Bioinformatics, 34, 2018, i23–i31
doi: 10.1093/bioinformatics/bty291
ISMB 2018

OXFORD

Viral quasispecies reconstruction via tensor factorization with successive read removal

Soyeon Ahn*, Ziqi Ke and Haris Vikalo

Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712, USA

*To whom correspondence should be addressed.

Abstract

Motivation: As RNA viruses mutate and adapt to environmental changes, often developing resistance to anti-viral vaccines and drugs, they form an ensemble of viral strains—a viral quasispecies. While high-throughput sequencing (HTS) has enabled in-depth studies of viral quasispecies, sequencing errors and limited read lengths render the problem of reconstructing the strains and estimating their spectrum challenging. Inference of viral quasispecies is difficult due to generally non-uniform frequencies of the strains, and is further exacerbated when the genetic distances between the strains are small.

Results: This paper presents TenSQR, an algorithm that utilizes tensor factorization framework to analyze HTS data and reconstruct viral quasispecies characterized by highly uneven frequencies of its components. Fundamentally, TenSQR performs clustering with successive data removal to infer strains in a quasispecies in order from the most to the least abundant one; every time a strain is inferred, sequencing reads generated from that strain are removed from the dataset. The proposed successive strain reconstruction and data removal enables discovery of rare strains in a population and facilitates detection of deletions in such strains. Results on simulated datasets demonstrate that TenSQR can reconstruct full-length strains having widely different abundances, generally outperforming state-of-the-art methods at diversities 1–10% and detecting long deletions even in rare strains. A study on a real HIV-1 dataset demonstrates that TenSQR outperforms competing methods in experimental settings as well. Finally, we apply TenSQR to analyze a Zika virus sample and reconstruct the full-length strains it contains.

Availability and implementation: TenSQR is available at https://github.com/SoYeonA/TenSQR.

Contact: soyeon.ahn@utexas.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

RNA viruses such as HIV, SARS, Zika and Ebola are characterized by high mutation rates that lead to new viral strains by means of point mutations, insertions and deletions. The resulting population of closely related yet non-identical viral genomes is known as a viral quasispecies (Lauring and Andino, 2010). Genetic heterogeneity of such viral populations enables the virus to adapt and proliferate in dynamically changing environments, e.g. over the course of an infection (Carroll *et al.*, 2013). Therefore, determining the genetic structure of viral quasispecies is of importance for effective anti-viral vaccine designs and the development of therapeutic treatments for viral diseases.

Quasispecies Spectrum Reconstruction (QSR) aims to reconstruct an a priori unknown number of viral sequences in a quasispecies and estimate their relative frequencies. To this end, QSR methods typically employ the following steps: (i) clustering together sequencing reads that originate from the same strain; (ii) reconstructing the strains using the clustered reads and (iii) determining relative frequencies of the reconstructed strains based on the corresponding cluster sizes (Beerenwinkel *et al.*, 2012). High-throughput sequencing (HTS) technologies have in principle enabled unprecedented studies of quasispecies populations. However, their precise reconstruction remains difficult due to the presence of sequencing errors and limited length of HTS reads. The QSR problem is particularly challenging when the frequencies of strains in a viral population are highly imbalanced, i.e. the quasispecies contains both strains having high and those having low abundances, and is further exacerbated if the genetic distances between strains are relatively small. In those settings, performance of clustering-based QSR methods suffers from erroneous attribution of the reads that have originated from rare strains to nearby (in terms of genetic distance) highly abundant strains; such errors lead to failures to discover

strains of low abundancy (Posada-Cespedes *et al.*, 2016) and thus may hinder the discovery of effective drug treatments (Le *et al.*, 2009; Simen *et al.*, 2009).

Methods for reconstruction of viral quasispecies from HTS data include a probabilistic clustering algorithm ShoRAH (Zagordi *et al.*, 2010, 2011), read-graph based path selection technique ViSpA (Astrovskaya *et al.*, 2011), a combinatorial algorithm QuRe (Prosperi and Salemi, 2012) and a hidden Markov model based scheme QuasiRecomb (Töpfer *et al.*, 2013). More recent softwares include PredictHaplo (Prabhakaran *et al.*, 2014), which relies on a probabilistic mixture model and specifically targets assembly of HIV populations; VGA (Mangul *et al.*, 2014), a scheme that employs a high-fidelity sequencing protocol to eliminate sequencing errors and assemble rare variants using deep sequencing and HaploClique (Töpfer *et al.*, 2014), the first method to attempt detection of long insertions and deletions (indels) by means of enabling insert-size compatibility in the max-clique enumeration procedure. Another recent technique, ViQuaS (Jayasundara *et al.*, 2015), adapts the combinatorial approach proposed by QuRe (Prosperi and Salemi, 2012) to a reference-assisted *de novo* assembly framework and generally outperforms existing state-of-the-art tools on a wide range of datasets. However, it has been demonstrated in (Schirmer *et al.*, 2014) and pointed out in (Posada-Cespedes *et al.*, 2016) that the existing methods generally struggle to reliably reconstruct quasispecies composed of strains having small mutual genetic distances. To specifically address reconstruction of quasispecies characterized by low diversity, a sequential Bayesian inference method, aBayesQR, was recently proposed in (Ahn and Vikalo, 2017). While aBayesQR is indeed more accurate than competing methods in low diversity ($\leq 5\%$) settings, reconstructing quasispecies characterized by both low diversity and highly uneven strain frequencies remains a challenge. More recently, methods for *de novo* quasispecies reconstruction, MLEHaplo (Malhotra *et al.*, 2015) and SAVAGE (Baaijens *et al.*, 2017), have also been proposed.

In this paper, we propose a reconstruction method that takes a step towards overcoming limitations of existing techniques and is capable of accurately assembling quasispecies characterized by a wide range of strain frequencies. The method, referred to as TenSQR (Tensor factorization with Successive removal for Quasispecies Reconstruction), represents sequencing data by means of a structured sparse binary tensor. Factorization of such objects, both matrices and tensors, was previously used to enable haplotype assembly of diploid (Cai *et al.*, 2016) and polyploid (Hashemi *et al.*, 2017) genomes. Note that the abundances of reads generated by sequencing diploid or polyploid haplotypes are near uniform—minor variations in those abundances are primarily due to imperfections of the sample preparation and sequencing steps. Matrix and tensor factorization schemes in (Cai *et al.*, 2016; Hashemi *et al.*, 2017) make an explicit assumption that the haplotype frequencies are equal. However, this assumption is clearly violated in the QSR problem. It therefore comes as no surprise that when matrix or tensor factorization is directly applied to reconstruct a heterogeneous mixture of sequences characterized by highly uneven frequencies (by means of forming imbalanced clusters, each collecting reads having originated from one sequence), dominant sequences (large clusters) are typically recovered correctly while the rare sequences (small clusters) are often either missed or reconstructed erroneously (Chaisson *et al.*, 2017). To address this concern, TenSQR successively infers strains in a quasispecies by repeatedly performing the following two steps. In the first step, the sparse tensor is factorized and its missing entries are inferred by alternatingly optimizing the factors; this step is completed by identifying and reconstructing the most abundant strain. In the second step, all the reads deemed to have originated from the reconstructed (the most abundant) strain are removed from the dataset and hence from the originally formed tensor; the number of such reads is indicative of the reconstructed strain's frequency. Then, the first step is performed anew on the reduced dataset to reconstruct the second most abundant strain and so on. These two steps are repeated until all the strains are reconstructed. Since the proposed scheme revisits tensor factorization multiple times, computational complexity of that step becomes a concern. To mitigate it, we exploit the special structure of the problem and propose a novel majority-voting based efficient alternating minimization scheme for sparse binary tensor factorization. We show that the convergence of the alternating minimization procedure is guaranteed, and that the proposed scheme allows detection of deletions in the reconstructed strains. The developed framework is augmented by an additional pipeline designed to detect insertions that may be present in some of the reconstructed strains. Our tests on simulated data demonstrate that, unlike the competing methods, the proposed tensor factorization framework for successive strain inference supports reliable discovery and accurate reconstruction of rare strains existing in highly imbalanced populations even when the population diversity is low. In particular, TenSQR compares favorably to state-of-the-art methods at diversities 1–10%, and detects deletions in strains with low abundance. Performance of TenSQR on a real HIV-1 dataset demonstrates TenSQR's ability to reliably reconstruct quasispecies in more general settings. Furthermore, we employ our method to reconstruct full-length strains in a Zika virus (ZIKV) sample.

2 Materials and methods

2.1 Problem formulation

Let $Q = \{q_i, i = 1, \ldots, k\}$ denote the set of k viral quasispecies strains that differ from each other at a number of variant sites, and let $R = \{r_j, j = 1, \ldots, m\}$ denote the set of reads generated by sequencing the strains in Q; relative ordering of the reads is determined by aligning them to a reference genome. Homozygous sites (i.e. the sites containing alleles common to all strains) are not utilized by our tensor model; instead, viral haplotypes are reconstructed using heterozygous sites that have abundance of alleles above a predetermined threshold (i.e. sites that are with high confidence declared to be variants). Note that the homozygous sites are later used to assemble full-length viral strains, as detailed in Section 2.3.

Let us organize the data (i.e. information about the variant sites provided by the paired-end reads) in an $m \times n$ read fragment matrix $\mathbf{F}'$, where the rows correspond to reads and columns correspond to variant positions in the sequences. A convenient numerical representation of $\mathbf{F}'$ is obtained by denoting nucleotides with 4D standard unit vectors $\mathbf{e}_i^{(4)}$, $1 \leq i \leq 4$, with 0s in all positions except the ith one that has value 1 (e.g. $\mathbf{e}_1^{(4)} = [1\,0\,0\,0]$, $\mathbf{e}_2^{(4)} = [0\,1\,0\,0]$ and so on). This leads to a representation of the read fragment matrix $\mathbf{F}'$ by means of a binary tensors $\mathcal{F}$ whose fibers represent nucleotides and horizontal slices correspond to reads. $\mathcal{F}$ can be thought of as being obtained by sparsely sampling an underlying tensor $\mathcal{T}$ whose fibers are standard unit vectors $\mathbf{e}_i^{(4)}$; sampling is potentially erroneous due to sequencing errors. To arrive at a tensor factorization formulation of the problem, it is useful to point out that $\mathcal{T}$ can be thought of as being obtained by multiplying a read membership indicator matrix $\mathbf{M}$ and a binary tensor $\mathcal{H}$ that encodes the true viral haplotype information—namely, fibers of $\mathcal{H}$ are standard unit vectors $\mathbf{e}_i^{(4)}$

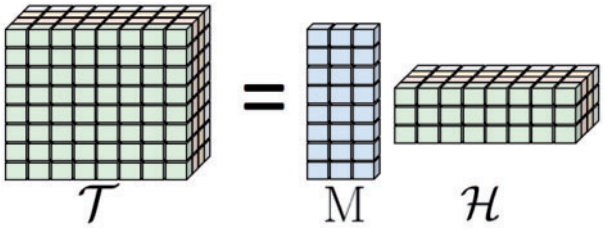

Fig. 1. An illustration of the tensor factorization representation of the viral quasispecies assembly problem

representing alleles while each lateral slice of $\mathcal{H}$ is one of the k viral haplotypes that need to be reconstructed. Moreover, the indicator matrix $\mathbf{M}$ has for rows the standard unit vectors $\mathbf{e}_i^{(k)}$, $1 \leq i \leq k$, with 0 s in all positions except the ith one that has value 1. If, for example, the jth row of $\mathbf{M}$ is $\mathbf{e}_l^{(k)}$, then that indicates the jth read was obtained by 'sampling' (i.e. via shotgun sequencing) the lth viral haplotype. Figure 1 illustrates the representation of $\mathcal{T}$ by means of a product of $\mathbf{M}$ and $\mathcal{H}$.

We formulate the task of reconstructing the set of viral haplotype sequences $\mathcal{H}$ from the observed reads $\mathcal{F}$ as a collection of k—1 tensor factorization problems; following each factorization, sequencing reads associated with the most dominant assembled strain are removed from $\mathcal{F}$ and the factorization is performed anew until all the reads remaining in $\mathcal{F}$ are of the same origin (i.e. come from the same viral haplotype). The tensor factorization procedure is formalized in the next section.

2.2 Structured tensor factorization using alternating minimization

Let $\mathbf{F} \in \{0,1\}^{m \times 4n}$ and $\mathbf{H} \in \{0,1\}^{4n \times k}$ be the mode-1 unfoldings of tensors $\mathcal{F}$ and $\mathcal{H}$, respectively. The QSR problem can be cast as the optimization

$$\min_{\mathbf{M},\mathbf{H}} \frac{1}{2} ||\mathcal{P}_\Omega(\mathbf{F} - \mathbf{M}\mathbf{H}^\top)||_F^2, \tag{1}$$

where Ω denotes the set of informative entries of $\mathbf{F}$ (i.e. positions of the information provided by the set of reads), $\mathcal{P}_\Omega$ is the projection operator (formalizing the sampling of viral strains by reads) and $||\cdot||_F$ denotes the Frobenius norm of its argument. This is a computationally challenging optimization problem that can be approximately solved by means of alternating minimization, i.e. alternately solving Equation (1) for either $\mathbf{M}$ or $\mathbf{H}$ while keeping the other one fixed (Jain *et al.*, 2013). In particular, given the current estimates $\mathbf{M}_t$ and $\mathbf{H}_t$, we alternately update

$$\mathbf{M}_{t+1} = \arg \min_{\mathbf{M} \in \{0,1\}^{m \times k}} \frac{1}{2} ||\mathcal{P}_\Omega(\mathbf{F} - \mathbf{M}\mathbf{H}_t^\top)||_F^2 \tag{2}$$

and

$$\mathbf{H}_{t+1} = \arg \min_{\mathbf{H} \in \{0,1\}^{4n \times k}} \frac{1}{2} ||\mathcal{P}_\Omega(\mathbf{F} - \mathbf{M}_{t+1}\mathbf{H}^\top)||_F^2 \tag{3}$$

until a termination criterion is met. Note that one can impose structural constraints on $\mathbf{M}$ to find it efficiently (Cai *et al.*, 2016). In particular, $\mathbf{M}$ can be found by examining for each read all possible haplotype associations and selecting the one that minimizes the number of base changes needed to be consistent with the observed information in $\mathbf{F}$ [i.e. minimizing the so-called minimum error correction (MEC) score, first proposed in (Lippert *et al.*, 2002)]. Therefore, the complexity of finding $\mathbf{M}$ given $\mathbf{H}$ is $O(mk)$, where m is the number of reads and k denotes the number of viral haplotypes. Now, optimization Equation (3) can be performed via, e.g. relaxing $\mathbf{H}$ and performing gradient descent as previously done in (Cai *et al.*, 2016; Hashemi *et al.*, 2017); however, we instead exploit the discrete nature of the problem to solve Equation (3) much more efficiently. In particular, we employ the majority voting rule to form consensus sequences among reads that belong to the same cluster, i.e. originate from the same viral haplotype. While the complexity of finding $\mathbf{H}$ given $\mathbf{M}$ is $O(mn)$ for both the majority voting and gradient descent schemes, the former solves Equation (3) directly while the latter only takes a step towards the solution. As a result, the convergence of the proposed alternating minimization scheme that employs majority voting to solve Equation (3) is significantly faster than that of the scheme relying on gradient descent (please see Supplementary Material B for numerical results illustrating this point). Moreover, we show that the convergence of the proposed alternating minimization procedure is guaranteed (the proof is provided in Supplementary Material A).

2.3 Successive reconstruction of viral sequences

The alternating minimization procedure described in Section 2.2 is expected to work well in settings where the abundances of different haplotypes are uniform (i.e. equal) and the ploidy is low (Cai *et al.*, 2016; Hashemi *et al.*, 2017). However, when the frequencies of components in the mixture are uneven, the described framework is capable of correctly reconstructing dominant sequences but struggles to assemble sequences having low abundances, as pointed out in (Chaisson *et al.*, 2017). The reason for such behavior is that Equation (1) emphasizes accurate recovery of dominant haplotypes which significantly contribute to the overall MEC score while neglecting rare ones whose contribution to the MEC score is relatively small. To address this concern, we propose an iterative scheme where upon performing optimization Equation (1), the most abundant viral strain is identified and reconstructed from $\mathbf{H}$; following this reconstruction, all the reads assessed to have originated from the reconstructed strain are removed from the dataset (and thus from $\mathbf{F}$). Optimization Equation (1) is repeated on the reduced $\mathbf{F}$ to recover the second most abundant strain and the procedure continues until all strains are reconstructed.

Let m_l be the number of rows having unit vectors $\mathbf{e}_l^{(k)}, 1 \leq l \leq k$, in $\mathbf{M}$. The most dominant haplotype can be identified as the wth lateral slice of $\mathcal{H}$, $\mathcal{H}_{w::}$, satisfying $w = \arg\max_l m_l$. While each row of $\mathbf{M}$ is essentially an indicator of the origin of the corresponding read, membership information obtained via optimization Equation (1) could be erroneous when the strains in a mixture have non-uniform frequencies. In fact, reads originating from a rare strain are likely to be assigned to a more abundant one, especially when those two strains are highly similar (i.e. in the low diversity setting). Motivated by this observation, we re-examine the reads in $\mathbf{F}$ to identify those originating from the reconstructed viral haplotype using statistical tests described next.

Assume that the sequencing errors are independent and identically distributed across all variant sites for all reads, and that they happen with probability ε. Let n_i denote the number of informative sites of the ith row in $\mathbf{F}'$, $f'_{i\cdot}$, and let d_i be the number of mismatches between $f'_{i\cdot}$ and the recovered haplotype (i.e. the most abundant haplotype reconstructed in the current iteration) counted over the observed nucleotides of $f'_{i\cdot}$. The probability $p_i(x)$ that x or more

sequencing errors occur in the ith fragment is given by the binomial distribution

$$p_i(x) = \mathrm{P}(X_i \geq x) = \sum_{z=x}^{n_i} \binom{n_i}{z} \varepsilon^z (1-\varepsilon)^{n_i - z}. \quad (4)$$

We first construct a significance test to infer if the aforementioned mismatch has been induced by mutations present in some of the informative sites. In particular, if $p_i(d_i)$ is smaller than or equal to a pre-specified P-value α, we declare that not all of the d_i mismatches are sequencing errors, implying at least one of them is due to mutation and therefore the ith read remains in $\mathbf{F}$. Otherwise, for a read such that $p_i(d_i) > \alpha$, we further examine its origin by considering the probability p_i that d_i sequencing errors occurs in the ith read,

$$p_i = \binom{n_i}{d_i} \varepsilon^{d_i} (1-\varepsilon)^{n_i - d_i}. \quad (5)$$

The reads which satisfy $p_i > \delta_i$ are assessed to have originated from the most dominant strain and thus eliminated from $\mathbf{F}$. The threshold δ_i is defined as $\delta_i = \prod_{j=1}^{n} p_{X_i[j]}$, where $X_i[j]$ is the nucleotide observed at position j of the ith read in $\mathbf{F}'$ and $p_{X_i[j]}$ denotes the probability of observing nucleotide $X_i[j]$ at position j; the latter probability is obtained as the empirical allele frequency distribution at position j. Note that if the mismatches between a read and the recovered sequence are due to mutations rather than sequencing errors, it is less likely that the read originated from the recovered haplotype and thus a higher threshold is used. To provide strong evidence against the null hypothesis of d_i sequencing errors, we set the P-value α to a small number; in particular, we set $\alpha = 10^{-5}$.

Finally, to reconstruct full-length strains q_i, we reinsert homozygous alleles into the reads and form a consensus sequence for each cluster. Abundances of the reconstructed strains are estimated by counting the number of reads in each cluster.

The performance comparison between TenSQR and the single-pass tensor factorization [i.e. AltHap (Hashemi *et al.*, 2017), an approach that does not employ successive data cancelation] can be found in the Supplementary Material B.

2.4 Determining the number of strains

The scheme outlined in this section requires that the number of strains (i.e. clusters) k be specified prior to performing tensor factorization and successive cancelation. To determine k, we consider the improvement rate of the MEC score defined as (Ahn and Vikalo, 2017)

$$\mathrm{MECimpr}(k) = \frac{\mathrm{MEC}(k) - \mathrm{MEC}(k+1)}{\mathrm{MEC}(k)}. \quad (6)$$

Recall that the MEC score counts the minimum number of nucleotides that need to be changed in the observed reads so that the modified reads are consistent with having originated from the reconstructed sequences; smaller MEC score indicates higher accuracy of clustering. As we increase k, the MEC score decreases monotonically; however, once k has reached the actual number of clusters, its further increase leads to only small improvements of the MEC score. Therefore, we approach the problem of detecting the number of strains by identifying k for which the MEC improvement rate (MECimpr) saturates. To detect this point, we compare the MEC improvement rate with a pre-defined threshold; while one can search for k by increasing it in steps of 1 until the MEC improvement rate saturates, we, for efficiency, update candidate k by relying on the so-called half-interval search. In particular, starting from an initial k_0, the number of clusters is updated as $k_\tau \leftarrow 2k_{\tau-1}$ until $\mathrm{MECimpr}(k_\tau) \leq \eta$; at this point the number of clusters starts to decrease as $k_{\tau+1} \leftarrow \lfloor (k_\tau + \max\{1, k_i\})/2 \rfloor$ where $\{i \in \{1, \ldots, \tau-1\} : k_i \leq k_\tau\}$. Once $\mathrm{MECimpr}(k_\tau) > \eta$, the number of clusters increases again as $k_{\tau+1} \leftarrow \lfloor (k_\tau + \min k_i)/2 \rfloor$ where $\{i \in \{1, \ldots, \tau-1\} : k_i > k_\tau\}$. If $k_\tau = k_{\tau-1}$, the search procedure stops by assigning $k_{\tau+1} \leftarrow k_\tau + 1$ which is our estimate of the number of strains. The recommended choice of the threshold η is discussed in (Ahn and Vikalo, 2017) where it has been demonstrated that the performance of estimating the number of strains via MEC improvement rate is robust with respect to the choice of the threshold. The described procedure will find the true number of strains starting from an arbitrary k_0; the closer k_0 is to the true number of strains, the fewer iteration will be needed. The proposed TenSQR method is formalized as Algorithm 1.

Algorithm 1: Tensor factorization with successive removal

Input: Set of reads R aligned to the reference genome
Output: Full length quasispecies Q and frequencies of k strains in Q
Pre-processing: From R, get mode-1 unfolding $\mathbf{F}$ of fragment tensor $\mathcal{F}$
Initial $\tau \leftarrow 0$, MECflag $\leftarrow 0$
while $\tau = 0$ or $k_\tau = k_{\tau-1}$ **do**
  for $k \in \{k_\tau, k_\tau + 1\}$ **do**
    $Q_k \leftarrow \varnothing$
    $\mathbf{F} \leftarrow$ mode-1 unfolding $\mathbf{F}$ of fragment tensor $\mathcal{F}$
    while $\mathbf{F} = \varnothing$ or $k \geq 1$ **do**
      Initialize $\mathbf{H}_0 \leftarrow V\Sigma^{\frac{1}{2}}$ where $U\Sigma V^\top = \mathrm{SVD}_k(\mathcal{P}_\Omega(\mathbf{F}))$
      Repeat
        $\mathbf{M}_{t+1} = \arg\min_{\mathbf{M}} \frac{1}{2} ||\mathcal{P}_\Omega(\mathbf{F} - \mathbf{M}\mathbf{H}_t^\top)||_F^2$
        $\mathbf{H}_{t+1} = \arg\min_{\mathbf{H}} \frac{1}{2} ||\mathcal{P}_\Omega(\mathbf{F} - \mathbf{M}_{t+1}\mathbf{H}^\top)||_F^2$
      Until termination criterion is met
      Identify $\mathcal{H}_{w::}$ s.t. $w = \arg\max_l m_l$
      Remove f_i s.t $p_i > \delta_i$ or $d_i = 0$ from $\mathbf{F}$
      Reconstruct $Q_k \leftarrow [Q_k; q_h]$ and estimate frequency of q_h
      $k \leftarrow k - 1$
    end while
    Calculate MEC(k)
  end for
  if $\mathrm{MECimpr}(k_\tau) \leq \eta$ **do**
    $k_{\tau+1} \leftarrow \lfloor (k_\tau + \max\{1, k_i\})/2 \rfloor, \{i \in \{1, \ldots, \tau-1\} : k_i \leq k_\tau\}$
    MECflag $\leftarrow 1$
  else do
    if MECflag $= 0$ **Do**
      $k_{\tau+1} \leftarrow 2k_\tau$
    else do
      $k_{\tau+1} \leftarrow \lfloor (k_\tau + \min k_i)/2 \rfloor, \{i \in \{1, \ldots, \tau-1\} : k_i > k_\tau\}$
    end if
  end if
  $\tau \leftarrow \tau + 1$
end while
$k \leftarrow k_\tau + 1$
$Q \leftarrow Q_k$

3 Results and discussion

3.1 Performance comparison on simulated data

We first test the performance of TenSQR on the synthetic data generated by emulating HTS of quasispecies samples. Viral strains in a quasispecies are generated by introducing independent mutations at uniformly random locations of a randomly generated reference genome of length 1300 bp (this is a typical length of a gene in the *pol* region of the HIV-1 genome). 2×250 bp-long Illumina's MiSeq reads with inserts that have average length and SD 150 bp and 30 bp, respectively, uniformly sample the mixture of viral strains. The reads are aligned to a reference using BWA-MEM algorithm with the default settings (Li and Durbin, 2009); reads shorter than 100 bp or having mapping quality score lower than 40 are filtered out. Simulated data is categorized into 40 different sets, each consisting of 50 samples, according to diversity (*div*%), sequencing error rate (ε) and the number of strains in a quasispecies (and hence frequencies of the strains). Diversity, defined as the average Hamming distance between different strains in a quasispecies, varies from 1% to 10%. Sequencing error rate is set to $\varepsilon = 7 \times 10^{-3}$ and $\varepsilon = 2 \times 10^{-3}$, the typical error rates in MiSeq datasets before and after quality trimming with error correction, respectively (Schirmer *et al.*, 2016). For each configuration of parameters we consider two mixture sets, each consisting of 5 and 10 viral strains. Frequencies of strains are chosen to approximately follow geometric distribution so as to emulate uneven populations which include strains with low abundance; relative strain frequencies for the 5-strain mix are $(0.5, 0.3, 0.15, 0.04, 0.01)$ while those for the 10-strain mix are $(0.36, 0.24, 0.16, 0.08, 0.055, 0.04, 0.03, 0.02, 0.01, 0.005)$. The coverage for the 5-strain and 10-strain population are $2500\times$ and $5000\times$, respectively, implying that strains having relative frequencies 0.0023 or higher in the 5-strain case and those with relative frequencies 0.0046 or higher in the 10-strain case are covered by sequencing reads with probability 0.99 (Eriksson *et al.*, 2008).

We compare the performance of TenSQR on the generated datasets with publicly available softwares PredictHaplo (Prabhakaran *et al.*, 2014), ShoRAH (Zagordi *et al.*, 2011), ViQuaS (Jayasundara *et al.*, 2015) and aBayesQR (Ahn and Vikalo, 2017), in terms of *Recall*, *Precision*, *Predicted Proportion*, *Reconstruction Rate* and *Jensen–Shannon divergence (JSD)*. To assess the ability of the compared methods to reconstruct viral strains perfectly (without errors), *Recall* is defined as the fraction of the reconstructed strains that match the true strains in a quasispecies, i.e. $Recall = \frac{TP}{TP+FN}$ while *Precision* reports the fraction of true sequences among the reconstructed strains, i.e. $\text{Precision} = \frac{TP}{TP+FP}$. We further report *Predicted Proportion*, defined as the ratio of the estimated and the true population size, thus measuring accuracy of the methods' population size prediction. Note that the proximity of the value of this metric to 1 indicates accuracy of the population size estimate. To assess the degree of accuracy of each reconstructed strain, we use

$$\text{Reconstruction Rate} = \frac{1}{k}\sum_{i=1}^{k}\left(1 - \frac{HD(q_i, \widehat{q}_i)}{G}\right),$$

where G denotes the length of a reference genome, k is the number of strains in a quasispecies, and q_i and $\widehat{q}_i$ denote the ith true strain and its nearest sequence among the k reconstructed ones, respectively. Finally, *JSD* measures accuracy of the estimated frequencies of the reconstructed strains, i.e. quantifies similarity between two inferred distributions. Formally, *JSD* between a true distribution P and its approximation Q is defined as

$$\text{JSD } (P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M),$$

where $D(\cdot||\cdot)$ denotes Kullback–Leibler (KL) divergence, M is defined as $M = \frac{1}{2}(P+Q)$ and the KL divergence is found as $D(P||Q) = \sum_{i=1}^{n} P(i) \log \frac{P(i)}{Q(i)}$.

Figures 2 and 3 compares the values of these five metrics computed for each of the considered reconstruction methods; the metrics are evaluated by averaging over 50 instances for each combination of the parameters and error rate $\varepsilon = 2 \times 10^{-3}$ and $\varepsilon = 7 \times 10^{-3}$, respectively. Note that PredictHaplo does not execute on all instances of the generated datasets and hence we report its performance only when all 50 instances are successfully processed. As can be seen in Figures 2 and 3, TenSQR outperforms the competing schemes. In particular, TenSQR performs the best at all levels of diversity in terms of both *Recall* and *Reconstruction Rate*. In terms of *Predicted Proportion* and *JSD*, at $div > 2\%$ TenSQR achieves superior or comparable performance to aBayesQR, which is designed to particularly excel at reconstructing low diversity populations. Note that while *Recall* quantifies the fraction of perfectly reconstructed viral strains, purpose of *Reconstruction Rate* is to assess quality of reconstruction when the assembled viral strains are allowed to have errors in some positions. Therefore, the fact TenSQR's *Recall* and *Reconstruction Rate* are close to 1 indicates that the proposed scheme is capable of reconstructing rare sequences (i.e. with low abundance) present in viral mixtures characterized by a wide range of strain frequencies. PredictHaplo underestimates the number of strains and reconstructs only those that have relative frequencies $\geq 15\%$, which explains its high *Precision* at $div \geq 5\%$. ViQuaS overestimates the population size at all levels of diversity, achieving the highest scores in *Predicted Proportion*; note that the only strains used in calculating ViQuaS performance metrics are those that ViQuaS estimated as having frequencies greater than $f_{\min}$, as recommended by (Jayasundara *et al.*, 2015). The strains reconstructed by ShoRAH are consistently shorter than the true strains, which appears to be due to the existence of low coverage regions in the synthetic datasets. ShoRAH completes missing sites on the reconstructed strains using bases from the reference genome, which partially explains why ShoRAH underperforms in terms of *Recall*, *Precision* and *Reconstruction Rate*. In conclusion, low *Predicted Proportion* of PredictHaplo and weak performance of other methods in terms of *Recall* and *Reconstruction Rate* indicate that existing techniques experience major difficulties when attempting to detect and reconstruct rare strains.

3.2 Evaluating identification of deletion

Following the comparison of performance of TenSQR to state-of-the-art methods, we next evaluate how accurate is TenSQR at estimating long deletions. In particular, we investigate TenSQR's ability to detect a fixed-length deletion in a strain over a range of strain frequencies and diversity levels. To this end, we generate sets of quasispecies consisting of two strains where the length of the abundant strain is 1300 bp and deletions of sizes $l_{\text{del}} = 100$ bp, 200 bp and 300 bp are placed into the strain of the lower abundance. We generated 40 benchmark sets of reads emulating sequencing of a mixture of two viral strains with diversity ranging from 1% to 10% and the lower of two frequencies taking values in {1%, 2%, 5% and 10%}. The coverage for the mixture of two viral stains is $1000\times$ and the

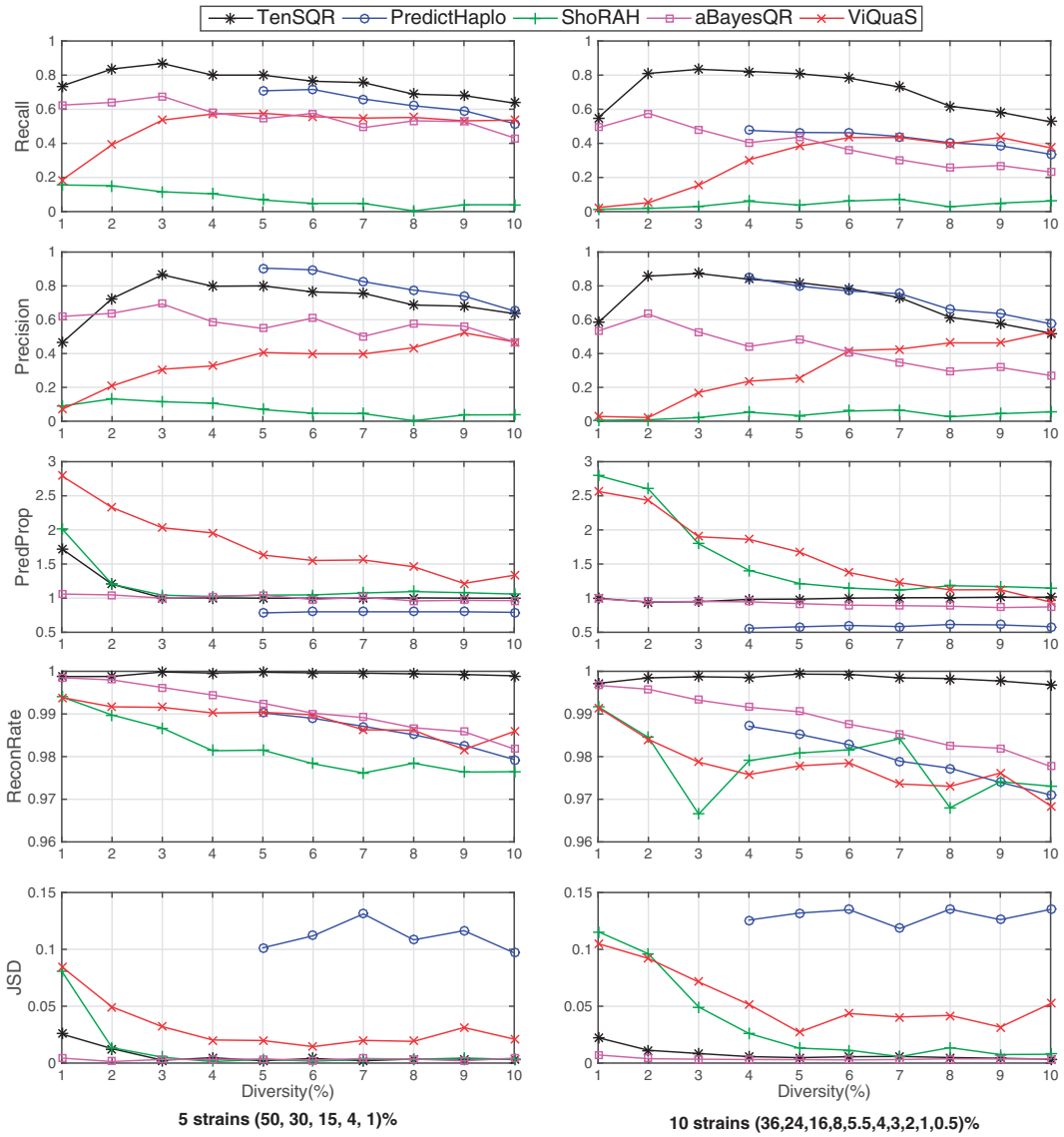

Fig. 2. Performance comparison of TenSQR, aBayesQR, ShoRAH, ViQuaS and PredictHaplo in terms of *Recall*, *Precision*, *Predicted Proportion* (*PredProp*), *Reconstruction Rate* (*ReconRate*) and *JSD* on the simulated data with $\varepsilon = 2 \times 10^{-3}$ for a mixture of (**a**) 5 viral strains and (**b**) 10 viral strains. (For the plots that include error bars, please see the corresponding Supplementary Fig. S2 in Supplementary Material B)

sequencing error rate is set to $\varepsilon = 2 \times 10^{-3}$. Instances of 50 are generated for each of the total 120 datasets. In this study, the performance of detecting deletions is characterized by means of the false negative rate evaluated over 50 instances of the experiment, and the deviations of the estimated length of deletions from the true length calculated by averaging the deviations over the 50 instances. Since the competing QSR methods considered in Section 3.1 are unable to detect deletions, we only show the performance of TenSQR. Note that while HaploClique (Töpfer *et al.*, 2014) can predict long deletions, the overlap assembly approach recovers many contigs shorter than true strains instead of reconstructing full-length strains and thus was not included in the benchmarking results. As evident from the results in Table 1, TenSQR is capable of detecting long deletions existing in the viral strains whose frequencies are as low as 1%. The performance of detecting long deletions is exceptional when the viral mixture is not particularly characterized by low diversity (div = 1%); in this setting, the performance under the short sequencing reads (2×250) tends to deteriorate as the length of deletions increases.

Remark 1: The proposed approach to detecting deletions is tested on a mixture of only two strains for the sake of clarity, so as to allow investigation of the interplay between diversity, strain frequency and deletion length. The method is, however, applicable to more general settings that involve multiple strains and/or multiple deleted regions (omitted for brevity).

Remark 2: An additional pipeline designed to detect insertions that may exist in some of the reconstructed strains and preliminary results demonstrating its performance are given in Supplementary Material C.

3.3 Performance comparison on gene-wise reconstruction of real HIV-1 data

We further test the performance of TenSQR on the real HIV dataset made publicly available by (Di Giallonardo *et al.*, 2014). An *in vitro* generated quasispecies population consists of 5 known HIV-1 strains (HIV-1_{HXB2}, HIV-$1_{89.6}$, HIV-1_{JR-CSF}, HIV-1_{NL4-3} and HIV-1_{YU2}) with pairwise distances between 2.61 – 8.45% and relative

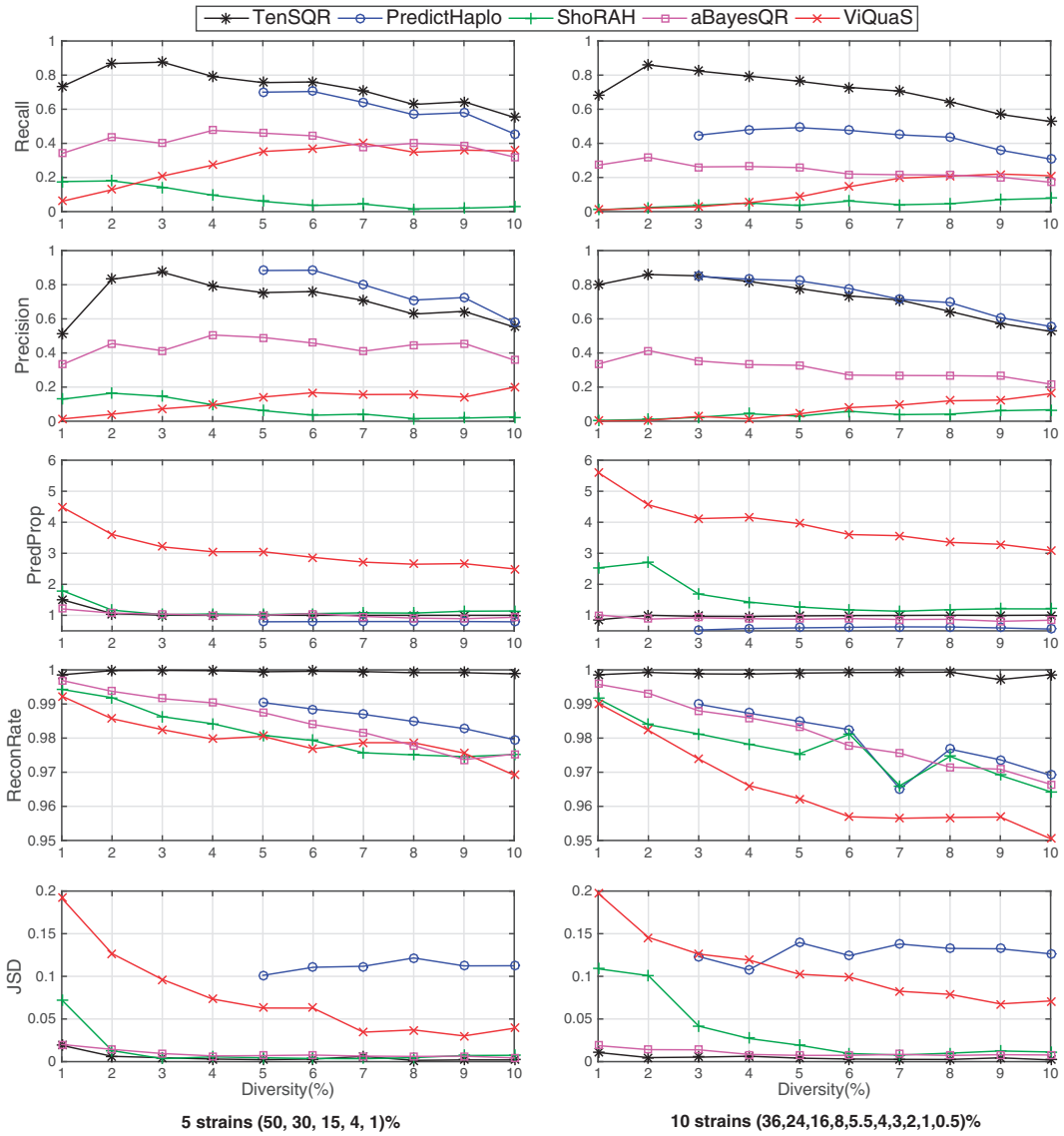

Fig. 3. Performance comparison of TenSQR, aBayesQR, ShoRAH, ViQuaS and PredictHaplo in terms of *Recall*, *Precision*, *Predicted Proportion* (*PredProp*), *Reconstruction Rate* (*ReconRate*) and *JSD* on the simulated data with $\varepsilon = 7 \times 10^{-3}$ for a mixture of (**a**) 5 viral strains and (**b**) 10 viral strains. (For the plots that include error bars, please see the corresponding Supplementary Fig. S3 in Supplementary Material B)

frequencies between 10% and 30% (Di Giallonardo *et al.*, 2014). Paired-end sequencing reads of length 2×250 bp generated by Illumina's MiSeq Benchtop Sequencer are aligned to the HIV-1_{HXB2} reference genome. Following Di Giallonardo *et al.*, 2014, to ensure reliable results, reads shorter than 150 bp and having quality scores of mapping lower than 60 are discarded. We first apply TenSQR to gene-wise reconstruction of the HIV population and compare its performance to that of aBayesQR (Ahn and Vikalo, 2017) and PredictHaplo (Prabhakaran *et al.*, 2014), shown to be the most competitive softwares in the benchmarking studies in Section 3.1. *Predicted Proportion*, defined in Section 3.1 as the ratio of the estimated and the true population size, is evaluated by setting the parameter needed to detect the number of strains in a mixture to $\eta = 0.09$, as recommended by Ahn and Vikalo, 2017. Since the ground truth information specifying the five HIV strains is available (http://bmda.cs.unibas.ch/HivHaploTyper/), we evaluate *Reconstruction Rate* for each recovered individual strain, along with the inferred strain frequencies. Note that the strains in the HIV-1 dataset are more evenly distributed than those in the simulated quasispecies in Section 3.1. The results reported in Table 2 show that TenSQR reconstructs all of the five HIV-1 strains correctly in six genes while the other considered methods accomplish the same in five genes. Consistent with the results in Section 3.1, PredictHaplo, designed for identification of HIV haplotypes, underestimates the number of strains by reconstructing three or four strains in the eight genes.

3.4 Assembly of HIV-1 gag-pol genomes

We further use TenSQR to reconstruct the HIV population on the 4036 bp long gag-pol region. Reconstruction of longer regions of viral quasispecies requires higher sequencing coverage than that needed for shorter ones. Therefore, for a reliable reconstruction of a viral population spanning long genome region, we fragment the long region into overlapping blocks, perform reconstruction of the blocks independently and merge the results to retrieve the full region of interest. Specifically, we split the HIV gag-pol region into a set of

Table 1. Performance of estimating deletion

ldel	f_{min}%	div% 1	2	3	4	5	6	7	8	9	10
	1	0.08 (0.80)	0 (0.88)	0 (1.02)	0 (2.16)	0 (2.64)	0 (2.20)	0 (2.52)	0 (3.38)	0 (4.76)	0 (4.08)
100	2	0.02 (1.18)	0 (1.26)	0 (1.70)	0 (1.42)	0 (2.66)	0 (1.58)	0 (2.78)	0 (3.14)	0 (2.72)	0 (4.22)
	5	0.06 (1.64)	0 (1.72)	0 (2.12)	0 (1.94)	0 (1.86)	0 (2.20)	0 (2.22)	0 (2.94)	0 (3.24)	0 (3.70)
	10	0.08 (2.18)	0 (2.62)	0 (2.20)	0 (1.90)	0 (2.42)	0 (1.92)	0 (2.00)	0 (3.34)	0 (3.00)	0 (2.50)
	1	0.16 (1.33)	0 (1.38)	0 (1.20)	0 (1.78)	0 (2.46)	0 (3.02)	0 (2.28)	0 (4.02)	0 (5.56)	0 (4.56)
200	2	0.14 (1.26)	0 (1.20)	0 (1.94)	0 (1.74)	0 (1.78)	0 (2.34)	0 (1.98)	0 (3.58)	0 (3.94)	0 (4.36)
	5	0.14 (2.86)	0 (1.67)	0 (1.90)	0 (2.10)	0 (2.20)	0 (1.96)	0 (2.54)	0 (3.16)	0 (3.48)	0 (4.98)
	10	0.16 (2.31)	0 (2.20)	0 (2.06)	0 (2.30)	0 (1.70)	0 (2.10)	0 (2.18)	0 (2.98)	0 (2.54)	0 (2.96)
	1	0.34 (1.00)	0 (1.62)	0 (1.38)	0 (2.24)	0 (2.04)	0 (3.32)	0 (2.52)	0 (4.56)	0 (5.14)	0 (5.00)
300	2	0.24 (2.00)	0.06 (1.64)	0 (1.78)	0 (1.70)	0 (2.02)	0 (2.24)	0 (2.28)	0 (3.74)	0 (5.08)	0 (3.92)
	5	0.30 (2.66)	0.02 (2.12)	0 (1.92)	0.02 (1.65)	0 (2.04)	0 (2.52)	0 (2.88)	0 (2.84)	0 (2.58)	0 (3.60)
	10	0.36 (2.72)	0 (2.60)	0 (2.62)	0 (2.10)	0 (2.28	0 (2.64)	0 (2.94)	0 (2.58)	0 (3.10)	0 (3.64)

Note: Performance of TenSQR of estimating deletion in terms of *False Negative* rate of detecting deletions and *Deviation* of estimated deletion length (in parenthesis) on the simulated data with $\varepsilon = 2 \times 10^{-3}$ and 1000× coverage for a mixture of two strains, depending on diversity (div) and frequency of the low abundant strain (f_{min}) which includes a deletion of length (l_{del}) 100 bp, 200 bp and 300 bp.

Table 2. Performance comparisons of TenSQR, aBayesQR and PredictHap on a real HIV-1 5-virus-mix data

		p17	p24	p2-p6	PR	RT	RNase	int	vif	vpr	vpu	gp120	gp41	nef
TenSQR	PredProp	1	1.6	1	1	1.4	1	1	1	1	1.6	2.2	1.2	0.8
	RR_{HXB2}	**100**	98.9	**100**	100	99.2	**100**	**100**	**100**	**100**	92.8	96.0	99.0	0
	$RR_{89.6}$	**100**	100	**100**	100	98.0	**100**	**100**	**100**	**100**	94.0	97.2	100	95.7
	$RR_{JR\text{-}CSF}$	**100**	100	**100**	100	100	**100**	**100**	**100**	**100**	100	98.3	97.7	99.8
	$RR_{NL4\text{–}3}$	**100**	99.3	**100**	100	99.5	**100**	**100**	**100**	**100**	100	99.8	99.5	99.7
	RR_{YU2}	**100**	99.3	**100**	99.7	99.7	**100**	**100**	**100**	**100**	100	94.9	100	98.6
aBayesQR	PredProp	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	1.2	**1**	0.8	0.8	1.2
	RR_{HXB2}	**100**	99.4	**100**	**100**	98.5	**100**	99.9	100	**100**	99.6	98	0	95.8
	$RR_{89.6}$	**100**	98.7	**100**	**100**	98.6	**100**	100	100	**100**	92	96.5	98.9	95.5
	$RR_{JR\text{-}CSF}$	**100**	99.6	**100**	**100**	99	**100**	100	100	**100**	98.8	97.7	99.1	98.2
	$RR_{NL4\text{–}3}$	**100**	100	**100**	**100**	98.9	**100**	100	99.8	**100**	100	96.3	98.8	100
	RR_{YU2}	**100**	99.7	**100**	**100**	99.2	**100**	99.5	99.7	**100**	100	0	98.6	99.2
PredictHaplo	PredProp	**1**	0.6	**1**	**1**	**1**	0.8	0.8	0.8	**1**	0.8	0.8	0.8	0.8
	RR_{HXB2}	**100**	0	**100**	**100**	**100**	98.9	100	100	**100**	93.2	0	0	0
	$RR_{89.6}$	**100**	100	**100**	**100**	**100**	100	99.8	100	**100**	0	97.8	100	98.87
	$RR_{JR\text{-}CSF}$	**100**	100	**100**	**100**	**100**	100	100	100	**100**	100	99.7	100	100
	$RR_{NL4\text{–}3}$	**100**	99.1	**100**	**100**	**100**	100	100	100	**100**	100	100	100	100
	RR_{YU2}	**100**	0	**100**	**100**	**100**	0	0	0	**100**	100	98.6	100	100

Note: Predicted Proportion (PredProp) and Reconstruction Rate [RR (%)] for TenSQR, aBayesQR and PredictHaplo applied to reconstruction of HIV-1HXB2, HIV-189.6, HIV-1JR-CSF, HIV-1NL4-3 and HIV-1YU2 for all 13 genes of the HIV-1 dataset. Values in the genes where all the strains are perfectly reconstructed without errors are denoted as boldface. (The inferred frequencies are shown in Table S2 in Supplementary Material B).

blocks of length 500 bp, where the consecutive blocks overlap by 250 bp. We run TenSQR to perform reconstruction of the viral strains in each of the 18 blocks independently, and merge the results in the consecutive blocks while testing consistency of the strains in the overlapping intervals. In particular, if there are mismatches between the reconstructed strains in the overlapping regions of consecutive blocks, we resolve them by performing majority voting using reads that are covering the mismatched positions. The number of strains retrieved by the global reconstruction procedure is decided via majority voting over the number of strains reconstructed in each block. Following this procedure on the 355 241 paired-end reads that remain after applying a quality filter, TenSQR perfectly reconstructed all 5 HIV-1 strains, achieving *Reconstruction Rate* of 100 for all 5 strains and *Predicted Proportion* of 1. Since the pairwise distances between the 5 HIV-1 strains are relatively high, we estimated frequencies of the viral strains by simply counting the number of reads assigned to the same strain according to the Hamming distance between the reads and the reconstructed strains. The resulting frequencies are 15.21%, 19.34%, 25.56%, 27.61% and 12.27%, which is consistent with the result obtained by aBayesQR (Ahn and Vikalo, 2017).

3.5 Assembly of the ZIKV strains

We apply TenSQR to reconstruct the full-length genome of an Asian-lineage ZIKV sample (accession SRR3332513) that is obtained from a rhesus macaque (animal 393422) on the 4th day of infection (Dudley *et al.*, 2016). 2 × 300 bp reads (∼30 000× coverage) are generated from the sample by the Illumina's MiSeq platform and aligned to the ZIKV reference genome (Genbank KU681081.3) of length 10 807 bp using the BWA-MEM algorithm with the default settings (Li and Durbin, 2009). The reads shorter than 100 bp and those having mapping quality scores smaller than 40 are discarded. For reliable reconstruction of the full-length

genome, we follow the strategy outlined in Section 3.4; the full region is split into a sequence of blocks of length 2500 bp where the consecutive windows overlap by 500 bp. We run TenSQR on those blocks and assemble the entire region by connecting reconstructed strains in the consecutive blocks. Relative strain frequencies are estimated using an expectation-maximization algorithm described in (Eriksson *et al.*, 2008). Applying the described procedure to 565 979 paired-end reads that pass the quality filter, TenSQR reconstructed two full-length viral sequences with relative abundances 74% and 26% that diverge from each other by 0.47% within regions between 200 bp and 5550 bp. Among all the competing methods considered in Section 3.1, PredictHaplo is the only one that completed reconstruction within 48 h. PredictHaplo, which typically underestimates the number of strains—especially in quasispecies characterized by low diversity (Ahn and Vikalo, 2017)—reconstructed only one strain which matches the dominant strain reconstructed by TenSQR.

4 Conclusion

In this article, we presented a novel tensor factorization based algorithm for the reconstruction of viral quasispecies from HTS data. In particular, sequencing data is represented by a sparse binary tensor and the viral strains in a quasispecies are reconstructed in an iterative manner; at each iteration, the most abundant sequence among those obtained by factorizing the tensor is selected and data originated from the most abundant sequence is removed from the tensor. Benchmarking tests on synthetic datasets demonstrate that the proposed scheme, referred to as TenSQR, is capable of reconstructing quasispecies characterized by imbalanced frequencies of strains, detecting and recovering low abundant strains more reliably than state-of-the-art algorithms. Further studies on a real HIV-1 and Zika dataset demonstrate that TenSQR outperforms existing methods in more general settings and is applicable to quasispecies reconstruction from virus-infected patient samples.

Future work will include the development of an improved methodology that permits accurate recovery of long insertions potentially present in rare viral strains.

Funding

This work was supported by the National Science Foundation under grants CCF 1320273 and CCF 1618427.

Conflict of Interest: none declared.

References

Ahn,S. and Vikalo,H. (2017) aBayesQR: a Bayesian method for reconstruction of viral populations characterized by low diversity. In: *International Conference on Research in Computational Molecular Biology*. Springer, Hong Kong, pp. 353–369.

Astrovskaya,I. *et al.* (2011) Inferring viral quasispecies spectra from 454 pyrosequencing reads. *BMC Bioinformatics*, **12**, S1.

Baaijens,J.A. *et al.* (2017) De novo assembly of viral quasispecies using overlap graphs. *Genome Res.*, **27**, 835–848.

Beerenwinkel,N. *et al.* (2012) Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Front. Microbiol.*, **3**, 329.

Cai,C. *et al.* (2016) Structured low-rank matrix factorization for haplotype assembly. *IEEE J. Selected Topics Signal Process.*, **10**, 647–657.

Carroll,S.A. *et al.* (2013) Molecular evolution of viruses of the family filoviridae based on 97 whole-genome sequences. *J. Virol.*, **87**, 2608–2616.

Chaisson,M.J. *et al.* (2017) Resolving multicopy duplications de novo using polyploid phasing. In: *International Conference on Research in Computational Molecular Biology*. Springer, Hong Kong, pp. 117–133.

Di Giallonardo,F. *et al.* (2014) Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. *Nucleic Acids Res.*, **42**, e115–e115.

Dudley,D.M. *et al.* (2016) A rhesus macaque model of asian-lineage zika virus infection. *Nat. Commun.*, **7**, 12204.

Eriksson,N. *et al.* (2008) Viral population estimation using pyrosequencing. *PLoS Comput. Biol.*, **4**, e1000074.

Hashemi,A. *et al.* (2017) Sparse tensor decomposition for haplotype assembly of diploids and polyploids. *bioRxiv*, 130930.

Jain,P. *et al.* (2013) Low-rank matrix completion using alternating minimization. In: *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, ACM, Palo alto, CA, USA, pp. 665–674.

Jayasundara,D. *et al.* (2015) Viquas: an improved reconstruction pipeline for viral quasispecies spectra generated by next-generation sequencing. *Bioinformatics*, **31**, 886–896.

Lauring,A.S. and Andino,R. (2010) Quasispecies theory and the behavior of rna viruses. *PLoS Pathogens*, **6**, e1001005.

Le,T. *et al.* (2009) Low-abundance hiv drug-resistant viral variants in treatment-experienced persons correlate with historical antiretroviral use. *PloS One*, **4**, e6079.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Lippert,R. *et al.* (2002) Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. *Brief. Bioinformatics*, **3**, 23–31.

Malhotra,R. *et al.* (2015) *Maximum Likelihood De Novo Reconstruction of Viral Populations Using Paired End Sequencing Data. arXiv preprint arXiv: 1502.04239*.

Mangul,S. *et al.* (2014) Accurate viral population assembly from ultra-deep sequencing data. *Bioinformatics*, **30**, i329–i337.

Posada-Cespedes,S. *et al.* (2016) Recent advances in inferring viral diversity from high-throughput sequencing data. *Virus Res.*, **239**, 17–32.

Prabhakaran,S. *et al.* (2014) Hiv haplotype inference using a propagating dirichlet process mixture model. *IEEE/ACM Trans. on Comput. Biol. Bioinform. (TCBB)*, **11**, 182–191.

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

Schirmer,M. *et al.* (2014) Benchmarking of viral haplotype reconstruction programmes: an overview of the capacities and limitations of currently available programmes. *Brief. Bioinformatics*, **15**, 431–442.

Schirmer,M. *et al.* (2016) Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*, **17**, 125.

Simen,B.B. *et al.* (2009) Low-abundance drug-resistant viral variants in chronically hiv-infected, antiretroviral treatment–naive patients significantly impact treatment outcomes. *J. Infectious Dis.*, **199**, 693–701.

Töpfer,A. *et al.* (2013) Probabilistic inference of viral quasispecies subject to recombination. *J. Comput. Biol.*, **20**, 113–123.

Töpfer,A. *et al.* (2014) Viral quasispecies assembly via maximal clique enumeration. *PLoS Comput. Biol.*, **10**, e1003515.

Zagordi,O. *et al.* (2010) Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. *J. Comput. Biol.*, **17**, 417–428.

Zagordi,O. *et al.* (2011) Shorah: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics*, **12**, 119.