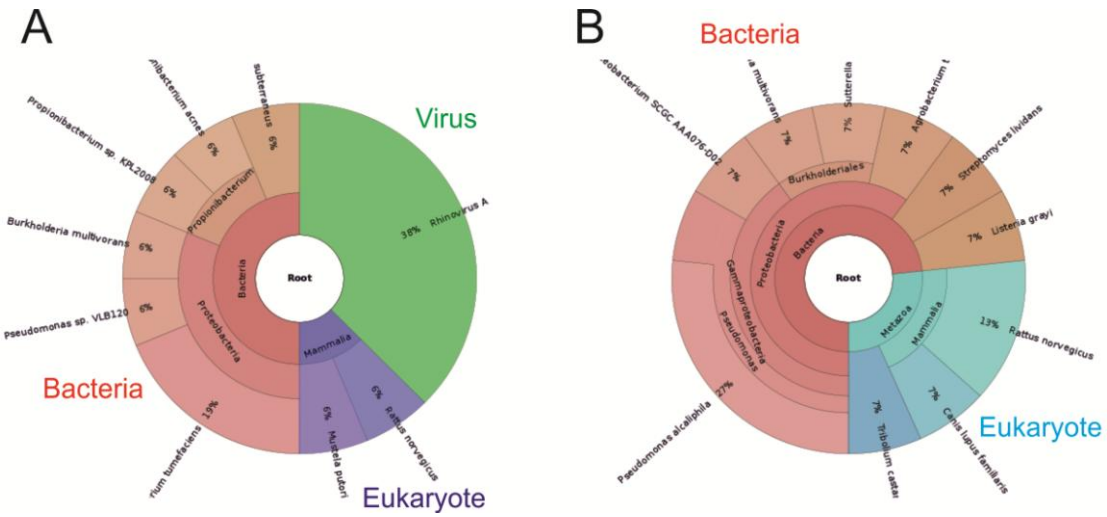


Figure 2-7. An example of a Krona output chart.
Panel A shows an output chart where viral reads were detected within the sample and panel B shows an output with no viral reads.

2.3.3 BLAST

The Basic Local Alignment Search Tool (BLAST) algorithm was run on a local server against the NCBI nucleotide (nt) or non-redundant protein (nr) database which is updated on a monthly basis. For all searches the Expect value (E) was set at 0.001. The output was saved in a text file and inspected manually.

2.3.4 Alignment of reads to reference database

Both BWA and Bowtie2 v 2.2.5 (Li and Durbin 2010; Langmead and Salzberg 2012) aligners were used to map sequenced reads to reference genomes, generating sequence alignment / map (SAM) files. The commands and settings used for these programs are documented in Appendix 1.

2.3.5 Manipulation of alignment files

Mapped reads were removed or extracted from SAM files using commands within SamTools. SAM files were also transformed to their binary counterparts, BAM, to reduce the computational memory required in analysis. This was also carried out

using SamTools. The commands and settings used for these processes are documented in Appendix 1.

2.3.6 Generating an alignment consensus

A consensus sequence was generated from the alignment file using a custom Perl script (J. Hughes, unpublished). A subset of these consensus sequences were manually compared with a visual output of the SAM file using Tablet, to ensure accuracy.

2.3.7 Phylogenetic analysis

Sequences used in phylogenetic analyses were initially aligned in MEGA6 (Tamura, Stecher et al. 2013) using Muscle. Neighbour-joining or maximum-likelihood trees were inferred using MEGA6 with a model determined by the Bayesian Information Criterion (BIC) score. Where possible, a similar viral sequence was included as an outgroup though in some cases this was not possible as the most closely related virus introduced too many gaps to the alignment. The bootstrapping method was used with 100 replicates for maximum-likelihood and 1000 for the neighbour joining method unless stated otherwise.

2.3.8 Statistical Analyses

All statistical analyses, unless stated, were carried out using GraphPad Prism Version 5.01 (GraphPad Software, Inc., California, USA)

Utilising Next Generation Sequencing as a diagnostic tool for viral respiratory tract infections

3.1 Introduction

Acute respiratory infections represent a significant cause of mortality and morbidity worldwide and are the commonest reason for primary care consultation in developed countries (Stanton, Francis et al. 2010). Viruses are the major contributing pathogens to infections of the human respiratory tract.

The common pathogens associated with respiratory infections are outlined in chapter 1. In brief, rhinoviruses, coronaviruses and influenza are the main contributing viruses to human respiratory infections but it is worth noting that contributing viruses will vary with population demographics and time of year as many circulate with distinctive seasonal patterns (Nickbakhsh, Thorburn et al. 2016). Please note that many other viruses can also cause respiratory symptoms as part of a further disease syndrome such as Epstein-Barr, measles and mumps viruses but these are not included in further discussion here as respiratory illness is not the main characteristic of infection.

As outlined previously the diagnostic tests for viral respiratory illness have changed significantly over the last 20 years. Until relatively recently, diagnostic testing in this field relied on immunofluorescence studies or culture and isolation of virus. Such methods are technically challenging and in the case of culture, result in a long turn-around time of up to 30 days for some pathogens. This is of little benefit in the management of respiratory tract infections as the individual will either recover or succumb prior to test results being issued. There is also difficulty in detecting some viruses through isolation, for example, rhinoviruses will not grow easily in cell lines. Many diagnostic virology laboratories now employ molecular based assays such as PCR or RT-PCR, which are now considered the gold standard in this field. Molecular assays have increased test sensitivity and specificity over previous methods (Aguilar, Perez-Brena et al. 2000; Ingram, Fenwick et al. 2006). Such tests can be developed quickly in reaction to the emergence of novel pathogens or strains. These tests can also be combined to test for multiple pathogens from a single specimen.

Respiratory virus testing is now a major part of most diagnostic virology laboratories' workload. The current method used to detect respiratory pathogens at the WoSSVC, Glasgow is a multiplex real time PCR test. This panel of five assays contains multiple specific primer and probe sets and is designed to detect 15 common pathogens associated with respiratory infections. The use of multiplex assays reduces overall workload and cost by reducing the number of tests carried out and volume of consumables required (Edwards and Gibbs 1994). The use of a single workflow also reduces the hands-on time required and generates results from a single workflow rather than the staggered results which would have been seen using a combination of methods. The syndromic approach of testing is particularly useful in respiratory infections; although many pathogens circulate with a seasonal pattern there may also be outbreaks, epidemics and even pandemics outwith the expected season. This blanket approach to testing will ensure that unexpected pathogens are detected in a timely manner.

Assay number				
1	**2**	**3**	**4**	**5**
Influenza A Parainfluenza 1 Human Metapneumovirus	Parainfluenza 2 Parainfluenza 3 Parainfluenza 4	Adenovirus Respiratory Syncytial Virus-A Respiratory Syncytial Virus-B Influenza B	Coronavirus 229E Coronavirus OC43 Coronavirus NL63 Rhino/Enterovirus	*Mycoplasma pneumoniae* *Bordetella pertussis* *Bordetella pertussis* toxin

Table 3-1. Pathogens detected by multiplex assays used in respiratory testing.

As discussed in chapter 1 there are pitfalls of using such tests. Two important examples would be firstly, established viral pathogens can undergo genetic mutations by various mechanisms, be that genetic drift, shift or recombination. It is important to consider that this evolution may affect PCR primer and probe target regions resulting in reduced efficacy of an assay or even false negative results. Also, in order to detect a pathogen it must be actively sought. As a

result, new or unknown viral infections will not be detected unless a specific assay has been designed.

The introduction of NGS in a research setting has been used in the discovery and characterisation of viral pathogens (Svraka, Rosario et al. 2010). We propose that this metagenomic approach (metagenomic referring to the study all genetic material within a specimen) could be used in the routine laboratory setting to detect respiratory pathogens in a single test/assay. The use of this technique would in theory, obviate the need for pre-test selection of a target pathogen. As all genetic material can be sequenced from a specimen in a single assay, this also raises the possibility of pan-microbial diagnostic test with the ability to detect viruses, bacteria and fungi from a single sample. The sequenced data generated could then be used for pathogen typing and subsequent epidemiological analyses. As a result these tasks could be carried out within the same test, something not currently achievable with the current real time PCR method.

3.2 Aim

The aim of this chapter was to compare the performance of NGS with that of real-time PCR using a panel of clinical respiratory samples. The performance points to be analysed included viral detection, quantitation and subsequent viral typing using sequence data generated by NGS. We compared the ability of the NGS pipeline to detect viral pathogens to RT-PCR. We also compared the number of sequenced viral reads with the RT-PCR Ct to determine if pathogen quantification would be possible. We then used the sequence information generated by NGS in the typing of detected viral pathogens.

3.3 Methods

3.3.1 Samples

Nasopharyngeal swabs (NPS) were collected as part of a cohort study into the use of vitamin D supplementation as a prophylactic treatment of upper respiratory tract infections (Murdoch, Slow et al. 2012). Healthy adults, working in a university hospital were recruited and asked to report any episodes of upper respiratory tract illness. Upon reporting symptoms swabs were collected for

diagnosis. When nasopharyngeal swabs were obtained from symptomatic individuals they were placed in a liquid viral transport media (VTM) and any cells or material obtained by swabbing a patient was then transferred from the swab to the VTM. From this point on the media was treated as the sample and stored at -80ºC until testing. Eighty-nine samples from 686 episodes were randomly selected to be included in this study.

3.3.2 Sample Preparation for Next Generation Sequencing

The sample preparation process is described in detail in chapter 2. In brief, VTM (sample) was thawed in a water bath at 37ºC and centrifuged at 500 g for 10 min to pellet any mucous or debris. A 200 µl aliquot of the supernatant was retained and nucleic acids extracted using the MagJET Viral DNA and RNA Purification Kit for the Kingfisher™ Flex platform (Thermofisher). Reverse transcription was carried out with the enzyme Maxima H Minus (Life Technologies) in the presence of primer FR26-RV. Second strand cDNA synthesis was carried out using NEBNext mRNA Second Strand Synthesis Module (New England BioLabs) and amplified with Advantage® 2 PCR Kit (Clontech) and primer FR20RV. The double stranded cDNA was purified with Agencourt Ampure XP beads (Beckman Coulter) and quantified with the Qubit® dsDNA HS Assay (ThermoFisher Scientific).

The cDNA was diluted to a concentration of 0.2 ng/µl and a 5 µl aliquot was used for sequencing library construction using the Nextera XT DNA Library Prep Kit (Illumina) as per the manufacturers' instructions. The resulting sequencing libraries were quantified using real-time PCR (Kappa Library Quantification Kit) and the concentration corrected for size using the average fragment size as determine by capillary electrophoresis (High Sensitivity D1K Reagents, Tape station, Agilent).

Multiple libraries were processed in each sequencing run therefore to ensure equal coverage of each the libraries were diluted to 4 nM and 5 µl of each diluted library pooled in a fresh tube. The pooled library was diluted and the internal control, PhiX, added at 1% before commencing the sequencing run. Sequencing was carried out on the MiSeq (Illumina) platform generating 150 bp paired-end reads.

3.3.3 Data Analysis

Multiple steps were taken in the analysis of the generated sequences, which are described in detail in chapter 2. An overview of these steps and the tools used for each are presented in Figure 2-5.

3.3.3.1 Quality Control Steps

Primers, sequencing adapters and low quality reads were removed from the dataset using Trim Galore! and Prinseq (Schmieder and Edwards 2011). The quality trimmed reads were mapped against a database of human reference sequences containing the human genome, human cDNA and endogenous retroviruses using the alignment tool Bowtie2 (Langmead and Salzberg 2012) and mapped reads removed from the dataset. The aim of this step was to reduce the computational power required for subsequent analysis. These steps were carried out using a Perl script, thus reducing the amount of input required by the user.

There is a possibility of cross-talk between adapters during the sequencing process. While the introduction of dual indexing has been shown to reduce the level of sample-to-sample contamination (Kircher, Sawyer et al. 2012) this can still occur in areas of high clustering on a flow cell. As it would not be possible to determine which sample these arose from any identical reads found in two or more samples were removed from the datasets.

To identify identical viral reads within samples the sequenced reads were aligned a database of viral reference sequences using bowtie2. The resulting SAM alignment files were then converted to BED files using BEDTools (Quinlan and Hall 2010). The BED file was compared to all samples on the same sequencing run, also using BEDTools, to retain reads mapping to unique portions of the reference.

3.3.3.2 Virus detection pipeline

The remaining unique non-human reads were entered into the MetAmos pipeline (Treangen, Koren et al. 2013), using the Bowtie2 alignment tool, multiple de novo assembly programs and annotation carried out using BLAST against a protein database. The presence of viral contigs was determined through manual

inspection of the Krona output chart (see Figure 2-7 for an example chart). These contigs were then used in a BLAST search against a viral nucleotide database. The top hit from this search as defined by E value and percentage identity was then used as a reference against which to align non-human reads, in order to generate an alignment and consensus sequence.

3.3.3.3 Real Time PCR Method and quantitation

In order to compare the efficacy of an NGS approach to diagnostics with that of current practices a 40 µl of the same nucleic acid extract was screened for human rhinovirus (HRV), influenza A/B (IFA/IFB), respiratory syncytial virus (RSV), adenovirus (ADV), human metapneumovirus (hMPV), parainfluenzavirus 1-4 (PIV 1-4), coronaviruses (HCoV) NL63, OC43 and 229E and Mycoplasma pneumonia using the routine diagnostic qRT-PCR at the West of Scotland Specialist Virology Centre (WoSSVC) as previously described (Gunson and Carman 2011).

The PCR results were compared with those of NGS, initially for virus identification. To assess quantitation the PCR Ct values were compared with the proportion of sequenced reads generated which were of viral origin to determine if there is a relationship between virus quantity and sequenced reads. Statistical analyses were carried out using GraphPad Prism 5 software.

3.4 Results

3.4.1 Virus detection by RT-PCR

To determine the efficacy of an NGS approach as a diagnostic tool, the same nucleic acid extract was subjected to both the current diagnostic RT-PCR test and the NGS method of viral detection. The results of this diagnostic screen could then be used as a standard from which the performance of a sequence independent NGS approach could be calculated.

Within the cohort, 48/89 samples tested positive for a virus by RT-PCR. One sample (1G1) was positive for two viruses, adenovirus and rhino/enterovirus therefore there were 49 viral detections in total. The viruses detected by RT-PCR were as follows: HRV 24/49, RSV 5/49, HCoV 229E 9/49, HCoV NL63 3/49,

HCoV OC43 3/49, hMPV 3/49, PIV-3 2/49, PIV-2 1/49, IFA 1/49 and AdV 1/49. The Ct of the detected viruses ranged from 14.63 to 36.07.

3.4.2 Virus detection by NGS

During optimisation of the protocol the nucleic acid concentration following extraction was found to be too low to quantify in most samples therefore the concentration was not routinely measured after extraction. cDNA was quantified after SISPA PCR and bead purification of the PCR product. The cDNA concentrations of the prepared samples are listed in Table 3-2. The concentration was then corrected to 0.2 ng/µl for subsequent library preparation.

Run 1		Run 2		Run 3		Run 4	
Sample	Conc. ng/µl	**Sample**	Conc. ng/µl	**Sample**	Conc. ng/µl	**Sample**	Conc. ng/µl
2A1	20.7	**1E6**	0.530	**1C1**	3.64	**1B1**	36.5
2A2	26.9	**1E7**	2.23	**1C2**	10.6	**1B2**	28.8
2A3	36.0	**1E8**	8.05	**1C3**	13.7	**1B3**	23.5
2A4	28.8	**1E9**	0.985	**1C4**	29.2	**1B4**	31.2
2A5	28.6	**1F1**	1.29	**1C6**	20.7	**1B5**	35.0
2A6	38.8	**1F2**	0.802	**1C7**	16.7	**1B6**	36.8
2A7	29.5	**1F3**	1.18	**1C8**	25.8	**1B7**	29.0
2A8	28.2	**1F5**	1.41	**1C9**	11.8	**1B8**	28.0
2A9	26.9	**1F7**	1.25	**1D1**	24.1	**1B9**	34.6
2B1	37.0	**1F8**	3.13	**1D2**	11.1	**1D7**	28.0
2B2	28.4	**1G1**	0.858	**1D3**	33.6	**1D8**	32.9
2B3	25.1	**1G2**	1.14	**1D4**	37.5	**1D9**	45.5
2B4	28.9	**1G3**	1.88	**1D5**	15.5	**1E1**	31.6
2B5	32.1	**1G5**	6.84	**1D6**	9.21	**1E2**	25.7
2B6	31.7	**1G6**	1.02	**1I2**	25.7	**1E3**	47.2
2B7	24.88	**1G7**	2.47	**1I3**	13.0	**1E4**	37.5
2B8	27.4	**1H1**	19.6	**1I4**	20.4	**1E5**	36.4
2B9	24.9	**1H3**	1.50	**1I5**	20.5	**2D1**	38.9
2C1	26.4	**1H4**	1.53	**1I6**	24.1	**2D2**	27.2
2C2	29.1	**1H5**	4.14	**1I7**	22.7	**2D3**	47.0
2C3	22.7	**1H6**	1.35	**1I8**	17.9	**2D4**	40.4
2C4	27.5	**1H7**	4.98	**1I9**	1.16	**2D5**	30.9
2C6	23.3			**1H8**	1.55	**2D6**	39.3

Table 3-2. Concentration of sample cDNA after RT, second strand synthesis and PCR clean up.

3.4.2.1 Sequenced Reads

Samples were processed in four runs of the Illumina MiSeq. The number of raw sequenced reads generated from each sample ranged from 206,898 – 2,235,798 (mean = 1,116,238) and after quality trimming 46,072 – 1278122 (mean = 661,799). Following mapping against the set of human reference sequences 5004 – 836868 (mean = 125,316) reads remained unmapped. The proportion of reads

removed by quality trimming and filtering per sample are detailed in Figure 3-1 with 15-60% of the total reads mapping to human reference sequences and over 90% of the raw reads being removed in the majority of samples.

One sample, 1E6, was removed from further analysis as the average read quality remained below Phred 20 following the quality control steps. As this correlates with an error rate of between 1 in 10 and 1 in 100 these reads could not be reliably assessed.

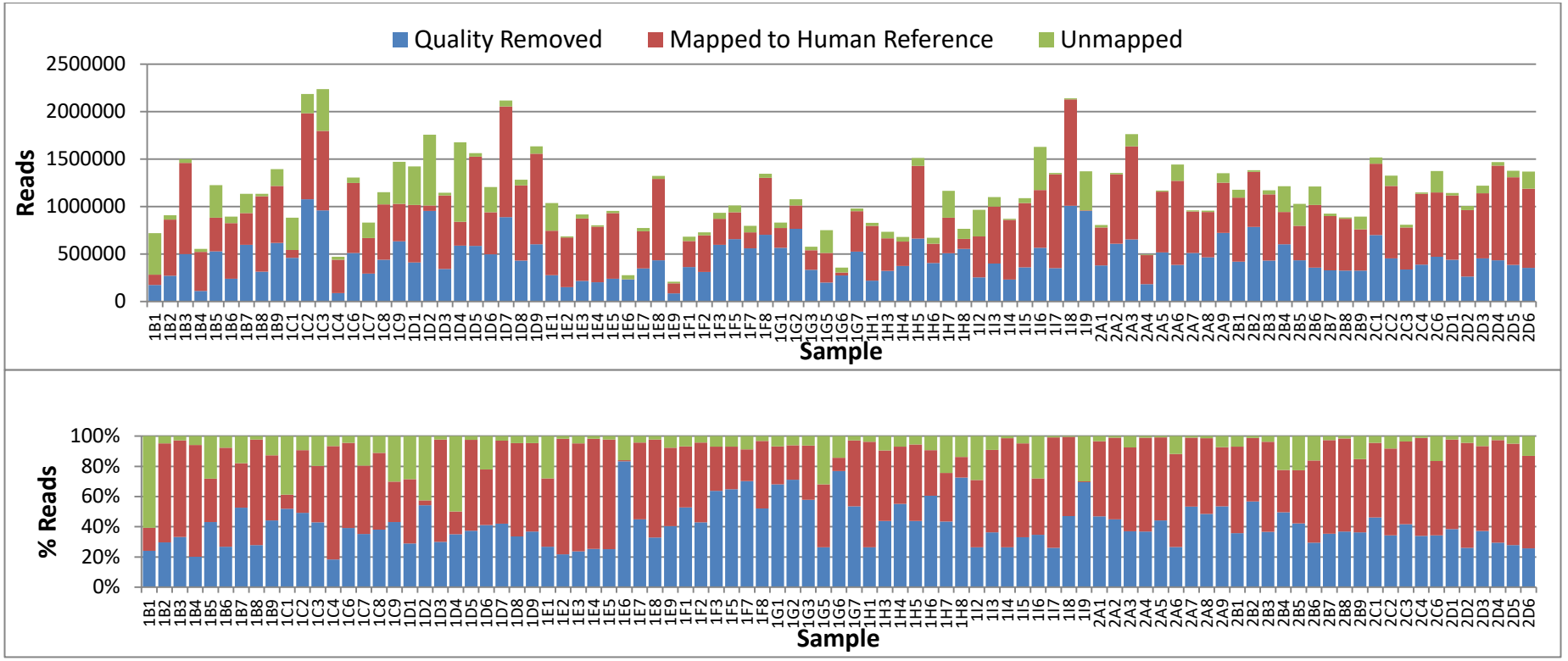


Figure 3-1. The distribution of sequenced reads.

The top panel shows the number of reads removed by quality trimming (blue), mapping to a human reference (red) and the remaining unmapped reads (green) which were entered into the MetAmos data analysis pipeline. The bottom panel shows the read distribution as a proportion of the total.

3.4.2.2 Multiple common respiratory viruses were detected by NGS

Using the virus detection pipeline, viral pathogens were detected in 38 of 89 clinical specimens. Picornaviruses were detected at the highest frequency and throughout the study period, with 21 detections. The next most frequently detected viral group was the coronaviridae. Nine samples were found to contain coronaviruses of the following types: HCoV 229E (4/9), HCoV NL63 (3/9) and HCoV OC43 (2/9). Paramyxoviruses were detected in seven samples, RSV (3/7), hMPV (2/7) and HPIV3 (2/7). One orthomyxovirus was detected and the consensus sequence identified as influenza A H3N2. All detections by NGS were also detected by RT-PCR. There were 49 viral detections by RT-PCR, 11 of these were not identified by NGS.

Using the quality trimmed and filtered reads it was possible to reconstitute partial or full genome consensus sequences from reference based alignments. The alignments showed between 1 and 10.8% nucleotide difference from the closest available reference sequences and in all but one case included a protein coding region. From this level of similarity the viral species could be inferred. The BLAST results with the accession number for the relevant BLAST hits and summary of subsequent reference alignments are shown in Appendix 2. Respiratory Specimens: Summary of Results. This information will be referred to throughout this section. Each virus group will be discussed individually below.

3.4.2.3 Rhinovirus

Multiple different rhinovirus species were detected: A (11/21), B (4/21) and C (5/21) as well as enterovirus D (1/21). The Ct of the viruses detected by RT-PCR ranged from 17.24 to 33.66.

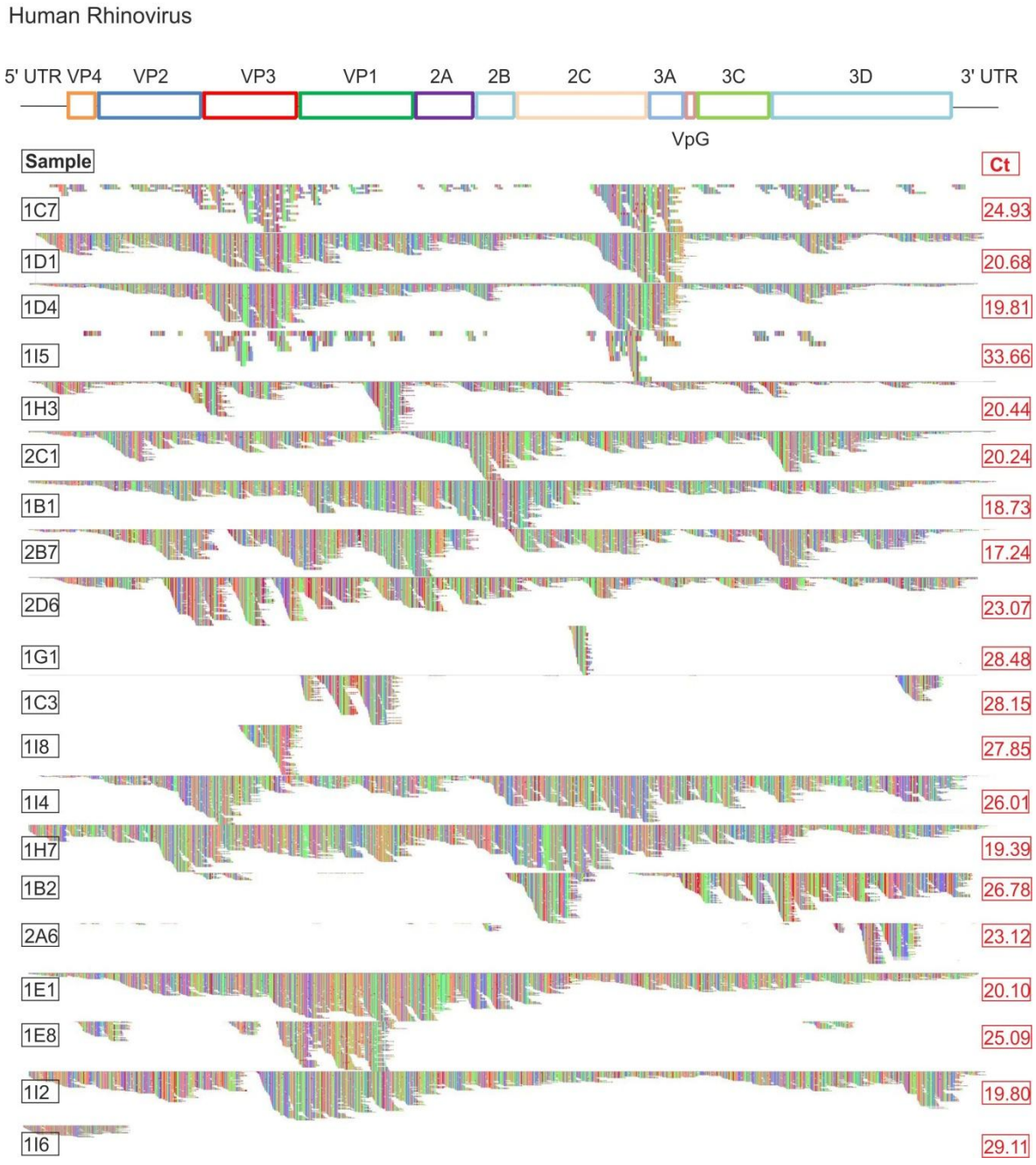


Figure 3-2. Sequencing coverage of the rhinovirus genome.
Alignments were visualised using Tablet.

A representation of the rhinovirus genome with the sequencing coverage of each clinical sample shown below, demonstrating the breadth and depth of coverage generated by NGS.

In general the coverage obtained of rhinoviruses was broad, with an average 68.3% reference genome coverage (range of 2.6 – 100%). The aligned reads showed a low level of nucleotide mismatch, an average of 4.3%, to the closest reference sequence as determined by BLAST (range of 1.2 – 10.8%). Sequence information was obtained from protein coding regions of the genome in all cases with the exception of 1I6 where only UTR sequence was generated. From this coverage of protein coding regions and the high level of nucleotide similarity to the reference in the remaining 20 cases, this information was then used to infer viral species. From this we identified the serotypes found in Table 3-3.

HRV A - Serotype	Number	HRV B - Serotype	Number	HRV C - Serotype	Number	Enterovirus	Number
A1	4	**B27**	1	**C11**	1	D68	1
A21	1	**B3**	1	**C15**	2		
A49	1	**B4**	1	**C17**	1		
A60	1	**B92**	1	**Untyped**	1		
A8	1						
A90	1						
A97	1						

Table 3-3. Rhinovirus and Enterovirus serotypes identified in clinical specimens.
NGS failed to detect virus in three samples which were RT-PCR positive, 1F7, 1B7 and 2B9. The Ct in these samples was 28.03, 28.94 and 33.31 respectively.

3.4.2.4 Respiratory Syncytial Virus

hRSV was detected in three clinical samples, 1E3, 1F1 and 2B4. These samples had a RT-PCR Ct of 20.84, 25.08 and 19.32 respectively.

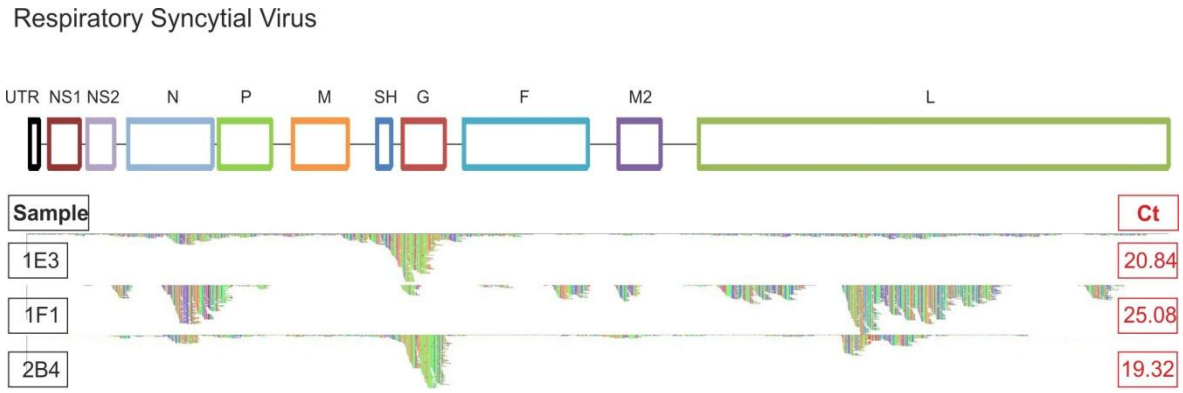


Figure 3-3. Sequencing coverage of the respiratory syncytial virus genome.
Alignments were visualised using Tablet.

Following assembly to the top BLAST hit, coverage was obtained for between 54.6 and 99.3% of a reference genome with a low level of nucleotide mismatch to the reference, 0.5 to 2%. The genome coverage is shown in Figure 3-3, shows at least partial G protein coverage in all cases and low level coverage of the F protein coding region. As both these regions have been used in the molecular typing of hRSV the top BLAST was used to infer type. Samples 1E3 and 1F1 were typed as hRSV-A and 2B4 was hRSV-B.

NGS failed to detect hRV in two cases which were RT-PCR positive, 1E5 and 2B6 with Cts of 30.01 and 33.96 respectively.

3.4.2.5 Human Metapneumovirus

Two cases of hMPV were identified with NGS in sample 1G6 and 2A9. These samples had Cts of 26.84 and 25.68 respectively. Genome coverage was generated against 18.5 and 48.4% of the reference with a low level of nucleotide mismatch, 2.0 and 3.1%. As shown in Figure 3-4, the reference coverage in these cases was not complete but partial coverage of the G protein coding region was obtained from both specimens. As this protein has been used in the molecular typing of hMPV the types for both specimens were inferred from the BLAST hit as type B.

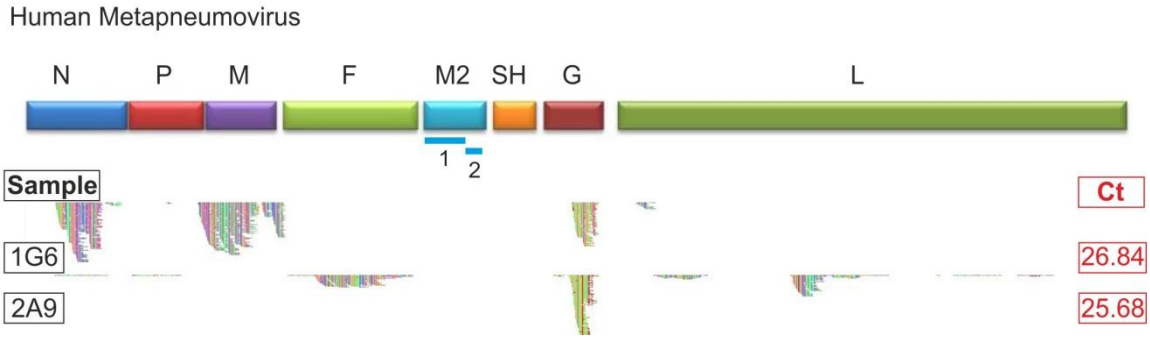


Figure 3-4. Sequencing coverage of the human metapneumovirus genome. Alignments were visualised using Tablet.

3.4.2.6 Parainfluenza Viruses

Parainfluenza viruses were detected in two clinical cases, 1B5 and 1G2. The alignments generated reference coverage of 5.0 and 2.8% only. The Cts of these

specimens were 25.57 and 28.09 respectively. The top BLAST for these specimens were both PIV-3 which were consistent with the RT-PCR results.

NGS failed to detect one specimen which was RT-PCR positive for PIV-2, specimen 2D4, with a Ct of 36.07.

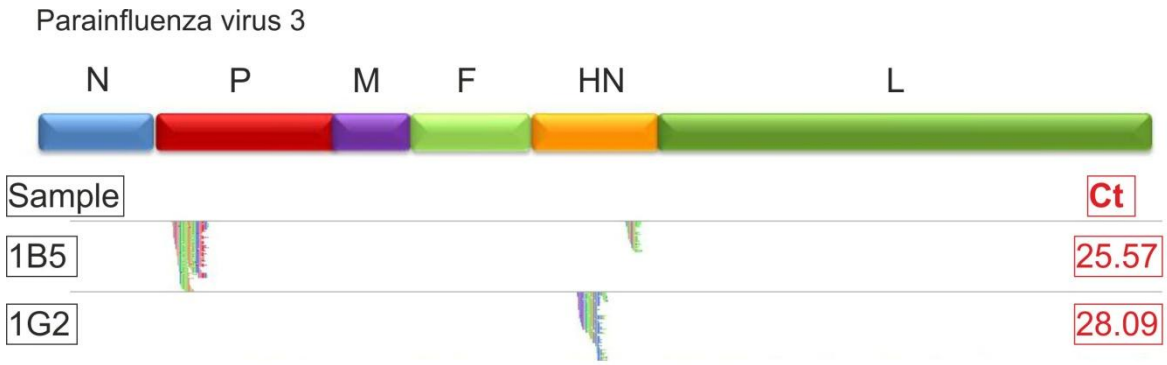


Figure 3-5. Sequencing coverage of the parainfluenza 3 genome.
Alignments were visualised using Tablet.

3.4.2.7 Coronaviruses

Using an NGS approach, three different species of coronaviruses were detected in this cohort, OC43, 229E and NL63.

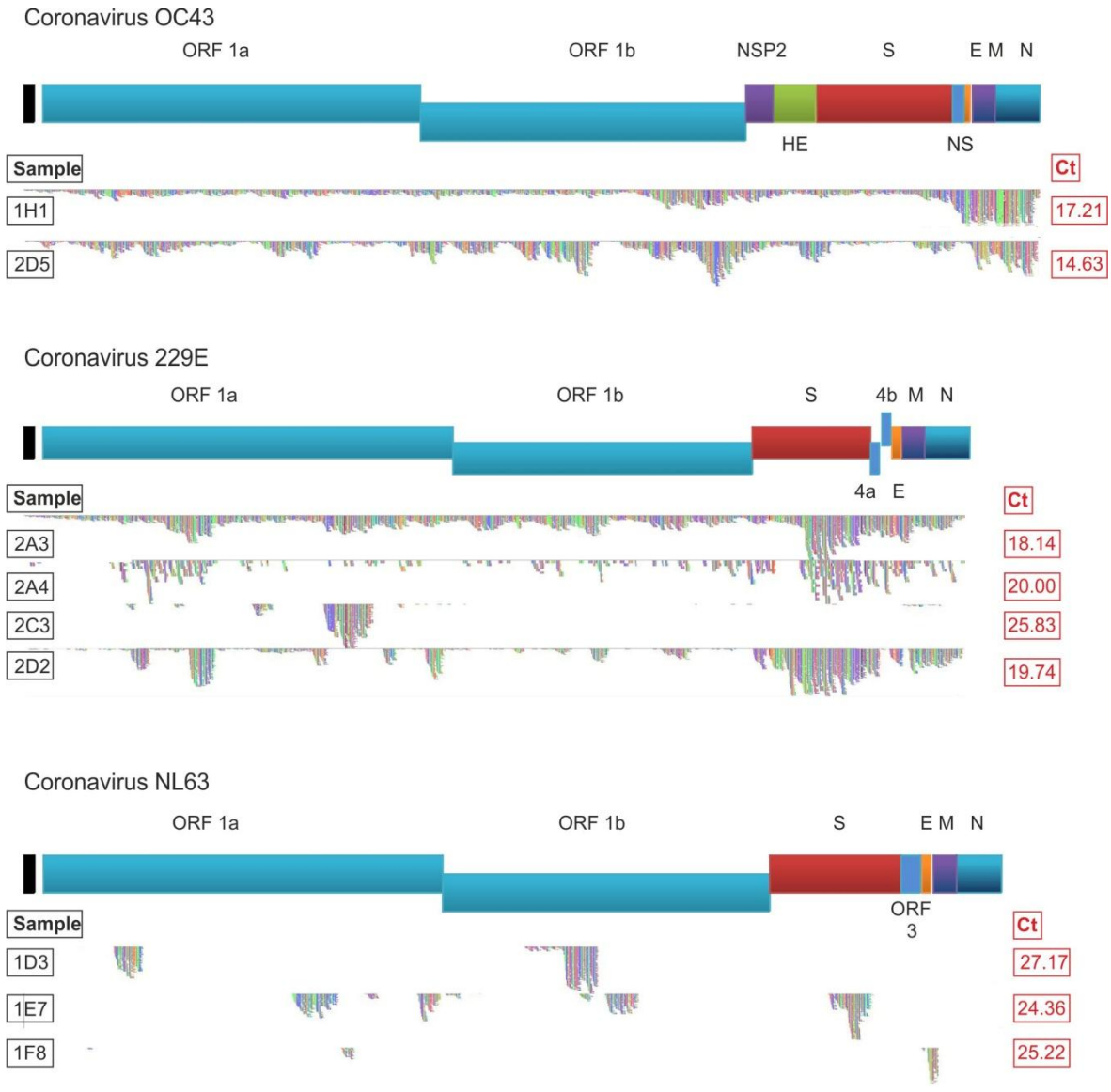


Figure 3-6. Sequencing coverage of the coronavirus genomes.
Alignments were visualised using Tablet.

3.4.2.7.1 Coronavirus OC43

Two cases of HCoV-OC43 were identified in this cohort by NGS, in sample 1H1 and 2D5, generating near full reference genome coverage at 99.9 and 97.6% respectively. The RT-PCR Ct for these samples was 14.63 and 17.21. The consensus sequences generated showed low levels of nucleotide mismatch at 0.7 and 1.6% respectively (accessions KF530099 and JN129835).

One RT-PCR positive case of HCoV OC43 was not detected by NGS. The Ct for this specimen was 24.05.

3.4.2.7.2 Coronavirus 229E

Four cases of HCoV-229E were identified in this cohort by NGS, in sample 2A3, 2A4, 2C3 and 2D2, generating reference genome coverage of 99.8, 55.3, 20.0 and 73.9% respectively. The RT-PCR Cts for these specimens were 18.14, 20.00, 25.83 and 19.74. The lowest genome coverage was in sample 2C3, the sample with the highest Ct value by PCR, implying a lower viral load in the initial clinical sample. Even though there was a low breadth of reference coverage in some cases there remained a high level of similarity to the closest reference sequence with nucleotide mismatch at 0.6 – 0.7% (accession JX503060).

3.4.2.7.3 Coronavirus NL63

Three cases of HCoV-NL63 were detected by NGS in this cohort, in sample 1D3, 1E7 and 1F8, generating reference genome coverage of 11.7, 23.3 and 3.7% respectively. The RT-PCR Cts for these specimens were 27.17, 24.36 and 25.22. In these cases the genome coverage was sporadic but the areas with coverage showed a high level of nucleotide similarity to the closest reference, with nucleotide mismatch rates of 0.4 – 0.5% (accessions KF530112 and JQ765569).

3.4.3 Orthomyxoviridae

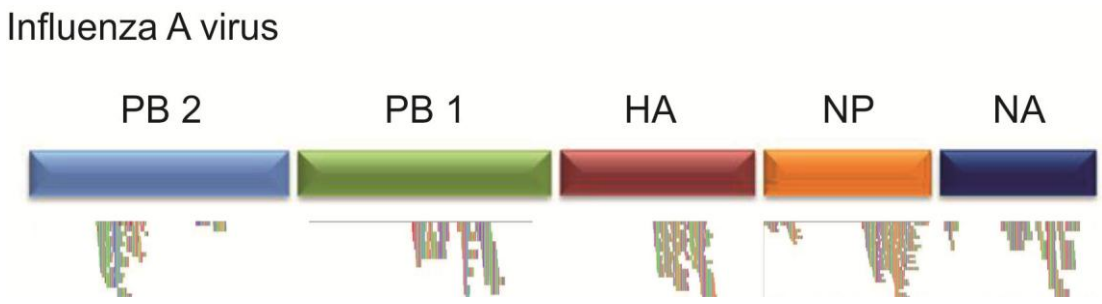


Figure 3-7. Sequencing coverage of the detected segments of the influenza genome. Alignments were visualised using Tablet.

Influenza A was detected in a single specimen. The RT-PCR Ct of this specimen was 22.34. Assembled contigs returned BLAST hits to five of the eight viral RNA segments (PB2, PB1, HA, NP and NA) arising from H3N2 subtype references. Following reference assembly only partial reference coverage was obtained at

low coverage, therefore the amino acid mutations associated with drug resistance or high pathogenicity could not analysed.

3.5 NGS reads detected in PCR negative specimens

As a further quality control step, where specimens were found to have NGS reads mapping to viral sequence but were negative by RT-PCR, the reads were compared to all other PCR positive samples on the same sequencing run. When reads mapped to identical sequence positions in a RT-PCR positive sample these reads were then removed as probable contaminants. The table below shows the number of remaining mapping reads following this step.

Sample	Total reads (after quality trimming)	HRV	RSV	PF
1B3	41390	*0*		
1B4	32376	*0*		*0*
1B5	343780	*0*		
1B6	68520	*3*		
1C1	341430	*1*		
1C8	128244	*0*		
1C9	441800	*0*		
1D3	26720	*0*		
1D6	265146	*3*		
1E1	289930		*0*	
1E3	43508	*1*		
1E4	13896	*2*		
1E5	22364	*1*		
1E9	15958	*10*		
1F8	43578	*1*		
1G7	27822	*0*		
2A1	26476	*3*	*3*	
2A3	130726	*3*		
2B5	232046		*0*	
2B9	135998		*0*	
2D1	25336	*1*		

Table 3-4. The number of unique mapping reads in PCR negative samples following duplicate removal.

In these cases the proportion of viral reads in RT-PCR negative specimens following duplicate removal ranged from less than 0.01 to 0.06%. These specimens were deemed negative by NGS.

3.6 Comparison of NGS with RT-PCR

3.6.1 Diagnostic Test Evaluation – sensitivity and specificity

A total of 13 viral pathogens were tested by RT-PCR in 89 samples, equating to 1157 tests in total. To compare the efficacy of NGS as a diagnostic test with that of PCR sensitivity and specificity were evaluated. A true positive was defined as corresponding virus detection by NGS and PCR. A true negative was defined as no virus detection by NGS or PCR. A false negative was defined as lack of virus detection by NGS where the sample was found to be PCR positive and a false positive was taken as detection of viral reads by NGS where the sample was PCR negative.

True positive (a)	38	False positive (c)	12
False negative (b)	11	True negative (d)	1098

The sensitivity of a test can be defined as the probability that a test will be positive when the disease is present i.e. the true positive rate:

$$\frac{\text{True positives}}{\text{True positives + False negatives}} = \frac{a}{a+b} = \frac{38}{38+11} = 77.6\%$$

(95% CI 63.4 to 88.2)

The specificity of a test can be defined as the probability that the test will be negative in the absence if disease i.e. the true negative rate:

$$\frac{\text{True negatives}}{\text{True negatives + False positives}} = \frac{d}{d+c} = \frac{1098}{1098+12} = 98.9\%$$

(95% CI 98.1 – 99.4)

Statistical analyses were performed using MedCalc for Windows, version 12 (MedCalc Software, Ostend, Belgium).

The number of tests carried out here would be too small to prove the efficacy of NGS as a diagnostic test but these calculations, even with the small numbers suggest that there is potential that the efficacy of NGS could be competitive with PCR.

3.6.2 Quantitation of Target with NGS

The number of PCR cycles required to enable target detection above a threshold relates to the initial target quantity within the sample. This allows an estimate of quantitation however without standardised samples run in parallel quantitation cannot be absolute. The threshold cycle value (Ct) of respiratory samples positive by RT-PCR which were either concordant or discordant with the NGS assay results are shown in Figure 3-8. The bars indicate the mean and one standard deviation. An unpaired t-test demonstrates a significant difference between the two groups ($p<0.0001$) suggesting a relationship between viral quantity and likelihood of detection with NGS.

The mean Ct of the discordant specimens was ~ 32, suggesting this could be a test sensitivity cut off.

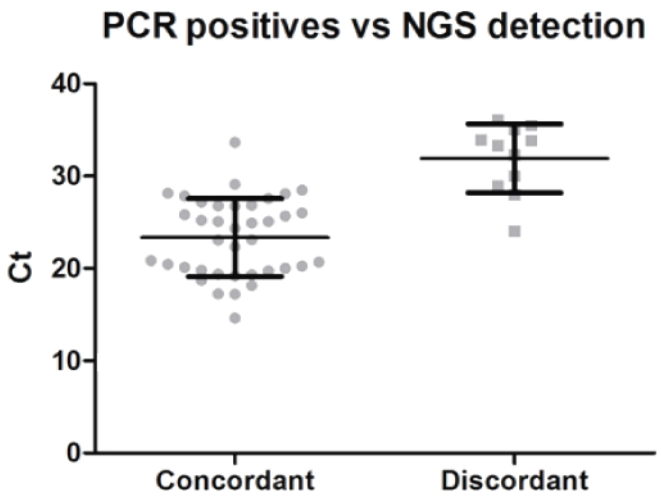


Figure 3-8. Ct value of PCR positive samples that were concordant or discordant with NGS diagnosis.

As the absolute number of sequenced reads varied between samples the overall proportion of reads mapping to a viral reference sequence was compared with the Ct value to determine if a similar semi-quantitative approach could be used in NGS (Figure 3-9).

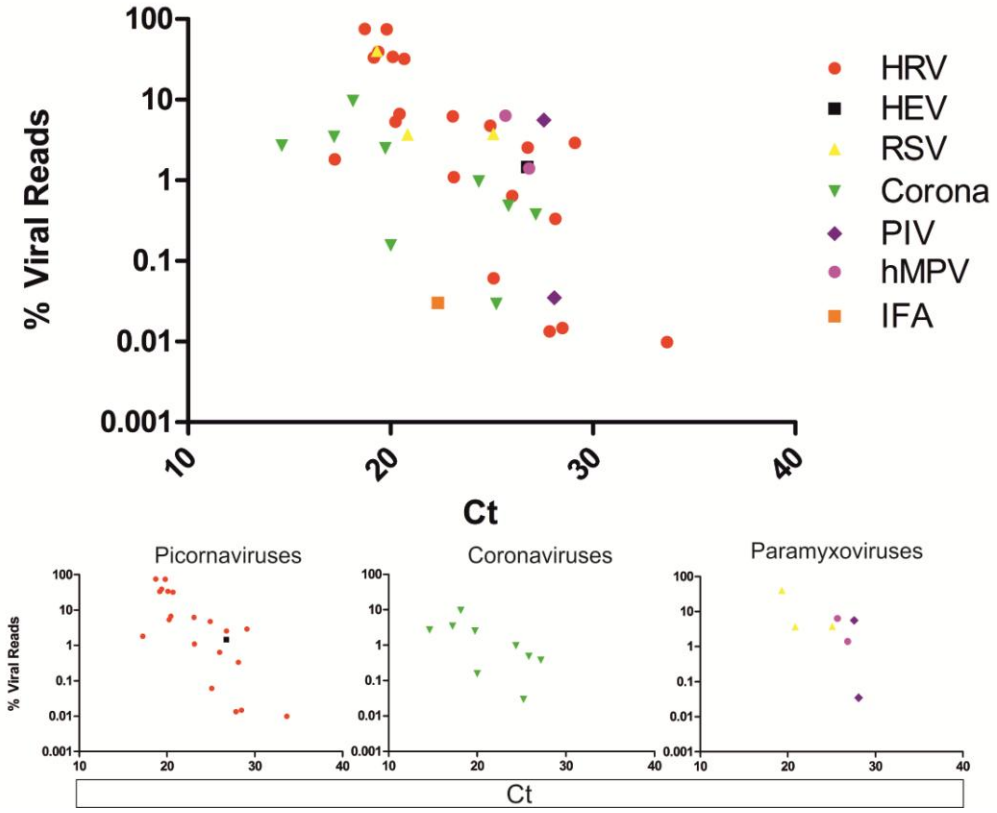


Figure 3-9. The proportion of sequenced reads mapping to viral reference.
The percentage of sequenced reads, after quality trimming, mapping to the taxonomic reference genome determined by BLAST. Linear regression of complete dataset, R^2 value = 0.21 (p = 0.0009). Linear regression of virus groups, Picornaviruses: R^2 = 0.37 (p = 0.0015), Coronaviruses: R^2 = 0.34 (p = 0.0471), Paramyxoviruses: R^2 = 0.40 (p = 0.0359).

The number of detections for each virus group was too small to prove any definitive correlation but the R^2 value following linear regression suggests there is a relationship between the proportion of reads mapping and the Ct value but this relationship varies between virus groups. As a consequence the lower limit of detection is also likely to vary between virus groups.

3.6.3 Detection of viral co-infection

The detection of viral co-infection using NGS could not be assessed in this cohort. A viral co-infection was identified by RT-PCR in a single sample (1G1).

HRV and AdV were detected in this sample with Ct values of 28.5 and 35.1 respectively. The AdV was not detected by NGS in this case however it could not be determined if this was due to sample processing methods or due to the low viral load as previous analysis suggested detection cut off in the region of Ct 32. The DNAse step early in sample preparation may preclude detection of viral genomic DNA; however the detection of mRNA using RT-PCR has been demonstrated previously (Ko, Cromeans et al. 2003) therefore it was hoped that the presence of a DNA virus could still be detected with this method.

3.7 Discussion

In this research sequence independent PCR with NGS was carried out with the aim being to determine if this method could be used to detect viral pathogens in clinical samples. The results from this method were compared with the current diagnostic method of RT-PCR to establish if it would be possible to use this as an alternative diagnostic test.

Viruses were detected with a high rate of concordance between the two methods (77.6%). Multiple different viruses were detected using a single NGS workflow with a sequence independent approach, in contrast with PCR which relies on targeting specific nucleotide sequences, and therefore requires numerous primer/probe sets to detect a panel of pathogens predetermined by the user.

The virus quantity within clinical samples, as determined by using the PCR Ct value as a proxy, correlated with both the likelihood of detection and the proportion of NGS sequenced reads mapping to the virus. A linear regression model of the small number of viral detections in this study shows there is a relationship between the proportion of sequenced reads of viral origin and the RT-PCR Ct. These data would suggest a lower level of detection with a cut off equivalent to a Ct of 32. This suggests the possibility of using NGS for quantitation as well as detection. In general quantitation of respiratory pathogens is not required in diagnostics yet in certain cases this can be valuable, such as monitoring response to therapy in adenovirus or influenza infections, particularly in those who are immunocompromised. Quantitation may also be important when considering the use of NGS in detecting bacterial pathogens

which are known to be part of the commensal microbiotia but presence at a high level may indicate pathogenic overgrowth.

The detection of viral pathogens by PCR relies on targeting conserved portions of the viral genome and as a result, in most cases, can only identify the virus to species level. Using a NGS approach, in some instances full genome sequences were obtained and in many cases protein coding regions were sequenced, aiding in molecular genotyping of the viral pathogen where appropriate. The value of this information in a diagnostic laboratory is not clear as the clinical impact of specific subtypes has not been proven.

In comparison with RT-PCR the sensitivity of NGS was lower, detecting fewer positive results from clinical samples. The detection of co-infections could not be analysed in this cohort of samples as only one sample contained greater than one virus by PCR.

The method used here enriched for RNA and would likely preclude the detection of DNA viruses. While the majority of known viral respiratory pathogens are RNA viruses there are DNA viruses which are established or proposed pathogens of the human respiratory tract such as adenoviruses and human bocaviruses. This has further implications when proposing to use this technique as a pan-microbial diagnostic test as bacterial and fungal pathogens possess a DNA genome. It is possible that RNA transcripts could be detected but adapting the methodology for the detection of DNA as well as RNA would be essential.

3.7.1 Future work

A major barrier to using this technique with the method outlined above is the sensitivity in comparison to RT-PCR. Many approaches could be applied to increase sensitivity and these require being trialled to determine what would be effective with clinical samples in a diagnostic setting. The introduction of further processing steps will increase both time and cost associated with sample processing but if this allows higher sample throughput it could offset the increased cost.

Alterations or additional steps would also be required to permit detection of DNA and RNA, ensuring a true representation of the known respiratory viral

pathogens. The ability to sequence both RNA and DNA in a single workflow would then make the prospect of pan-microbial diagnostic test a real possibility.

Analysis of samples containing more than one pathogen would be required to determine if there are further factors which may affect the likelihood of pathogen detection. It is probable that the presence of one pathogen at high levels may affect the lower limit of detection of further pathogens.

The established method of RT-PCR can generate results in a matter of hours. The speed of the NGS process would need to be improved before being implemented as a diagnostic test. When undertaken manually many of the steps, in particular clean-up processes are time-consuming. If undertaken in a diagnostic laboratory it is likely that automation could be applied, reducing manual intervention and streamlining the process.

The initial data analysis steps in this research were undertaken as part of a pipeline but the more in-depth analysis required user input. Much of this work could be implemented into a complete purpose-built diagnostic pipeline. A lot of information can be generated in this process, not all of which is useful to a diagnostician or clinician; in fact the output of processes such as BLAST could potentially be confusing for those without prior training in the area. Generation of an analysis pipeline which takes into account the salient information required for diagnostic purposes would ultimately be needed for the end users. In-depth analysis of viral subtypes could then be undertaken at a later stage as this is unlikely to have an impact on the patient in real time but is useful for epidemiological purposes.

The cost of NGS is decreasing but remains considerably higher than that of PCR. Increasing throughput, optimising protocols and reagents will contribute to reducing the cost of the process. The potential for pan-microbial detection from a single sample would greatly increase the cost effectiveness of NGS in a diagnostic setting and a single workflow could potentially combine the multiple processes which are currently required to detect viruses, bacteria and where appropriate, fungi, from clinical specimens.

The Use of an NGS pipeline for the Epidemiological Study of a Respiratory Pathogen

4.1 Introduction

The research presented in chapter 3 demonstrates that an NGS approach can be used to detect multiple viral pathogens from clinical respiratory samples using a single sample workflow without *a priori* selection of a target. The detection rate using this method was found to be similar to RT-PCR, the current gold standard diagnostic test. There was also evidence that the proportion of sequenced reads mapping to viral reference correlated with the RT-PCR Ct value. Although the number of specimens in the panel was too small to prove a definitive link this suggests that the NGS method may also be able provide semi quantitative data. The results outlined in chapter 3 also showed that the sequence data provided by the NGS method could be used for typing the detected virus and therefore may also be useful for epidemiological analysis. Unfortunately the panel tested in Chapter 3 was not large enough to determine the utility of the NGS method in this context.

This chapter describes the results of using the NGS method to study the epidemiology of Enterovirus 68 – a re-emerging pathogen – in the West of Scotland. Human enterovirus 68 (EV-D68) belongs to the genus Enterovirus and family Picornaviridae and possesses a positive sense single-stranded RNA genome approximately 7500 bases in length (Jacobs, Lamson et al. 2013). There are seven species of enterovirus known to cause infections in humans, enterovirus A, B, C and D and rhinovirus A, B and C. Within the group D enteroviruses there are four recognised serotypes, D68, D70, D94 and D111 which infect humans and a fifth, D120, associated with infection of gorillas (Holm-Hansen, Midgley et al. 2016). EV-D68 was first isolated in the 1960s in the United States from children with acute respiratory tract illness (Schieble, Fox et al. 1967). From then on only a small number of cases were reported annually. From 2000 onwards there has been a marked increase in the number of reported cases, now documented in several countries such as Germany, Japan and Mexico (Kaida, Kubo et al. 2011; Ly, Tokarz et al. 2014; Bottcher, Prifert et al. 2016; Vazquez-Perez, Ramirez-Gonzalez et al. 2016).

The clinical symptoms associated with EV-D68 were thought to range from asymptomatic to a mild respiratory illness. However, the recent outbreaks have highlighted a potential to cause more severe infections (Messacar, Schreiner et al. 2015). For example, during 2014 a large multistate EV-D68 outbreak occurred in the US, which resulted in a large number of severe lower respiratory infections in children and a possible association with acute flaccid paralysis.

During 2015 a pan-European study was carried out to examine the prevalence, clinical presentation and epidemiology of EV-D68 in European countries. The WoSSVC in Glasgow took part in this study. Between 1st July 2014 and 1st December 2014 all respiratory samples from children under 16 years of age found to be rhino/enterovirus positive by RT-PCR were screened for EV-D68 with a type specific PCR (Poelman, Scholvinck et al. 2015). During this time all enterovirus positive (by PCR) CSF samples were also screened for EV-D68. Sanger sequencing of the enterovirus VP1 segment was then carried out by the WoSSVC to subtype any positive samples.

In total EV-D68 was detected in 22 of the 488 respiratory samples (4.51%) screened. Interestingly nine samples also contained other viral respiratory pathogens. This level of EV-D68 prevalence is in line with those reported in other European studies (Poelman, Schuffenecker et al. 2015). The clinical details associated with these samples are outlined in appendix 3. EV-D68 was not detected in any CSF samples. This is in keeping with results from previous studies which did not demonstrate the presence of the virus in CSF, even in those with neurological symptoms (Greninger, Naccache et al. 2015).

The positive samples from the aforementioned work were then sequenced by the NGS method in order to provide a direct comparison.

4.2 Aim

The aim of this chapter was to examine the utility of the NGS pipeline (chapter 3) within an epidemiological context. EV-D68 positive samples from the work described above were then sequenced by the NGS method. We evaluated whether the NGS pipeline could correctly detect and type enterovirus positive samples. This panel also provided a further opportunity to assess whether the

NGS pipeline could detect additional respiratory pathogens from a clinical specimen. Finally we compared the sequence data generated by the NGS pipeline to that provided by the Sanger method. The results are presented and discussed below.

4.3 Methods

4.3.1 Samples

The clinical details associated with the samples are presented in Appendix 3. Nucleic acids were extracted from clinical samples using the EasyMag platform (BioMeriaeux) at WoSSVC and stored at -20°C until tested.

4.3.2 Real Time PCR

An EV-D68 specific RT-PCR was carried out using the primers, probes and parameters as described previously (Poelman, Scholvinck et al. 2015). This reaction utilises highly specific primers targeting the 5'-NTR of the EV-D68. This was carried out at WoSSVC.

4.3.3 Sanger Sequencing

A nested PCR approach was used to amplify and sequence a 316 bp segment of the VP1 coding region as described by Nix et al (Nix, Oberste et al. 2006). In samples where an appropriately sized amplicon was identified, Sanger sequencing was then carried out at WoSSVC.

4.3.4 Sample Preparation for Next Generation Sequencing

The methods used are described in detail in chapter 2. Briefly, the frozen extract was thawed in a water bath at 37°C and 10 µl was used as the template in a reverse transcription reaction. Second strand synthesis was carried out, followed by a SISPA step. The resulting DNA was clarified using AMPure XP and 1 ng of DNA used to prepare sequencing libraries (Nextera XT DNA Sample Prep kit, Illumina). Sample libraries were indexed and multiplexed using the Illumina Nextera indices and sequenced using an Illumina MiSeq to generate 150 base paired-end reads.

4.3.5 Data Analysis

4.3.5.1 Quality Trimming

Indices and primers were removed from sequenced reads using Trim Galore!. Reads with an average Phred quality score greater than 30 and length greater than 30 bp were retained and aligned to a database of human reference sequences (Bowtie2 (Langmead and Salzberg 2012)). The reads which did not map to the human references were extracted from the dataset using SamTools and used for further analysis.

4.3.5.2 Reference Based Detection and Consensus Generation

All EV-D68 full genome sequences freely available at the time of research (16/6/2015) were obtained and used as a database for a reference based assembly. The accession numbers of sequences used can be found in appendix 4. Sequenced reads remaining following quality trimming and alignment to human reference sequences were mapped to this database. The reference sequence with the greatest number of reads mapping was then used as a template for a reference based assembly against that sequence only. Multiple alignment tools were used for each sample to generate alignments (SAM files) and the alignment with the greatest reference coverage was used to generate a consensus sequence using a PERL script developed at the CVR (J. Hughes, unpublished). After choosing a virus reference sequence, all reads were then realigned to this single reference increasing both the number of reads mapping and the breadth of reference coverage.

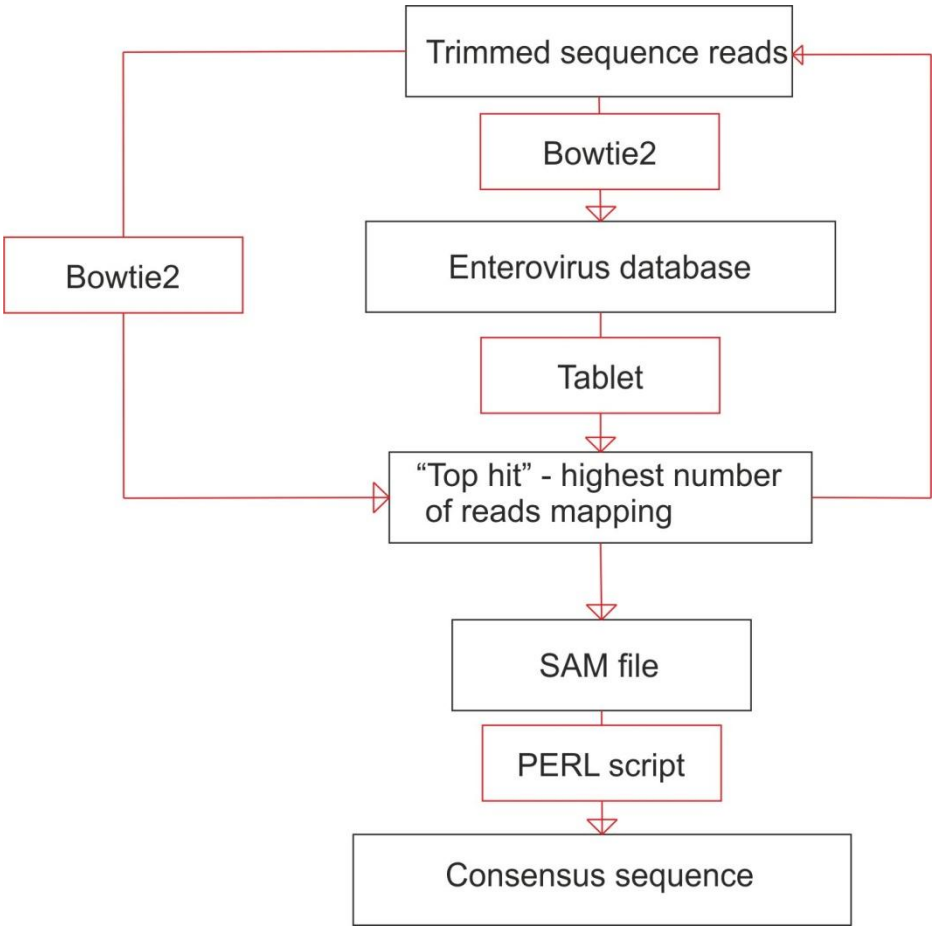


Figure 4-1. Workflow of reference based virus detection and consensus sequence generation.

4.3.5.3 Virus Detection Pipeline

Sequenced reads which did not map to a human reference were entered into the virus detection pipeline MetAmos (Treangen, Koren et al. 2013), using the Bowtie2 alignment tool, multiple *de novo* assembly programs (IDBA-UD, SparseAssembler and Velvet) and annotation carried out using BLAST. The presence of viral contigs was determined through visual inspection of the Krona output chart (see appendix 6 for charts).

4.3.6 Phylogenetic Analysis

Nucleotide sequences were aligned using MUSCLE as part of MEGA6 (Tamura, Stecher et al. 2013). Phylogenetic analyses were also carried out using MEGA6. Neighbour-joining and maximum likelihood trees were generated using the bootstrap test of reliability with 1000 bootstrap replicates. Trees were inferred using the best fit model as determined by the Bayesian Information Criterion score.

4.3.7 Recombination Detection

An alignment of all available complete genome sequences and the consensus sequences with greater than 80% coverage was generated in MEGA6. This alignment was then used in recombination detection using RDP4 (Darren P. Martin 2015).

4.3.8 Comparison of Sanger Sequencing and NGS

For specimens with both Sanger and NGS sequences available, these were aligned with each other using MEGA6. A pairwise comparison of nucleotides was calculated using CLC Genomics Workbench 6. The nucleotide alignments were then translated to amino acid sequences to assess for non-synonymous amino acids differences between the two methods. Phylogenetic trees were then inferred using the sequences from each method to determine if these differences resulted in a change of topology (MEGA6).

4.4 Results

4.4.1 Virus Detection and Identification

The metagenomics pipeline identified enterovirus contigs from the majority of clinical specimens tested, 19 of 22. The Cts from the clinical specimens where enterovirus contigs were not detected were 29.75, 32.76 and 32.82. A comparison of the Cts between specimens with concordant and discordant results by RT-PCR and NGS are shown in Figure 4-2. The number of reads mapped and breadth of reference coverage for each sample is detailed in Appendix 5. The top hit reference sequences determined by a BLAST search against a nucleotide database were the same in numerous cases, which would suggest similar viruses in these individuals. The reference with accession KP745768 was identified in 11 cases and KP745769 in eight cases. Of interest, from the remaining three where enterovirus contigs were not detected by the pipeline, enterovirus reads were identified by direct alignment to reference sequences. However the numbers were too small (e.g. 5. 35 and 51) to be considered relevant and were not studied further.

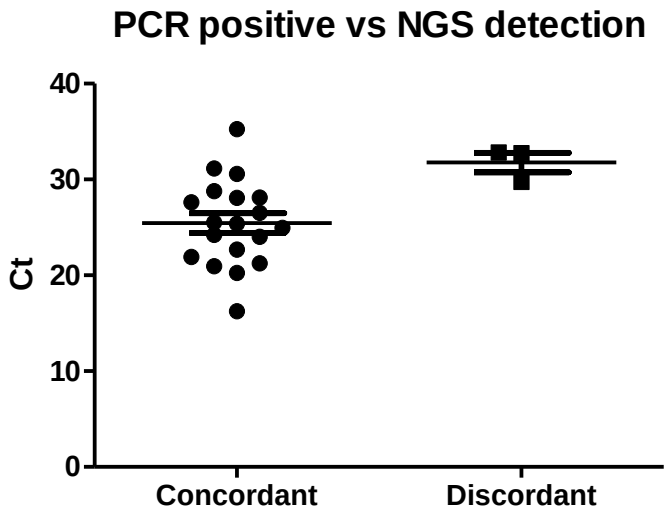


Figure 4-2. Ct value of EV-D68 RT-PCR positive specimens that were concordant or discordant with NGS diagnosis.
The EV-D68 RT-PCR positive specimens where viral contigs were not detected by NGS have a higher Ct with a significant difference between the two groups (P = 0.027). These data would suggest a sensitivity cut-off of Ct = 32.

The RT-PCR Ct related to the likelihood of detection of viral contigs by NGS. The specimens with concordant results by both methods had lower Cts, therefore a higher viral load. In specimens with discordant results i.e. no viral contigs detected by NGS the mean Ct was 31.78 which could be considered the lower limit of detection.

As mentioned above, the samples used in this study had initially been subjected to the full diagnostic panel for respiratory viruses. In 9 of the 22 an additional viral pathogen was detected by RT-PCR. The additional viruses detected are shown in Table 4-1 and include adenovirus, HHV-6, hMPV, PIV-4 and RSV. The NGS method failed to detect the DNA viruses AdV and HHV-6 with Cts of 22.34, 35.51 and 35.89. NGS also failed to detect some RNA viruses including one hMPV (Ct 37.73) and three HPIV4 detections (Cts of 26.08 – 34.53). In specimen 429519, EV-D68 was not detected by NGS though RSV-A was detected with greater than 97% reference genome coverage. EV-D68 and RSV-A were both detected in two cases, 430139 and 429319.

Interestingly, additional non EV-68D enterovirus reads were detected in a further four samples. In three of these, reads mapping to coding regions which are unique to the serotype were obtained. This suggested concurrent infection with

multiple enterovirus species, HRV-A80, HRV-C40 and HRV-C species. In one sample reads mapping to multiple 5’ UTR sequence of HRV-B were detected but as this region of the viral genome is conserved between serotypes it does not confirm the presence of an additional virus.

Sample	EV-D68 Ct	Additional Virus by RT-PCR	Ct	NGS EV-D68 detected	Virus detection pipeline	Accession	Coverage
428891	*21.92*	*Adenovirus*	*22.34*	**+**			
429110	*32.76*	*HHV-6*	*35.89*	**-**			
430146	*21.26*	hMPV	*37.73*	**+**			
430741	*20.24*	HPIV4	*26.08*	**+**			
430038	*24.96*	HPIV4	*33.54*	**+**			
428005	*20.94*	HPIV4 *Adenovirus*	*31.38* *35.51*	**+**			
429159	*32.82*	**RSV**	*18.27*	**-**	**RSV-A** Parecho-3	**JF920053**	**97.85%**
430139	*30.58*	**RSV**	*18.36*	**+**	**RSV-A**	**JF920053**	**99.98%**
429319	*25.50*	**RSV** HPIV4	*26.41* 34.53	**+**	**RSV-A**	**JF920052**	**16.93%**
429323	*16.25*			**+**	**HRV-B**	**Multiple 5' UTR**	
429660	*28.1*			**+**	**HRV-C**	**JN815251**	**100%**
430038	*24.96*			**+**	**HRV-A**	**FJ445156**	**4.62%**
429915	*28.07*			**+**	**HRV-C**	**HM236934**	**Partial (VP1)**

Table 4-1. Detection of additional viruses in EV-D68 positive samples.

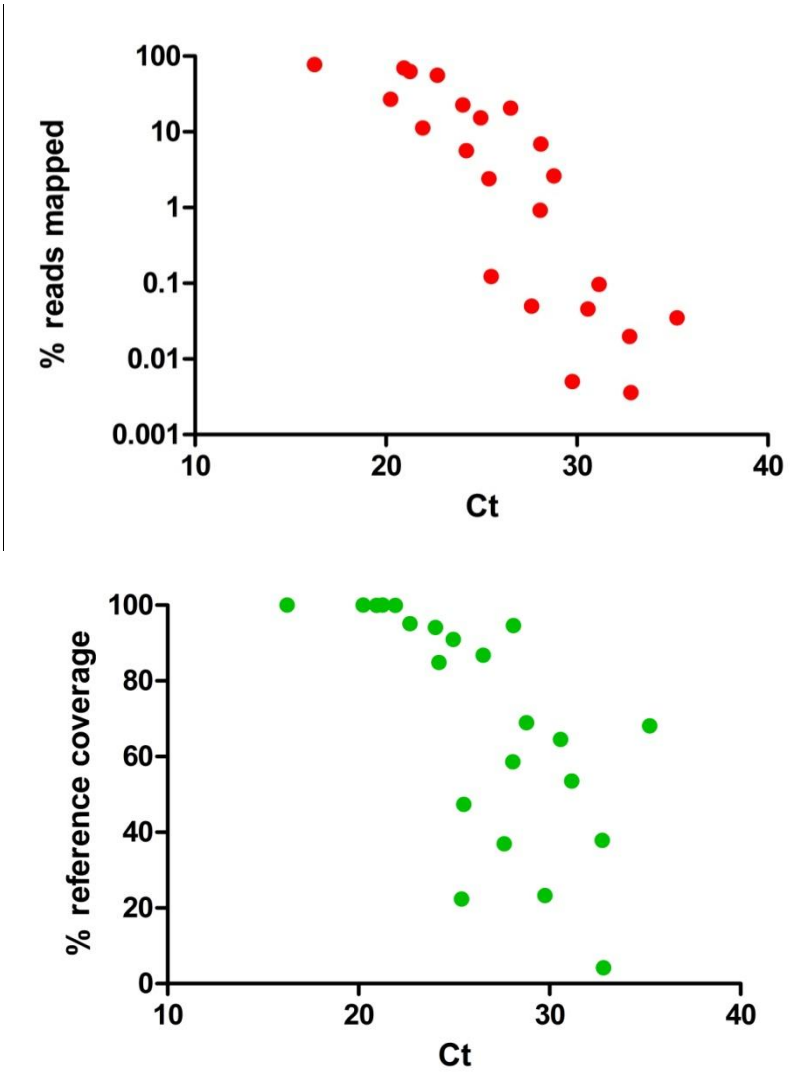


Figure 4-3. The correlation of Ct value with sequenced reads and reference genome coverage. The top panel shows the correlation of Ct value and the proportion of sequenced reads mapping to a viral reference. Linear regression, $R^2 = 0.60$ ($p < 0.0001$). The bottom panel shows the correlation of Ct value and the percentage of reference genome coverage obtained. Linear regression $R^2 = 0.45$ ($p = 0.0006$).

The relationships of the EV-D68 RT-PCR Ct with the proportion of mapped reads to EV-68D and reference coverage are shown in Figure 4-3. This demonstrates that in clinical samples with a low Ct i.e. a high viral load, a large proportion reads are of viral origin however this proportion decreases with a rise in Ct. Linear regression of this relationship shows an R^2 of 0.60 suggesting a relationship between these values. A similar trend is seen with the breadth of reference coverage. For example, in sample 429323 with a Ct of 16.25 100%

coverage of the reference genome was obtained, however in sample 428008 with a Ct of 28.79 this was 68.99%. This distribution of genome coverage is detailed in Figure 4-4. Again this shows that in specimens with a low Ct, for example 429323 and 430741 (Ct 16.25 and 20.24) there is broad coverage of the reference sequence with especially deep coverage of the VP1 region. In specimens with a higher Ct, for example 428129 and 428008 (Ct 25.39 and 28.79) the reference coverage is incomplete and sporadic.

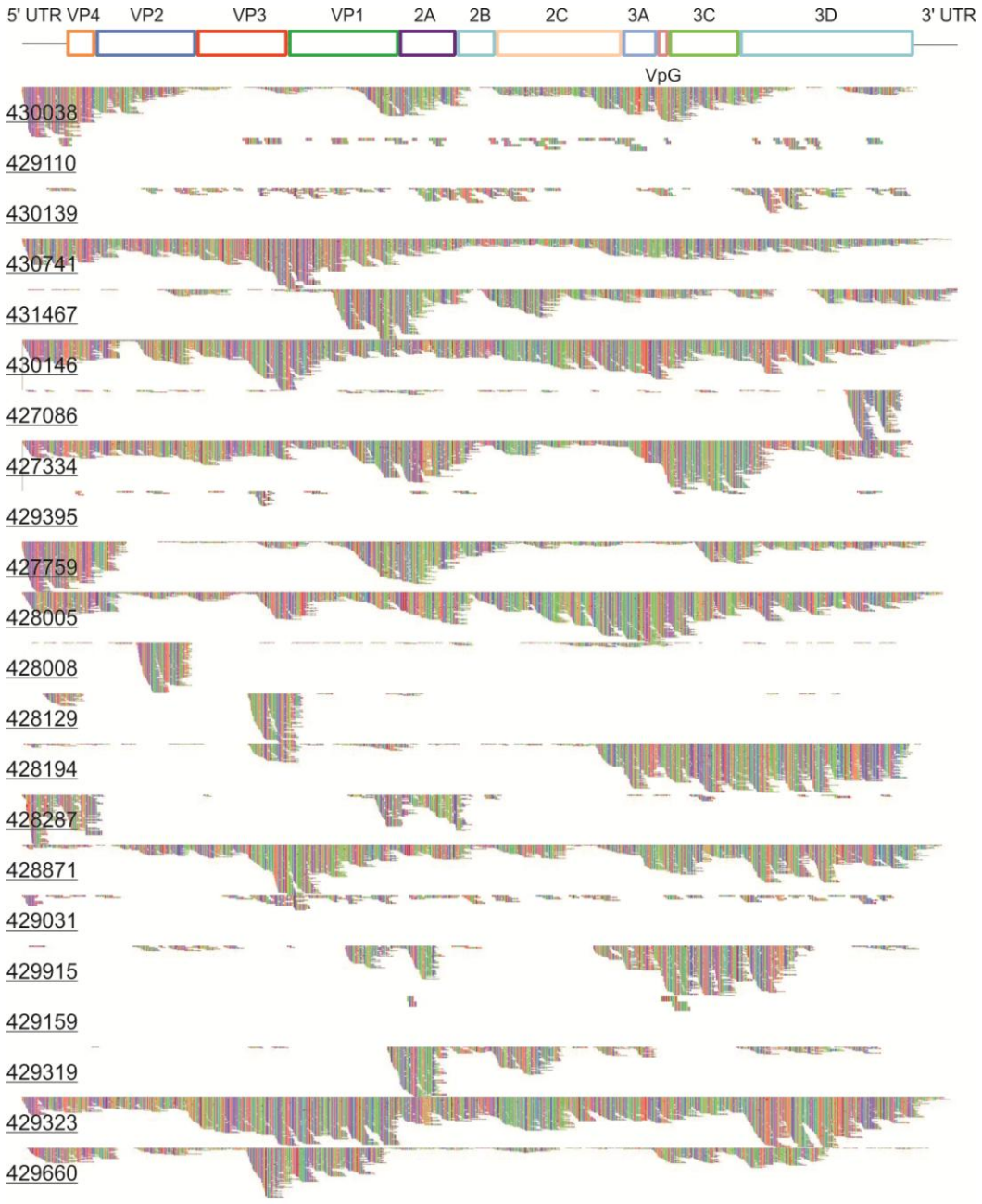


Figure 4-4. Coverage of enterovirus reference genome by sequenced reads.
This image shows a visual output of the alignment of sequenced reads against an annotated schematic of the reference genome for scale.

This was consistent with previous findings that the viral quantity in the sample correlates with likelihood of detection and proportion of viral reads sequenced. Greater than 80% reference genome coverage was obtained in cases with a Ct value lower than 25 (Figure 4-3). In samples with a higher Ct value, the breadth of coverage was sporadic and was not guaranteed to generate information from protein coding regions (Figure 4-4).

4.4.2 Phylogenetic Analysis

Assembled sequences with greater than 90% reference genome coverage were generated from eight of 22 clinical samples. These were added to an alignment of reference genomes and used for subsequent phylogenetic analysis. Figure 4-5 shows that all eight isolates belong to clade B and share a common origin.

Full VP1 coverage was available in 13 of the 22 samples. This included the 8 samples with >90% genome coverage (see above) and a further 5 samples. In a further four specimens a partial VP1 coverage was generated. Additionally, VP3 coverage was generated in a further case, 428192; therefore capsid protein coding information was generated in 18 of 19 cases detected by NGS. No capsid protein coding regions were sequenced from specimen 429319.

The VP1 sequences generated here were then aligned with reference VP1 sequences, including those available from the flaccid paralysis cases during the aforementioned US outbreak. This again confirmed that the majority of cases clustered in clade B. One case, 428194, was shown in clade B1 and was closely related to the sequences obtained from cases with acute flaccid paralysis in the recent US outbreak (Figure 4-6).

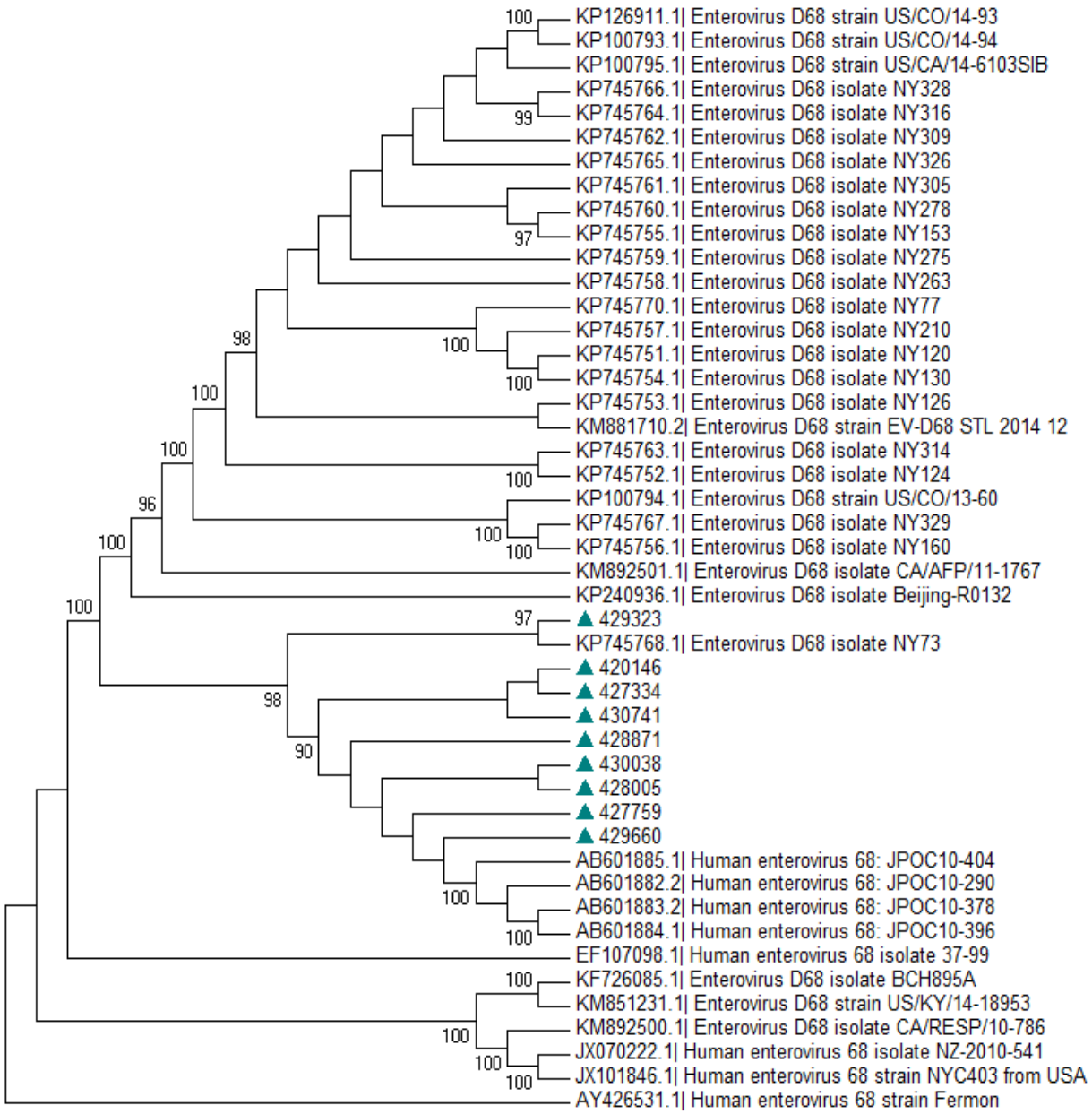


Figure 4-5. Enterovirus D68 complete genome phylogenetic analysis.

This neighbour joining tree was inferred using the Tamura-3 parameter model with 1000 bootstrap replicates. Bootstrap values greater than 90 are shown next to the nodes on the tree. The analysis involved 46 nucleotide sequences. There were a total of 7382 positions in the final dataset. Evolutionary analyses were conducted in MEGA6. The samples from this study are highlighted with teal triangles.

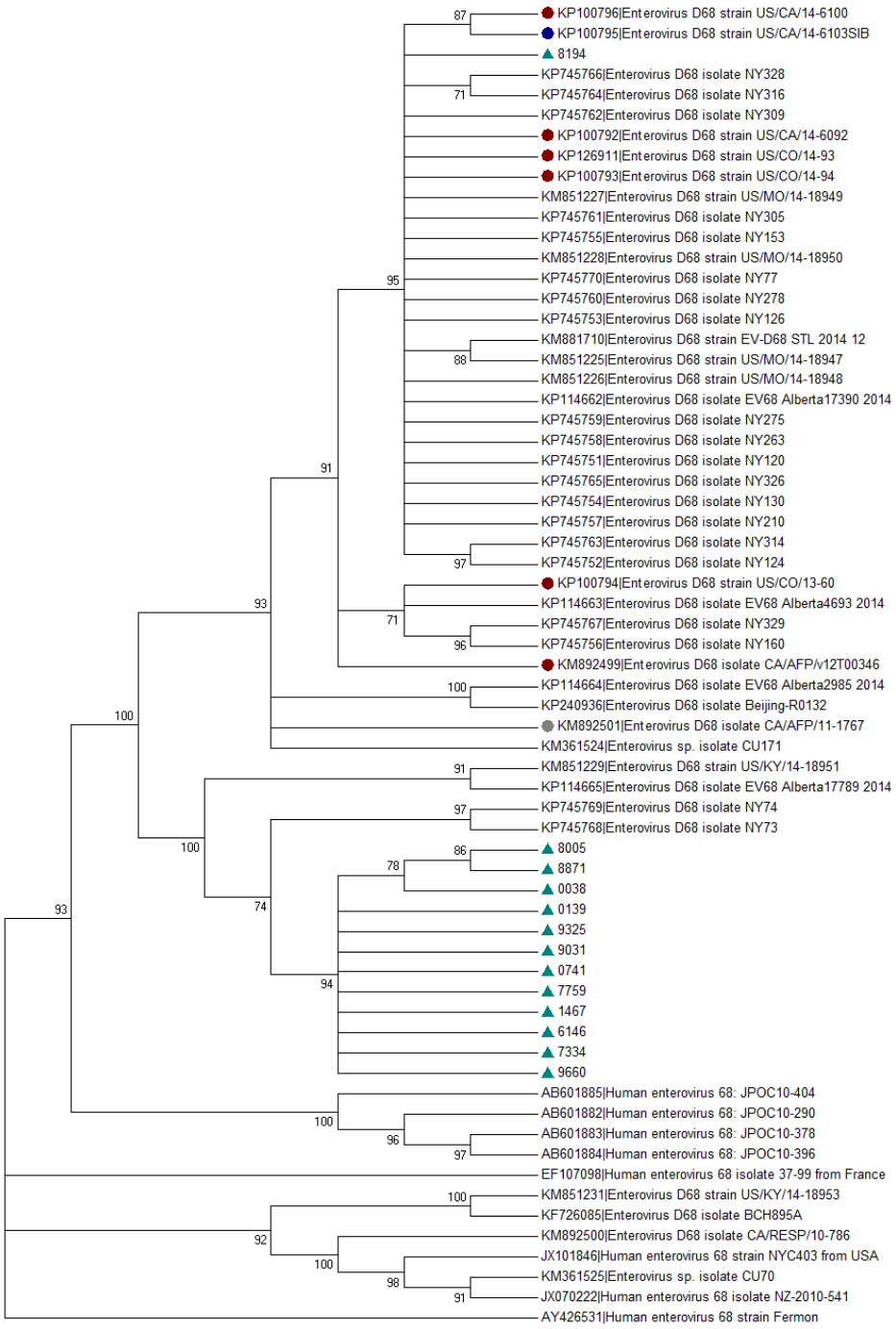


Figure 4-6. Enterovirus D68 VP1 phylogenetic analysis.

This Maximum Likelihood tree was inferred using the Tamura 3-parameter model with 1000 bootstrap replicates. Bootstrap values are shown next to the nodes of the tree. The analysis involved 65 nucleotide sequences. There were a total of 922 positions in the final dataset. Evolutionary analyses were conducted in MEGA6. The isolates from this study are highlighted with blue triangles. The isolates highlighted with red markers were from patients with neurological symptoms and those with blue or grey markers had respiratory symptoms only.

4.4.3 Recombinant Detection

The results obtained using RDP4 analysis indicated that no recombination events were detected among the EV-D68 virus sequences obtained from these clinical samples.

4.4.4 Comparison with Sanger sequencing

The Sanger sequencing method was applied to the same cohort of clinical samples as part of a separate study (Poelman, Schuffenecker et al. 2015). This method required nested PCR (two reactions) followed by a sequencing reaction, commonplace methods in large diagnostic centres. This targeted a 316 bp portion of the hypervariable region of VP1, commonly used in typing of enteroviruses (Nix, Oberste et al. 2006).

Sequence information was successfully obtained from 17 of 22 clinical samples. Phylogenetic analysis of this short fragment shows clinical samples within multiple clades (Figure 4-9). One lies within clade A and of the remaining 16, 14 cluster together in clade B and two cases fall within clade B1.

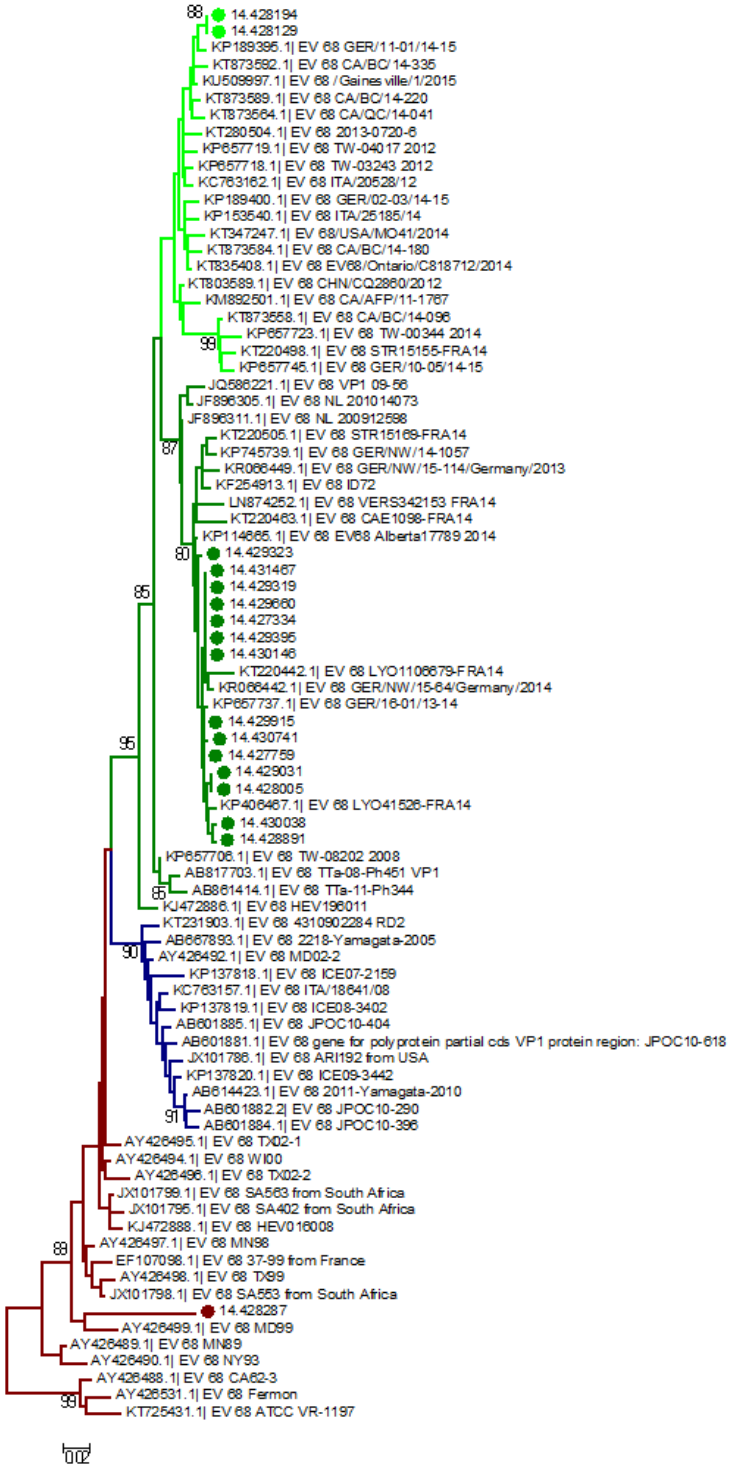


Figure 4-7. Phylogenetic analysis of EV-D68 VP1 sequences generated by Sanger sequencing.

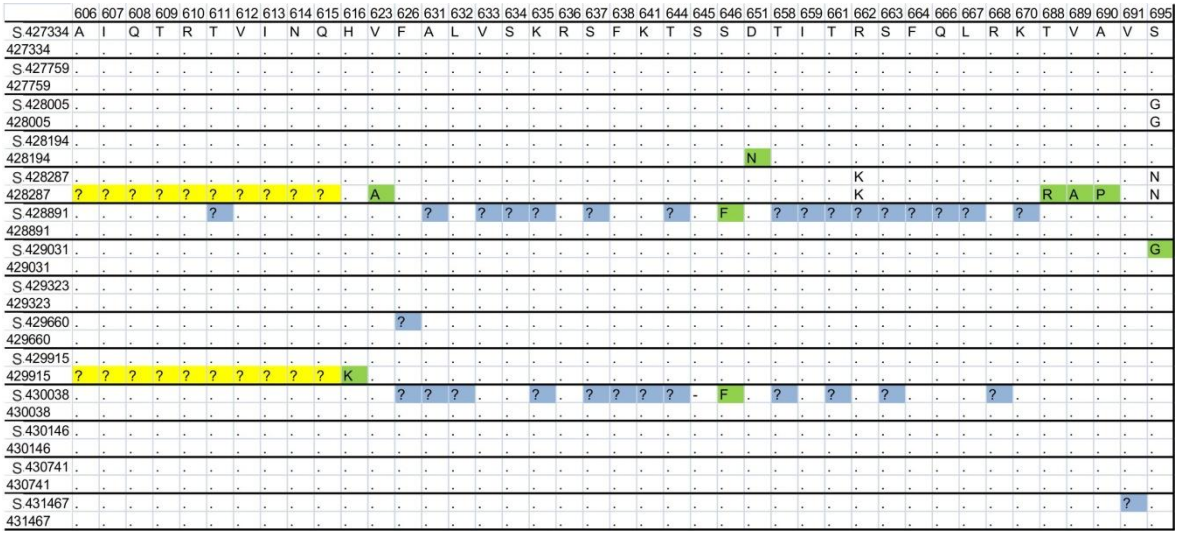
This Maximum Likelihood tree was inferred using the Tamura 3-parameter model with 1000 bootstrap replicates. Bootstrap values are shown next to the nodes of the tree. The analysis involved 65 nucleotide sequences. There were a total of 316 positions in the final dataset. Evolutionary analyses were conducted in MEGA6. The isolates from this study are highlighted with circles. Clade A sequences are in red, clade C in blue and clade B is shown in green with B1 in light green.

As discussed earlier, NGS identified full VP1 coverage from 13 specimens. The Sanger method used here targets a 316 bp hypervariable region of the protein. This region was identified from 17 specimens using NGS. Both methods identified the region from 14 specimens. Three were identified by Sanger alone (429395, 428192 and 429319) and three were identified by NGS alone (430139, 427086 and 428008). A direct comparison of the VP1 sequences generated by each method, where available, showed multiple differences at a nucleotide level (Table 4-2). Up to 6 nucleotide differences were identified in six of 14 cases.

Specimen	Nucleotide differences	Ambiguous base calls by Sanger	NGS gaps
427334			
427759			
428005			
428287	6		18
428891	2	17	
429031	2		
429323			
429660		1	
429915	2		18
430038	4	14	
430146			
430741			
431467		1	
428134	1		

Table 4-2. Nucleotide differences between NGS and Sanger sequences.

Analysis of the translated sequences shows non-synonymous differences in six of the 14 specimens with sequence coverage by both methods. Many amino acid changes could not be analysed due to lack of NGS coverage or ambiguous nucleotides generated by the Sanger method (Figure 4-8).



	606	607	608	609	610	611	612	613	614	615	616	623	626	631	632	633	634	635	636	637	638	641	644	645	646	651	658	659	661	662	663	664	666	667	668	670	688	689	690	691	695
S.427334	A	I	Q	T	R	T	V	I	N	Q	H	V	F	A	L	V	S	K	R	S	F	K	T	S	S	D	T	I	T	R	S	F	Q	L	R	K	T	V	A	V	S
427334	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
S.427759	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
427759	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
S.428005	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G
428005	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G
S.428194	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
428194	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	N	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
S.428287	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	K	.	.	.	.	.	.	.	.	.	.	.	N
428287	?	?	?	?	?	?	?	?	?	?	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	K	.	.	.	.	.	.	R	A	P	.	N
S.428891	.	.	.	.	?	.	.	.	.	.	.	.	?	.	?	?	?	.	?	.	.	?	.	.	F	.	?	?	?	?	?	?	?	?	.	?	.	.	.	.	.
428891	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
S.429031	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G
429031	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
S.429323	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
429323	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
S.429660	.	.	.	.	.	.	.	.	.	.	.	.	?	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
429660	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
S.429915	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
429915	?	?	?	?	?	?	?	?	?	?	K	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
S.430038	.	.	.	.	.	.	.	.	.	.	.	.	?	?	?	.	.	?	.	?	?	?	?	-	F	.	?	.	?	.	?	.	.	.	?	.	.	.	.	.	.	.
430038	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
S.430146	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
430146	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
S.430741	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
430741	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
S.431467	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	?	.
431467	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Figure 4-8. A comparison of non-synonymous nucleotide substitutions between Sanger and NGS methods.

The sequences with a prefix of S were generated with Sanger and those without a prefix were generated with NGS. Non-synonymous nucleotide substitutions are highlighted in green. Amino acids which could not be analysed due to ambiguous base calling are highlighted in blue and those due to lack of NGS coverage in yellow.

The NGS alignments were visually inspected using to tablet to determine if mixed bases were present at the same positions with ambiguous bases calls by Sanger sequencing. There were a number of bases with mixed populations at low levels, although not at these positions. Interestingly, as described above, in specimens 430038 and 429660 an additional enterovirus was detected which would be a possible explanation for ambiguous base calls. Unfortunately in the remaining two specimens with additional enteroviruses the VP1 sequences were not available from both methods to directly compare.

The phylogenetic relationship among the sequences generated from each method shows partners that fall within the same branch, therefore the small number of differences is not sufficient to alter the topology of the tree (Figure 4-9).

In comparison to analyses carried out using NGS sequences alone, the Sanger method provides a similar level of resolution into the relationship of the viruses in this cohort with those previously described.

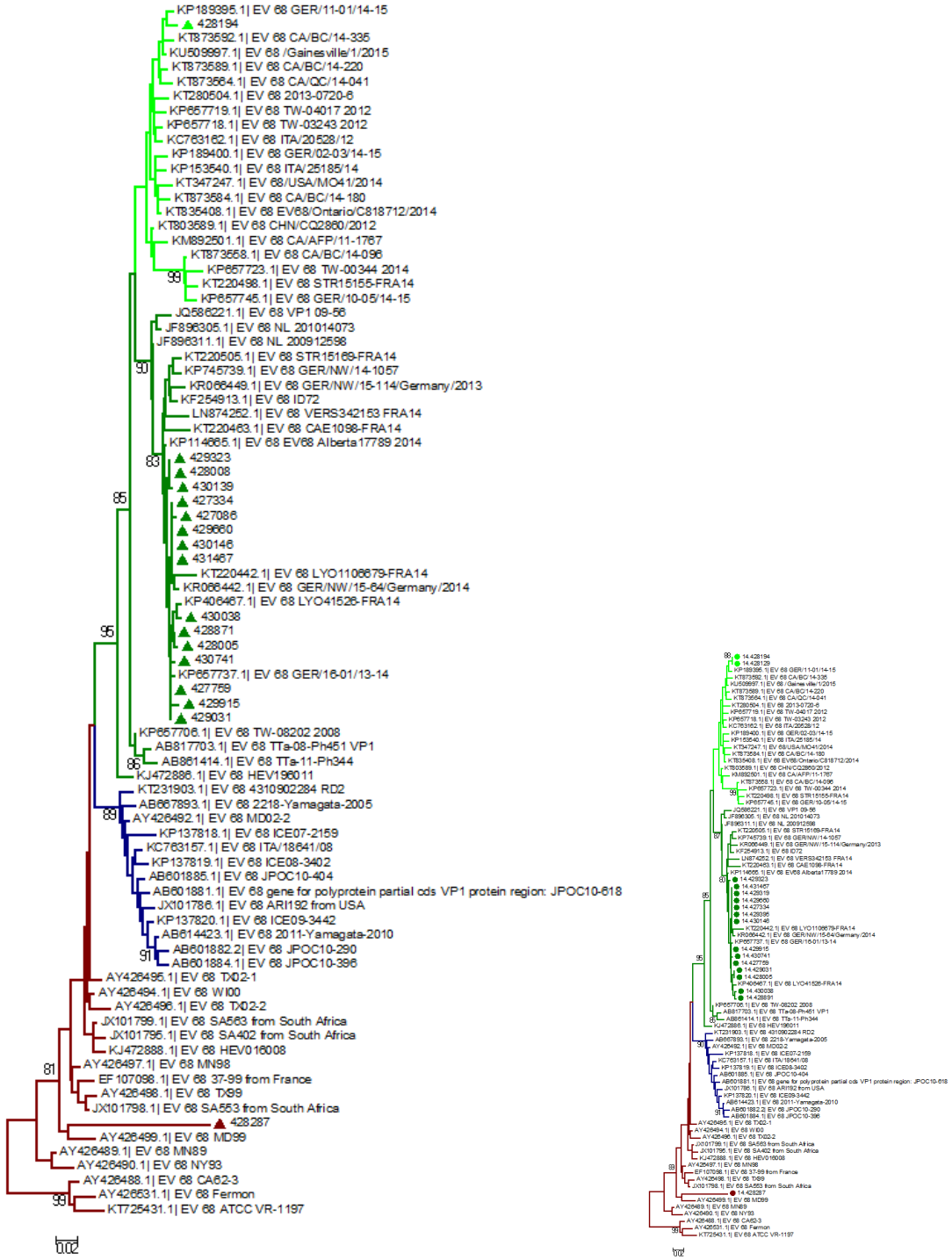


Figure 4-9. Comparison of phylogenetic relationship of Sanger and NGS sequences.
A Maximum Likelihood tree was inferred using the Tamura-3 model. A total of 316 bp from 84 sequences were included. Panel A shows the relationship of NGS sequences with VP1 references. The clinical isolates are highlighted with triangle markers. Panel B reiterates the tree inferred using Sanger sequences for comparison.

4.5 Discussion

In this research, diagnostic specimens which were screened for EV-D68 as part of an epidemiological study subsequently underwent NGS using the methods

optimised in the previous chapter. The aim of this being to determine if NGS could simultaneously detect pathogens and provide the sequence information required for additional typing and epidemiological analyses. The research presented here demonstrates that using an NGS approach in a diagnostic virology setting would allow simultaneous detection and in-depth sequence analysis utilising a single workflow and without specific pathogen targeting. Further to the work presented in the previous chapter, the NGS method used here was also capable of detecting more than one viral pathogen from a single clinical specimen.

Here we assessed the use of the NGS method for the detection and typing of EV-D68. We used a panel of 22 specimens found to be EV-D68 positive by the type specific real time PCR method used in the WoSSVC. The NGS method correctly identified EV-D68 in 19 of these specimens. A small number of EV-D68 sequenced reads were detected in the remaining three specimens by direct alignment to reference sequences however this would be equivalent to predefining a target and as the numbers were small there is less confidence in this result. As documented in chapter 3, detecting a small number of reads may not indicate a positive result.

As discussed in chapter 1, the NGS method can be carried out without a predefined target. Using this method it was possible to identify the virus under study including subtype using a single workflow, in comparison to the RT-PCR method where a genus specific RT-PCR followed by a type specific reaction was required. Whilst it would be technically possible to identify positive specimens using the type specific PCR alone this would not be an appropriate use of resources in a diagnostic laboratory since the prevalence of the virus is low. Instead pre-test probability is increased with the use of a genus specific RT-PCR as a screening tool.

The virus under study in this example, EV-D68, is highly variable and prone to both mutations and recombination events as are many viral pathogens. These mutations may result in a loss of primer binding efficiency over time and as a result reduce test performance of a type specific PCR.

The panel tested included nine specimens that were positive for an additional 11 RNA or DNA pathogens by real time PCR. This allowed us to assess the ability of the NGS pipeline to detect multiple infections – something that we couldn't ascertain in Chapter 3. Three of the 11 co-infections were detected by NGS. These were all RSV-A with Cts ranging from 18.27 to 26.41. Of the eight that were not confirmed, three cases were attributed to DNA viruses (two AdV and one HHV-6 with Cts of 22.34 – 35.89). It is unclear why these were missed and is in line with the limited data provided in Chapter 3 where an AdV (Ct = 35.1) was also missed by the NGS pipeline. As has been discussed previously, the method used in this research targets RNA and therefore may preclude the detection of DNA genomes. In three of the remaining five missed positives the Ct of the missed pathogen was greater than 32, which based on data from the previous chapter and supported by analysis of the smaller number of viral detections in this study, maybe the lower limit of detection for NGS. Interestingly, HPIV4 was detected as a co-infection in four cases by RT-PCR, however despite a Ct range of 26.08 to 34.53 no cases were identified by NGS. The reason for this is unclear. One explanation may be a variation in the reverse transcription and sequencing performance between virus groups (Prachayangprecha, Schapendonk et al. 2014).

In summary, the detection of co-infections using NGS was not comparable with RT-PCR, however of the 11 detections by RT-PCR only six had a Ct less than 32 which we found to be the lower limit of detection by NGS. Of these six, three were detected by NGS.

When considering the introduction of this method to a diagnostic laboratory the limit of detection for each pathogen would need to be assessed. This would require sequencing a panel containing serial dilutions with known viral quantities. To assess the ability of co-infection detection, a mix of viral pathogens at differing quantities should be sequenced. It may be that the presence of a pathogen in high concentration would inhibit the detection of subsequent pathogens. It was not possible to assess that relationship with the panel tested here.

As discussed in the previous chapter, the methodology used here would need to be assessed for the detection of DNA pathogens using more examples than were

available here. If the sensitivity for DNA pathogens is lacking then additional processing steps would be required.

The NGS method did detect co-infections that that were missed by the routine real time PCR method used at the WoSSVC. For example, in three cases the NGS method co-detection of other *Enteroviridae* pathogens, rhinovirus A and C, which would not be identified using current diagnostic practice. These would not be identified by the real time PCR in place at WoSSVC since the assay targets an area of the genome which is conserved between species. This is necessary as the number of distinct subtypes precludes targeting each individually. The EV-D68 RT-PCR and Sanger sequencing method would also fail to detect such mixtures. For example, the EV-D68 assay would only detect the presence or absence of this pathogen whereas the Sanger sequencing will only highlight a mixture if it is present in ~50:50 ratio. As most recent studies of viral respiratory infections have used a PCR based approach in the detection of pathogens there is little information on the clinical significance of infection with more than one of the *Enteroviridae*.

In this research, more than one virus and indeed more than one enterovirus was detected from a single clinical specimen. While multiple viruses can be detected by RT-PCR, this requires numerous primer sets and assays. It is not feasible to differentiate viral subtypes, this was not detected by either PCR or Sanger sequencing. The clinical significance of viral co-infection is not proven however this is clearly an opportunity for recombination events between serotypes and could shed light on the evolutionary dynamics of enteroviruses by linking events that take place within single patients with those observed at the population level. As discussed previously there are multiple clades of EV-D68 in circulation worldwide. The characterisation of a partial or single protein coding region is unlikely to highlight recombination events. Recombination is an important viral mechanism of evolution. These events can result in novel virus strains which are antigenically distinct and therefore not recognised by the host immune system. As has been discussed previously the *Enteroviridae* are responsible for multiple clinical syndromes and as a result of recombination involving neurotropic viruses, the potential exists for novel neurotropic strains or established strains to gain the ability to infect the CNS (Tapparel, Junier et al. 2009).

As in chapter 3 we found that the breadth of genome coverage and the number of reads with viral origin were again related to the RT-PCR Ct value, further evidence to suggest both that sequencing efficacy is linked to viral quantity and that NGS could be used in virus quantification. The detection cut-off was again in the region of Ct = 32, as the mean Ct of RT-PCR positive specimens which were not detected by NGS was 31.78. This correlates with findings from chapter 3, although a smaller number of specimens were included in this analysis. Near full genome coverage was generated in strongly positive cases by PCR (e.g. Cts < 25) but at lower viral quantities genome coverage was sporadic and unpredictable with no guaranteed coverage of the area of interest for comparison with Sanger or epidemiological analysis. As shown in chapter 3, the relationship between virus quantity and NGS sequencing varied between virus groups, for example, the breadth of sequencing coverage for the *Paramyxoviridae* was poor in comparison to the *Enteroviridae*, despite similar Cts. This would suggest that the test performance would require to be analysed for a full panel of pathogens to determine what initial viral load is likely to result in full genome coverage with NGS.

In comparison with Sanger sequencing which generated the targeted sequence in 17 of 22 cases, the same target area was also sequenced in 17 of the 22 clinical specimens. Both methods generated sequence information from 14 specimens. Each method generated sequence from three specimens which was not identified by the other. It is important to point out that following the highly specific amplification process utilised for Sanger sequencing no VP1 sequence was obtained in four cases and in the remaining case a sequence identified as parechovirus-3 was obtained. The perfect test would be 100% sensitive and specific however this is rarely the case. With regard to the VP1 coding region, NGS identified the same number as Sanger. Though, as this approach was target independent and other coding regions of interest, VP2 and VP3, were identified in a further two cases. These areas encode for the remaining capsid proteins and have been used in the molecular typing of *Enteroviridae*. VP2 and VP3 have been shown to give typing results comparable with those of VP1. This would indicate that the use of NGS in the detection, typing and epidemiological analysis of an enterovirus is comparable with that of nested PCR and Sanger sequencing. It is

difficult to extrapolate this information for other pathogens as molecular typing targets and methods vary between virus groups.

Additionally NGS sequencing provided near full genome information in eight of 22 cases. Is it possible to infer viral subtype from a single protein coding region however to study the evolution of the pathogen and detect possible recombination events would require greater genome coverage. To generate full genome coverage of the enterovirus genome would necessitate multiple PCR and sequencing reactions. There is a limit to the size of amplicon that can be reliably produced with RT-PCR and while it would be possible to generate an amplicon of 7.3kb there is a high likelihood of introducing errors, therefore it would not be appropriate to use this as a template for a sequencing reactions.

Comparable sequence regions between NGS and Sanger were identified in 14 cases. Nucleotide mismatches were found between these despite arising from the same clinical specimen. Where nucleotide differences occurred, many of these were as a result of Sanger sequencing allowing a mixed ambiguous nucleotide to be called. The generation of a consensus sequence by NGS with the programs used in this research will only allow an A, C, T or G base to be called. In a small number of cases the nucleotide differences were not to ambiguous bases, rather a different base. Phylogenetic analysis comparing the sequences generated in both methods showed that this small number of differences was not enough to alter the topology of the tree generated.

An interesting point to note is that of the four specimens noted to have ambiguous bases called by Sanger two had additional *Enteroviridae* detected by NGS. In a third case (428891) a co-infection with AdV was detected by PCR with a strong Ct, 22.34. It may be possible that this could affect NGS identification of a further enterovirus in this case as there may be a limit to the number of viruses detectable in a clinical specimen or this may preclude detection of low levels of other viruses. This would imply that despite species specific amplification reactions, the presence of a related virus in the clinical specimen is affecting the output of the Sanger sequencing.

The Application of an NGS Pipeline to the Detection and Epidemiological Investigation of Norovirus

5.1 Infectious Intestinal Disease

The research demonstrated in the previous chapters demonstrated that the implementation of an NGS method into a diagnostic setting can be used to identify respiratory viral pathogens from clinical specimens. The correlation between viral sequenced reads, proportion of reference coverage and the RT-PCR Ct suggests that this method could also be semi-quantitative. As demonstrated in chapter 4, multiple pathogens can be detected from a single specimen; however the sensitivity of this requires improvement to be used as a diagnostic tool.

The research presented in chapter 4 also demonstrates that the sequence information generated as part of this process can be used in for typing and epidemiologic analyses, which may not currently be carried out or, if studied, would necessitate further processing steps such as nested PCR and Sanger sequencing in order to generate similar levels of data.

The aim of this chapter was to apply the methods used in previous chapters to an alternative disease syndrome and specimen type to describe the efficacy of applying a single workflow to different specimen types for the identification of pathogens. The identification of gastrointestinal infections is a commonly requested test for most virology diagnostic services and was thus identified as an area which may benefit from such technology.

Multiple pathogens are associated with infectious intestinal diseases (IID) including bacteria and parasites but the majority are thought to be caused by viruses (Wikswo, Kambhampati et al. 2015). Within these cases noroviruses are a major contributor alongside sapovirus and rotavirus. Overall, such episodes of IID are responsible for over 1 million GP consultations and up to 17 million cases annually in the UK (Tam, Rodrigues et al. 2012). The introduction of the rotavirus vaccine to the routine schedule for children has seen norovirus become the leading reason for medical attendance related to IID (Payne, Vinje et al. 2013).

The research presented in this chapter will focus on norovirus, as this is an RNA pathogen associated with closed outbreaks and whose typing is important for both epidemiological and infection control purposes.

5.2 Norovirus

The Caliciviridae contains four genera; norovirus, sapovirus, vesivirus and lagovirus, of which only norovirus and sapovirus are human pathogens. They are non-enveloped viruses with a single-stranded positive sense RNA genome approximately 7.5 kb in length. The norovirus genome possesses three open reading frames. The first, ORF-1 is found at the 5' and codes for at least six non-structural proteins (p48, NTPase, p22, VPg, 3C like protease and the RdRp). This is followed by ORF-2 which codes for the major structural protein VP1 and ORF-3 coding for the minor structural protein VP2.

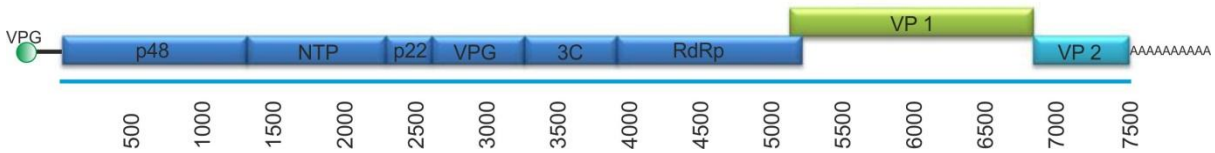


Figure 5-1. The norovirus genome.

VP1 can be divided into the shell (S) and the protruding (P) domain. The P domain is further subdivided into the P1 and P2 domain. P2 contains the hypervariable region of the capsid, responsible for antigenic and histo-blood group antigen binding sites. This variable region can be used to group and sub-type viral strains.

Norovirus was first described as the cause of a gastroenteritis outbreak in Norwalk, Ohio, US which had occurred in 1968 (Kapikian, Wyatt et al. 1972). The pathogenicity of the virus was confirmed using healthy human volunteers. In vitro study of pathogenicity is limited as to date there remains no permissive cell line in which to culture the virus in a laboratory setting. Studies carried out in Japan many years earlier had proven that bacteria-free stool filtrates could transmit gastrointestinal infections from one person to another, suggesting a virus as a likely causative agent. In 1972 Kapikian used bacteria-free samples derived from rectal samples obtained during the illness outbreak in 1968 which

were then serially passaged in adult volunteers. The stool from these volunteers was examined with immune electron microscopy (IEM). During this process a stool sample is incubated with serum obtained from a convalescing individual. The antibodies present in the convalescent serum resulted in aggregation of viral particles which could then be visualised with EM. This revealed the gross structure of the virus, with distinctive cup like structures on the capsid layer of the particle, from which the name is derived (Latin “kulix” meaning cup or goblet).

The genetic structure of the virus was determined in the 1980s at which point two distinct genogroups (I and II) were described. This has since been revised to include genogroup III, IV and V (and possibly VI) with over 30 genotypes within these. Only genogroups I, II and IV are associated with human infections. The lack of a cell culture model means typing of the noroviruses with neutralisation assays is not possible. There is no agreed standard method for the genotyping of norovirus although it is commonly based on the amino acid sequence of the VP1 protein (Kroneman, Vega et al. 2013). Accurate genotype identification can be problematic in the context of recombinant viruses possessing an established ORF-2 sequence with a novel polymerase, which has also been used in genotyping. The large number of variants likely arises from both genetic drift and recombination events, common methods used by viruses in escaping the host immune response.

The virus is spread by the faecal-oral route and can be transmitted through aerosols, environmental contamination and direct contact. A dose of only 18-1000 particles is enough to establish infection. Very high viral loads are found in vomitus and faeces of symptomatic individuals, who are the drivers of outbreaks, however asymptomatic infection also occurs in up to 7% of healthy individuals (Ahmed, Hall et al. 2014; Iturriza-Gomara and Lopman 2014). The immune response to norovirus is incomplete and is not life-long, leaving individuals vulnerable to repeated infections. The exact duration of immune protection is not clear as previous figures of six months to two years were based upon challenge studies using large doses of virus, far greater than those which result in natural infection thus it is unclear if the natural antibody response is protective against a smaller virus challenge (Simmons, Gambhir et al. 2013).

Host genetic factors play an important role in the development of illness. Investigations using virus like particles (VLP) demonstrated that the particles will only bind to host cells in the presence of genetically determined carbohydrates, the histo-blood group antigens (HBGA). This relationship is complex and the binding of noroviruses are strain and HBGA specific. Those who do not produce antigen H are immune to infection and those with histo-blood group B antigen are resistant to infection with norovirus GI strains (Nordgren, Nitiema et al. 2013).

5.3 Epidemiology of Norovirus Infections

Norovirus is estimated to be responsible for up to 18% of all cases of acute gastroenteritis and was the cause of 82% of the 137 reported outbreaks of infectious intestinal disease in Scotland in 2014 (HPS 2015). Infections occur both as sporadic events and outbreaks, occurring all year round but predominating in winter months. It is estimated that there are in the region of 3 million episodes of norovirus in the UK each year, resulting in up to 130,000 GP consultations (Tam, Rodrigues et al. 2012).

The most frequently detected genotype in outbreaks is GII.4 and this has remained the case for many years. This genotype contains multiple strains and the emergence of novel strains, occurring every few years, is often associated with increased virus activity and prolonged norovirus seasons (Bennett, MacLean et al. 2013). As is seen with influenza, this increase of viral activity is related to a lack of residual immunity within a population. Recombination events are also proposed to play an important role in the development of new variants (Bull, Eden et al. 2012) and a further reason for the genetic diversity of the virus.

5.4 Outbreaks and outbreak control

Norovirus outbreaks predominate in areas with high levels of contact, such as hospitals and care facilities. Such outbreaks result in a substantial amount of disruption to service at a great financial cost to health services (Lopman, Reacher et al. 2004).

Multiple factors contribute to risk of norovirus hospital outbreaks. The infectious dose required to result in infection is in the range of 10 to 1000 viral particles and there are multiple routes of infection, including direct contact with infected faeces or vomitus, but also aerosolised particles and fomite transfer (Robilotti, Deresinski et al. 2015). Infected individuals also continue to shed virus after resolution of symptoms. Contaminated hands can spread particles to up to seven subsequent surfaces (Barker, Vipond et al. 2004); therefore special attention must be pain to both hand washing and environmental decontamination. Viral particles can survive on surfaces for up to 12 days (Dalling 2004) therefore hygiene is essential in halting the spread of infections and outbreaks. The high predominance of the virus in the general population is a further risk for the introduction into a healthcare setting from the community.

The illness is usually short in duration with low mortality rates in the general population but the elderly and those with co-morbidities are at risk of becoming compromised by the infection. Individuals who are immunosupressed are also likely to shed the virus for a prolonged time. This has implications for the spread of disease. Variation of the virus within host in those with immunosupression may also play a role in emergence of viral strains (Vega, Donaldson et al. 2014).

The main strategy in the prevention of spread is to prevent direct contact between infected and susceptible individuals. In a hospital setting this is usually achieved through ward closures, patient isolation or cohorting and restriction of visitors. Additional measures include the use of personal protective equipment (PPE) such as gloves and aprons, effective hand hygiene and the control of staff movement i.e. preventing staff from affected areas entering non-affected areas and hand hygiene (Lopman, Reacher et al. 2004). The rapid identification and isolation of potentially infected individuals is imperative in the control of outbreaks.

5.5 Norovirus Treatment and Prevention

A greater understanding of viral spread and diversity is required to overcome some of the barriers which have thus far prevented the development of effective vaccines or therapies in the management of noroviruses (Bull, Eden et al. 2012).

No current therapies are available for the treatment of norovirus infections; however a proposed drug target is that of the VPg protein which is tethered to the 5’ end of viral genome. This protein is also relatively conserved between variants. The exact purpose of the protein is unclear but it is likely involved in the replication process. In murine models of infection the removal of this protein results in an attenuation of infection.

As with many viral pathogens the variability and large number of genotypes with lack of cross-protective antibody responses has made this a difficult process. Within the highly variable P2 domain of the VP1 protein there are relatively conserved regions associated with the histo-blood group antigen binding site which is proposed as a potential vaccination target. There are promising studies on the development of effective norovirus vaccines (Lindesmith, Ferris et al. 2015).

5.6 Diagnosis of Infection

The diagnosis of norovirus infection relies on the detection of the virus in stool samples. Norovirus cannot be grown in cell culture; requiring diagnosis to be confirmed using other methods. Transmission electron microscopy (TEM) has previously been the accepted method of viral identification however this method is both labour intensive and insensitive. Access to such a technique is now not available in most settings. Enzyme immunoassays (EIA) are commercially available and demonstrate ease of use with high throughput but many report sensitivity in the region of 50%. For this reason they are not recommended as a single diagnostic test and samples testing negative should be confirmed with an alternative method (Rabenau, Sturmer et al. 2003). As for respiratory viruses, in the diagnosis of norovirus, PCR has become the gold standard. High levels of both sensitivity and specificity have been demonstrated with the use of molecular methods. While gel based and nested PCR methods are available, a real-time assay used in the WoSSVC as it requires a single reaction therefore offers increased speed and ease of use. This assay is capable of detecting and differentiating genogroup I and II (Gunson and Carman 2005) but cannot distinguish between the genotypes within these. This results in a low resolution understanding of cases and outbreaks in real time. Genotyping can be carried

out but again this relies on the additional nested PCR reactions and Sanger sequencing.

5.6.1 The Role of NGS in Norovirus Diagnosis

The ability of NGS to detect norovirus from clinical specimens has been demonstrated, although to date there are no centres using NGS as a primary diagnostic method. There are many potential benefits from its implementation. The study of chronically infected individuals has revealed that nucleotide substitutions occur even within the P2 domain thus may affect the structure of the HBGA binding site (Carlsson, Lindberg et al. 2009). Such chronically infected individuals may be a source of new variants. Even minor variants at low frequencies can be transmitted and therefore play an important role in the increasing virus diversity (Bull, Eden et al. 2012).

Genotyping of norovirus strains is important in monitoring the epidemiology. A greater understanding of virus evolution and transmission events will be essential in the development of vaccinations.

5.7 Aim of Research

The aim of this chapter was initially to determine the efficacy of NGS as a method for the identification of noroviruses associated with IID from a panel RT-PCR positive clinical specimens. We wanted to evaluate whether this system could detect, type and provide data which could be used in potential epidemiological analyses such as detection of recombination events which may give rise to novel strains. As the RT-PCR Ct in norovirus positive specimens is generally low, it was hoped that the depth of sequencing coverage would be great enough to analyse single nucleotide polymorphisms and determine intra-host variability. It was acknowledged that the number of specimens and timeframe included would preclude analysis of within host evolution but as a proof of concept study, given norovirus is frequently associated with closed setting outbreaks, could such analyses which would be applied to determine the direction of spread between hosts be applied to these data.

5.8 Methods

5.8.1 Samples

Residual nucleic acids extracted from stool specimens submitted to the WoSSVC for IID diagnostics were used in this study. Two outbreaks were picked at random and all associated specimens from each outbreak were used. In total, 11 stool specimens which had tested positive for norovirus by RT-PCR (WoSSVC) were used in this study, five from outbreak 1 and six from outbreak 2. The RT-PCR Cts ranged from 9.06 to 24.54.

RT-PCR results			
Lab No	**Real-Time Ct**	Date Collected	Time
Outbreak 1			
500860	15.78	18/2/15	07:45
500863	12.51	17/2/15	15:45
500868	17.89	17/2/15	10:33
500920	12.22	19/2/15	N/A
500922	24.54	20/2/15	04:45
Outbreak 2			
500967	15.04	23/2/15	N/A
500968	10.05	23/2/15	N/A
500969	21.33	23/2/15	N/A
500970	9.06	22/2/15	N/A
500971	23.14	23/2/15	N/A
500973	9.93	23/2/15	N/A

Table 5-1. Date and time of collection of norovirus positive specimens.

5.8.2 Sample Preparation

The methods used are described in detail in chapter 2. In brief, the extracted nucleic acids from clinical specimens were treated with DNase, reverse

transcribed and second strand cDNA synthesised. Random whole genome amplification was carried out using the SISPA method.

The resulting PCR product was clarified using Ampure XP beads and sequencing libraries prepared using the Nextera XT DNA Sample Prep kit (Illumina). Sequencing libraries were quantified using the Qubit HS DNA assay and the DNA fragment size measured with the Tape Station HS DNA kit.

Libraries were then sequenced using the Illumina MiSeq V2 300 cycle reagents to generate 150 bp paired-end reads.

5.8.3 Data Analysis

Sequenced reads were quality trimmed and mapped to a database containing human reference sequences. The remaining sequences following this process were entered into the MetAmos pipeline where de novo assembly of contigs was carried and subsequent identification of contigs using a BLAST search against a protein database.

When norovirus contigs were identified in specimens, the assembled contigs were then used in a BLASTn search against a database of all norovirus full genome sequences available from NCBI. The top BLAST hit, based on coverage and identity scores, was the used as a template for a reference based assembly (Bowtie2).

Consensus sequences were generated from the alignment files using an in house tool (J. Hughes, unpublished). The consensus sequence was visually inspected and compared with the alignment file using Tablet to ensure correct consensus calling. The sequenced reads were then mapped to this consensus to increase coverage.

5.8.4 Phylogenetic Analysis

Consensus sequences were aligned with reference genome sequences in MEGA6. Phylogenetic analysis was carried out using MEGA6. Maximum likelihood trees

were inferred with 1000 bootstrap replicates. The model used for tree generation was chosen based on the BIC score.

5.8.5 Viral strain identification

The viral strain of each isolate was identified using the Norovirus Automated Genotyping Tool (Kroneman, Vennema et al. 2011). The consensus sequence was entered into this web based tool which determines viral strain based on a BLAST analysis of the ORF1 and/or ORF2 sequence.

5.8.6 Intrahost Virus Diversity

As single nucleotide polymorphisms can be used as markers to determine direction of viral spread within an outbreak, in order to assess if variable nucleotides could be detected within the consensus sequences reads were mapped back to the generated consensus. A pileup file was generated with DiversiTools. The SNP variants in the alignment were called using LoFreq where the depth of coverage was greater than 100 reads. Variants are assigned a p-value based on coverage and frequency. Those with a p-value greater than 0.05 were considered non-significant and excluded from analysis.

5.8.7 Recombination Analysis

Recombination analysis was carried out using RDP4 with sequence alignments generated using Muscle as part of MEGA6. A database of complete norovirus genome sequences was compiled from sequences available in PubMed (search term “norovirus”, sequence length 7300 – 8000 bp). The sequences were clustered at 99% homology (using cd-hit-est) to remove similar sequences and therefore reduce the overall size of database. This in turn reduces computation power required for analyses. The remaining nucleotide sequences were aligned with the consensus sequences using Muscle in MEGA6. The alignment file was entered into RDP4 to determine if the clinical isolates from these outbreaks were involved in recombination events.

5.9 Results

5.9.1 Virus identification

The distribution of reads generated by sequencing is shown in Figure 5-2. On average 1,000,000 reads were generated per specimen (range 808198 – 1615450). An average of 18% of reads was removed following quality trimming. Of the remaining reads 4% (range 0.03 – 32.36%) mapped to a human reference and 45% mapped to a viral reference (range 0.11 – 92.87%). Using the MetAmos pipeline norovirus contigs were detected in all specimens.

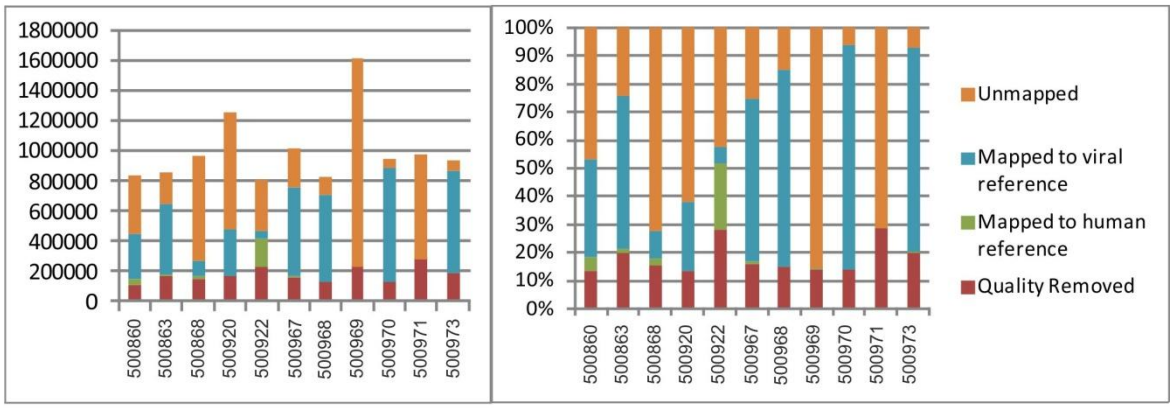


Figure 5-2. Distribution of sequenced reads from norovirus RT-PCR positive specimens.
The panel on the left shows the number of reads removed by quality trimming, mapping to a human reference, mapping to a viral reference and those remaining unmapped. The panel on the right expresses each of these as a proportion of the total number of sequenced reads.

RT-PCR results		NGS results							
Lab No	**Real-Time Ct**	**Human reads**	**Accession (genotype)**	**Reads mapping to viral reference**	**Reads mapping to NGS consensus**	**% mapped reads viral**	**% coverage**	**Average depth**	**Nucleotide mismatch**
500860	15.78	42138	KJ196292 (GI.3)	290300	323571	47.55	99.91	4766	7
500863	12.51	10827	KJ196292 (GI.3)	467031	518968	76.72	99.91	7658	6.8
500868	17.89	23677	KJ196292 (GI.3)	96486	108759	13.70	96.372	1595	6.9
500920	12.22	2463	KJ196292 (GI.3)	304077	343826	31.67	99.742	4979	7
500922	24.54	187303	KJ196292 (GI.3)	49655	56378	14.40	93.906	833	6.2
500967	15.04	7114	KC631827 (GII.4)	585627	590245	69.92	97.91	8903	2.9
500968	10.05	362	KC631827 (GII.4)	579716	585677	83.02	99.511	8657	3.1
500969	21.33	1021	KJ685412 (GII)	1859	1907	0.14	95.379	28	3.2
500970	9.06	228	KC631827 (GII.4)	754763	763037	93.88	98.836	11429	3.3
500971	23.14	592	KJ685412 (GII)	744	761	0.11	91.051	11	3.7
500973	9.93	4251	KC631827 (GII.4)	679042	686646	91.83	97.447	10015	3.1

Table 5-2. Sequenced reads mapping to norovirus reference genome.

Near full length genomes (>90% coverage) were generated in all cases (Table 5-2). The mean sequencing depth was 5352 and ranged from 11 to 11429 reads per base.

Viral contigs were detected in all clinical specimens. Following the nucleotide BLAST search five of 11 samples returned a norovirus of genogroup I, genotype 3, four of 11 returned genogroup II, genotype 4 and two of 11 returned genogroup II but no genotype documented. The specimens which failed to return a genotype had Cts of 21.33 and 23.14.

Assessment of both the proportion and number of sequenced reads mapping to the viral reference shows there is a relationship between these and the RT-PCR Ct (Figure 5-3). For example, in 500973 with a Ct of 9.93, 91.83% of non-human reads were viral and covered 97.45% of the reference genome. In contrast, in 500971 with a Ct of 23.14, 0.11% of the non-human reads were viral and covered 91.05% of the reference genome. The depth and breadth of reference coverage also correlated with the Ct. Linear regression of each of these parameters returned a high R^2 value, 0.68 or greater, suggesting a linear relationship.

Following assembly to the top nucleotide BLAST hit the depth of coverage ranged from 11 to 11429 reads. The level of nucleotide mismatch in the GII specimens was lower, 2.9 to 3.7%, than in the GI specimens, 6.2 to 7.0%.

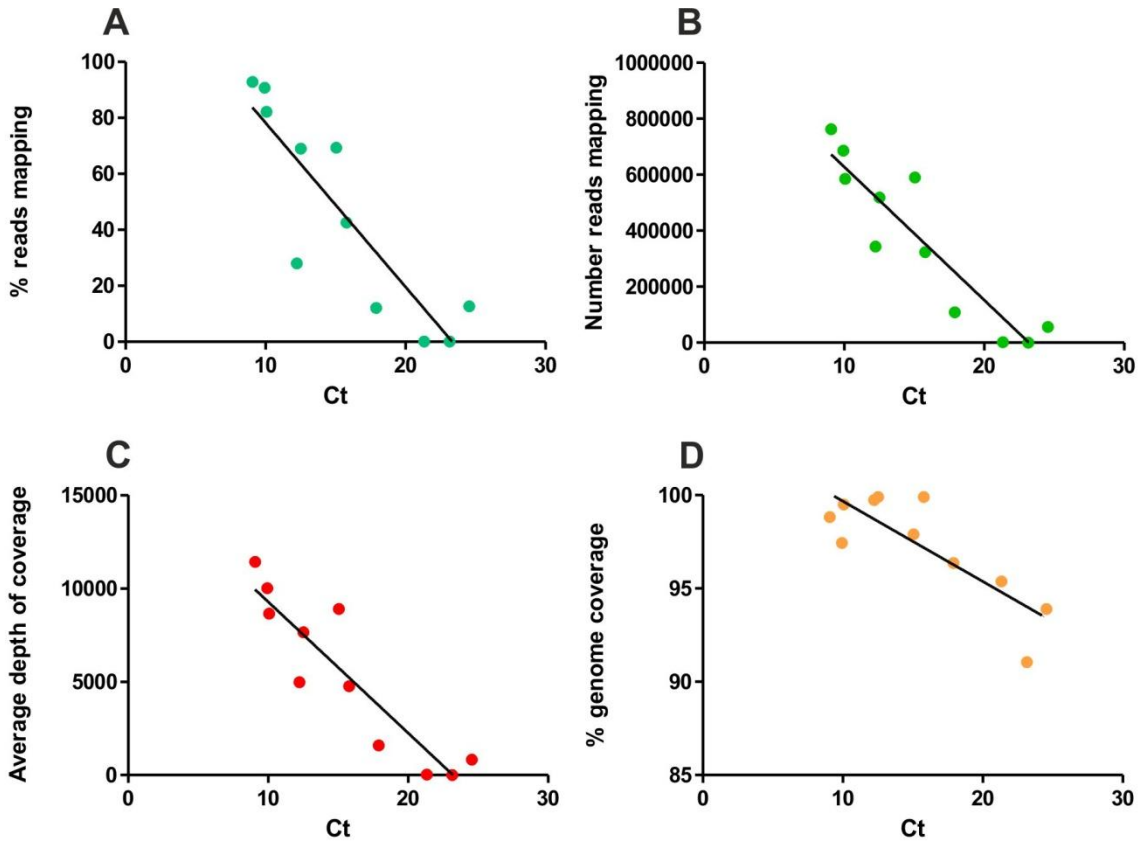


Figure 5-3. The relationship between mapped reads and RT-PCR Ct.
Panel A shows the proportion of unmapped reads following quality trimming and human reference alignment which mapped to the norovirus reference sequence and how this related with RT-PCR Ct. Linear regression $R^{2 =}$ 0.84 ($p<0.0001$). Panel B shows the number of unmapped reads mapping to the norovirus reference sequence. Linear regression $R^2 = 0.77$ ($p=0.0003$). Panel C shows the average depth of sequencing coverage following alignment of reads to the top hit reference genome. Linear regression $R^2 = 0.83$ ($p <0.0001$). Panel D shows the percentage of reference genome covered by sequenced reads following alignment. Linear regression $R^2 = 0.68$ ($p=0.0018$).

5.9.2 Assembly

Using a reference-based assembly with the top BLAST result as the template generated greater than 90% of reference genome coverage in all cases. The sequencing depth varied throughout the length of the genome but in general was greatest over the NTP and VP1 regions. In the genotype I specimens the depth of coverage was low out with these regions whereas in the genotype II specimens there was a further peak in coverage in the 3C coding region.

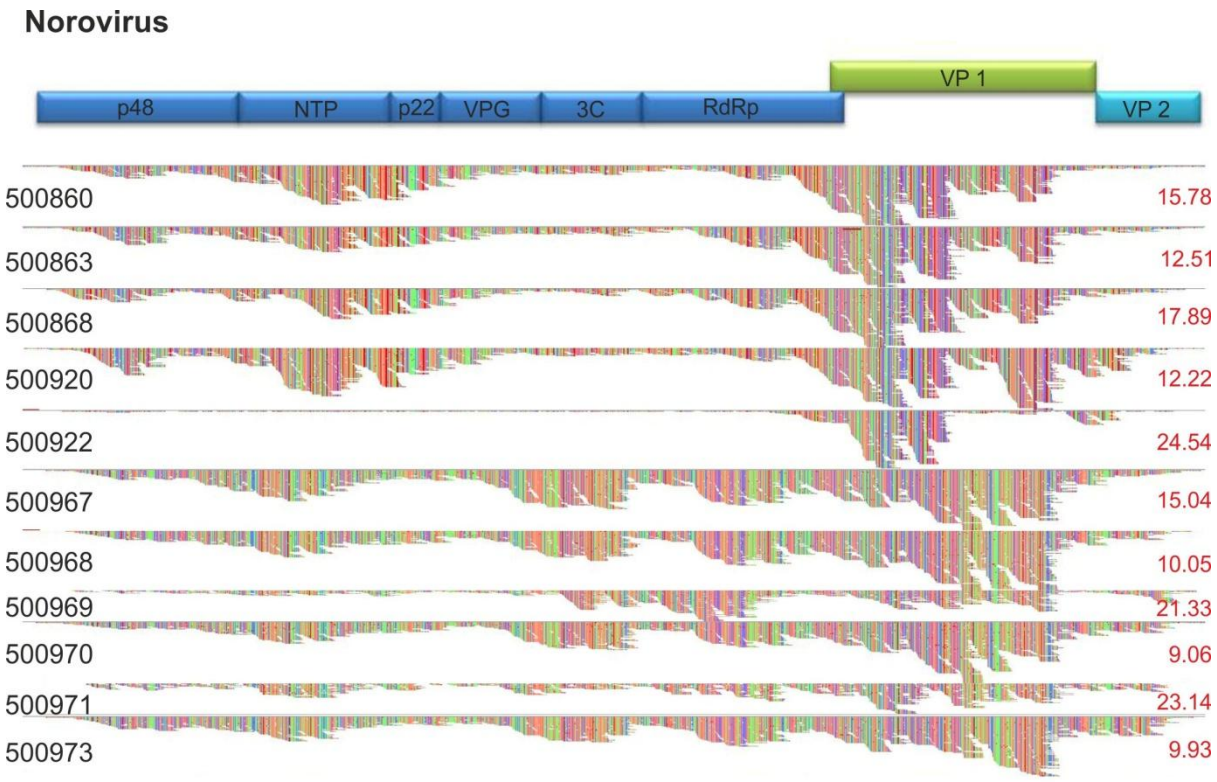


Figure 5-4. Reference based alignment of sequenced reads with a representation of the norovirus genome.

5.9.3 Strain Analysis

Using the Norovirus Automated Genotyping Tool, all specimens from outbreak 1 returned a result of GI.P3. All specimens from outbreak 2 returned a result of GII.4 Sydney 2012.

5.9.4 Phylogenetic Analysis

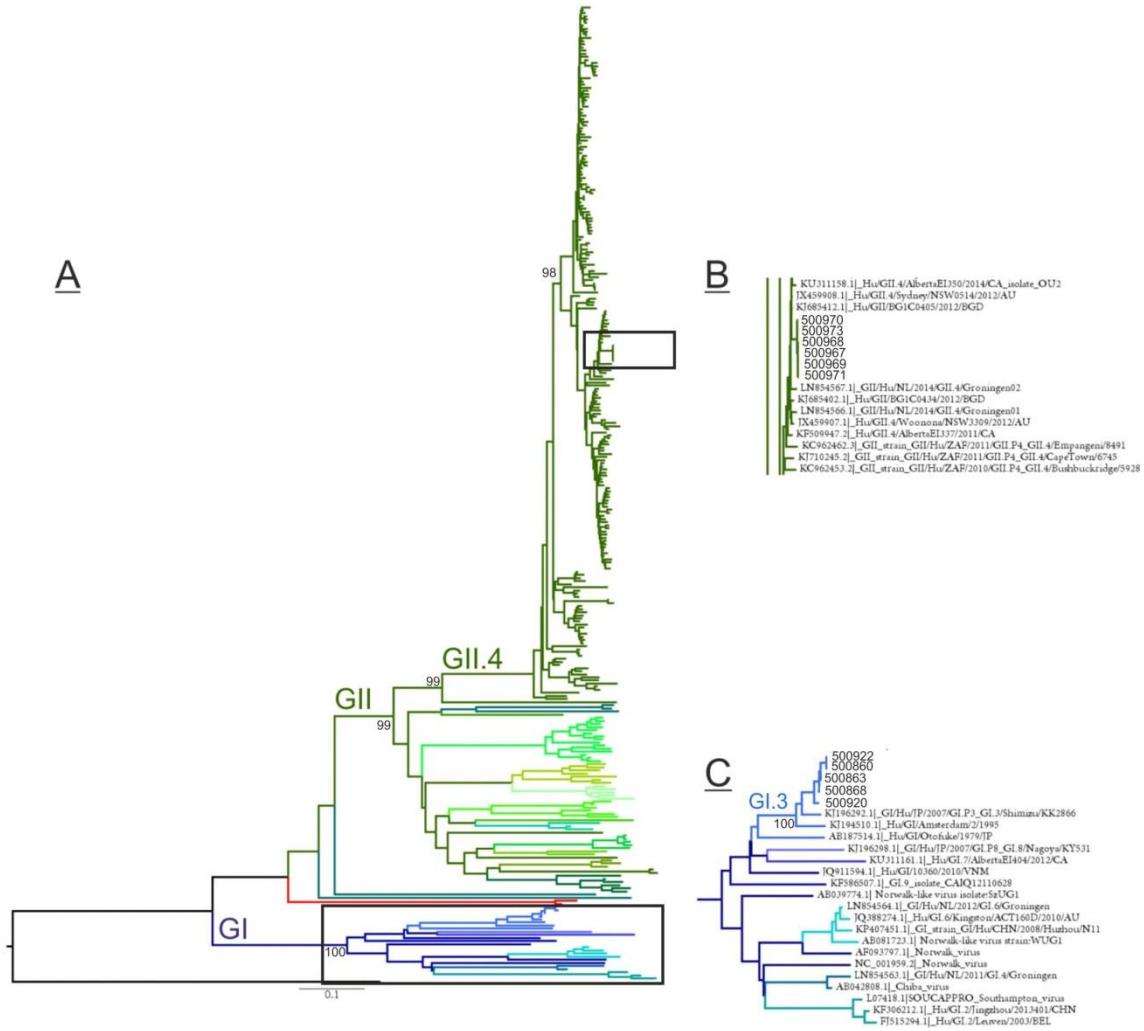


Figure 5-5. The phylogenetic relationship between clinical isolates and reference genomes, based on the VP1 segment.

The areas within the boxes are shown in more detail in panels B and C. This neighbour joining tree was inferred with the Tamura-nei model in MEGA6. 1000 bootstrap replicates were carried out and relevant values shown on the tree as percentages. The analysis involved 315 nucleotide sequences. There were a total of 2166 positions in the final dataset.

Phylogenetic analysis based on the VP1 coding region of the norovirus genome reveals that the sequences obtained from the clinical specimens in each outbreak were most closely related to each other.

The specimens from outbreak 2 are all on the same branch amongst genotype GII.4 sequences, suggesting a close relationship with each other. The specimens from outbreak 1 fall in a clade with genotype GI.3 sequences. In this outbreak one specimen, 500920, falls on a different branch from the remaining four specimens. This could suggest there are some differences in this virus.

5.9.5 Recombination Analysis

An alignment of the consensus and known full genome sequences was entered into RDP. The virus detected in outbreak 2 was result of a recombination event between KF72497 (US) and KJ685402 (Bangladesh). The breakpoint was identified between positions 4486 and 5146. This would be in keeping with the beginning of ORF2 which is in the region of position 5022. No recombination events occurred during the course of the outbreaks, although this would be the expected result given the short duration of each.

5.9.6 Intrahost Variability and SNP Analysis

The epidemiology of norovirus varies with that of the respiratory viruses described in previous chapters. As discussed earlier in this chapter, norovirus is commonly associated with outbreaks in closed settings such as healthcare facilities. SNPs can be identified and followed through the transmission chain to ascertain the direction of spread (Hughes, Allen et al. 2012).

Variable nucleotides detected in the alignment where the sequencing coverage was greater than a depth of 10 reads were analysed. The probability of these being significant was based on a p value < 0.05.

SNPs with a frequency below 1% were removed and the remaining sites plotted to highlight any areas with a high degree of variability.

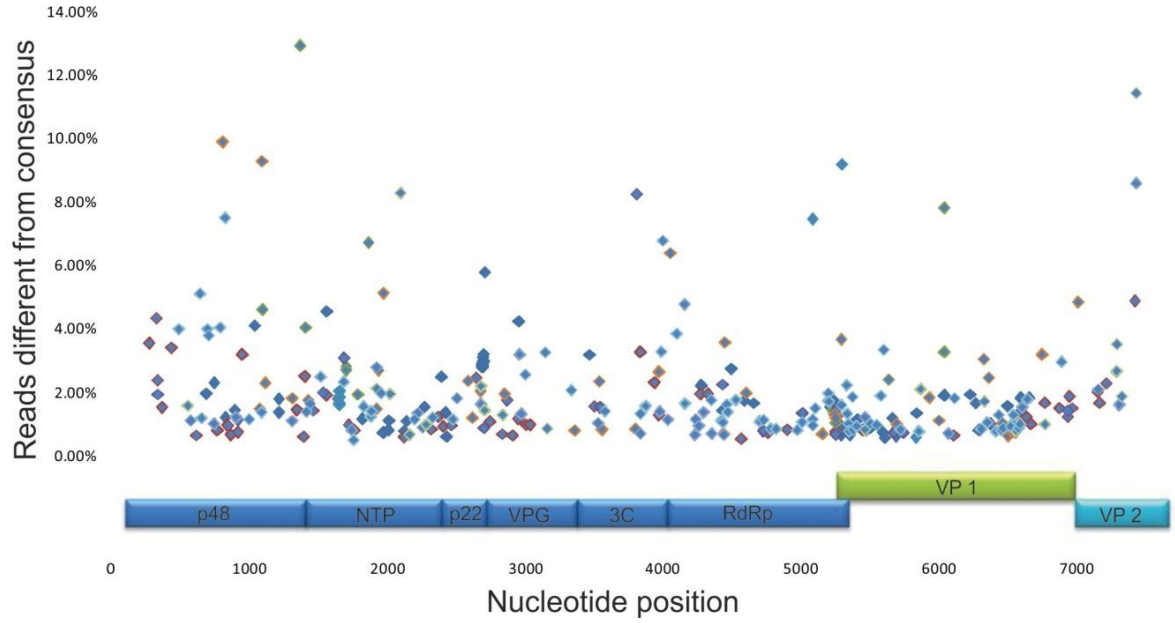


Figure 5-6. Nucleotide variants, showing the proportion of reads different from the consensus. The figure shows an amalgamation of all samples with the different coloured symbols representing each sample.

This shows that SNPs are found throughout the length of the virus genome, with the exclusion of the extreme ends of the alignment however this is a result of low coverage at these points. There is a suggestion of clusters of low frequency, less than 2%, SNPs at the beginning and end of the VP1 region.

In the specimens from outbreak 2, 500968, 500970 and 500973 demonstrated a low level variant at nucleotide position 6537 which is found in the coding region for VP1. Following translation this G to T switch would result in an amino acid change at position 485 from valine to phenylalanine.

A

Sample	500860	500863	500868	500920	500922
500860		99.96	99.97	98.94	99.81
500863	3		99.96	98.96	99.85
500868	2	3		98.94	99.81
500920	77	76	77		99.08
500922	14	11	14	67	

B

Sample	500967	500968	500969	500970	500971	500973
500967		99.99	99.99	100	99.82	99.99
500968	1		100	99.99	99.84	100
500969	1	0		99.99	99.84	100
500970	0	1	1		99.82	99.99
500971	12	11	11	12		99.84
500973	1	0	0	0	11	

Figure 5-7. A pairwise comparison of the consensus sequences generated from each clinical sample.

The analysis was restricted to sites with coverage in all samples. Panel A shows outbreak 1 with a total of 7279 bases analysed. Panel B shows outbreak 2 where a total of 6820 bases were analysed. The blue cells show the percentage identity between sequences. The cells in red show the number of nucleotide differences between sequences.

The specimens from each outbreak were collected over a short period of time, up to 72 hours. Specimens 500860 and 500868 were collected from the same individual as were 500968 and 500973. A small number of SNPs were demonstrated between these specimens. A large number of SNPs were seen in specimen 500920, up to 77. Despite these genetic differences, the virus isolated here was identified as the same genogroup and most closely related to the same reference strain as the others from this outbreak. A smaller number of SNPs were identified in samples 500922 and 500971, between 11 and 14. These may represent further introduction events but could be related to the lower depth of coverage generated from these specimens.

5.10 Discussion

The research presented in previous chapters demonstrated that NGS could successfully detect and type viral respiratory pathogens from clinical specimens and the resulting genetic sequence information generated as part of the process could be used for in-depth pathogen analyses, similar to those used for

epidemiological purposes. The aim of this chapter was therefore to demonstrate if these methods would be as effective when applied to different specimen types. To this end, NGS was carried out on a panel of norovirus RT-PCR positive faecal specimens with the initial aim of detecting norovirus, in line with the current diagnostic RT-PCR. As norovirus commonly occurs as outbreaks in healthcare facilities and these specimens regularly demonstrate low Ct values, therefore high viral load, it was hoped the depth and breadth of sequence information generated would allow analysis of mutations and variants within the viral population. These studies have been proposed as a mechanism to track viral spread during such outbreaks and may further our understanding of the outbreak process.

From this small panel of 11 specimens from two outbreaks, the presence of norovirus was confirmed in each using both a de novo assembly and reference based assembly method. The proportion of viral reads from these clinical specimens was greater than those in previous chapters, where the majority of mapped reads were of human origin. This may be due to a lower level of host contamination within this type of specimen or more likely relates to high viral load as demonstrated by the low Cts in this panel as further to the previous work, there is a relationship between Ct and proportion of viral reads sequenced. This relationship was more pronounced with this panel of specimens where a linear regression model returned an R^2 value of 0.84 compared with up to 0.40 in chapter 3. The percentage of reference genome covered by NGS reads was consistently above 90%. A difference in nucleotide mismatch levels was observed between the GI.3 and GII.4 samples. This is most likely explained by the fact that GII.4 is the most widely circulating virus strain and there are a greater number of reference sequences available. From the work in previous chapters, a lower limit of virus detection using NGS was proposed to be in the region of a Ct equal to 32. Using these norovirus positive specimens a lower limit of detection could not be identified. This again is likely a result of the high viral load in these specimens. Consequently this would imply that an NGS approach in the detection and typing of norovirus would be successful in the majority of cases.

Viral contigs were detected in all specimens. Following assembly to the top reference as determined by a BLAST search, great than 90% reference coverage was obtained in all cases. This shows a greater level of sensitivity and breadth of coverage in comparison with the respiratory specimens. Again this likely relates to the high viral load in these specimens. The number of specimens used in this research was considerably smaller than in previous chapters. The depth of coverage varied throughout the genome but returned particularly good coverage of the VP1 coding region. This may reflect the fact that this is the most abundant protein within the assembled viral particle as it constitutes the main component of the viral capsid.

This sequence information allowed robust identification and subsequent genotyping of the virus as good coverage of the genome region used in molecular typing, VP1/ORF2 was obtained. Over 25 genotypes from genogroup I, II and IV are associated with clinical disease but the current diagnostic assay is only capable of determining genogroup I and II. While these tests are highly sensitive they provide a low resolution understanding of outbreaks and the spread of disease. Genotyping can be carried out however this requires additional laboratory work, namely Sanger sequencing. As mentioned in the previous chapter in relation to the *Enteroviridae*, a genogroup specific RT-PCR is required to detect norovirus as it would neither be cost effective or feasible to carry out multiple genotype tests on each clinical specimen.

Determining genotype is not beneficial to the patient in real time as it will not change the course of management. This information is however useful from an epidemiological point of view. A new dominant norovirus strain circulating in a population emerges every two to four years and this emergence of a novel strain can be associated with a rise in the number of cases. Monitoring strain circulation in real time would provide early warning of a change in dominant strain and could assist in resource planning for potential outbreaks. Such measures may include additional ward space, isolation bays and cohorting of potentially infected individuals. As diagnostic techniques vary between laboratories, knowledge of a novel strain could ensure diagnostic services ensure their methods will be able to identify strains which are different to those currently circulating. This is applicable at a national and international level as

given the worldwide prevalence of norovirus, alerting neighbouring countries of emerging strains would be appropriate.

Vaccines and anti-viral therapies are in development however currently there are no specific options for norovirus. The mainstay of treatment for patients who require admission is supportive until symptoms subside. Should a vaccine become available in the future the sequence information generated could be used in monitoring efficacy. As is the case with influenza, multiple strains are known to circulate although a limited number can be included in vaccines. It is therefore essential to determine which strains are circulating and aim to predict what should be included in a vaccination programme.

Following detection and assembly, the sequence information generated in this process was used to assign the genogroup and genotype. In all cases the genogroup confirmed that found by RT-PCR i.e. the five specimens from outbreak 1 were GI and the six specimens from outbreak 2 were GII. In nine of 11 cases the genotype could also be assigned based on the top BLAST hit, which was then confirmed using a web based genotyping tool. In two specimens, 500970 and 500973, the top nucleotide BLAST hit did not specify genotype, only genogroup. It is worth noting however that the genotyping tool returned a result of GII.4.

In outbreak 1 the same viral strain was identified in all cases. Phylogenetic analysis of the VP1 segment revealed all isolates were most closely related to each other however specimen 500920 fell on a separate branch from the other specimens. A pairwise comparison of the resulting full genome consensus sequences identified that specimen 500920 had a large number of nucleotide differences, minimum 67, from all other cases. The number of nucleotide differences between all other specimens was substantially smaller with a maximum of 14. This would suggest that an unrelated virus was identified from this specimen. The implication here is that a second introduction event occurred during this outbreak.

In outbreak 2 the same viral strain was again identified in all cases, the current dominant Sydney 2012 strain. Similarly to outbreak 1, specimen 500971 stood

out as having more nucleotide differences than other specimen. In this case it may relate to the comparably lower depth of sequencing generated rather than a novel introduction.

The suggestion of multiple viral introductions would imply the need for a broader approach and more consideration of possible introduction routes. It is not possible to draw a conclusion on the co-circulation of multiple virus strains in a community or healthcare setting. The strains detected in this study did not undergo recombination however the isolates from outbreak 2 were the result of a recombination even between two geographically distinct viruses. It is proposed that the co-circulation of multiple strains would certainly contribute to this process but there is a lack of understanding into what drives such events to happen. The use of NGS in diagnostics and the routine generation of whole genome sequences from all clinical isolates is certainly something that would contribute to this.

The evolution of the virus within the host could not be analysed in this research as this would require multiple specimens from each individual over the course of the outbreak. We did however demonstrate that the methods required to carry this out could be employed with the NGS data generated in this study. Clinical details were not available for the cohort understudy here but the evolution within host is likely to be more significant in those who are known to shed virus for a long time and may have impaired viral clearance. Studies of outbreaks involving immunocompromised individuals have demonstrated that chronic infection can occur with detectable virus shedding for weeks or even months (Kundu, Lockwood et al. 2013). The lack of complete viral clearance along with a degree of immune pressure may have been responsible for the viral evolution that occurs within this group. Such information would not benefit individual patients rather it would further out understanding of viral evolution and diversity over time. This specialised data analysis could be carried out retrospectively for the subset of patients where it may be clinically significant.

The foundation of outbreak limitation is case isolation and appropriate hygiene measures of staff and visitors to prevent disease spread in a healthcare setting. Such knowledge would have implications relating to the infection control

measures in place to limit viral spread. Multiple introduction events may be related to staff, visitors or patients in other areas of the healthcare facility. Patient isolation, hand hygiene and personal protective equipment (PPE) are the main infection control measures put in place to prevent spread from patient to patient. These are unlikely to impact on further virus introductions from outside the healthcare facility. If multiple strains are isolated then the focus should be on limitation of movement between wards and consideration of admitting new patients to separate area until known not to be infected.

Testing a panel of samples with known concentrations would be needed to determine a detection cut off as this could not be addressed above. The ability to detect mixed infections would also need to be established. From this small panel we demonstrated that SNPs and variable nucleotides can be detected, however a well characterised outbreak with repeated sampling throughout the duration would need to be studied to determine their clinical utility.

The research presented here demonstrates some of the potentially achievable benefits of using a NGS approach in diagnostics. The additional generation of genetic sequence information would allow molecular epidemiological studies in real time including monitoring for the emergence of novel strains. Understanding the spread and evolution of norovirus within a local population could warn of such novel strains, allowing forward planning of resources. Increasing our understanding of the pathogen may in time aid in the development of targeted therapies and preventative measures which to date are lacking.

Discussion

6.1 Introduction

The research carried out in this thesis aimed to determine the feasibility alongside the potential benefits and challenges of using next generation sequencing technology in a viral diagnostic service. Given that respiratory diagnostics contribute substantially to the overall workload of most viral diagnostic laboratories; this was thought to be an ideal setting to test the potential of NGS.

As described in chapter 1, many viral respiratory pathogens cause human disease and the current diagnostic test used in Glasgow (which is similar to those in use in other laboratories in the United Kingdom) requires five multiplex RT-PCR assays to screen for an extensive but by no means exhaustive list of pathogens. These assays are highly sensitive and specific and enable laboratories to process large numbers of samples in a rapid turnaround time. Consequently the PCR-based approach to respiratory pathogen diagnosis is now considered the gold standard.

Respiratory viruses are in constant evolution and in recent decades a number of novel respiratory viruses have emerged (e.g. influenza, MERS, SARS and human metapneumovirus) and re-emerged (e.g EV-D68). As a result diagnostic PCRs must be constantly reassessed and in some cases altered to ensure these new targets are detected sensitively. The use of NGS to detect viruses from clinical specimens would not require such a targeted approach, rather a metagenomic approach where all genetic material is amplified and examined.

Alongside detection of viral pathogens, many diagnostic laboratories also play an important role in infection control and public health/epidemiology relating to respiratory infections. For example, laboratories will often be asked to investigate closed setting outbreaks of respiratory infection to determine what type of virus is involved (especially if the clinical outcome is unexpectedly severe) and to investigate the cause and spread of the outbreak (e.g. patient to patient spread or multiple introductions). In addition, many laboratories now monitor influenza subtypes during the winter season to detect the emergence of novel subtypes that may not be covered by the current vaccine. Currently, the

technology used to carry such investigations would be Sanger sequencing which would be carried out after the diagnostic PCR and is usually a retrospective request (i.e. requested after the outbreak has been detected or after a new clinical syndrome has thought to have emerged). Although some laboratories will do this on site, laboratories without access to this technology would have to refer samples away. NGS offers the potential of providing such epidemiological data at the same time as the diagnostic data thus having a major impact on infection control and public health responses.

The work described in this thesis aimed to evaluate the use of a metagenomic NGS pipeline for the diagnosis, subtyping and epidemiological study of respiratory pathogens. In each case the NGS method was compared to the existing gold standard (e.g. RT-PCR for diagnosis and Sanger sequencing for epidemiological analysis). The main benefit to the laboratory of introducing such a technique would be streamlining multiple processes into a single workflow, therefore in chapter 5 we aimed to determine if the NGS method could also be used for the detection of another non respiratory clinical syndrome, namely gastroenteritis caused by norovirus.

6.2 Summary of Research

In chapter 3, NGS was applied to a panel of respiratory specimens from individuals with symptoms of a respiratory infection to assess its usefulness when used in a diagnostic setting. The panel was tested by RT-PCR in parallel to allow a direct comparison of the two methodologies.

We confirmed the ability to detect multiple different viral pathogens, such as RSV, rhinoviruses and coronaviruses, from clinical specimens using NGS in combination with SISPA PCR without *a priori* sequence knowledge or enrichment for specific targets. The sensitivity demonstrated by NGS was less than that of the diagnostic RT-PCR panel in this study with 38 of 49 viral detections being identified by both methods. This could be explained to a certain extent by the lower limit of detection of the NGS method which for viral pathogens from respiratory specimens (as shown in chapter 3) was estimated to equate to a Ct of 32. This is approximately two logs less sensitive than RT-PCR which has the ability to reliably detect pathogens with Cts in the region of 38. Similar results

were obtained in chapter 4 when using NGS to test a panel of respiratory samples that had been typed as EV68 by RT-PCR. If NGS is to be used in place of RT-PCR then the sensitivity discrepancy will need to be addressed. The possible routes to increase NGS sensitivity will be discussed in more detail below.

Chapters 3 and 4 both outlined that there was a relationship between the viral load in a clinical specimen (as determined by RT-PCR Ct) and the proportion of viral reads sequenced. Both chapters also described a relationship between the viral load and the breadth of genome coverage. Similar findings have been described in other studies comparing NGS to RT-PCR (Prachayangprecha, Schapendonk et al. 2014). These results are encouraging as they suggest that NGS could provide semi-quantitative results, in a similar fashion to RT-PCR. Quantitation is a useful adjunct to respiratory pathogen diagnosis. It can be used to make a judgment call on the relevance of the infection (i.e. a weak positive may represent the remnant of a previous infection). Quantitation can also be used to monitor the progress of an infection and in the case of influenza the patient's response to therapy. It can also inform infection control aspect by determining a patient's infectious status. Quantitation is particularly informative in subsets of individuals such as those with a compromised immune system or those who are ventilated. These individuals frequently shed virus for extended periods of time and in such patients, emergence of treatment resistant influenza strains have been described (van der Vries, Stittelaar et al. 2013).

Respiratory viral co-infections are a well documented phenomenon, although the clinical significance of these events has not been established (Lim, de Klerk et al. 2016). Some studies suggest co-infections are more severe than single pathogen infections whereas others have failed to confirm this (Lim, de Klerk et al. 2016). Nonetheless it is essential for any novel diagnostic technique to be able to detect these events to a comparable level as is achieved with PCR. Only a single co-infection was detected by RT-PCR from the specimens understudy in chapter 3. This mixture was not identified by NGS but as the second pathogen was a DNA virus with Ct below the presumed detection cut-off it was not possible to draw any conclusions from the work in this chapter. However, the specimens under study in chapter 4 were all collected from children, who are known to have a greater incidence of respiratory virus co-infection (Nickbakhsh, Thorburn et al. 2016). This provided a further opportunity to assess the ability of

the NGS method to detect coinfection. Although the panel of samples was small, nine specimens contained an additional eleven viruses. Unfortunately, only three of these additional pathogens were detected by the NGS method, all RSV. Of the remaining eight additional pathogens that were missed by the NGS method, five had Cts greater than 32. A mentioned above, this is beyond the likely lower limit of detection of the NGS method and as a result is therefore not an unexpected finding.

Of interest, three of the additional samples missed by the NGS method were DNA pathogens. Taken together with the missed adenovirus positive in Chapter 3, this result suggests that the detection of DNA targets may be precluded by the initial reverse transcription process step of the NGS pipeline. The ability to detect DNA pathogens is an essential component of any respiratory virus diagnostic test. Firstly it is required to detect adenovirus a well documented cause of upper and lower respiratory tract illness. Secondly, DNA detection is required to detect bacterial pathogens such as *Mycoplasma pneumoniae* and *Bordetella pertussis* which are commonly tested for as part of many laboratories routine respiratory screen. Steps which could be taken to improve the performance of the NGS method in the detection of coinfections/DNA targets are discussed in more detail below.

In chapter 4 the NGS method detected 4 cases of what seem to be mixed enterovirus infections (i.e. samples that contained EV-D68 and another enterovirus subtype) that were not detected by the combination of RT-PCR and Sanger sequencing. In three of these cases an additional rhinovirus subtype could be assigned. In the remaining case sequenced reads which mapped to the 5'-UTR of multiple hRV-B references were identified but as this region of the genome is relatively conserved between species it was not possible to ascertain if these arose from the presence of an additional virus. This was an interesting finding which highlights an advantage of NGS versus current methods. This benefit would also be useful if applied to other sample types. One example is the surveillance of enterovirus infections using sewage samples. Current Sanger based methods would only reliably detect predominant types whereas the NGS method described here would be able to detect mixtures of subtypes thus giving a more accurate reflection of viral epidemiology.

As well as being highly sensitive, RT-PCR is highly specific. Although, based on the work described here, the NGS seemed to be specific a small number of sequence reads were detected in 12 RT-PCR negative specimens. Of these 12, 11 were hRV and 1 RSV. These could represent contamination during the preparation or sequencing process. The use of double indexing during sequencing aims to minimise reads being assigned to the wrong specimen however this can still occur at areas of high clustering (Kircher, Sawyer et al. 2012). These were most likely to represent false positive reads as the end point detection limit of the NGS was less than that of the RT-PCR.

Most RT-PCR panels detect the viral pathogen but do not offer simultaneous subtyping. As discussed in chapter 1, numerous subtypes exist within each respiratory virus group. For example there are three species of human rhinovirus, A, B and C comprised of over 100 known serotypes and two RSV subtypes, A and B, each containing over 10 genotypes. At present laboratories must use Sanger sequencing to type such viruses.

However, having rapid access to such information could be useful. For example, it could be used to identify and investigate outbreaks of respiratory illness far earlier than currently possible with the combination of RT-PCR and Sanger sequencing. The real time description of viral subtypes would also enable the detection of emerging viruses quicker than currently achievable. Such data would also aid in the investigation of the clinical relevance of some infections more effectively than currently. For example there is some evidence to suggest that hRV-C may be associated with more severe illness and a greater risk of post infection complications such a wheeze when compared to other rhinoviruses (Cox, Bizzintino et al. 2013). Similar variations may also exist with other virus subtypes.

The results described in chapters 3 and 4 suggest that an NGS approach would be possible for the diagnosis of respiratory viral infections and could provide the additional benefit of combining subtyping and real-time epidemiological analyses thus providing a high resolution understanding of the viral pathogens present within a population. For example, In Chapter 3, as mentioned above the subtypes or serotypes could be determined for hRV, RSV-A and B and hMPV-B. There was a suggestion that the generated sequences could be used for further

epidemiological analysis though the study was not set up to assess this aspect. Therefore when the occasion arose to study EV-D68 positive specimens, this was used as an opportunity to test this characteristic.

In chapter 4 we show that the sequence information generated by the NGS method as part of virus detection can be utilised in subsequent epidemiological analysis. The data generated by the established methods of type specific RT-PCR followed by nested PCR and Sanger sequencing was compared with that generated by NGS. We were able to show that identification of EV-D68 was comparable between both methods; the Sanger method targeted the VP1 coding region of the virus and was identified in 17 of 22 specimens. By NGS, the same region was sequenced in 17 of 22 specimens and an alternative coding region which can be used in viral typing, VP3, was identified in an additional case therefore the methods provided comparable results. Additionally near full genome sequences were generated from eight specimens.

Direct comparison of the sequences generated by both methods was available in 14 cases. Nucleotide differences were detected between methods in a small number of cases however this did not alter the phylogenetic analysis, which confirmed that EV-D68 isolated from individuals in the West of Scotland fell within multiple clades.

The generation of full genome sequences in a proportion of cases, 8 of 22, allowed recombination analysis. This would not be possible using Sanger sequencing unless multiple reactions were carried out to cover the entire genome. No recombination events were detected from the isolates tested.

Chapters 3 and 4 assessed the use of the NGS method on respiratory samples with the aim of determining whether such an approach could replace RT-PCR in that setting. However, the implementation of such a change in methodology would have a significant impact on the diagnostic service and workforce therefore it is important to consider if this same approach could be applied successfully to other areas of viral diagnostics. In chapter 5 we compared the NGS method to RT-PCR using a small panel of norovirus RT-PCR positive faecal specimens obtained from two healthcare associated outbreaks. In this proof of concept study, the presence of norovirus was confirmed in all samples, suggesting the NGS method had the potential to be used as a diagnostic assay in

this setting. As with chapters 3 and 4, the sequence reads correlated with RT-PCR Ct, highlighting the NGS method was semi-quantitative. Unfortunately the panel didn't contain other viral causes of gastroenteritis and no mixed infections were present. As a result we could not assess its ability to detect non norovirus targets of mixed infections.

In addition to virus detection, near full norovirus genome sequences were obtained in all cases. As was demonstrated in previous chapters, the breadth of genome sequencing was related to the initial viral load within a clinical specimen. In the case of these norovirus positive specimens, the average Ct was 16 (range 9.06 – 24.54) which is considerably lower than those demonstrated in the respiratory specimens. This would be expected given the high viral load associated with norovirus infection.

The generation of full genome coverage allowed genogroup, genotype and strain determination in all cases something which is not achievable with RT-PCR alone. This confirmed that in each outbreak the same strain was identified from all specimens.

Phylogenetic analysis of the ORF2 revealed that in each outbreak the clinical isolates were most closely related to each other, as would be expected, however in outbreak 1 this highlighted one specimen that fell on a different branch. A comparison of the nucleotide sequences generated showed that this sequence had a large number of nucleotide differences in comparison to the others from the same outbreak. This may indicate a second introduction event rather than a ward acquired case. Such information would not be readily available using RT-PCR, and although Sanger sequencing is carried out for epidemiological purposes this is not done in real time, therefore would provide retrospective understand of outbreak events.

The full genome sequences generated from clinical specimens could be analysed for recombination events. This revealed that no recombination occurred during the course of each outbreak although, again, this would be expected given the short duration and single circulating strain. The viruses isolated in outbreak 2 were the result of a previous recombination event with a break point between ORF1 and ORF2. Again, this analysis would not be possible with the current diagnostic or sequencing methods.

A theoretical benefit of using an NGS approach is the ability to study outbreak dynamics. One way of doing this would to look at single nucleotide polymorphisms (SNPs). The study of SNPs and variants has been used in determining the direction of virus spread in an outbreak setting (Hughes, Allen et al. 2012). The transmission of polymorphisms and low level variants are also a contributor to the diversity of norovirus strains in circulation. With the number of specimens and the time involved in this study it was not going to be possible to look at virus evolution over the course of the outbreak but this was carried as a proof of concept, to demonstrate that such things could be looked at using the type of data generated.

6.3 Impact to the laboratory

We have demonstrated that NGS is capable of detecting viral pathogens and in the process generates virus sequence information which can be utilised in subsequent typing and epidemiological studies. There are issues with sensitivity which have already been discussed and would need to be addressed prior instituting this in a diagnostic service. There are many other aspects to consider when implementing a new diagnostic test or method.

The cost of NGS is considerably greater than that of RT-PCR, around £70 per specimen in comparison to the multiplex assay used at WoSSVC which costs in the region of £20. This of course does not take in to account the cost of subsequent sequencing reactions which are carried out in a proportion of cases.

The ease of use must also be consider as applying complicated procedures to the throughput of specimens encountered by a regional diagnostic laboratory would not be feasible. The methods used here aimed to minimise the number of processes required with a view to reducing both cost and hands on time. Automation where possible would be the ideal option to both reduce the potential for error and decrease turnaround time.

There must also be standards and quality controls such as those out by Clinical Laboratory Improvement Amendments (CLIA) which include demonstrable reproducibility alongside positive and negative controls(Burd 2010). In the case of sequencing, there are measurable markers which are used in quality control. Using Illumina platforms as an example, the density of clusters generated during

the process must fall within a range; too low may indicate an issue with library preparation and too high can cause errors in base calling. A blank specimen i.e. water can be included to control for contamination, however the number of controls used should be considered in cost calculations as this will reduce the number of clinical specimens processed, increasing the overall cost per specimen. The addition of PhiX DNA is commonly used as a positive control for the sequencing reaction. This is added to the pooled libraries prior to sequencing, therefore does not affect the number of specimens which can be analysed. As a result, PhiX sequences can be detected within clinical specimens potentially resulting in erroneous results (Mukherjee, Huntemann et al. 2015).

The greatest obstacle to overcome with regards to introducing NGS to a diagnostic facility will not be specimen processing, rather data processing. As this panel of specimens was run in parallel with RT-PCR the potential false positive results were highlighted. Using NGS exclusively would require that a cut-off level of reads which may indicate a false positive result be determined. This number will vary depending on the sequencing platform used. The data from this study would suggest this is in the region of 10 reads however a greater number of specimens, including negative controls would be required.

There is no agreed consensus as to which are the most appropriate bioinformatic tools for the analysis of NGS data. The most effective tools will vary depending on the specimens, sequencing methods and the genomes under study (Bao, Jiang et al. 2011). The use of such tools requires a level of expertise and support from bioinformaticians. It is most likely that specific programs would be purpose built for laboratory and clinical staff that would not require in-depth knowledge of the bioinformatics process however a degree of support from someone with such expertise would be required.

6.4 Future work

Although the work outlined in this thesis outlines that NGS has the potential to be a diagnostic method in virology,…there are many pieces of future work that need to be carried out in order to improve and test the system still further before the NGS pipeline described could be considered “routine ready”.

The sensitivity of viral detection in this research was less than that of the standard diagnostic RT-PCR. As discussed above the lower limit of viral detection using NGS here were the equivalent of an RT-PCR Ct in the region of 32. To be implemented as a diagnostic test this would need to be improved by ~1 log in order to be in line with current methods. Various amendments to the pipeline could be used. For example, concentrating the nucleic acid extraction process could improve sensitivity. The current method used a specimen volume of 200 µl and eluted to 110 µl Increasing the input volume and reducing the output elution volume may be of benefit, for example 1 ml eluted in 60 µl.

The relationship between PCR Ct and viral reads sequenced would suggest that increasing the depth of sequencing could in turn increase the likelihood of virus detection. Many approaches could be undertaken to increase viral sequencing depth such as hybridisation and poly-A enrichment but these are essentially using targets in a similar fashion to PCR. Rather than enriching for the genomes of interest, an alternative approach would be to reduce the proportion of genomes that are not of interest. One method would be to reduce the quantity of human genetic material within a sample which can be achieved by exploiting the genetic differences between the host and the pathogen. Methods are available which target the methylated CpG sites within human DNA as these are rare in microbial genomes. Protein bound antibodies attached to magnetic particles bind to these sites, allowing them to be drawn out of a specimen by a magnet, resulting in an increased concentration of microbial genetic material in the remaining specimen (Figure 6-1.Microbiome enrichment by removing methylated CpG sites.). Ribosomal RNA can also be targeted in a similar fashion.

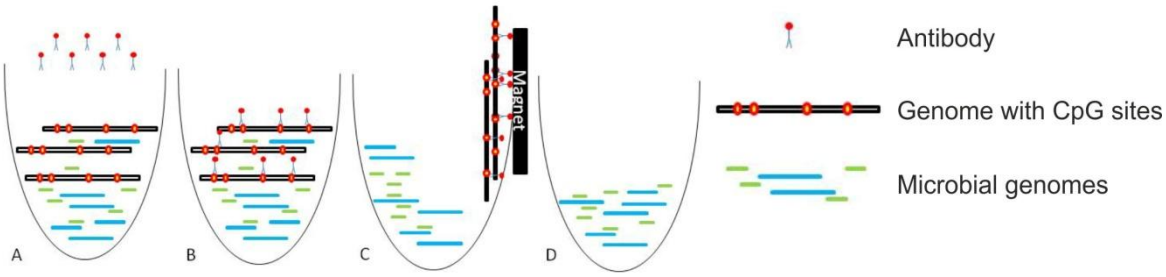


Figure 6-1.Microbiome enrichment by removing methylated CpG sites.
A) Addition of antibodies targeting methylated CpG sites, bound to magnetic beads. B) Antibodies bind to target if present in specimen. C) Application of a magnet, draws the antibodies and target out of solution. D) Remaining specimen, depleted of DNA containing methylated CpG sites.

However, it should be noted that the use of additional processing steps will in turn increase not just the cost and turnaround time but also the opportunity for errors. As the aim of this research was to explore the use of NGS in a diagnostic setting, these are all essential components to consider.

Another approach to increase sequencing depth would be to explore alternative platforms. In this research a benchtop sequencer, namely the MiSeq (Illumina) was used. Many platforms are available, some of which have greater throughput and output compared with this model, although the initial outlay for some of these may be prohibitive to a diagnostic service. It is important to note that even since the initial optimisation of this protocol there have been many upgrades to the reagents, hardware and software used here that have increased the volume of sequence data generation at no extra cost to the user.

The results presented in this research suggest that the sensitivity of detection may vary between viral pathogens. For example, in chapter 3, an instance of HCoV OC43 with a Ct of 24.05 was not detected however hRV were detected to a Ct of 33.66. This has also been suggested by others. To investigate this it would be useful to test dilution panels of each relevant pathogen by NGS and compare the results to the gold standard, RT-PCR. Only after this can we determine whether the detection limit is comparable across pathogens. Such a panel would also be useful in confirming that there is definitely a link between Ct and

sequence reads. This will only answer the sensitivity issues with known pathogens and any emergent pathogens would need to be assessed as they arise.

In chapter 3 a small number of specimens were found to have viral reads that were not confirmed by RT-PCR. This suggests there may an issue with false positivity using the NGS pipeline however such a problem was not encountered with specimens in chapters 4 and 5. To investigate further extended panels of known negative specimens should be tested to determine the extent of the issue. In the longer term, should this assay become a routine test, the use of negative controls, as per current testing, should be considered essential. The ability of the NGS pipeline to detect co infections is another aspect that needs to be improved. There are at least two issues relating to this. First the data from chapter 3 and 4 suggests that the NGS assay may not be able to detect DNA pathogens.

The protocol employed in this research required the use of reverse transcription and amplification as the quantity of DNA recovered directly from clinical specimens was not great enough to proceed to sequencing. The SISPA approach was chosen with the rationale that this would minimise any amplification bias as the aim was to use a non-targeted generic approach that could be used in any sample. As a result, in many cases full genome sequences could be reconstructed.

However, although successful at detecting and sequencing RNA targets the pipeline was poor at detecting DNA pathogens. This is probably a result of the initial reverse transcription step. It was initially hoped that RNA transcripts would be detected and thus allow the identification of pathogens with DNA genomes. This process was successful for the identification of bacterial genomes; however the same success was not seen for viral pathogens. While the majority of viruses affecting the respiratory tract are RNA viruses there are many important DNA pathogens that need to be detected. Furthermore if the method is to be applied to other viral disease syndromes then DNA detection is a must.

It is difficult to determine how an improvement in DNA detection could be achieved. Perhaps the methods employed in specimen preparation could be optimised for the identification of DNA alongside RNA. It could be that specimens are divided in two aliquots with DNA isolation from one and RNA isolation and

reverse transcription from the other. Specimens could then be merged for the remainder of the process. However this would have a significant impact on cost, throughput and turnaround time. As a result other solutions will need to be sought.

If this aspect can be improved then it is possible that this method could be used as a pan microbial test i.e. detection of viral, bacterial and fungal genomes from the same clinical specimen. Currently, the identification of these pathogens is carried out by separate laboratories requiring that multiple clinical specimens are sent to multiple locations. This is not always achieved therefore it is likely that diagnoses are being missed. Indeed in this research bacterial sequenced reads were detected in the majority of specimens however it was beyond the scope of this project to determine if these could be used for diagnostic purposes.

The other issue relating to missed co-infections may mostly be related to the detection limit of the NGS assay as most were weak positives with Cts >32. The latter aspect could be answered via testing the dilution panels outlined above. Further experiments using viral mixtures at differing concentrations would also be beneficial to rule out target competition as a cause of missed detection.

The turnaround time demonstrated in this research was in the region of seven days from the beginning of specimen preparation to initial results. This is obviously much greater than that of RT-PCR where results can be available in a matter of hours from the specimen arriving in the laboratory. Here, all preparation steps were carried out manually. This is both time consuming and leaves the process open to human error. To overcome this future work should examine whether we can automate these steps. The use of automation in the preparation of specimens would allow the setup process to be carried out 24 hours a day with minimal hands on steps. This will both reduce the overall turnaround time but also the potential for human error. It is also likely that in coming years improvements to the methodology described which will have an impact on turnaround time and ease of use. Using an alternative sequencing platform such the Ion Torrent or MinIon will also likely improve turnaround time.

Improvements into bioinformatic analysis are needed – although identification was automated and therefore straightforward, any further analysis required

more complex user input. Having a more automated process would be ideal, especially if this were to be considered for a routine service. As mentioned previously, custom built interpretation software would be necessary. An example of this would be DisCVR (Maabar 2016).

To test the true potential of NGS in the diagnosis of IID, it would ideal to examine outbreaks involving multiple pathogens and mixed infections over a long period of time. It would also be interesting to apply the technique to outbreaks involving those with compromised immune systems to determine virus evolution within hosts. Other syndromes which may benefit from an NGS approach would be those of CNS infections as the diagnostic samples obtained in these cases come from what should be an acellular sterile site therefore host DNA contamination should present less of an issue. The diagnostic specimens can be difficult to obtain, are often a small volume and may be required for multiple tests so the application of a single assay would be of a great benefit.

If successful then the next stage would be to focus on the detection of bacteria. The approach to infection management is changing to focus on delivering an infection service, combining clinical, microbiology and virology services. This is reflected in the fact that many virology services are carrying out molecular assays to detect bacteria which are difficult to identify by culture, such as *Mycoplasma pneumoniae* and *Legionella sp*..

The ideal scenario would be to implement NGS methods into routine practice to see how it goes.

6.5 Thesis Conclusion

The thesis expounded in this research is that an in house developed metagenomic Next Generation Sequencing pipeline has the potential to be used for both the diagnosis and epidemiological analysis of viral infection. I have used the aforementioned method to diagnose, quantitate and study the epidemiology of respiratory pathogens directly from a small panel of clinical samples. The method has also successfully been applied to the detection, quantitation and epidemiological analysis of norovirus directly from extracted stool samples. Although improvements are necessary, the pipeline has shown potential to be used as a future diagnostic test for both viral and non viral infectious diseases.

Appendix 1. Commands used in data analysis

Quality trimming

```
#!/bin/bash
############################### Name files ############################
for file1 in *L001_R1_001.fastq
do
        file2=${file1%L001_R1_001.fastq}L001_R2_001.fastq
echo $file1 $file2
###################### adapter and quality trimming ####################
        echo "Trimming ends with trim_galore";
        echo "Trimming Nextera adapters";
        echo "trim_galore --paired --length 30 --adapter CTGTCTCTTATACACATCT $file1 $fiel2;"
        trim_galore --paired --length 30 --adapter CTGTCTCTTATACACATCT $file1 $file2;
        echo "Trimming standard Illumina adapters"
        trim_galore --paired --length 30 ${file1%.fastq}_val_1.fq ${file2%.fastq}_val_2.fq;

        echo "Trimming FR20RV primer"
        trim_galore --paired --length 30 --adapter GCCGGAGCTCTGCAGATATC ${file1%.fastq}_val_1_val_1.fq ${file2%.fastq}_val_2_val_2.fq;
####################### Rename files ###################################
        mv ${file1%.fastq}_val_1_val_1_val_1.fq ${file1%.fastq}_validated_1.fq;
        mv ${file2%.fastq}_val_2_val_2_val_2.fq ${file2%.fastq}_validated_2.fq;
####################### Trimming ends with Prinseq #####################
        echo "Trimming ends with prinseq";
        prinseq -fastq ${file1%.fastq}_validated_1.fq -fastq2 ${file2%.fastq}_validated_2.fq -lc_method dust -lc_threshold 20 -trim_left 20 -derep 12345;
# -lc_method = trimming low entropy reads from dataset, threshold can be set using -lc_threshold
# -trim_left = removing first 20 bases from 5' end of sequence (this retains paired end information)
```

```
# -derep = removing dulpicate reads (exact duplicates, 5' and 3' duplicates)


        mv ${file1%.fastq}_validated_1_prinseq_good_singletons*
${file1%.fastq}_validated_1_singletons.fq;
        mv ${file2%.fastq}_validated_2_prinseq_good_singletons*
${file2%.fastq}_validated_2_singletons.fq;
        mv ${file1%.fastq}_validated_1_prinseq_good*
${file1%.fastq}_validated_1_good.fq;
        mv ${file2%.fastq}_validated_2_prinseq_good*
${file2%.fastq}_validated_2_good.fq;
        trim_galore --paired --length 30 ${file1%.fastq}_validated_1_good.fq
${file2%.fastq}_validated_2_good.fq;
        mv ${file1%.fastq}_validated_1_good_val_1.fq ${file1%.fastq}_clean.fq;
        mv ${file2%.fastq}_validated_2_good_val_2.fq ${file2%.fastq}_clean.fq;
file3=${file1%L001_R1_001.fastq}L001_R1_001_clean.fq;
file4=${file2%L001_R2_001.fastq}L001_R2_001_clean.fq;
echo $file3 $file4
################### Map to human references ############################
echo "Mapping trimmed reads to human reference files"
        weeMapper -1 $file3 -2 $file4 -a
######################## Remove report files ############################
rm *report.txt;
rm *val_1.fq;
rm *val_2.fq;
rm *bad*;
#################### Rename output files ##############################
mv ${file1%L001_R1_001.fastq}*unmapped_1.fq
${file1%L001_R1_001.fastq}unmapped_clean_1.fq;
mv ${file2%L001_R2_001.fastq}*unmapped_2.fq
${file1%L001_R1_001.fastq}unmapped_clean_2.fq;
###################### Aligned to viral databases ########################
echo "aligning to viral genome and cDNA databases"
bowtie2 -x ~/Reference/viral_genomes -1
${file1%L001_R1_001.fastq}unmapped_clean_1.fq -2
${file1%L001_R1_001.fastq}unmapped_clean_2.fq --very-sensitive -S
${file1%L001_R1_001.fastq}viral_genomes.sam
```

samtools view -bS ${file1%L001_R1_001.fastq}viral_genomes.sam > ${file1%L001_R1_001.fastq}viral_genomes.bam

echo "Get unmapped reads from viral genomes"

bam2fastq --no-aligned --force --strict -o ${file1%L001_R1_001.fastq}viral_genomes#.fq ${file1%L001_R1_001.fastq}viral_genomes.bam

bam2fastq --no-unaligned --force --strict -o ${file1%L001_R1_001.fastq}viral_genomes_unmapped#.fq ${file1%L001_R1_001.fastq}viral_genomes.bam

echo "Aligned to viral cDNA"

bowtie2 -x ~/Reference/viral_cDNA -1 ${file1%L001_R1_001.fastq}viral_genomes_unmapped_1.fq -2 ${file1%L001_R1_001.fastq}viral_genomes_unmapped_2.fq --very-sensitive -S ${file1%L001_R1_001.fastq}viral_cDNA.sam

echo "Get mapped reads from viral cDNA"

samtools view -bS ${file1%L001_R1_001.fastq}viral_cDNA.sam > ${file1%L001_R1_001.fastq}viral_cDNA.bam

bam2fastq --no-aligned --force --strict -o ${file1%L001_R1_001.fastq}viral_cDNA#.fq ${file1%L001_R1_001.fastq}viral_cDNA.bam

echo "get unmapped reads from viral cDNA"

samtools view -bS ${ file1%L001_R1_001.fastq }viral_cDNA.sam > ${file1%L001_R1_001.fastq}viral_cDNA.bam

bam2fastq --no-unaligned --force --strict -o ${file1%L001_R1_001.fastq}viral_cDNA_unmapped#.fq ${file1%L001_R1_001.fastq}viral_cDNA.bam

done

Commands used in Data analysis: Metamos pipeline

calculate average inert size of mapped read pairs – required for contigs assembly

Getinsertsize.py file.sam

run pipeline

initPipeline -q -1 file_human_unmapped_1.fq -2 file_human_unmapped_2.fq -d file_metamos_directory –i insert size

```
runPipeline -d file_metamos_directory -c  blast -m bowtie2 -a
sparseassembler,idba-ud,velvet -p 2
```

Appendix 2. Respiratory Specimens: Summary of Results

Sample	PCR	Ct	Total Reads	% mapped to human reference	Taxonomy	Accession	% reads mapping	% Genome Covered	Nucleotide Mismatch
1B1	HRV	18.73	546282	20.17	HRV-A49	JN798589	75.48	100	2.5
1B2	HRV	26.78	638257	93.07	HRV-B92	FJ445169	2.53	56.4	7.5
1B5	PIV-3	27.57	696929	50.67	PIV-3	KJ672618	5.62	5.0	0.5
1B7	HRV	28.94	537522	62.21	N/A				
1B8	Influenza A	22.34	818055	96.78	Influenza A H3N2	KJ942712	0.03	30.6 – 60.3	various
1C2	HRV	26.75	1107488	81.67	Enterovirus D-68	AB601885	1.46	14.9	1.4
1C3	HRV	28.15	1278307	65.56	HRV-A97	FJ445172	0.31	20.9	8.7
1C7	HRV	24.93	538412	69.88	HRV-A1	JN837694	4.76	87.4	1.2
1C9	hMPV	35.51	837310	47.23	N/A				
1D1	HRV	20.68	1010455	59.82	HRV-A1	JN837694	32.03	100	1.9
1D3	HCoV NL63	27.17	801021	96.66	HCoV NL63	KF530112	0.38	11.7	0.5
1D4	HRV	19.81	1089553	23.19	HRV-A1	JN837694	74.67	99.6	2.2
1E1	HRV	20.1	758165	61.76	HRV-C15	GU219984	33.93	100	2.0
1E3	RSV	20.84	700028	93.78	RSVA	KJ627329	3.71	91.0	0.7
1E5	RSV	30.01	714509	96.87	N/A				
1E7	HCoV NL63	24.36	426736	92.10	HCoV NL63	JQ765569	0.97	23.3	0.5
1E8	HRV	25.09	887501	96.50	HRV-C15	GU219984	0.06	29.7	2.5
1F1	RSV	25.08	322516	85.54	RSVA	KF826849	3.79	54.6	0.5
1F7	HRV	28.03	236596	70.59	N/A				
1F8	HCoV NL63	25.22	644660	93.24	HCoV NL63	JQ765567	0.03	3.7	0.4
1G1	HRV	28.49	265505	78.73	HRV-A90	FJ445167	0.02	2.6	10.8
1G1	Adeno	35.1	265505	78.73	N/A				
1G2	PIV-3	28.09	309778	78.95	PIV-3	KJ672606	0.04	2.8	0.6

1G6	hMPV	26.84	81749	37.68	hMPV-B	KJ627397	1.41	18.5	2.0
1H1	HCoV OC43	17.21	608153	94.92	HCoV OC43	KF530099	3.48	99.9	1.6
1H3	HRV	20.44	413108	83.13	HRV-A2	X02316	6.65	96.0	8.1
1H7	HRV	19.39	657627	56.80	HRV-B4	JN798573	39.08	99.7	1.4
1I2	HRV	19.80	710612	60.67	HRV-C17	JN815240	33.29	99.1	3.8
1I4	HRV	26.01	641357	98.15	HRV-B3	JF285331	0.64	97.2	2.4
1I5	HRV	33.66	728425	92.82	HRV-A1	JN837694	0.01	39.0	5.6
1I6	HRV	29.11	1063930	57.27	HRV-C (UTR)	JX129433	2.92	N/A	2.3
1I8	HRV	27.85	1132106	99.02	HRV-B27	JF285309	0.01	6.7	3.7
2A2	HCoV 229E	32.35	745105	98.00	N/A				
2A3	HCoV 229E	18.14	1110727	88.23	HCoV 229E	JX503060	9.63	99.8	0.6
2A4	HCoV 229E	20.00	311141	98.39	HCoV 229E	JX503060	0.16	55.3	0.6
2A6	HRV	23.12	1056405	83.90	HRV-C11	EU840952	0.29	27.2	8.9
2A9	hMPV	25.68	626886	84.12	hMPV-B	KF530171	6.35	48.4	3.1
2B4	RSV	19.32	612887	55.78	RSVB	JX576741	40.52	99.3	2.0
2B6	RSV	33.96	854593	77.09	N/A				
2B7	HRV	17.24	596971	95.94	HRV-A60	JN798590	1.82	95.3	3.7
2B9	HRV	33.31	569654	76.13	N/A				
2C1	HRV	20.24	816200	91.89	HRV-A21	JN837693	5.32	97.9	2.5
2C3	HCoV 229E	25.83	470840	93.93	HCoV 229E	JX503060	0.48	20.0	0.7
2C4	HCoV 229E	33.84	759226	98.06	N/A				
2D2	HCoV 229E	19.74	743683	94.07	HCoV 229E	JX503060	2.50	73.9	0.6
2D3	HCoV OC43	24.05	765913	89.42	N/A				
2D4	PIV-2	36.07	1034235	96.30	N/A				
2D5	HCoV OC43	14.63	992040	93.04	HCoV OC43	JN129835	2.70	97.6	0.7
2D6	HRV	23.07	1016390	82.29	HRV-A8	FJ445113	6.23	97.0	8.1

Table A-1. An overview of the total, human and viral sequenced reads per sample with a positive RT-PCR result.

Appendix 3. Clinical Details Associated with EV-D68 RT-PCR Positive Specimens

Sample ID	Age	Gender	Sample Date	Sample Type	Inpatient	First Results	EV D68 Ct
427086	49w	M	10/10/2014	THS	Y	Rhino/Entero	31.16
427334	1	M	14/10/2014	NPA	Y	Rhino/Entero	22.69
427759	1	M	17/10/2014	NPA	Y	Rhino/Entero	24.02
428005	1	M	21/10/2014	NPA	Y	Rhino/Entero, HPIV-4, Adeno	20.94
428008	18w	M	21/10/2014	NPA	Y	Rhino/Entero	28.79
428129	30w	F	22/10/2014	BAL	ICU	Rhino/Entero	25.39
428134	30w	F	22/10/2014	NS	ICU	Rhino/Entero	24.2
428287	39w	M	23/10/2014	THS	Y	Rhino/Entero	27.63
428871	1	F	29/10/2014	NPA	N	Rhino/Entero, Adeno	21.92
429031	11	F	29/10/2014	TNS	N	Rhino/Entero	35.24
429110	6	M	02/11/2014	Swab	Y	Rhino/Entero, HHV-6	32.76
429159	41w	F	02/11/2014	NPA	Y	Rhino/Entero, RSV	32.82
429319	41w	F	03/11/2014	NPA	Y	Rhino/Entero, RSV, HPIV 4	25.5
429323	22w	M	03/11/2014	NPA	Y	Rhino/Entero	16.25
429395	3	F	04/11/2014	THS	Y	Rhino/Entero	29.75
429660	1	F	07/11/2014	NPA	Y	Rhino/Entero	28.11
429915	1	M	07/11/2014	NS	Y	Rhino/Entero	28.07
430038	1	M	07/11/2014	NPA	Y	Rhino/Entero, HPIV-4	24.96
430139	1	M	11/11/2014	NPA	Y	Rhino/Entero, RSV	30.58
430146	1	M	11/11/2014	NPA	Y	Rhino/Entero, hMPV	21.26
430741	1	F	18/11/2014	NPA	Y	Rhino/Entero, HPIV-4	20.24
431467	4	M	23/11/2014	THS	Y	Rhino/Entero	26.53

Table A-2. Clinical information associated with study samples.

An overview of the clinical details associated with each EV-D68 positive sample along with the results of the initial diagnostic RT-PCR screen NPA = nasopharyngeal aspirate, NS = nose swab, THS = throat swab, BAL = bronchio-alveolar lavage and Swab = swab with site unspecified).

Appendix 4. EV-D68 Full Genome Sequences Used in Analyses.

Accession	Complete genomes as of 16/6/15
AY426531.1	Human enterovirus 68 strain Fermon
JX070222.1	Human enterovirus 68 isolate NZ-2010-541
JX101846.1	Human enterovirus 68 strain NYC403 from USA
AB601883.2	Human enterovirus 68 genomic RNA
AB601882.2	Human enterovirus 68 genomic RNA
EF107098.1	Human enterovirus 68 isolate 37-99 from France
KP114665.1	Enterovirus D68 isolate EV68_Alberta17789_2014
KP114664.1	Enterovirus D68 isolate EV68_Alberta2985_2014
KP114662.1	Enterovirus D68 isolate EV68_Alberta17390_2014
KF726085.1	Enterovirus D68 isolate BCH895A
KP745770.1	Enterovirus D68 isolate NY77
KP745769.1	Enterovirus D68 isolate NY74
KP745768.1	Enterovirus D68 isolate NY73
KP745767.1	Enterovirus D68 isolate NY329
KP745766.1	Enterovirus D68 isolate NY328
KP745765.1	Enterovirus D68 isolate NY326
KP745764.1	Enterovirus D68 isolate NY316
KP745763.1	Enterovirus D68 isolate NY314
KP745762.1	Enterovirus D68 isolate NY309
KP745761.1	Enterovirus D68 isolate NY305
KP745760.1	Enterovirus D68 isolate NY278
KP745759.1	Enterovirus D68 isolate NY275
KP745758.1	Enterovirus D68 isolate NY263
KP745757.1	Enterovirus D68 isolate NY210
KP745756.1	Enterovirus D68 isolate NY160
KP745755.1	Enterovirus D68 isolate NY153
KP745754.1	Enterovirus D68 isolate NY130
KP745753.1	Enterovirus D68 isolate NY126
KP745752.1	Enterovirus D68 isolate NY124
KP745751.1	Enterovirus D68 isolate NY120
KP126911.1	Enterovirus D68 strain US/CO/14-93
KP100794.1	Enterovirus D68 strain US/CO/13-60
KM881710.2	Enterovirus D68 strain EV-D68_STL_2014_12
KP240936.1	Enterovirus D68 isolate Beijing-R0132
KM851231.1	Enterovirus D68 strain US/KY/14-18953
KM851228.1	Enterovirus D68 strain US/MO/14-18950
KM851227.1	Enterovirus D68 strain US/MO/14-18949
KM851226.1	Enterovirus D68 strain US/MO/14-18948
KM851225.1	Enterovirus D68 strain US/MO/14-18947
KP100796.1	Enterovirus D68 strain US/CA/14-6100
KP100795.1	Enterovirus D68 strain US/CA/14-6103SIB
KP100793.1	Enterovirus D68 strain US/CO/14-94
KP100792.1	Enterovirus D68 strain US/CA/14-6092
KM892501.1	Enterovirus D68 isolate CA/AFP/11-1767
KM892500.1	Enterovirus D68 isolate CA/RESP/10-786
KM892499.1	Enterovirus D68 isolate CA/AFP/v12T00346

Table A-3. Full genome Enterovirus D68 genome sequences available from PubMed on 16/6/2016.

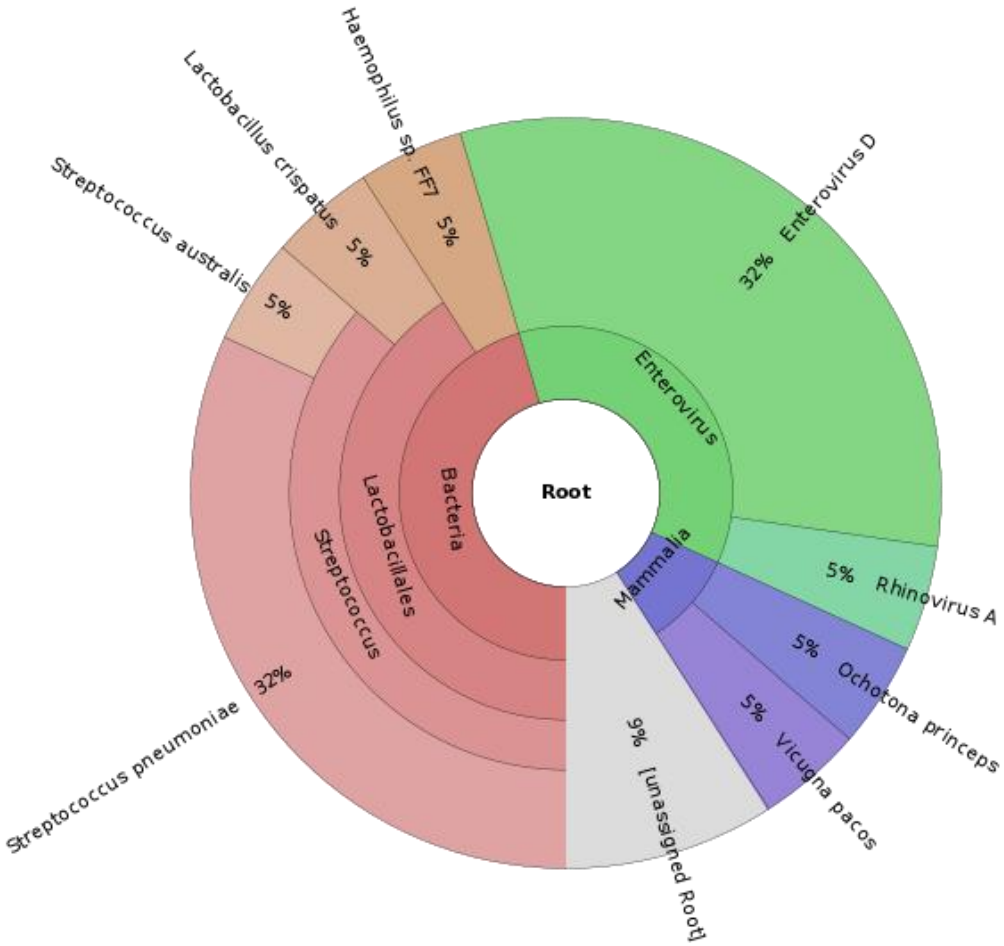
Appendix 5. EV-D68 Specimens: Summary of Results.

Sample	Ct Value	Top Database Hit (Accession)	Trimmed reads	Database alignment (Reads)	Coverage (%)	Single reference alignment (Reads)	Reads mapping (%)	Coverage (%)	Nucleotide Mismatch (%)
427086	**31.16**	KP745768	820028	143	**26.04**	374	0.05	**53.51**	1.3
427334	**22.69**	KP745769	154136	40833	**95.12**	107950	70.04	**95.12**	1.3
427759	**24.02**	KP745768	99208	8470	**93.51**	26678	26.89	**94.11**	1.5
428005	**20.94**	KP745768	365842	78674	**99.70**	205255	56.10	**99.97**	1.2
428008	**28.79**	KP745769	28980	733	**48.82**	1643	5.67	**68.99**	1.7
428129	**25.39**	KP126911	49208	61	**9.94**	1188	2.41	**22.40**	0.7
428194	**24.2**	KP100793	60660	421	**49.72**	6862	11.31	**84.90**	0.6
428287	**27.63**	KM851231	960540	254	**26.27**	336	0.03	**36.94**	2
428871	**21.92**	KP745769	78434	6505	**99.11**	17813	22.71	**99.93**	1.3
429031	**35.24**	KP745768	133912	49	**50.94**	129	0.10	**68.12**	2
429110	**32.76**	KP745769	257764	21	**19.99**	51	0.02	**37.90**	1.5
429159	**32.82**	KP745768	138358	3	**2.73**	5	0.00	**4.18**	4.8
429319	**25.5**	KP745768	67598	238	**41.89**	624	0.92	**47.35**	1.5
429323	**16.25**	KP745769	396444	109241	**100**	306593	77.34	**100**	1.5
429395	**29.75**	KP745769	692546	17	**20.85**	35	0.01	**23.30**	5.3
429660	**28.11**	KP745768	152066	3409	**93.81**	10544	6.93	**94.62**	1.4
429915	**28.07**	KP745768	593212	332	**44.71**	730	0.12	**58.60**	1.1
430038	**24.96**	KP745768	116842	6320	**87.52**	17976	15.38	**90.97**	1.3
430139	**30.58**	KP745768	403344	73	**43.94**	202	0.05	**64.55**	1.8
430146	**21.26**	KP745769	226678	52224	**99.52**	141956	62.62	**100**	1.5
430741	**20.24**	KP745769	335942	23898	**99.99**	69601	20.72	**100**	1.5
431467	**26.53**	KP745768	877546	8603	**78.68**	22945	2.61	**86.80**	1.5

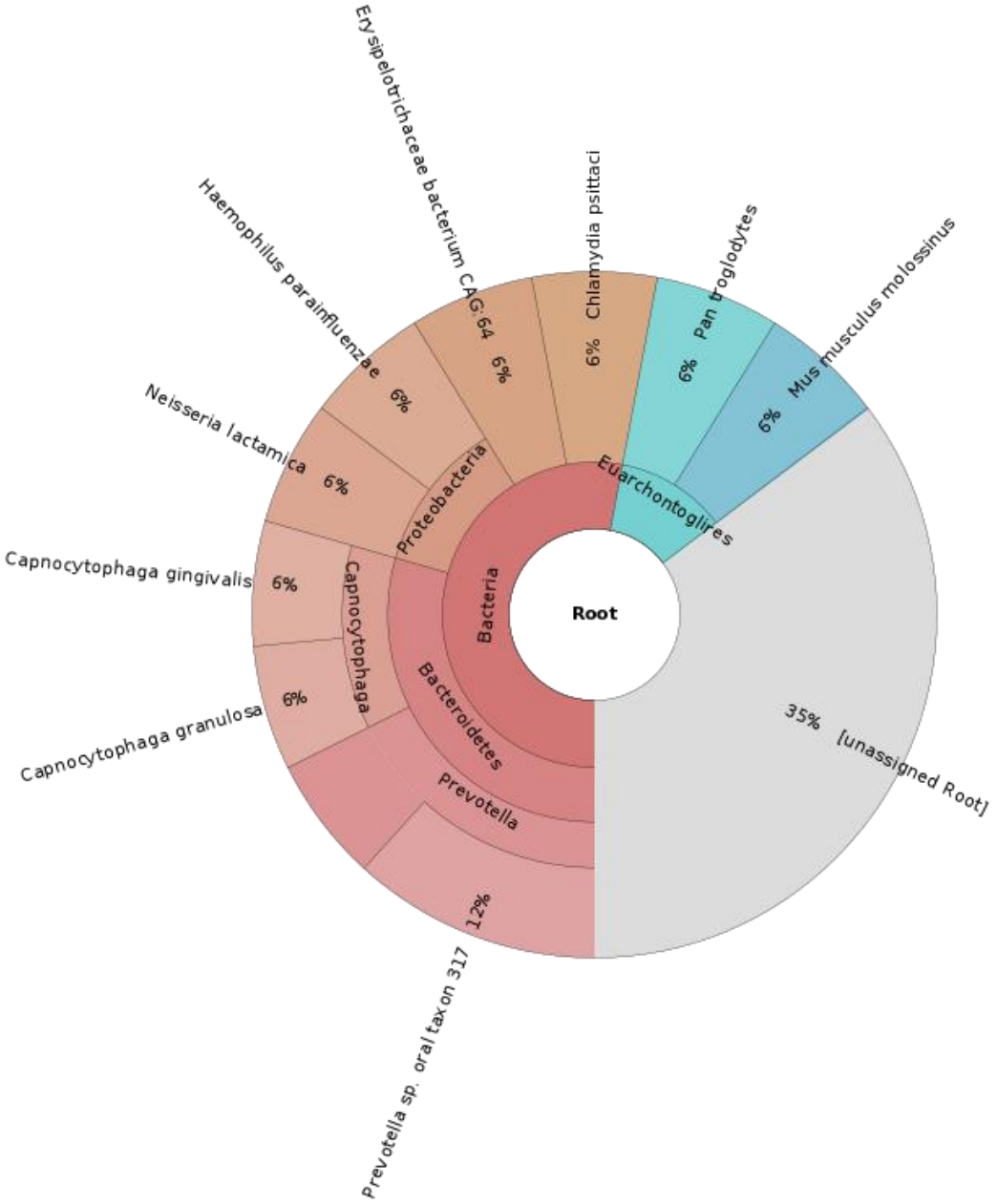
Table A-4. A summary of RT-PCR and NGS results.

Appendix 6. Krona Output Charts from EV-D68 RT-PCR Positive Specimens.

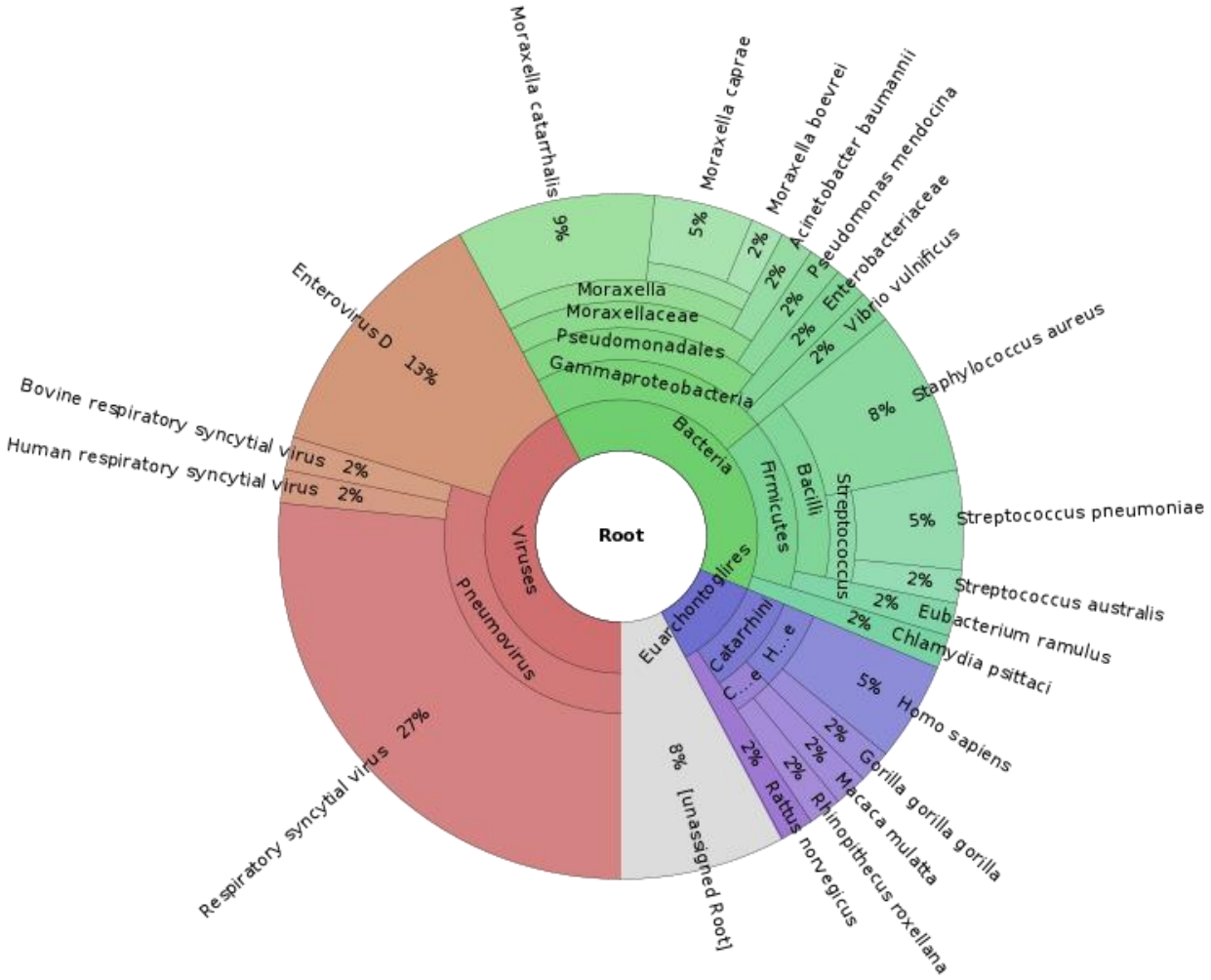
Specimen 430038



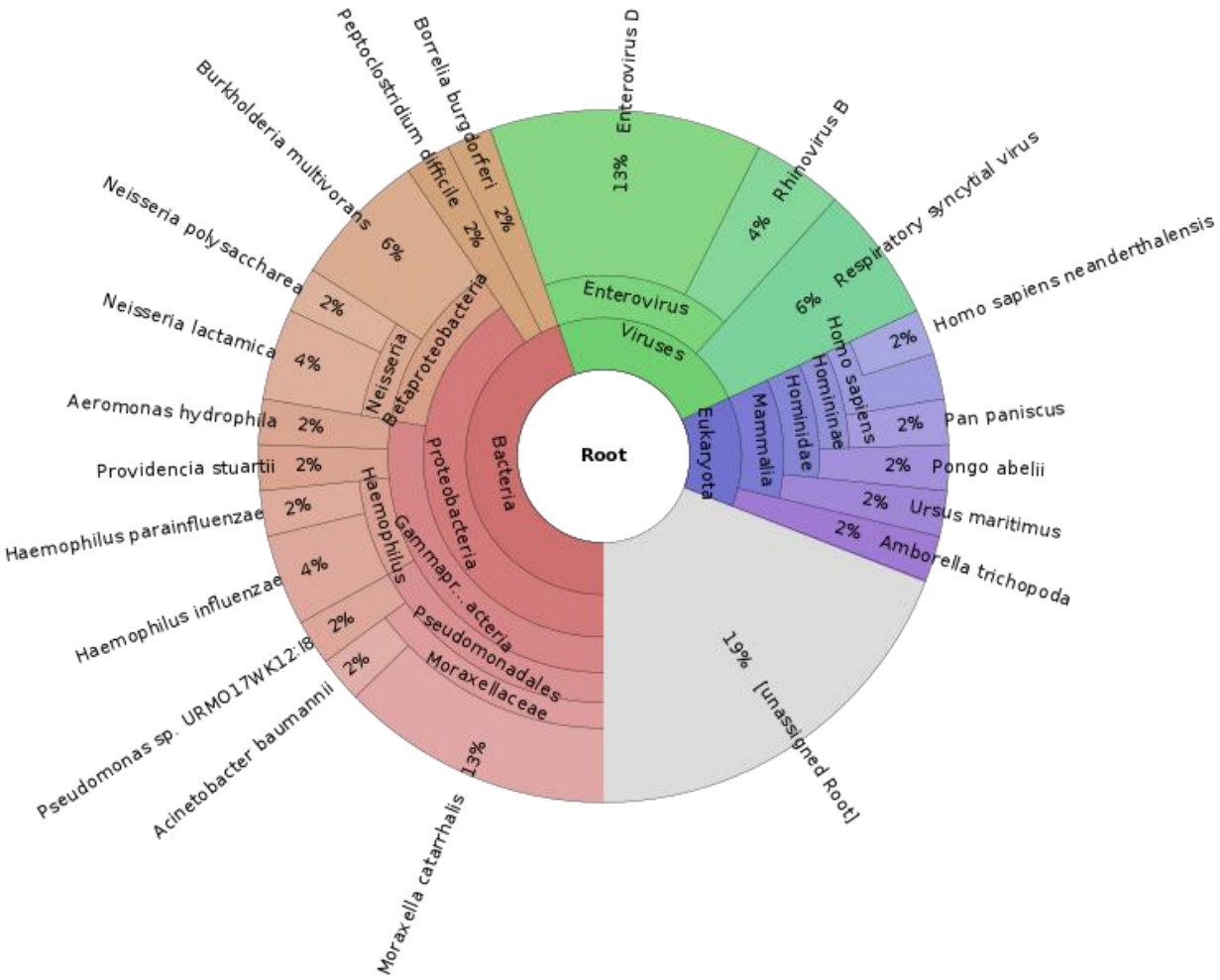
Specimen 429110



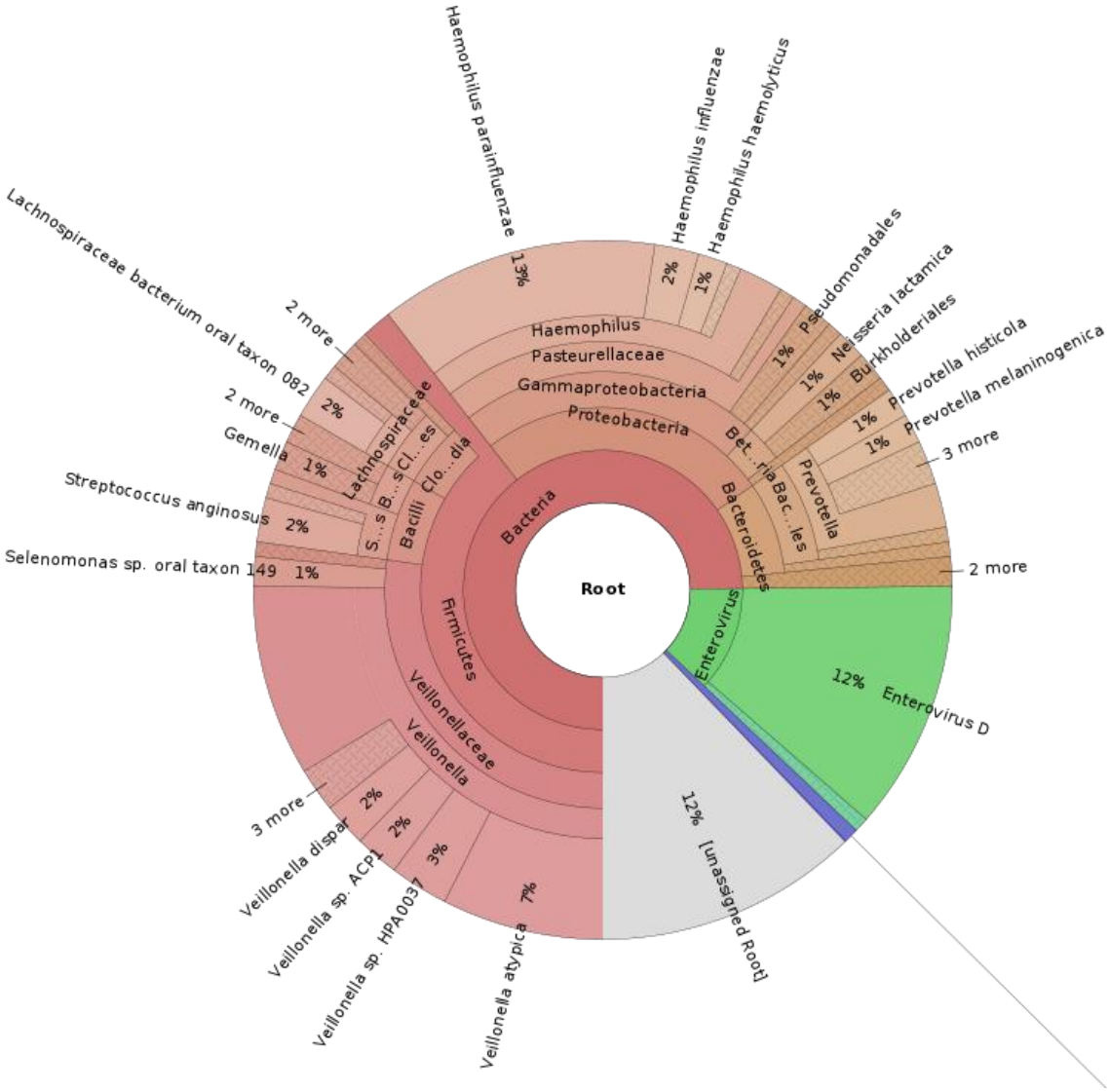
Specimen 430139



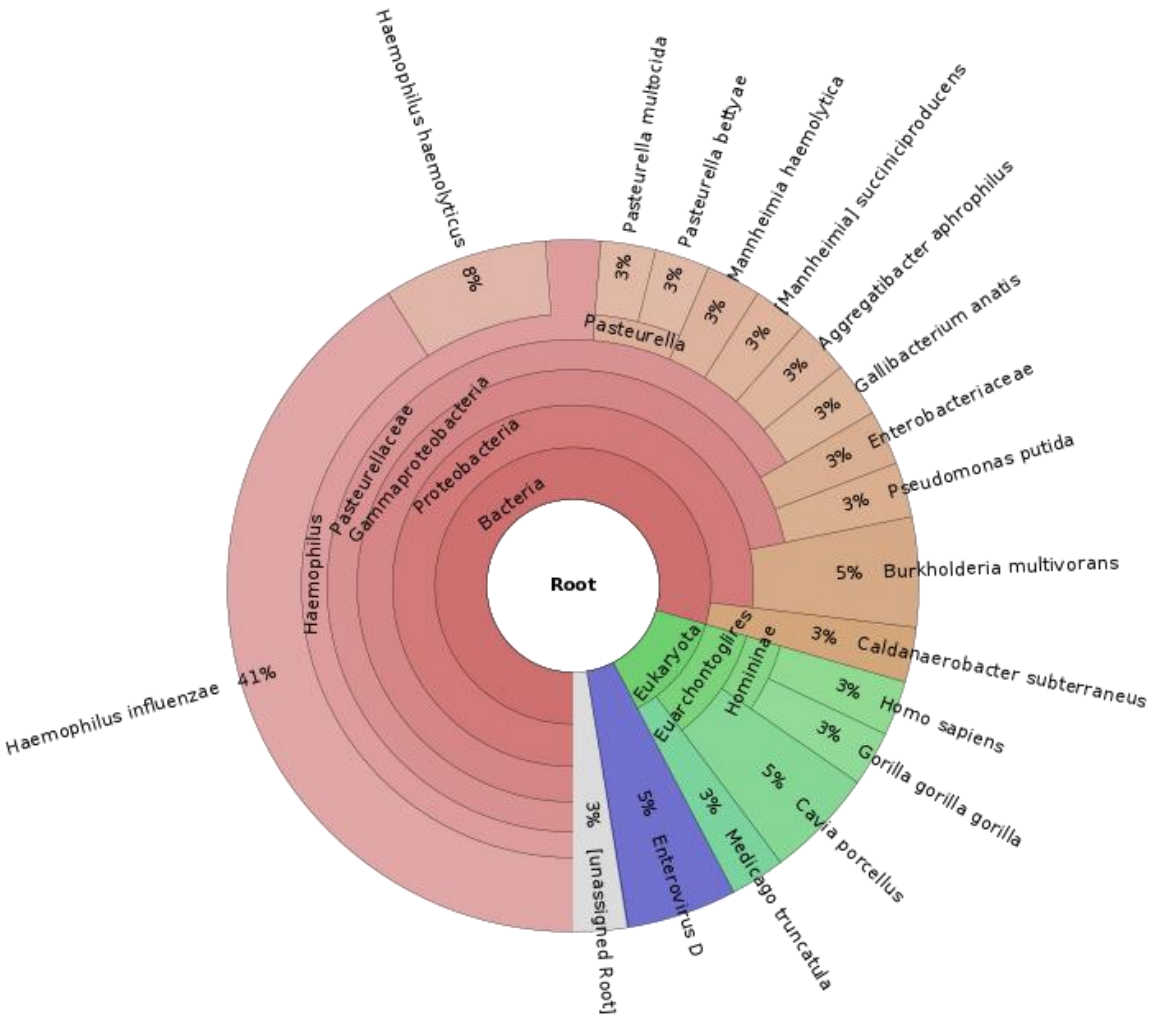
Specimen 430741



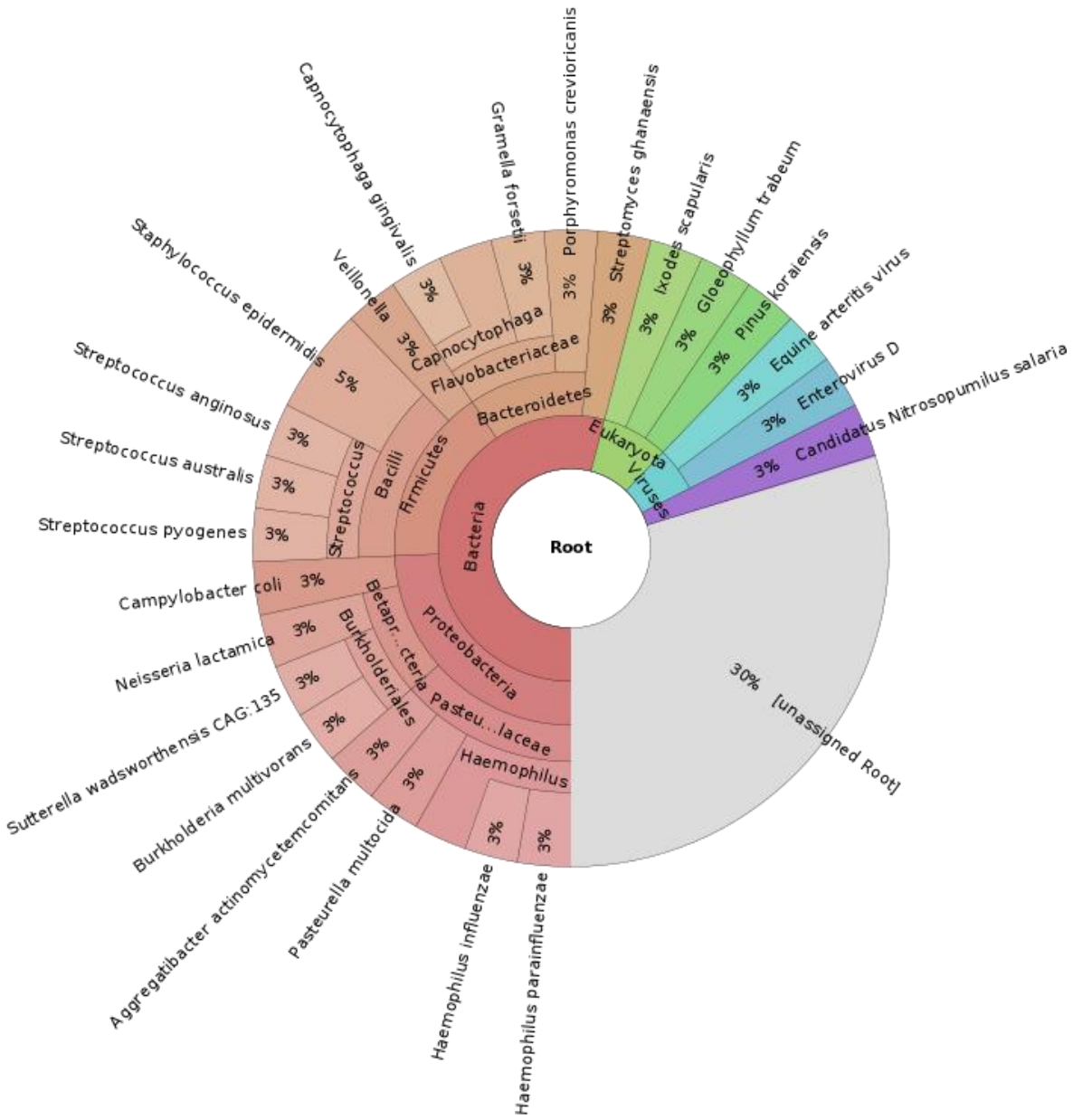
Specimen 431467



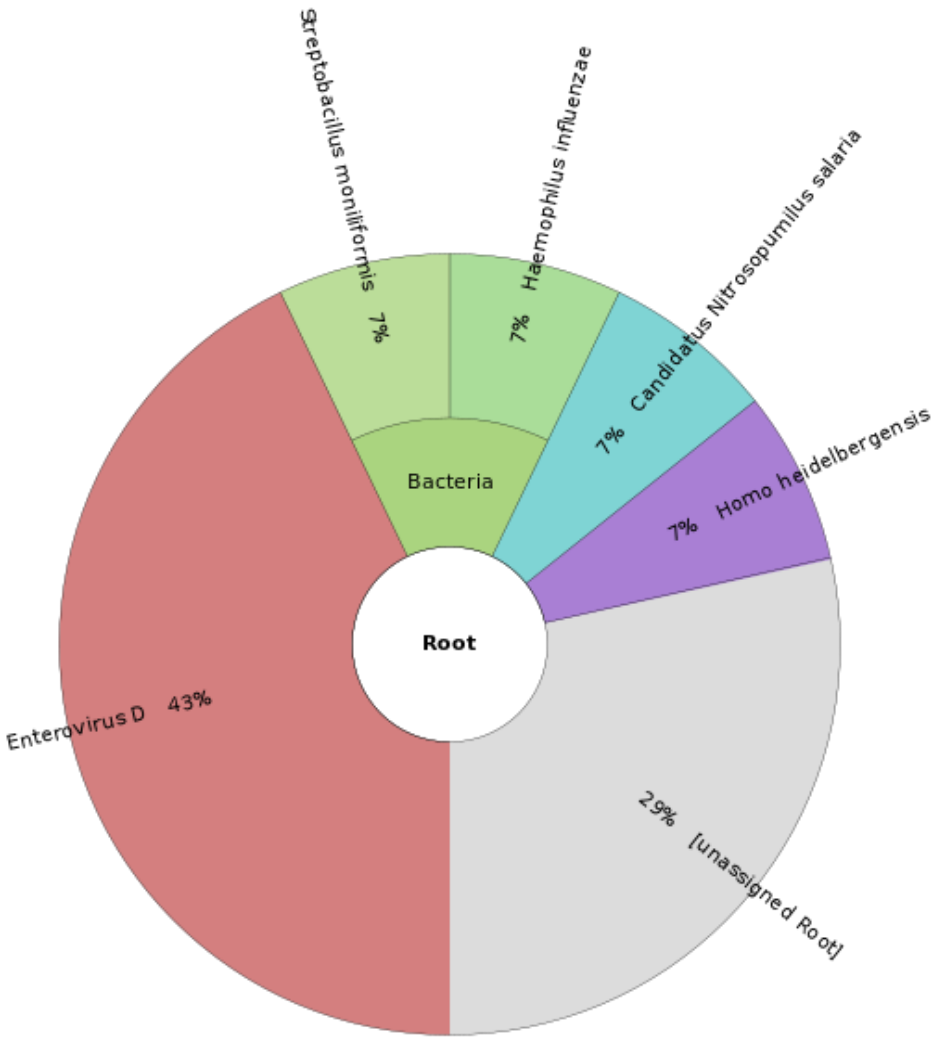
Specimen 430146



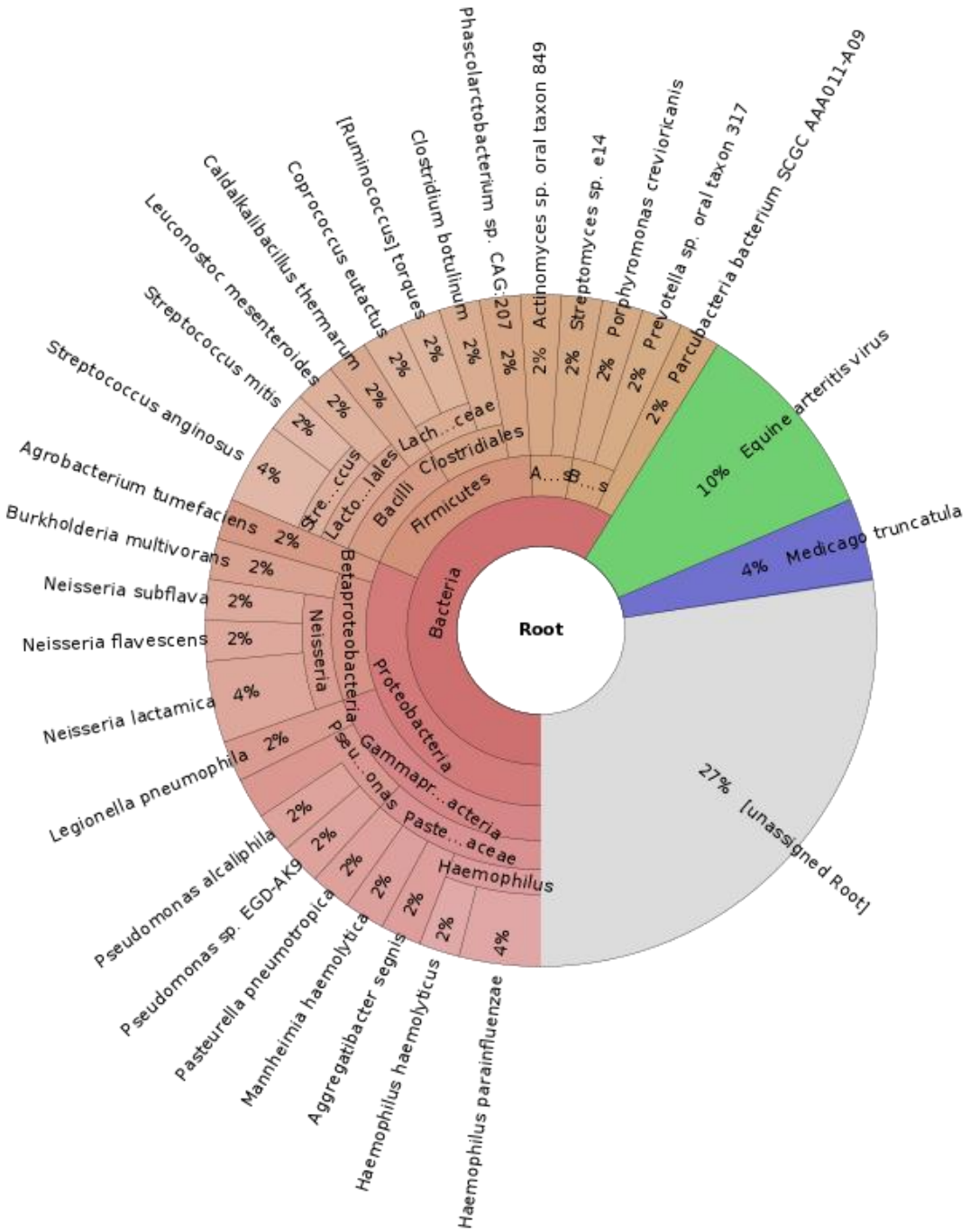
Specimen 427086



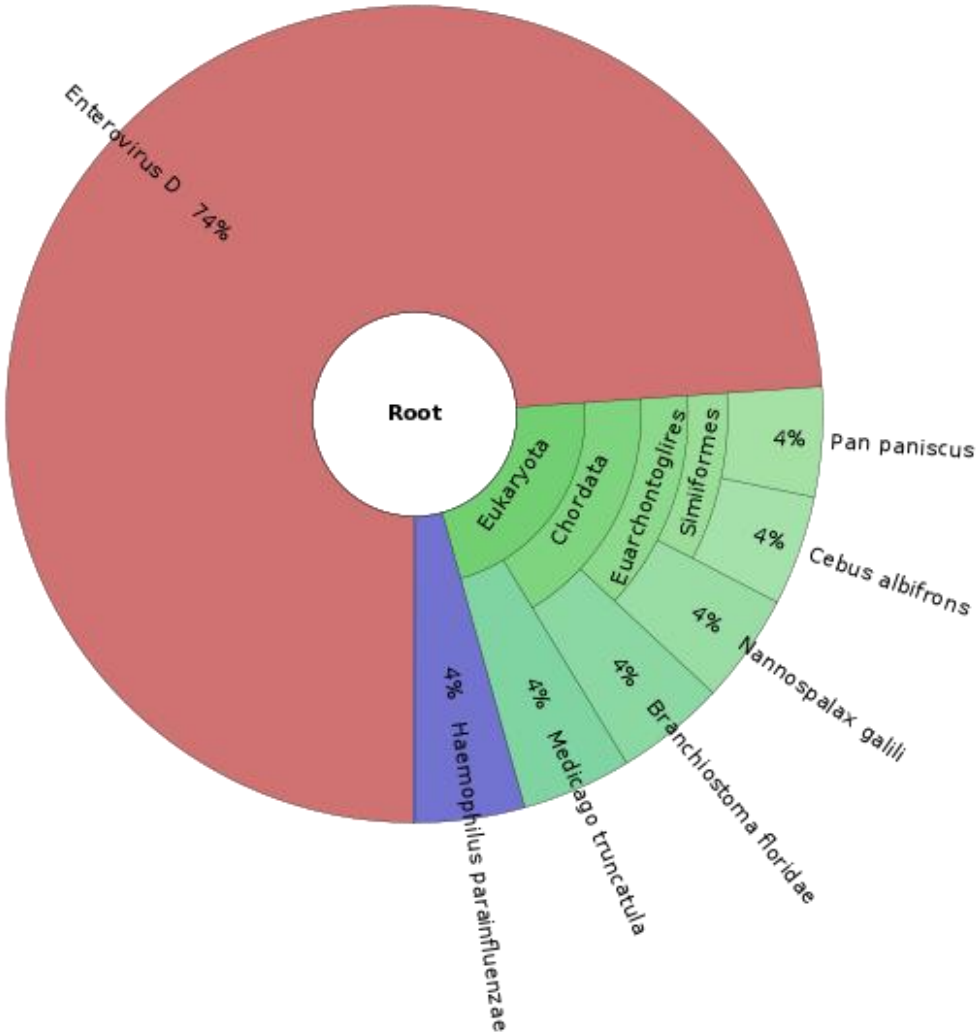
Specimen 427334



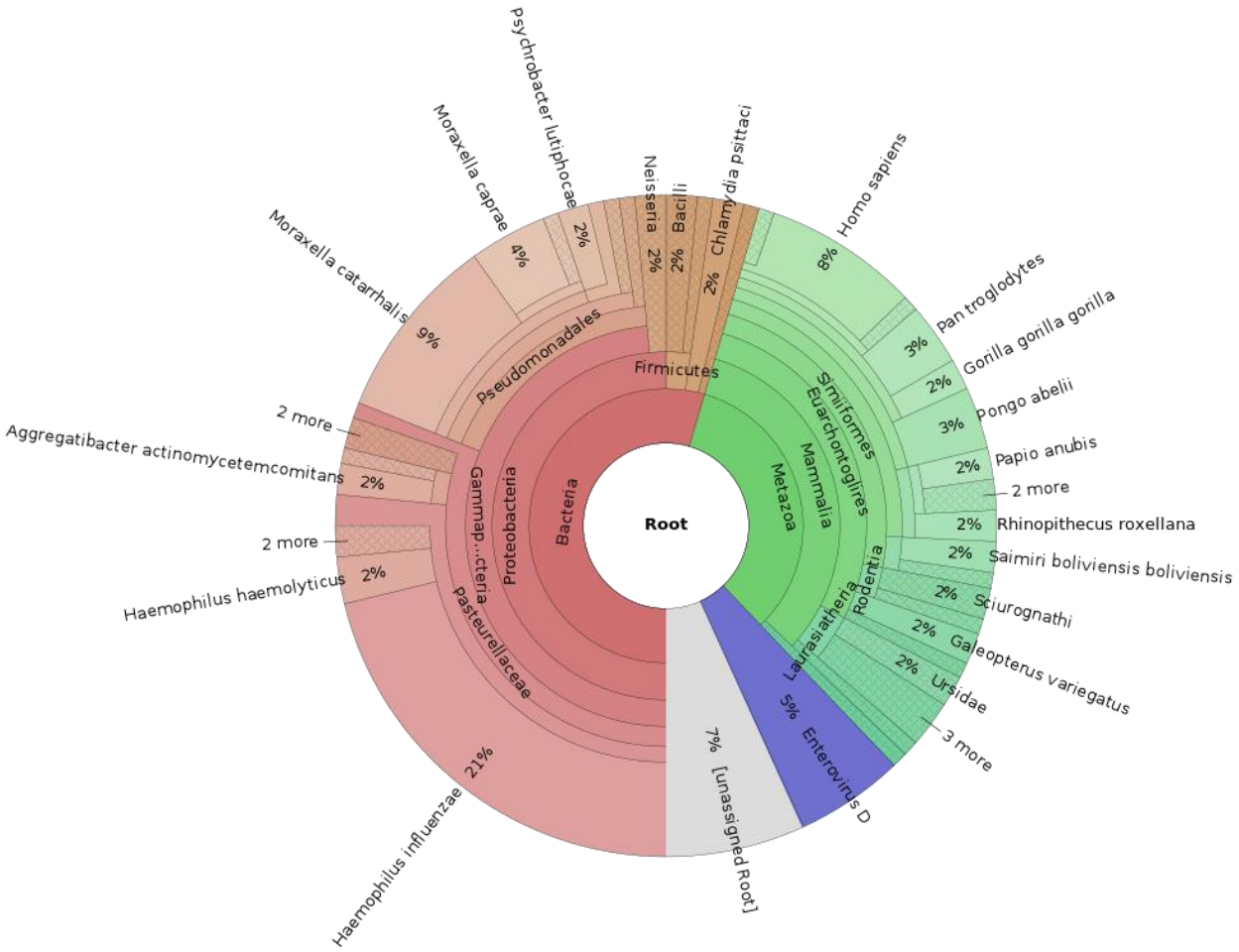
Specimen 429395



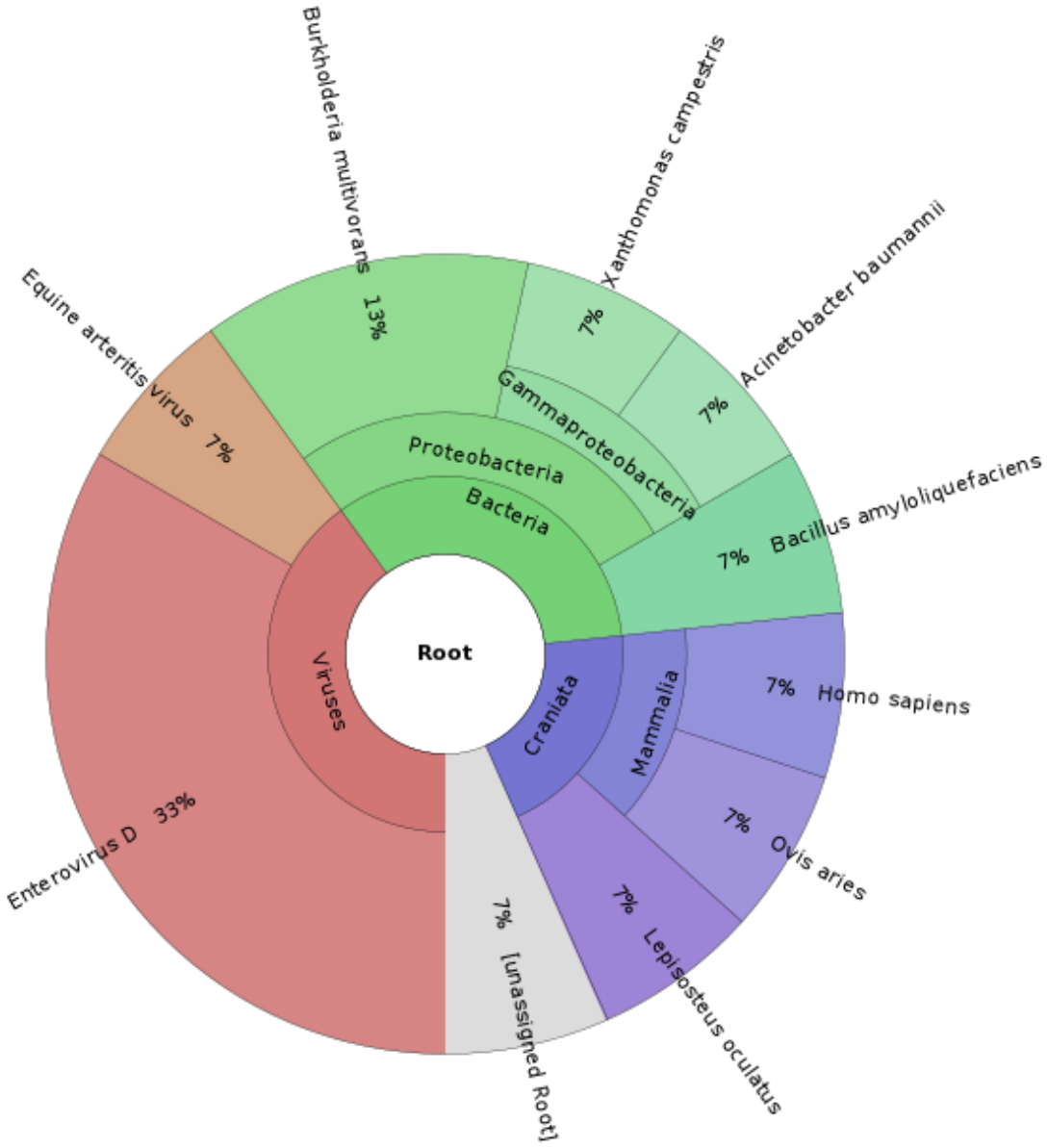
Specimen 427759



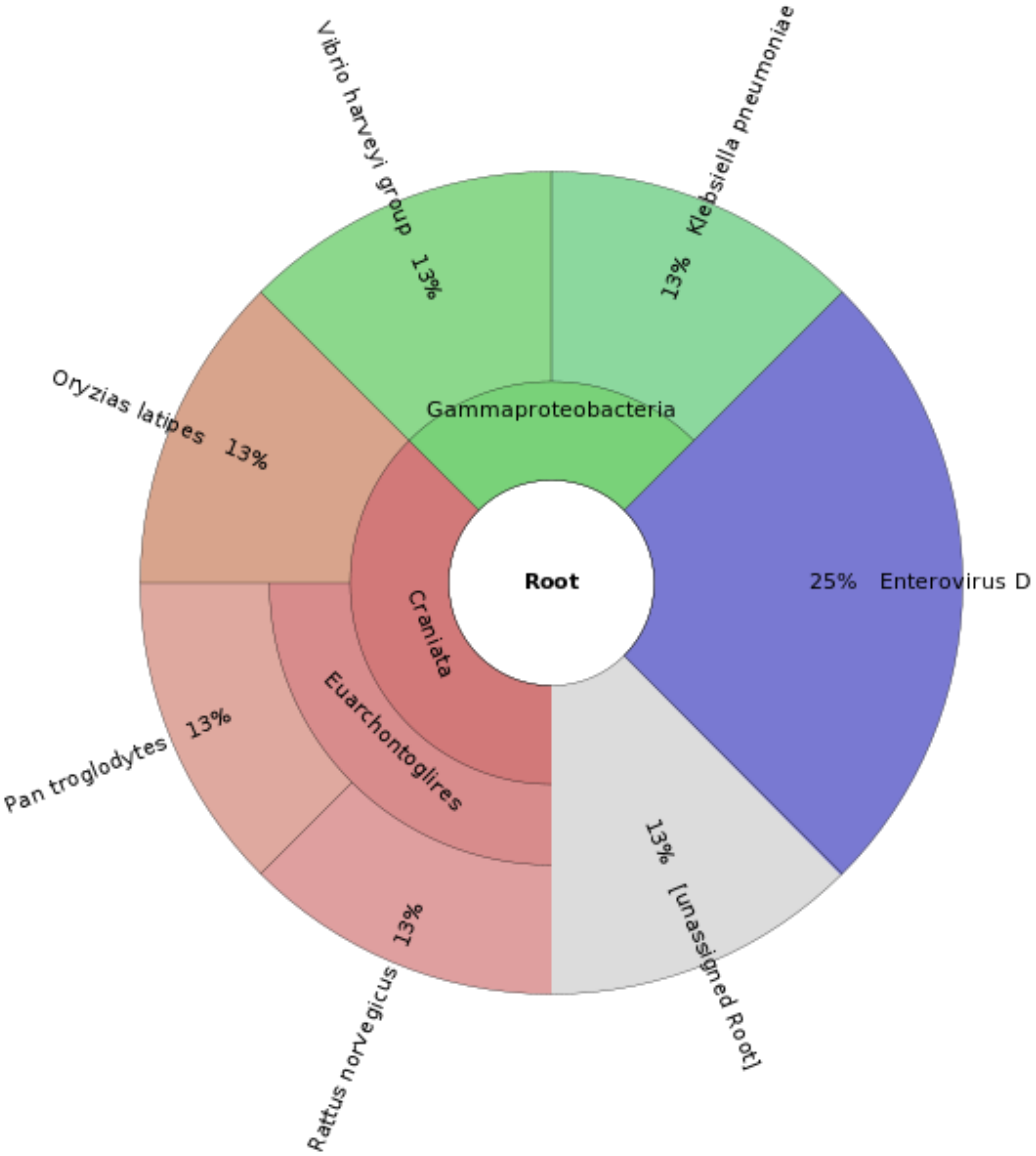
Specimen 428005



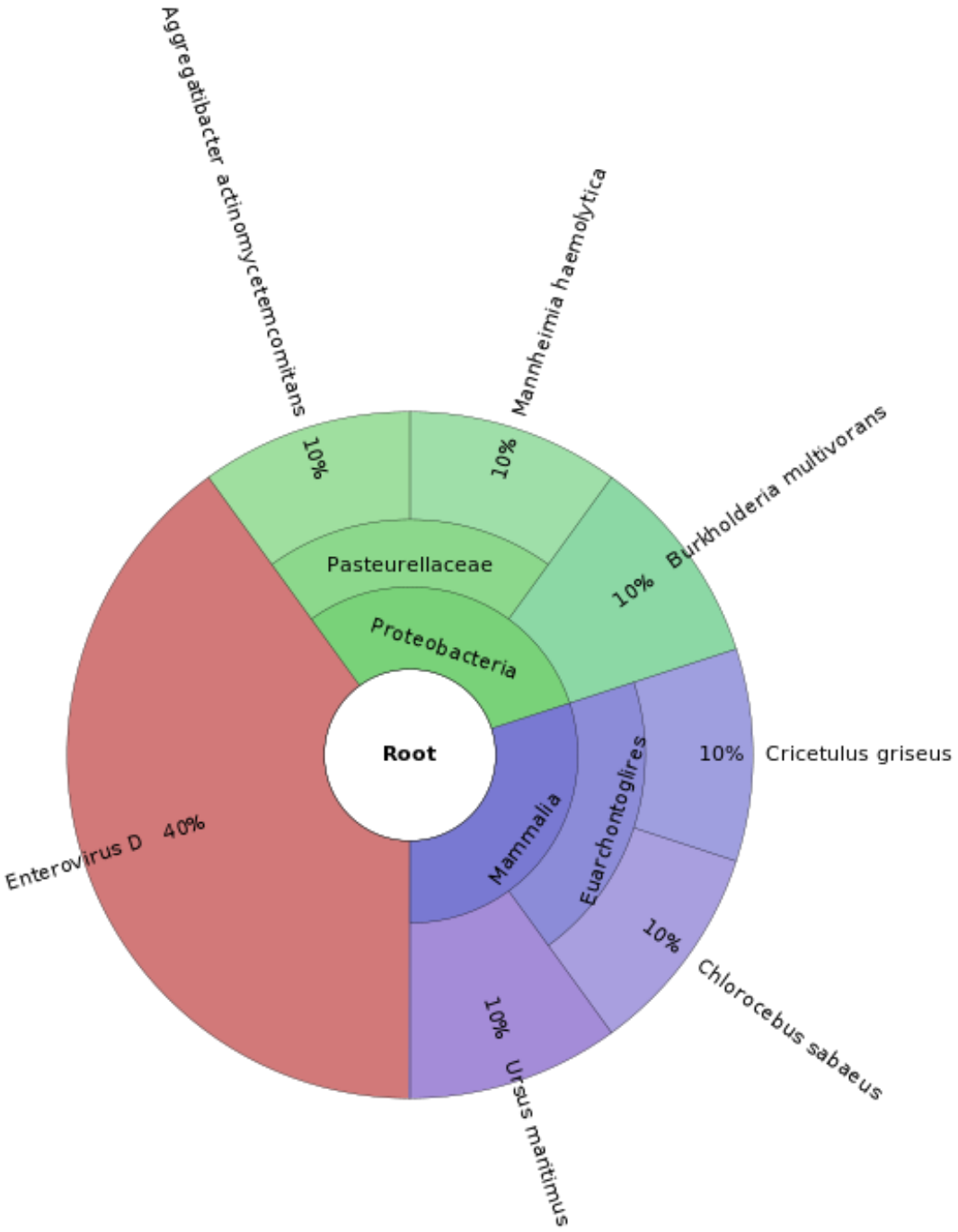
Specimen 428008



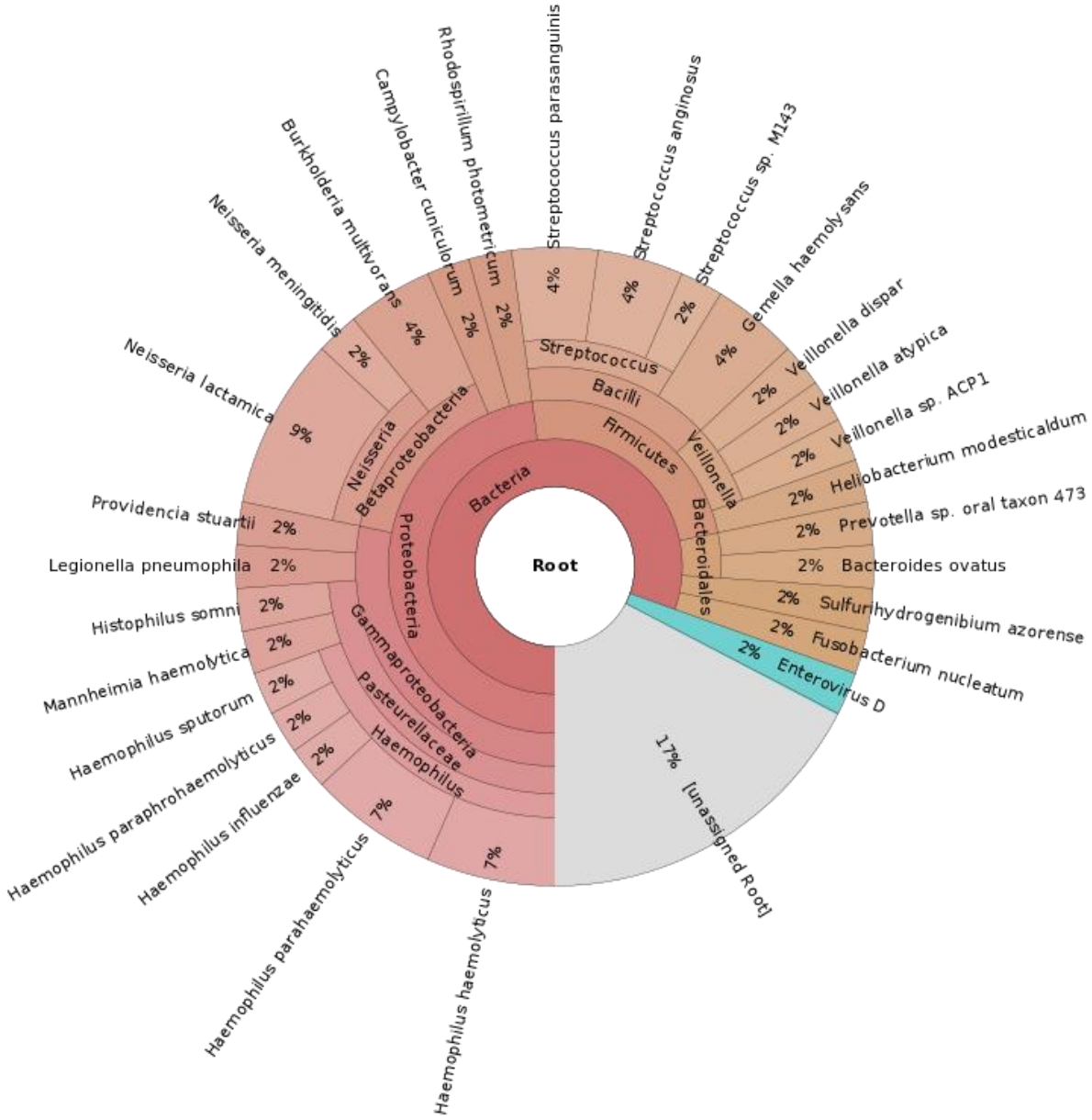
Specimen 428129



Specimen 428194

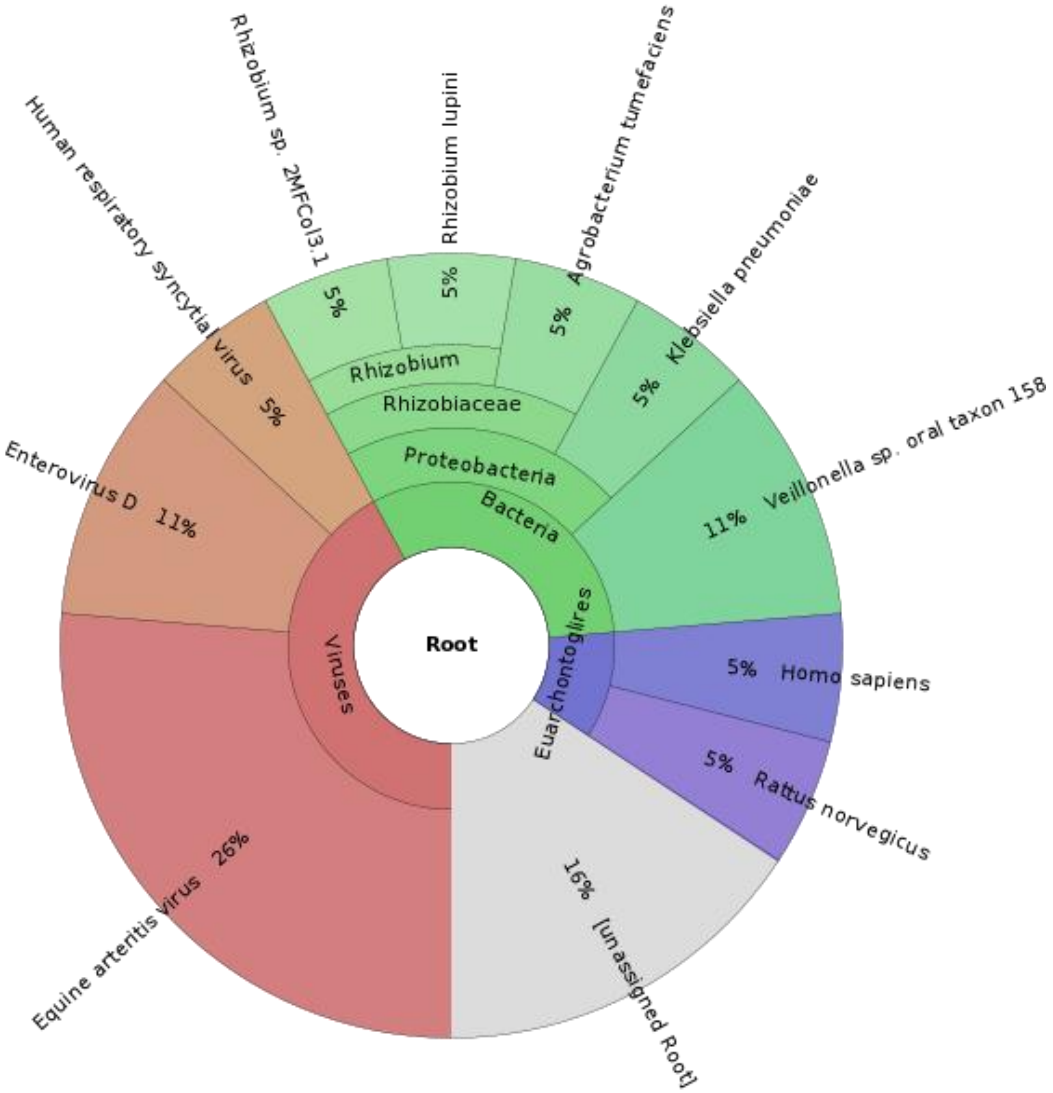


Specimen 428287

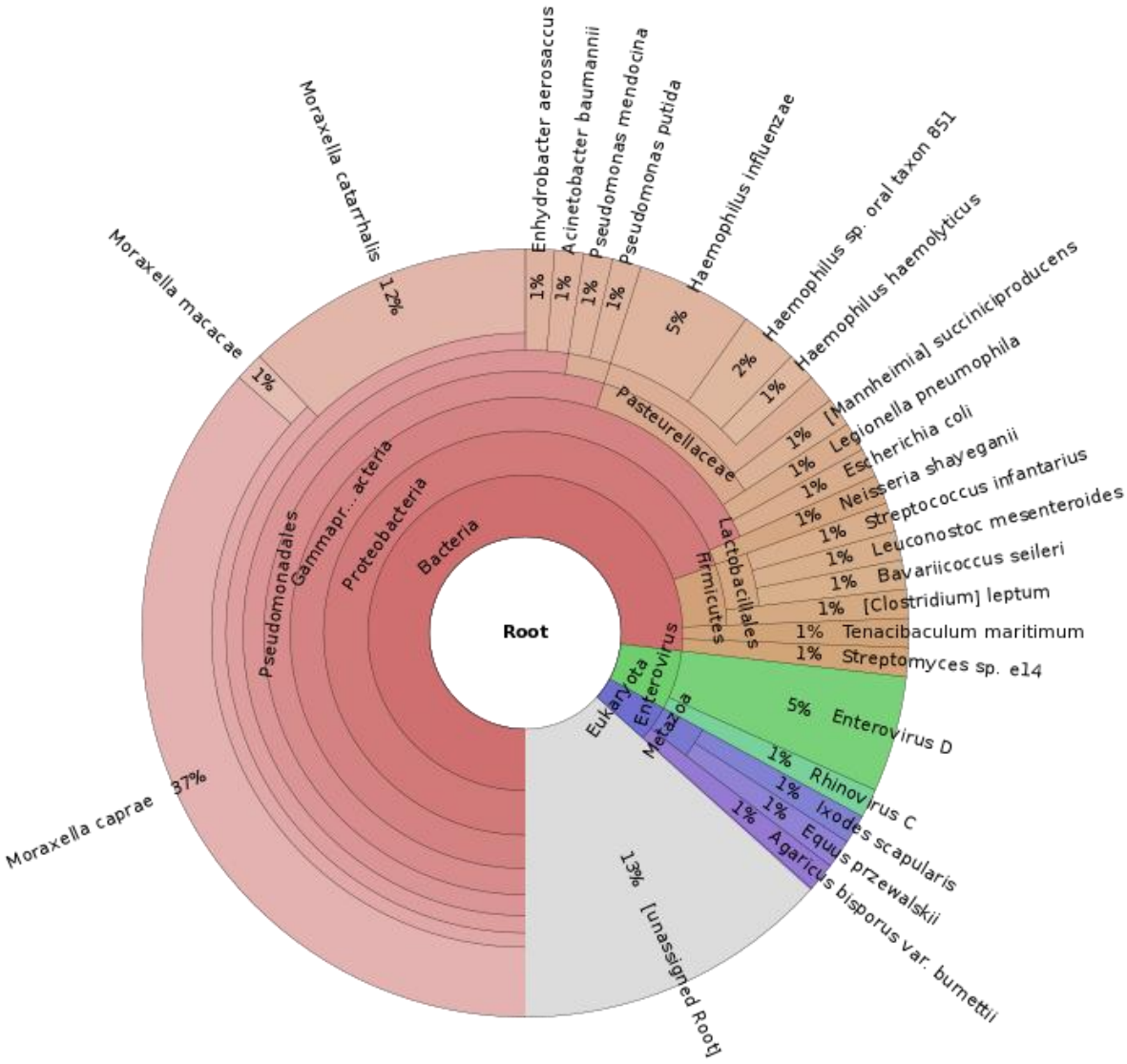


Specimen 428871

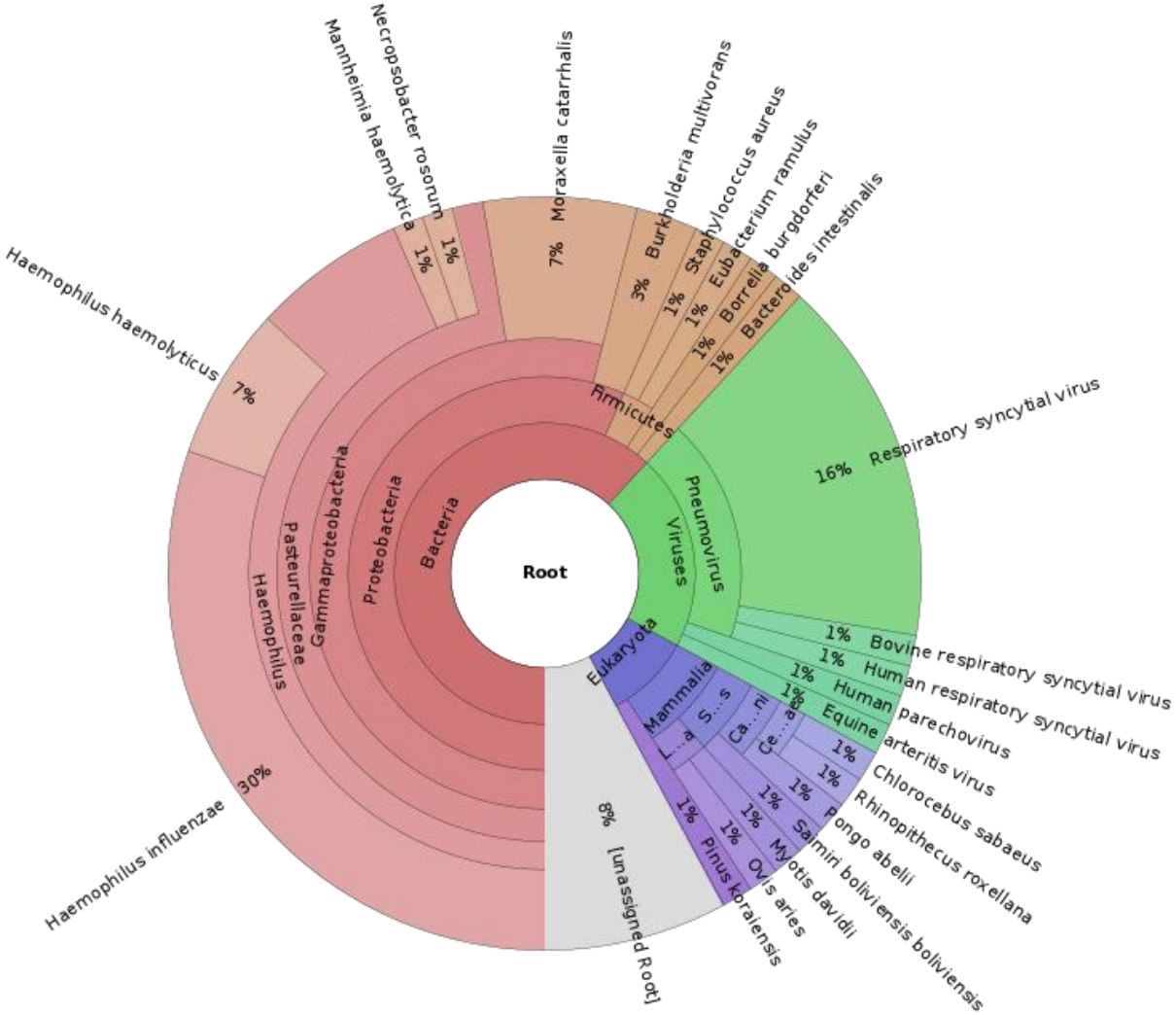
Specimen 429031



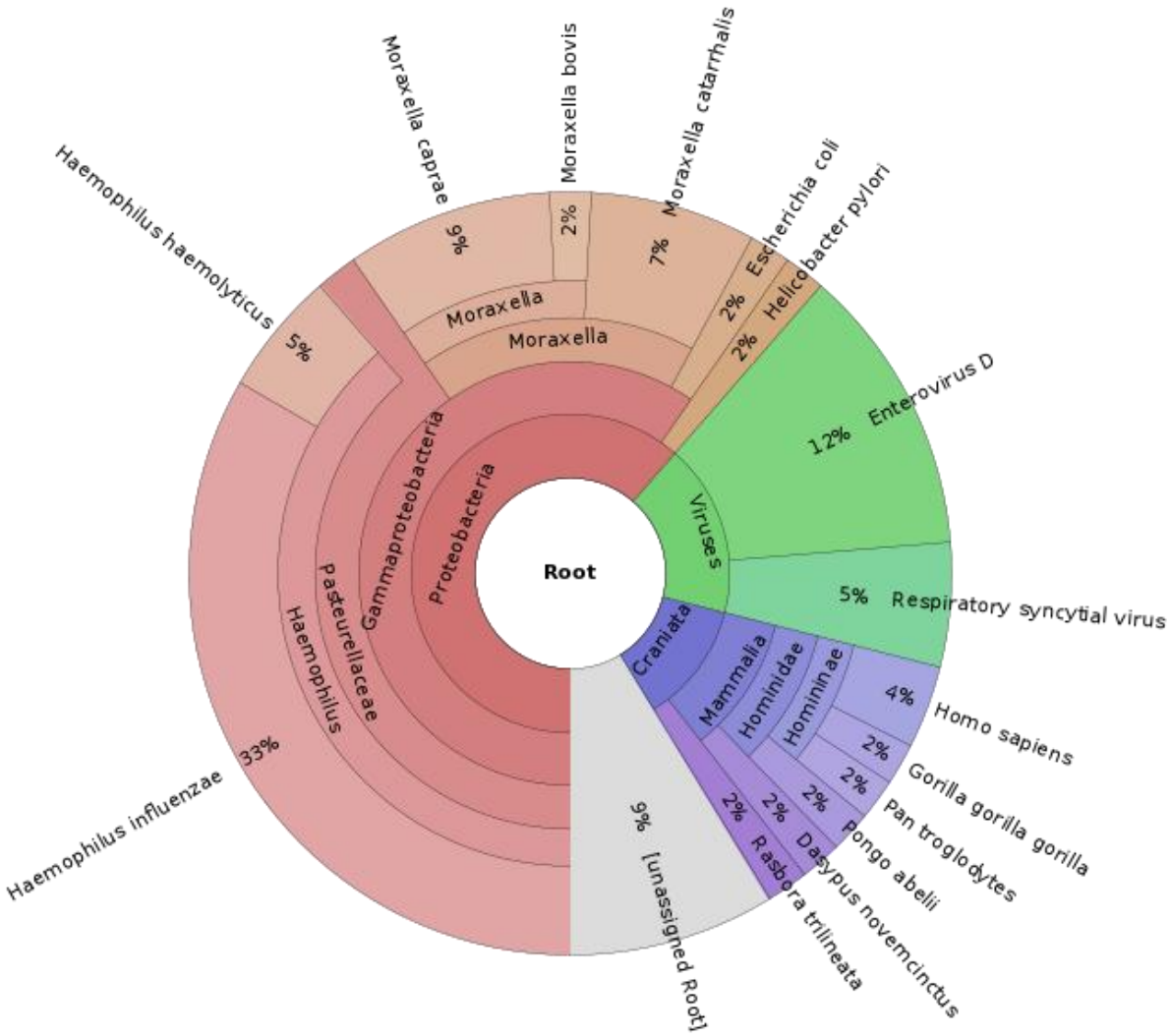
Specimen 429915



Specimen 429159

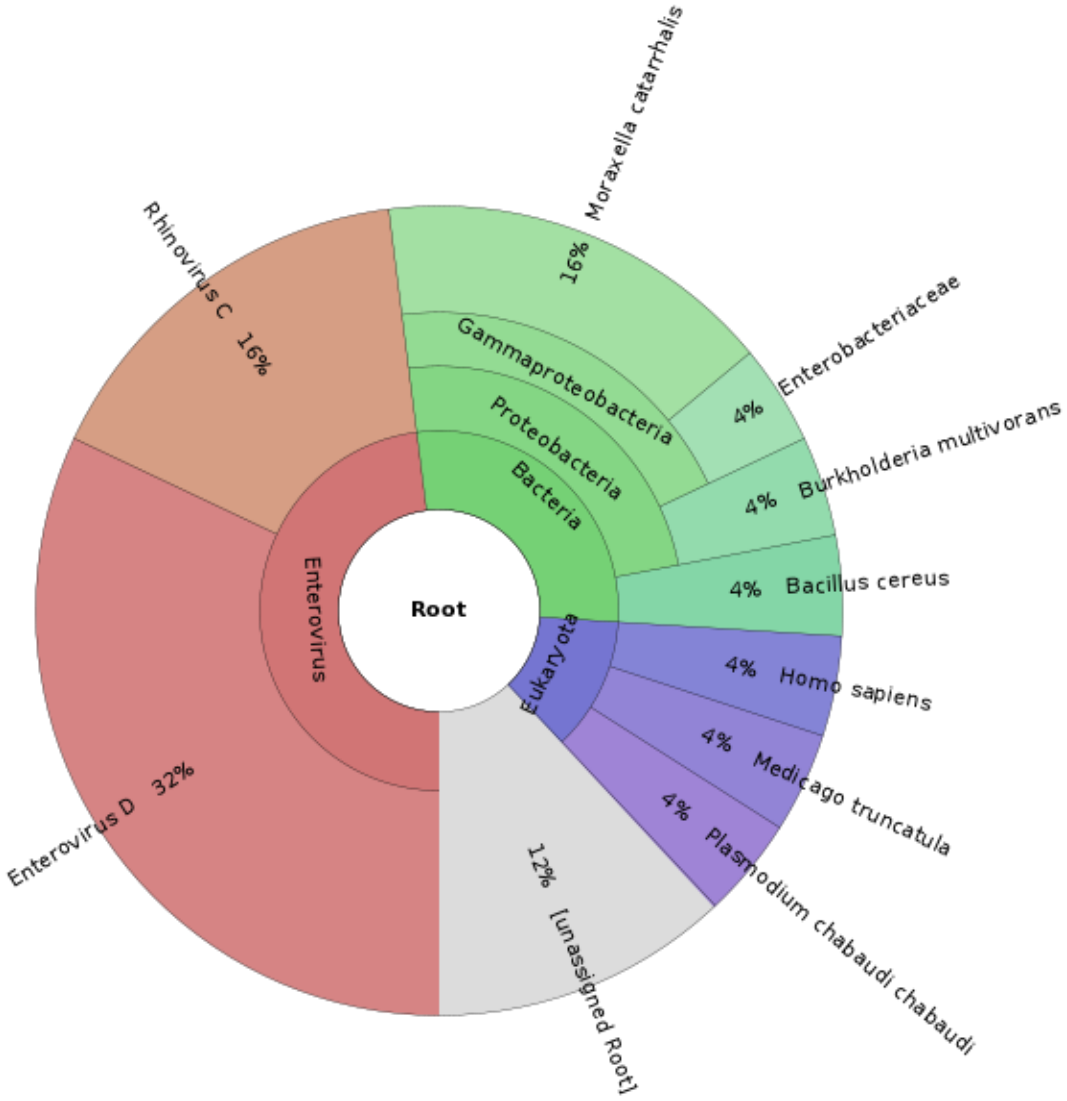


Specimen 429319



Specimen 429323

Specimen 429660



References

Abraham, E. P. and E. Chain (1940). "An enzyme from bacteria able to destroy penicillin." Nature **146**: 837-837.

Adalja, A. A., P. L. Sappington, et al. (2011). "Isolation of Aspergillus in three 2009 H1N1 influenza patients." Influenza Other Respir Viruses **5**(4): 225-229.

Aguilar, J. C., M. P. Perez-Brena, et al. (2000). "Detection and identification of human parainfluenza viruses 1, 2, 3, and 4 in clinical samples of pediatric patients by multiplex reverse transcription-PCR." J Clin Microbiol **38**(3): 1191-1195.

Ahmed, S. M., A. J. Hall, et al. (2014). "Global prevalence of norovirus in cases of gastroenteritis: a systematic review and meta-analysis." Lancet Infect Dis **14**(8): 725-730.

Alghamdi, I. G., Hussain, II, et al. (2014). "The pattern of Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive epidemiological analysis of data from the Saudi Ministry of Health." Int J Gen Med **7**: 417-423.

Allander, T., M. T. Tammi, et al. (2005). "Cloning of a human parvovirus by molecular screening of respiratory tract samples." Proc Natl Acad Sci U S A **102**(36): 12891-12896.

Ameyaw, E., S. B. Nguah, et al. (2014). "The outcome of a test-treat package versus routine outpatient care for Ghanaian children with fever: a pragmatic randomized control trial." Malar J **13**: 461.

Andrews, S. "FastQC A Quality Control tool for High Throughput Sequence Data." Babraham Bioinformatics Web site, from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Araujo, T. H., L. I. Souza-Brito, et al. (2012). "A public HTLV-1 molecular epidemiology database for sequence management and data mining." PLoS One **7**(9): e42123.

ARIA (2010). "World Lung Foundation - Acute Respiratory Infections Atlas."

Ashworth, M., J. Charlton, et al. (2005). "Variations in antibiotic prescribing and consultation rates for acute respiratory infection in UK general practices 1995-2000." Br J Gen Pract **55**(517): 603-608.

Badawi, A. and S. G. Ryoo (2016). "Prevalence of comorbidities in the Middle East respiratory syndrome coronavirus (MERS-CoV): a systematic review and meta-analysis." International Journal of Infectious Diseases **49**: 129-133.

Bao, S., R. Jiang, et al. (2011). "Evaluation of next-generation sequencing software in mapping and assembly." J Hum Genet **56**(6): 406-414.

Barker, J., I. B. Vipond, et al. (2004). "Effects of cleaning and disinfection in reducing the spread of Norovirus contamination via environmental surfaces." J Hosp Infect **58**(1): 42-49.

Bennett, S., A. MacLean, et al. (2013). "Increased norovirus activity in Scotland in 2012 is associated with the emergence of a new norovirus GII.4 variant." Euro Surveill **18**(2).

Bertino, J. S. (2002). "Cost burden of viral respiratory infections: issues for formulary decision makers." Am J Med **112 Suppl 6A**: 42S-49S.

Bossert, B. and K. K. Conzelmann (2002). "Respiratory syncytial virus (RSV) nonstructural (NS) proteins as host range determinants: a chimeric bovine RSV with NS genes from human RSV is attenuated in interferon-competent bovine cells." J Virol **76**(9): 4287-4293.

Bottcher, S., C. Prifert, et al. (2016). "Detection of enterovirus D68 in patients hospitalised in three tertiary university hospitals in Germany, 2013 to 2014." Euro Surveill **21**(19).

Bragg, L. M., G. Stone, et al. (2013). "Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data." PLoS Comput Biol **9**(4).

Bramley, T. J., D. Lerner, et al. (2002). "Productivity losses related to the common cold." J Occup Environ Med **44**(9): 822-829.

Bull, R. A., J. S. Eden, et al. (2012). "Contribution of intra- and interhost dynamics to norovirus evolution." J Virol **86**(6): 3219-3229.

Burd, E. M. (2010). "Validation of laboratory-developed molecular assays for infectious diseases." Clin Microbiol Rev **23**(3): 550-576.

Butler, C. C., K. Hood, et al. (2009). "Variation in antibiotic prescribing and its impact on recovery in patients with acute cough in primary care: prospective study in 13 countries." BMJ **338**: b2242.

Carlsson, B., A. M. Lindberg, et al. (2009). "Quasispecies dynamics and molecular evolution of human norovirus capsid P region during chronic infection." J Gen Virol **90**(Pt 2): 432-441.

Carr, J., J. Ives, et al. (2002). "Influenza virus carrying neuraminidase with reduced sensitivity to oseltamivir carboxylate has altered properties in vitro and is compromised for infectivity and replicative ability in vivo." Antiviral Res **54**(2): 79-88.

Clark, T. W., M. J. Medina, et al. (2014). "Adults hospitalised with acute respiratory illness rarely have detectable bacteria in the absence of COPD or pneumonia; viral infection predominates in a large prospective UK sample." J Infect **69**(5): 507-515.

Cox, D. W., J. Bizzintino, et al. (2013). "Human rhinovirus species C infection in young children with acute wheeze is associated with increased acute respiratory hospital admissions." Am J Respir Crit Care Med **188**(11): 1358-1364.

Dagan, R. and Y. Bar-David (1992). "Double-blind study comparing erythromycin and mupirocin for treatment of impetigo in children: implications of a high prevalence of erythromycin-resistant Staphylococcus aureus strains." Antimicrob Agents Chemother **36**(2): 287-290.

Dalling, J. (2004). "A review of environmental contamination during outbreaks of Norwalk-like virus." Journal of Infection Prevention **5**: 9-13.

Darren P. Martin, B. M., Michael Golden, Arjun Khoosal, Brejnev Muhire (2015). "RDP4: Detection and analysis of recombination patterns in virus genomes." Virus Evolution.

Deshpande, S. A. and V. Northern (2003). "The clinical and health economic burden of respiratory syncytial virus disease among children under 2 years of age in a defined geographical area." Arch Dis Child **88**(12): 1065-1069.

Doligalski, C. T., K. Benedict, et al. (2014). "Epidemiology of invasive mold infections in lung transplant recipients." Am J Transplant **14**(6): 1328-1333.

Dong, G. Y., C. Peng, et al. (2015). "Adamantane-Resistant Influenza A Viruses in the World (1902-2013): Frequency and Distribution of M2 Gene Mutations." PLoS One **10**(3).

Drancourt, M., A. Michel-Lepage, et al. (2016). "The Point-of-Care Laboratory in Clinical Microbiology." Clin Microbiol Rev **29**(3): 429-447.

Dykes, A. C., J. D. Cherry, et al. (1980). "A clinical, epidemiologic, serologic, and virologic study of influenza C virus infection." Arch Intern Med **140**(10): 1295-1298.

Eagle, H. (1955). "The specific amino acid requirements of a human carcinoma cell (Stain HeLa) in tissue culture." J Exp Med **102**(1): 37-48.
Edwards, M. C. and R. A. Gibbs (1994). "Multiplex PCR: advantages, development, and applications." PCR Methods Appl **3**(4): S65-75.
Eisfeld, A. J., G. Neumann, et al. (2015). "At the centre: influenza A virus ribonucleoproteins." Nat Rev Microbiol **13**(1): 28-41.
Escobar, C., V. Luchsinger, et al. (2009). "Genetic variability of human metapneumovirus isolated from Chilean children, 2003-2004." J Med Virol **81**(2): 340-344.
Falsey, A. R., P. A. Hennessey, et al. (2005). "Respiratory syncytial virus infection in elderly and high-risk adults." N Engl J Med **352**(17): 1749-1759.
Falsey, A. R. and E. E. Walsh (2003). "Novel coronavirus and severe acute respiratory syndrome." Lancet **361**(9366): 1312-1313.
Falzarano, D., E. de Wit, et al. (2013). "Treatment with interferon-alpha2b and ribavirin improves outcome in MERS-CoV-infected rhesus macaques." Nat Med **19**(10): 1313-1317.
Fehr, A. R. and S. Perlman (2015). "Coronaviruses: An Overview of Their Replication and Pathogenesis." Coronaviruses: Methods and Protocols **1282**: 1-23.
Fendrick, A. M., A. S. Monto, et al. (2003). "The economic burden of non-influenza-related viral respiratory tract infection in the United States." Arch Intern Med **163**(4): 487-494.
Fleming, D. M., G. E. Smith, et al. (2002). "Impact of infections on primary care--greater than expected." Commun Dis Public Health **5**(1): 7-12.
Froussard, P. (1992). "A random-PCR method (rPCR) to construct whole cDNA library from low amounts of RNA." Nucleic Acids Res **20**(11): 2900.
Fuentes, S., K. C. Tran, et al. (2007). "Function of the respiratory syncytial virus small hydrophobic protein." J Virol **81**(15): 8361-8366.
Gan, S. W., E. Tan, et al. (2012). "The small hydrophobic protein of the human respiratory syncytial virus forms pentameric ion channels." J Biol Chem **287**(29): 24671-24689.
Gaunt, E. R., A. Hardie, et al. (2010). "Epidemiology and clinical presentations of the four human coronaviruses 229E, HKU1, NL63, and OC43 detected over 3 years using a novel multiplex real-time PCR method." J Clin Microbiol **48**(8): 2940-2947.
Gaynor, A. M., M. D. Nissen, et al. (2007). "Identification of a novel polyomavirus from patients with acute respiratory tract infections." PLoS Pathog **3**(5): e64.
Gern, J. E., D. M. Galagan, et al. (1997). "Detection of rhinovirus RNA in lower airway cells during experimentally induced infection." Am J Respir Crit Care Med **155**(3): 1159-1161.
Ghildyal, R., C. Baulch-Brown, et al. (2003). "The matrix protein of Human respiratory syncytial virus localises to the nucleus of infected cells and inhibits transcription." Arch Virol **148**(7): 1419-1429.
Gierer, S., S. Bertram, et al. (2013). "The spike protein of the emerging betacoronavirus EMC uses a novel coronavirus receptor for entry, can be activated by TMPRSS2, and is targeted by neutralizing antibodies." J Virol **87**(10): 5502-5511.
Gonzales, R., J. G. Bartlett, et al. (2001). "Principles of appropriate antibiotic use for treatment of nonspecific upper respiratory tract infections in adults: background." Ann Emerg Med **37**(6): 698-702.

Gonzales, R., J. F. Steiner, et al. (1997). "Antibiotic prescribing for adults with colds, upper respiratory tract infections, and bronchitis by ambulatory care physicians." JAMA **278**(11): 901-904.
Greninger, A. L., S. N. Naccache, et al. (2015). "A novel outbreak enterovirus D68 strain associated with acute flaccid myelitis cases in the USA (2012-14): a retrospective cohort study." Lancet Infect Dis **15**(6): 671-682.
Gruber, C., T. Keil, et al. (2008). "History of respiratory infections in the first 12 yr among children from a birth cohort." Pediatr Allergy Immunol **19**(6): 505-512.
Gubbins, P. O., M. E. Klepser, et al. (2014). "Point-of-care testing for infectious diseases: opportunities, barriers, and considerations in community pharmacy." J Am Pharm Assoc (2003) **54**(2): 163-171.
Gunson, R. N. and W. F. Carman (2005). "Comparison of two real-time PCR methods for diagnosis of norovirus infection in outbreak and community settings." J Clin Microbiol **43**(4): 2030-2031.
Gunson, R. N. and W. F. Carman (2011). "During the summer 2009 outbreak of "swine flu" in Scotland what respiratory pathogens were diagnosed as H1N1/2009?" BMC Infect Dis **11**: 192.
Handforth, J., J. S. Friedland, et al. (2000). "Basic epidemiology and immunopathology of RSV in children." Paediatr Respir Rev **1**(3): 210-214.
Hause, B. M., E. A. Collin, et al. (2014). "Characterization of a novel influenza virus in cattle and Swine: proposal for a new genus in the Orthomyxoviridae family." MBio **5**(2): e00031-00014.
Hemila, H. and E. Chalker (2013). "Vitamin C for preventing and treating the common cold." Cochrane Database Syst Rev **1**: CD000980.
Higashi, T. and S. Fukuhara (2009). "Antibiotic prescriptions for upper respiratory tract infection in Japan." Intern Med **48**(16): 1369-1375.
Hoke, C. H., Jr. and C. E. Snyder, Jr. (2013). "History of the restoration of adenovirus type 4 and type 7 vaccine, live oral (Adenovirus Vaccine) in the context of the Department of Defense acquisition system." Vaccine **31**(12): 1623-1632.
Holm-Hansen, C. C., S. E. Midgley, et al. (2016). "Global emergence of enterovirus D68: a systematic review." Lancet Infect Dis.
HPS. (2015). "General outbreaks of infectious intestinal disease reported to HPS in 2014." from http://www.hps.scot.nhs.uk/ewr/article.aspx.
Huck, B., G. Scharf, et al. (2006). "Novel human metapneumovirus sublineage." Emerg Infect Dis **12**(1): 147-150.
Hughes, J., R. C. Allen, et al. (2012). "Transmission of equine influenza virus during an outbreak is characterized by frequent mixed infections and loose transmission bottlenecks." PLoS Pathog **8**(12): e1003081.
Illumina. (2016). "Sequencing Technology." from http://www.illumina.com/technology/next-generation-sequencing/sequencing-technology.html.
Ingram, R. E., F. Fenwick, et al. (2006). "Detection of human metapneumovirus in respiratory secretions by reverse-transcriptase polymerase chain reaction, indirect immunofluorescence, and virus isolation in human bronchial epithelial cells." J Med Virol **78**(9): 1223-1231.
Ison, M. G. (2006). "Adenovirus infections in transplant recipients." Clin Infect Dis **43**(3): 331-339.
Iturriza-Gomara, M. and B. Lopman (2014). "Norovirus in healthcare settings." Curr Opin Infect Dis **27**(5): 437-443.

Jacobs, S. E., D. M. Lamson, et al. (2013). "Human rhinoviruses." Clin Microbiol Rev **26**(1): 135-162.
Jain, S., D. J. Williams, et al. (2015). "Community-acquired pneumonia requiring hospitalization among U.S. children." N Engl J Med **372**(9): 835-845.
Jefferson, T., M. A. Jones, et al. (2014). "Neuraminidase inhibitors for preventing and treating influenza in healthy adults and children." Cochrane Database Syst Rev **4**: CD008965.
Jenson, A., L. Dize, et al. (2013). "Field evaluation of the Cepheid GeneXpert Chlamydia trachomatis assay for detection of infection in a trachoma endemic community in Tanzania." PLoS Negl Trop Dis **7**(7): e2265.
Johnson, P. R., M. K. Spriggs, et al. (1987). "The G glycoprotein of human respiratory syncytial viruses of subgroups A and B: extensive sequence divergence between antigenically related proteins." Proc Natl Acad Sci U S A **84**(16): 5625-5629.
Kahn, J. (2008). "Human bocavirus: clinical significance and implications." Curr Opin Pediatr **20**(1): 62-66.
Kaida, A., H. Kubo, et al. (2011). "Enterovirus 68 in children with acute respiratory tract infections, Osaka, Japan." Emerg Infect Dis **17**(8): 1494-1497.
Kapikian, A. Z., R. G. Wyatt, et al. (1972). "Visualization by immune electron microscopy of a 27-nm particle associated with acute infectious nonbacterial gastroenteritis." J Virol **10**(5): 1075-1081.
Khalid, M., F. Al Rabiah, et al. (2015). "Ribavirin and interferon-alpha2b as primary and preventive treatment for Middle East respiratory syndrome coronavirus: a preliminary report of two cases." Antivir Ther **20**(1): 87-91.
Ki, M. (2015). "2015 MERS outbreak in Korea: hospital-to-hospital transmission." Epidemiology and Health **37**: 4.
Kim, S. J., K. Kim, et al. (2015). "Outcomes of early administration of cidofovir in non-immunocompromised patients with severe adenovirus pneumonia." PLoS One **10**(4): e0122642.
Kircher, M., S. Sawyer, et al. (2012). "Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform." Nucleic Acids Res **40**(1): e3.
Ko, G., T. L. Cromeans, et al. (2003). "Detection of infectious adenovirus in cell culture by mRNA reverse transcription-PCR." Appl Environ Microbiol **69**(12): 7377-7384.
Kousha, M., R. Tadi, et al. (2011). "Pulmonary aspergillosis: a clinical review." Eur Respir Rev **20**(121): 156-174.
Kroneman, A., E. Vega, et al. (2013). "Proposal for a unified norovirus nomenclature and genotyping." Arch Virol **158**(10): 2059-2068.
Kroneman, A., H. Vennema, et al. (2011). "An automated genotyping tool for enteroviruses and noroviruses." J Clin Virol **51**(2): 121-125.
Kundu, S., J. Lockwood, et al. (2013). "Next-Generation Whole Genome Sequencing Identifies the Direction of Norovirus Transmission in Linked Patients." Clinical Infectious Diseases **57**(3): 407-414.
Kusel, M. M., N. H. de Klerk, et al. (2007). "Early-life respiratory viral infections, atopic sensitization, and risk of subsequent development of persistent asthma." J Allergy Clin Immunol **119**(5): 1105-1110.
Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nat Methods **9**(4): 357-359.
Lawrence, M. C., N. A. Borg, et al. (2004). "Structure of the haemagglutinin-neuraminidase from human parainfluenza virus type III." J Mol Biol **335**(5): 1343-1357.

Lazzarotto, T., P. Dal Monte, et al. (1992). "Lack of correlation between virus detection and serologic tests for diagnosis of active cytomegalovirus infection in patients with AIDS." J Clin Microbiol **30**(4): 1027-1029.
Lee, J. Y., H. J. Park, et al. (2015). "Cellular Profiles of Bronchoalveolar Lavage Fluid and Their Prognostic Significance for Non-HIV-Infected Patients with Pneumocystis jirovecii Pneumonia." J Clin Microbiol **53**(4): 1310-1316.
Lenaerts, L., E. De Clercq, et al. (2008). "Clinical features and treatment of adenovirus infections." Rev Med Virol **18**(6): 357-374.
Li, H. and R. Durbin (2010). "Fast and accurate long-read alignment with Burrows-Wheeler transform." Bioinformatics **26**(5): 589-595.
Li, H., M. He, et al. (2015). "The genomic and seroprevalence of human bocavirus in healthy Chinese plasma donors and plasma derivatives." Transfusion **55**(1): 154-163.
Liao, R. S., D. M. Appelgate, et al. (2012). "An outbreak of severe respiratory tract infection due to human metapneumovirus in a long-term care facility for the elderly in Oregon." J Clin Virol **53**(2): 171-173.
Lim, F. J., N. de Klerk, et al. (2016). "Systematic review and meta-analysis of respiratory viral coinfections in children." Respirology **21**(4): 648-655.
Linares, J., C. Ardanuy, et al. (2010). "Changes in antimicrobial resistance, serotypes and genotypes in Streptococcus pneumoniae over a 30-year period." Clin Microbiol Infect **16**(5): 402-410.
Lindbaek, M. (2006). "Prescribing antibiotics to patients with acute cough and otitis media." Br J Gen Pract **56**(524): 164-166.
Lindesmith, L. C., M. T. Ferris, et al. (2015). "Broad blockade antibody responses in human volunteers after immunization with a multivalent norovirus VLP candidate vaccine: immunological analyses from a phase I clinical trial." PLoS Med **12**(3): e1001807.
Lindgreen, S. (2012). "AdapterRemoval: easy cleaning of next-generation sequencing reads." BMC Res Notes **5**: 337.
Lopman, B. A., M. H. Reacher, et al. (2004). "Epidemiology and cost of nosocomial gastroenteritis, Avon, England, 2002-2003." Emerg Infect Dis **10**(10): 1827-1834.
Ly, N., R. Tokarz, et al. (2014). "Multiplex PCR analysis of clusters of unexplained viral respiratory tract infection in Cambodia." Virol J **11**(1): 224.
Maabar, M. (2016). DisCVR: a fast viral detection tool. Scottish Diagnostic Virology Group. Stirling.
Makela, M. J., T. Puhakka, et al. (1998). "Viruses and bacteria in the etiology of the common cold." J Clin Microbiol **36**(2): 539-542.
Mann, A. G., P. Mangtani, et al. (2013). "The impact of targeting all elderly persons in England and Wales for yearly influenza vaccination: excess mortality due to pneumonia or influenza and time trend study." BMJ Open **3**(8).
Maxam, A. M. and W. Gilbert (1977). "A new method for sequencing DNA." Proc Natl Acad Sci U S A **74**(2): 560-564.
Messacar, K., T. L. Schreiner, et al. (2015). "A cluster of acute flaccid paralysis and cranial nerve dysfunction temporally associated with an outbreak of enterovirus D68 in children in Colorado, USA." Lancet.
Miller, E. K., K. M. Edwards, et al. (2009). "A novel group of rhinoviruses is associated with asthma hospitalizations." J Allergy Clin Immunol **123**(1): 98-104 e101.
Mukherjee, S., M. Huntemann, et al. (2015). "Large-scale contamination of microbial isolate genomes by Illumina PhiX control." Stand Genomic Sci **10**: 18.

Munoz, P., I. Ceron, et al. (2014). "Invasive aspergillosis among heart transplant recipients: a 24-year perspective." J Heart Lung Transplant **33**(3): 278-288.
Murdoch, D. R., S. Slow, et al. (2012). "Effect of vitamin D3 supplementation on upper respiratory tract infections in healthy adults: the VIDARIS randomized controlled trial." JAMA **308**(13): 1333-1339.
NICE, G. C. (2008). Respiratory tract infections – antibiotic prescribing: Prescribing of antibiotics for self-limiting respiratory tract infections in adults and children in primary care. Respiratory Tract Infections - Antibiotic Prescribing: Prescribing of Antibiotics for Self-Limiting Respiratory Tract Infections in Adults and Children in Primary Care. London.
Nickbakhsh, S., F. Thorburn, et al. (2016). "Extensive multiplex PCR diagnostics reveal new insights into the epidemiology of viral respiratory infections." Epidemiol Infect: 1-13.
Nishio, M., J. Ohtsuka, et al. (2008). "Human parainfluenza virus type 2 V protein inhibits genome replication by binding to the L protein: Possible role in promoting viral fitness." J Virol **82**(13): 6130-6138.
Nix, W. A., M. S. Oberste, et al. (2006). "Sensitive, seminested PCR amplification of VP1 sequences for direct identification of all enterovirus serotypes from original clinical specimens." J Clin Microbiol **44**(8): 2698-2704.
Njenga, M. K., H. M. Lwamba, et al. (2003). "Metapneumoviruses in birds and humans." Virus Res **91**(2): 163-169.
Nordgren, J., L. W. Nitiema, et al. (2013). "Host Genetic Factors Affect Susceptibility to Norovirus Infections in Burkina Faso." PLoS One **8**(7).
Norja, P., I. Ubillos, et al. (2007). "No evidence for an association between infections with WU and KI polyomaviruses and respiratory disease." J Clin Virol **40**(4): 307-311.
Odell, I. D. and D. Cook (2013). "Immunofluorescence techniques." J Invest Dermatol **133**(1): e4.
Palmenberg, A. C., J. A. Rathe, et al. (2010). "Analysis of the complete genome sequences of human rhinovirus." J Allergy Clin Immunol **125**(6): 1190-1199; quiz 1200-1191.
Papadopoulos, N. G., G. Sanderson, et al. (1999). "Rhinoviruses replicate effectively at lower airway temperatures." J Med Virol **58**(1): 100-104.
Pavia, A. T. (2011). "Viral infections of the lower respiratory tract: old viruses, new viruses, and the role of diagnosis." Clin Infect Dis **52 Suppl 4**: S284-289.
Payne, D. C., J. Vinje, et al. (2013). "Norovirus and medically attended gastroenteritis in U.S. children." N Engl J Med **368**(12): 1121-1130.
Poelman, R., E. H. Scholvinck, et al. (2015). "The emergence of enterovirus D68 in a Dutch University Medical Center and the necessity for routinely screening for respiratory viruses." J Clin Virol **62**: 1-5.
Poelman, R., I. Schuffenecker, et al. (2015). "European surveillance for enterovirus D68 during the emerging North-American outbreak in 2014." J Clin Virol **71**: 1-9.
Prachayangprecha, S., C. M. Schapendonk, et al. (2014). "Exploring the potential of next-generation sequencing in detection of respiratory viruses." J Clin Microbiol **52**(10): 3722-3730.
Quinlan, A. R. and I. M. Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features." Bioinformatics **26**(6): 841-842.
Rabenau, H. F., M. Sturmer, et al. (2003). "Laboratory diagnosis of norovirus: which method is the best?" Intervirology **46**(4): 232-238.

Riley, M. A., S. M. Robinson, et al. (2012). "Resistance is futile: the bacteriocin model for addressing the antibiotic resistance challenge." Biochem Soc Trans **40**(6): 1438-1442.

Robert F Breiman, L. C., M Kariuki Njenga, John Williamson, Joshua A Mott, Mark A Katz, Dean D Erdman, Eileen Schneider, M Steven Oberste, John C Neatherlin, Henry Njuguna, Daniel M Ondari, Kennedy Odero, George O Okoth, Beatrice Olack, Newton Wamola, Joel M Montgomery, Barry S Fields, and Daniel R Feikin (2015). "Severe acute respiratory infection in children in a densely populated urban slum in Kenya, 2007–2011." BMC Infect Dis.

Roberts, R. J., M. O. Carneiro, et al. (2013). "The advantages of SMRT sequencing." Genome Biol **14**(7).

Robilotti, E., S. Deresinski, et al. (2015). "Norovirus." Clin Microbiol Rev **28**(1): 134-164.

Ruohola, A., M. Waris, et al. (2009). "Viral etiology of common cold in children, Finland." Emerg Infect Dis **15**(2): 344-346.

Salonen, E. M., A. Vaheri, et al. (1980). "Rheumatoid factor in acute viral infections: interference with determination of IgM, IgG, and IgA antibodies in an enzyme immunoassay." J Infect Dis **142**(2): 250-255.

Sanger, F. and A. R. Coulson (1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase." J Mol Biol **94**(3): 441-448.

Schieble, J. H., V. L. Fox, et al. (1967). "A probable new human picornavirus associated with respiratory diseases." Am J Epidemiol **85**(2): 297-310.

Schildgen, O., T. Glatzel, et al. (2005). "Human metapneumovirus RNA in encephalitis patient." Emerg Infect Dis **11**(3): 467-470.

Schmieder, R. and R. Edwards (2011). "Quality control and preprocessing of metagenomic datasets." Bioinformatics **27**(6): 863-864.

Seemungal, T., R. Harper-Owen, et al. (2001). "Respiratory viruses, symptoms, and inflammatory markers in acute exacerbations and stable chronic obstructive pulmonary disease." Am J Respir Crit Care Med **164**(9): 1618-1623.

Seemungal, T. A., G. C. Donaldson, et al. (2000). "Time course and recovery of exacerbations in patients with chronic obstructive pulmonary disease." Am J Respir Crit Care Med **161**(5): 1608-1613.

Siegrist, C. A. (2000). "Vaccination in the neonatal period and early infancy." Int Rev Immunol **19**(2-3): 195-219.

Simancas-Racines, D., C. V. Guerra, et al. (2013). "Vaccines for the common cold." Cochrane Database Syst Rev **6**: CD002190.

Simmons, K., M. Gambhir, et al. (2013). "Duration of immunity to norovirus gastroenteritis." Emerg Infect Dis **19**(8): 1260-1267.

Singh, M. (2013). "Heated, humidified air for the common cold." Cochrane Database Syst Rev **6**: CD001728.

Singh, M. and R. R. Das (2013). "Zinc for the common cold." Cochrane Database Syst Rev **6**: CD001364.

Sloots, T. P., P. McErlean, et al. (2006). "Evidence of human coronavirus HKU1 and human bocavirus in Australian children." J Clin Virol **35**(1): 99-102.

Smith, S. M., K. Schroeder, et al. (2014). "Over-the-counter (OTC) medications for acute cough in children and adults in community settings." Cochrane Database Syst Rev **11**: CD001831.

Stadler, K., V. Masignani, et al. (2003). "SARS--beginning to understand a new virus." Nat Rev Microbiol **1**(3): 209-218.

Stanton, N., N. A. Francis, et al. (2010). "Reducing uncertainty in managing respiratory tract infections in primary care." Br J Gen Pract **60**(581): e466-475.
Storch, G. A. (2000). "Diagnostic virology." Clin Infect Dis **31**(3): 739-751.
Svraka, S., K. Rosario, et al. (2010). "Metagenomic sequencing for virus identification in a public-health setting." J Gen Virol **91**(Pt 11): 2846-2856.
Tam, C. C., L. C. Rodrigues, et al. (2012). "Longitudinal study of infectious intestinal disease in the UK (IID2 study): incidence in the community and presenting to general practice." Gut **61**(1): 69-77.
Tamura, K., G. Stecher, et al. (2013). "MEGA6: Molecular Evolutionary Genetics Analysis version 6.0." Mol Biol Evol **30**(12): 2725-2729.
Tapparel, C., T. Junier, et al. (2009). "New respiratory enterovirus and recombinant rhinoviruses among circulating picornaviruses." Emerg Infect Dis **15**(5): 719-726.
Tayyari, F., D. Marchant, et al. (2011). "Identification of nucleolin as a cellular receptor for human respiratory syncytial virus." Nat Med **17**(9): 1132-1135.
Treangen, T. J., S. Koren, et al. (2013). "MetAMOS: a modular and open source metagenomic assembly and analysis pipeline." Genome Biol **14**(1): R2.
van Buul, L. W., R. B. Veenhuizen, et al. (2014). "Antibiotic Prescribing In Dutch Nursing Homes: How Appropriate Is It?" J Am Med Dir Assoc.
van den Hoogen, B. G., J. C. de Jong, et al. (2001). "A newly discovered human pneumovirus isolated from young children with respiratory tract disease." Nat Med **7**(6): 719-724.
van der Vries, E., K. J. Stittelaar, et al. (2013). "Prolonged influenza virus shedding and emergence of antiviral resistance in immunocompromised patients and ferrets." PLoS Pathog **9**(5): e1003343.
Vazquez-Perez, J. A., J. E. Ramirez-Gonzalez, et al. (2016). "EV-D68 infection in children with asthma exacerbation and pneumonia in Mexico City during 2014 autumn." Influenza Other Respir Viruses **10**(3): 154-160.
Vega, E., E. Donaldson, et al. (2014). "RNA populations in immunocompromised patients as reservoirs for novel norovirus variants." J Virol **88**(24): 14184-14196.
Watson, J. D. and F. H. Crick (1953). "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid." Nature **171**(4356): 737-738.
WHO (1980). "A Revision of the System of Nomenclature for Influenza-Viruses - a Who Memorandum." Bulletin of the World Health Organization **58**(4): 585-591.
WHO (2010). INTEGRATED DISEASE SURVEILLANCE AND RESPONSE IN THE AFRICAN REGION.
WHO (2015). "Middle East respiratory syndrome coronavirus (MERS-CoV) – Saudi Arabia." Global Alert and Response (GAR)
WHO (2016). Recommended composition of influenza virus vaccines for use in the 2016-2017 northern hemisphere influenza season. WHO Recommendation.
Wikswo, M. E., A. Kambhampati, et al. (2015). "Outbreaks of Acute Gastroenteritis Transmitted by Person-to-Person Contact, Environmental Contamination, and Unknown Modes of Transmission--United States, 2009-2013." MMWR Surveill Summ **64**(12): 1-16.
Williams, B. G., E. Gouws, et al. (2002). "Estimates of world-wide distribution of child deaths from acute respiratory infections." Lancet Infect Dis **2**(1): 25-32.

Zhang, G., Y. Hu, et al. (2012). "High incidence of multiple viral infections identified in upper respiratory tract infected children under three years of age in Shanghai, China." PLoS One **7**(9): e44568.

Zhao, S., C. Wan, et al. (2014). "Re-emergent human adenovirus genome type 7d caused an acute respiratory disease outbreak in Southern China after a twenty-one year absence." Sci Rep **4**: 7365.