Genomics 110 (2018) 375–381

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

QSdpR: Viral quasispecies reconstruction via correlation clustering

Somsubhra Barik[a,*], Shreepriya Das[b], Haris Vikalo[a]

[a] ECE Department, The University of Texas at Austin, Austin, TX 78712, United States
[b] Department of Systems Biology, Harvard Medical School, Boston, MA 02115, United States

ARTICLE INFO

Keywords:
Quasispecies
Clustering
Max *K*-cut
Next generation sequencing
RNA viruses

ABSTRACT

RNA viruses are characterized by high mutation rates that give rise to populations of closely related genomes, known as viral quasispecies. Underlying heterogeneity enables the quasispecies to adapt to changing conditions and proliferate over the course of an infection. Determining genetic diversity of a virus (i.e., inferring haplotypes and their proportions in the population) is essential for understanding its mutation patterns, and for effective drug developments. Here, we present QSdpR, a method and software for the reconstruction of quasispecies from short sequencing reads. The reconstruction is achieved by solving a correlation clustering problem on a read-similarity graph and the results of the clustering are used to estimate frequencies of sub-species; the number of sub-species is determined using pseudo F index. Extensive tests on both synthetic datasets and experimental HIV-1 and Zika virus data demonstrate that QSdpR compares favorably to existing methods in terms of various performance metrics.

1. Introduction

RNA polymerase, an enzyme responsible for viral genome replication, exhibits high error rates causing relatively frequent point mutations in viral genomic sequences. As a result, RNA viruses typically exist as collections of non-identical but closely related variants inside the host cells. The diversity of viral populations, often referred to as viral quasispecies [1], adversely affects antiviral drug therapies and renders vaccine designs challenging [2], thus motivating their close studies. The *quasispecies spectrum reconstruction* (QSR) problem involves both the reconstruction of individual sequences in a population as well as the estimation of their relative abundances. Presence of sequencing errors in high-throughput sequencing (HTS) reads, limited read lengths, and small genetic distances among viral sequences render QSR a hard problem to solve, even when sequencing coverage is high. Although conceptually similar to the single individual haplotyping problem, QSR has major additional challenges – the number of individual haplotypes is *a priori* unknown and the point mutations are in general poly-allelic rather than bi-allelic [3] (additionally, short indel errors may be present).

Existing approaches for solving the QSR problem include Bayesian inference methods such as ShoRAH [4], the non-parametric Bayesian approach based on a Dirichlet process mixture model in [5] named PredictHaplo, a hidden Markov model based Quasirecomb [6], max-clique enumeration technique on read alignment graphs named HaploClique [3], reconstruction method based on multinomial distributions named QuRe [7], graph-coloring based heuristic VGA [8], and the reference assisted *de-novo* assembly reconstruction method named ViQuaS [9]. Generally, these methods can be categorized as read-graph based [3,7,8,10], probabilistic inference based [4–6] and *de-novo* assembly based techniques [9]. Quasispecies reconstruction methods may employ high-fidelity sequencing protocols [8] and barcode-tagging of genomes [11] to facilitate grouping of reads in populations. Recently, an algorithm for single individual haplotyping which approximately solves a semidefinite programming relaxation of the max *K*-cut problem on a read similarity graph was proposed [12]. Building upon that framework, we here present a method and software for viral quasispecies reconstruction, QSdpR (Quasispecies assembly with Semidefinite program Relaxation), that accurately and efficiently detects the number of sequences in a viral population, reconstructs their genomes, and determines their frequencies. QSdpR processes data represented by a read-similarity graph where the nodes representing the reads are connected by edges representing read overlaps. We test the performance of QSdpR on synthetic datasets emulating varying frequency, coverage and nucleotide diversities, as well as on a real data set introduced in [13] comprising Illumina MiSeq reads from a mixture of 5 cloned HIV-1 strains present at non-uniform proportions. The performance of QSdpR in terms of the *minimum error correction* (MEC) scores [14], *Reconstruction Proportions*, *Reconstruction Errors* and *Frequency Deviation Errors* is compared with several existing methods. The

Abbreviations: RNA, Ribonucleic acid; HTS, High-throughput sequencing; MEC, Minimum error correction; SNV, Single nucleotide variant; HIV, Human immunodeficiency virus

[*] Corresponding author.
E-mail address: sbarik@utexas.edu (S. Barik).

https://doi.org/10.1016/j.ygeno.2017.12.007
Received 5 September 2017; Received in revised form 3 December 2017; Accepted 13 December 2017
Available online 19 December 2017
0888-7543/

benchmarking results demonstrate that QSdpR compares favorably with PredictHaplo, ShoRAH and ViQuaS. To demonstrate the applicability of QSdpR to virus-infected patient sample data, we analyze Illumina MiSeq reads obtained from rhesus macaques infected with Zika virus stock H/PF/2013 and perform its full-length genome reconstruction. The code for QSdpR is available for download from https://sourceforge.net/projects/qsdpr/. (Notation: Matrices are represented by uppercase bold letters and vectors by lowercase bold letters. For a matrix $\mathbf{M}$, $\mathbf{M}^{(i)}$ and $\mathbf{M}_i$ represent its ith row and ith column, respectively (regarded as column vectors). M_{ij} denotes the (i,j)th entry of the matrix $\mathbf{M}$ and u_i denotes the ith entry of the vector $\mathbf{u}$. Each vector is a column vector unless noted otherwise. x_+ denotes $\max(x,0)$. For the set S, the number of elements in S is denoted by $|S|$.)

2. Materials and methods

2.1. Materials

2.1.1. Synthetic data

In the first part of benchmarking tests, we synthesize datasets by emulating high-throughput sequencing of haplotypes present at uniform and non-uniform proportions, and with a range of population sizes. First, we consider a dataset *S1* of 5 viral strains at uniform proportions (20% each), generated by introducing SNVs in 1 out of 100 independent and uniformly random locations along a randomly generated reference sequence of length 10,000 bp. Paired-end reads of length 2 × 350 bp and 1000 ± 100 bp inserts are generated with *Grinder* [15] at an effective sequencing coverage 150 ×, where the effective sequencing coverage is defined as the total number of reads from all strains, times the fraction of reference covered by each read. Reads have an error rate of 1% (typical of Illumina's platforms [16]). A second dataset, *S2* is simulated that has reads at coverage 200 × from a uniform mixture of 10 viral strains, where other parameters are same as those used in *S1*. Two additional datasets, *S3* and *S4*, are simulated that contain 5 and 10 sequences with geometric (ratio 1/2) and (approximately) linearly spaced proportions, at 200 × and 400 × coverages, respectively, while keeping other parameters unchanged. To further test the performance of QSdpR, we simulated a dataset *S5* mimicking reads generated by sequencing an HIV-1 *in vitro* population [13] that consists of 5 sequences of length 1000 bp (typical length of a gene in the *pol* region of the HIV-1 genome) with a SNV rate of 0.0986[1]. Paired-end reads of length 237 ± 1% bp and 250 bp long inserts are simulated at an average coverage of 2000 ×; sequencing error rate is identical to that for datasets *S1–S4*. Frequencies of the haplotypes are set according to those in [13]. All datasets are summarized in Table S1 of Supplementary file 1.

In the second part of the simulation studies, we synthesized datasets to assess the performance of QSdpR for sequences with length varying between 1000 bp and 2500 bp, in steps of 500 bp. To this end, 2 × 150 bp long paired-end reads with 200 bp inserts are simulated; these reads sample a quasispecies population of 5 haplotypes present at uniform proportions at an effective coverage of 100 ×, keeping other parameters as before. For each value of the reference length, we synthesized 10 datasets and reported performance metrics averaged over 10 runs. These datasets are referred to as *L1–L4* (Table S2 of Supplementary file 1). Pre-processing steps for all datasets are described in Section S1.2 of Supplementary file 1.

2.1.2. HIV-1 virus mix data

For experimental data, we consider the HIV-1 *Five Virus Mix* experimental dataset from [13] and used the HTS reads generated by Illumina's MiSeq sequencing platform. The data set consists of reads from an *in vitro* mixture of 5 known HIV-1 strains, namely, HIV-1 89.6, HXB2, JR-CSF, NL4-3 and YU2. The paired-end reads are on average 237 bp long (standard deviation of 26 bp), with an average coverage of 23,000 reads per base; they are aligned to the *HXB2* genome and are obtained from the Genbank accession number **SRP029432**. After performing variant calling, 958 SNVs are identified along the reference genome; among these, 690 SNVs are located within the various genic regions of interest. Sequencing depth for this dataset is highly non-uniform, and the data is poly-allelic at a number of SNV sites (see Supplementary file 1, Figs. *S1–S2*). QSdpR is applied to the multiple gene regions comprising the HIV-1 genome (see Supplementary file 1, Table S2) for benchmarking of performance [13,17].

2.2. Methods

2.2.1. System model

Let $\mathscr{Q} = \{q_1, q_2, ..., q_K\}$ denote the set of K sequences present in a quasispecies. The q_i's are nucleotide strings of identical length that differ from each other at a number of variant sites. In our model, we assume that differences between the member sequences in a quasispecies are due to substitutions or *single nucleotide variants* (SNVs)[2]. However, performance of QSdpR in the presence of insertions/deletions (besides substitutions) is also presented in the Results section. Let $R = \{r_1, r_2, \ldots, r_{|R|}\}$ denote the set of reads acquired by a HTS platform in a shotgun sequencing experiment; relative ordering of the reads is determined by mapping them to a reference genome. Note that the HTS reads (e.g., from Illumina's MiSeq or HiSeq platforms) are typically much shorter than the sequences in the quasispecies. The homozygous sites (i.e., the sites at which all sequences contain the same allele) are not used by the proposed quasispecies reconstruction method and therefore the corresponding bases are removed from the read data. Following the variant calling step, the reads covering only one SNV are discarded since they are not helpful for the subsequent phasing of the strains. In particular, when multiple viral strains share the allele at a variant site, it is not possible to unambiguously assign a single-SNV read to a strain.[3] Let there be ℓ variant sites that are retained after performing the above pre-processing of sequencing data. Then, each q_k can be thought of as a string of alleles of length ℓ, while each read r_i is a short, randomly located, not necessarily contiguous (due to inserts between paired-end reads) and potentially erroneous sub-string of one of the q_k's. The essential goal of viral quasispecies reconstruction is to segment these reads into as many clusters as there are viral haplotypes (namely, K) so that each cluster consists of reads that originated from a specific sequence.

2.2.2. Quasispecies reconstruction as a correlation clustering problem

The previously mentioned clustering problem can be formalized by introducing a weighted and undirected correlation graph $\mathscr{G} = (\mathscr{V}, \mathscr{E})$, where each vertex in the set $\mathscr{V}$ corresponds to a read $r_i \in R$, and each edge e_{ij} in the edge set $\mathscr{E}$ exists due to an overlap between r_i and r_j. The weight or correlation associated with e_{ij}, denoted as ω_{ij}, is defined as

$$\omega_{ij} = \begin{cases} 0, & \text{if } r_i \text{ and } r_j \text{ do not share SNVs,} \\ \dfrac{s_{ij} - t_{ij}}{s_{ij} + t_{ij}}, & \text{otherwise,} \end{cases} \tag{1}$$

where s_{ij} and t_{ij} denote the number of matches and mismatches at the overlapping variant sites of r_i and r_j, respectively. Large w_{ij} implies that r_i and r_j originate from the same haplotype while small w_{ij} implies the opposite. Note that graph $\mathscr{G}$ is sparse since the reads are much shorter than the genomic region of interest and hence each read overlaps with relatively few other reads. The objective of clustering is to divide the

[1] SNV rate is inferred by analyzing haplotypes in the ground truth data [13].

[2] Throughout the manuscript, SNVs and "variant sites" are used interchangeably when referring to substitution errors.

[3] Note that in the experimental HIV-1 dataset that we analyzed, approximately 99.67% reads cover more than one SNV and thus the fraction of discarded reads is very small.

vertices $v \in \mathscr{V}$ into K clusters such that the sum of the weights of edges connecting vertices within clusters is maximized while the sum of the weights of edges connecting vertices across clusters is minimized. Note that sequencing errors in reads cause weights ω_{ij} to deviate from their true values, thereby making the clustering problem non-trivial and computationally challenging.

QSdpR solves a semidefinite relaxation of the max-K-cut formulation of the described correlation clustering problem. Since the number of components K in a quasispecies is a priori unknown, we rely on the so-called pseudo F index to infer it (Section 2.2.3). For a given graph with non-negative edge weights that indicate similarity (or dissimilarity) between vertex pairs, the max K-cut problem partitions the vertex set into K groups such that the sum total of the weights of edges crossing each pair of the groups is maximized. Formally, the max K-cut problem for a weighted graph is stated as

$$\underset{\mathscr{V}_1,\mathscr{V}_1,\ldots,\mathscr{V}_K}{\text{maximize}} \sum_{1\le r<s\le K} \sum_{i\in\mathscr{V}_r, j\in\mathscr{V}_s} \omega_{ij},$$
$$\text{subject to} \quad \mathscr{V} = \mathscr{V}_1 \cup \mathscr{V}_2 \ldots \cup \mathscr{V}_K, \quad \mathscr{V}_r \cap \mathscr{V}_s = \phi \ \forall\, r, s, \quad r \neq s, \tag{2}$$

where $\{\mathscr{V}_r\}$, $r = 1, \ldots, K$, represents a partition of the vertex set $\mathscr{V}$. This can be generalized to the setting of an incomplete undirected graph with signed edge weights, which describes the clustering task arising in the context of viral quasispecies reconstruction. Cluster membership of a read i can be represented by a K-dimensional vector with zeros at all entries except one entry, which corresponds to the cluster where i belongs. Let Y_{ij} be the correlation between the K-dimensional membership vectors corresponding to reads i and j. We define an $|R| \times |R|$ positive semidefinite matrix $\mathbf{Y} = \{Y_{ij}\}$ and using vector matrix notation, we can write this problem in the form of a semidefinite program (SDP) [12]

$$\max_{\mathbf{Y}} \ \langle \mathbf{W}, \mathbf{Y} \rangle, \quad \text{subject to Diag}(\mathbf{Y}) = \mathbf{1}, \quad \mathbf{Y} \succcurlyeq 0,$$
$$\text{and} \quad Y_{ij} \ge -\frac{1}{K-1}, \quad i, j = 1, \ldots, |R|, \tag{3}$$

where $\mathbf{W} = \{\omega_{ij}\}$ is the $|R| \times |R|$ edge weight matrix of $\mathscr{G}$, Diag($\mathbf{Y}$) denotes the diagonal vector of $\mathbf{Y}$, $\mathbf{1}$ is the $|R|$-dimensional vector of all 1's, and $\langle \mathbf{A},\mathbf{B} \rangle$ denotes the matrix dot product of $\mathbf{A}$ and $\mathbf{B}$; note that R is identical to the vertex set $\mathscr{V}$.

Detection of rare variants in a viral quasispecies population requires high read coverage, which in turn requires solving (3) for $\mathbf{Y}$ of high dimension. While approaches to solve (3) optimally can be computationally prohibitive in these settings, efficient approximate methods that exploit the structure of $\mathbf{Y}$ can be used. In particular, since $\mathbf{Y}$ can be interpreted as an erroneous version of a similarity matrix with underlying data originating from K clusters, $\mathbf{Y}$ can be factorized using low-dimensional rank-K matrices. This is notable since SDPs with low-rank solutions can be solved efficiently [18]. Therefore, to find computationally feasible solutions to QSR problem, it is beneficial to express $\mathbf{Y}$ as $\mathbf{Y} = \mathbf{V}\ \mathbf{V}^T$ where $\mathbf{V}$ is an $|R| \times K$ matrix, and re-phrase the optimization (3) in terms of this low-dimensional matrix $\mathbf{V}$ ($K \ll |R|$) such that $[\mathbf{V}^{(i)}]^T\mathbf{V}^{(j)} = Y_{ij}$. Using this decomposition of $\mathbf{Y}$, we can write the Lagrangian relaxation of (3) as

$$\min_{\lambda \ge 0} \max_{\mathbf{V}} \sum_{ij} \left(\omega_{ij} [\mathbf{V}^{(i)}]^T \mathbf{V}^{(j)} + \lambda_{ij} \left([\mathbf{V}^{(i)}]^T \mathbf{V}^{(j)} + \frac{1}{K-1} \right) \right), \tag{4}$$

where $\lambda = \{\lambda_{ij}\}$ is an $|R| \times |R|$ matrix of Lagrange multipliers for the inequality constraints in (3).

To solve (4), the objective function is alternately optimized with respect to $\mathbf{V}$ and λ, one at a time. With λ fixed, $\mathbf{V}$ is updated using gradient ascent [19] during which the ith row of $\mathbf{V}$ is found as $\mathbf{V}^{(i)} \leftarrow \sum_{j:e_{ij}\in\mathscr{E}} (\omega_{ij} + \lambda_{ij})\mathbf{V}^{(j)}$. With $\mathbf{V}$ fixed, λ is updated using sub-gradient descent [19] where the sub-gradient of the objective in (4) with respect to λ is given by $([\mathbf{V}^{(i)}]\ ^T\mathbf{V}^{(j)} + 1/(K-1))_+$. At each iteration of these alternating steps, $\mathbf{V}$ is augmented by adding a column vector having entries drawn from the normal distribution $\mathscr{N}(0, 1)$ until $\mathbf{V}$ becomes rank-deficient (i.e., until its rank becomes smaller than the number of its columns). Augmenting $\mathbf{V}$ with columns potentially increases value of the objective function in (4) which implies that the described procedure leads us to a locally optimal solution while making a parsimonious adjustment to the rank of the solution. However, the optimal solution $\mathbf{V}_{opt} \in \mathbb{R}^{|R|\times r_{opt}}$ has rank r_{opt} typically greater than K; therefore, to find a K-clustering of the vertices, a valid partition of $|R|$ reads into K clusters is obtained by

a. choosing K random vectors $\mathbf{z}_k \in \mathbb{R}^{r_{opt}}$, $k = 1, \ldots, K$ such that $\mathbf{z}_k \sim \mathscr{N}(0, \mathbb{I})$, and

b. assigning the ith read to the kth cluster by choosing k such that $\mathbf{z}_k$ is the closest to the ith row $\mathbf{V}_{opt}^{(i)}$ of $\mathbf{V}_{opt}$, i.e., $\hat{k} = \underset{1\le k\le K}{\arg\max}\, (\mathbf{V}_{opt}^{(i)})^T \mathbf{z}_k$, $\forall i = 1,2,\ldots,|R|$.

It has been shown in [20] that the output (in expectation) of the above random projection heuristic is a constant factor approximation of the optimal clustering objective.

For each cluster, a consensus sequence of length ℓ is created by position-wise majority voting among the reads assigned to that cluster. Next, we greedily explore if changing the cluster membership of reads may further improve the value of the objective function; if it does, the consensus sequences need to be re-evaluated. This greedy improvement step is repeated until no further improvement of the objective is possible. Finally, the consensus sequences are extended to full-length genomes by completing non-polymorphic (homozygous) sites with alleles excluded from the data in the pre-processing step; this result in K sequences. Moreover, frequencies of the sequences are estimated by computing relative fractions of the reads constituting corresponding clusters.

2.2.3. Determining the number of components in a quasispecies

The number of haplotypes in a quasispecies is typically not known before an experiment and thus needs to be inferred, i.e., a reconstruction procedure needs to determine the number of clusters K into which the vertices of $\mathscr{G}$ are to be partitioned. Finding the number of clusters is a major challenge for most of the existing clustering methods, regardless of application. Note that a clustering mechanism that relies on parsimonious cost objective functions (such as the minimum error correction score, which we define and use in the next section) favors larger number of clusters over smaller one (since the objective function monotonically decreases with K). Therefore, such approaches may lead to overestimating the number of clusters, i.e., they may generate a large number of false positives. The model selection problem remains an open and active research topic and is known to be difficult to solve, especially for objects with partial observations as is the case with the QSR problem. In this work, we determine the number of clusters by comparing the quality of clustering solutions quantified using the Caliński-Harabasz criterion [21], also known as the pseudo F index, defined as

$$F(K) = \frac{\text{inter cluster separation}/(K-1)}{\text{intra cluster separation}/(|R|-K)}. \tag{5}$$

In the context of quasispecies reconstruction, the terms in the numerator and denominator of (5) are defined as follows. Let $I_i \in \{1,\ldots,K\}$ denote the index of the cluster containing read $r_i, \forall i$. Let n_k denote the number of reads in the kth cluster and c_k denote the consensus of the reads in the kth cluster; moreover, let $\overline{c}$ be the consensus of $c_k, k = 1, \ldots, K$. If $D(\cdot,\cdot)$ denotes the generalized Hamming distance between two strings over the alphabet $\{A,C,G,T\}$ [12], then the inter-cluster separation is measured by $\sum_{k=1}^{K} n_k D(c_k, \overline{c})$ and the intra cluster separation is captured by $\sum_{k=1}^{K} \sum_{i:I_i=k} D(r_i, c_k)$. It has been observed that the large values of the pseudo F index indicate closely knit clusters [21,22] and, in practice, the value of K for which this index is maximized is a good candidate for the choice of the number of underlying clusters.

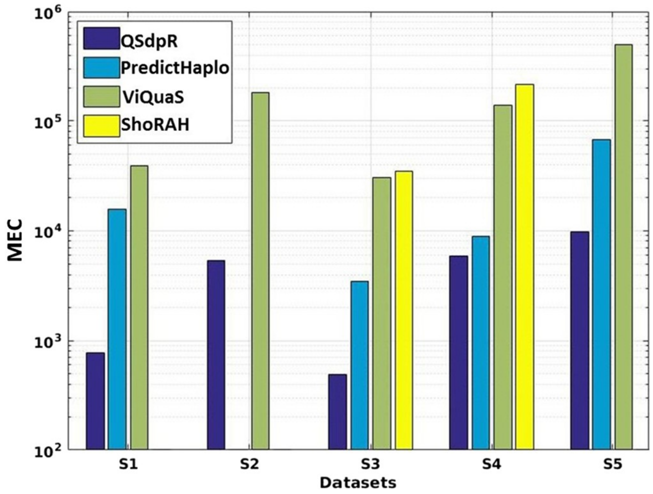

Fig. 1. MEC score comparison of QSdpR, PredictHaplo, ViQuaS and ShoRAH on the simulated datasets S1–S5. PredictHaplo could not run on S2. ShoRAH returned haplotypes with 72%, 44.6% and 93.6% of the reference genome lengths on sets S1, S2 and S5, respectively.

Therefore, the number of viral strains can be estimated by solving the clustering problem over a pre-selected range of K and choosing the value of K for which the pseudo F index has the highest value.

Note that in the classical correlation clustering framework [23], the choice of K is made within the actual clustering procedure. QSdpR decides on K in a different manner but we still refer to it as correlation clustering to emphasize the fact that the clustering is being performed on a read-to-read "correlation" graph.

3. Results and discussions

3.1. Results

3.1.1. Simulated data

We test QSdpR on the data sets *S1–S5* and characterize its performance in terms of the performance metrics discussed in Section S1.1 of Supplementary file 1. The mismatch errors (i.e., the MEC scores) are shown in Fig. 1, while Predicted Proportion, Reconstruction Error and Frequency Deviation values are reported in Table 1. For a meaningful and fair MEC comparison in Fig. 1, we excluded the cases where the reconstruction is partial. Among the considered schemes, our method achieves the lowest MEC scores for all of the data sets considered here, followed by PredictHaplo (PredictHaplo failed to run successfully on *S2* even though the sequencing coverage is as high as 200 ×), while ViQuaS and ShoRAH have much higher mismatch errors on all the datasets. The reason QSdpR performs so well in terms of MEC is due to the fact that correlation clustering seeks to form clusters collecting reads similar to each other in terms of the Hamming distance; therefore, an optimal correlation clustering solution naturally minimizes the overall MEC which can be interpreted as the total sum of the Hamming distances between the reads in clusters and the corresponding reconstructed strains. It is worthwhile pointing out that since the ground truth for a quasispecies is generally unavailable (discovering it is the entire purpose of QSR), performance metrics such as reconstruction error and frequency mismatch are in general not possible to compute in practice so one needs to use proxy measures such as the MEC score.

From Table 1, it can be seen that QSdpR infers the number of underlying sequences correctly for 3 out of 5 data sets, namely for *S1*, *S2* and *S5*, as indicated by the Predicted Proportion values. For datasets *S3* and *S4*, it overestimates the number of sequences by 1. On the other hand, PredictHaplo underestimates the number of sequences by 1 for sets *S3* and *S5*, and infers it correctly for *S1* and *S4*. ViQuaS and ShoRAH always overestimate this number except for set *S5* where ViQuaS reconstructs only one strain. In terms of Reconstruction Error, our method is able to recover all of the 5 sequences without a single mismatch for *S1* and *S5*; for *S3*, it matches the performance of PredictHaplo, which does not provide error-free reconstruction in any of the data sets on which it could successfully run. However, for set *S4*, PredictHaplo achieves the lowest error (it has 26 nucleotide mismatches fewer than our method). ShoRAH achieves better reconstruction than our method on *S4*. While our correlation clustering technique provides the most accurate spectra reconstruction for *S1* and *S5*, it is less accurate than PredictHaplo on sets *S3* and *S4*. This is due to overestimating the number of sequences for these 2 sets which leads to misclassification of some reads, thus causing discrepancy between the inferred frequencies and the correct ones. This effect is much more pronounced in ViQuaS and ShoRAH primarily because they significantly overestimate the number of sequences.

To demonstrate efficacy of the proposed approach for estimation of the number of viral haplotypes, we report normalized pseudo F indices $F(K)$ in Fig. 2 for the synthetic data sets *S1*, *S3* and *S5* and for all 13 genes of the HIV-1 dataset. Recall that the true number of haplotypes in each of these datasets is 5. It is evident from Fig. 2 that the correct value of K maximizes normalized pseudo F statistics for *S1* and *S5* though not for *S3* (where the metric is maximized for $K = 6$), and for all HIV-1 genes except *PR* and *nef* genes where the metric is maximum at $K = 7$ and $K = 4$, respectively. Results for runtime evaluation of QSdpR are discussed in Section S1.4 of Supplementary file 1. Performance of QSdpR on a broader spectrum of quasispecies can be found in Section S1.5 of the same.

To assess robustness of QSdpR to indel mutations, we consider the scenario where sequences in quasispecies harbor insertions and deletions, along with SNVs. A quasispecies dataset containing 5 sequences is simulated with parameters identical to those in datasets *S1*, *S3*; additionally, indels at 0.1% are planted at random positions of 4 among the 5 sequences. These indels have lengths between 1 and 3 bp. Paired-end reads are generated at 100 × coverage. QSdpR was able to correctly identify size of the population in this dataset (i.e., *Predicted Proportion* = 1), whereas ViQuaS and ShoRAH reported 3 and 4.2 respectively. PredictHaplo failed to execute on this dataset (citing insufficient coverage). In terms of Reconstruction Error, QSdpR (0.5263) performed better than ViQuaS (0.5383) and ShoRAH (5.5588). Frequency Deviation reported by QSdpR is 0.046, slightly better than

Table 1
Performance evaluation of QSdpR on the simulated datasets *S1–S5*. QS, PH, VQ and SH denote QSdpR, PredictHaplo, ViQuaS and ShoRAH, respectively. Boldface value in each row indicates the best performance for the given metric. PredictHaplo could not run on S2. ViQuaS reconstructed only one sequence for S5, hence is excluded from the comparison.

	Predicted proportion				Reconstruction error ($\times 10^{-3}$)				Frequency deviation ($\times 10^{-2}$)			
Dataset	QS	PH	VQ	SH	QS	PH	VQ	SH	QS	PH	VQ	SH
S1	**1**	**1**	6.2	13	**0**	9.2	5.7	6.7	**0.06**	0.78	1.68	7.2
S2	**1**	–	8.5	13	**0.4**	–	9.9	4.7	**0.06**	–	1.28	3.62
S3	1.2	0.8	6	9.6	**4**	**4**	6.5	6.6	0.09	**0.03**	8.2	5.21
S4	1.09	**1**	4.3	16.7	7.3	**4.7**	10.2	5	1.33	**0.88**	2.78	3.43
S5	**1**	0.8	–	56.4	**0**	0.25	–	6.6	**0.01**	3.07	–	8.96

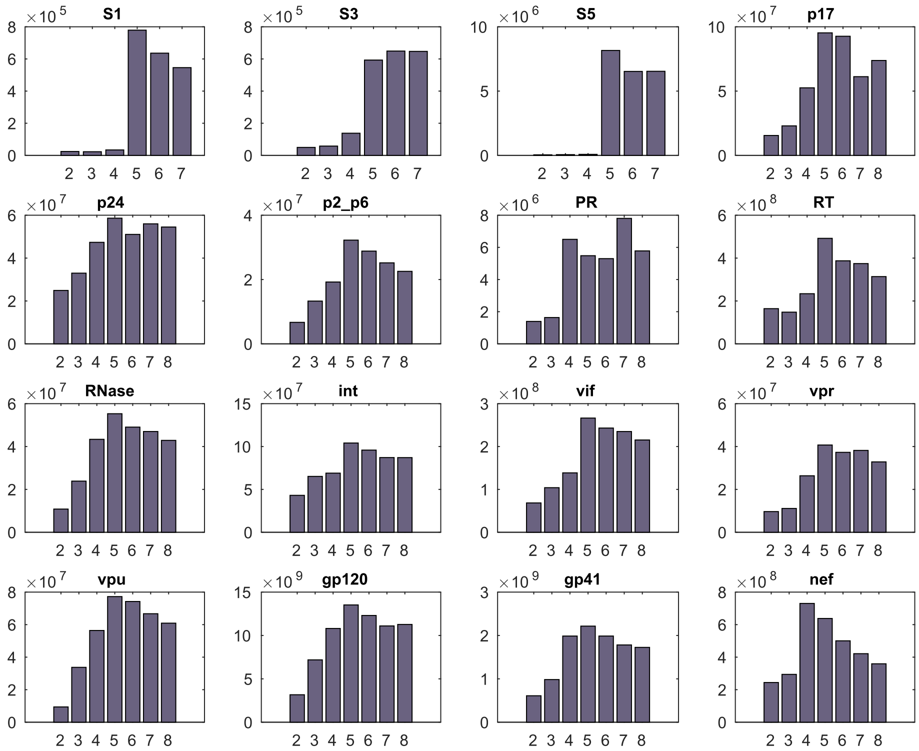

Fig. 2. Normalized pseudo F statistics as a function of the parameter K for simulated data sets *S1*, *S3*, *S5* and 13 HIV-1 genes *p17* through *nef*. The true number of species for each dataset is 5. Value of K is correctly inferred for synthetic sets *S1* and *S5* and for all HIV-1 genes except *PR* and *nef*.

those reported by ViQuaS (0.048) and ShoRAH (0.0587). On the other hand, for characterizing QSdpR performance on reads with a more realistic error profile containing short indels, 3 additional datasets *F1–F3* are simulated, each with 5 strains of 1000 bp at uniform proportions and respective diversities of 1%, 4% and 7%. 2 × 250 bp reads with 700 ± 20 bp inserts containing indels at 0.015% rate are simulated to contain no more than 5 indels per pair, at 1000 × coverage, using the Illumina MiSeqv3 read error profile of [24]. QSdpR accurately estimated K for *F1* and *F2*, and estimated 4 species for *F3*; it produced error-free reconstructions for all 5 species in *F1* and one species each for *F2* and *F3*. PredictHaplo, on the other hand, failed to execute on *F1* but predicted 5 species for *F2*, *F3*, with 2 and 4 error-free reconstructions respectively. These results suggest that QSdpR may perform reasonably well in the presence of indels in either member strains or reads. Results on the HIV-1 data, as discussed below, also support this observation since that mixture contains insertions and deletions along most of the genes.

3.1.2. HIV-1 virus mix data

Next, we report the results of comparing performance of QSdpR with existing algorithms in a application to HIV-1 Five Virus Mix data set. Gene-wise quasispecies reconstruction is performed on the major genic regions of the single strand HIV-1 RNA genome and performance metrics are computed for each of those regions. In order to determine the value of K to be used in the reconstruction, we analyze the 4036 bp long segment of the HIV-1 genome encompassing the *gag-pol* region. For cross-verification, we repeat the procedure of finding K with the 13 individual genes; in 11 cases, we obtain the correct number of clusters (see Fig. 2). Performance of QSdpR is here compared with that of PredictHaplo and ShoRAH. ViQuaS could not be used in this setting since the current version of that software does not support reconstruction over specific regions; upon trying to run it for genome-wide reconstruction, the program did not complete in 36 h on an 8-core machine. Other recent approaches such as *HaploClique* [3] and *VGA* [8] unfortunately experience code execution issues on this dataset.

Fig. 3 shows the MEC score comparison of QSdpR with PredictHaplo and ShoRAH. QSdpR achieves better MEC scores than both PredictHaplo and ShoRAH for all 13 genes of HIV-1 except for p17 gene where the MEC score of QSdpR is slightly outperformed by PredictHaplo.

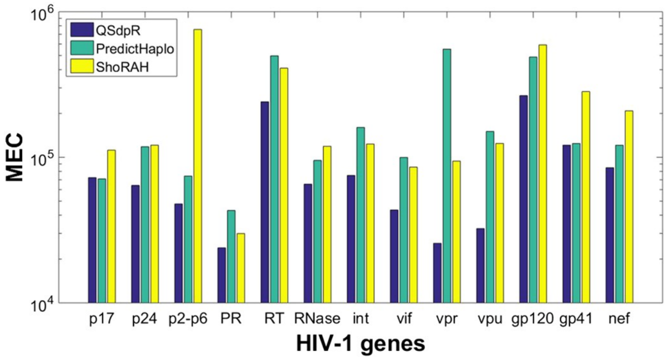

Fig. 3. The MEC score comparison of QSdpR, PredictHaplo and ShoRAH on the HIV-1 *Five Virus Mix* dataset.

QSdpR performance is further characterized using Predicted Proportion, Reconstruction Proportion, Reconstruction Error and Frequency Deviation metrics, all summarized in Table 2. As we can see from this table, Predicted Proportion of our method is better than that of the 3 competing methods for the 13 genes. Reconstruction Proportion of QSdpR is better than that of PredictHaplo on 5 genes and comparable to it on 3 genes. Compared to ShoRAH, QSdpR performs better on 9 out of 13 genes. In particular, QSdpR is able to recover the 5 true haplotypes for 3 out of 4 genes in the *pol* region comprising of *PR*, *RT*, *RNase* and *int* genes. It is interesting to note that QSdpR maintains high Reconstruction Proportion even for genes having low nucleotide diversities (see Table S2). In terms of Reconstruction Error, QSdpR outperforms PredictHaplo on 5 genes and is comparable in 1 gene, while for the remaining ones, PredictHaplo has better performance. As for ShoRAH, QSdpR has better performance on 9 genes and a comparable performance on 1 gene. Finally, in terms of Frequency Deviation, QSdpR has comparable or better performance than PredictHaplo on 8 genes and better performance than ShoRAH on 11 genes. The

Table 2

A comparison of *Predicted Proportion*, *Reconstruction Proportion*, *Reconstruction Error* and *Frequency Deviation* on the HIV-1 *Five Virus Mix* data. QS, PH and SH refer to our QSdpR, PredictHaplo and ShoRAH, respectively. *Reconstruction Error* is to be multiplied with 10^{-3} and *Frequency Deviation* error is to be multiplied with 10^{-2} to get the actual numeric value. Boldface value in each column indicates the best performance for the given metric in that column.

Metric	Gene	p17	p24	p2p6	PR	RT	RNase	int	vif	vpr	vpu	gp120	gp41	nef
Predict. prop.	QS	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
	PH	**1**	0.6	0.8	0.8	0.6	0.8	0.6	0.6	0.8	**1**	0.8	0.8	0.8
	SH	14.2	14.4	13.4	5.6	24.6	12.6	14.4	12.6	4.8	4.6	18.2	20.8	16.4
Recons. prop.	QS	0.4	**0.6**	0.6	**1**	**0.2**	**1**	**1**	**0.8**	0.2	0	0	0.4	**0.4**
	PH	**1**	0.4	**0.8**	0.6	**0.2**	0.6	0.4	0.4	**0.8**	**0.6**	0	**0.8**	**0.4**
	SH	0.8	0.2	0.4	0.8	0	0.8	0	0.2	**0.8**	0.4	0	0	0
Recons. error	QS	10.1	**2.9**	3.9	**0**	**7.3**	**0**	**0**	**0.35**	3.4	69.4	73.6	15.8	26.5
	PH	**0**	**2.9**	**0**	0.84	9.3	2.9	1.9	2.9	**0**	**34.7**	**28.6**	**0**	**5.2**
	SH	10.1	7.2	8.2	6.7	12.6	2.2	8.3	19	**0**	41.1	48.6	33.9	34.5
Freq. dev.	QS	**4.3**	4.9	**4.6**	**3.6**	**3.9**	**2.3**	**1.7**	**2.05**	2.48	3.6	5.7	4.3	**1.6**
	PH	**4.3**	**3.6**	6.8	5.8	7.3	3.4	3.6	4.8	3.2	**2**	**4.6**	**2.7**	2.3
	SH	5.42	5.89	5.67	5.38	7.05	4.41	6.13	6.01	**2.19**	3.33	6.39	7.11	6.19

genes where competing methods achieve slightly better performance than QSdpR are those in the gapped portion of the genome.

3.1.3. *Zika virus data*

In addition to benchmarking QSdpR on datasets with known ground truth, we here demonstrate its feasibility in applications to patient sample datasets. In particular, we employ QSdpR for full genome reconstruction of an Asian-lineage Zika virus (ZIKV) sequenced at ~30,000 × coverage using Illumina's MiSeq platform that generates 2 × 300 bp reads. The dataset was originally published in [25], where ZIKV strain H/PF/2013 (Genbank **KJ776791**) isolated from a human patient by European Virus Archive, Marseille, France, was used to infect a group of eight rhesus macaques for studying pathogenesis over the course of several days. We here focus on deep-sequenced samples obtained from one of the infected animals (animal 393422) on the 4th day of infection (accession **SRR3332513**) and apply the proposed method for the full genome assembly. The reference used to align the reads was the Asian-lineage ZIKV reference genome (Genbank **KU681081.3**) of length 10,807 bp [26]. QSdpR reconstructed 4 full length sequences; two of those were dominant with relative proportions 43% and 39.5%, diverging by 0.23% and 0.09% from the H/PF/2013 ZIKV strain that was used to infect the test animals. These results do not drastically differ from the findings of [27] that reported 2 major sequences at frequencies 61.3% and 38.7%. As for the competing methods, PredictHaplo reported only one strain of length that is 96% of the reference genome length and diverges from the H/PF/2013 ZIKV strain by 0.01%. ViQuaS did not complete reconstruction in 48 h while ShoRAH ran out of memory in multiple trials.

3.2. *Discussion*

QSdpR distinguishes a read from another on the basis of the SNVs on the reads. For extremely low divergence populations, sequencing errors can be mistakenly called as SNVs, which leads to an overestimation of the number of species. Therefore, the performance of QSdpR is closely tied to the quality of the variant caller used; furthermore, the processing time in QSdpR pipeline is also affected by the caller's efficiency. The QSdpR reconstruction error depends on reference indirectly since the homozygous sites are populated from this reference. Reads much shorter than average length of conserved regions in the species often convey no SNV linkage to facilitate proper clustering; therefore QSdpR is not guaranteed to perform well in spite of high coverage if the proportion of informative reads is low. QSdpR incurs complexity mainly due to repetitive clustering tasks of computing pseudo F indices. Although SNV calling and the computation of the read correlation graph is performed once, the max K-cut algorithm needs to be run for each value of K, which may be a bottleneck for quasispecies with large number of members.

Insertions and deletions (indels) are the other prevalent forms of mutations that cause viral sequences to differ from each other. While QSdpR does not directly incorporate indels into the clustering formulation, none of the existing QSR methods (except [3], which is no longer functional nor executable and thus not possible to compare with), considers indel mutations explicitly either, to the best of our knowledge, one of the reasons being that indels are known to be at least 4 times less likely than the point mutations among the viral populations [28][4]. Furthermore, our focus in this work is on datasets sequenced on Illumina platforms, which are more immune to indel errors compared to PacBio and 454 Roche technologies [29].

4. Conclusions

Inference of RNA viruses in heterogeneous populations and estimation of their relative proportion within the quasispecies has been an active area of research in recent years. In this paper, we proposed QSdpR, a framework for viral quasispecies reconstruction based on a correlation clustering formulation of the problem. The convex relaxation of this formulation is efficiently solved by exploiting the underlying sparse structure of the solution. We tested the method on synthetic data with uniform and non-uniform quasispecies spectra and varying diversity and mutation rate conditions. Moreover, the method was also tested on an experimental HIV-1 dataset having 5 known sequences. Finally, efficacy of QSdpR was demonstrated in an application to analyzing a real Zika virus dataset. It was shown that QSdpR compares favorably with the existing methods in most of the settings considered here, providing accurate estimation of the viral quasispecies spectrum. As a part of future work, it is of interest to devise an assembly framework that incorporates entire reads rather than only SNV information and potentially alleviates the dependency of the method on a reference genome.

Funding

This work was funded in part by the National Science Foundation CCF-1618427.

Availability of data and material

HIV-1 *Five Virus Mix* dataset can be obtained at https://github.com/cbg-ethz/5-virus-mix. Zika virus data can be obtained from https://www.ncbi.nlm.nih.gov/ using accession numbers given in the manuscript. All simulated datasets are available as part of the QSdpR

[4] In fact, the authors of [28] experimented with HIV-1 viruses and found that the average fraction of indels among all mutations combined is 0:07 to 0:35, and even lower for other viruses they studied.

software package.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2017.12.007.

References

[1] M.A. Nowak, What is a quasispecies? Trends Ecol. Evol. 7 (4) (1992) 118–121.
[2] A.S. Lauring, R. Andino, Quasispecies theory and the behavior of RNA viruses, PLoS Pathog 6 (7) (2010) e1001005.
[3] A. Töpfer, T. Marschall, R.A. Bull, F. Luciani, A. Schönhuth, N. Beerenwinkel, Viral quasispecies assembly via maximal clique enumeration, PLoS Comput. Biol. 10 (3) (2014) e1003515.
[4] O. Zagordi, A. Bhattacharya, N. Eriksson, N. Beerenwinkel, ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data, BMC Bioinf. 12 (1) (2011) 119.
[5] S. Prabhakaran, M. Rey, O. Zagordi, N. Beerenwinkel, V. Roth, HIV haplotype inference using a propagating Dirichlet process mixture model, IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB) 11 (1) (2014) 182–191.
[6] A. Töpfer, O. Zagordi, S. Prabhakaran, V. Roth, E. Halperin, N. Beerenwinkel, Probabilistic inference of viral quasispecies subject to recombination, J. Comput. Biol. 20 (2) (2013) 113–123.
[7] M.C. Prosperi, M. Salemi, QuRe: software for viral quasispecies reconstruction from next-generation sequencing data, Bioinformatics 28 (1) (2012) 132–133.
[8] S. Mangul, N.C. Wu, N. Mancuso, A. Zelikovsky, R. Sun, E. Eskin, Accurate viral population assembly from ultra-deep sequencing data, Bioinformatics 30 (12) (2014) 329–337.
[9] D. Jayasundara, I. Saeed, S. Maheswararajah, B. Chang, S.-L. Tang, S.K. Halgamuge, ViQuaS: an improved reconstruction pipeline for viral quasispecies spectra generated by next-generation sequencing, Bioinformatics 754 (2014).
[10] A. Huang, R. Kantor, A. DeLong, L. Schreier, S. Istrail, QColors: an algorithm for conservative viral quasispecies reconstruction from short and non-contiguous next generation sequencing reads, In Silico Biol. 11 (5, 6) (2011) 193–201.
[11] L.Z. Hong, S. Hong, H.T. Wong, P.P. Aw, Y. Cheng, A. Wilm, P.F. de Sessions, S.G. Lim, N. Nagarajan, M.L. Hibberd, et al., BAsE-Seq: a method for obtaining long viral haplotypes from short sequence reads, Genome Biol. 15 (11) (2014) 517.
[12] S. Das, H. Vikalo, SDhaP: haplotype assembly for diploids and polyploids via semi-definite programming, BMC genomics 16 (1) (2015) 260.
[13] F. Di Giallonardo, A. Töpfer, M. Rey, S. Prabhakaran, Y. Duport, C. Leemann, S. Schmutz, N.K. Campbell, B. Joos, M.R. Lecca, et al., Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations, Nucleic Acids Res. 42 (14) (2014) 115-115.
[14] G. Lancia, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, SNPs problems, complexity, and algorithms, European Symposium on Algorithms, Springer, 2001, pp. 182–193.
[15] F.E. Angly, D. Willner, F. Rohwer, P. Hugenholtz, G.W. Tyson, Grinder: a versatile amplicon and shotgun sequence simulator, Nucleic Acids Res. 251 (2012).
[16] M.A. Quail, M. Smith, P. Coupland, T.D. Otto, S.R. Harris, T.R. Connor, A. Bertoni, H.P. Swerdlow, Y. Gu, A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers, BMC genomics 13 (1) (2012) 341.
[17] A. Pandit, R.J. de Boer, Reliable reconstruction of HIV-1 whole genome haplotypes reveals clonal interference and genetic hitchhiking among immune escape variants, Retrovirology 11 (1) (2014) 56.
[18] G. Pataki, On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues, Math. Oper. Res. 23 (2) (1998) 339–358.
[19] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.
[20] A. Frieze, M. Jerrum, Improved approximation algorithms for MAXk-CUT and max bisection, Algorithmica 18 (1) (1997) 67–81.
[21] T. Caliński, J. Harabasz, A dendrite method for cluster analysis, Commun. Stat.-theory Methods 3 (1) (1974) 1–27.
[22] U. Maulik, S. Bandyopadhyay, Performance evaluation of some clustering algorithms and validity indices, IEEE Trans. Pattern Anal. Mach. Intell. 24 (12) (2002) 1650–1654.
[23] N. Bansal, A. Blum, S. Chawla, Correlation clustering, Foundations of Computer Science, 2002. Proceedings. The 43rd Annual IEEE Symposium on, IEEE, 2002, pp. 238–247.
[24] W. Huang, L. Li, J.R. Myers, G.T. Marth, ART: a next-generation sequencing read simulator, Bioinformatics 28 (4) (2011) 593–594.
[25] D.M. Dudley, M.T. Aliota, E.L. Mohr, A.M. Weiler, G. Lehrer-Brey, K.L. Weisgrau, M.S. Mohns, M.E. Breitbach, M.N. Rasheed, C.M. Newman, et al., A rhesus macaque model of Asian-lineage Zika virus infection, Nat. Commun. 7 (2016).
[26] D.W. Ellison, J. Ladner, R. Buathong, M. Alera, M. Wiley, L. Hermann, W. Rutvisuttinunt, C. Klungthong, P. Chinnawirotpisan, W. Manasatienkij, et al., Complete genome sequences of Zika virus strains isolated from the blood of patients in Thailand in 2014 and the Philippines in 2012, Genome Announc. 4 (3) (2016) e00359-16.
[27] J. Baaijens, A.Z. El Aabidine, E. Rivals, A. Schoenhuth, De novo assembly of viral quasispecies using overlap graphs, bioRxiv (2017) 080341.
[28] R. Sanjuan, M.R. Nebot, N. Chirico, L.M. Mansky, R. Belshaw, Viral mutation rates, J. Virol. 84 (19) (2010) 9733–9748.
[29] M. Schirmer, U.Z. Ijaz, R. D'Amore, N. Hall, W.T. Sloan, C. Quince, Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform, Nucleic Acids Res. 43 (6) (2015) 37.