APPLICATION OF NANOPORE DATA TO THE PIMMS (Pragmatic Insertional Mutation Mapping System) SEQ PIPELINE.

Matthew Carlile

Thesis submitted to the University of Nottingham
for the degree of
Master of Research

September 2022

Abstract

Significant developments in next generation sequencing technologies have over the past 15 years allowed novel and exciting opportunities to further our understanding of genomic science. Transposons are mobile regions of DNA that occur frequently throughout the genome, the insertion of these transposable elements play an essential role in both gene regulation and evolution. Locating the positions of transposon insertion into a genome can help our understanding of the role they play in such areas as gene expression and structural variation. Random transposon mediated mutagenesis has proven to be a useful tool in the identification of essential and conditionally essential genes within various bacterial species. Various methods have been generated to identify the regions within a given genome where transposon insertion has occurred. Some of these methods involve complex experimental design in the laboratory, and a high degree of competency in bioinformatics post sequencing.

PIMMS (Pragmatic Insertion Mutation Mapping System) Seq was introduced in 2016 to speed up and simplify the bioinformatic pipeline when mapping transposon insertion sites using random mutagenesis sequencing data. The bacterial samples in question undergoing random mutagenesis, followed by inverse PCR, library preparation and sequencing (using Illumina technology). The use of Nanopore sequencing technologies has steadily increased since its introduction in 2014. Although the data quality is not yet as accurate as Illumina, Nanopore sequencing does offer some unique advantages, including the ability to sequence native DNA, generate long sequencing reads, and facilitate real time data analysis.

Oxford Nanopore Technologies (ONT), the company that developed nanopore sequencing also offer some unique sample enrichment options such as Cas-9 targeted sequencing which, for PIMMS could potentially be used as an alternative to inverse PCR, targeting sequences that contain known transposon motifs that flank a genomic region of interest.

Nanopore sequencing data was generated using *S. agalactiae* insertion mutant libraries, and up-loaded to the PIMMS Seq bioinformatic pipeline to generate a database of transposon insertion sites. When comparing PIMMS output data generated using Nanopore sequencing to the established Illumina short read sequencing method, although the genomic distribution of insertions where consistent, differences where observed with Nanopore sequencing identifying 405 more unique insertion events, reducing the list of genes that were considered to be essential using short read sequencing. However, any advantages of this long read sequencing method must be off set against Nanopore's sequencing base calling accuracy, which was significantly lower than that generated using Illumina technologies.

The Cas-9 Targeted library preparation from ONT has been proven to enrich for transposon sequences that flank genomic regions of interest, with on average 68.1% of all reads mapping to the insert sequence. This establishes its potential as an attractive alternative to amplification based methods. This method may significantly reduce the total amount of sequencing required to run an experiment, which in turn will reduce time and costs. Any potential wet lab sample loses are minimised due to a far smaller amount of sample processing steps.

Moving forward, nanopore long read sequencing is a technology that offers some unique advantages over short read technologies, and its application to the PIMMS pipeline has been demonstrated to work well. The added option of Cas-9 Targeted nanopore sequencing, saves not only lab time and costs, but more importantly removes any potential amplification bias.

Acknowledgements

The fact that this report has been completed and handed in is an accomplishment in itself, and would not have been possible without real time sacrifices being made by members of my family, in particular my parents and my wife.

Within the School of Veterinary Sciences, I am very grateful to my supervisor Adam Blanchard, for his advice and always being on hand to answer my frequent questions, and to Sharon Egan for generating the extracted DNA samples for all sequencing runs.

Within the School of Life Sciences, I am indebted to my line manager, Matt Loose, for allowing me the time and space necessary to not only complete this research project, but to spend one day a week away from a very busy lab for the past two years. My work colleagues have been very understanding in my absence, a particular mention to Chris for running the Sanger sequencing service for me, and Nadine for her valuable advice when setting up the Cas-9 Targeted sequencing runs.

Three months after undertaking this Mres in 2020, my Aunt Susan passed away from lung cancer at the young age of 65. Her absence from our lives is keenly felt, and this report is dedicated to her memory.

Lastly, I must thank my wife Carrie, for her support and her strength, and for her continual belief in her husband.

Table of Contents

Abstract 2

Acknowledgements 3

List of Figures 7

List of Tables 9

1.Introduction 10

1.1 The Evolution of Genome Sequencing 10

1.1.1 Sanger Sequencing 10
1.1.2 High Throughput Sequencing technologies 12
1.1.3 Long Read Sequencing 14

1.2 Transposons 16

1.2.1 The essential genome 16
1.2.2 Transposon mediated mutagenesis 17
1.2.3 High Throughput Insertion Mutagenesis Sequencing 17
1.2.4 PIMMS Seq 17

1.3 Utilising Technological advances 19

1.3.1 PIMMS using Nanopore Sequencing 19
1.3.2 Cas-9 Targeted Sequencing 19

1.4 Aims of project 21

1.4.1 Assess the application of Nanopore long read sequencing to identify insertions in a mutagenised bacterial library 21
1.4.2 Objectives 22

2. Methods 23

2.1 Samples 23

2.2 Long Read PIMMS Laboratory Method 23

2.2.1 Sample QC 24
2.2.2 Library Preparation and sequencing 24
2.2.3 DNA Repair and End-Prep 24
2.2.4 First AMpure Clean-up step 24
2.2.5 Adapter ligation 25
2.2.6 Second AMpure Clean-up Step 25
2.2.7 Library and Flow Cell QC 25
2.2.8 Library Loading 25
2.2.9 Sequencing and Data Collection 26

2.3 Cas-9 Targeted Enrichment Laboratory Method 26

2.3.1 Sample QC 26
2.3.2 Library Preparation and Sequencing 26
2.3.3 crRNA design 27
2.3.4 Preparation of Ribonucleoprotein (RNP) Complex 27
2.3.5 Dephosphorylate Input DNA 27
2.3.6 Cleave and dA Tail target DNA 28
2.3.7 Adapter ligation 28
2.3.8 AMpure Clean-up step 28
2.3.9 Library and Flow Cell QC 28
2.3.10 Repeat Sequencing Run for ds1093_3 29

2.4 Initial Data Analysis 29

2.4.1 Read Quality Control 29
2.4.2 Manual Data Interrogation 29

2.5 PIMMS 30
2.5.1 Find Flank Module 31
2.5.2 BAM Extract Module 31
2.5.3 Comparison of Sequencing Methods 32
2.5.4 PIMMS Dashboard 32

3. Results 33

3.1 Cas-9 Targeted laboratory Method 33
3.1.1 Sample QC prior to library preparation 33
3.1.2 Library QC 34
3.1.3 Sequencing run QC 34

3.2 Manual Alignments, Cas-9 Targeted Method 35
3.2.1 Read Filtering 35
3.2.2 Data Set Merging 38
3.2.3 Manual confirmation of Insert-Chromosome alignments 39

3.3 PIMMS Inverse PCR Laboratory Method 40
3.3.1 Sample Quality Control prior to Library Preparation 40
3.3.2 Library Quality Control 41
3.3.3 Sequencing run QC 42

3.4 PIMMS Output – Illumina vs Nanopore 44
3.4.1 Insertion locations in the *S. agalactiae* genome 44
3.4.2 Normalised Insertion Mapped Scores 44
3.4.3 Identification of Essential Genes 47

3.5 PIMMS Output – Illumina vs Cas-9 Targeted Sequencing 48
3.5.1 Insertion locations in the *S. agalactiae* genome 48
3.5.2 Normalised Insertion Mapped Scores 49
3.5.3 Identification of Essential Genes 51
3.5.4 NIM scores, Illumina, Nanopore and Cas9 52
3.5.5 Normalised Reads Mapped Scores, Illumina, Nanopore and Cas-9 53

4. Discussion 55

4.1 Cas-9 Targeted enrichment of pGh9:IS*S1* insertion mutant libraries. 55
4.1.1 Amplification Bias 55
4.1.2 Expected observation of Cas-9 sequencing enrichment 55
4.1.3 Previous Cas-9 Targeted sequencing strategies 56
4.1.4 Optimisation of the Cas-9 Targeted Sequencing run 57

4.2 Identification of Transposon Insertions using Manual Alignment 59
4.2.1 Quality Control of Nanopore Sequencing Data 59
4.2.2 Mapping Long Reads 59
4.2.3 PAF Filtering 60

4.3 Utilizing Nanopore Sequencing for the PIMMS pipeline 60
4.3.1 Input Data Quality 60
4.3.2 Illumina and Nanopore PIMMS output using Inverse PCR 61
4.3.3 Cas-9 Targeted Sequencing and the PIMMS Pipeline 62
4.3.4 Comparison of Sequencing methods 63

4.4 Future Work 63
4.4.1 Further optimisation of the Cas-9 Targeted sequencing method 63
4.4.2 An alternative approach to crRNA design. 64
4.4.3 Nanopore technological innovation 65

5. Conclusions 67

Bibliography 68

Appendix 1 72
TERMINAL COMMANDS USED FOR MANUAL BIOINFORMATIC PIPELINE 72
COMMANDS USED IN R STUDIO 72
TERMINAL COMMANDS USED TO RUN PIMMS PIPELINE 73
Appendix 2 74
Appendix 3 77
Pimms2.config file 77

List of Figures

Figure 1.1 – Fluorescently labelled DNA fragments separated by capillary electrophoresis. 11

Figure 1.2 – Electropherogram of a basecalled DNA sequence using sanger sequencing. 11

Figure 1.3 – Standard Illumina library preparation – from IDT website, cited 2022. 12

Figure 1.4 – The reduced cost of genome sequencing. From National Human Genome research institute (cited 2022). 13

Figure 1.5 - Nanopore Sequencing - Image from Research Genome Limited (2022). 15

Figure 1.6 - Schematic of PIMMS bioinformatic pipeline (Blanchard et al, 2015). 18

Figure 1.7 - Representation of RNP complex interaction with ROI (Oxford Nanopore Technologies, 2018) 20

Figure 1.8 - Standard workflow for Cas-9 targeted sequencing using ONT (Oxford Nanopore Technologies, 2018). 21

Figure 2.1 - Laboratory workflow to generate PIMMS processed DNA. 23

Figure 2.2 – Laboratory workflow of a Cas-9 Targeted Library Preparation. 26

Figure 2.3 - Manual bioinformatic pipeline 30

Figure 3.1 – TapeStation profiles of extracted DNA sample ds1093_3 (UK15_BHI_WholeDNA). 33

Figure 3.2 – Bar chart showing the number of reads removed with quality filters set to Q9. 35

Figure 3.3 – Scatter plots showing the variation in reads length generated for each nanopore run vs the quality score of each read (post filtering of reads <Q9) Source NanoPlot. 35

Figure 3.4 – Porechop filtering of fastq reads. 36

Figure 3.5 – Filtered reads for ds1093_3. 37

Figure 3.6 – Filtered reads for ds1093_3_Rpt. 38

Figure 3.7 – Alignment Pairwise Venn Diagrams for ds1093_3 and ds1093_3_Rpt. 39

Figure 3.8 – TapeStation profile and trace for ds1093_1 DNA sample. 41

Figure 3.9 – ds1093_1 reads pre and post filtering. 42

Figure 3.10 – Quality control plots for ds10983_1 43

Figure 3.11 – Scatterplot showing depth and position of mutations plotted across the S. agalactiae genome. 44

Figure 3.12 – NIM (Normalised Insertions Mapped) scores represented across the S. agalactiae genome using Illumina and Nanopore PIMMS output. 45

Figure 3.13 – A magnified look at NIM scores for both Illumina and Nanopore PIMMS data sets. 45

Figure 3.14 – Histogram of NIM scores for Illumina and Nanopore PIMMS output. 46

Figure 3.15 – BoxPlot showing the NIM score distribution for Illumina and Nanopore PIMMs output. Source SPSS (V26). 47

Figure 3.17 – Venn Diagram showing incidence of insertion absence. 48

Figure 3.18 – Scatterplot showing depth and position of mutations plotted across the S. agalactiae genome. 49

Figure 3.19 – NIM (Normalised Insertions Mapped) scores represented across the *S. agalactiae* genome using Illumina and Cas-9 Targeted PIMMS output. 50

Figure 3.20 - A magnified look at NIM scores for both Illumina and Cas-9 PIMMS data sets. 50

Figure 3.21 - Histograms for Cas-9 NIM scores. 51

Figure 3.22 – BoxPlot showing the NIM score distribution for Illumina and Cas-9 PIMMs output. Source SPSS (V26). 51

Figure 3.23 – Venn Diagram showing incidence of insertion absence. 52

Figure 3.24 – BoxPlot showing the distribution of NIM scores for Illumina, Nanopore and Cas-9 methods. Source SPSS (v26) 53

Figure 3.25 – Boxplot of NRM (Normalised Reads Mapped) score distributions for all three methods. Source SPSS (v26). 54

Figure 4.1 – A representation of the Cas 9 'Single Cut and Run' approach (from ONT, 2022). 56

Figure 4.2 – Read length distribution of project previous Cas-9 project (from NanoPlot v1.38.1) 57

Figure 4.3 – Read length distribution of ds1093_3_Rpt. From NanoPlot (v1.38.1) 58

Figure 4.4 – Chop-Chop home page 65

Figure 4.5 – Output from Chop-Chop search tool using the pGh9:IS*S1* fasta sequence. 65

List of Tables

Table 2.1 - Summary of sample ID and library preparation method. 23

Table 2.2 – crRNA sequence used for both Cas-9 Targeted library preparations. 27

Table 2.3 – Requirements file for PIMMS 31

Table 3.1 – Summary of library preparation metrics, sequencing run conditions and output. 34

Table 3.2 – Summary of read alignments after Inner.Join of filtered paf files - for the insert (Pgh9:IS*S1*) and the reference (UK15 – *S. agalactiae).* 39

Table 3.3 – Sequencing read from run ds1093_3. 40

Table 3.4 – Results of BLAST search for read 493a4283-8f0e-4de7-a871-2f16e8c921e7. 40

Table 3.5 – ds1093_1 library preparation and run parameters. 42

Table 3.6 – NRM summary statistics for Illumina, Nanopore and Cas-9 experiments, generated using SPSS (v26). 53

Table 4.1 – Overview of run metrics from all three nanopore sequencing runs. Taken from NanoPlot (v1.38.1) html report files. 61

1.Introduction

1.1 The Evolution of Genome Sequencing

Automated DNA sequencing is a core research tool typically used within every research biochemistry laboratory. It is used to determine the sequence of DNA (Deoxyribonucleic Acid), or the genetic code, that serves as the blueprint of life for every organism on Earth. For young research scientists in the 21st century, genome sequencing, that is the sequencing of an organisms entire genetic code, is an accessible tool used within both universities and industrial settings. Over the past 20 years the cost of sequencing an organism's genome has significantly reduced, as has the time required to do so. Scattered throughout the scientific international community are platform facilities that house various sequencers, some of which use differing technologies to achieve their goals.
It is understandably difficult for early stage researchers to realise just how far the scientific community has advanced in such a small period of time; with regard to genome analysis.

Since the identification in 1953 of the double-helical structure of DNA (Watson et al, 1953), scientists have moved quickly forward in an almost obsessive manner, to generate the methods and technologies required to sequence, analyse and interpret the genetic code of an organism. Understanding this code, linking the 'genotype' to the 'phenotype' (the genetic traits to the observable traits) would facilitate greater study of an organisms function, and importantly, a greater understanding the organisms disease state.

Nucleic acid sequencing was a relatively late arrival for the analysis of biological macromolecules, prior to its arrival in the late 1970s protein sequencing was the primary tool for obtaining coding information found in the molecules of life (de Chadarevian, 1999). Protein sequencing was at the time a slow and expensive method, however with technological advances in DNA analysis the sequence of a protein can now be determined from a DNA sequence generated in hours.

1.1.1 Sanger Sequencing

Sanger sequencing, using dye terminator chemistry was developed by Frederick Sanger in 1977 (Sanger et al, 1977), sanger sequencing relies on a modified Polymerase Chain Reaction (PCR), a method for DNA amplification. During a typical PCR reaction, a template DNA is separated into its individual strands by denaturation at a high temperature, once separated a primer (small custom designed DNA sequence – referred to as an oligonucleotide) is annealed to the template strand at a position pre-designed by the user. An extension phase uses a DNA polymerase enzyme and individual nucleotide building blocks or dNTPs (dinucleotide triphosphates) to build a new template chain. This process is repeated a number of times (cycles) dependent on how much DNA you wish to amplify.

All components required for this method are placed together in a PCR reaction. For Sanger sequencing to work one extra component is added to make a sequencing reaction. Dye terminator chemistry relies on the addition of ddNTPs (dideoxynucleotide triphosphates) for the extension step. These molecules are incorporated randomly whilst a DNA chain is being extended. Once incorporated the chain extension stops as ddNTPs lack the 3'-hydroxyl group required for nucleotide bonding, each ddNTP will have a fluorescein attached, with each of the four bases represented by a different color. The proportion of dNTPs to ddNTPs is carefully managed to ensure the end result is a collection of DNA fragments of different sizes, each with a fluorescently labelled ddNTPs as its termination molecule. These fragments are separated by size using electrophoresis, and a laser is utilized to detect the fluorescence of each terminal molecule.

The electrophoresis of sequencing reaction products is currently carried out within very fine capillaries, the fluorescent signature of each base from a capillary electrophoresis run is recorded (Figure 1.1) in which each color represents one of the four bases, after which an electropherogram (Figure 1.2) is generated, each peak representing a base in the sequence.

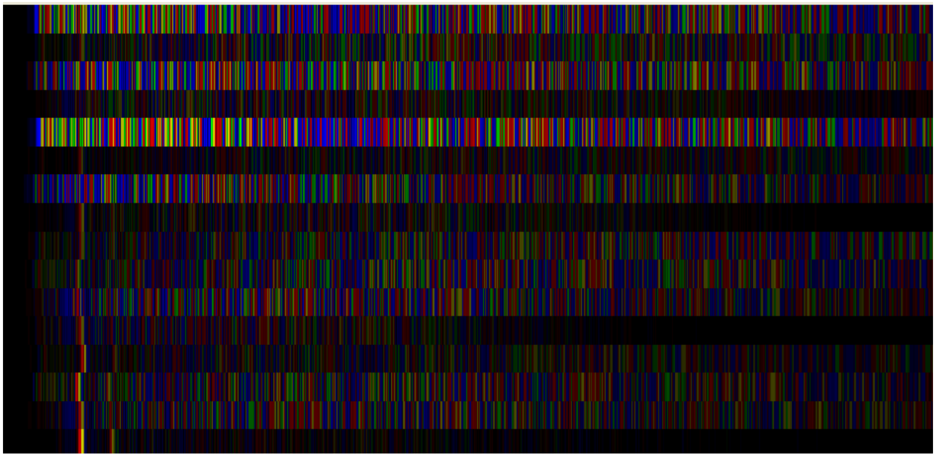


Figure 1.1 – Fluorescently labelled DNA fragments separated by capillary electrophoresis.
Each color represents a fluorescently labelled base, excited by an argon laser and recorded by the sequencing analysis software. (Source- Life Technologies)

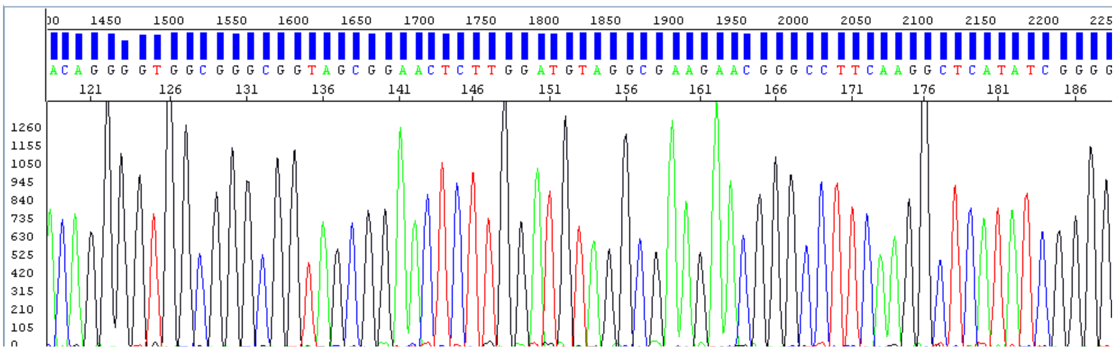


Figure 1.2 – Electropherogram of a basecalled DNA sequence using sanger sequencing.
Each peak represents a single base, blue confidence bars can be visualized at the top of the electropherogram, indicating how confident the base calling software is that the base call is correct.

Because of the straightforward and repetitive nature of Sanger sequencing, its application is now performed in centralised facilities where automated machines carry out the reactions and data analysis. The accuracy and accessibility of this core DNA sequencing technology means that it is still used heavily today and is referred to as the Gold standard (Arteche-López et al, 2021).

Sanger sequencing was used to sequence the first human genome (Human genome project, HGP) from 1990 to 2003, a global scientific endeavour that brought together multiple countries (IHGSC, 2001). Thirteen years and a cost of around 3 billion dollars, for one human draft genome. This draft was focused upon the sequencable euchromatin regions of the human genome which make up around 92.1%, around 2.9 out of 3.2 billion bases, A quality assessment of the draft human genome in 2004 identified that over 92% of these euchromatin regions had been sequenced to an accuracy of 99.99% (Schmutz et al, 2004). It was however obvious at the time that Sanger Sequencing was too expensive and took far too long. It was an impractical tool to use for large scale sequencing projects.

1.1.2 High Throughput Sequencing technologies

Significant investments in both time and money have seen the development of the next generation of genetic analysers. At the end of the 2000s one company in particular had come to dominate the market for genome sequencers. Illumina (established in 2006) helped develop and automate the sequencing by synthesis method (SBS) facilitating the generation of large amounts of sequence data, efficiently and at a fraction of the cost.

The fundamental steps for Illumina sequencing (Figure 1.3) involve, fragmenting your DNA of interest so that you have a collection of fragments that are compatible with the read lengths generated by the sequencer (mechanical or enzymatic fragmentation). Once sheared the DNA ends are prepared and adapter sequences (P5 and P7) ligated to each end. These adapters have a unique sequence that will bind to the Illumina Flow Cell. A second PCR step is used to amplify the amount of material, samples can also be multiplexed using indexing methods, the indexes themselves being typically added during the final PCR amplification step. After sequencing the data for each sample can be ascertained by demultiplexing.

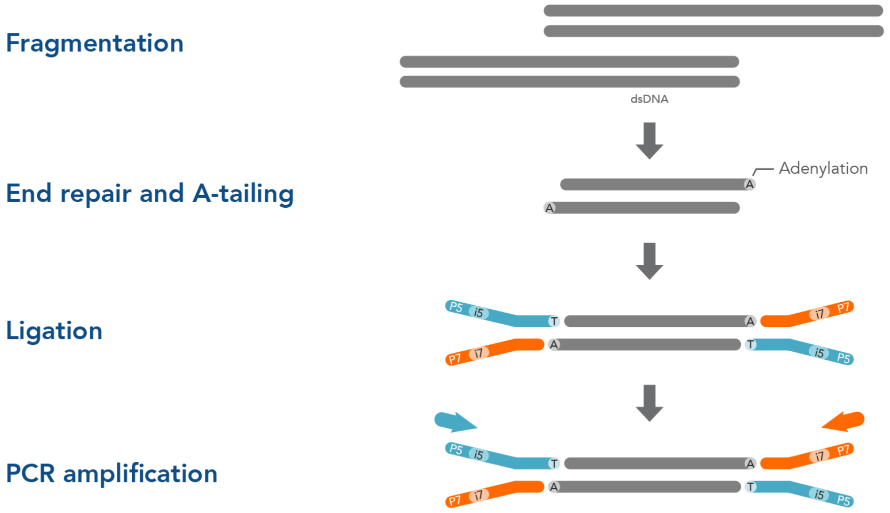


Figure 1.3 – Standard Illumina library preparation – from IDT website, cited 2022.
Genomic DNA is fragmented, after which the ends of each fragment are repaired and A tailed, sequencing adapters (P5 -blue and P7 -orange) are ligated to each library and indexes added (i5 and i7) using amplification (PCR).

The collection of DNA fragments that have been fragmented, end prepared, adapter ligated and amplified are called DNA libraries. These libraries are pooled together at a particular concentration and loaded onto a flow cell. An Illumina flow cell is coated will two oligonucleotides that the adapter sequences bound to DNA libraries (P5 and P7) will hybridise to. Once bound a strand is copied using bridge amplification, with multiple copies of each strand being generated on the surface of the flow cell (cluster generation). Sequencing then proceeds with fluorescently labelled nucleotides being added to each strand followed by imaging.

The first sequencer released by Solexa (a company Illumina purchased in 2007) was the Genome Analyzer in 2006 which had the capacity to sequence 1 gigabase (Gb) of data in a single run. Illumina have throughout the past 15 years frequently released enhanced and optimised sequencers to cater for both small, medium and large scale sequencing projects. The last of which in 2017 was the NovaSeq which can sequence terabases (Tb) of data in a single run.

The read lengths generated by Illumina sequencing are short compared to Sanger sequencing. A Sanger read will confidently read out to around 800-900 bases. Illumina sequencing has a maximum of 300 bases. However, it is the sheer volume of sequence reads generated that sets the Illumina method apart. Whereas Sanger sequencing will give you kilobases (Kb) of data in a single run. Illumina will typically generate Gbs. Add to this the fact that Illumina facilitates the reading of both ends of a sequencing read (PE – or paired end sequencing), having a sequence read from both ends designed to overlap gives real confidence in the accuracy of the sequence data generated.

This heavy investment in DNA sequencing technologies has significantly reduced the cost per genome, twenty years ago the average cost of sequencing a human genome would be around $100M, towards the end of the same decade this cost dropped to below $100,000 (Figure 1.4). In 2019 the National Human Genome Research Institute put the cost of sequencing a complete human genome at $942 (Adewale BA, 2020).

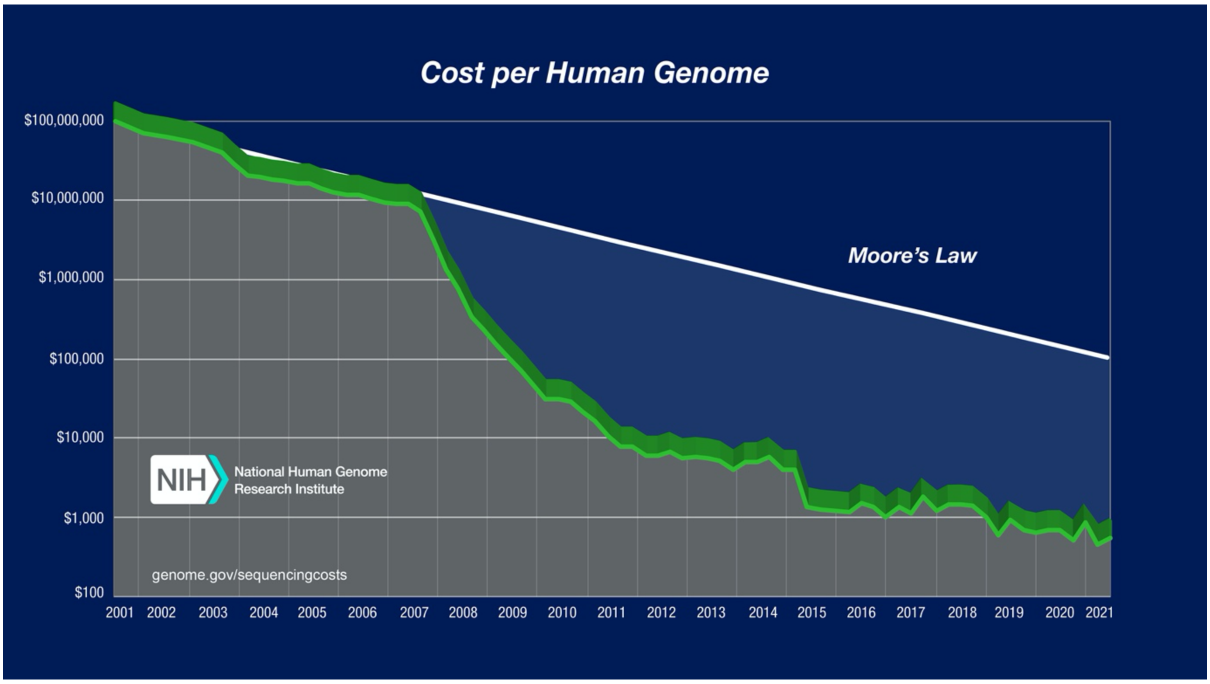


Figure 1.4 – The reduced cost of genome sequencing. From National Human Genome research institute (cited 2022).
The reduction in the cost of sequencing a human genome has been so dramatic since 2008, it had outpaced developments in computing, here represented by Moore's law.

The advantages of Illumina short read SBS sequencing are numerous, small and medium sized genomes can be sequenced in hours, allowing the study of an organisms genetic content to be performed efficiently. Low sample input concentrations can be accommodated by using amplification if necessary, plus the base calling accuracy of Illumina sequencing is strong, with large amounts of data having a Phred score of Q30 (the probability of an incorrect base call being 1 in 1000 times, or 99.9% accuracy, in comparison sanger sequencing Phred scores are on average 99.99%).

Illumina sequencing relies on PCR amplification and for low input sample types over-amplification can be a real concern. Typically, a sequencing run should have (as much as possible) uniform coverage of your genomic region(s) of interest. PCR bias can in some situations lead to under-representation, or complete drop out, of a genomic locus (Aird et al 2011).

Another common problem is the amount of processing steps in a library preparation protocol. Generating libraries to be run on Illumina platforms can be a complicated and lengthy process. DNA extraction and purification is followed by fragmentation to the desired length, end repair and adapter ligation leading to amplification of the library (with potential adding of indexes). The final libraries are then quality controlled (QC'd), diluted and pooled, quantitative PCR is performed to ensure adapter ligation has worked, after which the final library must be at a specific concentration when loaded onto the sequencer. So, if possible, longer reads would be preferable to work with, as would generating these reads with minimal, or preferably without, PCR amplification, using only the native DNA.

1.1.3 Long Read Sequencing

Two such sequencing methods are currently available, technologies that facilitate long read sequencing. Pacific Biosciences (PacBio) relies on a Single Molecule Real-Time (SMRT) sequencer that can generates read lengths in excess of 10,000 bases. A circular SMRTbell library is created using adapters bound to each end of an extracted DNA molecule, these libraries are loaded onto a SMRT flow cell which contains thousands of wells called Zero Mode Waveguides (ZMWs). Each library is immobilised in the ZMWs. A polymerase starts to incorporate randomly labelled nucleotides, each addition emitting a light that is recorded. In this instance each nucleotide incorporation can be measured in real time. Two different sequencing modes can be used depending on what suits a researcher more, Circular Consensus sequencing (CCS) will produce reads with a high degree of accuracy (HiFi read >99%). However, if your requirement is the longest reads possible then the Continuous Long Read sequencing (CLR) mode can be used, the data quality is reduced however the pay-off is the generation of long reads, a large proportion over 50Kb in length.

In 2014 Oxford Nanopore Technologies (ONT) released their first portable sequencing device, the MinION. The technology relies upon flow cells that contain a series of minute holes set within an array, the array being fixed within an electro resistant membrane (in grey, Figure 1.5a). The holes are referred to as nanopores. Each nanopore has its own electrode attached to a channel and sensor chip, and they monitor the electrical current that flows through the nanopore in real time.

When passing a DNA strand through a nanopore, the electrical current will be altered (disrupted) by each base that migrates through, this current disruption is called a Squiggle (Figure 1.5b) and will be different for each of the four bases. A continuous squiggle pattern, generated by multiple bases passing through the nanopore is decoded using basecalling algorithms to assemble a DNA sequence.

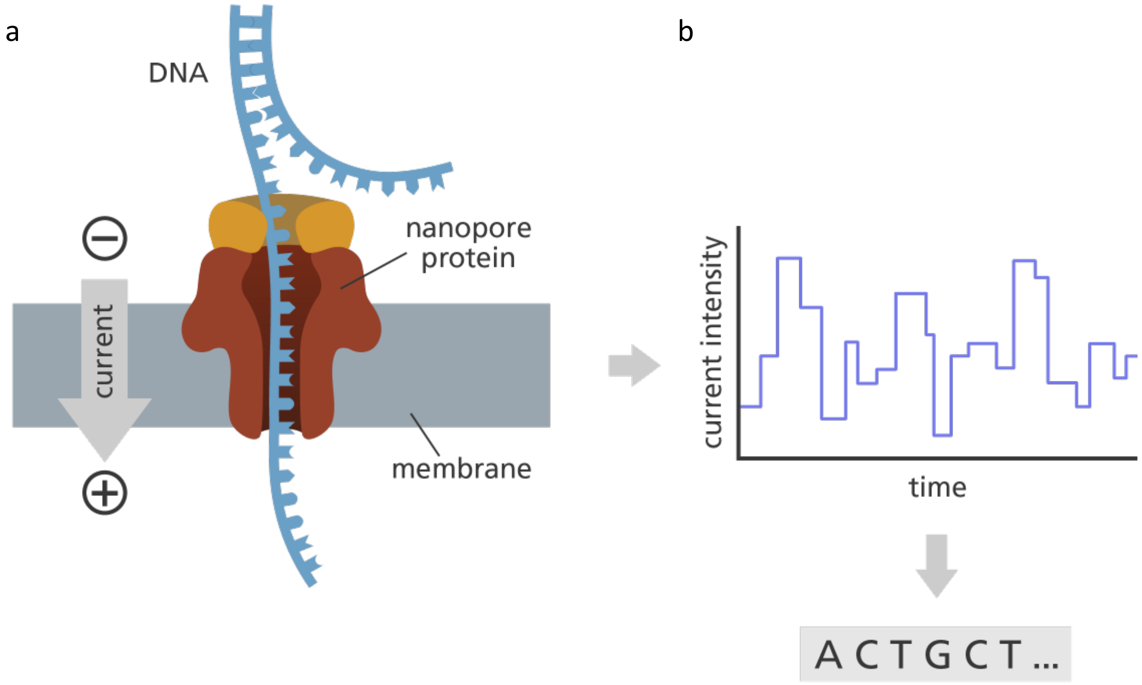


Figure 1.5 - Nanopore Sequencing - Image from Research Genome Limited (2022).
a) The Nanopore sits within electro resistant membrane, an electrical charge is applied to the membrane to facilitate movement of DNA through the nanopore. A sequencing tether is attached to the DNA strand during library prep to ensure the DNA molecule moves through the nanopore at a speed that allows squiggle detection. b) Changes in electrical current as DNA molecule travel through the nanopore are detected and translated using base calling algorithms to produce sequence.

Oxford Nanopore sequencing uses a software program called MinKNOW to both set up and monitor a sequencing run in real time. The raw squiggle data, generated during a sequencing run is stored in HDF5 files (Hierarchical Data Format version 5) developed by ONT and named fast5 files, these fast5 files can by basecalled to fastq files using a software tool called Guppy. Fastq files are text files that contain basecalled DNA sequence and its associated quality scores. Basecalling can be achieved either during, or after a sequencing run has finished. Two of the larger Platform instruments developed by ONT contain on-board basecalling capabilities, that will automatically generate fastq files from the fast5. A smaller portable device called the MK1c has recently been released, composing of a MinION with fully integrated compute and in-built monitor, allowing rapid sequencing to be performed in the field.

Read lengths generated using some nanopore library preparation kits often run close to a Megabase (Mb) (Jain et al, 2018) and in some cases over. A variety of library prep kits can work with unfragmented, native DNA input. ONT also have library prep kits that don't include any PCR steps and are rapid to use. However, there are limitations which include most significantly a poor read accuracy when compared to both Sanger and Illumina sequencing. Read accuracies have recently improved (Franka et al, 2018), from around 60% when first released in 2014, to 85% in 2018, and now using the current basecaller models over 95%. Furthermore, consensus sequences have been generated from homogenous DNA samples by genome assembly, with accuracies of more than 99% (Rang et al, 2018). This is however still problematic and is something that all bioinformaticians need to consider when performing any downstream data analysis. As rapid advances in sequencing technologies have developed, so have the application of these technologies to multiple biological fields.

1.2 Transposons

Transposons, also known as transposable elements (TE) or jumping genes are stretches of DNA sequence that have the ability to move position in the genome. Identified in 1953 by Barbara McClintok, for many years transposons were thought of as junk DNA that had no real function, even though studies identified that a large percentage of an organisms genome is made up of TEs, for example approximately 44% of the human genome (Pray L, 2008). The evidence for the potential importance of TEs came from the growing realization that many transposons were 'highly conserved among distantly related taxonomic groups', implying that they must be of some biological asset to the genome (Pennisi, 2007).

Transposons are typically categorised into two distinct groups, Class I (Retrotransposons) and Class II (DNA Transposons). The two groups differ on how a transposon integrates itself into a genome. Class I transposons use what is termed a 'copy and paste' method involving reverse transcription of an RNA intermediate followed by insertion of its copied cDNA at a new location in the genome (Goodier, 2016). Class II transposons use a 'cut and paste' method with no need for RNA intermediates, they move independently, inserting into, or excising themselves from a genome.

It is important to note, particularly for this project, that if a TE is inserted into a functional gene, it is likely to have a detrimental effect, leading to the deactivation of the gene. If this gene is essential for the survival of the cell, then the cell will die. This system has been utilized by the scientific community for the study of gene essentiality.

1.2.1 The essential genome

Every genome contains a set of functional genes, some of which will be crucial for the cell's survival. If as stated above, a TE is inserted into a functional essential gene, the cell will typically not survive, either that or it will exhibit an extreme loss of biological function that compromises the organism's viability (Hebing et al, 2020). A list of genes critical to the survival of a cell or organism (or living system) can be referred to as that organisms essentialome. Essential genes will partake in functions that vary from species to species, but are likely to include protein interaction networks, conserved biological systems and transcription events. Fundamentally, whatever is necessary for the cell's survival, development and proliferation.

Some genes can be conditionally essential, depending on what external factors the cell is being exposed to. For example, a specific gene may be essential when cells are being grown in a particular media, however a change in growth conditions might see that gene become non-essential (Zhao et al, 2017). Some studies have sub-classified essential genes even further into four groups adding 'redundant essential' and 'absolute essential'. In a situation where a mutation in two genes together could result in cell death, but a mutation in either gene alone does not, these genes would be described as redundant essential. The minimal collection of genes that are necessary for maintaining cellular life under a stress-free environment would be termed absolute essential. (Zhang et al, 2015).

It has become increasingly common over the past 10 years for transposons to be used as genetic tools to identify the essential genes within an organism. To achieve this end involves merging three distinct disciplines. Wet lab skills for transposon mutagenesis, technical skills to generate and run sequencing libraries and finally bioinformatic expertise to interpret the resulting data.

1.2.2 Transposon mediated mutagenesis

Transposons are frequently used as a biological mutagen in bacteria (Hamer et al, 2001), for transposon-mediated mutagenesis, TEs are randomly inserted into numerous positions within a genome and by doing so will disrupt the function of any essential genes they are inserted into. If no genes have been disrupted, cells will continue to grow (called insertion mutants).

1.2.3 High Throughput Insertion Mutagenesis Sequencing

There are multiple methods available for sequencing insertion mutant libraries, the large majority published and verified use Illumina sequencing data. Four of which were established in 2009, TraDIS, HITS, Tn-Seq and IN-Seq.

TraDIS (TRAnsposon Directed Insertion Sequencing) is likely to be the most popular method, first developed 13 years ago (Landgridge et al, 2009) and more recently optimised for high throughput sequencing (Barquist et al, 2016), it relies upon the PCR amplification of transposon containing fragments. Transposon enriched libraries are sequenced on an Illumina platform, and the fastq reads uploaded to a freely available Bio-TraDIS analytical pipeline (https://github.com/sanger-pathogens/Bio-Tradis). Having accessible software tools that are custom designed to perform the analysis for the user being a distinct advantage. However, this method is again reliant on PCR amplification which could potentially introduce bias. Also, the processing steps when setting up the sequencing libraries are numerous. It is worth noting, that to run TraDIS libraries on an Illumina platform – modifications to the standard Illumina run parameters and recipes need to be made. These changes are not supported by Illumina, and therefore it can be difficult to get a standard commercial sequencing facility to agree to modifying their (very expensive) Illumina systems.

The HITS (high-throughput insertion tracking by deep sequencing) method is quite similar to TraDIS, whereby Tn-Seq and IN-Seq are slightly different in approach with their own common similarities (Opijnen et al, 2013). In all instances, PCR is used alongside Illumina sequencing. Each method has advantages and disadvantages over the other, Tn-Seq and IN-Seq use enzymatic fragmentation during library preparation to yield uniform-length shorter reads (Cain et al, 2020) which could possibly remove PCR bias. TraDIS and HITS prefer using larger fragment sizes generated using random mechanical shearing, the longer sequencing libraries will supposedly improve the resulting transposon mapping (Barquist et al, 2016). All four methods use the mariner transposon Tn5, although the TraDIS and HITS methods do offer the added flexibility of changing the transposon tag (Cain et al, 2020). However, none of the above have been adapted for long read data, or allow the opportunity for native DNA input.

1.2.4 PIMMS Seq

The PIMMS (Pragmatic Insertional Mutation Mapping System) is a laboratory protocol and bioinformatic tool that can be used to identify and map transposon insertion events within a population of sequenced insertion mutants in gram positive bacteria. First reported in 2015 (Blanchard et al, 2015) the pipeline uses as input - sequence data generated from DNA libraries generated using inverse PCR.

The laboratory pipeline consist of multiple steps, genomic DNA is extracted from insertion mutants using a method described in Hill and Ligh 1989, after which the DNA is digested and purified using a PCR clean up kit. This DNA is then sheared using mechanical methods (Covaris sonication) producing short fragments with an average length of 3Kb, then purified using SPRI beads. Blunt ends resulting from the shearing are repaired and the DNA is again purified using SPRI beads. The end-repaired

DNA is then self-ligated to generate a re-circularised product which is then amplified using inverse PCR to enrich for sequences that flank the insertion element. A final bead purification step generates PIMMs processed DNA that is ready for sequencing library preparation.

Inverse PCR is specifically used to enrich for regions that contain transposon – chromosome junctions. The sequencing data generated being Illumina (short read) fastqs. These fastqs (paired end or single end reads) are loaded into the PIMMS bioinformatic pipeline (Figure 1.6). Reads that contain the appropriate transposon motif sequence are identified and mapped to a reference genome (mapping). The mapped SAM file is processed generating counts which are represented in output files containing information on genomic coordinates, insertion instances and mapped read scores plus first and last insert positions as a percentile.

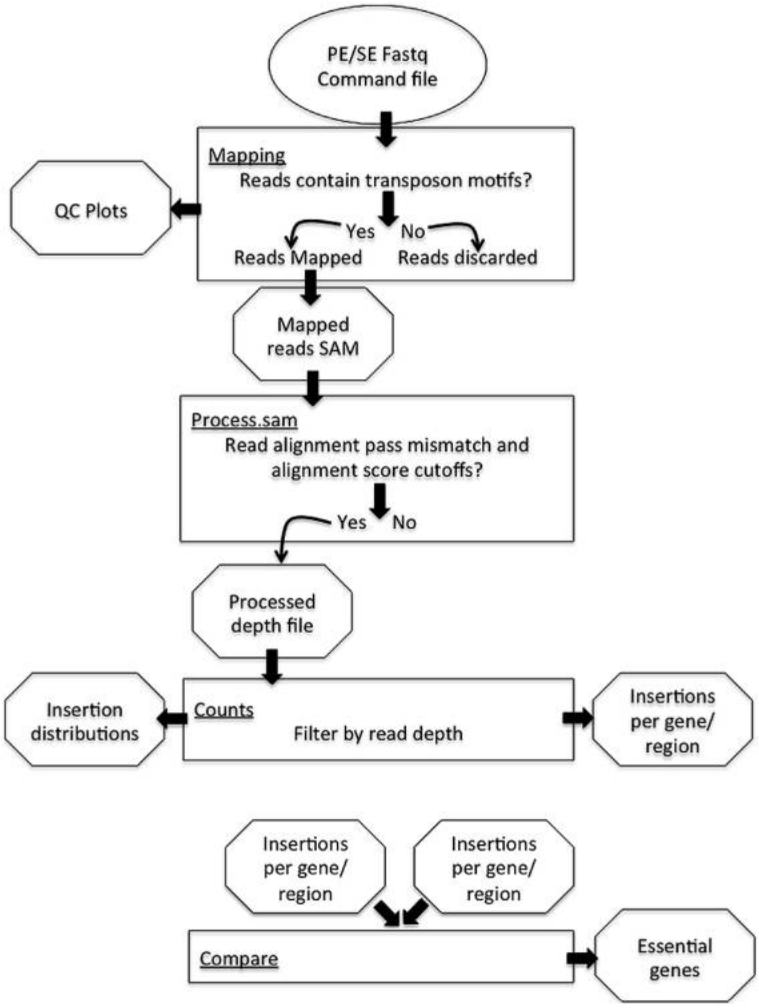


Figure 1.6 - Schematic of PIMMS bioinformatic pipeline (Blanchard et al, 2015).
Raw fastq data is uploaded, undergoing four bioinformatic processes, **mapping** identified reads to a reference fasta file, **processing** the resulting SAM file to identify the coordinates of insertion events, **counting** which reports the insertion outcomes and finally **comparing** the output to identify essential genes.

A PIMMS dashboard (https://pimms-dashboard-uon.azurewebsites.net) has recently been developed so that PIMMS output files can be uploaded and compared, with output summaries, metrics, statistics and publication ready figures being generated. Theoretically any transposon insertion sequencing data (providing the data is in the correct format) can be processed using the PIMMS pipeline (Blanchard et al, 2015), achievable with only small modifications to the existing commands. To ensure confidence in output, the pipeline itself enforces strict parameters when identifying reads that contain transposon motifs, each read must have a Phred score of at least Q30 (99.9% basecalling accuracy). The PIMMS pipeline has already been developed to identify essential and conditionally essential genes in *Streptococcus uberis* (Blanchard et al, 2016) and could potentially be used to achieve this end with other related bacterial species.

The pipeline processing time is fast, and all modules can be run using a standard laptop, the user friendly nature of its bioinformatic element acts as an incentive for researchers who are more at home in a wet lab environment, and don't have any advanced bioinformatic training. For the pipeline to work, it requires three external input files, firstly, raw fastq data files generated using high throughput sequencing of mutant libraries, in the case of PIMMS this has so far been Illumina short read data. Secondly, the fasta sequence of the bacterial species used for transposon mutagenesis, once reads containing transposon motifs have been identified this reference sequence will be used to align the reads to the bacterial genome. Lastly a general format file (GFF) is required. This tab delimited text file will contain a list of genomic features for the reference genome. The GFF is used to generate a report of insertion metrics related to each genomic feature.

1.3 Utilising Technological advances

In contrast to other companies that have developed methods of high throughput sequencing, Oxford Nanopore Technologies (ONT) have a history of continual innovation, searching for unique instances where the generation of long read sequence from native DNA might be of greatest use. The PIMMS pipeline has until now always been run with short read data, generated using inverse PCR. This pipeline has been proven to work well (Blanchard et al, 2016), however a potential opportunity has recently arisen for a new approach.

1.3.1 PIMMS using Nanopore Sequencing

The laboratory pipeline of the PIMMS contains multiple sample processing steps, and it is natural that there will be small, but potentially significant sample losses during each step. Illumina library preparation and sequencing has its own limitations, and not all of the extracted DNA will be sequenced, large and small fragments will be removed during the library preparation, largely for sequencing optimisation purposes, ensuring the final library loading concentration is accurate and measurable. A standard nanopore library preparation can retain fragments of all sizes, the end point library containing both short and long reads, the resulting sequences being more representative of the input DNA. It is also clear that the alignment of a long read to a genomic position will be more confident that the mapping of a short read. Moving forwards, there is the attractive option of combining both short and long reads to generate a hybrid assembly (Brown et al, 2021). In this context we would be merging the advantages of long read sequencing with the high accuracy and greater depth of short read sequencing.

1.3.2 Cas-9 Targeted Sequencing

For the PIMMS it would be desirable to remove as many of the initial sample processing steps as possible, in particular steps that might introduce amplification bias. Such a system would not only allow greater sample retention, it would reduce both time and costs. In 2012 a scientific paper published in Science proposed the use of the CRISPR (**C**lustered **r**egularly **i**nterspaced **s**hort **p**alindromic **r**epeats) Cas-9 system as a possible tool for gene editing (Jinek et al, 2012). The cleaving of site specific genomic regions could be achieved using a custom designed crRNA (or guide RNA) and the presence of a trans-activating crRNA (tracrRNA). These two units would combine with a Cas9 nuclease to introduce a location specific double stranded cut of the DNA.

In 2018 ONT established a protocol that used the CRISPR Cas-9 system to enrich for a pre-designated region of a genome. The goal was clear, targeted, amplification free DNA sequencing. The protocol incorporates the standard CRISPR Cas-9 system, with small modifications from what was described in 2012. During a standard ONT library prep, the ends of DNA fragments are prepared for dA tailing and adapter ligation. For a Cas-9 targeted approach all DNA ends are dephosphorylated (removal of 5’ phosphates from DNA). After which the DNA is incubated with an RNP (ribonuclease protein), this is a combination of crRNA, tracrRNA and cas9 nuclease (Figure 1.7).

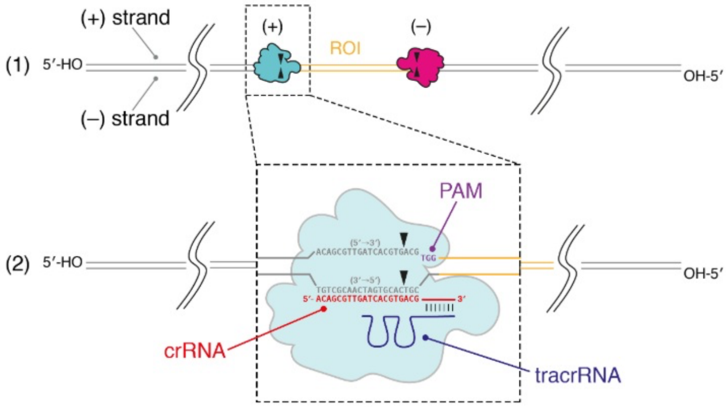


Figure 1.7 - Representation of RNP complex interaction with ROI (Oxford Nanopore Technologies, 2018)
The crRNA (in red) is a small RNA primer of around 20 bases, the sequence of which will be designed by the user, it will act as a ‘guide’ for the RNP – identifying specific locations at which the RNP complex will facilitate a double stranded cleave. The crRNA will bind to the complementary strand of the target region and will also contain a protospacer adjacent motif (PAM, in purple)), this is a 3 base sequence leading off from the 3’ end. Approximately 3 bases downstream of the PAM, the Cas-9 nuclease will enforce a double stranded break.

Moving forward, only the cleaved DNA (region of interest (ROI) in orange, Figure 1.8) will be available for dA tailing and adapter ligation, ready for sequencing. By this design we should see a significant enrichment of the desired ROI.

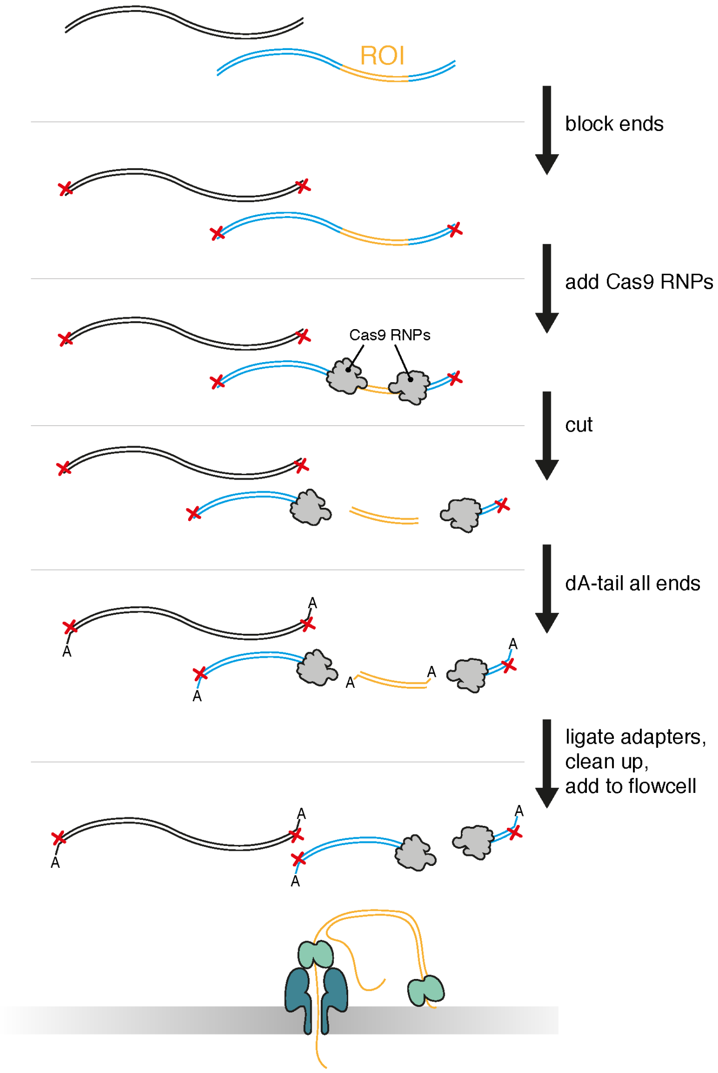


Figure 1.8 - Standard workflow for Cas-9 targeted sequencing using ONT (Oxford Nanopore Technologies, 2018).
The ends of all DNA strands are blocked using dephosphorylation, a Cas-9 RNP is added and cuts at the region of interest exposing the cleaved ends for dA tailing, adapter ligation and sequencing.

ONT offer three different methods of Cas-9 Targeted sequencing, the excision approach is used when a user knows the sequence they are trying to enrich for. In this case, crRNA's can be designed to excise the ROI, by cleaving at a known sequence at each end. Secondly, the single cut and run is when only a portion of the sequence is known, crRNA's can be designed to cut at a specific location, after which you sequence out into the unknown ROI. Both excision and single cut run methods rely on the ROI being <20Kb in length, anything above this and a Tiling approach, using multiple crRNA probes in 5-10 kb overlapping chunks, is recommended.

1.4 Aims of project

1.4.1 Assess the application of Nanopore long read sequencing to identify insertions in a mutagenised bacterial library

An effective transposon mutagenesis system has been developed, using *Streptococcus agalactiae* (*S. agalactiae*) This system uses a thermosensitive plasmid (pGh9+ host) to introduce the insertion sequence IS*S1* (Maguin et al, 2002) into a strain of streptococcus. This plasmid has previously been identified as a good tool for insertional mutagenesis (Blanchard et al, 2016) in gram-positive bacteria, due to its ability to integrate into the genome by a mechanism of replicative transposition (Thibessard et al, 2002). This plasmid-insertion sequence fusion (pGh9:IS*S1*) acts as a delivery vector, transporting the IS*S1* insertion sequence into the cell, where high frequency random insertion is achieved (estimated to be around 1% (Maguin et al, 2002). It is reported that this mutagenesis system will result in around 1 insertion per cell. Where insertion has occurred (replicative transposition) the insertion site will contain a duplicated IS*S1* sequence flanking the host plasmid.

DNA will be extracted from mutant libraries (generated using the transposon mutagenesis system) grown in BHI (Brain Heart Infusion).

1.4.2 Objectives

1. Generate sequencing data from PIMMS processed DNA using long read technologies

Source DNA will have been generated using the existing PIMMS wet laboratory pipeline, DNA libraries will be generated and sequenced using nanopore library preparation techniques and sequencing. The successful detection of transposon insertion will be achieved using the PIMMS and a custom designed bioinformatic pipeline.

2. Generate Cas-9 enriched DNA, sequenced using long read technologies

A crRNA guide will be designed to cut the DNA within the pGh9:IS*S1* insert during library preparation, the goal being the enrichment of sequences that read from the insert out into chromosomal DNA. This will be a single cut and run approach, and should hopefully reduce any less useful data and the sequencing of any off-target reads. Input will be native DNA, saving processing time and removing any potential amplification bias.

3. Compare the long read datasets to Illumina data

PIMMS output files will allow comparison of insertion events for three sequencing methods (Illumina, Nanopore and Cas-9 Targeted), and the potential identification of essential and conditionally essential genes. Manual alignments of the Cas-9 data sets will be run separately to the PIMMS seq pipeline, to be presented alongside the PIMMS pipeline results. These alignments will be used to confirm insert-chromosome events.

2. Methods

2.1 Samples

DNA samples used for this project were generated externally within the University of Nottingham's School of Veterinary Sciences. Transposon mutagenesis was used to create *Streptococcus algalactiae* bacterial culture libraries containing 10,000 different insertional mutations (approximately 1 per cell). Method for insertional mutation closely followed that used by Maguin et al 1996. Insertion mutants were grown in a standard bacterial media (Brain Heart Infusion – BHI) for six hours, sub-cultured onto agar, harvested and the DNA extracted. Three datasets were generated using two different library preparation methods. For all samples, sequence data was generated using Nanopore technology.

Sample ID	**ds1093_1**	**ds1093_3**	**ds1093_3_Rpt**
Sample source	BHI DNA	BHI DNA	BHI DNA
Library Prep method	Long Read PIMMS	Cas-9 Targeted	Cas-9 Targeted

Table 2.1 - Summary of sample ID and library preparation method.
DNA sequencing libraries will be generated for ds1093_1 using the standard PIMMS approach (inverse PCR), ds1093_3 and ds1093_3_Rpt will be generated using a Cas-9 Targeted method.

For a Cas-9 Targeted sequencing approach, DNA extracted from insertion mutations grown in BHI media required no further sample processing prior to library preparation and sequencing. For the Long Read PIMMS approach, extracted DNA required further sample processing.

2.2 Long Read PIMMS Laboratory Method

One extracted DNA sample (ds1093_1) was processed using a pre-existing wet laboratory pipeline, this method (Figure 2.1) is identical to that described in Blanchard et al 2016.

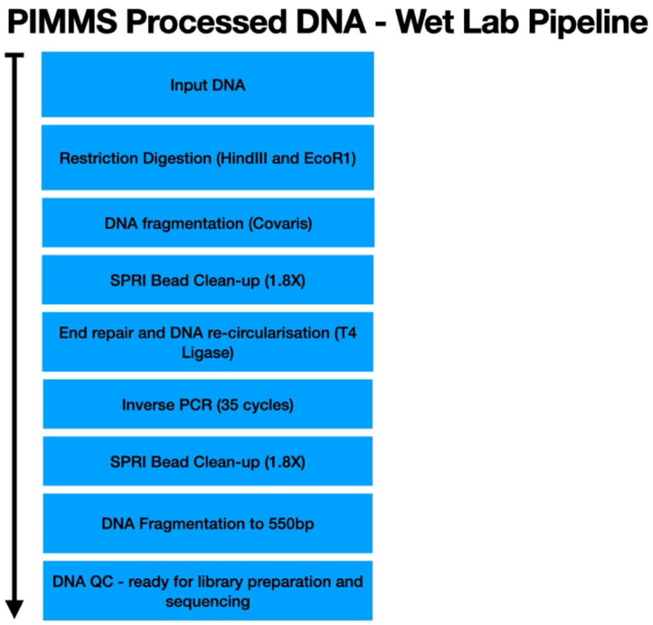

PIMMS Processed DNA - Wet Lab Pipeline
Input DNA
Restriction Digestion (HindIII and EcoR1)
DNA fragmentation (Covaris)
SPRI Bead Clean-up (1.8X)
End repair and DNA re-circularisation (T4 Ligase)
Inverse PCR (35 cycles)
SPRI Bead Clean-up (1.8X)
DNA Fragmentation to 550bp
DNA QC - ready for library preparation and sequencing

Figure 2.1 - Laboratory workflow to generate PIMMS processed DNA.

DNA is fragmented and purified using SPRI Beads, blunt ends are repaired followed by re-circularisation of the DNA ready for inverse PCR. Finally, DNA is purified and fragmented again, in preparation for library generation and sequencing.

2.2.1 Sample QC

DNA sample was stored at 4°C. DNA concentration generated using the Qubit v4 Fluorometer (Thermo Fisher Scientific) and the Qubit™ dsDNA BR Assay Kit (Thermo Fisher Scientific; Q32850). Optimal DNA fragmentation was confirmed using the Agilent 4200 TapeStation (Agilent Technologies) and D1000 DNA ScreenTape Assay (Agilent Technologies; 5067-5582 and 5067-5583).

2.2.2 Library Preparation and sequencing

Sequencing libraries were generated using Oxford Nanopore Technologies SQK-LSK110 library preparation kit (vGDE_9108_v110_revR_22May2022), for Ligation sequencing of genomic DNA. This protocol has been altered and optimised to use half reactions.

2.2.3 DNA Repair and End-Prep

As input, 500ng of DNA was made up to a volume of 24µl in molecular biology grade water (Sigma: W4502, Cas – 7732-18-5) in a 0.2ml thin walled PCR tube (Alpha labs: LW2570G). Added to this tube was 1.75µl of NEBNext FFPE DNA repair buffer, 1µl of NEBNext FFPE DNA repair mix (NEBNext: M6630), 1.75µl of NEBNext Ultra II End-prep reaction buffer and 1.5µl NEBNext Ultra II End-prep reaction mix (NEBNext: E7546). All components were mixed gently eight times using a 200µl wide bore tip (Fisher: 10089010). The 0.2ml thin walled PCR tube (Alpha labs: LW2570G) was then capped and placed into a DEEP well C1000 Touch Thermal Cycler (Biorad: 1851197) and incubated at 20°C for 30 minutes, followed by 65°C for 30 minutes.

2.2.4 First AMpure Clean-up step

The reaction mixture was pulse centrifuged and the contents transferred to a 1.5ml DNA low bind tube (Eppendorf: 0030108051), to which 30µl (1X) of AMpure XP beads (Beckman: A63881 – brought to room temperature for 30 minutes, and thoroughly resuspended) were added and mixed by flicking the tube. The tube was then incubated for 5 minutes at room temperature (RT) on a Hula mixer (Thermo Fisher Scientific: 15920D – rotation speed set to 10). After a pulse centrifuge the tube was placed onto a bead separation magnet (Invitrogen) for approximately 3 minutes. The supernatant was removed using a 200µl tip without disturbing the beads. 200µl of freshly made 70% Ethanol (Honeywell: 02875) was added to the tube whilst still on the magnet for 30 seconds, and then removed as before without disturbing the beads, this Ethanol wash step was performed twice. The tube was then pulse centrifuged and returned to the magnet, any remaining Ethanol was removed using a 10µl tip. Still on the magnet, the tube was left to air dry for approximately 3 minutes (at all times ensuring that the beads did not dry out or begin to crack). The tube was removed for the magnet and the beads resuspended in 31µl of molecular biology grade water (Sigma: W4502, Cas – 7732-18-5) using a low retention 200µl tip (Greiner Bio One: 775363). Beads where incubated at RT for 2 minutes and then returned to the magnet. Once the eluate appeared clear (2 minutes) 31µl was transferred into a new 1.5ml DNA low bind tube (without disturbing the beads). A qubit measurement was taken using the Qubit v4 Fluorometer (Thermo Fisher Scientific) and the Qubit™ dsDNA HS Assay Kit (Thermo Fisher Scientific; Q32854). This was to ensure sufficient recovery of DNA so we could proceed to Adapter ligation.

2.2.5 Adapter ligation

A 50µl adapter ligation reaction was assembled using 30µl of the End-Prepped DNA (from above step), 12.5µl of Ligation buffer (LNB – mixed by pipetting due to extreme viscosity), 5µl of NEBNext Quick T4 DNA Ligase (NEBNext: E6056) and 2.5µl of Adapter mix-F (AMX-F). Both LNB and AMX-F are components of the SQK-LSK110 Genomic DNA by ligation library preparation kit. Tube components were mixed 10 times by pipetting, pulse centrifuged and incubated at RT for 1 hour.

2.2.6 Second AMpure Clean-up Step

Adapter ligation reaction mixture was pulse centrifuged. 25µl (0.5X) of resuspended AMpure beads (Beckman: A63881) was added to the tube and mixed using gentle flicking. The tube was then incubated at RT for 5 minutes on a HulaMixer (Thermo fisher Scientific: 15920D – rotation speed set to 10). Tube contents were pulse centrifuged and placed onto a bead separation magnet (Invitrogen) for 3 minutes. Supernatant was discarded without disturbing the beads, after which the tube was removed from the magnet, and the beads resuspending in 125µl of Short Fragment Buffer (SFP – component of SQK-LSK110 Genomic DNA by ligation library prep kit) to retain DNA fragments of all sizes. Resuspension was achieved by flicking the tube, once fully resuspended the tube was pulse centrifuged and returned to the bead separation magnet. After 3 minutes the supernatant was removed; and the SFB step repeated. After the second SFP step the tube was pulse centrifuged and returned to the magnet, any remaining supernatant was eluted and disposed of using a 10µl tip. Tube was allowed to air dry for 1 minute, after which beads were resuspended in 15µl of Elution Buffer (EB – component of SQK-LSK110 Genomic DNA by ligation library prep kit) and incubated at RT for 10 minutes. The tube was returned to the magnet for 2 minutes, and the supernatant eluted into a fresh 1.5ml DNA low bind tube (Eppendorf: 0030108051).

2.2.7 Library and Flow Cell QC

End point sequencing libraries were quantified using the Qubit v4 Fluorometer (Thermo Fisher Scientific) and the Qubit™ dsDNA HS Assay Kit (Thermo Fisher Scientific; Q32854).
For both samples 100ng of library was made up to a volume of 12µl with Elution Buffer (EB) and placed on ice. A fresh MinION flow cell (Oxford Nanopore Technologies: R9.4.1) was removed from the fridge and allowed to reach RT (30 minutes). The flow cell was loaded onto the GridION platform (Oxford Nanopore Technologies: MK1) and QC was performed using the GridION's MinKNOW software to ascertain how many pores were available for sequencing. For sequencing run a note of Flow Cell ID, position on the GridION (platform has five loading positions 1-5) and pores available was recorded.

2.2.8 Library Loading

A volume of 30ul of FLT (Flush Tether - pipette mixed due to viscosity) was added to a flush buffer vial (FB – mixed by vortex), the contents then mixed using a p1000 pipette set to 800µl. The priming port of the MinION flowcell was then opened and a small amount of buffer extracted slowly using a p1000 pipette set to 200µl. 800µl of FLT/FB mixture was slowly pipetted into the priming port, ensuring no bubbles were added to the flow cell. The flow cell was then left for 5 minutes. During this time a loading mixture was made up in a 1.5ml DNA low bind tube (Eppendorf: 0030108051) to include 37.5µl of Sequencing Buffer (SBII), 25.5µl of loading solution (LS) and the 12µl of DNA library. Loading mixture was vortexed and pulse centrifuged. Following 5 minute wait step, the SpotON port of the flow cell was lifted, and 200µl of FLT/FB mixture again pipetted into the priming port using a P1000 pipette. Loading mixture was mixed using a 200µl pipette and all 75µl added (using a P200 pipette) to the SpotON port using a droplet pipetting technique. SpotON and priming ports were then closed.

2.2.9 Sequencing and Data Collection

MinKNOW software (v 21.11.7) was used to set up the sequencing. Each run was given a sequencing run name and sequencing ID, the run time was set to 72 hours and the software was instructed to collect both fast5 and fastq files. Quality threshold was set to Q9, fast5 files containing reads that achieved this threshold would be deposited into a fast5 pass folder, all other reads would be placed into a fast5 fail folder. Fastq files were then generated using the onboard basecaller Guppy (v 5.1.13). Leading to a fastq pass folder, and a fastq fail folder. Once run had completed, all data files were transferred to two separate servers for downstream bioinformatic processing.

2.3 Cas-9 Targeted Enrichment Laboratory Method

2.3.1 Sample QC

DNA sample stored at 4°C. DNA concentration generated using the Qubit v4 Fluorometer (Thermo Fisher Scientific) and the Qubit™ dsDNA BR Assay Kit (Thermo Fisher Scientific; Q32850).
For ds1093_3, DNA integrity (DIN score) ascertained using the Agilent 4200 TapeStation (Agilent Technologies) and the Genomic DNA ScreenTape Assay (Agilent Technologies; 5067-5366 and 5067-5365).

2.3.2 Library Preparation and Sequencing

Enriched sequencing libraries were generated using Oxford Nanopores Technologies SQK-CS9109 (Version: CAS_9106_v109_revE_16Sep2020) Cas-mediated PCR-free enrichment library preparation kit for sequencing.

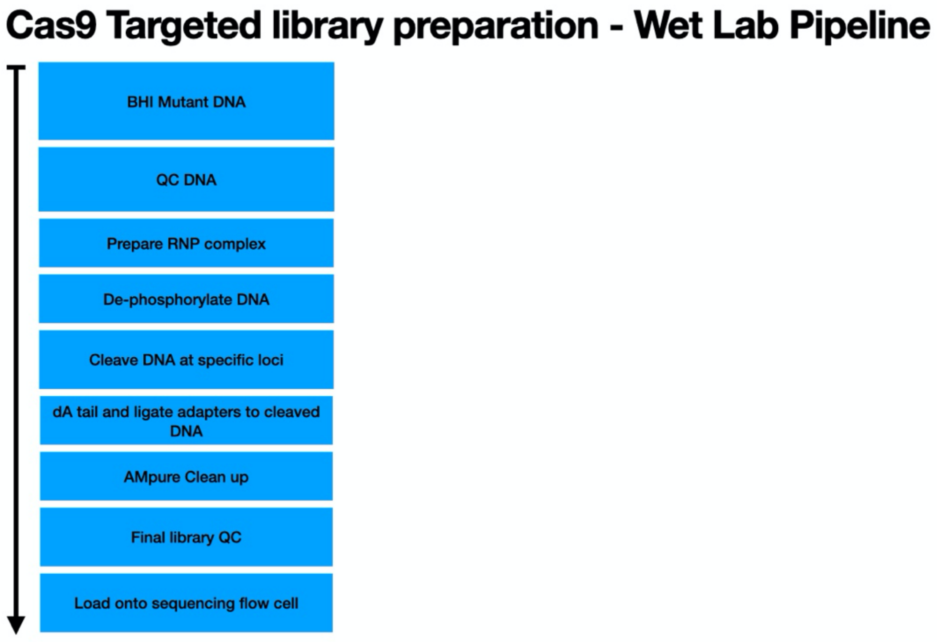

Cas9 Targeted library preparation - Wet Lab Pipeline
BHI Mutant DNA
QC DNA
Prepare RNP complex
De-phosphorylate DNA
Cleave DNA at specific loci
dA tail and ligate adapters to cleaved DNA
AMpure Clean up
Final library QC
Load onto sequencing flow cell

Figure 2.2 – Laboratory workflow of a Cas-9 Targeted Library Preparation.

Input DNA is quality controlled (QC) after which the RNP (ribonucleoprotein) is assembled. All DNA strands are dephosphorylated then cleaved at specific locations designated by the crRNA (within the RNP). Adapter sequences are ligated to the cleaved ends and DNA is purified using AMpure beads. Final library is quantified using a Qubit fluorometer before loading onto a flow cell for sequencing.

2.3.3 crRNA design

The crRNA (guide RNA) sequence was designed using IDT's (Integrated DNA Technologies) Custom Alt-R CRISPR-Cas9 software tool.

https://eu.idtdna.com/site/order/designtool/index/CRISPR_CUSTOM

This crRNA had already been designed and used in a previous Cas-9 experiment; in this instance the enrichment had worked well. The crRNA aliquots were still safety stored in a -80°C freezer on-site which allowed ease of access.

Design ID	Gene Symbol	Position	Strand	Sequence	PAM	On-Target Score	Off-Target Score
CD.Cas9.XPLP7325.AA		298	+	TTTGATTGGAGTTTTTTAAA	TGG	78	N/A

Table 2.2 – crRNA sequence used for both Cas-9 Targeted library preparations.
This crRNA (guide RNA) was designed using the Custom Alt-R CRISPR-Cas9 software tool from IDT. It was used for both Cas-9 Targeted runs to generate the RNP required for DNA cleaving.

2.3.4 Preparation of Ribonucleoprotein (RNP) Complex

A crRNA (IDT: CD.Cas9.XPLP7325.AA – 100µM) 5ul aliquot was removed from -80°C storage and thawed on ice, tracrRNA (IDT; 1072532 - 5nmol) was thawed, made up to 100µM in TE buffer (pH 7.5) and kept on ice. Cas9 nuclease v3 (IDT; 1081058) was also placed on ice. 8µl of Nuclease Free duplex buffer (IDT: 11-01-03-01), 1µl of crRNA and 1µl of tracrRNA were pipetted into a 0.2ml thin walled PCR tube (Alpha labs: LW2570G), tube contents were mixed by pipetting and pulse centrifuged. PCR tube was then placed into a deep well C1000 Touch Thermal Cycler (Biorad: 1851197) and incubated at 95°C for 5 minutes. After which tube contents were cooled to room temperature (RT) and pulse centrifuged.

To a 1.5ml DNA low bind tube (Eppendorf: 0030108051) the following components were added to form the RNP complex, the 10µl reaction from previous step, 10µl of Reaction Buffer (RB from SQK-CS9109 library preparation kit – thawed to RT, vortexed and placed on ice) 0.8µl of Cas9 nuclease v3 (IDT; 1081058) and 79.2µl of molecular biology grade water (Sigma: W4502, Cas – 7732-18-5). Tube contents were mixed by gentle flicking, pulse centrifuged and incubated at RT for 30 minutes.

2.3.5 Dephosphorylate Input DNA

As input, the maximum possible amount of DNA was used, input for ds1093_3 was 1.2µg. Genomic DNA (24µl) was added to 3µl of Reaction Buffer (RB from SQK-CS9109 library preparation kit – thawed to RT, vortexed and placed on ice) in a 0.2ml thin walled PCR tube (Alpha labs: LW2570G). Contents were mixed by flicking the tube gently followed by a pulse centrifuge. Added to this tube was 3µl of Phosphatase (PHOS - from SQK-CS9109 library preparation kit – thawed to RT, mixed using pipette) Contents were mixed gently by flicking the tube, pulse centrifuged and incubated in a deep well C1000 Touch Thermal Cycler (Biorad: 1851197) for 10 minutes at 37°C followed by 2 minutes at 80°C and holding at 20°C.

2.3.6 Cleave and dA Tail target DNA

To the 30µl of de-phosphorylated DNA the following components were added, 10µl of the Cas9 RNPs (from RNP complex step), 1µl of dATP (from SQK-CS9109 library preparation kit – thawed to RT, vortexed, pulse centrifuged and placed on ice) and 1µl of Taq Polymerase (TAQ – from SQK-CS9109 library preparation kit – pulse centrifuged and placed on ice). Tube contents were mixed by inversion and pulse centrifuged. PCR tube placed into DEEP well C1000 Touch Thermal Cycler (Biorad: 1851197) for 15 minutes at 37°C, 72°C for 5 minutes and held at 4°C (15 minutes was the recommended default cleavage time for the Cas-9 enzyme).

2.3.7 Adapter ligation

The 50µl contents of the above cleavage and dA tail reaction were placed into a 1.5ml DNA low bind tube (Eppendorf: 0030108051). A Ligation reaction mix was created using 20µl of ligation buffer (LNB – mixed by pipetting due to extreme viscosity), 3ul of molecular biology grade water (Sigma: W4502, Cas – 7732-18-5), 10µl of NEBNext Quick T4 DNA Ligase (NEBNext: E6056) and 5µl of Adapter mix (AMX). All components were added to a separate 1.5ml DNA low bind tube (Eppendorf: 0030108051). Contents were mixed by pipetting up and down 10 times. (LNB and AMX are components of the SQK-CS9109 library preparation kit). The ligation reaction mixture was added to cleavage and dA tail reaction in two deliveries, first 20µl (mixed by flicking tube) followed by 18µl, again mixed by flicking the tube. The tube was then pulse centrifuged and incubated at RT for 20 minutes.

2.3.8 AMpure Clean-up step

80µl of SPRI dilution buffer (SDB - component of the SQK-CS9109 library preparation kit) was added to the adapter ligation reaction, contents were mixed by gentle flicking of the tube.
48µl (0.3X) of resuspended AMpure beads (Beckman; A63882 – thawed to RT for 30 minutes) was added to the reaction tube and mixed by gentle inversion. Tube contents were incubated at RT for 10 minutes after which the tube was placed onto a bead separation magnet (Invitrogen). Once the beads had separated the supernatant was removed and discarded. Beads were then resuspended in 250µl of Long Fragment Buffer (LFB - component of the SQK-CS9109 library preparation kit – used to specifically retain larger DNA fragments) by gentle flicking of the tube. After a pulse centrifuge the tube was place back onto the magnet for 3 minutes, and the supernatant removed. This LFB wash step was then repeated. The tube was pulse centrifuged, placed back onto the magnet and all remaining supernatant removed using a 10µl pipette tip. The beads were then left for 1 minute to air dry before being resuspended in 13µl of Elution Buffer (EB - component of the SQK-CS9109 library preparation kit) and incubated at RT for 30 minutes. Beads were then pelleted using the bead separation magnet for 2 minutes, the eluate removed and placed into a new 1.5ml DNA low bind tube (Eppendorf: 0030108051).

2.3.9 Library and Flow Cell QC

End point sequencing libraries were quantified using the Qubit v4 Fluorometer (Thermo Fisher Scientific) and the Qubit™ dsDNA HS Assay Kit (Thermo Fisher Scientific; Q32854). The maximum amount of library possible was loaded onto the flow cell. This was 451ng for ds1093_3. A fresh MinION flow cell (Oxford Nanopore Technologies: R9.4.1) was removed from the fridge and allowed to reach RT (30 minutes). The flow cell was loaded onto the GridION platform (Oxford Nanopore Technologies: MK1) and QC was performed using the GridION's MinKNOW software (v 21.11.7) to ascertain how many pores were available for sequencing. For each run a note of Flow Cell ID, position on the GridION (platform has five loading positions 1-5) and available pores was retained. Library loading, sequencing and data collection steps identical to that for Long Read PIMMS method.

2.3.10 Repeat Sequencing Run for ds1093_3

The library preparation and sequencing for sample ds1093_3 was repeated (for reasoning please see discussion). Sample processing steps where identical to that above with two modifications to the existing protocol:

1. A 1X AMpure clean-up was performed on the input DNA before library preparation.
2. Adapter ligation time was doubled to 40 minutes.

The amount of end point library loaded onto the sequencing flow cell was 712.8ng.

2.4 Initial Data Analysis

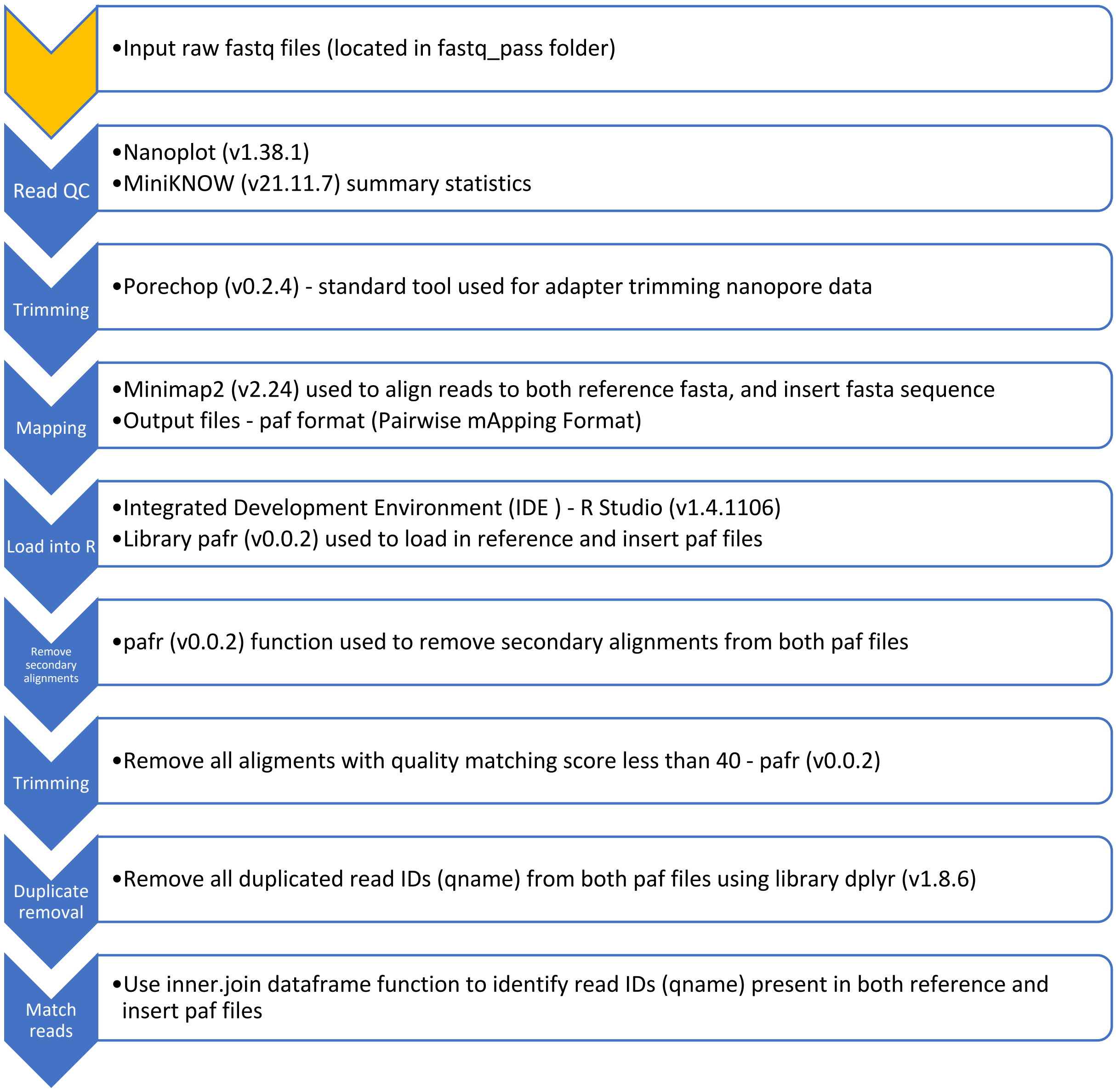
2.4.1 Read Quality Control

Sequencing data generation was monitored in real time using the GridION's MinKNOW software (v 21.11.7) to ensure no problems were encountered. Statistical summary reports were generated for each data set using NanoPlot (v1.38.1). Run reports (generated by MinKNOW) also used to assess data metrics.

2.4.2 Manual Data Interrogation

A bioinformatic pipeline (Figure 2.3) was used for both Cas9 data sets (ds1093_3, and ds1093_3_Rpt). In each case fastq files were concatenated to allow a more streamlined process. After which the software tool Porechop (v0.2.4) was used to trim all adapter sequences from fastq data sets. Each data set was aligned separately to both the insert sequence (Pgh9ISS1.fasta) and the reference (UK15.fasta – *S. agalactiae*). Both FASTA sequences were supplied by colleagues at the school of Veterinary Sciences. Software tool used for alignment was Minimap2 (v 2.24).

Alignment output from Minimap2 was generated in paf (**P**airwise m**A**pping **F**ormat) files. Output files loaded into R (R Studio v1.4.1106). Library pafr (v0.0.2) was used to remove secondary alignments and poor alignments with mapping quality scores less than 40. Library dplyr (v1.8.6) was then used to remove duplicated read ID's from each paf file.



- Input raw fastq files (located in fastq_pass folder)

Read QC

- Nanoplot (v1.38.1)
- MiniKNOW (v21.11.7) summary statistics

Trimming

- Porechop (v0.2.4) - standard tool used for adapter trimming nanopore data

Mapping

- Minimap2 (v2.24) used to align reads to both reference fasta, and insert fasta sequence
- Output files - paf format (Pairwise mApping Format)

Load into R

- Integrated Development Environment (IDE) - R Studio (v1.4.1106)
- Library pafr (v0.0.2) used to load in reference and insert paf files

Remove secondary alignments

- pafr (v0.0.2) function used to remove secondary alignments from both paf files

Trimming

- Remove all aligments with quality matching score less than 40 - pafr (v0.0.2)

Duplicate removal

- Remove all duplicated read IDs (qname) from both paf files using library dplyr (v1.8.6)

Match reads

- Use inner.join dataframe function to identify read IDs (qname) present in both reference and insert paf files

Figure 2.3 - Manual bioinformatic pipeline
Pipeline detailing the data manipulation from input raw fastq data (fastq_pass file) to the identification of reads that align to both reference and insert sequences. Example Rmarkdown file of code and output in Appendix 2.

Sequencing reads that matched to both insert and reference were identified used the Inner.join dataframe function. Summary statistics and Venn Diagrams were then generated for each manual alignment result. Confirmation that reads contained sequence reading from the insert into the chromosome was achieved by blast searching random identified reads against both Streptococcus, and the insert fasta sequence (NIH Blast - https://blast.ncbi.nlm.nih.gov/Blast.cgi).

2.5 PIMMS

PIMMS2 (Pragmatic Insertion Mutant Mapping System V2) was downloaded from the pimms2 github site (https://github.com/Streptococcal-Research-Group/PIMMS2). To create the required conda environment (for Macbook pro) all recommended install steps were undertaken. This included, ensuring python v3.6 was installed and actioning a requirements file in the command terminal, that ensured all necessary packages (Table 2.4) were downloaded.

Package	Version	Function
Pandas	0.25.1	Open source data analysis tool
ConfigArgParse	1	Handling environment variables and config file values
Fuzzysearch	0.7.0	Uses search algorithms to find strings that match patterns
gffpandas	1.2.0	Python library, used to work with annotation data
Pandasql	0.7.3	Facilitates the query of pandas DataFrames using SQL syntax
Pysam	0.19.0	Python module used for reading and manipulating files in BAM format
Urllib3	1.26.5	HTTP client for Python
XlsxWriter	1.2.1	Python module used to write text, numbers, formulas and hyperlinks to multiple worksheets
BWA	0.7.17	Short Read aligner
minimap2	2.24	Long Read aligner

Table 2.3 – Requirements file for PIMMS
A summary of the software packages that were downloaded to ensure error free running of the PIMMS bioinformatic pipeline.

Raw fastq files for all three data sets were concatenated and parsed through the PIMMS pipeline using the modules, Find_Flank and BAM_Extract.

2.5.1 Find Flank Module

Raw fastq files and the *S. algalactiae* reference fasta file (UK15.fasta) were uploaded to the PIMMS **Find Flank** module. The following arguments (flags) where specified for all Nanopore fastq input:

-Nano (Nanopore settings for the PIMMS pipeline, which introduces a Levenshtein distance of 1 to allow for potential base call errors in the reads when identifying the terminal insertion sequence.
-min (minimum read length set to 100)
-max (maximum read length set to 250000)
-mapper (minimap2 was used for the alignment of reads to the reference FASTA – for BAM file generation)

Fastq reads with motifs containing the Tn termini (of the insert) were identified, these reads were retained and mapped to the reference genome (UK15.fasta) with adherence to parameters specified in the pimms2 config text file (Appendix 3). Two output files were generated from this module, a sequencing summary text file containing an overview of insertions found per input fastq file, and a BAM file containing all the aligned sequencing reads that contained a Tn motif.

2.5.2 BAM Extract Module

The BAM file generated using the Find Flank tool was then parsed to the **BAM_Extract** module together with a GFF file (General Features Format) containing a list of genomic features and their coordinates within the *S. algalactiae* genome. To ensure continuity this GFF file was generated using the same UK15 fasta file entered into the Find Flank module.

Two output files were produced:

1. A detailed list of insertions per annotated genomic feature, generated in a standard spreadsheet format (xlsx file) – the information contained in this file includes percentiles of the first and last insertion locations, as well as NRM scores (Normalized Reads Mapped (total number of reads/length of gene in Kb)/(total mapped read count/10 6) and NIM scores (Normalized Insertions Mapped - total unique insertions mapped/Length of gene in Kb)/(total insertions mapped/10 6).

2. A GFF file detailing the coordinates of each insertion.

2.5.3 Comparison of Sequencing Methods

PIMMS output files for three sequencing methods were used for a comparative approach. The output files had been generated using:

Tests
Standard ONT sequencing of BHI DNA
Cas9 Targeted ONT sequencing of BHI DNA

Control
Illumina sequencing of BHI DNA (this data was generated external to this project)

Both Standard ONT (test) and Cas9 ONT (test) sequencing data, and their PIMMS output files were compared to the external Illumina sequencing data and output (control). NRM and NIM values for each sequencing method where uploaded to the software package SPSS Statistics (v26) to generate summary statistics and histograms.

2.5.4 PIMMS Dashboard

Output files from each data set were uploaded to the PIMMS dashboard (https://pimms-dashboard-uon.azurewebsites.net). Each test condition was separately uploaded to the PIMMS dashboard alongside the Illumina (control) PIMMS output. Once uploaded run metrics were assessed using the Dashboard to:

- Compare NIM scores across the genome
- Generate and export xlsx files containing insertion events identified using one sequencing method but not the other.
- Visualise the genomic features identified as essential for each test sequencing method (against the control) as a Venn
- Generate a scatterplot graph looking at the relative differences in the number of mutations identified at each position in the genome for test and control.

3. Results

3.1 Cas-9 Targeted laboratory Method

3.1.1 Sample QC prior to library preparation

It was important to establish the quantity and integrity of DNA sample ds1093_3 prior to library preparation, to ensure enough material was present, and that the DNA was not degraded. Sample concentration was 50.2ng/µl, in this instance the concentration was ascertained using the Qubit V4 Fluorometer (Thermo Fisher Scientific). Recommended input to the Cas-9 library preparation protocol is 1-10µg, so for ds1093_3 the input was within this range, albeit close to the minimum threshold at 1.2µg.

Sample ds1093_3 was repeated (ID - ds1093_3_Rpt) and in this instance a 1X AMpure (Beckman) clean-up was used for two purposes, firstly to elute the DNA in Milli Q water, and secondly to concentrate the DNA so more input could be fed into the library preparation protocol. Additionally, adapter ligation time was doubled (40 minutes) in an attempt to increase sequencing output.
The Agilent TapeStation profile of ds1093_3 (Figures 3.1a and b) recorded a good DIN (DNA integrity number) of 8.5 and showed little signs of degradation with a large collection of DNA fragments being high molecular weight (HMW). For our purposes it was important that the DNA was not degraded as for long reads to be generated the starting material must also be long.

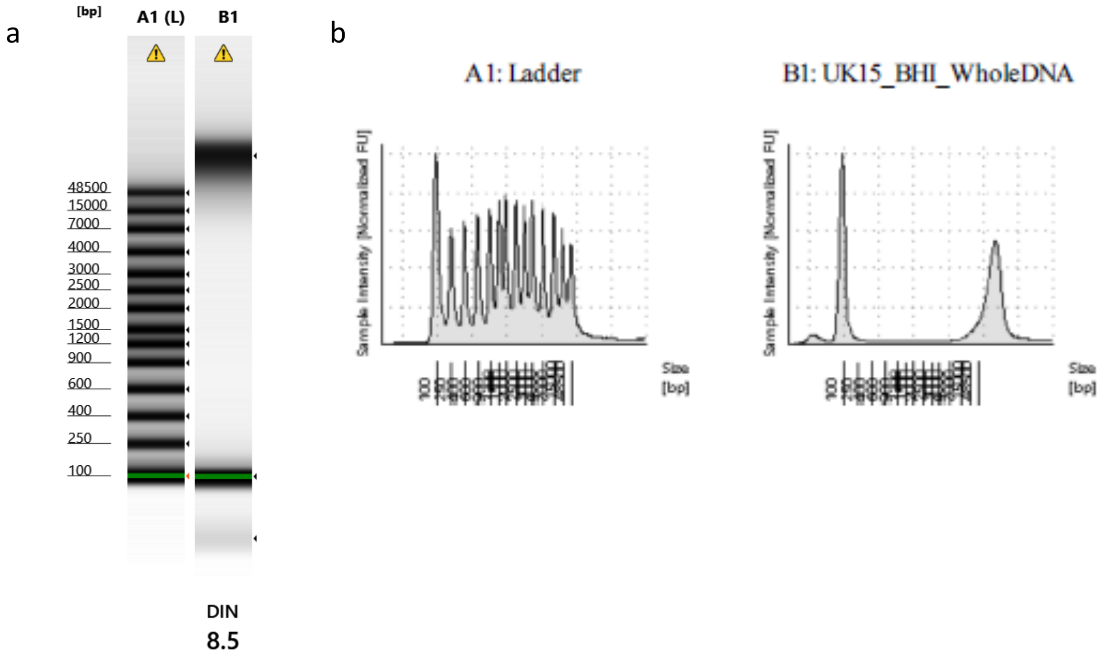


Figure 3.1 – TapeStation profiles of extracted DNA sample ds1093_3 (UK15_BHI_WholeDNA).
The Agilent TapeStation is a miniaturised electrophoresis system and migration profiles can be seen in 3.1a with the majority of source DNA (lane B1) being over 48.5Kb in length. The TapeStation traces in 3.1b gives a better visual representation of DNA sizing, the collection of fragments on the right hand side of trace B1 being the BHI DNA, the early sharper peak in the trace is the lower internal marker.

3.1.2 Library QC

Two data sets (Table 3.1) were generated using the Cas-9 targeted method. ds1093_3 and ds1093_3_Rpt.

Experiment ID	ds1093_3	ds1093_3_Rpt
Experiment type	Cas 9 enrichment	Cas 9 enrichment
Sample name	UK15 BHI DNA	UK15 BHI DNA
DNA Qubit	52.2ng/µl	78.8ng/µl
DIN	8.5	8.5
Input to library Prep	1.2µg	1.97µg
Cleavage time	15 minutes	15 minutes
Adapter ligation time	20 minutes	40 minutes
Final library Qubit	37.6ng/µl	59.4ng/µl
Input for sequencing	451ng	712.8ng
Flow cell ID	FAS93187	FAQ52641
Flow cell Position	x3	x3
Flow cell Pores	1347	1283
Output reads	34.32K	169.43K
Estimated bases	410MB	1.67GB
N50	25.28KB	18.94KB

Table 3.1 – Summary of library preparation metrics, sequencing run conditions and output. Modifications were made for the repeated library preparation in an attempt to increase the output. These changes included an increase in input DNA and adapter ligation time.

The Library preparation method for ds1093_3 generated 14µl of final library at a concentration of 37.6ng/µl. As only 12µl could be loaded onto the flow cell for sequencing, the total loading volume was 451ng. However, the wet lab modifications made to ds1093_3_Rpt increased the amount of end point library to a final library concentration of 59.4ng/µl, which in turn allowed a higher loading volume of 712.8ng. The effect of this increase can be clearly seen when looking at the sequencing output metrics. Focusing on estimated bases, the sequencing output has increased by over 300% (307.32%).

MinION flow cells were QC'd using the GridION (MinKNOW software v 21.11.7) to ensure enough pores were available for sequencing. As a rule; we wanted to use flow cells with pore availability > 1200.

3.1.3 Sequencing run QC

Run metrics were monitored in real time using ONTs MinKNOW software. For each sequencing run; output included pass and fail fast5 and fastq folders, a standard pdf run report and a sequencing_summary.txt file. Each run's sequencing_summary.txt file was used with software tool

NanoPlot (v 1.38.1) to generate graphical plots and a NanoStats file giving a summary of the key features of the dataset. NanoPlot reports were generated at first using all the data from each sequencing run, after which each data set was filtered to remove reads that had a quality threshold lower than Q9. This was necessary as only the Q9 data from the fastq pass folder would be used for downstream processing. The effects of filtering the raw data (Figure 3.2) resulted in 20-23% of reads being removed.

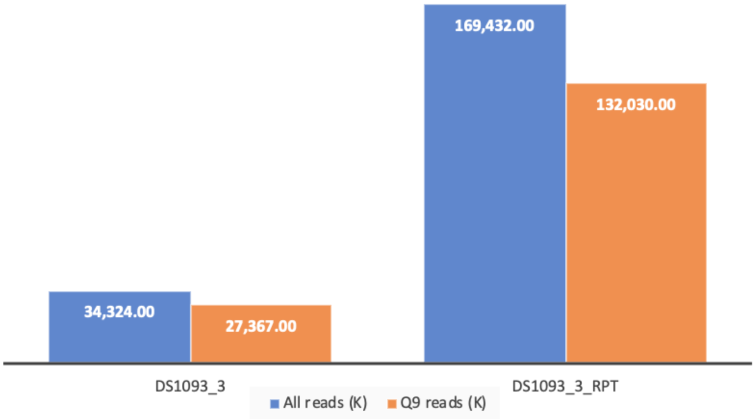


Figure 3.2 – Bar chart showing the number of reads removed with quality filters set to Q9.
Sequencing run ds1093_3 lost 6,957 reads which was 20.27% of its total reads. Ds1093_3_Rpt lost 37,402 reads, 22.07% of its total read count.

By viewing and assessing the various graphs generated by NanoPlot (Figures 3.3a and 3.3b), we can see large collections of long reads with length greater than 10Kb.

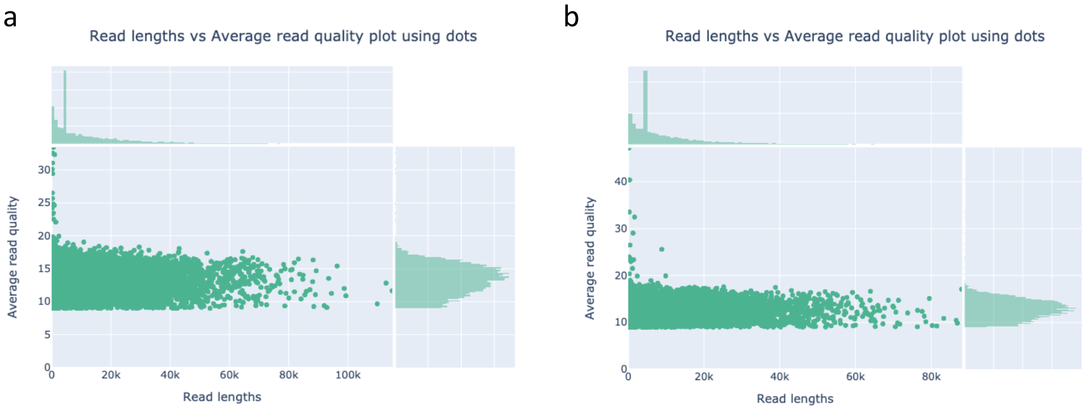


Figure 3.3 – Scatter plots showing the variation in reads length generated for each nanopore run vs the quality score of each read (post filtering of reads <Q9) Source NanoPlot.
Plot a shows the collective distribution of read lengths for ds1093_3 and the average read quality, the majority of data being between Q10 and Q20. Plot b which represents ds1093_3_Rpt displays a similar pattern.

3.2 Manual Alignments, Cas-9 Targeted Method

3.2.1 Read Filtering

For each run, following read QC, all fastq files were taken from each fastq pass folder and concatenated. The resulting fastq file was uploaded to Porechop (v0.2.4) and adapters were located and removed from each read. Adapter removal is now part of the Guppy toolkit and very few reads were filtered using Porechop (Figure 3.4). This gave a clear indication that Porechop adapter removal will not be required for this bioinformatic pipeline moving forward.

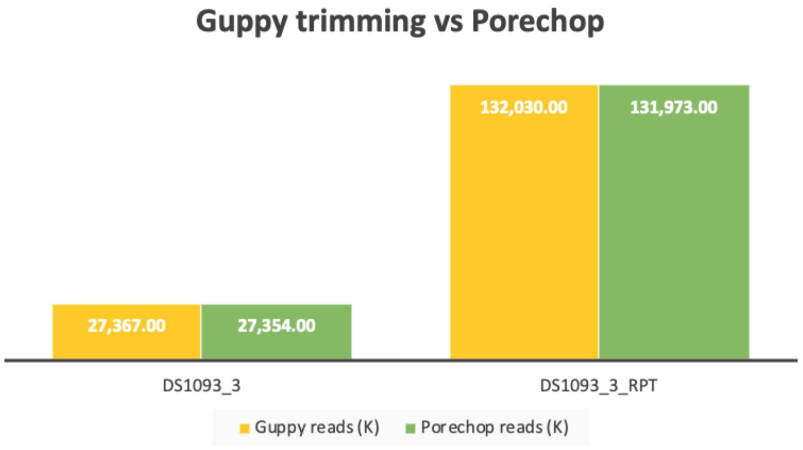


Figure 3.4 – Porechop filtering of fastq reads.
In yellow are sequencing reads that have been generated for each run using Guppy basecaller (v5.1.1.13), these reads were then uploaded to Porechop (v0.2.4) and the resulting filtered output reads are in green. Very few reads required filtering using Porechop, this being due to Guppy basecaller's in-built adapter removal application.

Trimmed fastq files were then aligned to the reference (UK15 – *S. agalactiae*) and insert (pGh9:IS*S1)* fasta sequences using Minimap2 v2.24 (code in Appendix 1). Output files were generated in PAF (**P**airwise m**A**pping **F**ormat) files for ease of downstream processing. A PAF is a plain text tabular format file, typically with 12 rows, each representing an alignment between two sequences. The library pafr (v0.0.2) was used to load each paf file into R, once loaded the paf file could then be manipulated like an R data frame which allowed greater flexibility, as we could filter the file using common data frame functions. To this end, after loading the two paf files (one for insert alignments and one for reference alignments) into R, the pafr (v0.0.2) library tool kit was used to remove all secondary alignments, followed by removal of all low mapping quality alignments (scoring <40). The final filtering method was achieved using the library dplyr (v1.8.6). The 'distinct' function removed all duplicated read IDs (qnames) from each paf file. This was necessary as both paf files would until now contain reads aligning to the fasta sequences in multiple places.

The total number of reads removed during these filtering steps for ds1093_3 (Figure 3.5) were noticeably different for insert and reference alignments. Each filtering step for reads that aligned to the reference (Figure 3.5a) removed a small proportion of reads, the final read count being 18,356, retaining 78.57% of reads first aligned. In contrast there was a large reduction in the number of reads aligning to the insert (Figure 3.5b) once all processing step were undertaken, 19,452 reads are retained, which only accounted for 7.85% of the total reads first aligned. The majority of these reads being lost when secondary alignments were removed.

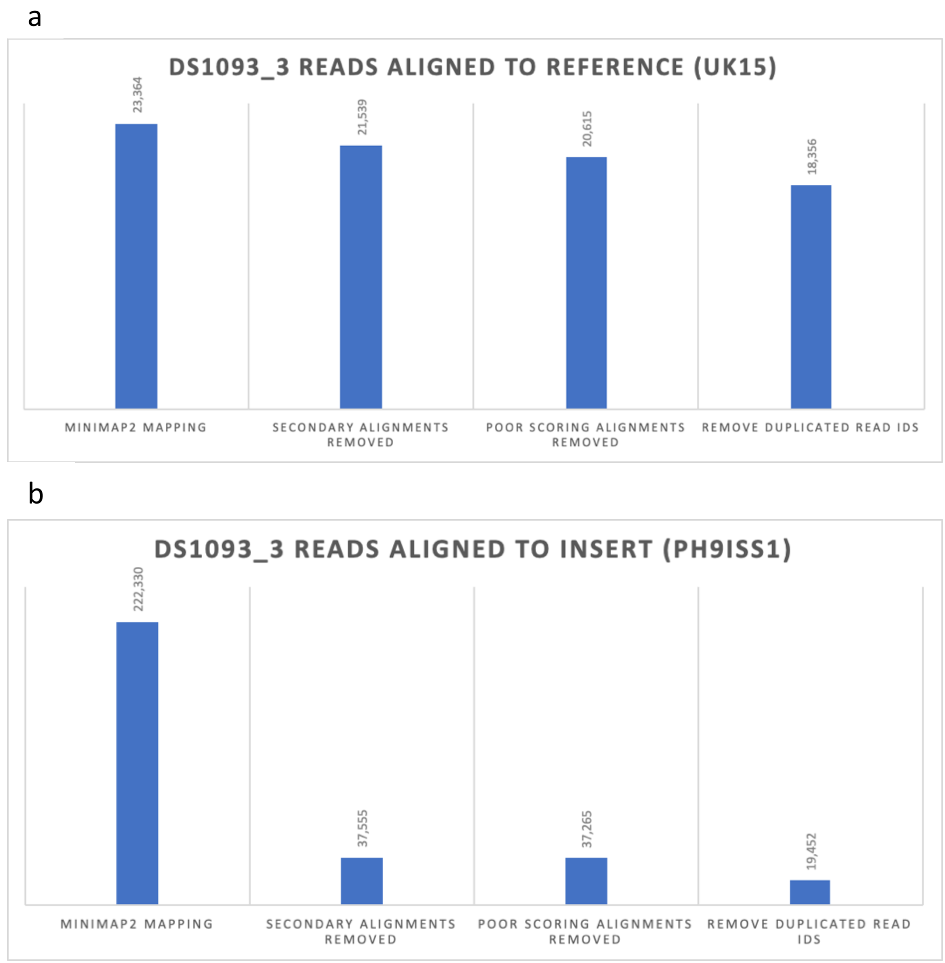


Figure 3.5 – Filtered reads for ds1093_3.
a) The result that each filtering process has on the number of reads aligned to the reference (UK15 – *S. agalactiae*), only 21.43% of reads being removed. b) Sequence reads aligning to the insert sequence (pGh9:IS*S1*) dramatically reduce when subjected to bioinformatic filtering, with 91.25% of reads being removed.

The results of filtering both reference and insert alignments for ds1093_3_Rpt follow a similar pattern, sequencing reads that align to the reference sequence reduce from 106,497 to 84,303 once taken through the manual bioinformatic filtering process (Figure 3.6a), a read retention of 79.16%. Sequencing reads that align to the insert sequence reduce from 1,047,876 to 85,915, a read retention of just 8.19%. As before, most insert aligned reads are lost when secondary alignments are removed (Figure 3.6b)

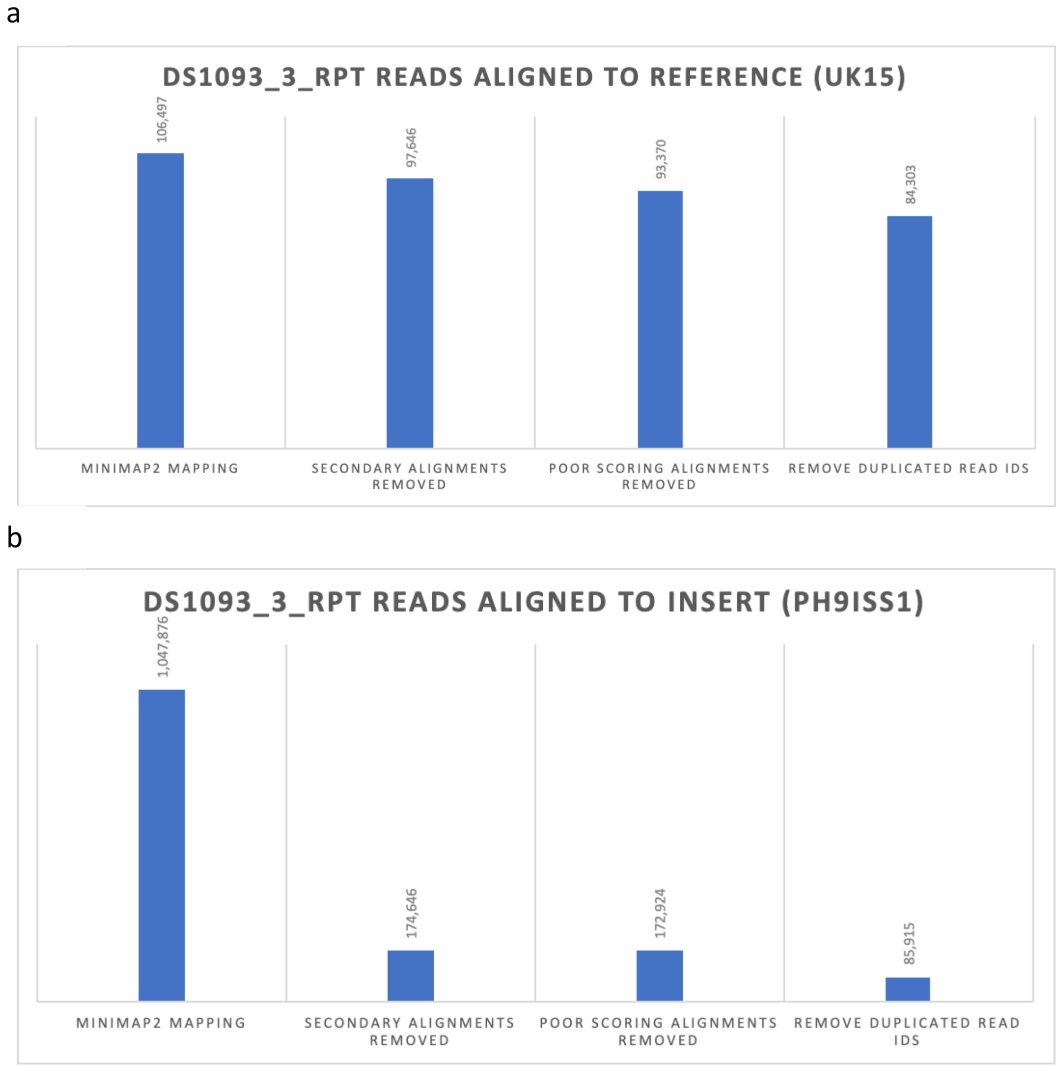


Figure 3.6 – Filtered reads for ds1093_3_Rpt.
a) The result that each filtering process has on the number of reads aligned to the reference (UK15 – *S. agalactiae*), only 20.84% of reads being removed. b) Sequence reads aligning to the insert sequence (pGh9:IS*S1*) dramatically reduce when subjected to bioinformatic filtering, with 91.81% of reads being removed.

Primary alignments were of greatest interest for this exercise, secondary alignments being classed as potential alignments between query and target sequences but not identified as the best alignments for those regions; were discarded. Looking at only the Q9 reads that had been processed through the above pipeline; sequence run ds1093_3_Rpt contains 85,915 reads that map to the insert, this is out of a total of 132,030 reads. An enrichment of 65.1%. Ds1093_3 recorded insert enrichment at 71.1%.

3.2.2 Data Set Merging

The data.frame function Inner.join (or natural join) located matching read ID's present in both insert and reference paf files. The output being the number of reads that aligned to both insert and reference. At least 50% of all reads that aligned to the insert, also aligned to the reference (Table 3.2). It also confirms that the total read count of each sequencing run contained a large portion of reads that had only aligned to insert or reference, 57.76% for ds1093_3 and 65.44% for ds1093_3_Rpt.

	Reads Aligned to Insert	Reads Aligned to Reference	Reads Aligned to Insert and Reference	Reads Aligned only to Insert	Reads Aligned only to Reference	Total reads
ds1093_3	19,452	18,356	11,227	8,225	7,129	26,581
ds1093_3_Rpt	85,915	84,303	43,721	42,194	40,582	126,497

Table 3.2 – Summary of read alignments after Inner.Join of filtered paf files - for the insert (Pgh9:IS*S1*) and the reference (UK15 – *S. agalactiae).*
For ds1093_3, 19,452 reads aligned to the insert using minimap2, 11,227 of which (57.72%) also aligned to the reference sequence. For ds1093_3_Rpt, 50.89% of insert aligned reads also aligned to the reference sequence (43,721 reads out of 85,915).

For the total reads generated by each sequencing run, an overview of the percentage alignments of each run is represented in pairwise Venn Diagrams (Figure 3.7).

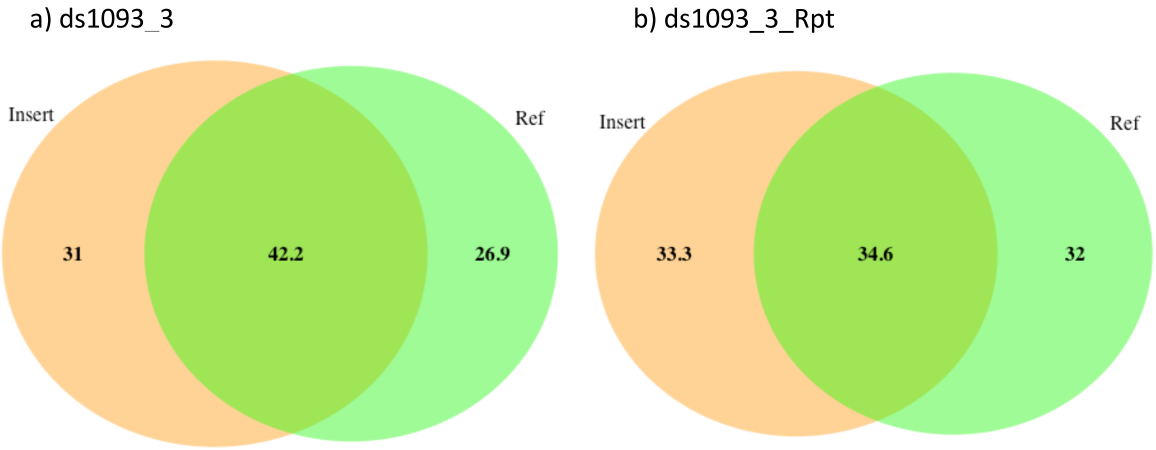


Figure 3.7 – Alignment Pairwise Venn Diagrams for ds1093_3 and ds1093_3_Rpt.
For both a and b, in orange are the percentage of sequencing reads that only aligned to the insert fasta file sequence using minimap2. In light green are the percentage reads that uniquely aligned to the reference fasta file. Centralised In dark green are the percentage sequencing reads that aligned to both insert and reference, in this instance suggesting transposon insertion.

3.2.3 Manual confirmation of Insert-Chromosome alignments

Data Quality Control and manual alignment confirmed the presence of sequencing reads that had aligned to both the insert and the reference sequence. To validate this conclusion, a sequencing read was taken at random from run ds1093_3 (Table 3.3). This sequencing read had been identified as having aligned to both insert and reference. The raw fastq sequence was uploaded to a NIH BLAST (Basic Local Alignment Search Tool: (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=bla

sthome). The BLASTn suite searched for nucleotide sequences that were similar to the sequencing read, two searches were performed, one selecting the organism *Streptococcus* (taxID – 1301), the other using Insertion vector pGh9:IS*S1* (taxid:481005)

Sequencing run	ds1093_3
Run ID	2066abf8b5900ada5871eb34889b2ef373de73d2
Read ID	493a4283-8f0e-4de7-a871-2f16e8c921e7
Alignment	Primary
Read length	15,025 bases

Table 3.3 – Sequencing read from run ds1093_3.
A long read of 15.025 bases was purposely selected to highlight the effectiveness of nanopore sequencing. A primary alignment was chosen to maximise confidence in BLAST results.

A BLAST search result will refer to the input nucleotide read as a query sequence, the nucleotide sequences within the BLAST database are referred to as subject sequences. The search results (Table 3.4) indicate that the input fastq sequence aligned to the insert from query position 29 - 2858, and then to the reference chromosome at query position 2859 – 15,025. We can surmise from this data that the sequencing read has read from insert sequence directly into the reference confirming transposon insertion.

Alignment	**Description**	**Sequence ID**	**Percentage ID**	**Query (read) Position**	**Subject position**
Insert	Insertion vector pGh9:ISS1, complete sequence	EU223008.1	96.2	29 - 2858	1530 - 4371
Reference	Streptococcus agalactiae strain 32790-3A chromosome, complete genome	CP029561.1	95.76	2859 - 15,025	2,048,658 - 2,060,935

Table 3.4 – Results of BLAST search for read 493a4283-8f0e-4de7-a871-2f16e8c921e7.
The query read has aligned to the insert from base positions 29-2858, and then to the reference from 2859-15,025. Both alignments having high base identification percentages of 96.2% and 95.76%. Confidence of both alignment is strengthened by their long read length of 2829 (insert) and 12,166 (reference) bases.

For extra confirmation this approach was repeated using three additional reads.

3.3 PIMMS Inverse PCR Laboratory Method

3.3.1 Sample Quality Control prior to Library Preparation

It was important to establish the quantity and quality of the input DNA sample prior to library preparation, to ensure enough material was present, and to confirm that size distributions were as expected. Sample ds1093_1 (PIMMS UK15 BHI) concentration was 24.2ng/ul, the concentration was ascertained using the Qubit V4 Fluorometer (Thermo Fisher). Recommended input to the SQK-LSK110 Genomic DNA by Ligation library preparation kit is 100+ng for fragmented DNA. So, for ds1093_1, input volume was more than adequate. Sample ds1093_1 had been prepared identically to samples that had previously been sequenced using short read methods, being sheared using covaries sonication. It was therefore expected that fragment distributions would be short (Figure 3.8).

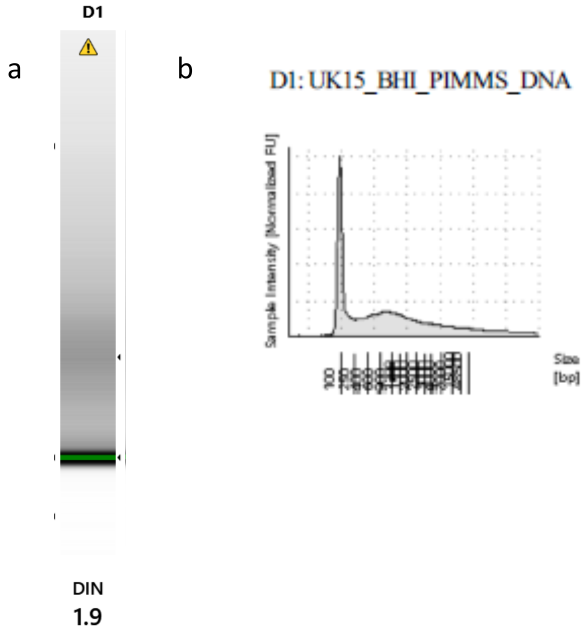


Figure 3.8 – TapeStation profile and trace for ds1093_1 DNA sample.
Sample has been generated using the PIMMS wet lab pipeline, ending in a DNA shearing step to generate a collection of fragments with an average length of ~600bp. The TapeStation profile in both a and b confirm this, as we can see a broad distribution of sizes peaking at around 800-900bp. The early sharp peak in trace b should be discounted as it represents an internal calibration marker.

3.3.2 Library Quality Control

The ds1093_1 data set (Table 3.5) was generated using a PIMMS ONT Sequencing method. Library preparation was achieved using Oxford Nanopore Technologies SQK-LSK110 library preparation kit (vGDE_9108_v110_revR_22May2022), for Ligation sequencing of genomic DNA.

Experiment ID	ds1093_1
Experiment type	Standard ONT
Sample name	PIMMS UK15 BHI
Library prep Method	SQK-LSK110
DNA Qubit (BR, 1ul)	24.2ng/ul
Input to Lib Prep	500ng
End Prep times	30 mins at 20°C 30 mins at 65°C

Adapter ligation time	1hr
Final lib Qubit (HS, 1ul)	16.4ng/ul
Input for sequencing	100ng
Flow cell ID	FAQ52406
Flow cell Position	x4
Flow cell Pores	1004
Output reads	18,282,606
Estimated bases	10,688,366,962
N50	695

Table 3.5 – ds1093_1 library preparation and run parameters.
Due to previous optimisation of the library preparation protocol, End Repair times where increased from 5 minutes at 20°C and 5 minutes at 65°C to 30 minutes at both temperatures. Adapter ligation time was also increased, from 10 minutes to 1 hour. These optimisation steps have been proven to increase end point library concentrations.

MinION flow cell was QC'd using the GridION (MinKNOW software v 21.11.7) to ensure enough pores were available for sequencing. As the DNA was fragmented prior to library preparation, we knew read lengths would be short, and therefore flow cells with pore availability ~1000 would still generate a large amount of data.

3.3.3 Sequencing run QC

A NanoPlot report was generated at first using all the data from each sequencing run, after which each data set was filtered to remove reads that had quality threshold lower than Q9. This was necessary as only the Q9 data from the fastq pass folder would be used for downstream processing.

Removing reads that gave Phred scores lower than Q9 resulted in the removal of 27.04% of all reads (Figure 3.9), this was significantly more than the volume of reads removed for the Cas-9 experiments.

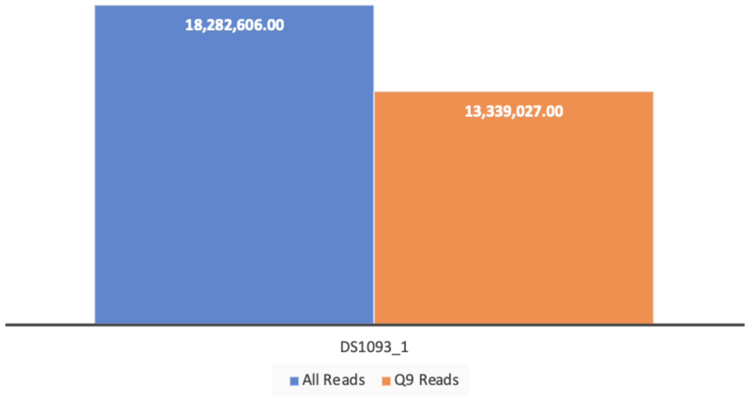


Figure 3.9 – ds1093_1 reads pre and post filtering.

Only 72.96% (in orange) of all reads have Phred scores equal to or higher than Q9. These reads alone are carried through to downstream bioinformatic processing.

By viewing and assessing the various plots generated by NanoPlot, we gain a strong sense of how well the sequencing run has worked. The majority of read lengths generated are short (Figure 3.10), with an N50 (at least 50% of the data containing reads greater in length than this figure) of 695 bases.

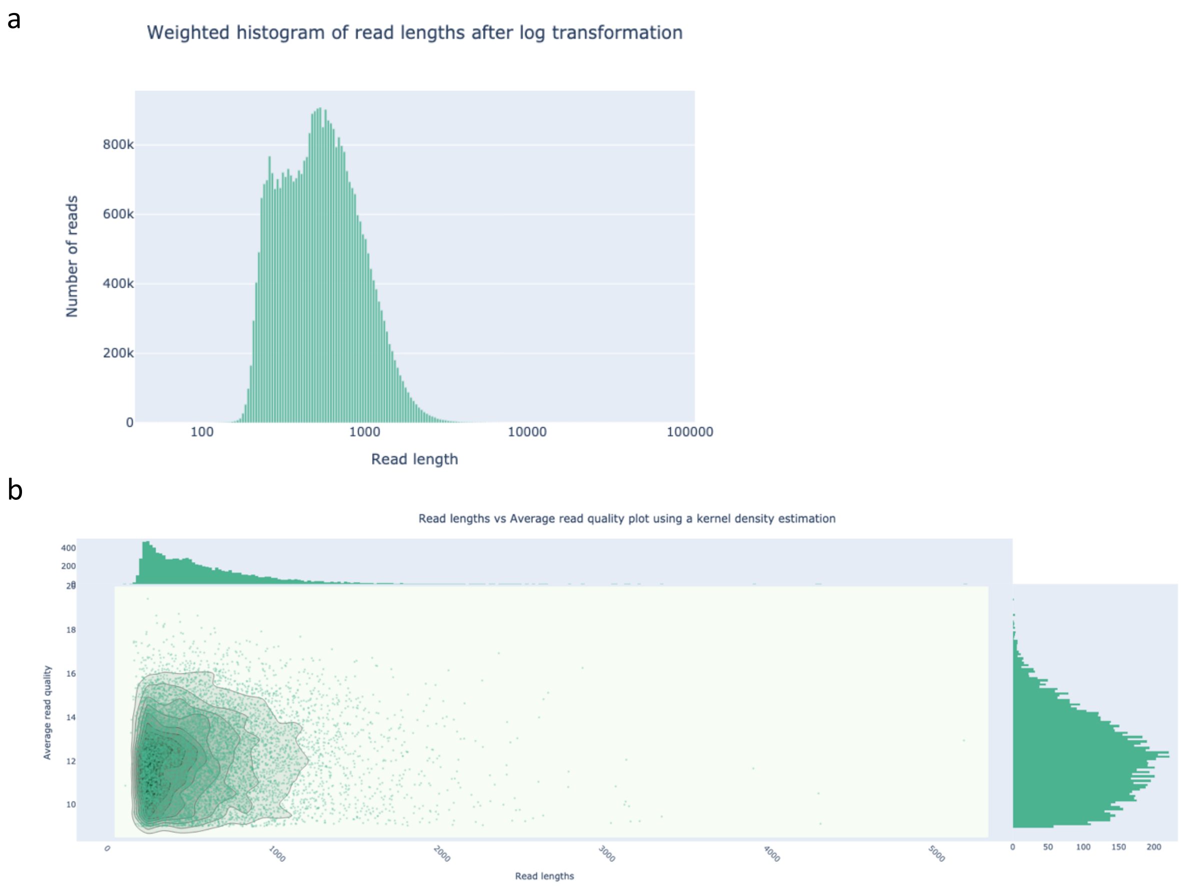


Figure 3.10 – Quality control plots for ds10983_1
Plot a is a weighted histogram of read lengths after log transformation, a normal bell shaped distribution showing a tight collection of short reads, with very little over 1000 bases in length. The relationship between read length and read quality is highlighted in Plot b, again the read quality has a normal distribution (observed on the right hand side of the plot) with a mean read quality of 12.1.

The sequencing output from a flowcell can be dependent on the length of the DNA libraries being sequenced. It is generally accepted that fragmented DNA libraries (short libraries) will generate more sequencing data than larger libraries (from native genomic DNA). Sequencing run ds1093_1 generated 11.1 gigabases (Gb) of data, 2.9Gb of which was removed as it didn't meet the quality threshold. This still leaves a fastq pass folder containing 7.79Gb of data. ONT product specifications suggest that a MinION should generate around 10Gb of data when sequencing short reads. We can therefore surmise that this sequencing run has performed optimally.

3.4 PIMMS Output – Illumina vs Nanopore

3.4.1 Insertion locations in the *S. agalactiae* genome

PIMMS output files (xlsx and GFF) generated from Illumina and Nanopore sequencing of cultured BHI DNA were uploaded to the PIMMS dashboard, using the Graphical User Interface (GUI). A scatterplot (Figure 3.11) gives a visual representation of the number of insertions mapped by location across the *S. agalactiae* genome, both plots show a broad and evenly distributed array of insertions, and no noticeable instances of location bias were observed for either sequencing method.

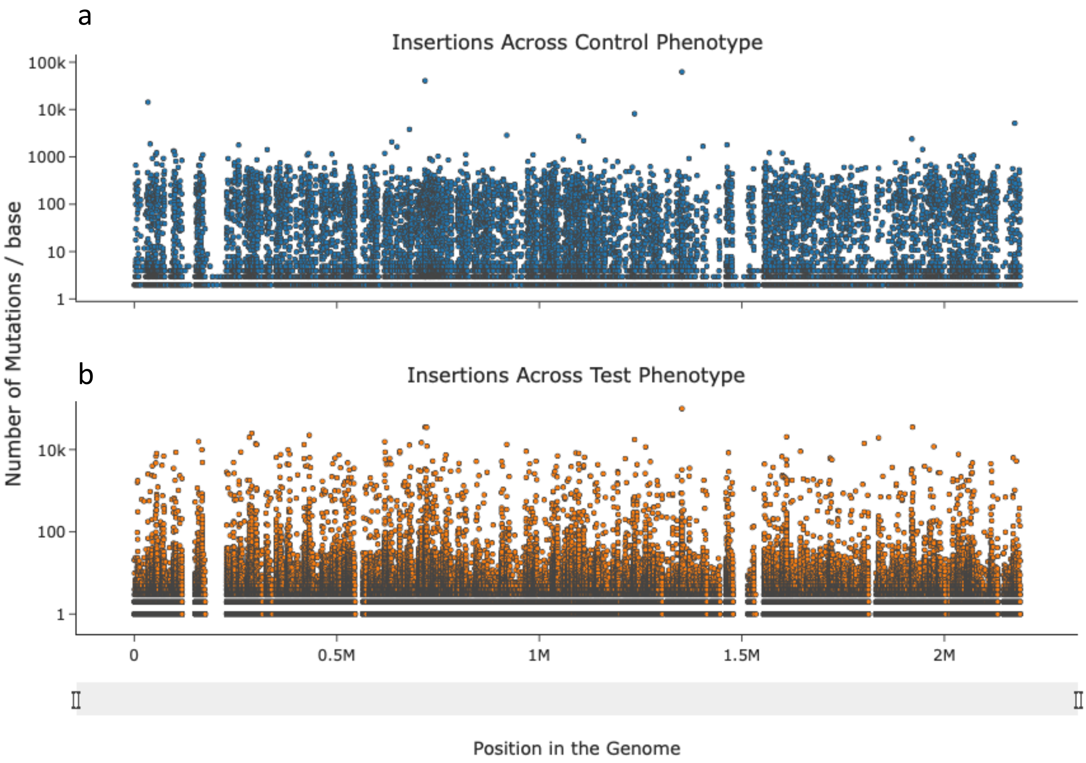


Figure 3.11 – Scatterplot showing depth and position of mutations plotted across the S. agalactiae genome. The number of insertions identified at the base level (y axis) using the PIMMS pipeline is plotted against genomic position (x axis). Both plots give an overall impression the location of mutated bases and the depth of mutations reported using each sequencing method, Figure 3.11a being generated using Illumina data, Figure 3.11b represents the Nanopore data.

3.4.2 Normalised Insertion Mapped Scores

Normalised Insertions Mapped (NIM) scores (total unique insertions mapped/Length of gene in Kb)/(total insertions mapped/10 6) give a number of insertions per gene value that can be used as a comparison between test and control. The NIM scores for the Illumina method were visually higher than those generated using the Nanopore method (Figure 3.12).

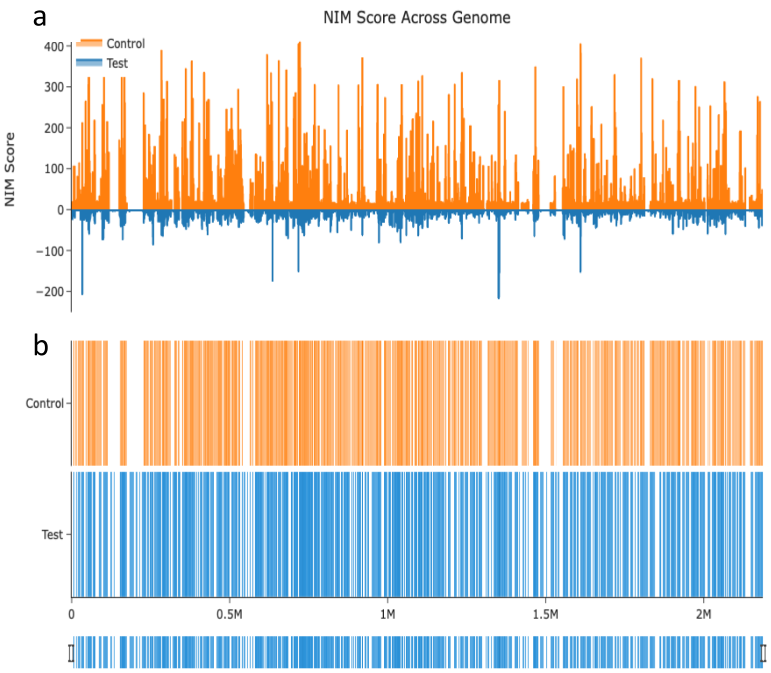


Figure 3.12 – NIM (Normalised Insertions Mapped) scores represented across the S. agalactiae genome using Illumina and Nanopore PIMMS output.
The Illumina NIM scores (orange) in Plot a are visually higher than those generated sing Nanopore data (blue). However, the coverage across the genome as seen in Plot b looks very similar.

Visually it was observed that some locations (Figure 3.13) identify positions in the genome that have no insertions with the control data sets (Illumina), but insertions are visible in the test (Nanopore) data set.

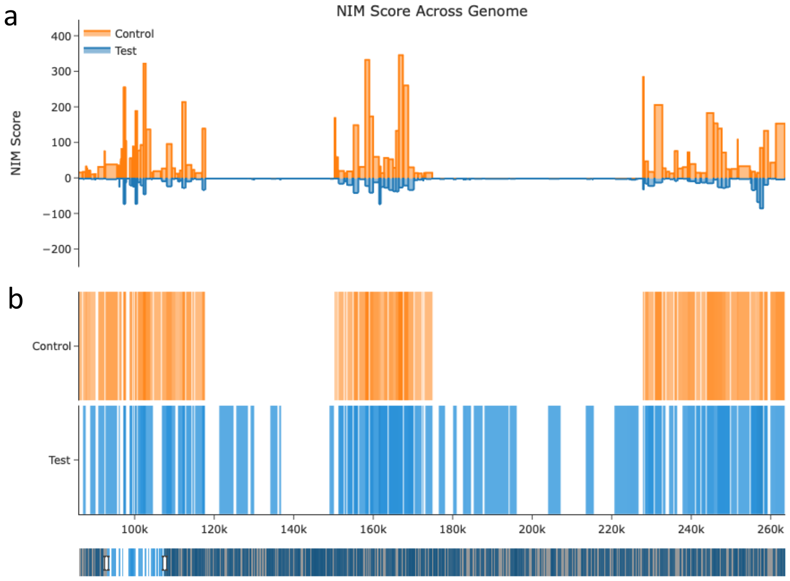


Figure 3.13 – A magnified look at NIM scores for both Illumina and Nanopore PIMMS data sets.
3.13b shows genomic locations where no insertions are detected using Illumina data PIMMS output (represented in orange), but insertions are visible using Nanopore data PIMMS output (blue).

The NIM score values for both the Illumina and Nanopore PIMMs output were used to generate summary statistics, and histograms to look at the frequency distributions of each as a continuous variable. NIM scores of 0 were removed from each data set before proceeding, this was so we could focus on the differences between NIM scores of insertion events for each sequencing method. The histograms for both Illumina and Nanopore NIM scores (Figures 3.14a and 3.14c were positively skewed due to a large collection of low NIM scores.

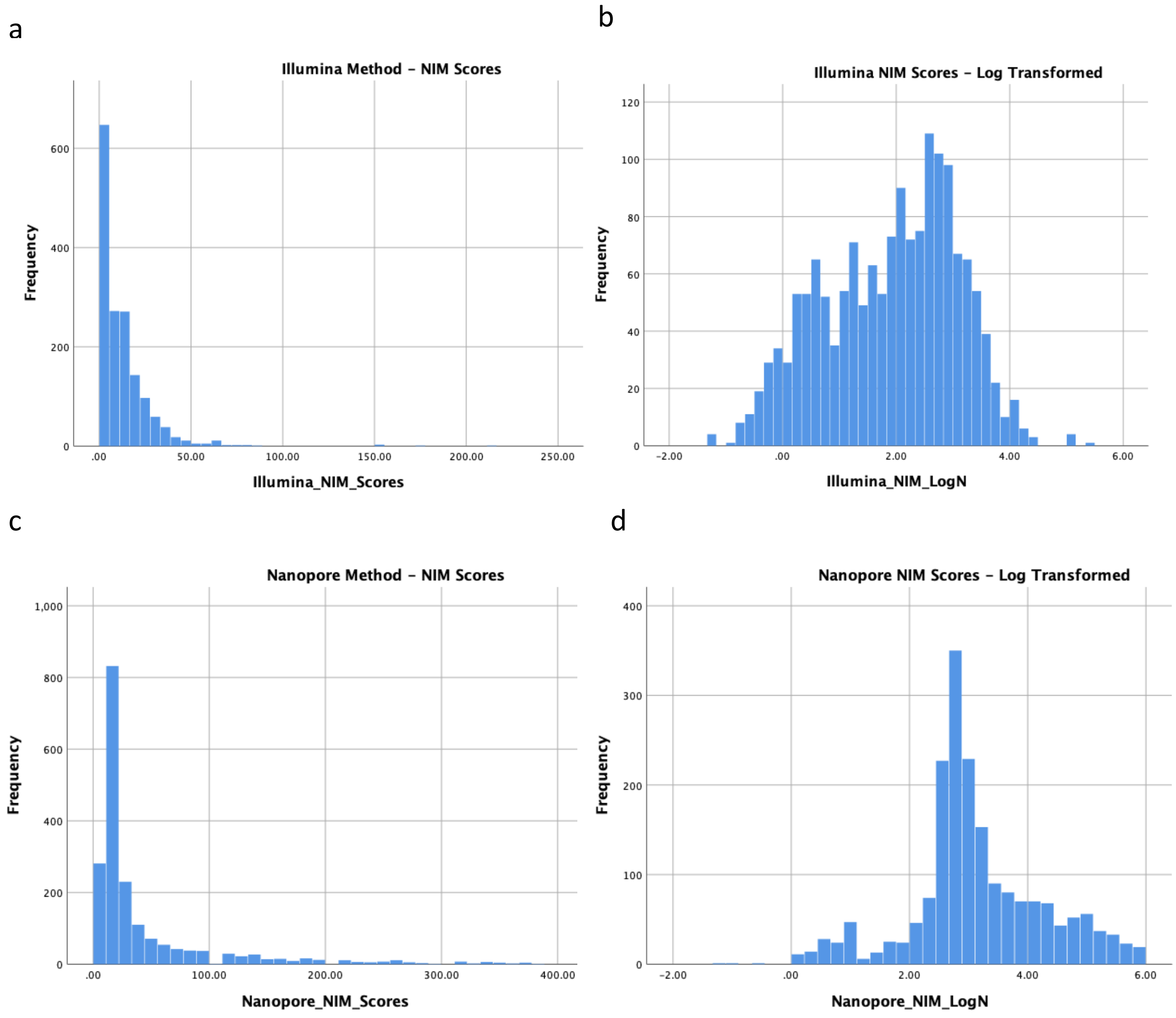


Figure 3.14 – Histogram of NIM scores for Illumina and Nanopore PIMMS output.
Both NIM data distributions where positively skewed as can be seen in histograms a and c, the Statistical software package SPSS was used to log transform both data sets, the resulting histograms b and d now show a roughly normal distribution. Nanopore method NIM scores are noticeably higher than those generated using the Illumina method.

The histograms show the NIM scores for both methods are quite low, albeit Illumina has noticeably lower NIM scores than those generated using the Nanopore Method. Summary statistics support this conclusion with a NIM score median difference of 10.52. The median value for Nanopore NIM scores being 18.68, 50% of the values falling between 13.55 and 42.73, in contrast the median value of Illumina NIM scores was 8.16, with 50% of the values between 2.74 and 16.74. Both data sets were log transformed, producing roughly normal distributions (Figures 3.14b and 3.14d), allowing a comparative approach to be used for the two outputs. A boxplot graph (Figure 3.15) shows the variation between NIM scores for both sequencing methods, with Nanopore NIM score distributions being noticeably higher than Illumina. This is in contrast to the PIMMS output scatterplot (Figure 3.12) which gave the visual impression that NIM scores for the Illumina data output (control in orange) were higher than that for Nanopore (Test in blue). It can also be observed that the nanopore NIM score data contains numerous outliers, which could potentially have an impact on any subsequent statistical analyses.

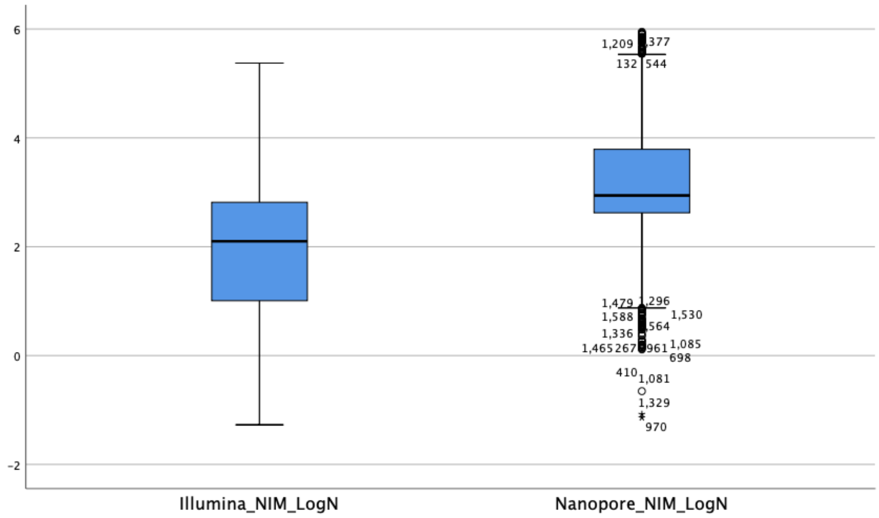


Figure 3.15 – BoxPlot showing the NIM score distribution for Illumina and Nanopore PIMMs output. Source SPSS (V26).

The average NIM scores for Nanopore being higher than that produced using Illumina data. Also observed is a number of data outliers in the nanopore data set. This boxplot was generated using log transformed data.

3.4.3 Identification of Essential Genes

A NIM score of 0 is a clear indication that the gene in question might be essential. The total number of all genes that recorded NIM scores of 0 across the two sequencing methods can be seen in the Venn compare diagram (Figure 3.16). The three interlinked groups show that 238 genes (displayed in purple) contained no transposon insertion in both Illumina and Nanopore group, the concordance of both sequencing methods strengthening the argument that they might be essential, or perhaps simply showing that both methods perform similarly. When observing the differences between each methods output, 405 genes were identified as being essential using the Illumina sequencing data (Control in blue) that the nanopore sequencing data identified as non-essential (containing insertions). In contrast Illumina sequencing located 79 non-essential genes that Nanopore sequencing indicated were essential.

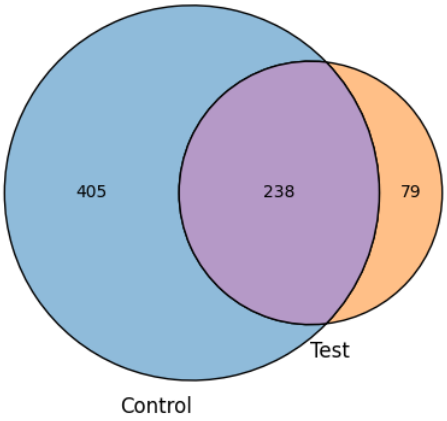


Figure 3.17 – Venn Diagram showing incidence of insertion absence.
Parameters included data that recorded 0 insertions (NIM scores) and were situated within the 0-100th percentile range. The control group in blue represents the number of genes that Illumina data recorded as being essential but where Nanopore data identified transposon insertions. The test group are genes that Nanopore recorded no insertions however Illumina data did find them. The Purple group represent genes that both Nanopore and Illumina data sets agreed where essential.

Three output xlsx files were generated, one containing gene's that were designated essential by both methods, one for genes only designated essential using Illumina sequencing, and one for genes only designated essential using Nanopore sequencing. NIM scores were uploaded to SPSS, and histograms of both data sets were skewed so required log transformation. The median value of insertions only identified using Nanopore data was 16.8 (with 50% of values being between 12.74 and 22.31), slightly lower than the median NIM score of 18.68 for all the Nanopore data, but still a strong average which gives confidence that the insertions identified are real. NIM scores for insertions only detected using Illumina data gave a median value of 2.86 (with 50% of values being between 1.29 and 7.53), again lower than the NIM score average for all Illumina data (4.14). Similar to the summary statistics for all the Illumina and nanopore NIM scores, we see a noticeable difference in median scores, this time being 13.94.

3.5 PIMMS Output – Illumina vs Cas-9 Targeted Sequencing

3.5.1 Insertion locations in the *S. agalactiae* genome

Output files generated using the PIMMs pipeline for Illumina data sets, and Cas-9 Targeted sequencing data sets were uploaded to the PIMMS dashboard. As before the insertion instances located using Illumina sequencing data would act as the control, and the Cas-9 insertion data, the test. A scatterplot (Figure 3.18) gives a visual representation of the number of insertions mapped by location across the *S. agalactiae* genome, as before we are looking at the number of insertions identified at the base level (X axis) using output data from the PIMMS pipeline, which is then plotted against genomic position (Y axis), even though the Cas-9 data does display an evenly distributed array of insertions, there is a noticeable difference in the depth of the data compared to the Illumina dataset. We can be more confident when looking at the Illumina data as each base mutation has been confirmed multiple times. The Cas-9 dataset chosen for this comparison is that from ds1093_3_Rpt as this was the Cas-9 sequencing run that produced the most amount of data. For each dataset, no noticeable instances of location bias were observed.

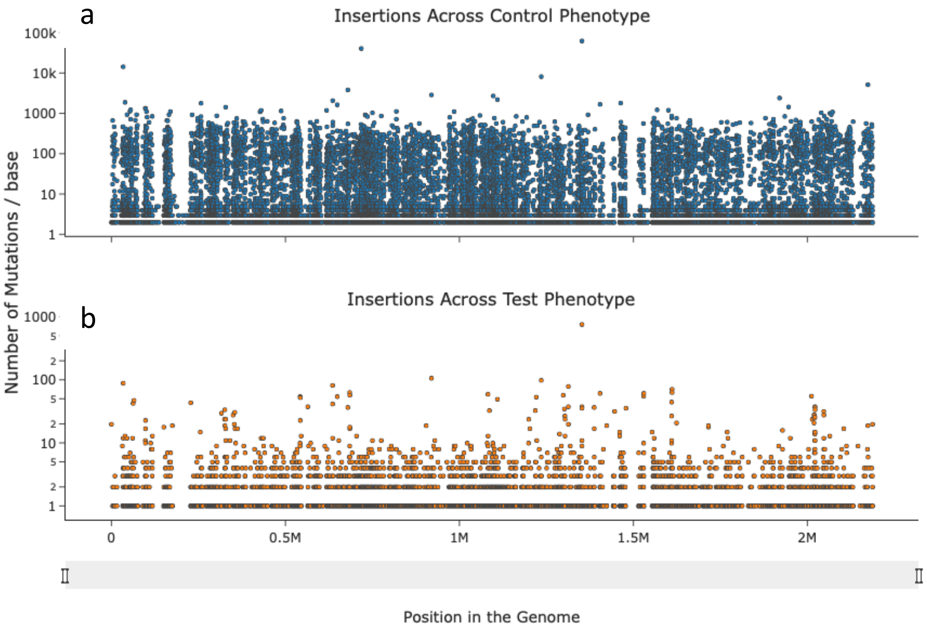


Figure 3.18 – Scatterplot showing depth and position of mutations plotted across the S. agalactiae genome. The X axis shows the genomic location, the Y axis records the number of mutations identified per base. Control phenotype in blue (Plot a) generated using Illumina PIMMS output data, orange is the Cas-9 PIMMS output data (Plot b). Depth of the Cas-9 identified mutations is substantially lower than the that generated using Illumina OIMMS output.

3.5.2 Normalised Insertion Mapped Scores

NIM scores generated using the Illumina method are visually far higher and in greater abundance compared to those generated using a Cas-9 Targeted sequencing approach (Figure 3.19). It is expected that the Illumina sequencing data was generated to a far greater depth than the Cas-9 sequencing, the DNA used for the Illumina approach would have been amplified using inverse PCR during the PIMMS Seq Laboratory pipeline, and then amplified further using PCR during the library preparation.

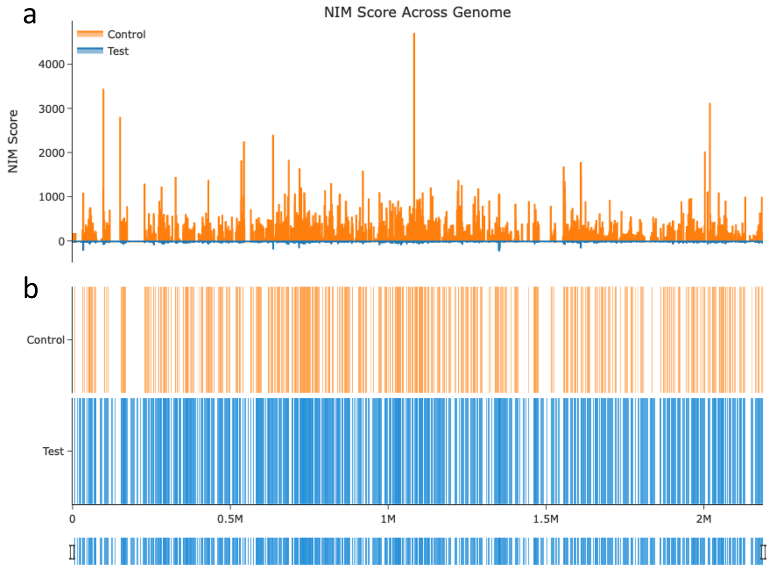


Figure 3.19 – NIM (Normalised Insertions Mapped) scores represented across the *S. agalactiae* genome using Illumina and Cas-9 Targeted PIMMS output.
Plot a shows a large contrast between the depth of mutation data generated using each method, the Illumina data set recording a much higher proportion of NIM scores across the genome. Plot b shows that the location of NIM scores per position are actually quite similar, in spite of the differences in data depth.

A closer look at NIM scores across the genome highlights insertions detected by Cas-9 sequencing that are not visually located using Illumina sequencing (Figure 3.20).

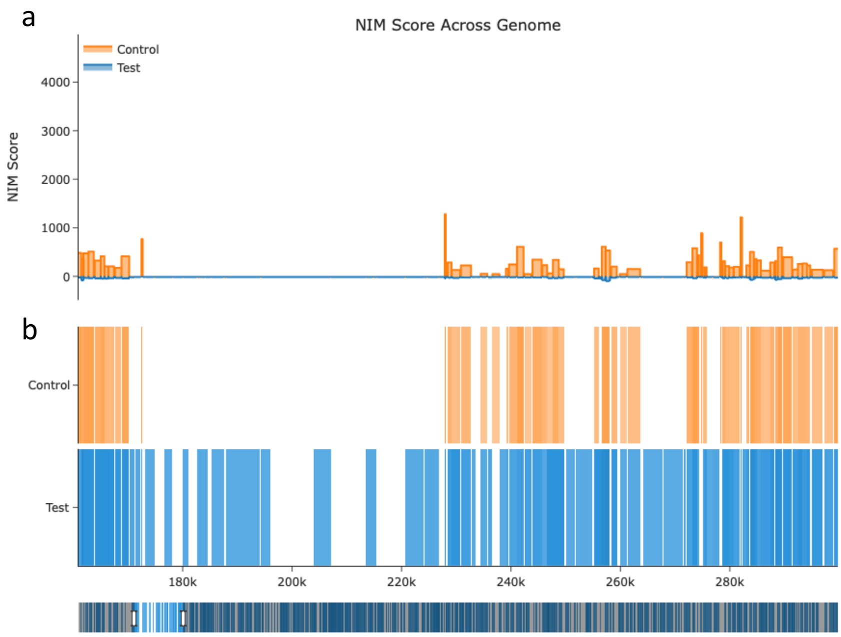


Figure 3.20 - A magnified look at NIM scores for both Illumina and Cas-9 PIMMS data sets.
3.20b identifies genomic locations where no mutations are detected using Illumina data PIMMS output (in orange), but mutation are clearly visible using Cas-9 data PIMMS output (blue).

In these instances, the NIM scores are low, but as were dealing with native DNA this is perhaps an expected observation. The NIM score values for the Cas-9 Targeted sequencing PIMMs output were used to generate summary statistics and a histogram (Figure 3.21a), so a comparison to the Illumina output could be drawn. As with all data output from the PIMMS pipeline, the Cas-9 NIM scores produced a histogram with a positively skewed distribution, to produce a roughly normal distribution the NIM scores where log transformed (Figure 3.21b).

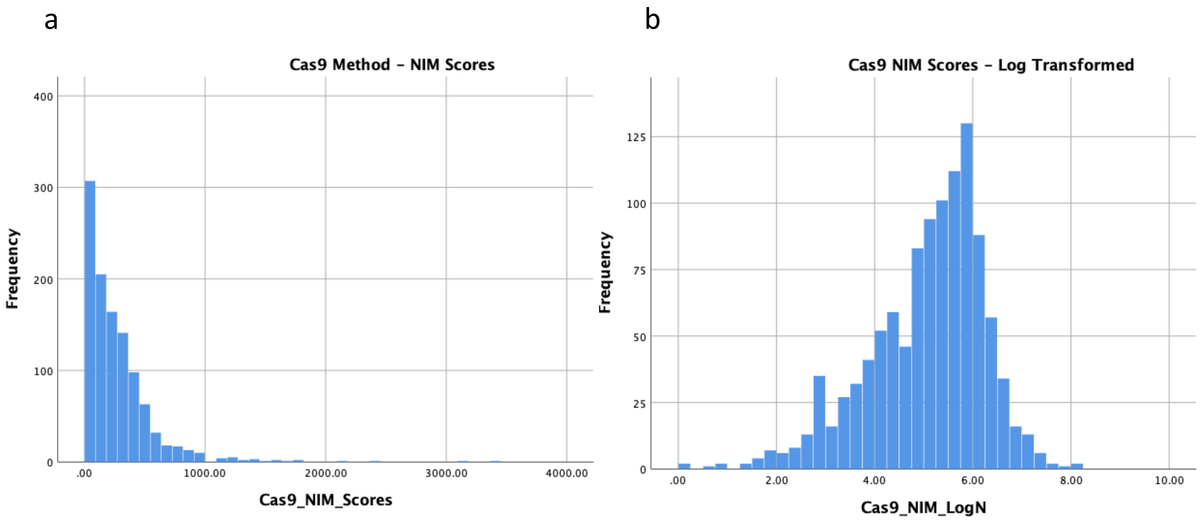


Figure 3.21 - Histograms for Cas-9 NIM scores.
The raw histogram (a) of NIM scores was positively skewed, it was log transformed (b) to produce a roughly normal distribution. Source SPSS (v26)

The median value of Cas-9 NIM scores is 197.85, with 50% of values falling between 78.89 and 364.27. This is a large increase from the median NIM scores for Illumina which was 8.16 (50% of values between 2.74 and 16.71). The median difference between the two data sets is 189.69. A BoxPlot (Figure 3.22) emphasises the difference in NIM score distributions.

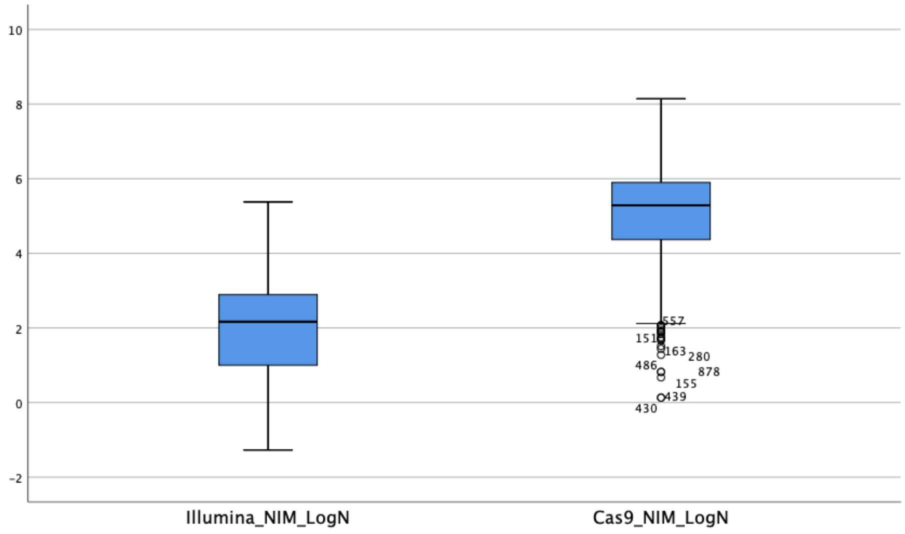


Figure 3.22 – BoxPlot showing the NIM score distribution for Illumina and Cas-9 PIMMs output. Source SPSS (V26).
The average NIM scores for Cas-9 are greater than that produced using Illumina data. Also observed is a number of data outliers in the Cas-9 data set (similar to Nanopore). This boxplot was generated using log transformed data.

3.5.3 Identification of Essential Genes

A NIM score of 0 would indicate that the gene in question might be essential. A Venn diagram (Figure 3.23) covers all the total genes that recorded NIM scores of 0 across the two sequencing methods in question. The three interlinked groups show that 593 genes (displayed in purple) contained no transposon insertion in both test and control group, the concordance of both sequencing methods strengthening the argument that they might be essential. 50 genes were identified as being essential using the Illumina sequencing data (Control in blue) that the Cas9 sequencing data identified as non-essential (containing insertions). Illumina sequencing located 547 non-essential genes (in orange) that Cas9 Targeted sequencing indicated were essential. Accepting that the depth of the Cas-9 sequencing data was extremely low compared to that generated using Illumina sequencing, it has still identified insertions events in 50 genes (49 of which are recognised as coding sequences) that Illumina output classified as essential.

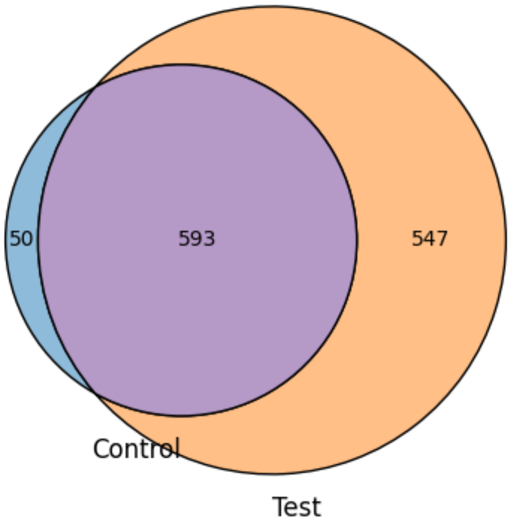


Figure 3.23 – Venn Diagram showing incidence of insertion absence.
Parameters included data that recorded 0 insertions (NIM scores) and were situated within the 0-100th percentile range. The control group in blue represents the number of genes that Illumina data recorded as being essential but where Nanopore data identified transposon insertions. The test group are genes that Nanopore recorded no insertions however Illumina data did find them. The Purple group represent genes that both Nanopore and Illumina data sets agreed where essential.

An output xlsx file containing gene's that were designated essential using the Cas-9 method was generated so that a comparison to the output xlsx file of unique insertion events detected using the Illumina method could be achieved. NIM scores were uploaded to SPSS, and on this occasion the median values could be generated and compared.

The median value of insertions only identified using Cas-9 data was 138.46 (with 50% of the data falling between 72.42(Q1) and 391.79(Q3)), NIM scores for insertions only detected using Illumina data gave a median value of 2.86 (with 50% of the data falling between 1.29 and 7.53). It is worth emphasising the significant difference in NIM scores between the two values. It is noticeable that the NIM scores for both Nanopore sequencing methods (Standard nanopore using PIMMS pipeline, and Cas9 Targeting sequencing) are higher than those generated using Illumina sequencing data.

3.5.4 NIM scores, Illumina, Nanopore and Cas9

The summary measures have shown that the NIM scores values for each sequencing method are statistically different, this can be further visually appreciated in a BoxPlot (Figure 3.24) which shows the complete NIM score distribution for all three methods.

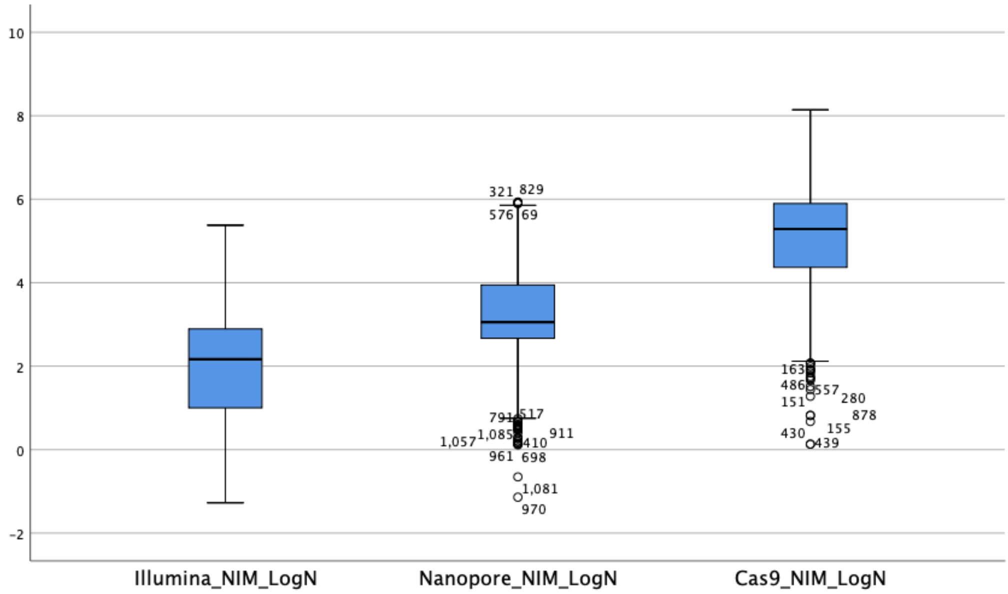


Figure 3.24 – BoxPlot showing the distribution of NIM scores for Illumina, Nanopore and Cas-9 methods. Source SPSS (v26)

The Illumina NIM scores are more tightly compact, whereas both Nanopore and Cas-9 NIM scores show more variation, containing multiple outliers.

3.5.5 Normalised Reads Mapped Scores, Illumina, Nanopore and Cas-9

The Normalised Reads Mapped (NRM) score, generated as output by the PIMMS bioinformatic pipeline can be used to compare each experimental condition, the NRM scores generated using the three different sequencing methods all gave skewed frequency distribution that required log transformation. Summary measures (Table 3.6) were generated using SPSS (v26).

Statistics

		Illumina_NRM_Scores	Nanopore_NRM_Scores	Cas9_NRM_Scores
N	Valid	1589	1915	1092
	Missing	326	0	823
Median		143.9000	32.2600	289.7800
Percentiles	25	12.9300	19.3100	116.8300
	50	143.9000	32.2600	289.7800
	75	498.9250	98.5300	663.0925

Table 3.6 – NRM summary statistics for Illumina, Nanopore and Cas-9 experiments, generated using SPSS (v26).

PIMMS output data generated Nanopore sequencing recorded the lowest median NRM score of 32.26. Followed by Illumina at 143.9 and Cas-9 at 289.78.

The distribution of NRM values (Figure 3.25) shows a similar outcome to that of NIM scores with respect to outliers, both the Nanopore and Cas-9 distributions containing multiple outliers scores compared to the Illumina values.

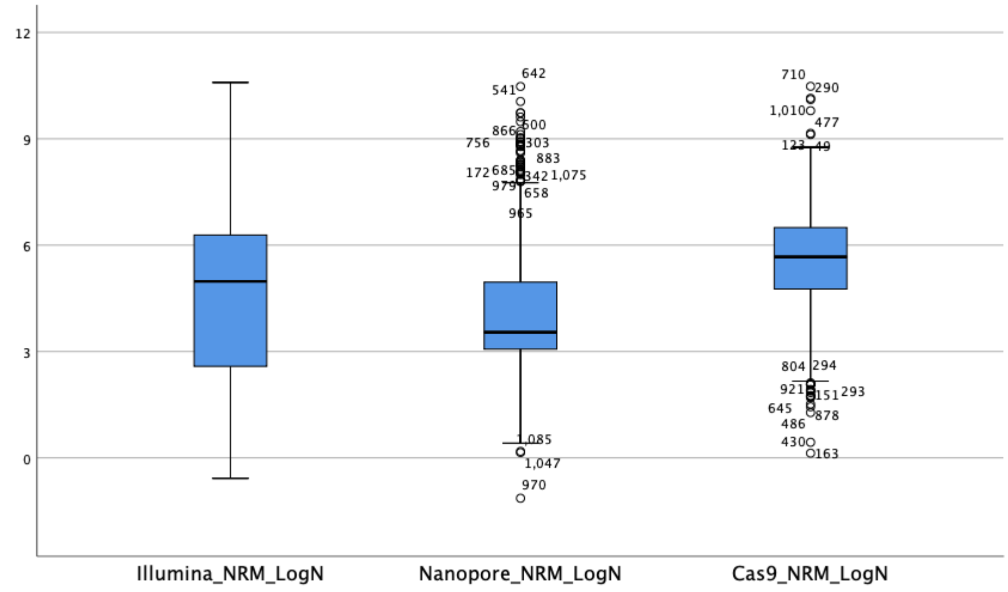


Figure 3.25 – Boxplot of NRM (Normalised Reads Mapped) score distributions for all three methods. Source SPSS (v26).

NRM scores across the three methods are not as variable as the NIM scores however, as with the NIM scores, both nanopore sequencing methods have generated a large collection of data outliers.

4. Discussion

Streptococcus Algalactiae (*S. algalactiae*) is a gram positive bacterium, also known as a Group B *Streptococcus* (GBS), it was first isolated in the 1930s from milk and cows with bovine mastitis (mammary gland inflammation in the breast or udder). It only began to be routinely diagnosed in humans from the 1960s onwards (Raabe et al, 2019) and is now recognized as being a leading cause of invasive disease in neonates and young infants, especially in high-income countries.

The study of the *S. agalactiae* essential genome has already been undertaken using traditional transposon sequencing techniques such as Tn-Seq (Hooven et al, 2016), the primary goal being to ascertain the contribution of individual genes towards the fitness of an organism, particularly when it is placed under different experimental pressures. It is customary that most transposon sequencing methods have until now utilised short read Illumina sequencing (Cain et al, 2020).

Recent development in transposon sequencing have seen the introduction of long read sequencing technologies, as an alternative to Illumina. One of these technologies, Nanopore sequencing, has been shown to be a useful tool in the detection of unique of transposon insertion sites (Yasir et al, 2022) with the potential to resolve entire Transposable Element (TE) insertions (Ewing et al, 2020).

4.1 Cas-9 Targeted enrichment of pGh9:IS*S1* insertion mutant libraries.

4.1.1 Amplification Bias

A reliance on amplification techniques such as PCR to generate DNA for library preparation will often lead to the under representation, or in some cases exclusion, of genomic locations that contain extreme base compositions (Aird et al, 2011). These sequence locations will be lost during many cycles of PCR due to the preferential amplification of DNA positions more amenable to the method parameters.

The impact of amplification bias on transposon insertion detection has become increasingly more scrutinized, particularly as the possibility of sequencing native DNA is now a realistic option. A recent study comparing Tn-Seq results with Nanopore Cas-9 sequencing (Alkam et al, 2021) concluded that even though Tn-Seq and Cas-9 results were comparable, the detection of a small number of genes previously identified as essential is sensitive to the number of PCR cycles used to amplify the source DNA. Their overall conclusion was to emphasise careful consideration of the number of PCR cycles required, particularly when attempting to identify putative essential genes. Additional to reducing PCR cycles, other method adaptations have been considered to reduce amplification bias, including the avoidance of locus specific amplification by directly sequencing genomic DNA (Krehenwinkel et al, 2017).

The fact that all these issues (and attempting to find solutions) could be removed by using native DNA and Cas-9 Targeted sequencing is a clear incentive to its utilization. Its's limited use so far can be attributed to the poor (in comparison to Illumina) accuracy of nanopore base calling, the lower sequencing depth, and the still largely accessible, robust and heavily tested alternative options available using Illumina sequencing, such as TraDIS and Tn-Seq. It must also be noted that many of the Illumina based Transposon sequencing methods have readily available bioinformatic pipelines set up for data analysis, a real advantage for researchers with limited bioinformatic knowledge.

4.1.2 Expected observation of Cas-9 sequencing enrichment

The Cas-9 approach used for all three runs was the Single cut and read out, Cas-9 targeted nanopore sequencing has already been proven to enrich for mobile elements (McDonald et al, 2021) and single genes (Bruijnesteijn et al, 2021) however, in these instances an excision approach was used as both side of the region of interest (ROI) where known. For this project a single cut and read out method was necessary as only one side of the (ROI) is known (Figure 4.1). All input genomic DNA was dephosphorylated, after which a single double stranded cut was introduced within the insert (pGh9:IS*S1*) upstream to the ROI (in orange). Sequencing adapters can only then be ligated to the end of the DNA molecules that have been exposed, the reads that are available to sequence will in theory only be those generated post cleavage. We should therefore see an enrichment of sequences that contain the insert, those sequences reading out into whatever position of the genome the insert has transposed itself into.

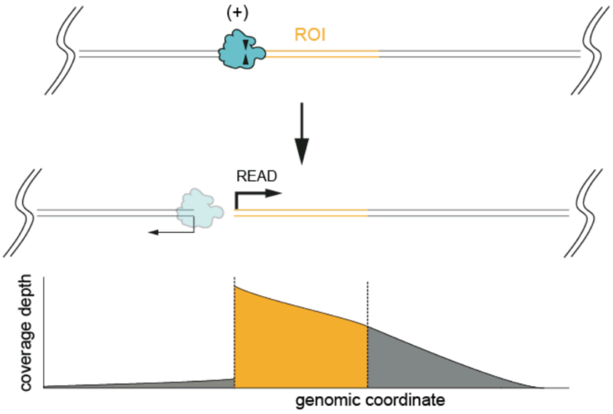


Figure 4.1 – A representation of the Cas 9 'Single Cut and Run' approach (from ONT, 2022)
A targeted double strand cleavage event is introduced upstream to the ROI. DNA containing insert sequence will be available for adapter ligation, followed by sequencing. Sequencing coverage should predominately contain these reads as shown by the coverage plot.

The size of the insert is around 4.6Kb (kilobases), we would therefore hope to see an enriched collection of DNA reads equal to and greater than this size. The read quantity would naturally decrease as the read lengths increased.

4.1.3 Previous Cas-9 Targeted sequencing strategies

The objective of a previous Cas-9 project was to successfully run the Cas-9 library preparation and sequencing method, and assess the results bioinformatically. Two Cas-9 runs were undertaken, using different crRNAs, one designed to cleave within the transposon itself, the other to cleave within the plasmid. Both experiments showed an enrichment for the insertion plasmid, however, the majority of reads generated were the insert sequence of 4.6Kb, very few reads were detected that read from insert into the chromosome (Figure 4.2).

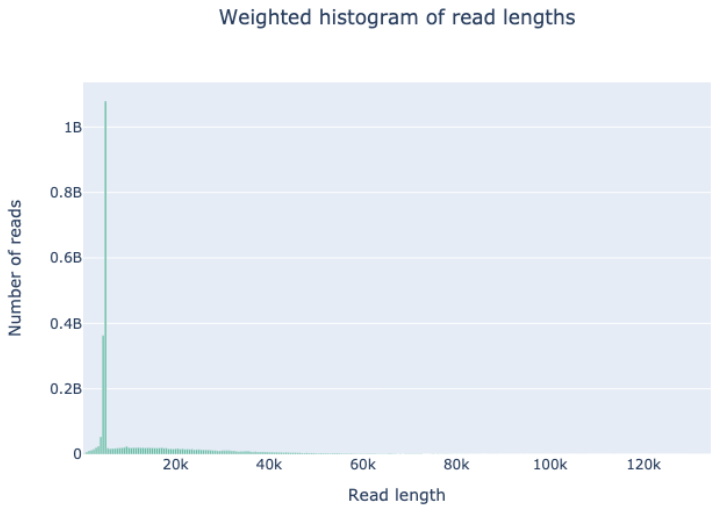


Figure 4.2 – Read length distribution of project previous Cas-9 project (from NanoPlot v1.38.1)
Histogram showing the number of sequencing reads (y axis) plotted against read length (x axis) The sharp peak early in the graph represents the amount of reads that were solely the insert (pGh9:IS*S1*) at ~ 4.6Kb.

The N50 of the run was (post base calling) estimated to be 4.59Kb. Looking at the run metrics and graphical plots generated by the GridION's onboard software, the enrichment did work with over 75% of the sequences mapping to the plasmid sequence. This showed that the crRNA design was very efficient, with excellent enrichment. However, bioinformatic analysis identified that 69.5% of sequencing reads were the insert plasmid pGh9:IS*S1.*

The sequencing results identified free plasmid (free copies of plasmid pGh9:IS*S1)* in the extracted input DNA sample, this was unexpected as any free plasmid should have been removed during the bacterial culture using a temperature shift from 28°C to 37.5°C during growth (Maguin et al, 1996), this is a fundamental component of the transposon mutagenesis system used to generate the insertion libraries. However, no short DNA fragments were identified when the extracted DNA was run on the Agilent TapeStation before library preparation. This could point towards the plasmid issue occurring during the Cas-9 library preparation. It has been recorded that during transposition using the IS*S1* insertion sequence, the integration of multiple tandem copies of the plasmid is a frequent occurrence (Thibessard et al, 2002). This might explain the result of the previous Cas-9 sequencing run, the crRNA was designed to enforce a double-stranded cleavage within every plasmid, if we have multiple tandem copies inserted within the genome, this could result is many single copies of pGh9:IS*S1* being generated, each plasmid sequence would be then be adapter ligated and sequenced. Problems have previous been recorded using IS*S1* for transposon mutagenesis due to some *Streptococcus* species containing endogenous IS*S1* elements that will recombine with the inserted ISS1 (Nillson et al, 2014). The recombination of in situ chromosomal IS*S1* and plasmid IS*S1* cannot be ruled out.

4.1.4 Optimisation of the Cas-9 Targeted Sequencing run

During the sequencing loading step for ds1093_3, a problem was encountered when introducing the final ds1093_3 library onto the MinION flow cell. A blockage was observed in the SpotON loading port and initially the library would not enter the flowcell using droplet pipetting. To resolve this issue the library was directly pipetted onto the flowcell, an action that is not recommended as it could introduce bubbles into the flow cell and air bubbles can irreversibly damage pores within the flow cell leading to a decrease in sequencing output. As the issue was due to a fault with the flowcell, it was replaced, and we were given the unexpected opportunity to repeat the run. This opportunity presented the option of adjusting some of the protocol steps to try and increase the amount of sequencing output.

Two modifications to the existing method were introduced, in both cases these changes were introduced to try and increase the sequencing output. Firstly, in an attempt to increase the input volume to the library preparation protocol, 60µl of the input DNA underwent a 1X AMpure purification step, and was eluted in 26µl of molecular grade water. This step was used to concentrate the DNA, increasing the amount of staring material fed into the library preparation protocol, more input material should theoretically lead to more end point library. Secondly, in an attempt to increase adapter efficiency, the adapter ligation time was doubled from 20 minutes to 40 minutes, as before this would hopefully lead to the generation of more sequencing library product. Another possible alteration might have been to increase the amount of Blunt/TA ligase added to the reaction mixture (Zascavage et al, 2019), an increase in incubation time was chosen as the cheaper option. The resulting end point library concentration was 59.4ng/µl, an improvement on the previous library preparation which had generated 37.6ng/µl. This allowed a higher library loading volume to be pipetted onto the flowcell, 712ng (the previous loading being 451ng). The read length distributions of the run (Figure 4.3) are similar to ds1093_3 however, the sequencing output had increased threefold with over 1.5 gigabases of data generated.

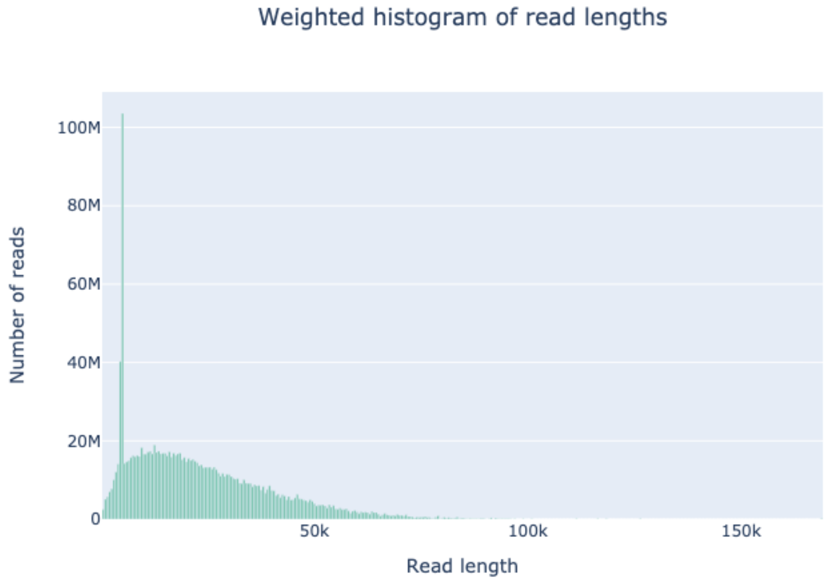


Figure 4.3 – Read length distribution of ds1093_3_Rpt. From NanoPlot (v1.38.1)
Histogram shows a visible free plasmid peak at around 4.6Kb, followed by a collection of longer reads. The free plasmid reads attribute 14.9% of the sequencing runs total read count.

The changes made to the original protocol have optimised the run conditions, and even though we increased the amount of sequencing data, the incidence of free plasmid did not increase, with 25,329 reads out of 169,432 being pGh9:IS*S1*. It was decided not to increase the default cleavage time (15 minutes), due to the risk of increasing any off-target cleavage events. In this instance 43,721 reads out of a total of 126,497 (post filtering using the manual bioinformatic pipeline) contain sequence reading from insert into the chromosome (34.6%).

4.2 Identification of Transposon Insertions using Manual Alignment

4.2.1 Quality Control of Nanopore Sequencing Data

A manual bioinformatic pipeline was generated to enable the analysis and interpretation of sequencing data generated using the Cas-9 Targeted approach. Unlike the PIMMS pre-existing bioinformatic pipeline, a read quality control process was used to assess the per base sequencing accuracy prior to any downstream bioinformatic processes and to remove poor quality reads from the dataset. This was necessary as the Nanopore reads once uploaded to the PIMMS pipeline are not subjected to any quality filtering. Sequencing reads were filtered based on their scored base calling accuracy, only reads with base calling accuracy scores greater than or equal to a Q9 Phred score (Phred scores are used to represent how confident we are in the assignment of each base call by a sequencer, a Q9 score indicates a confidence of 87.5%) were retained. Poor base calling accuracy is a common problem with nanopore sequencing, with a comparatively high base calling error rate compared to that of short read sequencing (Rang et al, 2018).

Low quality base calling accuracy can lead to a significant portion of sequencing reads being unusable for downstream data analysis, and more sequencing runs might have to be performed to ensure the desired depth of sequencing data required is attained. All three nanopore sequencing runs confirmed this with both Cas-9 runs losing 20.27% (ds1093_3) and 22.07% (ds1093_3_Rpt) of their data. This number was slightly higher for ds1093_1 with 27.04% of sequencing reads lost post filtering, and it would appear looking at all three runs that the more data generated, the greater the proportion of reads with a quality score lower than Q9.

In contrast to Illumina sequencing, the quality of nanopore sequencing is the same throughout a run and does not drop. As a run progresses the amount of sequence data generated will steadily reduce, as pores become blocked and as the amount of sequencing library decreases. The amount of data returned from a nanopore run is mostly limited by the ability to generate high molecular weight DNA (Amarasinghe et al, 2020). Illumina short read sequencing uses sequencing by synthesis (SBS) technology, which is heavily reliant on amplification and enzymatic processes, the quality of Illumina data will decrease as the reagents it uses become depleted.

4.2.2 Mapping Long Reads

The general purpose alignment program minimap2 was developed to facilitate the mapping of long reads generated using either Nanopore or PacBio sequencing. Existing short read aligner programs such as BWA and Bowtie2 are deemed unable or inefficient for the processing of large genomic contigs (Li, 2018). The continual introduction of dedicated analysis tools such as minimap2 that contain newly developed base calling algorithms designed to deal with long-read data has been essential in overcoming the early obstacles long read data analysis presented (Amarasinghe et al, 2020).

For both Cas-9 datasets minimap2 was used to align fastq sequencing reads to the insert and reference sequence. To optimise settings for the input of nanopore reads, an -ax map-ont flag was used to indicate that the input data would be nanopore long read data. This flag also ensured that the alignment program used ordinary minimizers as seeds, as alternative minimizers such as HPC are proven to be detrimental when aligning nanopore reads, but work well for PacBio data.
Alignment output files were generated as paf (Pairwise Mapping Format). This file type was chosen as it could be easily manipulated downstream using programming languages such as python or R.

4.2.3 PAF Filtering

Either Python or R could have been used to upload the paf files, both containing library's that have been written to achieve this goal. The choice to use R was not just user preference, but also because the library used to upload the paf files, pafr (v0.0.2) contained functions that would upload the paf file in a format that ensured it would subsequently behave almost exactly like a base R data.frame. This was useful due to a data frame object being a simple generic tabular format that can be manipulated using other R library's. The pafr library also permits the user to print a summary of each paf file, allowing the impact of each filtering step to be recorded (Appendix 2). Each summary prints the total of aligned genome sequence, plus the number of query and target sequences. Any extra tags (columns) separate from the original 12 are also recorded.

The pafr library contained two functions that were used to filter the data for low quality and secondary alignments. It was important to remove any non-ideal alignments from the data sets before proceeding any further, to improve the overall confidence in the results. The minimap2 aligner uses the tp tag (column) to identify the type of alignment, the function filter_secondary_alignments removed any non-primary alignments from each paf file (both reference and insert). This step, when used for both Cas-9 runs, removed a small proportion of reference reads, ds1093_3 being 7.8% (23.364 reads to 21,539) and ds1093_3_Rpt 8.3% (106,497 reads to 97,646), in contrast the removal of secondary alignments to the insert sequence extracted 83.1% of all reads for ds1093_3 (222,330 to 37,555) and 83.3% for ds1093_3_Rpt (1,047,876 to 174,646). Secondary alignments are defined as possible alignments between query and target sequence, but not the best possible alignments.

More strict parameters could have been used when using minimap2 to create the paf files, the aligner does have a --secondary=yes/no flag that would have filtered out the secondary alignments in the output paf files. The paf files were created using mostly minimap2 default settings and the large collection of secondary alignments to the insert could potentially have been further reduced by adjusting these setting, one such option, Gap opening penalties could have been minimized which would have created more strict alignments.

4.3 Utilizing Nanopore Sequencing for the PIMMS pipeline

4.3.1 Input Data Quality

The first step of the PIMMs pipeline, introduced in 2015 (Blanchard et al, 2015) is 'mapping' which involves the identification of transposon motifs, after which reads identified as having contained these terminal motifs are aligned to a reference sequence. For Illumina sequences, specific conditions have to be adhered to, to ensure confident high quality mapping. Sequencing read thresholds are set to a Q score of ≥ 30 (99.9% confidence that a base call is correct). This score is reached regularly using Illumina sequencing, and therefore doesn't restrict the amount of data uploaded to the pipeline. These strict parameters are implemented to increase our confidence that each read alignment is unambiguous.

If we applied the same strict parameters to nanopore data, it's likely that we wouldn't have any reads to upload to the PIMMS bioinformatic pipeline. The nanopore reads generated using both the Standard and Cas-9 library preparations gave average Q scores of 10.9 and 12.5 respectively (pre filtering). Interestingly the percentage of Cas-9 reads that gave Q scores over Q15 are a magnitude of 3-4 times higher than that for the standard nanopore reads, this could possibly be explained by the large volume of poor quality short read sequences generated by the PIMMS wet lab pipeline.

As previously mentioned, only Nanopore reads with a Q score greater than Q9 were uploaded to the PIMMS bioinformatic pipeline. The strength in nanopore sequencing is in the length of the reads generated, comparison of mean and median read lengths for all three sequencing runs (Table 4.1) give a stark contrast, the average mean read length of the two Cas-9 sequencing runs is 11,420.5, compared to the standard nanopore run which is 584. The sequencing of native DNA without any mechanical shearing has unsurprisingly produced significantly higher average read lengths.

Sequencing run ID	ds1093_1	ds1093_3	ds1093_3_Rpt
Run Method	Standard ONT	Cas-9	Cas-9
Mean read length	584	12,638.60	10,202.40
Mean read quality	12.1	13.3	12.8
Median read length	488	5,952.00	4,843.50
Median read quality	12	13.4	12.8
Number of reads	13,339,027.00	27,367.00	132,030.00
Read length N50	696	25,540.00	19,170.00
STDEV read length	369	14,520.10	11,301.90
Total bases	7,789,946,176.00	345,881,532.00	1,347,027,274.00

Table 4.1 – Overview of run metrics from all three nanopore sequencing runs. Taken from NanoPlot (v1.38.1) html report files.
These statistics are recorded post Q9 filtering. Both Cas-9 sequencing runs (ds1093_3 and ds1093_3_Rpt) record higher mean and median read lengths compared to the standard Nanopore sequencing run (ds1093_1) that used short amplified DNA as input generated using the PIMMS wet lab pipeline. Mean and median read quality is similar for all three sequencing runs, indicating that an increase in read length has not negatively affected read quality scores.

The base calling accuracy of nanopore reads uploaded to the PIMMs pipeline will (currently) be of inferior quality compared to Illumina short read sequences, for Cas-9 targeted sequencing an allowance for this discrepancy must be offset against the clear advantages achieved by using long reads with the pipeline.

4.3.2 Illumina and Nanopore PIMMS output using Inverse PCR

Looking at the genome scatterplots generated using the PIMMS dashboard, the Nanopore coverage of insertions across the genome is comparable to Illumina, and when looking at the NIM and NRM distributions across the *S .agalactiae* genome, we see location similarities at positions where no insertions have been detected, with both Illumina and nanopore output in agreement. Illumina PIMMS output identified 238 genes with no insertion events, nanopore PIMMS output also found no insertion events in the same 238 genes, a strong indicator that they might be essential genes.

It was perhaps expected that by using two different sequencing technologies we might see differences in output, and this was confirmed with transposon insertion identified in 405 genes using Nanopore sequencing which Illumina sequencing designated essential. In contrast Illumina output identified 79 insertions in genes that Nanopore sequencing did not detect.

Illumina sequencing is limited by a fixed read length, during an Illumina library preparation, short and long library fragments will be removed using a size selection, so only library products of a particular length are available for sequencing. This is necessary to ensure library fragment sizes are within the optimum range for the chosen sequencing kit and instrument. The Illumina data set for this study will have been generated using 250PE (paired end) reads. The mean read length of the nanopore data set generated for ds1093_1 was 584 bases, in contrast to the Illumina data set it will contain many longer reads, assessing the read quality control for this run (Results chapter 3.3.3) showed a large collection of reads over 10,000 bases in length. The longest read recorded was 105,103 bases with a Q value of 9.2. Perhaps the observed discrepancy in insertion data is due to the ability of nanopore sequencing to sequence all the DNA (irrespective of length) in a given sample.

However, it's worth remembering the strict quality control thresholds used for Illumina data, none of which are applied to the nanopore fastqs. In contrast to the manual alignment pipeline, the PIMMS output will include all secondary alignments. Including all these potential alignments in addition to the primary best fit alignments might explain the increase in insertion event identification using nanopore sequencing. It has already been discussed that nanopores lower base calling accuracy could lead to an increased incidence of errors when aligning reads to a reference, it's possible that some of the insertion events recorded might be due to variable mapping of reads to the reference. Perhaps installing nanopore data quality thresholds in the PIMMS pipeline would alleviate any doubts encountered regarding the insertions identified, as an exercise it would be interesting to see how filtering the nanopore reads based on quality would affect the number of insertion events detected, and how this might change the results of this study.

When visualising the distribution of NIM scores using the PIMMs compare dashboard, the scores for the Illumina output appear far greater than the Nanopore output scores. However, once the NIM scores were loaded into SPSS to generate summary statistics, all metrics pointed towards Nanopore sequencing output having higher NIM scores. Whilst uploading the Nanopore fastq files to the PIMMS bioinformatic pipeline, minimum and maximum read lengths are set as variable options. As the input was long read Nanopore data these options were set to a minimum read length of 100bp and maximum of 250,000bp. It is possible that due to the size of the reads uploaded to the pipeline we might (in some cases) be observing double counting of insertion positions where both ends of a single insertion are sequenced, one other option is that Nanopore data is struggling to uniquely map an insertion event to a specific location. Due to the variable quality of nanopore sequencing a unique insertion event sequenced multiple times might be recorded in separate positions using Nanopore sequencing, whereas only one position using Illumina sequencing data. This in theory could account for the larger NIM scores that were recorded by both standard and Cas-9 nanopore runs, and could potentially cause dispersions on the accuracy of each unique insertion event when dealing with nanopore data.

The boxplots of NIM and NRM scores indicated that both the standard nanopore and Cas-9 data sets contained numerous outliers which could potentially have an impact on any subsequent statistical analyses. A large collection of outliers has been proven to sometimes skew the results of any downstream hypothesis test (Aguinis et al, 2013).

4.3.3 Cas-9 Targeted Sequencing and the PIMMS Pipeline

The amount of data generated using Cas-9 sequencing was initially disappointing for sequencing run ds1093_3, however as already mentioned, once the protocol was modified for ds1093_3_Rpt, sequencing output increased threefold. The read lengths generated were very promising, and even though we observed free plasmid, when the fastq files were uploaded to the PIMMs bioinformatic pipeline is was clear that the data contained sequences that read from insert into the chromosome. The location of insertions mapped across the genome were comparable to the Illumina PIMMs output, but the coverage of those insertions was poor, and it is clear that more sequencing data is required to enable greater confidence when differentiating between control and test outputs.

The PIMMs pipeline output results for Cas-9 run ds1093_3_Rpt (BHI DNA) identified 50 insertion events in genes that the Illumina output had identified as essential. Taking into account the large discrepancy in sequencing input into the pipeline, this is an impressive result.

4.3.4 Comparison of Sequencing methods

Illumina sequencing offers accessible, rapid, and accurate data generation (De Maio et al, 2019), however short read sequencing used with PIMMS is heavily dependent on PCR during its wet lab pipeline (Blanchard et al, 2016), and this can introduce amplification bias to the results (Aird et al, 2011). Sequencing the PCR products generated by the current PIMMS wet lab pipeline using nanopore sequencing will generate data more representative of the input DNA, as all fragments present will be sequenced, irrespective of length. However, when sequencing inverse PCR fragments using nanopore sequencing, the amount of long reads generated will be hampered by the presence of overrepresented short reads which will have a higher efficiency for both adapter ligation and translocation through nanopores (Wang et al, 2021). Nanopore sequencing is still a more accessible technology (Lin et al, 2021) with smaller more affordable platforms available for low throughput projects.

Any method using the current PIMMS wet lab strategy will also be hampered by the numerous sample processing steps, the sample loses incurred during them, and the cost of the reagents required to run them. All of these extra processing steps (including any amplification) become unnecessary when using the Cas-9 Targeted sequencing method to sequence the extracted native DNA (Bruijnesteijn et al, 2019). Sequencing long reads using native DNA can improve mapping certainty (Amarasinghe et al, 2020) and for this project an increase in confidence when mapping an insertion event is key. A confidently mapped long read will always be preferable to a confidently mapped short read, with more of the genome covered in one continuous stretch. Long read sequencing of native DNA might also identify genomic regions that traditional short read sequencing methods struggle to resolve (Lin et al, 2021).

The Cas-9 Targeted library preparation method is a straightforward easy to follow protocol, however unlike the standard Nanopore library preparation method used for ds1093_1 it requires some additional customization by the user. This flexibility is a significant strength of the protocol, as you can choose single or multiple regions of interest (ROI) to enrich. The accuracy of long read sequencing data, although poor in comparison to Illumina, has continued to improve since its inception (Wang et al, 2021) and its increased application to multifarious genomic investigations is likely to continue.

4.4 Future Work

4.4.1 Further optimisation of the Cas-9 Targeted sequencing method

The Cas-9 targeted sequencing method from Oxford Nanopore Technologies (ONT) is still a relatively novel procedure, and contrasting from standard ONT library preparations, it's understandable and expected that each run methodology might require some optimisation. The repeated Cas-9 library preparation (ds1093_3_Rpt) demonstrated that increasing the volume of input, and doubling the adapter ligation time to 40 minutes led to a 307.32% increase in sequencing yield.

For both Cas-9 library preparations, the amount of input genomic DNA did not exceed 2µg, with 1.2µg and 1.97µg being used as input for ds1093_3 and ds1093_3_Rpt. The library preparation protocol guidelines allow for input between 1-10µg and it is reasonable to suggest that being at the lower end of the input scale has limited the amount of sequencing library generated, which is turn affected the amount of sequencing data each run generated. Producing a larger amount of input material can be quite problematic particularly if a researcher has limited sample availability. However, if it is possible to have input in the region of 5-10µg, it is likely that the end point library quantitation will be much higher, leading to more library being loaded onto the flowcell.

ONT recommend that the input genomic DNA is in nuclease free water (NFW), other buffers, in particular those containing EDTA might affect the efficiency of the library preparation. Moving forward, if a DNA extraction procedure has led to the DNA being eluted in solution other than NFW, it might be advisable to perform a 1X AMpure purification, not only will this allow a user to change the solution the DNA is eluted in, but they may wish to use this step to concentrate the DNA further buy eluting in a smaller volume. It is highly recommended to use wide bore tips when performing this purification step, to ensure the DNA does not become sheared and fragmented.

For ds1093_3_Rpt, doubling the adapter ligation time had a positive effect on the amount of library generated, increasing the concentration of end point sequencing library from 37.6ng/µl to 59.4ng/µl. It would make sense to explore this step in greater detail to try and maximise the ligation efficiency. A lengthier ligation time could be attempted however, longer ligation times could potentially increase the incidence of off-target read ligation, which would lead to more off target sequences being produced. For both Cas-9 sequencing runs, only one crRNA (guide RNA) was used. Primarily because we knew it had worked before and were therefore confident in its efficiency. However, another option would be to use two crRNAs, which is actually recommended by ONT. The reasoning being just in case one of the crRNAs yields an incomplete cleavage, overall the two probe approach might well provide a higher coverage of the ROI. Running the Cas-9 sequencing libraries on ONT's largest scale sequencing platform might be another option to consider, the PromethION flowcell's can generate terabases worth of data, and are compatible with Cas-9 libraries.

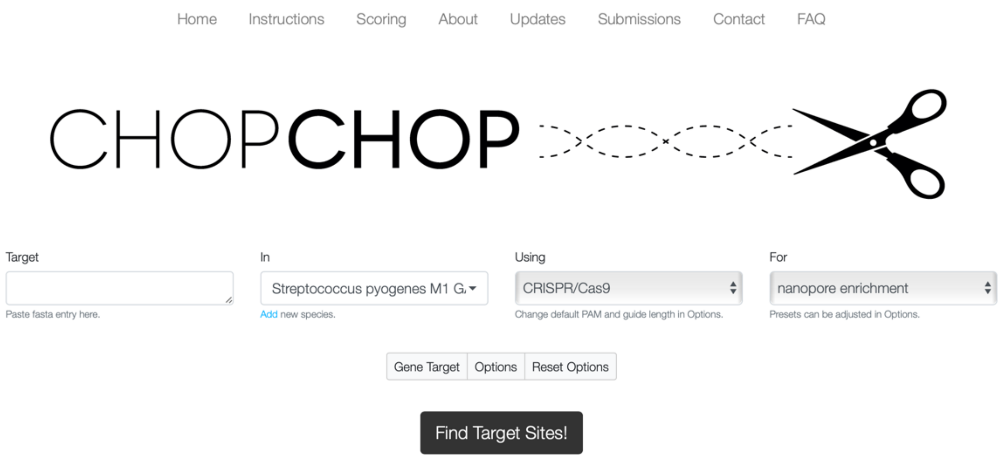
4.4.2 An alternative approach to crRNA design.

The design of crRNAs is one of the most important factors when designing a Cas-9 sequencing experiment. Each crRNA should be 20 bases long, and should be exactly three bases upstream of a protospacer adjacent motif (PAM), with sequence NGG**.** These parameters must be met for the RNP (ribonucleoprotein) complex to form and facilitate a double stranded cleave. The design of the crRNA was undertaken using a free source on-line crRNA design tool available on the IDT website:

https://eu.idtdna.com/site/order/designtool/index/CRISPR_CUSTOM

This software tool scans an input fasta sequence, and identifies locations for potential crRNA sites. It's output will give you a list of crRNA sequences, their position in the fasta sequence, and an on-target score. One issue found with this design tool is that fasta files of only 1000 bases can be uploaded at a time, this led to the 4.6Kb insert sequence being divided into 5 pieces before being uploaded to the design tool. This is obviously not ideal, not just from a time perspective, but more importantly by dividing the fasta sequence up, we may have lost potential crRNA candidate sites.

The ONT Cas-9 sequencing guide recommends using a webtool called Chop-Chop. (https://chopchop.cbu.uib.no). This software tool is more specifically tailored towards nanopore enrichment using the Cas-9 sequencing method, it gives the user the option of entering the whole fasta sequence and target species (Figure 4.4), after which you can directly select Cas-9 enrichment using nanopore data.



Target: (empty)
Paste fasta entry here.

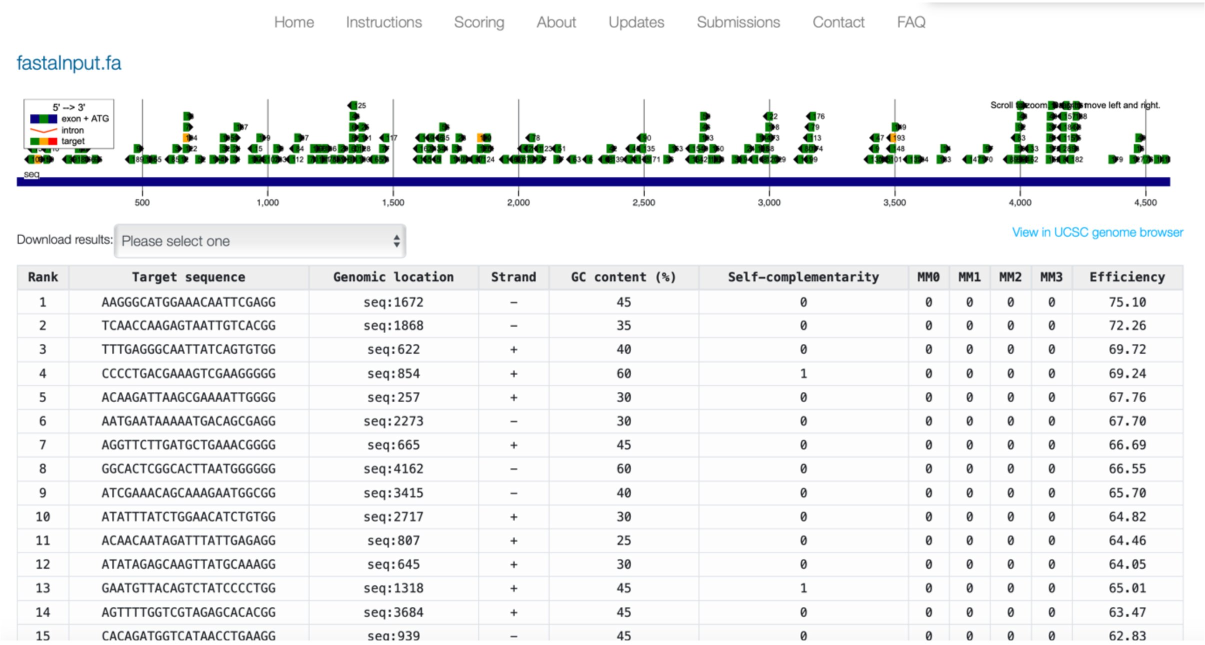
In: Streptococcus pyogenes M1 G.
Add new species.

Using: CRISPR/Cas9
Change default PAM and guide length in Options.

For: nanopore enrichment
Presets can be adjusted in Options.

Figure 4.4 – Chop-Chop home page
A website tool for crRNA design for ONT Cas-9 enrichment experiments. In this instance the tool is set to search a Streptococcus species sequence for potential crRNA candidate sites, specifically for nanopore enrichment.

Chop-Chop is a target prediction program which scans the whole FASTA sequence, and gives you a list of potential scored and ranked target sites (Figure 4.5). The output gives a visual display of the fasta sequence, and the location of each candidate crRNA. As with the IDT design tool, a list of crRNA's are generated alongside coordinates and efficiency scores.



Download results: Please select one

Rank	Target sequence	Genomic location	Strand	GC content (%)	Self-complementarity	MM0	MM1	MM2	MM3	Efficiency
1	AAGGGCATGGAAACAATTCGAGG	seq:1672	–	45	0	0	0	0	0	75.10
2	TCAACCAAGAGTAATTGTCACGG	seq:1868	–	35	0	0	0	0	0	72.26
3	TTTGAGGGCAATTATCAGTGTGG	seq:622	+	40	0	0	0	0	0	69.72
4	CCCCTGACGAAAGTCGAAGGGGG	seq:854	+	60	1	0	0	0	0	69.24
5	ACAAGATTAAGCGAAAATTGGGG	seq:257	+	30	0	0	0	0	0	67.76
6	AATGAATAAAAATGACAGCGAGG	seq:2273	–	30	0	0	0	0	0	67.70
7	AGGTTCTTGATGCTGAAACGGGG	seq:665	+	45	0	0	0	0	0	66.69
8	GGCACTCGGCACTTAATGGGGGG	seq:4162	–	60	0	0	0	0	0	66.55
9	ATCGAAACAGCAAAGAATGGCGG	seq:3415	–	40	0	0	0	0	0	65.70
10	ATATTTATCTGGAACATCTGTGG	seq:2717	+	30	0	0	0	0	0	64.82
11	ACAACAATAGATTTATTGAGAGG	seq:807	+	25	0	0	0	0	0	64.46
12	ATATAGAGCAAGTTATGCAAAGG	seq:645	+	30	0	0	0	0	0	64.05
13	GAATGTTACAGTCTATCCCCTGG	seq:1318	+	45	1	0	0	0	0	65.01
14	AGTTTTGGTCGTAGAGCACACGG	seq:3684	+	45	0	0	0	0	0	63.47
15	CACAGATGGTCATAACCTGAAGG	seq:939	–	45	0	0	0	0	0	62.83

Figure 4.5 – Output from Chop-Chop search tool using the pGh9:IS*S1* fasta sequence.
A list is generated of potential crRNA sites, with sequences, genomic locations and efficiency scores. Above list is a visual representation of the genomic position of each crRNA for pGh9:IS*S1*.

There is the potential here to use Chop-Chop to design crRNAs for future Cas-9 experiments.

4.4.3 Nanopore technological innovation

A new Guppy base caller has recently been made available on the ONT website (Guppy version 6, the most recent iteration being v6.2.1 released in July 2022). As base calling methods are continually being improved upon, an exercise moving forward might be to return to the raw fast5 files, and to manually base call them to fastq using the most recent Guppy version. This might lead to improved data quality, more data passing the Q9 threshold, and for our purpose the detection of more insertion events. Due to the processing power required for high accuracy base calling using Guppy, it would be advisable to perform this analysis using HPC (High Performance Computing). The GridION removes the need for extra IT expense as it has in-built compute to perform data processing for the user, the option to compress the resulting fastq files make transferring the output files to an external sever a fast efficient process.

Oxford Nanopore Technologies have also recently upgraded the nanopore used in their flow cells. The R9.4.1 nanopore giving way to the R10 series of nanopores. Typically, as a DNA molecule moves through the nanopore it passes a pinch-point which is a narrow hole in which the current of the molecule passing through is measured. The R10 nanopores will contain two pinch-points, therefore enabling two measurements, which should increase the overall base calling accuracy, and in particular might resolve the issues behind sequencing long homopolymer regions.

5. Conclusions

Significant enrichment of sequencing reads containing the insert pGh9:IS*S1* was observed when using the Cas-9 targeted library preparation protocol, followed by nanopore sequencing. For both Cas-9 runs, over 65% of all reads (post Q9 filtering) contained insert sequence. It is clear that with optimisation Cas-9 targeted sequencing could be a viable time saving alternative to inverse PCR, bypassing a significant portion of the processing steps used in the current PIMMS Seq web lab pipeline. Alongside this greater efficiency would be the removal of any amplification bias that PCR introduces during sample processing and library preparation steps. Long read sequencing of native DNA identified extra insertion events missed by Illumina sequencing, even when the sequencing was at a much lesser depth. A more detailed investigation of the Cas-9 sequencing method would be recommended, looking into the design of alternative crRNAs and the potential use of two to enhance the insert enrichment.

For PIMMs processed DNA, generated using inverse PCR, the Illumina and Nanopore PIMMs pipeline output is comparable. There are some explainable differences, but the coverage of insertions across the genome of *S .algalactiae* is notably similar. Nanopore sequencing however does offer some unique advantages over Illumina sequencing, such as longer reads which will generate more confident and conclusive alignments, ultimately improving the identification of essential and non-essential genes using the PIMMS bioinformatic pipeline. However, the base calling quality of Nanopore sequencing is still poor when compared to short read sequencing, and must be taken into consideration when comparing the PIMMS output for both methods.

Bibliography

Adewale BA. Will long-read sequencing technologies replace short-read sequencing technologies in the next 10 years? Afr J Lab Med. 2020 Nov 26;9(1):1340. doi: 10.4102/ajlm.v9i1.1340. PMID: 33354530; PMCID: PMC7736650.

Aguinis H, Gottfredson RK, Joo H. Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. Organizational Research Methods. 2013;16(2):270-301. doi: 10.1177/1094428112470848.

Aird, D., Ross, M.G., Chen, WS. et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol 12, R18 (2011). https://doi.org/10.1186/gb-2011-12-2-r18

Alkam D, Wongsurawat T, Nookaew I, Richardson AR, Ussery D, Smeltzer MS, Jenjaroenpun P. Is amplification bias consequential in transposon sequencing (TnSeq) assays? A case study with a Staphylococcus aureus TnSeq library subjected to PCR-based and amplification-free enrichment methods. Microb Genom. 2021 Oct;7(10):000655. doi: 10.1099/mgen.0.000655. PMID: 34596508; PMCID: PMC8627206.

Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biology. 2020;21(1):30. doi: 10.1186/s13059-020-1935-5.

Arteche-López, A., Ávila-Fernández, A., Romero, R. et al. Sanger sequencing is no longer always necessary based on a single-center validation of 1109 NGS variants in 825 clinical exomes. Sci Rep 11, 5697 (2021). https://doi.org/10.1038/s41598-021-85182-w

Barquist L, Mayho M, Cummins C, Cain AK, Boinett CJ, Page AJ, Langridge GC, Quail MA, Keane JA, Parkhill J. The TraDIS toolkit: sequencing and analysis for dense transposon mutant libraries. Bioinformatics. 2016 Apr 1;32(7):1109-11. doi: 10.1093/bioinformatics/btw022. Epub 2016 Jan 21. PMID: 26794317; PMCID: PMC4896371.

Blanchard AM, Egan SA, Emes RD, Warry A, Leigh JA. PIMMS (Pragmatic Insertional Mutation Mapping System) Laboratory Methodology a Readily Accessible Tool for Identification of Essential Genes in Streptococcus. Front Microbiol. 2016 Oct 25;7:1645. doi: 10.3389/fmicb.2016.01645. PMID: 27826289; PMCID: PMC5078762.

Blanchard AM, Leigh JA, Egan SA, Emes RD. Transposon insertion mapping with PIMMS - Pragmatic Insertional Mutation Mapping System. Front Genet. 2015 Apr 9;6:139. doi: 10.3389/fgene.2015.00139. PMID: 25914720; PMCID: PMC4391243.

Brown CL, Keenum IM, Dai D, Zhang L, Vikesland PJ, Pruden A. Critical evaluation of short, long, and hybrid assembly for contextual analysis of antibiotic resistance genes in complex environmental metagenomes. Scientific Reports. 2021;11(1):3753. doi: 10.1038/s41598-021-83081-8.

Bruijnesteijn J, van der Wiel M, de Groot NG, Bontrop RE. Rapid Characterization of Complex Killer Cell Immunoglobulin-Like Receptor (KIR) Regions Using Cas9 Enrichment and Nanopore Sequencing. Frontiers in Immunology. 2021;12. doi: 10.3389/fimmu.2021.722181.

Bruijnesteijn J, van der Wiel M, de Groot NG, Bontrop RE. Rapid characterization of complex genomic regions using Cas9 enrichment and Nanopore sequencing. bioRxiv. 2021:2021.03.11.434935. doi: 10.1101/2021.03.11.434935.

De Maio N, Shaw LP, Hubbard A, George S, Sanderson ND, Swann J, Wick R, AbuOun M, Stubberfield E, Hoosdally SJ, Crook DW, Peto TEA, Sheppard AE, Bailey MJ, Read DS, Anjum MF, Walker AS, Stoesser N, On Behalf Of The Rehab Consortium. Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. Microb Genom. 2019 Sep;5(9):e000294. doi: 10.1099/mgen.0.000294. Epub 2019 Aug 30. PMID: 31483244; PMCID: PMC6807382.

Ewing AD, Smits N, Sanchez-Luque FJ, Faivre J, Brennan PM, Richardson SR, et al. Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling. Molecular Cell. 2020;80(5):915-28.e5. doi: https://doi.org/10.1016/j.molcel.2020.10.024.

Gauthier J, Vincent AT, Charette SJ, Derome N. A brief history of bioinformatics. Brief Bioinform. 2019 Nov 27;20(6):1981-1996. doi: 10.1093/bib/bby063. PMID: 30084940.

Goodier, J.L. Restricting retrotransposons: a review. Mobile DNA 7, 16 (2016).

Hamer L, DeZwaan TM, Montenegro-Chamorro MV, Frank SA, Hamer JE. Recent advances in large-scale transposon mutagenesis. Current Opinion in Chemical Biology. 2001;5(1):67-73. doi: https://doi.org/10.1016/S1367-5931(00)00162-9.

Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016 Jan;107(1):1-8. doi: 10.1016/j.ygeno.2015.11.003. Epub 2015 Nov 10. PMID: 26554401; PMCID: PMC4727787.

Hebing Chen, Zhuo Zhang, Shuai Jiang, Ruijiang Li, Wanying Li, Chenghui Zhao, Hao Hong, Xin Huang, Hao Li, Xiaochen Bo, New insights on human essential genes based on integrated analysis and the construction of the HEGIAP web-based platform, Briefings in Bioinformatics, Volume 21, Issue 4, July 2020, Pages 1397–1410

Heng Li, Minimap2: pairwise alignment for nucleotide sequences, *Bioinformatics*, Volume 34, Issue 18, 15 September 2018, Pages 3094–3100, https://doi.org/10.1093/bioinformatics/bty191

Hill AW, Leigh JA. DNA fingerprinting of Streptococcus uberis: a useful tool for epidemiology of bovine mastitis. Epidemiol Infect. 1989 Aug;103(1):165-71. doi: 10.1017/s0950268800030466. PMID: 2776850; PMCID: PMC2249475.

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 409, 860–921 (2001). https://doi.org/10.1038/35057062
ISSN 0968-0004.

Jain, M., Koren, S., Miga, K. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol 36, 338–345 (2018).

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012 Aug 17;337(6096):816-21. doi: 10.1126/science.1225829. Epub 2012 Jun 28. PMID: 22745249; PMCID: PMC6286148.

Krehenwinkel, H., Wolf, M., Lim, J.Y. et al. Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. Sci Rep 7, 17668 (2017). https://doi.org/10.1038/s41598-017-17333-x

Langridge GC, Phan MD, Turner DJ, Perkins TT, Parts L, Haase J, Charles I, Maskell DJ, Peters SE, Dougan G, Wain J, Parkhill J, Turner AK. Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. Genome Res. 2009 Dec;19(12):2308-16. doi: 10.1101/gr.097097.109. Epub 2009 Oct 13. PMID: 19826075; PMCID: PMC2792183.

Leipzig, Jeremy. (2016). A review of bioinformatic pipeline frameworks. Briefings in Bioinformatics. 18. bbw020. 10.1093/bib/bbw020.

Lin B, Hui J, Mao H. Nanopore Technology and Its Applications in Gene Sequencing. Biosensors. 2021;11(7):214. PubMed PMID: doi:10.3390/bios11070214.

Maguin E, Prévost H, Ehrlich SD, Gruss A. Efficient insertional mutagenesis in lactococci and other gram-positive bacteria. J Bacteriol. 1996 Feb;178(3):931-5. doi: 10.1128/jb.178.3.931-935.1996. PMID: 8550537; PMCID: PMC177749.

Maguin, Emmanuelle & Prévost, Hervé & Ehrlich, S & Gruss, Alexandra. (1996). Efficient insertional mutagenesis in Lactococci and other Gram-positive bacteria. Journal of bacteriology. 178. 931-5. 10.1128/jb.178.3.931-935.1996.

McDonald, T.L., Zhou, W., Castro, C.P. *et al.* Cas9 targeted enrichment of mobile elements using nanopore sequencing. *Nat Commun* **12**, 3586 (2021). https://doi.org/10.1038/s41467-021-23918-y

Nilsson M, Christiansen N, Høiby N, Twetman S, Givskov M, Tolker-Nielsen T. A mariner transposon vector adapted for mutagenesis in oral streptococci. Microbiologyopen. 2014 Jun;3(3):333-40. doi: 10.1002/mbo3.171. Epub 2014 Apr 21. PMID: 24753509; PMCID: PMC4082707.

Pennisi, E. Jumping genes hop into the evolutionary limelight. Science 317, 894–895 (2007)
Pray, L. (2008) Transposons: The jumping genes. Nature Education 1(1):204

Raabe VN, Shane AL. Group B Streptococcus(Streptococcus agalactiae). Microbiol Spectr. 2019 Mar;7(2):10.1128/microbiolspec.GPP3-0007-2018. doi: 10.1128/microbiolspec.GPP3-0007-2018. PMID: 30900541; PMCID: PMC6432937.

Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biology. 2018;19(1):90. doi: 10.1186/s13059-018-1462-9.

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977 Dec;74(12):5463-7. doi: 10.1073/pnas.74.12.5463. PMID: 271968; PMCID: PMC431765.

Schmutz J, Wheeler J, Grimwood J, Dickson M, Yang J, Caoile C, et al. Quality assessment of the human genome sequence. Nature. 2004;429(6990):365-8. doi: 10.1038/nature02390.

Soraya de Chadarevian, Protein sequencing and the making of molecular genetics,
Thibessard A, Fernandez A, Gintz B, Decaris B, Leblond-Bourget N. Transposition of pGh9:ISS1 is random and efficient in Streptococcus thermophilus CNRZ368. Can J Microbiol. 2002 May;48(5):473-8. doi: 10.1139/w02-038. PMID: 12109889.

Thibessard, Annabelle & Fernandez, Annabelle & Gintz, Brigitte & Decaris, Bernard & Leblond-Bourget, Nathalie. (2002). Transposition of pGh9 : ISS1 is random and efficient in Streptococcus thermophilus CNRZ368. Canadian journal of microbiology. 48. 473-8. 10.1139/w02-038. Trends in Biochemical Sciences, Volume 24, Issue 5, 1999, Pages 203-206.

van Opijnen T, Camilli A. Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. Nat Rev Microbiol. 2013 Jul;11(7):435-42. doi: 10.1038/nrmicro3033. Epub 2013 May 28. PMID: 23712350; PMCID: PMC3842022.

Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Nature Biotechnology. 2021;39(11):1348-65. doi: 10.1038/s41587-021-01108-x.

Watson JD, Crick FHC. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature. 1953;171(4356):737-8. doi: 10.1038/171737a0.

Wratten, L., Wilm, A. & Göke, J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. Nat Methods 18, 1161–1168 (2021). https://doi.org/10.1038/s41592-021-01254-9

Yasir M, Turner AK, Lott M, Rudder S, Baker D, Bastkowski S, et al. Long-read sequencing for identification of insertion sites in large transposon mutant libraries. Scientific Reports. 2022;12(1):3546. doi: 10.1038/s41598-022-07557-x.

Zascavage, R.R., Thorson, K. and Planz, J.V. (2019), Nanopore sequencing: An enrichment-free alternative to mitochondrial DNA sequencing. ELECTROPHORESIS, 40: 272-280. https://doi.org/10.1002/elps.201800083

Zhang Z, Ren Q. Why are essential genes essential? - The essentiality of Saccharomycesgenes. Microb Cell. 2015 Jul 25;2(8):280-287. doi: 10.15698/mic2015.08.218. PMID: 28357303; PMCID: PMC5349100.

Zhao, L., Anderson, M.T., Wu, W. et al. TnseqDiff: identification of conditionally essential genes in transposon sequencing studies. BMC Bioinformatics 18, 326 (2017).

Appendix 1

TERMINAL COMMANDS USED FOR MANUAL BIOINFORMATIC PIPELINE

Read QC using NanoPlot (v1.38.1)

```
NanoPlot --summary <sequencing_summary_file location> -o <output location>
```

Concatenate fastq reads

```
cat *.fastq > <all_guppy.fastq>
```

Trim Adapter sequences using Porechop (v0.2.4)

```
porechop -i <input_reads.fastq.gz> -o <output_reads.fastq.gz>
```

Align reads to insert and reference fasta files using Minimap2 (v2.24)

```
minimap2 -x map-ont <location of fasta file> <location of fastq file> > <output.paf>
```

COMMANDS USED IN R STUDIO

```
# Load in pafr library (v0.0.2)

library(pafr)

# Upload alignment paf files

Ref_Alignment = read_paf(“<location of reference paf file>”)
Insert_Alignment = read_paf(“<location of insert paf file>")

# summary of edited paf files

Ref_Alignment
Insert_Alignment

# Filter both files to exclude secondary alignments

Ref_Prim = filter_secondary_alignments(Ref_Alignment)
Insert_Prim = filter_secondary_alignments(Insert_Alignment)

# summary of edited paf files

PIMMS_Ref_Prim
PIMMS_Insert_Prim

# Filter sequences, retaining reads with mapping quality scores above 40
```

```
Quality_Ref = subset(Ref_Prim, mapq > 40)
Quality_Insert = subset(Insert_Prim, mapq > 40)

# Summary of edited paf files

Quality_Ref
Quality_Insert

# Load in dlpyr library (v1.8.6)

library(dplyr)

# Remove duplicates read IDs from Col 1 (qname)

Ref_No_Dups = distinct(Quality_Ref, qname)
Insert_No_Dups = distinct(Quality_Insert, qname)

# Summary of edited paf files

nrow(Ref_No_Dups)
nrow(Insert_No_Dups)

#join the two data frames together using inner join function

Aligned_Ref_And_Insert = merge(x=Ref_No_Dups,y=Insert_No_Dups,by="qname")

# Record Read IDs identified in both reference and insert

nrow(Aligned_Ref_And_Insert)
```

TERMINAL COMMANDS USED TO RUN PIMMS PIPELINE

Concatenate all fastq files

```
cat *.fastq > <all_guppy.fastq>
```

Activate PIMMS2

```
$ conda activate pimms2
```

Run Find flank module

```
$ python ./pimms2.py find_flank -c pimms2.config --fasta <name>.fasta --mapper minimap2 --min 100 --max 250000 --nano --in_dir <location of fastq files> --out_dir <location of output files> --label <test>
```

Run Bam Extract module

```
$ python /pimms2.py bam_extract -c pimms2.config –bam <location of bam file from find flank module> --nano –label<test> --gff <location of GFF file> --gff_force
```

Appendix 2

ds1093_3 – Example of R studio file filtering.

Filtering of 'Pairwise mapping format' files for ds1093_3

Upload pafr libary (v0.0.2)

```
library(pafr)

## Loading required package: ggplot2

## Registered S3 methods overwritten by 'tibble':
##   method     from
##   format.tbl pillar
##   print.tbl  pillar
```

Upload alignment paf files

```
Cas9_Ref_Alignment = read_paf("/Users/matthewcarlile/Documents/Apprenticeship_Docs/New_Project_Data/ds1093_3/ds1093_Minimap_Results/ds1093_3_Ref_Alignment.paf")
Cas9_Insert_Alignment = read_paf("/users/matthewcarlile/Documents/Apprenticeship_Docs/New_Project_Data/ds1093_3/ds1093_Minimap_Results/ds1093_3_Insert_Alignment.paf")
```

summary of each paf file

```
Cas9_Ref_Alignment

## pafr object with 23624 alignments (282.4Mb)
##  18815 query seqs
##  5 target seqs
##  6 tags: tp, cm, s1, s2, dv, rl

Cas9_Insert_Alignment

## pafr object with 225330 alignments (351.3Mb)
##  19453 query seqs
##  1 target seqs
##  6 tags: tp, cm, s1, s2, dv, rl
```

Filter both files to exclude secondary alignments

```
Cas9_Ref_Prim = filter_secondary_alignments(Cas9_Ref_Alignment)
Cas9_Insert_Prim = filter_secondary_alignments(Cas9_Insert_Alignment)
```

summarize edited paf files

```
Cas9_Ref_Prim

## pafr object with 21539 alignments (273Mb)
##  18815 query seqs
##  4 target seqs
##  6 tags: tp, cm, s1, s2, dv, rl

Cas9_Insert_Prim
```

```
## pafr object with 37555 alignments (58.5Mb)
##  19453 query seqs
##  1 target seqs
##  6 tags: tp, cm, s1, s2, dv, rl
```

Filter sequences retaining reads with mapping quality above 40

```
Cas9_Quality_Ref = subset(Cas9_Ref_Prim, mapq > 40)
Cas9_Quality_Insert = subset(Cas9_Insert_Prim, mapq > 40)
```

Summarize edited paf files

```
Cas9_Quality_Ref

## pafr object with 20615 alignments (270.7Mb)
##  18356 query seqs
##  1 target seqs
##  6 tags: tp, cm, s1, s2, dv, rl

Cas9_Quality_Insert

## pafr object with 37265 alignments (58.5Mb)
##  19452 query seqs
##  1 target seqs
##  6 tags: tp, cm, s1, s2, dv, rl
```

Upload library dplyr (v1.8.6)

```
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Remove duplicate read ID from both alignment files

```
Ref_No_Dups = distinct(Cas9_Quality_Ref, qname)
Insert_No_Dups = distinct(Cas9_Quality_Insert, qname)
```

Summarize edited paf files

```
nrow(Ref_No_Dups)

## [1] 18356

nrow(Insert_No_Dups)

## [1] 19452
```

Join the two data frames together using inner join function.

```
Aligned_Ref_And_Insert = merge(x=Ref_No_Dups,y=Insert_No_Dups,by="qname")
```

Record Read IDs identified in both reference and insert

```
nrow(Aligned_Ref_And_Insert)

## [1] 11227
```

Appendix 3

Pimms2.config file

```
[general] # global settings (not in use)

[find_flank] # filter fastq files to find insertion site flanking sequence

 ## [default: illumina paired end]
 # nano = False

 fwdrev =_fastq.gz

 ## use Levenshtein distance (combined sub/insert/del score) OR set sub/insert/del separately
 ## lev overrides sub/insert/del (currently settings for nano are hardcoded)
 lev = 0
 sub = 0
 insert = 0
 del = 0

 ## directory containing input fastq files default=os.getcwd()
 # in_dir =

 ## result directory default=(['pimms2_' + time.strftime("%Y%m%d_%H%M%S")])
 # out_dir =

 ## min read length / clip reads to max (currently settings for nano are hardcoded)
 min = 25
 max = 50

 ## IS ends reference motifs
 motif1 = TCAGAAAACTTTGCAACAGAACC
 motif2 = GGTTCTGTTGCAAAGTTTAAAAA

 # fasta = suis.p17.fasta

 # cpus = 6

[bam_extract]
 # bam =

 # gff = suis.p17.gff3

 ## comma separated extra GFF3 annotation fields e.g 'note,translation'
 # gff_extra = 'note'
 min_depth = 3

 ## mismatch as fraction of bases (needs further testing use with caution)
 # mismatch = 0.05

 # force_gff = True
# noreps = True
```