THE UNIVERSITY *of* EDINBURGH
School of Biological Sciences

Dissertation Title

Using Nanopore Long-read Sequencing Data to Identify and Assemble PRRSV Quasispecies

Student Exam Number: B170490

In partial fulfilment of the requirement for the Degree of Master of Science in ***Biotechnology*** at the University of Edinburgh

2020 / 2021

Name of Dissertation Supervisor: Dr. Amanda Warr

Word Count*: XXX

**Word count should not include figure legends, tables, appendices and references*

Table of Contents

Abstract

Porcine Reproductive and Respiratory Syndrome Virus (PRRSV) has been spreading worldwide for over thirty years and has had a significant negative impact on the pig industry. However, genetic studies of PRRSV are extremely challenging due to the high mutation rate and the ease with which new strains can be generated by recombination which leads to populations of viruses, or “quasispecies” within a host. There is an urgent need for a method that can provide rapid identification and separation of virus quasispecies. In this study, I used Oxford Nanopore MinION long-read sequencing data to evaluate a novel pipeline for isolating quasispecies. The pipeline uses the principle of genome phasing but makes no assumptions about the number of genomes in the sample, to isolate the genomes of quasispecies and produce consensus genomes. Two almost full-length PRRSV genome simulated Nanopore sequencing datasets were successfully generated using bioinformatics tools to simulate PRRSV quasispecies. One dataset is a set of 10 viral genomes with even coverage, as a best-case scenario, and the other is the dataset with uneven coverage that is closer to a real scenario. The pipeline reliably isolated all quasispecies in the even coverage group, and for the uneven group also successfully detected species present in at least3% of the total population. In addition, I sequenced viruses amplified by two different polymerases for a dataset of authentic viruses and used the pipeline for quasispecies isolation, screening for the most suitable polymerase based on the isolation results. Overall, Nanopore long-read sequencing and the Untangle pipeline are proving to be promising methods for identifying and isolating PRRSV quasispecies and could be powerful tools for studying PRRSV genetics, providing more precise strategies for vaccine research and outbreak control.

1 Introduction

Pork is one of the world’s most consumed animal proteins, demand for pork increases with a country's wealth and as such pig farming plays an important role in feeding the globe [1]. It is necessary to increase meat production to meet rising demands of a growing global population. To meet the demand for pork, the pig industry has shifted from the traditional method of raising small herds to intensive rearing [1]. Although this dramatic change has increased the efficiency of pork production, it has also potentially encouraged the emergence of diseases. It is likely that these diseases could disrupt meat production, cause welfare problems for animals and even impact global economy.

1.1 Porcine Reproductive and Respiratory Syndrome

1.1.1 History of Mysterious Swine Disease

Porcine reproductive and respiratory syndrome (PRRS) is considered to be one of the world’s most economically significant and devastating diseases [2]. It can affect all stages of production, and perhaps most importantly, it may also dysregulate the immune response of pigs, thus it is even called the “AIDS” of pigs [3]. PRRS was first identified in North American pig farms in the 1980s, which was originally known as “mystery swine disease” due to the unknown aetiology [4]. Shortly after that, the disease spread rapidly around the world, including in Europe and Asia [5].

1.1.2 Symptoms

Pigs with symptomatic disease initially suffer from an acute disease involving anorexia, lethargy, pyrexia, respiratory distress, limb cyanosis, and often red-blue discolouration of the ears [5]. Even though there are some differences among pigs in the susceptibility, all ages, breeds, and sexes are susceptible, with sows and mid to late gestation fetuses being the most susceptible to this disease [3]. Pregnant sows typically fail to reproduce, including stillbirths, mummified or deformed fetuses [6]. Even if the sow can successfully give birth to piglets, the diseased piglets often will not survive [6]. Even though adult pigs may have no symptoms or very mild symptoms, pigs with the virus can still help spread the virus. Tests have shown that even wild ducks could also be infected with this disease and could be exposed to the outside world through their faeces [3]. Once pigs are infected with the disease, they will be continuously virulent, and contaminated pig farms may become the source of the disease. Highly pathogenic PRRS disease has been an explosive epidemic in the United States and China from the early twentieth century to the present, with high morbidity rates (50% to 100%) and mortality rates (20% to 100%) [5].

1.1.3 Economic Impact

PRRS has not only caused significant damage to the animal welfare of pigs but also had a major economic impact on the global pig industry. In 2005, it was estimated due to the disease in the US pig industry that the loss of production was as high as 560 million dollars per year [7]. In 2013, the research showed that the total loss of production in breeding herds in the United States was still growing at approximately 664 million dollars per year [8], with comparative losses in most other countries infected with the disease [9]. Thus, it can be seen that there is no time to lose in controlling PRRS, which continues to pose a threat to the pig industry even though combating and controlling the disease has taken years of global efforts.

1.2 Porcine Reproductive and Respiratory Syndrome Virus (PRRSV)

1.2.1 The Notorious PRRSV

The frequent recurrence of the PRRS epidemic in pig farms is one of the challenges of outbreak control [10], and 71% of pig farms in America are predicted and reported in 2019 to be re-infected with the disease in the next two years[11]. This is because PRRSV, the causative agent of PRRS, has genetically widespread variation, including mutations and recombination between different isolates. PRRSV is a single- and positive-stranded polyadenylated 15kb RNA virus, encoding a total of 8 open-read frames that can be translated into at least 8 structural proteins and 14 non-structural proteins [12]. PRRSV was discovered independently in Europe and the United States in 1991. Although they are the same virus, they belong to different species. Lelystad virus is the prototype strain of PRRSV-1 [4], while VR-2332 is the prototype strain of PRRSV-2 [13]. They only share about 60% nucleotide identity, which implies a high degree of evolutionary variation between the two species [14]. Current understanding of the full-genome sequencing results of PRRSV reveals the presence of considerable genetic variability, constituting a complex viral phylogenomic history [15]. Not only is there a significant genetic variation between the two species, but individual PRRSV isolates and subtypes in each species also exhibit considerable antigenic and clinical heterogeneity with up to 20% genetic variability at the nucleotide level [16].

PRRS recurrence becomes disastrous with the constant recombination of strains and the emergence of new strains, which recently appear to be more diverse and virulent than past variants. In 1995, PRRSV was first isolated in China under the name CH-la, and the PRRSV genome endemic to China has continued to mutate since then [17]. In 2006, there was a devastating outbreak of high mortality-related porcine hyperthermia in China, with a variant known as the highly pathogenic porcine reproductive and respiratory syndrome virus (HP-PRRSV) [17]. More than 2 million pigs were affected by the HP-PRRSV epidemic with approximately 400,000 deaths, including adult pigs [17]. Subsequently, it spread to other Southeast Asian countries [18]. In contrast, there were also serious outbreaks of the mutation-prone and recombinant NADC30-like PRRSV found in the USA [19]. Additionally, laboratory investigations have shown that multiple variants of PRRSV can co-exist in experimentally infected animals [20]. Recombination of each PRRSV genotype has been proven to be a potential for viral diversification [21]. Since 2013, several Chinese provinces have reported different NADC30 PRRSVs due to the recombination of NADC30 strains with classical PRRSVs, HP-PRRSV strains, or both in the field [19].

In recent years, the increase in the number of recombinant PRRSV strains and the frequency of recombination has complicated the global epidemiology of PRRSV. This suggests that PRRSV gene recombination and mutation play an important role in viral evolution and that the ongoing recombination and mutation of PRRSV have adverse effects on the diagnosis and control of PRRS [2]. Therefore, it is essential to determine appropriate control measures by understanding drivers of PRRS recurrence. The control of new introductions from off-farm sources suggests the demand for enhanced biosecurity, while re-introduction of resident strains indicates the need for better eradication or vaccination programme strategies [2].

1.2.2 Suboptimal Vaccine Control Methods

Due to the high genetic diversity of the virus, PRRSV vaccines are not fully effective to prevent and control infection, which is the second major challenge in PRRS control, and therefore outbreaks still occur in vaccinated herds [22]. Because of the high genetic and antigenic diversity of PRRSV, it is difficult to control and eradicate the disease. Even with over 30 years of research, a fully effective vaccine has not been produced [23]. The existence of multiple strains of PRRSV on the spectrum, the frequent recombination of strains, and the proliferation of recombinant strains resulted in a wide variety of strains in pig farms.

In 2007, new international standards for the next generation of PRRSV vaccines were developed, including rapid immune induction, protection of prevalent PRRSV strains, the ability to distinguish between immunised and infected animals, and no adverse effects on pig health [24]. To date, PRRSV vaccines developed include live attenuated, inactivated, subunit, DNA, and vector vaccines, but there is still no PRRSV vaccine on the market that meets all of these criteria [25]. Of these, live attenuated vaccines provide better protection but less cross-protection, in fact, the use of live vaccines may actually lead to the introduction of the virus into uninfected areas as live vaccines can revert to virulence; inactivated vaccines are safer but less effective in immunisation; vector vaccines offer partial protection but have not been used in practice [25].

Other problems with conventional PRRSV vaccines include the persistence of the virus and the possibility of regaining virulence, which contributes to higher recombination rates [26]. Due to the diversity of PRRSV strains [15], the varying ecology of PRRSV strains in pig farms [28], the persistent mutation and frequent recombination of PRRSV [19], and the return to strength of live attenuated PRRSV vaccines [24], it is currently unrealistic to rely solely on the means of vaccine immunisation to control PRRSV. The diagnostic method that provides genetic information about the strain causing the infection helps to identify potential causes of vaccination failure. For example, high genetic differences from the vaccine lead to limited cross-protection [27], or perhaps the virulence (escape mutation) returns from the vaccine itself in terms of genetic similarity to the vaccine [28]. Therefore, detection and genetic analysis of field strain through scientific testing and a combination of biosecurity and rational vaccination, followed by a selection of appropriate genotypic strains of vaccine for immunisation, is of great importance for the control of PRRS in the farms as well as in the region.

1.2.3 Suboptimal Diagnosis

The third reason that leads to unsatisfactory PRRS control is the limitations of diagnostic technology. Previously used to detect PRRSV, the main molecular diagnostic method was real-time quantitative reverse transcription-polymerase chain reaction (RT-qPCR), which allowed for rapid and accurate detection of the presence of PRRSV [38] but failed to identify specific strains or whether it is a highly pathogen strain unless strain-specific targeting primers were available [39]. Knowing exactly the type or species of the strain would help plan to vaccinate nearby farms with an appropriate vaccine, and for determining whether the endemic is to the country or has been imported, and for assessing the impact of the outbreak. To identify PRRSV paratypes, research methods previously used included reverse transcription, PCR amplification and cloning of PRRSV structural and non-structural proteins [16, 40], an additional step that can increase sequencing time and introduce bias, making rapid diagnosis difficult [41].

In recent years, nucleic acid sequencing techniques have also been used as a new detection tool to discover new strains [42]. However, most studies have been limited to one or two viral genes, mainly ORF5 and ORF7 [43]. This is because nucleotide variation in the PRRSV genome was concentrated in the ‘hotspots’ of sequence variability [36], and the technology available at the time had limitations for whole-genome sequencing. Moreover, the large amount of sequencing data related to the whole genome posed significant computational challenges [36], so only some major concentrations of variation could be selected at that time. However, phylogenomic studies based only on part of the PRRSV genome can be misleading. Although ORF5 and ORF7 have been widely used to understand the genetic variation of PRRSV, they only account for 5% of the whole genome [1]. If only ORF5 and ORF7 are used to predict the pathogenicity of the strain, 95% of the genome information could be lost. For example, the parts that have a functional role in pathogenicity happens to be in the remaining 95% of the genome. There is a necessity for deep sequencing of the entire PRRSV genome [37]. As a result, diagnostic tools that provide more genetic information are essential for the investigation, prevention, and control strategies of PRRSV.

1.3 PRRSV Quasispecies

The mutation rate and range of genetic variation in RNA viruses are usually determined by the fidelity and intrinsic error rate of the replication enzyme [29]. For example, viruses that have low fidelity RNA polymerases could generate every possible point mutation during each virus replication cycle and these may always exist in the population [30]. As small variants accumulate in the viral genome, these are amplified through subsequent rounds of replication. Some of these variants become extinct, but some persist leading to the formation of a population of different but related viral genomes, which is referred to as quasispecies [29]. In other words, quasispecies is a set of distinct viral genomes that are genetically related by mutations, interact synergistically at the functional level, and together contribute to the identity of the population. The theory of quasispecies has provided a population-based framework to understand the evolution of RNA viruses over the past 30 years. PRRSV is typical of RNA viruses as its high mutation rate and susceptibility to recombination allow PRRSV to evolve rapidly to adapt to dynamic environments [16]. This broad genetic diversity also reflects the low fidelity of RNA polymerases [29] and the lack of proofreading in PRRSV [31]. The combination of rapid replication kinetics and error-prone replication properties easily produces quasispecies mutation clouds in PRRSV [16].

Since the PRRSV genome is relatively small at only 15kb and contains reading frames or sequences overlapping with coding and structural functions [29], viruses rapidly accumulate potentially functionally relevant single nucleotide polymorphisms (SNPs) during the infection. These quasispecies variants may alter the phenotype and function of the virus [32]. For example, the protein function may be impaired by coding mutations that mediate evasion of host immune surveillance [29]. There is growing evidence that microevolution within PRRSV parapopulations can produce a high degree of sequence heterogeneity in viruses [33], and that genomic heterogeneity allows the rapid adaptation of parapopulations to changes in the microenvironment, such as the progression of the host immune response [34]. During PRRSV infection, viremia (virus levels in the blood) usually follows a cycle where it first drops and then rises one to three weeks after infection. However, after the initial peak, viremia continues to fall and then rises, which knows as “viral rebound” [new1]. The decline after the initial peak could be the result of the pig overcoming the dominant virus in the quasispecies, only for another virus in the quasispecies mutate into rebound viruses that escape the immune system and take over to reach a new adaptive peak.

Moreover the viruses in PRRSV quasispecies appear to have different tissue tropisms and have been found to vary in quantity in different tissues of the same infected individual[35]. Quasispecies differing in different tissues of the same individual may also reveal variants that change the success rate of the virus in different tissues, which may further reveal how different parts of the virus function [35]. Chen et al [36] found that PRRSV quasispecies are different in serum and tonsils. Mutations that occur during replication in organs such as the tonsils may not be detected if the virus quasispecies is only assessed in the serum [36]. Whereas the tonsils are one of the places that PRRSV hangs out during persistent infections, and persistent infections could be asymptomatic while spreading virus for many months [new1]. Understanding which dominant viruses persist in the tonsil could reveal the function potential of quasispecies in different tissues. The nature and extent of PRRSV quasispecies variation leads to genetic and antigenic drift, which may be clinically relevant in terms of virulence and pathogenesis [36, 37]. There is an urgent need for assays that can effectively identify and assemble the full range of quasispecies to distinguish between different variants of PRRSV to improve the understanding of the function of different parts of the PRRSV genome.

1.4 Long Read Sequencing

1.4.1 Whole Genome Sequencing Technologies

With the rapid development of next-generation sequencing technologies, whole-genome sequencing of PRRSV has been investigated using the next-generation short-read sequencing Illumina platform and traditional Sanger sequencing [44,45]. Although these two sequencing technologies can generate the entire PRRSV genome with an accuracy of more than 99.9%, the raw reads generated are usually less than 1500 bp. In fact, Illumina typically generates raw reads at 150-300bp [44]. Sanger is a lot more expensive, most of the sequencing will be in these even shorter fragments [45]. Moreover, to generate the whole PRRSV genome, Sanger sequencing as a time-consuming and laborious approach requires multiple individual sequencing reactions and primer sets [44]; while Illumina requires computationally resource-intensive genome assembly, which needs knowledge and time to perform efficiently [45].

Although sequencing throughput has increased dramatically over the last few years, there are still significant limitations in reading length that cause most genomes to be split into hundreds of fragments. Even though these large-scale sequencing techniques are able to capture the sequences of most genomes, their short-read lengths and scarce contextual information limit their utility in genome assembly and solving repetitive and complex genome regions [46]. Despite the suitability for small insertion deletions and calling single nucleotide variants (SNVs), accurate short reads are less useful for structural variation detection, haplotype dating and de novo assembly, all of which require information over longer sequence spans [47]. More importantly, short-read sequencing could only generate a consensus sequence that is a merged version of all of the genomes in the sample. So that, they could generate all SNPs, but could not show the relationship between SNPs. For example, the SNPs of PRRSV change their mutation frequency during infection depending on the tissue in which they are located [32]. Thus to understand how the mutational spectrum of PRRSV changes in time or space, long-read sequencing of the whole genome is also required, rather than short-read sequencing combined into the whole genome, as long-read sequencing can span some or all of the variants present within a quasispecies [37].

1.4.2 Advantages and Disadvantages of Long Read Sequencing

Long-read sequencing has been successfully applied to a wide variety of organisms including those with small genomes, such as viruses and bacteria [48,49]. There are many advantages of long-read sequencing over short-read sequencing. For example, a single long-read sequencing read can span the entire genome of a small genome such as that of PRRSV [50]. Additionally, long reads allow for de novo assembly and spanning of repetitive regions with low complexity to create accurate assemblies [50]. At present, two technologies are dominating long-read sequencing: Nanopore sequencing of Oxford Nanopore Technologies (ONT) and single molecule real time (SMRT) sequencing of Pacific Biosciences (PacBio). The SMRT sequencer reads a specific nucleotide by detecting a coloured fluorescence event from the incorporation of a tagged nucleotide by a polymerase attached to the bottom of a microwell [50,51]. The Nanopore sequencers measure current fluctuations of ions when single-stranded nucleic acids pass through biological nanopores [52]. Different nucleotides have different physical shapes, and so disrupt the current in a unique way, so base sequences can be inferred from specific patterns of current variation [53].

The high error rate of long-read sequencing has been much criticised by the scientific community, affecting their practical use, with an error rate of ONT tested at 15% in 2019 [54]. Recently, however, the base identification accuracy of reads generated by the two techniques has significantly improved, and the raw base identification accuracy of SMRT sequencers has been reported to reach up to 99% [55] and for nanopore sequencers to over 95% [52], although unpublished results suggest this is now higher than 95%. Moreover, long-read sequencing is a very new and fast improved technology with raw error rates that improve frequently, and recent improvements in physical and bioinformatics tools provide consensus accuracies over 99.9% [53].

The portable MinION nanopore sequencer may be more suitable for PRRSV sequencing than the SMRT sequencing platform [50], which requires a large number of high-quality RNAs as check materials. Although both sequencing platforms have read lengths that can accommodate the 15kb genome size of PRRSV [50], the nanopore has the ability to read individual viral molecules in real time; allows for rapid library preparation; and minimises the time between sample collection and data analysis [2]. Furthermore, the MinION offers the possibility to process samples immediately after site sampling for faster diagnosis and more rapid decision-making [2]. More importantly, the SMRT sequencer has very high throughput which does not need for PRRSV. The genome of PRRSV is 15kb, and this project requires only a few thousand reads. However, one SMRT run produces 0.5-1 billion bases, but we only want a few hundred thousand bases sequenced per sample. In addition, SMRT sequencing needs to send the samples at a sequencing facility, which would be quite expensive, while Nanopore could directly repeat the test in the laboratory many times.

However, in the field of bioinformatics, direct sequencing of samples with any technology is an expensive and time-consuming way to obtain data when there is a need to test new pipelines. There are tools that can simulate long read sequencing data for testing purposes, such as Badread [69], NanoSim [76], DeepSimulator [77]. Using simulated methods is faster, more economical and allows for more testing than using real sequencing data [75], and so this will be used to test the pipeline in this project. In addition, when using simulated reads, the reference nucleotide sequence provides a confident ground truth that may not be available with real data. Simulated reads from simulators such as NanoSim and DeepSimulator generally follow a realistic distribution and may draw read lengths similar to a real set of whole genome sequencing reads. Badread, however, allows the user to specify the mean and standard deviation, generating many quantitatively varying sets of reads, thus allowing us to assess the performance of the quasispecies assembly tool to be tested [69]. Also, Badread can simulate types of read errors that other tools cannot, including chimeras, adapters, faults, and rubbish reads [69]. In addition, Badread allows quality settings to be set. The other simulators are based on the error profile of older nanopore data, whereas Badread allows the use of data that simulate the most recent advances in reads accuracy. Therefore, the Badread simulator will be chosen for this project as the tool to simulate the data.

1.5 Quasispecies Separation

1.5.1 Haplotype Phasing

Whole-genome sequencing alone is not sufficient for a comprehensive analysis of mutation-prone viruses, and quasispecies separation is required for further understanding of quasispecies variation. In order to understand the separation of quasispecies, we first need to understand a very similar concept that is haplotype phasing. A diploid genome is defined as two similar but distinct versions of an individual's genome, or a "haplotype", and the problem of determining the complete sequence of two haplotypes is known as phasing [56]. In diploid species, one of these haplotypes is maternal and one is paternal and they are present in an equal quantity in the individual. Phasing is the process of linking variants together to determine which variants occur together on the same haplotype. The analysis of haplotypes is useful for understanding problems of genetic variation in populations [58], and for reducing errors during genome assembly. A quasispecies is more like a polyploid species in which there are numerous haplotypes, however these will not be present in even quantities in a quasispecies, and there is no prior knowledge of how many genomes to expect like in a polyploid species. A variety of sequencing strategies and analysis methods have been developed to phase genomes. The entire chromosome can be physically isolated before sequencing, or bioinformatically isolated after the entire genome, to obtain phased genomes [58,59].

There are currently three approaches for haplotype analysis. The first is population-based haplotype analysis [60]: genotypes of a large number of individuals are taken as input and haplotypes are inferred by using shared haplotype regions that exist due to common ancestry and limited recombination. This method produces highly accurate haplotypes, but it is only applicable to one individual and does not work for low-frequency variants [56]. The second is genetic haplotype analysis: genotype data from individuals with known pedigrees are used as input and haplotypes are determined according to the constraints imposed by Mendelian segregation [56]. This method is very successful when high-quality genotypes and large pedigrees are available. However, this method is unable to phase all individuals in the pedigree for heterozygous variants. The third is read-based phasing analysis: based on NGS sequencing, each read generated represents one of two haplotypes, and after mapped to a reference, these fragments are spliced into a longer haplotype fragment [61]. The ability of this method is limited by the length of sequencing reads, and only reads covering two or more heterozygous loci contribute to read-based fixation [56].

1.5.2 Long Reads Make Haplotype Phasing Better

To obtain longer phased blocks in the genome, long-read sequencing techniques are more suitable for this task than short-read NGS techniques. Long read sequencing is prone to insertion or deletion (indel) errors unlike short-read sequencing [61]. However, long reads usually contain multiple heterozygous sites to form stage information, which is a key advantage for haplotype reconstruction [56]. Moreover, compared with short reads, long reads are usually longer than repeat units, so the ambiguity of read mapping in repeat regions of the genome is reduced [56]. Therefore, long-read sequencing can solve many difficult areas that short-read sequencing cannot access. The key advantage of long read technologies over short read technologies in phasing is that they can link variants from the same haplotype unambiguously over a long range (Figure 1). In short, although long reads are subject to more sequencing errors, they align better to the reference genome and span a greater number of variant positions.

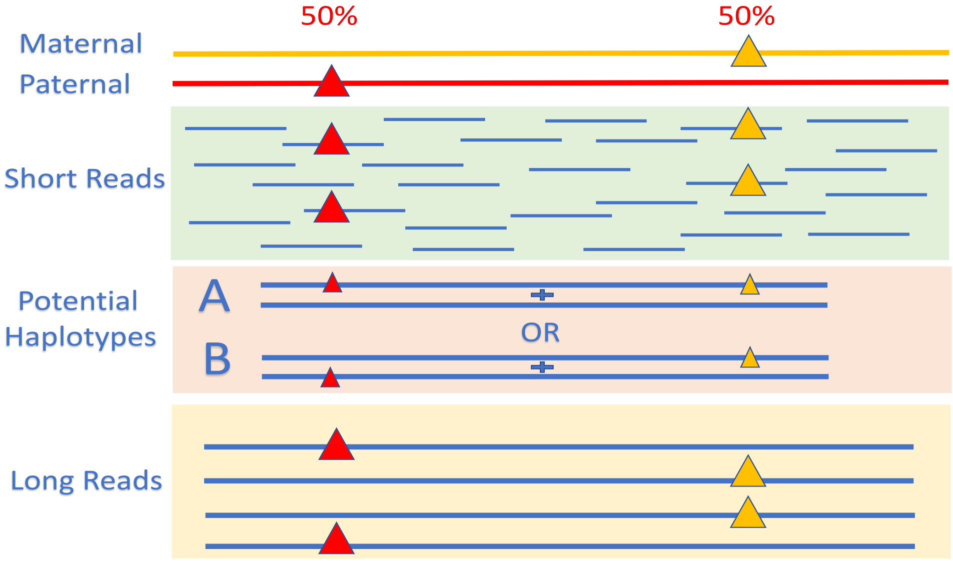


Figure 1: Comparison of Long Read Sequencing and Short Read Sequencing of Haplotype Phasing. Assume that the reference genome from the maternal and the paternal in the diploid organism show 50% of each SNP. The relationship between SNPs cannot be determined with short reads as they do not span the variants. In addition, the potential haplotypes based on short reads could be that the shown in A that SNPs come from the same haplotype, or the SNPs shown in B come from different haplotype. However, long read sequencing spans the entire region, showing the relationship between variants. Long read data shows the truth very clearly, whereas the short reads are ambiguous.

1.5.3 Limitations of Existing Haplotype Phasing Tools

The common approach to haplotype phasing is to call variants with a variant caller and then use a phasing tool for haplotype separation. Variant callers are designed to call SNVs and small indels, these include Samtools [72], Freebayes [] and GATK- HaplotypeCaller [79]. These tools are commonly used for downstream analysis of NGS short-read sequencing, which offers higher accuracy. Most of these variant callers tend to expect a diploid genome or a specified ploidy and will often filter based on variants proportions. Furthermore, existing long-read phasing tools are also essentially based on diploid fixation with uniform coverage of each haplotype, for example, the ratio-based phasing tool WhatsHap [62]. For diploids, knowledge of one haplotype can be inferred from the other, as each variable position can only be one of two possible bases [64]. This makes the method of phasing diploids relatively simple.

On the other hand, PRRSV is a haploid RNA virus and its analysis is more similar to polyploid phasing. For polyploid variants, the variant position can be one of many possible states, including multiple bases, deletions, or insertions [65,66], thus complicating the task of phasing. Although newer methods of polyploid typing are available, most expect a specified number of haplotypes and expect uniform coverage over each haplotype. However, it is not known how many PRRSV genomes there are or their frequency.

Variant callers do not tend to call non-diploid homozygous or heterozygous variants. In contrast to other variant callers, TAMA_go's variant calling module [74] is a very simple variant caller that makes no assumptions about the proportion of reads required to call a variant, allowing the user to customise the subsequent variant filtering. This is why in this project I use the TAMA_go variant caller rather than the more mainstream variant callers. For quasispecies isolation, we did not choose any mainstream phasing tool, but used a new unpublished pipeline designed by Amanda Warr and tested and evaluated it in this project.

1.6 Summary

PRRSV not only affects animal welfare, but also has a significant negative impact on pork production and the economy of the pig industry. Nanopore long-read sequencing technology can directly generate whole genome sequences in a single read, allowing for better identification and assembly of PRRSV quasispecies, leading to more accurate detection and control strategies during outbreaks. The hypothesis adopted in this project is that the complete genome sequence of PRRSV can be directly and accurately measured using nanopore long-read sequencing technology, and that quasispecies in viral samples can be reliably isolated by downstream bioinformatics tools and methods to better understand the formation of these quasispecies, and potentially their functional relevance. Here I have simulated quasispecies sequencing data and used it to test and optimise a new pipeline designed to separate the genomes in a quasispecies and produce a consensus genome for each genome in the sample. Using simulated methods to test this is faster, more economical and allows for more testing than using real sequencing data [75]. In addition, when using simulated reads, the reference nucleotide sequence provides a confident ground truth that may not be available with real data.

2 Methodology

This experimental project was divided into three phases to test the new pipeline based on the simulated PRRSV quasispecies data and then identify and assemble the real virus quasispecies based on the refined pipeline. The first phase involved the step-by-step testing of PRRSV quasispecies isolation through the simulated data and relevant bioinformatics tools. The aim was to test a new unpublished quasispecies separation pipeline and optimise its parameter settings. Specific steps included simulating the PRRSV genomes, simulating nanopore sequencing data, aligning the sequencing data to the reference sequence and pre-processing, the variant calling of the processed sequencing data, and finally separating the quasispecies through this new pipeline. The second phase was the one-click identification and assembly of the simulated viral quasispecies through an automated pipeline that integrated all the above steps. The purpose was to detect and evaluate this fully automated quasispecies separation pipeline. The third stage is sequencing the true PRRSV, amplified by two different polymerases, through the Nanopore long read platform to obtain the true genome. The automated pipeline is then used to identify and assemble the true quasispecies. In order to evaluate the effectiveness of the pipeline in isolating really genomic quasispecies and to identify the most suitable polymerase for PRRSV amplification.

2.1 Simulating PRRSV Quasispecies Data

The reference genome FASTA file from Lelystad virus PRRSV-1 (NC_043487) was obtained from the National Centre for Biotechnology Information (NCBI) []. The FASTA format provides a record of nucleic acid sequences in a single letter code while allowing sequence manipulation and analysis by text processing tools . In order to simulate single nucleotide variants, namely viral quasispecies in the viral genome, random point mutations including were introduced into the reference gene through the ‘msbar’ command in the EMBOSS tool [67] (version 6.6.0.0). Random point mutations were introduced to generate the original genomes containing mutations in order to simulate the genetic variation of PRRSV viruses in a quasispecies and approximate it to real genomic data as closely as possible. The number of introduced mutations was limited to under five to generate 10 FASTA files of simulated virus genomes. Phylogenetic trees were generated for of the simulated quasispecies through MEGA (Version 4.0) [68] to obtain a visual representation of the relationships between the different mutant genomes.

2.2 Simulating Nanopore Sequencing Data

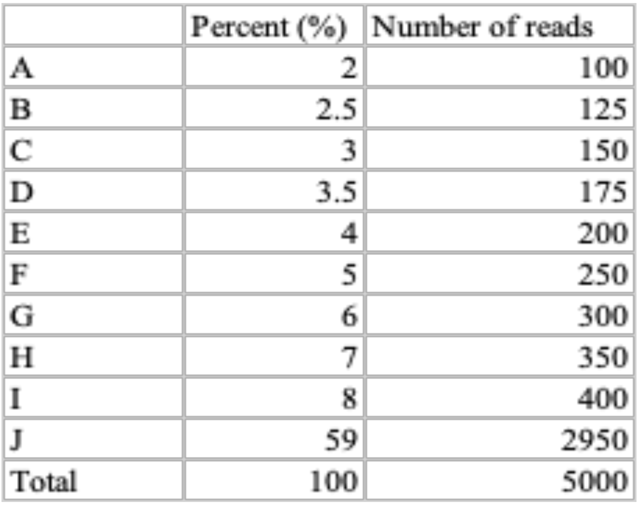
To simulate nanopore sequencing data including long-read sequencing errors, data was generated for each simulated viral genome with the Badread tool [69] (version v0.2.0) (https://github.com/rrwick/Badread). The number of reads was set to 800x. The mean of the fragment length distribution was set to 14000 to the PRRSV genome size in order to simulate full-length sequencing, and the standard deviation was set to 13000, while other parameters were set to the official recommendation of 'very good reads' (error_model random --qscore_model ideal --glitches 0,0,0 --junk_reads 0 --random_reads 0 --chimeras 0 --identity 95,100,4 --start_adapter_seq --end_adapter_seq) to generate the FASTQ files that approximate “super accuracy” reads, the current most accurate error profile of real Nanopore data. While the genome is longer than 14,000 bases, existing primer sets within the group target a slightly smaller region and the data was designed to approximate the coverage of these amplicons as a text format that holds information about nucleic acid sequences and their sequencing quality scores, FASTQ format is now the standard format for saving high-throughput sequencing results[].

2.2.1 Setting up an even dataset

PRRSV sequencing datasets with even coverage were produced to generate the “best case scenario” even dataset, which simulates an unrealistic distribution of equally represented genomes in a quasispecies similar to a polyploid scenario. The even dataset consists of ten simulated PRRSV strains with 24 variants. The ten simulated Nanopore datasets from Badread were mixed directly into one FASTQ file through the "cat" (concatenate) command. The total yield, total read number, read quality, and read length of the simulated genome with even coverage were analysed with NanoPlot (version 1.38.0) [70] for quality control.

2.2.2 Setting up an uneven dataset

The sequencing dataset of uneven coverage also consists of the same ten viral genomes in different proportions to simulate real PRRSV gene sequences more realistically and accurately. Full-length PRRSV reads were first filtered from the FASTQ file generated by Badread through the "awk" command to ensure high coverage of full-length genomes without the shorter reads that Badread also produces. During sequencing of real data the reads represent amplicons of a consistent length and this filtering allows for a dataset more similar to the real data. Afterwards the simulated full-length genome sequence set was subsampled at different proportions through the "sample" command in the seqtk tool (version 1.3-r106) (https://github.com/lh3/seqtk) to generate the FASTQ files. This resulted in a total number of 5000 reads and a coverage proportion for each genome ranging from 1% to 59% (Table 1). If the number of full-length genomes did not reach the required coverage number to reach the correct proportion, for example in the most abundant genome, the Badread tool was used to re-simulate the data set at a higher coverage. Then the ‘seqstats’ command was used to check the number of full-length genes filtered. The ten FASTQ files generated by the 'sample' command were again mixed into one FASTQ file with the 'cat' command for uneven coverage of the simulated reads. The uneven-coverage dataset was quality-checked through NanoPlot [70]. The comparison of even and uneven data sets allowed a more visual observation of the difference between the ideal result and the result more closely resembling the real gene.



	Percent (%)	Number of reads
A	2	100
B	2.5	125
C	3	150
D	3.5	175
E	4	200
F	5	250
G	6	300
H	7	350
I	8	400
J	59	2950
Total	100	5000

Table 1:Uneven Dataset Coverage Distribution. The total number of reads is 5000, with a minimum coverage of only 2% (100 reads). The maximum coverage is 59% with 2950 reads. the first four genomic sequences are in 0.5% size increments and the last five genomes are in 1% size increments.

2.3 Alignment and Pre-processing

Simulated sequencing data has to be compared with a reference sequence in order to obtain raw error rates and detect variants. The simulated sequencing reads were mapped to the PRRSV-1 reference sequence obtained previously from NCBI with the Minimap2 tool (version 2.18-r1015) [71], so as to generate a SAM file. As a common format for storing sequence alignment results, the SAM format provides the flexibility to store all alignment information[]. However, it was also necessary to convert SAM files into BAM files with the Samtools tool (version 1.12) [72] as SAM files are plain text files that are not conducive to the storage and manipulation of information. BAM file is a binary alignment or mapping format that compresses the file size for more efficient storage. The bam file can then be sorted by coordinate and indexed for compatibility with other bioinformatics tools [].Alignments to the original reference sequence were carried out for both even and uneven sets of sequences to produce sorted and indexed BAM files. A visual check was then performed in IGV (version 2.10.0) [73] to visualise all introduced simulated random variants and their proportions. This step allowed for quality control of the BAM files and of the process of combining the simulated data in different proportions.

2.4 Variant Calling

Once the processed simulated sequencing data was obtained, variant calling was performed on them with the TAMA_go variant calling module(version 2.2), which is part of the TAMA tool [74]. The TAMA_go variant caller was used with the input of the BAM files for the sorted even and uneven sets as well as using the original unmutated Lelystad reference genome as a reference. The TAMA variant caller output text files containing SNV positions (prefix_variants.txt), and information regarding the quality of the reads and their alignment to the reference (prefix_read.txt). The command parameters were set to reduce computational effort relating to non-essential additional outputs as recommended by the tools' author (additional parameters are -d merge_dup -log log_off -cv no_clips_var).

2.5 Separation of PRRSV Quasispecies

The variants were isolated and assembled through the Untangle python script (by Amanda Warr, currently unpublished). The script takes the following steps:

1. Identify “informative SNVs” in the TAMA output file (prefix_variants.txt) that reach a minimum coverage threshold (CV) specified by the user.
2. Identify reads that passed TAMA’s quality filters and that are at least 90% of the length of the reference.
3. For each read passing quality filtering, genotype the read for every informative SNV and produce a string representing these genotypes.
4. Group reads into clusters of matching strings.
5. Output lists of read names that occur in each cluster that contains more reads than a user specified read coverage threshold (CR).
6. Output a histogram showing coverage over all clusters and the position of CR.

The identified variants were separated by Untangle into clusters (genomes) that had reached a threshold by parameters, including the CV which is the percentage of reads that support each variant and the CR which is the minimum number of reads that share a genotype string. During this project these parameters were changed to see what results were produced at different thresholds and how these related to the ground truth. The output of Untangle includes the number of variants identified, the number of full-length reads identified, the number and the reads of potential clusters that the variants might form, the number of clusters that met the read count threshold and a histogram of the clusters.

Quasispecies segregation was performed separately for even and uneven data sets using the Untangle script. The same parameters were set for the even and uneven datasets, respectively, with attempts at values of CV from 1 to 5 and CR from 25 to 150. Thus, the results produced by Untangle at different coverage rates were obtained. In turn, the limits of the effective detection of the Untangle procedure and the accurate assembly of PRRSV quasispecies were inferred, and a reliable set of parameters was determined for the subsequent use with real data.

In order to check the reliability of the results produced by Untangle on the simulated data, the reads from the clusters in the TXT format file also needed to be extracted. The ‘seqsub’ command in the seqtk tool was used to obtain the FASTQ file of the reads from the list of reads produced by Untangle. The reads separated by Untangle were aligned to the original genome through the Minimap2 tool [71], and a position-ordered indexed BAM file was generated with the Samtools [72]. The reads were then visualised in IGV [73] to check whether the reads had been separated as expected and approximated the data originally simulated for each genome. Finally, the results and differences using Untangle were compared in IGV for even and uneven datasets.

2.6 Automated pipeline

The fully automated pipeline by Amanda Warr is currently unpublished. The working steps of the automatic pipeline can be divided into three stages (figure 2). The first stage is to process the raw genome sequencing data to generate a consensus sequence to use as a reference. The second stage is the isolation of quasispecies clusters using variant caller-TAMA and phasing tool- Untangle pipeline. The third stage is to polish the reads in each quasispecies cluster and produce a genome representing that cluster. In order to evaluate the functionality of the automated pipeline, this pipeline was used to perform the quasispecies separation of nanopore sequencing data generated by the simulation software. Quasispecies identification and separation were first performed with the parameters CV of 5 and CR of 50 for both even and uneven sequencing datasets. The two datasets were compared to produce the desired results and the parameters were adjusted as necessary. The MEGA tool was used to generate a phylogenomic tree of the results obtained and verify whether all quasispecies could be found with high coverage. IGV tool was used to compare the difference between the even and uneven set results

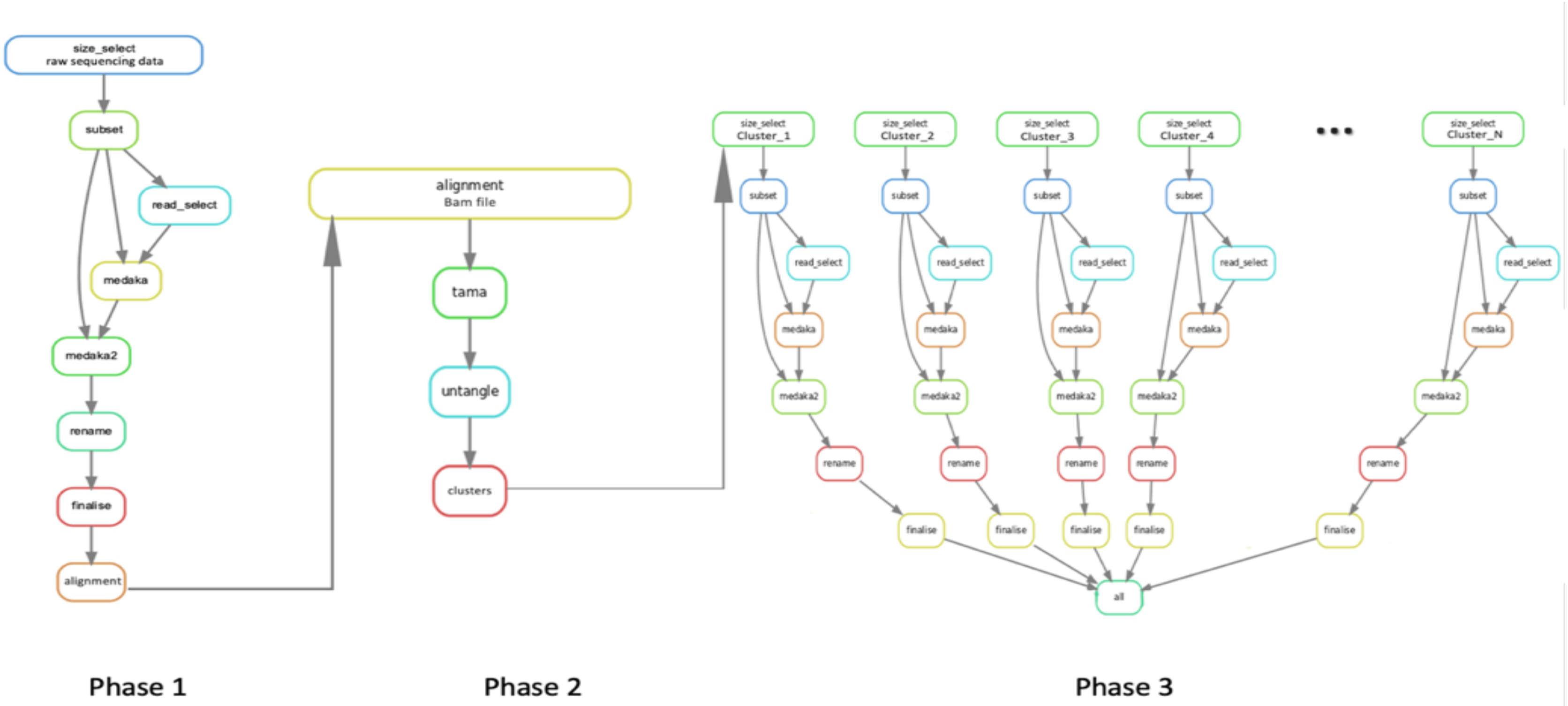


Figure2: Three Phases of Automated Untangle Process. The work steps of the automated pipeline can be divided into three phases, namely the processing of the raw viral genomes for the initial separation and generation of the consensus genome, the separation of quasispecies clusters using the TAMA and Untangle, and the further polishing of the quasispecies clusters with the same steps of phase one.

Due to the rapid mutation of the real PRRSV genome mutates, its reference genome is often difficult to align to due to a hypervariable region. Thus, the first step in the first phase was aimed at generating a reference genome sequence that closely represents the viral genome in the sample. The simulated Nanopore reads were used as the raw data in the automated pipeline, and the genome length reads were first filtered selected from the dataset. The full-length read with the highest average base quality was selected as the reference genome, and full-length reads with 100-fold coverage were used as the material for polishing the reference genome. The fourth step was to polish the reference sequence twice with the Medaka tool and generate a consensus genome from the reads representing the final reference genome sequence for the dataset.

In the second phase, the full-length reads were aligned to the consensus genome, and a Bam file was generated. Then the TAMA_go module of TAMA tool was used to make variant calling for the full-length PRRSV genome. Finally, the Untangle tool was used to separate quasispecies clusters. In the third phase, a reference genome was generated in the automated pipeline for each cluster after Untangle generated a series of quasispecies clusters. The same steps of the automated pipeline were conducted in this phase as in the first phase. The full-length reads with the highest average base quality were selected first to represent a genome in the reference sequence, then two polishing steps were performed using medaka, and a consensus genome was generated from the reads.

2.7 Sequencing and analysis of real data

2.7.1 MinION Sequencing

While simulated data can help to optimise the pipeline, the real test is using real data. Real data has the added complication of errors introduced by polymerases during PCR. In order to assess the impact of this, two samples of amplified whole PRRSV genome cDNA were obtained directly from the Tait-Burkard group at the Roslin Institute. Amplifications had been carried out on the same sample, a lab strain called SU1-Bel, with two polymerases of different accuracy, namely LongAmp (New England Biolabs, UK), a relatively low accuracy polymerase that amplifies long fragments well, and Phusion (New England Biolabs, UK), a high accuracy polymerase that is more suited to amplify shorter fragments. In order to compare the detectable quasispecies produced by accurate and inaccurate polymerases for the same sample, I sequenced both samples using Nanopore sequencing.

As per the manufacturer's SQK-LSK109 protocol (ONT, UK) for sequencing Genomic DNA.. Briefly, the DNA was repaired using NEBNext FFPE DNA Repair Mix (NEB, UK), and end prepped using the Ultra II End-prep kit (NEB, UK) and incubated at 20°C for 30 minutes and 65°C for 30mins followed by a clean-up with AMPure XP Beads (Agencourt, UK). Unique barcodes were ligated to the DNA ends of each sample with the Native barcoding expansion kit (ONT, UK), and Blunt T/A ligase master mix (NEB, UK) and the reaction was incubated at room temperature for 20 minutes, this was followed by an AMPure bead clean up. Thirdly, sequencing adapters were ligated to the DNA ends with NEBNext Quick Ligation Reaction Buffer (NEB, UK) and Quick T4 ligase (NBEB, UK) and incubated for 10 minutes at room temperature, and AMPure XP beads clean-up was done. Finally, an R9.4 flow cell was primed following manufacturer's instructions, and a 15μl DNA library was combined with sequencing buffer and library loading beads and loaded into the flow cell. Sequencing was carried out using ONT's MinKNOW software on the MinION mk1c.

2.7.2 Downstream Analysis

Basecalling and demultiplexing was carried out by the Guppy software developed by Oxford Nanopore Technologies (UK) using the "super accuracy” basecalling model. This part was done by another lab partner as it required GPU access. Once the basecalled data was acquired, further analysis and comparison of results were carried out by me with automated Untangle. MEGA and IGV software was used to compare the results produced by two different polymerases.

3 Results

3.1 Quality control of simulated datasets

3.1.1 The relationship between simulation PRRSV genomes

Ten mutant genomes were generated from the original reference sequence of the PRRSV-1 virus from NCBI by the EMBOSS tool, with a phylogenetic tree generated by the MEGA tool (Figure 3). Three of these genomes (sequence mutation 6/7/8) were generated using the original PRRSV-1 as a reference for the first generation of virus quasispecies genomes. Furthermore, the second generation of quasispecies genomes (mutation-related 6/7/8/9) was generated by introducing random mutations based on the first generation of genome. The third generation of genome (mutation-related 6a/7a/8a) was generated based on the second generation of genome to generate a quasispecies gene with a random point mutation.

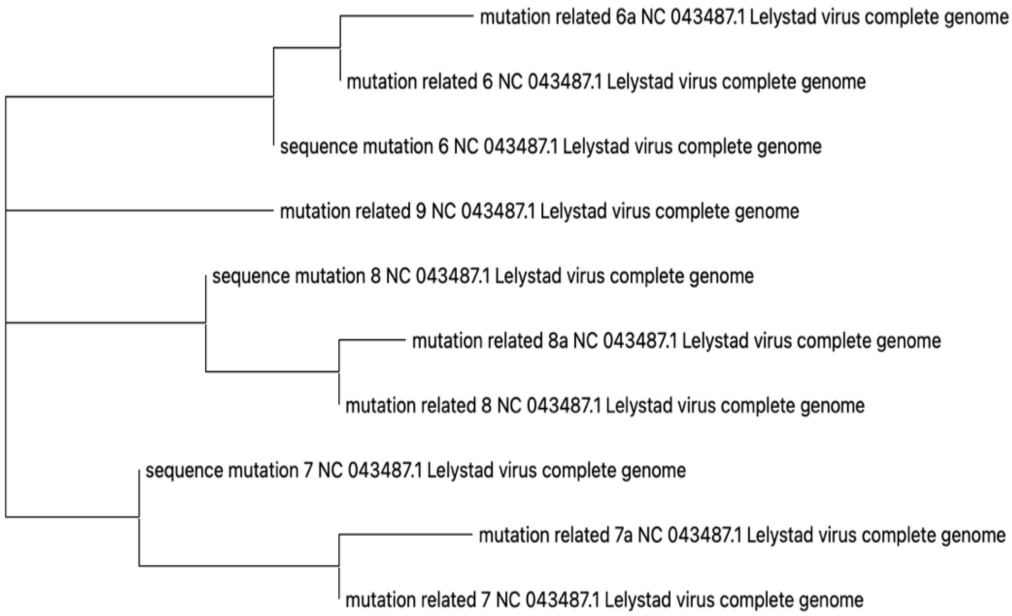

mutation related 6a NC 043487.1 Lelystad virus complete genome
mutation related 6 NC 043487.1 Lelystad virus complete genome
sequence mutation 6 NC 043487.1 Lelystad virus complete genome
mutation related 9 NC 043487.1 Lelystad virus complete genome
sequence mutation 8 NC 043487.1 Lelystad virus complete genome
mutation related 8a NC 043487.1 Lelystad virus complete genome
mutation related 8 NC 043487.1 Lelystad virus complete genome
sequence mutation 7 NC 043487.1 Lelystad virus complete genome
mutation related 7a NC 043487.1 Lelystad virus complete genome
mutation related 7 NC 043487.1 Lelystad virus complete genome

Figure 3: Phylogenomic tree of the simulated genomes. Genetic relationships between ten simulated genome sequences, form a simulated PRRSV paraphyletic genome sequence.

3.1.2 Quality control of simulated nanopore sequencing data

A series of graphs were produced by quality control analysis of the even dataset which aligned with reference genomes using the NanoPlot tool (Figure 4). The length distribution can be clearly observed from the weighted length distribution plot (A) where the peak of the data is mainly located around 15K. It is clear in the length output plot (B) shows that all of the data is 15kb or less, as we would expect because that is the length of the genome. It also shows that a large portion of the data are full length reads. Graph C is a plot of the read length versus the average quality point which is how identical the reads are to a truth. The quality scores are Phred scores which estimate the probability of an incorrect base call (Phred quality score 10 presents 90% base call accuracy, Phred 20 presents 99%accuracy). It is clearly shows that the quality of this set of mock sequencing is evenly distributed, with an average value of around 13.

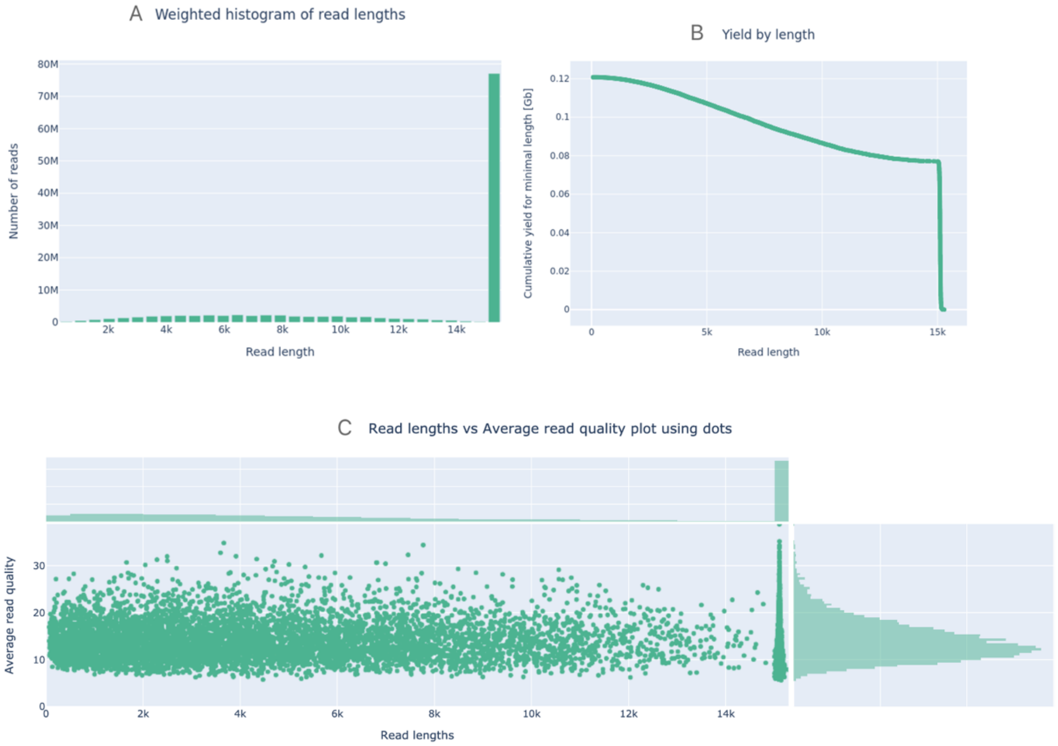


Figure 4: NanoPlot main graph for even dataset. Graph A is the weighted histogram of read lengths, which are mainly distributed at 15 K, while the X-axis and the Y-axis represent the sequence length and the number of reads, respectively. Graph B is Yield by Length. Graph C is read lengths vs average read quality plot using dots, which indicate the distribution of lengths and qualities, and the histogram on either side further presents the distribution of lengths and qualities, indicating that the sequence has a large number of short reads and 15K read lengths.

Since full-length genes were screened from the uneven dataset, it can be seen from Figure 5 that there are hardly short genomes in the uneven dataset and the read length distribution is still around 15K, which is consistent with the expected results. The yield of the 15K read length does not fall vertically, either, but shows a curvilinear fall to 15.2K. This is because the X-axis is smaller, the curves are likely very similar.

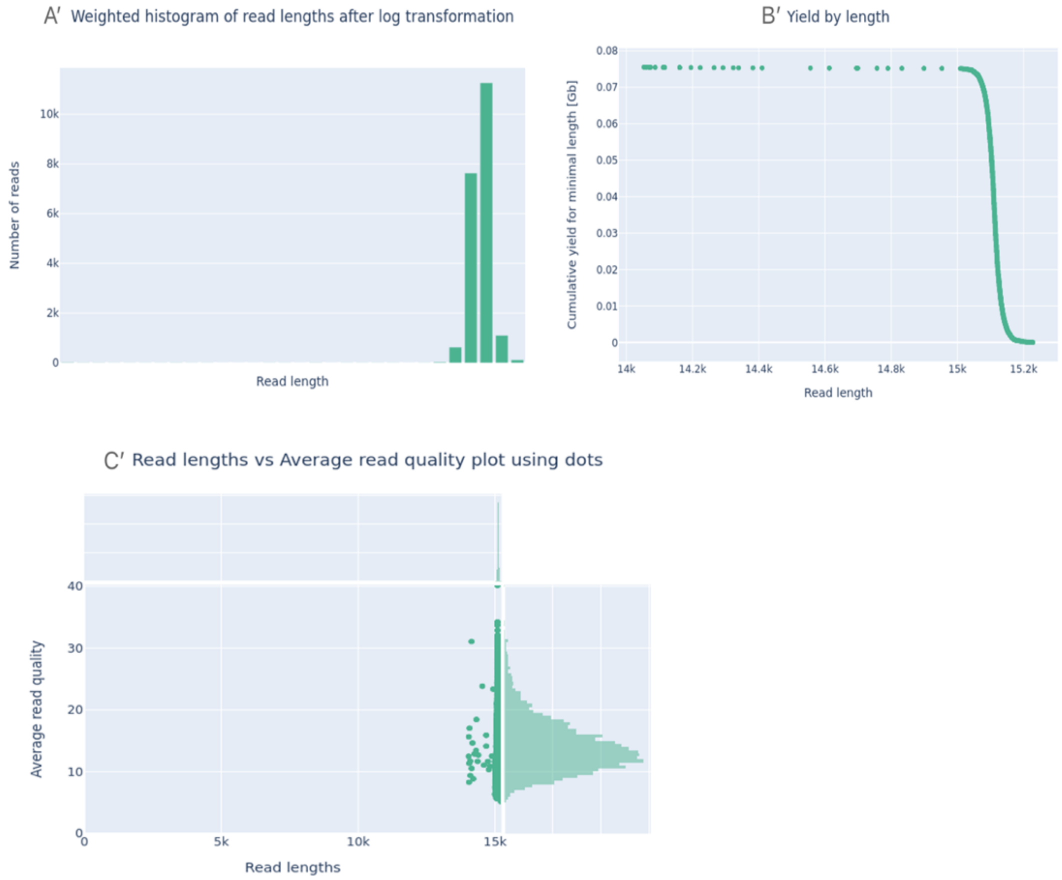


Figure 5: Nanoplot main plot for the uneven dataset. Graph A is the weighted histogram of read lengths, which are mainly distributed between 14K-15K. The X-axis and the Y-axis represent the sequence length and the number of reads, respectively. Graph B is Yield by Length where X-axis and Y-axis represent the length and the frequency of the yield. Furthermore, the yield decreases from 15K to 15.2K, and other read lengths are unevenly distributed. Graph C is Read Lengths Vs Average Read Quality Plot Using dots, which indicate the distribution of lengths and qualities. Reads are mainly concentrated at 15K, while qualities are mainly distributed at 13.

3.2 Visualisation of alignment and pre-processing results

Figure 6 shows a visual inspection of the IGV of the even dataset (A) and the uneven dataset (B) prior to the segregation of the quasispecies. The positions of the variants in the even dataset and the uneven dataset are almost identical as both sets of data are sequencing data generated from the same simulated paraphyletic genome. However, the SNPs represented by the coloured lines in both plots show that the two sets of data have different SNP frequencies. For example, as can be seen in the top bold grey bar (coverage track), the SNP represented by the blue or green allele on the far right (near the 14 Kb position) has an allele frequency in the uniform coverage group (A) different than that in the uneven coverage group (B). This is due to the design of the uneven dataset and the difference in the coverage of different genomes that some genomes have a large number of reads, while others have fewer reads. In the even dataset, however, each genome has roughly the same coverage of reads, and this can also be seen in the read alignment tracks below where some of the genomes present in the even dataset are very sparsely represented in the uneven dataset.

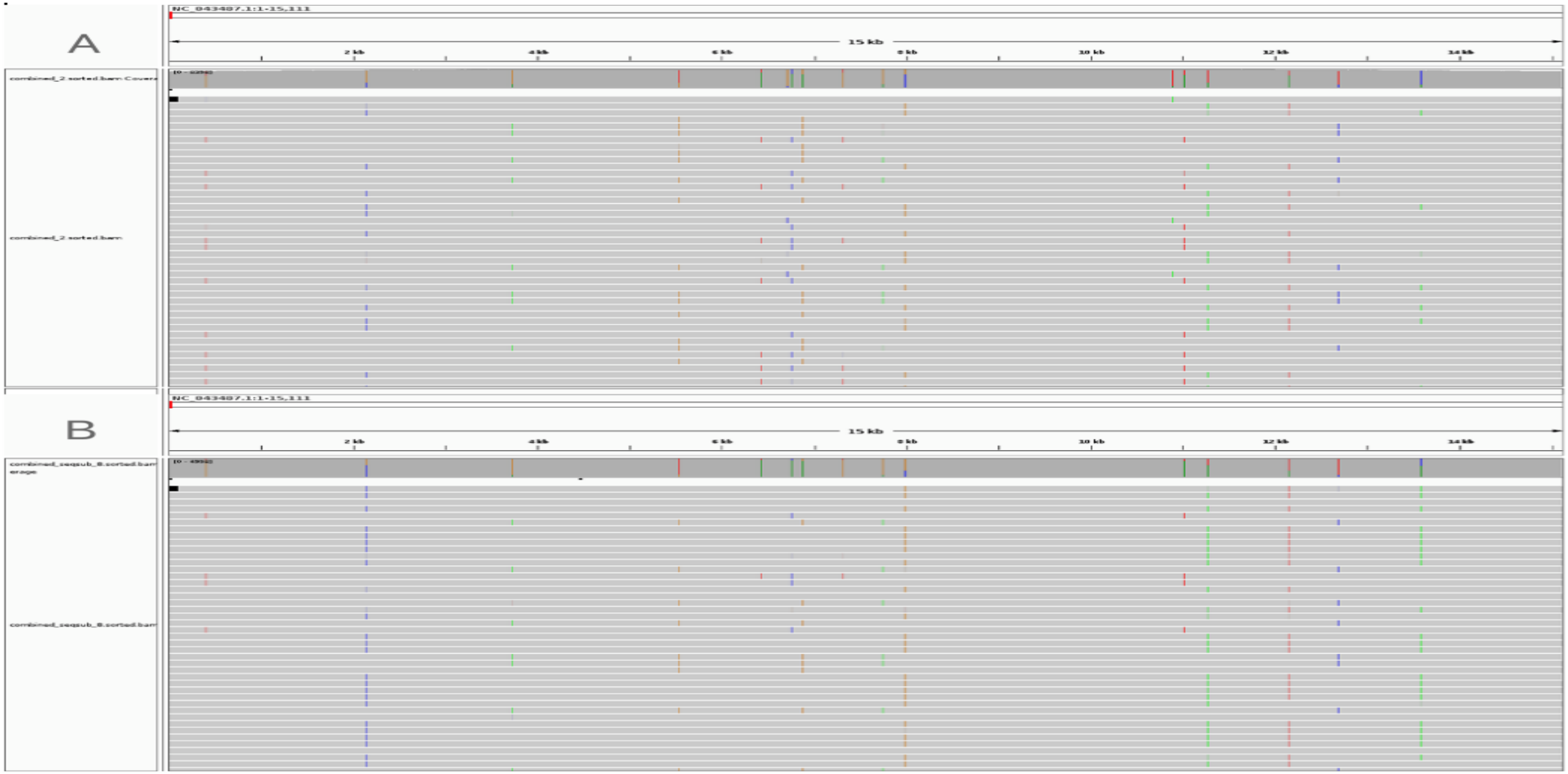


Figure 6: IGV visualisation of simulated nanopore sequencing data. Graph A is the even dataset, Graph B is the uneven dataset where the top thick grey bar shows the coverage of each base. When it is solid grey, it means that almost all bases at that position are the same as the reference value in each read. In addition, the colours show the proportion of each base where bases in the reads differ from the reference. One of the bases is the same as the reference base, while the other is a variant. If there is a coloured line, it represents a base that is different from the reference base. If the line has two different colours, it means that two possible bases occur at that position. Below this bar the alignment of individual reads can be seen.

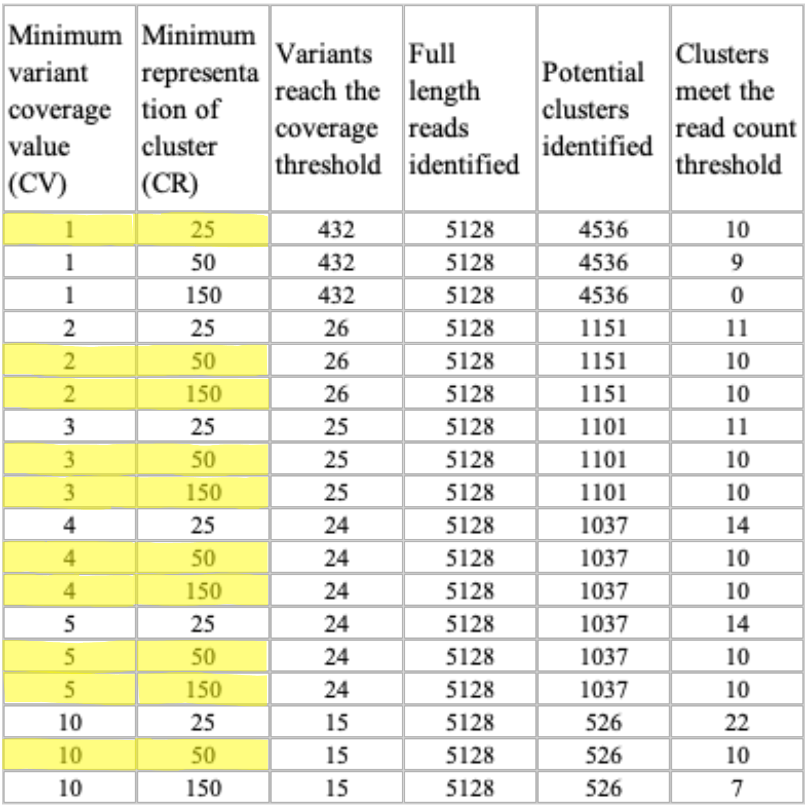
3.3 Quasispecies separation in simulated data

3.3.1 Improvement the Code of Untangle

When running a quasispecies separation on the even dataset using the first version of Untangle, the results revealed that there were deletions in some of the reads exactly where the informative SNPs were located. There might have been some issues in the Untangle script causing deletions that overlapped with informative variants to cause clusters to be split when they shouldn't have been. Amanda Warr then modified the code of Untangle to not consider reads where there is a deletion in the same position as SNPs and from that edit Untangle was able to work perfectly on the dataset.

3.3.2 Selection of Untangle parameters

Different settings of the Untangle parameter were tried for the uniform coverage dataset (Table 1), and it was known that the correct separation should result in 10 clusters reaching the read threshold and there were 24 variants. The CV parameters of 1, 2, 3, 4, 5 and 10 were tried, the CR parameters of 25, 50 and 150 were tried. When the CV was 1, the number of variants reaching the threshold was the highest, likely due to the false inclusion of more Nanopore sequencing errors as “informative variants”. When the CV was 1 and the CR was 25, 10 clusters were produced but the number of reads in each cluster was very small. When the CR was 150 (the CV of 1), there was no cluster left to reach the threshold at this point. When the CV was greater than 1, there were more reads per cluster. When the CV was 10, only 15 reachable coverage variants were obtained. When the CV was between 2 and 5, an approximately correct number (24) of variants was obtained, and also when the CV increased, more accurate number of variants was obtained. For most levels of coverage (CR) when the CV was between 2 and 5, the Untangle pipeline gave the same correct answer of 10 per cluster. The only case where we failed to get the correct answer was when the coverage was 25 per cluster, which again likely accepted clusters that were the result of Nanopore sequencing errors. In addition, the CR was set to 30, and a number of clusters of 10 were obtained as well.



Minimum variant coverage value (CV)	Minimum representation of cluster (CR)	Variants reach the coverage threshold	Full length reads identified	Potential clusters identified	Clusters meet the read count threshold
1	25	432	5128	4536	10
1	50	432	5128	4536	9
1	150	432	5128	4536	0
2	25	26	5128	1151	11
2	50	26	5128	1151	10
2	150	26	5128	1151	10
3	25	25	5128	1101	11
3	50	25	5128	1101	10
3	150	25	5128	1101	10
4	25	24	5128	1037	14
4	50	24	5128	1037	10
4	150	24	5128	1037	10
5	25	24	5128	1037	14
5	50	24	5128	1037	10
5	150	24	5128	1037	10
10	25	15	5128	526	22
10	50	15	5128	526	10
10	150	15	5128	526	7

Table 1: Comparison of the results for different parameter settings of the Untangle pipeline. The minimum variant coverage value (CV) of 1-10 and the minimum representation of clusters (CR) of 25-150 yielded the results including the number of variants reaching coverage, the number of full-length reads identified, the number of potential clusters and the number of clusters reaching coverage.

A histogram generated by separating quasispecies of even datasets using Untangle with parameter CV of 2 and CR of 50 (Figure 7 A). The histogram shows the distribution of the error (left side) and the real genome (right side), and select a cut off (red line) that are confident the errors end and the real genome begins. There was a very tight distribution on the left, which the errors are very obviously separated from genomes where all of the real genomes are high coverage. It is clear where errors end (near 50) and real genomes begin (near 200), therefore, the cut off (CR value) is between 50-200 could both produce the correct answer. Figure 7 B is from uneven dataset with the same Untangle setting of even dataset. It is a little tougher to see where to cut errors and real data off since the low abundance genomes are so much closer to the errors. The histogram could help to choose CR value (the position of cut off) which determines whether miss real genomes or get all of genomes and some errors.

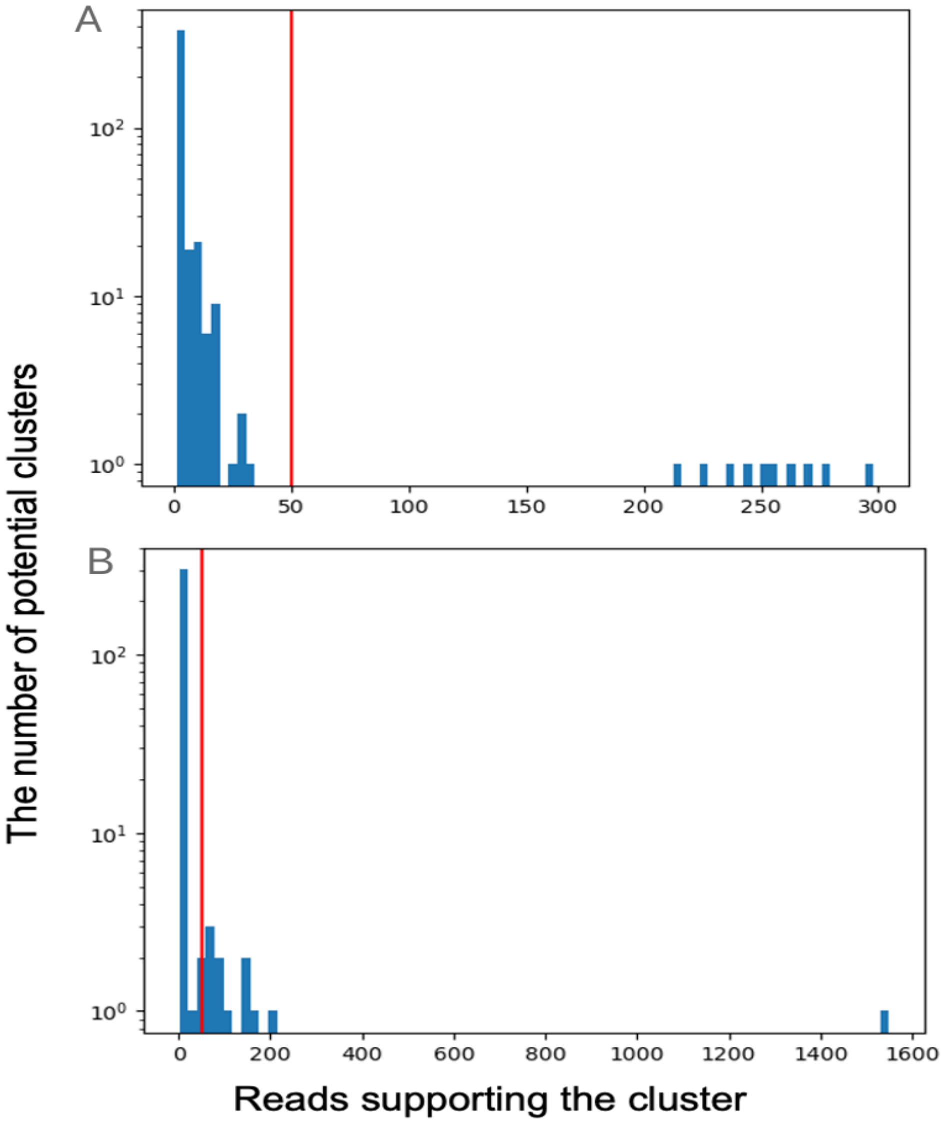


Figure 7: The histograms produced by the Untangle pipeline. Graph A is from even dataset, Graph B is from uneven dataset. For each genotype string that could be seen from the histogram generated by Untangle and how many reads support the string, basically is the coverage of each potential genome. The X-axis is

the read count which supporting the clusters, and the Y-axis is the number of potential clusters with that read count. When there are errors from either Nanopore or PCR, Untangle pipeline will generates a bunch of unreal strings. They have relatively low support in the reads as they are unreal, which means on the histogram they fall on the left side, towards the lower read count value. The real genomes have higher coverage because they have lots of support in the reads. Therefore, the histogram shows the distribution of the error and the real genome, which likely represents almost completely error, and select a cut off that are confident the errors end and the real genome begins (red line).

3.3.3 Testing for uneven coverage

Tests of the quasispecies separation of unevenly covered viral genomes were conducted for CR2-5 and CV50. At first, the Untangle pipeline was assumed to be unable to separate the genomes with the lowest coverage at certain parameter settings. However, after testing in uneven groups, it was found that Untangle did identify all variants for all parameters of all sizes. The problem, however, was that Untangle repeatedly isolated more quasispecies clusters than the standard answer of 10. A comparison of the IGV images (Figure 7) from the even dataset (A) with a CV of 2 and a CR of 50 to those from the uneven dataset (B) revealed that Untangle not only successfully identified the low-coverage genome in the uneven dataset but also produced an erroneous set of clusters. After zooming in on that erroneous cluster (C), it was found that the cluster was the highest coverage genome split, which perhaps increased the chance of error stacking splitting it. However, it is not clear what caused Untangle to separate these specific clusters, and thus further investigation is required.

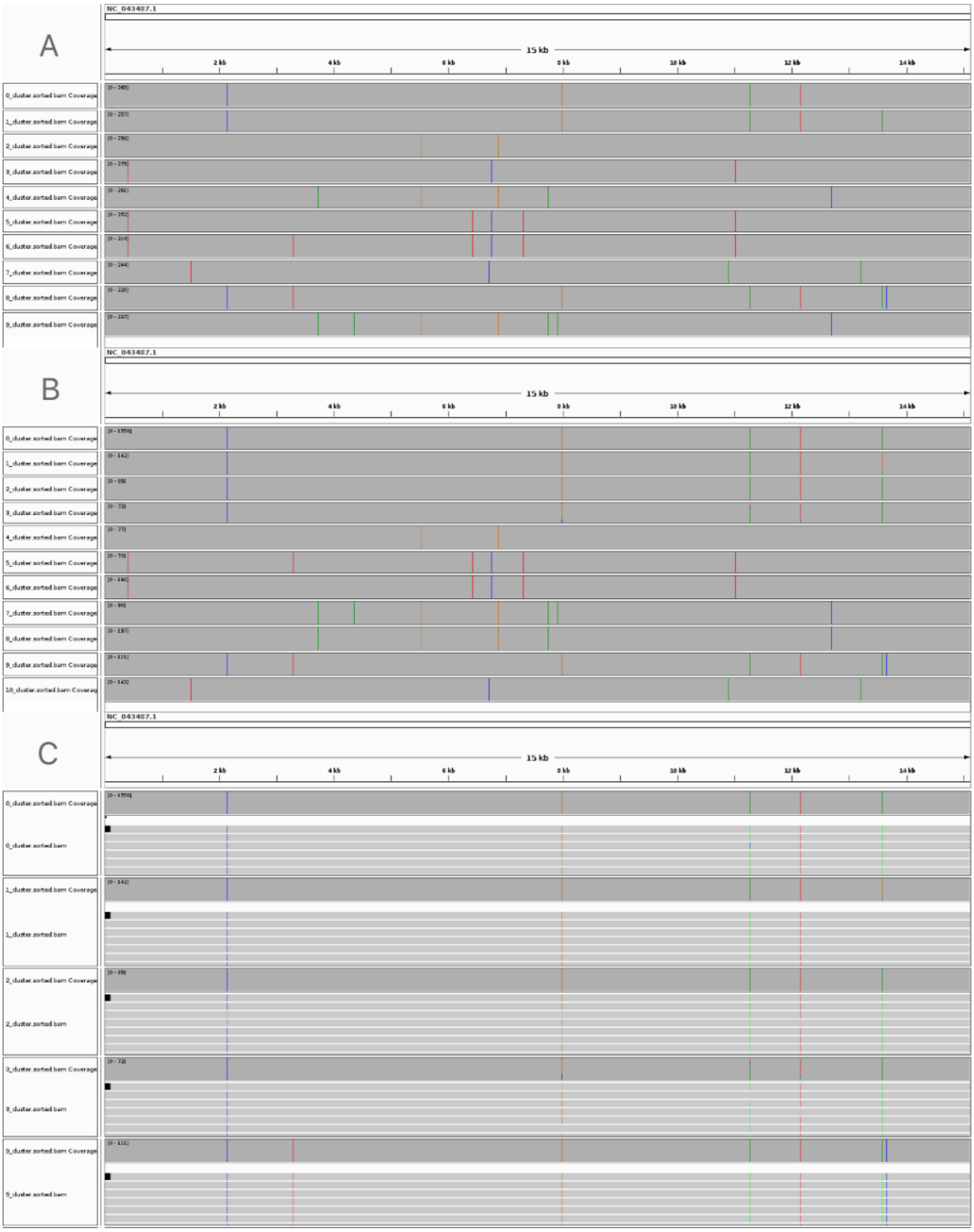

A
B
C
NC_043487.1
15 kb
2 kb
4 kb
6 kb
8 kb
10 kb
12 kb
14 kb
0_cluster.sorted.bam Coverage
1_cluster.sorted.bam Coverage
2_cluster.sorted.bam Coverage
3_cluster.sorted.bam Coverage
4_cluster.sorted.bam Coverage
5_cluster.sorted.bam Coverage
6_cluster.sorted.bam Coverage
7_cluster.sorted.bam Coverage
8_cluster.sorted.bam Coverage
9_cluster.sorted.bam Coverage
10_cluster.sorted.bam Coverag
0_cluster.sorted.bam
1_cluster.sorted.bam
2_cluster.sorted.bam
3_cluster.sorted.bam
9_cluster.sorted.bam

Figure7: IGV visualisation of separated clusters. Graph A is correctly separated clusters in the even dataset, Graph B is the separated clusters in the uneven dataset, and Graph C is a zoomed-in version of an error cluster in the uneven set. In Graph C, the first, third and fourth tracks show match the genome with the highest coverage in the dataset.

3.4 Automated Quasispecies separation of simulated data

The raw Fastq reads prior to separating quasispecies were observed by IGV as they were entangled prior to classification (figure 8). The results of the mixed reads from the homogeneous group are clear (A), and the results generated from the inhomogeneous group (B) are also clear for the dominant genome within it.

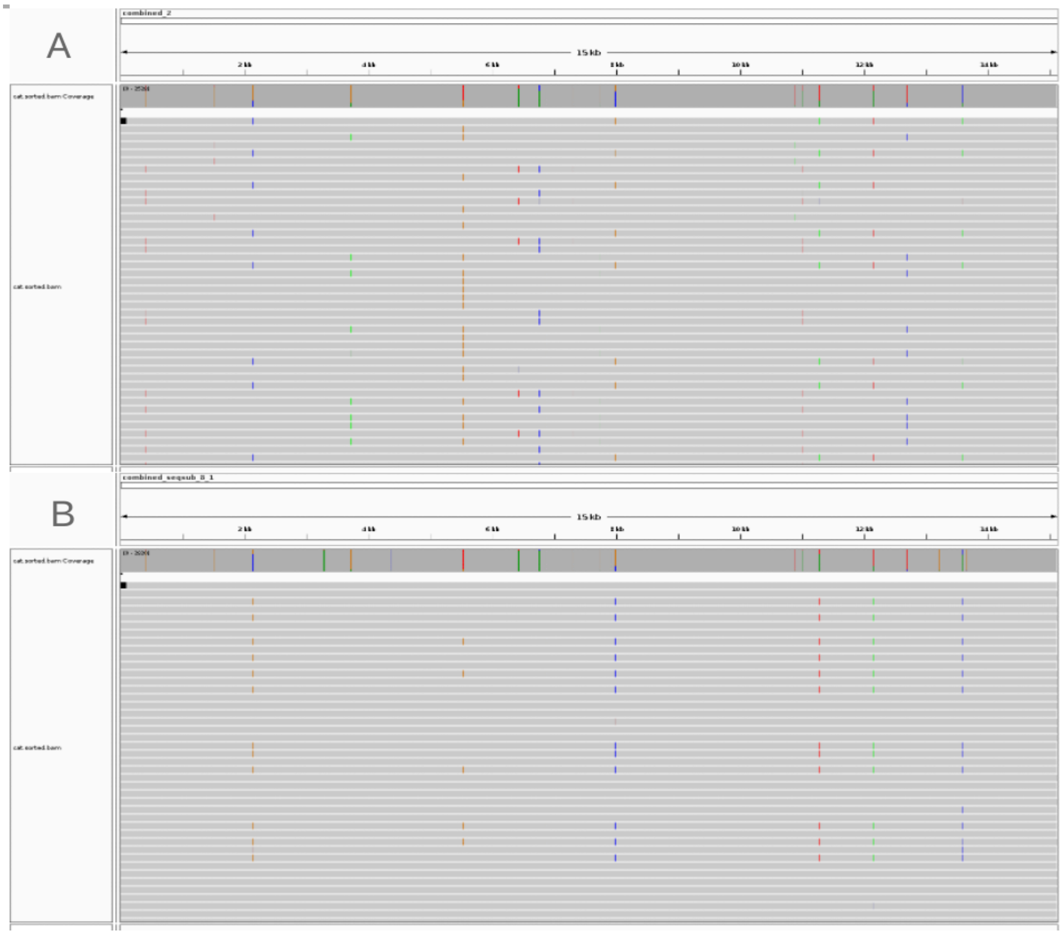


Figure 8: IGV plots before quasispecies separation. Graph A is from the even group, and Graph B is from the uneven group.

The automated pipeline produced good results for both even and uneven coverage sets at the parameters of a CV of 2 and a CR of 50 (Figure 9). Both of them identified all variants and were separated into 10 clusters. Besides, automated Untangle fixed the previous error of identifying high coverage reads as multiple clusters. In the Untangle group (B), cluster #0 is consistent with the reference genome which was polished by the polishing tool to look like the richest cluster. However, some reads from cluster 0 appeared in cluster 2 (marked by a grey circle), but the cause is unclear and requires further investigation.

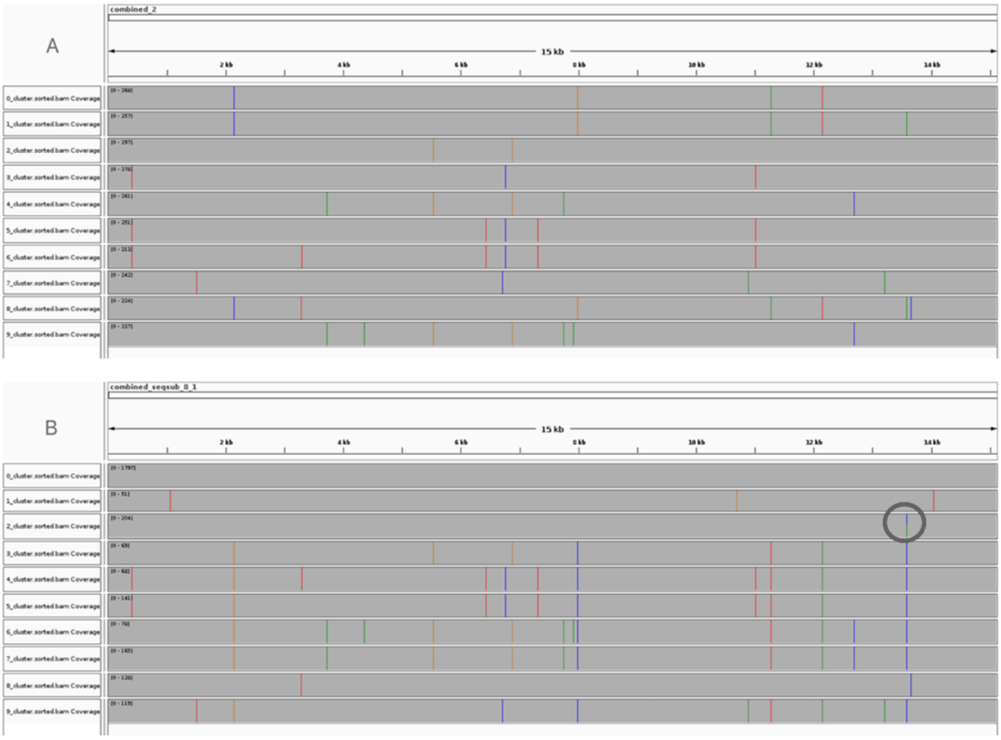

A
combined_2
15 kb
2 kb
4 kb
6 kb
8 kb
10 kb
12 kb
14 kb
0_cluster.sorted.bam Coverage
1_cluster.sorted.bam Coverage
2_cluster.sorted.bam Coverage
3_cluster.sorted.bam Coverage
4_cluster.sorted.bam Coverage
5_cluster.sorted.bam Coverage
6_cluster.sorted.bam Coverage
7_cluster.sorted.bam Coverage
8_cluster.sorted.bam Coverage
9_cluster.sorted.bam Coverage
B
combined_seqsub_0_1
15 kb
2 kb
4 kb
6 kb
8 kb
10 kb
12 kb
14 kb
0_cluster.sorted.bam Coverage
1_cluster.sorted.bam Coverage
2_cluster.sorted.bam Coverage
3_cluster.sorted.bam Coverage
4_cluster.sorted.bam Coverage
5_cluster.sorted.bam Coverage
6_cluster.sorted.bam Coverage
7_cluster.sorted.bam Coverage
8_cluster.sorted.bam Coverage
9_cluster.sorted.bam Coverage

Figure 9: IGV plots of the automatic pipeline separation quasi-species. Graph A is the even dataset. Graph B is the uneven dataset. The locations marked by circles in Graph B show error variants. These results are each mapped to different reference genomes produced by the automated pipeline and so should not be compared directly to one another. However the variants shown match the expected variants for the input genomes.

When the value of CV was increased to 5 (with a CR still at 50), it was found that the pipeline identified only 7 clusters for the inhomogeneous data set. The inspection by IGV visualisation (Figure 10) revealed the same problem that occurred in the inhomogeneous set when the CV was 2, with some reads from cluster 0 ending up in cluster 1 (a). In the other two tracks, the third SNP in cluster 3 is the variant that should have been split (b), appearing in less than 5% of the total number of reads. And in cluster 4, 2 different variants can be seen, namely the 3rd one (d) and the 7th one (c), which might also have been too small in coverage to be identified separately by Untangle using these parameters.

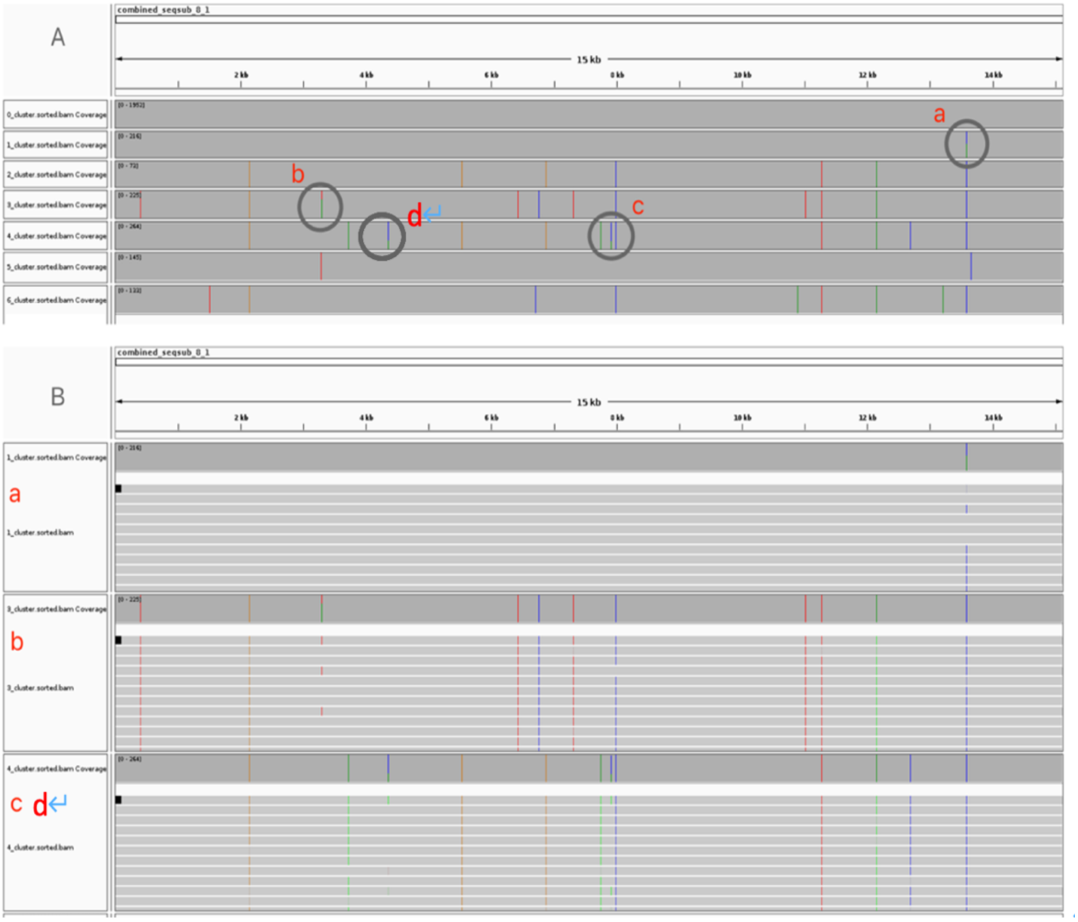

A
combined_seqsub_8_1
15 kb
2 kb
4 kb
6 kb
8 kb
10 kb
12 kb
14 kb
0_cluster.sorted.bam Coverage
[0 - 1952]
1_cluster.sorted.bam Coverage
[0 - 216]
2_cluster.sorted.bam Coverage
[0 - 72]
3_cluster.sorted.bam Coverage
[0 - 225]
4_cluster.sorted.bam Coverage
[0 - 264]
5_cluster.sorted.bam Coverage
[0 - 145]
6_cluster.sorted.bam Coverage
[0 - 132]
a
b
d
c
B
combined_seqsub_8_1
15 kb
2 kb
4 kb
6 kb
8 kb
10 kb
12 kb
14 kb
1_cluster.sorted.bam Coverage
a
1_cluster.sorted.bam
3_cluster.sorted.bam Coverage
b
3_cluster.sorted.bam
4_cluster.sorted.bam Coverage
c d
4_cluster.sorted.bam

Figure 10: IGV plots of the Untangle pipeline separating the uneven groups with a CV of 2 and a CR of 50. Graph A is the seven clusters isolated; Graph B is the IGV plot with the clusters containing erroneous SNPs enlarged. a represents the erroneously assigned SNPs, while b, c and d are the SNPs not identified by the pipeline.

To assess whether the missing clusters were from the least covered sequences, a phylogenetic tree was generated for these seven identified clusters as well as the original reference sequences (EMBOSS simulation data) to identify each cluster (figure11). The comparison of the phylogenetic trees revealed that the missing clusters were from the three least covered genomes (blue triangles), with coverage rates of 2%, 2.5% and 3% of the total reads, respectively.

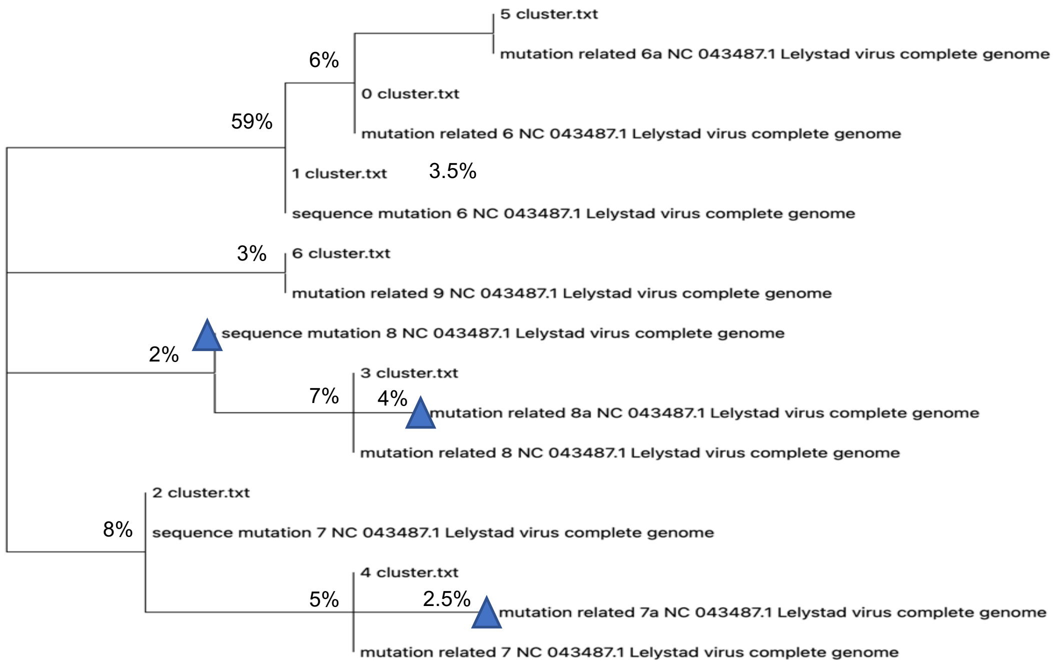

5 cluster.txt
mutation related 6a NC 043487.1 Lelystad virus complete genome
6%
0 cluster.txt
59%
mutation related 6 NC 043487.1 Lelystad virus complete genome
1 cluster.txt
3.5%
sequence mutation 6 NC 043487.1 Lelystad virus complete genome
3%
6 cluster.txt
mutation related 9 NC 043487.1 Lelystad virus complete genome
sequence mutation 8 NC 043487.1 Lelystad virus complete genome
2%
3 cluster.txt
7%
4%
mutation related 8a NC 043487.1 Lelystad virus complete genome
mutation related 8 NC 043487.1 Lelystad virus complete genome
2 cluster.txt
8%
sequence mutation 7 NC 043487.1 Lelystad virus complete genome
4 cluster.txt
5%
2.5%
mutation related 7a NC 043487.1 Lelystad virus complete genome
mutation related 7 NC 043487.1 Lelystad virus complete genome

Figure 11: Phylogenetic tree generated by the seven clusters with the original reference sequences. The segregated clusters correspond to the reference sequences, respectively, characterising the relationships. The coverage levels of each cluster of the original datasets were annotated on the left side. Three original genomes which annotated by blue triangles (sequence_mutation 8, mutation_related 8a and mutation_related7a) lack the corresponding paraphyletic clusters.

3.5 Quasispecies separation of real data

The sequencing results for the real data were not ideal, with a large number of full-length reads in the LongAmp polymerase group (11,365) but very few full-length reads in the Phusion polymerase group (only 443 in total). Since the number of reads in the Phusion group is quite small, it may be impossible to separate quasispecies with sufficient coverage. I have tried using several parameters including CV 5,8,10,15 with CR 20,50, 75, 200, but no good results were produced. In order to see the genome as true as possible, Phusion would only produce clusters with the CV set to 15, but as LongAmp had higher coverage, so it could go as low as 8. The choice of the subsequent CR values was determined based on the histogram generated by the Untangle pipeline (Figure 12), with the red line being the cut-off that was chosen to be the CR. Real data have to contend with not only Nanopore errors but PCR errors. So the distribution on the left is much wider than in the simulated data, which makes picking that cut off just a little more difficult. The Phusion group (A) produced four separated clusters at the parameters of CV 15 and CR 20, and the LongAmp group (B) chose to separate five clusters at the parameters of CV 8 and CR 75.

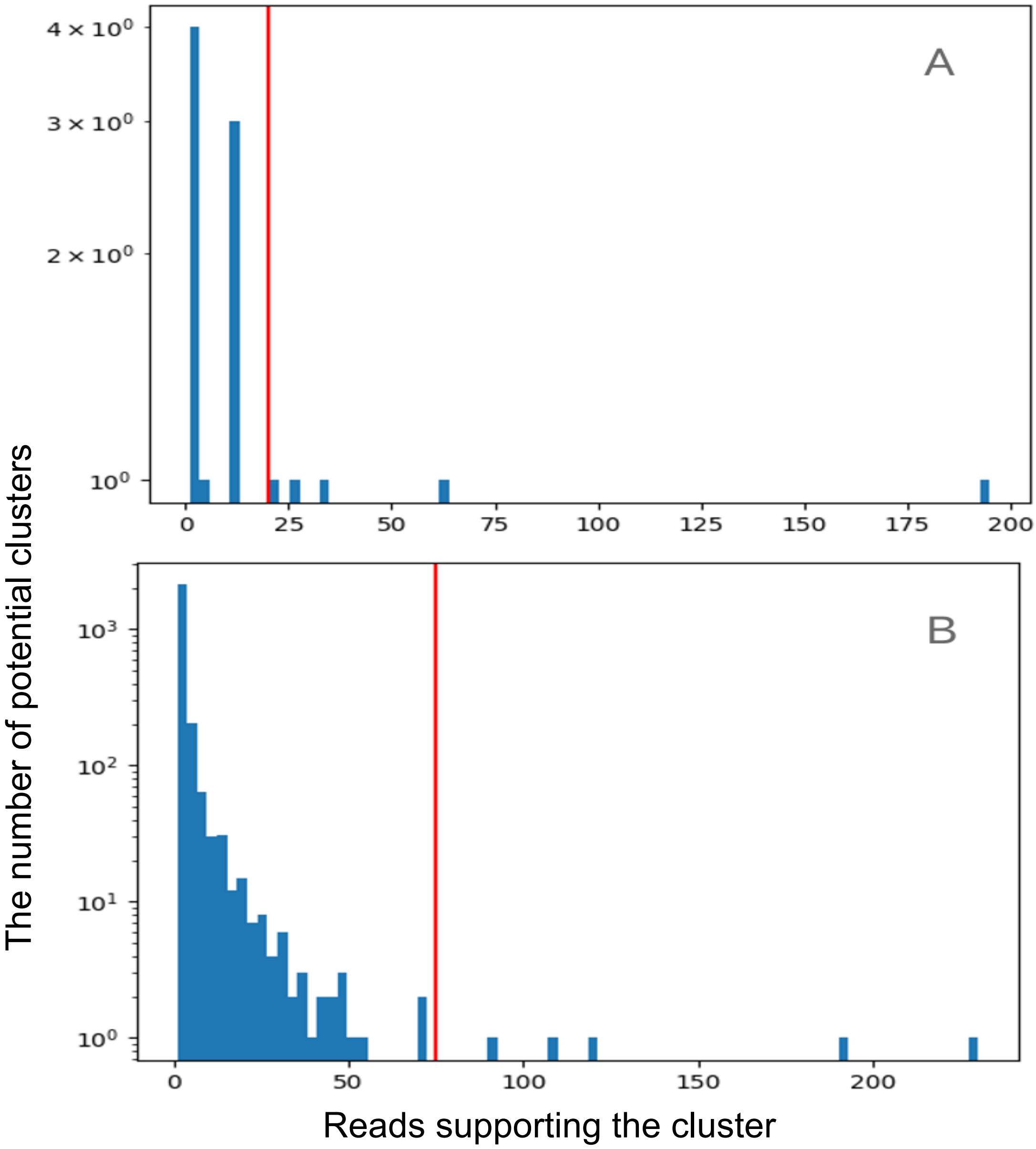


Figure12: The histogram produced by the Untangle pipeline. Graph A is the Phusion group, and Graph B is the LongAmp group. The X-axis shows the number of reads supporting the cluster, and the Y-axis shows with that coverage level. The red line is the dividing line indicated by the CR value. The left side of the red line shows the incorrect clusters excluded, and the right side shows the likely real genomes output by the pipeline.

The Untangle separation of the two polymerases did not give very good results. The LongAmp group simply split the most abundant genome into five clusters (figure 13A). Although the Phusion group has low coverage, it split the SNPs of the genome (figure 13B), while the clusters of the LongAmp group were the same genome.

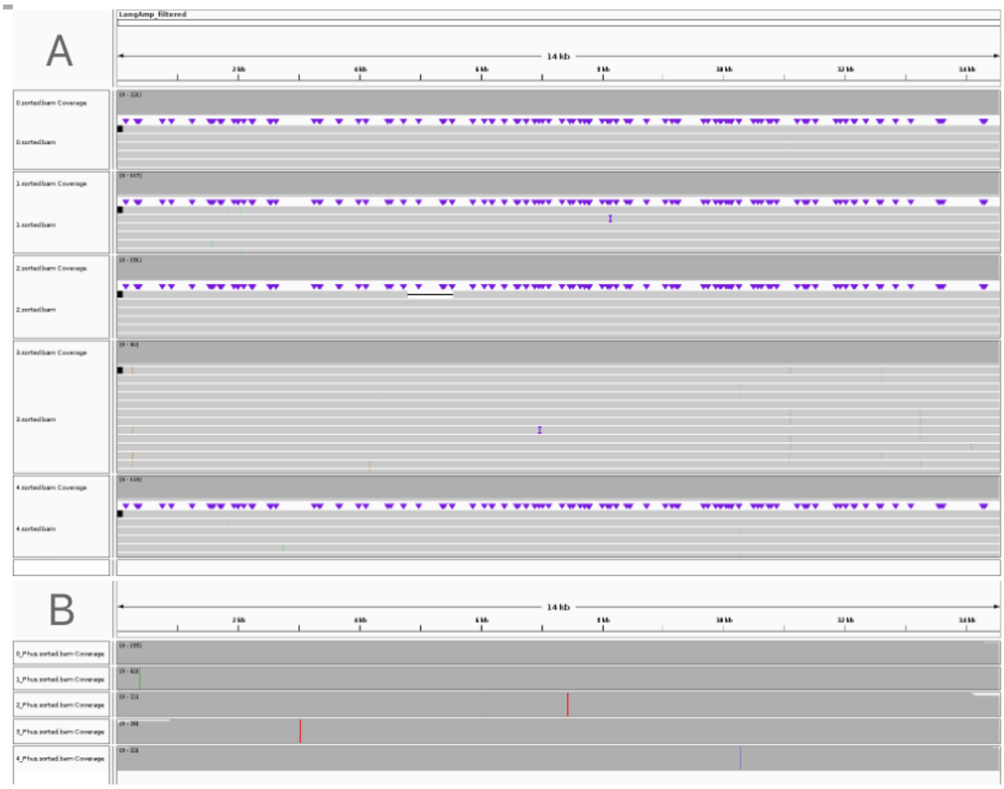


Figure 13: Diagram of two polymerase generating sequences isolating quasi-species IGVs. Graph A is the five clusters isolated from the LongAmp group, which do not appear to be different, and Graph B is the four clusters isolated from the Phusion group.

4. Conclusion and Discussion

PRRSV has been a serious threat to the global pig industry since its discovery late last century [1]. The susceptibility of PRRSV to mutation and recombination makes outbreaks more difficult to control, but has profound implications for the economic development and animal welfare of the pig industry. The accurate and rapid identification and isolation of PRRSV paratypes plays a key role in the development of vaccines and the genetics of the virus. Several methods have been developed and applied for PRRSV diagnosis [64], mainly including antibody detection by serology and nucleic acid detection using PCR-based assays. sequencing of PRRSV has previously focused on the short regions of open reading frame 5 (ORF5) and ORF7 to differentiate strains. Due to technical and financial constraints, complete genomes are rarely included [65 , 66] . PRRSV ORF5 shows extensive genetic diversity and has been used to provide insight into PRRSV epidemiology, but it represents only 5% of the entire genome, so 95% of the genomic information is still used to predict genetic variation.

In the recent few years, the advent of third-generation sequencing technologies for long-read length molecular sequencing has become an important tool in the field of viral infectious disease research. With the advancement in NGS, PRRSV research and diagnosis have benefited greatly. Studies have emerged to use direct RNA sequencing to detect the presence of PRRSV strains without amplification [1]. While these methods can be very efficient and accurate in detecting the presence of the virus, it is impossible to use them to reveal quasi-species and mutations of the virus by high-depth sequencing. Although the PCR amplification step introduces unwanted bias, the long-read cDNA sequencing approach is capable of detecting viral variants at higher depths. To my knowledge, this study is the first to identify and isolate PRRSV quasispecies by high-depth sequencing. Both the rapid detection of the virus in real time and the accurate identification and isolation of the quasispecies are important for outbreak control. Further insights into their genetic variability, especially that in relation to potential functional associations, are needed. This could be particularly important for the development of vaccines and the control over the spread of viruses.

The aim of this study was to assess the feasibility of the PRRSV quasispecies isolation pipeline Untangle using simulated data to identify and isolate real virus quasispecies and generate a consensus genome for each quasispecies genome. The main interests addressed in this study include whether the simulation tool can accurately generate PRRSV genome sequences and Nanopore sequencing sequences, whether the variant calling tool can correctly identify variants without making assumptions about variant reads, at what confidence intervals the Untangle pipeline can isolate quasispecies, the impact of different polymerases on long read sequencing to generate whole genome sequences for subsequent sequencing of real selection of the appropriate polymerase and generation of full-length genes with sufficient coverage when sequencing the real viral genome. Current approaches to genome phasing expect the genome to be diploid or expect a specified ploidy [62, 64], and none of the current phasing tools are applicable in the face of genetic variation in haploid RNA viruses where accurate genome ploidy is not available. In this study, we successfully simulated full-length nanopore sequencing data for PRRSV and tested Untangle, a bioinformatics pipeline developed by Amanda Warr that uses full-length reads with the highest average base mass as a scaffold to identify and isolate viral quasispecies to generate consensus genomes.

A distinct advantage of this project is the use of bioinformatics tools to simulate viral sequences and long-read sequencing, followed by testing of the pipeline. Direct sequencing of samples is an expensive and time-consuming way of obtaining data. Furthermore, obtaining viral genes is difficult. The original viral genome must then be cultured and amplified.

However, they mutate rapidly in culture due to the mutable nature of PRRSVs. Similarly, the choice of the polymerase during PCR amplification affects the accuracy of quasispecies identification. In addition, using simulated methods is easier and more efficient than using real sequencing data and allows multiple attempts at experimentation.

The genome consisting of ten mutation-containing genes was used to simulate the parent-offspring-grandchild genetic relationship, which more accurately represents the rapid genetic variation of PRRSV paraphyletic species in real-life situations, bringing the simulated data closer to the real genome. Interestingly, one group of genes (mutation-related 9) in the second generation of mutation genome was classified as the first generation of genome in the phylogenetic tree, which was probably due to the fact that the first generation of mutation-related 9 (sequence mutation 9) is missing from these 10 mutation genomes. Therefore, MEGA was unable to identify it as a second generation of genome. Long-read Nanopore sequencing of PRRSV was successfully simulated by the Badread tool and split into two groups with uniform and non-uniform coverage. The desired results were obtained through the simulated sequencing of the even dataset by Nanoplots quality testing. Although the uniform distribution was different from the real sequencing, it could be compared with the uneven dataset as the ideal case. The simulated sequencing coverage of the uneven dataset was very heterogeneous, reaching the desired settings and being closer to the real sequencing results. And the quality averages of the uneven dataset reads were not very different from those of the even dataset. So, apart from the differences in full-length genes and the total number of reads, the uneven dataset and even dataset settings were essentially identical.

The success of the simulation data provides a strong basis for evaluating Untangle, however, the setting of the CV and CR of Untangle determines whether the identification and separation of paratypes is successful. With lower CVs, it is less likely to miss true barcodes, but this also means that they may be incorrectly assigned to a barcode that does not exist. This includes the inevitable nanopore sequencing errors in PCR reactions. When the CV becomes higher, there will be more reads per cluster, but there is also a risk of missing a large number of true barcodes. When the CR is low, the pipeline will become more flexible for different numbers of taxa in the sample. However, it may also introduce more clusters of errors. The error correction tool (Medaka) is the most accurate when it has coverage of more than 50 times on the sequences it corrects. Therefore, the cluster was expected to have at least 50 reads. If only the highest confidence genome is needed, the CV and the CR should be increased. If true genomes are required as many as possible, the CV and the CR should be lowered. It is important to maintain a balance between the CV and the CR to find the most appropriate combination.

The Untangle pipeline was very successful in separating quasispecies for enevn dataset, but for uneven dataset or real genomes, separating quasispecies is more complex through the Untangle pipeline. This is because Untangle has to identify and separate not only quasispecies but also nanopore errors and PCR errors. Therefore, the Untangle pipeline has certain requirements for sequence coverage. If the coverage is below 3%, the pipeline may miss true barcodes. However, if the coverage is sequenced to a higher depth or the Untangle pipeline is run multiple times on the same sequence, it may be able to recover them. Untangle can at least accurately identify and separate quasispecies when the coverage reaches 3% of the total population.

Unfortunately, the quasispecies separation of LongAmp polymerase and Phusion polymerase from strain SU1-Bel used in the real data was not very satisfactory. The first difficulty is that Phusion is a polymerase with high accuracy but produces mostly short-read sequences, resulting in a lack of enough full-length genomes in sequencing data to meet the coverage requirement. The second difficulty is that LongAmp, while producing enough full-length sequences, has too much noise due to errors. Thus, it is difficult to reliably find informative SNPs through LongAmp due to its low accuracy, even though noise can be reduced to obtain more accurate clusters by using Phusion polymerase to produce enough full-length sequences, with the same variant threshold as LongAmp. According to the results obtained in this study, of the two polymerases, LongAmp and Phusion, Phusion was more suitable for amplifying PRRSV viruses.

The current evaluation of the simulation tests of the Untangle pipeline showed that the pipeline was able to successfully identify and isolate the quasispecies of the virus for at least 3% of the total population. However, further research and improvement are still needed for tests with real data, including the incorrect isolation of high-coverage genomes and the omission of low-coverage genomes. Further plans include the use of nanopore-length read cDNA for high-depth sequencing of real viruses to obtain high coverage full-length genes, the identification and isolation of real sequencing quasispecies via the pipeline, the test of the limits of the pipeline in the face of real data and the determination of confidence intervals where more quasispecies could be obtained.

Acknowledgments

I would like to express my gratitude to all those who helped me during the writing of this thesis. My deepest gratitude goes first and foremost to Professor. Amanda Warr, my supervisor, for her constant encouragement and guidance. She has walked me through all the stages of the writing of this thesis. Without her consistent and illuminating instruction, this thesis could not have reached its present form. Especially during the tough times of the epidemic, when I was still far away in China, it was Amanda who guided me remotely on how to learn Linux systems. I will never forget the first time I met Amanda at the Roslin Institute, her kind smile helped to ease my nervousness and anxiety. As a freshman to bioinformatics, Amanda's encouragement kept me going when I was feeling anxious about the difficulties.

Second, I would like to express my heartfelt gratitude to Professor Andrew Free, who helped me to define the direction of my choice of topic in this difficult time of the Covid-19. I also owe a special debt of gratitude to all the professors in Roslin Institute and the school of Biological Science, from whose devoted teaching and enlightening lectures I have benefited a lot and academically prepared for the thesis.

Last my thanks would go to my beloved family for their loving considerations and great confidence in me all through these years. I also owe my sincere gratitude to my friends and my fellow classmates who gave me their help and time in listening to me and helping me work out my problems during the difficult course of the thesis.

Reference