5-8-2020

Data Preprocessing for Haplotype Calling from Viral NGS Data

Sai Sudheep Reddy Kaidapuram

DATA PREPROCESSING FOR HAPLOTYPE CALLING FROM VIRAL NGS DATA

by

SAI SUDHEEP REDDY KAIDAPURAM

Under the Direction of Alex Zelikovsky, PhD

ABSTRACT

For viral outbreaks like the recent COVID-19 outbreak, medical professionals in many areas require to know who infected whom, which variants are drug resistant and what therapy should be selected. To answer these questions, it is necessary to identify viral variants (haplotypes and SNP's) in patients. A haplotype refers to a combination of alleles or a set of single nucleotide polymorphisms (SNPs) found on the same chromosome. This thesis describes the development and assessment of several pipelines and tools for viral NGS and read data analysis and the effect on the accuracy of the haplotype identification.

INDEX WORDS: Haplotypes, Raw reads, Aligning, Single nucleotide polymorphisms

DATA PREPROCESSING FOR HAPLOTYPE CALLING FROM VIRAL NGS DATA

by

SAI SUDHEEP REDDY KAIDAPURAM

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2020

DATA PREPROCESSING FOR HAPLOTYPE CALLING FROM VIRAL NGS DATA

by

SAI SUDHEEP REDDY KAIDAPURAM

Committee Chair: Alex Zelikovsky

Committee: Pavel Skums

Robert Harrison

Electronic Version Approved:

Office of Graduate Services

College of Arts and Sciences

Georgia State University

May 2020

DEDICATION

I dedicate this thesis to my family and friends who supported me, for always being there for me.

I also dedicate this to Kajal Agarwal for motivating me.

ACKNOWLEDGEMENTS

I wish to record my sincere thanks to my advisor, Dr. Alex Zelikovsky, for his patient guidance, insightful suggestions, and support which made my graduate study a rewarding experience. It is a great honor for me to be able to participate in this challenging research. Appreciation is also extended to my co-advisor, Dr. Pavel Skums, and to Dr. Robert Harrison for serving on my advisory committee and taking precious time to review this thesis.

Sincere thanks are extended to Sergey Knyazev for his kind help to my research work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SNV: Single Nucleotide Variant

SNP: Single Nucleotide Polymorphism

NGS: Next Generation Sequencing

BWA: Burrows Wheeler Aligner

IVA: Iterative Virus Assembler

Shiver: Sequences from HIV Easily Reconstructed

WGS: Whole Genome Sequencing

DNA: Deoxyribonucleic acid

RNA: Ribonucleic acid

1 INTRODUCTION

Medical professionals need information about infection direction, drug resistance, and therapy selection for viral outbreaks. All of these require knowledge of viral variants in patients. Viral variants are usually described by haplotypes. A haplotype refers to a combination of alleles or to a set of single nucleotide polymorphisms (SNPs) found on the same chromosome. This thesis describes the work on three pipelines to improve the accuracy of the haplotypes with different methods for cleaning the raw data, aligning the data, processing the data.

The below sections describe Next Generation Sequencing and some of its methods, RNA viruses and quasispecies.

1.1 Next Generation Sequencing

DNA sequencing is the process of determining the sequence of nucleotides in a section of DNA. Next-generation sequencing (NGS) refers to the deep, high-throughput, in-parallel DNA sequencing technologies developed a few decades after the Sanger DNA sequencing method. The NGS technologies are different from the Sanger method in that they provide massively parallel analysis, extremely high-throughput from multiple samples at much reduced cost. Millions to billions of DNA nucleotides can be sequenced in parallel, yielding substantially more throughput and minimizing the need for the fragment-cloning methods that were used with Sanger sequencing. The second-generation sequencing methods are characterized by the need to prepare amplified sequencing libraries before undertaking sequencing of the amplified DNA clones, whereas third-generation single molecular sequencing can be done without the need for creating the time-consuming and costly amplification libraries. The parallelization of a high number of sequencing reactions by NGS was achieved by the miniaturization of sequencing reactions. The time needed

to generate the gigabase sized sequences by NGS was reduced from many years to only a few days or hours, with an accompanying massive price reduction.

The massively parallel sequencing technology known as next-generation sequencing (NGS) has revolutionized the biological sciences. With its ultra-high throughput, scalability, and speed, NGS enables researchers to perform a wide variety of applications and study biological systems at a level never before possible.

Today's complex genomic research questions demand a depth of information beyond the capacity of traditional DNA sequencing technologies. Next-generation sequencing has filled that gap and become an everyday research tool to address these questions.

Using capillary electrophoresis-based Sanger sequencing, the Human Genome Project took over 10 years and cost nearly $3 billion. Next-generation sequencing, in contrast, makes large-scale whole-genome sequencing (WGS) accessible and practical for the average researcher. It enables scientists to analyze the entire human genome in a single sequencing experiment, or sequence thousands to tens of thousands of genomes in one year.

NGS-based RNA-Seq is a powerful method that enables researchers to break through the inefficiency and expense of legacy technologies such as microarrays. Microarray gene expression measurement is limited by noise at the low end and signal saturation at the high end. In contrast, next-gen sequencing quantifies discrete, digital sequencing read counts, offering a broader dynamic range. Below paragraphs describes some of the types of next generation sequencing.

Illumina sequencing works by simultaneously identifying DNA bases, as each base emits a unique fluorescent signal, and adding them to a nucleic acid chain. This sequencing method is based on reversible dye-terminators that enable the identification of single bases as they are introduced into DNA strands.

PacBio sequencing captures sequence information during the replication process of the target DNA molecule. The template, called a SMRTbell, is a closed, single-stranded circular DNA that is created by ligating hairpin adaptors to both ends of a target double-stranded DNA molecule.

Roche 454 sequencing method is based on pyrosequencing, a technique which detects pyrophosphate release, again using fluorescence, after nucleotides are incorporated by polymerase to a new strand of DNA. Roche 454 sequencing can sequence much longer reads than Illumina. Like Illumina, it does this by sequencing multiple reads at once by reading optical signals as bases are added.

Ion Torrent sequencing measures the direct release of H+ (protons) from the incorporation of individual bases by DNA polymerase and therefore differs from the previous two methods as it does not measure light.

Nanopore sequencing is a unique, scalable technology that enables direct, real-time analysis of long DNA or RNA fragments. It works by monitoring changes to an electrical current as nucleic acids are passed through a protein nanopore. The resulting signal is decoded to provide the specific DNA or RNA sequence.

1.2 RNA Viruses and Quasispecies

1.2.1 RNA Viruses

An RNA virus[15] is a virus that has RNA as its genetic material. This nucleic acid is usually single-stranded RNA but may also be double-stranded RNA. Human diseases causing RNA viruses include Orthomyxoviruses, Hepatitis C Virus (HCV), Ebola disease, SARS, influenza, polio measles and human immunodeficiency virus (HIV) etc. Viruses may exploit the presence of RNA-dependent RNA polymerases for replication of their genomes or, in retroviruses,

with two copies of single strand RNA genomes, reverse transcriptase produces viral DNA which can be integrated into the host DNA under its integrase function. Among human retroviruses, HIV-1 is a lentivirus with an RNA genome formed by two copies of a single-stranded, positive-sense RNA. Upon entry into the target cell, the viral RNA genome is reverse transcribed into double-stranded DNA by a virally encoded reverse transcriptase that is transported along with the viral genome into the virus particle. The viral DNA is imported into the cell nucleus and integrated into the cellular DNA by a virally encoded integrase and host co-factors. Once integrated, the virus may become latent, or may be transcribed, producing new RNA genomes and viral proteins that are packaged and released from the infected cell as new virus particles that will infect other cells to begin the new replication cycle. Many aspects of the life cycle of retroviruses are intimately linked to the functions of cellular proteins and RNAs.

1.2.2 Quasispecies

A viral quasispecies is a population structure of viruses with many variant genomes. Quasispecies result from high mutation rates as mutants arise continually and change in relative frequency as viral replication and selection proceeds. Viral quasispecies are the mutant distributions (also termed mutant swarms or clouds) that are generated upon replication of RNA viruses, and some DNA viruses in infected cells and organisms. The quasispecies concept originated in a theoretical formulation of molecular evolution that emphasized error-prone replication of simple RNA or RNA-like replicons as an essential feature of self-organization and adaptability of primitive life forms. An important aspect of the quasispecies concept is that the large size of virus populations enables positive and negative interactions between individual viruses to establish a quasi-equilibrium of the variant proportions. Genetic variation is generated

by the accumulation of mutations during replication and their re-arrangement by genetic recombination, and genome segment reassortment in the case of segmented genomes.

1.2.3 Haplotypes

A haplotype is a group of genes within an organism that was inherited together from a single parent. The word "haplotype" is derived from the word "haploid," which describes cells with only one set of chromosomes, and from the word "genotype," which refers to the genetic makeup of an organism.

Biologists, immunologists, epidemiologists, pharmacologists, bio-medical specialists need information like who infected whom and how are they related, drug resistant variants, therapy selection. All of them require knowledge of haplotypes. Information about haplotypes is used to investigate the influence of genes on disease. Knowledge of the haplotype structure would therefore make it possible to type the minimum number of SNPs that would be needed to uniquely tag all the haplotypes to search for disease mutations.

Haplotypes are an allelic configuration of multiple markers that are present on a single chromosome of a given individual. Recombination will break up haplotypes when they are passed on to the subsequent generation. The size of ancestral haplotypes will therefore have been reduced considerably after many generations. However, if recombinations are more likely at specific locations or recombination hotspots, a block structure will arise. If within a block there have been no or few recombinations, variations within blocks will mainly be caused by mutation. As a result, it will be possible to characterize a large percentage of the subjects by a few common haplotypes which are parts of ancestral haplotypes that are conserved in the general population.

1.3 Contribution

Our contributions include:

1) Developed the three pipelines for haplotype construction
    a. Pipeline 1: includes BWA aligner and CliqueSNV
    b. Pipeline 2: includes BWA aligner, BBduk to trim the reads below a certain quality value and CliqueSNV
    c. Pipeline 3: includes BWA aligner, BBduk, Shiver to map reads to a constructed reference and CliqueSNV

    More about the pipelines, tools used in pipelines, input data, intermediate data is described the sections later.

2) Run the tools with the NGS read samples from CDC on the 3 pipelines.
3) Assess the haplotypes using metrics like Earth Mover's Distance (EMD), the edit distance to the closest predicted variant (ECP), the frequency of the closest predicted variant (FCP) and the explanation error of T (EEV).

1.4 Problem Formulation

Given: NGS read samples collected from CDC.

Assess: This thesis describes and assess more accurate procedure for finding the haplotypes and the three pipelines for haplotype reconstruction from the given NGS read samples from CDC. The tools used in the three pipelines and the steps in these pipelines are described in the following sections.

1.5 RoadMap

This section describes the roadmap of this report. Section 2 describes software tools used in the 3 pipelines for NGS reads data analysis. Section 3, the experimental design describes the 3 pipelines and the steps in each pipeline, the input for each pipeline and the intermediate input for each step in the pipeline and the output obtained from the pipelines. The metrics section describes several metrics used to validate reconstructed haplotypes from the three pipelines.

2 REVIEW OF SOFTWARE TOOLS

2.1 CliqueSNV

CliqueSNV [14], is a novel reference-based method for reconstruction of viral variants from NGS data. It efficiently constructs an allele graph based on linkage between single nucleotide variations and identifies true viral variants by merging cliques of that graph using combinatorial optimization techniques. CliqueSNV outperforms existing methods in both accuracy and running time on experimental and simulated NGS data for titrated levels of known viral variants.

CliqueSNV eliminates the need for preliminary error correction and assembly and infers haplotypes from patterns in distributions of SNVs in sequencing reads. It is suitable for long single-molecule reads (PacBio) as well as short paired reads (Illumina). CliqueSNV uses linkage between single nucleotide variations (SNVs) to distinguish them from sequencing errors efficiently. It constructs an allele graph with edges connecting linked SNVs and identifies true viral variants by merging cliques of that graph using combinatorial optimization techniques.

Previous tools such as V-phaser [3], V-phaser2 [4] and CoVaMa [5] exploited linkage of nucleotide variants, but they did not take into account sequencing errors when deciding whether two variants are linked. Results of these tools show that they were unable to reliably detect variants of frequency even higher than the error rate of sequencing.

The 2SNV algorithm [6] accommodated errors in links and was the first such tool to be able to detect haplotypes with a frequency below the error rate correctly. CliqueSNV method keeps the basic idea of 2SNV linkage analysis but develops a novel approach for collecting multiple SNV's and inference of true haplotypes. Unlike 2SNV, which hierarchically clusters together reads containing pairs of linked SNVs, CliqueSNV identifies true viral variants in a single clustering using an efficient merging of cliques of the allele graph. 2SNV is designed only for single amplicon

data whereas CliqueSNV can handle short paired reads from shotgun experiments. Finally, CliqueSNV identifies linked SNVs and constructs allele graphs using highly efficient data structures. As a result, CliqueSNV is more accurate and significantly faster than 2SNV and capable of rapidly handling millions of reads in of minutes.

CliqueSNV was validated on simulated and experimental data and compared with Savage, PredictHaplo, aBayesQR and 2SNV tools. The tools are benchmarked using the results of a PacBio sequencing experiment on a sample containing a titrated level of known Influenza A (IAV) viral variants, on similar data sets for experimental HIV-1 single-read and paired-end Illumina data and simulated Illumina HIV-1 and IAV data. In addition to standard algorithm performance measures, the CliqueSNV method developers used a new measure based on earth mover's distance between real and reconstructed haplotype distributions. In this validation study, CliqueSNV significantly outperformed these other methods in both accuracy and running time.

CliqueSNV algorithm consists of the following six steps:

1. Finding linked SNV pairs
2. Constructing the allele graph
3. Finding maximal cliques in the allele graph
4. Merging cliques in the clique graph
5. Finding consensus viral variants for merged cliques
6. Estimating frequencies of the viral variants.

2.2 BWA

BWA is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome. BWA aligner is used to align forward reads and reverse reads of the data. The reason why we need aligner is, in conventional paired-end sequencing, we simply sequence using the adapter for one end, and then once you're done you start over sequencing using the adapter for the other end. This means the two reads are the reverse complement of each other. This is how the forward read and reverse read are represented.

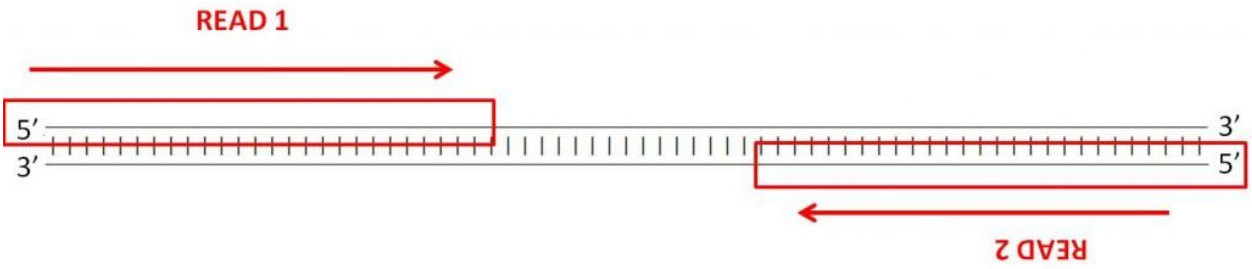

Figure 1 Forward and reverse reads representation

Therefore when you open your FASTQ files and look at a pair of reads, the sequences you see are, conceptually, pointing towards each other on opposite strands. When you align them to the genome, one read should align to the forward strand, and the other should align to the reverse strand, at a higher base pair position than the first one so that they are pointed towards one another. This is known as an "FR" read – forward/reverse read, in that order.

2.3 BBduk

BBDuk [13] is a fast and accurate tool for trimming and filtering sequencing data that is part of the BBTools package. BBduk stands for decontamination using kmers. BBDuk was developed to combine most common data-quality-related trimming, filtering, and masking operations into a single high-performance tool. It is capable of quality-trimming and filtering, adapter-trimming, contaminant-filtering via kmer matching, sequence masking, length filtering, entropy-filtering, format conversion, histogram generation, subsampling, quality-score recalibration, kmer cardinality estimation, and various other operations in a single pass.

Each sequence read in FASTQ file is associated with a quality score. Using BBduk we can trim the reads below a quality score we want to define.

2.4 Shiver

Shiver [7] is a tool for mapping paired-end short reads to a custom reference sequence constructed using do novo assembled contigs, in order to minimize the biased loss of information that occurs from mapping to a reference that differs from the sample.

Paired-end short reads and contigs assembled from those reads are required as input for each sample; also required is a set of existing reference genomes, chosen by the user. This thesis uses HXB2. HXB2 is a subtype B HIV-1 isolate used as the reference strain for aligning and numbering HIV-1 sequences.

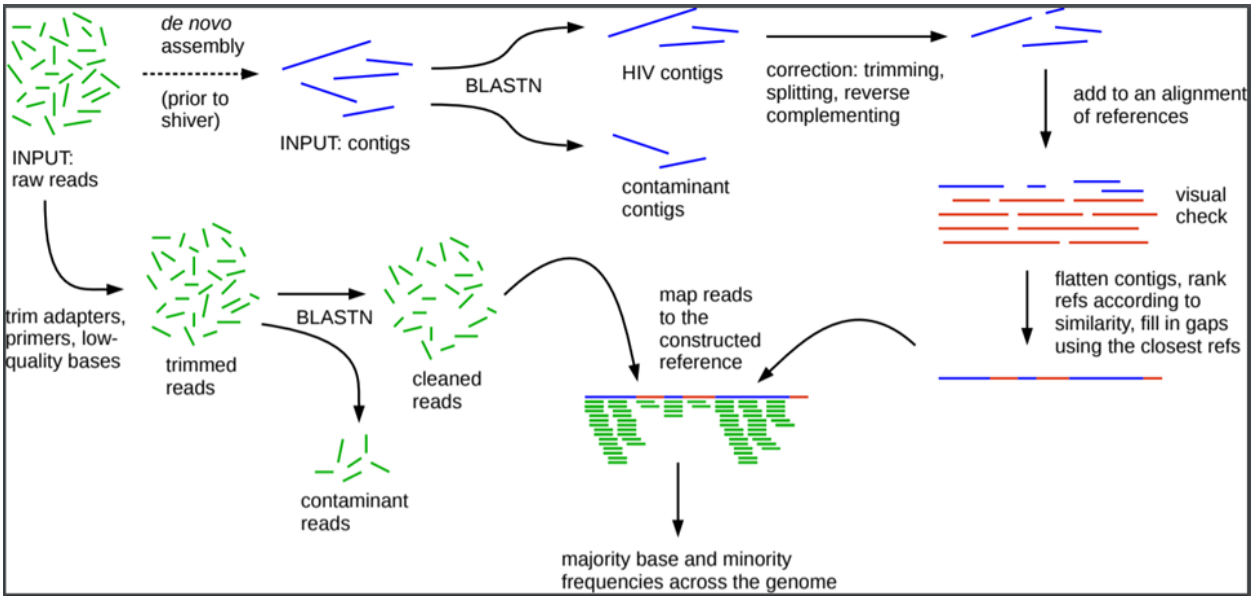

Figure 2: Figure explaining the flow of shiver process

2.4.1 Contig Flow

The first step is to find contigs using IVA. IVA is explained below.

2.4.1.1 IVA

IVA [8] is a de novo assembler designed to assemble virus genomes that have no repeat sequences, using Illumina read pairs sequenced from mixed populations at extremely high and

variable depth. IVA approach is similar to that of PRICE [9] one of the few assemblers designed for viral data in which the aligning process begins with seed sequences, which are iteratively extended by generating new sequence from local assemblies of reads at contig ends. But in IVA they extend contigs more conservatively using consensus kmers from the reads instead of using local assemblies. Also, IVA (Iterative Virus Assembler), is a completely *de novo* assembler, whereas PRICE must be provided with seed sequences to be extended into contigs.

Before assembling, adapter sequences are removed from the reads using Trimmomatic [10] followed by the trimming of polymerase chain reaction primer sequences. After trimming the reads, the most abundant kmer among the reads is found using kmc. This short seed kmer is iteratively extended into a contig using reads that have a perfect match to that kmer, treating the reads as unpaired. A list of all possible extension sequences is made (one sequence per overhanging read). IVA identifies the kmer of length *k* among prefixes of the possible extension sequences, for largest possible *k*, such that the kmer appears at least 10 times and is at least four times as abundant as the next most common kmer of length *k*. In this way, the seed is iteratively extended until its length reaches the insert size of the read pairs.

Contigs are extended in a similar manner to that of seed kmers. Instead of using perfect string matches, reads are mapped to the contigs with SMALT. During mapping, IVA also uses SAMtools [11]. Reads mapped as part of a perfect pair (in the correct orientation and separated by the correct distance) and hang off a contig end are used to extend the contig. The sequence added to a contig end is constructed using the method described above for kmer extensions.

When no more contigs can be extended, they are cleaned as follows before generating a new seed. Contig ends are trimmed for quality and overlapping contigs are merged based on sequence

similarity found at their ends using nucmer [12]. Assembly stops either when a pre-defined maximum contig number is reached or no new seeds can be made.

2.4.2 Read Flow

Before mapping, reads are trimmed for low-quality bases, adapter and primer sequences using Trimmomatic and fastaq. Contaminant read pairs are diagnosed as those matching contaminant contigs more closely than the tailored reference and are removed. The remaining reads are mapped to the tailored reference. By default shiver does mapping using smalt with a minimum read identity (the fractional agreement between a read and the reference to be considered mapped) of 70%, independent mapping of mates in a pair, a maximum insert size of 2,000 bp, and discarding improperly paired reads.

Following mapping, each position in the genome is considered in turn using SAMtools to find the frequencies of different bases. At positions where some reads have deletions relative to the mapping reference, it counts the frequency of the gap character together with actual bases. At positions where some reads have insertions relative to the mapping reference, for the consensus it uses the most common insertion size (which may be 0, i.e. no insertion). By default, the most common base is called to give the consensus; optionally ambiguity codes can be used more readily, when the frequency of the most common base(s) is below a threshold. A consensus base is only called if the coverage equals or exceeds a minimum threshold specified by the user, to protect against the effect of residual low-coverage contaminant reads in genomic regions lacking genuine HIV reads. By default this is 15, but this is likely to need adjusting for different datasets. A tool contained in shiver helps the user to explore appropriate values.

By default, once the consensus is called, the cleaned reads are re-mapped to it (with any missing coverage in the consensus filled in with the corresponding part of the original tailored reference) for a second iteration of calling the base frequencies and the consensus. shiver also produces a 'global alignment' of all consensuses it generates by coordinate translation, without need for an alignment algorithm.

Shiver tool was developed to preprocess and map reads from each sample to a custom reference, constructed using *de novo* assembled contigs supplemented by existing reference genomes. Tailoring the reference to be as close as possible to the expected consensus before mapping maximises the accuracy of the mapping, and therefore of the resulting consensus. shiver's identification, ranking, and use of the closest existing references to fill in gaps between contigs boosts data recovery for samples with amplification failure or assembly failure. Such partial-genome samples, which are inevitable in large diverse data sets, are processed with exactly the same two commands; this simplifies scripted application of shiver to all samples in a data set. shiver also produces a global alignment containing all of the consensuses separately generated for each sample, which is usually required for comparative analysis of the sequences such as for phylogenetics.

Mapping to shiver's constructed reference instead of mapping the same reads to the closest identified real reference gives a median increase in consensus sequence length of 205 bp, with thirteen of the original bases called differently and more accurately. This shows the importance of tailoring the reference to the sample before mapping. shiver's consensus, obtained by mapping reads to a reference constructed from the contigs, has a median of 7 bases called differently from

the contigs even after correcting structural problems in the contigs and trimming suspicious sequence from their ends. This illustrates the need for mapping in addition to assembly.

3 EXPERIMENTAL DESIGN

3.1 Metrics

Validation of different haplotype reconstruction methods should simultaneously answer two general questions: (i) how close are the reconstructed and true variants and (ii) how narrow is the reconstructed and true variant frequency distribution. Previous studies report high variation in results addressing these questions likely due to the challenge of simultaneously addressing them. This thesis uses the Earth Mover's Distance (EMD) as a distance measure for populations, which generalizes edit distances between genomes of individual variants.

3.1.1 Earth Movers Distance

Let $\mathcal{T} = \{T_i, t_i\}_{i=1}^{|\mathcal{T}|}$ be the true viral population, where *Ti* is the *i*th true variant with frequency *ti*, and let $\mathcal{P} = \{P_j, p_j\}_{j=1}^{|\mathcal{P}|}$ be the predicted viral population, where *Pj* is the *j*th predicted variant with frequency *pj*. Let $dij = d(T_i, P_j)$ be the edit distance between variants *Ti* and *Pj*. The EMD measures the total error of explaining true variants with predicted variants. If we decide to explain *fij* copies of *Ti* with *fij* copies of *Pj* then we will make an error of $f_{ij}d_{ij}$. The total error of explaining $\mathcal{T}$ with $\mathcal{P}$ equals $\Sigma_{i,j} f_{ij}d_{ij}$. Of course, the total amount of *Pj* used cannot exceed available *pj*, $\Sigma i\ f_{ij} \leq pj$, and all the amount *ti* of *Ti* should be explained, i.e. $\Sigma j\ fij = ti$. EMD (i.e., the minimum explanation error) could be efficiently computed as an instance of the transportation problem using network flows. We can also compute the explanation error for any particular true variant *Ti* which is defined as $EEV(Ti) = (\Sigma i,j\ f_{ij}d_{ij})/ti$. Note that EMD equals to the sum of frequency-weighted explanation errors: $EMD(\mathcal{T}, \mathcal{P}) = \sum_i t_i EEV(T_i)$.

3.2 Pipelines and Results

3.2.1 Pipeline1

Raw reads are aligned using BWA aligner. Alignment gives sam file. Running CliqueSNV on this sam file with different t and tf values gives the output haplotypes.

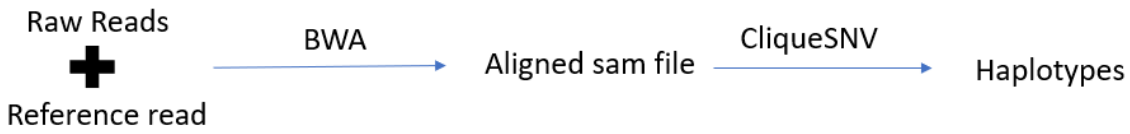

Figure 3: Figure explaining steps in pipeline 1

The sample reads are fastq files with two files the forward read and the reverse read. BWA aligner aligns the forward read and reverse read and gives output a sam file which contains the reads after alignment using the reference read. The reference read used is HXB2. The sam file obtained with aligned reads is the input for CliqueSNV. CliqueSNV is run on the aligned reads with different parameter values. It is run with 2 parameters, t and tf.

t is the minimum threshold for O22 value. Default is 100 (only for Illumina reads). tf is minimum threshold for 022 frequency relative to the reads coverage. Default value is 0.003%. For more sensitive algorithm work decrease this parameter (may significantly increase runtime for diverse samples).

3.2.1.1 Results

Comparison in done between the three pipelines i.e cliqueSNV, BBduk + cliqueSNV and BBduk + shiver + cliqueSNV. All the methods are run on Intel(R) Xeon(R) CPU E7-4850 v4 @ 2.10GHz with 16 cores and 2 threads per core. The maximum speed of CPU is 2800 MHz and minimum speed of CPU is 1200 MHz. The sample of data to be compared is a mixture data taken from CDC.

It contains 9 true variants. The true variants are named A, B, C, D, E, F, G, H, J and the frequencies of the true variants are: 50%, 25%, 12.5%, 6.3%, 3.2%, 1.6%, 0.8%, 0.4%, 0.2% respectively.

Table 1: ECP, FCP, EEV, EMD values for sample mixture data set with t=50 and tf=0.5% by running pipeline1

		CliqueSNV_MIX1_t_50_tf_0.5		
variant	frequency	ECP	FCP%	EEV
A	50	0	8.46	4
B	25	1	7.72	4.15
C	12.5	2	0.85	3.4
D	6.3	1	3.22	1
E	3.2	1	2.18	1.33
F	1.6	5	0.35	6.49
G	0.8	1	1.8	1
H	0.4	1	0.71	1
J	0.2	4	0.28	4
EMD				3.694

Table 2: ECP, FCP, EEV, EMD values for sample mixture data set with t=100 and tf=1% by running pipeline1

		CliqueSNV_MIX1_t_100_tf_1		
Variant	frequency	ECP	FCP%	EEV
A	50	1	5.94	3.8
B	25	1	10.46	4.12
C	12.5	0	11.42	0.32
D	6.3	1	15.68	1
E	3.2	1	1.91	2
F	1.6	7	11.42	11
G	0.8	1	2.68	1
H	0.4	2	1.56	2
J	0.2	8	3.01	8
EMD				3.307

Table 3: ECP, FCP, EEV, EMD values for sample mixture data set with t=200 and tf=2% by running pipeline1

		CliqueSNV_MIX1_t_200_tf_2		
variant	frequency	ECP	FCP%	EEV
A	50	0	27.13	2.54
B	25	1	13.9	4.1
C	12.5	0	25.7	0
D	6.3	1	15.8	1
E	3.2	2	5.3	2
F	1.6	7	25.7	7
G	0.8	2	2.8	2
H	0.4	4	27.1	5
J	0.2	8	5.3	9
EMD				2.591

Table 4: ECP, FCP, EEV, EMD values for sample mixture data set with t=500 and tf=5% by running pipeline1

		CliqueSNV_MIX1_t_500_tf_5		
variant	frequency	ECP	FCP%	EEV
A	50	0	31.3	1.9
B	25	1	14.4	3.9
C	12.5	0	30.1	0
D	6.3	1	15.7	1
E	3.2	3	31.3	5
F	1.6	7	30.1	7
G	0.8	6	30.1	6
H	0.4	4	31.3	14
J	0.2	9	8.3	9
EMD				2.4

3.2.2 Pipeline2

Raw reads are aligned using BWA aligner. BBduk is run on aligned reads to trim off the reads below a certain quality. This gives the trimmed reads. CliqueSNV is run on these trimmed reads gives the haplotypes.

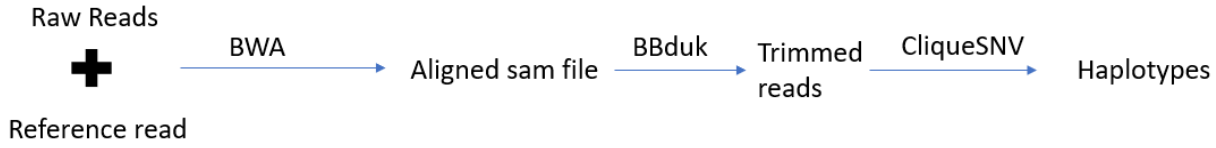

Figure 4: Figure explaining steps in pipeline 2

The sample reads are fastq files with two files the forward read, and the reverse read. BWA aligner aligns the forward read and reverse read and gives output a sam file which contains the reads after alignment using the reference read. The reference read used is HXB2. The sam file obtained with aligned reads is the input for BBduk. BBduk trims the reads below a certain quality value we define while running the tool from the command prompt. In this thesis the BBduk is run with three different quality values 20,25 and 30. CliqueSNV is run on the trimmed reads to get the haplotypes. CliqueSNV is run with different t and tf values (CliqueSNV parameters).

Table 5: Number of aligned reads after BBduk trimming with different quality trimming values of 20,25 and 30.

	Number of reads
Original file	5,604,910
Quality trim = 20	2,682,384
Quality trim = 25	1,048,476
Quality trim = 30	220,542

3.2.2.1 Results

Table 6: ECP, FCP, EEV, EMD values for sample mixture data set with t=10 and tf=0.033% by running pipeline 2 with quality trim 20

		bbduk_qtrim20_CliqueSNV_MIX1_t_10_tf_0.033		
variant	frequency	ECP	FCP%	EEV
A	50	1	0.14	3.9
B	25	1	5.3	3.5
C	12.5	0	9.6	0.36
D	6.3	0	9.1	0
E	3.2	1	1.2	1
F	1.6	4	0.1	6.86
G	0.8	1	1.6	1
H	0.4	1	0.7	1
J	0.2	1	0.6	1
EMD				3.09

Table 7: ECP, FCP, EEV, EMD values for sample mixture data set with t=50 and tf=0.5% by running pipeline 2 with quality trim 20

		bbduk_qtrim20_CliqueSNV_MIX1_t_50_tf_0.5		
variant	frequency	ECP	FCP%	EEV
A	50	1	0.5	3.2
B	25	1	12.5	4.7
C	12.5	0	20.3	0
D	6.3	1	0.5	1
E	3.2	1	1.4	1.5
F	1.6	6	0.9	6
G	0.8	3	4.3	3
H	0.4	0	0.8	0
J	0.2	0	0.7	0
EMD				3.03

Table 8: ECP, FCP, EEV, EMD values for sample mixture data set with t=100 and tf=1% by running pipeline 2 with quality trim 20

		bbduk_qtrim20_CliqueSNV_MIX1_t_100_tf_1		
variant	frequency	ECP	FCP%	EEV
A	50	2	3.3	4.3
B	25	1	14.3	3.9
C	12.5	0	26.7	0
D	6.3	1	23.4	1
E	3.2	1	3.3	1
F	1.6	7	26.7	7
G	0.8	1	4.2	1
H	0.4	5	12.5	14
J	0.2	8	12.5	9
EMD				3.44

Table 9: ECP, FCP, EEV, EMD values for sample mixture data set with t=200 and tf=2% by running pipeline 2 with quality trim 20

		bbduk_qtrim20_CliqueSNV_MIX1_t_200_tf_2		
variant	frequency	ECP	FCP%	EEV
A	50	0	26.08	2.6
B	25	1	13.7	4.14
C	12.5	0	26.7	0
D	6.3	1	15.3	1
E	3.2	2	4.9	2
F	1.6	7	26.7	7
G	0.8	1	3.3	1
H	0.4	4	26.08	5
J	0.2	8	4.9	9
EMD				2.666

Table 10: ECP, FCP, EEV, EMD values for sample mixture data set with t=500 and tf=5% by running pipeline 2 with quality trim 20

		bbduk_qtrim20_CliqueSNV_MIX1_t_500_tf_5		
variant	frequency	ECP	FCP%	EEV
A	50	0	30.9	2.1
B	25	1	14.1	1.04
C	12.5	0	31.5	0
D	6.3	1	15.2	1
E	3.2	3	30.9	5
F	1.6	7	31.5	7
G	0.8	6	31.5	6
H	0.4	4	30.9	14
J	0.2	9	8.1	9
EMD				2.55

Table 11: ECP, FCP, EEV, EMD values for sample mixture data set with t=10 and tf=0.033% by running pipeline 2 with quality trim 25

		bbduk_qtrim25_CliqueSNV_MIX1_t_10_tf_0.033		
variant	frequency	ECP	FCP%	EEV
A	50	0	1.1	3.3
B	25	1	6.4	4.5
C	12.5	3	8.2	3.06
D	6.3	0	13.4	0
E	3.2	0	1.3	2.06
F	1.6	3	0.5	8.05
G	0.8	3	1.5	3
H	0.4	1	0.8	1
J	0.2	3	0.6	3
EMD				3.43

Table 12: ECP, FCP, EEV, EMD values for sample mixture data set with t=50 and tf=0.5% by running pipeline 2 with quality trim 20

		bbduk_qtrim25_CliqueSNV_MIX1_t_50_tf_0.5		
variant	frequency	ECP	FCP%	EEV
A	50	2	4.07	4.23
B	25	1	9.7	3.56
C	12.5	0	11.09	3.52
D	6.3	0	12.2	4
E	3.2	0	7.3	0
F	1.6	5	0.6	7.3
G	0.8	4	2.03	4
H	0.4	1	2.6	1
J	0.2	3	0.4	3
EMD				3.464

Table 13: ECP, FCP, EEV, EMD values for sample mixture data set with t=100 and tf=1% by running pipeline 2 with quality trim 25

		bbduk_qtrim25_CliqueSNV_MIX1_t_100_tf_1		
variant	frequency	ECP	FCP%	EEV
A	50	1	3.06	2.61
B	25	2	11.3	5.1
C	12.5	3	11.5	3.13
D	6.3	0	14.4	0
E	3.2	2	0.42	2
F	1.6	9	14.1	9
G	0.8	4	14.1	4
H	0.4	1	2.3	1
J	0.2	7	1.8	8
EMD				3.735

Table 14: ECP, FCP, EEV, EMD values for sample mixture data set with t=200 and tf=2% by running pipeline 2 with quality trim 25

		bbduk_qtrim25_CliqueSNV_MIX1_t_200_tf_2		
variant	frequency	ECP	FCP%	EEV
A	50	2	13.7	3.8
B	25	1	12.9	4.6
C	12.5	0	23.1	0
D	6.3	0	19.8	0
E	3.2	1	13.7	1
F	1.6	6	1.9	6
G	0.8	1	3.3	1
H	0.4	5	5.5	8
J	0.2	8	1.1	9
EMD				3.245

Table 15: ECP, FCP, EEV, EMD values for sample mixture data set with t=500 and tf=5% by running pipeline 2 with quality trim 25

		bbduk_qtrim25_CliqueSNV_MIX1_t_500_tf_5		
variant	frequency	ECP	FCP%	EEV
A	50	2	27.2	3.78
B	25	1	14.1	3.6
C	12.5	1	27.4	1
D	6.3	0	27.2	0
E	3.2	5	20.8	8.25
F	1.6	8	27.4	8
G	0.8	7	27.4	7
H	0.4	6	27.2	7
J	0.2	8	27.4	8
EMD				3.411

Table 16: ECP, FCP, EEV, EMD values for sample mixture data set with t=10 and tf=0.033% by running pipeline 2 with quality trim 30

		bbduk_qtrim30_CliqueSNV_MIX1_t_10_tf_0.033		
variant	frequency	ECP	FCP%	EEV
A	50	0	13.8	3.3
B	25	2	10.7	4.88
C	12.5	3	10.6	3.8
D	6.3	0	10.3	0
E	3.2	2	0.92	3.8
F	1.6	6	5.06	6
G	0.8	3	6.9	3
H	0.4	4	2.7	4
J	0.2	3	0.58	3
EMD				3.66

Table 17: ECP, FCP, EEV, EMD values for sample mixture data set with t=50 and tf=0.5% by running pipeline 2 with quality trim 30

		bbduk_qtrim30_CliqueSNV_MIX1_t_50_tf_0.5		
variant	frequency	ECP	FCP%	EEV
A	50	0	10.7	3.34
B	25	1	3.6	3.39
C	12.5	3	3.1	3.5
D	6.3	0	13.5	0
E	3.2	2	3.1	3.19
F	1.6	4	0.5	7.83
G	0.8	0	2.5	0
H	0.4	4	2.5	4
J	0.2	3	0.56	3
EMD				3.211

Table 18: ECP, FCP, EEV, EMD values for sample mixture data set with t=100 and tf=1% by running pipeline 2 with quality trim 30

		bbduk_qtrim30_CliqueSNV_MIX1_t_100_tf_1		
variant	frequency	ECP	FCP%	EEV
A	50	0	13.03	3.34
B	25	1	10.4	3.11
C	12.5	3	12.7	3.17
D	6.3	0	13.6	0
E	3.2	3	7.6	3.8
F	1.6	8	12.7	8
G	0.8	3	12.7	3
H	0.4	4	13.03	6
J	0.2	7	1.7	7
EMD				3.156

Table 19: ECP, FCP, EEV, EMD values for sample mixture data set with t=200 and tf=2% by running pipeline 2 with quality trim 30

		bbduk_qtrim30_CliqueSNV_MIX1_t_200_tf_2		
variant	frequency	ECP	FCP%	EEV
A	50	1	2.9	3.74
B	25	1	12.1	4.5
C	12.5	0	27.2	0
D	6.3	0	15.1	0
E	3.2	2	1.2	2
F	1.6	7	27.2	8.9
G	0.8	1	2.3	1
H	0.4	4	12.7	11
J	0.2	7	5.9	9
EMD				3.299

Table 20: ECP, FCP, EEV, EMD values for sample mixture data set with t=500 and tf=5% by running pipeline 2 with quality trim 30

		bbduk_qtrim30_CliqueSNV_MIX1_t_500_tf_5		
variant	frequency	ECP	FCP%	EEV
A	50	2	22.7	4.1
B	25	1	14.8	2.3
C	12.5	3	23.07	3
D	6.3	0	22.7	0
E	3.2	5	22.7	10
F	1.6	10	23.07	10
G	0.8	9	23.07	9
H	0.4	6	22.7	8
J	0.2	8	23.07	8
EMD				3.654

3.2.3 Pipeline3

Raw reads are aligned with BWA aligner using the reference read. Mapping gives aligned sam files. Running BBduk on these sam files with a quality barrier drops the reads below a specified quality barrier. Shiver is run on these trimmed reads. It has 2 Steps. The first step is to get contigs and the second step is to map the reads to the constructed reference. Running CliqueSnv on these mapped reads with different t and tf values gives the haplotypes.

Figure 5: Figure explaining steps in pipeline 3

Table 21: Number of aligned reads after BBduk trimming and after shiver, with different quality trimming values of 20,25 and 30.

	Number of reads in sam file	Number of reads after running shiver
Without BBduk quality trim	5,604,910	-
Quality trim = 20	2,682,384	2,431,605
Quality trim = 25	1,048,476	957,537
Quality trim = 30	220,542	197,949

The sample reads are fastq files with two files the forward read, and the reverse read. BWA aligner aligns the forward read and reverse read and gives output a sam file which contains the reads after alignment using the reference read. The reference read used is HXB2. The sam file

obtained with aligned reads is the input for BBduk. BBduk trims the reads below a certain quality value we define while running the tool from the command prompt. In this thesis the BBduk is run with three different quality values 20,25 and 30. Shiver takes the trimmed reads i.e input from BBduk and gives the bam file of mapped reads to the constructed reads as output. The bam is converted to sam as input for CliqueSNV, with t and tf values, gives the haplotypes.

3.2.3.1 Results

Table 22: ECP, FCP, EEV, EMD values for sample mixture data set with t=10 and tf=0.033% by running pipeline 3 with quality trim 20

		bbduk_qtrim20_shiver_CliqueSNV_MIX1_t_10_tf_0.033		
variant	frequency	ECP	FCP%	EEV
A	50	1	5.01	3.3
B	25	1	11.6	1.06
C	12.5	0	16.8	0
D	6.3	1	5.01	1
E	3.2	1	1.76	1
F	1.6	3	0.4	5.6
G	0.8	3	3.3	3
H	0.4	1	1.1	1
J	0.2	1	1.1	1
EMD				2.93

Table 23: ECP, FCP, EEV, EMD values for sample mixture data set with t=50 and tf=0.5% by running pipeline 3 with quality trim 20

		bbduk_qtrim20_shiver_CliqueSNV_MIX1_t_50_tf_0.5		
variant	frequency	ECP	FCP%	EEV
A	50	2	3.7	4.6
B	25	1	14.4	3.95
C	12.5	0	21.8	0
D	6.3	1	22.6	1
E	3.2	1	3.7	1
F	1.6	6	1.4	6.08
G	0.8	3	4.4	3
H	0.4	0	1.3	0
J	0.2	1	1.8	1
EMD				3.545

Table 24: ECP, FCP, EEV, EMD values for sample mixture data set with t=100 and tf=1% by running pipeline 3 with quality trim 20

		bbduk_qtrim20_shiver_CliqueSNV_MIX1_t_100_tf_1		
variant	Frequency	ECP	FCP%	EEV
A	50	0	26.08	2.9
B	25	1	13.8	4.12
C	12.5	0	28.7	0
D	6.3	1	15.3	1
E	3.2	2	4.7	2
F	1.6	7	28.7	7
G	0.8	1	4.1	1
H	0.4	4	26.8	5
J	0.2	8	4.7	9
EMD				2.765

Table 25: ECP, FCP, EEV, EMD values for sample mixture data set with t=200 and tf=2% by running pipeline 3 with quality trim 20

		bbduk_qtrim20_shiver_CliqueSNV_MIX1_t_200_tf_2		
variant	frequency	ECP	FCP%	EEV
A	50	0	26.03	2.8
B	25	1	13.8	4.11
C	12.5	0	27.2	0
D	6.3	1	15.35	1
E	3.2	2	4.7	2
F	1.6	7	27.2	7
G	0.8	1	3.1	1
H	0.4	4	26.03	5
J	0.2	8	4.7	9
EMD				2.727

Table 26: ECP, FCP, EEV, EMD values for sample mixture data set with t=500 and tf=5% by running pipeline 3 with quality trim 20

		bbduk_qtrim20_shiver_CliqueSNV_MIX1_t_500_tf_5		
variant	frequency	ECP	FCP%	EEV
A	50	0	30.6	2.32
B	25	1	14.2	4.02
C	12.5	0	32.3	0
D	6.3	1	15.3	1
E	3.2	3	30.6	5
F	1.6	7	32.3	7
G	0.8	6	32.3	6
H	0.4	4	30.6	14
J	0.2	9	7.4	9
EMD				2.625

Table 27: ECP, FCP, EEV, EMD values for sample mixture data set with t=10 and tf=0.033% by running pipeline 3 with quality trim 25

		bbduk_qtrim25_shiver_CliqueSNV_MIX1_t_10_tf_0.033		
variant	frequency	ECP	FCP%	EEV
A	50	2	13.4	4.5
B	25	2	9.1	4.6
C	12.5	3	0.7	3.1
D	6.3	0	13.4	0
E	3.2	0	3.2	0
F	1.6	4	0.5	8.6
G	0.8	1	2.1	1
H	0.4	1	2.6	1
J	0.2	3	0.62	3
EMD				3.984

Table 28: ECP, FCP, EEV, EMD values for sample mixture data set with t=50 and tf=0.5% by running pipeline 3 with quality trim 25

		bbduk_qtrim25_shiver_CliqueSNV_MIX1_t_50_tf_0.5		
variant	frequency	ECP	FCP%	EEV
A	50	2	14.5	4.7
B	25	2	9.5	4.4
C	12.5	1	1	3
D	6.3	0	14.5	0
E	3.2	0	3.03	0.2
F	1.6	5	0.9	7.4
G	0.8	1	2.4	1
H	0.4	1	2.5	1
J	0.2	3	0.6	3
EMD				3.98

Table 29: ECP, FCP, EEV, EMD values for sample mixture data set with t=100 and tf=1% by running pipeline 3 with quality trim 25

		bbduk_qtrim25_shiver_CliqueSNV_MIX1_t_100_tf_1		
variant	frequency	ECP	FCP%	EEV
A	50	1	13.5	4.3
B	25	1	10.7	3.24
C	12.5	3	9.7	3.25
D	6.3	0	15.1	0
E	3.2	1	3.3	1
F	1.6	10	4.02	10
G	0.8	3	4.02	3
H	0.4	1	2.9	1
J	0.2	8	5.4	8
EMD				3.617

Table 30: ECP, FCP, EEV, EMD values for sample mixture data set with t=200 and tf=2% by running pipeline 3 with quality trim 25

		bbduk_qtrim25_shiver_CliqueSNV_MIX1_t_200_tf_2		
variant	frequency	ECP	FCP%	EEV
A	50	1	19.1	3.8
B	25	1	10.6	2.66
C	12.5	0	22.04	0
D	6.3	0	18.7	0
E	3.2	2	19.1	8.2
F	1.6	7	22.04	7
G	0.8	3	2.1	3
H	0.4	5	19.1	11
J	0.2	8	1.8	9
EMD				3.049

Table 31: ECP, FCP, EEV, EMD values for sample mixture data set with t=500 and tf=5% by running pipeline 3 with quality trim 25

		bbduk_qtrim25_shiver_CliqueSNV_MIX1_t_500_tf_5		
variant	frequency	ECP	FCP%	EEV
A	50	2	25.2	4.5
B	25	1	10.6	3.15
C	12.5	0	24.05	0
D	6.3	0	25.26	0
E	3.2	5	25.26	10
F	1.6	7	24.05	7
G	0.8	6	24.05	6
H	0.4	6	25.26	11
J	0.2	7	10.61	9
EMD				3.601

Table 32: ECP, FCP, EEV, EMD values for sample mixture data set with t=10 and tf=0.033% by running pipeline 3 with quality trim 30

		bbduk_qtrim30_shiver_CliqueSNV_MIX1_t_10_tf_0.033		
variant	frequency	ECP	FCP%	EEV
A	50	0	11.77	3.75
B	25	2	1.6	4.16
C	12.5	2	0.5	3.04
D	6.3	0	11.9	0
E	3.2	2	4.2	6.25
F	1.6	5	0.6	8.32
G	0.8	1	1.6	1
H	0.4	3	2.09	3
J	0.2	4	0.35	4
EMD				3.659

Table 33: ECP, FCP, EEV, EMD values for sample mixture data set with t=50 and tf=0.5% by running pipeline 3 with quality trim 30

		bbduk_qtrim30_shiver_CliqueSNV_MIX1_t_50_tf_0.5		
variant	frequency	ECP	FCP%	EEV
A	50	1	12.3	4.38
B	25	2	0.36	3.44
C	12.5	2	3.37	3.6
D	6.3	1	12.01	1
E	3.2	1	2.65	1.34
F	1.6	5	0.85	6.39
G	0.8	2	1.9	2
H	0.4	3	2.3	3
J	0.2	4	0.61	4
EMD				3.757

Table 34: ECP, FCP, EEV, EMD values for sample mixture data set with t=100 and tf=1% by running pipeline 3 with quality trim 30

		bbduk_qtrim30_shiver_CliqueSNV_MIX1_t_100_tf_1		
variant	frequency	ECP	FCP%	EEV
A	50	0	3.8	4.49
B	25	2	3.9	3.06
C	12.5	2	4.2	3.95
D	6.3	1	12.01	1
E	3.2	1	2.5	1.94
F	1.6	9	4.2	11.32
G	0.8	1	1.8	1
H	0.4	1	0.77	1
J	0.2	8	3.8	8
EMD				3.84

Table 35: ECP, FCP, EEV, EMD values for sample mixture data set with t=200 and tf=2% by running pipeline 3 with quality trim 30

		bbduk_qtrim30_shiver_CliqueSNV_MIX1_t_200_tf_2		
variant	frequency	ECP	FCP%	EEV
A	50	1	0.21	3.72
B	25	1	10.4	4.2
C	12.5	0	22.2	0
D	6.3	0	13.8	0
E	3.2	2	0.21	3
F	1.6	7	22.2	8.9
G	0.8	2	1.8	2
H	0.4	4	11.9	11
J	0.2	8	1.3	8
EMD				3.23

Table 36: ECP, FCP, EEV, EMD values for sample mixture data set with t=500 and tf=5% by running pipeline 3 with quality trim 30

		bbduk_qtrim30_shiver_CliqueSNV_MIX1_t_500_tf_5		
variant	frequency	ECP	FCP%	EEV
A	50	2	20.5	4.14
B	25	1	13.8	2.33
C	12.5	5	30.1	5
D	6.3	0	20.5	0
E	3.2	5	20.5	8
F	1.6	12	30.1	12
G	0.8	9	30.1	9
H	0.4	6	20.5	6
J	0.2	8	30.1	8
EMD				3.84

4 CONCLUSIONS

This thesis developed 3 pipelines for inference of rare genetically-related viral variants, which allows for accurate haplotyping in the presence of high sequencing error rates and which is also suitable for both single-molecule and short-read sequencing. Using experimental data, CliqueSNV demonstrates that CliqueSNV can detect haplotypes with frequencies as low as 0.1%, which is comparable to the sensitivity of many deep sequencing-based point mutation detection methods. CliqueSNV has its limitations. For example, with large number of reads cliqueSNV doesn't run with default parameters. It throws memory out of bounds exception. The processing speeds of the CliqueSNV can be increased by adding BBduk and shiver to its pipeline. The results shoe that with BBduk quality parameter of 20 pipeline 3 is shown to produce more accurate haplotypes. Future work might include adding the pipelines to galaxy.

REFERENCES

[1] 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 (TCBB), **11**(1): 182–191, 2014.

[2] M.A. Quail, M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, and Y. Gu. A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina miseq sequencers. BMC genomics, **13**(1): 341, 2012.

[3] A. R. Macalalad, M. C. Zody, P. Charlebois, N. J. Lennon, R. M. Newman, C. M. Malboeuf, E. M. Ryan, C. L. Boutwell, K. A. Power, D. E. Brackney, et al. Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data. PLoS computational biology, **8**(3):e1002417, 2012.

[4] X. Yang, P. Charlebois, A. Macalalad, M. R. Henn, and M. C. Zody. V-phaser 2: variant inference for viral populations. BMC genomics, **14**(1): 674, 2013.

[5] A. Routh, M. W. Chang, J. F. Okulicz, J. E. Johnson, and B. E. Torbett. Covama: Co-variation mapper for disequilibrium analysis of mutant loci in viral populations using next-generation sequence data. Methods, **91**: 40–47, 2015.

[6] A. Artyomenko, N. C. Wu, S. Mangul, E. Eskin, R. Sun, and A. Zelikovsky. Long single-molecule reads can resolve the complexity of the influenza virus composed of rare, closely related mutant variants. In *International Conference on Research in Computational Molecular Biology*, pages 164–175. Springer International Publishing, 2016.

[7] Chris Wymant, François Blanquart, Tanya Golubchik, Astrid Gall, Margreet Bakker, Daniela Bezemer, Nicholas J Croucher, Matthew Hall, Mariska Hillebregt, Swee Hoe Ong, Oliver

Ratmann, Jan Albert, Norbert Bannert, Jacques Fellay, Katrien Fransen, Annabelle Gourlay, M Kate Grabowski, Barbara Gunsenheimer-Bartmeyer, Huldrych F Günthard, Pia Kivelä, Roger Kouyos, Oliver Laeyendecker, Kirsi Liitsola, Laurence Meyer, Kholoud Porter, Matti Ristola, Ard van Sighem, Ben Berkhout, Marion Cornelissen, Paul Kellam, Peter Reiss, Christophe Fraser, BEEHIVE Collaboration, Easy and accurate reconstruction of whole HIV genomes from short-read sequence data with shiver, *Virus Evolution*, Volume 4, Issue 1, January 2018, vey007, https://doi.org/10.1093/ve/vey007

[8] Martin Hunt, Astrid Gall, Swee Hoe Ong, Jacqui Brener, Bridget Ferns, Philip Goulder, Eleni Nastouli, Jacqueline A. Keane, Paul Kellam, Thomas D. Otto, IVA: accurate *de novo* assembly of RNA virus genomes, *Bioinformatics*, Volume 31, Issue 14, 15 July 2015, Pages 2374–2376, https://doi.org/10.1093/bioinformatics/btv120

[9] Ruby J.G.et al. (2013) PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. *G3*,3, 865–880.

[10] Anthony M. Bolger, Marc Lohse, Bjoern Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, *Bioinformatics*, Volume 30, Issue 15, 1 August 2014, Pages 2114–2120, https://doi.org/10.1093/bioinformatics/btu170.

[11] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, Volume 25, Issue 16, 15 August 2009, Pages 2078–2079, https://doi.org/10.1093/bioinformatics/btp352.

[12] Kurtz, S., Phillippy, A., Delcher, A.L. *et al.* Versatile and open software for comparing large genomes. *Genome Biol* **5,** R12 (2004). https://doi.org/10.1186/gb-2004-5-2-r12.

[13] BBMap – Bushnell B. – sourceforge.net/projects/bbmap/

[14] CliqueSNV: Scalable Reconstruction of Intra-Host Viral Populations from NGS Reads Sergey Knyazev, Viachaslau Tsyvina, Andrew Melnyk, Alexander Artyomenko, Tatiana Malygi n, Yuri B. Porozov, Ellsworth Campbell, William M. Switzer, Pavel Skums, Alex Zelikovsky bioRxiv 264242; doi: https://doi.org/10.1101/264242.

[15] Poltronieri P, Sun B, Mallardo M. RNA Viruses: RNA Roles in Pathogenesis, Coreplication and Viral Load. *Curr Genomics*. 2015;16(5):327–335. doi:10.2174/1389202916666150707160613.