



University  
of Glasgow

Calus, Szymon Tomasz (2018) *Evaluation of nanopore-based sequencing technology for gene marker based analysis of complex microbial communities. Method development for accurate 16S rRNA gene amplicon sequencing*. PhD thesis.

<https://theses.gla.ac.uk/41086/>

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses

<https://theses.gla.ac.uk/>  
[research-enlighten@glasgow.ac.uk](mailto:research-enlighten@glasgow.ac.uk)

# Evaluation of nanopore-based sequencing technology for gene marker based analysis of complex microbial communities

Method Development For Accurate 16S rRNA Gene Amplicon Sequencing

by

Szymon Tomasz Calus

---

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

School of Engineering  
College of Science and Engineering  
University of Glasgow

December 18th, 2018

© Szymon T Calus (2018)

## Abstract

Nucleic acid sequencing can provide a detailed overview of microbial communities in comparison with standard plate-culture methods. Expansion of high-throughput sequencing (HTS) technologies and reduction in analysis costs has allowed for detailed exploration of various habitats with use of amplicon, metagenomics, and metatranscriptomics approaches. However, due to a capital cost of HTS platforms and requirements for batch analysis, genomics-based studies are still not being used as a standard method for the comprehensive examination of environmental or clinical samples for microbial characterization.

This research project investigated the potential of a novel nanopore-based sequencing platform from Oxford Nanopore Technologies (ONT) for rapid and accurate analysis of various environmentally complex samples. ONT is an emerging company that developed the first-ever portable nanopore-based sequencing platform called MinION™. Portability and miniaturised size of the device gives an immense opportunity for de-centralised, in-field, and real-time analysis of environmental and clinical samples. Nonetheless, benchmarking of this new technology against the current gold-standard platform (i.e., Illumina sequencers) is necessary to evaluate nanopore data and understand its benefits and limitations.

The focus of this study is on the evaluation of nanopore sequencing data: read quality, sequencing errors, alignment quality but also bacterial community structure. For this reason, mock bacterial community samples were generated, sequenced and analysed with use of multiple bioinformatics approaches. Furthermore, this study developed sophisticated library preparation and data analyses methods to enable high-accuracy analysis of amplicon libraries from complex microbial communities for sequencing on the nanopore platform. Besides, the best performing library preparation and data analyses methods were used for analysis of environmental samples and compared to high-quality Illumina metagenomics data. This work opens a new possibility for accurate, in-field amplicon analysis of complex samples with the use of MinION™ and for the development of autonomous biosensing technology for culture-free detection of pathogenic and non-pathogenic microorganisms in water, soil, food, drinks or blood.

## Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisors: Prof. Ameet J. Pinto and Dr. Umer Z. Ijaz, for giving me the opportunity to undertake this PhD and their continuous support of my research study, their patience, motivation, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisors and mentors for my PhD study. Besides my supervisors, I would like to thank the rest of my thesis committee: Prof. Barbara Mable and Dr. David Werner for their insightful comments and encouragement, but also for the difficult questions, which encourage me to widen my research from various perspectives.

Moreover, I would like to thank The Engineering and Physical Sciences Research Council (EPSRC) for supporting my PhD project and giving me the opportunity to present this research at the international conferences.

I thank my colleagues at the University of Glasgow for the stimulating discussions, the time we were working together before deadlines. Many of them have become good friends over the years and made this PhD a great journey: Alex, Fabien, Jill, Karol, Rosana, Mathieu, Sissy, Marta, Emanuele and for all the fun we have had in the last four years.

Last but not the least, I would like to thank my family: my parents (Beata and Darek), grandmother (Krystyna) and to my brother (Hubert) for supporting me spiritually throughout writing this thesis.



# Contents

Abstract .....	II
Acknowledgements .....	III
Contents .....	1
List of figures .....	3
Declaration .....	5
Nomenclature .....	6
Glossary .....	7
 <b>1 Project Description &amp; Research Objectives</b>	
1.1 Research objectives .....	11
1.2 Overview of Chapters .....	15
1.3 Overview of the public disclosures .....	18
 <b>2 Introduction to DNA Sequencing</b>	
2.1 History of DNA discovery .....	20
2.2 Description the DNA Sequencing Methods .....	23
2.3 History of Nanopore Sequencing .....	44
2.4 Description of various nanopore concepts .....	45
2.5 Future of DNA Analysis and Biosensing .....	65
 <b>3 Evaluation of Nanopore Technology for Amplicon Sequencing</b>	
3.1 Abstract .....	72
3.2 Introduction .....	74
3.3 Methods and Analysis .....	77
3.4 Results .....	86
3.5 Conclusions and Future Work .....	99

<b>4</b>	<b>Protocol Development for Microbial Genomics</b>	
4.1	Abstract . . . . .	101
4.2	Introduction . . . . .	102
4.3	Method development . . . . .	106
4.4	Conclusion and future Work . . . . .	111
<b>5</b>	<b>Benchmarking of novel bioinformatics algorithms</b>	
5.1	Abstract . . . . .	113
5.2	Introduction . . . . .	115
5.3	Algorithm 1: chopSEQ . . . . .	119
5.4	Algorithm 2: nanoClust . . . . .	124
5.5	Conclusions and Future Work . . . . .	128
<b>6</b>	<b>Comparison of Oxford Nanopore vs. Illumina</b>	
6.1	Abstract . . . . .	131
6.2	Introduction . . . . .	133
6.3	Drinking water samples . . . . .	135
6.4	Conclusions and Future Work . . . . .	144
<b>7</b>	<b>Conclusions and Future Work</b>	
7.1	Major Discoveries and Limitations . . . . .	146
7.2	Future work . . . . .	153
7.3	Conclusion . . . . .	158
<b>8</b>	<b>References</b> . . . . .	160
	<b>Appendix I</b> . . . . .	175
	<b>Appendix II</b> . . . . .	179

## List of figures

1.1 Main aims of the research project . . . . .	15
1.2 Layout of 16S rRNA experiments . . . . .	17
2.1 Structure of deoxyribonucleic acid . . . . .	21
2.2 Scientists awarded Nobel Prize for discovery of DNA . . . . .	21
2.3 Inventors of DNA sequencing . . . . .	24
2.4 Overview of three sequencing generations . . . . .	24
2.5 Maxam-Gilbert laboratory protocol . . . . .	25
2.6 Plus and Minus sequencing method . . . . .	27
2.7 Capillary dideoxynucleotide sequencing . . . . .	28
2.8 454-pyrosequencing method . . . . .	31
2.9 Dye-terminated sequencing-by-synthesis . . . . .	34
2.10 DNA nanoballs sequencing . . . . .	36
2.11 Ion semiconductor sequencing . . . . .	38
2.12 Single molecule fluorescent sequencing-by-synthesis . . . . .	39
2.13 Single molecule real time sequencing . . . . .	41
2.14 Polymerase enhanced nanopore sequencing . . . . .	42
2.15 Single molecule nanopore sequencing . . . . .	43
2.16 Schematic diagram of a nanopore concept . . . . .	48
2.17 Exonuclease-assisted nanopore sequencing . . . . .	50
2.18 Opti-pore nanopore sequencing . . . . .	52
2.19 Hybridisation-associated nanopore sequencing . . . . .	53
2.20 NanoTag sequencing by synthesis . . . . .	55
2.21 Transverse electron tunnelling nanopore sequencing . . . . .	56
2.22 Sequencing by expansion . . . . .	57
2.23 Representation of biological and synthetic nanopore technologies . . . . .	60
2.24 Schematic representation of a biosensor . . . . .	66
3.1 Three sequencing runs . . . . .	72
3.2 Structure of data analysis . . . . .	73
3.3 Sequencing flowcell . . . . .	81
3.4 Classification of reads . . . . .	84
3.5 Total amount of reads . . . . .	86
3.6 Average Phred scores . . . . .	87

3.7 Size distribution of reads . . . . .	88
3.8 Percentage of aligned reads . . . . .	89
3.9 Error rates . . . . .	91
3.10 Indel rates . . . . .	92
3.11 Mismatch rates . . . . .	93
3.12 Mapping quality vs. read base quality . . . . .	95
3.13 Diversities: Shannon, Simpson and Pielou's evenness . . . . .	96
3.14 Rank abundance distribution . . . . .	97
3.15 Heatmap representing reads realignment . . . . .	98
4.1 Loop-mediated isothermal AMplification (LAMP) . . . . .	104
4.2 Rolling Circle Amplification (RCA) . . . . .	105
4.3 NanoAmpliSeq protocol . . . . .	108
4.4 Size distribution of NanoAmpliSeq libraries . . . . .	110
5.1 General overview of the bioinformatics pipeline . . . . .	114
5.2 Sequence variability of different 16S rRNA gene regions . . . . .	115
5.3 Phylogenetic trees of <i>Legionella spp.</i> . . . . .	116
5.4 Schematic representation of chopSeq workflow . . . . .	120
5.5 Reduction of base identity vs. increase of tandem repeat length . . . . .	121
5.6 Percent identity of reads with INC-Seq and chopSeq algorithms . . . . .	122
5.7 Size of tandem repeats . . . . .	123
5.8 Proportion of the read aligned for longest contiguous alignment . . . . .	124
5.9 Schematic representation of nanoClust workflow . . . . .	125
5.10 Consensus sequence accuracy . . . . .	126
5.11 Number of OTUs and average sequence quality . . . . .	127
5.12 Amount of OTUs and quality detected for certain genus. . . . .	128
6.1 Experimental workflow of newest NanoAmpli-Seq workflow . . . . .	136
6.2 Average Phred scores from Oxford Nanopore sequencing run . . . . .	137
6.3 Average Phred scores from Illumina sequencing run . . . . .	138
6.4 Data analysis with Albacore and INC-Seq algorithms . . . . .	140
6.5 Data analysis with nanoCLUST algorithm . . . . .	141
6.6 Size distribution of 16S rRNA genes extracted with MATAM . . . . .	142
6.7 Size distribution of 16S rRNA genes assembled with MATAM . . . . .	142

# Declaration



University  
of Glasgow

## Declaration of Originality Form

This form **must** be completed and signed and submitted with all assignments.  
Please complete the information below (using BLOCK CAPITALS).

Name.....
Student Number.....
Course Name .....
Assignment Number/Name .....

An extract from the University's Statement on Plagiarism is provided overleaf. Please read carefully THEN read and sign the declaration below.

### I confirm that this assignment is my own work and that I have:

Read and understood the guidance on plagiarism in the Student Handbook, including the University of Glasgow Statement on Plagiarism	<input type="checkbox"/>
Clearly referenced, in both the text and the bibliography or references, <b>all sources</b> used in the work	<input type="checkbox"/>
Fully referenced (including page numbers) and used inverted commas for <b>all text quoted</b> from books, journals, web etc. (Please check with the Department which referencing style is to be used)	<input type="checkbox"/>
Provided the sources for all tables, figures, data etc. that are not my own work	<input type="checkbox"/>
Not made use of the work of any other student(s) past or present without acknowledgement. This includes any of my own work, that has been previously, or concurrently, submitted for assessment, either at this or any other educational institution, including school (see overleaf at 31.2)	<input type="checkbox"/>
Not sought or used the services of any professional agencies to produce this work	<input type="checkbox"/>
In addition, I understand that any false claim in respect of this work will result in disciplinary action in accordance with University regulations	<input type="checkbox"/>

### DECLARATION:

I am aware of and understand the University's policy on plagiarism, and I certify that this assignment is my own work, except where indicated by referencing, and that I have followed the good academic practices noted above

Signed .....

## Nomenclature

1D	One directional (read)
1D <sup>2</sup>	One directional squared (read)
16S rRNA	16 Subunit ribosomal DNA
2D	Two directional (read)
A	Adenine
BAC	Bacterial Artificial Chromosome
bp	Base Pair
C	Cytosine
ss	Single-stranded
ds	Double-stranded
dNTP	Deoxynucleotide triphosphates
DNA	Deoxyribonucleic acid
Gb	Gigabyte
HMW	High-Molecular Weight
HMM	Hidden Markov Model
kbp	Kilo base pairs
MAP	MinIon Access Program
Mb	Megabyte
ONT	Oxford Nanopore Technologies
PCR	Polymerase Chain Reaction
RPM	Revolutions Per Minute
RT	Room Temperature
rRNA	Ribosomal Ribonucleic Acid
RNN	Recurrent Neuron Network
SNV	Single Nucleotide Polymorphism
T	Thymine
ZMW	Zero-Mode Waveguide

# Glossary

## 16S rRNA

The 16S ribosomal RNA gene encodes for one of the two small subunits (the 30S and 16S) present in prokaryotic organisms. Ribosomes are essential for a protein synthesis process called translation and are present in all living organisms. Amongst small and larger subunits the 16S rRNA gene is one of the best-known gene markers used for phylogenetical identification of microorganisms.

## adapter

An adapter is very often a double-stranded DNA molecule in a linear or hairpin-like structure that is being attached to the end of the nucleic acid by ligation or PCR process. Addition of adapters to the DNA allows for barcoding of the samples but also converts nucleic acid molecules into sequencing libraries that in turn provides for immobilisation of the particles on the flowcell surface.

## aligner

The aligner is a bioinformatics program that is used to match sequencing reads to the reference database, e.g. SILVA. Aligners have various matrices and switches that allow for analysis of data generated with multiple types of sequencers that in turn produce higher or lower accuracy results.

## basecaller

A basecaller is a mathematical algorithm used to convert a raw electronic signal from a sequencer into a computer simplified human-readable data format. Their local or server-based installation requirements characterise various algorithms. Moreover, algorithms behind basecallers include statistical models such as HMM or more complex artificial neural networks like RNN.

## base pair (bp)

Guanine (G) and Cytosine (C) are complementary nucleotide bases bound by three hydrogen bonds while two hydrogen bonds connect Adenosine (A) and Thymine (T). Length of sequencing reads is measured in base pairs while contigs or genomes can be measured in kilo (Kbps) or megabase pairs (Mbps).

#### concatemer

Concatemer refers to a long continuous amplicon or DNA molecule made of repetitive DNA sequences like tandem repeats. This type of read can be generated during rolling circle amplification.

#### deletion

Deletion is a type of a mutation in the part of the chromosome or a sequence error that has been incorporated during DNA replication or data generation. Various numbers of nucleotides can be deleted, from a single base pair to a large piece of chromosome.

#### DNA

The deoxyribonucleic acid (DNA) also referred to as double helix, is a self-replicating molecule made of complementary nucleotide base pairs. The DNA contains genetic information and is present in almost all living organisms, including animals, plants or bacteria. The DNA is a blue print for RNA and proteins,

#### error rate

Error rate refers to the number of base substitutions, insertions/deletions that are being incorporated during PCR amplification but also during the sequencing process. Higher or lower error rate characterises various enzymes and sequencers. Amount of errors can be verified by comparison of generated data to their representative reference sequence.

#### fragmentation

The process of fragmentation requires DNA molecules and is used to break up long strands of nucleic acid into smaller pieces. This process is used during the process of library preparation and may be achieved by mechanical methods (centrifugation shearing, nebulisation, sonication) or enzymatic digestion (transposase, endonuclease).

#### hyper-branch

Hyper-branch is a DNA molecule structure that was synthesised with use of multiple displacement amplification methods such as rolling circle amplification. Construction of branches creates a complex DNA structure called a nanoball that cannot be analysed with any known DNA sequencer. Use of debranching protocols based on DNA fragmentation is used to remove unnecessary hyper-branches from concatemer DNA structure.



## INC-Seq

Intramolecular-ligated Nanopore Consensus Sequencing is a library preparation technique based on RCA. This method was designed for 16S rRNA amplicon sequencing on the nanopore MinION platform from ONT. INC-Seq also stands for software that is being used for concatemer consensus calling of the data generated with INC-Seq library preparation.

## indel

Indel refers to a type of an error that takes into consideration both insertions and deletions. This type of error is used for statistical descriptions of processed data and may be more prominent on certain sequencing platforms.

## insertion

Insertion is a type of an error that occurs during the addition of a single or more nucleotide base pair into a DNA sequence.

## nucleotides (dNTPs)

Nucleotides are organic molecules (guanine, cytosine, adenine and thymine) that serve as building blocks of DNA during polymerase chain reaction.

## phi29

Phi29 is a type of polymerase enzyme from *Bacillus subtilis* phage phi29. This polymerase performs very well at room temperature and has got great strand displacement activity.

## polymerase

The polymerase is an enzyme that synthesises DNA and RNA molecules by assembling nucleotides into a pair of complementary strands. The enzyme is present in every living cell and is required for the cell division and transcription processes.

## polymerase chain reaction (PCR)

Polymerase chain reaction is a molecular biology technique used to amplify a single or multiple parts of a DNA segments. PCR allows for exponential amplification of the particular DNA sequence across several orders of magnitude. This technique uses polymerase enzyme and is based on the ability to synthesise a new complementary strand of a DNA by adding new nucleotides to the pre-existing 3'-OH group at the end of the primer.

#### primer

Primer is a short (15-25bps) string of nucleotides in single-stranded DNA format that is designed to be complementary to a certain target of DNA sequence.

#### read

A read is a chain of DNA nucleotide molecules in a computer or human readable format (e.g. FASTA/FASTQ). Reads are generated during DNA sequencing process by the sequencer machine.

#### ribosome

The ribosome is a cell structure that serves as a molecular machine for the synthesis of proteins (translation). A prokaryotic ribosomal molecule is built of 30S small and 50S ribosome subunits. The small subunit of the RNA molecule is responsible for binding and reading the mRNA while the large subunit joins amino acids to form a polypeptide chain.

#### RNA

The ribonucleic acid (RNA) is a type of nucleic acid present in all living cells. The principal role of the RNA is to act as a messenger that carries information between DNA to form protein.

#### rolling circle amplification (RCA)

Rolling circle amplification is a molecular biology technique in which circularised single-stranded DNA, or RNA molecules are being amplified. This technique can be performed with use of an isothermal polymerase such as phi29 and with or without random hexamers. Moreover, a generated product of the RCA is in concatemer form, which means that the molecule is made of tens of tandem repeats complementary to the circular template.

#### sequencing library

Sequencing library is a collection of the DNA molecules suitable for analysis on a sequencing platform. Preparation of libraries is specific for a particular type of sequencer and design of the experiment. Very often libraries require fragmentation of the DNA material, PCR amplification and attachment of adaptors to the ends of the input molecules.

# 1 Project Description and Research Objectives

## 1.1 Research Objectives

### *Motivation*

Discovery of the DNA structure by James Watson, Francis Crick, and Rosalind Franklin in the 1950s and subsequent revolutionary developments in the area of biotechnology allowed for rapid innovations in many research fields such as microbial ecology, human health and engineering (Watson et al., 1953). Technological advancements in nucleic acid analysis (i.e. microarrays or sequencing) such as application of high-tech electronic components for molecular biology analysis reduced the time of analysis and increased data generation capacity (Heather et al., 2016). In turn, these influential technologies allowed scientists to solve various experimental, clinical, and environmental questions that until recently would have been impossible. For example, advances in understanding the human genome (Human Genome Project) and associated technological developments allowed for deep insight into the broad diversity of microorganisms in previously uncharted habitats (Venter et al., 2001; Ursell et al., 2012). Broad-scale microbiome studies such as the Human Microbiome Project (HMP) launched in 2008 (Turnbaugh et al., 2007) and the Earth Microbiome Project (EMP) launched in 2010 (Gilbert et al., 2014; Thompson et al., 2017) were designed to collect and characterise samples across different ecosystems. The HMP tested oral cavity, nasal passages, urogenital and gastrointestinal tract while EMP analysed microbial community around the globe such as soil, desert sand, deep ocean and freshwater lakes. Scientists began to interpret correlations among organisms detected in these environments but also environments themselves and how they affect the human and environmental processes. Sequencing technologies enable these studies by catalysing research in high priority areas such antimicrobial resistance and drug and vaccine development (Schwartz et al., 2003; Laddy et al., 2018). Moreover, sequencing

technologies are used in environmental research in drinking water and wastewater treatment, to both monitor and develop process strategies to improve these critical engineering applications (Pinto et al., 2012; Bautista-de los Santos et al., 2016). Furthermore, genomic investigations in the plant and agricultural sciences fields may allow for the development of drought-resistant or pest-resistant plants that would enable longer seasons and increased overall yield in agriculture, making it more sustainable in both developed and developing countries (Takken et al., 2000).

Key developments in the field of nucleic sequencing occurred in the early 2000s and these developments made metagenomics and amplicon sequencing accessible and affordable for a vast number of research laboratories (Metzker et al., 2005; Mardis et al., 2011; Heather et al., 2016). Due to a range of technological developments, high throughput sequencing has made computer scientists and mathematicians an essential component of data interpretation to answer questions in new interdisciplinary research fields. Bioinformatics is rapidly developing and is deploying a range of new statistical algorithms run on supercomputers (Pop et al., 2008). However, the rapid development in sequencing platforms means that often new data processing approaches and bioinformatics programs need to be developed for these newly emergent sequencing technologies. Novel algorithms and a better understanding of sequencing platform related errors and biases is necessary for improved and accurate data analysis (Schirmer et al., 2015). Moreover, comprehensive knowledge of the advantages and disadvantages of the most recent bioinformatics software is also crucial if we are to accurately interpret the outputs of these programs. Performing this benchmarking on different sequencing platforms and data analyses often relies on the use of pre-constructed synthetic communities with known sequencing composition and community structure (Shakya et al., 2013).

### *Main aims and objectives*

The purpose of this research project was to evaluate the potential of the MinION<sup>TM</sup> Mk1B nanopore sequencing device for amplicon sequencing based characterization of complex microbial communities. This includes estimation of bacterial community structure but also sequencing errors such as mismatch, insertion, deletion errors for standard amplicon library preparation and a modified version of the developed protocol. Analysis of data was initially performed with the help of reference-based workflow using multiple aligner algorithms while in subsequent steps reference-free (i.e., *de novo*) analysis of the data was introduced, including two novel bioinformatics softwares. The developed sequencing process can provide extensive information on the microorganisms in any given sample. However low quality of the raw data generates a large amount of false positive results including under or overestimation of microorganisms in the synthetic community sample, which can be corrected by using the developed protocol. Objectives of the research have been outlined in Figure 1.1 and indicate description for each of the respective chapters.

#### 1. Technology evaluation (Chapter 3):

- Identification of advantages and disadvantages of the nanopore technology:
  - Generation of a bacterial mock community sample made of 16S rRNA molecules and preparation of standard amplicon sequencing libraries.
  - Validation of multiple alignment algorithms on simulated data and estimation of their performance by examination of error rates and read types generated by the MinION<sup>TM</sup> Mk1B sequencer.
  - Detailed analysis and reconstruction of bacterial community structure using multiple bioinformatics programs and different read types.

#### 2. Improvement of data quality (Chapter 4):

- Rolling Circle Amplification (RCA)-based library preparation method for amplicon studies:

- Design and optimisation of novel sample preparation protocol for reduction of total error rates present in the amplicon data.

3. Benchmarking study (Chapter 4 and 5):

- Analysis of high-quality data generated with an RCA-based library protocol:
  - Generation of RCA-based sequencing libraries using a single organism and simple synthetic bacterial community for two sequencing chemistries (2D and 1D<sup>2</sup> sequencing chemistries).
  - Assessment of high-quality data with use of available basecallers and multiple concatemerisation algorithms (INC-Seq).
  - Comparison of data from former 2D library preparation chemistry and novel 1D<sup>2</sup> chemistry.
  - Estimation of false positive/negative and associated mismatch, insertion and deletion profiles and bacterial community structure for both sequencing chemistries.

4. Novel bioinformatics tools (Chapter 5):

- Development of bioinformatics tools to improve data quality:
  - Design and implementation of an advanced program (chopSeq) for precise detection of incorrectly oriented post-INC-Seq 16S rRNA amplicons and correction of the sequences.
  - Identification of limitations in currently used algorithms for *de novo* analysis of long-16S rRNA nanopore reads.
  - Program development for accurate binning of near full-length 16S rRNA sequences in to operational taxonomic units (OTUs) (nanoCLUST).

5. Environmental samples (Chapter 6):

- Analysis of wastewater samples with RCA-based library protocol:
  - Amplification of 16S rRNA genes with use of protocols: NanoAmpli-Seq and Illumina Nextera XT.
  - Sequence samples with the use of two different sequencing platforms, i.e. ONT MK1b and Illumina HiSeq 2500.
  - Comparison of data generated from ONT and Illumina sequencers, with statistical evaluation of errors and community structure.

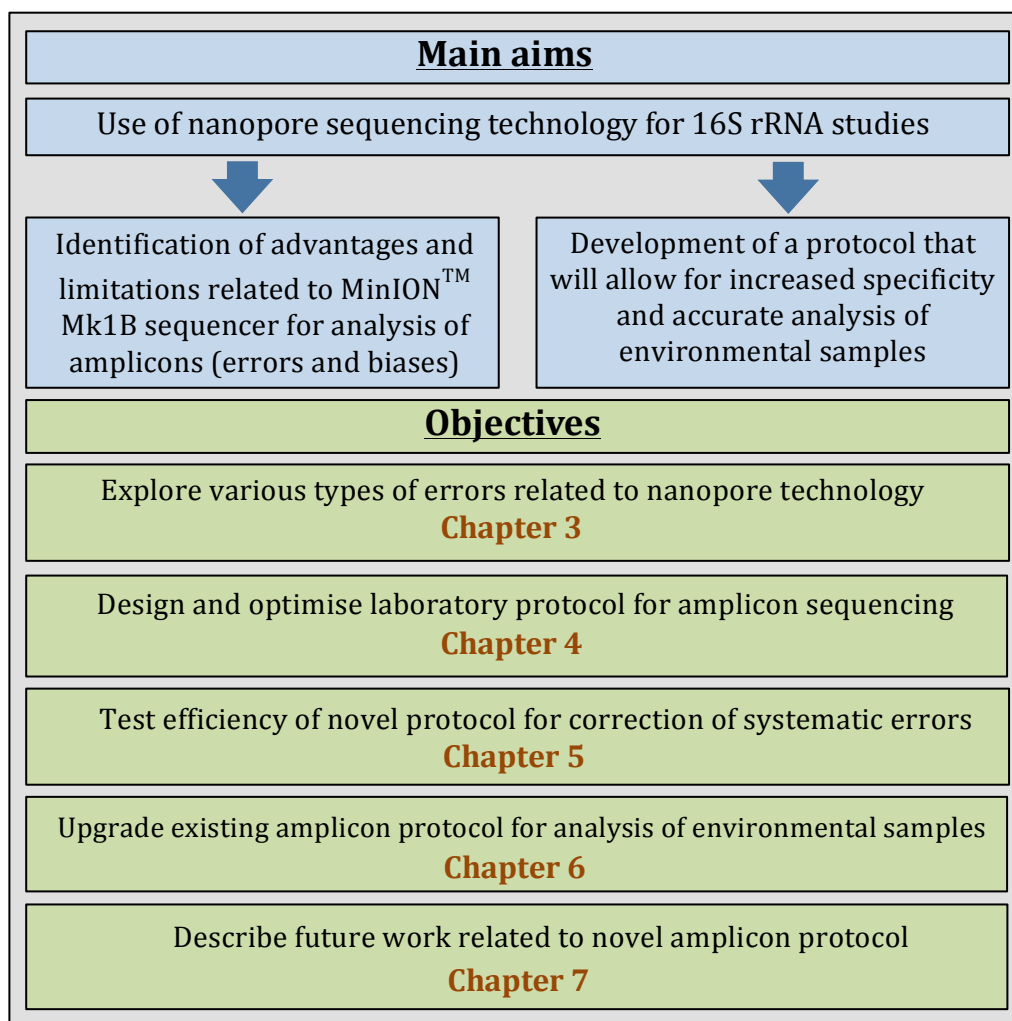


Figure 1.1 Figure outlines the main aims of the research project presented in this PhD thesis. Moreover, the chart summarises the objectives described in each chapter, more details of each objective is available in section 1.2.

## 1.2 Overview of Chapters

This section provides a list of chapters described in the dissertation including a short description of each part.

**Chapter 1** provides a general description of the PhD research project and highlights the importance of DNA sequencing, its potential applications, and issues related to data analysis. This chapter also illustrates the main aims and objectives of the project, which are associated with an evaluation of nanopore sequencing technology and improvement in sequencing data quality. Subsequent sections provide a list and overview of the chapters reported in the thesis. The first chapter ends with a list of conference posters, talks and peer-reviewed publications I carried out all along this PhD research.

**Chapter 2** describes a history of DNA discovery followed by an overview of first, second, and third generation sequencing technologies with a detailed description of technological concepts related to the development of nanopore sequencing. Moreover, this chapter outlines the significance of bioinformatics in high-throughput data analysis and promising applications of advanced high throughput sequencing technologies and their use for real-time clinical and environmental monitoring of pathogens.

**Chapter 3** describes initial nanopore sequencing experiments to assess the quality of the 16S rRNA gene amplicons and community structure of a sequenced bacterial mock community and various algorithms designed for error-prone data. This chapter starts with a short description of sequencing platforms and an overview of recent nanopore sequencing studies. It continues with an experimental workflow utilized for direct sequencing of 16S rRNA gene amplicons on the nanopore sequencing platform and is followed by data analysis. Results highlight the advantages and disadvantages of the nanopore sequencing technology but also discuss the requirement for improvement in library preparation methodology and a potential need for novel bioinformatics software. Results of this section are significant as they shaped further chapters described in my PhD research, which represent a significant advance in the quality of nanopore sequencing technology.

**Chapter 4** this section discusses the design of a novel library preparation method (i.e. NanoAmoli-Seq) for accurate amplicon sequencing on the nanopore sequencing platform. Innovative ways based on loop-mediated isothermal amplification (LAMP) and rolling-circle amplification (RCA) were investigated and the best-performing assay was optimised for enhanced performance of bacterial marker gene analysis.

**Chapter 5** discusses the experimental design of a benchmarking study to test mock communities containing 16S rRNA genes from one or ten bacterial organisms with two different sequencing chemistries: 2D and 1D<sup>2</sup> (Fig. 1.2). Samples were prepared using the previously validated and optimised RCA protocol and sequenced on multiple flowcells. This chapter introduces a bioinformatics program for correction of nanopore sequencing data processed to achieve amplicons of high sequence quality (i.e., chopSEQ).



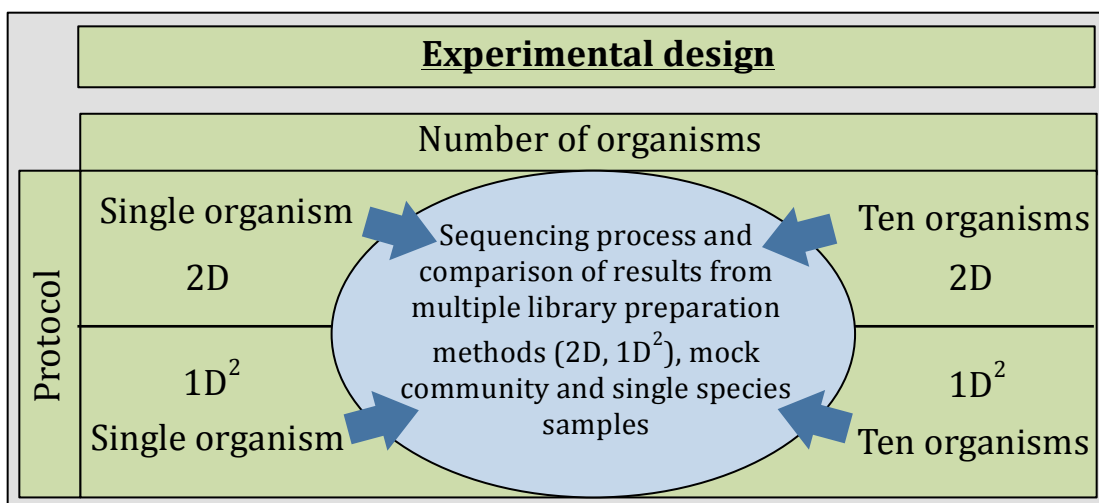


Figure 1.2 A schematic layout of experiments for 16S rRNA amplicon analysis with MinION™ Mk1b by Oxford Nanopore Technologies.

The software is based on detection of forward and reverse primer sequences, read orientation, and final collapse of long tandem repeats generated during read re-orientation. Correction of data is a significant part of the amplicon analysis workflow of NanoAmpli-Seq 16S rRNA nanopore data. Next, this chapter introduces a sophisticated algorithm for short multi-window binning that was incorporated (i.e. nanoCLUST) to develop high quality consensus sequences to minimize post-chopSeq residual errors resulting for OTU construction.

**Chapter 6** presents results of the last experiment, which was based on the previously described new library preparation (NanoAmpli-Seq) method and bioinformatics programs on samples from a wastewater treatment plant (i.e., complex microbial community). This chapter validates all the advances outlined in the previous chapters but also adds additional steps (i.e. addition of molecular barcoding for single molecule DNA tagging) and resolves data analysis bottlenecks (i.e. wrapper for multithreading data analysis) to improve speed of data analysis and sequence quality.

**Chapter 7** is the final section of the PhD thesis which summarises the conducted research and the application of the NanoAmpli-Seq protocol but also outlines future work that will be continued in current or expected collaborative projects. This section ends with a brief overview of how nucleic acid sequencing analysis may change clinical and environmental studies in the near future.

### 1.3 Overview of the public disclosures

#### **Conferences:**

**Calus S. T.**, Sevillano-Rivera M. C., Ijaz U. Z. and Pinto A. J., (2015). MinION-enabled and customer-led drinking water quality monitoring for pathogen detection. Conference: London Calling, London, UK; [Poster].

**Calus S. T.**, Ijaz U. Z. and Pinto A. J., (2016). MinION-enabled and customer-led drinking water quality monitoring for pathogen detection. Conference: Achieving Zero Bacteriological Failures in Water Supply Systems, Glasgow, UK; [Talk].

**Calus S. T.**, Ijaz U. Z. and Pinto A. J., (2016). Evaluation of multiple DNA aligners for the analysis of full-length 16S rRNA gene from mixed microbial communities from the MinION nanopore-based sequencing technology. Conference: American Society of Microbiology – Microbe, Boston, USA; [Poster].

**Calus S. T.**, Ijaz U. Z. and Pinto A. J., (2016). Evaluation of multiple DNA aligners for the analysis of full-length 16S rRNA gene sequences from mixed microbial communities using the MinION nanopore-based sequencing technology. Conference: Microbial Ecology and Water Engineering, Copenhagen, Denmark; [Talk].

Sevillano-Rivera M., Knapp C. W., **Calus S. T.** and Pinto A. J., (2016). Does water stress increase the incidence of antibiotic resistance genes in drinking water supplies. Conference: American Society of Microbiology - Microbe, Boston, MA; [Poster].

Sevillano-Rivera M. C., Knapp C. W., **Calus S. T.**, Dai Z. and Pinto A. J., (2016). Does water stress increase the incidence of antibiotic resistance genes in water supplies. Conference: Microbial Ecology and Water Engineering, Copenhagen, Denmark; [Talk].

Bautista Q. M., Dai Z., **Calus S. T.**, Sevillano-Rivera M. C., Ijaz U. Z., Sloan W. T. and Pinto A. J., (2016). Impact of source water on the structure and metagenomic profile of drinking water communities. Conference: Microbial Ecology and Water Engineering. Copenhagen, Denmark; [Talk].

Dai Z., Sevillano-Rivera M. C., Bautista Q. M., **Calus S. T.**, Ijaz U. Z. and Pinto A. J., (2016). Elucidating the long-term impact of disinfection strategies on the drinking water microbiome. Conference: Microbial Ecology and Water Engineering, Copenhagen, Denmark; [Poster].

**Calus S. T.**, Ijaz U. Z. and Pinto A. J., (2018) NanoAmpli-Seq: A workflow for amplicon sequencing from mixed microbial communities on the nanopore sequencing platform. Conference: ASM, Atlanta, USA and ISME Leipzig, Germany; [Poster].

Dai Z., Sevillano-Rivera M. C., Bautista Q.M., **Calus S. T.**, Ijaz U. Z. and Pinto A. J., (2018) Elucidating the long-term impact of disinfection strategies on the drinking water microbiome. Conference: ASM, Atlanta, USA and ISME Leipzig, Germany; [Poster].

**Published articles:**

**Calus S. T.**, Ijaz U. Z. and Pinto A. J., (2018). NanoAmpli-Seq: A workflow for amplicon sequencing from mixed microbial communities on the nanopore sequencing platform bioRxiv 244517, 2018. DOI:10.1101/244517 and GigaScience, Volume 7, Issue 12, 1 December 2018, giy140.

**Articles in preparation:**

Sevillano-Rivera M. C., Dai Z., **Calus S. T.**, Bautista Q. M. and Pinto A. J. Metagenomic analysis of antibiotic resistance genes in drinking water systems in the United Kingdom.

Dai Z., Sevillano-Rivera M. C., **Calus S. T.**, Ijaz U. Z. and Pinto A. J., Elucidating the long-term impact of disinfection strategies on the drinking water microbiome.

## 2 Introduction to DNA Sequencing

### 2.1 History of DNA discovery

Early discoveries of DNA occurred in the second half of the 19<sup>th</sup> century by a Swiss physician and biologist named Friedrich Miescher (Dahm, 2005; Dahm, 2008). In 1869, Miescher isolated genetic material called “nuclein”, later called DNA. Albrecht Kossel built on Miescher’s work to reveal that nuclein was composed of four bases and sugar structures (Jain et al., 2014), for which he received a Nobel Prize in the field of Physiology or Medicine in 1910 (Choudhuri, 2003; Dahm, 2008). Alongside Kossel, Phoebus Levene described components of DNA (adenine [A], thymine [T], guanine [G] and cytosine [C], Fig. 2.1) and showed that they are linked to each other in a phosphate-sugar-base structure. However, his idea that tetranucleotide structure of nucleic acid were bound together into a ring structure was later found to be wrong (Hunter, 1999; Simoni et al., 2002; Dahm, 2008). The mid 1940’s saw an extensive resurgence of interest in the field of nucleic acids after Oswald T. Avery, Colin MacLeod, and Maclyn McCarthy published results of their study (Avery et al., 1944) describing induced transformation of pneumococcal organisms from non-virulent into virulent strains with DNA as a carrier of microbial genetic information. In 1951, Rosalind Franklin obtained X-ray crystallographic picture of the DNA structure (Klug, 1968; Rapoport, 2002; Gibbons, 2012). Franklin’s X-ray crystallographic photographs of the DNA provided the impetus for the development and refinement of potential DNA models in an attempt to correctly unravel DNA structure by James Watson and Francis Crick (Watson et. al. 1953).

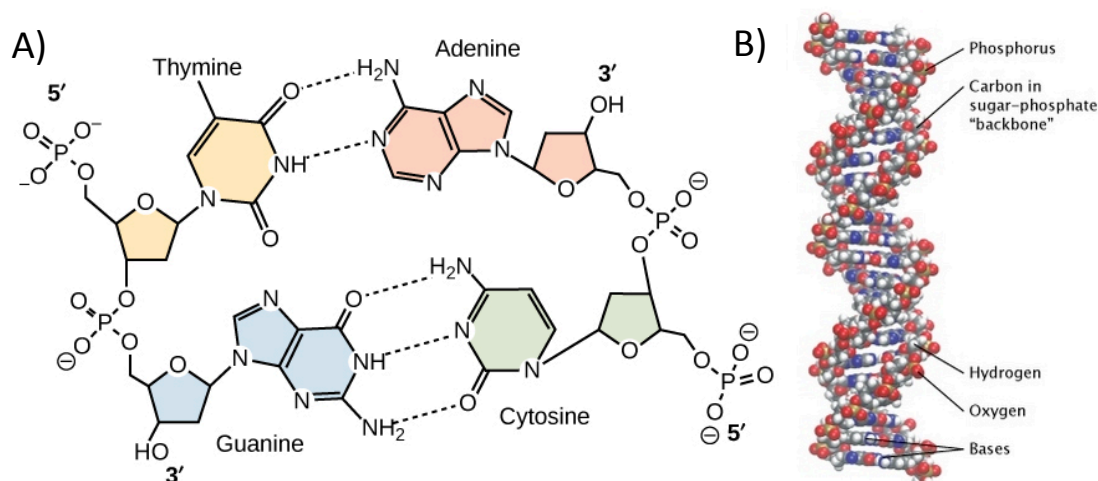


Figure 2.1: The structure of deoxyribonucleic acid (DNA). A) Arrangement of complementary nucleotides: Thymine – Adenine (two hydrogen bonds) and Guanine – Cytosine (three hydrogen bonds). When these nucleotides are connected with phosphate backbone, this creates a nucleic acid chain also called DNA. Each base is composed of three subunits: a nitrogenous base, five-carbon sugar, i.e., deoxyribose, and phosphate group. B) Structure of two biopolymer strands wrapped around each other that in turn forms a DNA double helix. The atoms used for the construction of four nucleotide bases are coloured-coded (Pray, 2008).

In 1962, Crick, Watson and Wilkins received a Nobel Prize in Physiology and Medicine "for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material" (Slobodkin, 2003; Chadarevian, 2003).



Figure 2.2: Pictures of Francis Harry Compton Crick, James Dewey Watson and Maurice Hugh Frederick Wilkins who were jointly awarded (1/3<sup>rd</sup> each) a Nobel Prize in Physiology or Medicine (1962) for the first correct description of a double-helix model of DNA structure. Rosalind Franklin, the co-author of a famous paper was not awarded a Nobel Prize because she died in 1958. Unfortunately, the Royal Swedish Academy of Sciences does not award posthumous prizes (Watson, Crick, Wilkins and Franklin, 1953). Picture: [https://www.nobelprize.org/nobel\\_prizes/medicine/laureates/1962/](https://www.nobelprize.org/nobel_prizes/medicine/laureates/1962/)

In 1957, Francis Crick also proposed a concept of the “central dogma” of biology, which described that information encoded in DNA is first converted into RNA (transcription) and later into a protein (translation) (Crick, 1970). One year later, Matthew Meselson and Franklin Stahl released results of their study, describing details on semiconservative replication of DNA (Meselson et al., 1958). The milestone discovery in the field of DNA sequencing was accomplished by Ray Jui Wu (Xue et al., 2016), who in 1971 used a polymerase enzyme for location-specific primer extension to sequence part of lambda phage. In 1972, Walter Fiers from Ghent University completed the RNA genome of bacteriophage MS2 (Remaut et al., 1972) and in 1977, Frederick Sanger, Allan Maxam, and Walter Gilbert developed the chain-termination DNA sequencing method (Sanger et al., 1977). Due to the low output of capillary-based Sanger sequencers, microorganisms with small genomes were analysed initially (e.g. *Haemophilus influenza*, *Saccharomyces cerevisiae* and *Caenorhabditis elegans*), while analysis of complex eukaryotic genomes (i.e., *Homo sapiens sapiens*) required 15 years of international collaboration (Sawicki et al., 1993; Collins et al., 1998). In 1983, Kary Mullis invented a revolutionary protocol called Polymerase Chain Reaction (PCR), however, it took him another 2 years to prove and publish his idea in Science magazine (Saiki et al. 1985). His discovery of PCR revolutionised a field of molecular biology (e.g. qPCR) and allowed for amplification of short DNA sections, which nowadays is crucial for high throughput screening (Shampo et al., 2002). Use of DNA sequencing methods is still not commonly used and cultivation of bacteria remains the gold standard for many clinical and environmental studies. Nonetheless, it is well known that a large proportion of bacteria are not cultivatable by existing methods (Harper-Owen et al., 1998; Sharma et al., 2005). For this reason, the majority of the microorganisms remain unexplored

using traditional methods. However, advances in DNA sequencing technologies could solve the aforementioned problems (Meyer et al., 2010). DNA sequencing-based studies have become widespread due to increase in sample throughput due to multiplexing capability and an overall reduction in price (Rothberg et al., 2008; Cox et al., 2010) and thus provided a fundamentally novel approach to study microbial communities in various environmental (e.g., soil, water) and clinical samples (e.g., skin, gut) (Mardis, 2008; Morozova, 2008 et al.; Shokralla et al., 2012).

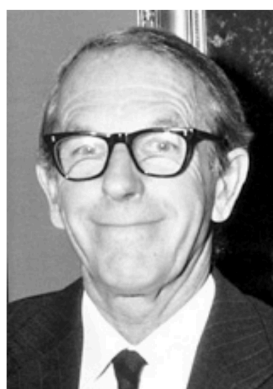
## 2.2 Description of DNA Sequencing Methods

The first attempt to sequence DNA was published in 1968 by plant geneticist Wu et al (Wu et al., 1968; Xue et al., 2016). Between 1969 and 1974 Wu's group published several papers on primer-extension polymerase-based DNA sequencing (Wu et al., 1971; Padmanabhan et al., 1972, Donelson et al., 1972). Much of this enzymatic primer-extension technique work was published before Fredric Sanger's first paper, 1975 (Sanger, 1975) on DNA sequencing. Between 1973-77, Allan Maxam and Walter Gilbert published a method based on nucleobase-specific partial chemical modification of DNA also called chemical sequencing (Gilbert et al., 1973; Maxam et al., 1977). This method used chemicals to cut DNA at the specific bases containing radioactive labels followed by polyacrylamide gel electrophoresis to determine the sequence of the DNA chain (Maxam et al., 1980). In 1977, British biochemist Fredrick Sanger published his sequencing procedure so called dideoxynucleotides chain-terminator sequencing, often named the Sanger-sequencing method (Sanger et al., 1977; Sanger et al., 1978). Fredrick Sanger and Walter Gilbert shared a Nobel Prize in Chemistry "for their contributions concerning the determination of base sequences in nucleic acids". In 1986 Applied Biosystems released automated

technology called ABI 370 (Chan, 2005; Czosnek et al., 2014), which was based on Sanger's dideoxy sequencing method.



**Walter Gilbert  
(1932 - )**



**Fredrick Sanger  
(1918 - 2003)**

Figure 2.3: Pictures of Walter Gilbert and Fredrick Sanger who were awarded the Nobel Prize in Chemistry for their discoveries in the field of DNA sequencing. It is often forgotten that in 1980 another scientist, Paul Berg shared half of the Nobel Prize for his discovery in DNA recombination while Gilbert and Sanger only 1/4<sup>th</sup> each. Picture: [https://www.nobelprize.org/nobel\\_prizes/chemistry/laureates/1980/](https://www.nobelprize.org/nobel_prizes/chemistry/laureates/1980/)

Since then, technology and sequencing methodology has progressed rapidly with second and third generation of DNA sequencing platforms, which have revolutionised biological research in many areas (Fig. 2.4).

Generation	Name	Release year	DNA read length
1st	Maxam-Gilbert Method	1976	200-300bp
	Sanger Sequencing	1977	<300bp
	ABI 370	1986	<1000bp
2nd	454 Life Sciences	2005	<800bp
	Polony Sequencing	2005	<2x13bp
	Solexa/Illumina	2006	<2x300bp
	SOLiD	2007	<2x35bp
	Complete Genomics	2009	<2x35bp
	Ion Torrent	2011	<200bp
3rd	Helicos SMS	2008	<100bp
	Pacific Biosciences	2011	<30,000bp
	Oxford Nanopore Tech.	2014	<1,000,000bp
	Genia	2014	<40,000bp

Figure 2.4: Overview of three generations of sequencing platforms, their commercial release and maximum DNA read length.



## First Generation Sequencing

### Maxam-Gilbert Sequencing

The Maxam-Gilbert sequencing process requires initial denaturation of double-stranded DNA into a single-stranded form by increasing temperature. Single-stranded molecules are then radioactively labelled with gamma-<sup>32</sup>P phosphate molecules which are added to 5' end of the DNA strand (Maxam et al., 1977), which is followed by chemical cleavage of DNA at specific positions. Dimethyl sulphate or hydrazine chemicals allowed for selective breakage of glycosidic linkage between the ribose sugar and the purine or pyrimidine bonds. Piperidine is then added to the reaction to catalyse phosphodiester bond cleavage and to cleave a base from the nucleic acid chain. Dimethyl sulphate and piperidine would cleave DNA at A or G's while piperidine with hydrazine would break C and T nucleotides.

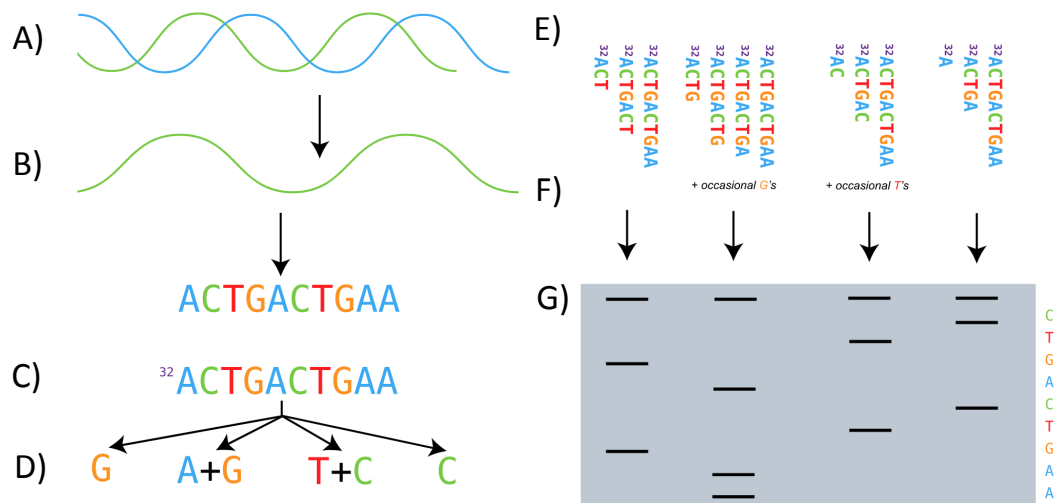


Figure 2.5: Simplified laboratory protocol of Maxam-Gilbert, chemical sequencing. A) extraction of dsDNA from the sample; B) denaturation of the dsDNA into single-stranded form; C) addition of radioactive <sup>32</sup>P phosphate to 5' end of ssDNA; D) cleavage of DNA at specific nucleotides i.e., G, C, A+G and T+C; E) different sizes of ssDNA molecules; F) use of high resolution acrylamide gel electrophoresis for size separation of four reactions; G) decoding of DNA sequence. Image source: <https://snipcademy.com/binf/img/lessons/sequencing-techniques/maxam-gilbert.svg>

Fragmented DNA molecules are loaded onto a high-resolution polyacrylamide gel and electrophoresed to achieve size separation. Subsequently, the gel is placed under X-ray film to use 5' radioactive tags to image the cleaved fragments. The sequences are confirmed by running multiple replicate samples on a single gel. Using this method a scientist could analyse only 200-300bp of DNA sequence every few days.

### *Plus -Minus and 2'-deoxynucleotide Chain Terminator or Sanger Sequencing*

The first version of Sanger's chain terminator sequencing is also called the 'Plus and Minus' method (Fig. 2.6), (Schuster, 2008). While the second version of the protocol is known as 'Capillary-based Chain Terminator Sequencing' or just Sanger sequencing (Fig. 2.7), (Sanger, 1977). Both methods require purified nucleic acids that are fragmented and cloned into bacterial vectors or PCR amplified. These methods are based on nucleotide chain termination with use of modified dNTPs. The 'classical' Sanger sequencing method relies on radioactively labelled nucleotides (radioactive phosphorus or sulfur isotopes) being added to four separate tubes. Each reaction contains a polymerase, primers and 2'-deoxynucleotide triphosphates (dNTPs) and modified chain-terminating 2',3'-dideoxynucleotide triphosphates (ddNTPs). One specific ddNTP nucleotide is added to every reaction e.g., ddCTP to the cytosine "C" reaction and so on. The polymerase-based extension of the primer is terminated when ddNTP is incorporated into a newly synthesised DNA (Sanger, 1982). As various dNTPs and single ddNTP are present in the reaction, the termination occurs rarely and at random positions in the DNA. It results in amplicons of varying lengths with 3' end always terminated by a nucleotide base. In case of "Plus and Minus" sequencing, the results of the PCR amplification are visualised using high-resolution agarose gel electrophoresis. Radioactive labelled nucleotides

incorporated into an amplicon are separated on four lanes. The sequence of the nucleic acid is deduced according to molecular weight of the fragments (Fig. 2.6), this process was manual and time consuming in compare to newer versions.

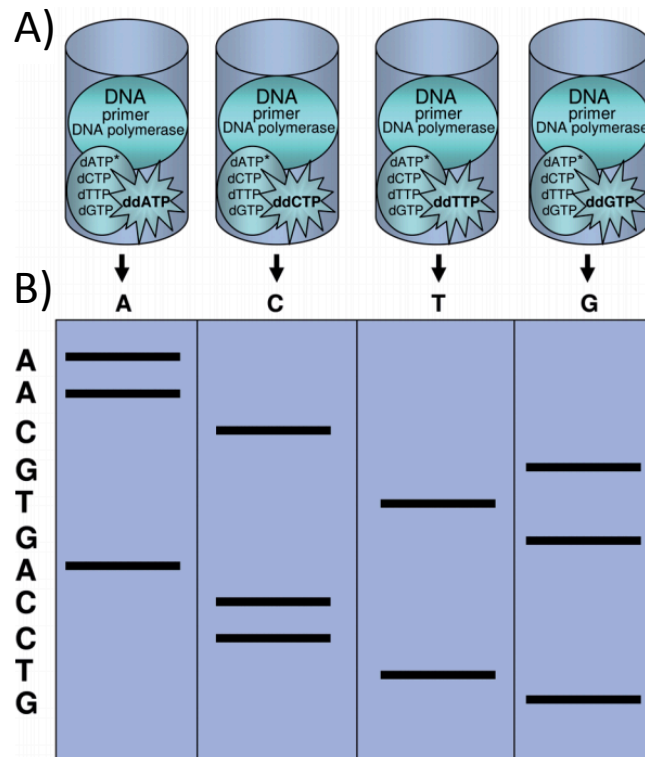


Figure 2.6: The ‘Plus and Minus’ sequencing method was the first version of the Sanger sequencing. A) the process is conducted in four separate tubes where PCR amplification was prepared with separate dideoxy chain-terminator nucleotides (i.e., ddATP, ddCTP, ddTTP, or ddGTP); B) this is followed by gel electrophoresis and visualised on a polyacrylamide gel to manually decode nucleic acid sequence. Image source: [https://application.wiley-vch.de/books/sample/3527320903\\_c01.pdf](https://application.wiley-vch.de/books/sample/3527320903_c01.pdf)

The second version of the Sanger sequencing included fluorescently labelled nucleotides; each ddNTP was tagged with a different fluorescent tag, which allowed for single reaction per sample instead of four (Fig. 2.7), (Sanger, 1980). Incorporation of automation decreased time of analysis, improved data quality and read length (up to 1,000bp) and allowed for multi-sample analysis with up to 96 and 384-well channels parallelizing the process. Low error rates and relatively long reads remain an advantage of the capillary sequencing and for this reason this type of sequencing is still used in various research projects.

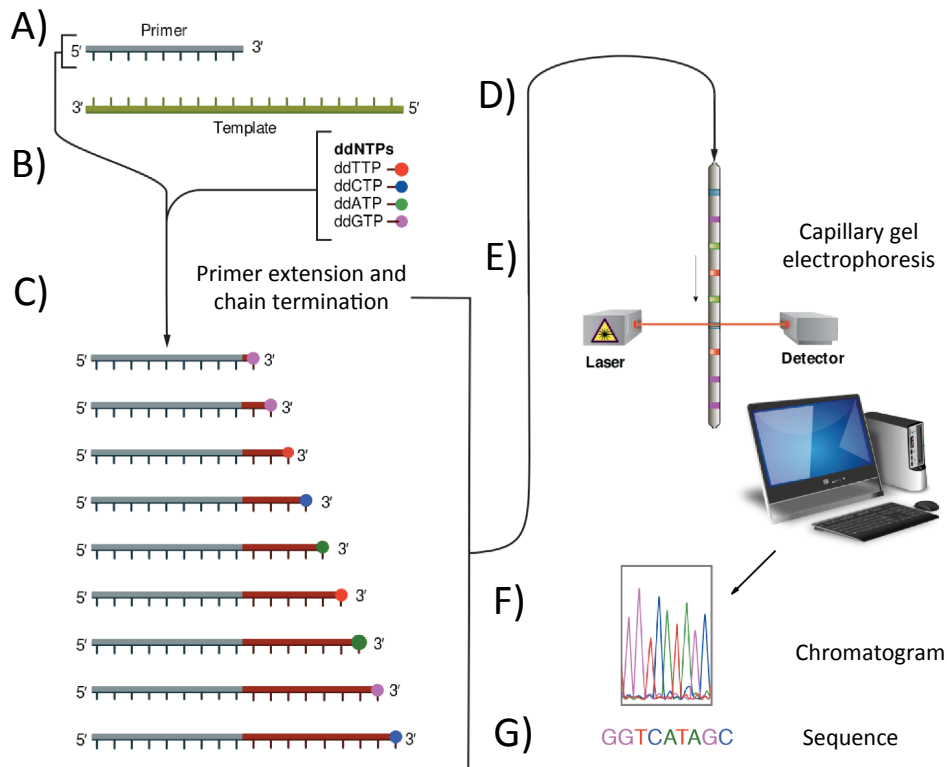


Figure 2.7: A schematic representation of fluorescently labelled dideoxynucleotide chain-terminator sequencing. A) denaturation of DNA into single-stranded form and attachment of sequencing primer; B) addition of fluorescently labelled nucleotides; C) primer extension and chain termination of polymerisation process; D) capillary gel electrophoresis of DNA fragments; E) detection of fluorescently labelled nucleotides with use of computer software; F) decoding of the chromatogram picture; G) final DNA sequence of the target molecule. Image source: <https://upload.wikimedia.org/wikipedia/commons/b/b2/Sanger-sequencing.svg>

## Second Generation Sequencing

The Second Generation Sequencing technology is often referred to as Next Generation Sequencing (NGS) or High-Throughput Sequencing (HTS). The main difference between First and Second Generation sequencing is automation and incorporation of massive parallelisation, which makes high data output possible. High throughput nucleic acid sequencing also allowed for a large variety of methods to be implemented i.e., amplicon, genome sequencing, genome resequencing, DNA-protein interactions (e.g., chromatin immune-precipitation), transcriptome profiling (RNA-Seq) and epigenome characterization (Wang et al., 2009; Anon, 2008).

### *454-Pyrosequencing – Pyrosequencing AB/Roche*

In 1993, Bertil Pettersson, Mathias Uhlen and Pål Nyren combined the solid phase sequencing method with streptavidin coated magnetic beads (Ronaghi et al., 1998) and recombined DNA polymerase without proofreading activity (3' to 5' exonuclease). The signal from each nucleotide was detected with a mixture of enzymes (i.e. DNA polymerase, firefly luciferase and ATP sulfurylase) and standard dNTPs (Ronaghi, 1998). Pyrosequencing process required ssDNA and sequencing occurred by incorporation of nucleotides during synthesis of the complementary strand. A fluorescence signal is detected by a camera and the light intensity is used to predict the number of dNTPs incorporated into a DNA strand during each cycle (Ronaghi et al., 1996). In 1998, the second version of Pyrosequencing included a new enzyme called apyrase. This enzyme removes unincorporated nucleotides during the DNA extension. Implementation of the enzyme mix at the beginning of the reaction provides a chance for simple automation of the sequencing process. In 2005, 454 Life Sciences introduced the third generation of Pyrosequencing also called 454-Pyrosequencing (Rothberg, 2008). Their method was based on the initial principle of solid phase sequencing with use of magnetic beads. Rothberg incorporated highly parallel Pyrosequencing using microarray nanofabrication (Hardiman et al., 2003). With this improvement, the 454-instrument became the first commercially available second-generation sequencing platform and initiated a new era in genomics. Application of nanofabricated chips allowed for a rapid reduction in the cost of DNA sequencing. The 454-Pyrosequencing method was established on a principle called “sequencing-by-synthesis” (SBS); this means a single stranded nucleic acid is sequenced by enzymatic synthesis of the complementary DNA strand (Rothberg et al., 2008). This requires fragmentation of DNA using sonication or nebulisation.

Subsequently, DNA fragments are inserted into oil microdroplets and attached to magnetic beads inside of the droplets (Fig. 2.8), within millions of droplets. Each microdroplet contains the polymerase and standard dNTPs. An amplification reaction occurs inside of the microdroplets and is called emulsion PCR or emPCR. Oil droplets are digested with chemical detergent, magnetic beads are purified and loaded on to the micro-fabricated chip containing millions of micro wells. The magnetic beads are immobilised inside of the microwells, then the sequencing process commences (Ronaghi, 2001). A solution of polymerase and various dNTPs are sequentially loaded (one dNTP at a time) and washed over the flow cell. Incorporation of any of the four dNTPs into the DNA strand releases pyrophosphate (PPi). ATP sulfurylase converts PPi into ATP when adenosine 5' phosphosulfate is available in the reaction. However, the light signal is produced when ATP acts as a catalyst for the luciferase-mediated modification of luciferin to oxyluciferin. Unincorporated nucleotides are degraded by apyrase enzyme and then the reaction restarts with new dNTPs. 454-Pyrosequencing generates DNA read lengths up to 700-800bp of high quality; while longer reads were attainable they suffer from poor sequence quality (Barbazuk et al., 2007). Other disadvantages of the platform included generation of chimeras, difficulties in homopolymer sequencing (i.e., systematic homopolymer errors), tag switching and challenges during emPCR if more than one template molecule hybridized to the bead. In 2013, Roche announced discontinuation of the 454-Pyrosequencing technology.

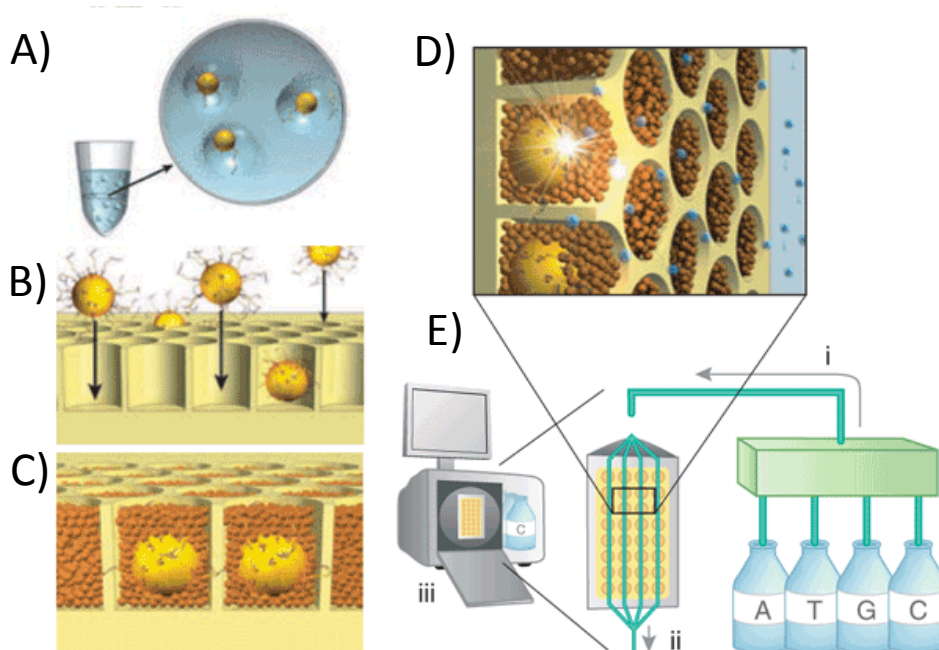


Figure 2.8: Simple representation of 454-pyrosequencing. A) DNA molecules are amplified in million of oil micro-droplets. This step is called emulsion-PCR and uses beads for DNA amplification; B) oil droplets are cracked with chemical detergent and beads containing hybridised DNA are loaded onto the flow cell; C) loaded beads are immobilised on the chip; D) sequencing process occur on highly parallel chip; E) automated software loads reagents (i) onto the sequencing flow cell (ii) computer collects images (iii) of the whole process for subsequent signal decoding. Image source: <https://www.nature.com/articles/nbt1485/figures/2>

### *Polymerase colony or Polony sequencing – Harvard University*

The Polony sequencing method was developed by Dr George Church's group from Harvard Medical School, Harvard University, USA (Mitra et al., 2004). This technology is the only one with open source downloadable software, protocols and could be operated using an epifluorescent microscope. The Polony sequencing protocol requires fragmentation of dsDNA to ~1,000bp length. Sheared genomic DNA is then inserted into T30 vector and self-ligated into a plasmid-like structure. Circularised DNA is subject to rolling circle amplification (RCA), followed by DNA cleavage with use of the MmeI restriction enzyme. Fragmented DNA particles contain 17-18bp flanking regions due to the cut distance from enzyme recognition sites. These paired-end DNA fragments are mixed with paramagnetic streptavidin-coated beads and processed using emPCR (Shendure et al., 2005). The emPCR solution is then

enriched for droplets with successful PCR amplification and the magnetic beads are finally loaded onto a glass slide with an aminosilane coating and acrylamide gel which results in them spreading out in a monolayer. The sequencing process relies on discriminatory capacities of polymerases and ligases (sequencing by ligation). Addition of sequencing reagents such as degenerate nanomers (anchor primers) is used for sequencing. Each of the anchor primers contains an attached fluorophore and allows for detection of the last added nucleotide base. The read length of the Polony sequencing is only 26bp per amplicon. In 2009, Polonator G.007 platform (Danaher, Hamamatsu, Leica) was released; however, advancements in the Polony sequencing technique contributed to development of a more competitive sequencing platform; i.e., ABI SOLiD sequencing (Shendure et al., 2017).

#### *Reversible Dye-Terminated Sequencing-By-Synthesis – Solexa/Illumina*

In 1994, Bruno Canard and Simon Sarfati from the Pasteur Institute in Paris designed a reversible-terminated nucleotide chemistry (Canard et al., 1994). Around the same time, scientists at the University of Cambridge (Shankar Balasubramanian and David Klenerman) used fluorescently labelled nucleotides to observe kinetics of the polymerase at the single molecule level while the enzyme was synthesising the complementary strand of DNA (Bentley et al., 2008). These advances resulted in the emergence of solid phase clonal array amplification of DNA, which in turn allowed for massively parallel analysis of short DNA fragments. This method is often referred to as “Sequencing by Synthesis” or SBS (Van Dijk et al., 2014). Further, the SBS process was enhanced by reversible-terminated nucleotides and in turn allowed for automation of the process. It was estimated that this method might enhance speed of DNA analysis and increase data output by a factor of 100,000-fold (Voelkerding,



2009). Early research and technology development of the Solexa sequencing platform was carried out in the Department of Chemistry at the University of Cambridge. The protocol involves initial fragmentation (sonication, nebulisation or tagmentation) of DNA into small pieces (100-600bp depending on sequencing chemistry). DNA fragments are clonally or enzymatically (PCR) amplified to increase template concentration. This step involves addition of adapters to the DNA molecules that are complementary to oligonucleotides immobilized on the sequencing flow cell. Hybridisation of DNA fragments to oligonucleotides is required for bridge formation, which is followed by solid phase bridge amplification of DNA molecules to create monoclonal clusters on the flow cell. The sequencing process begins with addition of sequencing primers complementary to one end of the DNA molecules and synthesis of the complementary strand polymerase enzyme extends the primer by incorporating fluorescently labelled nucleotides. The fluorescent signal due to the incorporation of nucleotides is captured by a camera, followed by de-blocking of terminated nucleotide and washing off of unused reagents and the cycle begins once more. Generation of clusters was crucial in the sequencing process as this increases the fluorescence signal and improves data quality (Cock et al., 2009). In 2006, the first commercially available sequencer was released by Solexa and was called the Genome Analyser. This instrument was able to generate 1Gb of data in a single run, which was significantly higher than the capillary sequencing method. In 2007, Illumina acquired Solexa and since then introduced multiple new sequencing platforms such as MiniSeq, MiSeq or HiSeq. This platform development reduced the price of DNA sequencing by roughly million-fold when compared to 1997 technology.

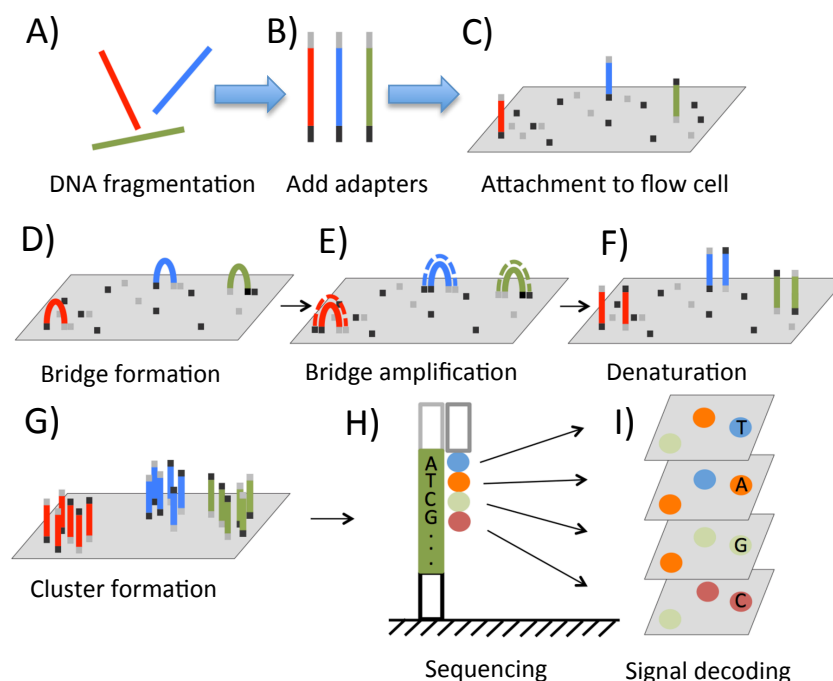


Figure 2.9: Schematic representation of SBS Illumina sequencing. A) DNA is fragmented into short ~500bp fragments; B) adaptors complementary to oligos on the sequencing flow cell are added to fragmented DNA; C) the DNA fragment is attached to the flow cell; D) generation of bridge structures E) solid phase bridge amplification; F) denaturation of amplified DNA clonal molecules; G) bridge amplification is repeated multiple time to create clusters of molecules; H) sequencing process is initiated with addition of primers, polymerase and fluorescently labelled nucleotides; I) images of the sequencing process are taken after addition of each base and later decoded with use of basecalling software.

Image source: [http://www.3402bioinformaticsgroup.com/wp-content/uploads/2016/07/590x452xNGS.png.pagespeed.ic.Ax4pdH\\_qJ0.png](http://www.3402bioinformaticsgroup.com/wp-content/uploads/2016/07/590x452xNGS.png.pagespeed.ic.Ax4pdH_qJ0.png)

### *Sequencing by Oligonucleotide Ligation and Detection (SOLiD) - Life Technologies*

In 2006, the Life Technologies released a novel platform called SOLiD with a library construction approach based on ligation followed by sequencing. This method significantly differs from 454 or Illumina sequencing because it uses a DNA ligase instead of DNA polymerase (Mardis, 2008). Samples are initially fragmented and ligated with P1 adapter molecules and these are immobilised to agarose beads. The beads containing DNA molecules are PCR amplified in the emulsion oil. Subsequently, amplified beads are attached to the glass flow cell and immersed in liquid containing fluorescently labelled di-base probes (2 base pairs). If one set of nucleotides matches the ssDNA sequence then bases are extended by ligation of the

probe. The ligated probe is cleaved to release base pairs containing the fluorescent molecule. Finally, the reagents are washed away and a new cycle of probe ligation begins. While this method only allows for sequencing of 25 to 35bp, the primary advantage was that SOLiD generates around 40 million reads in a single run, which was equivalent to 2-4Gbs. With this increase in throughput, SOLiD reduced the sequencing cost by 100-fold in 2 years and allowed to increase capacity by 5000-times in 5 years when measured in bases/machine/day. Initially, SOLiD sequencing generated more data (3Gb) than Illumina platform, however, this technology was characterised by a generation of very short sequencing reads (2x35bp). Moreover the sequencing process was time consuming and could take up to 14 days, depending on used chemistry.

#### *DNA nanoballs sequencing - Complete Genomics/ BGI*

In 2006, Clifford Reid, Radoje Drmanac and John Curson funded Complete Genomics. This company commercialised DNA sequencing technology that specialises in human genome analysis. The technology is based on generation of DNA nanoballs and subsequent sequencing them on a microarray-alike flow cell. In early 2009, Complete Genomics announced completion of the first human genome and by the end of the 2009 they had analysed 50 human genomes (Carnevali et al., 2012). The library preparation process includes fragmentation (via sonication or nebulisation) of high-molecular weight DNA into 400-500bp fragments, size selection if necessary, adapters are ligated to both ends of the DNAs and molecules are self-ligated into plasmids. RCA amplification is performed for template concentration, and subsequently restriction enzymes are added to cleave the amplicons 13bp to the right of the right adapter (Drmanac et al., 2010).

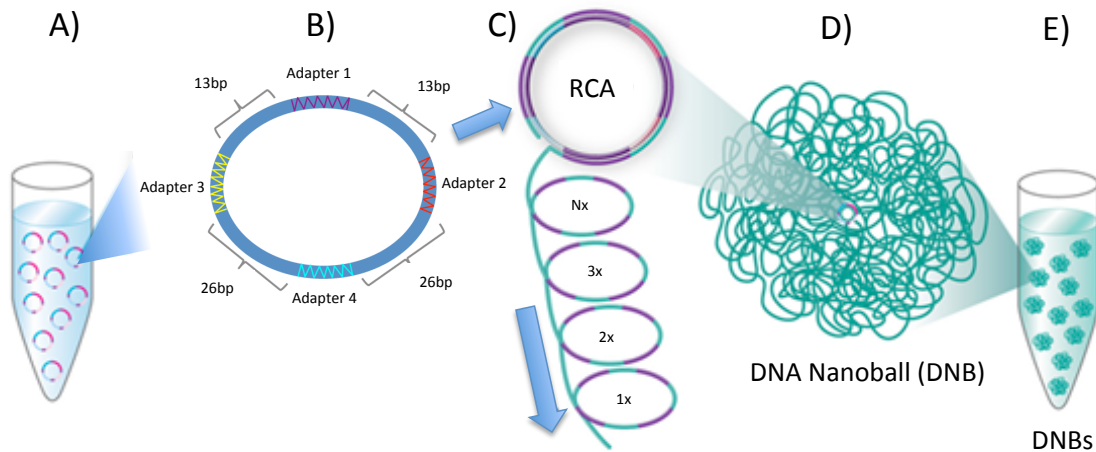


Figure 2.10 Complete Genomics, a library preparation of DNA. A) fragmented DNA pieces are ligated with adapters for multiple RCA assays and enzymatic fragmentations; B) short pieces of nucleic acid from previous step are ligated into a plasmid with four adapters and four DNA pieces; C) RCA amplification of the DNA plasmid is performed with use of phi29 isothermal polymerase; D) reaction creates very long chain of DNA that forms DNA nanoballs (DNBs); E) thousands of nanoballs present in the liquid are later loaded onto a microarray-like sequencing flow cell. Image source: [http://2.bp.blogspot.com/\\_VyTCyizqrHs/SOow5HldTqI/AAAAAAAAABco/5DjwO06ftgs/s1600-h/completegenomics.jpg](http://2.bp.blogspot.com/_VyTCyizqrHs/SOow5HldTqI/AAAAAAAAABco/5DjwO06ftgs/s1600-h/completegenomics.jpg)

This forms linear dsDNA molecules, which are again ligated with an adapter to form a circularised plasmid structure. RCA is performed again and a new restriction enzyme is added to the reaction that cleaves molecules 13bp to the left side of the left adapter. The reaction is repeated twice more with use of another restriction enzyme that generates breaks 26bp from each side of the adapters. The result of the amplification is circularised DNA containing four adapters and four DNA fragments made of 13 and 26bp. An additional step of RCA is performed, in which the long chain of highly concentrated nucleic acid creates the DNA nanoball structure (DNBs). These DNBs are finally loaded onto a microarray-like sequencing flow cell where anchor-probe ligation takes place. In 2013, BGI-Shenzhen acquired Complete Genomics to form the largest genomics service company in the world. They also simplified library preparation methodology with single DNA insert being ligated into a plasmid vector. Moreover, BGI-Shenzhen commercialised a platform called

BGISEQ-500 as competitor to the Illumina platform. Nonetheless, Illumina remains the leading sequencing platform for analysis of human, plant or microbial genomes.

### *Ion semiconductor sequencing – Ion Torrent Systems*

This method is very similar to 454-Pyrosequencing and was developed by Jonathan Rothberg, a scientist who previously worked for 454 Life Sciences (Rothberg et al., 2008). The ion semiconductor technology is based on fragmentation of DNA into short 200bp pieces and subsequent emPCR inside of oil microdroplets (Quail et al., 2012). The DNA fragments are released from oil droplets with use of chemical detergents and loaded into microwells where the sequencing by synthesis process taken place. However, this method does not measure fluorescence emission like in the pyrosequencing platform but rather the release of hydrogen ions ( $H^+$ ) that are discharged during dNTP polymerisation.

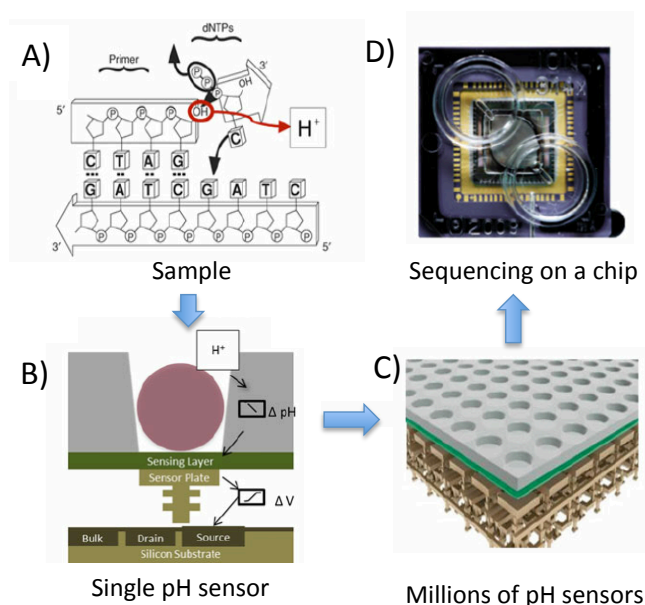


Figure 2.11: Schematic workflow of the ion semiconductor DNA sequencing with Ion Torrent. A) release of hydrogen atoms during the polymerisation process; B) atoms are recognised by a sensor present at the bottom of the micro-well; C) that can be scaled up into a set of micro-wells; D) sequencing is performed on a large microchip

Image source: <https://www.ncbi.nlm.nih.gov/pubmed/23929110>

This method was commercialised by Rothberg in 2007 and the sequencing machine was called the Personal Genome Machine (PGM), the second version of the platform was named Ion Proton and was meant to compete with Illumina HiSeq platform (Merriman et al., 2012). This technology suffered by high amount of errors ( $<Q20$ ), especially insertions and deletions caused by homopolymers but also chimeras due to emPCR and clonal amplification. Both Ion Torrent and Ion Proton generated low amount of data and were not competitive against Illumina HiSeq platforms.

### Third Generation Sequencing

#### *Single molecule fluorescent sequencing-by-synthesis - Helicos Biosciences*

Helicos single-molecule-sequencing is a method derived from the Sanger approach, first described in 2003 by Dr. Steve Quake from the California Institute of Technology (CalTech), (Ozsolak et al., 2011). This platform was the first commercially accessible technology for single molecule fluorescent DNA sequencing. The procedure requires fragmentation of dsDNA into short 100-200bp particles and ligation of polyA tails to the 3' ends of each molecule. Subsequently, dsDNA is denatured and loaded onto a flowcell, which contains complementary polyT oligonucleotides – the immobilisation phase. The DNA molecules are packed at a very high density, which allowed for high throughput data generation. The sequencing process requires polymerase and fluorescently labelled nucleotides that are incorporated and visualised during complementary strand synthesis with PCR (Fig. 2.12. A). Excess reagents are washed away (Fig. 2.12. B), followed by illumination of the flowcell surface (Fig. 2.12. C) and scanning with use of a laser that shows the location of each DNA template. The labels are enzymatically cleaved (Fig. 2.12. D) and the flow cell is loaded with fresh reagents containing different nucleotide. Every

single DNA strand is sequenced and visualised independently, which reduce chances of phasing and pre-phasing biases, which are present in technologies based on bridge amplification; e.g. Illumina, MiSeq. The Helicos SMS platform produces billions of short reads that are around 55bp in length. Longer sequences are not desirable because they will contain more errors (mismatch rate 0.2% while indel rate is 1-3%).

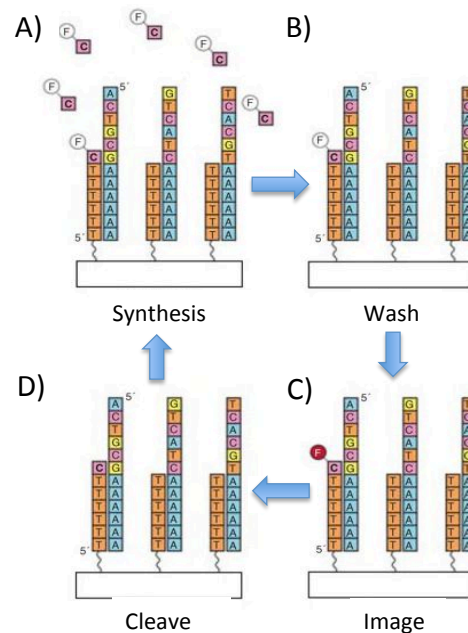


Figure 2.12: Workflow of Helicos single-molecule-sequencing. A) synthesis i.e. incorporation of fluorescently labelled nucleotide; B) removal of unused reagents; C) scanning surface of the flow cell to determine which strands incorporated fluorescently labelled nucleotides; D) enzymatic removal of fluorescent tag and repetition of the cycle. Image source: <http://slideplayer.com/5818924/18/images/49/Helicos+Biosciences%3A+Single+molecule+sequencing-by-synthesis.jpg>

### *Single-molecule real time sequencing (SMRT) – Pacific Biosciences (PacBio)*

The Pacific Biosciences of California was funded in 2004 and was established by researchers based at Cornell University. Their technology combines semiconductor and photonics sensors for DNA sequencing and real-time observation. The first commercial release of the PacBio RS platform was in 2010 and it was available to a limited number of customers. This platform uses real-time sequencing, which means, compared to previous technologies there is no need to pause the DNA polymerase to

identify incorporated nucleotide bases. The sequencing flow cells or so called SMRT cells contain thousands of miniaturised wells called zero-mode waveguide (ZMW) that are lit up from beneath. The ZMW is a well-like structure that leads the electromagnetic waves of the fluorescent light to the microscope camera aperture. The results are observed in real-time by a video recording of a polymerase incorporating each individual nucleotide. By scaling up the number of ZMW's the machine can monitor simultaneously tens of thousands reactions at a time. The initial library preparation requires 2000-5000ng of High-Molecular Weight (HMW) DNA to achieve 14-40kb fragments (Fig. 2.13 A). Subsequently, both ends of the long DNA particles are ligated with hairpin or so called SMRTbell adapters, which are used as a priming site for the polymerase enzyme present at the bottom of the wells (Fig. 2.13 B). The DNA complex is then immobilised at the bottom of each ZMW chamber (Fig. 2.13 C) and uniquely labelled nucleotides (i.e. fluorophores) are added to the ZMWs (Fig. 2.13 D). The synthesis of complementary DNA structure takes place and allows for the DNA complex to roll-around the polymerase. The fluorophores are clipped off the nucleotides by physical forces of polymerase that in turn allows for emission of light that is excited by the laser and recorded by the camera. Release of different colours indicates incorporation of specific nucleotide into the DNA chain. This method allows for sequencing of the same DNA library multiple times, which in turn increases data quality by consensus calling with PBDAGCon (Pacific Biosciences Directed Acyclic Graph Consensus), (Mosher et al., 2014). The generated raw signal is collected in real-time (Fig. 2.13 E) with sensitive a microscope camera and finally, the collected video is converted (i.e. basecalled) into FASTQ format (Fig. 2.12 F). The recent version of the PacBio RS II platform was able to generate reads with an average read length of ~8.5kb and up to 50,000 reads. The raw error rate was reported



to be ~11%, however, which could be reduced by use of aforementioned PBDAGCon and reduced down to ~0.001%. In 2015, Pacific Biosciences in partnership with Roche Diagnostics released a new version of the sequencer called the Sequel System. This platform is able to generate 7-times more data, has a smaller footprint and is 50% cheaper than its predecessor.

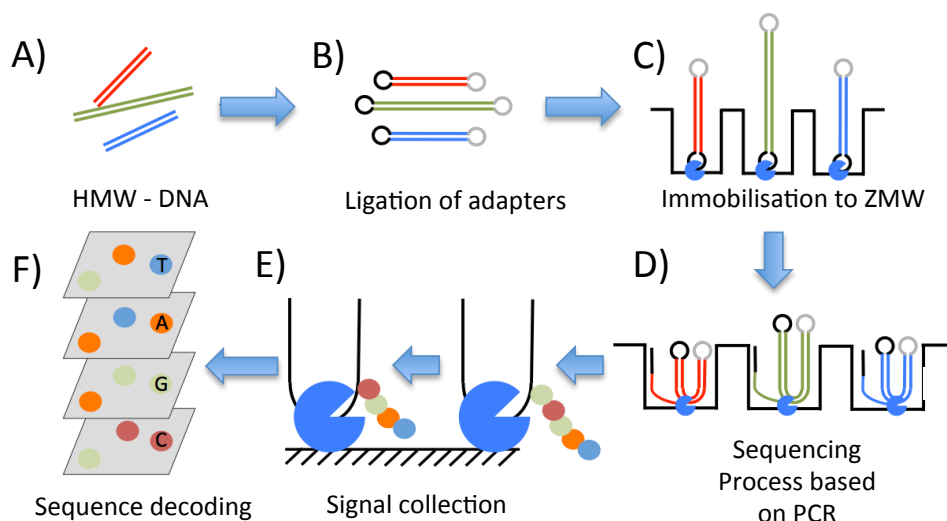


Figure 2.13: Workflow for the Pacific Biosciences library preparation and sequencing process. A) High-Molecular Weight DNA is sheared into 14-40kb fragments; B) hair-pin loop adapters are ligated; C) immobilisation on the flow cell takes place; D) circularised DNA libraries travel around the polymerase, generating fluorescence; E) signal is generated and collected in real-time, i.e. video; F) decoding of the signal (i.e. basecalling) follows later with specific software. Image source: <http://www.3402bioinformaticsgroup.com/wp-content/uploads/2016/07/641x380xpacbio.png.pagespeed.ic.CMSqyRLmny.png>

### *Polymerase enhanced nanopore sequencing - Genia Technologies*

One of the first nanopore-based sequencing platforms was introduced by Genia Tech. In 2014 the company was acquired by Roche, the pharmaceutical corporation that previously released the 454-pyrosequencing platform. The sequencing technology is still under development but is described in the subsequent chapter under “NanoTag sequencing by synthesis”

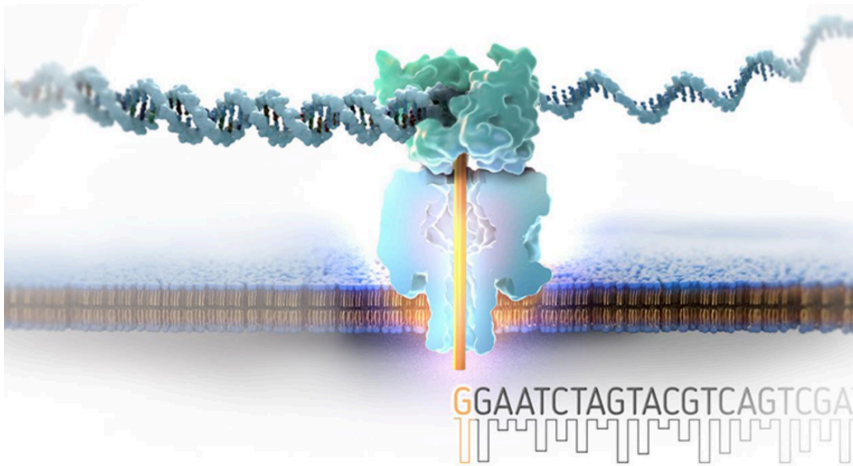


Figure 2.14 The sequencing process with use of the Genia/Roche sequencing technology. The DNA molecules do not move through the nanopore but travel through the polymerase and release the NanoTag, which is latter recognised by the electric current. Each nucleotide contains different NanoTags that create a characteristic signal for each base pair.  
Image source: <https://sequencing.roche.com/en/technology-research/technology/nanopore-sequencing.html>

#### *Single-molecule nanopore sequencing – Oxford Nanopore Technologies*

The most recent approach for single-molecule sequencing is being developed by the Oxford Nanopore Technologies (ONT). The company was founded in 2005 as a start-up from the University of Oxford by Spike Wilcocks, Hagan Bayley, and Gordon Sanghera (Eisenstein , 2012). The first MinION sequencer was available in early 2014 to limited group of researchers through the early entry plan, called MinION early Access Program (MAP). This technology utilises nanopores that are small tunnels with a diameter of around one nanometre. These tunnels can be made of either biological transmembrane proteins or artificial holes in a silicon oxide layer. Currently, ONT platforms utilise biological pores embedded into a synthetic membrane. These in turn are immersed in conductive liquid with electric current passing through the top to bottom chambers, similar to electrophoresis. Figure 2.14 represents a cross-section of a single nanopore (blue) showing ssDNA being translocated through the pore and dsDNA outside of the pore; the green protein is

used to slow the movement of DNA through the pore. While the sensing point is in the middle of the nanopore, the results are sent out in real-time to a computer as visualised in the grey box.

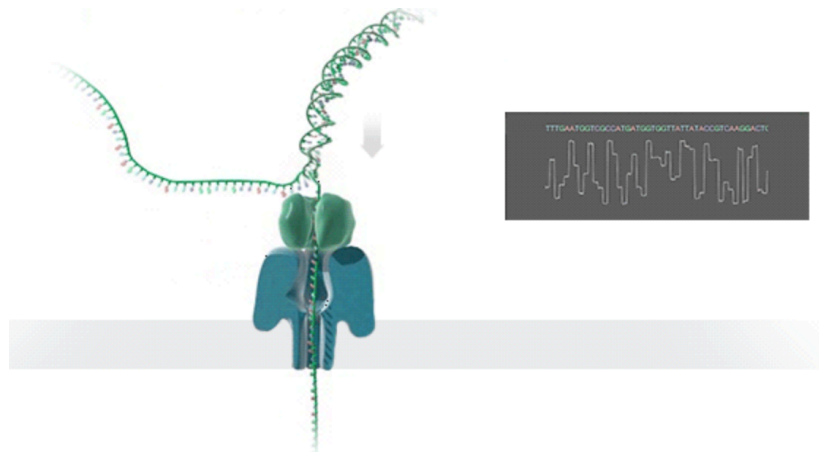


Figure 2.15: Oxford Nanopore Technologies workflow, based on single stranded-nucleic acid passing through the nanometer size hole, i.e. nanopore. The biological nanopore senses multiple bases (k-mer) at a time and collects information about them, visible in the grey chart. Image source: [https://mcic.osu.edu/sites/mcic/files/imce/images/MinION\\_nanopore.jpg](https://mcic.osu.edu/sites/mcic/files/imce/images/MinION_nanopore.jpg)

Each nucleotide causes a distinctive change in the intensity of the electric current and so the DNA sequence can be decoded. This technology does not require fluorescently labelled nucleotides or expensive microscope cameras and for these reason it is the only portable DNA sequencer available on the market (Laver et al., 2015). Moreover, rapid translocation of the DNA molecule through the nanopore can allow for real-time data generation and close to real-time data analysis.

## 2.3 History of Nanopore Sequencing

The idea of nanopore sequencing originated in multiple research groups during the 1980s. The first available notes related to the nanopore concept are dated to 1989 and are available thanks to handmade drawings of David Deamer (Deamer et al., 2016). Deamer suggested that it might be feasible to decode a sequence of DNA molecules by pulling them through a nanometer hole with use of electrophoresis-like forces. The nanometer-size tunnel can be made of natural protein embedded into a biological or synthetic polymer membrane or nanohole created in a solid-state sheet. In 1995 George Church, David Deamer, and Daniel Branton applied for a patent, which described nanopore sequencing. One year later for the first time scientists managed to transport DNA molecules through an alpha-haemolysin pore (Wang et al., 2015). Thanks to further research and development of the technology, in 1999 decoding of electrical traces allowed distinguishing between purine and pyrimidine nucleotide bases from a single RNA molecule (Bundschuh et al., 2005). Two years later, alpha-hemolysin pores were embedded into a lipid-bilayer, which separated two chambers containing a potassium chloride liquid solution. In 2005, Oxford Nanopore Technologies (ONT) became established, whereas at the same time Bayles and Ghadiris (Ashkenasy et al., 2005) showed that all four bases present in the DNA chain could be recognised and distinguished with use of the nanopore technology. This discovery of single nucleotide base discrimination was a critical point and allowed for further development of the strand-sequencing approach instead of exonuclease-nanopore sequencing. Since then, various improvements have been introduced that have increased sensitivity, accuracy, and scale of the device. In February 2012 during the Advances in Genome Biology and Technology conference, ONT presented a nanopore-sequencing device called MinION<sup>TM</sup> for the first time. Two years later,

ONT opened the MAP to early users of the MinION<sup>TM</sup> sequencing technology. The first ever publicly released data from ONT technology was released in June 2014 based on the sequencing of the O6-antigen determining fragment from the *Pseudomonas aeruginosa* bacterial genome (Loman et al., 2014). Since then, hundreds of researchers across the world have tested this portable nanopore device and multiple papers have been published describing various research projects.

## 2.4 Description of various nanopore concepts

A general description of nanopore-based technology is relatively uncomplicated and could be stated as a nanometer-size hole through which molecules pass and generate changes in ionic current flowing through the pore (Siwy et al., 2010; Taniguchi et al., 2009). The size of these pores ranges from one to a few nanometers in diameter and could be made of biological proteins; e.g. alpha-hemolysin, MspA or CsGg embedded into a biological or synthetic membrane made of materials such as graphene, silicon nitride, silicon oxide or metal oxides (Schneider et al., 2010). Biological nanopores are proteins used in devices such as MinION<sup>TM</sup>, GridION<sup>TM</sup> X5 or PromethION<sup>TM</sup> by Oxford Nanopore Technologies (ONT). However, these natural nanopores possess various limitations such as pore blocking (Aksimentiev et al., 2004; Branton et al., 2008; Deamer, 2010) during the sequencing run that in turn decreases the number of active pores and total quantity of generated data. Further, the tunnel depth of biological pores and some synthetic pores is relatively large when compared to the size of a single nucleotide base. This means that multiple DNA bases (4, 5 or even 6 bases) are present inside of the sensing nano-groove while a DNA molecule travels through the pore (Schneider et al., 2010). The electric signal is then generated from multiple bases as the signal recognises k-mer sequences instead of a single base (Jain

et al., 2015). This in turn may cause problems while analysing homopolymer and tandem repeat structures, resulting in insertion or deletion errors. These high error rates have been reported in multiple papers describing ONT technology, which results in limitations for various protocols; e.g. accurate environmental 16S rRNA analysis or metagenomic sequencing (Goodwin et al., 2015; Laver et al., 2015). Multiple bioinformatics programs and laboratory protocols have been developed in the past few years to tackle the problem of high error rates (Loman et al., 2015; Goodwin et al., 2015). These algorithms improved the accuracy of the generated data with more sophisticated basecallers; e.g. use of Recurrent Neural Network instead of Hidden Markov Models or improved assembler and aligner programs designed for error-prone reads; i.e.. LAST, BWA ONT2D or GraphMap. Use of natural proteins may be also disadvantageous due to their natural degradation. This can cause problems with long-term storage of sequencing flow cells, as the active pores decrease with storage time, resulting in lower total data output when the flow cell is used (Branton et al., 2009). Furthermore, long-term application of electric current on biological pores may be detrimental to protein structure. Prolonged exposure of the nanopore proteins to the electrical current decreases sensing accuracy and allows for a higher error rate of data generated in a later stage of sequencing run when compared to the beginning of the analysis; synthetic materials could resolve these issues. For example synthetic materials, with the thickness of a single atom (e.g. graphene) could allow for single-base resolution of the nucleotide chain (Sint et al, 2009) and the size of the pore could be changed depending on the width of the analyte. For example, nucleic acid molecules could be analysed using pores with a 1nm diameter while for larger molecules like proteins, the size of the hole can be increased. Different sizes of solid-state nanopores can be formulated in silicon or metal bases with use of atomic layer

deposition, e-beam drilling, a beam sculpting (Li et al., 2001; Chen et al., 2004). An array of nanoholes is placed into an electrolytic liquid containing conductive molecules and electrodes immersed in each chamber. Movement of the analytes through the nanopores is achieved by application of electric-field-induced translocation and detected by pore-surrounding sensors as during translocation the particle changes the electric current inside of the nanopore (Deamer et al., 2000). When voltage electromotive force is applied, electrolyte ions in solution move electrophoretically only through the hole generating an ionic current signal. When negatively charged DNA particles travel from the *cis* chamber through the pore it generates a blockage and interrupts the signal. Amplitude duration, and current blockages from translocation events are used for calculation of the physical and chemical DNA properties (Shendure et al., 2005; Aksimentiev et al., 2004). The addition of graphene would allow for insulation of two separate chambers containing conductive liquid and the graphene membrane could also conduct electrical signals and send raw information directly to electrodes and then to the hard drive for analysis (Siwy et al., 2010). The signals generated during translocation are decoded via a basecalling process by pattern recognition algorithms (Hidden Markov Model or Recurrent Neural Network) and converted into standard (FASTQ/A) sequence record nucleotide format (Leggett et al., 2015). This sequence must be later analysed with use of multiple bioinformatics algorithms for a *de novo* or reference based search against high-quality databases (e.g. GreenGenes or SILVA).

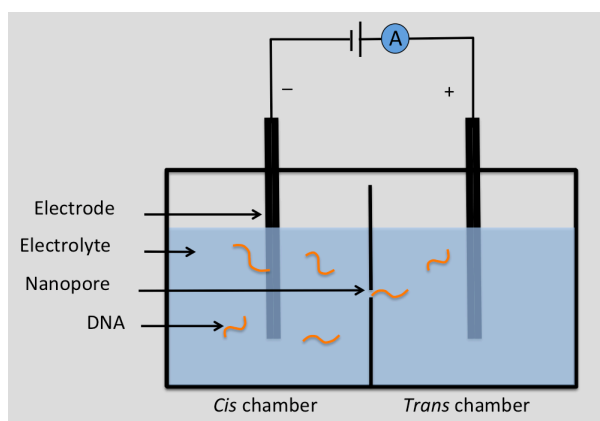


Figure 2.16 Cross section representing a schematic diagram of the nanopore concept. The device contains: negative (anode) and positive (cathode) electrodes, electrolytic liquid present in *cis* and *trans* chambers, nanopore, electric supply, DNA molecules and materials measuring electrical resistance and conductance i.e. Ohmmeter or Multimeter (A). Image source: own work.

Potential advantages of the nanopore-based analysis of DNA and RNA over sequencing by synthesis technologies include avoiding the polymerase chain reaction and their biases (Turner et al., 2015). Eliminating this step also decreases the time for sample preparation but also reduces sequence errors caused by the polymerase reaction and amplification biases when analysing environmental microorganisms (Polz et al., 1988; Pinard et al., 2006; Klindworth et al., 2013). Further, lack of PCR amplification allows for preservation of epigenetic base modifications such as methylations (Jones et al., 2001; Liu et al., 2015; Simpson, 2017). Recognition of these inherited modifications can be identified as modified nucleotide bases that generate a different current signal (Shim et al., 2013; Laszlo et al., 2013). Other DNA sequencing technologies are based on initial PCR amplification and fluorescence detection and are not feasible for direct analysis of epigenetic marks (Rhoads et al., 2015; McCarthy, 2010). Additionally, elimination of fluorescently labelled nucleotides and the need for extensive microscope cameras to detect light signals allows for miniaturising the device, which is incredibly advantageous for analysis of environmental samples in the field (Quick et al., 2014). Another benefit of this biosensing capability of the nanopore technology is its scalability and potential for



high throughput output (Tarraga et al., 2016). This means that the biosensing device can be enlarged to generate more and faster results by including a higher number of pores. High throughput lowers the cost of the analysis and would allow for data output for tens or hundreds of multiplexed samples. There are also challenges and disadvantages related to nanopore biosensing technology. One of them is the high amount of the analyte (i.e., nucleic acid material) required for analysis. Lack of PCR amplification and direct sequencing of native DNA strands limits the types of samples that are suitable for analysis, especially for low biomass environments like air or drinking water, which could require an additional step of cultivation or PCR amplification. Furthermore, use of high amounts of DNA may limit scope for potential re-sequencing (Branton et al., 2009).

## Biological nanopores

Exonuclease-assisted nanopore sequencing was proposed by two independent groups (i.e. Bayley and Wang) and was meant to overcome a problem of signal interference from nearby bases (Wang et al., 2015). Sensors measure modulation of the ionic current from individual dNTPs while they are sequentially cleaved off by an exonuclease from the leading nucleotide chain strand and traversed through the nanopore. Furthermore, it was recently shown that this technique can be used for both DNA and RNA analysis (Henley et al., 2016). Further, in addition to four dNTP's (adenosine, thymine, guanine and cytosine), this technique could also detect epigenetically modified bases such as 5-methyl-2'-deoxycytosine (5-mdC) on DNA or uridylation of 3' ends of RNA molecules (Ayub et al., 2012). Detection of genomic and post-transcriptomic expression modifications could allow for direct analysis at the single-molecule level.

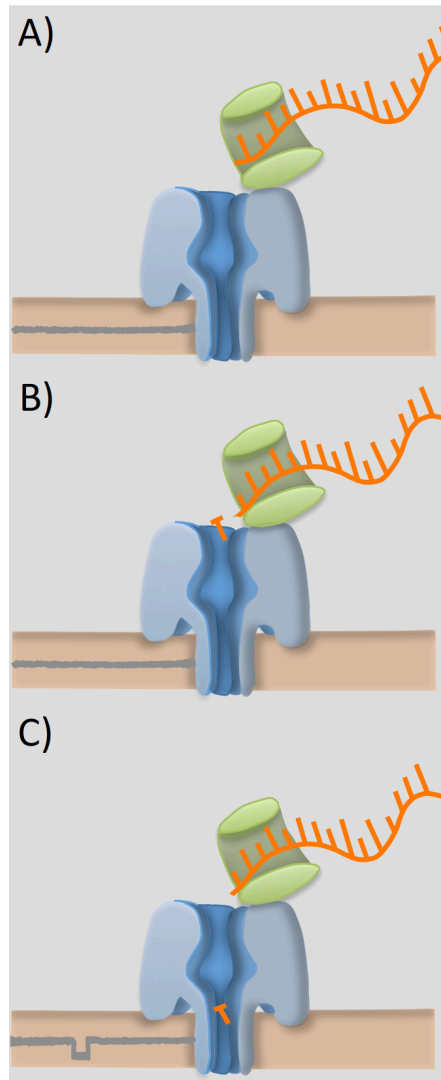


Figure 2.17. Cross-section of nanopore embedded into a membrane. A) sample is prepared by denaturation of the dsDNA into single-stranded form and then molecules are placed into an enzyme capable of cleaving phosphodiester bonds at the end of DNA chain; B) nuclease enzyme is placed on the top of the nanopore and it cleaves a single nucleotide at a time; C) the ionic current is applied across the membrane which translocates single nucleotides down the nanopore which turn allows for accurate resolution of the DNA sequence. Image source: own work.

It has also been shown that exonuclease-assisted nanopore sequencing generates a lower error rates (mean accuracy 99.8%) as it allows for single base analysis instead of multiple nucleotides being present in the pore (Astier, 2006; Clarke, 2009; Wang, 2015). However, the dwell time for each nucleotide base in the alpha-hemolysin pore is longer when compared to the commonly used strand-sequencing method. This makes this method slower and in turn would generate less data in the same experimental time. Nonetheless, this method is purely based on sensing of translocating events and does not require microscope camera for detection of fluorescently labelled nucleotides. This allows for use of the sequencing instrument at a miniaturised size.

Opti-pore is a nanopore-based sequencing technique that uses optical readout of synthetic DNA strands passing through the nanopore (Singer et al., 2012). This method is designed to enhance signal-to-noise ratio by use of fluorescently labelled molecule probes. Combination of optical sensing with electrical detection of translocating ssDNA moving through the pore could generate contrast between nucleotides and could potentially allow for detection of epigenetic base modification if the electric current is measured (Wang et al., 2015). This protocol requires changes to the DNA to convert it into a synthetic molecule that is a hybrid of the original ssDNA and synthetically designed DNA polymers (DDP) made of binary coded oligonucleotides. One of the disadvantages of the opti-pore technology may be a requirement for a microscope camera required to detect each DNA molecule as it moves through the nanopore. The use of camera recognition of DNA molecules would cause an increase in price and size of the device that in turn would create a

limit its portability. Furthermore, cross talking of fluorescent signals could reduce accuracy and increase error the rate of the data.

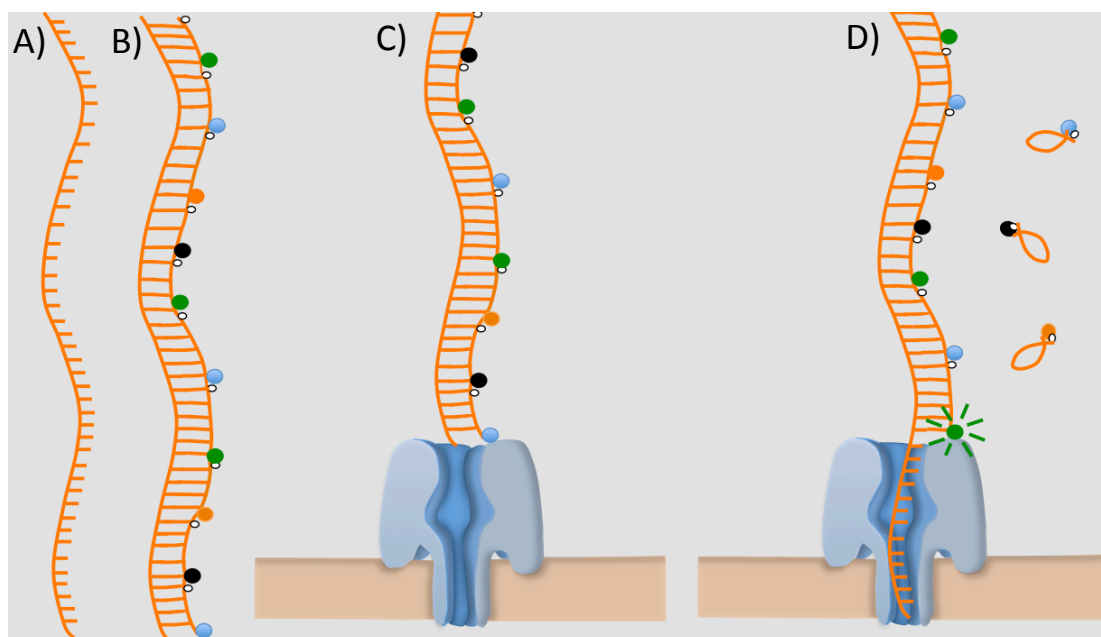


Figure 2.18. Cross-section of nanopore embedded into a membrane. A) DNA fragment is first denatured to create a single-stranded molecule; B) subsequently the DNA particle is converted into synthetic double-stranded DNA containing fluorescently labelled probes; C) prepared sequencing libraries are loaded onto the membrane with nanopores or a nanoholes; D) when ionic current drives DNA molecule through the pore then fluorescently labelled probes are unwound. What allows for the camera to record the colour of the dye and convert it into a nucleotide format. Image source: own work.

The hybridisation-associated nanopore sequencing (HANS) platform was developed by NABsys Inc. The protocol for this type of sequencing involves fragmentation of dsDNA into very long fragments of ~100kb on average and denaturation to form a single-stranded forms. Molecules of ssDNA are hybridised with short oligonucleotide probes, between 4 to 8bp. This step generates hybrid DNA molecules containing the ssDNA backbone with hybridised dsDNA sites. Genomic libraries are then pulled through the nanopore by applying an electrical current. While the dsDNA hybrid site moves through the pore, it will change the signal of ionic flow as compared to ssDNA (Ling, 2007). Stronger induction creates a certain blockage signal that allows for recognition of complementary sites of the probes on the nucleic acid strand. When the

DNA hybrid molecule passes through the pore, it generates a specific linear map of the hybrid molecule.

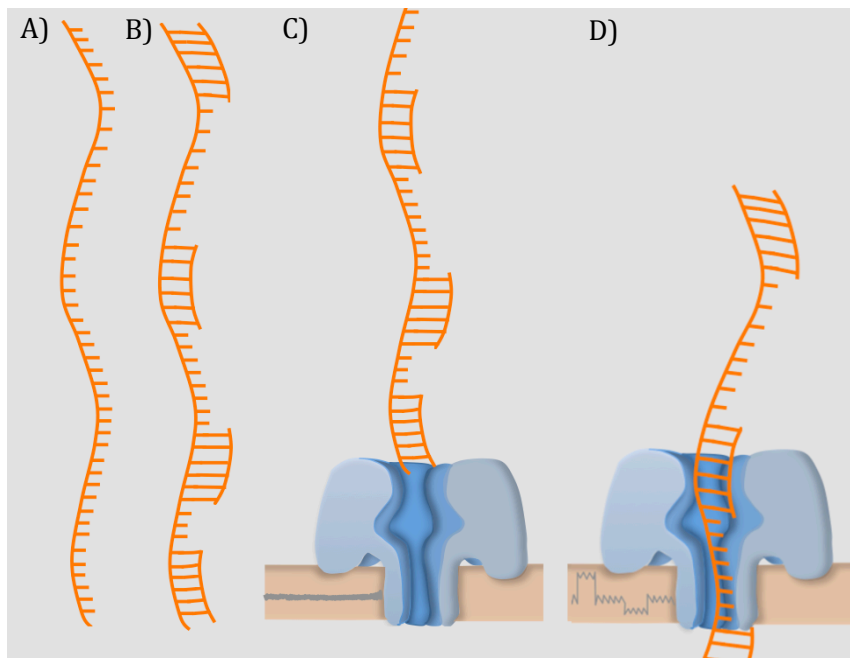


Figure 2.19. Cross-section of a nanopore embedded into a membrane. A) protocol involves denaturation of dsDNA into single-stranded molecules; B) next step involves ligation of hexamers libraries to the ssDNA to form synthetic molecule; C) prepared sequencing libraries are then loaded onto the membrane containing nanopores and, an ionic current drives semi-double stranded molecules through the pores; D) electric current is measured while the particles pass through the hole/pore. The hexamer structure is recognised by the nanopore and the basecalling software decodes the signal pattern and converts it into a readable format. Image source: own work.

This distinct signal of ssDNA-dsDNA molecules creates a map distribution of the probes. Long genomic fragments are then assembled and a probe map is created. As this whole process is prepared in parallel for the entire library of probes e.g. 256 or 4096 different probes depending on the protocol (i.e.  $4^4$  or  $8^4$ ). These maps are aligned using a bioinformatics program to generate consensus sequences (<https://youtu.be/HV0aWVrDM2U>). Nonetheless, this method contains multiple limitations; for example, probes can bind in random non-complementary regions of the ssDNA and cause errors during decoding of the probe map. Furthermore, accurate analysis of probe sites is highly dependent on the constant speed of the molecule

while passing through the nanopore. However, changes in speed while ssDNA moves through the nanochannel may affect subsequent mapping results.

NanoTag sequencing by synthesis is a method for real-time data analysis, which uses Genia's complementary metal-oxide semiconductor (CMOS) integrated circuit and NanoTag sequencing chemistry. NanoTag SBS is a nanopore-based method that uses an immobilised polymerase enzyme on top of the entry to the nanopore. As the complementary strand to the template DNA is synthesized, the nanopore measures four molecular tags made of polyethylene glycol released from modified dNTPs nucleotides. Polymerase incorporated at the *cis* entrance of the pores extends the growing strand of the sequenced library. The tag discharged from the nucleotide is transferred through the nanopore under voltage, which results in a modified current signal. Genia also proposed another way of labelling dNTPs, which involve base specific tags to enhance the signal across the pore and potentially improve accuracy. Recently published results indicate that this technology produced only 20bp reads on a synthetic template with no homopolymers (Ansorge, 2016). Similar to other types of nanopore sequencing, the current signal is basecalled and converted into standard text-based nucleotide sequence format (Wang et al., 2015). Challenges related to this technology may include immobilisation of a single polymerase on top of the pore such that the tag labels are directly captured and transferred across the nanopore membrane. Furthermore, it has been shown that assembly of the polymerase on top of the nanopore can increase the noise signal. Other types of limitations and progressed achieved by Genia are described by Ansorge et al. (2016). In 2014 Roche Ltd. acquired Genia Technologies, their platforms is still under the development. For the moment this technology is not available for clinical or research testing.

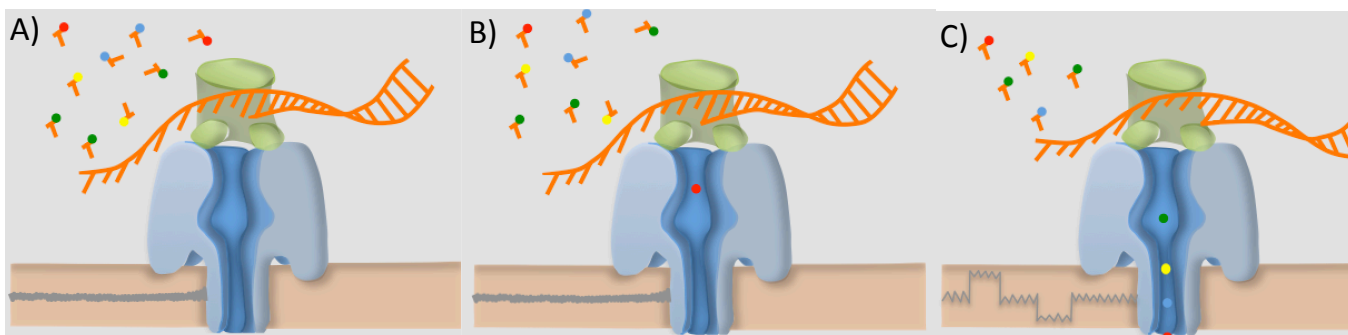


Figure 2.20. Cross-section of polymerase enzyme assembled atop a nanopore embedded into a membrane. A) first step requires immobilisation of ssDNA into polymerase protein. Reagents present in the reaction contain polyethylene glycol (PEG) modifications on dNTPs; B) reaction initiates when polymerase begins amplification, the enzyme cleaves the PEG-label of the nucleotide which then travels through the pore; C) amplification rate of the polymerase sets the speed of PEG-labels being released from the dNTPs and the whole sequencing process. Image source: own work.

### Solid-state nanopores

Strand sequencing by transverse electron tunnelling is a method that is similar to strand sequencing. However, it does not use biological molecules but moves the DNA strand through nanogate containing probes, such as an electron emitter and collector. Tunnelling is a quantum-mechanical effect, based on movement of an electronic wave. Electrons can leap across the gap between two probes separated by a few nanometres. The tunnelling process is distance-dependent and separation of probes is related to signal decay. Despite the small gap junction and the probes being fabricated with use of electron-beam-induced-deposition (EBID), the signal-to-noise ratio is high and significantly affects accuracy. Signal strength also depends on electric properties of the media in the gap and the molecule passing through. Changes in tunnelling current vary for nucleotides according to each nucleotide base of the DNA strand as the molecule is driven by electrophoretic forces (Prasongkit et al., 2015).

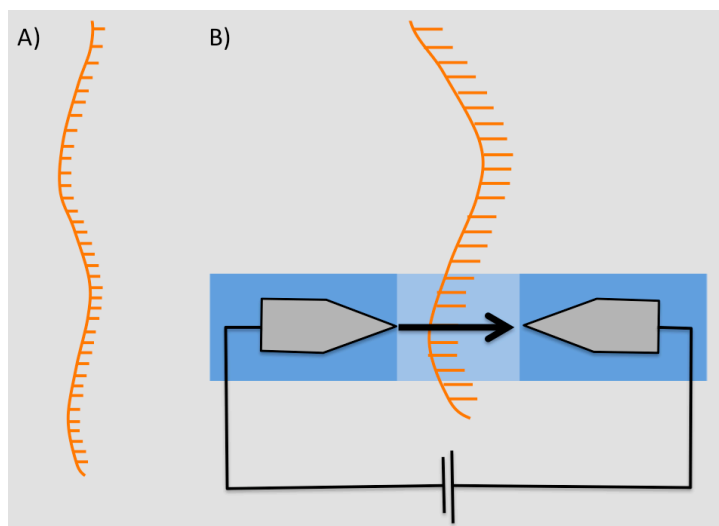


Figure 2.21 Cross-section of nanopore strand sequencing by transverse electron tunnelling. A) denaturation of dsDNA is a first step in sample preparation; B) subsequently DNA molecules of interest converted into single-stranded form are loaded onto the device. When electric current is applied, the DNA molecule moves across the tunnelling electrodes. Platinum electrodes (grey), emitter and collector immobilised into silicon nitride ( $\text{Si}_3\text{N}_4$ ) membrane (blue) detect changes in ionic current blockade and tunnelling current modulations. Image source: own work.

Sequencing By Expansion (SBX) by Stratos Genomics Inc. is being developed to improve accuracy caused by the small size of nucleotide bases (Figure 2.22), as a biological nanopore DNA sequencing method cannot easily resolve below dimensions of  $3.4\text{\AA}$ . The SBX protocol requires ssDNA molecules and random tetra or hexamer probes called X-Probes (Kokoris, 2011). A complementary strand of a synthetic dsDNA molecule is generated with use of a modified replication process where X-Probes are added and form a double-stranded helix. Each molecule from the library of X-Probe hexamers has a unique sequence that will precisely match the template DNA sequence. A tether loop containing signal reporters is attached in the middle of the X-Probe. The signal reports a mirror to the X-Probe nucleotide base sequence coding. Extension of hexamer X-Probes continue along the full length of the ssDNA molecule. Subsequently, the dsDNA hybrid is denatured and a synthetic molecule is released from the dsDNA helix. Addition of a cleaving reagent causes formation of gaps in the middle of the X-Probes in between the tether loops, which allows the



ssDNA backbone to spread into a 50-times longer molecule than the initial target DNA, called an X-Pandamer. This surrogate signalling molecule codes for the same DNA sequence as the X-Probe; however, signal reporters contain large signal-to-noise ratio. Scaling up the size of molecules and signal with use of X-Pandamers allows for high accuracy, long DNA molecule nanopore sequencing.

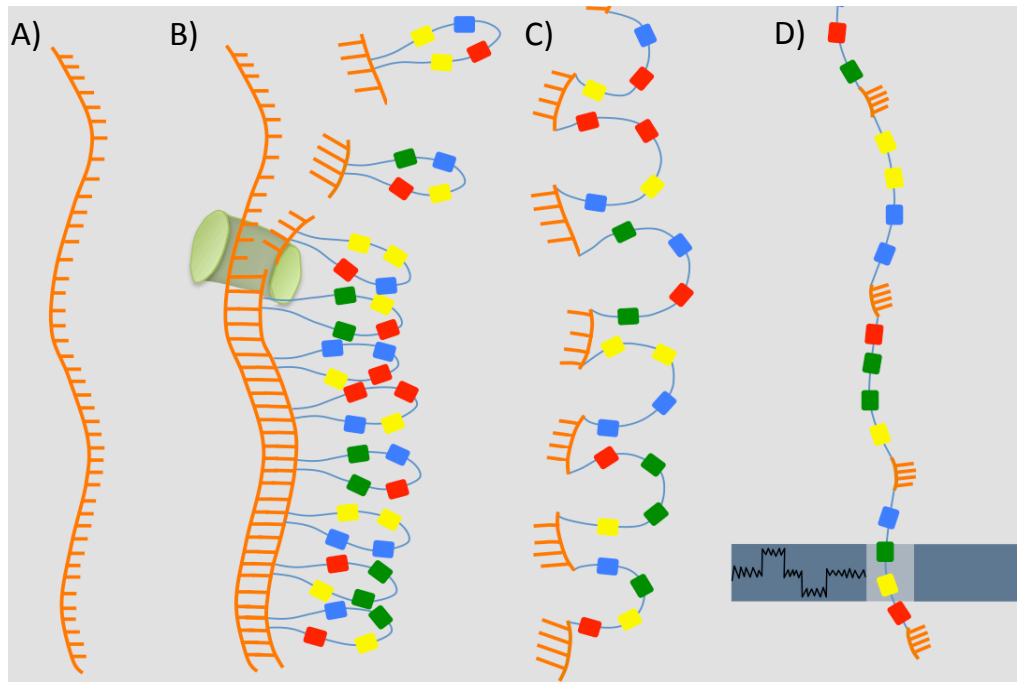


Figure 2.22. Protocol for SBX requires A) denaturation of dsDNA molecules into single-stranded nucleic acid form; B) subsequently polymerase (green) incorporates random tetramers (X-Probe) containing enhancer molecules adequate to DNA sequence at the strand of interest; C) dsDNA hybrid is denatures into single-stranded form then nuclease enzyme is added to cleave the single-stranded chain in the middle of each probe; D) sequencing libraries prepared in this way are subsequently loaded onto the flowcell, molecules move through the hole in synthetic membrane. Each nanopore contains a sensor that detects signals generated only from X-probes (enhanced signal); however, the detector ignores ssDNA fragments. That is turn increases reliability and accuracy of generated data as X-Probes improve signal-to-noise ratio. Image source: own work.

Electric forces move the long synthetic X-Pandamer molecule through the silicon nanopore chip. The SBX technology only recognises the electric current for one of four reporters, which uniquely identifies the nucleotide base. Because library preparation of X-Pandamer is based on recognition of four bases, it is not possible to use it for detection of epigenetic marks.

Sequencing By Electron Tunnelling (SBET) is based on an idea by James W. Lee and Thomas G. Thundat based at the Oak Ridge National Lab (ORNL-USA), which was published in 2003 and patented in 2005 (Lee et al., 2005). Their proposition was based on the assumption that each nucleotide base pair has its own recognisable structure and pattern signal when translocating through the gate; i.e. the nanoelectrode. In their publication Lee and Thundat theorised that each nucleotide will have considerable charge conductance that hypothetically could be detected when moving through a  $\sim 1.5$  nm hole placed between two electrodes. Moreover, SBET technology analyses single-stranded DNA without a use of proteins or enzymes, however. Another group reproduced their calculations (Zwolak et al., 2005) and speculated that the sequencing speed could reach up to  $10^6$ - $10^7$  bases per second (Lagerqvist et al., 2006), which attracted more scientists and groups to pursue a new technology and a commercial device. Nonetheless, fabrication of such a device has been very challenging due to the tiny gap between electrodes required for DNA translocation, which is at a nanometer or even Angstrom scale. For this reason there is no available technology for sequencing by electronic tunnelling that has been made available so far, however, generation of such a technology could out perform protein based nanopore sequencing due to single nucleotide resolution.

Direct tunnelling is a method, which allows inserting nanopores into a 10-nm thick chromium nanoelectrode and 2-50 nm gap. Other groups have shown that gold electrode spacing can be controlled in a range of 0.5 – 10nm when mechanically controllable break-junction method is applied. Moreover, in 2010 scientists from Osaka University, Japan published results indicating successful development of a 1nm gap based on gold electrodes (Tsutsui et al., 2010) and were able to distinguish

between three of the nucleotide base pairs, but not adenosine. Their research indicated possible recognition of epigenetical marks present on the DNA molecules; i.e. 5-mdC and 8-oxo-deoxyguanosine (8-oxo-dG). Further developments in the technology allowed for ssDNA and RNA sequencing but also establishment of a Quantum Biosystems in 2014 (Karow, 2014)

In 2005, scientists from University of Tokyo, Japan published their results on DNA tunnelling with use of a gold surface tip and a scanning tunnelling microscope (STM), also called Hydrogen based-mediated nanopore tunnelling (Xu et al., 2005). Two years later, they reported that four nucleotides and their epigenetic modifications have surface specific conductance and can be recognised despite their signal overlapping (Xu et al., 2007). At the same time, another group from the University of Tokyo reported that STM with thiol can be used for recognition of hydrogen bonds, which in turn allows for identification of the DNA sequence (Ohshiro et al., 2006). Further work allowed for generation of nanogaps between electrode tips with an Angstrom precision (Chang et al., 2011).

	Biological			Synthetic					Other
Pore type	Protein tunnels e.g. $\alpha$ -Hemolysin, MspA, Phi29 or CsGg			Solid-state nanopores are made as a hole in a synthetic substrate					Biomimetic or hybrid nanopores e.g. $\alpha$ -Hemolysin embedded into silicon nitride membrane
Diameter of channel	$\alpha$ -Hemolysin 1.4–2.6nm MspA 1.2nm Phi29 3.6–6nm			Varies but can be precisely controlled in nm					
Depth of channel	$\alpha$ -Hemolysin 5.2nm MspA 3.7nm Phi29 7nm			Depends on the thickness of the membrane 0.3-60nm					
Membrane	Lipid bilayer, polymer			Si3N4, SiO2, Al2O3, MoS2, Al2O3A, Al2O3, BN or Graphene				Glass or quartz	
Fabrication method	Self-assembly			Ion beam sculpting, ion milling track-etch method, e-beam drilling, atomic layer deposition, TEM or FIB				Laser heating and pooling	
Type of sequencing protocol	<u>Direct strand tunnelling</u>	<u>Exonuclease assisted nanopore</u>	<u>Nano Tag Sequencing by Synthesis</u>	<u>Transverse electronic transport</u>	<u>Sequencing by expansion</u>	<u>Optical nanopore (Optipore)</u>	<u>Hydrogen bond-mediated tunnelling</u>	<u>Hybridization-assisted pore-based</u>	
								Nanopipette	

Table 2.23 Detailed representation of biological, synthetic and other types of nanopores, their size, type of membranes, fabrication methods, sequencing protocols, type of molecules being detected, necessity for amplification, amount of bases creating signal, detection of epigenetic marks, homopolymer errors, signal detection and translocation forces. Different types of nanopores possess various advantageous or disadvantageous features so, depending on application i.e. lower or higher error rate.

References	Translocation forces	Signal detection	Homopolymer errors	Detection of epigenetic marks	Amount of bases creating signal	Requirement for amplification	Detecting molecules
Branton, 2008; Wang 2013	Electrophoretic like-forces, pressure, optical and magnetic tweezers, tethered probes or hydrophobic adsorption of membrane molecules e.g. graphene	Changes in ionic current caused by translocation of molecule chain	YES	YES	Multiple	NO	Potential for various
Branton, 2008; Reiner, 2012		Changes in ionic current caused by translocation of single nucleotide	NO	YES	Single	NO	DNA, RNA and proteins
Kumar, 2012		Changes in ionic current caused by translocation of molecular tag	NO	NO	Single	YES	DNA, RNA only
Lagerqvist, 2006; Haque 2013		Changes in electric current caused by direct translocation of ssDNA	YES	YES	Single	NO	Potential for various
Gierhart 2008; Kokoris, 2011; Ivanov, 2011		Changes in electric current caused by translocation of Xpandomers	NO	NO	Single	YES	DNA, RNA only
Huang, 2015; McNally, 2010; Jonsson, 2012		Recognition of pulled out fluorescently labelled beacons	YES	NO	Multiple	YES	DNA, RNA only
Chang, 2009; Zwolak, 2012; Maitra, 2012		Signal is created through interaction of hydrogen bonds with DNA bases	YES	YES	Single	NO	Nucleic acid only
Wash, 2008		Changes in ionic current caused by translocation of DNA-probe hybrid	YES	NO	Multiple	YES	Nucleic acid only
Karhanek, 2004; Fu, 2009; Villozny, 2011; Bell 2013		Changes in ionic current caused by translocation of ssDNA	YES	YES	Multiple	NO	Potential for various
Banerjee, 2010; Feng, 2015; Hall, 2010		Molecular nanopores for signal desoising e.g. cyclodextrin (CD); ssDNA-rotaxane library moves through $\beta$ -CD sugar units					

The aforementioned technologies demonstrate a solid track record of improvements being made in the field of nanopore sequencing. Yet challenges with high error rates, complex library preparation, complex fabrication technologies, low amount of generated data and high running costs remain. Figure 2.23 includes a detailed comparison (i.e. advantages and disadvantages) of the above-mentioned technologies.

### **Recent applications of the nanopore**

One of the most advantageous features of nanopore technology is the generation of long (1kb – 99kb), (Storm et al., 2005) and ultra-long reads with >100kb range (Urban, 2015). This makes it highly attractive for assembly and completion of genome drafts of various organisms like bacteria, yeast and algae (Karlsson et al., 2015; Istace et al., 2016; David et al., 2016). Initially, the use of high quality reads (2D) was necessary for accurate, low error rate results (Loman et al., 2015). Due to technological developments and a decrease in nanopore error rates, low quality reads (1D) can also be used for accurate assembly of single genomes. However, the use of lower quality reads may require more data as compared to high quality reads. On the other hand, hybrid assembly uses long nanopore reads and high accuracy, high amount of short Illumina reads (Sovic et al., 2016). This approach in turn allows for very high coverage of the genome with short reads but also allows to resolve the problem of homopolymers and repetitive regions where short read assemblies would fail (Pop et al., 2004). Nanopore sequencing technology has also been used for identification of bacterial pathogens and antibiotic resistance genes from clinical samples (Schmidt et al., 2016). Due to the miniaturised size of the device, other applications of nanopore sequencing include field deployment for rapid detection of pathogens (Quick et al., 2015). An example for this is the detection and

characterisation of the *Ebola* virus in Mid-West Africa and investigation of the *Zika* virus in Brazil (Quick et al., 2017). Furthermore, ONT proposed a protocol for direct analysis of RNA molecules. Lack of reverse a transcription step on ssRNA particles will allow not only for transcriptomic but also for epi-transcriptomics analysis of eukaryotic organisms (Garalde et al, 2016). Analysis of post-transcriptional modifications may give us better understanding of various genetic diseases and protein related abnormalities. Most recently, scientists have become more interested in using nanopores for analysis of protein structures (Nivala, 2013). This technique involves use of unfoldase (ClpXP), an enzyme for linearization of secondary and tertiary protein structures. Subsequently, denatured proteins are moved through the pore similar to the nucleic acid chain. This step generates specific pattern changes in electrical current that later can be used for basecalling of the amino acids and then polypeptide chain structure and further analysis.

#### Automation of nanopore biosensing analysis

An advantageous feature of the nanopore technology is the miniaturised size of the device, which can provide for field or office based sample analysis, depending on the type of sequencing protocol (Edwards et al., 2016). However, to achieve complete portability, the equipment required for collection and preparation of sequencing libraries also have to be miniaturised. Use of liquid handling robots is still widely applied in laboratories for full- or semi- automation of various protocols (Meldrum, 2000; Wu et al., 2003). Nonetheless, liquid handling robots require a constant supply of electricity, which would not allow for field deployment. Moreover, size and weight of these robots is large and not suitable for easy transportation. One way of achieving portability is use of manual or semi-automated devices that could have rechargeable

batteries combined with a solar panel. Such instruments could be used for sample collection and later nucleic acid extraction by heat treatment or bead beating. An example of a semi-automated device that has a great potential for portability may be the PureLyse® Bacterial gDNA Extraction Kit (ClaremontBio Solutions). This system uses a disposable micro-scale mechanical ultra-rapid lysis kit for gDNA extraction with a very simple design. TerraLyzer™ from Zymo Research is another device that could be used for in-field or point of care application. This instrument is a hand-held, cordless device that was designed for rapid vigorous in-field sample disruption for in-field DNA extraction (Urbina, 2016). Another device required for the DNA extraction is a portable centrifuge. Various scientists have developed multiple types of techniques: one of the simplest centrifuge-like devices is modification of a cordless drill with a microcentrifuge tubes adapter. Vendors such as Bento Lab have developed extensive portable sample preparation capabilities. Their instruments contain multiple apparatus like: centrifuge, PCR thermocycle, gel electrophoresis and transilluminator. Nonetheless, none of the aforementioned devices are fully comprehensive, for complete end-to-end sample collection and preparation.

One of the promising systems that would allow for complete and comprehensive mobile end-to-end sample preparation is use of microfluidics technology (Kim et al., 2013). This uses small quantities of reagents and samples to perform amplification, detection or separation at high accuracy and sensitivity. This could allow for low cost of reagents, faster analysis, reproducibility and small size of the device (Whitesides, 2006). Automation and miniaturization of various biological and chemical processes for autonomous sample collection and analysis would significantly simplify the



process and allow for various environmental and clinical applications on earth but also in space.

## 2.5 Future of DNA Analysis and Biosensing

Development of modern technologies in the past decades allowed for miniaturisation of computers, expansion of electronics and successful automation of laboratory workflows. That in turn gives immense opportunity for development of innovative technologies in areas such as clinical, pharmaceutical, agriculture, military, environmental but also food and water safety. Each one of these aforementioned fields contains various obstacles that in turn create more challenges for development of single biosensing device applicable in all of the areas. Various approaches have been evaluated for improvement of biosensing assays; e.g. optical sensing such as fluorescence, luminescence, absorbance for immunological sensing based on complementarity of an antibodies to an antigen; molecular biosensing which allows for detection of nucleic acids; amperometric sensing, which detects ions based on electric currents and others such as piezoelectric sensors (Peng et al., 2011). Despite great technological developments and increasing demand, there is a limited amount of comprehensive biosensor devices available on the market. Currently available biosensors suffer from multiple aspects such as short-life times due to consumption of reagents, limitation to few sterilisation methods, responsive to small amount of analytes, requirement of frequent calibration and maintenance due to biofouling, debris and sediments aggregation. Because of the aforementioned reasons deployment of instruments for long-term environmental monitoring is difficult and instruments that are meant to work months or weeks produce high quality data for only few days.

What is a biosensor?

Biosensor is a self-contained integrated device that typically but not always (exceptions may include physical thermometer, blood pressure cuff or medical stethoscope) consisting of hardware and software. A biosensing device should allow for presence/absence detection but also quantitative or semi-quantitative analysis of various objects or substances and transform their biological, chemical or physical signals into measurable and readable digital signal format for diagnostic purposes (Cornell et al., 1997; Keusgen, 2002). The device is usually built out of multiple elements such as a recognition component called a bioreceptor, which may include a selectively permeable membrane that recognises certain types of molecules or a group of molecules of interest (chemicals, microorganisms etc.). A common example includes portable glucose biosensors (Wang, 2006; Turner, 2013). Another element is a transducer that converts biochemical event into an electrical signal. The most common types of transducers are: electrochemical, optical, electronic, gravimetric, pyroelectric and piezoelectric (Thévenot, 2001).

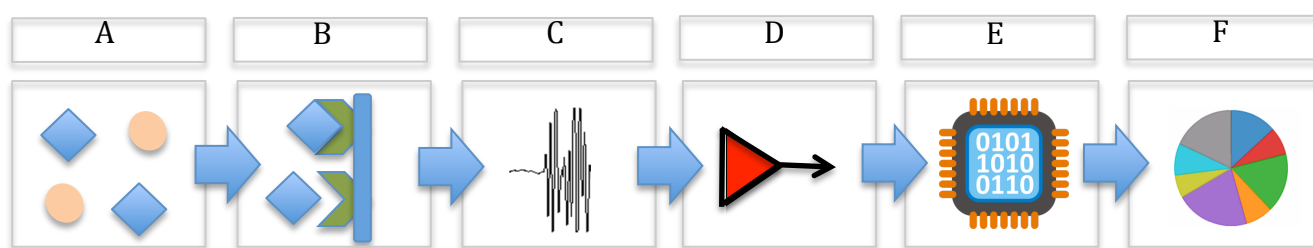


Figure 2.24 Schematic representation of a typical biosensor. A) complex mixture of analyte sample is added to onto the device; B) bioreceptor made of proteins, enzymes or antibodies detects molecule or molecules of interest; C) physical transducer converts physical, optical or chemical signals into electric traces; D) subsequently amplification of the signal and denoising step takes place; E) processor; F) generates results that are displayed in the form of plots or raw numbers, which subsequently can be processed according to standard operating procedures. Image source: own work.

The subsequent element of a biosensor device is amplification that is responsible for increase of a signal with an external power supply (Redwood, 1961; Chang et al., 2016). The signal processor is a component of the device that filters and interprets the electric signals into readable, meaningful values (Thévenot, 2001). The last element is a display that will project previously generated values for results interpretation.

General requirements for a comprehensive biosensor:

- Speed-to-result: rapid detection of environmental markers is key for early detection, prevention and response to various chemical contaminants, pathological organisms but also the chance for real-time monitoring of changes in the habitat. For this reason a biosensing device should generate results relatively fast so collected samples and produced data could be rapidly interpreted and action applied to solve encountered problems.
- Sensitivity: high sensitivity and low limit of detection are crucial requirements in development of biosensors. Demand for devices that allow for detection of very low concentrations of chemicals, small molecules or dormant microorganisms would allow for faster response and avoid false negative results.
- Accuracy: all electronical devices produce a 'noise'; that is introduction of an undesired signal to the system from the inner or outer disturbance of the device. Interruptive signals are superimposed onto the authentically generated signal and eventually decreases accuracy of the signal level. There is a need for filters which would reduce the generated noise and maximise the signal to noise ratio. Denoised data would then be available for normalisation according to the noise of the sensor in turn, results would be more centred on biological signal of interest. Accuracy of biosensing is also very important in the case of false positives or false negatives

- Reliability: biosensor devices should be easy in operation so could be used by non-specialists to deliver results in decentralised locations such as home or even in the field.
- Output: results should be simplified to the point where they are recognisable by specialist and non-specialist users. This would allow to shorten training times for the device but also to eliminate potential mistakes during sample analysis.
- Coverage: low price and simplicity of the biosensor should allow for vast deployment of the technology, as the more places where we can put the tool, the more relevant results we could generate about the environmental system.

#### What needs to be sensed?

Drinking water compounds, which should be accurately monitored, include substances that can be harmful to human health and cause irritation, sensitisation, carcinogenicity or poisonous (McCarthy et al., 1989; Rodriguez-Mozaz et al., 2004).

- An organic chemical is a broad group of compounds containing at least one carbon forming its base. Examples of such organic substances may include living matter such as plants, animals, methane or DNA. Organic chemicals can be characterised into four groups: synthetic organic chemicals (e.g. pesticides and herbicides), disinfection by-products, polychlorinated biphenyls (PCBs) and commercial and industrial organics, also described as volatile organic chemicals (Mohan et al., 2014).

- An inorganic chemical is another group of pollutants present in the aquatic environment that include heavy metals such as arsenic, boron, fluoride. These materials are very often naturally occurring metals however, remaining ones such as chromium, mercury, cyanide are introduced by industrial, agricultural and domestic waste disposal (Moore, 2012). Moreover, distribution of drinking water may

contaminate it with inorganic substances such as copper, iron, lead or zinc. Industrial, domestic and agricultural wastes pollute surface water in ponds, rivers or lakes; however, naturally occurring inorganic substances contaminate groundwater (Järup, 2003).

- Pathogenic microorganisms such as bacteria, viruses, amoeba and worms present in drinking water cause concerns when considering organisms harmful to human health. These disease-causing organisms are responsible for: intestinal infections, Hepatitis A, dysentery, Cholera, Amoebiasis, Cryptosporidiosis, Poliomyelitis. Primary source of pathogens in water come from human and animal wastes, pastures, runoff from feedlots, sewage, treatment facilities but also can be introduced by insects or rodents entering water supplies (McFeters, 1990; Buse et al., 2014).

- Radiological contaminants can be present or deposited as radionuclide substances in liquids, gasses, on surfaces or within solids. This type of contamination is typically a result of an undesired event when producing or using radioisotopes with unstable nuclei that are vulnerable to radioactive decay. Threats related to radiological contaminants are dependent on concentration of these materials, type of radiation, amount of energy being emitted and proximity of the radioactive particles to the organs in the body. That in turn can cause damage to vital molecules, DNA, cells or even tissues or organs. Examples can include: mutation at the DNA level causing cancers, leukaemia, immunological, endocrine system disorders or congenital disabilities (Lytle et al., 2014).

The majority of these aforementioned substances present in water cannot be detected by sight, taste or smell. For this reason, drinking water has to be regularly tested for presence of all of these substances. Modern technological developments are mainly

focused on detection and analysis of human genetic disorders or microbial diseases. Nonetheless, often devices developed for hospital or clinical researches are not suitable for environmental or in-field analysis (Goldman et al. 1996). Comprehensive biosensing technology is not yet very well established for routine monitoring of broad environmental factors. However, demand for such a technology steadily increases. Focus on rapid detection of environmental hazards could prevent spread of human threats. For example, routine biosensing analysis of drinking water could prevent spread of water-borne *Zika* viruses in Brazil and avoid developing encephalopathy in thousands of newborn babies or allow for monitoring of recycled drinking water at the board of the International Space Station (ISS), (Castro-Wallace et al., 2017). Moreover, biosensing of food could prevent various human diseases such as *Salmonellosis* or more severe infections causing thousands of deaths such as *Ebola* virus, which is well known to be transmitted by eating meat from wild animals such as bats or monkeys (Hoenen et al., 2016). Prevention of infection by monitoring of the environment should be a priority instead of identifying already developing diseases. That approach could reduce the amount of infected or hospitalised patients and by that decrease amount of used antibiotics or other expensive drugs (Quick et al., 2016).

The natural habitat is a complicated structure containing hundreds of different factors that continuously keep changing at various rates. Some of them change multiple times a day while others much slower. Use of advanced, novel technological systems containing multiple biosensors; i.e. temperature, pH, salinity, drugs, proteins, microbial detection. would shorten and simplify analysis of the samples. Rapid analysis of environmental changes could be applied by introduction of Real-Time Data Monitoring (RTDM) technology based on electronic biosensors. That would allow for fast generation of “big data” from multiple sites without interruption or

delays in measurements. The biosensing process could be used for monitoring of various pollutants, detection of contaminant sources, rapid response to them and administration of environmental contaminants. Information gathered during monitoring of environmental systems such as water, soil or air could be automatically streamed and analysed with mathematical models or compared to historical data. Moreover, introduction of wireless communication from the environmental biosensor would allow for the results to be immediately shared with consultant scientists, groups of researches or even general society via web-based tracking systems or mobile software applications. Furthermore, electronically stored data could be used for long-term monitoring of the environment and allow for recognition of various patterns and motifs in the examined surroundings. This way of research could provide accurate examination of multiple factors over long-time but also a broader picture that would allow for better understanding of discrete changes. Results obtained from these observations could be directly applied to protect the ecosystem and indirectly human health. For example, by monitoring the amount of pathogenic microorganisms present in lakes or rivers or airborne organisms such as anthrax spores, rapid detection of biohazardous organisms. Sensor affordability would allow for vast deployment of devices into tens or hundreds of locations and real-time monitoring of numerous environmental factors. Long-term understanding of behavioural patterns from various environmental points and systems could lead to development of simulation algorithms that in turn would allow forecasting forthcoming behaviour of the habitat or climate system.

### 3 Evaluation of Nanopore Technology for Amplicon Sequencing

#### 3.1 Abstract

Remote and real-time analysis of bacterial organisms is highly desirable for rapid diagnostic applications to detect environmentally and clinically relevant microorganisms. While multiple protocols have been developed over the years for this purpose, none of them provide the opportunity for comprehensive in-field characterization of microbes. This chapter describes our efforts to test the MinION™ device, a miniaturised nanopore-based sequencing device, for characterisation of microbial communities using a mock community generated using full-length 16S rRNA gene amplicons (~1.4kb). Experiments were designed to test the same mock community consisting of fifteen 16S rRNA gene amplicons in triplicate and three molecular barcodes (Fig. 3.1). Detailed analysis of data with use of four different read aligners allowed us to evaluate the run-to-run reproducibility of the sequencing platform and its sequencing error rates (Fig. 3.2).

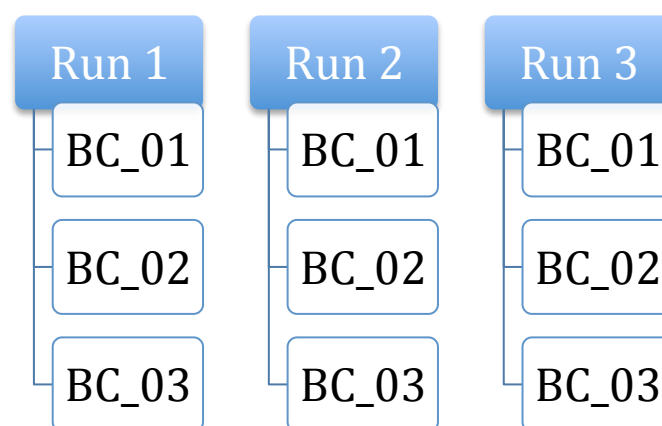


Figure 3.1 Graph represents three sequencing runs (Run1, Run2 and Run 3) containing three barcoded (BC\_01, BC\_02 and BC\_03) mock samples, 15 organisms in each.



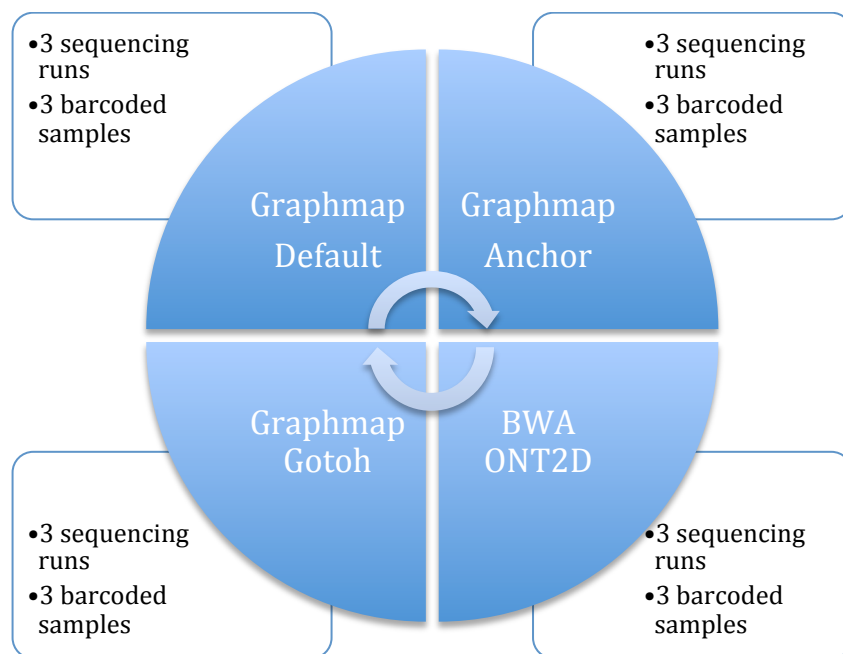


Figure 3.2 Plot outlining the strategy for data analysis: three sequencing runs (Run 1, Run 2 and Run 3) containing three barcoded (BC\_01, BC\_02 and BC\_03) samples. All these samples were analysed with four aligners (Graphmap Default, Graphmap Anchor, Graphmap Gotoh and BWA ONT2D) and results were compared against each other.

We also evaluated the relationship between mapping quality scores and read base quality for all four mappers. Moreover, we estimated diversity indexes such as Shannon, Simpson and Pielou's evenness to determine the impact of the sequencing process on deviation from theoretical community metrics. Additionally, we tested fidelity in recovering the theoretical microbial community using evenly distributed microbial community structure and membership-based metrics. Results indicate a high level of reproducibility across sequencing runs in recovering the microbial community structure. However, the MinION<sup>TM</sup> sequencing platform suffers from high error rates, which compromises its ability for reference-based and reference-free data analyses. Our results indicate that direct amplicon sequencing using the MinION<sup>TM</sup> device is limited in its application due to high these sequence error rates. Results presented in this chapter were crucial towards shaping further work undertaken during this PhD research project that are described in subsequent chapters.

Results of this chapter were presented at the conferences:

Calus S.T., I.U.Z. and P.A.J., (2016). MinION-enabled and customer-led drinking water quality monitoring for pathogen detection. Conference: Achieving Zero Bacteriological failures in Water Supply Systems, Glasgow, UK. [Oral Presentation]

Calus S.T., I.U.Z. and P.A.J., (2016). Evaluation of multiple DNA aligners for the analysis of full-length 16S rRNA gene from mixed microbial communities from the MinION nanopore-based sequencing technology. Conference: American Society of Microbiology – Microbe, Boston, USA. [Poster Presentation]

Calus S.T., I.U.Z. and P.A.J., (2016). Evaluation of multiple DNA aligners for the analysis of full-length 16S rRNA gene sequences from mixed microbial communities using the MinION nanopore-based sequencing technology. Conference: Microbial Ecology and Water Engineering, Copenhagen, Denmark. [Oral Presentation]

### 3.2 Introduction

Amplicon sequencing of marker genes, particularly the bacterial and archaeal 16S rRNA gene, is a widely used method for profiling microbial community structure and membership across a range of different sample types, from clinical to environmental (Zhou, 2011). The wide-scale application of amplicon sequencing has been driven by the ability to multiplex 10-100's of samples on a single sequencing run and obtain in-depth coverage of the target community by sequencing tens of thousands to millions of reads per sample (Caporaso et al., 2012). The most commonly used platforms for marker gene sequencing currently include Illumina MiSeq (Caporaso et al., 2012; Salipante et al., 2014; Zhou et al., 2011), Ion Torrent (Whiteley et al., 2012; Daum et al., 2012) and 454-pyrosequencing (Armougom et al., 2009; Griffen et al., 2012). Though powerful, all of these sequencing approaches are limited in the read lengths

that can be sequenced. The aforementioned three methods can sequence amplicons ranging from 300-400 bp (for Illumina MiSeq and Ion Torrent) to 700-800 bp (for 454-pyrosequencing) (Loman et al., 2012). Lack of the full-length sequence of the marker gene, in particular the 16S rRNA gene, restricts our ability to either confidently classify the microorganisms detected beyond the family or genus level or prevents us from differentiating between sequences that arise from closely related organisms (i.e., species) due to highly similar partial gene sequence (Eren et al., 2013). Recently, the development of single-molecule real-time sequencing (SMRT) on the PacBio platform (Gupta, 2008; Greenleaf, 2014) and single molecule sensing technologies on the Oxford Nanopore Technologies (ONT) MinION™ platform (Eisenstein, 2012; Goodwin, 2015) have opened the possibility of obtaining ultra-long reads, with reported maximum read length in excess of 100kb (Urban et al., 2015). Of the two platforms, ONT's MinION™ device is particularly attractive due to its miniaturised design, allowing its use in a conventional microbiological laboratory or even directly in the field (Quick et al., 2016). Additionally, in contrast to all existing platforms, the MinION™ is not a DNA sequencer, where the DNA sequence is determined through the incorporation of nucleotides (often fluorescently labelled) in the molecular process of polymerase chain reaction. Rather, it is a DNA sensor, where the MinION™ device measures the modulation of ionic current passing through the nanopore while a DNA molecule is being translocated through each nanopore (Hoenen et al., 2016). The current signals can then be decoded to obtain nucleotide sequences of very long DNA molecules, over 1Mb (Jain et al., 2018). The current version of the MinION™ sequencing device was made available through the MAP (MinION™ Access Program) but the associated chemistry (R7.3 and SQK-MAP005) was characterised by a multiple limitations e.g. data quality and limited

output. Specifically, it requires high levels of DNA (>1ug), suffers from maintaining stable activity of nanopores for the duration of the sequencing run, and provides reads with relatively high error rates compared to currently available short-read sequencers (Karlsson et al., 2015). Nonetheless, the ability for real-time sequencing, run-time flexibility, and long reads have already made a significant impact on DNA sequencing based studies; primarily related to whole genome assembly of pure cultures of organisms. For example, data from the MinION<sup>TM</sup> device has been used to produce single contig *Escherichia Coli* assemblies (Loman et al., 2015), and hybrid assemblies involving nanopore and Illumina reads of *Saccharomyces cerevisiae* (Goodwin et al., 2015). It has also been used for determination of antibiotic resistance chromosomal cassette re-arrangements (Ashton et al., 2015), reconstruction of short viral genomes (Kilianski et al., 2015) and rapid detection of Salmonella outbreaks (Quick et al., 2015). However, results of the aforementioned studies describing nanopore data very often avoided detailed description of the sequencing errors and were designed to test single organisms i.e., *E.coli*. The initial goal of this experimental chapter was to evaluate the accuracy of the current version of the MinION MK1b device and associated chemistry (2D and 1D<sup>2</sup>). Secondary objective was to reconstruct full-length 16S rRNA sequences and assess the effect of multiple data processing algorithms on its ability to reconstruct the community structure using a simple evenly distributed mock-community consisting of 15 organisms. Conclusions described in this chapter meant to clarify whether accurate analysis of complex microbial samples is possible to obtain with MinION MK1b.

### 3.3 Methods and Analysis

#### *The MinION™ Access Programme*

We enrolled into the MinION™ Access Programme (MAP) organised by Oxford NanoPore Technologies (ONT) in January 2015 and were provided with the MinION™ device, two sequencing flow cells (R.73) and reagents for standard library preparation (MAP-SQK005).

#### *Mock community construction*

The mock community consisted of 15 near full-length 16S rRNA genes (~1.5kb) obtained from a previous experimental clone library from an anaerobic digester (Connelly et al., 2017) and included: 1. *Lactobacillus harbinensis* (B2); 2. *Listeria monocytogenes* (B3); 3. *Streptococcus castoreus* (B12); 4. *Flectobacillus major* (B17); 5. *Acinetobacter baumannii* (B22); 6. *Moorella thermoacetica* (B20); 7. *Desulfosporosinus orientis* (B29); 8. *Pseudomonas\_aeruginosa* (B35); 9. *Meiothermus\_ruber* (B40); 10. *Neisseria\_meningitidis* (B45); 11. *Aquimarina intermedia* (B54); 12. *Bacteroides vulgatus* (B63); 13. *Moorella thermoacetica* (B74); 14. *Propionibacterium acne* (B77); 15. *Beijerinckia indica subsp. Indica* (B83). This set of organisms was chosen for analysis mainly due to genomic variability i.e. closely and distantly related organisms, which mean to represent true environmental sample. Amplified and purified ribosomal genes of interest were ligated into 15 separate pCR4-TOPO plasmids (Thermo Fischer Scientific, K457502) and transformed into competent *E. coli* cells according to manufacturer instructions. Cells containing plasmids were cultured overnight at 35°C in 10 ml of LB Broth media. Overnight cultures were centrifuged for 5 minutes at 5000rpm, the supernatant was discarded, and the pellet was resuspended in phosphate buffered saline (PBS) media

before plasmid extraction using the Qiagen MiniPrep Kits (Qiagen, 27106) according to the manufacturer's instructions. Extracted plasmids were quantified using the Qubit dsDNA HS (Life technologies, Q32854) and analysed on 1% agarose gel to verify the quality of plasmids. Extracted plasmids containing 15 different near full-length 16S rRNA clones were submitted for Sanger sequencing at Edinburgh Genomics, UK and sequenced using four universal primers: 8F: AGRGTTTGATCMTGGCTCAG, 518R: ATTACCGCGGCTGCTGG, 1100F: CAACGAGCGCAACCCT and 1387R: GGGCGGWGTGTACAAGRC, (Marchesi et al., 1998; Dorn-in et al., 2015). The nanopore mock community was constructed by amplification of the near full-length 16S rRNA gene using primers 8F and 1387R, with the forward primer fused to barcodes provided by ONT which are compatible with the Metrichor Agent basecaller software. Each 16S rRNA gene was independently PCR amplified from the plasmid using the primers mentioned above at a final concentration of 0.5 $\mu$ M, template plasmid (0.1ng) and Q5 Master Mix polymerase (New England BioLabs Inc., M0492L) with a PCR reaction volume of 25  $\mu$ l. The target gene was amplified using the following thermocycling conditions: initial denaturation at 98°C for 30 seconds, followed by 30 cycles of 98°C for 10 seconds, 57°C for 25 seconds, and 72°C for 45 seconds, followed by a 2-minute final extension at 72°C. The PCR product from each reaction was visualised by 1% agarose gel electrophoresis to ensure the presence of a target amplicon of approximately 1.3-1.4kb. Subsequently, amplicons were cleaned-up to remove primer dimers and potential unwanted short amplicons using High-Prep PCR (MagBio, AC-60050). The process was carried out according to the manufacturer's protocol and 0.45x concentration of magnetic beads. Purified 16S rRNA gene amplicons were eluted into 50 $\mu$ l of 10mM Tris-HCl pH 8.5, and once more 1 $\mu$ l of liquid was analysed by gel electrophoresis to ensure that all short

fragments had been removed and amplicons of interest were still present. Each 16S rRNA gene template was quantified using Qubit dsDNA HS kit and normalised to 20ng/μl with 10mM Tris-HCl pH 8.5. Normalised products for all 15, 16S rRNA templates were then combined in equimolar proportions (even distribution) into a single pool to construct the mock community used for MinION™ library preparation. The simulated communities were independently built by repeating all steps post-plasmid extraction, and each of the below outlined library preparation, flow-cell preparation, and sequencing steps were carried out separately for each sequencing run.

#### *Preparation of sequencing libraries*

The mock community amplicon pool was subjected to end-repair using NEBNext End Repair Module (NEB, E6050L). Briefly, 85μl of the amplicon sequence pool was mixed with 10μl NEBNext End Repair Reaction Buffer (10x), and 5μl NEBNext End Repair Enzyme Mix (NEB, E6050L) and incubated at room temperature for 30min. The end-repaired amplicon pool was cleaned using High-Prep PCR (MagBio, AC-60050) and eluted in 25μl of 10mM Tris-HCl (pH 8.5). Subsequently, the end-repaired amplicon pool was Adenosine-tailed using dA-Tailing Module (NEB, E6053L) according to ONT recommendations as follows: 25μl of the end-repaired amplicon pool was mixed with 3μl of NEBNext dA-Tailing reaction buffer (10x) and 2μl of Klenow Fragment (3'→ 5' exo-) then the reaction mix was incubated for 30min at 37°C. The dA-Tailed amplicon pool was purified using High-Prep PCR (MagBio, AC-60050) and eluted into 32μl of 10mM Tris-HCl (pH 8.5), quantified using the Qubit dsDNA HS kit and transferred to a Protein LoBind tube (Eppendorf, 925000092). Subsequently, 250ng of the dA-tailed amplicon pool (0.2pmol) was

combined with 10µl Adapter Mix (ONT, SQK-MAP005), 10µl hairpin (HP) adapter (ONT, SQK-MAP005) and 50µl 2x Blunt/TA Ligase Master Mix (NEB, M0367L). The reaction was briefly vortexed and incubated for 30min at room temperature, according to ONT instructions. At the same time, 550µl of 2x Wash Buffer (ONT, SQK-MAP005) was diluted with 550µl of ultrapure water (Roche Ltd., 03315843001), generating 1100µl of 1x Wash Buffer. The diluted buffer was mixed 10 times by inverting the tube then briefly spun down for 5s in a microcentrifuge. Second, Dynabeads® His-Tag Isolation and Pulldown (Life Tech., 10104D) were prepared by vortexing at maximum speed for >2min until beads were homogeneously mixed. Subsequently, 10µl of re-suspended His-Tag beads were transferred to a 1.5ml Protein LoBind Tube (Eppendorf, E0030108116) and diluted with 250µl of 1x Wash Buffer. The mix was gently pipetted up and down to re-suspend beads and placed on a magnetic rack for 2-3 min to allow separation of the beads. The supernatant was discarded with a sterile pipette tip and beads were washed a second time with 250µl of 1x Wash Buffer. Double-washed and pelleted His-Tag beads were resuspended in 100µl of concentrated 2x Wash Buffer. The double washed His-Tag beads (100µl) were then added to the amplicon pool after initial incubation of the DNA with the adapter mix. Reagents were gently pipetted to homogenise magnetic beads and incubated for 5 minutes at 22°C. The post-incubation amplicon pool was placed on a magnetic rack for 10min to pellet the magnetic beads. Subsequently, the supernatant was discarded, and the beads were resuspended in 25µl of Elution Buffer and left for 10min before pelleting a second time, again using the magnetic rack. The liquid containing sequencing libraries was transferred into a fresh Protein Lo-Bind tube and left on the magnetic rack for another 10min to remove leftover His-Tag beads. The



final supernatant (i.e., the pre-sequencing mix) was transferred into a new Protein Lo-Bind tube and stored at 4°C until processed for sequencing.

### *Flow cell preparation*

On arrival, the sequencing flow cells (ver. R7.3) were stored at 10°C for no more than a week to minimize deterioration of biological nanopores, which would reduce the quantity of output data. While preparation of the pre-sequencing mix was completed, the single flow cell was removed from the fridge, and a new heat pad sticker was applied to the bottom of the flow cell, which was then secured to the MinION™ instrument. The flow cell inlet port was opened and inspected with a P1000 pipette (Gilson, Inc., F167700) for the presence of any air bubbles (Fig. 3.3). If air bubbles were detected, they were gently drawn into the pipette tip. Following this, the flowcell was primed twice using 150µl of priming mix (6.5µl Fuel Mix, 162.5µl Running Buffer and 156µl nuclease-free water), with 10min incubation between each addition of the priming mix.

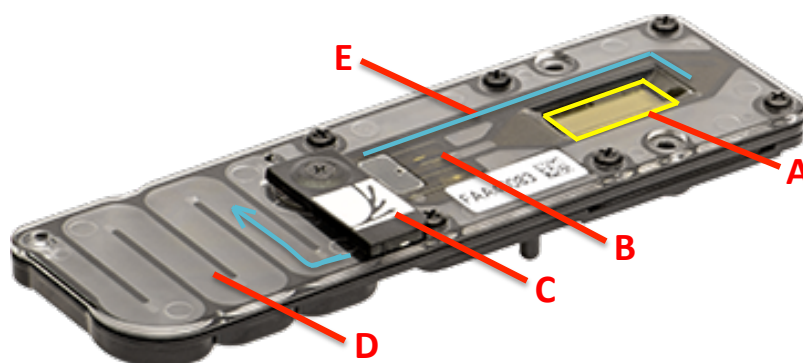


Figure 3.3 Picture is representing the sequencing flowcell R7.3v from ONT: A) array of 2048 wells containing biological nanopores (protein tunnels) embedded into a polymer membrane. Below the array there is an Application Specific Integrated Circuit (ASIC) with 512 signal amplifiers; B) electrode C) sample inlet port; D) waste reservoir; E) channel transporting liquid from membrane array to waste chamber.

ONT MinKNOW v0.50.2.13 software was switched on, and each primed flow cell was tested separately with the Run Protocol Script: MAP\_Platform\_QC. Only flowcells that contained between 900-1000 active pores were used for the sequencing run.

#### *MinION<sup>TM</sup> sequencing*

The amount of DNA template in the pre-sequencing mix was quantified using the HS dsDNA mix using a Qubit 2.0 Fluorometer (Life Tech., Q32866). The 10µl of the pre-sequencing mix was then combined with 75µl 2x Running Buffer, 60µl nuclease-free water, and 3µl Fuel Mix. This sequencing library was then loaded onto the primed flowcell using a P1000 pipette and incubated for 10 minutes. Afterwards, the protocol for a 48h sequencing process was set-up on the ONT MinKNOW v0.50.2.13 software with the real-time “Barcoding plus 2D Basecalling” ver. 1.25 algorithm, Metrichor ver. 2.31.1. Moreover, real-time information output from the MinKNOW software and Metrichor real-time basecaller was used to evaluate the number of reads being generated, and this information was used for sequencing library top-up, as appropriate. If at the conclusion of the 48h sequencing run, the flow cell still indicated a large number of active pores, then the software was restarted and the process repeated until no more nanopores were deemed “available”. The sequencing process generated Fast5 files stored on the local computer, which were transferred to a cloud-based basecaller (Metrichor), (Boza et al., 2017). Data was processed, then basecalled results containing read sequence and base quality were downloaded onto the local computer and split between ‘fail’ and ‘pass’ folders generated according to ONT criteria, related to data quality.

*Data analysis*Generation of reference genes

Sequencing data from the capillary-based Sanger platform was manually trimmed using FinchTV v1.4 package. Removal of low quality bases was necessary for accurate generation of reference genes. Curated reads were aligned with a pairwise nucleotide aligner (MUSCLE) to generate long 16S rRNA reference gene sequences (Edgar, 2004). The resulting long reads were then aligned against the SILVA 1.6v database to obtain taxonomic information and multisequence alignment. The MEGA 6 software tool was used with unweighted pair group method with arithmetic mean (UPGMA) algorithm to build rooted and hierarchical phylogenetic tree (Tamura, 2013). Generation of such phylogenetic tree was necessary to estimate similarity between organisms and their potential cross-contamination during analysis.

Sequencing quality and mapping analyses

Raw nanopore data was basecalled with Metrichor ver. 2.31.1 and subsequently converted from FAST5 to FASTQ format with the use of poretools software (ver. 0.5.1) (Loman et al., 2014). Reads were processed with script: poretools yield\_plot -q --theme-bw, which created a plot of the number of generated reads over sequencing time. All reads ('pass' and 'fail') were mapped to the Sanger sequencing generated reference sequences using multiple algorithms: BWA (Burrows-Wheeler Aligner) with ONT2D setting (ver. 0.7.12) (Li, 2013) and GraphMap software with three settings; i.e. Default, Anchor, and Gotoh (Sovic et al., 2015). We aligned all three variants of MinION<sup>TM</sup> reads, specifically the template (T), complement (C) and the consensus read (2D). Results generated from each aligner in SAM format were size-

selected (1300bp - 1500bp) to further filter out reads and minimise PCR artefacts as well as those generated due to incorrect basecalling.

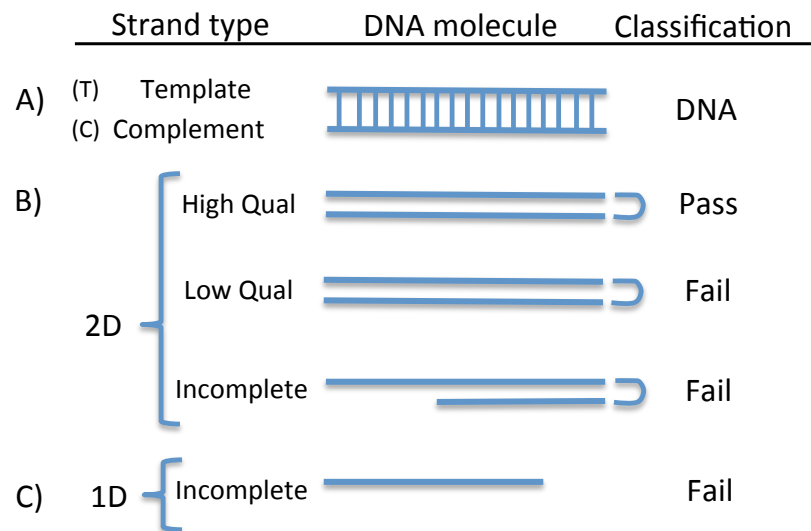


Figure 3.4. This figure demonstrates how the Metrichor basecaller classifies nanopore reads: A) DNA strand is made of two reverse complementary molecules called Template and Complement; B) when both template and complement reads ‘pass’ through the nanopore then they are called 2D. However, only those that have a threshold Phred quality ( $>9$ ) are saved in the ‘pass’ folder, while the remaining low quality, and incomplete 2D reads are saved in a “fail” folder. C) Incomplete reads (both 1D and 2D) that are missing part of complement or template read are transferred into the ‘fail’ folder with no difference in Phred quality scores of them.

The resulting reads were then converted to BAM format and processed as follows. For each of the parameters, i.e., three sequencing runs, three read types (T, C, 2D), and two categories (‘pass’, ‘fail’) returned by the basecaller (Fig. 3.4.). Moreover, we used Picard tools to estimate the total number of reads, read length, number of aligned/unaligned reads, percentage of aligned/unaligned reads, insertion-deletion (i.e., indel) rates, mismatch rates, and overall error rates. We also estimated the relationship between mapping quality scores and base quality score by varying the mapping quality threshold from 0-50 and obtaining the average base quality for the reads (T, C, and 2D), in both the aligned and unaligned categories.

Effect of sequence type on the reconstruction of mock community structure

The reads mapping to the fifteen reference sequences were assigned taxonomy using three configurations: (1) All same (AS): where T, C, and 2D for each fragment assigned to the same reference sequence; (2) Template 2D (T2D): where only T and 2D reads for each molecule mapped to a unique reference sequence; (3) Complement 2D (C2D): where only C and 2D reads assigned to a single reference sequence and (4) Combined, where ‘fail’ and ‘pass’ reads for AS, or T2D were merged to create a single category. This was done for both datasets returned by the Metrichor real-time basecalling software (‘pass’ and ‘fail’). The number of reads mapping to each reference were then counted within each read type (both ‘pass’ and ‘fail’ separately), as well as collated counts for each read type inclusive of ‘pass’ and ‘fail’, resulting in the following read categories: T2D (‘pass’), AS (‘pass’), C2D (‘pass’), T2D (‘fail’), AS (‘fail’), C2D (‘fail’), T2D (‘pass’ + ‘fail’ collated), AS (‘pass’ + ‘fail’ collated) and C2D (‘pass’ + ‘fail’ collated). The abundance table in each category was then used in the downstream statistical analysis to estimate both alpha and beta-diversity metrics. Specifically, we estimated the Shanon, Simpson and Pielou's diversity indices and generated Non-metric distance scaling (NMDS) plots using Bray-Curtis distance; all this was done in R using the Vegan package.

Testing reads alignment to closely related reference sequences

There may be a possibility that reads will align to closely related sequences that are present in the reference database. For this purpose, we extracted the reads that mapped to a particular reference sequence and re-aligned them to a new reference database (14 sequences in total), which did not contain the reference sequence against which it was mapped initially. This was repeated for all the 15 reference sequences.

### 3.4 Results

Reads processed with `poretools yield_plot -q --theme-bw`, created a plot of the number of generated reads over sequencing time (Fig. 3.5). The number of generated reads was highest in the first 8 hours of the flowcell run and more 1D reads ('fail' or low quality) were generated compared to 2D reads ('pass', high quality). In total, there was around 18% (+/- 8) 2D reads across the three sequencing runs, with the remaining ~80% of data assigned to the low quality 'fail' folder. The rate of read generation plateaued at around 40-48h, which may be caused by pore blockage, nanopore protein structure damage due to electric current applied to pores but also due to the cholesterol tether structure, which tends to migrate away from the active nanopore, and that reduces the number of DNA molecules in close proximity to the pores. The total number of reads for individual sequencing runs was: Run 1: 79950 'fail' and 14342 'pass', Run 2: 26403 'fail' and 6975 'pass' and Run 3: 28814 'fail' and 2826 'pass' files.

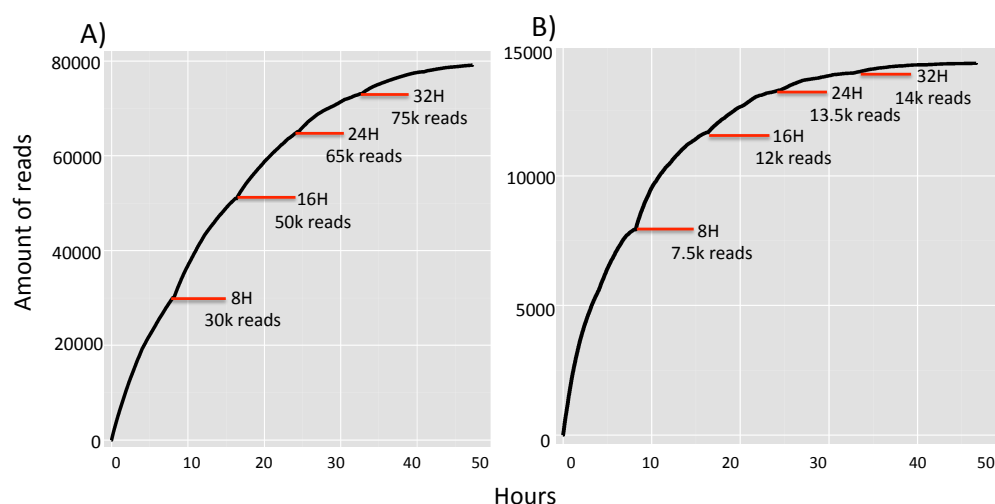


Figure 3.5 Plots representing the total number of reads generated during amplicon sequencing run 1. Results indicate that there was over 5 times more of low quality 'fail' reads A) in comparison to high quality 'pass' data B) . The highest number of reads were generated during the first 8h for both 'fail' (30k), and 'pass' (7.5k) reads. Moreover, data generation plateaued at 40-48 hour for the 'pass' and 'fail' reads accordingly.

Data in Fastq format was loaded into FastQC to visualize sequence quality. Quality scores across all bases indicated that the nanopore data was characterised by low Phred scores, with the majority of reads in the Q8 range (Fig. 3.6).

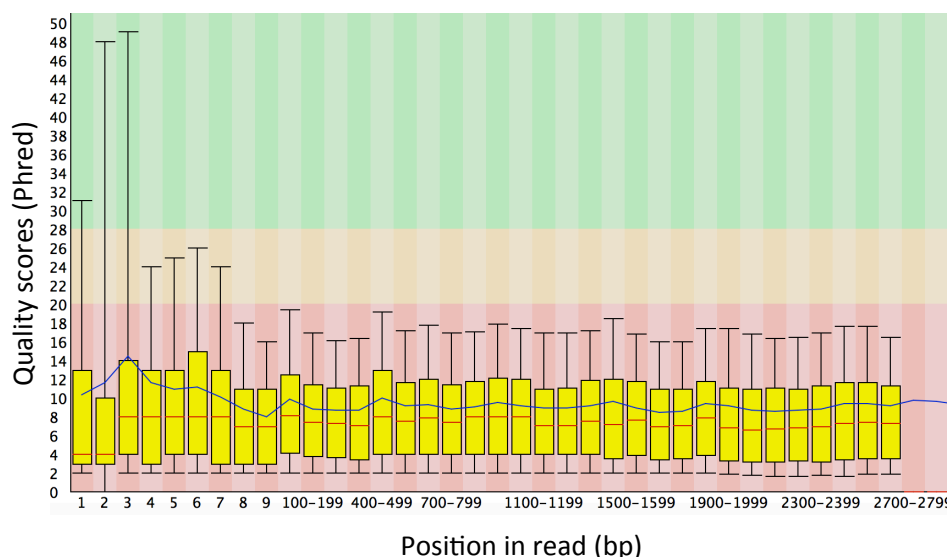


Figure 3.6 Phred scores assigned to each nucleotide by the basecaller for assessment of the sequence quality. Low Phred scores are characteristic of nanopore data with a mean value at around Q8. The FastQC program used for generation of this plot was designed for high-quality Illumina data. For this reason, 16S rRNA nanopore data was placed into the ‘red zone’ of the plot, while high-quality Illumina data (~Q30) would be in the top ‘green zone’. Quality scores are logarithmically correlated to error probabilities, so Q10 indicates 1 error every 10 bp (90% confidence), Q20 indicates error every 100bp (99% confidence) while Q30 gives 99.9% accuracy with 1 error every 1000bp. Data quality thresholds have been automatically assigned by the FastQC software, which is used for analysis of Illumina reads. The X axis goes to 2800bp when library preparation (i.e., 2D) combines template and complement molecules with hairpin adapter. That in turn creates pseudo-long molecule of ~2800bp instead of 1400bp long.

All reads across the three read types (T, C and 2D) and two read categories (‘pass’, ‘fail’) were mapped to the reference sequences using four different mapping approaches. We used these mapping approaches because they have been specifically designed for long, error-prone nanopore reads. Aligners used for data analysis (BWA ONT2D, Graphmap Default, Anchor and Gotoh) have fixed scoring matrices that were specifically designed for nanopore data. These algorithms allowed us to evaluate error rates on the data mentioned above (T, C and 2D for AS, T2D and C2D). The mean read length for the three experiments was approximately 1250bp (+/-45) for the

‘fail’ template, 1100bp (+/- 15) for the ‘fail’ complement and 1170bp (+/-40) for ‘fail’ 2D reads. The ‘pass’ molecules were characterised by a longer read length of 1365bp (+/-35) for template, 1300bp (+/-20) for complement and 1320bp (+/-10) for ‘pass’ 2D reads (Fig. 3.7).

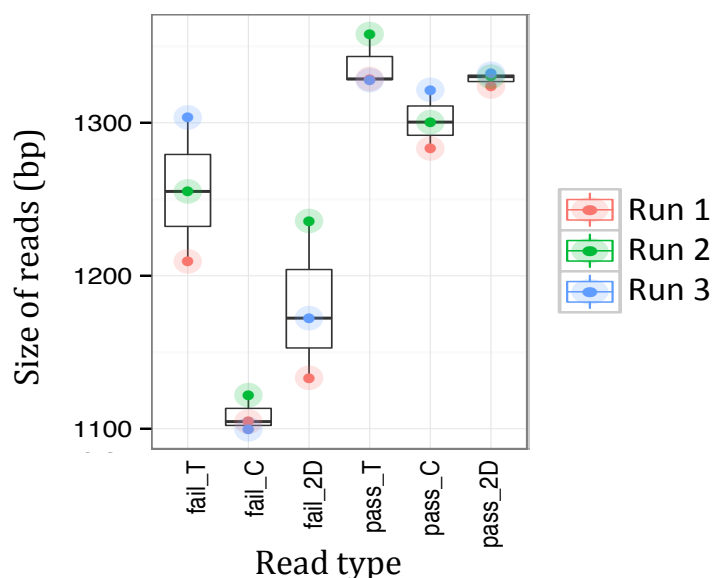


Figure 3.7 Plot representing the size distribution of various read types (template, complement, 2D) for both classifications (‘pass’, ‘fail’) from three independent sequencing runs. Template molecules tend to be longer compared to complement molecules in both ‘pass’ and ‘fail’ classifications. Additionally, size the of molecules from ‘pass’ folder is similar to the original amplicons while size fragments of molecules from the ‘fail’ folder are markedly smaller.

Successfully aligned reads were filtered out, counted and plotted with RStudio (Fig. 3.8). Results indicate that both ‘fail’ T and C reads were aligned to the reference sequences at a much lower proportion in comparison to T and C ‘pass’ molecules. Moreover, nearly 100% of the 2D ‘pass’ molecules aligned to the reference sequences with all 4 algorithms. This indicates that 2D consensus sequences have significantly lower sequence error compared to either T, C or 1D reads.



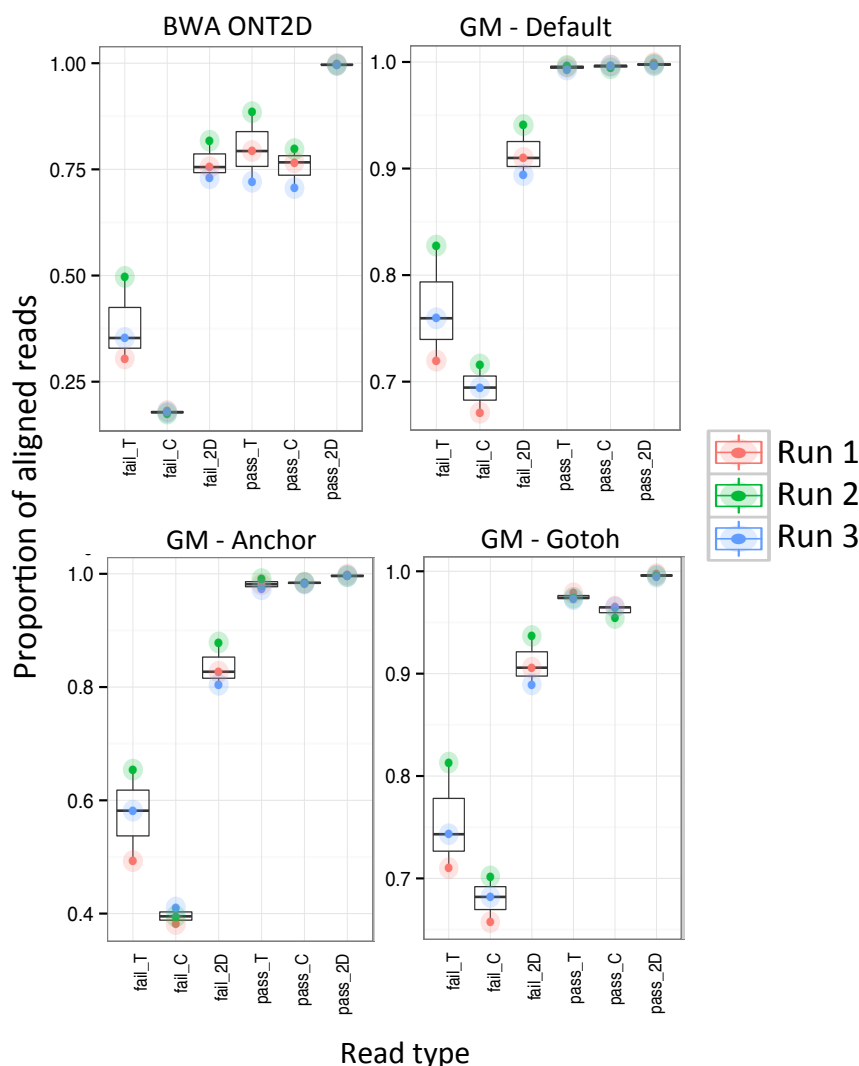


Figure 3.8 Proportion of aligned reads passing alignment filtering for three sequencing runs with use of four different aligners BWA ONT2D, Graphmap Default, Anchor and Gotoh, respectively. Results indicate that ‘fail’ reads aligned at a lower rate in comparison to ‘pass’ reads, which may indicate that data from the ‘fail’ folder contains lower quality reads that should not be used for analysis of complex environmental samples. Moreover, results indicate that GM-Default is the best performing aligner for nanopore ‘pass’ data, as it managed to assign almost 100% of all three types of reads: T, C and 2D. In contrast BWA ONT2D performed worst with T and C reads assigned only at 75% but 2D at ~100%.

The low proportion of successfully aligned reads from the ‘fail’ category suggests that data from ‘fail’ reads were overall of lower quality when compared to ‘pass’ data.

Comparison of aligners using high-quality reads indicated that the GM-Default performed best, as it correctly assigned almost all T, C, and 2D reads (~100%) from the ‘pass’ category. The second best aligner was GM-Anchor, then GM-Gotoh, while BWA ONT2D performed worst as, it aligned only 75% of T and C molecules to the

reference genes. The average sequence accuracy of the ‘pass’ 2D reads for BWA, Graphmap Default, Anchor and Gotoh were approximately 93.5%, 89%, 89.1% and 85%, respectively. In contrast for ‘fail’ 2D reads, sequence accuracy was around 88.8%, 82.7%, 84.7%, 77.2% respectively. However, single template or complement reads had lower accuracy with around 80%, 77.6%, 79.7%, 71.2% for ‘fail’ template and 80.3%, 75.9%, 79.9%, 69% for ‘fail’ complement, as compared to the ‘pass’ template at 79.9%, 79.7%, 80.1%, 73.9% and ‘pass’ complement at 80.9%, 80.8%, 81.8%, 75.4% respectively (Fig. 3.9). These results indicate that BWA ONT2D aligned reads at highest accuracy levels for ‘pass’ and ‘fail’ 2D but also for ‘pass’ and ‘fail’ template molecules. GM-Default and GM-Anchor performed very similarly, while GM-Gotoh aligned reads with the highest error rate. The highest quality reads were generated with ‘pass’ 2D reads and depending on the algorithm, with insertion-deletion errors of 6.5%, 12%, 10.7%, 7.5% error rate for ‘pass’ 2D reads and 9.7%, 18.4%, 16.2%, 11.2% for ‘fail’ 2D reads, respectively (Fig. 3.10). Mismatch rates for ‘pass’ 2D reads were the lowest at 14%, 12%, 12% and 15% while ‘fail’ 2D had 19%, 18%, 16% and 23%, respectively. Error rates at such a levels generated over hundreds of false positive signals (depending on used software) when raw data was analysed against high quality SILVA database. Moreover, these errors (e.g. 10%) does not allow for accurate *de novo* analysis of the reads and OUT generation.

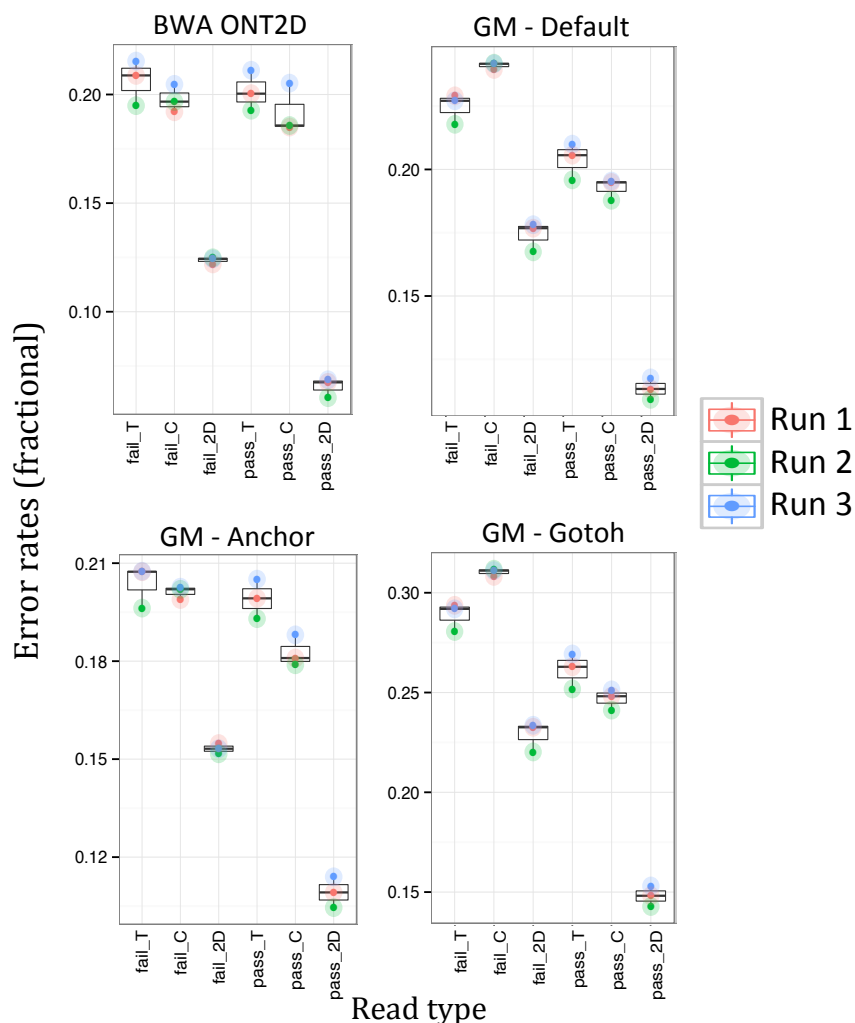


Figure 3.9 Plots demonstrating error rates (passing filter) for all runs and aligners that were estimated with Picard tools. The overall error rate for ‘fail’ data was around 20-30% in case of T and C molecules. However, 2D ‘fail’ reads had the lowest error rate (~12.5%) for BWA ONT2D while remaining algorithms had ~15.5% for GM-Anchor, 17.5% for GM-Default and 22.5% for GM-Gotoh. ‘pass’ 2D molecules were characterised by the lowest error profile that is below 15%. The best result was achieved with BWA ONT2D (6.5%), followed by GM-Anchor (10.9%), GM-Gotoh (11%) and GM-Default (15%).

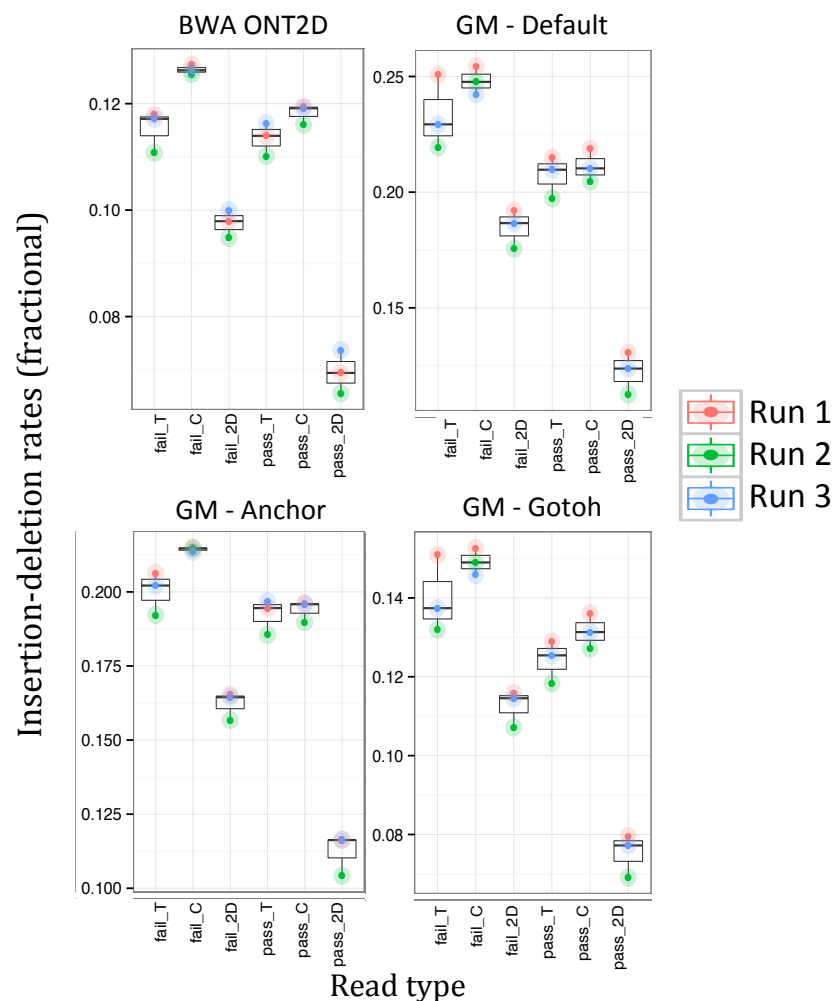


Figure 3.10 Proportion of indels (insertion-deletion) for all 3 runs, 4 aligners and various read types, estimated with use of Picard tools. Results indicate that ‘pass’ 2D molecules contained the smallest amount of indel errors; BWA ONT2D performed best (~7%), while GM-D and GM-A achieved the worst results (8-12%).

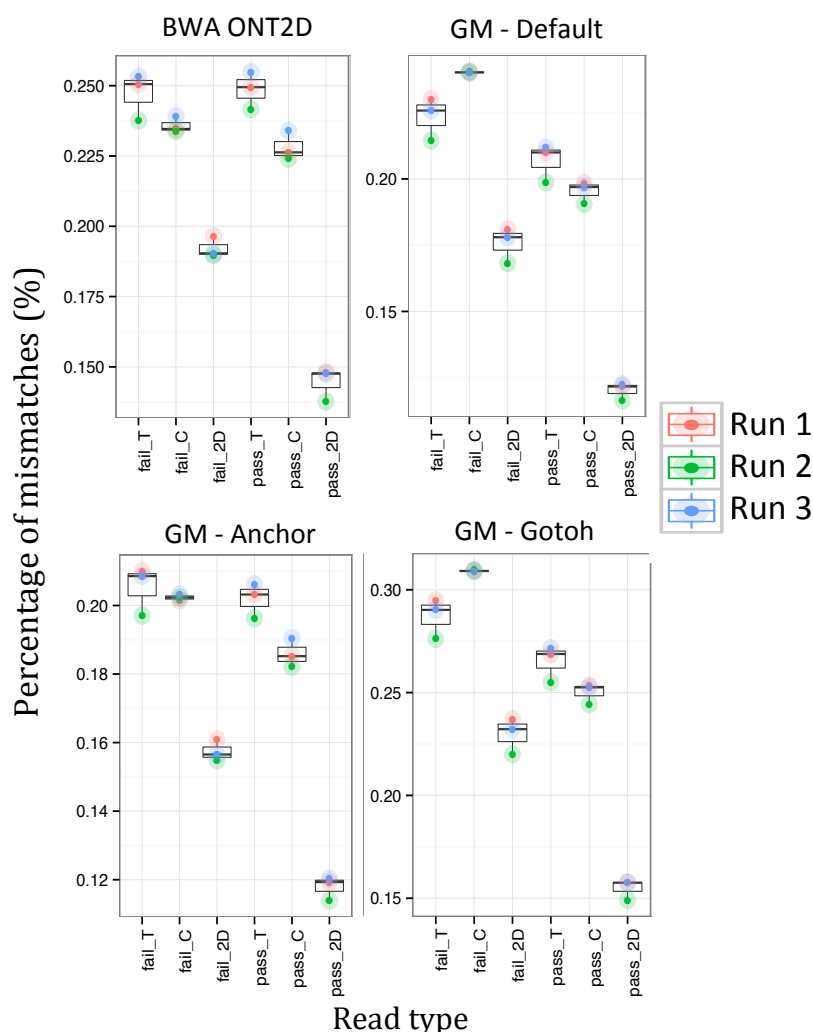


Figure 3.11 Proportion of mismatches present in aligned reads for 3 runs, 4 aligners and various read types evaluated with Picard tools. Results indicate that the mismatch rate for 'fail' T and C reads were in the range of ~25-30% while 'pass' T and C molecules had ~20-25% mismatch rates. The 'pass' 2D reads were characterised by the lowest error rate of 12-16%. The

All assigned and unassigned reads (T, C and 2D) from both ('pass' and 'fail') folders were used to estimate the relationship between mapping quality scores and read base quality for four different aligners and three separate runs. This procedure was undertaken to evaluate whether Phred quality scores can be used as an additional filtration step before mapping to reference sequences (Fig. 3.12). Results for BWA ONT2D indicate that the mean Phred quality scores of assigned reads decreased against a mean mapping quality threshold of 1-15. However, mean Phred quality

increased for all three runs indicating that a higher alignment threshold uses top quality reads. For the first and third runs, the unassigned reads had significantly lower mean Phred quality scores in comparison to assigned reads along the whole alignment quality threshold. However, unassigned reads from the 2nd run had overlapping mean alignment quality with assigned reads that in turn could cause problems with separation of correct data from false low-quality alignments.

Moreover, each run was characterised by a different mean Phred quality, which makes data filtration challenging due to lack of run-to-run reproducibility and limits the use of a single base quality threshold for data filtering. The results from the other three aligners (GM-D, GM-A and GM-G) looked similar to each other. Overall, these results indicate that mean Phred scores for unassigned reads were higher in comparison to assigned reads. This result was noticed in all three runs and demonstrates that Graphmap algorithms were assigning reads with much lower Phred quality into the overall alignment. For this reason, total Phred quality of assigned reads dropped in comparison to unassigned reads.

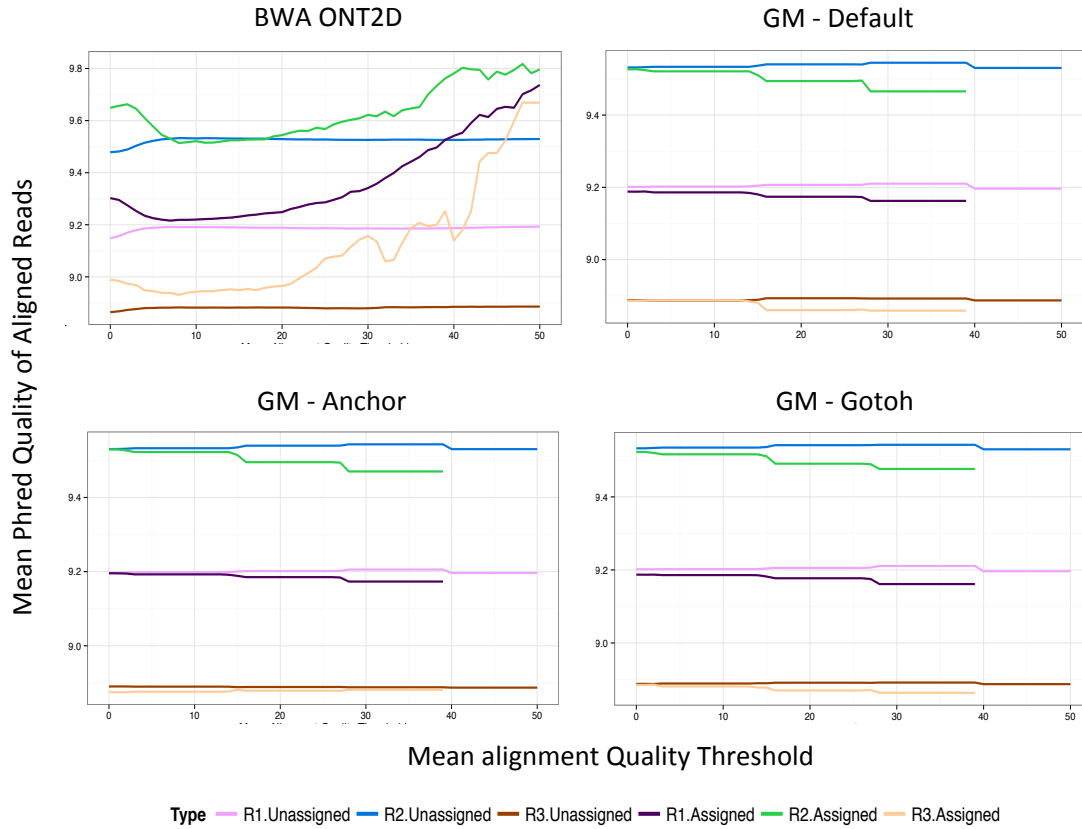


Figure 3.12 Plots representing the relationship between mean Phred quality scores and mean alignment quality threshold for all runs, read types (assigned and unassigned), generated by BWA ONT2D, GM-Default, MG-Anchor and GM-Gotoh, accordingly. Results indicate that all Graphmap aligners performed almost identically, with unassigned reads having higher Phred quality in compare to assigned reads. Moreover, assigned reads from Graphmap aligners did not achieve alignment quality above the 39 point threshold. Results from BWA ONT2D indicated that Phred quality scores of assigned reads increased when compared to unassigned and a mean alignment threshold reached 50 points.

Furthermore, none of the Graphmap aligners generated assigned reads with mean alignment quality above the 39th threshold point. Subsequently, reads aligned to reference sequences were used to estimate the diversity indexes i.e. Shannon, Simpson and Pielou's evenness (Fig. 3.13). All three indexes illustrated that each mapper estimated highest diversity values from the first data run, especially for read types: 'pass' C2D and 'pass' T2D, while the lowest diversity values were estimated for the second run, especially with 'fail' AS. The BWA ONT2D aligner generated higher Shannon entropy values, which indicate for higher species diversity but also

higher evenness of the samples in the pool. Nonetheless, there was no statistical difference between any of the aligners in case of the diversity metrics.

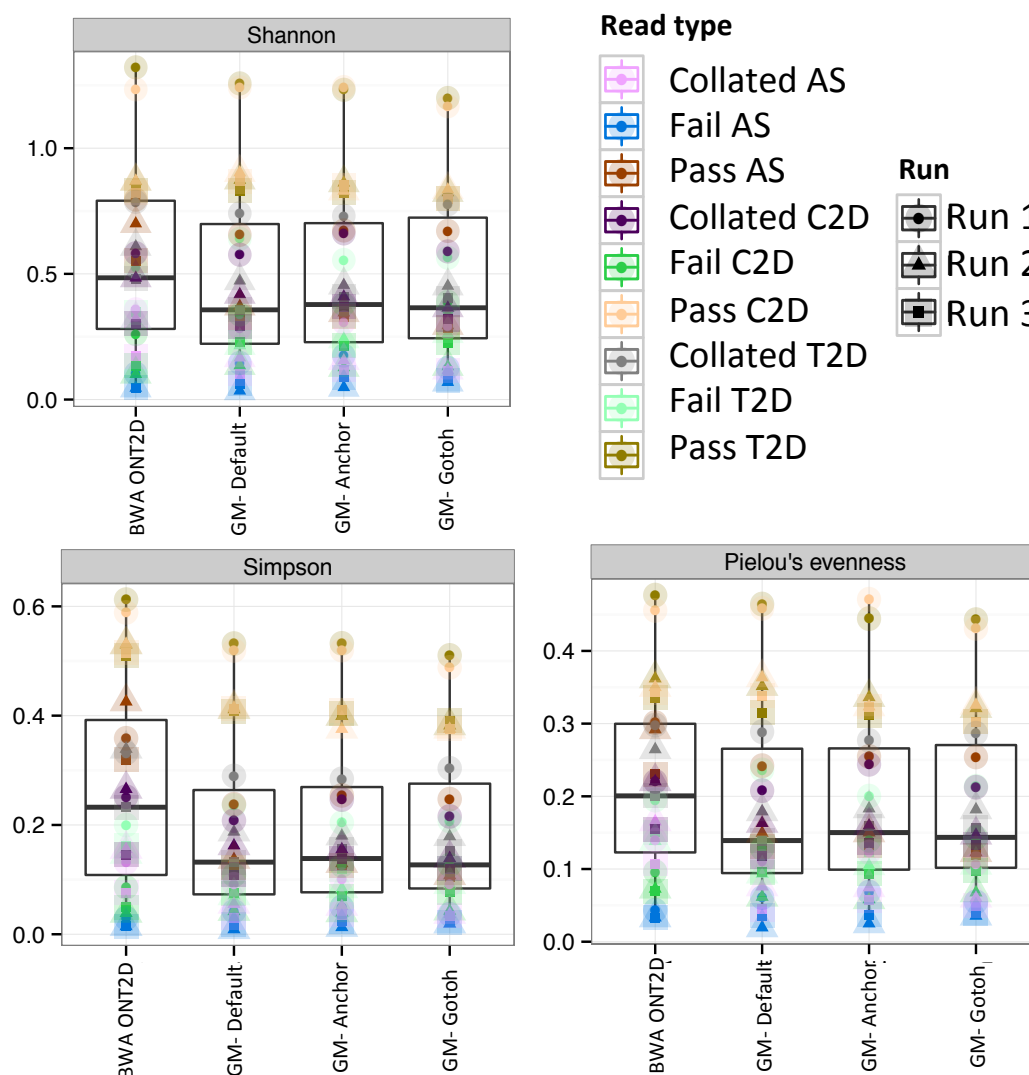


Figure 3.13 Plots representing Shannon, Simpson diversities and Pielou's evenness. All the data was rarefied and unassigned reads were removed. Results indicate that the lowest diversity was achieved with 'Fail AS' type of reads, while the highest diversity was generated with 'Pass AS' for all three statistical classifications. Moreover, the BWA ONT2D aligner generated the highest overall diversity when compared to remaining aligners and all three statistical approaches.

The data was further rarefied and the relative abundance of each reference sequence was estimated for each run, aligner and read type (Collated All Same – AS, Collated Complement 2D – C2D and Collated Template 2D – T2D) were evaluated (Fig. 3.15).



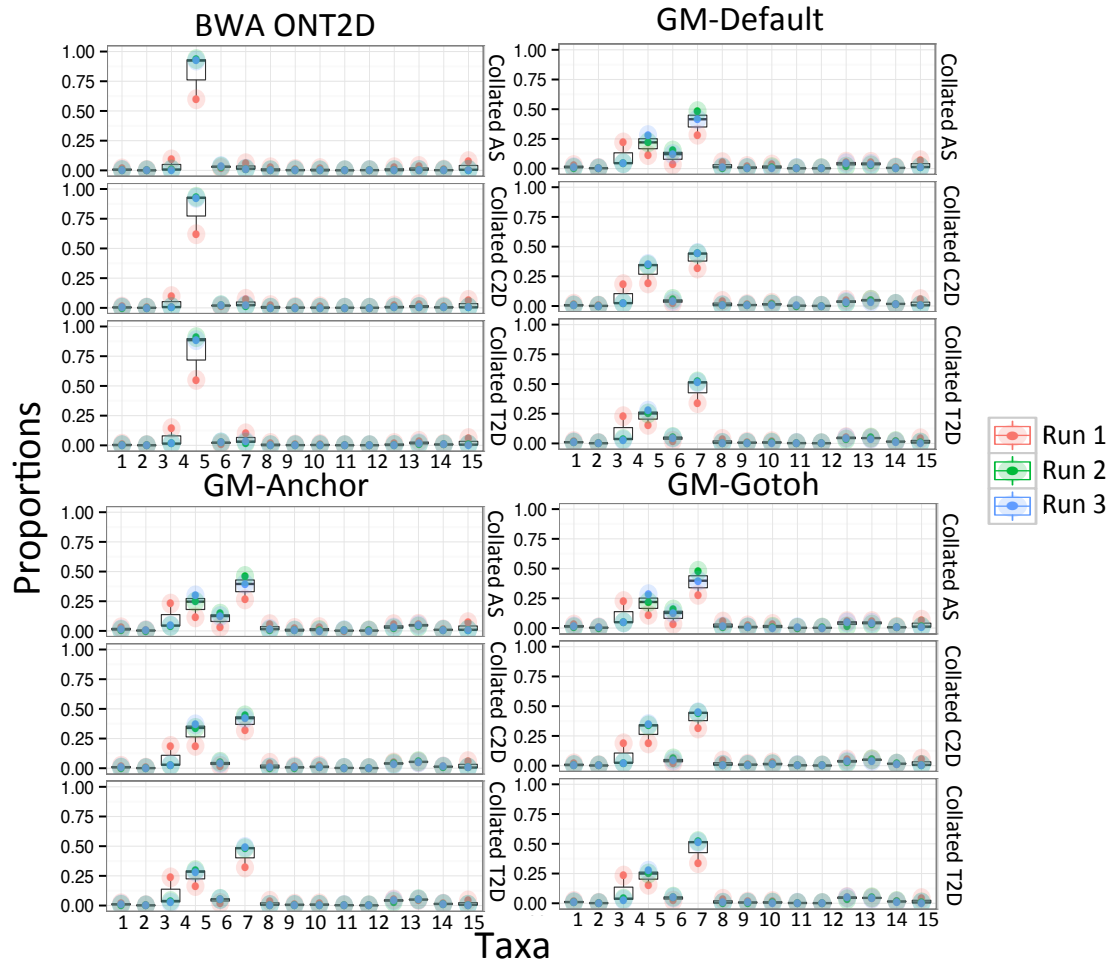


Figure 3.14 Summary plots of microbial abundance for different read types with four different aligners: BWA ONT2D, GM-Default, Anchor and Gotoh accordingly. Results indicate that the BWA ONT2D aligner generated the highest accuracy of microbial community structure when compared to the other aligner algorithms. One of the 16S rRNA genes was overrepresented for the BWA ONT2D aligner, while other algorithms (i.e. Graphmap) had multiple i.e. 4 organisms being overrepresented.

Aligner programs generated different abundance profiles whereas BWA ONT2D indicated overrepresentation of single organism (number 4) in all runs and all read types. The remaining three aligners estimated relative abundance profiles, which were very similar to each other without significant difference between them. A comparison of the relative abundance plots generated by Graphmap algorithms to BWA ONT2D indicates that all approaches estimated the relative abundance of 1-2 and 8-15 taxa correctly (i.e. 0.07 ratio) while overestimating the relative abundance 3-4 taxa in different ways (e.g. 0.9 ratio). Considering BWA ONT2D as the most accurate

aligner, we decided to test why taxa number 4 was over abundant. To do this, the raw reads were aligned with BWA ONT2D to 14 out of the 15 reference genes in an iterative manner to assess whether removal of one reference sequence results in the re-assignment of reads previously aligned to it to another reference sequence. This was done for read groups AS, T2D and C2D, as mentioned previously (Fig. 3.15). Results indicate that a small number of reads were misaligned when their corresponding reference was missing, with the exception on taxa 4 (B20). A significant number of reads previously aligned to reference B20 have realigned to another reference (B29), which is very highly phylogenetically related (Fig. 3.15). This indicates that the sequencing error can have an impact of reference base alignment for closely related reference sequences.

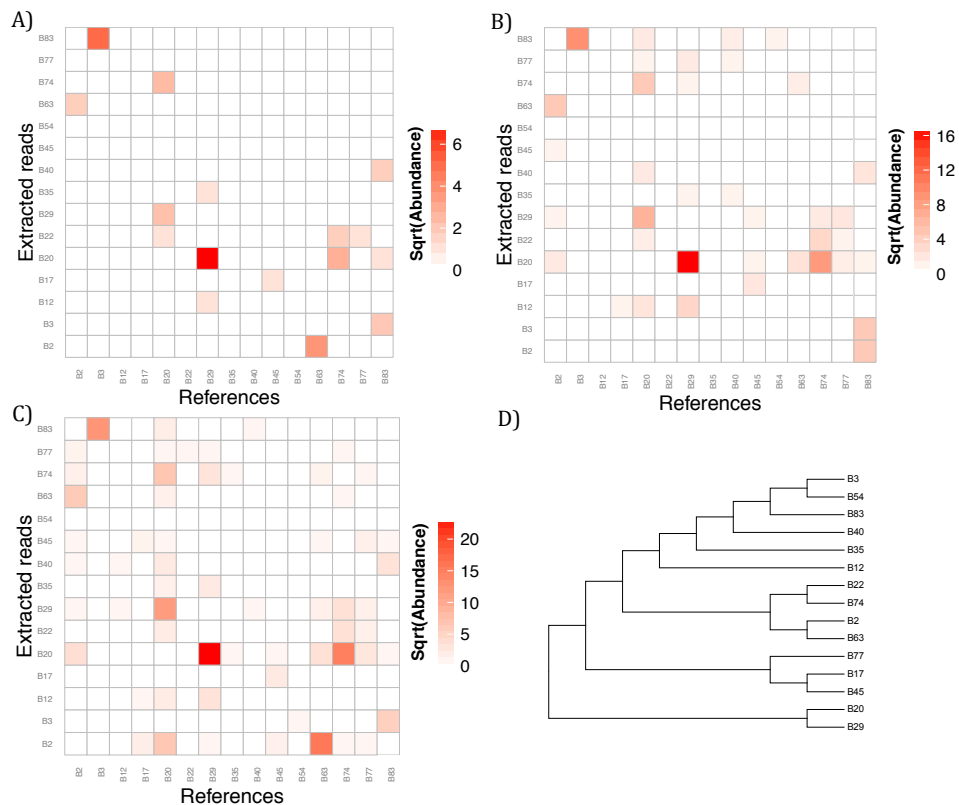


Figure 3.15 Heatmap representing reads realigned to reference gene sequences with BWA ONT2D by iteratively removing one reference sequence for runs 1, 2, and 3 (A, B, and C). D) Phylogenetic tree of all 15-reference genes was generated with maximum parsimony and shows that that reference sequence B20 and B29 are very closely related (Class: Clostridia), which explains the incorrect alignment of reads between these two reference sequences.

### 3.4 Discussions and Future Work

Initial results of the research indicated that all 15 reference sequences were detected for all 3 runs, with the highest quantity of reads aligned for the 1<sup>st</sup> run and the smallest number for the 3<sup>rd</sup> run as a result of the amount of total generated data. Error profiles for MinION<sup>TM</sup> sequencing for T, C and 2D molecules were consistent across sequencing runs and was not related to the proportion of aligned reads, however, differed depending on the aligner algorithm used. The lowest error profile was obtained for 2D pass reads: BWA ONT2D 6.5%, GM-D 11.3%, GM-A 10.9% GM-G 14.7%, respectively. The highest read quality was observed with the BWA ONT2D aligner, which indicates that this aligner effectively discarded most of the low quality reads compared to Graphmap aligners. Estimation of the relationship between mean mapping quality scores and mean Phred quality indicated that in most cases BWA ONT2D aligned reads with higher mean Phred quality values. In contrast, all unaligned reads for Graphmap aligners had higher Phred mean quality scores over aligned reads, due to higher tolerance for sequence errors. Diversity indices such as Shannon, Simpson and Pielou's evenness did not indicate a significant difference between any of the four alignments, which may be due to the low number of reference sequences in the mock community sample. The BWA aligner overestimated presence of only a single sample (B20) in the evenly distributed mock pool (prepared independently for each run), while abundance for all three Graphmap aligners indicated overrepresentation of 3-4 taxa. This discrepancy was investigated by reanalysis of reads previously aligned to B20. Results indicated that overrepresentation of taxa B20 by BWA ONT2D algorithm was due to misalignment of a large number of reads associated with taxa B29. Sequencing errors on these reads were large enough to result in misalignment of reads between B20 and B29 and thus

skew the relative abundance profile from the theoretical estimate. The outcome of this analysis indicates that the nanopore sequencing device and library preparation are not directly reliable for community composition or structure estimation. However, in this experiment, we did not test the impact of PCR amplification, which could also impact community structure.

The results of this chapter very precisely described error rates of the nanopore sequencing technology. Conclusions of this experiment are not published, mainly due to release of similar work by Benítez-Páez et al. Nonetheless, his study did not compare multiple different aligners, mentioned a problem with false positive detection of organisms when reads were tested against high quality ribosomal databases (i.e. SILVA) or problems with *de novo* binning of reads while in this study we tested all of these issues. Nanopore sequencing with the tested chemistry (R7.3 and SQK-MAP005) had error rates similar to that of other long-read sequencing platforms, i.e. PacBio (Koren et al., 2012). Nonetheless, PacBio SMARTbell chemistry allows for the generation of high accuracy consensus sequences out of the single sequencing strand and to use it for precise *de novo* 16S rRNA analysis in comparison to reference based analysis.

## 4 Protocol development for amplicon sequencing of mixed microbial communities

### 4.1 Abstract

The MinION™ nanopore technology is a miniaturised and inexpensive sequencer for real-time analysis of long DNA molecules. This technology could revolutionise the field of environmental and clinical microbiology. Nonetheless, high error rates (2-15%) of recent nanopore chemistry (R7 and R9) results in limitations for multiple applications, i.e., retention of errors in genomic or metagenomic data but also limited scope in analysis of SSU rRNA from complex samples. These limits are mainly caused by insertion and deletion errors in homopolymer regions that cannot be fixed using a simple consensus based approach. Moreover, these errors cannot be easily fixed in low sequence variability regions, such as ribosomal genes, due to presence of multiple conserved regions. To allow for accurate analysis of long ribosomal genes on the nanopore platform there is a need to develop novel library preparation protocol that would allow for reduction of the aforementioned errors. This chapter describe the development of the NanoAmpli-Seq laboratory protocol, which is a significant expansion and improvement in the previously described Intramolecular-ligated Nanopore Consensus Sequencing (INC-Seq) workflow. Development of the library workflow reduced time needed for construction of sequencing libraries, increased size of amplicon molecules and boosted total data output.

This chapter is partly based on the publication:

Calus S.T., I.U.Z. and P.A.J., (2016)., **NanoAmpli-Seq: A workflow for amplicon sequencing from mixed microbial communities on the nanopore-sequencing platform.** GigaScience, giy140, <https://doi.org/10.1093/gigascience/giy140>

Original contributions:

Tested two laboratory protocols Loop-mediated isothermal AMplification (LAMP) and Rolling Circle Amplification (RCA) for reduction of nanopore sequencing errors rates i.e., mismatches, deletions and insertions. Both protocols were experimentally evaluated but only one best performing method was used the new NanoAmpli-Seq protocol. Consideration to simplify the library preparation protocol was paramount and the protocol is characterised by a significant reduction in manual labour (when compared to a competitive protocol; i.e., INC-Seq). Optimisation of the protocol was achieved by improvement of the RCA procedure; i.e., addition of random hexamer-free isothermal amplification. Moreover, introduction of additional steps (e.g., enzymatic cleavage using T7 endonuclease I) resulted in a significant increase in total data output and allowed for a more precise analysis of the bacterial genetic markers.

## 4.2 Introduction

Advances in various sequencing platforms (i.e., first, second and third generation) has revolutionised DNA sequencing. Nonetheless, each sequencing generation, platform or even experiment relies on different protocols for DNA or RNA preparation (e.g., amplicon PCR, whole genome DNA amplification, reverse transcription for RNA). Often the sequencing process is time and resource intensive, requiring that a rational and reliable library preparation protocol is utilized. The development and validation of these protocols involved their testing with initially simple and then complex microbial communities (e.g., single organism or mock community sample). Numerous laboratory-based modifications are tested at various stages of the protocols to benchmark the method being developed. Use of mock community samples can also help in further evaluation of and/or improvement to existing bioinformatics

algorithms and new programs can be developed for improved analysis of the natural samples using insights from mock community experiments.

This chapter describes the incorporation of new approaches for improved library preparation for the nanopore sequencing platform, with emphasis on full-length 16S rRNA gene (~1400bp) sequencing. Initially two laboratory protocols were tested: Loop-mediated isothermal AMPlification (LAMP) and Rolling Circle Amplification (RCA), (Parida et al., 2008). Isothermal PCR with use of LAMP is a method for generation of synthetic concatamerised long amplicons (Fig. 4.1). This method uses isothermal DNA polymerase from *Bacillus subtilis*, phage  $\Phi 29$  (phi 29). This polymerase possesses strong strand displacement activity and 3' - 5' proofreading activity. The enzyme is active at 30°C and can be inactivated by heating up to 65 °C for 10min. Preliminary results of the LAMP protocol indicated that this method generated false positive signals in every negative control samples (DNA-free sample), which could be caused due to high concentration of multiple primers in the reaction, resulting in unspecific product formation. Moreover, accurate amplification of long templates was not possible due to a high level of unorganised hyperbranching structures. Results indicated that LAMP amplification could be successfully used for accurate amplification of very short fragments only (e.g., 100-200bp). Despite various attempts at optimisation (i.e., lower temperature, reduced volume of phi29 polymerase and random hexamers primers) none of the reactions successfully amplified long-16S rRNA gene. For this reason, development and optimisation of the LAMP protocol was discontinued and another (RCA) laboratory protocol was tested.

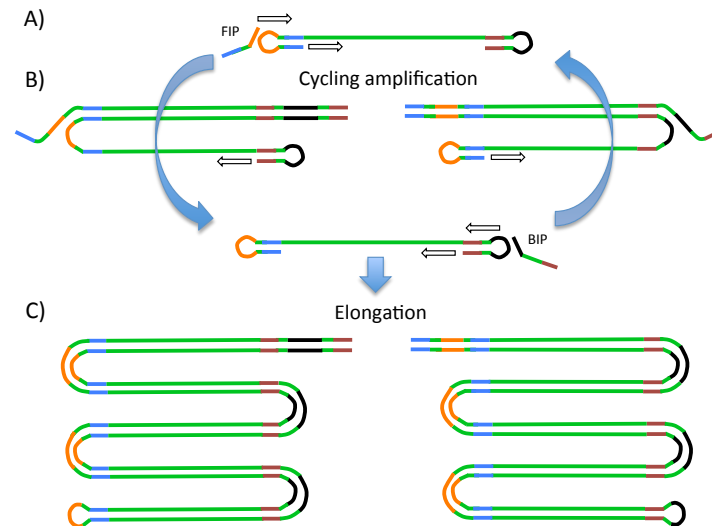


Figure 4.1 Schematic representation of the Loop-mediated isothermal AMPLification (LAMP) assay: A) requires ligation of hairpin adapters to both ends of an amplicon and subsequent addition of two primers; Forward Inner Primer (FIP) and Backward Inner Primer (BIP); B) isothermal polymerase phi29 is added to the reaction and these primers allow for the first amplification cycle, depending on the primer there are two versions of the concatamerisation; C) further elongation of FIP and BIP primers to DNA amplicons generates, long concatamerised gene of interest.

Rolling Circle Amplification is another type of isothermal amplification that uses phi29 polymerase (Wu et al., 2003). This method requires initial DNA fragmentation or amplification for later self-ligation into a plasmid like structure (Fig. 4.2). The standard RCA protocol requires the addition of random hexamer primers to the reaction and 4-8h incubation at room temperature. Nonetheless, use of these primers may cause self-amplification and induce false signals in negative control samples. For this reason, modifications to the RCA protocol were necessary. In 2016, Li et al. published a protocol called Intramolecular-ligated Nanopore Consensus Sequencing (INC-Seq), which was based on the RCA method (Li et al., 2016). The INC-Seq library preparation for MinION™ device allows obtaining of long (~680bp) concatamerised amplicon molecules for accurate (i.e., 97-98%) bacterial profiling. However, the INC-Seq protocol is not suitable for long ribosomal amplicon



sequencing, containing 2-3% errors and did not allow for *de novo* analysis (OTU binning) of the data.

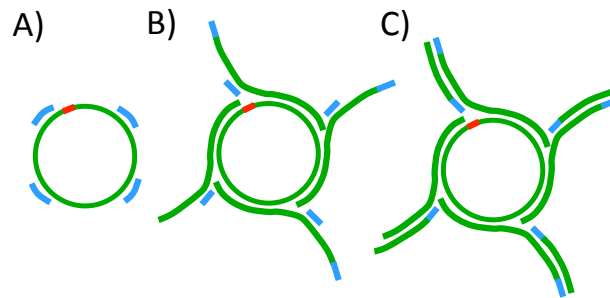


Figure 4.2 Schematic representation of the RCA assay that was tested and used for library preparation for 16S rRNA amplicons. The protocol includes amplification of a gene of interest (i.e., 16S rRNA), then A) circularisation of amplicons by self-ligation and addition of random hexamer primers, B) addition of isothermal polymerase phi29 to allow for elongation primers and generation of multiple branches, C) prolonged incubation of the reaction causes ligation of random hexamers at various positions on new branches and allows form hyperbranches made of repeated ribosomal gene sequences.

The protocol described in this chapter (NanoAmpli-Seq) was developed around the same time as Li's et al (2016) INC-Seq method. However, the NanoAmpli-Seq method represents a significant improvement over the INC-Seq protocol as it allows for twice as long marker gene amplicons (up to 1380bp), and reduces library preparation time by 70%. Moreover, addition of the TthPrimPol enzyme allowed for elimination of random hexamers and in turn eliminated primer-self amplification. Additional fragmentation steps were added (e.g., T7 endonuclease I) into the protocol to enhance debranching of RCA products, which in turn increases total data output by minimizing pore blockage. Finally, the sequencing data was further improved with two novel bioinformatics algorithms, which reduced error rates further when compared to INC-Seq analysis (Chapter 6).

### 4.3 Method development

#### Mock community description and preparation

Two different mock samples consisting of 16S rRNA genes from one and ten organisms were constructed for the experiments outlined in this study. First, a single organism mock sample was constructed by amplifying the near full-length of the 16S rRNA gene from genomic DNA of *Listeria monocytogens*, using primers sets 8F (5'-AGRGTTTGATCMTGGCTCAG-3') and 1387R (5'-GGGCGGWGTGTACAAG-3'), both with 5' phosphorylated primers (Eurofins Genomics). Phosphorylated ends were essential for the subsequent self-ligation step. The PCR reaction mix was prepared in 25µl volumes with use of 12.5µl of Q5® High-Fidelity 2X Master Mix (New England BioLabs Inc., M0492L), 0.8µl of 10pmol of each primer, 9.9µl of nuclease-free water, (Roche Ltd.) and 1ng of bacterial DNA in total followed by PCR amplification as described previously in Chapter 3. PCR amplicons from replicate PCR reactions were combined and purified with use of HighPrep™ PCR magnetic beads (MagBio, AC-60050) at 0.45x ratio. The ten organism mock community was constructed from purified near full-length 16S rRNA amplicons of 10 organisms. Genomic DNA from 10 bacteria was obtained from DSMZ, Germany (Appendix I, Figure 1.1), the aforementioned primers, PCR reaction mix and thermocycling conditions were used to independently PCR amplify the near full length 16S rRNA gene, followed by purification using HighPrep™ PCR magnetic beads, as detailed above. The purified amplicons from each organism were quantified on the Qubit 2.0 fluorometer using the dsDNA HS kit, normalized to 4ng/µl, and combined to generate an amplicon pool consisting of equimolar proportions of the 16S rRNA gene amplicons of the 10 organisms. Multiple titration experiments have been performed to

test amplification rate, centrifugation speed and purification of long DNA fragments with magnetic beads (Appendix I).

#### Library preparation

To circularize the linear amplicons into plasmid-like structures, 5 $\mu$ l of Blunt/TA Ligase Master Mix (New England Biolabs, M0367L) was added to 55 $\mu$ l of amplicon pool at a concentration of 1ng/ $\mu$ l and incubated for 10min at 15°C then 10min at room temperature (total time = 20 minutes). Not all linear amplicons self-ligate into plasmid-like structures, but some are likely to cause long chimeric linear amplicons (Fig. 4.3.C). These long chimeric structures were removed using magnetic bead based purification, with the following modifications. HighPrep™ PCR magnetic beads were homogenized by vortexing, followed by aliquoting 50 $\mu$ l into sterile 2ml tubes and placing on a magnetic rack for 3min. A total of 25 $\mu$ l of supernatant was carefully removed using a sterile pipette to concentrate the beads to 2x their original concentration. The tube was removed from the magnetic rack and vortexed vigorously to resuspend the beads. This concentrated bead solution was used at a ratio of 0.35x to remove any amplicons greater than 2000 bp in the post-ligation reaction mix. Briefly, the post-ligation product was mixed with concentrated bead solution at 0.35x ratio by vortexing, followed by incubation for three minutes at room temperature. The tube was placed on the magnetic rack to separate the beads from solution, followed by transferring of clear liquid containing DNA structures less than 2000 bp into new sterile tubes. Samples containing short self-ligated molecules were subject to another round of concentration using standard magnetic beads at 0.5x ratios, according to manufacturer instructions and eluted in 15 $\mu$ l of warm nuclease-free water.

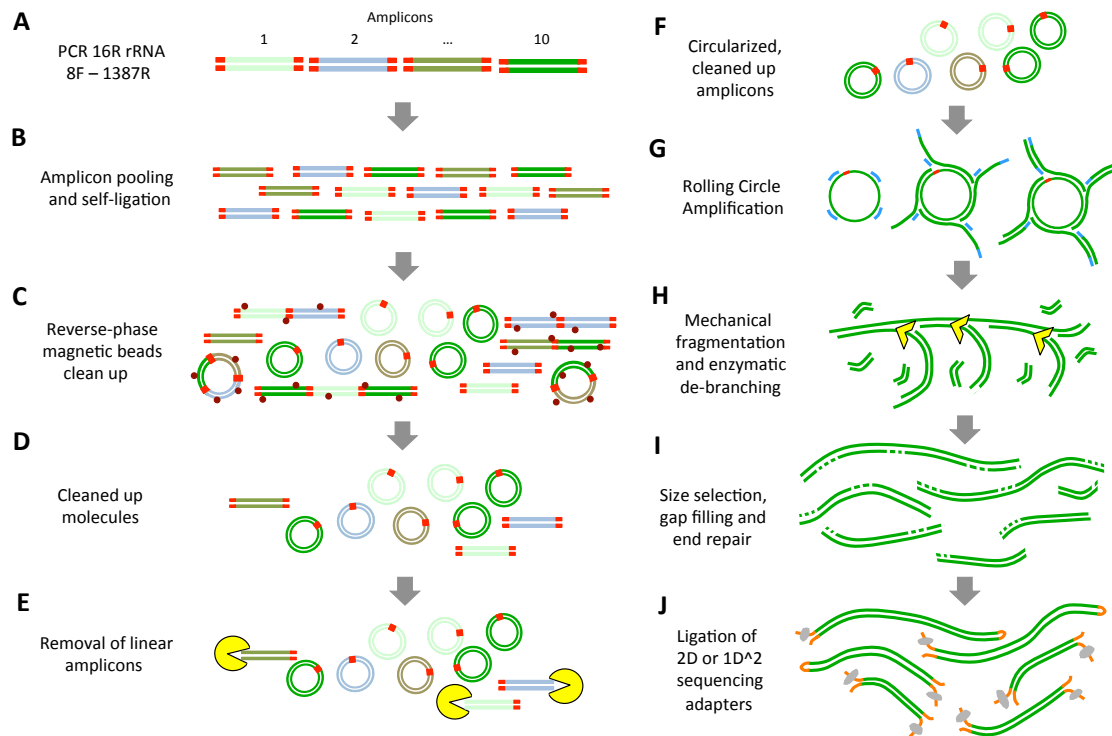


Figure 4.3 Schematic representation of NanoAmpli-Seq laboratory protocol: A) PCR amplification of 16S rRNA gene with use of 8F and 1387R primers, B) purified amplicons were combined in equimolar concentration and subjected for self-ligation, C) since not all amplicons circularised correctly, with some of them forming long multi-molecule-hybrid made of 2 or more amplicons had to be removed with use of magnetic beads. D) Cleaned up plasmids may still contain short amplicons in a linear form so in the next step E) linear molecules are enzymatically degraded, F) only correctly self-ligated amplicons were taken forward for Rolling Circle Amplification. G) At this step Phi29 isothermal polymerase is used to generate long concatemerised molecules H) hyperbranched molecules are initially fragmented with use of g-TUBE then with T7 endonuclease I, I) molecules are size selected and gaps in DNA molecules are filled up with NEBNext® FFPE DNA Repair Mix. J) Final step of library preparation involving ligation of 2D or 1D2 sequencing adapters.

The concentrated and cleaned DNA pool consisting of plasmid-like structures and remaining linear amplicons was then processed with Plasmid-Safe™ ATP-Dependent DNase (Epicentre, E3101K) reagents to digest linear amplicons using the mini-prep protocol according to manufacturer instructions this was followed by another round of cleanup with magnetic beads at 0.45x ratio as described before, and then samples were eluted in 15µl of warm nuclease-free water. The pool containing plasmid-like structures was subject to RCA with use of the TruPrime™ RCA Kit (Sygnis, 390100)

random hexamer-free protocol. Samples were prepared in triplicate and processed according to the manufacturer's protocol with all incubations performed in triplicate for 120-150min, depending on the assay efficiency (Appendix I, Figure 1.2). The progress of RCA was monitored by measuring the concentration of DNA using a Qubit® 2.0 Fluorometer at 90, 120 or 150min time points. Negative control samples containing reagents without circularized plasmid-alike amplicons were processed and analysed concomitantly with the samples. The final concentration of the RCA product after 150min of incubation was typically ~70ng/μl when using a starting DNA concentration of 0.5ng/μl, with no detectable unspecific product formation in the negative control. Replicate RCA products were combined (~4.5μg of DNA in total) and subject to de-branching and fragmentation of post-RCA molecules to remove hyperbranching structures generated during RCA. The RCA product was first treated with T7 endonuclease I enzyme (New England BioLabs, M0302S) by adding 2μl of the reagent to 65μl of RCA product followed by vortexing and incubation, as recommended by the manufacturer. Subsequently, the reaction mix was transferred into a g-TUBE (Covaris, 520079) and centrifuged at 1800 rpm for 4min or until the entire reaction mix passed through the fragmentation hole. The g-TUBE was reversed and the centrifugation process was repeated. Post debranching and fragmentation, short fragments were removed using modified bead based clean up step using concentrated bead solution. Speed of centrifugation was optimised and the results are available in Appendix I, Figure 1.3. Concentrated beads were mixed with the fragmented RCA product at 0.35x ratio, vortexed for 15sec, and incubated at room temperature for 3min then placed on a magnetic rack until beads separated and the supernatant was then removed. The beads were subsequently washed with 70% freshly prepared ethanol according to manufacturer protocols. Size selected amplicons

bound to the beads were eluted in 41µl of warm nuclease-free water. Preliminary experiments indicated that one round of de-branching did not completely resolve the hyperbranching structure, which later inferred with poor sequencing yield likely caused due to pore blocking by hyperbranched DNA. As a result, a second round of enzymatic de-branching using T7 endonuclease I was added and the de-branched product was cleaned a second time using the bead based clean-up step. Figure 4.4 shows an example of BioAnalyzer traces of the RCA product post-debranching/fragmentation and post-cleanup using the magnetic bead based protocol. Removal of short DNA particles has been necessary to increase the efficiency of the assay. Moreover, modified version of magnetics beads clean up proves to be successful and could be used in other molecular biology workflows.

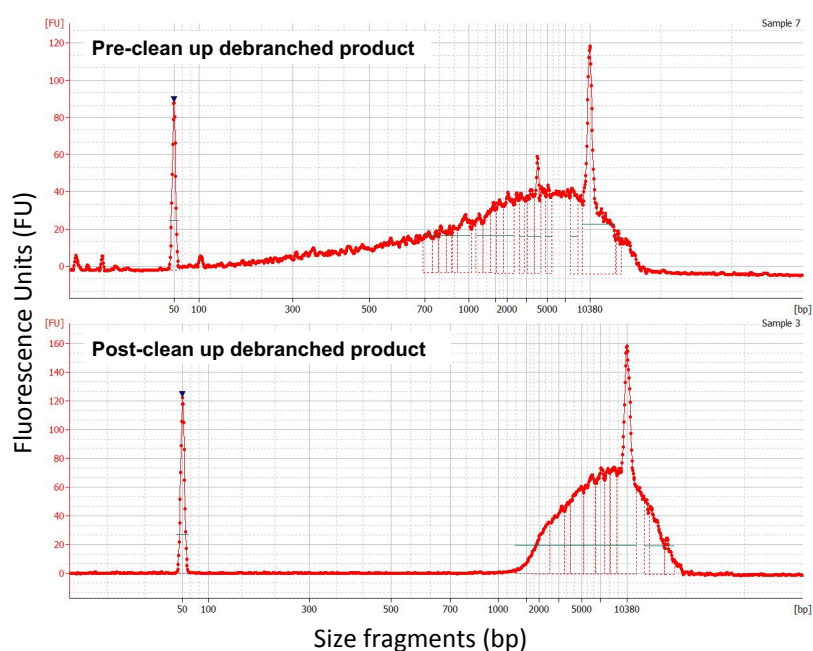


Figure 4.4 Plots generated with Bioanalyzer 2100, DNA 12000 Kit (Agilent Genomics). Electropherograms indicate size distribution of hyper-branched DNA molecules before (pre) and after (post) magnetic beads clean up. Two peaks at 50k and 10k base pairs represent internal standards of the reaction. Results demonstrate that short size amplicons have been removed from the sample, with a cut off point at 2,000bp. The magnetic bead clean up is a cheap and portable method that was modified and introduced into the protocol to substitute for Blue Pippin Prep (Sage Science) size selection, which is an expensive and labour intensive method.

Finally, the de-branched RCA product was treated with NEBNext® FFPE DNA Repair Mix (New England BioLabs, M6630S) for gap filling and repairing DNA damages caused during g-TUBE fragmentation and with T7 endonuclease I enzyme digestion. All reagent components were combined with the de-branched RCA product according to the manufacturer's recommendations and incubated at 12°C for 10min then at 20°C for another 10min. Post incubation, the repaired RCA product was cleaned using standard magnetic beads at 0.5x ratio, washed with 70% ethanol, and eluted in 46µl of warm nuclease-free water. Concentration of the DNA product was measured using Qubit and was approximately 20-25ng/µl, with a total yield of ~1000ng of DNA, with a product size typically ranging from 1,800bp to 20,000bp. Optimisation of bead to liquid ratio was performed with titration assay and results are available in Appendix I Figure 1.4. A total of 45µl DNA pool of concatamerized amplicons was prepared for sequencing using the standard 2D and 1D2 library preparation protocol of Oxford Nanopore Tech. (SQK-LSK208, SQK-LSK308), according to manufacturer specifications to obtain the pre-sequencing mix. The final concentration for prepared libraries was determined using the dsDNA HS kit on the Qubit instrument. A detailed step-by-step protocol is provided in the supplementary text (Appendix II).

#### **4.4 Conclusions and future work**

This chapter describes improved and optimised NanoAmpli-Seq protocol that is a combination of INC-Seq with novel RCA-based laboratory workflow for bacterial 16S rRNA marker identification. The protocol was specifically designed for nanopore-based sequencing technology, such as MinION™ and GridION™ X5. The entire protocol can be found at: <http://dx.doi.org/10.17504/protocols.io.u26eyhe> This

protocol can generate long (~1400bp), organised, concatamerised ribosomal molecules that can be later used to reduce nanopore error rates by consensus data calling (INC-Seq software). The protocol was optimised to generate long-concatamerised ribosomal genes with random hexamer-free isothermal amplification (PrimPol Tth) and significantly reduce preparation time by around 70% when compared to a competitive INC-Seq or R2C2 workflow. Improvements related to DNA hyperbranching such as use of T7 endonuclease I allowed for better DNA fragmentation and in turn increased overall data output. Detailed description of the protocol with step-by-step explanation is provided in Appendix II. The protocol is specifically optimised for long-16S rRNA bacterial genes and may be varied at multiple stages if different bacterial or fungal gene markers would be analysed. Use of mock samples allowed for further improvement and generation of two novel bioinformatics algorithms described in Chapter 6. Currently, real-time nanopore sequencing is still limited by various errors and time consuming manual labour. However, forthcoming improvements introduced by Oxford Nanopore Tech. may reduce raw error rates and allow for automation of laboratory workflow. The NanoAmpli-Seq protocol is recommended for generation of high-quality 16S rRNA data. Future work will involve further simplification of the protocol and application of the workflow on the microfluidics chip rather than liquid handling robots. Automation and miniaturisation of the devices for sample analysis is necessary for high-throughput in-field analysis of environmental and clinical samples, which in turn may allow for autonomous analysis of environmental samples such as water or air but also clinical samples alike blood or urine.



## 5 Benchmarking of novel bioinformatics algorithms

### 5.1 Abstract

Accurate reconstruction of complete bacterial and archaeal SSU rRNA genes from third generation sequencing technology (i.e., Illumina – MiSeq, Thermo Fischer – Ion Torrent) is highly desirable but is often associated with multiple difficulties in laboratory and bioinformatics workflows. However, the recent advances in nanopore-based, single molecule sequencing technology (Oxford Nanopore Tech. – MinION<sup>TM</sup>) provides an opportunity for relatively inexpensive analysis of long DNA molecules in comparison to other single molecule sequencing technologies; i.e., PacBio, which requires significant capital investments. Initial results with sequencing on the MinION<sup>TM</sup> MKI device and Rolling Circle Amplification protocol indicated multiple problems with the sequencing data. Issues included high error rate, incorrect read orientation, and tandem repeat insertions in the sequencing; all of these issues had a significant impact on the sequence accuracy and community structure reconstruction of the tested mock community samples. This chapter describes two novel algorithms (chopSeq and nanoClust) that were developed to improve the accuracy of INC-Seq concatemerized data and allow for accurate *de novo* reconstruction of long 16S rRNA ribosomal genes.

This chapter is based on the publication:

Calus S.T., I.U.Z. and P.A.J., (2016). **NanoAmpli-Seq: A workflow for amplicon sequencing from mixed microbial communities on the nanopore-sequencing platform.** GigaScience, giy140, <https://doi.org/10.1093/gigascience/giy140>

Main contributions:

Figure 5.1 presents a general overview of the study design undertaken for analysis of 16S rRNA genes with use of the Rolling Circle Amplification (RCA) protocol. We conducted the first ever near-complete 16S rRNA gene sequencing (~1380bp) with use of the RCA based protocol. In this chapter, we describe and solve three main errors related to sequencing of concatamerized amplicons on the nanopore-sequencing platform. First, reads corrected with the INC-Seq consensus calling algorithm created long-ribosomal genes with incorrect intra-read orientation. Second, anchor alignment and consensus construction using INC-Seq resulted in insertion error formation inside of the ribosomal gene sequence, which had to be programmatically removed. These two issues were tackled by the development of the chopSeq software.

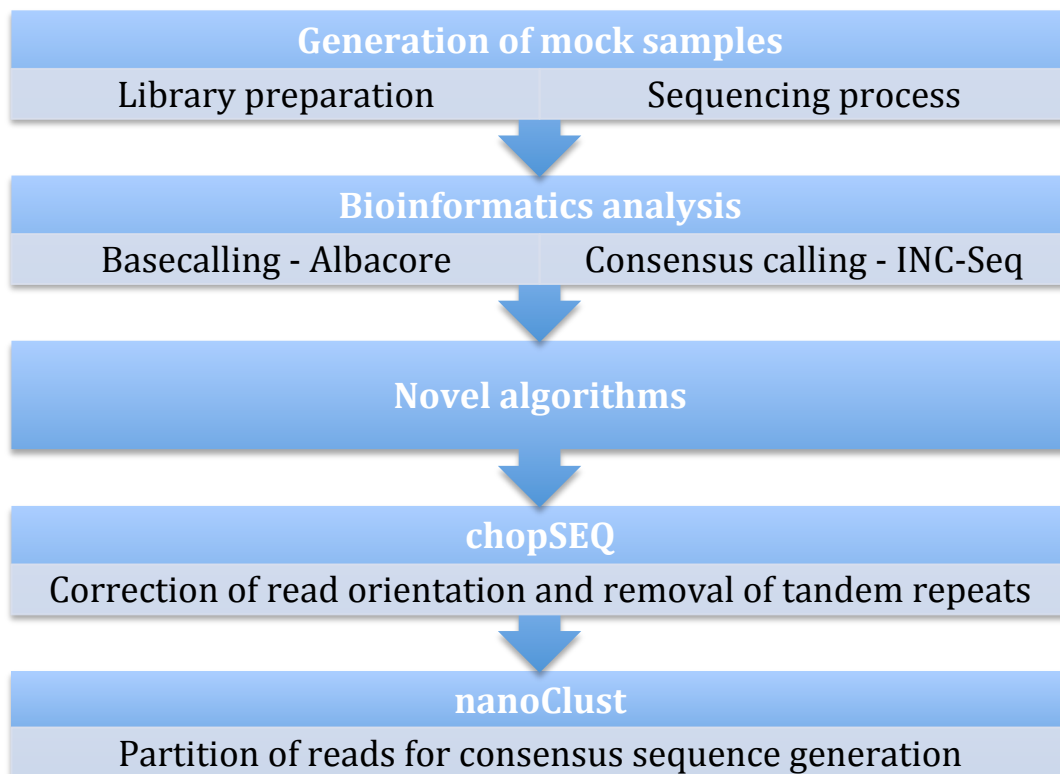


Figure 5.1 General overview of the study and steps (generation of mock samples and bioinformatics analysis) prior to the development of the novel bioinformatics programs, i.e. chopSeq and nanoClust. The figure indicates the order of the data processing and a brief explanation from each of the novel algorithms.

The third error indicated that standard *de novo* amplicon data analysis (i.e., Vsearch, Uclust) significantly underestimated the diversity of the mock samples. A study was designed on a simple bacterial mock sample made of 1 and 10 organisms only, which allowed for easier error detection, development of a novel OTU clustering method (i.e., nanoClust), and its benchmarking.

## 5.2 Introduction

The application of 16S rRNA gene sequencing for the identification of bacterial and archaeal phylogeny and taxonomy is the most common approach in environmental and clinical studies (Janda, 2007). The primary reason is that the 16S rRNA gene is ubiquitous in bacteria, and archaea, exhibits strong evolutionary signal, allowing for classification of microorganisms into taxonomic groups, and has conserved regions, which allow for simultaneous amplification of this gene from a diverse microbial populations. The 16S rRNA gene consists of conserved and hypervariable sequence regions.

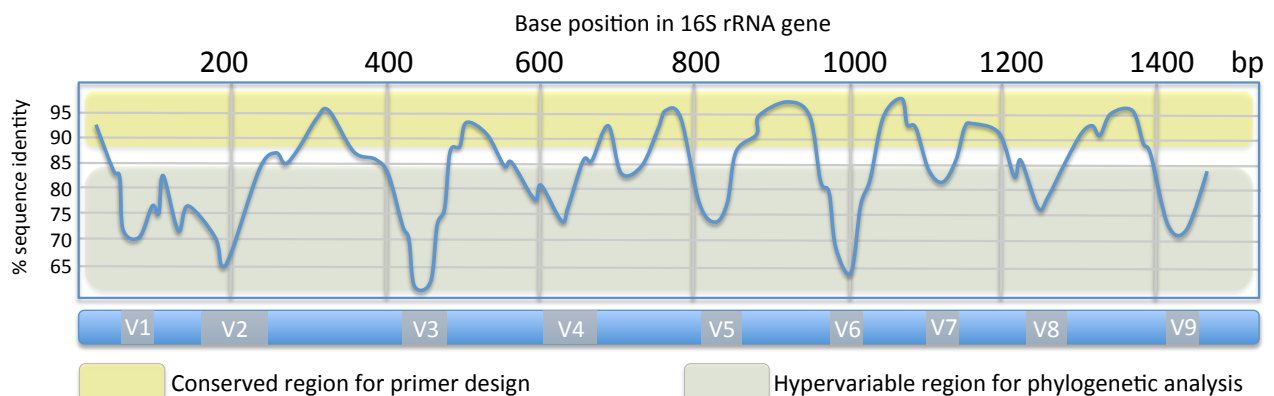


Figure 5.2 Relative sequence variability of different 16S rRNA gene regions. At the top of the plot, there is a scale that indicates sequence position in the gene (0-1500bp) while the bottom scale represents hypervariable regions better known as ‘V’ e.g. V1. The yellow region indicates conserved sections in ribosomal genes that are suitable for primer design (e.g., 8F or 1387R), while the green region indicates hypervariable sections of the ribosomal genes that are commonly used for sequence comparison and assignment of bacterial phylogeny (e.g., V3-V4).

Conserved sections of the ribosomal gene allow primer design to target all bacteria or archaea (Fig. 5.2), while hypervariable regions contain modifications in the sequence of bases that permit for determination of taxonomic classification e.g., order, family, genus. Commonly used sequencing technologies (MiSeq, Ion Torrent) for high-throughput analysis are often limited to small hypervariable ribosomal regions (i.e. 200-350bp) of the 16S rRNA gene and thus allow for identification of organisms down to family and sometimes to the genus level using publicly available and easily searchable, large databases such as SILVA or NCBI, but also can be processed with a reference-free approach also called *de novo*. Despite high-output of short amplicons, use of complete ribosomal genes is essential for high accuracy classification, especially as it relates to clinically relevant microorganisms (Fig. 5.3). As a result of technological developments and reduction in sequencing costs, analysis of the 16S rRNA gene has become standard practice for identification of novel bacterial and archaeal organisms.

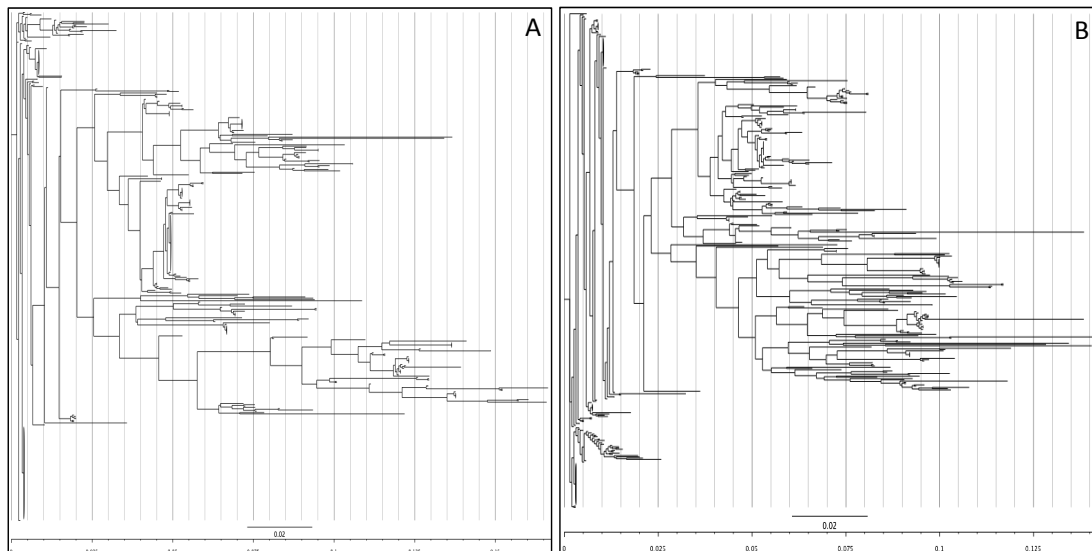


Figure 5.3 Two phylogenetic trees made of available *Legionella* spp. 16S rRNA genes (NCBI). The panel A plot uses only the V4 (300bp) region for phylogenetic placement of *Legionella* sequences while panel B plot is based on V1-V9 (1400bp) hypervariable ribosomal regions. Results indicate that while shorter ribosomal fragments can generate similar tree topology, they can significantly underestimate phylogenetic complexity.

Nonetheless, in the near future sequencing of small DNA fragments may become redundant as whole genome sequencing may become feasible at the same cost. Until whole genome sequencing becomes more cost effective, 16S rRNA gene sequencing will continue to be used widely for microbial profiling. Amplicon data generated with second-generation sequencing platforms (i.e., MiSeq, Ion Torrent) tend to be analysed using two main bioinformatics approaches; i.e., reference-based and a reference-free (i.e., *de novo*) approach. Each approach can be processed with multiple programs (e.g., Qiime, Mothur) for operational taxonomic unit (OTU) binning (Plummer, 2016). The fundamental difference between these two approaches includes the reliance on 16S rRNA gene databases for clustering of sequences in the phylotypes/OTUs. Application of reference-based 16S rRNA analysis is often applied for well-sequenced environments (i.e., human gut microbiome, thanks to the Human Microbiome Project) while *de novo* classification is recommended for under-sequenced environments (e.g. soil microbiome) primarily due to the poor representation of microbes from these environments in 16S rRNA gene databases (Turnbaugh, et al. 2007). *De novo* OTU construction remains the ‘gold standard’ method for analysis of 16S rRNA data because it does not rely on any underlying assumption of sequence similarity to known microorganisms.

The majority of work focussed on 16S rRNA gene sequencing on the nanopore platform has focussed on reference-based analysis (Benítez-Páez et al., 2016). Moreover, Oxford Nanopore Tech. released a GUI-based program named EPI2ME (initially WIMP – What’s In My Pot) that theoretically allows for real-time identification of microbial organisms down to genus level (Juul, et al., 2015 by mapping reads to an underlying database. However, reference-based analysis of raw

amplicon reads from the nanopore platform remains challenging due to the high sequence error rates (i.e., between 2-10%), which could result in misalignment to the reference database and misclassification of sequenced reads. Additionally, use of reference-based approaches precludes the detection of microorganisms that are not part of reference databases. Yet, nearly all amplicon sequencing data generated on the nanopore platform is still done using a reference-based approach. *De novo* analysis is not yet feasible with nanopore sequencing data, as the high error rates result in significant inflation in diversity estimates during OTU clustering. To this point, results generated with INC-Seq approach, which relies on generation and sequencing of amplicon concatemers (i.e., physically linked amplicon molecules) followed by python/Biopython based workflow to generate consensus sequences from concatemerised nanopore amplicons (Li. et al. 2016). The INC-Seq approach is based on PBDAGCON software released by Pacific-Biosciences, which has been widely used for analysis of PacBio data, involving circular consensus sequencing. This program was tested and validated on short (770bp) concatemerised 16S rRNA amplicons only. Use of the INC-Seq program returned significantly better results when compared to standard 16S rRNA amplicon sequencing, with average sequence accuracy of 97-98%. Nonetheless, use of the INC-Seq algorithm on long (1380bp) concatemerised amplicons produced multiple issues that had to be resolved before processing to reference-free and reference-based analysis.

To this end, laboratory protocol was re-designed (Chapter 4) for concatemerisation of amplicons and their data analysis. The novel laboratory protocol was based on a pre-existing workflow and uses the INC-Seq software for concatemer consensus calling. This chapter describes the development of two algorithms (chopSeq and nanoClust)

that were designed to improve the data quality of INC-Seq concatemerised reads and facilitate *de novo* clustering of full length 16S rRNA genes and increase sequence accuracy. These algorithms were validated on two mock samples consisting of 16S rRNA genes of a single organism and of 10 organisms. This protocol was tested on data that were analysed with use of two library preparation chemistries, which are 2D (R9 and R7 pores) and 1D<sup>2</sup> (R9.5 pore).

### 5.3 Algorithm 1: chopSeq

The chopSEQ program is python based and uses multiple open source libraries (os, argparse, sys, getopt, numpy, subprocess, and math) but also requires the Biopython package and libraries (Seq, SeqIO, and pairwise2). This software was designed to fix problems generated during consensus calling of short and long 16S rRNA reads, produced by the INC-Seq software. These issues include rearranging the direction of the reads to produce amplicons with the forward primer at the beginning and reverse primer at the end. INC-Seq generated over 95% consensus reads, which had incorrect sequence orientation with forward and reverse primers occurring at random locations along the length of the amplicon (Fig. 5.4). To fix that error, INC-Seq concatemerised reads were processed by the chopSEQ algorithm. Data must be provided (-i) in fasta format, primer sequences (forward -f and reverse -r) are necessary for the pairwise2 aligner from the Biopython package. It uses the provided primer sequences (accepts ambiguous bases; e.g. W or S) to detect them in various orientations in an INC-Seq sequence, then primer match scores for each orientation (e.g. forward complementary or forward-reverse complementary) are determined. Regions on the amplicon matching the primer sequences with the highest mean score are processed for removal of right-overhang (e.g., to the right of the primer 5' site) and its orientation is changed

to the start of the read. The same process is repeated with the reverse primer. Li et al. (2016) did not describe this type of issue, however, we noticed presence of these errors (wrong amplicons orientation), while analysing short amplicons (~725bp) from their (INC-Seq) data (PRJEB12294). The necessity for read re-orientation was initially noticed and became obvious when ~1380bp INC-Seq corrected reads were aligned (BLASTN) against the on-line 16S ribosomal RNA reference sequences (NCBI database).

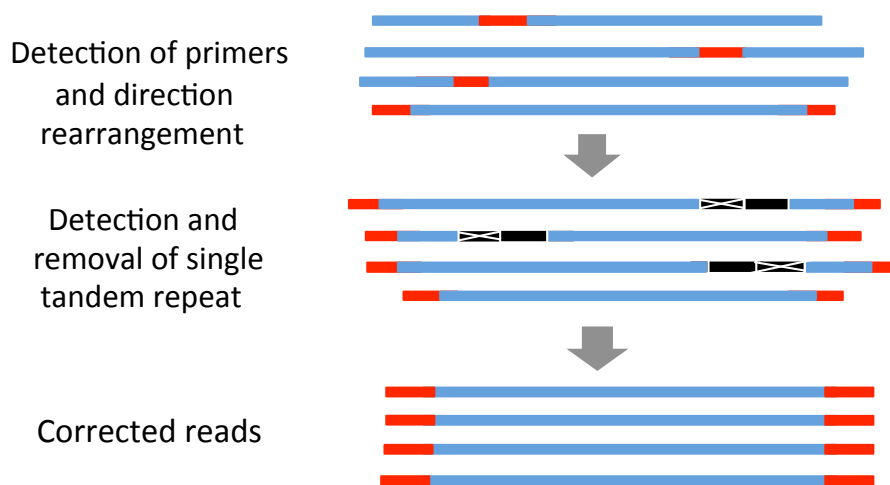


Figure 5.4 chopSeq requires INC-Seq corrected data (input) and uses the pairwise2 aligner for detection of primers inside of the wrongly oriented ribosomal genes. Correctly oriented reads possess insertion error (tandem repeats) in the position of read stitching. One of the repeats is removed with the use of the etandem (EMBOSS package) library. Finally, corrected reads are saved in fasta format or can be size filtered for subsequent *de novo* analysis of data with nanoClust.

The primary steps such as detection of incorrectly oriented reads, read fragmentation and stitching two read sections back in a correct way produces large insertions in the reads (Fig.5.4). These insertions are caused by repeated fragments, which come from both ends of the INC-Seq corrected molecules and were a result of the restricted window threshold used for intra-molecule alignment. These insertions vary in size and read position (Fig 5.5). Detection and removal of tandem repeats was achieved with the use of the etandem algorithm from EMBOSS open source analysis software



package (<http://emboss.bioinformatics.nl/cgi-bin/emboss/etandem>). The process of tandem repeat identification recognizes various features such as `tandem_min_repeat=10`, `tandem_max_repeat=350`, `tandem_threshold=10`, `tandem_mismatch=5`, `tandem_identity_threshold=85`. The tandem identity threshold is calculated with the highest value for short tandem fragments while identity threshold for longer reads gradually decreases (5%) and read size increases every 10bp (Fig. 5.6). A dynamic identity threshold was applied, as longer tandem fragments tend to have lower similarity between themselves in comparison to short molecules.

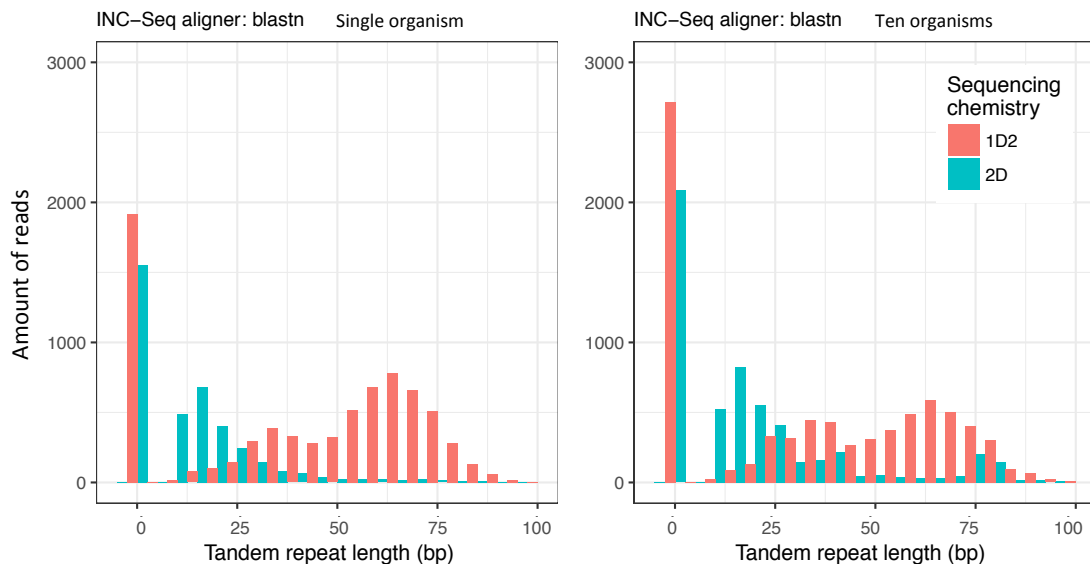


Figure 5.5 The size of tandem repeats for both chemistries (2D, 1D<sup>2</sup>) and single (blastn) switch from the INC-Seq algorithm. Results indicate that the majority of reads were in the correct orientation and did not contain tandem repeats (0 bp length). The older nanopore chemistry “2D” is characterised by a shorter length of tandem repeats (blue) while the newer chemistry “1D<sup>2</sup>” is defined by longer repeats (red). These differences may be related to changes in library preparation where two molecules (template, complement) are connected to each other with hair-pin adapter (2D) while lack of physical connection of template and complement between DNA strands (1D<sup>2</sup>) generates longer insertion errors after direction re-orientation with chopSeq.

Lack of the dynamic identity threshold to remove longer-tandem repeats (Figure 5.6) would result in a decrease in the quality of the data and affect *de novo* gene clustering. The final correction of reads with the chopSeq software includes a size filtration step

(e.g., between 1300 and 1450bp; i.e., -l 1300 -m 1450), applied with the standard string counting bash command 'len'.

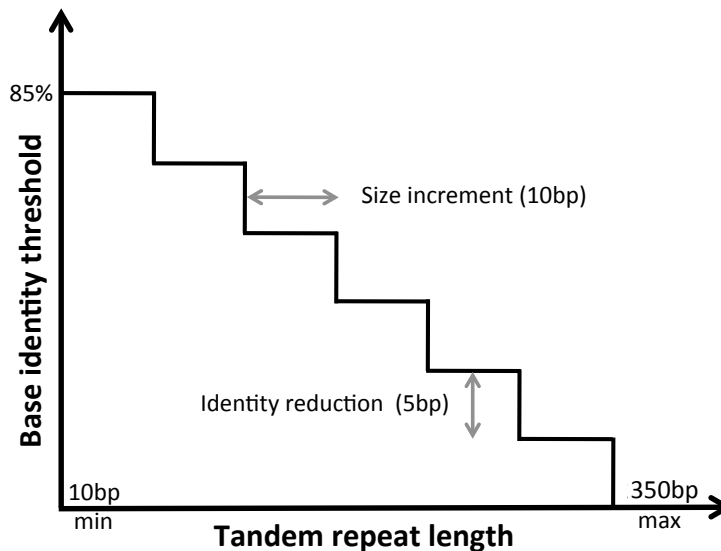


Figure 5.6 Gradual identity threshold reduction, which is applied for detection of tandem repeats (etandem – EMBOSS) inside of the INC-Seq corrected ribosomal reads. Base identity is reduced by 5% every 10bp; this step is necessary to increase tandem repeat identification. Application of dynamic base threshold recognition increased the quantity of etandem-corrected reads that in turn caused general reduction of error rates.

The chopSeq read-correction and size filtration can be run in silent or verbose modes (-v); the latter visualises the primers and tandem repeats. chopSeq based read re-orientation and tandem repeat removal allowed for nearly complete alignment of processed reads against reference sequences as compared to INC-Seq consensus reads (Fig 5.7) without affecting the overall sequence accuracy (Fig 5.8). The chopSEQ software is open source and available on the GitHub repository: <https://github.com/umerijaz/nanopore/blob/master/chopSEQ.py>. Data processed with chopSeq can be used for both reference-based and reference-free analysis of 16S rRNA genes. However, in this study, we focused our interest on reference-free (nanoClust) analysis, as this type of data analysis is not well studied for 16S rRNA analysis using nanopore data. Moreover, application of *de novo* amplicon analysis will allow for investigation of bacterial (16S rRNA) and fungal (ITS) marker genes

from complex, unknown or under-sequenced environments; use of which in turn could allow generation of long-reference genes for high-quality ribosomal databases such SILVA or GreenGenes.

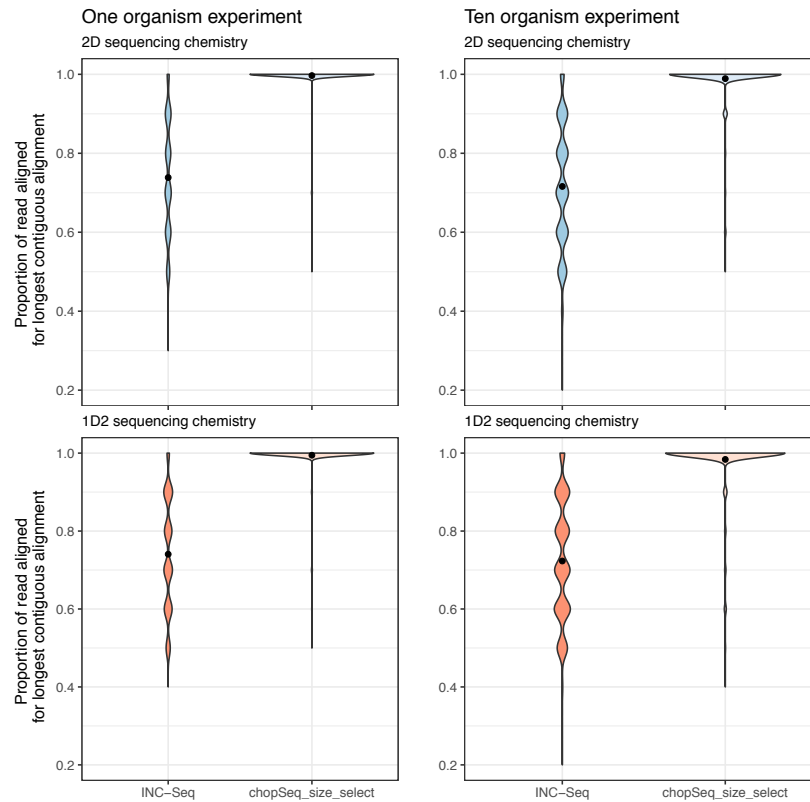


Figure 5.7 The proportion of reads aligned for the longest contiguous alignment. Results indicate the distribution of reads correctly aligning to the reference gene; i.e. 16S rRNA, containing one and ten mock samples, for 2D and 1D<sup>2</sup> sequencing technology for INC-Seq corrected only and INC-Seq and chopSeq correction. Addition of the chopSeq algorithm for read reorientation, significantly improved distribution of long reads aligning at nearly 100% when compared to pre-chopSeq data for both 2D and 1D<sup>2</sup> data.

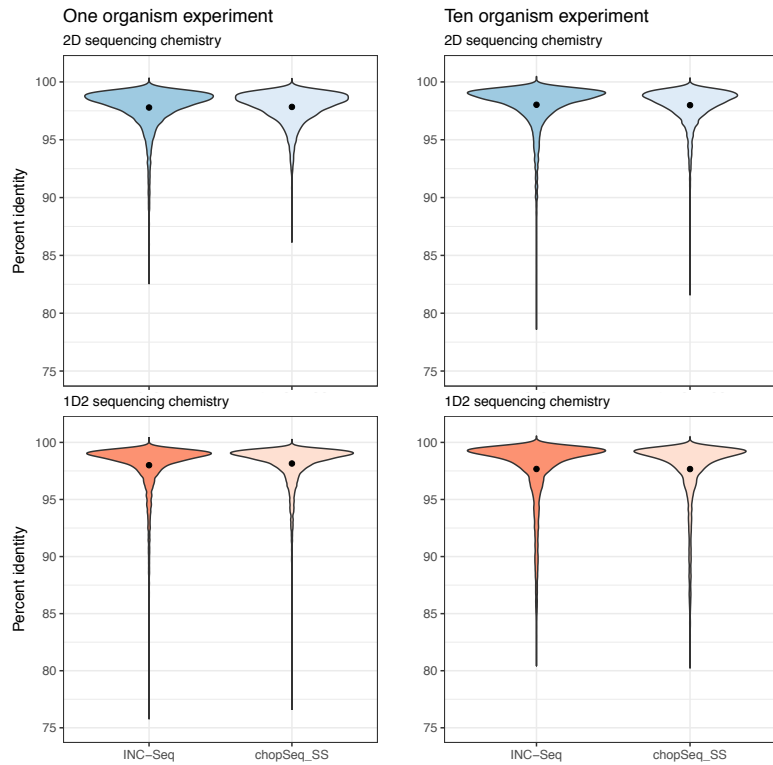


Figure 5.8 The percent identity of reads with INC-Seq and chopSeq algorithms for 2D and 1D<sup>2</sup> data, i.e. one and ten organisms. Results indicate that there is no significant difference between INC-Seq and chopSeq corrected data as the majority of reads have around 98% identity.

## 5.4 Algorithm 2: nanoClust

The nanoClust program was designed specifically for a novel type of *de novo* clustering approach of long 16S rRNA genes from an RCA-laboratory protocol. The residual 2-3% sequencing errors in INC-Seq and chopSeq processed reads does not allow for standard OTU clustering using tools such as VSearch. The remaining errors (2-3%) result in significant inflation in community diversity and results in a large number of singletons. To remedy this issue, we developed another approach for read-partitioning based OTU clustering and consensus calling within the nanoClust algorithm. The nanoClust algorithm works by initially partitioning the reads into multiple partitions of user-defined lengths. Post partitioning VSearch is used for

dereplication, singleton removal, and clustering of reads into OTUs for each partition. Post OTU binning, the partition with the maximum number of OTUs is selected as the optimal partition and read ID's for each OTU within the optimal partition are used to recruit full-length reads into OTUs. Subsequently, the full-length reads within each OTU are aligned using MAFFT-GINS-i and the alignment is used to determine the consensus sequence per OTU. The nanoClust algorithm is based on python libraries i.e. os, os.path, sys, getopt, time, datetime, numpy, subprocess and math but also uses open source libraries from Biopython package: Seq, SeqIO, AlignIO, AlignInfo, and pairwise2, as well as Vsearch and MAFFT-GINS-i (Figure 5.9). The nanoClust software is deposited and available for open source usage:

<https://github.com/umerijaz/nanopore/blob/master/nanoCLUST.py>

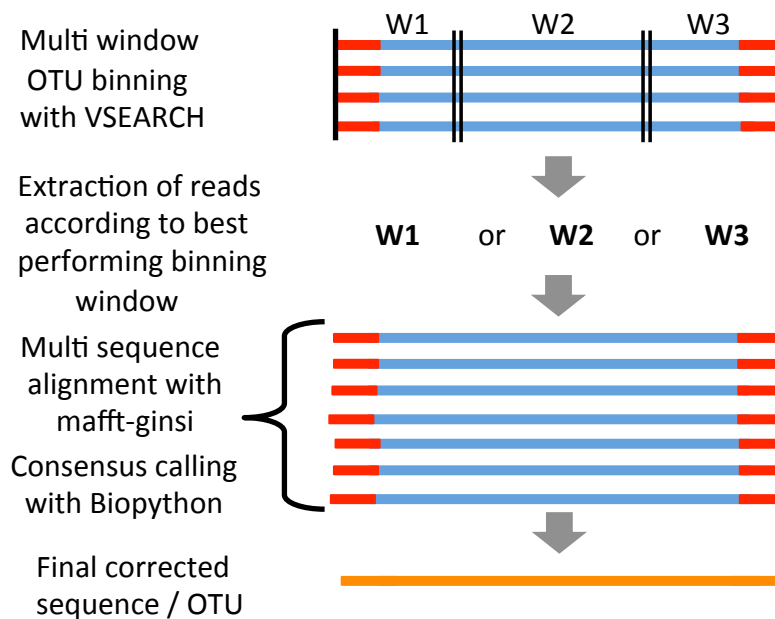


Figure 5.9 The nanoClust algorithm requires data in fasta format processed initially with INC-Seq, then chopSeq algorithms. The nanoClust algorithm uses Vsearch for partial gene binning (window 1, 2 or 3). The best performing window is used for multisequence alignment with a MAFFT-GINS-i algorithm for read consensus calling. Reads processed in this way create final corrected reads.

The partition size chosen for nanoClust and the number of reads used for consensus sequence construction can have a strong impact on both the number of OTUs and the

overall sequence accuracy. We experimented with these two parameters by varying the window partition size (and as a result number of partitions) used for read partitioning and the number of sequences used for consensus. A decrease in the partition size (i.e., increase in number of partitions) was associated with an inflation in community diversity as compared to the theoretical community diversity and a decrease in overall sequence accuracy (Fig 5.10). These experiments indicated that partitioning the 16S rRNA reads into two to three partitions allowed for both accurate determination of community diversity and the highest consensus sequence accuracy.

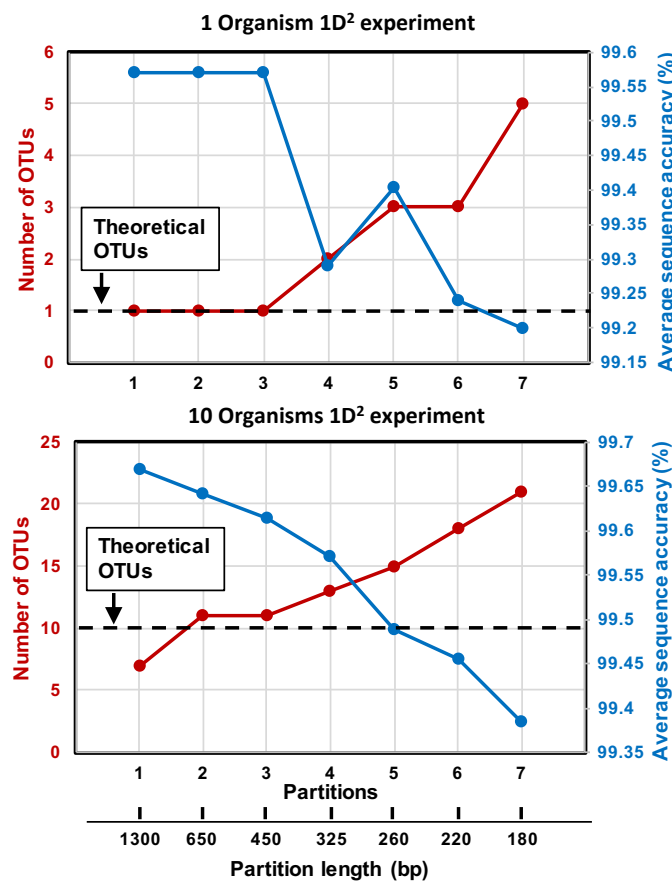


Figure 5.10 Plots representing the number of OTUs and average sequence quality when 16S rRNA reads are split into smaller partitions. Results indicate that higher length of the read corresponds with higher overall accuracy of final data; however, this can underestimate the amount of true OTUs. Moreover, increase of partitions reduces accuracy and increases number of OTUs. Results are reproducible for one and ten organisms while the best outcome of the OTU binning was achieved with two or three partitions.

Further, analysed the impact of the number of reads recruited per OTU for MAFFT-GINS-i alignment and consensus sequence construction. Figure 5.10 results were based on utilizing a total of 50 reads per OTU for alignment and consensus sequence construction. Tested the effect of five to 100 reads used for final error correction and found that the benefit of increasing the number of reads for consensus sequence construction diminishes above 50 reads and stabilizes at a sequence accuracy of ~99.5%, while fewer than five reads are required for sequence accuracy of 99% (Figure 5.11). The combination of three partitions and 50 reads for consensus sequence construction resulted in highly accurate estimates of community diversity and an average sequence quality of 99.5% for both the one and ten organism mock community samples for both 2D and 1D<sup>2</sup> sequencing chemistry (Fig 5.12).

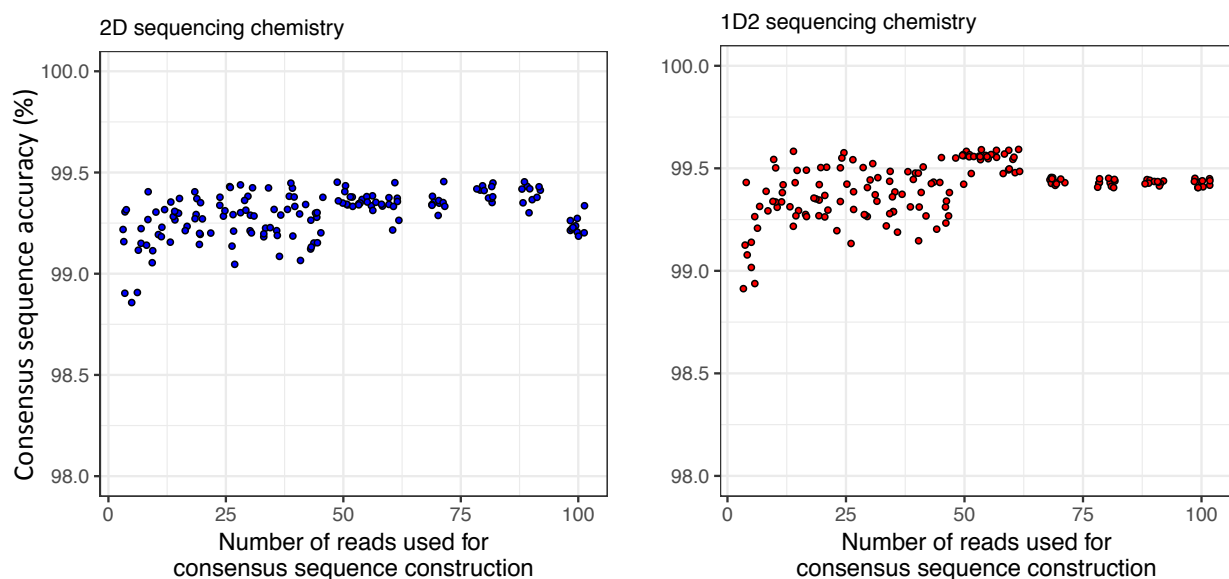


Figure 5.11 Consensus sequence accuracy for *Listeria monocytogenes* reads generated with two sequencing chemistries (2D and 1D<sup>2</sup>) and a various number of reads for 16S rRNA gene consensus construction. Results indicate that the lowest accuracy is generated with 3-10 consensus reads while the highest accuracy consensus sequences were achieved with around 50 reads (+/-10). Nonetheless, use of a very high number of reads (e.g. 100) for consensus sequence calling resulted in slightly lower accuracy. Increase in error rate with the high number of reads occurs due to indel errors, which appear in positions of homopolymers. Use of a higher amount of reads assures multisequence aligner that occurring errors are authentic and in turn retains them as true variations. For this reason, the nanoClust algorithm was optimised to use up to 50 ribosomal reads for final multisequence alignment correction with MAFFT-GINS-i.

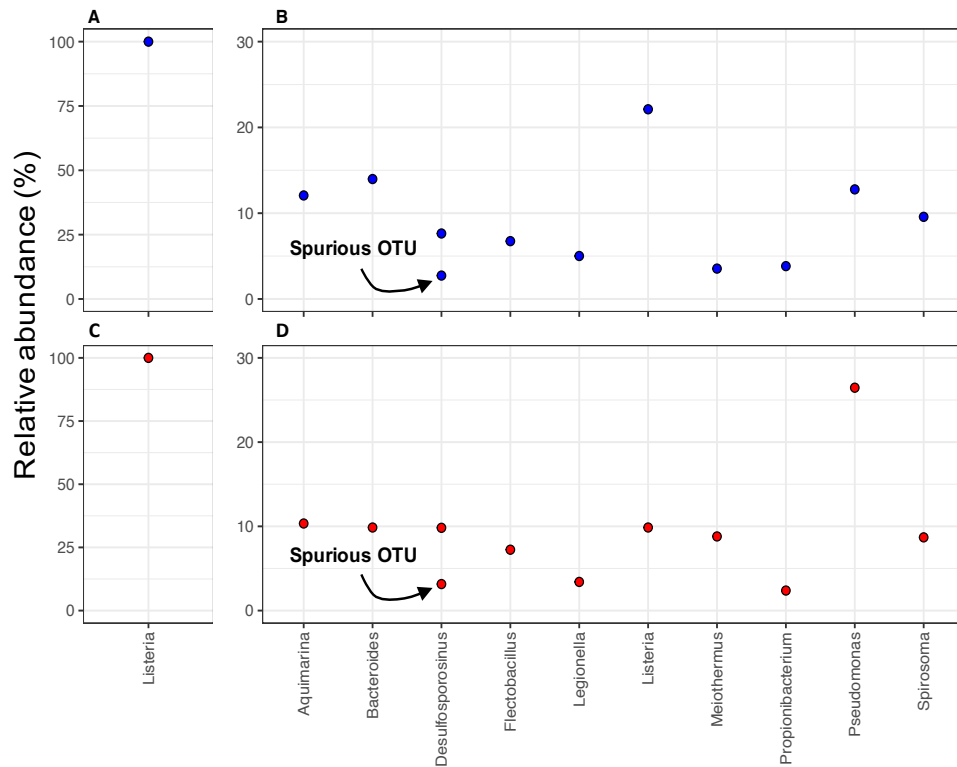


Figure 5.12 Plots indicate amount of OTUs detected for certain genera of bacteria within mock samples and their abundance for single a organism: A and C, mock samples: B and D adequately for 2D and 1D<sup>2</sup> sequencing chemistry. Results indicate that relative abundance for single organisms was 100% and did not produce false positive signals. While samples containing mock community with ten organisms are characterised by skewed microbial distribution e.g. *Pseudomonas aeruginosa* or *Listeria monocytogenes*. Additionally, in both cases 2D and 1D<sup>2</sup> chemistries the *Desulfosporosinus Orientis* generated spurious OTUs, which may be caused due to homopolymer regions in the ribosomal gene structure.

## 5.5 Conclusion and future work

In this chapter, we present two novel algorithms (chopSeq and nanoClust) for additional analysis of long 16S rRNA (INC-Seq corrected) reads, produced with the nanopore sequencing platform. The first program was able to resolve all issues related to incorrect read orientation removal of tandem repeats but also filtering out (size selection) unwanted reads that do not conform to the expected amplicon length. The reads with correctly reconstructed orientation still contained many mismatches (between 1 to 5%) and indels that could not be used directly for OTU construction. The second program was this designed for precise OTU binning, which prevented



under or overestimation in the community diversity. Construction of the nanoClust algorithm was critical as it allowed for accurate diversity estimation and significantly reduced overall error rates (~0.5%). The majority of the errors present in final OTU consensus sequences were caused by homopolymer errors and could not be resolved by increasing the number of reads used for consensus construction. The nanoClust software was optimised to generate consensus reads from 50 reads per OTU bin. None of the previously designed programs for amplicon analysis (e.g. Mothur or Qiime) were constructed to detect wrongly oriented reads as presented with chopSeq, that allowed subsequent OTU binning with nanopore sequencing data.

The detailed bioinformatics workflow is deposited on the protocols.io website (<http://dx.doi.org/10.17504/protocols.io.u25eyg6>). The algorithms presented here are the first officially released version of the software, however, we have identified multiple opportunities to further improve these tools. One modification to the chopSeq algorithm would be the application of multi-threaded data analysis for parallel multi-sequence processing. The current version of the program analyses single reads at a time, which significantly extends the time of analysis. Due to a data analysis bottleneck, the chopSeq algorithm processes a thousand reads in a matter of hours. Nonetheless, future improvements of library preparation and nanopore sequencing chemistry will increase total data output, which could further slow down overall data analysis of the 16S rRNA concatemer reads. Slow processing of reads will cause delays in data processing. The second algorithm (nanoClust) will be improved as well, however, subsequent optimisation is mainly related to the analysis of a complex environmental sample (Chapter 6). The presented optimization of the nanoClust algorithm was conducted using simple bacterial mock samples (1 and 10

organisms) while issues such as multi-species binning may occur when analysing environmental samples with closely related organisms. To circumvent this issue the library preparation protocol for NanoAmpli-Seq itself may be optimized to allow for molecular tagging of individual amplicon molecules prior to PCR amplification and concatemerization. This process would essentially eliminate the need for sequence similarity based clustering prior to consensus sequence construction, but rather rely on clustering of INC-Seq consensus reads based on the molecular tags, thus eliminating the possibility of binning sequences from different but closely related organisms into a single OTU.

## 6 Comparison of Oxford Nanopore vs. Illumina

### 6.1 Abstract

The initial objective of this PhD was to use a portable DNA sequencer from Oxford Nanopore for rapid analysis of microorganisms present in tap water. Nonetheless, preliminary study unravelled high level of error rates, which are discussed in Chapter 3. For this reason the development of the NanoAmpli-Seq protocol was necessary and has resulted in a significant reduction of sequencing errors for nanopore 16S rRNA gene sequencing data. Moreover, development of two novel algorithms (i.e., chopSeq and nanoClust) allowed for accurate *de novo* identification of bacterial organisms from simple mock communities. Precise identification and classification of the organisms is essential to characterize natural and complex bacterial communities. This chapter describes further modifications to the NanoAmpli-Seq library preparation and related software. Advanced modifications to the protocol included: addition of sample barcoding and de-multiplexing steps as well as multi-threaded analysis. Addition of barcodes enhanced performance of the PCR assay and allowed for sequencing of multiple samples on a single flow cell. Additionally, chapter presents the results of a study comparing 16S rRNA data generated with MinION MKIb against Illumina whole genome shotgun sequencing from complex environmental samples. Results of this analysis allowed for accurate estimation of microbial communities in low diversity drinking water samples and proved that this protocol can be used for rapid identification of pathogenic organisms from complex samples. From the biological perspective this chapter aimed to detect all microorganisms present in the drinking tap water samples. The main reason

for that is precise and rapid identification of pathogenic organisms with NanoAmpli-Seq protocol, which could be applied in drinking water filtration facilities.

Original contributions:

This experiment was conducted to test performance of the newly developed NanoAmpli-Seq protocol and its accompanying software on drinking water samples. This study is the first time ever reporting comparison of environmental samples between nanopore concatemer corrected reads against Illumina WGS data. None of the previously designed and published journal articles (i.e. INC-Seq or R2C2) released results or data indicating how their consensus calling algorithms perform on real samples for 16S rRNA identification (Li et al., 2016; Volden et al., 2018). Additionally, none of the previously released nanopore data based on RCA included barcodes for sample multiplexing. Moreover, INC-Seq and chopSeq algorithms were updated by multithreaded GNU parallel wrapper, which in turn significantly reduced time needed for data analysis. Thanks to advancements in the NanoAmpli-Seq protocol, its reliability, reproducibility and fast sample turn over the samples can now be prepared according to the NanoAmpli-Seq workflow, sequenced overnight and results can be generated in around 8-10h. That makes it the most accurate and reliable protocol for rapid identification of bacterial organisms (i.e. 16S rRNA) when compared to other methods; i.e. INC-Seq and R2C2 or even Illumina, V4 16S rRNA or WGS sequencing.

## 6.2 Introduction

The first known epidemic due to drinking water microorganisms was Typhoid fever caused by *Salmonella typhi* and *S. paratyphi* in 430-424 BC, which caused death of around 30% of the population of Athens, Greece (Papagrigorakis et al., 2007). The disease was probably transmitted by poor hygiene and public sanitation conditions and the resulting contamination of drinking water sources. This theory was confirmed in 2007 by a study that analysed dental pulp remains found at a burial place dated to the time of the outbreak. Bioinformatics analysis of sequenced teeth detected DNA reads homologous to the bacterium causing Typhoid fever. It is also believed that Typhoid fever caused on English colony (Jamestown, Virginia) of 6000 settlers to collapse between 1607 and 1624 (Earle et al., 1979). Cholera is another water-borne disease, caused by the gram-negative, comma-shaped, facultative anaerobic bacterium *Vibrio cholerae* (Sharma et al., 2003). This disease emerged within the Indian continent as a result of poor living conditions and presence of stagnant water pools, which were ideal conditions for cholera to thrive. The first cholera pandemic erupted near the Bengal region of India through 1817-24 and spread, affecting almost every Asian country. Trade routes and technological advancements at the beginning of 19th century allowed spread of the disease from India to all over Asia and Russia. As the disease spread across the rest of the Europe, quarantine failed and the disease was reported across Britain and then spread further to North America and finally to the rest of the world. Lack of hygiene and knowledge about sanitation caused expansion of cholera but also diseases such as Typhoid and Typhus (Lilienfeld et al., 1984). Outbreak of cholera was the greatest threat caused by contaminated drinking water. The best known cholera outbreak was in 1854 and was called the Broad Street Cholera Outbreak that erupted near Broad Street in the London district of Soho. This

outbreak was well known because of the study carried out by physician John Snow and his discovery that disease was spread via contaminated water (Snow, 1855; Cameron et al., 1983). At this time, the germ theory proposed by Louis Pasteur in 1861, was not known yet, resulting in lack of awareness of the mechanisms behind disease transmission. However, Snow was sceptical of the then well known miasma theory, which assumed that cholera was caused by pollution or noxious substances from “bad air”. The burden of evidence led him to theorize that the disease was not spread via foul air but by contaminated water with the identification of public water pump at Broad Street as the source of the disease. His studies were enough to convince the authorities to disable the pump by removing its handle. This action caused rapid decline of infection and the end of the outbreak. This finding had a significant impact on public health and facilitated the construction of improved sanitation and drinking water treatment facilities in the 19th and 20th centuries.

Recently, rapid and accurate detection of microorganisms from complex, low volume samples is one of the most desired microbiological techniques for prevention of disease spread (Deamer et al., 2000). Comprehensive assessment of microbial taxonomy and correlated with it toxin or antibiotic resistance genes is difficult and may be time consuming, costly or inaccurate (Ashton et al., 2015). Multiple different companies (i.e. eDNA) are working on precise and rapid detection of low volume pathogens from difficult sources (e.g. blood, drinking water). Nonetheless, at the moment there is no single, optimal detection method that would fulfil all the requirements for real-time, meticulous biosensing technology for clinical or environmental samples.

### **6.3 Drinking water samples**

This chapter was designed to test reproducibility of results between Illumina, Nextera XT and Oxford Nanopore, NanoAmpli-Seq workflows. Moreover, chapter describes the analysis of environmental samples, which was necessary to understand errors related to real, complex microbial communities. Moreover, analysis of multiple samples with use of two sequencing technologies will allow for comparison of reproducibility of the newly developed NanoAmpli-Seq protocol against a gold standard Illumina method. Evaluation of these factors was necessary prior to wide spread of the protocol for broad analysis of bacterial samples.

#### Sample collection

Multiple drinking water samples were collected from Scotland, United Kingdom. Maria Sevillano Rivera, a PhD student working within School of Engineering, University of Glasgow undertook the initial collection of samples for her experimental work in detection of antibiotic resistance in drinking water pathogens. All the samples (between 50 to 100L) were subjected to filtration with use of 0.22µm Sterivex filters (Millipore, Z359912) and a peristaltic pump (Watson Marlow 323). DNA from microorganisms collected on the filter was extracted using the FastDNA™ SPIN Kit (MP Biomedicals, 116560200) with use of the FastPrep-24™ bead beating machine (MP Biomedicals, 116004500). Bacterial DNA was eluted in 50µl of molecular grade water and subjected to Qubit HS dsDNA assay (Thermo Fisher Scientific, Q32854), then stored at -20°C freezer until needed.

### Sample preparation

The extracted DNA was initially diluted down to 0.2ng/μl and prepared according to the Nextera XT DNA library preparation kit (Illumina, 15031942) for whole genome shotgun sequencing. The library preparation process was followed according to the manufacturer's protocol and prepared DNA libraries were sent to the Centre for Genomic Research - University of Liverpool. The sequencing process was conducted on the Illumina, HiSeq 2500 instrument with use of the 2x250bp rapid output workflow. The remaining DNA from the stock samples were diluted down to 1ng/μl and prepared by the previously designed NanoAmpli-Seq laboratory protocol (Chapter 4) with a few exceptions described in Fig. 6.1. e.g., addition of indexes sequence (ID) to the PCR primer for sample multiplexing.

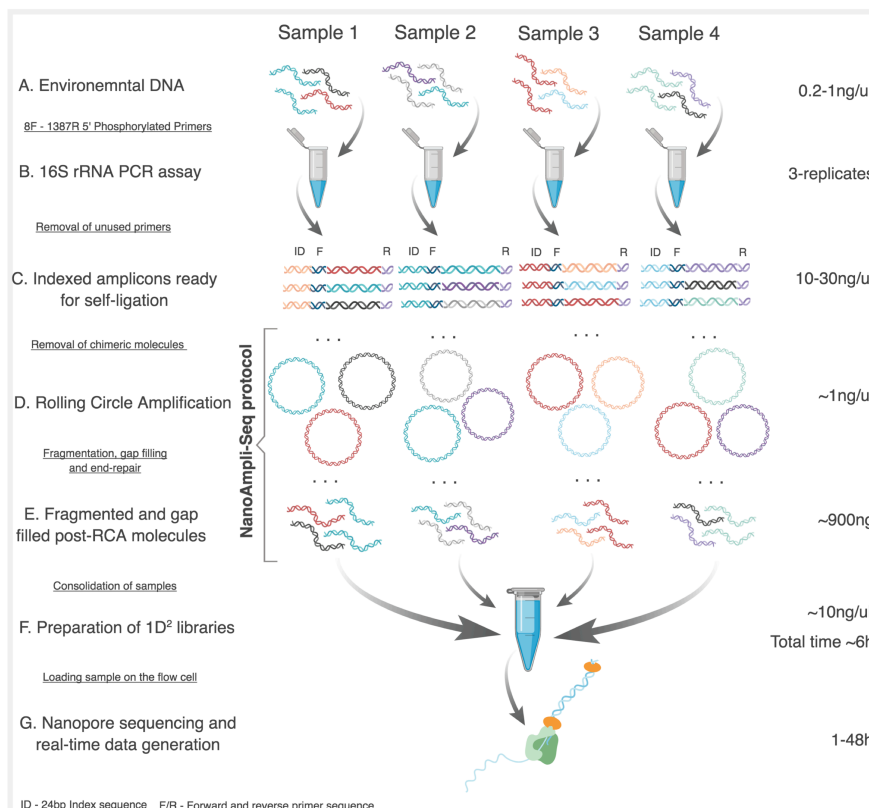


Figure 6.1 Experimental workflow of improved NanoAmpli-Seq workflow. A) Experiment included collection of drinking water samples and extraction of DNA. B) Amplification of 16S rRNA genes with indexed primers i.e. 8F and 1387R. C) Amplicons were purified, self-ligated and D) processed according to the TruePrime RCA protocol. The subsequent steps were followed according to the initial NanoAmpli-Seq workflow; i.e. E) fragmentation of long DNA molecules, F) preparation of 1D<sup>2</sup> sequencing libraries, G) sequencing process and real-time data generation.



### Data analysis

The 48h sequencing run generated 11,281 failed reads and 33,601 passed reads for all 4 drinking water samples. The Illumina sequencing platform generated paired end reads: 151,398 for sample 1; 2,797,359 for sample 2; 3,448,671 for sample 3 and 332,435 for sample 4. Data was generated in a FASTQ format (there was no need for local data basecalling) and was initially subjected to quality trimming of nucleotide bases with a Phred scores below Q30, using Trimmomatic (Bolger, 2014). Results of the quality-trimmed data for both nanopore and Illumina were visualised with FastQC (Fig. 6.2 and 6.3).



Figure 6.2 Phred scores for high quality nanopore reads with the improved 1D<sup>2</sup> sequencing chemistry and R9.5 flow cell. Results indicate that raw data had higher Phred quality scores when compared to the previously described sequencing chemistry in Chapter 3. The quality scores now exceeded Q30 for a few hundred bases, while the majority of the data is still below Q20, which means 1 error in 100bp. This improvement in data quality was achieved thanks to the most advanced version of the basecalling algorithm, which used Recurrent Neural Network and machine learning approaches instead of Hidden Markov Models.

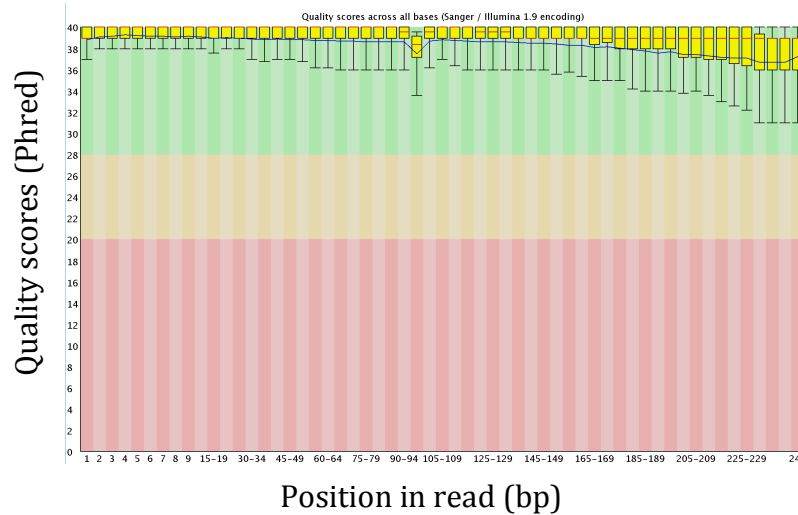


Figure 6.3 Phred scores for Illumina, WGS data after quality trimming with Trimmomatic. Results indicate that reads have sequence quality above Q30, which is 1 error in 1000bp for Q30 scores and 1 error in 10,000bp for Q40 quality scores. Difference in data quality between ONT and Illumina data is in a range of 10 to 100 fold, which makes Illumina data highly reliable when compared to raw nanopore data. Despite the reproducibility of Illumina data, the difference in size and cost of these two sequencing machines is significant i.e. ~\$1000 for MinION and ~\$200,000 for HiSeq 2500.

Subsequently, extraction of partial 16S rRNA reads and reconstruction of them from metagenomics data was performed with Mapping-Assisted Targeted-Assembly for Metagenomics; i.e. MATAM (Pericard et al., 2017). This program used shotgun metagenomics reads for reference-based assignment and identification of fragmented ribosomal genes that are aligned with use of SortMeRNA to the SILVA database (Fig. 6.6). Aligned reads were compared to form an overlap graph, which latter was compressed and assembled to form contigs and then scaffolds for full-length 16S rRNA reconstruction. The ribosomal reads extracted in such a way had the highest recovered fraction of ribosomal genes when compared to other software e.g. REAGO, SPAdes, EMIRGE (Pericard et al., 2017). The main objective and results of this

The analysis of data generated with the nanopore platform was initially subjected for basecalling with use of Albacore software, 1D<sup>2</sup> option. Generated data was in FASTA format and high quality reads from the ‘pass’ folder were processed with INC-Seq

software (Fig. 6.4). Nanopore data does not require use of quality trimming algorithms when compared to the Illumina workflow. To increase the quality of reads the INC-Seq software used consensus calling of concatamerized reads. The reads corrected in this way have between 93% and 98% identity and could not be used for alignment-based identification of microorganisms due to errors and wrong orientation of reads. Use of the chopSeq algorithm was necessary to correct orientation of the reads and demultiplex the data for each index. The chopSeq program was improved to detect index sequences provided in a separate 'csv' file. Demultiplexed data was processed with the nanoCLUST algorithm and standard window thresholds (i.e. 0,350,350,900,900, -1). Additionally, both INC-Seq and chopSeq algorithms were updated with use of GNU parallel-based wrappers. The algorithms have an optional switch that allows for multithreaded data processing. That in turn allows reducing the time of consensus calling and reading reorientation up to 10-fold, which is highly-desirable during analysis of clinical samples.

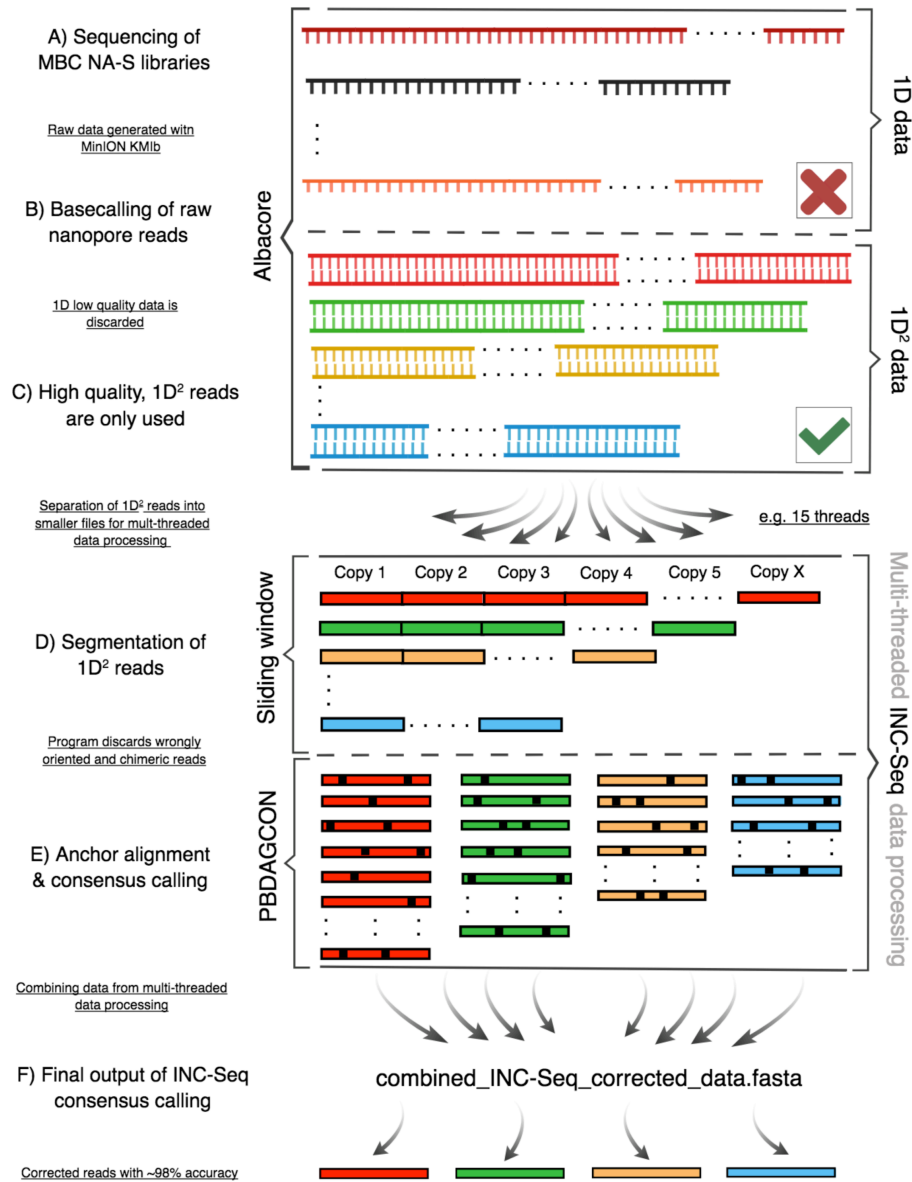


Figure 6.4 Schematic representations of the nanopore data analysis with Albacore and INC-Seq algorithms. A) The NanoAmpli-Seq libraries were sequenced on MinION MK1b. B) The basecalling process took place and converted the HDF5 data into FASTA reads. C) Only high-quality 1D<sup>2</sup> reads were used further as the low quality data would generate false positive results. Reads were split into multiple files and processed on multiple threads; e.g. 15. D) Use of INC-Seq algorithm for segmentation of reads. E) Reads were anchor aligned with use of blastn, graphmap or poa algorithms and consensus sequence was generated. F) Finally reads combined into a single file and then were ready for subsequent analysis.

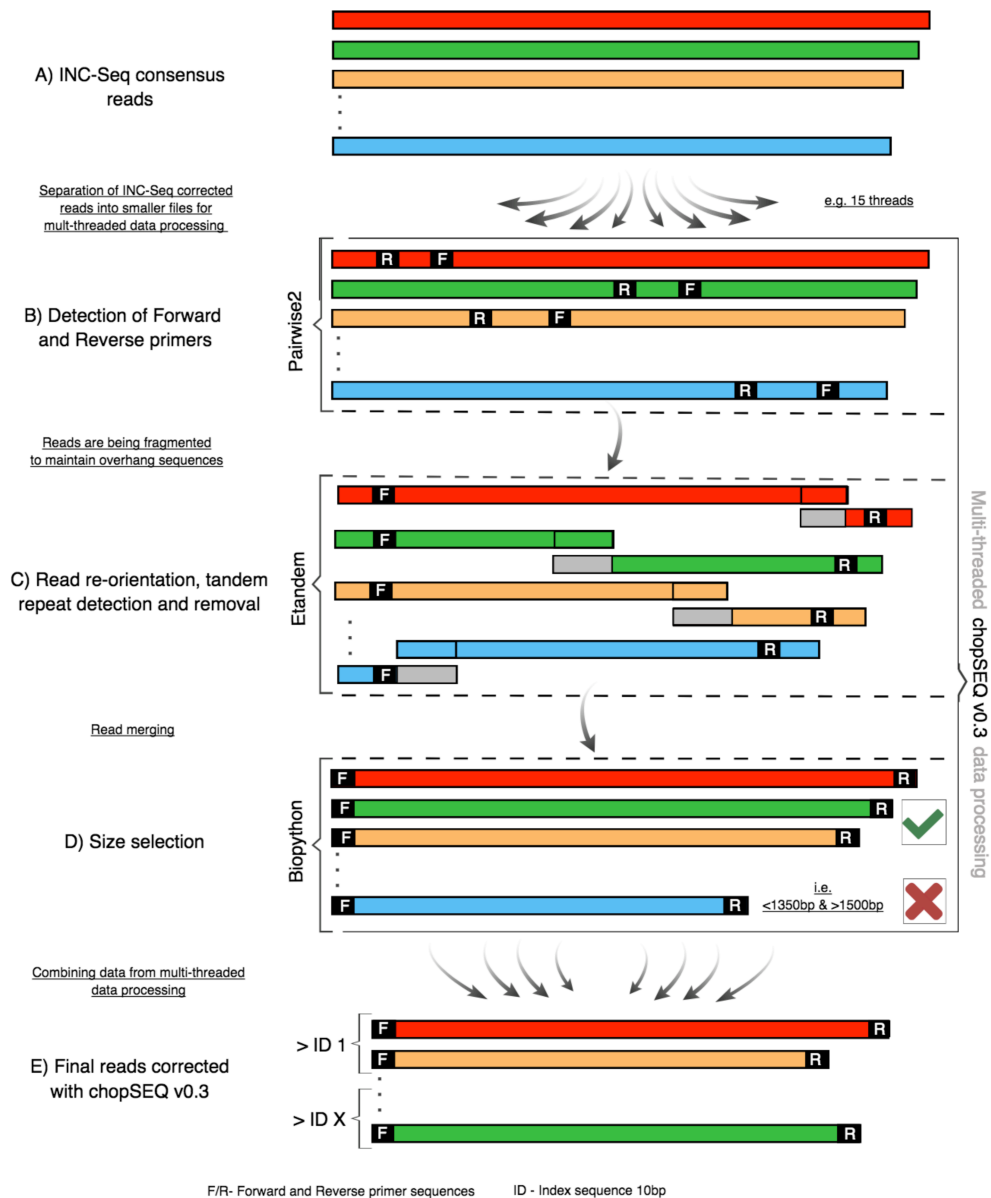


Figure 6.5 Schematic representations of the nanopore data analysis with the chopSeq algorithm. A) Data corrected with the INC-Seq software is necessary for the chopSeq data input. Reads are separated into multiple sub-files, which allow for multithreaded data analysis. B) Detection of forward and reverse primer sequences is applied with the pairwise2 algorithm. C) Read reorientation and removal of tandem repeats is performed to correct wrongly assembled 16S rRNA reads with INC-Seq. D) The final step filters out too short or too long reads and combines the reads with the correct gene size. E) Reads generated in such a way have their index sequences recognised and are renamed for further nanoCLUST data analysis.

The length of the 16S rRNA genes from both sequencing technologies was compared against each other (Fig. 6.7) and indicated that Illumina data followed by analysis with MATAM created significantly shorter reads when compared to the NanoAmpliSeq protocol.

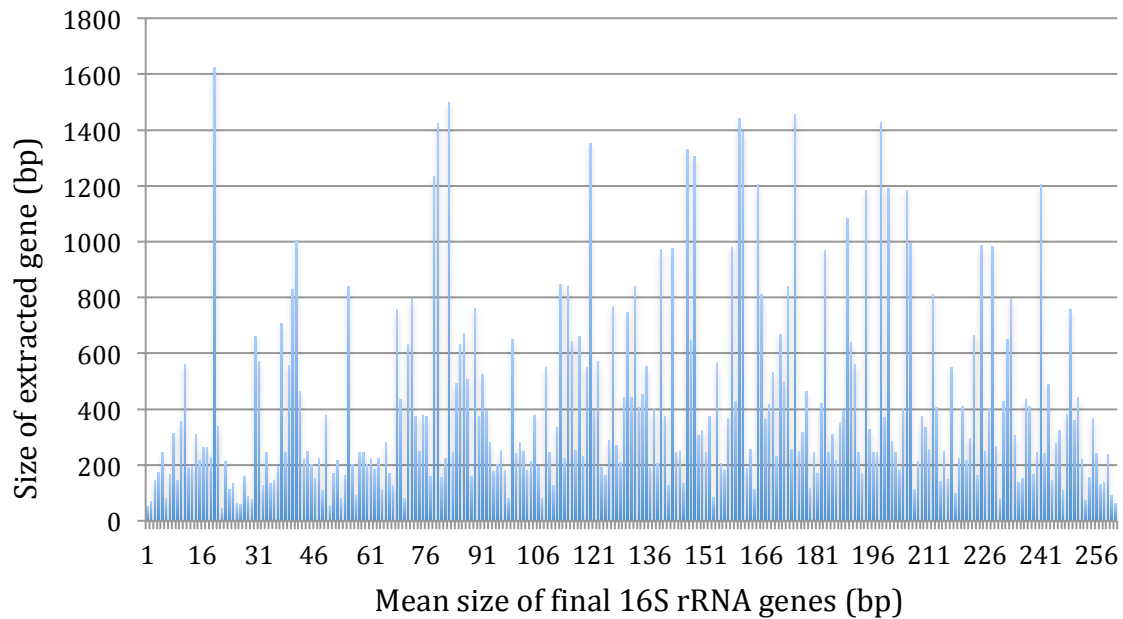


Figure 6.6 Chart representing the size distribution of 16S rRNA genes extracted with MATAM prior to the final assembly. Results indicate that the majority of the extracted reads were short i.e. 150-600bp. Only 4.7% of reads were longer than 1200bp, 12% of reads were longer than 600bp while 87.5% were shorter than 600bp. The subsequent step in MATAM analysis was assembly for contig and scaffold generation to increase the size of the short ribosomal reads.

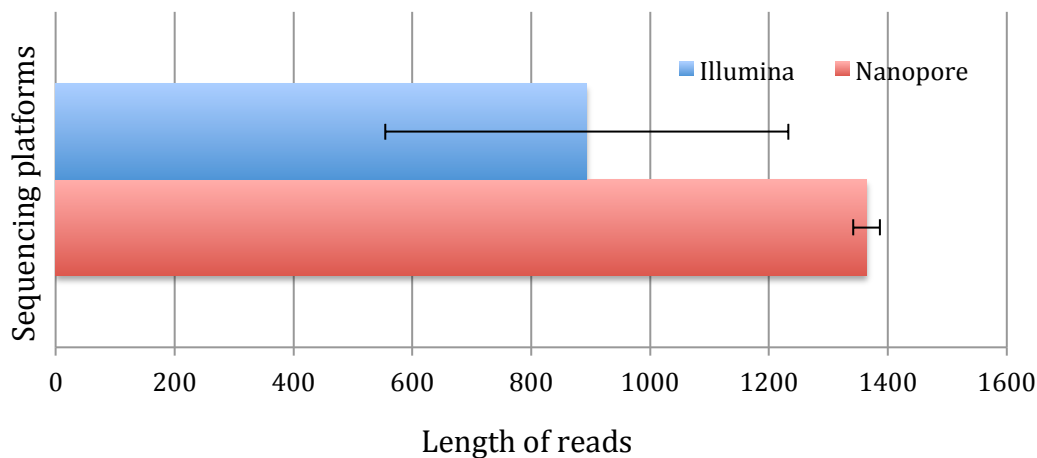


Figure 6.7 Length of the 16S rRNA genes after extraction and assembly with MATAM and the NanoAmpli-Seq workflow for all 4 samples. Results indicate that 16S rRNA genes generated with Illumina data had significantly lower mean length with very high standard deviation i.e. 37%. However, nanopore data was characterised by near-complete size of the ribosomal gene with very low error bars; i.e 1.6% of the total size. Despite initial lower quality scores the ONT data become superior to high quality Illumina reads.

The low quality reads from the nanopore platform were improved to the level of ~99.5% accuracy, which allows for higher resolution analysis when compared to shorter Illumina 16S rRNA reads. Subsequently, data from both platforms were

compared against each other with use of Mantel test (R Studio), (Manly et al., 2016). The results of the Mantel test observed correlation at a level of 0.989, which indicates significant similarity between the communities from the two different sequencing technologies (i.e. Illumina HiSeq 2500 vs. ONT MinION MKIb) and various library preparation techniques; i.e. Nextera XT vs. NanoAmpli-Seq, (McElhoe et al., 2014). This test was used to verify the distance between microbial community structure to evaluate the reproducibility of aforementioned sequencing platforms and laboratory methods. Another type of statistical analysis called Procrustes, performed with R Studio and was used for verification of the community structure by attempting to transform the data into a state of a multidimensional superimposition. Results of this test indicated significant similarity between two datasets by accepting the null hypothesis ( $p=0.998$ ). The results of the data analysis with two sequencing technologies and two different laboratory protocols (i.e. Illumina, Nextera XT and ONT NanoAmpli-Seq) indicated that addition of a fusion indexing strategy to 8F forward primer worked well. Moreover, demultiplexing of the data with the newest version of the chopSeq software was successful and allowed for separation of reads. However, one of the 24bp long indexes contained two homopolymer sequences, i.e. “GGGG” and “AAAA”. These four-mer sequences were not successfully corrected during INC-Seq and chopSeq data consensus calling. That in turn caused partial loss of reads containing the indices with these two homopolymers. Failure in demultiplexing step was caused due to high error rate caused by indels. That in turn did not allow for correct identification of the index sequences by the chopSeq algorithm. Nonetheless, the remaining reads containing other indexes were successfully demultiplexed.

## 6.4 Conclusions and Future Work

The first set of samples included 4 drinking water specimens that were sequenced with the NanoAmpli-Seq protocol and compared to high quality Illumina, whole genome shotgun data. The initial assessment of the study showed that the NanoAmpli-Seq protocol and its related softwares are able to achieve important enhancements for analysis of 16S rRNA genes from complex environmental samples. Results indicated that the NanoAmpli-Seq protocol is superior to WGS for 16S rRNA based analysis, Illumina sequencing for full-length 16S rRNA analysis and could be used for accurate phylogenetic analysis and enhancement of high quality ribosomal databases. In this study we were also able to demonstrate that sample-indexing method was successfully applied with 24bp. However, the length of the barcodes can be shortened down to 10bp in the future. Additionally, the barcode sequences require improvements to avoid presence of short homopolymers, which will increase demultiplexing efficiency and will allow for a higher percentage of data retention. Preservation of a large amount of high quality INC-Seq and chopSeq corrected reads is significant for detection of low abundant organisms. Especially, that sequencing libraries produced with the NanoAmpli-Seq protocol generates convoluted DNA molecules that can partially clog the nanopores present at the sequencing flowcells. That in turn produces smaller amount of data when compare with standard 1D nanopore protocols i.e. amplicon or genomic DNA. Nonetheless, the statistical analysis with Mantel and Procrustes tests performed on results presented in this chapter indicated that generated data with the NanoAmpli-Seq protocol is reproducible and comparable to gold standard Illumina WGS data. Results from this chapter have not yet been published or presented at any scientific conference. To



publish accomplishments (i.e. indexing, multithreading) described in this chapter, a larger number of samples are required for analysis.

Future improvements to the NanoAmpli-Seq protocol will include addition of molecular barcodes made of 10 random nucleotide basepairs to the reverse primer. This method was previously described by multiple scientists for assignment of single DNA molecules and reduction of PCR artefacts (Ståhlberg et al., 2016). Use of a molecular barcoding (MBC) strategy could allow for further clustering of molecules coming from the same DNA strand. Moreover, assignment of OTUs with the aforementioned strategy would allow for species resolution or oligotyping data analysis. That in turn would allow for more accurate estimation of bacterial community structure but also could allow for modification of the NanoAmpli-Seq protocol for detection of Single Nucleotide Variations (SNV). One of the fields of interest where MBC NanoAmpli-Seq would be applied is in rapid identification of antimicrobial resistance genes and deleterious mutations from clinical samples (Peng et al., 2016). Enhancement of the laboratory protocol with use of MBC and a microfluidics device along, with establishment of a semi-automated bioinformatics pipeline would allow generation of a autonomous or semi-autonomous biosensor for a very quick recognition of disease-causing organisms and mutations.

## 7 Conclusions and Future Work

In this chapter, I reflect on the objectives of this PhD research and summarise the important advances along with obstacles, limitations, and scope of future work. Additionally, I provide a short overview of the two innovative bioinformatics programs for DNA analysis and their limitations. Further, I provide an assessment of additional research and development that needs to be done for wider dissemination of the NanoAmpli-Seq protocol. In the end, I explain the benefits of this PhD research for the broad scientific community and include short and long-term perspectives for nanopore sequencing related to laboratory workflows and data storage.

### 7.1 Major Discoveries and Limitations

Real-time DNA sequencing with use of the nanopore technology has great potential and ability to critically transform various aspects of biological research and clinical diagnostics. Real-time data production and processing will be widely tested and used in medical research and clinical diagnostics. Previous generations of DNA sequencing platforms already allow for detailed analysis of pathogenic microorganisms and their antibiotic resistance genes, complex microbiomes and community structure biases during the health vs. disease states. However, faster detection of disease-related organisms is necessary to prevent their spread. For this reason, I tested the Oxford Nanopore sequencing platform i.e. MinION MK1b and developed protocols to allow for its broader and accurate use.

*Error profiles of nanopore technology*

The first objective of this PhD project was to evaluate error rates associated with nanopore sequencing technology to improve our understanding of the biases and errors related to 2D and 1D<sup>2</sup> data while using reference-based amplicon sequencing analysis. Determination of the error rates is important for benchmarking of all new sequencing and data analysis approaches prior to application in complex environmental and/or clinical samples. Moreover, knowledge of systematic and non-systematic errors can help in the development of effective algorithms for data processing to either error correct or discard poor quality sequences that disproportionately contribute to high error rates. These evaluations have been previously performed on second-generation sequencing technologies (e.g., Ion Torrent, Roche-454 or MiSeq) and contributed significantly to understanding error profiles and developing strategies to minimize them or their impact, to the extent possible (Harismendy, et al., 2009). Additionally, in certain cases, different library preparation methods or samples can generate different error profiles and should be evaluated separately.

This project involved assessing the error profiles of nanopore sequencing data and then developing methods to correct these errors using simple mock communities constructed from full-length 16S rRNA gene amplicons from multiple organisms. Raw nanopore sequencing error profiles were evaluated by constructing mock communities of 15 16S rRNA gene amplicons prepared in triplicate and sequenced in three independent runs. The raw sequencing data from these runs was evaluated with four alignment algorithms (Burrows-Wheeler Aligner – ONT2D, GraphMap – Default, GraphMap – Anchor, GraphMap – Gotoh) for a reference-based data analysis

approach. The outcomes of this study indicate low sequencing accuracy, with mean Phred scores of Q8-Q10. Short-read sequencing technologies (e.g. Illumina) are characterised by much higher accuracy (~Q35), often with lowest cut off points are set for Q20 scores during quality trimming (Bolger, et al., 2014). Furthermore, various read types (T-template, C-complement and 2D-two direction) from high-quality “pass” and low-quality “fail” categories were investigated for their overall quality. The outcome of the study indicated that the highest quality reads were the ones categorized as “pass” 2D while the lowest were in “fail” C. Further, the BWAONT2D aligner resulted in the best overall performance due to more effective removal of low-quality data. While the GraphMap algorithms aligned a larger proportion of reads to the reference sequences, they also had the low-quality reads, which in turn lowered overall data accuracy.

The mean alignment quality was also tested for all four aligners for the highest quality data from all three runs. In all three runs, all three GraphMap aligners performed similarly, indicating that these algorithms had comparable scoring matrices. That in turn means, software included low quality reads into the overall assignment, which in turn lowered the mean alignment scores. In contrast, BWA ONT2D discarded most of the lowest quality reads, and for this reason, mean alignment quality had higher values for assigned reads than for unassigned. Further, the ability of each alignment approach to recover the microbial community structure was evaluated for each run, replicate, and alignment approach. While the variability in relative abundance of individual reference sequences and the overall community structure was consistent between replicate sequencing runs, the type of aligner and read-type had the largest impact.

Considering the low Phred quality scores and the high mismatch and indel rates, there is no ideal approach to align reads against complex reference gene databases (e.g., SILVA database for SSU rRNA) without resulting in several false positives. Moreover, use of *de novo* clustering approach using available tools (e.g., Mothur, Qiime, Vsearch) was also unreliable, since the gold standard OTU binning method recommends a 97% similarity cut off point. Nevertheless, the identification of errors and the resultant challenges with data processing was crucial towards understanding the challenges with nanopore sequencing technology and developing novel laboratory and bioinformatics approaches for accurate amplicon sequencing analysis.

#### *Development of novel laboratory protocol for accurate amplicon sequencing*

The second objective of this PhD project was the development of a method that would reduce the error rates associated with nanopore DNA sequencing. Single organism and simple mock community 16S rRNA amplicon pools were used for benchmarking study for developing a reliable method for nanopore library preparation. Benchmarking using known reference sequences is crucial to draw logical and authentic conclusions based on the outcome of the analysis (D'Amore et al., 2016). Unfortunately, countless nanopore experiments and peer-reviewed articles lack the fundamental understanding of error rates of complex amplicon sequencing experiments (Benítez-Páez et al., 2016). For this reason, many academic research groups use 1D or 1D2 error-prone reads that generates vague or highly biased results that rely primarily on mapping of raw error prone reads to reference databases (Ma, et al., 2017). I have established and optimised an efficient library preparation protocol (i.e. NanoAmpli-Seq) using a Rolling Circle Amplification based concatameter

generation for accurate sequencing near complete 16S rRNA gene. Optimisation of the NanoAmpli-Seq protocol required multiple iterations; e.g. incubation times, centrifugation speeds, volumes, and concentrations of the reagents. There are two other library preparation protocols (i.e. INC-Seq and R2C2) that use the RCA-based method for nanopore library preparation (Li et al., 2015; Volden et al., 2018). However, both protocols are labour intensive and take around 18 or more hours to complete the protocol. That would require two or even three days of active participation, involving analyses of relatively short amplicons (400-800 bp), and resulting in error rates ranging from 4% (R2C2) to 2-3% (INC-Seq). Currently, the NanoAmpli-Seq is the only protocol that can generate a concatamerised DNA library of long (~1380bp) 16S rRNA genes. Additionally, this study introduced a novel RCA procedure that relied on use of the PrimPol enzyme for *in situ* primer generation rather than using random hexamers, which results in a high level of unspecific amplification. These modifications in the laboratory protocol shortened the library preparation step, which can now be completed in approximately 7 hours. This makes it available for a standard 8-hour workday schedule including breaks for lunch and other pause points during the laboratory work; these are important criteria for commercial laboratories. Comprehensive optimisation of the NanoAmpli-Seq protocol has attracted interest from various research groups interested in the implementation of the protocol in their laboratories for comprehensive detection of fungi marker genes (i.e., ITS1, ITS2) and agricultural studies (Wurzbacher, et al., 2018; Theuerl, et al. 2018). Other groups are interested in use of NanoAmpli-Seq for identification changes in the direct evolution of recombinase enzymes for change in their site-specific target or amplicon sequencing for rapid detection of food-borne pathogens and their corresponding antibiotic resistance marker genes; the latter

interest was from the United States Centers for Disease Control and Prevention. In the future, the protocol can be improved depending on the user requirements, including automation or semi-automation of the protocol. The NanoAmpli-Seq library protocol is comprehensive, reproducible and open-source, which makes it available for academics, industry and governmental labs use or modification of the existing protocol for their purposes.

#### *Evaluation of NanoAmpli-Seq data and design of algorithms for data correction*

The third objective of my research project was to evaluate data generated using the NanoAmpli-Seq protocol; (i.e. error profiles) and develop methods to correct these errors to result in high accuracy sequencing. Detailed analysis of the data revealed sequencing issues that were previously unknown and which required the development of read correction algorithms. The first algorithm, named *chopSeq*, was designed to recognise PCR primers in reads, and correct direction of the wrongly oriented INC-Seq consensus reads. Additionally, *chopSeq* was designed to remove long tandem repeat insertions caused by the consensus calling procedure in INC-Seq. This procedure (i.e., INC-Seq followed by chopseq) resulted in reads with an accuracy of 97-98% and where nearly the entire length of the read was correctly aligned with a reference. However, the residual errors did not allow for direct application of previously developed OTU binning approaches. Despite the lower error rates from consensus calling of concatemerized amplicons, binning of 16S rRNA reads at full-length was still inaccurate and required optimisation of the Vsearch algorithm. To this end, I developed nanoCLUST, for custom read partitioning based OTU clustering and consensus calling of chopSeq corrected 16S rRNA reads. As a result of these two novel algorithms, this project was able to obtain an overall sequencing accuracy of

99.5% for 2D and 1D2 reads. Currently, the NanoAmpli-Seq workflow is the only *de novo* approach for amplicon sequencing on the nanopore platform and has the highest reported sequence accuracy of any amplicon sequencing analysis reported thus far. The NanoAmpli-Seq workflow and associated results have been presented at various international conferences (i.e. ISME, ASM) and disseminated through the pre-print server (i.e. BioRxiv) and social media profiles (i.e. Twitter). That in turn allowed for the establishment of a joint project with the Center for Disease Control and Prevention (CDC, Atlanta, USA) and further improvements of both the laboratory protocol and the bioinformatics pipeline.

#### *Analysis of environmental samples using NanoAmpli-Seq workflow*

The final experiment during my PhD involved analysis of multiple environmental samples with the use of the NanoAmpli-Seq laboratory protocol and its accompanying programs developed and optimized using mock communities samples. Moreover, experimental work involved further improvements of the laboratory protocol, i.e. addition of a 10bp, single-ended barcode to the 5' end of 16S rRNA amplicon (fusion PCR primer). This improvement allowed for multiplexing multiple samples on a single nanopore flow cell and thus reduces cost of the sequencing per sample. Further, modifications were made to (1) the chopSeq software to, demultiplex generated data and (2) parallelisation of the algorithms used for data analysis; i.e. INC-Seq and chopSeq. Introduction of multithreaded data analysis allowed for significant reduction in data processing time, which will be crucial when rapid data processing. Other small improvements to the laboratory protocol have also been tested but have not been completed successfully and thus not included in the current thesis and could be pursued in the future. One of these improvements was addition of random decamers



(ten ‘N’ basepairs) between PCR primer and barcode sequence. This improvement is being tested for single molecule barcoding and consensus calling that will allow for detection of Single Nucleotide Polymorphism from highly related amplicons. The nanoCLUST algorithm is currently being improved for detection of random decamers, filtration, separation, and binning of reads according to the N-bases.

## 7.2 Future Work

The miniaturisation of DNA sequencing platforms and their potential for real-time data processing has a great potential to revolutionise various areas of environmental research and clinical diagnostics in the near future. However, various technological and bioinformatics challenges limits the widespread application of nanopore sequencing and these obstacles have to be resolved first. Moreover, the ability to sequence various genes and genomes in real-time can lead to concerns related to data storage and security, especially in the case of clinical diagnostics and human genome analysis (Khan, 2011).

### *Medical diagnostics and environmental sensing*

One of the most desired and promising functions for the nanopore sequencing is rapid analysis of bacterial pathogens from clinical and environmental settings, e.g., sepsis or drinking water pathogens (Greninger, et al., 2015). Current gold-standard methods are still based on culture, microscopy, and biochemical tests, which are time consuming. Further, these methods are restricted to the analysis of live organisms that successfully grow on agar or liquid media, while the majority of microorganisms are difficult to grow in well-defined culture media (Fournier, et al., 2013). Moreover, these methods are often limited in their capacity for high-throughput analysis of

samples. However, DNA sequencing technologies (i.e., Illumina or Oxford Nanopore Tech.) can revolutionise the way pathogenic organisms are detected. Current sequencing platforms (i.e., Illumina MiniSeq or MiSeq) require tens or even hundreds of samples to be processed on a single sequencing run which can in turn cause delays. Moreover, second-generation sequencing platforms require several independent stages of sample processing, sequencing, and data analyses, all of which are time-consuming and thus do not allow for real-time data generation and analysis. While portable genomic sequencers, (i.e., MinION) or benchtop versions of the device, (i.e., GridION and PromethION) can provide rapid *in situ* amplicon-based or metagenomic sequencing analysis. One of the main advantages of DNA sequencing is the analysis of both cultured and uncultured organisms with no need for prior knowledge or assumptions on the microbial composition of samples (Yarza, et al., 2014). Further, various protocols can allow for amplicon or whole genome analysis, which allows for a simple or complex understanding of the organisms and related pathogenicity. Use of DNA sequencing would allow for sophisticated analysis of the majority or even all microbial infections and outbreaks. A long-term analysis would allow for identification of unknown pathogenic microorganisms, characterisation of the genomic features such as antibiotic resistance or virulence factors and generation of comprehensive databases (Guo, et al., 2010; McArthur, et al., 2013). Moreover, identification of the origin of infection is essential as the elimination of the source can reduce the spread of the pathogens. Examples of epidemiological outbreaks where nanopore sequencing technology contributed to outbreak control include Salmonella, Ebola and Zika outbreaks (Faria, et al., 2016; Hoenen, et al., 2016). Currently, the cost of reagents for bacterial growth and microscopy is still low when compared to sequencing reagents and high-performance computers for analysis of data. However, I

believe bacteria culturing methods will diminish in prominence in the next 10 to 15 years as the cost the DNA sequencing will further decline, and analysis of the data becomes simplified.

#### *Automated library preparation for biosensing*

Rapid analysis of tens or hundreds of independent specimens from medical patients or environmental samples will require a large number of technicians to collect, prepare, and analyse these sample. While sample collection likely has to be performed by the trained technician, nurse, medical doctor or a water utility operator, the preparation of the sequencing libraries could be simplified or even automated. Manual preparation of samples is a major bottleneck which causes significant delays in sample processing but also results in some loss in reproducibility. The use of liquid handling robots or cheap microfluidics devices could reduce hands-on requirements and reduce the time of sample preparations. The use automated liquid handling devices by Beckman Coulter or Hamilton Robotics tends to be expensive and allows for analysis of up to 96 samples at a time. Unfortunately, batch analyses of clinical samples for diagnostic purposes is not ideal due to resultant diagnosis delays and a sample-by-sample independent processing approach is desirable. For this reason, cheap miniaturised microfluidics chips that treat each sample separately are much more desirable. Technologies such as the NeoPrep System (Illumina) or VolTRAX (ONT), that offer this approach, have to be further improved and yet to be tested in independent laboratories. The optimal microfluidics technology would have library preparation and sequencing flowcell combined, which would significantly simplify the laboratory protocol by minimizing sample handling steps and time. Integration of these two technologies together would not only simplify laboratory protocol but also allow for

remote analysis of various environmental samples (e.g., soil, river) without a need for a full-scale laboratory.

*Real-time data generation and analysis with cloud computing*

Giga or even Terabytes of data generated with sequencing platforms (i.e., NovaSeq or PromethION) have to be stored and analysed on high-performance computing units. While real-time data generation is highly desirable especially, real-time data analysis may not be feasible due to lack of computing infrastructure. Data analysis and storage increases the costs, as computing units are expensive and must expand continually as generated data needs to be stored for multiple years. Raw data storage is needed in case of data reanalysis caused by algorithm upgrades (e.g., basecallers: NanoNet, Albacore, Guppy). It is possible that through technological improvements and reduction in cost, some of these limitations may be resolved in the next 5 to 10 years from now. One way to resolve the problem of data storage will be the progression towards decentralised Internet (e.g., anonymous distributed ledgers) similar to cryptocurrencies and block chain system or cloud computing. A decentralized Internet system is based on file partitioning, encryption, and distribution of files across the decentralised network. This would significantly increase the level of security when compared to centralised data storage while also spread the burden of data storage across millions of systems. Data protection is particularly critical when considering population-level studies, involving the human genomes. Recent reports indicate that data leaks are still prevalent (e.g., in 2018 MyHeritage leaked data of 92 million users), (Manogaram, et al., 2017). Hacking thousands of random computing nodes, with random encryptions is much more difficult as compared to a single server. Moreover, cloud computing for data analysis would mean that data analysis software

could be updated on all distributed nodes simultaneously. Currently, cloud-based programs are available (i.e., Dropbox, Netflix or GeForce NOW) and have successfully substituted conventional hard drives, DVD and gaming computers. I believe that extensive implementation of DNA sequencing technologies in clinical and environmental settings is possible in the next 10 to 15 years, however, will require development in the field of distributed data storage and analysis.

#### *Next-generation sequencing technologies*

Improvements in the field of nanopore sequencing are crucial to further reduce error rates, increase reproducibility and shelf life of the reagents; e.g. flow cells. For this reason, advances in the area of solid-state nanopores is crucial and may bring the pivotal switch away from second generation (i.e., Illumina) to third generation sequencing platforms such as Oxford Nanopore Tech. Further, development of new sequencing platforms that will allow for analysis of long reads with low error rates are still highly desirable. One of the companies working on that is Base4 Ltd., which is developing an innovative device that utilises a microfluidics chip for single base DNA sequencing resolution. Sequencing technology based on the microfluidics chip may have miniaturised size and have low production cost. The small size of the device and its portability will be necessary to compete with MinION<sup>TM</sup> nanopore sequencer and a wide range of Illumina products. However, their device is not commercially available yet.

## 7.3 Conclusions

In the past decade, the cost of sequencing has dropped significantly, which has resulted in increase in popularity of nucleic acid analysis in diverse scientific fields. These reductions were caused by market competition and associated technological advancements. Generation of Giga and Terabytes of data sparked development of a new field of science called bioinformatics. In the past, multiple sequencing technologies were evaluated for their throughput, error rate, and overall speed of data generation (Mardis, 2017). Specific sequencing platforms are characterised by a characteristic pattern of errors and biases, which needs to be considered during data analysis. These inherent inaccuracies present in complex datasets can significantly affect conclusions and interpretation of data. Meaningful results can only be drawn when high-quality data is analysed with use of appropriate bioinformatics programs, which have previously been tested on various samples and verified to be reproducible.

This PhD thesis describes the evaluation of nanopore sequencing technology for accurate detection of the SSU rRNA gene (i.e., 16S rRNA gene), with emphasis on amplicon size, error rate estimation, and community diversity estimation. Preliminary results of the 16S rRNA sequencing on the MinION<sup>TM</sup> platform suggested the necessity for the development of a laboratory protocol to reduce sequencing errors. I developed laboratory protocol called NanoAmpli-Seq, which includes laboratory protocol improvements and novel bioinformatics programs (i.e., chopSeq, nanoCLUST), which significantly improved the final results when analyzing mock and environmental samples. This analysis provides a reference point for future improvements and could help in improving data accuracy in other protocols and strategies. When SSU rRNA gene datasets are analysed accurately, it can facilitate

generation of long marker gene references that could augment and improve current high-quality databases (i.e., SILVA or RDP). The NanoAmpli-Seq workflow represents the most accurate approach for SSU rRNA gene analysis on the nanopore platform, to date.

Nonetheless, nanopore sequencing and related bioinformatics pipelines (including NanoAmpli-Seq) do not allow for imminent real-time analysis of environmental or clinical samples. Despite the improvements described in this research thesis, the real-time data generation and high-accuracy data analysis approach is not yet feasible. Growing demand for fast and accurate analysis of clinical and environmental samples will continue to contribute to development of various new sequencing technologies (e.g., Oxford Nanopore Tech., Genia Tech. Inc.) and programs (i.e., MinKNOW or MinoTour). Further improvements (e.g., graphene instead of biological proteins) in nanopore technology are crucial and need to be applied to reduce overall error rates, increase data throughput, and allow for possible real-time data analysis.

## References:

- Aksimentiev, A. et al., 2004. Microscopic Kinetics of DNA Translocation through synthetic nanopores. *Biophysical journal*, 87(3), pp.2086–97.
- Amore, R.D. et al., A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. , pp.1–40.
- Anon, 2008. Epigenetics meets next-generation sequencing AU - Park, Peter J. *Epigenetics*, 3(6), pp.318–321.
- Ansorge, W.J., 2016. Next-generation DNA sequencing (II): techniques, applications. *Next Generat. Sequenc. & Applic*, 1, pp.1-10. Yet challenges with high error rates, complex library preparation, complex fabrication technologies, low amount of generated data and high running costs remain.
- Armougom, F. and Raoult, D., 2009. Exploring microbial diversity using 16S rRNA high-throughput methods. *J Comput Sci Syst Biol*, 2(1), pp.74-92.
- Ashkenasy, N., Sánchez - Quesada, J., Bayley, H. and Ghadiri, M.R., 2005. Recognizing a single base in an individual DNA strand: a step toward DNA sequencing in nanopores. *Angewandte Chemie*, 117(9), pp.1425-1428.
- Ashton, P.M., Nair, S., Dallman, T., Rubino, S., Rabach, W., Mwaigwisya, S., Wain, J. and O'grady, J., 2015. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nature biotechnology*, 33(3), p.296.
- Avery, O.T., 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* type III. *Journal of Experimental Medicine*, 79(2), pp.137–158.
- Ayub, M. & Bayley, H., 2012. Individual RNA base recognition in immobilized oligonucleotides using a protein nanopore. *Nano Letters*, 12(11), pp.5637–5643.
- Barba, M., Czosnek, H. & Hadidi, A., 2014. Historical Perspective, Development and Applications of Next-Generation Sequencing in Plant Virology. *Viruses*, 6(1), pp.106–136.
- Barbazuk, W.B. et al., 2007. SNP discovery via 454 transcriptome sequencing. *Plant Journal*, 51(5), pp.910–918.
- Bautista-de los Santos, Q.M. et al., 2016. Emerging investigators series: microbial communities in full-scale drinking water distribution systems – a meta-analysis. *Environ. Sci.: Water Res. Technol.*,
- Benítez-Páez, A., Portune, K.J. & Sanz, Y., 2016. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. *GigaScience*, 5(1), p.4.
- Bentley, D.R. et al., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218), pp.53–59.
- Berlin, K. et al., 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology*, 33(6), pp.623–630.
- Bolger, A.M., Usadel, B. & Lohse, M., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), pp.2114–2120.



- Boza, V., Brejova, B. and Vinar, T., 2017. DeepNano: deep recurrent neural networks for base calling in MinION nanopore reads. *PloS one*, 12(6), pp.e0178751-e0178751.
- Branton, D. et al., 2008. The potential and challenges of nanopore sequencing. *Nature Biotechnology*, 26(10), pp.1146–1153.
- Branton, D. et al., 2009. The potential and challenges of nanopore sequencing. In *Nanoscience and Technology*. Co-Published with Macmillan Publishers Ltd, UK, pp. 261–268.
- Burke C., 2014. Resolving microbial microdiversity with high accuracy, full length 16S rRNA Illumina sequencing.
- Bundschuh, R. & Gerland, U., 2005. Coupled Dynamics of RNA Folding and Nanopore Translocation. *Physical Review Letters*, 95(20), p.208104.
- Buse, H.Y., Lu, J., Lu, X., Mou, X. and Ashbolt, N.J., 2014. Microbial diversities (16S and 18S rRNA gene pyrosequencing) and environmental pathogens within drinking water biofilms grown on the common premise plumbing materials unplasticized polyvinylchloride and copper. *FEMS microbiology ecology*, 88(2), pp.280-295.
- Calus, N.L.J.Q.S., 2014. A *P. aeruginosa* serotype-defining single read from our first Oxford Nanopore run. *FigShare*.
- Cameron, D. and Jones, I.G., 1983. John Snow, the Broad Street pump and modern epidemiology. *International journal of epidemiology*, 12(4), pp.393-396.
- Canard, B. & Sarfati, R.S., 1994. DNA polymerase fluorescent substrates with reversible 3'-tags. *Gene*, 148(1), pp.1–6.
- Caporaso, J.G. et al., 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal*, 6(8), pp.1621–1624.
- Carnevali, P. et al., 2012. Computational Techniques for Human Genome Resequencing Using Mated Gapped Reads. *Journal of Computational Biology*, 19(3), pp.279–292.
- Castro-Wallace, S.L., Chiu, C.Y., John, K.K., Stahl, S.E., Rubins, K.H., McIntyre, A.B., Dworkin, J.P., Lupisella, M.L., Smith, D.J., Botkin, D.J. and Stephenson, T.A., 2017. Nanopore DNA sequencing and genome assembly on the International Space Station. *Scientific reports*, 7(1), p.18022.
- Cavalli-Sforza, L.L., 1998. The DNA revolution in population genetics. *Trends in Genetics*, 14(2), pp.60–65.
- Chadarevian, S. De, 2003. Portrait of a Discovery. *Isis*, 94(1), pp.90–105.
- Chan, E.Y., 2005. Advances in sequencing technology. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 573(1–2), pp.13–40.
- Chang, S. et al., 2011. Gap distance and interactions in a molecular tunnel junction. *Journal of the American Chemical Society*, 133(36), pp.14267–14269.
- Chang, Y.F. et al., 2016. Use of liposomal amplifiers in total internal reflection fluorescence fiber-optic biosensors for protein detection. *Biosensors and Bioelectronics*, 77, pp.1201–1207.
- Chen, P. et al., 2004. Atomic layer deposition to fine-tune the surface properties and diameters of fabricated nanopores. *Nano Letters*, 4(7), pp.1333–1337.

- Choudhuri, S., 2003. The path from nuclein to human genome: A brief history of DNA with a note on human genome sequencing and its impact on future research in biology. *Bulletin of Science, Technology and Society*, 23(5), pp.360–367.
- Chu, Y. & Corey, D.R., 2012. RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation. *Nucleic Acid Therapeutics*, 22(4), pp.271–274.
- Cock, P.J.A. et al., 2009. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6), pp.1767–1771.
- Collins, F.S., Morgan, M. & Patrinos, A., 2003. The Human Genome Project: Lessons from large-scale biology. *Science*, 300(5617), pp.286–290.
- Collins, F.S. et al., 1998. New goals for the U.S. Human Genome Project: 1998–2003. *Science*, 282(5389), pp.682–689.
- Connelly, S. et al., 2017. Bioreactor scalability: Laboratory-scale bioreactor design influences performance, ecology, and community physiology in expanded granular sludge bed bioreactors. *Frontiers in Microbiology*, 8(MAY).
- Cornell, B. a et al., 1997. A biosensor that uses ion-channel switches. *Nature*, 387(6633), pp.580–583.
- Cox, M.P., Peterson, D.A. & Biggs, P.J., 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*, 11.
- Crawford, B.L. et al., 2015. Factors Influencing Risk of Homelessness among Youth in Transition from Foster Care in Oklahoma: Implications for Reforming Independent Living Services and Opportunities. *Child welfare*, 94(1), pp.19–34.
- Crick, F., 1970. Central dogma of molecular biology. *Nature*, 227(5258), pp.561–3. 4913914.
- D’Amore, R. et al., 2016. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics*, 17(1), p.55.
- Dahm, R., 2005. Friedrich Miescher and the discovery of DNA. *Developmental Biology*, 278(2), pp.274–288.
- Dahm, R., 2008. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Human Genetics*, 122(6), pp.565–581.
- Daum, L.T. et al., 2012. Next-generation ion torrent sequencing of drug resistance mutations in *Mycobacterium tuberculosis* strains. *Journal of Clinical Microbiology*, 50(12), pp.3831–3837.
- David, M. et al., 2016. Nanocall: An Open Source Basecaller for Oxford Nanopore Sequencing Data. *bioRxiv*, p.046086.
- Deamer, D., 2010. Nanopore analysis of nucleic acids bound to exonucleases and polymerases. *Annual review of biophysics*, 39, pp.79–90.
- Deamer, D.W. & Akeson, M., 2000. Nanopores and nucleic acids: Prospects for ultrarapid sequencing. *Trends in Biotechnology*, 18(4), pp.147–151.
- Deamer, D., Akeson, M. & Branton, D., 2016. Three decades of nanopore sequencing. *Nature Biotechnology*, 34(5), pp.518–524.

- Donelson, J.E. & Wu, R., 1972. Nucleotide Sequence Analysis of Deoxyribonucleic Acid. *The journal of Biological Chemistry*, 247(14), pp.4654–4660.
- Dorn-in, S. et al., 2015. Specific amplification of bacterial DNA by optimized so-called universal bacterial primers in samples rich of plant DNA. *Journal of Microbiological Methods*, 113, pp.1–7.
- Drmanac, R. et al., 2010. Human {Genome} {Sequencing} {Using} {Unchained} {Base} {Reads} on {Self}-{Assembling} {DNA} {Nanoarrays}. *Science*, 327(5961), pp.78–81.
- Earle, C., 1979. Environment, disease and mortality in early Virginia. *Journal of historical geography*, 5, pp.365–390.
- Edwards, A. et al., 2016. Extreme metagenomics using nanopore DNA sequencing: a field report from Svalbard, 78 N. *bioRxiv*, p.073965.
- Edgar, R.C., 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), pp.1792–1797.
- Ee, L.T. et al., 2008. Implantable biosensors for real-time strain and pressure monitoring. *Sensors*, 8(10), pp.6396–6406.
- Eisenstein, M., 2012. Oxford Nanopore announcement sets sequencing sector abuzz. *Nature Biotechnology*, 30(4), pp.295–296.
- Engle, K. M.; Mei, T-S.; Wasa, M.; Yu, J.-Q., 2008. NIH Public Access. *Accounts of Chemical Research*, 45(6), pp.788–802.
- Eren, A.M., Maignien, L., Sul, W.J., Murphy, L.G., Grim, S.L., Morrison, H.G. and Sogin, M.L., 2013. Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods in Ecology and Evolution*, 4(12), pp.1111–1119.
- Faria, N.R. et al., 2016. Mobile real-time surveillance of Zika virus in Brazil. *Genome Medicine*, 8(1), p.97.
- Fondriest, S., 2007. Recent Developments in Real-time Environmental Monitoring Technology Water Quality – YSI Water Flow / Velocity - Sontek.
- Fournier, P. E. et al., 2013. Modern clinical microbiology: new challenges and solutions. *Nature Reviews Microbiology*, 11, p.574. Available at: <https://doi.org/10.1038/nrmicro3068>.
- Garalde, D.R. et al., 2016. Highly parallel direct RNA sequencing on an array of nanopores. *bioRxiv*, p.068809.
- Gibbons, M.G., 2012. Reassessing Discovery: Rosalind Franklin, Scientific Visualization, and the Structure of DNA\*. *Philosophy of Science*, 79(1), pp.63–80.
- Gilbert, J.A., Jansson, J.K. & Knight, R., 2014. The Earth Microbiome project: Successes and aspirations. *BMC Biology*, 12(1).
- Gilbert, W. & Maxam, A., 1973. The Nucleotide Sequence of the lac Operator. *Proceedings of the National Academy of Sciences*, 70(12), pp.3581–3584.
- Goldmann, D.A., Weinstein, R.A., Wenzel, R.P., Tablan, O.C., Duma, R.J., Gaynes, R.P., Schlosser, J., Martone, W.J., Acar, J., Avorn, J. and Burke, J., 1996. Strategies to prevent and control the

- emergence and spread of antimicrobial-resistant microorganisms in hospitals: a challenge to hospital leadership. *Jama*, 275(3), pp.234-240.
- Goodwin, S. et al., 2015. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Research*, 25(11), pp.1750–1756.
- Grabow, W.O.K. & Du Preez, M., 1979. Comparison of m-Endo LES, MacConkey, and Teepol media for membrane filtration counting of total coliform bacteria in water. *Applied and Environmental Microbiology*, 38(3), pp.351–358.
- Greninger, A.L. et al., 2015. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis.
- Griffen, A.L. et al., 2012. Distinct and complex bacterial profiles in human periodontitis and health revealed by 16S pyrosequencing. *The ISME Journal*, 6(6), pp.1176–1185.
- Guo, J. et al., 2010. Short- and long-term effects of temperature on partial nitrification in a sequencing batch reactor treating domestic wastewater. *Journal of Hazardous Materials*, 179(1), pp.471–479.
- Hardiman, G., 2003. Microarray technologies 2003-an overview, 251-256.
- Harismendy, O. et al., 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, 10(3), p.R32.
- Harper-Owen, R. et al., 1999. Detection of unculturable bacteria in periodontal health and disease by PCR. *Journal of Clinical Microbiology*, 37(5), pp.1469–1473.
- Heather, J.M. & Chain, B., 2016. The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1), pp.1–8.
- Heerema, S.J. & Dekker, C., 2016. Graphene nanodevices for {DNA} sequencing. *Nature Nanotechnology*, 11(2), pp.127–136.
- Henley, R. Y., Carson, S., & Wanunu, M. 2016. Studies of RNA Sequence and Structure Using Nanopores. *Progress in molecular biology and translational science*, 139, 73-99.
- Heinz, E. & Domman, D., 2017. Reshaping the tree of life. *Nature Reviews Microbiology*, 15(6), p.322.
- Hiltemann, S. et al., 2014. CGtag: Complete genomics toolkit and annotation in a cloud-based Galaxy. *GigaScience*, 3(1), pp.1–6.
- Hoenen, T. et al., 2016. Nanopore Sequencing as a rapidly deployable Ebola outbreak tool. *Emerging infectious diseases*, 22(2), pp.331–334.
- Hoffmann, C. et al., 2013. Archaea and Fungi of the Human Gut Microbiome: Correlations with Diet and Bacterial Residents. *PLoS ONE*, 8(6).
- Huang, W. et al., 2012. ART: A next-generation sequencing read simulator. *Bioinformatics*, 28(4), pp.593–594.
- Hug, L.A. et al., 2016. A new view of the tree of life. *Nature Microbiology*, 1(5).
- Hunter, G.K., 1999. Phoebus Levene and the Tetranucleotide Structure of Nucleic Acids. *Ambix*, 46(2), pp.73–103.

- Istace, B. et al., 2016. de novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *bioRxiv*, p.066613.
- Jain, M. et al., 2015. Improved data analysis for the MinION nanopore sequencer. *Nature Methods*, 12(4), pp.351–356.
- Jain, M. et al., 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, 36(4), pp.338–345.
- Jain, S. & Bhatnagar, V., 2014. Analogy of various DNA based security algorithms using cryptography and steganography. In *Proceedings of the 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques, ICICT 2014*. pp. 285–291.
- Järup, L., 2003. Hazards of heavy metal contamination. *British Medical Bulletin*, 68, pp.167–182.
- Jones, P.A. & Takai, D., 2001. The role of DNA methylation in mammalian epigenetics 8 227. *Science*, 293(0036–8075 (Print)), pp.1068–1070.
- Lee, J.W. and Thundat, T.G., UT-Battelle LLC, 2005. DNA and RNA sequencing by nanoscale reading through programmable electrophoresis and nanoelectrode-gated tunneling and dielectric detection. U.S. Patent 6,905,586.
- Kahn, S.D., 2011. On the Future of Genomic Data. *Science*, 331(6018), p.728 LP-729.
- Karlsson, E. et al., 2015. Scaffolding of a bacterial genome using MinION nanopore sequencing. *Scientific reports*, 5(CLiC), p.11996.
- Karow, J., 2014. Japan's Quantum Biosystems shows raw read data from single-molecule nanopore sequencer. *GenomeWeb*.
- Keusgen, M., 2002. Biosensors: New approaches in drug discovery. *Naturwissenschaften*, 89(10), pp.433–444.
- Kilianski, A. et al., 2015. Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. *GigaScience*, 4.
- Kim, H. et al., 2013. A Microfluidic DNA Library Preparation Platform for Next-Generation Sequencing. *PLoS ONE*, 8(7), pp.1–9.
- Kircher, M., Stenzel, U. & Kelso, J., 2009. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biology*, 10(8).
- Klindworth, A. et al., 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research*, 41(1), pp.1–11.
- Klug, A., 1968. Rosalind Franklin and the discovery of the structure of DNA. *Nature*, 219(5156), pp.808–844.
- Kokoris, M.S. & McRuer, R.N., 2011. High throughput nucleic acid sequencing by expansion.
- Koren, S., Schatz, M.C., Walenz, B.P., Martin, J., Howard, J.T., Ganapathy, G., Wang, Z., Rasko, D.A., McCombie, W.R., Jarvis, E.D. and Phillippy, A.M., 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature biotechnology*, 30(7), p.693.
- Kürekcı, G.K. & Dinçer, P., 2014. Next-generation DNA sequencing technologies. *Erciyes Tip Dergisi*, 36(3), pp.99–103.

- Laddy, D.J. et al., 2018. Toward tuberculosis vaccine development: Recommendations for nonhuman primate study design. *Infection and Immunity*, 86(2), pp.1–6.
- Lagerqvist, J., Zwolak, M. & Di Ventra, M., 2006. Fast DNA sequencing via transverse electronic transport. *Nano Letters*, 6(4), pp.779–782.
- Laszlo, A.H. et al., 2013. Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. *Proceedings of the National Academy of Sciences of the United States of America*, 110(47), pp.18904–9.
- Laszlo, A.H. et al., 2014. Decoding long nanopore sequencing reads of natural DNA. *Nature biotechnology*, 32(8), pp.829–834.
- Laver, T. et al., 2015. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification*, 3, pp.1–8.
- Leggett, R.M. et al., 2015. NanoOK: Multi-reference alignment analysis of nanopore sequencing data, quality and error profiles. *Bioinformatics*, 32(1), pp.142–144.
- Li, C., Chng, K.R., Boey, E.J.H., Ng, A.H.Q., Wilm, A. and Nagarajan, N., 2016. INC-Seq: accurate single molecule reads using nanopore sequencing. *GigaScience*, 5(1), p.34.
- Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv*, 00(00), p.3.
- Li, H., 2015. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *arXiv*, 32(March), pp.1–7.
- Li, J. et al., 2001. Ion-beam sculpting at nanometre length scales. *Nature*, 412(6843), pp.166–169.
- Lilienfeld, a M. & Lilienfeld, D.E., 1984. John Snow, the Broad Street pump and modern epidemiology. *International journal of epidemiology*, 13(4), pp.376–378.
- Ling, X. & Bready, B., 2007. Hybridization assisted nanopore sequencing.
- Liu, N. & Pan, T., 2015. RNA epigenetics. *Translational Research*, 165(1), pp.28–35.
- Loman, N.J. et al., 2012. Performance comparison of benchtop high-throughput sequencing platforms. 30(5).
- Loman, N. & Quinlan, a., 2014. Poretools: a toolkit for analyzing nanopore sequence data. *bioRxiv*, 30(23), p.007401.
- Loman, N.J., Quick, J. & Simpson, J.T., 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*, 12(8), pp.733–735.
- Lytle, D.A., Sorg, T., Wang, L. and Chen, A., 2014. The accumulation of radioactive contaminants in drinking water distribution systems. *Water research*, 50, pp.396–407.
- Ma, X., Stachler, E., Bibby, K., 2017. Evaluation of Oxford Nanopore MinION Sequencing for 16S rRNA Microbiome Characterization. *GigaScience*, doi: <https://doi.org/10.1101/099960>
- Malcovati, P., Baschiroto, A. & Di, C., 2009. *Sensors and mycosystems*.

- Manly, B.F. and Alberto, J.A.N., 2016. Multivariate statistical methods: a primer. Chapman and Hall/CRC.
- Manogaran, G. et al., 2017. Big Data Knowledge System in Healthcare BT - Internet of Things and Big Data Technologies for Next Generation Healthcare. In C. Bhatt, N. Dey, & A. S. Ashour, eds. Cham: Springer International Publishing, pp. 133–157.
- Marchesi, J.R., Sato, T., Weightman, A.J., Martin, T.A., Fry, J.C., Hiom, S.J. and Wade, W.G., 1998. Design and evaluation of useful bacterium-specific PCR primers that amplify genes coding for bacterial 16S rRNA. *Applied and environmental microbiology*, 64(2), pp.795-799.
- Mardis, E.R., 2008. The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3), pp.133–141.
- Mardis, E.R., 2011. A decade's perspective on DNA sequencing technology. *Nature*, 470(7333), pp.198–203.
- Mardis, E.R., 2017. DNA sequencing technologies: 2006–2016. *Nature Protocols*, 12, p.213.
- Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17, p.10.
- Maxam, A.M. & Gilbert, W., 1977. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2), pp.560–4.
- Maxam, A.M. & Gilbert, W., 1980. [57] Sequencing End-Labeled DNA with Base-Specific Chemical Cleavages. *Methods in Enzymology*, 65(C), pp.499–560.
- McArthur, A.G. et al., 2013. The Comprehensive Antibiotic Resistance Database. *Antimicrobial Agents and Chemotherapy*, 57(7), p.3348 LP-3357.
- McCarthy, A., 2010. Third generation DNA sequencing: Pacific biosciences' single molecule real time technology. *Chemistry and Biology*, 17(7), pp.675–676.
- McCarthy, J.F. & Zachara, J.M., 1989. Subsurface transport of contaminants: binding to mobile and immobile phases in groundwater aquifers. *Environmental Science & Technology*, 23(5), pp.496–502.
- McElhoe, J.A., Holland, M.M., Makova, K.D., Su, M.S.W., Paul, I.M., Baker, C.H., Faith, S.A. and Young, B., 2014. Development and assessment of an optimized next-generation DNA sequencing approach for the mtgenome using the Illumina MiSeq. *Forensic Science International: Genetics*, 13, pp.20-29.
- McFeters, G.A., 1990. *Drinking Water Microbiology: Progress and Recent Developments*, Springer-Verlag.
- Me, M., 2018. Ten simple rules for developing good reading habits during graduate school and beyond. , pp.1–4.
- Meldrum, D., 2000. Automation for genomics, part one: preparation for sequencing. *Genome research*, 10(8), pp.1081-1092.
- Merriman, B., Torrent, I. & Rothberg, J.M., 2012. Progress in Ion Torrent semiconductor chip based sequencing. *Electrophoresis*, 33(23), pp.3397–3417.



- Meselson, M. & Stahl, F.W., 1958. The replication of DNA in *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 44, pp.671–682.
- Metzker, M.L., 2005. Emerging technologies in DNA sequencing. *Genome Research*, 15(12), pp.1767–1776.
- Meyer, M. & Kircher, M., 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, 5(6).
- Mitra, R.D. et al., 2004. Erratum: Fluorescent in situ sequencing on polymerase colonies (Analytical Biochemistry (2003) 320 (55-65) DOI: 10.1016/S0003-2697(03)00291-4). *Analytical Biochemistry*, 328(2), p.245.
- Mohan, D. et al., 2014. Organic and inorganic contaminants removal from water with biochar, a renewable, low cost and sustainable adsorbent - A critical review. *Bioresource Technology*, 160, pp.191–202.
- Moore, C.B., 2012. Chemical and biochemical applications of lasers (Vol. 1). Elsevier.
- Morozova, O. & Marra, M.A., 2008. Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92(5), pp.255–264.
- Mosher, J.J. et al., 2014. Improved performance of the PacBio SMRT technology for 16S rDNA sequencing. *Journal of Microbiological Methods*, 104, pp.59–60.
- Null et al., 2014. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nature Biotechnology*, 33(3).
- Ohshiro, T. & Umezawa, Y., 2006. Complementary base-pair-facilitated electron tunneling for electrically pinpointing complementary nucleobases. *Proceedings of the National Academy of Sciences*, 103(1), pp.10–14.
- Padmanabhan, R., Wu, R. & Bode, V.C., 1972. Arrangement of DNA in lambda bacteriophage heads. III. Location and number of nucleotides cleaved from  $\lambda$  DNA by micrococcal nuclease attack on heads. *Journal of Molecular Biology*, 69(2), pp.201–207.
- Papagrigorakis, M.J., Synodinos, P.N. & Yapijakis, C., 2007. Ancient typhoid epidemic reveals possible ancestral strain of *Salmonella enterica* serovar Typhi. *Infection, Genetics and Evolution*, 7, pp.126–127.
- Paulechka, E. et al., 2016. Nucleobase-functionalized graphene nanoribbons for accurate high-speed DNA sequencing. *Nanoscale*, 8(4), pp.1861–1867.
- Pericard, P., Dufresne, Y., Couderc, L., Blanquart, S. and Touzet, H., 2017. MATAM: reconstruction of phylogenetic marker genes from short sequencing reads in metagenomes. *Bioinformatics*, 34(4), pp.585–591.
- Pinard, R. et al., 2006. Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC genomics*, 7, p.216.
- Pinto, A.J., Xi, C. & Raskin, L., 2012. Bacterial community structure in the drinking water microbiome is governed by filtration processes. *Environmental Science and Technology*, 46(16), pp.8851–8859.
- Polz, M.F. & Cavanaugh, C.M., 1998. Bias in template-to product ratios in multitemplate PCR. *Appl.Environ.Microbiol.*, 64(10), pp.3724–3730.



- Pop, M., Phillippy, A. & Delcher, A.L., 2004. Comparative genome assembly. *Bioinformatics*, 5(3), pp.237–248.
- Pop, M. & Salzberg, S.L., 2008. Bioinformatics challenges of new sequencing technology. *Trends in Genetics*, 24(3), pp.142–149.
- Pray, L.A., 2008. Discovery of DNA Structure and Function: Watson and Crick The First Piece of the Puzzle: Miescher Discovers DNA. *Nature Education*, 1(1), pp.1–8.
- Pray, L.A., 2008. Discovery of DNA Double Helix: Watson and Crick. *Nature Education*, 1(1), p.100.
- Pruesse, E. et al., 2007. SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, 35(21), pp.7188–7196.
- Quail, M.A. et al., 2012. A tale of 3 NGS sequencing platforms.
- Quast, C. et al., 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, 41, pp.1–7.
- Quick, J. et al., 2015. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome biology*, 16(1), p.114.
- Quick, J., Quinlan, A.R. & Loman, N.J., 2015. Erratum: A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. *GigaScience*, 4(1), p.6.
- Quick, J. et al., 2016. Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 530(7589), pp.228–232.
- Quick, J., Grubaugh, N.D., Pullan, S.T., Claro, I.M., Smith, A.D., Gangavarapu, K., Oliveira, G., Robles-Sikisaka, R., Rogers, T.F., Beutler, N.A. and Burton, D.R., 2017. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *nature protocols*, 12(6), p.1261.
- Ozsolak, F. & Milos, P.M., 2010. RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, 12, p.87.
- Parida, M., Sannarangaiah, S., Dash, P.K., Rao, P.V.L. and Morita, K., 2008. Loop mediated isothermal amplification (LAMP): a new generation of innovative gene amplification technique; perspectives in clinical diagnosis of infectious diseases. *Reviews in medical virology*, 18(6), pp.407–421.
- Peng, H., Stolovitzky, G.A., Rossmagel, S.M., Polonsky, S., Luan, B. and Martyna, G.J., International Business Machines Corp, 2011. Piezoelectric-based nanopore device for the active control of the motion of polymers through the same. U.S. Patent 8,039,250.
- Peng, Q., Satya, R.V., Lewis, M., Randad, P. and Wang, Y., 2015. Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes. *BMC genomics*, 16(1), p.589.
- Prasongkit, J., Feliciano, G.T., Rocha, A.R., He, Y., Osotchan, T., Ahuja, R. and Scheicher, R.H., 2015. Theoretical assessment of feasibility to sequence DNA through interlayer electronic tunneling transport at aligned nanopores in bilayer graphene. *Scientific Reports*, 5, p.17560.
- Rapoport, S., 2002. Rosalind Franklin: Unsung hero of the dna revolution. *History Teacher*, 36(1), p.116.

- Redwood, M., 1961. of Practical Problems Involve the. *the Journal of the Acoustical Society of America*, 33(4), pp.527–536.
- Remaut, E. & Fiers, W., 1972. Studies on the bacteriophage MS2. XVI. The termination signal of the a protein cistron. *Journal of Molecular Biology*, 71(2), pp.243–261.
- Rhoads, A. & Au, K.F., 2015. PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics*, 13(5), pp.278–289.
- Robertson, K.D. & Jones, P.A., 2000. DNA methylation: past, present and future directions. *Carcinogenesis*, 21(3), pp.461–467.
- Rodriguez-Mozaz, S. et al., 2004. Biosensors for environmental applications: Future development trends. *Pure and Applied Chemistry*, 76(4), pp.723–752.
- Ronaghi M., 2001. Pyrosequencing Sheds Light on DNA Sequencing References  
<http://genome.cshlp.org/content/11/1/3.full.html#related-urls> Pyrosequencing Sheds Light on DNA Sequencing. , (650), pp.3–11.
- Ronaghi, M. et al., 1996. Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. *Analytical Biochemistry*, 242(1), pp.84–89.
- Ronaghi, M., Uhlén, M. & Nyrén, P., 1998. A sequencing method based on real-time pyrophosphate. *Science*, 281(5375), pp.363–365.
- Rothberg, J.M. & Leamon, J.H., 2008. The development and impact of 454 sequencing. *Nature Biotechnology*, 26(10), pp.1117–1124.
- Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A. and Arnheim, N., 1985. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, 230(4732), pp.1350–1354.
- Salipante, S.J. et al., 2014. Performance Comparison of Illumina and Ion Torrent Next-Generation Sequencing Platforms for 16S rRNA-Based Bacterial Community Profiling. *Applied and Environmental Microbiology*, 80(24), pp.7583–7591.
- Sanger, F. & Coulson, A.R., 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3).
- Sanger, F. et al., 1980. Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. *Journal of Molecular Biology*, 143(2), pp.161–178.
- Sanger, F., 1982. Sequence of Bacteriophage. *J. Mol. Biol*, 162, pp.729–773.
- Sanger, F. et al., 1978. The nucleotide sequence of bacteriophage  $\phi$ X174. *Journal of Molecular Biology*, 125(2), pp.225–246.
- Sanger, F., Nicklen, S. & Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12), pp.5463–5467.
- Sawicki, M.P. et al., 1993. Human Genome Project. *The American Journal of Surgery*, 165(2), pp.258–264.
- Schirmer, M. et al., 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*, 43(6).

- Schmidt, K. et al., 2016. Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. *Journal of Antimicrobial Chemotherapy*, p.dkw397.
- Schneider, G.F. et al., 2010. DNA Translocation through Graphene Nanopores. *Nano Letters*, 10(8), pp.3163–3167.
- Schuster, S.C., 2008. Next-generation sequencing transforms today's biology. *Nature Methods*, 5(1), pp.16–18.
- Schwartz, T. et al., 2003. Detection of antibiotic-resistant bacteria and their resistance genes in wastewater, surface water, and drinking water biofilms. *FEMS Microbiology Ecology*, 43(3), pp.325–335.
- Sender, R., Fuchs, S. & Milo, R., 2016. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biology*, 14(8).
- Shakya, M. et al., 2013. Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environmental Microbiology*, 15(6), pp.1882–1899.
- Shampo, M.A. and Kyle, R.A., 2002, July. Kary B. Mullis—Nobel Laureate for procedure to replicate DNA. In Mayo Clinic Proceedings (Vol. 77, No. 7, p. 606). Elsevier.
- Sharma, R. et al., 2005. “Unculturable” bacterial diversity: An untapped resource. *Current Science*, 89(1), pp.72–77.
- Sharma, S., Sachdeva, P. and Viridi, J.S., 2003. Emerging water-borne pathogens. *Applied Microbiology and Biotechnology*, 61(5-6), pp.424–428.
- Shendure, J. et al., 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309, pp.1728–1732.
- Shendure, J. et al., 2017. DNA sequencing at 40: Past, present and future. *Nature*, 550(7676).
- Shendure, J. et al., 2004. Advanced sequencing technologies: methods and goals. *Nature reviews. Genetics*, 5(5), pp.335–344.
- Shim, J. et al., 2013. Detection and quantification of methylation in DNA using solid-state nanopores. *Scientific reports*, 3, p.1389.
- Shokralla, S. et al., 2012. Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology*, 21(8), pp.1794–1805.
- Simoni, R.D., Hill, R.L. & Vaughan, M., 2002. The Structure of Nucleic Acids and Many Other Natural Products: Phoebus Aaron Levene. *J. Biol. Chem.*, 277(22), pp.22–24.
- Simpson, J.T. et al., 2017. Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods*, 14, p.407.
- Singer, A. et al., 2012. DNA Sequencing by Nanopore-Induced Photon Emission BT - Nanopore-Based Technology. In M. E. Gracheva, ed. Totowa, NJ: Humana Press, pp. 99–114.
- Sint, K., Wang, B.Y. & Kral, P., 2009. Selective Ion Passage through Functionalized Graphene Nanopores (vol 130, pg 16448, 2008). *Journal of the American Chemical Society*, 131(27), p.9600.

- Siwy, Z.S. & Davenport, M., 2010. Nanopores: Graphene opens up to DNA. *Nature nanotechnology*, 5(10), pp.697–698.
- Slobodkin, L.B., 2003. Just before Watson and Crick. *Nature Genetics*, 33(4), pp.451–452.
- Snow, J., 1855. *On the mode of communication of cholera*. John Churchill.
- Sović, I. et al., 2016. Evaluation of hybrid and non-hybrid methods for de novo assembly of nanopore reads. *Bioinformatics (Oxford, England)*, (May), p.030437.
- Sovic, I. et al., 2015. Fast and sensitive mapping of error-prone nanopore sequencing reads with GraphMap. *bioRxiv*, p.020719.
- Sovic, I. et al., 2015. Fast and sensitive mapping of error-prone nanopore sequencing reads with GraphMap.
- Ståhlberg, A., Krzyzanowski, P.M., Jackson, J.B., Egyud, M., Stein, L. and Godfrey, T.E., 2016. Simple, multiplexed, PCR-based barcoding of DNA enables sensitive mutation detection in liquid biopsies using sequencing. *Nucleic acids research*, 44(11), pp.e105-e105.
- Stoddart, D. et al., 2009. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proceedings of the National Academy of Sciences of the United States of America*, 106(19), pp.7702–7.
- Storm, A.J. et al., 2005. Fast DNA translocation through a solid-state nanopore. *Nano Letters*, 5(7), pp.1193–1197.
- Takken, F.L.W. & Joosten, M.H.A.J., 2000. Plant resistance genes: Their structure, function and evolution. *European Journal of Plant Pathology*, 106(8), pp.699–713.
- Tamura, K. et al., 2013. MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, 30(12), pp.2725–2729.
- Taniguchi, M. et al., 2009. Fabrication of the gating nanopore device. *Applied Physics Letters*, 95(12), pp.13–16.
- Tarraga, J. et al., 2016. HPG pore: an efficient and scalable framework for nanopore sequencing data. *BMC Bioinformatics*, 17(1), p.107.
- Tarrand, J.J. & Groschel, D.H.M., 1982. Rapid, modified oxidase test for oxidase-variable bacterial isolates. *Journal of Clinical Microbiology*, 16(4), pp.772–774.
- Theuerl, S. et al., 2019. The Future Agricultural Biogas Plant in Germany: A Vision. *Energies*, 12(3).
- Thévenot, D.R. et al., 2001. Electrochemical biosensors: Recommended definitions and classification. *Biosensors and Bioelectronics*, 16(1–2), pp.121–131.
- Thompson, L.R. et al., 2017. A communal catalogue reveals Earth’s multiscale microbial diversity. *Nature*, 551(7681), pp.457–463.
- Thundat, T.G. et al., 2005. DNA AND RNASEQUENCING BY NANOSCALE READING THROUGH PROGRAMMABLE ELECTROPHORESS AND NANOELECTRODE-GATED TUNNELING AND DIELECTRIC DETECTION, 2(12).
- Timp, W. et al., 2010. Nanopore sequencing: Electrical measurements of the code of life. *IEEE Transactions on Nanotechnology*, 9(3), pp.281–294.

- Tsutsui, M. et al., 2010. Identifying single nucleotides by tunnelling current. *Nature Nanotechnology*, 5(4), pp.286–290.
- Turnbaugh, P.J. et al., 2007. The Human Microbiome Project. *Nature*, 449(7164), pp.804–810.
- Turner, A.P.F., 2013. Biosensors: sense and sensibility. *Chemical Society Reviews*, 42(8), pp.3184–3196.
- Turner, D.J. et al., 2015. Complete assembly of novel environmental bacterial genomes by MinION™ sequencing. *bioRxiv*, p.026930.
- Urban, J.M., Bliss, J., Lawrence, C.E. and Gerbi, S.A., 2015. Sequencing ultra-long DNA molecules with the Oxford Nanopore MinION. *bioRxiv*, p.019281.
- Urbina, H., Scofield, D.G., Cafaro, M. and Rosling, A., 2016. DNA-metabarcoding uncovers the diversity of soil-inhabiting fungi in the tropical island of Puerto Rico. *mycoscience*, 57(3), pp.217–227.
- Ursell, L.K. et al., 2012. Defining the human microbiome. *Nutrition Reviews*, 70(SUPPL. 1).
- Van Dijk, E.L. et al., 2014. Ten years of next-generation sequencing technology. *Trends in Genetics*, 30(9).
- Vartoukian, S.R., Palmer, R.M. & Wade, W.G., 2010. Strategies for culture of “unculturable” bacteria. *FEMS Microbiology Letters*, 309, pp.1–7.
- Venkatesan, B.M. & Bashir, R., 2001. Nanopore sensors for nucleic acid analysis. *Nature nanotechnology*, 6(10), pp.615–24.
- Venter, J.C. et al., 2001. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507), pp.1304–51.
- Vickery, H.B. & Schmidt, C.L.A., 1931. The history of the discovery of the amino acids. *Chemical Reviews*, 9(2), pp.169–318.
- Voelkerding, K. V., Dames, S.A. & Durtschi, J.D., 2009. Next-generation sequencing: from basic research to diagnostics. *Clinical Chemistry*, 55(4), pp.641–658.
- Volden, R., Palmer, T., Byrne, A., Cole, C., Schmitz, R.J., Green, R.E. and Vollmers, C., 2018. Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proceedings of the National Academy of Sciences*, 115(39), pp.9726–9731.
- Wang, J.X. et al., 2006. Zinc oxide nanocomb biosensor for glucose detection. *Applied Physics Letters*, 88(23), pp.38–41.
- Wang, Y., Yang, Q. & Wang, Z., 2015. The evolution of nanopore sequencing. *Frontiers in Genetics*, 5(JAN), p.449.
- Wang, Z., Gerstein, M. & Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10, p.57.
- Watson, J.D. & Crick, F.H., 1953. The structure of DNA. *Cold Spring Harbor symposia on quantitative biology*, 18, pp.123–131.

- Whitehead, M.I. & Hillard, T.C., 1990. The role and use of progestogens. *Obstetrics and Gynecology*, 75(4 SUPPL.), p.59S–76S.
- Whiteley, A.S. et al., 2012. Microbial 16S rRNA Ion Tag and community metagenome sequencing using the Ion Torrent (PGM) Platform. *Journal of Microbiological Methods*, 91(1), pp.80–88.
- Whitesides, G.M., 2006. The origins and the future of microfluidics. *Nature*, 442(7101), p.368.
- Wu, H.C., Shieh, J., Wright, D.J. and Azarani, A., 2003. DNA sequencing using rolling circle amplification and precision glass syringes in a high-throughput liquid handling system. *Biotechniques*, 34(1), pp.204–207.
- Wu, R. & Taylor, E., 1971. Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage lambda DNA. *Journal of Molecular Biology*, 57(3), pp.491–511.
- Wu, R. & Kaiser, A.D., 1968. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *Journal of Molecular Biology*, 35(3), pp.523–537.
- Wu, R. & Taylor, E., 1971. Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage  $\lambda$  DNA. *Journal of Molecular Biology*, 57(3), pp.491–511.
- Wurzbacher, C. et al., 2019. Introducing ribosomal tandem repeat barcoding for fungi. *Molecular Ecology Resources*, 19(1), pp.118–127.
- Xu, M.S. et al., 2005. Conformation and local environment dependent conductance of DNA molecules. *Small*, 1(12), pp.1168–1172.
- Xu, M., Endres, R.G. & Arakawa, Y., 2007. The electronic properties of DNA bases. *Small*, 3(9), pp.1539–1543.
- Xue, Y., Wang, Y. & Shen, H., 2016. Ray Wu, fifth business or father of DNA sequencing? *Protein & Cell*, 7(7), pp.467–470.
- Yarza, P. et al., 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology*, 12, p.635.
- Yildirim, N. et al., 2018. Silymarin ameliorates uterine and ovarian damage in streptozotocin induced diabetic rat model. *Indian Journal of Biochemistry and Biophysics*, 55(2), pp.137–142.
- Yin, S., Liu, J. & Teng, L., 2016. An improved ant colony algorithm used for unscented Kalman filter. *ICIC Express Letters, Part B: Applications*, 7(11), pp.2411–2417.
- Zhou, J. et al., 2011. Reproducibility and quantitation of amplicon sequencing-based detection. *The ISME journal*, 5(8), pp.1303–1313.
- Zwolak, M. & Di Ventra, M., 2005. Electronic signature of DNA nucleotides via transverse transport. *Nano Letters*, 5(3), pp.421–424.

## Appendix I

**Figure 1.1:** Names and DSMZ catalog numbers of bacteria used to construct mock communities for the single organism and ten organism experiments and their corresponding accession numbers are shown below. The 16S rRNA genes were extracted from genome assemblies using RNAmmer<sup>1</sup> for bacteria for use in estimation of sequencing accuracy. Where genome assemblies were unavailable, the 16S rRNA gene sequence in Genbank was utilized.

Organism name	DSMZ catalog number	Genbank assembly accession number	16S rRNA gene accession number
One organism experiment			
<i>Listeria monocytogenes</i>	19094	HE999705.1	
Ten organism experiment			
<i>Aquimarina intermedia</i>	17527	-	AM113977
<i>Bacteroides vulgatus</i>	1447	CP000139	-
<i>Desulfosporosinus orientis</i>	765	CP003108	-
<i>Flectobacillus major</i>	103	ATXY00000000	-
<i>Legionella pneumophila</i>	7513	AE017354	-
<i>Listeria monocytogenes</i>	19094	HE999705.1	-
<i>Meiothermus ruber</i>	1279	CP001743	-
<i>Propionibacterium acnes</i>	16379	AE017283	-
<i>Pseudomonas aeruginosa</i>	1128	NC_009656	-
<i>Spirosoma linguale</i>	74	CP001769	-

**Figure 1.2** Titration experiment validating TruePrime RCA kit, included various incubation times (i.e., 30, 60, 90, 120, 150 and 180min) and various concentrations of plasmid like DNA molecules (i.e., 0.1, 0.5 and 1ng/μl), each sample was prepared in triplicate and the final concentration was measured with Qubit 2.0 HS dsDNA assay. The optimal results were generated with 0.5ng/μl of DNA incubated for around 150min. The samples with smaller amount of nucleic acid have not been reproducible, while the more concentrated (i.e., 1ng/μl), tend to over amplify the 16S rRNA fragment and clog the g-Tubes in the fragmentation step.

Amount of DNA	Incubation time	30min	60min	90min	120min	150min	180min
	0.1ng/μl	0	3	15	30	60	Too high
	0.5ng/μl	1	8	25	46	70	Too high
	1ng/μl	15	26	39	60	Too high	Too high
Concentration of the DNA (Qubit HS dsDNA assay)							

**Figure 1.3** Results of experiment that tested how various centrifugations speeds affects post-RCA fragment size. The optimal centrifugation speed was observed with 1800rpms, which generated DNA molecules at mean size of ~10kpbs. Faster (e.g., 5000rps) centrifugation fragmented the DNA fragments to a size that would not allow for concatemer analysis. While the lower centrifugation speed (i.e., 1000rpms) did not generate enough energy to push the nanoballs through the fragmentation hole. The same problem with clogging of the g-Tube was observed when used highly concentrated (>100ng/μ) post-RCA samples.

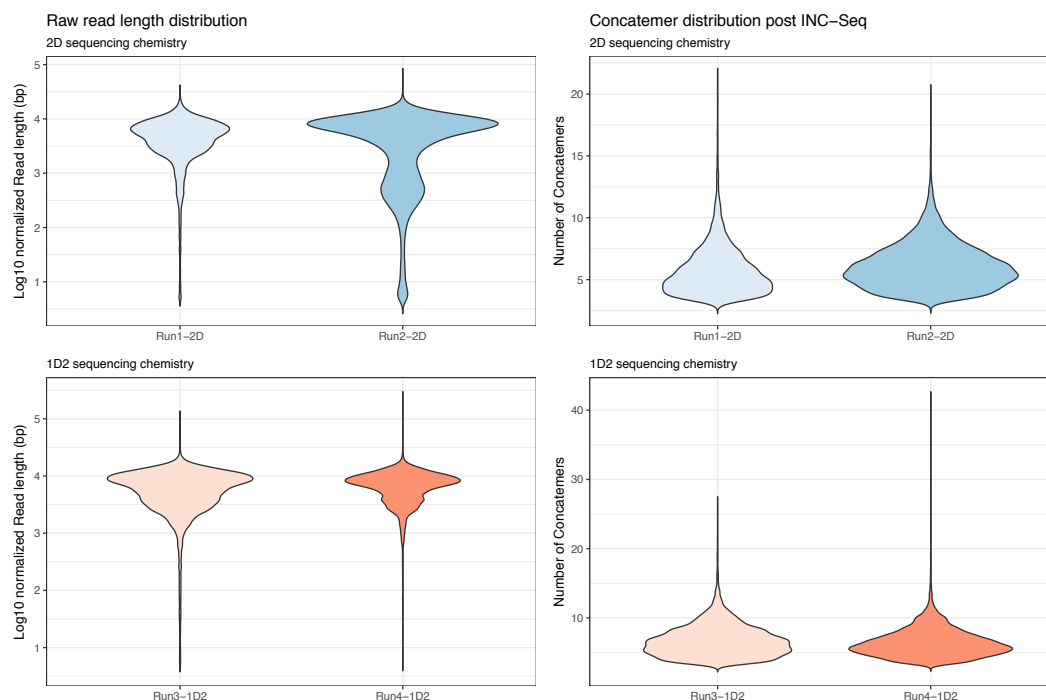


Centrifugation speed (rpms)	5000	4000	3000	2000	1800	1000
Mean fragment size (kbps)	1	2.5	5	8	10	Failed

**Figure 1.4** Table represents results from the optimization process that was performed on magnetic beads. Same amount of DNA was used for each experiment while the magnetic beads had various amounts of buffer removed, that step meant to concentrate them. That has changed their affinity towards DNA molecules size selection. Results of each experiment were verified with Qubit HS dsDNA assay and on Bioanalyser DNA chip. The best performing assay was obtained when 70% of buffer was removed that allowed for size selection of molecules longer than 1500bp. The other assays retained short unwanted DNA molecules or failed to retain any nucleic acid (i.e., -80%).

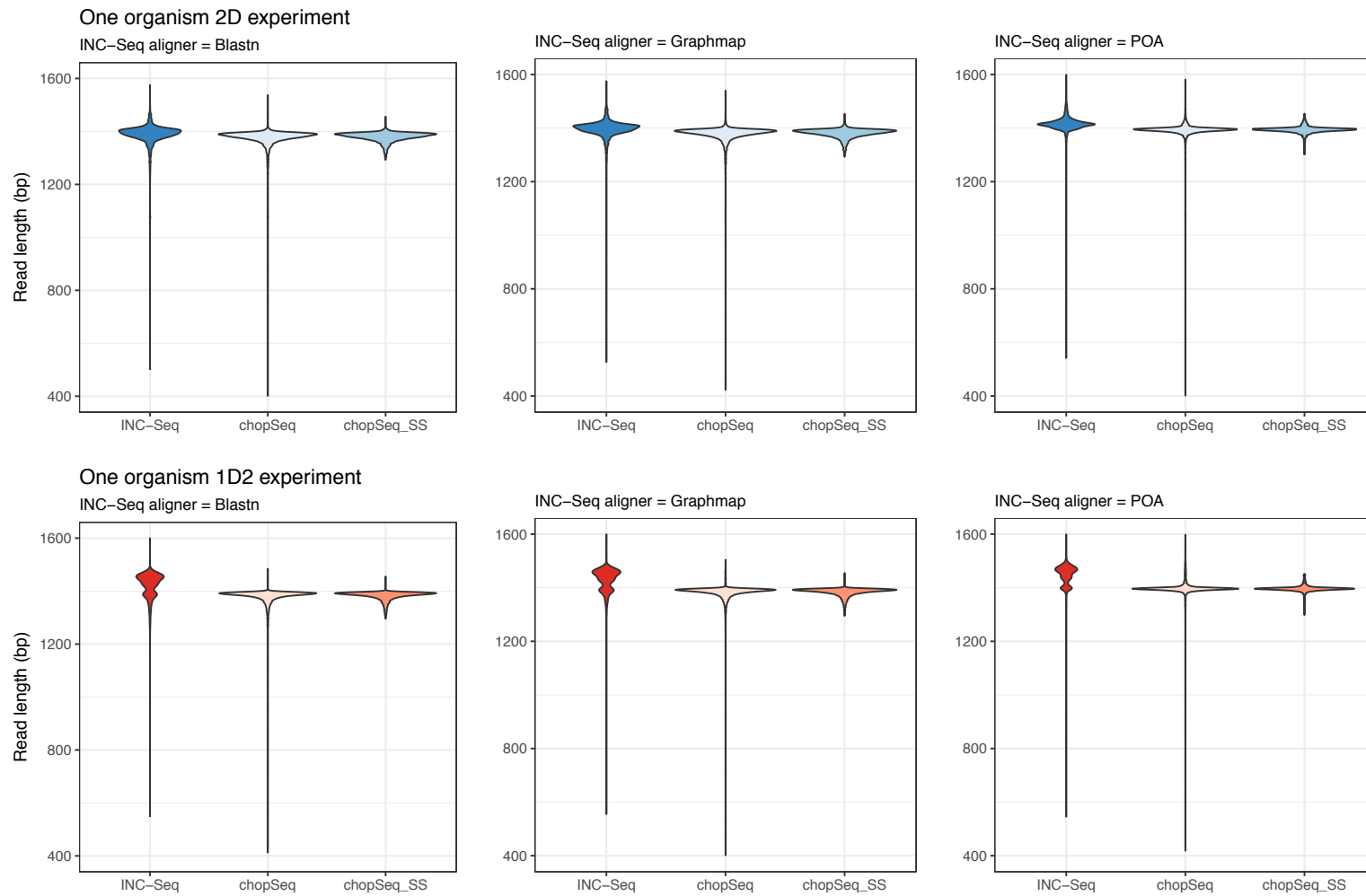
Buffer removed	Concentrated beads ratio	Initial amount of DNA	Final amount of DNA	Size of retained DNA molecules
-40%	0.35x	800ng	700ng	>800bp
-50%	0.35x	800ng	650ng	>1000bp
-60%	0.35x	800ng	600ng	>1200bp
-70%	0.35x	800ng	500ng	>1500bp
-80%	0.35x	800ng	200ng	N/A

**Figure 1.5:** Violin plots showing the read length distribution of raw data (i.e., post base calling with Albacore 1.2.4), and the number of concatemers on each base called read estimated using read lengths from INC-Seq processing using the “blastn” alignment approach for all four experiments involving both 2D and 1D2 chemistry.

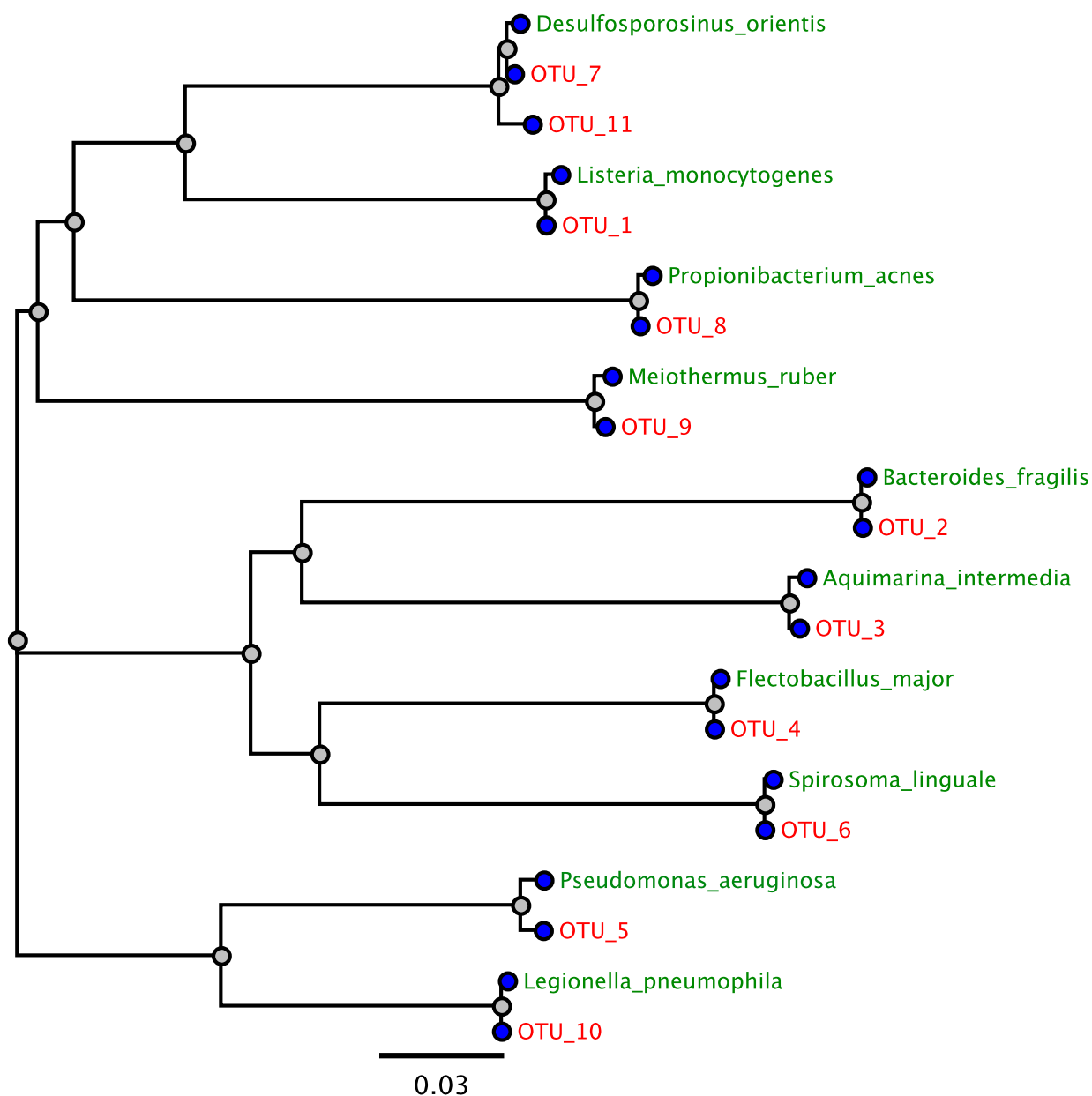




**Figure 1.6:** Violin plot of read length distributions for INC-Seq processed reads, chopSeq corrected reads after tandem repeat removal, and post-size-selection of chopSeq corrected reads and size selected reads (chopSeq\_SS) for one organism experiments using both 2D and 1D2 sequencing chemistry.



**Figure 1.7:** Consensus sequences from OTUs from Run 3 (1D2 sequencing chemistry, INC-Seq aligner: blastn) were combined with reference sequences and aligned using muscle (default parameters) and Neighbor-Joining tree using Jukes-Cantor model was constructed in Geneious (version 8) using 100 bootstraps. Reference sequences are labelled in green, and OTUs are labelled in red.



## Appendix II

### NanoAmpli-Seq - Sample processing and sequencing library preparation workflow dx.doi.org/10.17504/protocols.io.u26eyhe



NanoAmpli-Seq - Sample processing and sequencing library preparation workflow [↗](#)

Szymon T Calus<sup>1</sup>, Umer Zeeshan Ijaz<sup>1</sup>, Ameet Pinto<sup>2</sup>

<sup>1</sup>University of Glasgow, <sup>2</sup>Northeastern University

[dx.doi.org/10.17504/protocols.io.u26eyhe](https://dx.doi.org/10.17504/protocols.io.u26eyhe)

Pinto Lab



Szymon T Calus  
University of Glasgow



#### ABSTRACT

#### TAGS

sequencing

amplicon

Show tags

#### EXTERNAL LINK

<https://www.biorxiv.org/content/early/2018/01/07/244517>

#### PROTOCOL STATUS

**Working**

We use this protocol in our group and it is working very well.

#### MATERIALS

NAME	CATALOG #	VENDOR
T7 Endonuclease I - 250 units	M0302S	New England Biolabs
Blunt/TA Ligase Master Mix - 50 rxns	M0367S	New England Biolabs
NEBNext FFPE DNA Repair Mix - 24 rxns	M6630S	New England Biolabs
Magnetic stand for microcentrifuge tubes	12321D	Life Technologies
ethanol		
NEBNext Ultra II Q5 Master Mix - 50 rxns	M0544S	New England Biolabs
NEBNext End repair / dA-tailing Module (E7546)		
Corning® Filtered Pipette Tips, 1000 µL 1000 Tips	38031	Stemcell Technologies
Corning® Filtered Pipette Tips, 200 µL 960 Tips	38032	Stemcell Technologies
Corning® Filtered Pipette Tips, 10 µL 960 Tips	38034	Stemcell Technologies
Qubit dsDNA HS Assay Kit	Q32851	Thermo Fisher Scientific
DNA LoBind Tubes	#022431021	Eppendorf
PCR tubes, strips or plates		
HighPrep™ PCR	AC-60050	
Plasmid-Safe™ ATP-Dependent DNase	E3101K	Epicentre

NAME ▾	CATALOG # ▾	VENDOR ▾
Sygnis TruePrime™ RCA Kit	<a href="#">SYG390100</a>	<a href="#">Lucigen</a>
g-TUBE	<a href="#">520079</a>	<a href="#">Covaris</a>
1D <sup>2</sup> Sequencing Kit	<a href="#">SQK-LSK308</a>	
Flow Cell R9.5	<a href="#">View</a>	
Molecular Grade Water	<a href="#">60-2450</a>	<a href="#">ATCC</a>

## MATERIALS TEXT

**Equipment required:**

1. PCR thermocycler from any vendor
2. PCR cabinet/hood with UV sterilization
3. Thermal mixer with appropriate blocks from any vendor
4. Pipettes with varying volumes range from any vendor
5. Centrifuge for 2 ml and 0.2 ml tubes from any vendor
6. MinION™ MK1b device and compatible personal computer

Primers for PCR amplification of 16S rRNA gene can be ordered from any provider.

Experiment started on Nov 06, 2018 15:54:54

Finished steps: 0

## BEFORE STARTING

Make sure you have all necessary equipment, reagents and PCR primers before beginning the protocol.

For preparation of multiple samples please, use 'START EXPERIMENT' then 'SCALE PROTOCOL'.

Step 1 has not been completed

## PCR amplification of 16S rRNA gene

- 1 Master Mix - combine the following reagents using volumes below:

 **9.9 µl Molecular Grade Water**

 **12.5 µl NEBNext Ultra II Q5 Master Mix**

 **0.8 µl of 10 µM primer: Forward\_PHO+**

 **0.8 µl of 10 µM primer: Reverse\_PHO+**

Total volume:  **24 µl of reagents** in 2ml tube.

Aliquot **24 µl of Master Mix** reagents in **0.2ml PCR tubes** then add **1 µl of DNA** to each tube.

Final volume: **25 µl of reagents** in **0.2ml PCR tubes**.

Incubate the reaction at the thermocycler according to the following conditions:

 **00:00:30 sec**  **98 °C Initial denaturation**

 **00:00:05 sec**  **98 °C Denaturation**

 **00:00:10 sec**  **62 °C Annealing**

 **00:00:35 sec**  **72 °C Extension**

 **00:02:00 min**  **72 °C Final extension**

 **00:00:00**  **8 °C Hold**

This PCR assay requires **20 cycles** of Denaturation, Annealing and Extension.

During the PCR assay running go to Step 2 (prepare 70% ethanol) then to Step 3 (prepare Qubit reagents). Finally remove tubes from PCR machine and continue with Step 2.

**NOTE**

Amplification of samples can be performed in duplicates or triplicates as use of replicates will help reduce PCR biases.

Make sure both primers are 5' PHO positive. If protocol is used for the first time or was modified to amplify different gene/s than 16S rRNA, we recommend verifying the results of the PCR assay. To confirm the correct size of the amplicons, we advise using standard agarose gel or automated capillary electrophoresis.

Step 2 has not been completed

**PCR product clean up**

- 2 Prepare **5-10ml of 70% ethanol** depending on the number of samples, e.g. **3.5ml of 100% ethanol** with **1.5ml of Molecular Grade Water** and keep it on ice.

Combine replicates (3x20µl) into single **2ml tube**, then add **30µl of magnetic beads** (0.5x ratio) to the **60µl of PCR product** and incubate for **00:02:00 min at room temp**. Then place tube in the magnetic rack, allow beads to set then discard the liquid. Subsequently, wash the pellet twice with **200µl of freshly prepared ice-cold 70% ethanol**. Discard the ethanol and briefly centrifuge the tube at low speed, place the tube back on the magnetic rack, remove residuals of the ethanol and place the tube at the heat block for around **00:00:10 sec at 50°C**. Be careful not to overdry the beads. Use **20µl Molecular Grade Water** kept at the heat block **50 °C** to resuspend the beads, incubate the tube for **00:02:00 min at room temp**. Then place the tube back on the rack, allow beads to set and transfer the **20µl PCR product** into a fresh **2ml tube**.

**NOTE**

To expand the lifespan of HighPrep™ reagents and reduce the chance of contaminating 50ml stock reagents, we recommend to aliquot the beads in 2ml tubes, e.g. 1500µl. That will decrease a need for moving the stock reagents from the fridge every time a small volume of magnetic beads is used and reduce a time needed for the reagents to reach room temperature.

Step 3 has not been completed

**PCR product concentration estimation**

- 3 Prepare Qubit™ dsDNA HS reagents and related standards, according to the vendor instructions (1). i.e., **199µl dsDNA HS Buffer** with **1µl dsDNA HS Reagent** e.g., **995µl dsDNA HS Buffer** and **5µl dsDNA HS Reagent** for 5 reactions

To determine the concentration of cleaned amplicons combine:  
**1µl of PCR product** with **199µl Qubit reagents**.

**EXPECTED RESULT**

The concentration of samples should be ~50ng in a total (depending on assay efficiency) of cleaned amplicons per sample.

1) [https://assets.thermofisher.com/TFS-Assets/LSG/manuals/Qubit\\_dsDNA\\_HS\\_Assay\\_UG.pdf](https://assets.thermofisher.com/TFS-Assets/LSG/manuals/Qubit_dsDNA_HS_Assay_UG.pdf)

Step 4 has not been completed

#### Self-ligation for the formation of plasmid-like structure

- 4 If multiple amplicon pools are being used, normalized their concentration to 2ng/μl and transfer 45μl of each sample to new **0.2ml PCR grade tube**. Then add **5μl of Blunt/TA Ligase Master Mix** to the **0.2ml PCR** containing purified amplicons. Gently mix the reagents by pipetting up and down and incubate the tube for **00:15:00 min at 10°C** and **00:10:00 min at 25°C** in a PCR thermocycler (make sure lid heating is turned off).

Initiate Step 5 while the ligation reaction is ongoing in the thermocycler.

Step 5 has not been completed

#### Reverse phase cleanup

- 5 Vortex HighPrep™ reagents and transfer **150μl of magnetic beads** into a clean **2ml tube**. Place the tube on a magnetic rack for **00:02:00 min** or until beads separate from the liquid. While the tube is on the magnetic rack, remove **90μl of clear buffer** without disturbing the beads and place it inside the lid of the **2ml tube**. Then discard remaining **60μl** of the buffer without disturbing the beads. Subsequently, transfer exactly **75μl of clear buffer** (concentrate beads by 50%) from the lid into the tube and discard the remaining wastes from the tube's lid. Remove the tube from magnetic rack and gently vortex to resuspend the beads. Keep 50% **concentrated beads** at room temperature until needed.

Add precisely **17.5μl concentrated beads** (0.35x ratio) to **50μl of a self-ligation mix** from Step 4. Gently mix the tube by pipetting up and down then, incubate the mixture for **00:02:00 min at room temp**. Place the tube on a magnetic rack and allow the beads to separate. The beads will bind to long linear amplicons (i.e., chimeric amplicons) while the liquid contains short linear amplicon and plasmid-like structures. Finally, transfer **67.5μl of clear liquid** to the new **2ml tube**.

Remaining **57.5μl of concentrated beads** will be used at Step 11 and 13 until that time keep the beads at room temp.

#### NOTE

Concentration of beads by 50% described for Step 5 has to be done every time reaction is performed. We do not recommend preparing a high volume of concentrated beads and storing them. Magnetic beads stored at high-concentration (i.e. -50%) for a long time clump and lose their efficiency.

#### SAFETY INFORMATION

This approach of concentrating magnetic beads was optimised to remove long DNA structures i.e. >2000bp. However, the beads need to be prepared fresh and mixed with an **exact volume of PCR products**. That is why we recommend **using a pipette wheel** to measure the volume of liquids and **recalculate the number of reagents** added to the tube when necessary. This approach of concentrating beads is an improved version of (Additional file 6) <https://doi.org/10.1186/1471-2164-15-645>

Step 6 has not been completed

#### Plasmid and short amplicon clean-up

- 6 Add **33.75μl of magnetic beads** (0.5x ratio) to **67.5μl of clear liquid** from Step 5, gently vortex and incubate the tube for **00:02:00 min at room temp**. Place the tube on the magnetic rack and allow beads to set, discard the liquid. Subsequently, wash the pellet twice with **200μl of freshly prepared ice-cold 70% ethanol**. Discard the ethanol and briefly centrifuge the tube at low speed, place tube back in the magnetic rack, remove residuals of the ethanol and place the tube

at the heat block for around **00:00:10 sec at 50°C**. Remember to do not overdry the beads. Use **15µl of**

**Molecular Grade Water** kept at the heat block **50 °C** to resuspend the beads, incubate the tube for

**00:02:00 min at room temp**. Then place the tube back on the rack, allow beads to set and transfer the **15µl of PCR product** into a fresh **2ml tube**.

Step 7 has not been completed

#### Removal of linear molecules from plasmid mix

- 7 Combine the following reagents using volumes below:
- 15µl of self-ligated and purified amplicons**
  - 2µl of Molecular Grade Water**
  - 2µl of 25mM ATP**
  - 5µl of 10x Reaction Buffer**
  - 1µl of Plasmid-Safe DNase (10U)**

Total volume: **25µl of reagents** in **0.2ml PCR grade tube**.

For mini-preparation, incubate the reaction in a thermocycler according to the following conditions:

**37 °C Incubation** **00:15:00 min**

After incubation, clean the products as described in Step 6 (elute in **10µl of Molecular Grade Water**) and determine the concentration of DNA in the cleaned product as described in Step 3.

#### NOTE

Do not deactivate the Plasmid-safe DNase enzyme with 30min incubation at 70°C as recommended in the protocol.

This process would take additional 30min. The subsequent magnetic bead cleanup will remove ATP, Reaction Buffer, nucleotide debris and active DNase enzyme.

Reagent volumes are followed according to a mini-preparation protocol (2), for higher volumes, please check Lucigen and Epicentre instructions.

2) <https://www.lucigen.com/docs/manuals/MA044E-PlasmidSafe-DNase.pdf>

Step 8 has not been completed

#### Rolling Circle Amplification (RCA)

- 8 Perform RCA in triplicate for each pooled sample and include negative controls using **Molecular Grade Water** instead of cleaned product from Step 7. All reagents are included in the TruePrime™ RCA kit and should be kept on ice unless specified otherwise.

#1 Below are conditions and volumes for a 1 and 5 RCA reactions:

Combine **3µl of cleaned product** from Step 7 with **2.5µl of Buffer D** (provided with TruePrime™ RCA kit) in **0.2ml PCR grade tube**. Pipette up and down to mix and incubate at room temperature for **3-5 minutes**.

#2 While the sample is being incubated, prepare the amplification mix and keep on ice until needed.

1 sample	5 samples	Reagents
<b>9.3µl</b>	<b>46.5µl</b>	<b>of Molecular Grade Water</b>
<b>2.5µl</b>	<b>12.5µl</b>	<b>of Reaction buffer</b>
<b>2.5µl</b>	<b>12.5µl</b>	<b>of dNTPs</b>
<b>2.5µl</b>	<b>12.5µl</b>	<b>of Enzyme 1</b>
<b>0.7µl</b>	<b>3.5µl</b>	<b>of Enzyme 2</b>
Total volume: <b>17.5µl</b>	<b>87.5µl</b>	<b>of Amplification Mix</b>

#3 After **3-5 minutes**, add **2.5µl of Buffer N** to **5µl of incubated sample #1** and pipette up and down to mix.

Finally, add **17.5µl of amplification mix #2** to the **7.5µl of DNA mix #3**, pipette up and down to homogenise the reaction then incubate the **0.2ml PCR grade tube** at **29.5°C** on a heat block or thermocycler for **120-150 minutes** - depending on assay efficiency.

After **90min**, the efficiency of the assay can be tested according to Step 3. If RCA resulted in appropriate concentration of amplification product, then the amplification mix can be followed by enzymatic fragmentation (Step 9). However, if the concentration of the RCA products is low, then reagents can be incubated for another **30-60min** and quantified according to Step 3.

Incubation time takes around 2 hours. Take a break!

#### NOTE

Reagent volumes are used according to TruePrime™ RCA kit (3). An increase of incubation temperature will boost the efficiency of phi29 polymerase, however, that in turn will trigger an increase of unspecific product formation in the negative controls. Higher incubation temperatures (i.e. 32-36°C) may be investigated to reduce incubation time (by 40-60%) and shorten overall workflow. However, we recommend to perform reaction at ~29.5°C that makes the RCA protocol slightly longer (120-150min) but more reliable as negative control sample will have no unwanted amplification.

3) <https://p2v6h7b4.stackpathcdn.com/wp-content/uploads/2018/01/TruePrime%E2%84%A2-RCA-Kit.pdf>

#### EXPECTED RESULT

After 120-150min, the concentration of samples should be ~30-40ng/µl that in turn gives:  
25µl x triplicate x 30-50ng = 2250-3000ng in total and is sufficient for subsequent steps.  
The bare minimum concentration of RCA product at this step is 2000ng before fragmentation.

\*\* If the protocol is used for the first time or was modified to amplify different gene/s than 16S rRNA, we recommend verifying the results of the RCA assay. To confirm correct assay efficiency, we advise using standard agarose gel or automated capillary electrophoresis.

Step 9 has not been completed

#### First step enzymatic de-branching

- 9 Combine all three **21 µl RCA replicates** from Step 8 into a single **0.2ml PCR grade tube**. Mix **63µl of RCA product** with **2µl of T7 endonuclease I** and mix with use of wide bore tips then incubate for **00:05:00 min at room temp**.

Step 10 has not been completed

#### Mechanical fragmentation

- 10 Transfer **65µl of RCA product** into a g-TUBE using wide bore pipette tips. Centrifuge the tube for **00:03:00 min at 1800rpm** or until the entire reaction mix passes through the fragmentation hole. Reverse the g-TUBE and centrifuge it for **00:03:00 min at 1800rpm** or until the entire reaction mix passes through the fragmentation hole.

Step 11 has not been completed

#### Cleanup of fragmented RCA products



Gently vortex the concentrated beads (~50%) and add exactly **22.75µl of concentrated beads** (0.35x ratio) to **65µl of RCA fragmented products** (or recalculate the volume of concentrated beads if necessary). Mix the tube by pipetting up and down then, incubate the mixture for **00:02:00 min at room temp**. Subsequently, place the tube on a magnetic rack and allow the beads to separate. The beads contain long linear, concatemeric amplicons while the liquid contains short (<2000bp) fragments. Subsequently, wash the pellet twice with **200µl of freshly prepared ice-cold 70% ethanol**. Discard the ethanol and briefly centrifuge the tube at low speed, place back the tube at the magnetic rack, remove residuals of the ethanol and place the tube at the heat block for around **00:00:10 sec at 50°C**. Remember to do not overdry the beads. Use **65µl of Molecular Grade Water** kept at the heat block **50 °C** to resuspend the beads, incubate the tube for **00:02:00 min at room temp**. Then place the tube back on the rack, allow beads to set and transfer the **63µl of RCA product** into a fresh **2ml tube**. Remaining **2µl of RCA product** can be used for Qubit dsDNA HS assay.

Remaining **34.75µl of concentrated beads** will be used at Step 13 until that time keep the beads at room temp.

Step 12 has not been completed

#### Secondary enzymatic de-branching

- 12 Mix **63µl of fragmented RCA products** with **2µl of T7 endonuclease I** and incubate for **00:05:00 min at 37°C**.

Step 13 has not been completed

#### Cleanup of post fragmented RCA products

- 13 The **34.75µl of concentrated beads** from Step 11 is used for cleanup of RCA fragmented products at this stage.

Add precisely **29.25µl of concentrated beads** (0.45x ratio) to **65µl of RCA fragmented products** (or recalculate the volume of concentrated beads if necessary). Gently mix the tube by pipetting up and down then, incubate the mixture for **00:02:00 min at room temp**. Then, place the tube on a magnetic rack and allow the beads to separate. The beads contain long linear, concatemeric amplicons while the liquid contains short fragments. Subsequently, wash the pellet twice with **200µl of freshly prepared ice-cold 70% ethanol**. Discard the ethanol and briefly centrifuge the tube at low speed, place the tube back on the magnetic rack, remove residuals of the ethanol and place the tube at the heat block for around **00:00:10 sec at 50°C**. Remember to do not overdry the beads. Use **55µl of Molecular Grade Water** kept at the heat block **50 °C** to resuspend the beads, incubate the tube for **00:02:00 min at room temp**. Then place the tube back on the rack, allow beads to set and transfer the **53µl RCA product** into a fresh **0.2ml PCR grade tube**. Remaining **2µl of RCA product** can be used for **Qubit dsDNA HS assay** or verification of size fragments\*\*.

#### NOTE

\*\* If the protocol is used for the first time or was modified to amplify different gene/s than 16S rRNA, we recommend verifying the results of the PCR assay. To confirm the correct size of the amplicons, we advise using standard agarose gel or automated capillary electrophoresis.

Step 14 has not been completed

#### Gap-filling and dA-tailing of fragmented RCA products

- 14 Combine the following reagents using volumes below:  
**53µl of RCA product**

**3.5µl of FFPE DNA Repair Buffer**  
**3.5µl of NEBNext Ultra II End Prep Buffer**  
**2µl of NEBNext FFPE DNA Repair Mix**  
**3µl of Ultra II End Prep enzyme mix**

Total volume: **65µl** in **0.2ml PCR grade tube**.

Mix the reaction mix by pipetting gently up and down (10-times) using wide bore tips then incubate the reaction at the thermocycler according to the following conditions:

20 °C	00:10:00 min
65 °C	00:10:00 min
4 °C	00:00:00 Hold

#### NOTE

Reagent volumes are followed according to NEBNext® FFPE DNA Repair Mix  
<https://www.neb.com/-/media/catalog/datacards-or-manuals/manualm6630.pdf> - page 5

Step 15 has not been completed

#### Cleanup of end-repaired and dA-tailed RCA products

- 15** Add **32.5µl of magnetic beads** (0.5x ratio) to **65µl of RCA product** from Step 14, gently vortex and incubate the tube for **00:02:00 min at room temp**. After that time place tube on the magnetic rack and allow beads to set, discard the liquid. Subsequently, wash the pellet twice with **200µl of freshly prepared ice-cold 70% ethanol**. Discard the ethanol and briefly centrifuge the tube at low speed, place back the tube at the magnetic rack, remove residuals of the ethanol and place the tube at the heat block for around **00:00:10 sec at 50 °C**. Remember to do not overdry the beads. Use **35µl of Molecular Grade Water** kept at the heat block **50 °C** to resuspend the beads, incubate the tube for **00:02:00 min at room temp**. Then place the tube back on the rack, allow beads to set and transfer the **33µl of RCA product** into a fresh **2ml tube**.

Follow Step 3 to determine the concentration of DNA in samples and negative controls.

Step 16 has not been completed

#### Library preparation and nanopore sequencing (Fig.1I)

- 16** Prepare **1D<sup>2</sup> libraries** for nanopore sequencing with **SEQ-LSK308** kit by Oxford Nanopore Technologies (4).

Combine the following reagents using volumes below:

**33µl of ~500ng RCA products** from Step 15  
**2.5µl of 1D<sup>2</sup> Adapter**  
**14.5µl of Blunt/TA Ligase Master Mix**

Total volume: **50µl** in a **2ml tube**

Homogenise the reagents by pipetting gently up and down (10-times) using wide bore tips then incubate for

**00:10:00 min at room temp**. Then add **24µl of magnetic beads (0.4x ratio)** and incubate it for

**00:02:00 min at room temp**. Subsequently, place the tube on the magnetic rack and allow beads to set, discard the liquid. Wash the pellet twice with **200µl of freshly prepared ice-cold 70% ethanol**. Discard the ethanol and briefly centrifuge the tube at low speed, place back the tube at the magnetic rack, remove residuals of the ethanol and place the tube at the heat block for **00:00:10 sec at 50 °C**. Use **46µl of Molecular Grade Water** kept at the heat block

**50 °C** to resuspend the beads, incubate the tube for **00:02:00 min at room temp**. Then place the tube back

on the rack, allow beads to set and transfer the **45µl of RCA product** into a fresh **2ml tube**.

Combine the following reagents using volumes below:

**45µl of 1D<sup>2</sup> adapted DNA**  
**5µl of Barcode Adapter Mix (BAM)**  
**50µl of Blunt/TA ligase**


Total volume: **100µl** in a **2ml tube**

Homogenise the reagents by pipetting gently up and down (10-times) using wide bore tips then incubate for

 **00:10:00 min at room temp**

. After that time add **40µl of magnetic beads (0.4x ratio)** and incubate it for

 **00:02:00 min at room temp**

. Subsequently, place the tube on the magnetic rack and allow beads to set, discard the liquid. Wash the pellet twice with **140µl of ABB buffer**. Discard the ABB buffer and briefly centrifuge the tube at low speed, place the tube back in the magnetic rack, remove residuals of the ABB buffer. Use **15µl of Elution Buffer** kept at the room temperature to resuspend the beads, incubate the tube for  **00:02:00 min at room temp**. Then place the tube back on the rack, allow beads to set and transfer the **15µl of Elution Buffer** into a fresh **2ml tube**.

#### EXPECTED RESULT

Eluted 1D<sup>2</sup> libraries can be quantified with the use of Qubit reagents according to Step 3.  
 Expected concentration should be around 8-10ng/µl or 120-150ng in total.

Prime **R9.5 nanopore flow cell** (suitable for 1D<sup>2</sup> libraries) using protocols outlined by Oxford Nanopore Technologies (4). Load the libraries on the device and initiate 48h sequencing process.

#### NOTE

Some of the reagent volumes at Step 16 have been modified when compared to 1D<sup>2</sup> genomic protocol.  
 4) [https://community.nanoporetech.com/protocols/1d%5E2-genomic-sequencing/v/lzd\\_9032\\_v11\\_revo\\_23mar2017/checklist-protocol](https://community.nanoporetech.com/protocols/1d%5E2-genomic-sequencing/v/lzd_9032_v11_revo_23mar2017/checklist-protocol)

## NanoAmpli-Seq – Bioinformatics workflow

[dx.doi.org/10.17504/protocols.io.u25eyg6](https://dx.doi.org/10.17504/protocols.io.u25eyg6)



### NanoAmpli-Seq - Bioinformatics Workflow

Szymon T Calus<sup>1</sup>, Umer Zeeshan Ijaz<sup>1</sup>, Ameet Pinto<sup>2</sup>

<sup>1</sup>University of Glasgow, <sup>2</sup>Northeastern University

[dx.doi.org/10.17504/protocols.io.u25eyg6](https://dx.doi.org/10.17504/protocols.io.u25eyg6)

Pinto Lab



Szymon T Calus

University of Glasgow



#### ABSTRACT

#### TAGS

amplicon

nanopore

Show tags

#### EXTERNAL LINK

<https://www.biorxiv.org/content/early/2018/07/04/244517>

#### PROTOCOL STATUS

##### Working

We use this protocol in our group and it is working very well.

Experiment started on Nov 06, 2018 15:53:54

Finished steps: 0

#### BEFORE STARTING

Make sure all the necessary programs and dependencies are installed on your PC or server and work correctly.

Step 1 has not been completed

Download and install all the required software.

1

#### SOFTWARE

##### Albacore v2.3.3

Linux

[source](#) by Oxford Nanopore Tech.

#### SOFTWARE

##### INC-Seq

Linux

[source](#) by Genome Institute of Singapore

 SOFTWARE

### chopSeq v0.3

Linux

[source](#) by University of Glasgow

 SOFTWARE

### nanoCLUST v0.4

Linux

[source](#) by University of Glasgow

Step 2 has not been completed

Basecalling of raw nanopore data with Albacore software.

2

#### COMMAND

```
# Program requires input data (-i), version of the flow cell (-f),
# version of the sequencing kit (-k), output file (-o),
# amount of cores used for analysis (-t) and saving directory (-s).
```

```
/home/opt/.pyenv/versions/3.5.0/bin/full_1dsq_basecaller.py -i data/ -f FLO-MIN107 -k SQK-LSK308 -o fasta -t 20 -s .
```

Raw data (HDF5) generated with MinKNOW has to be basecalled with Albacore v2.3.3 (or newer) software. The output of the basecalling should be in FASTA format. Further analysis requires 1D2 data only so, full\_1dsq\_basecaller.py algorithm must be used.

Step 3 has not been completed

Consensus calling of long 16S rRNA concatemerized reads with use of the INC-Seq algorithm.

3

**COMMAND**

```
# Export all necessary PATH's for the required programs.
# These PATH's are specific to our cluster and may differ
# to yours, depending on where you have installed these programs.
```

```
export PYENV_ROOT="/home/opt/.pyenv"
export PATH="$PYENV_ROOT/bin:$PATH"
eval "$(pyenv init -)"
export PYTHONPATH=/home/opt/INC-Seq/utlis:$PYTHONPATH
export PATH=/home/opt/pacb/bin:$PATH
export PATH=/home/opt/pbdagcon/src/cpp:$PATH
export PATH=/home/opt/ncbi-blast-2.2.28+/bin:$PATH
export PATH=/home/opt/INC-Seq:$PATH
export PATH=/home/opt/.pyenv/versions/3.4.0/bin:$PATH
```

```
# INC-Seq consensus calling requires input data (-i),
# aligner (-a) e.g. poa, output file name (-o),
# minimum number of concatemers (--copy_num_thre) and --iterative.
```

```
inc-seq.py -i input.fasta -a poa -o incseq.fasta --iterative --copy_num_thre 3
```

The INC-Seq software requires basecalled data (e.g. Albacore) from Step 2. Correction of the data with INC-Seq algorithm uses only 1D2 data and is divided into two main steps:

- 1) Identification of segments made of 16S rRNA genes.
  - 2) Anchor alignment of concatamerised amplicons and consensus calling with PBDAGCon.
- Corrected reads have got ~98% accuracy and can be directly used as an input for chopSEQ software.

Linux

Step 4 has not been completed

Correction of wrongly oriented reads and size filtration with a chopSeq algorithm.

4

**COMMAND**

```
# Algorithm requires input data (-i) from previous step,
# forward (-f) and reverse (-r) primer sequence,
# lower (-l) and maximum (-m) size filtration range,
# and new file destination (> new_file.fasta),
# while verbosity (-v) mode is optional.
```

```
chopSEQ.py -i incseq.fasta -f "AGRGTTCGATCMTGGCTCAG" -r "GGGCGGWTGTACAAGRC" -l 1250 -m 1500 -v > chopseq.fasta
```

The chopSeq requires INC-Seq corrected data from Step 3.

Correction of the data is divided into multiple steps:

- 1) Identification of forward and reverse primers (e.g. 8F and 1387R) with pairwise2 aligner.
- 2) Re-orientation of incorrectly concatamerised reads and removal of tandem repeats recognised with use of etandem (EMBOSS) and subsequent merging of reads.
- 3) Size filtration with Biopython.

Now reads are qualified for nanoClust OTU binning and consensus calling.

Linux

Step 5 has not been completed

Read binning and generation of OTUs with a nanoCLUST algorithm.

5

#### COMMAND

```
# Export all necessary PATH's for the required programs.
# These PATH's are specific to our cluster and may differ
# to yours, depending on where you have installed these programs.

export PATH=/home/opt/vsearch/bin:$PATH
export PATH=/home/opt/mafft-7.273-without-extensions/core/bin:$PATH
export MAFFT_BINARIES=/home/opt/mafft-7.273-without-extensions/core/libexec/mafft

# Provide chopSeq corrected data (-i) and window split
# range (-s) for read partitioning and output folder (-o).

nanoCLUST.py -i chopSEQ.fasta -s 0,450,451,900,901,1300,-1 -o nanoclust_output/

The nanoCLUST requires chopSEQ corrected data from Step 4.
Correction of the data is divided into multiple steps:
1) Data is partitioned (i.e. 1-450,451-900, 901-1300bp).
2) Reads from each partition are grouped according to 97% similarity with VSEARCH.
3) VSEARCH partition dereplication, singleton removal and binning are performed on split data.
4) Optimal read sections are used for clustering.
5) MAFFT G-INS-i is used for within OTU alignment and consensus calling of data.
6) Consensus sequences are generated (~99.5% accuracy).
7) Abundance table is generated.
```

Linux