RESEARCH ARTICLE

Long single-molecule reads can resolve the complexity of the Influenza virus composed of rare, closely related mutant variants

Alexander Artyomenko[1†], Nicholas C Wu[2†], Serghei Mangul[3†], Eleazar Eskin[3], Ren Sun[4] and Alex Zelikovsky[1*]

Abstract

As a result of a high rate of mutations and recombination events, an RNA-virus exists as a heterogeneous "swarm". The ability of next-generation sequencing to produce massive quantities of genomic data inexpensively has allowed virologists to study the structure of viral populations from an infected host at an unprecedented resolution. However, high similarity and low frequency of the viral variants impose a huge challenge to assembly of individual full-length genomes. The long read length offered by a single-molecule sequencing technologies allows each mutant variant to be sequenced in a single pass. However, high error rate limits the ability to reconstruct heterogeneous viral population composed of rare, related mutant variants. In this paper, we present 2SNV, a method able to tolerate the high error-rate of the single-molecule protocol and reconstruct mutant variants. The proposed protocol is able to eliminate sequencing errors and reconstruct closely related viral mutant variants. 2SNV uses linkage between single nucleotide variations to efficiently distinguish them from read errors. To benchmark the sensitivity of 2SNV, we performed a single-molecule sequencing experiment on a sample containing a titrated level of known viral mutant variants.

Our method is able to accurately reconstruct clone with frequency of 0.2% and distinguish clones that differed in only two nucleotides distantly located on the genome. 2SNV outperforms existing methods for full-length viral mutant reconstruction. With a high sensitivity and accuracy, 2SNV is anticipated to facilitate not only viral quasispecies reconstruction, but also other biological questions that require detection of rare haplotypes such as genetic diversity in cancer cell population, and monitoring B-cell and T-cell receptor repertoire. The open source implementation of 2SNV is freely available for download at `http://alan.cs.gsu.edu/NGS/?q=content/2snv`

Keywords: Long SMRT reads; Quasispecies; RNA viral variants; Single Nucleotide Variation

Introduction

Majority of the emerging and re-emerging diseases (influenza, hantaviruses, Ebola virus, and Nipah virus), which represent a global threat to the public health, are caused by RNA viruses [1]. Human immune system provides several layers of defense mechanisms against viral infection [2], which may clear the virus from the body. However, in some cases, such as HIV, HCV, and HBV infection, the virus is able to persist in the human body. RNA viruses can be featured by their robust adaptability and evolvability due to their high mutation rates and rapid replication cycles [3,4]. This enables a within-host RNA virus population to organize as a complex and dynamic mutant swarm of many highly similar viral genomes. This mutant spectrum, also known as quasispecies [5], is continuously maintained and regenerated during viral infection, which permits RNA viruses to readily hide from the host immune surveillance, and to acquire vaccine escape and drug resistance [6,7]. Deep sequencing has provided a new lens to monitor individual viral variants. It accelerates the understanding of escape and resistance mechanisms in both laboratory and clinical settings [8,9], in addition to providing insights about the

[*] Correspondence: alexz@cs.gsu.edu
[1] Computer Science Department, Georgia State University, 30302-3994 Atlanta, GA
Full list of author information is available at the end of the article
[†] Equal contributor

viral evolutionary landscape and the genomic interactions [10–12].

Fragmentation-based protocols is a common practice in high-throughput sequencing [13], which breaks the genetic material into a small fragments and destroy the linkage between genetic mutations. Short reads offered by these protocols are well suited to detect discrete genome components, such as the frequency of each single-nucleotide polymorphism. However, high similarity of the individual viral genomes imposes a huge challenge to assemble discrete components into a population of full-length viral genomes. In particular, mutations are often located on the distances unreachable by the short reads. The read length offered by a single-molecule sequencing protocol [14] is comparable to the genome size of most RNA viruses. It allows each genome variant to be sequenced in a single pass, providing an accurate phasing of the distant mutations. The main drawbacks of the long single-molecule technologies are the high error rate and comparatively low throughput, limiting ability of those technologies to study the heterogeneous viral populations. Thus, a complete profiling of all viral genomes within a mutant spectrum is not yet possible.

Recently, this problem has been addressed using various computational and statistical approaches implemented in Quasirecomb [15], PredictHaplo [16], VGA [17], and *k*GEM [18]. These methods perform reasonably well on short reads with high coverage and low error rate, but our experimental validation shows far from satisfactory performance on the sequencing data provided by single-molecule technologies. Also a workflow for reconstruction of closely related variants from raw reads generated during SMRT sequencing was proposed in [19]. Note that a recent method for haplotyping using Pacbio reads proposed in [20] is only applicable for diploid organisms and is not suitable for viral haplotyping with numerous variants.

In this paper, we present **two** **S**ingle **N**ucleotide **V**ariants (2SNV), a comprehensive method for the accurate reconstruction of the heterogeneous viral population from the long single-molecule reads. The 2SNV method hierarchically clusters together reads containing pairs of correlated (i.e., linked) SNV's until no cluster has correlated SNV's left and outputs consensus of each cluster. It allows to reduce error rate and differentiate true biological variants from sequencing artifacts, thus providing increased accuracy to study diversity and composition of the viral spectrum. To bench-

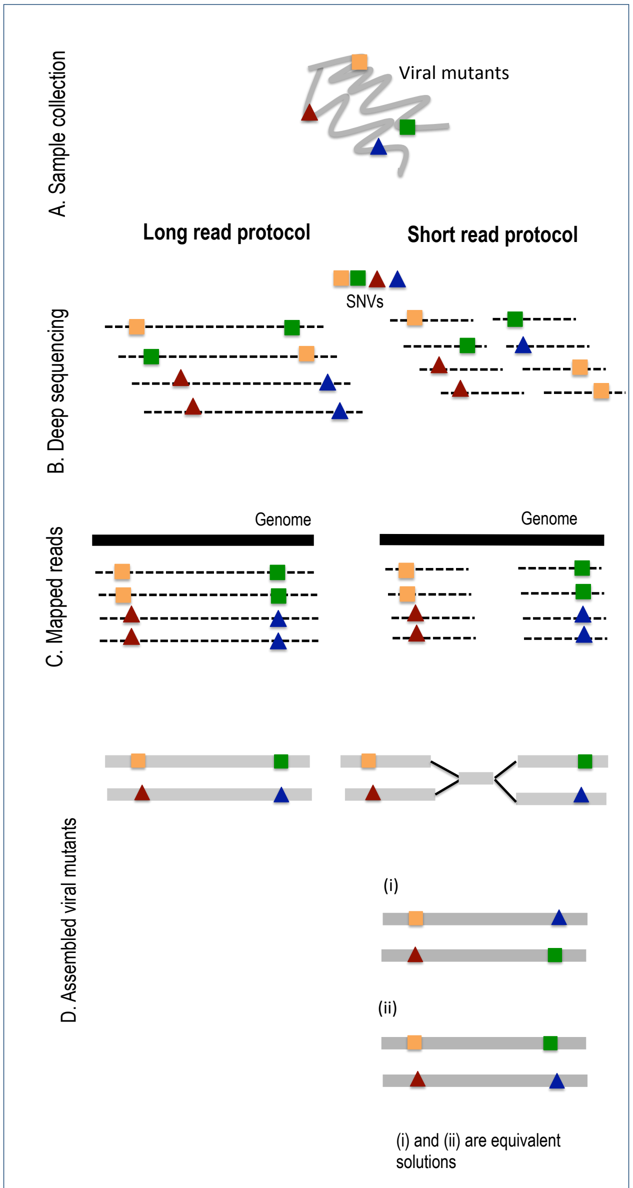

Figure 1 Overview of the long single-molecule sequencing protocol. (a) Extract the viral genomic DNA from the whole blood sample. (b) DNA material from the viral mutants is cleaved into sequence fragments using any suitable restriction enzyme. Amplified fragments are sequenced. (c) Long single-molecule reads are mapped to the reference genome. (d) SNVs are detected and assembled into the viral mutant variants. The short read protocol produces equivalent solutions.

mark the sensitivity of 2SNV, we performed a single-molecule sequencing experiment on a sample containing a titrated level of known viral mutant variants. We were able to reconstruct a haplotype with a frequency of 0.2% and distinguish clones that differed in only two nucleotides. We also showed that 2SNV outperformed existing haplotype reconstruction tools. With a high sensitivity and accuracy, 2SNV is anticipated to facilitate not only viral quasispecies reconstruction, but also other biological questions that require detection of rare haplotypes such as genetic diversity in cancer cell population, and monitoring B-cell and T-cell receptor repertoire.

METHODS

Any method for reconstruction of viral variants from single-molecule reads should overcome low volume and high error rate of sequencing data combined with very high similarity and very low frequency of viral variants. This challenge is equivalent to extraction of an extremely weak signal from very noisy background with signal-to-noise ratio approaching zero. However impossible this task may seem, a satisfactory solution can be based on distinguishing randomness of the noise from systematic signal repetition.

Indeed, since all reads are from the same RNA region of very similar sequences, they can be reliably aligned to each other. In general, the errors in different positions are independent from each other and the further these positions are from each other the less likely any dependency can be caused by systematic errors. Therefore, even slightly more than expected co-occurrence of two rare alleles in non-adjacent positions may serve as a trustful signature of one or more rare variants having the both rare alleles. Such single nucleotide variations (SNVs) are called linked.

The proposed 2SNV method recursively clusters reads containing pairs of linked SNV's until no pair of SNVs exhibits statistically significant linkage in any cluster. Then each cluster should contain just a single viral variant which can be simply reconstructed as the consensus of all reads in the cluster.

In the remainder of the section we derive statistical conditions of SNV linkage and then give detailed description of the 2SNV method which identifies rare variants based SNV pairs satisfying these conditions.

Linkage of SNV pairs

In this section we analyze statistical significance of the linkage between a pair of SNVs which allows to distinguish reads emitted by a rare variant from background errors.

We assume that errors are random and a rare variant has at least 2 mismatches with other variants. Let us consider an arbitrary pair of two distinct positions $I, J \in \{1, \ldots, L\}, I \neq J$, where L be the length of the amplicon (see Figure 2b). Let I_1 and J_1 be the alleles of the most frequent 2-haplotype $(I_1 J_1)$. Note that $(I_1 J_1)$ should be a 2-haplotype from at least one true viral variant assuming that the error rates in the I-th and J-th positions are small and independent.

Let $I_2 \neq I_1$ and $J_2 \neq J_1$ be the alleles of another 2-haplotype. Let E_{kl}, $k, l \in \{1, 2\}$, be the expected number of reads with 2-haplotypes $(I_k J_l)$. The following theorem can be used to decided if the haplotype $I_2 \neq I_1$ exists.

Theorem 1 *Assume that the sequencing error is random, independent and does not exceed 50%. If no viral variant with the haplotype $(I_2 J_2)$ exists, then the expected value of E_{22} is at most*

$$E_{22} \leq \frac{E_{21} \cdot E_{12}}{E_{11}} \qquad (1)$$

The inequality (1) becomes an equality if at least one of 2-haplotypes $(I_1 J_2)$ or $(I_2 J_1)$ also does not exist.

Proof. Let ε_I^{kl} and ε_J^{kl}, $k, l \in \{1, 2\}$, be the probabilities to observe the allele l instead of the true allele k in the positions I and J, respectively. Let T_{kl}, $k, l \in \{1, 2\}$, be the true count of 2-haplotypes $(I_k J_l)$. Then error randomness and independence imply that

$$E_{kl} = \sum_{m,n=1,2} \varepsilon_I^{mk} \varepsilon_J^{nl} T_{mn}$$

In order to prove (1), it is sufficient to show that $E_{11} \cdot E_{22} \leq E_{12} \cdot E_{21}$ assuming that $T_{22} = 0$. Indeed,

$$\begin{aligned}
E_{11} \cdot E_{22} \quad = \quad & \sum_{m,n=1,2} \varepsilon_I^{m1} \varepsilon_J^{n1} T_{mn} \cdot \sum_{m,n=1,2} \varepsilon_I^{m2} \varepsilon_J^{n2} T_{mn} \\
= \quad & \varepsilon_I^{11} \varepsilon_J^{11} \varepsilon_I^{12} \varepsilon_J^{12} T_{11}^2 + \varepsilon_I^{21} \varepsilon_J^{11} \varepsilon_I^{22} \varepsilon_J^{12} T_{21}^2 \\
& + \varepsilon_I^{11} \varepsilon_J^{21} \varepsilon_I^{12} \varepsilon_J^{22} T_{12}^2 \\
& + (\varepsilon_I^{11} \varepsilon_J^{11} \varepsilon_I^{12} \varepsilon_J^{22} + \varepsilon_I^{11} \varepsilon_J^{21} \varepsilon_I^{12} \varepsilon_J^{12}) T_{11} T_{12} \\
& + (\varepsilon_I^{11} \varepsilon_J^{11} \varepsilon_I^{22} \varepsilon_J^{12} + \varepsilon_I^{21} \varepsilon_J^{11} \varepsilon_I^{12} \varepsilon_J^{12}) T_{11} T_{21} \\
& + (\varepsilon_I^{11} \varepsilon_J^{21} \varepsilon_I^{22} \varepsilon_J^{12} + \varepsilon_I^{21} \varepsilon_J^{11} \varepsilon_I^{12} \varepsilon_J^{22}) T_{12} T_{21}
\end{aligned}$$

$$\begin{aligned}
E_{12} \cdot E_{21} \quad = \quad & \sum_{m,n=1,2} \varepsilon_I^{m1} \varepsilon_J^{n2} T_{mn} \cdot \sum_{m,n=1,2} \varepsilon_I^{m2} \varepsilon_J^{n1} T_{mn} \\
= \quad & \varepsilon_I^{11} \varepsilon_J^{12} \varepsilon_I^{12} \varepsilon_J^{11} T_{11}^2 + \varepsilon_I^{21} \varepsilon_J^{12} \varepsilon_I^{22} \varepsilon_J^{11} T_{21}^2 \\
& + \varepsilon_I^{11} \varepsilon_J^{22} \varepsilon_I^{12} \varepsilon_J^{21} T_{12}^2 \\
& + (\varepsilon_I^{11} \varepsilon_J^{12} \varepsilon_I^{12} \varepsilon_J^{21} + \varepsilon_I^{11} \varepsilon_J^{22} \varepsilon_I^{12} \varepsilon_J^{11}) T_{11} T_{12} \\
& + (\varepsilon_I^{11} \varepsilon_J^{12} \varepsilon_I^{22} \varepsilon_J^{11} + \varepsilon_I^{21} \varepsilon_J^{12} \varepsilon_I^{12} \varepsilon_J^{11}) T_{11} T_{21} \\
& + (\varepsilon_I^{11} \varepsilon_J^{22} \varepsilon_I^{22} \varepsilon_J^{11} + \varepsilon_I^{21} \varepsilon_J^{12} \varepsilon_I^{12} \varepsilon_J^{21}) T_{12} T_{21}
\end{aligned}$$

Note that only coefficients for $T_{12}T_{21}$ are different for these products. Therefore, if either $T_{12} = 0$ or $T_{21} = 0$, then $E_{11} \cdot E_{22} = E_{12} \cdot E_{21}$. Otherwise, let all three 2-haplotypes $(I_1 J_1)$, $(I_1 J_2)$, and $(I_2 J_1)$ exist. Then

$$\begin{aligned}
& E_{12}E_{21} - E_{11}E_{22} \\
= \; & (\varepsilon_I^{11}\varepsilon_J^{22}\varepsilon_I^{22}\varepsilon_J^{11} + \varepsilon_I^{21}\varepsilon_J^{12}\varepsilon_I^{12}\varepsilon_J^{21} \\
& -\varepsilon_I^{11}\varepsilon_J^{21}\varepsilon_I^{22}\varepsilon_J^{12} - \varepsilon_I^{21}\varepsilon_J^{11}\varepsilon_I^{12}\varepsilon_J^{22})T_{12}T_{21} \\
= \; & \left(1 + \frac{\varepsilon_J^{12}}{\varepsilon_J^{11}}\frac{\varepsilon_J^{21}}{\varepsilon_J^{22}}\frac{\varepsilon_I^{12}}{\varepsilon_I^{11}}\frac{\varepsilon_I^{21}}{\varepsilon_I^{22}} - \frac{\varepsilon_J^{12}}{\varepsilon_J^{11}}\frac{\varepsilon_J^{21}}{\varepsilon_J^{22}} - \frac{\varepsilon_I^{12}}{\varepsilon_I^{11}}\frac{\varepsilon_I^{21}}{\varepsilon_I^{22}}\right) \times \\
& \times \; \varepsilon_I^{11}\varepsilon_J^{22}\varepsilon_I^{22}\varepsilon_J^{11}T_{12}T_{21} \\
= \; & \left(1 - \frac{\varepsilon_I^{12}}{\varepsilon_I^{11}}\frac{\varepsilon_I^{21}}{\varepsilon_I^{22}}\right)\left(1 - \frac{\varepsilon_J^{12}}{\varepsilon_J^{11}}\frac{\varepsilon_J^{21}}{\varepsilon_J^{22}}\right)\varepsilon_I^{11}\varepsilon_J^{22}\varepsilon_I^{22}\varepsilon_J^{11}T_{12}T_{21} \\
> \; & 0
\end{aligned}$$

The last inequality holds since observing the true allele is more probable than observing the erroneous allele and, therefore, $\varepsilon_I^{kl} < \varepsilon_I^{kk}$ and $\varepsilon_J^{kl} < \varepsilon_J^{kk}$, $k, l \in \{1, 2\}$. QED

The 2SNV method uses Theorem 1 to decide if the alleles I_2 and J_2 are linked as follows. Let O_{kl}, $k, l \in \{1, 2\}$, be the observed number of reads with 2-haplotypes $(I_k J_l)$. Let n be the total number of reads covering the both positions I and J, then

$$p = \frac{O_{21} \cdot O_{12}}{O_{11} \cdot n} \qquad (2)$$

is the largest probability of observing the 2-haplotype $(I_2 J_2)$ among these n reads. The probability to observe at least O_{22} reads in the (n, p) binomial distribution equals

$$Pr(X \geq O_{22}) = 1 - \sum_{i=1}^{O_{22}-1} \binom{n}{i} p^i (1-p)^{n-i} \qquad (3)$$

Since we are looking for a pair of SNV's among $\binom{L}{2}$ possible pairs, we also adjust to multiple testing using Bonferroni correction. Finally, the 2SNV method decides that there exists a true 2-haplotype $(I_2 J_2)$, i.e., I_2 is linked with J_2, if O_{22} is larger than empirically defined value (by default equal 30) and

$$1 - \sum_{i=1}^{O_{22}-1} \binom{n}{i} p^i (1-p)^{n-i} \leq \frac{\mathcal{P}}{\binom{L}{2}} \qquad (4)$$

where p is defined in (2) and $\mathcal{P}$ is the user-defined P-value, by default $\mathcal{P} = 0.01$.

2SNV method for viral variant reconstruction

Algorithm 1 2SNV Algorithm

```
procedure 1: constructing the consensus haplotype for all reads:
  Initialize the set of all clusters with a single cluster with all reads $\mathcal{C} \leftarrow \{R\}$
  For each position $i$ find allele of highest frequency $a_i$
  $Consensus(C) \leftarrow (a_1, \ldots, a_L)$
procedure 2: partitioning reads into simple clusters
  while not all clusters are simple do
    for each non-simple cluster $C \in \mathcal{C}$ do
      if no pair SNVs is linked according to (2-4) then
        Regard $C$ as a simple cluster
      else
        Find a pair of linked SNV's $I_2$ and $J_2$ minimizing (3)
        Find the set $C_1$ of all reads with the 2-haplotype $(I_2 J_2)$
        Find the consensus $c_1 \leftarrow Consensus(C_1)$
        $C_1 \leftarrow Voronoi(c_1)$
        $C_2 \leftarrow C \setminus C_1$, $c_2 \leftarrow Consensus(C_2)$
        $\mathcal{C} \leftarrow \mathcal{C} \cup \{C_1\} \cup \{C_2\} \setminus \{C\}$
procedure 3: estimating frequencies of the consensuses of simple clusters
  Run $k$GEM algorithm for the set of haplotypes $\{Consensus(C), C \in \mathcal{C}\}$.
```

The input to 2SNV consists of a set of aligned PacBio reads (see Figure 2(a)). Alignment required to be in a form of multiple sequence alignment (MSA). The MSA algorithms are too slow to handle PacBio datasets, so instead, we use pairwise alignment by BWA [21] and b2w from Shorah [22] to transform pairwise alignment to MSA format.

The main novel step of the 2SNV algorithm identifies a pair of linked SNV's (see Figure 2(b)). with higher than expected portion of reads containing the 2-haplotype with the both minor alleles according to (2-4).

The 2SNV method maintains a partition of all reads into clusters each containing at least one variant (see Figure 2(c)). Until no pair of SNVs is linked, we recursively

(i) find reads with the linked pair of SNVs and make a new cluster C,
(ii) find consensus c of C, $c \leftarrow Consensus(C)$,
(iii) replace C with the $Voronoi(c)$, and
(iv) update all other clusters and their consensuses

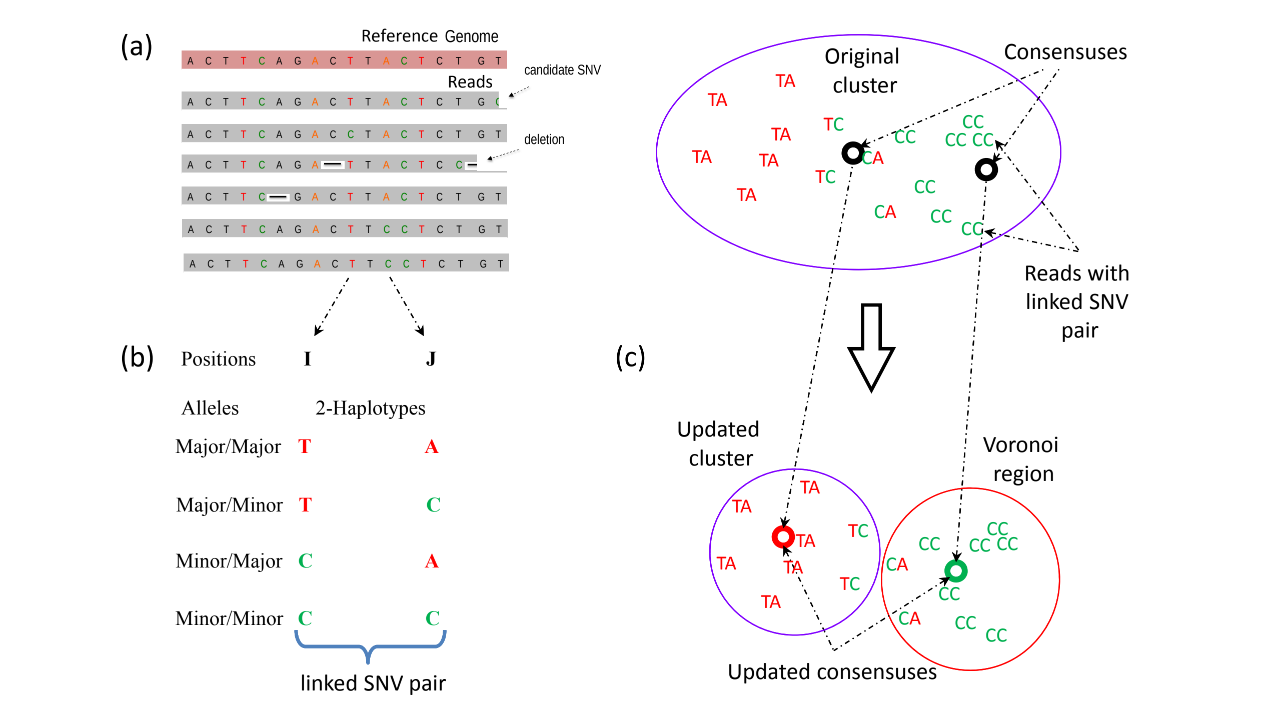
(a)
Reference Genome
A C T T C A G A C T T A C T C T G T
Reads
candidate SNV
deletion
A C T T C A G A C T T A C T C T G C
A C T T C A G A C C T A C T C T G T
A C T T C A G A — T T A C T C C —
A C T T C — G A C T T A C T C T G T
A C T T C A G A C T T C C T C T G T
A C T T C A G A C T T C C T C T G T
(b)
Positions I J
Alleles 2-Haplotypes
Major/Major T A
Major/Minor T C
Minor/Major C A
Minor/Minor C C
linked SNV pair
(c)
Original cluster
Consensuses
Reads with linked SNV pair
TA
TC
CA
CC
Updated cluster
Voronoi region
Updated consensuses

Figure 2 Overview of the 2SNV method: (a) Multiple sequence alignment of reads from the same amplicon; (b) Identification of a linked SNV pair in positions I and J; (c) Recursive cluster splitting: (i) finding consensus of reads with the linked SNV pair, (ii) finding Voronoi region of this consensus, (iii) update the original cluster and the consensuses for the two new clusters.

Here $Voronoi(c)$ is *the Voronoi region of* c which consists of reads that are closer to c than to any other cluster consensus.

The consensus of each resulted cluster defines a single variant. Finally, kGEM estimates variant frequencies based on the maximum likelihood, filters out unlikely variants and fixes incorrect alleles in likely variants.

Formally, the 2SNV algorithm consists of three major procedures (see Algorithm 1):

1. Constructing the consensus haplotype for a set of reads, i.e., a haplotype with the most frequent alleles.
2. Recursive clustering of reads containing a pair of linked SNV's until all clusters are simple, i.e., do not contain linked SNV's.
3. Estimating frequencies of consensuses of simple clusters.

RESULTS

Datasets

Error-prone PCR was performed on the influenza A virus (A/WSN/33) PB2 segment using GeneMorph II Random Mutagenesis Kits (Agilent Technologies, Westlake Village, CA) according to manufacturer's instruction. In the first experiment, a single independent clone has been selected. In the second experiment, 10 independent clones, ranging from 1 to 13 mutations, were selected. These 10 clones were mixed at a geometric ratio with two-fold difference in occurrence frequency for consecutive clones starting with the maximum frequency of 50% of and the minimum frequency of 0.1%. The pairwise edit distance between clones are given in the heatmap on Figure 3.

The 2kb region was amplified from the viral population and subjected to PacBio RS II sequencing using 2 SMRT cells with P4-C2. The average read length was 1973bp and ranges from 200bp to 5kb. In the first experiment there were 11907 reads and in the second experiment there were 33558 reads. Raw sequencing data have been submitted to the NIH Short Read Archive (SRA) under accession number: BioProject PRJNA284802. The nucleotide sequences of the 10 clones are freely available at http://alan.cs.gsu.edu/NGS/?q=content/2snv.

The dataset with a single clone. The average Hamming distance between the recovered haplotype and reads is 14.4%. The 2SNV has been applied for reads. The result of this run perfectly matches the original clone.

The dataset with 10 clones. We ran 2SNV on reads obtained from 10 IAV clones. Our method reported

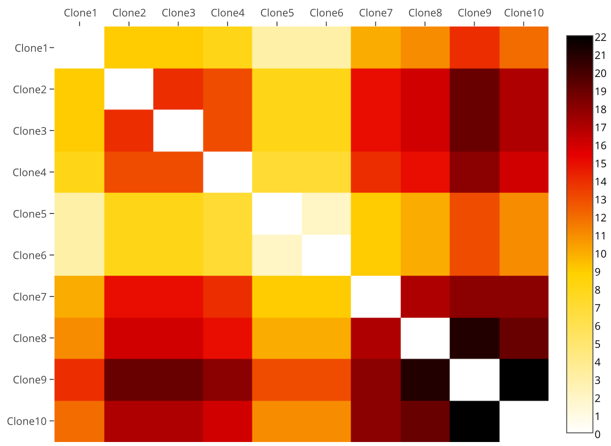

Figure 3 The heatmap representing pairwise edit distance between the 10 IAV clones.

10 haplotypes: the 9 most frequent haplotypes exactly match 9 most frequent clones and the least frequent haplotype does not exactly match any clone. The correlation between the estimated and true frequencies of the 9 correctly reconstructed haplotypes is 99.4% .

Reconstruction of viral variants

Viral variants were reconstructed from the original and sub-sampled datasets. 2SNV was compared with PredictHaplo [16], Quasirecomb [15], and kGEM [18]. Quasirecomb and kGEM were able to reconstruct no more than two most frequent true variants. A workflow [19] reports reduction of error rate to 0.007%. It can distinguish variants with at least 5 mutations away from each other but cannot reconstruct a variant with frequency 0.78%.

For the original data containing all 33.5K reads, 2SNV reconstructed 9 true variants and was not able to reconstruct the least frequent variant represented just by 17 reads (<0.1%). It also reported a single false positive variant with estimated frequency less than 1%. PredictHaplo was able to reconstruct only 6 true variants missing 4 variants with total frequency of 8% while not having any false positives. In order to reliably compare the reconstruction rate of two methods, we have applied them to 40 sub-samples of the original data (each subsample consists of 33558 reads randomly selected with repetition from the original data). The results are presented on Figure 4 and Table 1 in Supplementary materials.

In order to estimate how accuracy of reconstruction methods depends on the coverage, we have randomly sub-sampled N reads (N = 500, 1000, 2000, 4000, 8000, 16000) from the original 33558 reads and run 2SNV and PredictHaplo. The results are shown on Figure (2) in Supplementary Materials . For each coverage and each clone (except Clone5), 2SNV more accurately estimates the frequency. Clone6 and Clone8 for all sub-samples, Clone4 for $N \leq 8000$ and Clone 3 for $N \leq 1000$ are missed by PredictHaplo but reconstructed by 2SNV. Clone6 which is only two mutations away from the more frequent Clone5 was successfully reconstructed for $N \geq 4000$ while PredictHaplo was never able to reconstruct Clone6. Note that since these 2 SNVs between Clone5 and Clone6 are far apart, only long reads can reconstruct this rare variant. From the last plot one can see that the false positive rate for PredictHaplo is also higher than for 2SNV, e.g. 2SNV does not report false positives for $N \leq 8000$. The averages of all runs are given in Table 2 in Supplementary Data.

Runtime

The runtime of 2SNV as well as PredictHaplo is linear with respect to number of reads (see Figure (1) in Supplementary Materials) and quadratic with respect to number of positions. 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 operating system CentOS 6.4.

DISCUSSION

Haplotype phasing represents one of the biggest challenges in next-generation sequencing due to the short read length. The recent development of single-molecule sequencing platform produces reads that are sufficiently long to span the entire gene or small viral genome. It not only benefits the assembly of genomic regions with tendem repeat [23–25], but also offers the opportunity to examine the genetic linkage between mutations. In fact, it is shown that the long read in single-molecule sequencing aids haplotype phasing in diploid genome [26], and in polyploid genome [27]. Nonetheless, the sequencing error rate of single-molecule sequencing platform is extremely high (≈ 14% as estimated by this study), which hampers its ability to reconstruct rare haplotypes. This drawback prohibits single-molecule sequencing platform from applications in which a high sensitivity of haplotypes are needed, such as quasispecies reconstruction. In this study, we have developed 2SNV, which allows quasispecies reconstruction using single-molecule sequencing despite the high sequencing error rate. The high sensitivity of 2SNV permits the detection of extremely rare haplotypes and distinguish between closely related haplotypes. Based on titrated levels of known haplotypes, we demonstrates that 2SNV is able to

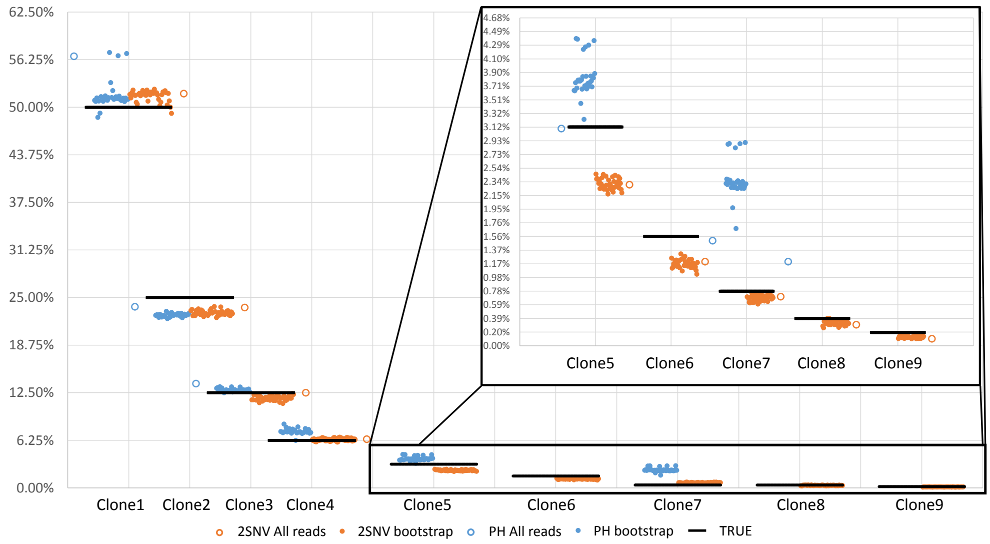

Figure 4 The results of running 2SNV and PredictHaplo (PH) on the original sample with all 33558 reads and on 40 bootstrapped samples (only 35 runs of PH were successful). Clones 6 and 9 were never reconstructed by PH and clone 8 was reconstructed only on full data.

detect a haplotype that has a frequency as low as 0.2%. This sensitivity is comparable to many deep sequencing-based point mutation detection methods [28–31]. In addition, 2SNV successfully distinguishes between Clone5 and Clone6 in this study, which are only two nucleotides away from each other. It highlights the sensitivity of 2SNV to distinguish closely related haplotypes. Our results also show that the sensitivity is coverage-dependent, implying that the sensitivity of 2SNV may further improve when sequencing depth increases. Therefore, the constant increase of sequencing throughput offered by single-molecule sequencing technology provides the unprecedented resolution promising to increase number of discovered rare haplotypes.

The ability to accurately determine the genomic composition of the viral populations and identify closely related viral genomes makes our tool applicable for dissecting evolutionary trajectories and examining mutation interactions in RNA viruses. Evolutionary trajectories and mutation interactions have been shown to play an important role in viral evolution, such as drug resistance [8–10,32], immune escape [33], and cross-species adaptation [34,35]. An unbiased and accurate understanding of the genomic composition of the RNA viruses opens a new avenue to study the underlying mechanism of adaptation, persistence and virulence factors of the pathogen, which are yet to be comprehended.

While viral quasispecies reconstruction is used as a proof-of-concept in this study, the application of 2SNV can be extended to detect haplotype variants in any sample with high genetic heterogeneity and diversity, such as B-cell and T-cell receptor repertoire, cancer cell populations, and metagenomes. It is shown that monitoring B-cell and T-cell receptor repertoire helps investigate virus-host interaction dynamics [36–40]. Furthermore, examining the genetic composition of the cancer cell populations in high sensitivity can facilitate diagnosis and treatment [41]. Therefore, we anticipate that 2SNV will benefit different subfields of biomedical research in the genomic era. We also propose that 2SNV can be applied to increase the resolution of metagenomics profiling from species level to strain level. In summary, 2SNV is a widely applicable tool as single-molecule sequencing technology being popularized.

Software availability

2SNV software is freely available at http://alan.cs.gsu.edu/NGS/?q=content/2snv.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

AA developed, implemented and tested 2SNV, prepared figures and wrote the manuscript, SM wrote the manuscript, prepared figures, and conceived the project, NW wrote the manuscript and provided data, RS, EE, AZ wrote the manuscript and supervised the project, AZ developed the method and conceived the project.

ACKNOWLEDGEMENTS

We would like to thank H. Hao for performing the PacBio sequencing at Johns Hopkins Deep Sequencing & Microarray Core Facility. A.A. was supported by GSU Molecular Basis of Disease Fellowship. S.M. and E.E were supported by National Science Foundation grants 0513612, 0731455, 0729049, 0916676, 1065276, 1302448 and 1320589, and National Institutes of Health grants K25-HL080079, U01-DA024417, P01- HL30568, P01-HL28481, R01-GM083198, R01-MH101782 and R01-ES022282. S.M. was supported in part by Institute for Quantitative & Computational Biosciences Fellowship, UCLA.

Author details

[1]Computer Science Department, Georgia State University, 30302-3994 Atlanta, GA. [2]Department of Integrative Structural and Computational Biology, The Scripps Research Institute, 92037 La Jolla, CA. [3]Computer Science Department, University of California, Los Angeles, 90095 Los Angeles, CA. [4]Molecular and Medical Pharmacology, University of California, Los Angeles, 90095 Los Angeles, CA.

References

1. Murphy, F.A., Kingsbury, D.W.: Virus taxonomy. Fields Virology **2**, 15–57 (1996)
2. Jenner, R.G., Young, R.A.: Insights into host responses against pathogens from transcriptional profiling. Nature Reviews Microbiology **3**(4), 281–294 (2005)
3. Domingo, E.: Mutation rates and rapid evolution of RNA viruses. The evolutionary biology of viruses, 161–184 (1994)
4. Holland, J., Spindler, K., Horodyski, F., Grabau, E., Nichol, S., VandePol, S.: Rapid evolution of RNA genomes. Science **215**(4540), 1577–1585 (1982)
5. Eigen, M.: Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften **58**(10), 465–523 (1971)
6. Domingo, E., Holland, J.: RNA virus mutations and fitness for survival. Annual Reviews in Microbiology **51**(1), 151–178 (1997)
7. Lauring, A.S., Andino, R.: Quasispecies theory and the behavior of rna viruses. PLoS Pathogens **6**(7), 1001005 (2010)
8. Bushman, F.D., Hoffmann, C., Ronen, K., Malani, N., Minkah, N., Rose, H.M., Tebas, P., Wang, G.P.: Massively parallel pyrosequencing in hiv research. Aids **22**(12), 1411–1415 (2008)
9. Margeridon-Thermet, S., Shulman, N.S., Ahmed, A., Shahriar, R., Liu, T., Wang, C., Holmes, S.P., Babrzadeh, F., Gharizadeh, B., Hanczaruk, B., *et al.*: Ultra-deep pyrosequencing of hepatitis b virus quasispecies from nucleoside and nucleotide reverse-transcriptase inhibitor (nrti)–treated patients and nrti-naive patients. Journal of Infectious Diseases **199**(9), 1275–1285 (2009)
10. Wang, C., Mitsuya, Y., Gharizadeh, B., Ronaghi, M., Shafer, R.W.: Characterization of mutation spectra with ultra-deep pyrosequencing: Application to HIV-1 drug resistance. Genome Research **17**(8), 1195–1201 (2007)
11. Liu, J., Miller, M.D., Danovich, R.M., Vandergrift, N., Cai, F., Hicks, C.B., Hazuda, D.J., Gao, F.: Analysis of low-frequency mutations associated with drug resistance to raltegravir before antiretroviral treatment. Antimicrobial agents and chemotherapy **55**(3), 1114–1119 (2011)
12. Palmer, S., Boltz, V., Maldarelli, F., Kearney, M., Halvas, E.K., Rock, D., Falloon, J., Davey Jr, R.T., Dewar, R.L., Metcalf, J.A., *et al.*: Selection and persistence of non-nucleoside reverse transcriptase inhibitor-resistant HIV-1 in patients starting and stopping non-nucleoside therapy. Aids **20**(5), 701–710 (2006)
13. Cheval, J., Sauvage, V., Frangeul, L., Dacheux, L., Guigon, G., Dumey, N., Pariente, K., Rousseaux, C., Dorange, F., Berthet, N., *et al.*: Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples. Journal of clinical microbiology **49**(9), 3268–3275 (2011)
14. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., *et al.*: Real-time dna sequencing from single polymerase molecules. Science **323**(5910), 133–138 (2009)
15. Töpfer, A., Zagordi, O., Prabhakaran, S., Roth, V., Halperin, E., Beerenwinkel, N.: Probabilistic inference of viral quasispecies subject to recombination. Journal of Computational Biology **20**(2), 113–123 (2013)
16. Prabhakaran, S., Rey, M., Zagordi, O., Beerenwinkel, N., Roth, V.: HIV haplotype inference using a propagating Dirichlet process mixture model. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) **11**(1), 182–191 (2014)
17. Mangul, S., Wu, N.C., Mancuso, N., Zelikovsky, A., Sun, R., Eskin, E.: Accurate viral population assembly from ultra-deep sequencing data. Bioinformatics **30**(12), 329–337 (2014)
18. Skums, P., Artyomenko, A., Glebova, O., Ramachandran, S., Mandoiu, I.I., Campo, D.S., Dimitrova, Z., Zelikovsky, A., Khudyakov, Y.: Computational framework for next-generation sequencing of heterogeneous viral populations using combinatorial pooling. Bioinformatics **31**(5), 682–690 (2015)
19. Dilernia, D.A., Chien, J.-T., Monaco, D.C., Brown, M.P.S., Ende, Z., Deymier, M.J., Yue, L., Paxinos, E.E., Allen, S., Tirado-Ramos, A., Others: Multiplexed highly-accurate DNA sequencing of closely-related HIV-1 variants using continuous long reads from single molecule, real-time sequencing. Nucleic Acids Research (2015)
20. Tilgner, H., Grubert, F., Sharon, D., Snyder, M.P.: Defining a personal, allele-specific, and single-molecule long-read transcriptome. Proceedings of the National Academy of Sciences **111**(27), 9869–9874 (2014)
21. Li, H., Durbin, R.: Fast and accurate short read alignment with burrows–wheeler transform. Bioinformatics **25**(14), 1754–1760 (2009)
22. Zagordi, O., Bhattacharya, A., Eriksson, N., Beerenwinkel, N.: Shorah: estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinformatics **12**(1), 119 (2011)
23. Ummat, A., Bashir, A.: Resolving complex tandem repeats with long reads. Bioinformatics **30**(24), 3491–3498 (2014)
24. Krsticevic, F.J., Schrago, C.G., Carvalho, A.B.: Long-Read Single Molecule Sequencing to Resolve Tandem Gene Copies: The Mst77Y Region on the Drosophila melanogaster Y Chromosome. G3 (Bethesda) **5**(6), 1145–1150 (2015)
25. Doi, K., Monjo, T., Hoang, P.H., Yoshimura, J., Yurino, H., Mitsui, J., Ishiura, H., Takahashi, Y., Ichikawa, Y., Goto, J., Tsuji, S., Morishita, S.: Rapid detection of expanded short tandem repeats in personal genomics using hybrid sequencing. Bioinformatics **30**(6), 815–822 (2014)
26. Pendleton, M., Sebra, R., Pang, A.W., Ummat, A., Franzen, O., Rausch, T., Stutz, A.M., Stedman, W., Anantharaman, T., Hastie, A., Dai, H., Fritz, M.H., Cao, H., Cohain, A., Deikus, G., Durrett, R.E., Blanchard, S.C., Altman, R., Chin, C.S., Guo, Y., Paxinos, E.E., Korbel, J.O., Darnell, R.B., McCombie, W.R., Kwok, P.Y., Mason, C.E., Schadt, E.E., Bashir, A.: Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat. Methods (2015)
27. Aguiar, D., Istrail, S.: Haplotype assembly in polyploid genomes and identical by descent shared tracts. Bioinformatics **29**(13), 352–360 (2013)
28. Flaherty, P., Natsoulis, G., Muralidharan, O., Winters, M., Buenrostro, J., Bell, J., Brown, S., Holodniy, M., Zhang, N., Ji, H.P.: Ultrasensitive detection of rare mutations using next-generation targeted resequencing. Nucleic Acids Res. **40**(1), 2 (2012)
29. Harismendy, O., Schwab, R.B., Bao, L., Olson, J., Rozenzhak, S., Kotsopoulos, S.K., Pond, S., Crain, B., Chee, M.S., Messer, K., Link, D.R., Frazer, K.A.: Detection of low prevalence somatic mutations in solid tumors with ultra-deep targeted sequencing. Genome Biol. **12**(12), 124 (2011)
30. Forshew, T., Murtaza, M., Parkinson, C., Gale, D., Tsui, D.W., Kaper, F., Dawson, S.J., Piskorz, A.M., Jimenez-Linan, M., Bentley, D., Hadfield, J., May, A.P., Caldas, C., Brenton, J.D., Rosenfeld, N.: Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. Sci Transl Med **4**(136), 136–68 (2012)
31. Li, M., Stoneking, M.: A new approach for detecting low-level mutations in next-generation sequence data. Genome Biol. **13**(5), 34 (2012)
32. Beerenwinkel, N., Schmidt, B., Walter, H., Kaiser, R., Lengauer, T., Hoffmann, D., Korn, K., Selbig, J.: Diversity and complexity of hiv-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. Proceedings of the National Academy of Sciences **99**(12), 8271–8276 (2002)
33. Goepfert, P.A., Lumm, W., Farmer, P., Matthews, P., Prendergast, A., Carlson, J.M., Derdeyn, C.A., Tang, J., Kaslow, R.A., Bansal, A., *et al.*: Transmission of hiv-1 gag immune escape mutations is associated

with reduced viral load in linked recipients. The Journal of experimental medicine **205**(5), 1009–1017 (2008)

34. Herfst, S., Schrauwen, E.J., Linster, M., Chutinimitkul, S., de Wit, E., Munster, V.J., Sorrell, E.M., Bestebroer, T.M., Burke, D.F., Smith, D.J., *et al.*: Airborne transmission of influenza a/h5n1 virus between ferrets. Science **336**(6088), 1534–1541 (2012)
35. Imai, M., Watanabe, T., Hatta, M., Das, S.C., Ozawa, M., Shinya, K., Zhong, G., Hanson, A., Katsura, H., Watanabe, S., *et al.*: Experimental adaptation of an influenza h5 ha confers respiratory droplet transmission to a reassortant h5 ha/h1n1 virus in ferrets. Nature **486**(7403), 420–428 (2012)
36. Wu, X., Zhou, T., Zhu, J., Zhang, B., Georgiev, I., Wang, C., Chen, X., Longo, N.S., Louder, M., McKee, K., O'Dell, S., Perfetto, S., Schmidt, S.D., Shi, W., Wu, L., Yang, Y., Yang, Z.Y., Yang, Z., Zhang, Z., Bonsignori, M., Crump, J.A., Kapiga, S.H., Sam, N.E., Haynes, B.F., Simek, M., Burton, D.R., Koff, W.C., Doria-Rose, N.A., Connors, M., Mullikin, J.C., Nabel, G.J., Roederer, M., Shapiro, L., Kwong, P.D., Mascola, J.R.: Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. Science **333**(6049), 1593–1602 (2011)
37. Zhu, J., Ofek, G., Yang, Y., Zhang, B., Louder, M.K., Lu, G., McKee, K., Pancera, M., Skinner, J., Zhang, Z., Parks, R., Eudailey, J., Lloyd, K.E., Blinn, J., Alam, S.M., Haynes, B.F., Simek, M., Burton, D.R., Koff, W.C., Mullikin, J.C., Mascola, J.R., Shapiro, L., Kwong, P.D., Becker, J., Benjamin, B., Blakesley, R., Bouffard, G., Brooks, S., Coleman, H., Dekhtyar, M., Gregory, M., Guan, X., Gupta, J., Han, J., Hargrove, A., Ho, S.L., Johnson, T., Legaspi, R., Lovett, S., Maduro, Q., Masiello, C., Maskeri, B., McDowell, J., Montemayor, C., Mullikin, J., Park, M., Riebow, N., Schandler, K., Schmidt, B., Sison, C., Stantripop, M., Thomas, J., Thomas, P., Vemulapalli, M., Young, A.: Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. Proc. Natl. Acad. Sci. U.S.A. **110**(16), 6470–6475 (2013)
38. Zhu, J., Wu, X., Zhang, B., McKee, K., O'Dell, S., Soto, C., Zhou, T., Casazza, J.P., Mullikin, J.C., Kwong, P.D., Mascola, J.R., Shapiro, L., Becker, J., Benjamin, B., Blakesley, R., Bouffard, G., Brooks, S., Coleman, H., Dekhtyar, M., Gregory, M., Guan, X., Gupta, J., Han, J., Hargrove, A., Ho, S.L., Johnson, T., Legaspi, R., Lovett, S., Maduro, Q., Masiello, C., Maskeri, B., McDowell, J., Montemayor, C., Mullikin, J., Park, M., Riebow, N., Schandler, K., Schmidt, B., Sison, C., Stantripop, M., Thomas, J., Thomas, P., Vemulapalli, M., Young, A.: De novo identification of VRC01 class HIV-1-neutralizing antibodies by next-generation sequencing of B-cell transcripts. Proc. Natl. Acad. Sci. U.S.A. **110**(43), 4088–4097 (2013)
39. Miconnet, I.: Probing the T-cell receptor repertoire with deep sequencing. Curr Opin HIV AIDS **7**(1), 64–70 (2012)
40. Klarenbeek, P.L., Remmerswaal, E.B., ten Berge, I.J., Doorenspleet, M.E., van Schaik, B.D., Esveldt, R.E., Koch, S.D., ten Brinke, A., van Kampen, A.H., Bemelman, F.J., Tak, P.P., Baas, F., de Vries, N., van Lier, R.A.: Deep sequencing of antiviral T-cell responses to HCMV and EBV in humans reveals a stable repertoire that is maintained for many years. PLoS Pathog. **8**(9), 1002889 (2012)
41. Mardis, E.R., Wilson, R.K.: Cancer genome sequencing: a review. Hum. Mol. Genet. **18**(R2), 163–168 (2009)