CliqueSNV: Scalable Reconstruction of Intra-Host Viral Populations from NGS Reads

Sergey Knyazev[1,2,3]*, Viachaslau Tsyvina[1]*, Andrew Melnyk[1], Alexander Artyomenko[4], Tatiana Malygina[5], Yuri B. Porozov[5,6], Ellsworth Campbell[2], William M. Switzer[2], Pavel Skums[1,2]†, and Alex Zelikovsky[1,6]†

[1]Department of Computer Science, Georgia State University, Atlanta, GA, 30303, USA
[2]Centers for Disease Control and Prevention, Atlanta, GA, 30333, USA
[2]Oak Ridge Institute for Science and Education, Oak Ridge, TN, 37830, USA
[4]Guardant Health Inc., Redwood City, CA, 94063, USA
[5]The laboratory of bioinformatics, ITMO University, St. Petersburg, 197101, Russia
[6]The laboratory of bioinformatics, I.M. Sechenov First Moscow State Medical University, Moscow, 119991, Russia

ABSTRACT

Highly mutable RNA viruses such as influenza A virus, human immunodeficiency virus and hepatitis C virus exist in infected hosts as highly heterogeneous populations of closely related genomic variants. The presence of low-frequency variants with few mutations with respect to major strains may result in an immune escape, emergence of drug resistance, and an increase of virulence and infectivity. Next-generation sequencing technologies permit detection of sample intra-host viral population at extremely great depth, thus providing an opportunity to access low-frequency variants. Long read lengths offered by single-molecule sequencing technologies allow all viral variants to be sequenced in a single pass. However, high sequencing error rates limit the ability to study heterogeneous viral populations composed of rare, closely related variants.

In this article, we present CliqueSNV, a novel reference-based method for reconstruction of viral variants from NGS data. It efficiently constructs an allele graph based on linkage between single nucleotide variations and identifies true viral variants by merging cliques of that graph using combinatorial optimization techniques. The new method outperforms existing methods in both accuracy and running time on experimental and simulated NGS data for titrated levels of known viral variants. For PacBio reads, it accurately reconstructs variants with frequency as low as 0.1%. For Illumina reads, it fully reconstructs main variants. The open source implementation of CliqueSNV is freely available for download at `https://github.com/vyacheslav-tsivina/CliqueSNV`

1 INTRODUCTION

Highly mutable RNA viruses such as influenza A virus, human immunodeficiency virus (HIV) and hepatitis C virus (HCV) continue to be major public health threats [20, 17, 22]. The hallmark of RNA viruses is their high intra-host genetic diversity originated from error-prone replication [13]. As a result, such viruses exist in infected hosts as heterogeneous populations of closely genetically related variants, known to virologists as *quasispecies* [23, 27, 40, 11, 34]. The composition and structure of intra-host viral populations plays a crucial role in disease progression and epidemic spread. In particular, low-frequency variants with few mutations with respect to dominant haplotypes may be responsible for viral transmission, immune escape, drug resistance, increase of virulence and infectivity [7, 12, 15, 19, 33, 10, 37].

Next-generation sequencing (NGS) technologies now provide versatile opportunities to study viral quasispecies. In particular, the popular Illumina MiSeq/HiSeq platforms produce 25-320 million reads which allow multiple coverage of highly variable viral genomic regions. This high coverage is essential for capturing rare variants. However, *haplotyping* of heterogeneous populations (i.e., reconstruction of full-length genomic variants and their frequencies) is complicated due to the vast number of sequencing reads, the need to assemble an unknown number of closely related viral sequences and to identify and preserve low-frequency variants. Single-molecule sequencing technologies, such as PacBio, provide an alternative to short-read sequencing by allowing all viral variants to be sequenced in a single pass. However, the high level of sequence noise (background or platform specific sequencing errors) produced by all currently available platforms makes inference of low-frequency highly genetically related variants especially challenging, since it is required to distinguish between real and artificial genetic heterogeneity produced by sequencing errors.

In the recent years, a number of computational tools for inference of viral quasispecies populations from "noisy" NGS data have been proposed, including Savage [5], PredictHaplo [30], aBayesQR [1], QuasiRecomb [42], HaploClique [41], VGA [26], VirA [39, 25], SHORAH [48], ViSpA [4], QURE [31] and others [49, 38, 36, 6, 45]. Even though these algorithms proved useful in many applications, accurate and scalable viral haplotyping remains a challenge.

In this paper, we present CliqueSNV, a novel method designed to reconstruct closely related low-frequency intra-host viral variants from noisy next-generation and third-generation sequencing data. CliqueSNV eliminates the need for preliminary error correction and assembly and infers haplotypes from patterns in distributions of SNVs in sequencing reads. It is suitable for long single-molecule reads (PacBio) as well as short paired reads (Illumina). CliqueSNV uses linkage between single nucleotide variations (SNVs) to distinguish them from sequencing errors efficiently. It constructs an allele graph with edges connecting linked SNVs and identifies true viral variants by merging cliques of that graph using combinatorial optimization techniques.

*Equal contributor

†Corresponding authors, e-mails: {pskums, alexz}@cs.gsu.edu

Previously, several tools such as V-phaser [24], V-phaser2 [47] and CoVaMa [35] exploited linkage of nucleotide variants, but they did not take into account sequencing errors when deciding whether two variants are linked. Results of these tools show that they were unable to reliably detect variants of frequency even higher than the error rate of sequencing. The 2SNV algorithm [2] accommodated errors in links and was the first such tool to be able to detect haplotypes with a frequency below the error rate correctly. The proposed CliqueSNV method keeps the basic idea of 2SNV linkage analysis but develops a novel approach for collecting multiple SNV's and inference of true haplotypes. Unlike 2SNV, which hierarchically clusters together reads containing pairs of linked SNVs, CliqueSNV identifies true viral variants in a single clustering using an efficient merging of cliques of the allele graph. Furthermore, unlike 2SNV, which is designed only for single amplicon data, CliqueSNV can handle short paired reads from shotgun experiments. Finally, the new method identifies linked SNVs and constructs allele graphs using highly efficient data structures. As a result, CliqueSNV is more accurate and significantly faster than 2SNV and capable of rapidly handling millions of reads in of minutes.

Several previously published methods (e.g., HaploClique [41], Savage [5]) reconstructed viral haplotypes using maximal cliques in a graph, where vertices represent reads. These methods infer haplotypes by iteratively merging these read cliques, thus heavily relying on the correct order of merging. In contrast, our proposed approach finds maximal cliques in a graph with vertices corresponding to alleles. This facilitates a significant performance increase since for viruses the size of the allele graph is significantly smaller than the size of the read graph. Furthermore, the clique merging problem is formulated and solved as a combinatorial problem on the auxiliary graph of cliques of the allele graph, thus allowing an increase of the algorithm's accuracy.

CliqueSNV was validated on simulated and experimental data and compared with Savage [5], PredictHaplo [30], aBayesQR [1] and 2SNV [2]. We benchmark the tools using the results of a PacBio sequencing experiment on a sample containing a titrated level of known Influenza A (IAV) viral variants, on similar data sets for experimental HIV-1 single-read and paired-end Illumina data and simulated Illumina HIV-1 and IAV data. In addition to standard algorithm performance measures, we used a new measure based on earth mover's distance between real and reconstructed haplotype distributions. In this validation study, CliqueSNV significantly outperformed these other methods in both accuracy and running time.

2 METHODS

CliqueSNV algorithm

Data input for CliqueSNV consists of a set of N PacBio long reads or Illumina paired reads from an intra-host viral population aligned to a genomic region of interest. Output is the set of inferred viral variants with their frequencies. Algorithm 1 describes the formal high-level pseudocode of the CliqueSNV algorithm. CliqueSNV consists of the following six major steps detailed below:

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

```
Algorithm 1 CliqueSNV Algorithm
procedure 1: finding linked SNV pairs
    Split the read alignment $M_{L\times N}$ into binary matrix $4M$
    Construct a compact representation of the binary matrix $4M$
    For each $I, J \in \{1, \dots, 4L\}$ find $O^{IJ}$ and $O_{22}^{IJ}$, where
        $O^{IJ}$ = # of reads covering both $I$ and $J$
        $O_{22}^{IJ}$ = # of reads with both minor alleles
        If $O_{22}^{IJ} > \epsilon O^{IJ}$ compute $p$-value (3) (default $\epsilon = 0.0003$)
    Find all linked SNV pairs with the adjusted p-value $< 1\%$
procedure 2: constructing the allele graph
    Filter out $10\%$ of the most erroneous PacBio reads
    Construct the allele graph $G = (V, E)$, where
        $V = \{1, \dots, 4L\}$, and $E$ are links between minor alleles
procedure 3: finding maximal cliques in the allele graph using Bron-Kerbosch algorithm [9]
procedure 4: merging cliques in the clique graph
    Find the clique graph $C_G$ with forbidden pairs.
    Find all maximal connected subgraphs in $C_G$.
    Merge all cliques inside each maximal connected subgraph.
procedure 5: finding consensus viral variants
    Find the set $S$ of all positions that belong to at least one clique.
    Make an empty clique on $S$.
    Assign each read to the closest clique.
    Find the consensus $v(q)$ of all assigned reads for each $q$.
procedure 6: estimating frequencies of the viral variants
```

1: Finding linked SNV pairs. CliqueSNV uses pairs of linked SNVs which have been previously introduced for the 2SNV method [3]. Let the major (minor) allele at a given genomic position be the allele observed in the majority (minority) of reads covering this position. The pair of alleles at two positions will be referred to as a 2-haplotype. Assuming that errors are random, it has been proved that in any two positions I and J with major alleles denoted 1, and minor alleles denoted 2, if the variant (22) does not exist, then the expected number E_{22} of reads containing minor alleles should not exceed

$$E_{22} \leq \frac{E_{21} \cdot E_{12}}{E_{11}} \tag{1}$$

where E_{21}, E_{12}, and E_{11} are expected numbers of reads containing minor allele in the first position and major in the second, major allele in the first position and minor in the second, major alleles in both position, respectively. To determine if the minor alleles in positions I and J are linked we need to estimate the probability that the observed counts of 2-haplotypes $O_{11}, O_{12}, O_{21}, O_{22}$ in the reads covering I and J are produced by counts satisfying (1).

Let n be the total number of reads covering both positions I and J. Then

$$p = \frac{O_{21} \cdot O_{12}}{O_{11} \cdot n} \tag{2}$$

is the largest probability of observing the 2-haplotypes (22) among these n reads given that the variant (22) does not exist. It has been shown in [3] that after Bonferroni correction to multiple testing the value of p should satisfy the following inequality

$$1 - \sum_{i=0}^{O_{22}-1} \binom{n}{i} p^i (1-p)^{n-i} \leq \frac{\mathcal{P}}{\binom{L}{2}} \qquad (3)$$

where $\mathcal{P}$ is the user-defined P-value, by default $\mathcal{P} = 0.01$. Note that we compute the probability of existence of a 2-haplotype with minor alleles rather than probability of observing such 2-haplotype.

Let M be a $L \times N$ matrix of a multiple sequence alignment of all reads, where L is the length of the reference, and N is the number of reads. CliqueSNV splits each column of M into 4 columns each corresponding to one of the 4 minor alleles (including deletions). Let $4M$ be the resulted matrix. Since the average minor allele frequencies (MAF) λ are usually small ($\lambda < 3\%$), the matrix $4M$ is very sparse and can be compactly represented as follows – each row is represented by a sorted list of all columns with minor alleles, and each column is represented by a sorted list of all rows with minor alleles.

For each pair of columns I and J, we find O_{22}^{IJ}, the number of common rows with minor alleles. The compact data structure for the sparse matrix $4M$ only requires at total runtime of $O((\lambda L)(\lambda N)L)$. This gives us 1000x computational acceleration compared with the straightforward $O(L^2N)$ for $\lambda \approx 0.03$. If O_{22}^{IJ} is large enough (the default threshold is 0.03% of coverage), then the remaining statistics of O_{11}^{IJ}, O_{21}^{IJ}, and O_{12}^{IJ} are computed for the corresponding pair of columns I and J. For each such pair of columns we the calculate p-value according to (3) and determine if these alleles are linked. Note that the number of I, J pairs with large enough O_{12}^{IJ} is relatively small and the total time to compute the p-value is proportional to the O_{22} computation.

2: Constructing the allele graph. The allele graph $G = (V, E)$ consists of vertices corresponding to minor alleles and edges corresponding to linked pairs of minor alleles from different positions. There are no isolated vertices in G since the minor alleles are only considered if they are linked to other minor alleles. If the intra-host population consists of very similar haplotypes, then the graph G is very sparse. Indeed, the PacBio dataset for Influenza A virus encompassing $L = 2500$ positions is split into 10000 vertices while the allele graph contains only 700 edges, and, similarly, the simulated Illumina read data set for the same haplotypes contains only 368 edges.

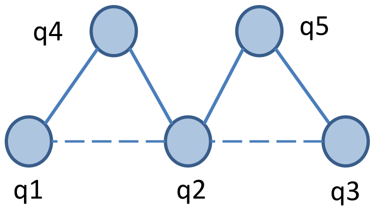

Fig. 1. The clique graph C_G with 5 vertice corresponding to cliques in G, 4 edges and two forbidden pairs (q_1, q_2) and (q_2, q_3). There 3 maximal connected subgraphs avoiding forbidden pairs: $\{q_1, q_4\}$ $\{q_4, q_2, q_5\}$ $\{q_5, q_3\}$

Note that the isolated minor alleles correspond to genotyping errors unless they have a significant frequency. This fact allows us to estimate the number of errors per read assuming that all isolated alleles are errors. As expected, the distribution of the PacBio reads has a heavy tail which implies that most reads are (almost) error-free while a small number of heavy-tail reads accumulate most of the errors. Our analysis allows the identification of such reads which can then be filtered out. By default, we filter out $\approx 10\%$ of PacBio reads but we do not filter out any Illumina reads. The allele graph is then constructed for the reduced set of reads. Such filtering allows the reduction of systematic errors and refines the allele graph significantly.

3: Finding cliques in the allele graph G. Ideally, the individual minor alleles distinguishing a viral haplotype from the consensus should all be pairwise adjacent to the allele graph $G = (V, E)$. Therefore, CliqueSNV looks for maximal cliques in G. Although the MAX CLIQUE is a well-known NP-complete problem and there may be an exponential number of maximal cliques in G, a standard Bron - Kerbosch algorithm requires little computational time since G is very sparse [9].

Unfortunately, cliques corresponding to individual viral haplotypes frequently miss edges. Indeed, the short span of Illumina reads results in sparsity or even complete absence of coverage of paired SNV positions. Similarly, PacBio reads have lower coverage of the ends of the sequencing region resulting in missing links between SNVs from the opposite ends. Therefore, it is necessary to merge cliques into larger cliques corresponding to true haplotypes simultaneously avoiding over-merging cliques corresponding to different haplotypes.

4: Merging cliques in the clique graph C_G with forbidden pairs. We first find all pairs of cliques p and q which are unlikely to come from the same haplotype. For each pair of positions respectively in p and q, we check whether they share sufficiently many reads (default is 50) and P-value (3) is large enough (by default $P > 0.1$), i.e., it is extremely unlikely that the minor alleles are in the same haplotype. If this is the case, we say that p and q form a *forbidden pair*. If p and q do not form a *forbidden pair*, then we check if it is likely that they come from the same haplotype, namely, if there exists at least one edge in G between a pair of positions in p and q. In this case, we say that p and q are adjacent in the clique graph C_G with vertices representing cliques. Any true haplotype corresponds to a maximal connected subgraph of C_G that does not contain a forbidden pair (see Fig. 1).

Unfortunately, even deciding whether there is a p-q-path avoiding forbidden pairs is known to be NP-hard [21]. We solve the problem of finding all maximal connected subgraphs without forbidden pairs for C_G as follows: we connect all pairs of vertices except forbidden pairs obtaining a graph C'_G, find all maximal super-cliques (i.e., cliques in C'_G) using [9], split each super-clique into connected components in C_G, and filter out the connected components which are proper subsets of other maximal connected components.

5: Finding consensus viral variants for merged cliques. Let S be the set of all positions that belong to at least one clique. Let q_S be an *empty clique* corresponding to a haplotype with all major alleles in S. For each read r restricted to the positions in S, we assign r to the closest clique q (which can be q_S), i.e. clique q which differs from r in the minimum number of positions in S. In case of a tie, we assign r to all closest cliques.

For each clique q, CliqueSNV finds the consensus $v(q)$ of all restricted reads assigned to q. Then $v(q)$ is extended from S to a full-length haplotype by setting all non-S positions to major alleles.

6: Expectation-maximization (EM) Algorithm for Estimating Variant Frequencies. We estimate the frequencies of the reconstructed viral variants via an EM algorithm (see, e.g., IsoEM [28]).

3 RESULTS

We compared the performance of CliqueSNV with state-of-the-art haplotyping methods SAVAGE [5], aBayesQR [1], PredictHaplo [30], and 2SNV [2] on two experimental and two simulated NGS datasets. Below we describe the datasets, the validation metrics, and comparison results.

Benchmarks

IAV PacBio data (IAV_Pacbio) [2]. In the experimental data (see [2]) 10 influenza A virus clones were mixed with the frequency range 0.1%- 50%. The Hamming distances between clones was in the range 0.1-1.1%(2-22 bp difference). The 2kb-long amplicon was sequenced using the PacBio platform, yielding a total of 33,558 reads of an average length 1973 nucleotides.

Simulated IAV Illumina MiSeq data (IAV_Sim_MiSeq). The IAV clones and their frequencies are the same as in IAV_Pacbio dataset. 10K coverage by paired Illumina MiSeq platform was simulated using SimSeq [8].

Simulated HIV Illumina MiSeq data (HIV_Sim_MiSeq). This benchmark contains simulated Illumina MiSeq reads with 10k-coverage of gag/polymerase *(pol)* region of length 1kb which includes important drug-resistant mutations. The reads were simulated from seven equally distributed HIV-1 variants chosen from the NCBI database with Hamming distances between clones in the range from 0.6-3.0%(6 to 30 bp differences).

Reduced HIV Illumina lab mixture (Reduced_labmix) [16]. Illumina MiSeq (2×250-bp) data set with an average read coverage of 20,000×, obtained from a lab mixture of five HIV-1 strains with pairwise Hamming distance in the range from 2-3.5%(27 to 46 bp difference). The original sequencing length was 9.3Kb, but was reduced to the same *gag/pol*-region of length 1.3Kb.

Validation via Earth Mover's Distance

Validation of different haplotype reconstruction methods should simultaneously answer two general questions: (i) how close are the reconstructed and true variants and (ii) how narrow is the reconstructed and true variant frequency distribution. Previous studies report high variation in results addressing these questions likely due to the challenge of simultaneously addressing them. Here we propose to use the Earth Mover's Distance (EMD) [29] as a distance measure for populations, which generalizes edit distances between genomes of individual variants.

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

Comparison of Haplotype Reconstruction Methods

Tables 1 and 2 describe the datasets (true variant ID's and their frequencies) and report for each true variant T the quality of its prediction: the edit distance to the closest predicted variant (ECP), the frequency of the closest predicted variant (FCP) and the explanation error of T (EEV). In row EMD, we report the EMD distance from the population of the true variants to the read consensus (underscored) and to the population of variants predicted by the corresponding method. Note that the EMD to the read consensus is a measure of the benchmark diversity. CliqueSNV is intended to work with a population of closely related genetic variants which are expected to be in a single patient sample.

We compare only three methods (CliqueSNV, 2SNV, and PredictHaplo) on the IAV_PacBio benchmark since the other two methods can only use Illumina reads (see Table 1). CliqueSNV managed to correctly recover all 10 true variants including Clone8 whose frequency is significantly below the error rate. 2SNV was able to recover 9 true variants but reported one false positive. PredictHaplo recovered only 7 true variants. In addition, we created low-coverage datasets by randomly subsampling $n = 16K, 8K, 4K$ reads from the original dataset. For each dataset, CliqueSNV found at least one true variant more than both 2SNV and PredictHaplo.

We compare four methods (CliqueSNV, SAVAGE, PredictHaplo, and aBayesQR) on three Illumina benchmarks (see Table 2). Note that SAVAGE results are not fully comparable with other methods since SAVAGE (with the reference option) reports contigs rather than complete haplotypes. Therefore, when finding edit distance to closest predicted haplotype, it is necessary to decide how to count the uncovered positions in the true variant. We do not count uncovered position as mismatches and report ECP and EEV, which are significantly underestimated for SAVAGE. Table 2 shows that CliqueSNV outperforms all other methods on two simulated benchmarks and better than PredictHaplo and aBayesQR on the remaining datasets. For IAV, it reconstructs 7 IAV haplotypes without mismatches and 3 haplotypes with a single mismatch, and accurately identifies all haplotypes for the HIV_Sim_MiSeq dataset. The distances between variants from the dataset Reduced_labmix are significantly higher than expected from the real HIV population sampled from a single host [43], resulting in a high EMD to read a consensus of 19.4. Such populations are more difficult to handle by CliqueSNV. Nonetheless, for that benchmark CliqueSNV outperformed all other tools and reconstructed three variants without errors.

Runtime

For all experiments, we used the same PC (Intel(R) Xeon(R) CPU X5550 2.67GHz x2 8 cores per CPU, DIMM DDR3 1333 MHz RAM 4Gb x12) with the CentOS 6.4 operating system. The runtime of CliqueSNV is sublinear with respect to the number of reads while the runtime of PredictHaplo and 2SNV exhibit super-linear growth. For 33k PacBio reads CliqueSNV needs 21 seconds, while PredictHaplo and 2SNV require around 30 minutes. The runtime of CliqueSNV is quadratic with respect to the number of SNV rather than the length of the sequencing region. We generated five HIV variants within 1% Hamming distance from each other, which is the expected distance between related HIV variants from the same person [44]. Then we simulated 1M Illumina reads for sequencing regions of length 566, 1132, 2263 and 9181 for which CliqueSNV required 37, 144, 227, 614 seconds, respectively. CliqueSNV is significantly faster than the other tools in our study. For the Reduced_labmix benchmark the runtimes of aBayesQR and SAVAGE were more than 10h, PhedictHaplo's runtime was 24 min, and CliqueSNV took 79 seconds.

4 CONCLUSIONS

Reconstruction of quasispecies populations from noisy sequencing data is one of the most challenging problems of computational genomics. High-throughput sequencing technologies, such as Illumina MiSeq and HiSeq, provide deep coverage, but short reads require assembly of unknown numbers of closely related haplotypes of various frequencies. Furthermore, the reads from these instruments contain a significant amount of sequencing errors with frequencies comparable with true minor mutations [36]. The recent development of single-molecule sequencing platforms such as PacBio produces reads that are sufficiently long to span entire genes or small viral genomes. However, the sequencing error rate of single-molecule sequencing is exceptionally high and could reach $13 - 14\%$ [32], which hampers its ability to reconstruct rare viral variants.

We developed CliqueSNV, a new method for inference of rare genetically-related viral variants, which allows for accurate haplotyping in the presence of high sequencing error rates and which is also suitable for both single-molecule and short-read sequencing. CliqueSNV infers viral haplotypes by detection of clusters of statistically linked SNVs rather than through assembly of overlapping reads. Using experimental data, we demonstrate that CliqueSNV can detect haplotypes with frequencies as low as 0.1%, which is comparable to the sensitivity of many deep sequencing-based point mutation detection methods [14, 18]. Furthermore, CliqueSNV can successfully infer viral variants, which differ by only a few mutations, thus demonstrating the high sensitivity of identifying closely related variants. Another significant advantage of CliqueSNV is its low computation time, which is achieved by fast searching of linked pairs of SNVs and the application of the special graph-theoretical approach to SNV clustering.

Besides the aforementioned advantages, CliqueSNV has its limitations. Unlike Savage [5], it is not a *de novo* assembly tool and requires a reference viral genome. This obstacle could be addressed by using Vicuna [46] or other analogous tools to assemble a consensus sequence, which can be used as a reference. Another limitation consists in the possibility that the variants which differ only by isolated SNVs separated by long conserved genomic regions longer than the read length may not be accurately inferred. Such situations usually do not occur for viruses, where mutations are densely concentrated in different genomic regions. We are planning to address this problem in the next version of CliqueSNV.

The ability to accurately infer the structure of intra-host viral populations makes CliqueSNV applicable for studying evolution and examining genomic compositions in RNA viruses. However, we envision that the application of our method can be extended to other highly heterogeneous genomic populations, such as metagenomes, immune repertoires, and cancer cells.

5 ACKNOWLEDGEMENTS

A.Z. has been partially supported by NSF Grants DBI-1564899 and CCF-1619110 and NIH Grant 1R01EB025022-01. P.S. has been partially supported by NIH Grant 1R01EB025022-01. S.K. and A.M. has been partially supported by MBD.

REFERENCES

[1] 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*, pages 353–369. Springer, 2017.

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

[3] 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. *Journal of Computational Biology*, 24(6):558–570, 2017.

[4] 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*, 12(Suppl 6):S1, 2011.

[5] J. A. Baaijens, A. Z. El Aabidine, E. Rivals, and A. Schönhuth. De novo assembly of viral quasispecies using overlap graphs. *Genome Research*, 27(5):835–848, 2017.

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

[7] 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*, 21:3943–3950, 2005.

[8] S. Benidt and D. Nettleton. Simseq: a nonparametric approach to simulation of rna-sequence datasets. *Bioinformatics*, 31(13):2131–2140, 2015.

[9] C. Bron and J. Kerbosch. Algorithm 457: Finding all cliques of an undirected graph. *Commun. ACM*, 16(9):575–577, Sept. 1973.

Table 1. Comparison of three haplotype reconstruction methods on real PacBio data

IAV_Pacbio		CliqueSNV			2SNV			PredictHaplo		
Variant	**TF, %**	**ECP**	**FCP, %**	**EEV**	**ECP**	**FCP, %**	**EEV**	**ECP**	**FCP, %**	**EEV**
fv3	50	0	52.6	0	0	51.8	0	0	56.7	0
Clone1	25	0	23.7	0.51	0	23.7	0.51	0	23.7	0.62
Clone2	12.5	0	12.6	0	0	12.5	0.04	0	13.7	0
flu1-Dmut	6.26	0	6.41	0	0	6.39	0	0	6.01	0.36
Clone3	3.13	0	2.32	2.13	0	2.3	2.13	0	3.01	0
fv2	1.56	0	1.17	0.5	0	1.19	0.48	2	56.7	9.57
Clone4	0.78	0	0.69	0.89	0	0.7	0.84	0	2.9	0
Clone6	0.39	0	0.35	0.92	0	0.34	1.13	0	1.2	0
Clone7	0.19	0	0.12	2.56	0	0.12	2.56	7	56.7	13
Clone8	0.1	0	0.05	5.79	12	1	12	12	56.7	17
EMD	4.22	0.22			0.23			0.38		
# variants	10	10			11			7		

TF = true frequency, ECP = editing distance to the closest predicted variant, FCP = frequency of the closest predicted variant, EEV = explanation error for the true variant

Table 2. Comparison of four haplotype reconstruction methods on simulated and real Illumina data

IAV_Sim_MiSeq		CliqueSNV			SAVAGE			PredictHaplo			aBayesQR		
Variant	**TF, %**	**ECP**	**FCP, %**	**EEV**	**ECP**	**FCP, %**	**EEV**	**ECP**	**FCP, %**	**EEV**	**ECP**	**FCP, %**	**EEV**
fv3	50	0	50.13	0	0*	0.1	0.844	0	76.3	0	1	35.2	2.35
Clone1	25	0	24.91	0.0493	0*	0.1	0.213	4	18.5	5.31	1	14	3.2
Clone2	12.5	0	12.43	0.07	0*	0.1	0.059	6	5.27	8.89	6	8.11	9
flu1-Dmut	6.25	1	6.3	1	0*	0.1	0.073	3	76.3	3	2	35.2	3.89
Clone3	3.13	0	3.12	0.0132	0*	0.1	0.063	8	76.3	8	0	4.24	0
fv2	1.56	0	1.6	0	0*	0.1	0.143	2	76.3	2	3	35.2	6
Clone4	0.78	1	0.78	1.014	0*	0.1	0	8	76.3	8	9	35.2	13.37
Clone6	0.39	0	0.41	0	0*	0.1	0	8	76.3	8	9	35.2	14
Clone7	0.19	1	0.2	1	0*	0.1	0	7	76.3	7	8	35.2	13
Clone8	0.1	0	0.1	0	0*	0.1	0	12	76.3	12	13	35.2	16
EMD	4.22	0.0939			0.492**			3.03			3.64		

HIV_Sim_MiSeq		CliqueSNV			SAVAGE			PredictHaplo			aBayesQR		
Variant	**TF, %**	**ECP**	**FCP, %**	**EEV**	**ECP**	**FCP, %**	**EEV**	**ECP**	**FCP, %**	**EEV**	**ECP**	**FCP, %**	**EEV**
AY835778	14.3	0	14.3	0	0*	23.4	0	7	39	7	1	14.4	1
AY835770	14.3	0	14.3	0	0*	7.9	0	5	28.7	5	1	15.1	1
AY835771	14.3	0	14.3	0	0*	2.1	0.13	1	28.7	1	1	12.1	2.85
AY835777	14.3	0	14.3	0.02	0*	6.7	0.57	2	39	2	1	15.5	1
AY835763	14.3	0	14.3	0	0*	6.1	6.06	3	32.3	3	0	14.3	0
AY835762	14.3	0	14.2	0.1	0*	2	10.3	10	32.3	10	0	14.4	0
AY835757	14.3	0	14.3	0.004	7*	0.9	14.9	12	39	13.1	0	14.2	0.05
EMD	11	0.018			4.56**			5.87			0.84		

Reduced_labmix		CliqueSNV			SAVAGE			PredictHaplo			aBayesQR		
Variant	**TF, %**	**ECP**	**FCP, %**	**EEV**	**ECP**	**FCP, %**	**EEV**	**ECP**	**FCP, %**	**EEV**	**ECP**	**FCP, %**	**EEV**
89.6	20	0	12.5	10.8	0*	1.1	2.22	0	21.8	0	18	9.94	20.8
HXB2	20	5	6.9	11	0*	3.4	1.68	22	22.7	29	15	9.08	23.1
JRCSF	20	1	7.55	6.58	0*	0.4	0.55	0	29	0	14	8.16	14.6
NL43	20	0	16.9	6.62	0*	0.2	0.16	0	26.6	0	16	7.36	16.6
YU2	20	0	10.8	5.13	0*	0.7	2.27	5	22.7	5	19	7.36	21
EMD	19.4	6.52			1.37**			6.8			19.2		

TF = true frequency, ECP = editing distance to the closest predicted variant, FCP = frequency of the closest predicted variant, EEV = explanation error for the true variant. The underscored value is the EMD distance to the population consisting of a single variant coinciding with the read consensus. *The ECP value for SAVAGE is significantly underestimated since it does not generally reconstruct full haplotypes. **The EMD distance for SAVAGE is significantly underestimated.

[10]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*, 95(6):627–635, 2014.

[11]E. Domingo, J. Sheldon, and C. Perales. Viral quasispecies evolution. *Microbiology and Molecular Biology Reviews*, 76(2):159–216, 2012.

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

[13]J. W. Drake and J. J. Holland. Mutation rates among rna viruses. *Proceedings of the National Academy of Sciences*, 96(24):13910–13913, 1999.

[14]P. Flaherty, G. Natsoulis, O. Muralidharan, M. Winters, J. Buenrostro, J. Bell, S. Brown, M. Holodniy, N. Zhang, and H. P. Ji. Ultrasensitive detection of rare mutations using next-generation targeted resequencing. *Nucleic Acids Res.*, 40(1):e2, Jan 2012.

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

[16]F. D. Giallonardo, A. Tpfer, M. Rey, S. Prabhakaran, Y. Duport, C. Leemann, S. Schmutz, N. K. Campbell, B. Joos, M. R. Lecca, A. Patrignani, M. Dumer, C. Beisel, P. Rusert, A. Trkola, H. F. Gnthard, V. Roth, N. Beerenwinkel, and K. J. Metzner. Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. *Nucleic Acids Research*, 42(14):e115, 2014.

[17]B. Hajarizadeh, J. Grebely, and G. J. Dore. Epidemiology and natural history of hcv infection. *Nature Reviews Gastroenterology and Hepatology*, 10(9):553–562, 2013.

[18]O. Harismendy, R. B. Schwab, L. Bao, J. Olson, S. Rozenzhak, S. K. Kotsopoulos, S. Pond, B. Crain, M. S. Chee, K. Messer, D. R. Link, and K. A. Frazer. Detection of low prevalence somatic mutations in solid tumors with ultra-deep targeted sequencing. *Genome Biol.*, 12(12):R124, 2011.

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

[20]P. H. Kilmarx. Global epidemiology of hiv. *Current Opinion in HIV and AIDS*, 4(4):240–246, 2009.

[21]J. Kováč. Complexity of the path avoiding forbidden pairs problem revisited. *Discrete Appl. Math.*, 161(10-11):1506–1512, July 2013.

[22]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*, 380(9859):2095–2128, 2012.

[23]M. E. M, J. McCaskill, and P. Schuster. The molecular quasi-species. *Adv Chem Phys*, 75:149–263, 1989.

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

[25]N. Mancuso, B. Tork, P. Skums, L. Ganova-Raeva, I. Măndoiu, and A. Zelikovsky. Reconstructing viral quasispecies from ngs amplicon reads. *In silico biology*, 11(5):237–249, 2011.

[26]S. Mangul, N. C. Wu, N. Mancuso, A. Zelikovsky, R. Sun, and E. Eskin. Accurate viral population assembly from ultra-deep sequencing data. *Bioinformatics*, 30(12):i329–i337, 2014.

[27]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*, pages 3225–3229, 1992.

[28]M. Nicolae, S. Mangul, I. Mandoiu, and A. Zelikovsky. Estimation of alternative splicing isoform frequencies from rna-seq data. *Algorithms for Molecular Biology*, 6:9, 2011.

[29]S. Peleg, M. Werman, and H. Rom. A unified approach to the change of resolution: Space and gray-level. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):739–742, 1989.

[30]S. Prabhakaran, M. Rey, O. Zagordi, N. Beerenwinkel, and V. Roth. HIV haplotype inference using a propagating Dirichlet process mixture model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 11(1):182–191, 2014.

[31]M. C. Prosperi and M. Salemi. Qure: software for viral quasispecies reconstruction from next-generation sequencing data. *Bioinformatics*, 28(1):132–133, 2012.

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

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

[34]F. Rodriguez-Frias, M. Buti, D. Tabernero, and M. Homs. Quasispecies structure, cornerstone of hepatitis b virus infection: mass sequencing approach. *World J Gastroenterol*, 19(41):6995–7023, 2013.

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

[36]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.

[37]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*, 112(21):6653–6658, 2015.

[38]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*, 13(S-10):S6, 2012.

[39]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*, 14(Suppl 9):S2, 2013.

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

[41]A. Töpfer, T. Marschall, R. A. Bull, F. Luciani, A. Schönhuth, and N. Beerenwinkel. Viral quasispecies assembly via maximal clique enumeration. *PLoS Computational Biology*, 10(3), 2014.

[42]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*, 20(2):113–123, 2013.

[43]J. O. Wertheim, A. J. L. Brown, N. L. Hepler, S. R. Mehta, D. D. Richman, D. M. Smith, and S. L. K. Pond. The global transmission network of hiv-1. *Journal of Infectious Diseases*, 209(2):304–313, 2014.

[44]J. O. Wertheim, A. J. Leigh Brown, N. L. Hepler, S. R. Mehta, D. D. Richman, D. M. Smith, and S. L. Kosakovsky Pond. The global transmission network of hiv-1. *The Journal of Infectious Diseases*, 209(2):304–313, 2014.

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

[46]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*, 13(1):475, 2012.

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

[48]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*, 12(1):119, 2011.

[49]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, pages 342–354, Berlin, Heidelberg, 2012. Springer-Verlag.