Spring 5-4-2021

Algorithms for analysis of next-generation viral sequencing data

Andrii Melnyk

ALGORITHMS FOR ANALYSIS OF NEXT-GENERATION VIRAL SEQUENCING DATA

by

ANDRII MELNYK

Under the Direction of Alexander Zelikovsky, PhD

ABSTRACT

RNA viruses mutate at extremely high rates, forming an intra-host viral population of closely related variants, which allows them to evade the host's immune system and makes them particularly dangerous. Viral outbreaks pose a significant threat for public health. Progress of sequencing technologies made it possible to identify and sample intra-host viral populations at great depth. Consequently, the contribution of sequencing technologies to molecular surveillance of viral outbreaks becomes more and more substantial. Genome

sequencing of viral populations reveals similarities between samples, allows to measure viral genetic distance and facilitate outbreak identification and isolation. Computational methods can be used to infer transmission characteristics from sequencing data. However, due to the specifics of next-generation sequencing (NGS) approaches, and the limited availability of viral data, existing methods lack accuracy and efficiency. In this dissertation, I present a novel, flexible methods, that allow tackling crucial epidemiological problems, such as identification of transmission clusters, sources of infection, and transmission direction.

INDEX WORDS: Genetic relatedness, transmission networks, outbreak investigations, clustering.

ALGORITHMS FOR ANALYSIS OF NEXT-GENERATION VIRAL SEQUENCING DATA

by

ANDRII MELNYK

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

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2021

ALGORITHMS FOR ANALYSIS OF NEXT-GENERATION VIRAL SEQUENCING DATA

by

ANDRII MELNYK

Committee Chair: Alexander Zelikovsky

Committee: Pavel Skums

Robert Harrison

Yury Khudyakov

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

May 2021

DEDICATION

To my parents and my girlfriend Lindsey Green.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Alex Zelikovsky. Working under his supervision was an extremely valuable experience, that allowed me to grow both as a scientist and an individual. I would also like to thank Dr. Pavel Skums for his help. Finally, I would like to thank other Ph.D. candidates I was lucky to work with - Sergey Knyazev, Pelin B. Icer, Seth Sims and Viachaslau Tsyvina.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

- NGS - Next Generation Sequencing
- EMD - Earth Mover's Distance
- IAV - Influenza A virus
- HCV - Hepatitis C virus
- HIV - Human immunodeficiency virus
- SARS-CoV2 - Severe acute respiratory syndrome coronavirus 2
- RNA - Ribonucleic acid
- EPLD - End-point limiting-dilution
- ROC - Receiver operating characteristic
- SNV - Single nucleotide variant

PART 1

INTRODUCTION

To tackle a virus, we first need to identify and study it. However, this is complicated by the fact that most viruses are too small for the light microscope. Sequencing, on the other hand, doesn't have this limitation. As a consequence, it is being extensively used during basic and clinical research, infection diagnosis, molecular epidemiology and drug-resistance testing.

Introduction of next-generation sequencing and development of new sequencing technologies, such as 454, Illumina, SOLiD and Ion Torrent, fundamentally changed the field of biological and medical sciences, and drastically increased the role of bioinformatics. Significant decrease in the cost of sequencing resulted in rapid increase in the amount of data available, thus posing new problems, that require development of new computational methods. For instance, Sanger [2], one of the earliest sequencing technologies, that is known for low error rates [3], became impractical due to high sequencing costs, and was recently surpassed by NGS by number of viral sequences in NIH genetic sequence database GenBank [4].

As recent advances in sequencing allowed to identify viral populations at great depth, new opportunities for dealing with crucial epidemiological tasks, such inference of relatedness between viral samples, identification of transmission clusters and sources of infection, were introduced.

This dissertation presents novel algorithms for analysis of intra- and inter-host viral populations for NGS data, aimed to tackle essential epidemiological tasks. In particular, I designed an algorithm, that doesn't rely on read assembly and allows to cluster genetically related samples, infer transmission directions and predict sources of outbreaks.

1.1 Background

A large number of medically important viruses, including HIV, hepatitis C virus, and influenza, have RNA genomes [1]. These viruses replicate with extremely high mutation rates and exhibit significant genetic diversity, which allows viral populations to rapidly adapt to dynamic environments, and evolve resistance to vaccines and antiviral drugs.

Viral quasispecies is a viral population, that is represented by a cloud of diverse variants, that are genetically linked through mutation and interact on a functional level, collectively contributing to the characteristics of the population [1].

Figure 1.1 illustrates a virus replicating with a high mutation rate. Over a course of a few generations, a diverse mutant repertoire is generated. In the demonstrated trees, each branch indicates two variants, linked by a point mutation, and the concentric circles represent serial replication cycles.The resulting distribution is often represented as a cloud centered on a master sequence. This two dimensional schematic is a vast oversimplification of the intraquasispecies connectivity. In the mathematical formulations of quasispecies theory, sequence space is multidimensional, with numerous branches between variants.

It has been shown, that the structure of viral quasispecies affects virulence [5] and pathogenesis [6]. Furthermore, certain low-frequency genetic variants may contain mutations, which allows viruses to be stay unaffected by the selective pressure of host immune responses [7] and anti-viral drug treatment [8]. Even though NGS is currently being introduced into clinical diagnostics, single-nucleotide variant (SNV) calling is still widely used for assessing of viral quasispecies structure. However, this approach is limited, because it ignores patterns of co-occurrence among mutations, which is critically important for RNA viruses, which have abundant epistatic interactions [9]. Thus, inferring the underlying mix of haplotypes (viral quasispecies assembly) is necessary for viral phenotypes prediction [10].

Genome sequencing of viral populations reveals similarities between samples, allows to measure viral genetic distance, and to facilitate outbreak identification and isolation. Computational methods can be used to infer transmission characteristics from sequencing

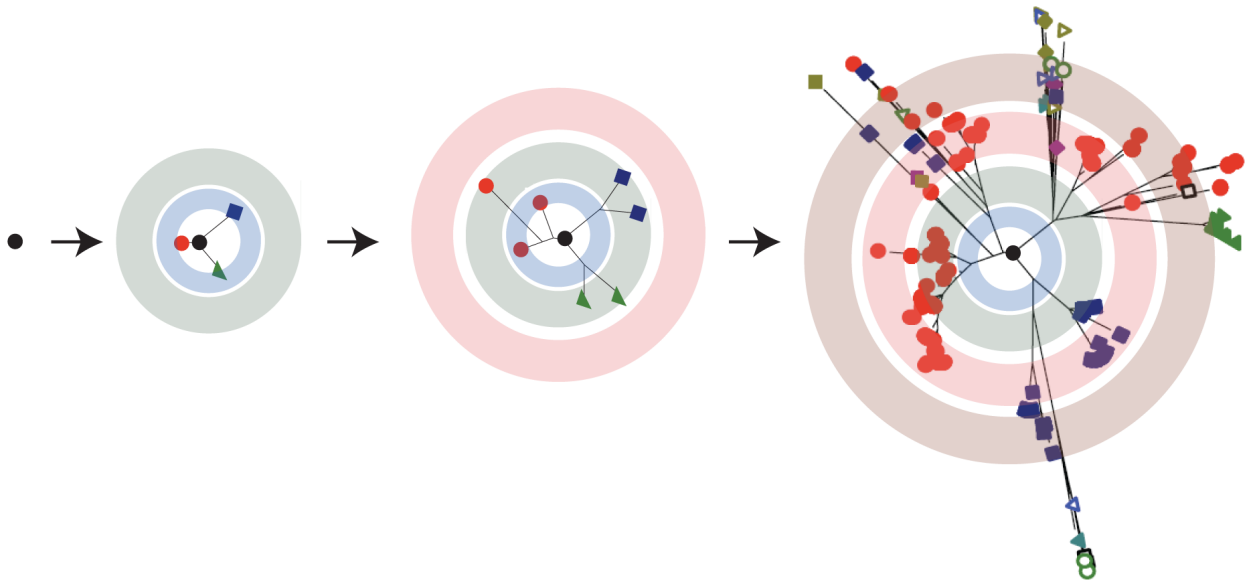

Figure (1.1) RNA viruses exist as quasispecies [1]

data. These methods typically utilize the simple observation that all samples from the same outbreak are genetically related, so they use some measure of genetic relatedness as a predictor for epidemiological relatedness [11–13]. They usually rely on assembled sequences, thus requiring an extra step preprocessing step when dealing with raw NGS reads. MiSeq [14] is a popular NGS technology, that is used to sequence viral samples and detect rare viral mutations. Since MiSeq reads are short, their alignment and assembly for rapidly mutating RNA viruses is error-prone and complicated, which makes it appealing to develop an approach, that will allow to skip alignment and assembly steps.

1.2 Problems

In this dissertation, several problems, related to viral NGS data analysis are addressed.

1.2.1 Viral outbreak investigation

Essential epidemiological tasks (T1-T5) were considered, where T1-T2 are applied to 2 hosts, and T3-T5 are applied to multiple hosts.

T1. **Identification of relatedness:**

Given: NGS reads from hosts A and B

Decide: Whether A and B are related (whether they belong to the same outbreak)

T2. **Identification of transmission direction:**

Given: NGS reads from hosts A and B

Decide: Whether host A infected B or B infected A

T3. **Identification of transmission clusters:**

Given: NGS reads from a set of hosts

Find: The transmission clusters corresponding to individual outbreaks

T4. **Presence of outbreak source:**

Given: NGS reads from a set of hosts

Decide: Whether outbreak source is present among sequenced hosts

T5. **Identification of outbreak source:**

Given: NGS reads from a set of hosts

Find: Outbreak source

Identifying whether 2 hosts belong to the same outbreak (T1) and transmission direction between them (T2) are tasks, that have to be solved in order to find transmission chains. Another important task is to discover boundaries of an outbreak (T3). Once hosts, that belong to an outbreak are obtained, it is critical to design whether the source is among them (T4). Finally, identifying the main spreader of an outbreak (T5) is a crucial epidemiological task, by solving which outbreak spreading can be prevented.

1.2.2 Viral quasispecies assembly

In formal definitions, we adopt a terminology, based on papers [15] and [10]. Let R be a collection of NGS sequencing reads, which are sequences of the alphabet of nucleotides $\{A,C,G,T,N\}$, where N is a common placeholder for unknown nucleotides. Let $A := A(R)$ be the set of their alignments to a reference genome, as computed by a read aligner.

Problem 1. *Given R and A, find a set of master contigs, each representing a group of very closely related viral haplotypes.*

Problem 2. *Given R, find a set of master contigs, each representing a group of very closely related viral haplotypes.*

It can be easily seen that the problem in [15] (Problem 2) is a more general case of the problem in [10] (Problem 1), because it implies inference of viral haplotypes without the alignments to a reference sequence.

1.3 Contributions

This dissertation presents multiple contributions to the analysis of viral NGS data. These contributions include new algorithms for viral outbreak investigation and viral haplotype assembly.

Introduced outbreak investigation methods allow to cluster genetically related samples, infer transmission directions and predict sources of outbreaks. Among the main advantages of proposed algorithms is the ability to bypass cumbersome read assembly, thus eliminating the chance to introduce new errors, and allowing to save processing time by using raw NGS reads (k-mers EMD). Additionally, while some viral outbreak investigation algorithms involve building transmission networks or phylogenetic trees, introduced algorithms for clustering of viral outbreak data provide an efficient alternative, that uses cluster entropy to capture the underlying process of viral mutation.

All algorithms are applicable to the analysis of outbreaks highly heterogeneous RNA viruses.

Proposed haplotype assembly method allows for accurate haplotyping in the presence of high sequencing error rates, which is also suitable for both single-molecule and short-read sequencing. In contrast to other haplotyping methods, it infers viral haplotypes by detection of clusters of statistically linked SNVs rather than through assembly of overlapping reads used with methods such as Savage [16] and can successfully infer and reconstruct viral variants,

which differ by only a few mutations, thus demonstrating the high precision of identifying closely related variants. Another significant advantage of CliqueSNV is its low computation time, which is achieved by a very fast searching of linked SNV pairs and the application of the special graph-theoretical approach to SNV clustering.

1.4 Roadmap

This dissertation is organized as follows. Chapter 1 presents a highlight of viral quasispecies assembly problem along with main epidemiological problems, that arise during viral outbreak investigation and existing methods in these contexts.

In the following chapters novel and efficient algorithms related to viral NGS data analysis are presented. In particular, Chapter 2 presents a novel intra-host viral data analysis algorithms. Clustering-based identification of SARS-CoV-2 subtypes is as a viable and scalable alternative to unveiling trends in the spread of SARS-CoV-2. k-mers EMD provides competitive performance and allows more flexibility compared to existing approaches. VOICE is an evolutionary simulation method for genetic relatedness inference. Chapter 3 proposes a viral haplotype assembly method for rapid and accurate inference of viral populations, applicable to clinical and epidemiological NGS data.

Discussion and future directions are provided in the Chapter 4.

1.5 Products

1.5.1 Publications

Journal Papers

1. S. Knyazev, V. Tsyvina, A. Shankar, **A. Melnyk**, A. Artyomenko, T. Malygina, Y. Porozov, E. Campbell, S. Mangul, W. Switzer, P. Skums, and A. Zelikovsky (under revision) Accurate Assembly of Minority Viral Haplotypes from Next-Generation Sequencing through Efficient Noise Reduction. Nucleic Acids Research

2. **A. Melnyk**, F. Mohebbi, S. Knyazev, B. Sahoo, R. Hosseini, P. Skums, A. Zelikovsky, M. Patterson (to appear) Clustering based identification of SARS-CoV-2 subtypes. 10th International Conference on Computational Advances in Bio and medical Sciences (ICCABS 2020), Virtual Conference, December 10-12, 2020, Revised Selected Papers

3. **A. Melnyk**, S. Knyazev, F. Vannberg, L. Bunimovich, P. Skums, A. Zelikovsky (2020) Using Earth Mover's Distance for Viral Outbreak Investigations. BMC, DOI: https://doi.org/10.1186/s12864-020-06982-4

4. O. Glebova, S. Knyazev, **A. Melnyk**, A. Artyomenko, Y. Khudyakov, A. Zelikovsky, P. Skums (2017) Inference of genetic relatedness between viral quasispecies from sequencing data. BMC Genomics, 18(Suppl 10):918, DOI: 10.1186/s12864-017-4274-5

Conference Abstracts

1. **A. Melnyk**, S. Knyazev, Y. Khudyakov, F. Vannberg, L. Bunimovich, P. Skums, A. Zelikovsky (2019) Using Earth Mover's Distance for Viral Outbreak Investigations. 15th International Symposium on Bioinformatics Research and Applications (ISBRA)

2. S. Knyazev, V. Tsyvina, **A. Melnyk**, A. Artyomenko, T. Malygina, Y. Porozov, E. Campbell, W. Switzer, P. Skums, and A. Zelikovsky (2018) CliqueSNV: Scalable Reconstruction of Intra-Host Viral Populations from NGS Reads. The 8th RECOMB Satellite Workshop on Massively Parallel Sequencing (RECOMB-Seq)

1.5.2 Presentations

1. **A. Melnyk**, F. Mohebbi, S. Knyazev, B. Sahoo, R. Hosseini, P. Skums, A. Zelikovsky, M. Patterson (2021) Clustering based identification of SARS-CoV-2 subtypes. 10th International Conference on Computational Advances in Bio and medical Sciences (ICCABS)

2. **A. Melnyk**, S. Knyazev, Y. Khudyakov, F. Vannberg, L. Bunimovich, P. Skums, A. Zelikovsky (2019) Using Earth Mover's Distance for Viral Outbreak Investigations. 15th International Symposium on Bioinformatics Research and Applications (ISBRA)

1.5.3 Software Packages

- **k-mers EMD** - Inference of gentic relatedness for viral samples using Earth Mover's Distance. https://github.com/amelnyk34/kemd.

PART 2

ALGORITHMS FOR VIRAL OUTBREAK INVESTIGATION

2.1 Introduction

RNA viruses mutate at extremely high rates, forming an intra-host viral population of closely related variants (or quasi-species). Their high variability [17] allows them to evade the host's immune system and makes them particularly dangerous. Viral outbreaks pose a significant threat for public health, and, in order to deal with it, it is critical to infer transmission clusters, i.e., decide whether two viral samples belong to the same outbreak.

The progress of sequencing technologies made it possible to identify and sample intra-host viral populations at great depth [18–23]. Consequently, contribution of sequencing technologies to molecular surveillance of viral outbreaks becomes more and more substantial. Genome sequencing of viral populations reveals similarities between samples, allows to measure viral genetic distance, and to facilitate outbreak identification and isolation. Computational methods can be used to infer transmission characteristics from sequencing data. The first question usually is whether two viral populations belong to the same outbreak. The methods typically utilize the simple observation that all samples from the same outbreak are genetically related, so they use some measure of genetic relatedness as a predictor for epidemiological relatedness [11–13].

The second question is which samples constitute isolated outbreaks. For this purposes, we define a transmission cluster as a connected set of genetically related viral populations. The third questions we address in this article is "Who is the source of infection?". This questions is the most difficult to answer, and there were only a few attempts to do it computationally using solely genomic data [24] without invoking additional epidemiological information [25]. To the best of our knowledge, there is still no freely available computational tool for this problem.

Computational methods for detection of viral transmissions and inference of transmission clusters are often consensus-based, i.e. they analyze only a single representative sequence per intra-host population (for example, consensus sequence). Such methods assign two hosts into one transmission cluster, if the distances between corresponding sequences do not exceed a predefined threshold [11,12]. Although consensus-based methods proved to be useful, they do not take into account intra-host viral diversity. Inclusion of whole intra-host populations into analysis is important, because minor viral variants are frequently responsible for transmission of RNA viruses [26,27].

Recently published computational approach (further referred to as MinDist) [13] uses the minimal genetic distance between sequences of two viral populations as a measure of genetic relatedness of intra-host viral populations. Since minimal genetic distances between different pairs of populations can be achieved on various pairs of sequences, this approach takes into account intra-host diversity.

However, both consensus-based and MinDist approaches have further limitations. First of all, they do not allow to detect directions of transmissions, which is crucial for detection of outbreak sources and transmission histories. Secondly, distance thresholds utilized by both approaches could be derived from analysis of limited or incomplete experimental data and highly data- and situation-specific, with different viruses or even different genomic regions of the same virus requiring specifically established thresholds. Additionally, MiSeq [14], a popuar NGS technology, that is used to sequence viral samples and detect rare viral mutations, produces short reads. Their alignment and assembly for rapidly mutating RNA viruses is error-prone and complicated, which makes it appealing to develop an approach, that will allow to skip alignment and assembly steps. Finally, existing clustering approaches involve building transmission networks or phylogenetic trees, which, due to high computational complexity, makes their application problematic when it comes to rapidly growing datasets.

In this dissertation, several novel algorithms, that address above limitations are proposed. The new algorithms allow to infer important epidemiological characteristics, including

genetic relatedness, directions of transmissions and transmission clusters.

- SARS-CoV-2 clustering method that applies CliqueSNV to inter-host SARS-CoV-2 viral sequences and uses cluster *entropy* to measure the clustering quality.
- $k-mers$ Earth Mover's Distance ($k-mers$ EMD) method applies an alignment- and assembly-free k-mer strategy to intra-host viral sequencing data analysis.
- Viral Outbreak InferenCE ($VOICE$) is a simulation-based method which imitates viral evolution as a Markov process in the space of observed viral haplotypes.

Proposed algorithms were validated on the experimental data obtained from HCV outbreaks. Comparative results suggest that introduced methods are efficient in epidemiological characteristics inference.

2.2 Methods

2.2.1 Clustering based identification of SARS-CoV-2 subtypes

Proposed algorithm clusters sequences of SARS-CoV-2 based purely on sequence content, and under no *a priori* hypothesis about the relationships between these sequences, *i.e.*, it is unsupervised. Additionally information from the clustering is used to patch gaps in the sequences, so that the aim is to fill gaps in sequences with the objective of minimizing the entropy of the result.

CliqueSNV-based clustering

For clustering viral subtypes, introduced approach proposes to use existing tools for identification of the intra-host viral populations subtypes from NGS data reviewed in [28], *e.g.*, Savage [16], PredictHaplo [29], aBayesQR [30], etc. However, in this context the setting is different, since the data consists of large collections of *inter-host* consensus sequences, gathered from different regions and countries around the world [31, 32]. Thus, the "host" is now an entire region or country, and variants and their dynamics within these regions or

countries are reconstructed. The SARS-CoV-2 sequences in GISAID are consensus sequences of approximate length 30K. Such sequences by quality and length have similar properties as PacBio reads. The algorithm uses CliqueSNV since it performed very well on PacBio reads [33]. Default parameters are used to run CliqueSNV, setting the minimum cluster frequency to be at least 1% of the population.

k-modes clustering

Since proposed algorithm clusters sequences, which are on the *categories* A, C, G, T (and –, a gap), is uses k-modes [34, 35] for this purpose. This approach is almost identical to k-means [36, 37], but it is based on the notion of *mode* (rather than Euclidean mean), making it appropriate for clustering categorical data. Indeed, the Euclidean mean of three nucleotides has little meaning in this context, and may not even be well-defined, *e.g.*, in cases where the "distance" from A to G is different than from G to A. A similar observation was made in the context cancer mutation profiles [38], in the form of absence/presence information. Treating these as categories, in using k-modes (rather than as 0's and 1's, in using k-means) resulted in a clustering approach [39] that, when used as a preprocessing step, allowed cancer phylogeny building methods to attain a higher accuracy [40], and in some cases with much lower runtimes [41].

The *mode* q of a cluster C of sequences is another "sequence" (on A, C, G, T, –) which minimizes

$$D(C, q) = \sum_{s \in C} d(s, q) \tag{2.1}$$

where d is some distance (*e.g.*, Hamming) between the sequences we are considering. Note that q is not necessarily an element of C. Aside from finding finding modes instead of Euclidean means, the k-modes algorithm operates similarly to k-means, following the same iteration:

Algorithm uses k-modes with the following six combinations of different settings. First, cluster centers (1.) are initialized by:

Algorithm 1 k-modes clustering

Input: Viral sequences from a set of hosts.
Output: Transmission clusters.
1: Initialize cluster centers (or centroids);
2: Assign each sequence to the closest center based on distance d;
3: For each cluster resulting from this assignment, find its (new) center (Eq. 2.1); and
4: Return to step 2. until convergence (clusters do not change between 2. and 3.).

(a) choosing k random sequences from the dataset;

(b) choosing k centers that are maximally pairwise distant from each other; or

(c) using the centers (the subtypes) that were found by CliqueSNV.

Then, the distance d that is used is either the (i) Hamming distance, or (ii) TN-93 distance [42].

Cluster entropy

In the proposed approach, various clusterings of the SARS-CoV-2 data without a ground truth are compared. Thus, an *internal* evaluation criteria should be considered. Most of the commonly used criteria require some notion of a *distance* (or dissimilarity measure) between the objects being clustered. For example, criteria such as the Calinski-Harabasz Index [43] or the Gap Statistic [44] rely on the Euclidean distance, while the Davies-Bouldin Index [45] or the Silhouette Coefficient [46] require this distance to be a *metric*. In the setting of viral sequences, with the categories A, C, G, T and also the *gap* (–), it is unsure even what the distance between two categories (*e.g.*, A to G) would be, let alone whether this distance is Euclidean, or even a metric.

The cluster *entropy* [47], a criterion that was shown to generalize any distance-based criterion, does not require a distance at all. This is ideal in this context, since it does not make any assumptions about the relationships between the categories A, C, G, T, –. Indeed, since the information about such relationships is so lacking, forcing an arbitrary set of assumptions in using a distance-based criterion may only bias the resulting analysis.

Moreover, cluster entropy very naturally captures out setting: that the population of viral sequences comes from a number of subtypes. Indeed, cluster entropy can be formally derived using a likelihood principle based on Bernoulli mixture models. In mixture models, the observed data are thought of as coming from a number of different latent classes. In [47], the authors prove that minimizing cluster entropy is equivalent to maximizing the likelihood that set of objects are generated from a set of (k) classes. This is very akin to the setting here: indeed the set of objects are viral sequences, and they come from a set of k subtypes.

This relates closely to the widely-used notion of *sequence logo* [48]: a graphical representation of a set of aligned sequences which conveys at each position both the relative frequency of each base (or residue), and the amount of information (*i.e.*, how low is the entropy) in bits. So indeed, a clustering of viral sequences of low entropy gives rise to a confident set of sequence logos (in terms of information), and can hence shed light on the possible biological function of viral subtype that each such logo (or related motif) represents.

Formally, a set S of *aligned* sequences over a set X of columns is considered. A given column is then also a (vertical) "sequence" on the categories A, C, G, T, –. Let = {A, C, G, T}, the four nucleotides, not counting the gap (–) character. Using the notation of [47], the entropy $_x(C)$ of a set C of rows (a cluster of sequences) in this column x is then

$$_x(C) = -\sum_{s \in C} \sum_{a \in} p_x(s = a) \log p_x(s = a) \tag{2.2}$$

Note that $p_x(s = a)$ — the probability that a sequence $s \in C$ has nucleotide a in column x — essentially amounts to the relative *frequency* of nucleotide $a \in$ in C in this column x. The entropy $_X(C)$ of set C of rows in a set X of columns is then

$$_X(C) = \sum_{x \in X} (x) \tag{2.3}$$

that is, sums of entropies of the columns are computed. Since the set of columns will always the set of SNV sites of our sequences, (C) will be used for the entropy of this set of rows from hereon in. This way, (C) is understood to be the entropy of a set (a cluster) of sequences.

The *expected* entropy [47] of a clustering $= C_1, \ldots, C_k$ of sequences is then

$$H() = \frac{1}{n} \sum_{i=1}^{k} n_i(C_i) \tag{2.4}$$

where $n_i = |C_i|$, the number of elements in cluster C_i, and n is the total number of sequences. For completeness, the total entropy of a clustering is simply the sum

$$T() = \sum_{i=1}^{k} (C_i) \tag{2.5}$$

of the individual entropies of each cluster (not weighted by n_i).

Fitness

Here we propose a novel notion of the *fitness* of a cluster, based on how its size (number of sequences it contains) changes over a series of time steps. For a given set of clusters $C_1, \ldots C_k$, $X_i(t)$ denotes the size of cluster C_i at a particular time point t. The fitness coefficient is calculated using X_i by first computing

$$v_i(t) = \frac{X_i(t)}{\sum_{i=1}^{k} X_i(t)} \tag{2.6}$$

$$u_i(t) = \frac{v_i(t)}{\sum_{i=1}^{k} v_i(t)} \tag{2.7}$$

which are the the frequency and normalized frequency respectively, of cluster C_i at time point t. The *fitness function* g_i, for each cluster C_i is then

$$g_i(t) = \frac{\dot{u_i}(t)}{u_i(t)} + \frac{\dot{X}_i(t)}{X_i(t)} \tag{2.8}$$

Using cubic splines, $u_i(t)$ and $X_i(t)$ are interpolated over the time period and the derivatives $\dot{u_i}(t)$ and $\dot{X}_i(t)$ are calculated. The *fitness coefficient* r_i, which is the average fitness over

the time period T (composed of the time points t) for cluster C_i is then

$$r_i = \frac{1}{T} \int_1^T g_i(t)dt \tag{2.9}$$

In order to reduce sampling error, we use the Poisson distribution to draw random samples. For each cluster at each time step, a sufficiently large number of random samples are drawn from the Poisson distribution on $X_i(t)$ as the expectation of the interval. Then $X_i(t)$ is replaced by the mean value of these random samples. This is repeated a sufficiently large number of times (*e.g.*, 100) to calculate a set of Poisson-distributed sizes. The fitness coefficient calculation is then applied on each separately and a (*e.g.*, 95%) confidence interval of this fitness coefficient is obtained.

2.2.2 k-mers EMD

Proposed algorithm is based on finding the distance between populations using *Earth Movers' Distance (EMD)* between distributions of k-mers in NGS data. The general pipeline of the algorithm (see Figure 1) includes obtaining k-mer distributions from NGS reads for corresponding hosts and computing EMD between them. As a result, we obtain mean of hosts A and B $Mean(A, B)$ and EMD $EMD(A, B)$ between them. We first describe how we find distances between k-mers and then describe how we find distance between samples.

Finding distances between k-mers in the De Bruijn graph

k-mer refers to a substring of length k. In our work, we use *De Bruijn graph* to calculate distance between k-mers. De Bruijn graph is the graph, that is constructed so that vertices represent every string over a finite alphabet of length l, and edges are added between vertices that have overlap of $l - 1$.

Once De Bruijn graph is constructed, distance between k-mers can be calculated as a length of shortest path between corresponding vertices using *breadth-first search* algorithm. In our algorithms, obtained graph is converted to undirected before shortest path computation.

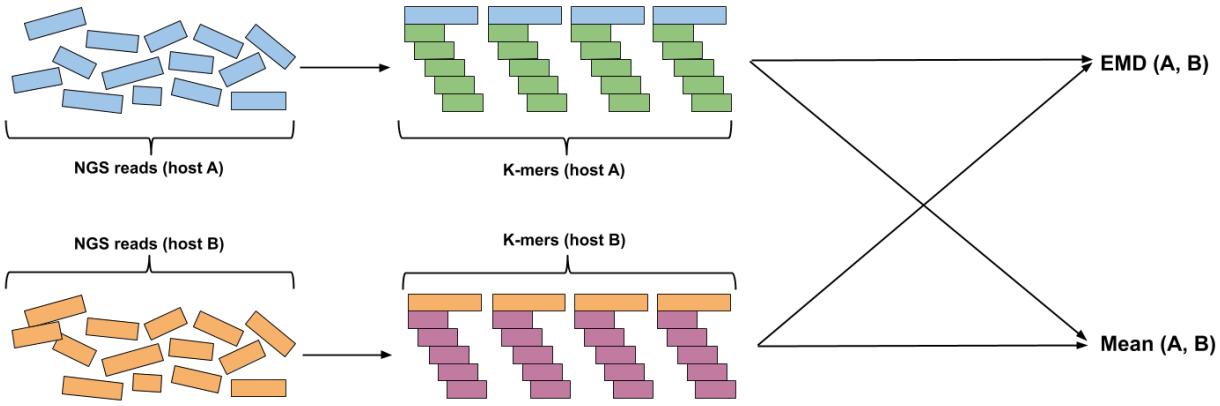

Figure (2.1) Algorithm pipeline. k-mer distributions for hosts, that need to be compared, are obtained from NGS reads. Then, EMD is computed and mean is obtained using k-mer distributions.

Finding EMD between viral samples

Viral populations can be compared by comparing the corresponding k-mer distributions using EMD. First, k-mer distributions are obtained for each sample, so that they contain all k-mers and normalized frequencies.

EMD is a method, that allows to evaluate dissimilarity between two multi-dimensional distributions in some feature space where a distance measure between single features (*ground distance*) is given [49]. Distributions can be represented as *signatures* - sets of clusters, so that each cluster is represented by its mean and by the fraction of distribution that belongs to that cluster. Computation of EMD is based on solving the *transportation problem*, which can be formulated as following: for several suppliers, each with a given amount of goods, several consumers, each with limited capacity, and a cost of transporting a single unit of goods between each supplier-consumer pair, find a least-expensive flow of goods from the suppliers to the consumers that satisfies the consumers' demand. EMD is calculated as the following $EMD(P,Q) = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}d_{ij}$ where f_{ij} is the minimum-cost flow between supplier i and consumer j, and d_{ij} is the distance between i and j. It should also be noted that EMD is usually normalized by the total flow, but we perform normalization of frequencies in k-mer distributions before EMD computation, which results in total flow always being equal to 1.

Algorithm 2 k-mers EMD

Input: Sets of sequencing reads from hosts A and B (S_A and S_B).
Output: k-EMD distance between A and B.

1: Produce k-mers from S_A and S_B:

$$KM_A \leftarrow \text{k-mer distribution from} S_A$$

$$KM_B \leftarrow \text{k-mer distribution from} S_B$$

2: Initialze distance matrix $D(A, B)$: for any pair of k-mers $x \in KM_A$ and $y \in KM_B$, find $dist(A, B)$ in De Bruijn graph;
3: Compute $EMD(KM_A, KM_B, D(A, B))$.

Example of EMD computation

Constructing of the De Bruijn graph between two sequences $CGATTCTAAGT$ and $CGATTGTAAGT$ is shown on Figure 2. Once original graph is obtained, directions are removed and pairwise distances are computed for all k-mers. Figure 3 describes an example of k-EMD distance computation. After k-mer distributions are generated for input sequences, EMD is computed as the work $\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}d_{ij}$, where f_{ij} is the flow between histogram(k-mer distribution) elements i and j and d_{ij} is the corresponding distance between k-mers, which is obtained from De Bruijn graph (Figure 2). This way, $EMD = 0.88$.

Mean k-mer distribution

Representing samples as k-mer distributions allows to estimate the center from a group of samples by introducing a mean host. We use the **maximum mean** k-mer distribution, which is obtained by finding the maximum observed frequency for each k-mer k_i $f_i^{max} = \max\limits_{1\leq i\leq n} f_i$ and normalization $f_i' = \frac{f_i^{max}}{\sum\limits_{1\leq i\leq n} f_i^{max}}$

Identification of relatedness

Algorithm is trained on all given outbreaks, so that minimal EMD between 2 unrelated hosts (relatedness threshold t is obtained). To identify whether 2 hosts A and B are

related, we compute EMD between them $EMD(A, B)$ and predict that they are related if $EMD(A, B) < t$, and unrelated otherwise.

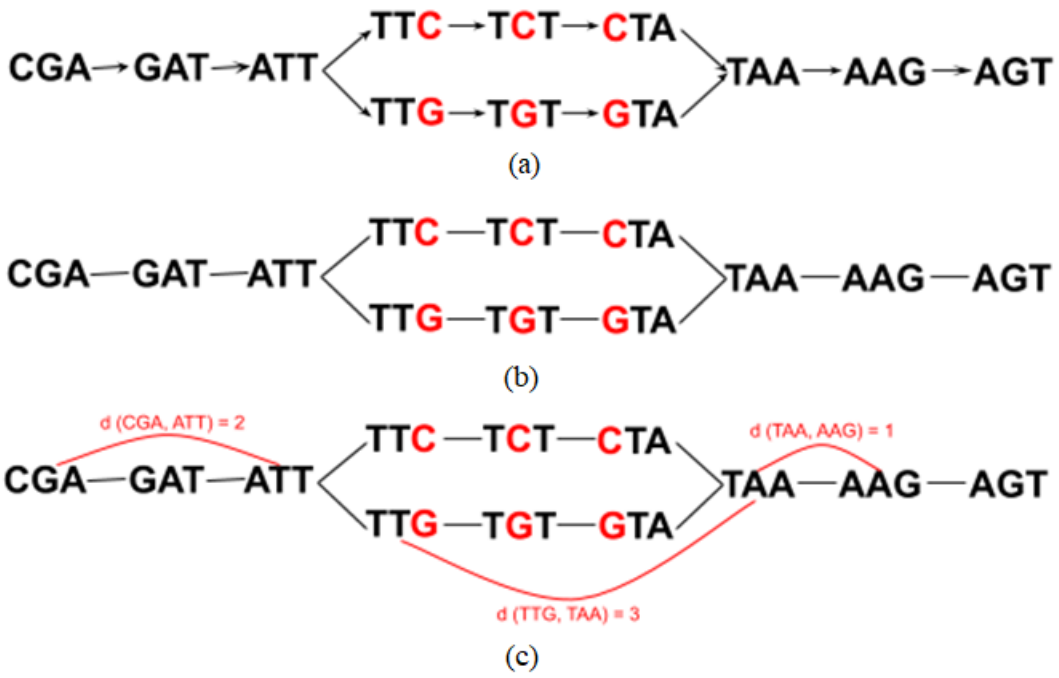

Figure (2.2) De Bruijn graph for 3-mers, obtained from sequences $CGATTCTAAGT$ and $CGATTGTAAGT$. Once original graph is obtained (a), directions (b) are removed and pairwise distances are computed for all k-mers (c).

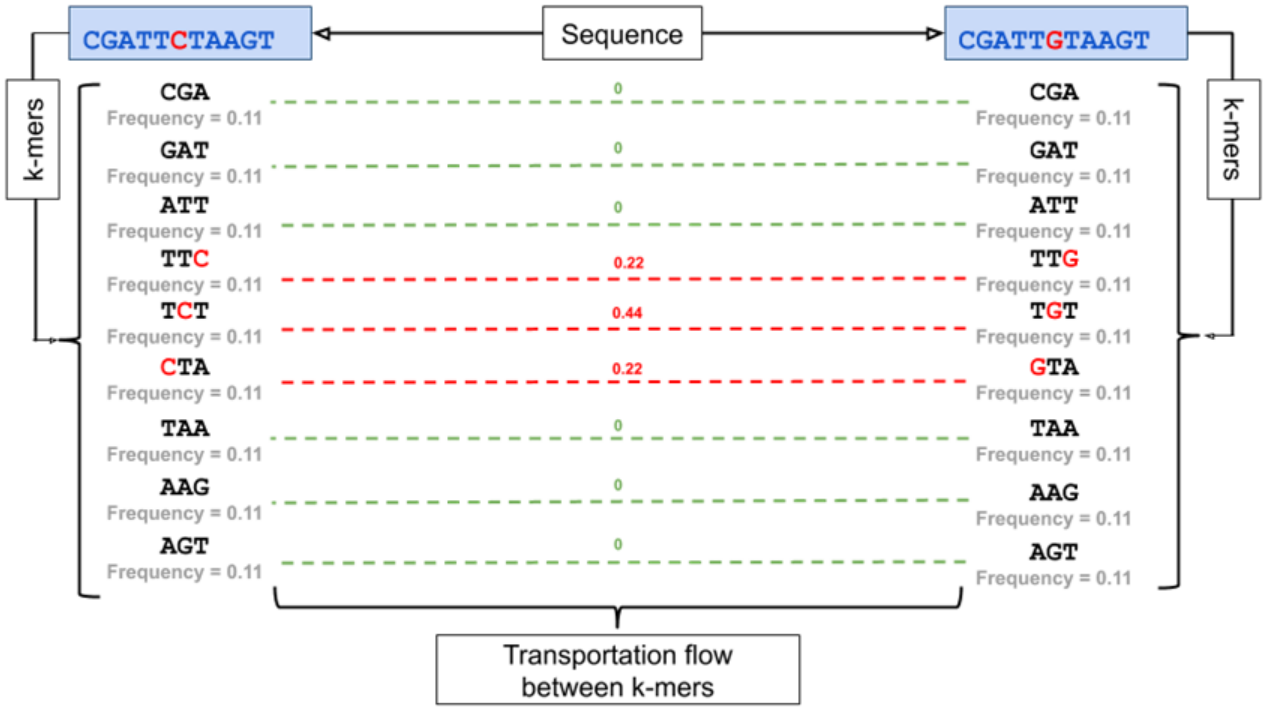

Figure (2.3) Finding EMD distance between k-mers of sequences $CGATTCTAAG$ and $CGATTGTAAGT$. The k-mer distributions are on the left and right sides. Dashed lines represent transportation flow between k-mers; corresponding flow values are shown in green. Red values on top of the lines represent distance between corresponding k-mers in the De Bruijn graph.

Identification of transmission direction between hosts

To infer transmission direction between a pair of samples X and Y, we first compute a mean host $Mean(A, B)$.

Once $Mean(A, B)$ is obtained, we calculate EMD between mean host and hosts A and B $EMD(Mean(A, B), A)$ and $EMD(Mean(A, B), B)$. Host, that is closer to the maximum mean is assumed to be the transmission source, so that if $EMD(Mean(A, B), A) < EMD(Mean(A, B), B)$, we predict that the transmission happened from A to B (Figure 4).

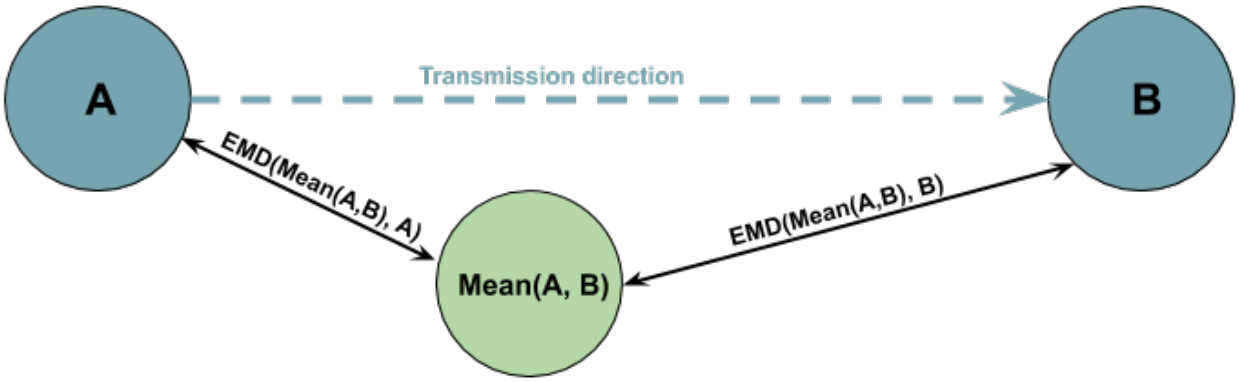

Figure (2.4) Inference of transmission between hosts A and B. First, mean host $Mean(A, B)$ is introduced. Then EMD is computed between $Mean(A, B)$ and hosts A and B. Finally, $EMD(Mean(A, B), A)$ is compared with $EMD(Mean(A, B), B)$. If $EMD(Mean(A, B), A) < EMD(Mean(A, B), B)$, then transmission direction is predicted as the one that happened from A to B.

Identification of transmission clusters

To test hierarchical clustering, *single-linkage* algorithm was used. This method evaluates the similarity of two clusters based on their most similar members [**?**] and groups clusters in bottom-up order until certain termination condition is satisfied. In our algorithm, we use a distance criteria, so clusters are merged until distance between them exceeds a predefined distance threshold, which represents EMD between two closest unrelated samples in the dataset. This way, we obtain a partition, where some of the related hosts remain in different clusters. At this point, we proceed to the second stage of the algorithm, that allows to improve the clustering quality by merging the clusters, that contain related hosts by performing the following steps:

1. For each cluster, obtained from hierarchical clustering, compute center as the mean of all hosts within the cluster;

2. For each center, obtained at the previous step:

- Find distances to the furthest in-cluster host and closest host, that belongs to the different cluster;
- If for cluster A there exists an 'overlap' (there is a host from cluster B, that is closer to the center than the furthest host, belonging to the same cluster (A)), merge A and B

Example of the algorithm is demonstrated in Figure 5. a) shows output of threshold-based hierarchical clustering, where circles represent hosts, that are connected with an edge if distance between them doesn't exceed a threshold. There are 2 clusters that belong to the same outbreak. b) shows how clusters are merged based on circle overlap. For each cluster, mean host of all hosts within the cluster is calculated (shown in the center). Circles with dashed borders have centers in respective mean hosts; their radiuses are calculated as distances between mean hosts and furthest in-cluster hosts. In the example, Mean 1 is closer to host A that to the furthest host from the same (left) cluster. This way, according to our algorithm, intersecting clusters collapse.

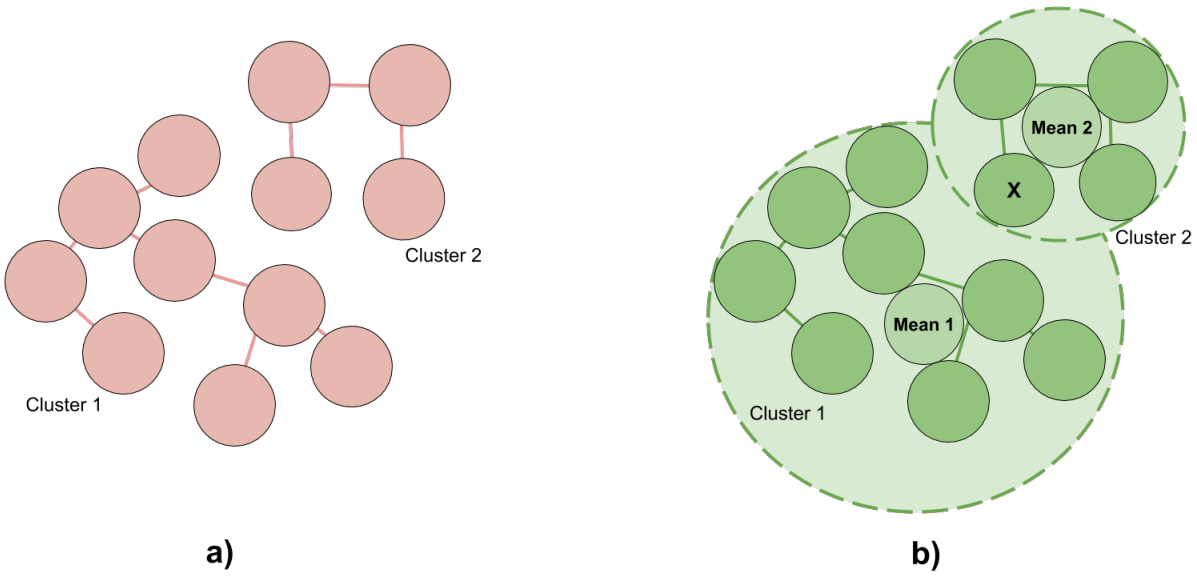

Figure (2.5) Example of overlap-based cluster merging. a) Output of threshold-based hierarchical clustering, where circles represent hosts (k-mer distributions), that are connected with an edge if distance (EMD) between them doesn't exceed a threshold (so that no unrelated hosts are connected). There are 2 clusters that belong to the same outbreak, which means that some related hosts are treated as unrelated. b) For each cluster, circles were build, so that mean hosts reside in the center of the circle, and radius is defined as the distance between mean host and furthest host in an outbreak. Circle of cluster 1 intersects with cluster 2 since host X is closer to Mean 1 than furthest host in cluster 1. Therefore, clusters 1 and 2 are merged.

Deciding whether source is present in a set of hosts

To decide whether source is present in a set of sequenced hosts S, the following algorithm is applied:

1. Calculate mean $Mean(S)$ for all hosts within an outbreak;
2. For every host H, calculate EMD between H and mean $Mean\ EMD(Mean(S), H)$;
3. If there exists a host, for which $EMD(Mean(S), H) < t$, source is present.

To obtain threshold t, we train the algorithm on all outbreaks with known sources. For every such outbreak, we first calculate the mean host $Mean(S)$ and distances between mean and every host H in the outbreak $EMD(Mean(S), H)$, find the smallest distance and normalize it by the median distance from mean to host in an outbreak. After this, we repeat the procedure for the same outbreak, but discard the source. We define t as the minimal $EMD(Mean(S), H)$ for an outbreak without source, which maximizes accuracy, so that outbreaks, where source is present, have $EMD(Mean(S), H) < t$.

Source identification

To identify sources, a maximum mean host for an outbreak $Mean$ is computed, and EMD is calculated between every host and $Mean$. Host with minimum $EMD(H, Mean)$ is assumed to be the source.

2.2.3 VOICE

$VOICE$ [50] is a non-deterministic algorithm for analysis of NGS data from viral outbreaks. The algorithm uses Markov process to simulate the process of viral population evolution from source to recipient.

Identification of relatedness and transmission direction Given two hosts A and B, VOICE infers times t_{AB} and t_{BA}, that represent evolution time for a corresponding direction of infection. Based on obtained times, algorithm decides whether viral populations from hosts are related and infers transmission direction.

Data normalization Due to biases, that can be introduced at sampling and sequencing steps, sizes of observed viral populations may vary significantly, which, in turn, may affect simulation time. To compensate for this, $VOICE$ performs normalization step, where each viral population is clustered, and each cluster is replaced with consensus of its members. During subsampling normalization, q sequences are randomly chosen from each population, and procedure is repeated r times.

Identification of clusters and sources of outbreaks To identify outbreak clusters, $VOICE$ produces a weighted directed relatedness graph, where $G = (V, A, w)$ with $V = \mathcal{P}$. Viral populations P_A and P_B are connected with an edge if value $\min\{t_{AB}, t_{BA}\}$ is less than a threshold, so that A and B are considered to be related. Transmission clusters are computed as weakly connected components of G. Outbreak sources are inferred by building a Shortest Paths Tree (SPT) for every vertex in the corresponding cluster. Vertex with SPT of minimal weight is assumed to be the source.

2.2.4 Results

k-mers EMD and VOICE were validated on a publicly available dataset obtained from an epidemiological study of HCV outbreaks [13]. SARS-CoV-2 clustering algorithm was validated on two publicly avalilable datasets, obtained from GISAID [31] and EMBL-EBI [32, 51] databases.

HCV data set

The data consists of 368 sequenced hosts where 175 of them belong to 34 annotated outbreaks. Among these annotated outbreaks, 11 have a known main spreader (Table 1). All outbreaks contain from 2 to 33 hosts. Every host is represented as an HCV intra-host population, obtained with end-point limiting-dilution (EPLD). All viral sequences represent a fragment of E1/E2 genomic region of length 264bp. Data samples annotation consists of host and outbreak id along with abundance for every sequence. This way, we were able to interpret obtained experimental results.

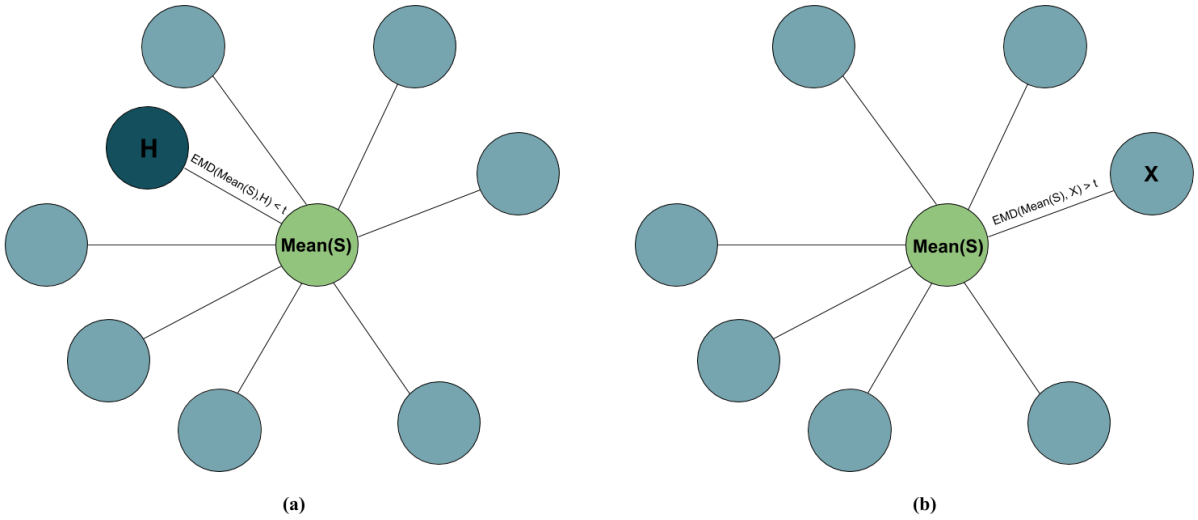

Figure (2.6) Deciding whether source is present in a given set of hosts. Here, every circle represents a host, belonging to an outbreak, and green circle represents mean. Edges represent distances between mean hosts and hosts in an outbreak. If there is a host, that is close to mean (so that the distance is smaller than a threshold, case (a)), we conclude, that source is present in an outbreak. Otherwise, analyzed set of hosts doesn't include the outbreak source (case (b)).

We simulated MiSeq reads from known haplotypes by SimSeq [52] and created mixtures using abundances from original data to test k-mers EMD method.

SARS-CoV-2 data sets

The first data set consists of sequences submitted to the GISAID [31] database from December 2019 to November 2020. This data set contains sequences from all over the world. The second data set consists of sequences submitted to the EMBL-EBI [32, 51] database from the beginning of October 2020 to the middle of December 2020. For both data sets, we align the sequences and trim the first and last 50bp of the aligned sequences.

k-EMD and VOICE validation

Identification of relatedness Viral populations from two samples are genetically related if they belong to the same outbreak and unrelated, otherwise. The genetic relatedness is validated on the union of both collections containing all outbreaks and unrelated samples. There are 67528 host pairs (obtained from all 368 hosts). Among these pairs, 1007 represent related cases (so that both hosts in pair belong to the same annotated outbreak). We used EMD as predictor for relatedness. We measured the sensitivity of our method as following. First we determining the EMD value for all unrelated pairs, the minimum value we have chosen as a threshold which prohibits false-positive relatedness detection, the pairs which have EMD below the threshold are considered as related. Precision of our algorithm is 100%. We calculated the recall as a proportion of correctly predicted related pairs among all known related pairs. Results are described in Table 2. Relatedness ROC is shown on Figure 7.

Identification of transmission direction between hosts Performance of algorithm when identifying transmission direction was calculated as a ratio of pairs of hosts with correctly predicted directions to all host pairs, where direction is known. Results are shown in Table 2.

Identification of transmission clusters Precision for our algorithm is equal to 100%, since we don't merge hosts from different outbreaks. Similarities between true and estimated partitions were evaluated using an editing metric [53]. Given metric is defined as the minimum number of elementary operations, required to transform one partition into another, such as joining or partition of clusters [53]. Clustering recall was calculated similarly to [50], so that editing distance E was normalized by dividing it by the number of elementary operations N, required to transform trivial partition into singleton sets into true partition, which is equal to $n - k$, where n is the number of samples and k is the number of true clusters [50]:

$$Recall = \frac{E}{n-k} \times 100\%$$

Deciding whether outbreak source is present Source presence recall was calculated as the proportion of outbreaks with present source, that were correctly identified as such; precision - as the proportion of correctly identified outbreaks, where source is not present. Finally, specificity was calculated as the total number of outbreaks with present source, divided by the sum of total number of outbreaks with present source and the number of outbreaks, that were incorrectly identified to have a source present. For our algorithm, precision = 90%, specificity = 80%, and recall = 85%. ROC curve for source presence detection is shown on Figure 8.

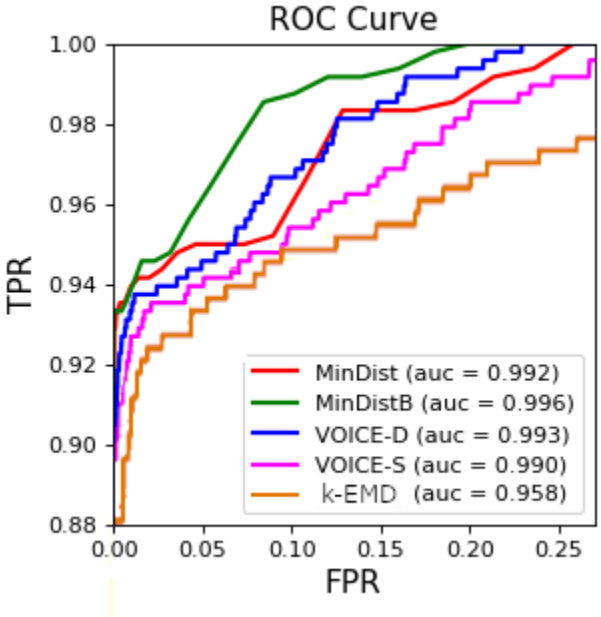

Figure (2.7) ROC curve for prediction of source presence. AUROC = 0.8

Identification of outbreak sources Source identification accuracy is calculated as the percentage of outbreaks with correctly predicted sources for outbreaks with known sources. ROC curve for source presence detection is shown on Figure 9.

Outbreak	AA	AC	AI	AJ	AQ	AW	BA	BB	BC	BJ	NH
# samples	3	4	15	3	9	19	6	7	2	4	33

Table (2.1) Outbreaks with known sources

Method	**k-EMD**	**MinDist**	**MinDisB**	**ReD**	**VOICE-D**	**VOICE-S**
Relatedness sensitivity, %	80.4 (90)	90	92.9	55.3	85.2	86.8
Clustering sensitivity, %	100 (100)	100	100	96.3	98.2	98.2
Direction accuracy, %	88.7 (90.4)	N/A	N/A	87.1	83.9	87.1
Source accuracy, %	80 (81.8)	50	40	90	80	90

SARS-CoV-2 clustering validation

GISAID dataset Using our technique involving CliqueSNV, GISAID [31] dataset was clustered to identify at most 66 subtypes, which vary in proportion between December 2019 and November 2020. Indeed, a k of 66 was needed in order for the minimum cluster frequency to be at least 1% of the population in this case. Relative distributions of these different subtypes is reported in Fig. 2.8 and Fig. 2.9, in a similar way to that of Fig. 3 of [54].

Table 2.2 gives an assessment of the various clusterings computed, in terms of both the expected entropy (Eq. 2.4) and total entropy (Eq.2.5). While any form of clustering achieves a better expected (and total) entropy than not clustering at all, introduced CliqueSNV-based approach tends to outperform all other forms of clustering using either Hamming or TN-93 distance. Finally, by filling gaps in sequences based on the closest cluster center, an even lower expected (and total) entropy is achieved. This illustrates the appropriateness of this cluster-based approach for filling gaps: indeed the entropy of the dataset without clustering remained high after filling gaps (based on the consensus for the entire dataset), for example. Finally, Table 2.3 reports runtimes of the various stages of this analysis, and Table 2.4 compares runtimes of CliqueSNV and k-modes clustering. Given the latter table, it should be noted that CliqueSNV-based method had a slightly lower runtime than k-modes,

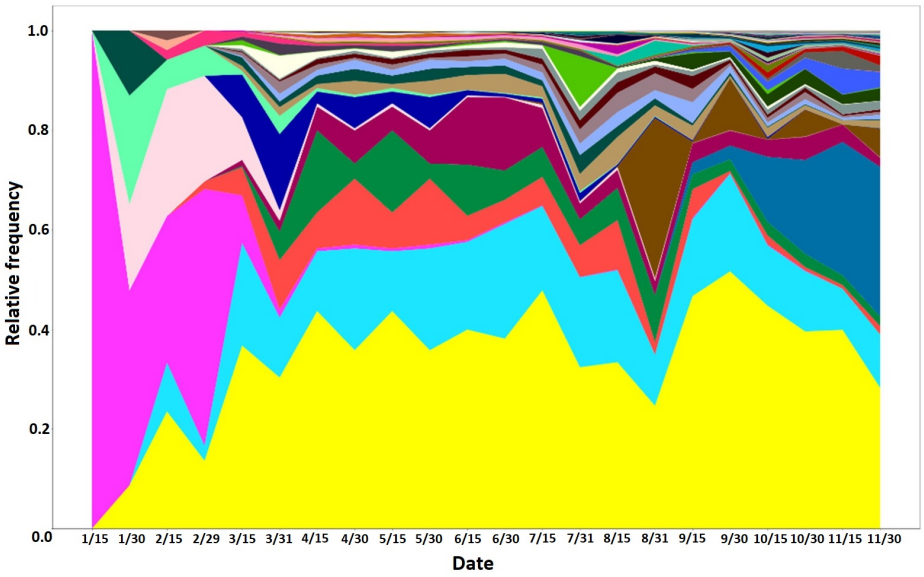

Figure (2.8) Subtype distribution (GISAID dataset, 15-day window, relative count)

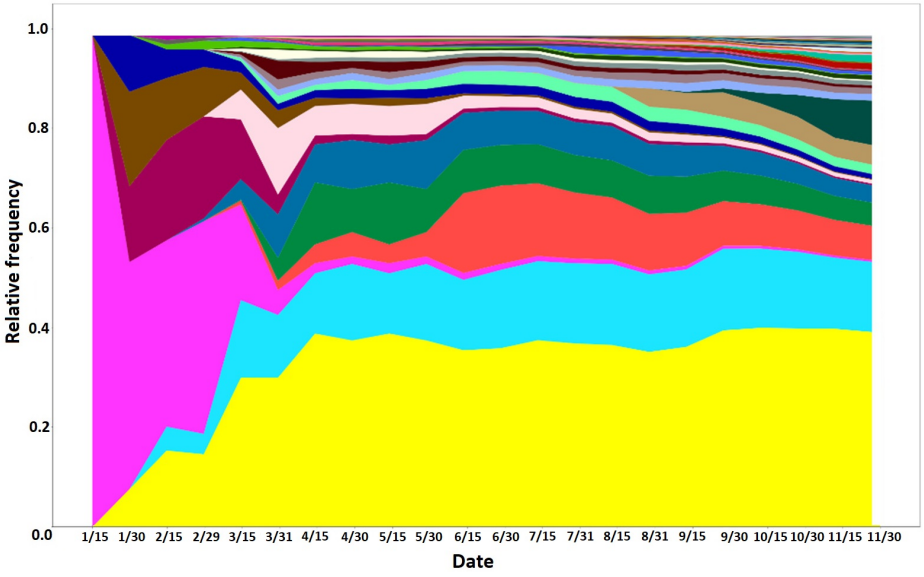

Figure (2.9) Subtype distribution (GISAID dataset, cumulative, relative count).

k-modes setting (initialization, distance)	without gap filling		with gap filling	
	expected entropy	total entropy	expected entropy	total entropy
without clustering	9536.89	9536.89	8417.89	8417.89
random centers, Hamming	123.00	3170.60	109.21	2474.30
random centers, TN-93	127.32	4401.18	111.05	3470.03
pairwise distant, Hamming	422.65	4651.23	294.98	3629.47
pairwise distant, TN-93	273.34	3500.14	256.44	3007.07
CliqueSNV, Hamming	110.58	2585.29	90.42	2308.95
CliqueSNV, TN-93	121.87	2379.46	100.85	2117.40

Table (2.2) The expected entropy (Eq. 2.4) and total entropy (Eq. 2.5) of the GISAID sequences without clustering (*i.e.*, considered as a single cluster containing all sequences), and when clustering using each of the six combinations of settings mentioned in Sec. 2.2.1, both without filling gaps and with gap filling.

algorithm stage	time in seconds
CliqueSNV (inferring subtypes)	2405.08
CliqueSNV (finding closest subtypes)	2324.34
gap filling	2740.32
entropy computation	1254.22
Total	8723.96

Table (2.3) Runtimes of the different stages of the algorithm for the GISAID dataset, which contains 199240 sequences. All stages were executed on a PC with an Intel(R) Xeon(R) CPU X5550 2.67GHz x2 with 8 cores per CPU, DIMM DDR3 1333 MHz RAM 4Gb x12, and running the CentOS 6.4 operating system.

despite it performing best overall.

EMBL-EBI dataset Data from the EMBL-EBI database was clustered to identify 15 subtypes which vary in proportion between the beginning of October 2020 and the middle of December 2020. Since the data here are over a shorter time span (*i.e.*, are smaller), and are more uniform, a k of 15 was sufficient for the minimum cluster frequency to be at least 1% of the population in this case. The relative distributions of these different subtypes is reported in Fig. 2.10 using a weekly moving average, since a weekly oscillation in SARS-CoV-2 data has been noted in [55]. One will notice, in Fig. 2.10, the sharp increase of the relative proportion of a certain subtype (in red) to more than a third of the population. We

clustering method	time in seconds
CliqueSNV	4729.42
k-modes	4922.44

Table (2.4) Runtimes of CliqueSNV and k-modes clustering using random centers and Hamming distance, for the GISAID dataset, which contains 199240 sequences. Both methods were executed on a PC with an Intel(R) Xeon(R) CPU X5550 2.67GHz x2 with 8 cores per CPU, DIMM DDR3 1333 MHz RAM 4Gb x12, and running the CentOS 6.4 operating system.

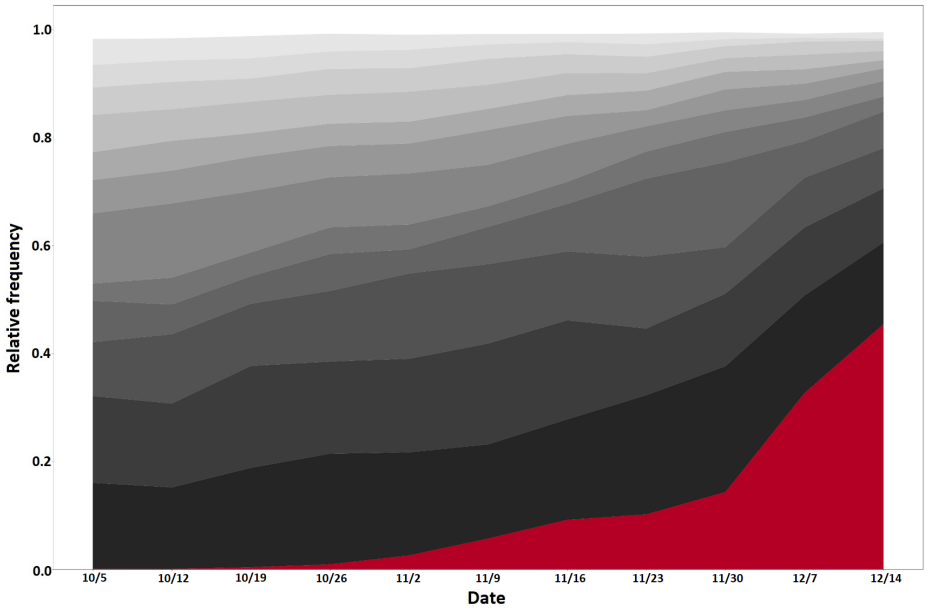

Figure (2.10) Subtype distribution (UK dataset, weekly window, relative count), produced by CliqueSNV. Red subtype contributes to 99.86 % of the sequences that correspond to B.1.1.7 lineage.

confirm from metadata, that this indeed corresponds to the B.1.1.7 variant that was first identified in studies such as [56]. Fig. 2.11 gives the number of sequences from Fig. 2.10 which belong to this B.1.1.7 lineage by mid December 2020, which shows how accurately our approach has detected this subtype. This illustrates the ability of our clustering to identify subtypes which are known in the literature. Interestingly enough, the study of [56] is based on an approach of building a phylogenetic tree — this demonstrates our approach, which is based on clustering sequences, as a viable alternative.

Because our method detected one subtype which tends to dominate the population in this UK data, we wanted to see if this is consistent with a cluster-based fitness coefficient,

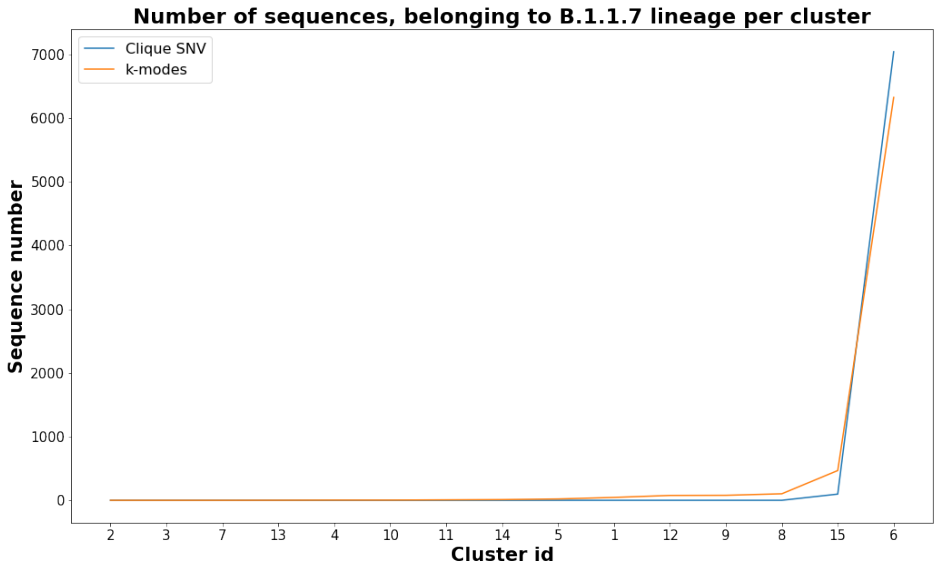

Figure (2.11) Number of sequences belonging to the B.1.1.7 lineage per cluster for CliqueSNV and k-modes clustering. For CliqueSNV, all sequences are contained in 2 clusters (out of a total of 15): 7044 in cluster 6 and 97 in cluster 15. The k-modes clustering, on the other hand, reported that B.1.1.7 sequences are contained in 13 out of 15 clusters, with counts ranging from 1 to 6327 sequences per cluster. Expected entropy for gap-filled CliqueSNV clustering is 75.73, and 94.16 for k-modes. (Total entropy is 986.48 for CliqueSNV and 2074.12 for k-modes.)

Cluster	Interval lower bound	Interval Upper bound
1	-0.012703573	-0.012649641
2	0.086896596	0.087769609
3	0.354406762	0.368566931
4	0.048125146	0.048717893
5	-0.020624053	-0.020552015
6	1.342718732	1.504481237
7	0.016457622	0.016783626
8	0.086911634	0.088104572
9	-0.421159711	-0.406373723
10	0.030043564	0.030574889
11	-0.019298934	-0.018901415
12	0.018863127	0.019330401
13	0.025013817	0.025309291
14	0.284365966	0.323771427
15	-0.002840931	-0.002727241

Table (2.5) The 95% confidence interval of the fitness coefficient of each of the 15 clusters of the UK data obtained using CliqueSNV centers and Hamming distance.

i.e., that of Sec. 2.2.1. In this case, we have $k = 15$ clusters, and we chose our time points t to be intervals of one week over the period of the beginning of October to the middle of December. The size $X_i(t)$ of each cluster C_i in every week t was obtained, and each fitness coefficient r_i was computed accordingly (Eq. 2.9). In order to reduce sampling error, we drew 2000 random samples from the Poisson distribution on $X_i(t)$ according to Sec. 2.2.1. We repeated this 100 times, and we report the 95% confidence interval of the resulting coefficients of the clusters obtained with CliqueSNV centers using Hamming distance in Table 2.5, and using TN-93 distance in Table 2.6. We note that similar results are obtained with either distance. In either case, these coefficients confirm that the cluster with ID 6, identified in Fig. 2.11 to corresponding to this this B.1.1.7 variant, is by far the most fit. This highlights the ability of our clustering-based approach for detecting, based purely on sequence content, novel subtypes which have the potential of becoming dominant in the population.

Cluster	Interval lower bound	Interval Upper bound
1	-0.013470426	-0.013402308
2	0.788873024	0.794633526
3	0.350509061	0.363602609
4	0.066572874	0.067456665
5	-0.017051993	-0.016967840
6	1.389827068	1.509878915
7	0.015262998	0.015514851
8	0.086021497	0.086922187
9	-0.380903006	-0.372241770
10	0.032476523	0.033215182
11	-0.021344261	-0.021102968
12	0.000508028	0.000937268
13	0.042481718	0.043032257
14	0.353102235	0.390211682
15	-0.032297606	-0.032036793

Table (2.6) The 95% confidence interval of the fitness coefficient of each of the 15 clusters of the UK data obtained using CliqueSNV centers and TN-93 distance.

PART 3

VIRAL QUASISPECIES ASSEMBLY

3.1 Introduction

RNA viruses, such as IAV, HIV and HCV, are known for their high mutation rates and exist in infected hosts as highly heterogeneous populations of closely related genomic variants called quasispecies [57–64].

Structure and composition of quasispecies is an important factor, that influences disease progression and epidemic spread. In particular, low-frequency variants may result in immune escape, emergence of drug resistance and an increase of virulence [23,65–70]. Therefore, accurate characterization of viral mutation profiles sampled from infected individuals is essential for viral research, therapeutics and epidemiological investigations [33].

Recent advances in NGS technologies provide new opportunities when it comes to analysis of viral populations, and allow to produce strong coverage of highly variable viral genomic regions, which is crucial for capturing of rare variants. Nonetheless, haplotype reconstruction problem remains challenging due to several reasons, such as large number of sequencing reads, unknown number of true haplotypes, and need to preserve low-frequency variants. While there exist sequencing solutions, that provide long reads, their applicability to haplotype reconstruction problem is challenged by the need to distinguish between real and artificial genetic heterogeneity produced by sequencing errors [33].

A number of computational tools for inference of viral quasispecies populations from noisy NGS data have been proposed recently. These methods include PredictHaplo [29], Savage [16], aBayesQR [30], QuasiRecomb [71], HaploClique [72], VGA [73], VirA [74, 75], SHORAH [76], ViSpA [77], QURE [78] and others [79–83]. While given algorithms showed strong performance in many applications, they still struggle when it comes to accurate and scalable reconstruction of viral haplotypes, especially when it comes to low-frequency variants

and large datasets produced by modern sequencing protocols.

While some of existing methods, such as V-phaser [84], V-phaser2 [85] and CoVaMa [86] use motations linkage for SNV calling, they don't account for sequencing errors, which makes them unable to detect mutations of frequency above sequencing error rates [87]. 2SNV algorithm [88] was the first tool to correctly detect haplotypes with a frequency below the sequencing error rate by accommodating errors in links.

Alternatively, other existing methods, such as HaploClique [72], Savage [16] rely on finding maximal cliques in a graph, where nodes represent sequencing reads. To infer haplotypes, they iteratively merge cliques, which makes them dependend on order of merging.

Instead of relying on a read graph, CliqueSNVinds maximal cliques in a graph with nodes corresponding to SNVs. This allows to drastically increase performance when compared to methods, based on read graphs.

Furthermore, the clique merging problem is formulated and solved as a combinatorial problem on the auxiliary graph of cliques of the SNV graph, thus allowing an increase of the CliqueSNV algorithm's accuracy [33].

3.2 Methods

3.2.1 Clique SNV algorithm

The pipeline of the CliqueSNV algorithm is shown in Figure 3.1 [33]. The algorithm takes aligned reads as input and outputs haplotype sequences with their frequencies. The method consists of six main steps.

Step 1 Consensus sequence is built from aligned reads and all SNVs are identified. All pairs of SNVs are tested for dependency and divided into three groups: *linked*, *forbidden*, or *unclassified*. In case there is enough reads that have two SNVs simultaneously, they are tested for dependency and independency, and algorithm classifies the SNV pair as *linked* or *forbidden*.

Step 2 Graph $G = (V, E)$ with a set of nodes V representing SNVs, and a set of edges E connecting linked SNV pairs is constructed.

Step 3 *Maximal cliques* in graph G are computed, so that each maximal clique represents groups of pairwise-linked SNVs that potentially belong to a single haplotype.

Step 4 Overlapping cliques are merged if they contain a forbidden SNV pair.

Step 5 Each read is assigned to a merge clique with which it shares the largest number of SNV; consensus haplotype from all reads assigned to a single merged clique is constructed.

Step 6 Haplotype frequencies are estimated via an expectation-maximization algorithm.

3.2.2 Validation metrics for viral population inference

Precision and recall The quality of inference is usually measured by precision and recall.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

where TP is a number of true predicted haplotypes, FP is a number of false predicted haplotypes, and FN a number of undiscovered haplotypes. Precision and recall were initially measured either by treating a predicted haplotype with a single mismatch as a FP or by introducing an acceptance threshold [29], so that s a number of mismatches is permitted in a predicted haplotype, and it can still be counted as TP.

Matching errors between populations However, precision and recall do not account for distances between true and inferred viral variants and their frequencies. For this reason, an analogous index is proposed for analysis of viral haplotype reconstruction tools [33]:

Let $T = \{(t, f_t)\}$, be the true haplotype population, where f_t is the frequency of the true haplotype t, $\sum_{t \in T} f_t = 1$. Similarly, let $P = \{(p, f_p)\}$, be the reconstructed haplotype

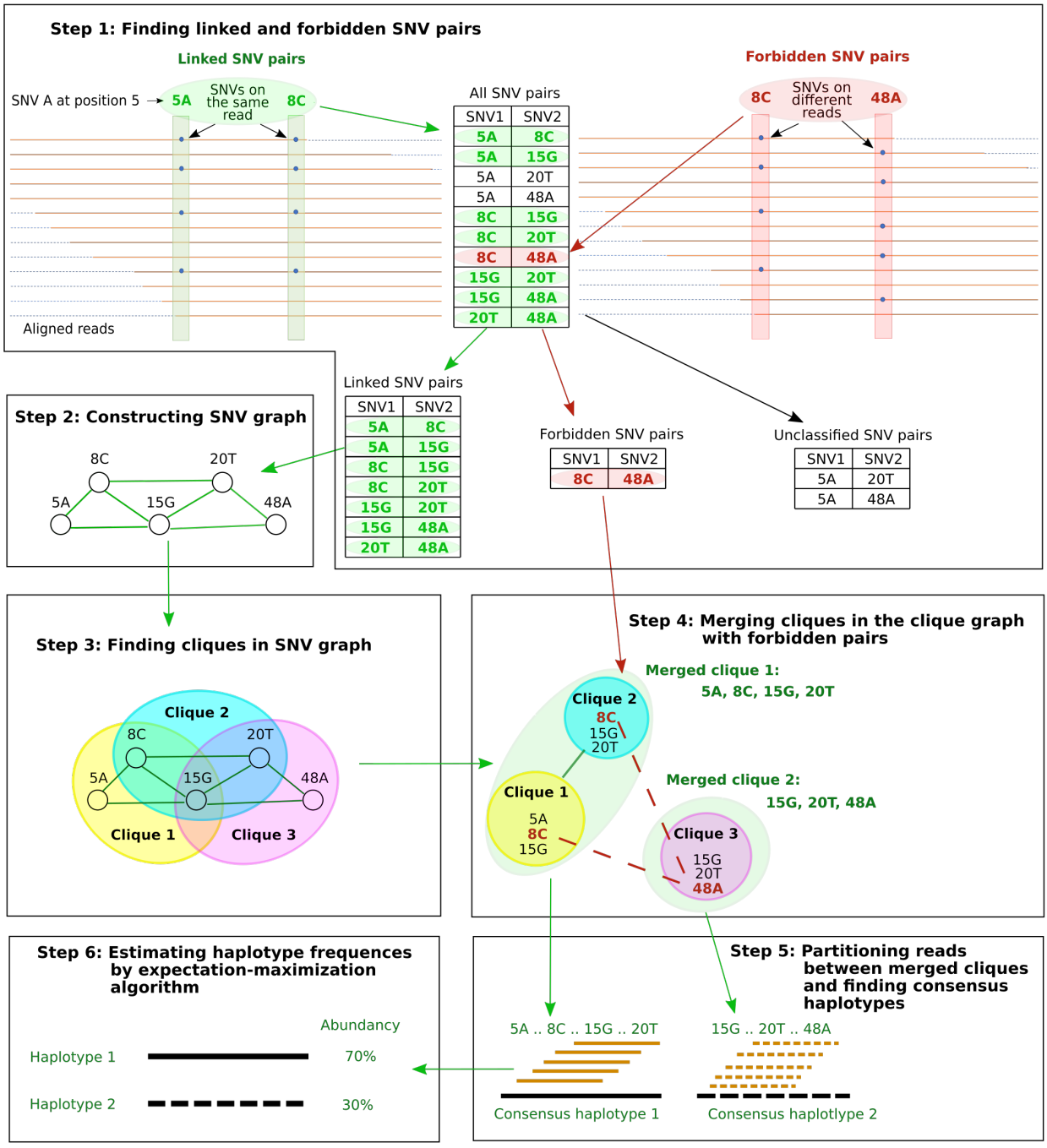

Figure (3.1) Schematic representation of the CliqueSNV algorithm, where SNV is single nucleotide variation.

population, where f_p is the frequency of the reconstructed haplotype p, $\sum_{p \in P} f_p = 1$. Let d_{pt} be the edit distance between haplotypes p and t. Thus, instead of precision, *matching*

error $E_{T\to P}$ is used to measure how well each reconstructed haplotype $p \in P$ weighted by its frequency is matched by the closest true haplotype.

$$E_{T\to P} = \sum_{p\in P} f_p \min_{t\in T} d_{pt}$$

Precision increases while $E_{T\to P}$ decreases and reaches 100% when $E_{T\to P} = 0$. Instead of recall, *matching error* is used $E_{T\leftarrow P}$ to measure how well each true haplotype $t \in T$ weighted by its frequency is matched by the closest reconstructed haplotype. [89]

$$E_{T\leftarrow P} = \sum_{t\in T} f_t \min_{p\in P} d_{pt}$$

Recall increases while $E_{T\leftarrow P}$ decreases and reaches 100% when $E_{T\leftarrow P} = 0$.

Earth mover's distance (EMD) between populations While matching errors match haplotypes of true and reconstructed populations, they do not match their frequencies. In order to simultaneously match haplotype sequences and their frequencies, fractional matching needs to be used, so that portions of a single haplotype p of population P are matched to portions of possibly several haplotypes of T and *vice versa* [33]. This way f_p is separated into f_{pt}'s each denoting portion of p matched to t such that $f_p = \sum_{t\in T} f_{pt}$, $f_{pt} \geq 0$. Symmetrically, f_t's are also separated into f_{pt}'s, i.e, $\sum_{p\in P} f_{pt} = f_t$. Finally, f_{pt}'s that minimizes minimizing the total error of matching T to P needs to be chose. This problem is known as Wasserstein metric or EMD between T and P [90, 91].

$$EMD(T,P) = \min_{f_{pt}>0} \sum_{t\in T}\sum_{p\in P} f_{pt} d_{pt}$$

$$\text{s.t.} \sum_{t\in T} f_{pt} = f_p, \text{ and} \sum_{p\in P} f_{pt} = f_t$$

EMD is efficiently computed as an instance of the transportation problem using network flows. EMD varies significantly for different benchmarks, since they have various number of

true variants and their frequency distribution, similarity between haplotypes, and sequencing parameters, such as depth, error rate, etc. Given this, the complexity of a benchmark can be measured as the EMD between the true population and a population consisting of a single consensus haplotype [92].

3.2.3 Results

CliqueSNV was tested using four real (experimental) and two simulated datasets from HIV and IAV samples (Table 3.1 [33]). Datasets contain two to ten haplotypes with frequencies 0.1 to 50%.

Name	Type	Virus	#haplotypes	Haplotype frequencies	Hamming distance
HIV9exp	experimental	HIV-1	9	0.2-50%	0.22-2.1%
HIV2exp	experimental	HIV-1	2	50-50%	1.2%
HIV5exp	experimental	HIV-1	5	20-20%	2-3.5%
IAV10exp	experimental	IAV	10	0.1-50%	0.1-1.1%
HIV7sim	simulated	HIV-1	7	14.3-14.3%	0.6-3%
IAV10sim	simulated	IAV	10	0.1-50%	0.1-1.1%

Table (3.1) Four experimental and two simulated sequencing datasets of human immunodeficiency virus type 1 (HIV-1) and influenza A virus (IAV). The datasets contain MiSeq and PacBio reads from intra-host viral populations consisting of two to ten variants each with frequencies in the range of 0.1-50%, and Hamming distances between variants in the range of 0.1-3.5%.

Experimental datasets:

1–2. *HIV-1 subtype B plasmid mixtures and MiSeq reads (HIV2exp and HIV9exp).* Nine *in silico* plasmid constructs comprising a 950-bp region of the HIV-1 polymerase *(pol)* gene were designed, synthesized and then cloned into pUCIDT-Amp (Integrated DNA Technologies, Skokie, IL). Given region at the beginning of *pol* is known to contain protease and reverse transcriptase drug-resistant mutations, and is monitored with sequence analysis for patient care. Designed plasmids contain point mutations chosen from real clinical study [93]. Plasmids were mixed in varaious ratios and then sequenced

using an Illumina MiSeq protocol. HIV2exp dataset is based on a mixture of two variants, and HIV9exp is based on nine.

3. *HIV-1 subtype B mixture and MiSeq reads (HIV5exp).* This dataset consists of Illumina MiSeq 2×250-bp reads obtained from a mixture of five HIV-1 isolates: 89.6, HXB2, JRCSF, NL43, and YU2 available at [94]. Pairwise Hamming distances of isolates are in the range from 2-3.5%(27 to 46-bp difference). HIV-1 sequence was reduced to the beginning of *pol* with length of 1.3Kb.

4. *IAV mixture and PacBio reads (IAV10exp).* Given benchmark consists of ten IAV clones, mixed at a frequency of 0.1-50%. Hamming distances between clones range from 0.1-1.1% [88].

Simulated datasets:

1. *HIV-1 subtype B mixture and MiSeq reads (HIV7sim).* This benchmark contains simulated Illumina MiSeq reads with 10k-coverage of 1-kb *pol* sequences. Reads were simulated from seven equally distributed HIV-1 variants chosen from the NCBI database: AY835778, AY835770, AY835771, AY835777, AY835763, AY835762, and AY835757. Hamming distances between clones are in the range from 0.6-3.0%(6 to 30-bp differences). SimSeq [52] was used to generate reads.

2. *IAV mixture and MiSeq reads (IAV10sim).* This benchmark contains simulated IAV Illumina MiSeq reads with IAV haplotypes from IAV10exp benchmark. Paired Illumina MiSeq reads were simulated by SimSeq [52] using default error profile.

Performance of haplotyping methods

CliqueSNV was comapred to 2SNV, PredictHaplo, and aBayesQR. CliqueSNV, PredictHaplo and aBayesQR handle Illumina reads and were compared on HIV9exp, HIV2exp, HIV5exp, HIV7sim, and IAV10sim datasets. CliqueSNV, 2SNV, and PredictHaplo were also

tested on the IAV10exp PacBio dataset. Additionally, consensus sequences [92] were used in validation to evaluate sequences most similar to those generated by the Sanger method [95].

Haplotype reconstruction results for compared methods are shown in Table 3.2 [33]. For five out of six datasets, CliqueSNV demonstrated the best precision and recall. For the HIV5exp dataset, PredictHaplo outperformed CliqueSNV in prediction of false positive variants. CliqueSNV demonstrated 100% precision and recall for three datasets, including the HIV2exp and IAV10exp and HIV7sim.

Benchmark	CliqueSNV		aBayesQR		PredictHaplo	
	Precision	Recall	Precision	Recall	Precision	Recall
HIV9exp	**0.50**	**0.33**	0.08	0.11	0.00	0.00
HIV2exp	**1.00**	**1.00**	0.08	0.50	0.33	0.50
HIV5exp	0.50	**0.60**	0.00	0.00	**0.75**	**0.60**
HIV7sim	**1.00**	**1.00**	0.43	0.43	0.00	0.00
IAV10sim	**0.70**	**0.70**	0.13	0.10	0.33	0.10

(a)

Benchmark	CliqueSNV		2SNV		PredictHaplo	
	Precision	Recall	Precision	Recall	Precision	Recall
IAV10exp	**1.00**	**1.00**	0.82	0.90	0.70	0.70

(b)

Table (3.2) Prediction statistics of haplotype reconstruction methods using experimental and simulated (a) MiSeq and (b) PacBio data. The precision and recall was evaluated stringently such that if a predicted haplotype has at least one mismatch to its closest answer, then that haplotype is scored as a false positive.

Figure 3.2 [33] shows the EMD distance between inferred and true haplotypes for MiSeq datasets, and exact EMD values are provided in Table 3.3 [33].

In terms of EMD, CliqueSNV showed better results than other tools on all benchmarks, and made almost ideal predictions in some cases, where EMD was close to zero. PredictHaplo outperformed aBayesQR on four out of five MiSeq datasets. As for aBayesQR, it showed almost zero-EMD on HIV7sim, but performed significantly worse than other methods on HIV5exp.

Benchmark	Consensus	CliqueSNV		aBayesQR		PredictHaplo	
	EMD	EMD	Improvement	EMD	Improvement	EMD	Improvement
HIV9exp	4.18	**2.47**	**40.83 %**	5.09	-21.85 %	3.58	14.30 %
HIV2exp	5.50	**1.71**	**68.95 %**	3.53	35.80 %	2.91	47.08 %
HIV5exp	19.40	**4.03**	**79.20 %**	19.22	0.91 %	6.80	64.97 %
HIV7sim	11.00	**0.02**	**99.84 %**	0.84	92.34 %	5.87	46.68 %
IAV10sim	4.22	**0.09**	**97.77 %**	3.64	13.73 %	3.03	28.15 %
Mean Improvement			**77.32 %**		24.19 %		40.23 %

(a)

Benchmark	Consensus	CliqueSNV		2SNV		PredictHaplo	
	EMD	EMD	Improvement	EMD	Improvement	EMD	Improvement
IAV10exp	4.22	**0.22**	**94.69%**	0.23	94.46%	0.38	91.02%

(b)

Table (3.3) Earth Movers' Distance from predicted haplotypes to the true haplotype population and haplotyping method improvement. Four haplotyping methods(aBayesQR, CliequeSNV, Consensus, PredictHaplo) are benchmarked on five MiSeq datasets (a) and IAV10exp dataset (b). The improvement shows how much better is prediction of haplotyping method over inferred consensus, and it is calculated as $\frac{(EMD_c - EMD_m)\times 100\%}{EMD_c}$, where EMD_c is an EMD for consensus, and EMD_m is an EMD for method. CliqueSNV outperformed all other methods in accuracy on all datasets.

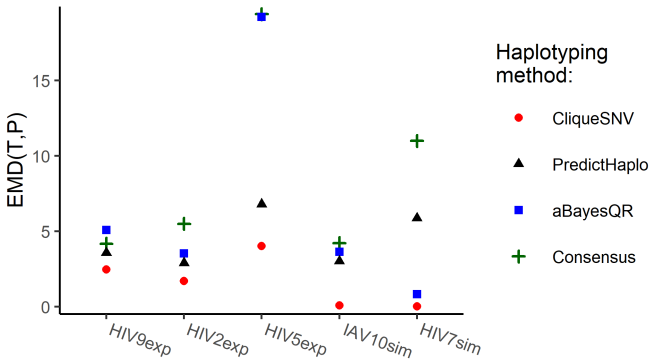

Figure (3.2) Earth Movers' Distance (EMD) between true and reconstructed haplotype populations. Four haplotyping methods (CliqueSNV, aBayesQR, PredictHaplo, Consensus) are benchmarked using three experimental and two simulated datasets for human immunodeficiency virus type 1 (HIV-1) and influenza A virus (IAV). For all benchmarks the CliqueSNV predictions are the closest to the true populations.

Runtime comparison

Each method was executed on a cluster (Intel(R) Xeon(R) CPU X5550 2.67GHz x2 8 cores per CPU, DIMM DDR3 1,333 MHz RAM 4Gb x12) with the CentOS 6.4 operating

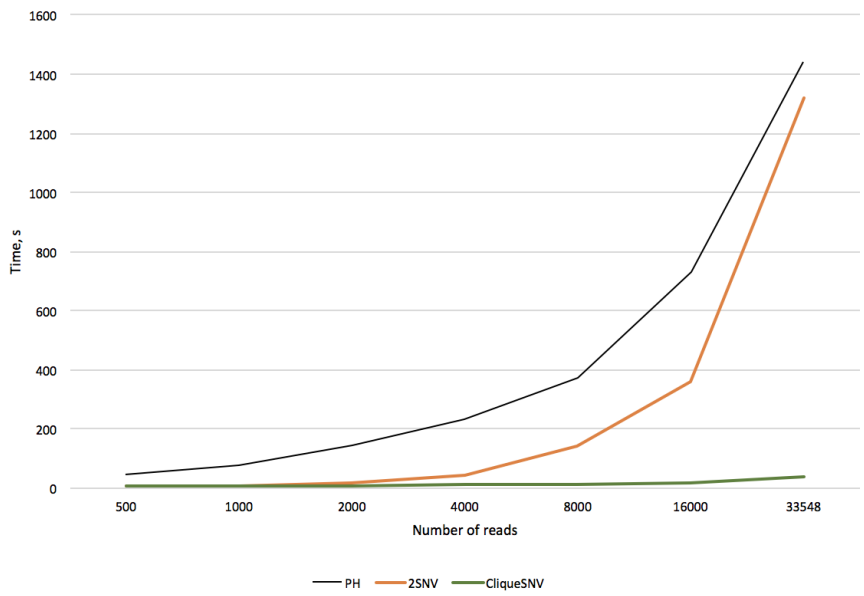

Figure (3.3) Runtime of PredictHaplo (PH), 2SNV and CliqueSNV on datasets with different sizes.

system. CliqueSNV demonstrated sublinear (with respect ot the number of reads) runtime, as opposed to PredictHaplo and 2SNV. Runtime complexity of CliqueSNV is quadratic with respect to the number of SNVs rather than by the length of the sequencing region. CliqueSNV is significantly faster than aBayesQR and PredictHaplo. In particular, HIV2exp dataset took over ten hours for aBayesQR, 24 minutes for PhedictHaplo, and 79 seconds for CliqueSNV (see Figures 3.3 [33]).

PART 4

DISCUSSION AND FUTURE WORK

Application of molecular viral analysis to investigation of outbreaks is a promising research area, that also generates novel computational challenges. Methods, that are mentioned in this dissertation can be extended in several directions.

In particular, for entropy-based SARS-CoV-2 clustering, only one-column entropy is currently used. However, other entropies can be computed, which should improve clustering results and provide more insights into the dynamics of the virus, such as potential predecessors of the novel B.1.1.7 strain.

EMD k-mer-based viral outbreak analysis tool can also be extended in several ways. Given approach needs additional attention when dealing with unstable datasets, such as some of the datasets, produced by PANGEA study [96, 97]. In particular, when coverage is low or varies greatly between samples, some extra steps, such as multiple window analysis may be required. This way, next step is to improve the algorithm so that it is applicable to a wider range of datasets. Another future direction is the application of this method to the assessment of infection stage as recent or chronic [98] to analyze the correlation between distance to outbreak source and time elapsed since the infection event. This, in turn, should provide insights into the mutation rate of the virus after the infection event.

REFERENCES

[1] A. S. Lauring and R. Andino, "Quasispecies theory and the behavior of RNA viruses," *PLoS pathogens*, vol. 6, no. 7, p. e1001005, 2010.

[2] F. Sanger, S. Nicklen, and A. R. Coulson, "Dna sequencing with chain-terminating inhibitors," *Proceedings of the national academy of sciences*, vol. 74, no. 12, pp. 5463–5467, 1977.

[3] X. V. Wang, N. Blades, J. Ding, R. Sultana, and G. Parmigiani, "Estimation of sequencing error rates in short reads," *BMC bioinformatics*, vol. 13, no. 1, p. 185, 2012.

[4] R. E. F. N. T. Castro C, Marine R, "The effect of variant interference on de novo assembly for viral deep sequencing," *bioRxiv*, 2019.

[5] H. Töpfer, "Sequencing approach to analyze the role of quasispecies for classical swine fever.," *Virology*, no. 438, pp. 14–19, 2013.

[6] S. Vignuzzi, "Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population.," *Nature*, no. 439, pp. 344–348, 2006.

[7] B. Henn, "Whole genome deep sequencing of hiv-1 reveals the impact of early minor variants upon immune recognition during acute infection.," *PLoS pathogens*, no. 8.

[8] G. Beerenwinkel, "Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data.," *Microbio*, no. 3, p. 239, 2012.

[9] M. Hinkley, "A systems analysis of mutational effects in hiv-1 protease and reverse transcriptase.," *Natgenet*, no. 43, pp. 487–489, 2011.

[10] A. Töpfer, "Viral quasispecies assembly via maximal clique enumeration.," *PLoS computational biology*, vol. 10, no. 3, 2014.

[11] J. O. Wertheim, A. J. L. Brown, N. L. Hepler, , and S. L. K. Pond, "The global transmission network of hiv-1," *Journal of Infectious Diseases*, vol. 209, no. 2, pp. 304–313, 2014.

[12] J. O. Wertheim, S. L. K. Pond, L. A. Forgione, S. R. Mehta, B. Murrell, S. Shah, D. M. Smith, K. Scheffler, and L. V. Torian, "Social and genetic networks of hiv-1 transmission in new york city," *PLoS pathogens*, vol. 13, no. 1, p. e1006000, 2017.

[13] D. S. Campo, G.-L. Xia, Z. Dimitrova, Y. Lin, J. C. Forbi, L. Ganova-Raeva, L. Punkova, S. Ramachandran, H. Thai, P. Skums, *et al.*, "Accurate genetic detection of hepatitis c virus transmissions in outbreak settings," *Journal of Infectious Diseases*, vol. 213, no. 6, pp. 957–965, 2016.

[14] K. W. RK, Ravi and K. M, "Miseq: A next generation sequencing platform for genomic analysis," pp. 223–232, February 2018.

[15] J. Baaijens, "De novo assembly of viral quasispecies using overlap graphs.," *Genome Research*, vol. 5.

[16] 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, pp. 835–848, 2017.

[17] J. W. Drake and J. J. Holland, "Mutation rates among rna viruses," *Proceedings of the National Academy of Sciences*, vol. 96, no. 24, pp. 13910–13913, 1999.

[18] N. Eriksson, L. Pachter, Y. Mitsuya, S.-Y. Rhee, C. Wang, B. Gharizadeh, M. Ronaghi, R. W. Shafer, and N. Beerenwinkel, "Viral population estimation using pyrosequencing," *PLoS computational biology*, vol. 4, no. 5, p. e1000074, 2008.

[19] J. Archer, M. S. Braverman, B. E. Taillon, B. Desany, I. James, P. R. Harrigan, M. Lewis, and D. L. Robertson, "Detection of low-frequency pretherapy chemokine

(cxc motif) receptor 4-using hiv-1 with ultra-deep pyrosequencing," *AIDS (London, England)*, vol. 23, no. 10, p. 1209, 2009.

[20] C. Hoffmann, N. Minkah, J. Leipzig, G. Wang, M. Q. Arens, P. Tebas, and F. D. Bushman, "Dna bar coding and pyrosequencing to identify rare hiv drug resistance mutations," *Nucleic acids research*, vol. 35, no. 13, p. e91, 2007.

[21] W. Wang, X. Zhang, Y. Xu, G. M. Weinstock, A. M. Di Bisceglie, and X. Fan, "High-resolution quantification of hepatitis c virus genome-wide mutation load and its correlation with the outcome of peginterferon-alpha2a and ribavirin combination therapy," *PloS one*, vol. 9, no. 6, p. e100131, 2014.

[22] P. Skums, D. S. Campo, Z. Dimitrova, G. Vaughan, D. T. Lau, and Y. Khudyakov, "Numerical detection, measuring and analysis of differential interferon resistance for individual hcv intra-host variants and its influence on the therapy response," *In silico biology*, vol. 11, no. 5, pp. 263–269, 2011.

[23] D. S. Campo, P. Skums, Z. Dimitrova, G. Vaughan, J. C. Forbi, C.-G. Teo, Y. Khudyakov, and D. T. Lau, "Drug resistance of a viral population and its individual intrahost variants during the first 48 hours of therapy," *Clinical Pharmacology & Therapeutics*, vol. 95, no. 6, pp. 627–635, 2014.

[24] E. O. Romero-Severson, I. Bulla, and T. Leitner, "Phylogenetically resolving epidemiologic linkage," *Proceedings of the National Academy of Sciences*, vol. 113, no. 10, pp. 2690–2695, 2016.

[25] N. De Maio, C.-H. Wu, and D. J. Wilson, "Scotti: Efficient reconstruction of transmission within outbreaks with the structured coalescent," *arXiv preprint arXiv:1603.01994*, 2016.

[26] G. E. Fischer, M. K. Schaefer, B. J. Labus, L. Sands, P. Rowley, I. A. Azzam, P. Armour, Y. E. Khudyakov, Y. Lin, G. Xia, *et al.*, "Hepatitis c virus infections from unsafe

injection practices at an endoscopy clinic in las vegas, nevada, 2007–2008," *Clinical infectious diseases*, vol. 51, no. 3, pp. 267–273, 2010.

[27] A. Apostolou, M. L. Bartholomew, R. Greeley, S. M. Guilfoyle, M. Gordon, C. Genese, J. P. Davis, B. Montana, and G. Borlaug, "Transmission of hepatitis c virus associated with surgical procedures-new jersey 2010 and wisconsin 2011.," *MMWR. Morbidity and mortality weekly report*, vol. 64, no. 7, pp. 165–170, 2015.

[28] S. Knyazev, L. Hughes, P. Skums, and A. Zelikovsky, "Epidemiological data analysis of viral quasispecies in the next-generation sequencing era," *Briefings in Bioinformatics*, June 2020.

[29] 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)*, vol. 11, no. 1, pp. 182–191, 2014.

[30] S. Ahn and H. Vikalo, "abayesqr: A bayesian method for reconstruction of viral populations characterized by low diversity," in *International Conference on Research in Computational Molecular Biology*, pp. 353–369, Springer, 2017.

[31] S. Elbe and G. Buckland-Merrett, "Data, disease and diplomacy: GISAID's innovative contribution to global health," *Global Challenges*, vol. 1, pp. 33–46, 2017.

[32] "EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK."

[33] S. Knyazev, V. Tsyvina, A. Melnyk, A. Artyomenko, T. Malygina, Y. B. Porozov, E. Campbell, W. M. Switzer, P. Skums, and A. Zelikovsky, "CliqueSNV: Scalable reconstruction of intra-host viral populations from ngs reads," *bioRxiv*, 2018.

[34] Z. Huang, "A fast clustering algorithm to cluster very large categorical data sets in data mining," in *the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, pp. 1–8, 1997.

[35] Z. Huang, "Extensions to the k-modes algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[36] M. Anderberg, *Cluster Analysis for Applications*. Academic Press, 1973.

[37] J. McQueen, "Some methods for classification and analysis of multivariate observations," in *the 5th Berkely Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.

[38] S. Ciccolella, M. Soto, M. D. Patterson, G. D. Vedova, I. Hajirasouliha, and P. Bonizzoni, "gpps: An ILP-based approach for inferring cancer progression with mutation losses from single cell data," *BMC Bioinformatics*, vol. 21, no. 413, 2020.

[39] S. Ciccolella*, M. Patterson*, P. Bonizzoni, and G. D. Vedova, "Effective clustering for single cell sequencing cancer data," in *the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB, Niagara Falls, NY, USA, 2019)*, ACM, pp. 437–446, 2019.

[40] S. Ciccolella, C. Ricketts, M. S. Gomez, M. Patterson, D. Silverbush, P. Bonizzoni, I. Hajirasouliha, and G. D. Vedova, "Inferring cancer progression from single-cell sequencing while allowing mutation losses," *Bioinformatics*, vol. btaa722, 2020.

[41] K. Jahn, J. Kuipers, and N. Beerenwinkel, "Tree inference for single-cell data," *Genome Biology*, vol. 17, no. 1, p. 86, 2016.

[42] K. Tamura and M. Nei, "Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees," *Molecular Biology and Evolution*, vol. 10, no. 3, p. 512–526, 1993.

[43] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[44] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society*, vol. 63, no. 2, pp. 411–423, 2001.

[45] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[46] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[47] T. Li, S. Ma, and M. Ogihara, "Entropy-based criterion in categorical clustering," in *Twenty-First International Conference on Machine Learning*, 2004.

[48] T. D. Schneider and R. Stephens, "Sequence logos: a new way to display consensus sequences," *Nucleic Acids Research*, vol. 18, no. 20, pp. 6097–6100, 1990.

[49] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," *1998 IEEE International Conference on Computer Vision*, 1998.

[50] O. Glebova, S. Knyazev, A. Melnyk, A. Artyomenko, Y. Khudyakov, A. Zelikovsky, and P. Skums, "Inference of genetic relatedness between viral quasispecies from sequencing data," *BMC Genomics*, no. 18, 2017.

[51] T. C.-U. Consortium, "An integrated national scale SARS-CoV-2 genomic surveillance network," *Lancet Microbe 2020*, vol. 1, no. 3, pp. 99–100, 2020.

[52] S. Benidt and D. Nettleton, "Simseq: a nonparametric approach to simulation of rna-sequence datasets," *Bioinformatics*, vol. 31, no. 13, pp. 2131–2140, 2015.

[53] M. M. Deza and E. Deza, "Encyclopedia of distances," 2009.

[54] L. du Plessis, J. T. McCrone, A. E. Zarebski, V. Hill, C. Ruis, B. Gutierrez, J. Raghwani, J. Ashworth, R. Colquhoun, T. R. Connor, N. R. Faria, B. Jackson, N. J. Loman,

Á. O'Toole, S. M. Nicholls, K. V. Parag, E. Scher, T. I. Vasylyeva, E. M. Volz, A. Watts, I. I. Bogoch, K. Khan, D. M. Aanensen, M. U. G. Kraemer, A. Rambaut, and O. G. Pybus, "Establishment and lineage dynamics of the sars-cov-2 epidemic in the uk," *Science*, 2021.

[55] Q. Bukhari, Y. Jameel, J. Massaro, R. D'Agostino, and S. Khan, "Periodic oscillations in daily reported infections and deaths for coronavirus disease 2019," *JAMA network open*, vol. 3, no. 8, p. e2017521, 2020.

[56] E. Volz, S. Mishra, M. Chand, J. C. Barrett, R. Johnson, L. Geidelberg, W. R. Hinsley, D. J. Laydon, G. Dabrera, Á. O'Toole, R. Amato, M. Ragonnet-Cronin, I. Harrison, B. Jackson, C. V. Ariani, O. Boyd, N. Loman, J. T. McCrone, S. Gon calves, D. Jorgensen, R. Myers, V. Hill, D. K. Jackson, K. Gaythorpe, N. Groves, J. Sillitoe, D. P. Kwiatkowski, S. Flaxman, O. Ratmann, S. Bhatt, S. Hopkins, A. Gandy, A. Rambaut, and N. M. Ferguson, "Transmission of sars-cov-2 lineage b.1.1.7 in england: Insights from linking epidemiological and genetic data," *medRxiv*, 2021.

[57] P. H. Kilmarx, "Global epidemiology of hiv," *Current Opinion in HIV and AIDS*, vol. 4, no. 4, pp. 240–246, 2009.

[58] B. Hajarizadeh, J. Grebely, and G. J. Dore, "Epidemiology and natural history of hcv infection," *Nature Reviews Gastroenterology and Hepatology*, vol. 10, no. 9, pp. 553–562, 2013.

[59] R. Lozano, M. Naghavi, K. Foreman, S. Lim, K. Shibuya, V. Aboyans, J. Abraham, T. Adair, R. Aggarwal, S. Y. Ahn, *et al.*, "Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010," *The lancet*, vol. 380, no. 9859, pp. 2095–2128, 2012.

[60] M. Eigen, J. McCaskill, and P. Schuster, "The molecular quasi-species," *Advances in chemical physics*, vol. 75, pp. 149–263, 1989.

[61] M. Martell, J. Esteban, J. Quer, J. Genesca, A. Weiner, R. Esteban, J. Guardia, and J. Gomez, "Hepatitis c virus (hcv) circulates as a population of different but closely related genomes: quasispecies nature of hcv genome distribution," *Journal of Virology, 66*, pp. 3225–3229, 1992.

[62] D. Steinhauer and J. Holland, "Rapid evolution of rna viruses," *Annual Review of Microbiology, 41*, pp. 409–433, 1987.

[63] E. Domingo, J. Sheldon, and C. Perales, "Viral quasispecies evolution," *Microbiology and Molecular Biology Reviews*, vol. 76, no. 2, pp. 159–216, 2012.

[64] F. Rodriguez-Frias, M. Buti, D. Tabernero, and M. Homs, "Quasispecies structure, cornerstone of hepatitis b virus infection: mass sequencing approach," *World J Gastroenterol*, vol. 19, no. 41, pp. 6995–7023, 2013.

[65] N. Beerenwinkel, T. Sing, T. Lengauer, J. Rahnenfuehrer, and K. Roomp, "Computational methods for the design of effective therapies against drug resistant HIV strains.," *Bioinformatics*, vol. 21, pp. 3943–3950, 2005.

[66] N. G. Douek DC, Kwong PD, "The rational design of an AIDS vaccine.," *Cell*, vol. 124, pp. 677–681, 2006.

[67] B. Gaschen, J. Taylor, K. Yusim, B. Foley, and F. Gao, "Diversity considerations in HIV-1 vaccine selection," *Science*, vol. 296, pp. 2354–2360, 2002.

[68] J. Holland, J. De La Torre, and D. Steinhauer, "RNA virus populations as quasispecies," *Curr Top Microbiol Immunol*, vol. 176, pp. 1–20, 1992.

[69] S.-Y. Rhee, T. Liu, S. Holmes, and R. Shafer, "HIV-1 subtype B protease and reverse transcriptase amino acid covariation," *PLoS Comput Biol*, vol. 3, p. e87, 2007.

[70] P. Skums, L. Bunimovich, and Y. Khudyakov, "Antigenic cooperation among intrahost hcv variants organized into a complex network of cross-immunoreactivity," *Proceedings of the National Academy of Sciences*, vol. 112, no. 21, pp. 6653–6658, 2015.

[71] A. Töpfer, O. Zagordi, S. Prabhakaran, V. Roth, E. Halperin, and N. Beerenwinkel, "Probabilistic inference of viral quasispecies subject to recombination," *Journal of Computational Biology*, vol. 20, no. 2, pp. 113–123, 2013.

[72] 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.

[73] 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, pp. i329–i337, 2014.

[74] P. Skums, N. Mancuso, A. Artyomenko, B. Tork, I. Mandoiu, Y. Khudyakov, and A. Zelikovsky, "Reconstruction of viral population structure from next-generation sequencing data using multicommodity flows," *BMC bioinformatics*, vol. 14, no. Suppl 9, p. S2, 2013.

[75] N. Mancuso, B. Tork, P. Skums, L. Ganova-Raeva, I. Măndoiu, and A. Zelikovsky, "Reconstructing viral quasispecies from ngs amplicon reads," *In silico biology*, vol. 11, no. 5, pp. 237–249, 2011.

[76] O. Zagordi, A. Bhattacharya, N. Eriksson, and N. Beerenwinkel, "Shorah: estimating the genetic diversity of a mixed sample from next-generation sequencing data," *BMC bioinformatics*, vol. 12, no. 1, p. 119, 2011.

[77] I. Astrovskaya, B. Tork, S. Mangul, K. Westbrooks, I. Măndoiu, P. Balfe, and A. Zelikovsky, "Inferring viral quasispecies spectra from 454 pyrosequencing reads," *BMC bioinformatics*, vol. 12, no. Suppl 6, p. S1, 2011.

[78] M. C. Prosperi and M. Salemi, "Qure: software for viral quasispecies reconstruction from next-generation sequencing data," *Bioinformatics*, vol. 28, no. 1, pp. 132–133, 2012.

[79] O. Zagordi, A. Töpfer, S. Prabhakaran, V. Roth, E. Halperin, and N. Beerenwinkel, "Probabilistic inference of viral quasispecies subject to recombination," in *Proceedings of the 16th Annual international conference on Research in Computational Molecular Biology*, RECOMB'12, (Berlin, Heidelberg), pp. 342–354, Springer-Verlag, 2012.

[80] P. Skums, Z. Dimitrova, D. S. Campo, G. Vaughan, L. Rossi, J. C. Forbi, J. Yokosawa, A. Zelikovsky, and Y. Khudyakov, "Efficient error correction for next-generation sequencing of viral amplicons," *BMC Bioinformatics*, vol. 13, no. S-10, p. S6, 2012.

[81] P. Skums, A. Artyomenko, O. Glebova, D. S. Campo, Z. Dimitrova, A. Zelikovsky, and Y. Khudyakov, "Error correction of ngs reads from viral populations," *Computational Methods for Next Generation Sequencing Data Analysis*, 2016.

[82] S. Barik, S. Das, and H. Vikalo, "Viral quasispecies reconstruction via correlation clustering," *bioRxiv*, p. 096768, 2016.

[83] K. Westbrooks, I. Astrovskaya, D. Campo, Y. Khudyakov, P. Berman, and A. Zelikovsky, "Hcv quasispecies assembly using network flows," *Bioinformatics Research and Applications*, pp. 159–170, 2008.

[84] 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*, vol. 8, no. 3, p. e1002417, 2012.

[85] X. Yang, P. Charlebois, A. Macalalad, M. R. Henn, and M. C. Zody, "V-phaser 2: variant inference for viral populations," *BMC genomics*, vol. 14, no. 1, p. 674, 2013.

[86] 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*, vol. 91, pp. 40–47, 2015.

[87] B. M. Verbist, K. Thys, J. Reumers, Y. Wetzels, K. Van der Borght, W. Talloen, J. Aerssens, L. Clement, and O. Thas, "Virvarseq: a low-frequency virus variant detection pipeline for illumina sequencing using adaptive base-calling accuracy filtering," *Bioinformatics*, vol. 31, no. 1, pp. 94–101, 2014.

[88] 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*, pp. 164–175, Springer International Publishing, 2016.

[89] E. Gerasimov, *Analysis of NGS Data from Immune Response and Viral Samples*. PhD thesis, Georgia State University, 2017.

[90] E. Levina and P. Bickel, "The earthmover's distance is the mallows distance: Some insights from statistics," *Proceedings of ICCV 2001*, pp. 251–256, 2001.

[91] C. L. Mallows, "A note on asymptotic joint normality," *Annals of Mathematical Statistics*, vol. 43, no. 2, pp. 508–515, 1972.

[92] X. Yang, P. Charlebois, S. Gnerre, M. G. Coole, N. J. Lennon, J. Z. Levin, J. Qu, E. M. Ryan, M. C. Zody, and M. R. Henn, "De novo assembly of highly diverse viral populations," *BMC genomics*, vol. 13, no. 1, p. 475, 2012.

[93] F. Zanini, J. Brodin, L. Thebo, C. Lanz, G. Bratt, J. Albert, and R. A. Neher, "Population genomics of intrapatient hiv-1 evolution," *eLife*, Dec 2015.

[94] F. D. Giallonardo, A. Töpfer, M. Rey, S. Prabhakaran, Y. Duport, C. Leemann, S. Schmutz, N. K. Campbell, B. Joos, M. R. Lecca, A. Patrignani, M. Däumer, C. Beisel, P. Rusert, A. Trkola, H. F. Günthard, V. Roth, N. Beerenwinkel, and K. J. Metzner, "Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations," *Nucleic Acids Research*, vol. 42, no. 14, p. e115, 2014.

[95] D. E. Kireev, A. E. Lopatukhin, A. V. Murzakova, E. V. Pimkina, A. S. Speranskaya, A. D. Neverov, G. G. Fedonin, Y. S. Fantin, and G. A. Shipulin, "Evaluating the accuracy and sensitivity of detecting minority HIV-1 populations by Illumina next-generation sequencing," *J. Virol. Methods*, vol. 261, pp. 40–45, 11 2018.

[96] D. Pillay, J. Herbeck, M. S. Cohen, T. de Oliveira, C. Fraser, O. Ratmann, A. L. Brown, P. Kellam, and P.-H. Consortium, "Pangea-hiv: phylogenetics for generalised epidemics in africa," *Infectious diseases*, vol. 15, March 2015.

[97] L. Abeler-Dörner, M. Grabowski, A. Rambaut, D. Pillay, C. Fraser, and P.-H. Consortium, "Pangea-hiv 2: Phylogenetics and networks for generalised epidemics in africa," *Curr Opin HIV AIDS*, vol. 14, May 2019.

[98] I. P., L. J., Y. Khudyakov, A. Zelikovsky, and P. Skums, "Quantitative differences between intra-host hcv populations from persons with recently established and persistent infections," *Virus Evolution*, vol. 7, January 2021.