Accurate assembly of minority viral haplotypes from next-generation sequencing through efficient noise reduction

Sergey Knyazev 1,2,3 *, Viachaslau Tsyvina 1, Anupama Shankar 2, Andrew Melnyk 1, Alexander Artyomenko 1, Tatiana Malygina 7, Yuri B. Porozov 5,7,8, Ellsworth M. Campbell 2, Serghei Mangul 6, William M. Switzer 2, Pavel Skums 1, and Alex Zelikovsky 1,5 †

1 Department of Computer Science, Georgia State University, Atlanta, GA, 30302, USA, 2 Division of HIV/AIDS Prevention, Centers for Disease Control and Prevention, Atlanta, GA, 30333, USA, 3 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, I.M. Sechenov First Moscow State Medical University, Moscow, 119991, Russia, 6 lab of Clinical Pharmacy, School of Pharmacy, University of Southern California, Los Angeles, CA, 90033, USA, 7 The laboratory of bioinformatics, ITMO University, St. Petersburg, 197101, Russia and 8 Department of Computational Biology, Sirius University of Science and Technology, Sochi, 354340, Russia

Received YYYY-MM-DD; Revised YYYY-MM-DD; Accepted YYYY-MM-DD

ABSTRACT
Rapidly evolving RNA viruses continuously produce minority haplotypes that can become dominant if they are drug-resistant or can better evade the immune system. Therefore, early detection and identification of minority viral haplotypes may help to promptly adjust the patient’s treatment plan preventing potential disease complications. Minority haplotypes can be identified using next-generation sequencing (NGS), but sequencing noise hinders accurate identification. The elimination of sequencing noise is a non-trivial task that still remains open. Here we propose CliqueSNV based on extracting pairs of statistically linked mutations from noisy reads. This effectively reduces sequencing noise and enables identifying minority haplotypes with the frequency below the sequencing error rate. We comparatively assess the performance of CliqueSNV using an in vitro mixture of nine haplotypes that were derived from the mutation profile of an existing HIV patient. We show that CliqueSNV can accurately assemble viral haplotypes with frequencies as low as 0.1% and maintains consistent performance across short and long bases sequencing platforms.

INTRODUCTION
Rapidly evolving RNA viruses such as human immunodeficiency virus (HIV), hepatitis C virus (HCV), influenza A virus (IAV), SARS, and SARS-CoV-2 form populations of closely related genomic variants inside infected hosts (1, 2, 3, 4, 5, 6, 7, 8, 9, 10). The intra-host viral populations include minority viral variants that are frequently responsible for drug resistance, immune escape, and disease transmission (11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23). Therefore, accurately predicting minority viral populations from extremely large and noisy viral genomic data is important for biomedical research, epidemiology, and clinical applications. Although this problem has recently attracted significant interest from the biomedical research community (24, 25, 26), numerous obstacles still delay NGS integration into the viral studies. The last decade witnessed numerous attempts to employ NGS and bioinformatics methods for reconstructing intra-host viral populations. These methods are not accurate enough for clinical and epidemiological applications since they cannot reliably identify haplotypes accounting for a substantial portion of the population. Existing methods are ill-equipped to assemble closely related haplotypes and have elevated false-positive rates. Additionally, there is only one in vitro viral sequencing benchmark for validation of haplotyping tools (26), and to convincingly demonstrate that such tools are ready for clinical and epidemiological applications, new comprehensive sequencing benchmarks are urgently required (27).

Next-generation sequencing (NGS) technologies now provide versatile opportunities to study viral populations. 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. Ability of NGS technologies to efficiently identify minority variants have recently gained FDA approval (28). However, haplotyping of heterogeneous viral populations (i.e., assembly of full-

*equal contribution
†To whom correspondence should be addressed. SK: sergey.n.knyazev@gmail.com AZ: Tel: +1 404 4135730; Fax: +1 404 4135717; Email: alexz@gsu.edu

© YYYY The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
length genomic variants and estimation of their frequencies) is extremely 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 full-length viral variants to be sequenced in a single pass. However, the high level of sequence noise due to background or platform-specific sequencing errors produced by all currently available platforms makes inference of low-frequency genetically close variants especially challenging, since it is required to distinguish between real and artificial genetic heterogeneity produced by sequencing errors.

Recently, a number of computational tools for inference of viral quasispecies populations from NGS reads have been proposed (27), including Savage (24), PredictHaplo (29), aBayesQR (30), QuasiRecomb (31), HaploClique (32), VGA (33), VirA (34, 35), SHORAH (36), ViSpA (37), QUET (38) and others (39, 40, 41, 42, 43). Even though these algorithms proved useful in many applications, accurate and scalable viral haplotyping remains a challenge. In particular, inference of low-frequency viral variants is still problematic, while many computational tools designed for the previous generation of sequencing platforms have severe scalability problems when applied to datasets produced by state-of-the-art technologies.

Previously, several tools such as V-phaser (44), V-phaser2 (45) and CoVaMa (46) exploited linkage of mutations for single nucleotide variant (SNV) calling rather than haplotype assembly, but they do not accommodate sequencing errors when deciding whether two variants are linked. These tools are also unable to detect the frequency of mutations above sequencing error rates (47). The 2SNV algorithm (48) accommodates errors in links and was the first such tool to be able to correctly detect haplotypes with a frequency below the sequencing error rate.

We propose a novel method that can accurately identify minority haplotypes from NGS reads consisting of three steps. First, we extract pairs of statistically linked mutations. Second, we find maximal sets of pairwise linked mutations (cliques) where each clique corresponds to a set of mutations in a minority haplotype. Finally, we assign each read to the closest clique, and for each clique, we form a haplotype as a consensus of reads assigned to it.

All haplotyping tools require solid and convincing validation benchmarks (49, 50). The true viral variants and their distribution are only known for simulated data (51), but sequencing errors, variation of coverage depth, PCR bias, and systematic noise cannot be reliably simulated. Therefore experimental sequencing benchmarks that provide an adequate evaluation of haplotyping tools are necessary.

By now, there are only two experimental sequencing benchmarks -- (i) Illumina sequencing reads consisting of a mixture of five HIV-1 strains (HIV5exp, see Table 1) (52) and (ii) PacBio sequencing reads from a sample consisting of ten IAV viral variants (IAV10exp, see Table 1) (48). In the HIV5exp, five different HIV-1 strains each having 20% frequency were prepared to mimic an intra-host viral population. Unfortunately, this benchmark is not realistic enough since the observed intra-host viral populations consist of variants that are much closer to each other than different strains and contain both frequent and rare variants (53). The IAV10exp benchmark significantly better mimics the intra-host viral population since its variants are very similar to each other and the variant frequencies are realistically non-uniform. Thus, similar to the IAV10exp benchmark, it would be beneficial to develop Illumina benchmarks which adequately imitate intra-host viral populations containing closely related minority variants.

To validate our method’s performance, we have introduced two novel in vitro sequencing HIV-1 benchmarks, which consist of Illumina MiSeq experiments on haplotype mixtures based on the mutation profile from an existing patient.

Finally, there is a essential gap in existing quality measures of intra-host viral population assembly. Up-to-date, instead of populations (i.e. haplotypes with their frequencies), only sets of reconstructed and the ground truth haplotypes are compared (29). Here we propose to measure differences between haplotype populations using Matching Error and the Earth Mover’s Distance which account for both the distances between haplotypes and their frequencies.

MATERIALS AND METHODS

CliqueSNV algorithm idea

A schematic diagram of the CliqueSNV algorithm is shown in Figure 1. The algorithm takes aligned reads as input and infers haplotype sequences with their frequencies as output. The method consists of six steps:

- Step 1 uses aligned reads to build the consensus sequence and identifies all SNVs. Then all pairs of SNVs are tested for dependency and are then divided into three groups: linked, forbidden, or unclassified. Each SNV is represented as a pair (p, n) of its position p and nucleotide value n in the aligned reads. If there are enough reads that have two SNVs (p, n) and (p', n') simultaneously, then they are tested for dependency. If the dependency test is positive and statistically significant (see CliqueSNV algorithm details for more information), then the algorithm classifies these two SNVs as linked. Otherwise, these two SNVs are tested for independence. If the independency test is positive and statistically significant (see Detailed description for details), then these two SNVs are classified as a forbidden pair.

- In Step 2, we build a graph $G = (V, E)$ with a set of nodes $V$ representing SNVs, and a set of edges $E$ connecting linked SNV pairs.

- Ideally, SNVs of each true minority haplotype form a clique in $G$. A maximal clique $C \subseteq V$ is a set of nodes such that $(u, v) \in E$ for any $u, v \in C$ and for any $x \notin C$ there is $u \in C$ such that $(x, u) \notin E$. Step 3 finds all maximal cliques in $G$.

- For real sequencing data, the linkage between some SNV pairs may be undetected due to sequencing noise, uneven coverage, or the shortness of the NGS reads. As a result, a single clique corresponding to a haplotype will be split into several overlapping cliques. Step 4
Figure 1. Schematic representation of the CliqueSNV algorithm. Where SNV is single nucleotide variation.

merges such overlapping cliques. In order to avoid merging distinct haplotypes, two cliques are not merged if they contain a forbidden SNV pair.

- Step 5 assigns each read to a merged clique with which it shares the largest number of SNVs. Then CliqueSNV builds a consensus haplotype from all reads assigned to a single merged clique.

- Finally, haplotype frequencies are estimated via an expectation-maximization algorithm in Step 6.
### Intra-host viral population sequencing benchmarks

We tested the ability of CliqueSNV to assemble haplotype sequences and estimate their frequencies from PacBio and MiSeq reads using four real (experimental) and two simulated datasets from HIV and Influenza A virus (IAV) samples (Table 1). Each dataset contains between two to ten haplotypes with frequencies of 0.1 to 50%. The Hamming distances between pairs of variants for each dataset are shown in Figure S1.

#### Experimental datasets:

1–2. **HIV-1 subtype B plasmid mixtures and MiSeq reads (HIV2exp and HIV9exp).** We designed nine in silico plasmid constructs comprising a 950-bp region of the HIV-1 subtype B polymerase (pol) gene that were then synthesized and cloned into pUCIDT-Amp (Integrated DNA Technologies, Skokie, IL). Each clone was confirmed by Sanger sequencing. This 950-bp region at the beginning of pol contains known protease and reverse transcriptase genes that are monitored for drug-resistant mutations and is monitored with sequence analysis for patient care. Each of these plasmids contains a specific set of point mutations chosen using mutation profiles of patient p7 from a real clinical study (53) to create nine unique synthetic HIV-1 pol haplotypes. Different proportions of these plasmids were mixed and then sequenced using an Illumina MiSeq protocol to obtain 2x300-bp reads (see Supplementary Methods). HIV2exp and HIV9exp are mixtures of two and nine variants, respectively.

3. **HIV-1 subtype B mixture and MiSeq reads (HIV5exp).** This dataset consists of Illumina MiSeq 2 × 250-bp reads with an average read coverage of ∼20,000× obtained from a mixture of five HIV-1 isolates: 89.6, HXB2, JRCSF, NL4.3, and YU2 available at (52). Isolates have pairwise Hamming distances in the range from 2.3.5% (27 to 46-bp differences). The original HIV-1 sequence length was 9.3kb, but was reduced to the beginning of pol with a length of 1.3kb.

4. **Influenza A mixture and PacBio reads (IAV10exp).** This benchmark contains ten IAV virus clones that were mixed at a frequency of 0.1-50%. The Hamming distances between clones ranged from 0.1-1.1% (2-22-bp differences) (48). The 2kb-amplicon was sequenced using the PacBio platform yielding a total of 33,558 reads with an average length of 1973 nucleotides.

#### Simulated datasets:

1. **HIV-1 subtype B mixture and MiSeq reads (HIV7sim).** This benchmark contains simulated Illumina MiSeq reads with a 10k-coverage of 1-kb pol sequences. The reads were simulated from seven equally distributed HIV-1 variants chosen from the NCBI database: AY835778, AY835770, AY835771, AY835777, AY835763, AY835762, and AY835757. The Hamming distances between clones are in the range from 0.6-3.0% (6 to 30-bp differences). We used SimSeq (54) for generating reads.

2. **Influenza A mixture and MiSeq reads (IAV10sim).** This benchmark contains simulated IAV Illumina MiSeq reads with the same IAV haplotypes and their frequencies as for the IAV10exp benchmark. The sequencing of a 2kb-amplicon with 40k coverage with paired Illumina MiSeq reads was simulated by SimSeq (54) with the default sequencing error profile in SimSeq.

### Validation metrics for viral population inference

**Precision and recall** Inference quality is typically measured by precision and recall.

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

where \(TP\) is the number of true predicted haplotypes, \(FP\) is the number of false predicted haplotypes, and \(FN\) is the number of undiscovered haplotypes.

Initially we measured precision and recall strictly by treating a predicted haplotype with a single mismatch as an \(FP\). Additionally, like in (29) we introduced an acceptance threshold, which is the number of mismatches permitted for a predicted haplotype to count as a \(TP\).

**Matching errors between populations** However, precision and recall do not take into account (i) distances between true and inferred viral variants as well as (ii) the frequencies of the true and inferred viral variants. Instead, we chose to use analogues of precision and recall defined for populations as follows.

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 population, where \( f_p \) is the frequency of the reconstructed haplotype \( p \). Let \( d_{pt} \) be the distance between haplotypes \( p \) and \( t \). Thus, instead of precision, we used the matching error \( E_{T \rightarrow P} \) which measures how well each reconstructed haplotype \( p \in P \) weighted by its frequency is matched by the closest true haplotype.

\[
E_{T \rightarrow P} = \sum_{p \in P} f_p \min_{t \in T} d_{pt}
\]

Indeed, precision increases while \( E_{T \rightarrow P} \) decreases and reaches 100% when \( E_{T \rightarrow P} = 0 \). Similarly, instead of recall, we propose to use the Earth mover’s distance (EMD) between populations which measures how well each true haplotype \( t \in T \) weighted by its frequency is matched by the closest reconstructed haplotype. (55)

\[
E_{T \rightarrow P} = \sum_{t \in T} f_t \min_{p \in P} d_{pt}
\]

Note that recall increases while \( E_{T \rightarrow P} \) decreases and reaches 100% when \( E_{T \rightarrow P} = 0 \).

The Earth mover’s distance (EMD) between populations is a measure of the dissimilarity between two probability distributions. It is defined as the minimum cost of moving one distribution to another. In the context of viral population, it is estimated as the cost to move each true haplotype to a reconstructed haplotype, weighted by their frequencies.

\[
E_{I \rightarrow O} = \sum_{J \in O} E_{I \rightarrow J} = \frac{E_{11} - E_{21} \cdot E_{12}}{E_{11}}
\]

where \( E_{21}, E_{12}, E_{11} \) are the expected numbers of reads containing the 2-haplotype (21), (12) and (11), respectively. To determine if a pair of SNVs (the minor variants in positions \( I \) and \( J \)) are linked, we need to estimate the probability that the observed counts of 2-haplotype reads covering \( I \) and \( J \) are produced by 2-haplotype counts satisfying equation 1.

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

\[
p = \frac{O_{21} \cdot O_{12}}{O_{11} \cdot n}
\]

is the probability of observing \( O_{22} \) reads with the both minor variants given that the variant (22) does not exist. The 2-haplotype (22) exists with high probability 1 – \( p \) and the corresponding pair of SNVs is linked if the value of \( p \) satisfies the following inequality (59).

\[
1 - \sum_{i=0}^{O_{22}-1} \binom{n}{i} p^i (1-p)^{n-i} \leq P \left( \frac{1}{2} \right)
\]

where \( P \) is the user-defined \( P \)-value (by default \( P = 0.01 \)) and dividing by \( \left( \frac{1}{2} \right) \) is the Bonferroni correction for multiple testing.

CliqueSNV algorithm details

Data input for CliqueSNV consists of PacBio or Illumina reads from an intra-host viral population aligned to a reference genome. Output is the set of inferred viral variant RNA sequences with their frequencies. The formal high-level pseudocode of the CliqueSNV algorithm is described in the supplementary materials. Below we describe in detail the six major steps of CliqueSNV that are schematically presented in Figure 1.

**Step 1: Finding linked and forbidden SNV pairs.** At a given genomic position \( I \), the most frequent nucleotide is referred to as a major variant and is denoted 1. Let us fix one of the less frequent nucleotide (referred to as a minor variant) and denote it 2. A pair of variants at two distinct genomic positions \( I \) and \( J \) is referred to as a 2-haplotype. There are four 2-haplotypes with major and minor variants at \( I \) and \( J \): (11), (12), (21), and (22). Let \( O_{11}, O_{12}, O_{21}, O_{22} \) be the observed counts of 2-haplotype reads covering \( I \) and \( J \). In this step, CliqueSNV tries to decide whether the \( O_{22} \) reads are sequencing errors or they are produced by an existing haplotype containing the 2-haplotype (22).

The pairs of minor variants (referred to as SNV pairs) are classified into three categories: linked, forbidden, and unclassified. An SNV pair is linked if it is highly probable that there exists a haplotype containing both minor variants. On the contrary, an SNV pair is forbidden if it is extremely unlikely that the corresponding minor variants belong to the same haplotype. All other SNV pairs are referred to as unclassified.

Assuming that errors are random, it has been proven in (59) that if the 2-haplotype (22) does not exist, then the expected number of reads \( E_{22} \) containing the 2-haplotype (22) should not exceed

\[
E_{22} \leq \frac{E_{21} \cdot E_{12}}{E_{11}}
\]

where \( E_{21}, E_{12}, E_{11} \) are the expected numbers of reads containing the 2-haplotypes (21), (12) and (11), respectively. To determine if a pair of SNVs (the minor variants in positions \( I \) and \( J \)) are linked, we need to estimate the probability that the observed counts of 2-haplotype reads \( O_{11}, O_{12}, O_{21}, O_{22} \) are produced by 2-haplotype counts satisfying equation 1.
Pairs of SNVs passing this linkage test are classified as a linked SNV pair. For every other pair of SNVs, we check whether they can be classified as a forbidden SNV pair, i.e., whether the probability of observing at most 0.02 reads is low enough (< 0.05) given that the variant (22) has frequency $T_{22} \geq t$ (by default $t = 0.001$).

$$P(x \leq O_{22}|T_{22} \geq t) \leq \sum_{i=0}^{O_{22}} \binom{n}{i} (1-t)^{n-i}$$  \hspace{1cm} (4)

**Step 2: Constructing the SNV graph.** The SNV graph $G = (V, E)$ consists of vertices corresponding to minor variants and edges corresponding to linked pairs of minor variants from different positions. If the intra-host population consists of very similar haplotypes, then graph $G$ is very sparse. Indeed, the PacBio dataset for IAV encompassing 2,500 positions is split into 10,000 vertices, while the SNV graph contains only 700 edges, and, similarly, the simulated Illumina read dataset for the same haplotypes contains only 368 edges.

Note that the isolated minor variants 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 SNVs are errors. As expected, the distribution of the PacBio reads has a heavy tail (see Figure S4), 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 SNV graph is then constructed for the reduced set of reads. Such filtering allows the reduction of systematic errors and refines the SNV graph significantly.

**Step 3: Finding cliques in the SNV graph $G$.** 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 (60).

**Step 4: Merging cliques in the clique graph $C_G$.** The clique graph $C_G = (C, F, L)$ consists of vertices corresponding to cliques in the SNV graph $G$ and two sets of edges $F$ and $L$. A forbidden edge $(p, q) \in F$ connects two cliques $p$ and $q$ with at least one forbidden pair of minor variants from $p$ and $q$ respectively. A linking edge $(p, q) \in L$ connects two cliques $p$ and $q$, $(p, q) \not\in F$, with at least one linked pair of minor variants from $p$ and $q$ respectively. Any true haplotype corresponds to a maximal $(L, F)$-connected subgraph $H$ of $C_G$ which is connected with edges from $L$ and does not contain any edge from $F$ (see Fig. 1 (4)).

Unfortunately, even deciding whether there is a $L$-path between $p$ and $q$ avoiding forbidden edges is known to be NP-hard (61). We find all subgraphs $H$ as follows (see Figure S5): (i) connect all pairs of vertices except connected with forbidden edges, (ii) find all maximal super-cliques in the resulted graph $C_G' = (C, C(2) - F)$ using (60), (iii) split each super- clique into $L$-connected components, and (iv) filter out the $L$-connected components which are proper subsets of other maximal $L$-connected components.

**Step 5: Partitioning reads between merged cliques and finding consensus haplotypes.** Let $S$ be the set of all positions containing at least one minor variant in $V$. Let $q_S$ be an major clique corresponding to a haplotype with all major variants in $S$. The distance between a read $r$ and a clique $q$ equals the number of variants in $q$ that are different from the corresponding nucleotides in $r$. Each read $r$ is assigned to the closest clique $q$ (which can possibly be $q_S$). In case of a tie, we assign $r$ to all closest cliques.

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

**Step 6: Estimating haplotype frequencies by using the expectation-maximization (EM) algorithm.** CliqueSNV estimates the frequencies of the assembled intra-host haplotypes via an expectation-maximization algorithm similar to the one used in IsoEM (62). Let $K$ be the number of assembled viral variants, and let $\alpha$ be the probability of sequencing error. EM algorithm works as follows:

1. Initialize frequencies of viral variants $f_j^{(0)} \leftarrow \frac{1}{N}$.
2. Compute the probability of $l$-long read $r_i$ $i=1, N$, being emitted by viral variant $j = 1, K$,

   $h_{ji} = \prod_{l=1}^{l} ((1-\alpha)M_{ji,l} + \frac{\alpha}{3} (1-M_{ji,l}))$, where $M_{ji,l}$ - indicator if $i$-th read coincides with $j$-th viral variant in the position $l$.
3. (Expectation) Update the amount of read $r_i$ emitted by the $j$-th viral variant $p_{ij} \leftarrow \frac{f_j^{(n-1)}h_{ji}}{\sum_{k=1}^{K} f_k^{(n-1)}h_{ki}}$
4. (Maximization) Update the frequency of the $j$-th viral variant $f_j^{(n)} \leftarrow \frac{\sum_{i=1}^{N} p_{ij}}{\sum_{k=1}^{K} \sum_{i=1}^{N} p_{ik}}$

4. if $|f_j^{(n-1)} - f_j^{(n)}| > \varepsilon$, then $n \leftarrow n + 1$ and go to step 2
5. Output estimated frequencies $f_j^{(n)}$

**RESULTS**

**Performance of haplotyping methods**

We compared CliqueSNV to the 2SNV, PredictHaplo, and aBayesQR haplotyping methods. Since CliqueSNV, PredictHaplo and aBayesQR use Illumina reads, we compared them using the HIV9exp, HIV2exp, HIV5exp, HIV7sim, and IAV10sim datasets. Since CliqueSNV, 2SNV, and PredictHaplo can also use PacBio reads, we compared them using the IAV10exp dataset. We also used consensus sequences in the comparisions (58) because of its simplicity and to evaluate sequences most similar to those generated by the Sanger sequencing method (63).

The precision and recall of haplotype discovery for each method is provided in Table 2. CliqueSNV had the best precision and recall for five of the six datasets. For the HIV9exp dataset, PredictHaplo was more conservative and predicted less false positive variants (better precision) than CliqueSNV.
Table 2. Prediction statistics of haplotype reconstruction methods using experimental and simulated (a) MiSeq and (b) PacBio datasets. 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.

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>CliqueSNV</th>
<th>aBayesQR</th>
<th>PredictHaplo</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV9exp</td>
<td>0.60</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>HIV2exp</td>
<td>0.66</td>
<td>0.11</td>
<td>0.50</td>
</tr>
<tr>
<td>HIV5exp</td>
<td>0.18</td>
<td>0.00</td>
<td>0.33</td>
</tr>
<tr>
<td>HIV7sim</td>
<td>1.00</td>
<td>1.00</td>
<td>0.45</td>
</tr>
<tr>
<td>IAV10sim</td>
<td>0.75</td>
<td>0.11</td>
<td>0.33</td>
</tr>
</tbody>
</table>

The exact EMD values are provided in Table 3. CliqueSNV is sublinear with respect to the number of reads while the runtime of PredictHaplo and 2SNV exhibit super-linear growth. For the 33k IAV10sim reads the CliqueSNV analysis took 21 seconds, while PredictHaplo and 2SNV took around 30 minutes. The runtime of CliqueSNV is quadratic with respect to the number of SVs rather than by the length of the sequencing region (Fig. S2).

We also generated five HIV-1 variants within 1% Hamming distance from each other, which is the estimated genetic distance between related HIV variants from the same person (64). Then we simulated 1M Illumina reads for sequence regions of length 566, 1132, 2263 and 9181 nucleotides for which CliqueSNV required 37, 144, 227, and 614 seconds, respectively, for analyzing these datasets (Fig. S3). For the HIV2exp benchmark, aBayesQR, PredictHaplo, and CliqueSNV required over ten hours, 24 minutes, and only 79 seconds, respectively.

### DISCUSSION

Assembly of haplotype populations from noisy NGS data is one of the most challenging problems of computational genomics. High-throughput sequencing technologies, such as Illumina MiSeq and HiSeq, provide deep sequence coverage that allows discovery of rare, clinically relevant haplotypes. However, the short reads generated by the Illumina technology require assembly that is complicated by sequencing errors, an unknown number of haplotypes in a sample, and the genetic similarity of haplotypes within a sample. Furthermore, the frequency of sequencing errors in
Figure 2. The number of true and false predicted haplotypes depending on the number of accepted mismatches for five benchmarks: (A) HIV9exp; (B) HIV2exp; (C) HIV5exp; (D) HIV7sim; (E) IAV10sim. Two haplotypes are regarded identical if the Hamming distance between them is at most the number of accepted mismatches.

Illumina reads is comparable to the frequencies of true minor mutations (41). The recent development of single-molecule sequencing platforms such as PacBio produce reads that are sufficiently long to span entire genes or small viral genomes. Nonetheless, the error rate of single-molecule sequencing is exceptionally high reaching 13 – 14% (65), which hampers PacBio sequencing to detect and assemble rare viral variants.

We developed CliqueSNV, a new reference-based assembly method for reconstruction of rare genetically-related viral variants such as those observed during infection with rapidly evolving RNA viruses like HIV, HCV and IAV. We demonstrated that CliqueSNV infers accurate haplotyping in the presence of high sequencing error rates and is also suitable for both single-molecule and short-read sequencing. In contrast to other haplotyping methods, CliqueSNV infers viral haplotypes by detection of clusters of statistically linked SNVs rather than through assembly of overlapping reads used with methods such as Savage (24).
Figure 3. Matching distances $E_{T \rightarrow P}$ and $E_{T \rightarrow P}$ between the true haplotype population $T$ and the reconstructed haplotype population $P$ for five benchmarks.

Figure 4. Earth Movers’ Distance (EMD) between true and reconstructed haplotype populations for five benchmarks.

Table 3. Earth Movers’ Distance from predicted haplotypes to the true haplotype population and haplotyping method improvement. Four haplotyping methods (aBayesQR, CliqueSNV, Consensus, PredictHaplo) are benchmarked using five MiSeq (a) and one PacBio datasets (b). The column Impr. (improvement) shows how much better is prediction of haplotyping method over inferred consensus, and it is calculated as $\frac{EMD_c}{EMD_m}$, where $EMD_c$ is an EMD for consensus, and $EMD_m$ is an EMD for method.
Applied to the novel in vitro sequencing HIV-1 benchmark, CliqueSNV correctly reconstructed 87% of the intra-host haplotype population. At the same time, other state-of-the-art tools were not able to recover even a single haplotype without errors. Additionally, we have used the only previously known and commonly used in vitro benchmark (26) and simulated datasets to evaluate the accuracy of existing haplotyping methods. In contrast to the existing methods, CliqueSNV was able to detect minority haplotypes at a low 0.1% frequency and distinguish minority haplotypes differently in only two base pairs.

Although very accurate and fast, CliqueSNV has some limitations. Unlike Savage (24), CliqueSNV is not a de novo assembly tool and requires a reference viral genome. This obstacle could easily be addressed by using Vicuna (58) or other analogous tools to first assemble a consensus sequence from the NGS reads, which can then be used as a reference. Another limitation is for variants that differ only by isolated SNVs separated by long conserved genomic regions longer than the read length which may not be accurately inferred by CliqueSNV. While such situations usually do not occur for viruses, where mutations are typically densely concentrated in different genomic regions, we plan to address this limitation in the next version of CliqueSNV.

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

DATA AVAILABILITY

The datasets HIV2exp and HIV9exp have been deposited in the Sequence Read Archive under accession number SRR12042289 and SRR12042290, respectively.

The links to the data sets and the consensus sequences of the individual strains are available at https://github.com/Sergey-Knyazev/CliqueSNV-validation/blob/master/relevant_haplotypes/HIV9exp.fasta

Software availability

CliqueSNV is available at https://github.com/vtsyvina/CliqueSNV

Validation scripts availability

All scripts and configuration files that were used for validation of the tools are available at https://github.com/Sergey-Knyazev/CliqueSNV-validation

ACKNOWLEDGEMENTS

Use of trade names is for identification only and does not imply endorsement by the U.S. Department of Health and Human Services, the Public Health Service, or the Centers for Disease Control and Prevention (CDC). The findings and conclusions in this paper are those of the authors and do not necessarily represent the official position of the CDC.

Conflict of interest statement. None declared.


55. Gerasimov, E. Analysis of NGS Data from Immune Response and Viral Samples PhD thesis Georgia State University (2017).


