Master Thesis

Benchmarking of viral quasispecies assembly algorithms

By Rucha Narkhede

Master of Science in Computer Science – Data Science and Technology track

at Delft University of Technology, Faculty of Electrical Engineering, Mathematics and Computer Science , The Netherlands

25th May 2022

Author

Name: Rucha Narkhede

Student number: 5152240

Master Program: Msc in Computer Science

Faculty: Faculty of Electrical Engineering, Mathematics and Computer Science

University: Delft University of Technology

Thesis Committee

Dr. Jasmijn Baaijens (Pattern Recognition and Bioinformatics, TU Delft)

Dr. Thomas Abeel (Pattern Recognition and Bioinformatics, TU Delft)

Dr. Julián Urbano (Multimedia Computing, TU Delft)

Table of Contents

PREFACE

The time when I started my research for master thesis topic, I was surprised by the wide spectrum of research topics that bioinformatics has and the pace at which this field is evolving. After the outbreak of SARS-CoV-2 in 2019 and following pandemic, RNA viruses gained a lot of attention. I was really drawn towards viruses and keen on spending my research period exploring about RNA viruses and why is it still so challenging to tackle them. I have spoken more about what RNA viruses are in detail for the readers with no biological background in Chapter 1. The following report talks about my master thesis in which eight months (October 2021 – May 2022) were spent in exploring and learning about viral quasispecies and creating benchmarking datasets along with evaluating the performance of various viral quasispecies assembly tools. All these months have been no less than a roller-coaster ride but exciting journey in both research and personally.

I would like to express my gratitude towards my supervisor Dr. Thomas Abeel for addressing the broader questions related to the thesis and preparing me for future challenges. I am also very thankful to my second supervisor Dr. Jasmijn Baaijens who always supported and motivated me and pushed me to do things beyond my capabilities, to look at the bigger picture even when the timeline was tight. I would also like to thank you for inspiring me during the challenging periods. She has always been really calming and helped me focus on completing my task list. I would also like to express my sincere thanks to Dr. Julián Urbano for being a part of my thesis defense committee and looking forward to getting your valuable feedback on the work done.

Lastly, I would like to specially thank my family and friends who supported and believed in me and listened to my crazy ideas during the rough phases of this journey. It would have been really difficult without them.

I wish you enjoy what you read!

Rucha Narkhede

ABSTRACT

Viral quasispecies refers to viral populations that comprises of numerous viral strains closely related to each other due to within-host evolution or co-infection. The reconstruction of viral strain-specific genomes using sequencing reads is referred to as viral quasispecies assembly, and it is also crucial to determine the relative abundances of the viral strains in the mixture for various treatments. There are currently many software tools available to transform NGS sequencing reads into haplotypes but earlier benchmarks of viral quasispecies reconstruction tools were only tested using simulated datasets but do not reflect closely on the real-world scenarios and on virus evolution. In this research, using realistic evolutionary viral populations, we assessed six viral quasispecies assembly tools. The existing real dataset mix that is still being used for experiments is a decade old, so it has become important to create broader and complex high quality real datasets as a new standard for future haplotype caller experiments. We introduce a new high quality benchmarking dataset for viral quasispecies assembly from real samples. The aim of this research is to evaluate extensive performance of six tools approaches that allow for reconstruction of unique viral haplotypes which are necessary to research complex and heterogeneous virus communities thoroughly. A comparative study of the performance of these tools has been done. Based on the results achieved, to improve the haplotypes generated, an existing de novo method is used for reconstructing full-length haplotypes from pre-assembled contigs of challenging mixed samples. In general, this improved the overall accuracy of the assembly and abundance estimations.

1 Introduction

RNA viruses like SARS-CoV, HIV, Hepatitis and influenza are main causes of infection and illness in human[1], [2]. Because of their genetic variety and the infections, they induce, RNA viruses have gained a lot of attention. The exceptional propensity of RNA viruses to acquire resistance to treatment is one of the most challenging aspects of therapy based on viral inhibitor chemicals[2]. For example, a single point mutation has been demonstrated to provide considerable resistance to Human Immunodeficiency Virus HIV after a minimal exposure to several non-nucleoside reverse transcriptase inhibitors[3], [4]. Severe acute respiratory syndrome coronavirus (SARS-CoV) virus is mutating in real time[5]. As the SARS-CoV-2 epidemic unfolds, it is posing a challenge to existing containment techniques. Before we discuss about the rapid mutation rates, it is first important to understand the structure of RNA.

Ribonucleic Acid (RNA) consists of four nucleotides (Adenine(A), Cytosine(C), Guanine(G) and Uracil (U) instead of thymine(T))[1]. Usually, RNA molecules are single stranded unlike the double stranded structure of DNA molecules. Due to this, the RNA molecules are prone to errors during any damages. These errors result in genomic mutations of the virus. These errors are so common that the mutation rates are also high. RNA viruses' fast mutation rate helps them to quickly adapt to novel surroundings, such as host immunological challenges and therapeutic treatments. RNA Viral genomes contains all genetic information for the virus to exist and replicate. Higher error rates, along with fast replication, rapidly result in the establishment of a community of genomes that are remarkably identical also known as viral quasispecies[6]. Viral quasispecies is explained further in Section 1.1.

1.1 Viral Quasispecies

In comparison to bacteria and eukaryotes, viral genomes are comparatively small, yet they are vulnerable to extremely rapid mutation rates [7]. Within individual hosts, virus populations can have a lot of genetic variability. A typical characteristic observed in RNA viruses is sequence heterogeneity[2], [8]. The population of RNA viruses that possess such a characteristic is commonly referred to as quasispecies. Some common diseases and infections caused by viruses like human immunodeficiency virus (HIV), COVID-19, Hepatitis C virus (HCV) are influenced by the dynamic and diverse nature of quasispecies[7]. Quasispecies provide significant challenges in terms of treatment and immune response resistance, resulting in viral evolution in animals and human species[9]. In every replication cycle, the mutation rate in quasispecies can be as high as 10^{-4} substitutions per nucleotide(nt) copied. This happens due to the lack of repair and proofreading mechanisms. Viruses such as HIV multiply rapidly after infection and frequently incorporate mutations into their offspring's genomes. In HIV, due to the absence of error checking mechanisms, the error rate is approximately 3.4 x 10^{-5} per nucleotide due to high mutation rates during replication cycles[6], [10].

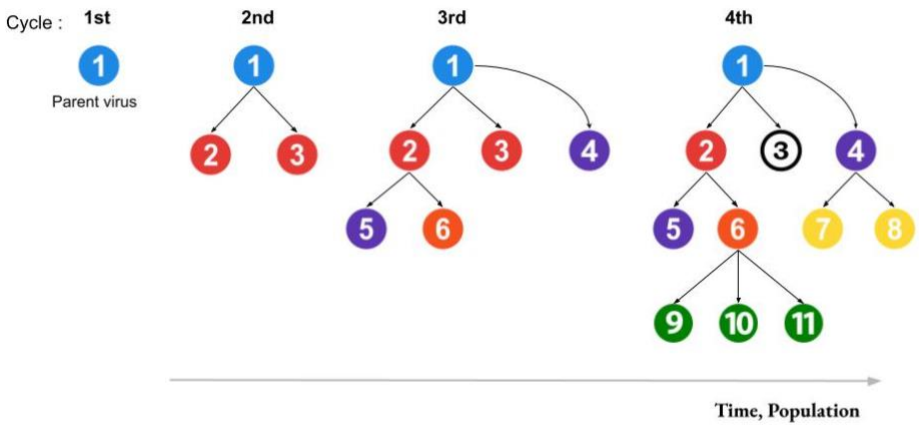

Figure 1:Hypothetical representation of Quasispecies formation

Figure 1 depicts the development of a hypothetical quasispecies, originating at cycle 1(infection genome). The abundance of every individual virus in the quasispecies cloud is

determined by its own replication ability as well as the likelihood that it will originate through mutation of existing individuals of quasispecies cloud. The quasispecies population exists as a whole structure so any selective pressures cannot be applied to an individual mutant but to the whole cloud. In the first cycle, it starts with a parent viral genome(infected). The initial parent virus replicates as the time progresses from cycle 1 to cycle 4. Simultaneously, the population size increases along with the introduction of new viruses to the quasispecies cloud. The new viruses added to the community, mutates and replicates in a similar way producing new offspring, rapidly resulting in a population of extremely similar viruses. The fitness of these descendants is related to the host's fitness. The fitness of descendant mutants is based on the survival of the host. In the given figure, node number 3 is eventually going to die out with time. The parent genome and the offspring's genomes will continue to replicate independently if the mutant offspring is suitable for existence in the host, along with relative abundances being proportional to their fitness. At a molecular scale, determining the genetic diversity of a viral quasispecies would aid researchers in better understanding the behavior and development of quasispecies-forming RNA viruses.

1.2 Genome sequencing

A machine called a sequencer can process physical DNA molecules and determine the corresponding sequence of nucleotides. Unfortunately, sequencersare unable to process a full genome at this time. Moreover, the reads produced by a sequencer may have errors that do not exist in the genomic information. A mismatch error is one form of mistake that can be introduced while sequencing using a sequencer. For instance, a sequencer could report an A when in the genomic sequence it is a G.

In the last decade, with the development of next-generation sequencing (NGS), it was possible to produce genome wide data in a short period of time and at a relatively lower cost. This transformed the fields of transcriptomics, genomics, medical research and evolutionary biology[11]. It is now possible to sequence complete genomes of practically any organism at a sufficient coverage. The identification of haplotypes and their occurrence in diverse virus species is progressing thanks to the advent of NGS methods[12]. Each of the

three methods can be used to examine viral genomes, but none of them can reliably sequence sections longer than a few hundred bases.

Thereafter, a new set of sequencing technologies known as third generation sequencing (TGS) or long-read sequencers have been available. These technologies can generate reads up to 900 Kilobase pairs(kbp) length[13].

Paired-end reads are created when the fragment size employed in the sequencing procedure is substantially larger (usually between 250 and 500 bp) and the ends of the fragment are read in towards the middle[14]. There are two "paired" readings as a result of this. One from the left end of a piece and one from the right end, separated by a given distance. While long read sequencing has a lot of potential but they also increase the error rate significantly (approximately 13%) and there is a lot of NGS data in the sequencing reads archives. In this thesis, we have worked with paired-end reads. The algorithms directly used in this thesis were developed to deal with NGS datasets, particularly Illumina sequencing technologies.

1.3 Assembly

The crucial task is to rebuild a genome using the sequencing data. This process is termed as genome assembly. Usually, it is done in two steps. Firstly, the sequencing reads produced by the technologies are utilized to construct continuous sequences of the maximum length, known as contigs. Secondly, these contigs are further connected together to form a sequence of contigs[15]. The order of these contigs depends on the order of these contigs as they appear in the genome wherever possible. These sequence of contigs are known as scaffolds.

A viral quasispecies assembly, in theory, describes an infection's genetic diversity by showing all the viral haplotypes. Along with the haplotypes, it also shows the abundance rates of each strain. There are two key obstacles to overcome. In most cases, the number of distinct strains is unknown[16]. Furthermore, strains can be different from each other by only tiny amounts of mutations. The abundance rates might be as low as sequencing error rates, making it difficult to discover real mutations that occur at low frequencies. Reference genomes with high-quality consensus genome sequences may be outdated at the time of

the disease epidemic due to the large variety and high mutation rates[15]. As sequencing reads produced by technologies discussed earlier contain errors, with error rates changing by sequencing technology, error correction is a crucial part of genome assembly[17]. By using specific tools for correcting errors on the datasets before actually assembling, some assembly tools presume the input is free on any errors. While some assemblers don't require any preprocessing and deal with sequencing problems during the assembly process. One of the most challenging aspects of genome assembly is distinguishing between sequencing mistakes and actual genomic variation. Many viral quasispecies assembly procedures are hampered by the absence of a relevant reference genome. We will be talking about the two types of genome assembly methods: Reference-free (de novo) and reference-guided assembly in detail in Chapter 2.

1.4 Research Objectives

The aim of this research is to evaluate performance of available approaches that allow for thorough reconstruction of unique viral haplotypes which are necessary to research complex and heterogeneous virus communities thoroughly. Over the last decade, several genome assembly methods have been developed to help with the problem of determining sequence variations from NGS data. Until now, these methods have only been evaluated on the old 5-virus mix which contains real samples[18] or simulated datasets. The work done in [19]also evaluates the performance of the tools using simulated datasets. We studied the challenges faced by these methods on real datasets. It is also important to understand why this research focuses on real datasets. In general, real datasets are messier and much more difficult to deal with compared to simulated datasets. Simulated datasets or synthetic datasets mimic the real datasets, but they do not completely reflect the real samples. There is a wide range of methods to simulate datasets, but they come with limitations. Usually, tools perform better on simulated datasets. To evaluate performance of these assembly tools on real dataset, we created benchmarking datasets from real dataset samples.

The first part of the research is to create a high-quality benchmarking dataset for viral quasispecies analysis which is obtained from real samples. This study is highly motivated by the work done in [18]. The existing dataset in [18] is still being used which is older than nine

years. Due to the fast-evolving nature of viruses, this dataset is ancient. There are a lot of other issues along with this which are discussed in Chapter 5. The second task in this research is to perform benchmarking experiments of six haplotype reconstruction tools (de-novo and reference based) on the benchmarked dataset in the first step and on the simulated dataset. Once the evaluation of these existing methods has been done, the next task is to figure out if the results obtained can be improved. For that, we use VG-Flow method implemented in [20]. VG-Flow method consists of two steps. The first step is to build a variation graph followed by haplotype reconstruction. The original algorithm uses VG toolkit[21] for constructing variation graph which comes with limitations. VG toolkit is updated frequently which makes it difficult to work with the older versions of the tool as they are not compatible with the new versions. There are various other issues faced while using VG. To overcome these issues, we implemented Seqwish[22] instead of VG to build the variation graph followed by the second step of VG-Flow.

This research and its contributions can be divided into three parts and are briefly listed below :

1. A benchmarking dataset for viral quasispecies assembly from real data.

2. Perform benchmarking experiments results of different assembly tools on the above dataset and simulated datasets obtained from and evaluate on different metrics using QUAST[23].

3. Perform experiments using VG-Flow with pre-assembled contigs produced by the assembly tools in part 2.

1.5 Report Outline

To make it easier to achieve a natural flow, the report is divided into four parts. Each part is presented as chapter. All the chapters are linked to each other.

Chapter 2 talks about various viral quasispecies genome assembly methods and how they work. Chapter 3 describes about the benchmarking real datasets design for the experiments as well as existing simulated datasets. Chapter 4 and 5 talk about the metrics used for evaluation, results and interpretation of the results.

INTRODUCTION	VIRAL QUASISPECIES GENOME SEQUENCING GENOME ASSEMBLY RESEARCH QUESTION REPORT OUTLINE
VIRAL QUASISPECIES GENOME ASSEMBLY METHODS	DE-NOVO ASSEMBLY REFERENCE BASED ASSEMBLY
BENCHMARKING OF DATASETS	BENCHMARKING OF REAL DATASETS BENCHMARKING OF
BENCHMARKING RESULTS	QUAST ANALYSIS AND INTERPRETATION OF RESULTS
DISCUSSION	IMPLICATION OF THE RESEARCH LIMITATIONS OF THE RESEARCH FURTHER RECOMMENDATIONS FOR FUTURE RESEARCH

Table 1: Report Framework

2 Genome Assembly Tools

To support the ever-growing amount of sequencing data, a range of sequence assembly techniques and tools exist. Furthermore, the diverse nature of data generated by various sequencing platforms necessitates the development of specific algorithms for each type of sequencing data.

2.1 De-Novo Assembly

De novo genome assembly presumes no prior knowledge of the length, architecture, or composition of the original DNA sequence. The DNA of the target organism is broken up into millions of little fragments and read on a sequencing machine in a genome sequencing study. Depending on the sequencing method, the length of these "reads" can range from 20 to 1000 nucleotide base pairs (bp). A read's position is determined by identifying reads that overlap because of similar sequences when compared to the true sequence (true positives) and overlaps because of sequencing errors (False Positives). Until a low coverage sequence region is met, the true positives are clustered into a continuous set of read sequences. The advantage of a de novo assembler is that it can predict placement of reads within a continuous region, but it can be difficult for a de novo assembler to place the reads in fragmented regions. For this purpose, some de novo assemblers employ the mate-pair reads' information to aid in allocation of placement locations and fragmented regions.

For de novo assembly approach generally sequence graphs are used. All the de novo assembly methods use sequence graphs mostly like de bruijn graph or overlap graph for viral quasispecies reconstruction. These assemblers depending on its type either focus on reconstructing all the strains in the population or are consensus-based. The primary purpose of de novo (consensus-based) methods can also be to produce a better reference genome that can then be utilized as a template sequence for further detailed studies. In our experiments, we have only used strain-specific de novo assemblers.

The final genome sequences produced by viral quasispecies assemblers like SAVAGE and Haploflow do not always reflect the full-length haplotypes. Haploflow integrates resolving strain-level sequences and assembling haplotypes for a strain-resolved viral genome assembly[24]. This is achieved by combining de bruijn graph assembly (fast metagenome assembler) along with obtaining a customized sequence flow algorithm that captures variations in the strains. This also helps in linking strain-variants which does not co-occur[24].

While Haploflow is based on de bruijn graphs, SAVAGE is based on overlap graphs. After constructing an overlap graph, SAVAGE joins overlapped read pairs. At the next step, SAVAGE iteratively merges reads into contigs and contigs into scaffolds using clique enumeration and contig formation. Finally, the tool uses Kallisto to estimate frequencies of the resulting haplotype. SPAdes was traditionally developed for bacterial genomes. SPAdes also uses de bruijn assembly approach. The recent development of assemblers like VG-Flow are developed to complete strain-specific assemblies using pre-assembled contigs produced by assemblers. VG-flow is based on flow-variation graphs and tries to convert strain-specific contigs into full-length haplotypes taking into account their abundances. VG-Flow is divided into two steps. Firstly, it constructs a variation graph from the pre-assembled contigs provided as an input and secondly, reconstructing the individual haplotypes present in an assembly[20]. Originally, the VG-Flow algorithm uses vg toolkit to build the variation graphs. Following table shows a brief description of the de novo assemblers that we have used for our experiments.

Software Tool	Published year	Last Updated	Programming language	Abundances of the strains
Haploflow	2018	2021	Java 6	yes
SAVAGE	2014	2014	C ++	Yes
SPAdes (Generic Assembler)	2013	2022	Java 7	yes

Table 2 : De-novo assemblers used for experiments in this research

2.2 Reference-Based Assembly

In reference-guided assembly, as the name suggests, one or several known true genome (reference) sequences are used to assemble the genome under examination[15]. In reference-based assembly, a couple or several genome sequences are aligned to check for similarity; this process is called sequence alignment or read mapping. For accurate reconstruction of sequences, a high-quality reference genome is required[25]. The assembled genome sequences are often biased towards the genome sequence used for the assemblies. Rather than utilizing a single genome sequence, this bias can be minimized by employing a set of genome sequences capturing variants in the given population. Reference-guided assembly is substantially more computationally efficient than de novo. This only holds true given if the existing assembly is sufficiently comparable to the genome that is being assembled[15]. The capability to rebuild full-length haplotypes is the fundamental benefit of reference-based assembly approaches over de novo assemblers.

Reference based assemblers use variety of ways to reconstruct the haplotypes. CliqueSNV being a reference based assembler, assembles a graph using the information of links among variations of single nucleotide. Further, it identifies true strain variants then merges cliques while assembling the graph[26].

QuasiRecomb and HaploClique both being pretty old tools use distinct techniques and applying these techniques to the problem of viral variant reconstruction were novel. HaploClique allows for huge insertions and deletions. It is also built in a way that detects point mutations. QuasiRecomb tries to incorporate the existing knowledge of recombining of sequencing as events into viral mutations and evolution. HaploClique first reconstructs reads that may potentially represent haplotypes. This is done by enumerating maximum cliques and inserting a size distribution in a given viral network[27]. The complex implementation of Maximal clique enumeration makes it computationally expensive which in turn affects the resource requirements for HaploClique on data sets with coverage more than or equal to 1,000×. Finally, QuasiRecomb utilizes data parameters of a hidden Markov model for estimating point mutations and recombination events. These parameters allow estimation of the probability of each possible haplotype with respect to the observed read data[19].

Software Tool	Published year	Last Updated	Programming language	Abundances of the strains
CliqueSNV	2018	2018	Java 6	yes
HaploClique	2014	2014	C ++	Yes
QuasiRecomb	2013	2013	Java 7	yes

Table 3 : Reference-based assemblers used for experiments in this research

It should be noted that there are various de novo as well as reference-based assemblers that supports reconstructing of viral quasispecies sequences. We did an extensive research on most of the tools based on graph used, publications, what type of input does it support, if error correction is needed for the input reads, output generated and how regularly is the tool updated. The qualifying factors for the tools were if the tools were developed purely for viral quasispecies assembly, whether the tools are used in practice or has many issues (if it does it was discarded as an option), how frequently is the tool updated or how well is the repository maintained. We also tried to include tools which were recent rather than older tools like QuRe which requires a really high memory limit and does not work on very complex datasets. For a better comparison, we tried to include both de novo and reference-based assemblers. Based on this, we shortlisted six assembly tools which are discussed above.

3 Benchmarking Datasets

3.1 Description of the dataset used

For the purpose of evaluation and benchmarking experiment results of different assembly tools, we have conducted experiments on both benchmarking real data mixtures and simulated datasets. For the benchmarking of real dataset for viral quasispecies assembly, we obtained the dataset sample from Harvard lab. In addition to the real datasets, we have also done our experiments on existing simulated datasets simulated using SimSeq[28]. SimSeq simulates paired-end short read sequences for Illumina[28]. We have used 6-strain Poliovirus Mixture, 10-strain HCV mixture and 15-strain ZIKV mixture obtained from (paper). These datasets have been described in detail below.

3.1.1. Real Benchmarking Dataset

We created three real dataset mixtures for benchmarking of five, four and two strains with varying complexity. It is challenging to deal with real datasets. The dataset is obtained from Harvard Lab of 61 different bacteriophages of about 15kb each. Bacteriophages, in short, phages, are viruses thus having the same structural properties as any other viruses. The only difference is these viruses only infect and multiply in bacteria. We have used phages because of the following reasons. Firstly, it exhibits similar structural features to animal and human viruses. Secondly, phages are also easier to produce in large quantities resulting into larger complex datasets. Lastly, they possess morphological and genetic diversity which makes it a good choice for viral quasispecies analysis.

These bacteriophages were sequenced using NovaSeq[29]. NovaSeq can perform whole-genome sequencing efficiently and is also budget friendly[29]. It can generate up to 6TB and 20 billion reads[29]. The dataset consists of 7 different samples pf read length 2 x 250bp. The average coverage of the dataset is ~50,000x. To analyze the effect of the level of divergence and of different abundance of the strains, we constructed three datasets. From the first sample we created a five-strain mixture of strain divergence of about 2 – 4%. The strain abundance varies from 1 – 60%; 1%, 2%, 7%, 22% and 68% abundance distribution increasing exponentially for all the five strains. For observing the performance of assembly tools with lower divergence, we constructed an additional dataset with four strains with strain divergence of maximum of 4%. The abundance distribution varies from 2-78%. To assess the

performance of the genome assemblers for lower abundances, we created a 2- strain mixture with varying abundances thus producing thirteen datasets: (1%,99%), (2%,98%), (3%,97%), (4%,96%), (5%,95%), (6%,94%), (7%,93%), (8%,92%), (9%,91%), (10%,90%), (15%,85%), (20%,80%), (30%,70%). These datasets provided interesting insights about the performance of the assembly tools. The main aim was to produce a high quality, realistic dataset, for this we only trimmed the adapters to produce realistic results. We constructed datasets for low(100x), medium(1000x) and high(10000x) coverages for all the strain mixtures.Before conducting any experiments using these datasets, the unprocessed NGS Illumina reads were trimmed from ends using bbduk[30] and Skewer[31] to get rid of any adapters. Adapters are short synthetic oligonucleotides that are covalently attached to the ends of RNA or DNA sequences[31]. Some tools require error-corrected reads as an input, for this purpose we have used MultiRes[32].

The following table 2 gives a brief description of the characteristics of the real benchmarking datasets created.

Obtained From	Genome Length(bp)	No. of strains	Abundance Distribution	Pairwise Divergence
Boston wastewater (2021-09-21)	14707-15079	5	1-60%	1-4%
Boston wastewater (2020-07-07)	14689 -15578	4	2-78%	1-5%
Boston wastewater (2021-09-21)	14707-15079	2	1-99%	4%

Table 4 : Characteristics of viral Quasispecies of benchmarking real datasets.

3.1.2. Simulated Datasets

Along with real dataset benchmarks, to expand our research and examine the performance of the six assembly tools on various datasets with varying complexity we also considered three simulated datasets: 6-strain Poliovirus mixture, 10-strain HCV mixture and 15-strain ZIKV mixture. Along with varying complexity, some experiments have already been performed on these datasets so it might also help in better comparison of performance of the tools. HCV and ZIKV are the two most challenging simulated datasets presented in[33]. These datasets were simulated using SimSeq (2 x 250bp) Illumina MiSeq reads.

Poliovirus Mixture

This is a combination of six Poliovirus strains with a total sequencing depth of ~20 000. The haplotypes were derived from the NCBI database of Poliovirus genomes[33]. The simulation of paired-end reads was done at 1.6 -50.8% relative frequencies. Abundance distribution increases exponentially like the real benchmarking datasets: 1.6%, 3.2%, 6.3%, 12.7%, 25.4%, 50.8%.

HCV Mixture

This is a combination of ten hepatitis C virus (HCV) strains from Subtype 1a, with a total sequencing depth of 20 000x (400 000 reads)[33]. The haplotypes were derived from the NCBI database of HCV genomes and pairwise divergence ranges from 6% to 9%. The simulation of pair-ended reads was done at 5-13% relative frequencies, sequencing depth varying from 1000x- 4600x.

ZIKV Mixture

This is a combination of fifteen strains of Zika Virus (ZIKV) retrieved from NCBI database which includes three parent strains along with four mutations per parent strain [33]. This dataset consists of Illumina MiSeq 2 x 300bp reads with a pairwise divergence ranging from 1-12%. The simulation of pair-ended reads was done at 2-13.3% relative frequencies at a sequencing depth of 20,000x.

The following table 2 gives a brief description of the characteristics of the real benchmarking datasets created.

Simulated Dataset	Genome Length(bp)	No. of strains	Abundance Distribution	Pairwise Divergence
Poliovirus mix	7428-7460	6	1.6-51%	1.2-7%
HCV mix	9273-9311	10	5-19%	5-19%
ZIKV mix	10251-10269	15	2-13%	2-13%

Tabel 5 : Characteristics of viral Quasispecies of existing simulated datasets.

4 Benchmarking Results and Analysis

This section presents the results obtained on different datasets presented in Chapter 3 to gain insight into the performance of different genome assemblers, challenges faced with increasing complexity of the datasets. This chapter is divided into three main parts. First part talks about the results obtained on datasets with real mixtures with varying number of strains followed by the results obtained on simulated datasets described above. The last part provides insights on how VG-Flow can be used to improve the quality of pre-assembled contigs by producing full length haplotypes and improving the contiguity of the assembly.

4.1 Performance Metrics

For evaluating the performance of the assembly tools and the quality of haplotype assembly produced by these tools we have used MetaQUAST (Meta Quality Assessment Tool)[34], which is widely used to analyze metagenomic assemblies (genetic composition of any collection of microorganisms is called metagenome) and gives valuable metrics was used. It is a tool that assesses and compares metagenome assemblies by comparing them to the closest reference genome. MetaQUAST is ideal for comparison since it can be used with various assemblies simultaneously. For our experiments we have evaluated the assembly tools on the following eight metrics. These metrics are described as follows:

- *Number of contigs* - Reports the number of contigs present in the assembly produced by the tools.
- *Genome Fraction* - the percentage of target or reference genome covered
- *N50* - N50 score reflects on the completeness of an assembly
- *NGA50* - NGA50 predicts the largest alignment and continuity
- *Absolute Frequency Error Rates* - Absolute Frequency Errors as $\sum_{i \epsilon I} \frac{|x_i - x_i'|}{|I|}$, where x_i and x'_i represent the estimates and true abundances. The strain abundance estimates(x_i) was calculated by adding the abundance estimates of each sequence allocated to each actual true haplotype.
- *Error Rate* – Error Rate is calculated as Summation of mismatches, indels and N-rate
- *Misassemblies*

- *Precision* - To reflect upon the False Positives (FP), we have also calculated the precision ($TP \div (TP + FP)$) where, TP (True Positives) specifies the percentage of accurately assembled strain genomes of all the ground truth genome assembly.

Number of contigs, N50 and NGA50 are important metrics which tells about the correctness and if the assembly is continuous or fragmented. The genome length for which the total assembly length produced by the assembly tools of the blocks that are aligned of length or greater is at least 50% of the total genome length of actual haplotypes is reported by the NGA50 metric. If a genome coverage of 50% is not achieved, the NGA50 value is undetermined. If a contig has a single misassembly, that means a location where the left and right bordering sequences correspond to the actual genomes with an overlap or gap (>1kbp) or align to other strains or strands. Error rate helps to show how accurate the assembly is. It is relative to the genome size. MetaQUAST[23], [34] was also used to check the contigs against the consensus strains of the ground truth. This was helpful in the evaluation of frequency estimates. To evaluate the tools on abundance estimates, we compared the estimated abundances with the true abundance of the strains. For this we computed Absolute Frequency Errors as mentioned above. As we cannot expect an assembly method to predict the abundances of absent strains, we only took into account the strains present in the assembly along with the abundances (i.e., $x_i > 0$). For every assembly, we have evaluated the assemblies by comparing the constructed contigs with the ground truths. The ground truth was constructed using SPAdes[35].

4.2 High coverages result in improved assembly quality

First, we evaluated performance of the de novo and reference-based assemblers on real benchmarking dataset mixtures. Dealing with real datasets is the most challenging for these tools as discussed in previous chapters. For benchmarking purposes, the performance of the genome assembly tools was evaluated on three real datasets with varying abundances and strains. A mixture of five, four and two strains were used. The description of these datasets is mentioned in the previous chapter. The experiments were performed on low, medium and high coverages: 100x, 1000x and 10000x for every mixture.

Haplotype reconstruction of PRD mixtures

1. 5 strain PRD mixture

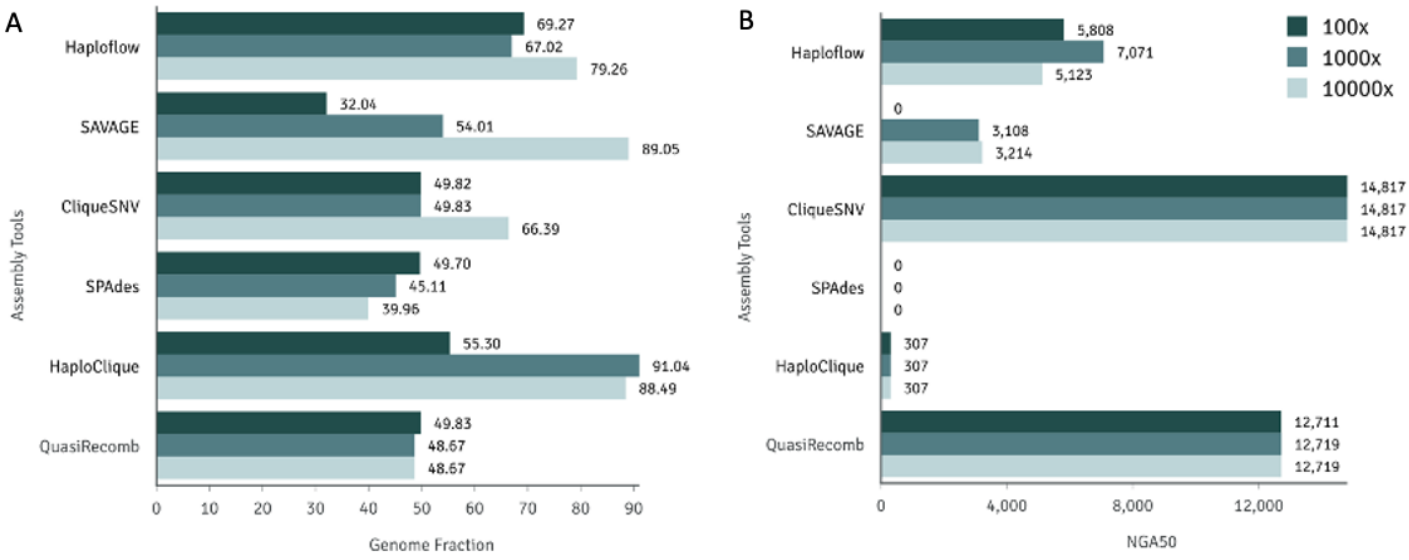

*Figure 2 : Performance of the assembly tools for real 5-strain mixture on all the three coverages on two metrics. **A** show the target genome covered (Genome Fraction (%)) of the ground truth. **B** shows the continuity of the assembly constructed (NGA50)*

Overall, out of these methods, based on all the performance metrics all the tools perform decently well except for QuasiRecomb (Appendix 1). Figure 2 gives a visual idea of how these tools perform on different coverages with respect to the viral quasispecies assembled and the contiguity which also reflects on the length of the contigs. It is quite clear that with increasing coverage, the quality of the assembly also improves for almost all the tools except SPAdes. In case of SPAdes, the NGA50 value is unknown because the genome target covered is less than 50% for all the three coverages.

Haploflow being a de novo assembler could assemble at least more than half of the quasispecies for all the coverages with a decent performance for other metrics as well with a low error rate. The mismatch rate/kb was less than 2 with no False Positives. The quality of the assembly tends to improve with increasing coverage in case of SAVAGE. This shows that the size of the population does play an important role in the results obtained with SAVAGE. As the size of the population increases, performs of SAVAGE also improves. In terms of error rate, SAVAGE outperforms all the other tools. CliqueSNV gives consistent results for low and medium coverages (genome fraction: 49.83%, NGA50: 14817). Even though the assembled genome for 10000x coverage is 66.388% covered but it underestimates the number of strains in the sample. This is because CliqueSNV fails in assembling strains having low frequency. It misses out on the low-frequency strains (<5%) completely. This is further reflected in the

results using 2-strain mixture samples where we study the performance of the tools for low abundances. SPAdes being a generic genome assembler could not assemble half of the target genome. The lower N50 value shows that it produces fragmented assembly. It also produced shorter length contigs for medium and high coverages. HaploClique produced an immense number of fragmented contigs due to which the assembly was substantially fragmented but is also able to assemble more than 90% of the genome assembly for 1000x and 10000x coverage. It's important to note here that, after careful consideration we filtered out the contigs with frequency below 0.05% without losing on the quality of assembly. Firstly, to get rid of the excess contigs and secondly, an attempt to get rid of the fragmented overlapping contigs.

In general, out of all the reference-based assemblers, CliqueSNV performed the best across all the metrics. In terms of misassemblies, all the tools were equally good with zero misassemblies except for HaploClique with one misassembly and QuasiRecomb reported more than eleven misassemblies for coverages equal to or more than 1000x. QuasiRecomb was not able to assemble even half of the genome assembly (48.67% - 49.83%).

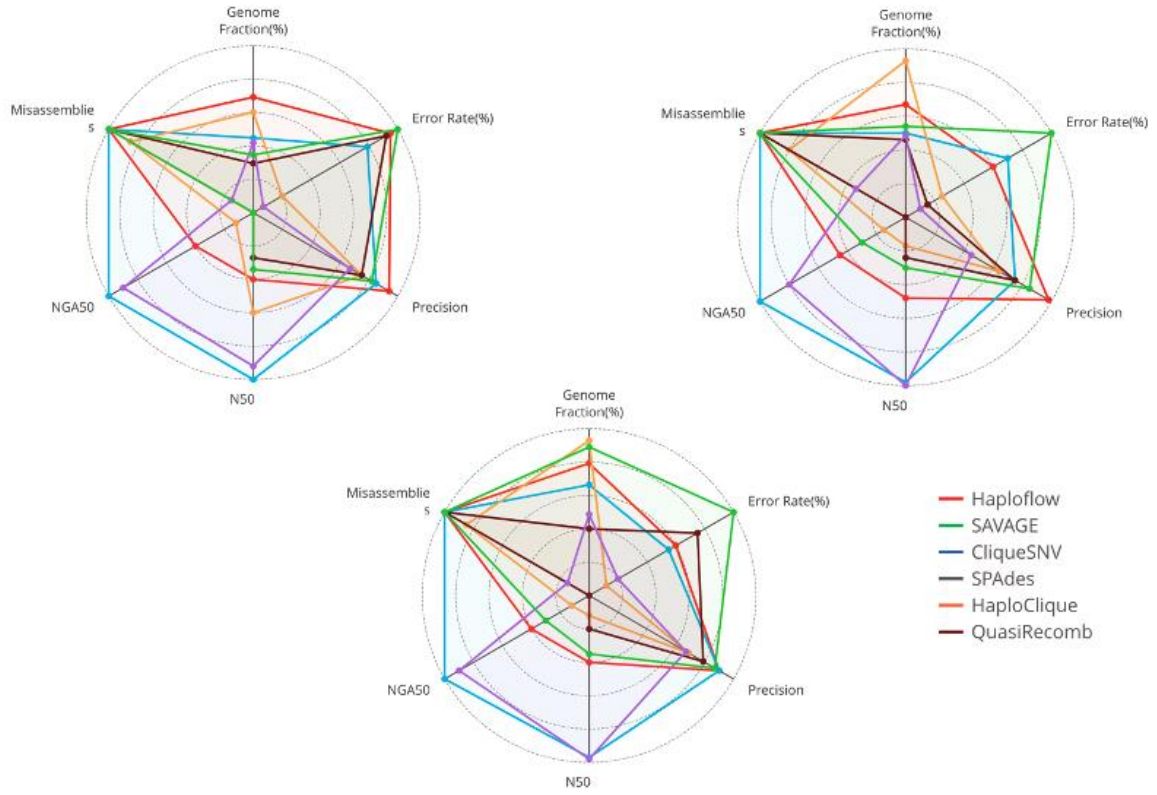

Figure 3 : Radar plot depicts an informative comparison of the performance of the six tools on all the metrics for real 5-strain mixture dataset. The best values are at 100%. ***A, B*** *and* ***C*** *represent 100x, 1000x and 10000x coverage datasets respectively.*

Figure 3 gives an overall comparison of the performance of the tools on all the metrics. The best values are on the outsides, the worst values are in the centerpoint. For 100x coverage, CliqueSNV ranks first in N50, NGA50, misassemblies whereas Haploflow ranks first in Precision, Genome Fraction and misassemblies. HaploClique and QuasiRecomb rank really low in strain precision. For 1000x, similar pattern is observed for CliqueSNV and Haploflow. HaploClique outperforms in terms of total assembly covered but as discussed before produces more number of contigs (lower NGA50) and a strain genome with more mismatches.

2. 4 strain PRD mixture

We next evaluated the performance of these six tools on a four-strain mixture with characteristics described previously. This dataset was comparatively complex to deal with because of more low-frequency strains. Figure 5 gives a visual representation of the genome target reconstructed and the contiguity (NGA50) of the assemblies. Similar trends are observed in both the 5-strain and 4-strain real dataset mixtures, the performance of most of the tools improve with increasing coverage.

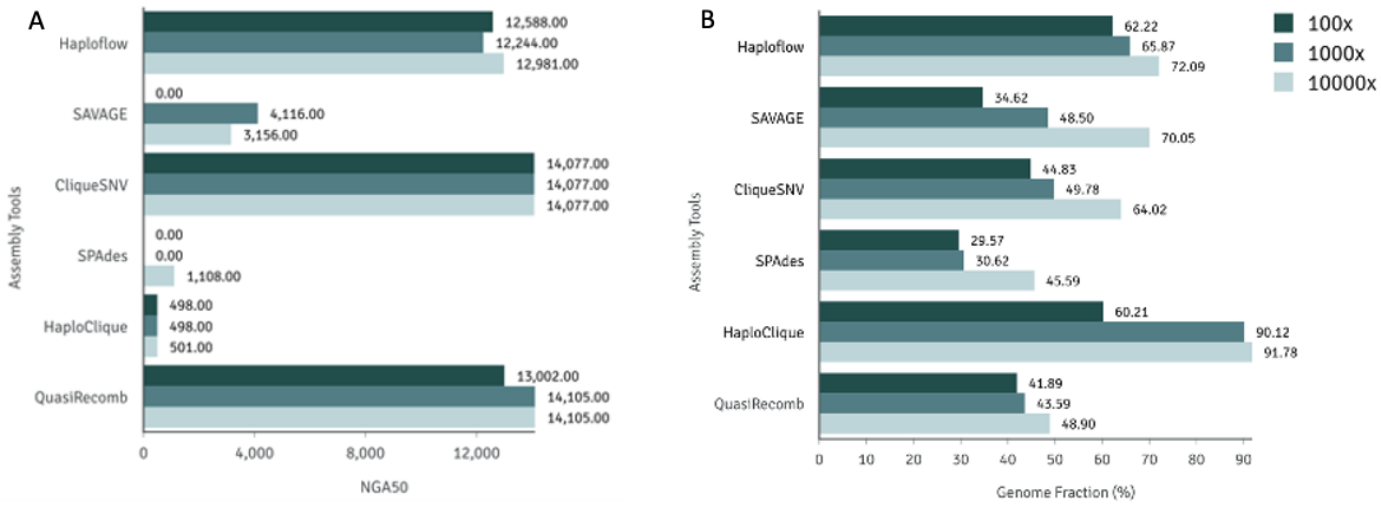

*Figure 4 : Performance of the assembly tools for real 4-strain mixture on all the three coverages on two metrics. **A** shows the continuity of the assembly constructed (NGA50) **B** shows the target genome covered (Genome Fraction (%)) of the ground truth.*

Out of these six assemblers, Haploflow's performance was consistent across all the metrics for all the coverages (Appendix 1). The overall score of Haploflow was better than other tools. In case of SAVAGE, HaploClique and QuasiRecomb, similar trends were observed as in 5-strain mix. SAVAGE was able to reconstruct genome fraction of more than 40% of the total assembly of quasispecies for coverages higher than 1000x. Compared to the 5-strain mix, Haploflow reconstructs a better-quality assembly for 4-strain mix in terms of contiguity (N50:13545- 14841, NG50 :12244-12981) and length of contigs which is quite higher than the 5-strain mix. This is possibly because the strain-wise divergence for 4 strain mixture varies from 2-6% and for 5 strains mix it varies from 2-4%. So, it becomes easier for the assembler to distinguish between the strains. Haploflow and CliqueSNV assemblies were of really high quality for 1000x and 10000x coverage, recovering the most correct strain genomes (2 out of 4 strains). Haploflow and HaploClique was also able to recover more than 60% of the low abundant strain but for HaploClique the assembly was fragmented resulting in poor contiguity along with high error rates (more than 1.5%).

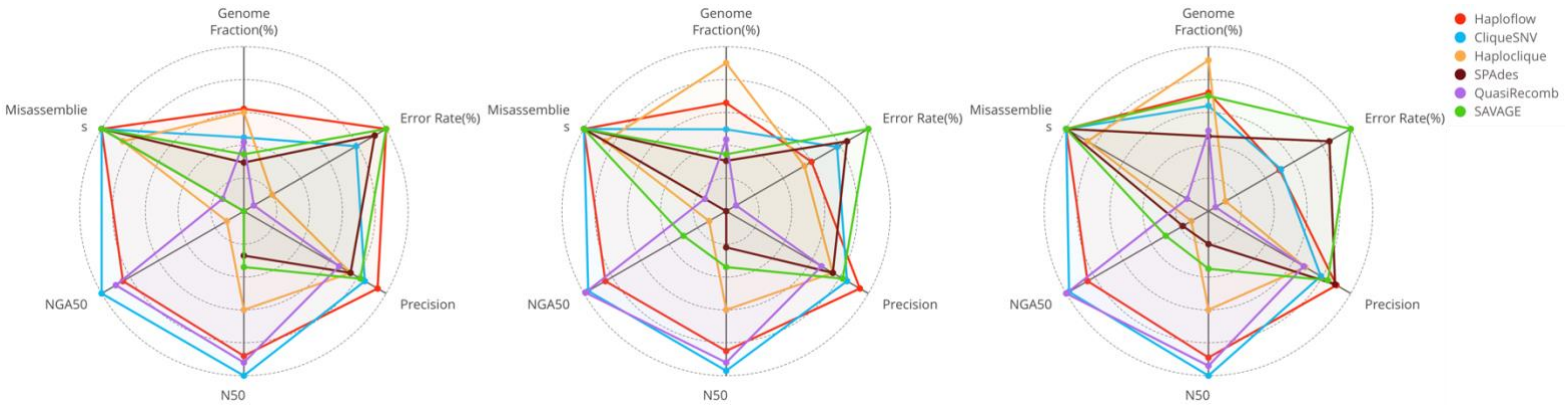

*Figure 5 : Radar plot depicts an informative comparison of the performance of the six tools on all the metrics for real 4-strain mixture dataset. The best values are at 100%. **A, B** and **C** represent 100x, 1000x and 10000x coverage datasets respectively*

Figure 6 shows the comparison of the performance of all the assembly tools on all the metrics for 4-strain mixture. For 100x coverage, CliqueSNV ranks first in N50, NGA50, misassemblies whereas Haploflow ranks first in Precision, Genome Fraction and misassemblies (similar trend was observed for 5-strain mixture). HaploClique and QuasiRecomb rank really low in strain precision. For 1000x, similar pattern is observed for CliqueSNV and Haploflow. SAVAGE outperforms all the tools in error rate for all the coverages. HaploClique outperforms in terms

of total assembly covered but as discussed before produces more number of contigs (lower NGA50) and a strain genome with more mismatches also more false positives.

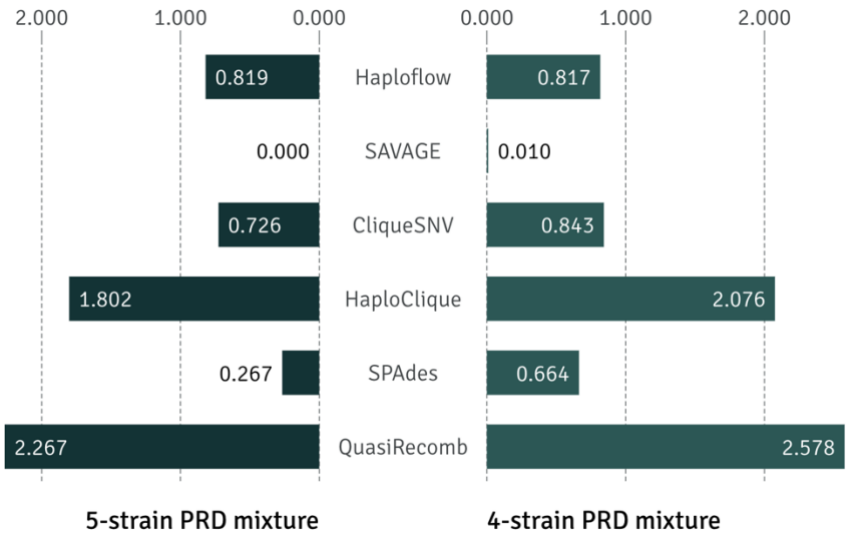

Figure 6 : Comparison of error rates produced by the assembly tools on two different real dataset mixes (5-strain and 4-strain mixture)

The error rate reflects upon the mismatches rate/kb, indels and the N-rate. These metrics measure the correctness of the assembly. SAVAGE performs extremely well with an error rate of almost 0% for both the mixtures. This is illustrated in Figure 4. This chart gives a good comparison of the average error rates produced by the assemblers for both the 4-strain and 5-strain mix for all the coverages. Second best is CliqueSNV and Haploflow (<0.9%). CliqueSNV has a slightly lower error rate than Haploflow. The trend observed in error rates for the both the mixtures is similar for all the tools. QuasiRecomb performance is extremely poor with an error rate more than 2%.

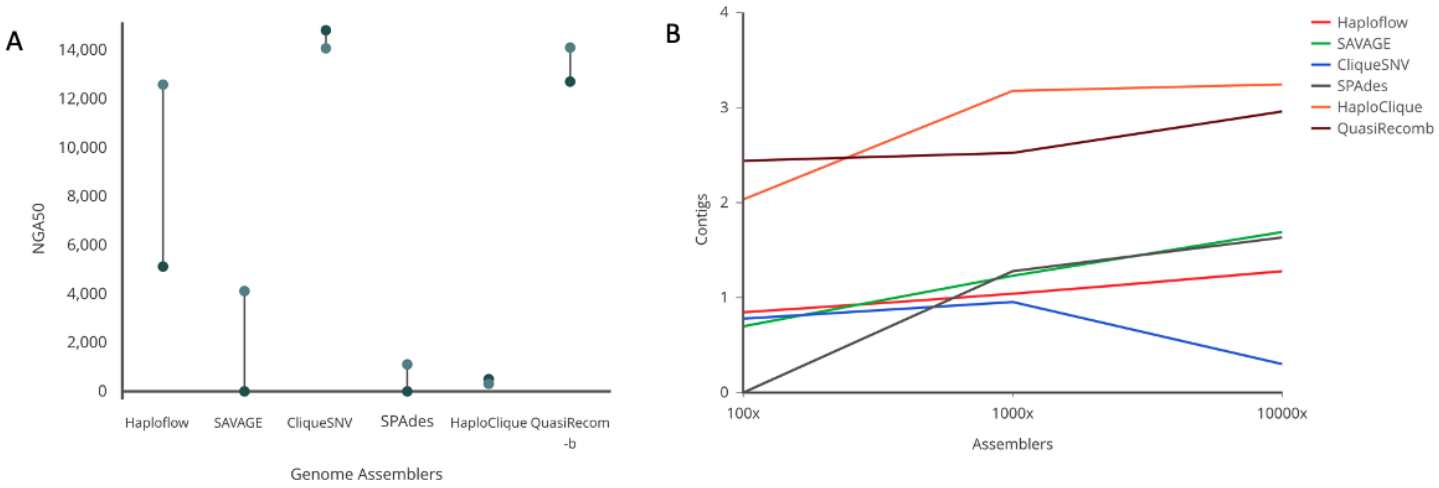

*Figure 7 :Comparison of continuity and quality of the assembly. **A** Range Plot to show the range of average NGA50 observed for 4-strain and 5-strain real dataset mixtures. **B** Line graph to depict the average number of contigs (all the values are taken in log normal form) produced by the assemblers for real datasets on low to high coverages.*

The NGA50 alone reflects upon the contiguity of the assembly but along with the number of contigs taken in logarithmic form, it also reflects upon the quality of the assembly. The comparison of number of contigs and NGA50 produced by different tools for real datasets (4-strain and 5-strain) is shown in figure 8. CliqueSNV performs the best on these two metrics. It produces a smaller number of full length contigs. SAVAGE performs average and produces smaller contigs. NGA50 for Haploflow varies a lot with coverage. This can be because the number of contigs produced for high coverage is also higher than the low coverage. HaploClique produces the greatest number of contigs but lowest NGA50 which also shows that the assembly produced is fragmented (in pieces). If we look at QuasiRecomb's NGA50 values alone, it looks like it performs really well but the number of contigs is also really high for QuasiRecomb which shows that the quality of assembly is really low as the contigs produced are overlapping or erroneous.

3. Performance of assembly tools on low-abundant strains in a mixture

As discusses earlier, abundance estimation and reporting the frequencies of the strains present in a viral quasispecies population is crucial for various treatment and immunity related concerns. In reality, these populations also consist of strains of low frequencies ~1%. To test if these six tools can report or assemble the low abundant strains, we performed the below experiments on a 2-strain mixture. To examine the performance of the assembly tools on low frequency strains we created various benchmarking datasets of 2 strains with varying abundance distributions ((1,99), (2,98), (3,97), (4,96), (5,95), (6,94), (7,93), (8,92), (9,91), (10,90), (15,85), (20,80), (30,70)) at 1000x coverage. The summarized results are mentioned in Appendix 1.

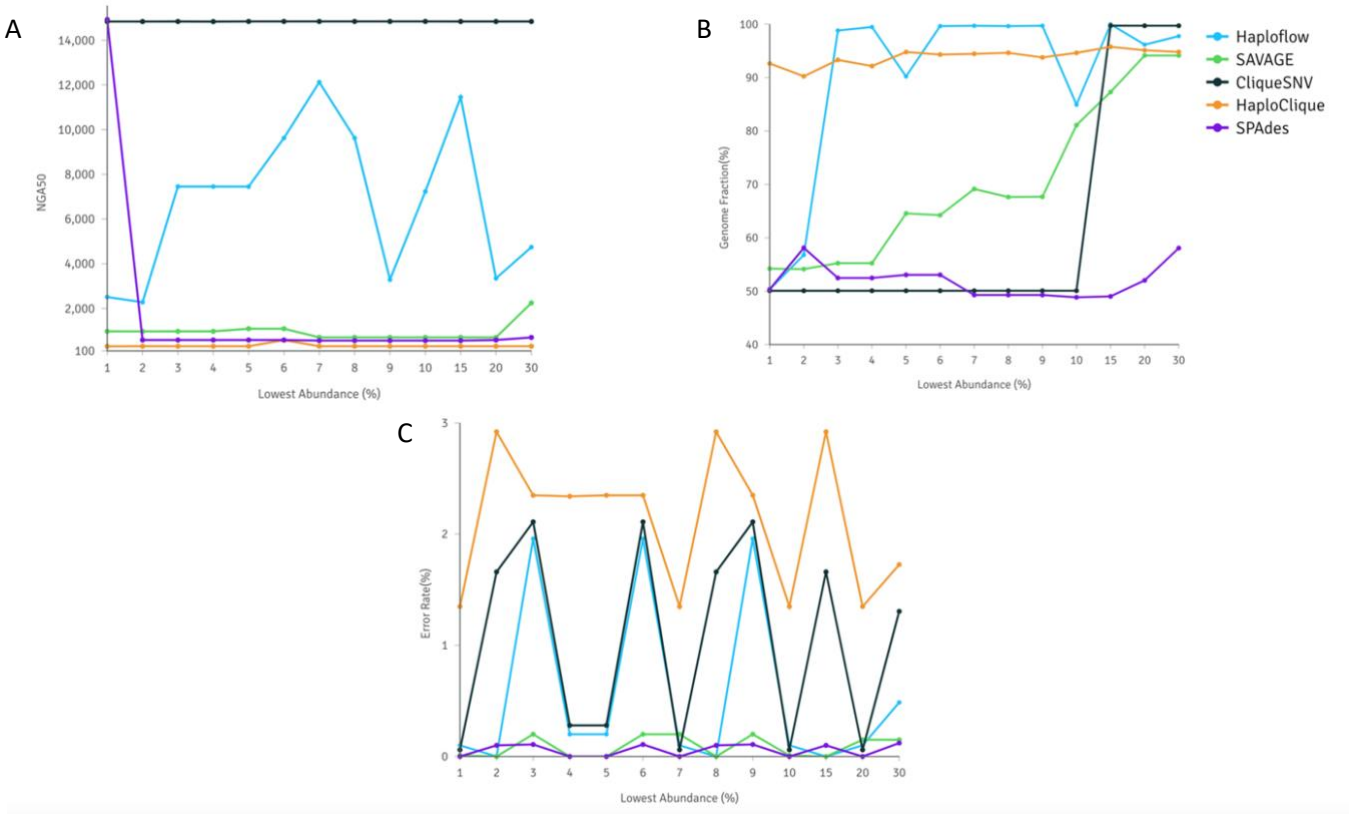

Figure 8 : Performance of the assembly tools on lowest abundant strain from 2-strain mixture real dataset samples. A shows the continuity of the assembly (NGA50). B shows the target genome recovered (genome graction (%)). C depicts the trends in error rate(%)

It is comparatively difficult for the assemblers to assemble sequences of strains with low abundances. We compared the performance of all the assembly methods on all the metrics (Appendix 1). We investigated that Haploflow performed the best in handling the low abundant strains with good quality assemblies. The continuity (NGA50) and error rate vary a lot with different abundances. Haploflow was able to reconstruct the genome more than 90% when the lowest abundant strain is >2%. Whereas

SAVAGE tends to improve linearly after 5% abundance reaching more than 90% for 20% lowest abundant strain. SAVAGE and SPAdes produces approximately negligible error rate. CliqueSNV and SPAdes highly underestimates the lower abundant strain. The 50% target genome reconstructed reflects that it only reconstructs one of the two strains, that is the higher abundant strain with the best NGA50 for CliqueSNV. CliqueSNV produced contigs of length more than 14000bp which did not change with changing abundances. HaploClique was able to recover the target genome of more than 90% for all the abundances but it produces fragmented assembly with huge number of contigs as observed before. The error rate is also pretty high for HaploClique.

Runtime and Memory consumption

The increase in average runtime (seconds) with low to high coverages for 5-strain and 4-strain mixtures is given in Figure 10. All the experiments were performed on TU Delft's HPC cluster (on the same cluster node). Haploflow was the fastest compared to all the other tools with an average memory peak of 15 GB. SPAdes being a generic assembler still was the second best followed by CliqueSNV. HaploClique and SAVAGE took comparable runtime, but SAVAGE's memory usage was higher than HaploClique. QuasiRecomb took the most runtime and Memory Usage.

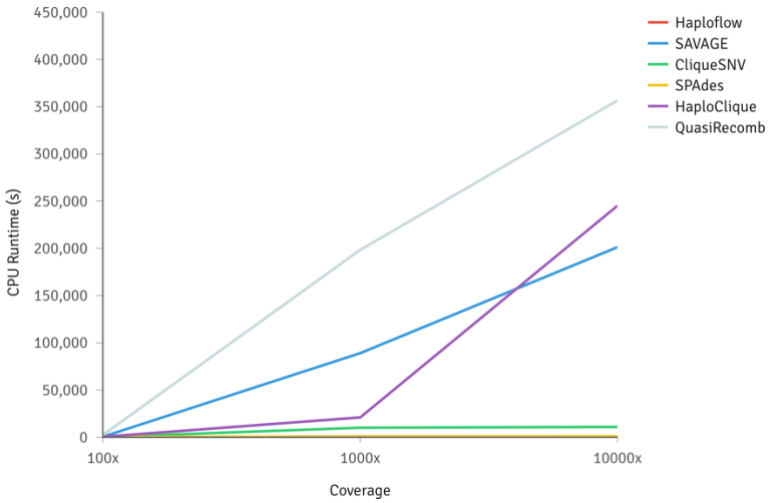

Figure 9 : Runtime comparison of all the six assembly tools on real datasets on all the coverages: 100x, 1000x and 10000x.

4.3 Assembly tools perform better on simulated datasets compared to real datasets

We also performed our experiments on three simulated datasets (2 x 250bp Illumina Miseq reads) with varying complexity. The datasets are from different viral quasispecies infections like Poliovirus, Hepatitis C virus (HCV), Human Immunodeficiency virus (HIV). The detailed description of the datasets is given in Chapter 3. The following section gives valuable insights on how the genome assemblers performed on these datasets.

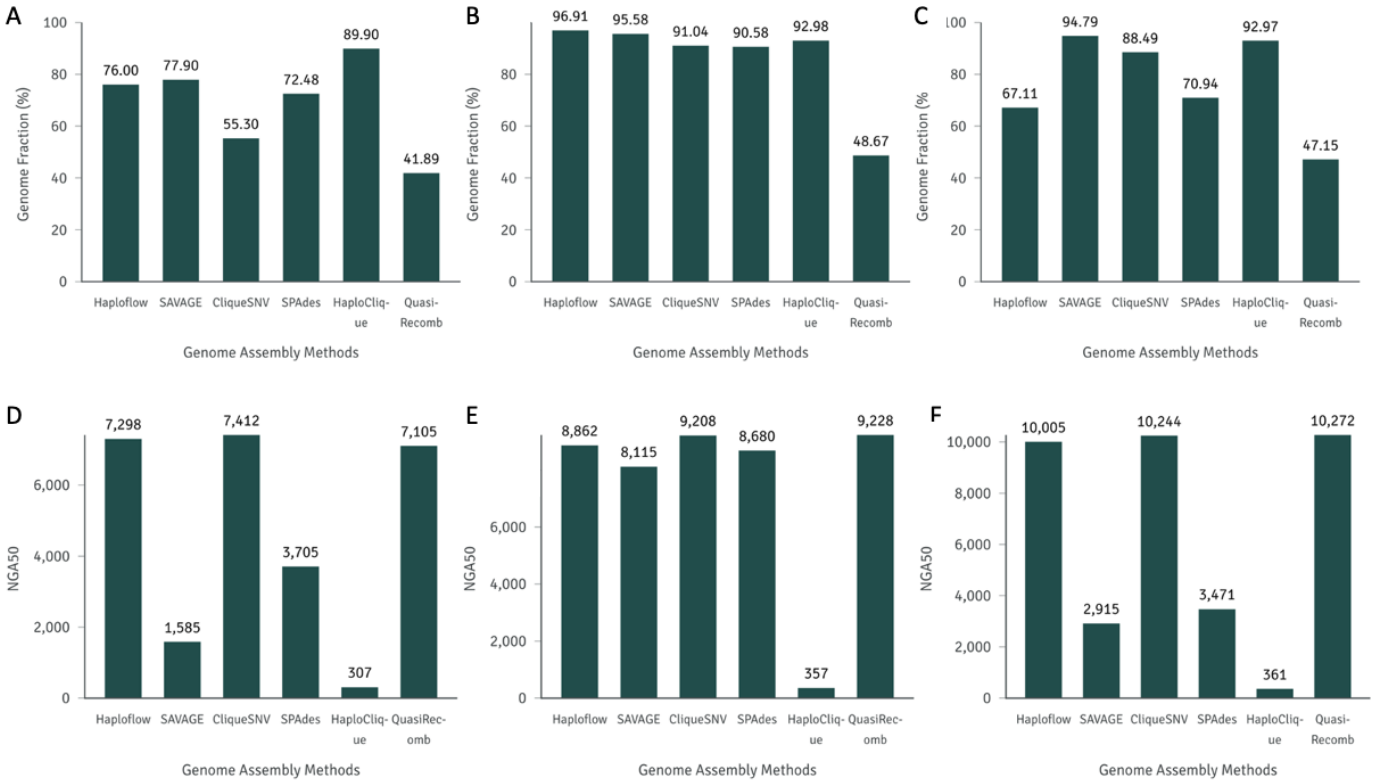

Figure 10 : Performance comparison of all the assembly tools on all the three simulated datasets. ***A, B and C*** *shows the target genome covered.* ***D, E and F*** *depicts the NGA50 achieved by the assembly tools on 6-strain Poliovirus, 10-strain HCV, 15-strain ZIKV respectively.*

The results have been summarized in Appendix 1. Overall, HaploClique and SAVAGE are able to reach a higher target (>75%) for all the datasets compared to the other assembly tools. Overall, the error rate for SAVAGE is also approximately zero whereas HaploClique has a really high error rate that is more mismatches, N-rate and indels (Appendix 1). SPAdes was capable of reaching a genome fraction more than 50% which is assembling more than half of the viral quasispecies with a really low error rate. The target genome covered is the least (~70%) for 15-strain ZIKV mixture as it contains more low abundant strains. As observed before, SPAdes misses out on low abundant strains. The contiguity of the assembly is average compared to Haploflow and CliqueSNV which do not reach a genome target of more than 78% for the 6-

strain Poliovirus mixture and 15-strain ZIKV mixture (Figure 11). For Haploflow, we also observe that the number of total output sequences is two to three magnitudes greater than the true value of number of strains in the sample. On all the simulated datasets, CliqueSNV ranks first in the continuity and length of the contigs which shows that the assembly produced in continuous and not fragmented. Even though QuasiRecomb ranks the second best in continuity, it produces huge number of contigs with the highest error rate (>2%). Among the datasets, all the datasets perform the best on 10-strain HCV mixture. This is because as datasets get more complicated or complex i.e., low abundant strains, more strains and less pairwise divergence we observed higher error rates. It is also important to note that the performance of reference-based assembly tools (CliqueSNV, HaploClique and QuasiRecomb) is highly dependent on the quality of the reference genome.

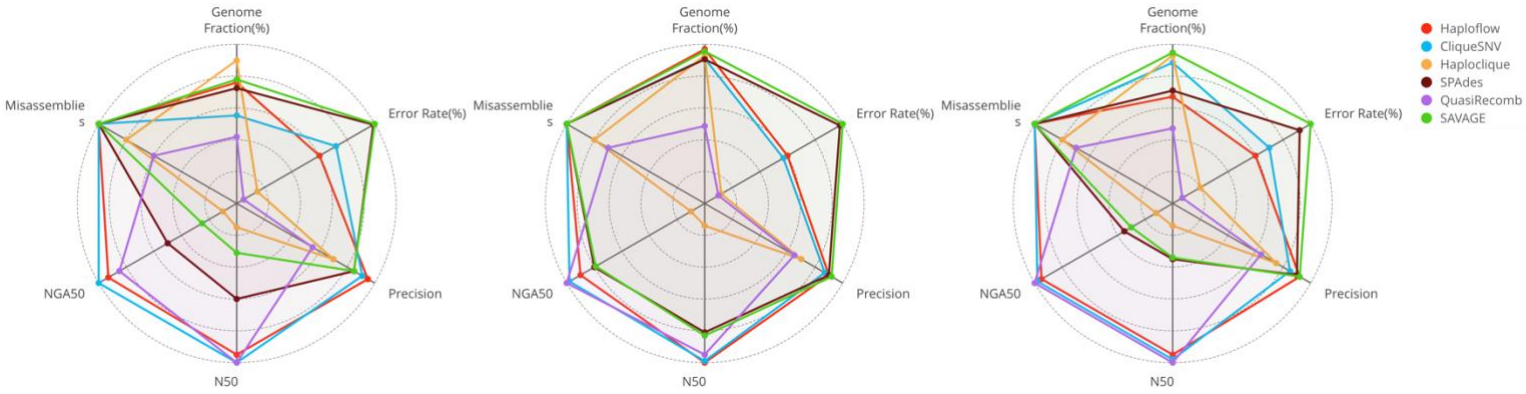

Figure 11 : Radar plot depicts an informative comparison of the performance of the six tools on all the metrics for real 5-strain mixture dataset. The best values are at 100%. A, B and C represent 6-strain Poliovirus, 10-strain HCV, 15-strain ZIKV datasets respectively

Figure 12 gives a comparison of performance of all the tools for all the three simulated datasets. There is no clear winner in this one. As we can observe that some tools perform better on some metrics, and some perform better on other. In general, there are zero misassemblies in the assemblies produced by Haploflow, SAVAGE, CliqueSNV and SPAdes. CliqueSNV, followed by Haploflow produce the most continuous assembly for all the datasets compared to other tools. As observed for real datasets as well, Haploflow performs the best in precision. SAVAGE, CliqueSNV and SPAdes perform equally well on precision metrics (i.e., less False Positives) on simulated datasets compared to real datasets.

4.4 Improved assembly quality on at least one performance metric using VG - Flow

After gaining insights from the results obtained using these six viral genome assembly tools, we see that the length of the reconstructed contigs can be improved. Along with the length, the contiguity of the assembly also can be improved. Most of these tools focus on reconstructing relatively shorter regions of the genomic assembly. Also, some of the tools do not report the abundance estimation of the strains. To achieve better results, we applied VG-Flow. VG-Flow is a two-step assembly method. Firstly, it employs vg toolkit for the construction of variation graphs. Instead of using vg toolkit we have used Seqwish to build the variation graph. VG- Flow is a de novo assembly tool for reconstructing full-length haplotypes from pre-assembled contigs of complex mixtures.

We compared the performance of some VG-Flow + contigs produced by genome assemblers (Haploflow, SAVAGE and CliqueSNV) with the results obtained without VG-Flow. We evaluated the performance of VG-Flow on three benchmarked real datasets: 5-strain mixture (1000x coverage), 5-strain mixture(10000x coverage) and 4-strain mixture(1000x coverage) and three simulated datasets : Poliovirus, HCV and ZIKV mixtures as mentioned before. For input, VG-Flow uses contigs obtained from Haploflow, SAVAGE and CliqueSNV.

Simulated Datasets

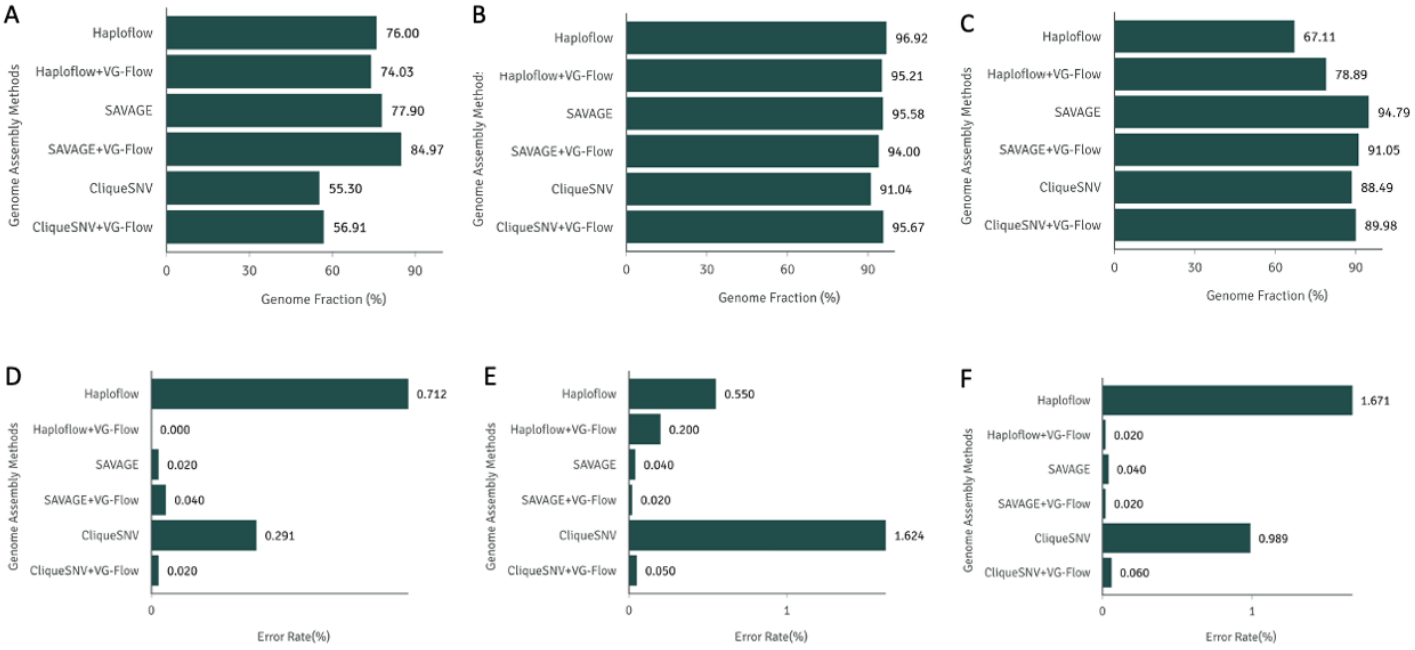

Figure 12 : Performance comparison of Haploflow, SAVAGE and CliqueSNV alone and along with VG-Flow tools on all the three simulated datasets. ***A, B and C*** *shows the target genome covered.* ***D, E and F*** *depicts error rate on 6-strain Poliovirus, 10-strain HCV, 15-strain ZIKV respectively.*

We observed that VG-Flow significantly improves the continuity (Figure 13) of the assembly and lowers the error rate to almost zero for all the assemblies. On all the datasets, we observe that the target genome recovered also improves in most of the cases or remains roughly the same (1% decrease in case of Haploflow for Poliovirus and HCV datasets). This is shown in Figure 12(A, B and C). This can be because VG-Flow algorithm filters out contigs with abundances below a threshold value to ensure correctness of the assembly. On the ZIKV dataset, Haploflow covers 67.11% of the target genome covered which improved with VG-Flow (78.89%). Overall, the number of contigs also reduces while improving the continuity of the assembly. As observed before, SAVAGE produces shorter contigs. With VG-Flow, the number of contigs reduces for all the datasets with an increase in NGA50 (contiguity of the assembly). In general, CliqueSNV produces a continuous assembly along with longer contigs but with a higher error rate going up to 1.624%. VG-Flow doesn't improve the assembly so much in case of CliqueSNV, but it decreases the error rate to ~0.05%.

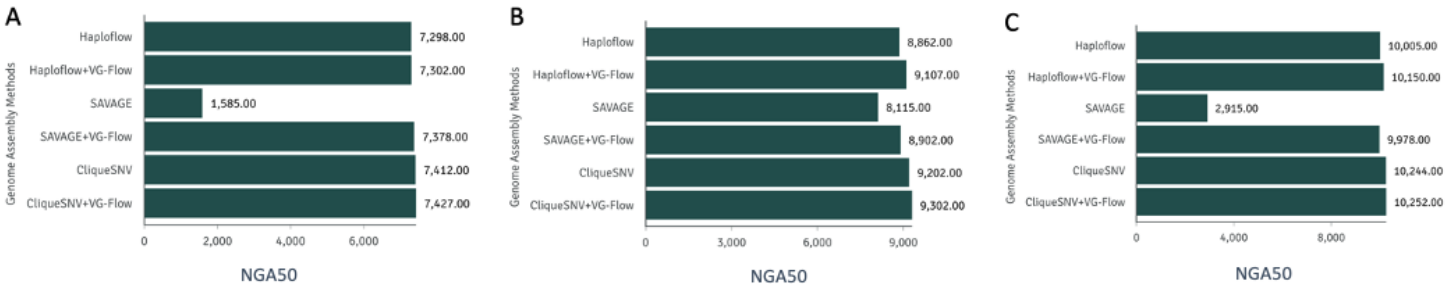

Figure 13 : Performance comparison of HaploFlow, SAVAGE and CliqueSNV alone and along with VG-Flow tools on all the three simulated datasets. ***A, B and C*** *shows the NGA50 value achieved on 6-strain Poliovirus, 10-strain HCV, 15-strain ZIKV respectively.*

Real Datasets

We also evaluated the performance of VG-Flow on real datasets. We observe similar pattern as observed for simulated datasets in the results obtained for real datasets. The contiguity of the assembly improves significantly while decreasing the number of contigs which improves the overall quality of the assembly. This is illustrated in Figure 14. For 5-strain mixture (10,000x coverage) the error rate for CliqueSNV lowers from 1.38% to 0%. The assembly also recovers more than half of the true genome assembly (56.14%). With improved continuity of the assembly for Haploflow on 5-strain mixture (1000x coverage), the target genome covered has slightly reduced by ~3 which is not a significant difference. Overall, the error rate for SAVAGE is really low, but with VG-Flow the contiguity and the length of the contigs drastically improve (10074 -14112bp) along with slight increase in the genome target covered.

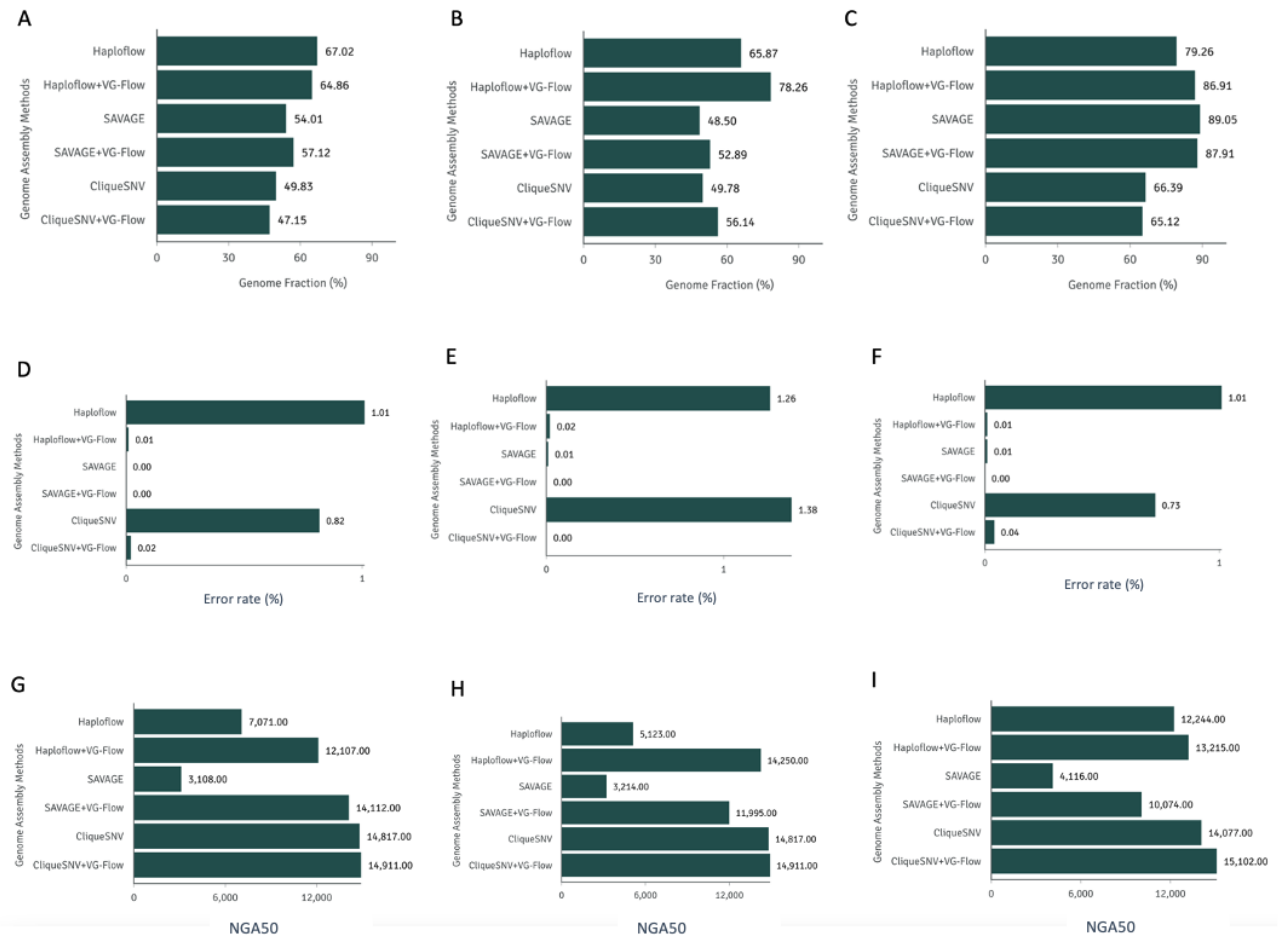

Figure 14 : Bar charts shows the improvement in the performance of Haploflow, SAVAGE and CliqueSNV alone and along with VG-Flow on real datasets. ***A, B and C*** *show the genome fraction reconstructed by the tools;* ***D, E and F*** *show the error rate produced by the assemblers;* ***G, H and I*** *show the NGA50 achieved by the assembly tools on 5-strain mixture (1000x coverage), 5strain mixture(10000x coverage) and 4-strain mixture(1000x coverage) respectively.*

Abundance Estimations

The absolute frequency errors per assembler on 5-strain real dataset mixture is shown in Figure 15. This figure highlights that VG-Flow performs the best which has smaller error values than CliqueSNV, HaploClique and QuasiRecomb. QuasiRecomb highly overestimated the higher abundant strains. VG-Flow + Haploflow performs the best with ~0.01 error across all the datasets. All the tools except for HaploClique highly underestimated the lower abundant strain. Similar conclusions were drawn from the 2-strain mixture analysis.

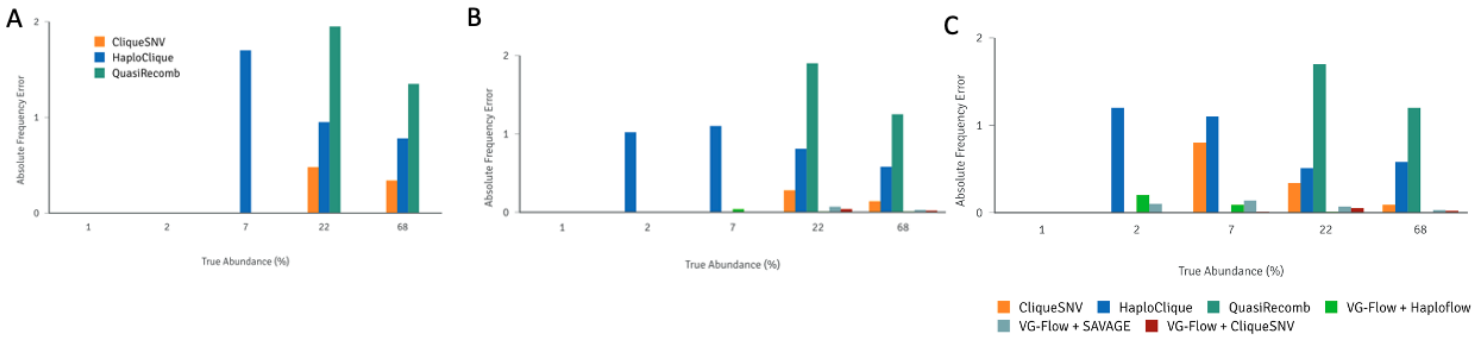

Figure15 : Abundance estimation errors on 5-strain real mixture dataset. Absolute Frequency errors compared to the true abundances. A 100x : VG-Flow was not implemented on this dataset so the comparison is done for only three tools; B 1000x; C 10000x coverage

5 Conclusions and Discussions

5.1 Why is reconstructing of Viral quasispecies genomes important?

Recent viral diseases outbreak like SARS-COV2 and earlier outbreaks like Hepatitis C virus (HCV) has indicated the urgent need for techniques or ways to examine genetic diversity of viral diseases. As mentioned before, viral quasispecies can rapidly evolve which results in multiple infectious strains either evolution within the same host or intra-host. Strains can have different phenotypes, like resistance, or the level of immunity resistance against the host immunity or virulence all of which is crucial for treatment related concerns[24]. The goal of haplotype-aware genome assembly is to use sequencing reads to reassemble an organism's copies of original genomic sequences called haplotypes.

5.2 Importance of creating broad and diverse gold-standard datasets

Until now, all the experiments are mostly performed using the virus mix in [18]or simulated datasets with similar characteristics presented in [15], [27], [36]. The major drawback of the 5-virus mix presented in [18] is that the least abundant strain is about 20% which is unlikely to see in real. Even though datasets presented in [15], [27] contain haplotypes ranging from 6-20%, they are still simulated. Usually, the pairwise divergence between the strains is about 1-10% but the 5-virus mix consists of strains at pairwise divergence varying from 2.6% to 8.4%. They do not or poorly reflect viral intra-host evolution as seen in previous studies. Here we created multiple benchmarking real datasets with varying number of strains (five, four and two) and coverages (100x, 1000x and 10000x) to overcome these challenges. The pairwise divergence varies from 1-5% for most of the datasets and the least abundant strain is also about 1% which reflects more on how the haplotypes frequencies exist in nature.

5.3 None of the six assembly tools outperforms other tools on all the performance metrics

When compared to simulated datasets, real dataset was more challenging to handle for the assembly tools. The results indicate that *none of the six assembly tools perform the best on all the performance evaluation metrics.* The trends observed in the performance of all the six assembly tools was similar for both real and simulated datasets. As we saw earlier, as the complexity of the datasets increase, the quality of the produced haplotypes also reduces. For

real datasets, the assembly tools perform better for 4-strain mixture than 5-strain mixture in terms of contiguity and produces a shorter length assembly. This can be explained by 4-strain mixture is less complex to handle as it contains lesser strains and the pairwise divergence is approximately 1% more than 5-strain mixture. Similarly for simulated datasets the assembly tools performed the best on 10-strain HCV mixture as it is the least complex dataset to handle compared to 6-strain Poliovirus and 15-strain ZIKV mixtures. The genome target achieved for 15-strain mixture was the least for almost all the tools compared to other simulated datasets because of high number of strains and the sample has a greater number of less abundant strains (~1%). If we look into the individual performance of the tools, Haploflow could handle datasets with substantial variation in terms of all the metrics for real datasets but does not report anything about the abundances. SAVAGE performs significantly better as coverage increases for real datasets but performs the best in genome coverage for all the simulated datasets but produces relatively shorter contigs. SAVAGE outperformed all the tools in terms of error rate (mismatches, indels and N-Rate). CliqueSNV and QuasiRecomb perform the best in terms of continuity, but QuasiRecomb performs poorly on all the other metrics. It also crashed a lot of times for complex datasets. Even though it achieved a high genome coverage CliqueSNV failed to assemble the strains with lower abundances. SPAdes being a generic assembler performs decently well on all the datasets but fails to assemble the low abundant strains. HaploClique produced a substantially fragmented assembly with huge number of contigs for all the real and benchmarking datasets but performs really well in terms of genome coverage. Overall, in the beginning of the research we expected that reference-based assemblers to produce more accurate haplotypes compared to de-novo assemblers but that was not the case as the assembly produced by reference-based assemblers.

5.4 Improved strain-specific genome assemblies with VG-Flow using pre-assembled contigs

Now the question is if these assembled contigs further be used to produce a high quality and accurate assembly and if so, how? To test this, we used VG-Flow assembly method with some tweaks. VG-Flow improved the quality and accuracy of the haplotypes. One feature of VG-Flow is deriving frequencies of output contigs which is crucial in viral quasispecies assemblies. As many assembly tools for example SAVAGE and Haploflow in our case do not report any frequencies for the assembled contigs, therefore, implementing VG-Flow along with these

assemblers is highly important. Even if reference-based tools like CliqueSNV, HaploClique and QuasiRecomb report the frequencies of the assembled contigs, VG-Flow outperforms all the tools in terms of abundance estimates. The type of pre-assembled contigs given to VG-Flow also plays a major role. So, the final conclusion is, to achieve a better continuous assembly we recommend using VG-Flow along with the pre-assembled contigs because it does improve the performance on at least one of the metrics.

5.5 Limitations and Future Work

The main goal of this research if put in layman terms was to produce a high-quality real benchmarking datasets which can be a primary standard or inspire researchers to focus on broader real datasets. Along with that, examining the performance of viral quasispecies assemblers at strain-level and to find alternative ways to improve the quality of the produced haplotypes. We focused our research on developing a high-quality challenging benchmarking real datasets and how the six assembly tools perform on these real and existing simulated datasets. As mentioned earlier, we have used bacteriophages (phages) for our benchmarking datasets. Well, phages are viruses and exhibit same properties as any other viruses, but these are not human or animal viruses. Though the structural properties are like that of higher order viruses like human viruses.

Our results also show that there is a need to create assembly tools which can handle more complex and diverse datasets with different strain variants and varying coverages. It might be interesting to investigate long reads rather than short reads and see how the tools perform. With Long-read sequencing technologies like PacBio and Nanopore advancements and price reductions present a different set of problems for haplotype reconstruction, along with the creation of novel sequencing techniques and tools. This emerging technology is capable of sequencing amplicons or even complete viral genomes sequences in a single run, eliminating the requirement for sequencing read assembly. Upcoming simulation research should also focus on handling datasets which are prone to errors with haplotype reconstruction tools that can incorporate handling these errors as well as the impact of average coverage and recombination on haplotype reconstruction. In the second part of our research where we use VG-Flow method to improve the quality of the haplotypes.

However, VG-Flow has not been tested for the threshold of maximum number of pre-assembled contigs that is given to it as input.

APPENDIX 1

1. 5-strain mixture

The results of all the six assembly tools for 5-strain mixture(100x, 1000x and 10000x) are summarized in the tables below.

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	7	69.265	5849	5808	0.186	0
SAVAGE	5	32.036	4555	-	0	0
CliqueSNV	6	49.824	14882	14817	0.977	0
SPAdes	1	49.7	14719	-	0	0
HaploClique	108	57.224	553	307	0.910	0
QuasiRecomb	278	49.83	14836	12711	2.38	0

Table 5: 100x coverage

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	11	67.02	7100	7071	1.01	0
SAVAGE	17	54.012	4471	3108	0	0
CliqueSNV	9	49.83	14927	14817	0.819	0
SPAdes	19	46.114	994	-	0.12	0
HaploClique	1500	92.97	562	307	1.529	1
QuasiRecomb	334	48.67	14984	12719	2.17	12

Table 6 : 1000x coverage

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	19	79.258	5164	5123	1.262	0
SAVAGE	49	89.045	4555	3214	0.01	0
CliqueSNV	3	66.388	14926	14817	1.383	0
SPAdes	43	39.96	1092	-	0.683	0
HaploClique	1757	92.979	562	307	2.968	2
QuasiRecomb	912	48.67	14984	12719	2.17	27

Table 7 : 10000x coverage

2. 4-strain mixture

The results of all the six assembly tools for 4-strain mixture(100x, 1000x and 10000x) are summarized in the tables below.

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	10	62.221	13733	12588	0.033	0
SAVAGE	8	34.62	3341	-	0.01	0
CliqueSNV	6	44.827	15216	14077	0.383	0
SPAdes	8	29.569	3061	-	0.716	0
HaploClique	278	60.21	9685	498	1.78	1
QuasiRecomb	927	41.89	14105	13002	2.178	19

Table 8 : 100x - 4strain mixture

Assembly Tools	#contigs	Target (%)	N50	NGA50	ER (%)	#Misassemblies
Haploflow	11	65.87	13545	12244	1.01	0
SAVAGE	12	48.504	4641	4116	0.01	0
CliqueSNV	8	49.778	15226	14077	0.727	0
SPAdes	18	30.621	1505	-	0.573	0
HaploClique	924	90.118	9685	498	1.480	1
QuasiRecomb	1025	43.592	14108	14105	2.519	15

Table 9 : 1000x - 4strain mixture

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	14	72.092	14841	12981	1.41	0
SAVAGE	57	70.058	4685	3156	0.01	0
CliqueSNV	4	64.02	15226	14077	1.42	0
SPAdes	48	45.594	1947	1108	0.705	0
HaploClique	1214	91.782	9685	501	2.968	4
QuasiRecomb	2289	48.901	14105	14105	3.105	27

Table 10 : 10000x coverage

3. 2 strain mixtures

The results of five assembly tools for 2 -strain mixture(1000x) coverage for different abundance distribution are summarized in the tables below

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	6	50.347	4413	2511	0.1	0
SAVAGE	13	54.233	3441	975	0.01	0
CliqueSNV	1	50.081	14882	14837	0.06	0
HaploClique	1078	92.634	588	305	1.349	1
SPAdes	1	50.347	14928	14928	0	0

Table 11 : 1,99

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	8	56.786	4352	2281	0	0
SAVAGE	12	54.14	3441	972	0	0
CliqueSNV	1	50.081	14881	14837	1.66	0
HaploClique	1022	90.245	583	311	2.92	3
SPAdes	12	58.118	1563	588	0.1	0

Table 12 : 2,98

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	7	98.813	7453	7453	1.959	0
SAVAGE	13	55.251	3441	975	0.2	0
CliqueSNV	2	50.081	14840	14840	2.11	0
HaploClique	1171	93.319	583	310	2.349	1
SPAdes	11	52.482	4084	588	0.108	0

Table 13 : 3,97

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	8	99.477	7452	7453	0.2	0
SAVAGE	13	55.255	3441	975	0	0
CliqueSNV	1	50.081	14884	14836	0.282	0
HaploClique	1121	92.15	584	310	2.349	2
SPAdes	11	52.482	1984	588	0	0

Table 14 : 4,96

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	6	90.219	7452	7451	0.2	0
SAVAGE	15	64.556	3792	1094	0	0
CliqueSNV	2	50.088	14884	14842	0.282	0
HaploClique	1195	94.779	584	310	2.349	2
SPAdes	11	53.062	1984	588	0	0

Table 15 : 5,95

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	4	99.642	9725	9626	1.959	0
SAVAGE	14	64.244	3792	1094	0.2	0
CliqueSNV	7	50.088	14884	14842	2.11	0
HaploClique	1125	94.319	583	310	2.349	1
SPAdes	11	53.062	1984	588	0.108	0

Table 16 : 6.94

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	4	99.7	12165	12124	0.1	0

SAVAGE	16	69.14	3441	707	0.2	0
CliqueSNV	6	50.088	14884	14842	0.06	0
HaploClique	1198	94.454	584	310	1.349	1
SPAdes	10	49.295	1918	562	0	0

Table 17 : 7,93

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	4	99.642	9725	9626	0	0
SAVAGE	14	67.639	3441	707	0	0
CliqueSNV	6	50.088	14884	14842	1.66	0
HaploClique	1199	94.645	583	310	2.92	3
SPAdes	10	49.295	1918	562	0.1	0

Table 18 : 8,92

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	9	99.703	3291	3290	1.959	0
SAVAGE	14	67.68	3441	707	0.2	0
CliqueSNV	6	50.088	14884	14842	2.11	0
HaploClique	1127	93.774	583	310	2.349	1
SPAdes	10	49.295	1918	562	0.108	0

Table 19 : 9,91

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	7	84.927	7273	7232	0.1	0
SAVAGE	14	81.094	3441	707	0.01	0
CliqueSNV	6	50.091	14885	14843	0.06	0
HaploClique	1031	94.634	583	310	1.349	1
SPAdes	11	48.863	1563	562	0	0

Table 20 : 10,90

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	4	99.98	11558	11453	0	0
SAVAGE	17	87.291	3441	707	0	0
CliqueSNV	4	99.723	14885	14840	1.66	0
HaploClique	1054	95.761	583	310	2.92	1
SPAdes	10	49.035	1918	562	0.1	0

Table 21 : 15,85

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	9	96.162	3494	3348	0.1	0
SAVAGE	15	94.145	3441	707	0.15	0
CliqueSNV	6	99.713	14882	14840	0.06	0
HaploClique	1180	95.12	583	310	1.349	1
SPAdes	12	52.024	1563	588	0	0

Table 22 : 20,80

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	7	97.757	4730	4743	0.486	0
SAVAGE	11	94.145	3041	2245	0.15	0
CliqueSNV	3	99.713	14885	14840	1.305	0
HaploClique	1157	94.794	582	309	1.726	1
SPAdes	13	58.071	1501	702	0.121	0

Table 23 : 30,70

4. Simulated datasets

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	5	76.001	7356	7298	0.712	0
SAVAGE	44	77.9	2154	1585	0.02	0
CliqueSNV	2	55.3	7428	7412	0.291	0
SPAdes	8	72.476	4216	3705	0.014	0
HaploClique	552	89.901	557	307	1.924	10
QuasiRecomb	2411	41.89	7422	7105	2.064	14

Table 24 : 6-Strain Poliovirus Mixture

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	40	96.915	9311	8862	0.55	0
SAVAGE	24	95.58	8680	8115	0.01	0
CliqueSNV	14	91.04	9208	9201	1.624	0
SPAdes	10	90.582	8680	8070	0.013	0
HaploClique	1162	92.98	512	357	2.14	8
QuasiRecomb	5492	48.67	9245	9228	2.17	10

Table 25 : 10 Strain HCV Mixture

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	25	67.114	10081	10005	1.671	0
SAVAGE	100	94.792	3724	2915	0.01	0
CliqueSNV	22	88.491	10251	10244	0.989	0

SPAdes	50	70.941	3473	3471	0.242	0
HaploClique	1757	92.979	524	361	2.106	7
QuasiRecomb	7785	47.15	10278	10272	2.78	11

Table 26 : 15-strain ZIKV mixture

5. VG-Flow Results

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	11	67.02	7100	7071	1.01	0
Haploflow+VG-Flow	5	64.861	13343	12107	0.01	0
SAVAGE	17	54.012	4471	3108	0	0
SAVAGE+VG-Flow	3	57.115	12909	12112	0.0	0
CliqueSNV	9	49.83	14927	14817	0.819	0
CliqueSNV+VG-Flow	3	47.15	14928	14911	0.02	11

Table 27 : VG-Flow results obtained on 5-strain real dataset sample mixture (1000x coverage).

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	19	79.258	5164	5123	1.262	0
Haploflow+VG-Flow	7	86.912	14267	14250	0.02	0
SAVAGE	49	89.045	4555	3214	0.01	0
SAVAGE+VG-Flow	15	87.91	12115	11995	0.0	0
CliqueSNV	3	66.388	14926	14817	1.383	0
CliqueSNV+VG-Flow	3	65.117	14926	14911	0.0	0

Table 28 : VG-Flow results obtained on 5-strain real dataset sample mixture (10000x coverage).

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	11	65.87	13545	12244	1.01	0
Haploflow+VG-Flow	5	78.26	13789	13215	0.01	0
SAVAGE	12	48.504	4641	4116	0.01	0
SAVAGE+VG-Flow	5	52.891	11945	10074	0.0	0
CliqueSNV	8	49.778	15226	14077	0.727	0
CliqueSNV+VG-Flow	4	56.142	15226	15102	0.04	0

Table 29 : VG-Flow results obtained on 4-strain real dataset sample mixture (1000x coverage)

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	5	76.001	7356	7298	0.712	0
Haploflow+VG-Flow	4	74.028	7443	7302	0.0	0
SAVAGE	44	77.9	2154	1585	0.02	0
SAVAGE+VG-Flow	14	84.97	7415	7378	0.04	0
CliqueSNV	2	55.3	7428	7412	0.291	0
CliqueSNV+VG-Flow	2	56.91	7445	7427	0.02	0

Table 30 : VG-Flow results on 6-strain Poliovirus simulated dataset

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	40	96.915	9311	8862	0.55	0
Haploflow+VG-Flow	14	95.21	9311	9107	0.02	0
SAVAGE	24	95.58	8680	8115	0.01	0
SAVAGE+VG-Flow	11	94.00	9205	8902	0.02	0
CliqueSNV	14	91.04	9208	9202	1.291	0
CliqueSNV+VG-Flow	12	95.67	9311	9302	0.05	0

Table 31 : VG-Flow results on 10-strain HCV mixture

Assembly Tools	#contigs	Target(%)	N50	NGA50	ER(%)	#Misassemblies
Haploflow	25	67.114	10081	10005	1.671	0
Haploflow+VG-Flow	13	78.89	10267	10150	0.02	0
SAVAGE	100	94.792	3724	2915	0.01	0
SAVAGE+VG-Flow	19	91.05	10176	9978	0.02	0
CliqueSNV	22	88.491	10251	10244	0.989	0
CliqueSNV+VG-Flow	12	89.98	10269	10252	0.05	0

Table 32 : VG-Flow results on 15-strain ZIKV mixture dataset

6 References

[1] P. Poltronieri, B. Sun, and M. Mallardo, "RNA Viruses: RNA Roles in Pathogenesis, Coreplication and Viral Load," *Current Genomics*, vol. 16, no. 5, 2015, doi: 10.2174/1389202916666150707160613.

[2] E. Domingo, J. Sheldon, and C. Perales, "Viral Quasispecies Evolution," *Microbiology and Molecular Biology Reviews*, vol. 76, no. 2, 2012, doi: 10.1128/mmbr.05023-11.

[3] E. Ghedin *et al.*, "Mixed Infection and the Genesis of Influenza Virus Diversity," *Journal of Virology*, vol. 83, no. 17, 2009, doi: 10.1128/jvi.00773-09.

[4] M. Vignuzzi, J. K. Stone, J. J. Arnold, C. E. Cameron, and R. Andino, "Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population," *Nature*, vol. 439, no. 7074, 2006, doi: 10.1038/nature04388.

[5] F. Sun *et al.*, "SARS-CoV-2 Quasispecies Provides an Advantage Mutation Pool for the Epidemic Variants," *Microbiology Spectrum*, vol. 9, no. 1, 2021, doi: 10.1128/spectrum.00261-21.

[6] E. Domingo, "Quasispecies Theory in Virology," *Journal of Virology*, vol. 76, no. 1, 2002, doi: 10.1128/jvi.76.1.463-465.2002.

[7] A. S. Lauring and R. Andino, "Quasispecies theory and the behavior of RNA viruses," *PLoS Pathogens*, vol. 6, no. 7. 2010. doi: 10.1371/journal.ppat.1001005.

[8] M. C. F. Prosperi and M. Salemi, "QuRe: Software for viral quasispecies reconstruction from next-generation sequencing data," *Bioinformatics*, vol. 28, no. 1, 2012, doi: 10.1093/bioinformatics/btr627.

[9] I. N. Lu, C. P. Muller, and F. Q. He, "Applying next-generation sequencing to unravel the mutational landscape in viral quasispecies," *Virus Research*, vol. 283. 2020. doi: 10.1016/j.virusres.2020.197963.

[10] S. Prabhakaran, M. Rey, O. Zagordi, N. Beerenwinkel, and V. Roth, "HIV haplotype inference using a propagating dirichlet process mixture model," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 1, 2014, doi: 10.1109/TCBB.2013.145.

[11] H. E. L. Lischer and K. K. Shimizu, "Reference-guided de novo assembly approach improves genome reconstruction for related species," *BMC Bioinformatics*, vol. 18, no. 1, 2017, doi: 10.1186/s12859-017-1911-6.

[12] R. Pereira, J. Oliveira, and M. Sousa, "Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics," *Journal of Clinical Medicine*, vol. 9, no. 1. 2020. doi: 10.3390/jcm9010132.

[13] T. Xiao and W. Zhou, "The third generation sequencing: The advanced approach to genetic diseases," *Translational Pediatrics*, vol. 9, no. 2. 2020. doi: 10.21037/TP.2020.03.06.

[14] N. Beerenwinkel and O. Zagordi, “Ultra-deep sequencing for the analysis of viral populations,” *Current Opinion in Virology*, vol. 1, no. 5. 2011. doi: 10.1016/j.coviro.2011.07.008.

[15] J. A. Baaijens, A. Z. el Aabidine, E. Rivals, and A. Schönhuth, “De novo assembly of viral quasispecies using overlap graphs,” *Genome Research*, vol. 27, no. 5, 2017, doi: 10.1101/gr.215038.116.

[16] S. Mangul, N. C. Wu, N. Mancuso, A. Zelikovsky, R. Sun, and E. Eskin, “Accurate viral population assembly from ultra-deep sequencing data,” *Bioinformatics*, vol. 30, no. 12, 2014, doi: 10.1093/bioinformatics/btu295.

[17] Y. Lin, J. Li, H. Shen, L. Zhang, C. J. Papasian, and H. W. Deng, “Comparative studies of de novo assembly tools for next-generation sequencing technologies,” *Bioinformatics*, vol. 27, no. 15, 2011, doi: 10.1093/bioinformatics/btr319.

[18] F. di Giallonardo *et al.*, “Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations,” *Nucleic Acids Research*, vol. 42, no. 14, 2014, doi: 10.1093/nar/gku537.

[19] A. Eliseev *et al.*, “Evaluation of haplotype callers for next-generation sequencing of viruses,” *Infection, Genetics and Evolution*, vol. 82, p. 104277, Aug. 2020, doi: 10.1016/J.MEEGID.2020.104277.

[20] J. A. Baaijens, L. Stougie, and A. Schönhuth, “Strain-aware assembly of genomes from mixed samples using flow variation graphs,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020, vol. 12074 LNBI. doi: 10.1007/978-3-030-45257-5_14.

[21] G. Hickey *et al.*, “Genotyping structural variants in pangenome graphs using the vg toolkit,” *Genome Biology*, vol. 21, no. 1, 2020, doi: 10.1186/s13059-020-1941-7.

[22] E. Garrison and A. Guarracino, “Unbiased pangenome graphs,” *bioRxiv*, p. 2022.02.14.480413, 2022, [Online]. Available: https://www.biorxiv.org/content/10.1101/2022.02.14.480413v1

[23] A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler, “QUAST: Quality assessment tool for genome assemblies,” *Bioinformatics*, vol. 29, no. 8, 2013, doi: 10.1093/bioinformatics/btt086.

[24] A. Fritz *et al.*, “Haploflow: strain-resolved de novo assembly of viral genomes,” *Genome Biology*, vol. 22, no. 1, 2021, doi: 10.1186/s13059-021-02426-8.

[25] X. Luo, X. Kang, and A. Schönhuth, “Strainline: full-length de novo viral haplotype reconstruction from noisy long reads,” *Genome Biology*, vol. 23, no. 1, 2022, doi: 10.1186/s13059-021-02587-6.

[26] S. Knyazev *et al.*, “CliqueSNV: Scalable Reconstruction of Intra-Host Viral Populations from NGS Reads,” *bioRxiv*, vol. xx, 2018.

[27] A. Töpfer, T. Marschall, R. A. Bull, F. Luciani, A. Schönhuth, and N. Beerenwinkel, “Viral Quasispecies Assembly via Maximal Clique Enumeration,” *PLoS Computational Biology*, vol. 10, no. 3, 2014, doi: 10.1371/journal.pcbi.1003515.

[28] S. Benidt and D. Nettleton, “SimSeq: A nonparametric approach to simulation of RNA-sequence datasets,” *Bioinformatics*, vol. 31, no. 13, 2015, doi: 10.1093/bioinformatics/btv124.

[29] Illumina, “NovaSeq 6000 Sequencing System,” *770-2016-025-H*, vol. 4, no. February. 2016.

[30] A. Kechin, U. Boyarskikh, A. Kel, and M. Filipenko, “CutPrimers: A New Tool for Accurate Cutting of Primers from Reads of Targeted Next Generation Sequencing,” *Journal of Computational Biology*, vol. 24, no. 11, 2017, doi: 10.1089/cmb.2017.0096.

[31] H. Jiang, R. Lei, S. W. Ding, and S. Zhu, “Skewer: A fast and accurate adapter trimmer for next-generation sequencing paired-end reads,” *BMC Bioinformatics*, vol. 15, no. 1, 2014, doi: 10.1186/1471-2105-15-182.

[32] R. Malhotra, M. Jha, M. Poss, and R. Acharya, “A random forest classifier for detecting rare variants in NGS data from viral populations,” *Computational and Structural Biotechnology Journal*, vol. 15, 2017, doi: 10.1016/j.csbj.2017.07.001.

[33] J. A. Baaijens, B. van der Roest, J. Köster, L. Stougie, and A. Schönhuth, “Full-length de novo viral quasispecies assembly through variation graph construction,” *Bioinformatics*, vol. 35, no. 24, 2019, doi: 10.1093/bioinformatics/btz443.

[34] A. Mikheenko, V. Saveliev, and A. Gurevich, “MetaQUAST: Evaluation of metagenome assemblies,” *Bioinformatics*, vol. 32, no. 7, 2016, doi: 10.1093/bioinformatics/btv697.

[35] A. Bankevich *et al.*, “SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing,” *Journal of Computational Biology*, vol. 19, no. 5, 2012, doi: 10.1089/cmb.2012.0021.

[36] D. Jayasundara, I. Saeed, S. Maheswararajah, B. C. Chang, S. L. Tang, and S. K. Halgamuge, “ViQuaS: An improved reconstruction pipeline for viral quasispecies spectra generated by next-generation sequencing,” *Bioinformatics*, vol. 31, no. 6, 2015, doi: 10.1093/bioinformatics/btu754.