Bioinformatics, 31(6), 2015, 886–896
doi: 10.1093/bioinformatics/btu754
Advance Access Publication Date: 13 November 2014
Original Paper

OXFORD

Genetics and population analysis

ViQuaS: an improved reconstruction pipeline for viral quasispecies spectra generated by next-generation sequencing

Duleepa Jayasundara[1,*], I. Saeed[1], Suhinthan Maheswararajah[2], B.C. Chang[3], S.-L. Tang[4] and Saman K. Halgamuge[1]

[1]Optimisation and Pattern Recognition Research Group, Department of Mechanical Engineering, Melbourne School of Engineering, The University of Melbourne, Parkville, VIC 3010, [2]Portland House Research and Advisors Ltd., Melbourne, VIC 3000, Australia, [3]Yourgene Bioscience, No. 376-5, Fuxing Rd., Shu-Lin District, New Taipei City, Taiwan and [4]Biodiversity Research Center, Academia Sinica, Nan-Kang, Taipei 11529, Taiwan

*To whom correspondence should be addressed.
Associate Editor: Gunnar Ratsch

Received on February 3, 2014; revised on October 28, 2014; accepted on November 11, 2014

Abstract

Motivation: The combined effect of a high replication rate and the low fidelity of the viral polymerase in most RNA viruses and some DNA viruses results in the formation of a viral quasispecies. Uncovering information about quasispecies populations significantly benefits the study of disease progression, antiviral drug design, vaccine design and viral pathogenesis. We present a new analysis pipeline called ViQuaS for viral quasispecies spectrum reconstruction using short next-generation sequencing reads. ViQuaS is based on a novel reference-assisted *de novo* assembly algorithm for constructing local haplotypes. A significantly extended version of an existing global strain reconstruction algorithm is also used.

Results: Benchmarking results showed that ViQuaS outperformed three other previously published methods named ShoRAH, QuRe and PredictHaplo, with improvements of at least 3.1–53.9% in recall, 0–12.1% in precision and 0–38.2% in F-score in terms of strain sequence assembly and improvements of at least 0.006–0.143 in KL-divergence and 0.001–0.035 in root mean-squared error in terms of strain frequency estimation, over the next-best algorithm under various simulation settings. We also applied ViQuaS on a real read set derived from an *in vitro* human immunodeficiency virus (HIV)-1 population, two independent datasets of foot-and-mouth-disease virus derived from the same biological sample and a real HIV-1 dataset and demonstrated better results than other methods available.

Availability and implementation: http://sourceforge.net/projects/viquas/

Contact: d.jayasundara@student.unimelb.edu.au

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

1 Introduction

Inside an infected host, most RNA viruses such as human immunodeficiency virus (HIV), hepatitis C, foot-and-mouth-disease virus (FMDV) and some DNA viruses like hepatitis B exhibit very high replication rates (Brunetto *et al.*, 1999; Lauring and Andino, 2010). The low fidelity of the proof-reading function of the viral polymerase (Brunetto *et al.*, 1999; Lauring and Andino, 2010) coupled with this high replication rate results in the formation of a viral

quasispecies population. More specifically, a quasispecies is a set of genetically related but non-identical viral mutant types (which can also be referred to as strains) that are able to co-exist within the host (Lauring and Andino, 2010).

As anti-viral vaccines and drugs make use of the knowledge on various stages of the host-viral interaction process and viral replication process (Carter and Saunders, 2007; Dimmock *et al.*, 2007), the presence of different viral strains in the population affects the natural host-immune response as well as the antiviral drug therapy (Brunetto *et al.*, 1999; Dimmock *et al.*, 2007; Lauring and Andino, 2010; Vignuzzi *et al.*, 2006). Therefore, information about these strains, or quasispecies spectra, is of critical importance to antiviral drug design and vaccine design. In particular, such information would enhance research into viral evolution, disease progression and provide insights into the biology and pathogenesis of viruses (Baldick *et al.*, 2008; Brunetto *et al.*, 1999; Carter and Saunders, 2007; Dimmock *et al.*, 2007; Nishijima *et al.*, 2012).

During the analysis of quasispecies, identifying the co-occurrence of nucleotide level mutations is as important as identifying individual mutations. As such, there are two main challenges in this analysis: assembling nucleotide sequences of each individual strain to identify co-occurring mutations and to estimate the relative frequency of each strain within the quasispecies. Collectively, these two challenges constitute *Quasispecies Spectrum Reconstruction* (QSR).

The most similar problem to QSR is the metagenomic *binning* problem as applied to bacterial metagenomes (Kunin *et al.*, 2008), but Kunin *et al.* (2008) emphasizes the inability of binning methods based on sequence similarity and sequence composition to discriminate between strains of the same species due to high nucleotide compositional similarity.

Single-genome *de novo* assembly is another similar problem to QSR (Astrovskaya *et al.*, 2011). These algorithms are optimized to overcome technical sequence errors assuming the existence of a single species in the sequenced sample (Astrovskaya *et al.*, 2011; Beerenwinkel *et al.*, 2012). Hence given a metagenomic dataset as input, most assemblers work towards generating a single assembly of reads misinterpreting genetic variability as potential technical sequencing errors.

One of the earliest methods for viral population estimation using next-generation sequencing (NGS) was proposed in Eriksson *et al.* (2008). It uses a multiple read alignment and clustering strategy for read error correction, a combinatorial algorithm for strain reconstruction and a statistical model-based expectation-maximization strategy for strain frequency estimation. The method formulated in Zagordi *et al.* (2011), named ShoRAH, extends the method of Eriksson *et al.* (2008) and uses a Bayesian inference algorithm for read error correction and a probabilistic clustering algorithm for strain reconstruction (Zagordi *et al.*, 2010). Three other interesting approaches to solve the QSR problem are the combinatorial method proposed in Prosperi *et al.* (2011), which was later published as a software named QuRe (Prosperi and Salemi, 2012), the method proposed in Astrovskaya *et al.* (2011) named ViSpA and the method proposed in Prabhakaran *et al.* (2010, 2014) named PredictHaplo.

Current methods that address the QSR problem rely on the fact that the NGS reads of a quasispecies sample can be assigned with a unique alignment position to a known reference genome or a consensus sequence. This has become possible due to the rarity of repeats in viral genomes (Astrovskaya *et al.*, 2011). Following read alignment, a typical quasispecies analysis pipeline consists of three major steps: (i) First, the genetic diversity is determined on either overlapping or non-overlapping local windows of the reference genome by clustering reads within each window together with strategies to eliminate or correct technical sequencing errors. This is also referred to as constructing local haplotypes. (ii) Second, the global diversity is inferred by connecting the local haplotypes to form the nucleotide sequences spanning the entire genomic region of interest (i.e. strain reconstruction). (iii) Third is the step of estimating the relative frequencies of the reconstructed strains. Comprehensive reviews on different techniques used within these main steps of an analysis pipeline are found in Beerenwinkel *et al.* (2012) and Beerenwinkel and Zagordi (2011).

The most critical component within this pipeline is global diversity estimation, where previous studies have relied on one of two key concepts (Beerenwinkel *et al.*, 2012). The most popular strategy of inferring global diversity is based on graph-based methods, as has been implemented in Zagordi *et al.* (2011), Prosperi and Salemi (2012), Astrovskaya *et al.* (2011), Mancuso *et al.* (2011), Huang *et al.* (2011) and O'Neil and Emrich (2012). Alternatively, the problem has also been approached using probabilistic clustering methods (Prabhakaran *et al.*, 2010, 2014; Quince *et al.*, 2011; Zagordi *et al.*, 2012a). Less common, but a very useful third approach for quasispecies identification based on *de novo* assembly has been published in Ramakrishnan *et al.* (2009). Even though Ramakrishnan *et al.* (2009) does not address the QSR problem from strain assembly to frequency estimation, it provides an interesting initiation point for the total strain assembly problem using short NGS reads. We have further developed the idea of using a *de novo* assembly algorithms for quasispecies assembly in this work.

An analysis of the effects of average read length, depth of coverage and technical sequencing errors on total strain reconstruction is presented in Zagordi *et al.* (2012b), and it highlights the fact that inferring quasispecies population diversity of larger genomic regions using short NGS reads remains a challenging problem.

In this article, we present a new analysis pipeline named ViQuaS for viral QSR using short NGS reads. Our main focus is on inferring quasispecies spectrum within a certain genomic region of a viral sample using sequence reads that are several folds shorter in length compared with the region, in the presence of a moderate level of technical sequencing errors. The novelty of our method is 2-fold. (i) We propose a novel reference-assisted *de novo* assembly algorithm for defining local haplotypes. (ii) We propose a significant extension to the global reconstruction algorithm presented in Prosperi *et al.* (2011), based on relative coverage depths and overlap agreement of local haplotypes. Moreover, we present the first benchmark study of ViQuaS and three other existing methods named ShoRAH (Zagordi *et al.*, 2011), QuRe (Prosperi and Salemi, 2012) and PredictHaplo (Prabhakaran *et al.*, 2010, 2014), on a broad spectrum of simulated data allowing for a thorough comparison of performance between the four methods.

The correctness of the nucleotide sequences reconstructed is primarily evaluated based on the measures *Recall* and *Precision*. Under different simulation conditions, the four methods used in the benchmark study outperformed each other in *Recall* and *Precision* without a particular pattern. Hence, the geometric mean of *Recall* and *Precision* termed as *F-score* is used to measure the performance using a single measure. On the other hand, the accuracy of the estimated relative frequency distribution of the reconstructed strains is measured using *KL-divergence* (Kullback and Leibler, 1951) and *root mean-squared error*. ViQuaS outperformed ShoRAH, QuRe and PredictHaplo, demonstrating improvements of at least 3.1–53.9% in *Recall*, 0–12.1% in *Precision* and 0–38.2% in *F-score* in terms of strain sequence assembly and improvements of at least 0.006–0.143 in *KL-divergence* and 0.001–0.035 in *root*

mean-squared error over the next-best algorithm under various simulation settings.

2 Materials and Methods

2.1 Datasets

We evaluated ViQuaS on a broad spectrum of simulated viral quasispecies NGS read samples. The complexity of a sample of reads primarily depends on three independent factors; the number of different strains in the population (N_s), the average hamming distance between all strains (*Diversity*) and the extent of technical sequencing errors present in the NGS reads.

An important factor in measuring the performance of the reconstruction algorithm is the number of theoretically reconstructible strains (N_p) out of N_s. Given a set of n_{total} number of NGS reads with a mean read length L_r, N_p depends on the minimum relative frequency (f_{min}) that a strain can have, in order to have a probability of at least p_{min} of being completely covered during the sequencing process. f_{min} is calculated according to the relationship between n_{total}, L_r, p_{min} and the length of the reference genome, derived with the assumption of Lander–Waterman model of sequencing (Eriksson *et al.*, 2008). In calculating f_{min} values corresponding to the read sets described below, we set $p_{min} = 0.99$. In performance evaluation of the method, we used N_p instead of N_s as the expected number of strains to be retrieved through the method.

2.1.1 Platform-independent simulated datasets

The samples used have *Diversity* values in the range of 1–10% ($Diversity \in \{1\%, 2\%, 3\%, 4\%, 5\%, 6\%, 7\%, 8\%, 9\%, 10\%\}$), N_s in the irregularly spaced range of 3–100 ($N_s \in \{3, 5, 7, 10, 25, 50, 75, 100\}$) and substitutional technical error probability (e) of 0% (error-free reads) and 0.1% (theoretical maximum substitutional error probability of quality trimmed reads with a PHRED threshold of 30). An irregularly spaced range between 3 and 100 was used for N_s to evaluate the behaviour of methods under both small and moderately high number of strain numbers without unnecessarily increasing the number of simulated datasets. Different strains were derived by mutating a reference sequence at randomly chosen locations to achieve desired *Diversity* values. A total of 6400 read samples categorized in to eight sets (*SS1–SS8*) were simulated using the read simulating software named Grinder (Angly *et al.*, 2012) (version 0.5.3). Supplementary Table S1 of Supplementary File S4 summarizes the simulation parameters of *SS1–SS8* and Supplementary File S4 provides further details on the simulation settings.

2.1.2 Platform-independent simulated dataset with real HBV mutations

To evaluate the performance of ViQuaS on 'near real' data, we designed a simulated dataset (*SS9*) where the strains contain mutations at nucleotide positions from 1814 to 1956 (a region of 143 nt length in the *pre-C/core* gene), summarized under Table 1 of Brunetto *et al.* (1999). Supplementary Table S2 of Supplementary File S4 summarizes the simulation parameters of *SS9*.

2.1.3 Platform-independent simulated datasets with real HIV-1 strains

We simulated two psudo-real (simulated with a mix of real and artificial parameters) quasispecies sample datasets (*SS10* and *SS11*) by using 10 real HIV-1 strains published in Zagordi *et al.* (2010). Supplementary File S4 provides further details on *SS10* and *SS11* and Supplementary Table S3 summarizes the simulation parameters used.

2.1.4 HIV-1 *in vitro* population real dataset

We used the *5-virus-mix* dataset published in Giallonardo *et al.* (2014) to evaluate the performance of ViQuaS and other existing method on real Illumina sequence reads derived from a quasispecies population generated *in vitro* using five known HIV-1 strains named as $HIV-1_{89.6}$, $HIV-1_{HXB2}$, $HIV-1_{JR-CSF}$, $HIV-1_{NL4-3}$ and $HIV-1_{YU2}$. Sequencing has been performed on an Illumina MiSeq Benchtop sequencer. The reads from the sequencer are paired end with 2×250 bp length. We selected a 5000-bp-long genomic region encompassing the 4036-bp-long *gap-pol* region of the HIV-1 genome as the region of interest for the global reconstruction of haplotypes. Supplementary File S4 provides further details on data pre-processing.

2.1.5 Real Illumina FMDV dataset

We used two Illumina Genome Analyzer IIx datasets (GenBank accession numbers ERR180978 and ERR180979) derived from the same biological sample (GenBank accession number ERS182429) presented in Morelli *et al.* (2013) to evaluate the performance of the four methods on real data. Supplementary File S4 provides further details on this dataset and pre-processing.

2.1.6 Real Roche 454 HIV-1 dataset

Roche 454 was the first new generation sequencing platform used for rare variant analysis. Even though our main focus is on data generated from Illumina platform, we used V11909 dataset derived from an antiretroviral-experienced HIV-1 infected patient, published in Wang *et al.* (2007), to evaluate the applicability of ViQuaS on Roche 454 sequences. Supplementary File S4 provides further details on this dataset and pre-processing.

2.2 ViQuaS analysis pipeline

The proposed analysis pipeline assumes the availability of a set of NGS reads (R_{total}) generated out of a viral quasispecies sample and the availability of an appropriate reference genome (G_{ref}). G_{ref} could either be a reference sequence of the viral type of interest obtained from a public sequence database or the consensus sequence of the multiple sequence alignment of R_{total}. Both R_{total} and G_{ref} are the inputs to the pipeline, and we are interested in identifying the co-occurring mutations in each strain and the relative frequency of each strain in the population. Following is a description of the main steps of the pipeline. A detailed description of ViQuaS analysis pipeline with an example is presented in Supplementary File S1.

2.2.1 Read filtering

R_{total} is aligned to G_{ref} and is partitioned into two sets denoted by R_{pa} and R_{mu} based on whether or not each read has a perfect alignment with G_{ref} or not [i.e. *mu*tated (*mu*)], respectively.

2.2.2 *de novo* assembly

This step involves unsupervised partitioning of R_{mu} into contigs based on the greedy graph-based short NGS read assembly algorithm presented in Warren *et al.* (2007), named SSAKE (version 3.8) [a brief review of this algorithm can be found in Miller *et al.* (2010)].

A major concern in using a *de novo* assemblers for quasispecies assembly is that they produce chimeric contigs when the strains are closely related in genetic content. SSAKE, which is a greedy graph-based *de novo* assembler, tries to overcome this by reusing similar reads that do not agree with an existing extension iteration at a later stage to generate new contigs (Miller *et al.*, 2010). This property of

Table 1. Performance comparison of ViQuaS, ShoRAH and QuRe under different quasispecies population characteristics and NGS sequencing characteristics when Diversity > 3%

Sample set name	ViQuaS	ShoRAH	QuRe	PredictHaplo
Recall (strain reconstruction)				
SS1	**0.903**	0.732	0.338	0.750
SS2	**0.954**	0.831	0.498	0.735
SS3	**0.682**	0.454	0.016	0.651
SS4	**0.888**	0.802	0.300	0.678
SS5	**0.900**	0.737	0.341	0.760
SS6	**0.695**	0.472	0.015	0.654
SS7	**0.838**	0.000	0.317	0.000
SS8	**0.848**	0.000	0.309	0.055
Precision (strain reconstruction)				
SS1	0.782	0.657	0.628	**0.840**
SS2	0.786	0.791	**0.866**	0.852
SS3	0.499	0.381	0.038	**0.712**
SS4	**0.830**	0.776	0.604	0.821
SS5	0.778	0.664	0.628	**0.854**
SS6	0.565	0.405	0.033	**0.740**
SS7	**0.697**	0.000	0.617	0.000
SS8	**0.713**	0.000	0.592	0.064
F-score (strain reconstruction)				
SS1	**0.830**	0.682	0.421	0.773
SS2	**0.851**	0.805	0.612	0.763
SS3	0.560	0.403	0.022	**0.670**
SS4	**0.853**	0.778	0.385	0.718
SS5	**0.827**	0.687	0.421	0.784
SS6	0.611	0.420	0.019	**0.681**
SS7	**0.753**	0.000	0.401	0.000
SS8	**0.767**	0.000	0.385	0.058
KL-divergence (frequency estimation)				
SS1	**0.005**	0.042	0.284	0.213
SS2	**0.004**	0.010	0.222	0.247
SS3	**0.043**	0.186	0.619	0.209
SS4	**0.005**	0.019	0.252	0.205
SS5	**0.004**	0.047	0.328	0.212
SS6	**0.036**	0.144	0.492	0.198
SS7	**0.018**	0.060	0.326	0.206
SS8	**0.015**	0.064	0.383	0.207
Root mean-squared error (frequency estimation)				
SS1	**0.011**	0.021	0.047	0.132
SS2	**0.009**	0.010	0.063	0.129
SS3	**0.024**	0.059	0.128	0.138
SS4	**0.018**	0.021	0.106	0.153
SS5	**0.010**	0.022	0.100	0.131
SS6	**0.030**	0.062	0.142	0.148
SS7	**0.016**	0.028	0.079	0.131
SS8	**0.015**	0.027	0.111	0.131

In comparison to *SS1*, where $L_r = 200$ bp, $e = 0\%$, only SNPs are present ($T = P$) and $f_{min} = 0.7\%$, *SS2* has a longer read length ($L_r = 300$ bp), *SS3* has a shorter read length ($L_r = 100$ bp), *SS4* has a reduced depth of coverage ($f_{min} = 2.1$), *SS5* contains sequencing errors ($e = 0.1\%$), *SS6* has both a shorter read length ($L_r = 100$ bp) and a reduced depth of coverage ($f_{min} = 1.4$), *SS7* has indels ($T = P, I, D$) and *SS8* contains sequencing errors ($e = 0.1\%$) and has indels ($T = P, I, D$). Each point indicates the mean value of the measure. Bold face figures show the best performance measure in each row.

SSAKE was the main reason to choose it for this step over the well-known *de bruijn* graph-based assemblers. This algorithm may still produce chimeric contigs and we address this issue under the *Chimeric error correction* step. The pipeline is developed in a modular architecture, such that SSAKE is replaceable with any other better performing algorithm as far as it keeps track of the individual reads contributing the contigs.

SSAKE is designed and tested for assembling high-throughput Illumina short reads. Let N_{CE} be the number of contigs produced by SSAKE. Each contig consists of a subset of overlapping reads from R_{mu}.

Moreover, within the *de novo* assembly algorithm, SSAKE eliminates probable technical sequencing errors. The two user defined parameters, o (which stands for the minimum number of reads needed to call a base during an extension) and r (which stands for the minimum base ratio used to accept an overhang consensus base), are used to determine the probable bases with technical sequencing errors. Further details regarding these parameters are found in the support documents provided with the SSAKE software package. Parametrization of SSAKE (i.e. setting o and r values) is done, such that mutation arising from low abundant strains is not misinterpreted as technical sequencing errors. Accordingly, o gets a value in the range of 3–5 and r gets the default value of 0.7. An accurate assessment of the technical sequencing error profile of the input read set will enable specific parametrization of SSAKE. We observed that $o = 3$ and $r = 0.7$ produces the best results under the ideal scenario of error free reads. For an appropriately quality-controlled input read set with an Illumina technical error profile (average error rate$<0.12\%$), $o = 5$ and $r = 0.7$ produces the best results.

2.2.3 Mutation calling

Each read in R_{mu} consists of at least one mutation with respect to G_{ref}. Hence, appropriate parametrization of SSAKE produces contigs from R_{mu}, such that each contig contains at least one mutation. The *Mutation calling* step involves aligning each *contig* to G_{ref} using an altered version of Smith–Waterman algorithm (Smith and Waterman, 1981), which performs a semi-global alignment without end gap penalties to identify the nucleotide variations (i.e. mutation) with respect to G_{ref}. The output of this step is N_{CE} number of mutation sets where each contig produces one mutation set.

2.2.4 Chimeric error correction

The *de novo assembly* algorithm does not guarantee that all the mutations present in one contig correspond to a single strain. During the graph generation step of SSAKE, reads corresponding to two different strains can have significant common overlap regions resulting such reads being included in a single contig. This phenomenon leads to mutations corresponding to different strains to be present in a single contig, which we consider as a *chimeric error*. On the contrary, in a technical error-free experiment, the NGS technology ensures that the nucleotide sequence captured in each read in R_{total} corresponds to a genomic region of only one of the strains in the sample (Zagordi *et al.*, 2010). Because R_{mu} is a subset of R_{total}, the nucleotide sequence (and hence the set of mutations) captured in a single read of R_{mu} corresponds to a single strain. In other words, we assume that no chimeric reads occur during the sequencing process. Therefore, the *Chimeric error correction* algorithm is based on the assumption that the set of mutations captured in a single read of R_{mu} corresponds to a single strain.

Based on the above assumption, each of the N_{CE} number of mutation sets and the corresponding contigs will be split into two or more subsets if insufficient evidence is found in the reads of R_{mu}

supporting the co-occurrence of any pair of consecutive mutations. Assume that this step results in a total of N_{CC} number of error-free contigs ($N_{CC} \geq N_{CE}$). Each of the N_{CC} number of error-free contigs (and the corresponding mutation sets) represents a set of closely located mutations of a single strain, which is also referred to as a *local haplotype*.

The three steps: *de novo assembly*, *Mutation calling* and *Chimeric error correction* constitute the proposed new reference-assisted *de novo* assembly algorithm.

2.2.5 Local haplotype frequency estimation

Assume that l_s and l_e be the start and end alignment positions, respectively, of a local haplotype to G_{ref}. Then, the relative frequency of the particular local haplotype is the ratio between the number of reads in R_{mu} contributing to it and the total number of reads in R_{total} having an alignment to G_{ref}, entirely within l_s and l_e.

At the end of this step, we obtain N_{CC} number of local haplotype sequences (HS_i where $i \in 1,\ldots,N_{CC}$) and the corresponding local haplotype frequencies (HF_i where $i \in 1,\ldots,N_{CC}$).

2.2.6 Global spectrum reconstruction

The two goals of *Global spectrum reconstruction* are combining local haplotypes, such that all mutations from the same strain are grouped together forming each of the original strains and estimating the corresponding relative frequencies. The algorithm used here was derived from the algorithm described in Prosperi *et al.* (2011), which is based on matching multinomial distributions. The *Global spectrum reconstruction* algorithm iteratively connects overlapping (HS_i, HS_j) sequence pairs, where $i \neq j$, if the overlapping regions agree with each other. If two or more such pairs show overlap agreement, the pair with the closest (HF_i, HF_j) value pair will take precedence. At the end of each iteration, the connected set of overlap-agreeing HS_i's gives the nucleotide sequence of the reconstructed strain. The lowest value out of the corresponding HF_i's gives the relative frequency of the reconstructed strain.

The three major extensions we propose to the existing algorithm in Prosperi *et al.* (2011) are as follows:

1. Instead of partitioning the nucleotide space of G_{ref} into windows having a predefined constant width (the window width is typically set equal to the average read length) to infer the local diversity (i.e. local haplotypes), we allow contigs to grow along the nucleotide space of the G_{ref} without restricting to a predefined nucleotide width. This approach reduces the number of partition boundaries at which heuristic decisions are made to connect overlapping HS_i's during the reconstruction process.
2. We use the HF_i values to determine the closeness of prevalence of two overlapping HS_i's. In contrast, the existing algorithm employs the read numbers contributing to the read clusters corresponding to HS_i's. This alteration leads to accurate strain reconstruction and frequency estimation without relying on the assumption of a uniform coverage depth.
3. We hypothesize that the lowest value out of the HF_i's corresponding to the connected set of HS_i's gives the most accurate estimate to the relative frequency of a reconstructed strain. In contrast, the existing algorithm in Prosperi *et al.* (2011) calculates it based on the number of reads contributing to the local haplotype that belongs to both the set of connected local haplotypes and the partition window of G_{ref} through which the maximum number of haplotypes pass.

Supplementary File S1 describes the iterative algorithm in detail with an example.

3 Results

Performance evaluation of QSR is 2-fold. The performance in terms of total strain reconstruction is evaluated based on the metrics *Recall*, *Precision* and *F- score* as defined below.

$$Recall = \frac{\text{True positive strains with a relative frequency} > f_{min}}{\text{Expected number of strains } (N_p)} \quad (1)$$

$$Precision = \frac{\text{True positive strains with a relative frequency} > f_{min}}{\text{Total number of reconstructed strainswith a relative frequency} > f_{min}} \quad (2)$$

$$F\text{-}score = 2 \times \frac{(Recall \times Precision)}{(Recall + Precision)} \quad (3)$$

The performance in terms of frequency estimation is evaluated based on *KL-divergence* and *root mean-squared error* between the input and output relative frequency distributions. We compared ViQuaS with three other previously published software, named ShoRAH (Zagordi *et al.*, 2011), QuRe (Prosperi and Salemi, 2012) and PredictHaplo (Prabhakaran *et al.*, 2010, 2014), based on entirely simulated, psudo-real and real data.

A noteworthy factor in performance comparison is the distinction between the objectives of the global spectrum reconstruction algorithms used by these three methods. ShoRAH aims to output the most parsimonious set of strains that is consistent with a given set of reads (Eriksson *et al.*, 2008), which in general produces a larger number of strains than the actual number of strains present in the population. This leads to a better chance of producing true strains, while also producing many false *in silico* recombinants. In contrast, QuRe, PredictHaplo and ViQuaS aim to output a selective set of strains that is likely to be totally covered by the given set of reads, minimizing the production of false *in silico* recombinants. Our choice to consider only the number of theoretically reconstructible strains (N_p) as the expected number of strains in calculating *Recall* (eq. 1) and to disregard the number of reconstructed strains having a relative frequency less than f_{min} in calculating *Precision* (eq. 2) provide practical and fair grounds to compare the performance in terms of the metrics *Recall*, *Precision* and *F-score*. Also, in calculating *KL-divergence* and *root mean-squared error*, we only considered the input and output strains with relative frequency values greater than f_{min}.

In Sections 3.1–3.3, we validate ViQuaS and benchmark it against ShoRAH, QuRe and PredictHaplo using the simulated datasets *SS1*–*SS11* using the metrics *Recall*, *Precision*, *F-score*, *KL-divergence* and *root mean-squared error*. In Section 3.4, we validate and benchmark the methods using average pair-wise alignment cost between the known and reconstructed haplotypes. Finally, in Sections 3.5 and 3.6, we compare the results of the four methods on real datasets. As a result of the fact that the exact nucleotide content of the strains and their relative frequencies are unknown in these real datasets, we have used other criteria as described under the respective subsections to evaluate the results.

3.1 Performance comparison on platform-independent simulated data

Sample set *SS1* was selected as the reference dataset to compare the performance of ViQuaS, ShoRAH, QuRe and PredictHaplo. The intention of selecting *SS1* was to nullify any effect on performance from technical sequencing errors and the presence of indels, focusing solely on the capability of the four methods in strain reconstruction and frequency estimation. ShoRAH and QuRe were initially

published with pyrosequencing reads (Prosperi and Salemi, 2012; Zagordi *et al.*, 2011) (>300 bp), and ShoRAH was later used in a comparison study (Zagordi *et al.*, 2012b) on shorter read lengths (36 bp, 75 bp and 150 bp). Hence, we chose an intermediate mean read length of 200 bp for *SS1* and platform-independent simulation parameters for *SS1*–*SS8*.

On the reference dataset *SS1*, ViQuaS clearly outperformed ShoRAH, QuRe and PredictHaplo in total strain reconstruction when *Diversity* > 3% [Fig. 1(a)]. We observe that the higher the complexity of a quasispecies population, the better the performance of ViQuaS compared with ShoRAH, QuRe and PredictHaplo in terms of total strain reconstruction. ViQuaS produced the highest number of correct strains out of the four methods when *Diversity*≥3 % as shown by the *Recall* graph [Fig. 1(a)]. ShoRAH and PredictHaplo performed at comparable levels (with ShoRAH performing marginally better) in terms of *Recall*, while QuRe showed the lowest levels of *Recall* out of the four methods. Because of the highly conservative reconstruction of strains by PredictHaplo, it performed better in terms of *Precision* [Fig. 1(b)] over the other three methods, but ViQuaS performed comparably with PredictHaplo at higher *Diversity* levels. When considering the *F-score* values [Fig. 1(c)], which measures the trade-off between *Recall* and *Precision*, ViQuaS performed better than the competitors, whereas PredictHaplo held an advantage over ShoRAH due to its much superior *Precision* values compared ShoRAH. PredictHaplo showed higher *Precision* due to its inherent drawback of underestimating the number of strains in a population (i.e. richness) as confirmed by a previous study by Schirmer *et al.* (2012). Therefore, PredictHaplo performed better than ShoRAH in terms of *Precision* and *F-score*, but on average ShoRAH managed to produce a higher number of true strains than PredictHaplo.

Figure 1(d) shows that both ViQuaS and ShoRAH perform at comparable levels to each other, while significantly outperforming QuRe and PredictHaplo in terms of frequency estimation. Moreover, we calculated *root mean-squared error* between the input and output frequency distributions and observed similar patterns to *KL-divergence* (refer Supplementary Figs. S23(d) and S24(d) of Supplementary File S3).

Considering the performance of the four methods on the reference dataset *SS1*, we observe that ViQuaS performs the best in both aspects of total strain sequence reconstruction and strain frequency estimation. ShoRAH reconstructs marginally higher number of true strains than PredictHaplo and significantly outperforms PredictHaplo in strain frequency estimation. We assume that the comparable levels of performance of ShoRAH and PredictHaplo is due to the fact that both methods use of a Dirichlet process mixture model for read clustering (Prabhakaran *et al.*, 2010; Zagordi *et al.*, 2011). Even though QuRe demonstrates the lowest levels of performance on *SS1*, we observe that QuRe is the only existing method out of the three used in our study that is capable of performing a successful spectrum reconstruction when the strain contain mutations in the form of indels in addition to substitutions (Fig. 2I). Furthermore, the results of PredictHaplo are not reproducible.

Figure 2 presents the variations in performance of the four methods under different sequencing and population parameters simulated in *SS1*–*SS8* in terms of both total strain reconstruction measured by *Recall* (Fig. 2I) and strain frequency estimation measured by *KL-divergence* (Fig. 2II). Barring the samples within the *Diversity* range 1–3%, under which all three methods performed poorly, Table 1 summarizes the performance of the three methods on *SS1*–*SS8*.

In comparison to *SS1* [$L_r = 200$ bp, $e = 0\%$ and only SNPs are present ($T = P$)], the performance of all four methods was negatively affected by a shorter read length (*SS3*: $L_r = 100$ bp) [Fig. 2I(b) and Fig. 2II(b)], the presence of indels (*SS7*: $T = P, I, D$) [Fig. 2I(e) and Fig. 2II(e)] and the presence of technical sequencing errors (*SS5*: $e = 0.1\%$) [Fig. 2I(g) and Fig. 2II(g)]. Even though a reduced read length has a significant negative impact, the performance measures tend to converge for higher *Diversity* and N_s values [Fig. 2I(a), Fig. 2II(a), Fig. 2I(b) and Fig. 2II(b)].

It was observed that the number of true positives reduces under a reduced depth of coverage. Nevertheless, in terms of *Recall*, the performance of the method under a reduced depth of coverage (*SS4*: $n_{total} = 10\ 000$) was comparable with *SS1* ($n_{total} = 30\ 000$) [Fig. 2I(d)]. This observation is due to the fact that the inverse proportionality between f_{min} and n_{total} makes *Recall*, *Precision* and *F-score* values comparable, as the number of strains to be reconstructed (N_p) reduces with an increased f_{min}.

SS6 was employed to evaluate the combined effect of a shorter read length and a reduced depth of coverage. Performance patterns of all four methods under *SS6* [Fig. 2I(a) and Fig. 2II(a)] were comparable to the performance patterns under *SS3* [Fig. 2I(b) and Fig. 2II(b)] providing evidence for the conclusion that a shorter read length has a significant negative effect on performance compared with a reduced coverage depth. Comparison of performance patterns between *SS3* [Fig. 2I(b) and Fig. 2II(b)], *SS1* [Fig. 2I(d) and Fig. 2II(d)] and *SS2* [Fig. 2I(f) and Fig. 2II(f)] shows that for a given depth of coverage, performance of all four methods increases with an increased read length.

As the mean read length (L_r) and the total number of reads (n_{total}) of a sequenced sample are user-controllable parameters of an NGS experiment, the presence of indels and technical sequencing errors become the two inevitable negative influential factors of a sample dataset. *SS8* was employed to evaluate the combined effect of indels and technical sequencing errors on the performance of the four methods and ViQuaS showed clear advantage over ShoRAH, QuRe and PredictHaplo [Fig. 2I(h) and Fig. 2II(h)].

Having observed the performance patterns of the four methods over a broad range of parameters, we further evaluated the performance of them under a more adverse sequencing error rate including errors in the form of indels in addition to mismatches. We chose two sets of 10 sample populations from *SS8* having parameters (*Diversity* = 1%, $N_s = 3$) and (*Diversity* = 7%, $N_s = 75$). These two sets represent comparatively simple and complex population structures, respectively. We generated 20 sample datasets from these populations keeping all parameters except the error rate unchanged from the values of *SS7* and *SS8*. The error rate was set at a level of 0.3% with a 4:1 ratio between mismatch errors and indel errors. Table 2 summarizes the performance of the four methods under different levels of technical sequencing errors. The observed performance pattern under $e = 3\%$ shows similarity to the pattern observed in Figure 2I(h).

In summary, ViQuaS managed to maintain its superior performance compared with ShoRAH, QuRe and PredictHaplo under the negative influences of a shorter read length, the presence of technical sequencing errors, the presence of indels and a reduced coverage depth. Under all simulated settings used in this benchmark study, ViQuaS managed to produce the highest average number of true strains (refer *Recall* measures in Table 1) and the most accurate frequency estimates (refer *KL-divergence* and *root mean-squared error* measures in Table 1). On average, PredictHaplo performed better than ShoRAH, QuRe and ViQuaS in terms of *Precision* (refer *Precision* measures in Table 1). This superiority in *Precision* becomes prominent under shorter read lengths. Nevertheless, on average, ViQuaS managed better *F-score* values over PredictHaplo, ShoRAH

and QuRe. Variations in performance of the four methods in terms of *Recall*, *Precision*, *F-score*, *KL-divergence* and *root mean-squared error* against different values of *Diversity* and N_s are presented in Supplementary File S3 (Supplementary Figs. S15–S24). A detailed table carrying the raw results and performance measures of all 6400 sample in sample sets *SS1–SS8* is presented in Supplementary File S2.

To compare the time efficiency, we analyzed the running times of the four methods when the most complex samples are given as input. Table 3 summarizes the running times of 10 samples from each of the sample sets *SS1–SS8*. All measurements were done on a Ubuntu 12.04 (64-bit) operating system running on a Intel® Core™ i5-4200U CPU at 1.60 GHz × 4 with 5.6-GB random access memory. It is important to note that for ViQuaS and QuRe, input reads were given in FASTA format, whereas ShoRAH was provided with BAM alignment format, and PredictHaplo was provided with SAM alignment format. Therefore, indicated running times of ShoRAH and PredictHaplo do not include the time taken to perform sequence alignment. ViQuaS demonstrated the best time efficiency under all simulation configurations. The presence of indels and the level of technical sequencing errors are observed to have minimal effect on the time efficiency of all methods except ShoRAH. For a given read length, running time is observed to have a direct proportionality to the number of reads, but a simple relationship between the running time and the read length under a constant coverage level is not observed in these results.

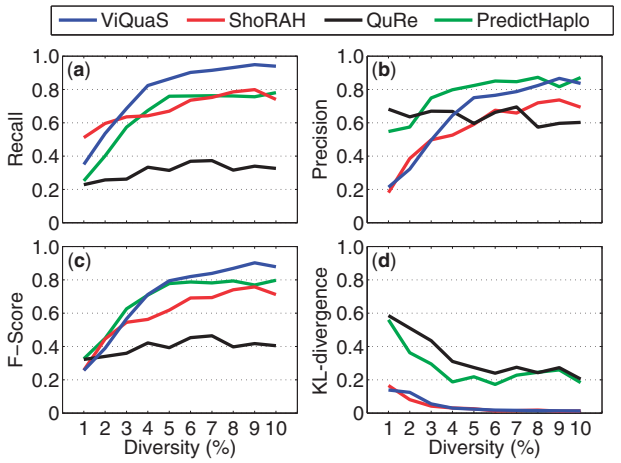


Fig. 1. Performance comparison of ViQuaS, ShoRAH, QuRe and PredictHaplo on *SS1* ($L_r = 200$ bp, $e = 0\%$, only SNPs are present ($T = P$) and $n_{total} = 30\,000$). L_r is the mean read length (read length $\sim \mathcal{N}(L_r, 25)$), e is the substitutional error probability of the reads and n_{total} is the total number of reads. Each point indicates mean value of the measure

3.2 Performance comparison on real HBV mutations

Figure 3 shows the performance comparison of ViQuaS, ShoRAH, QuRe and PredictHaplo on the dataset simulated with real HBV mutations (*SS9*). The performance patterns with varying *Diversity* and N_s values observed for entirely simulated datasets (*SS1–SS8*) were similarly observed for *SS9* as well. Also, ViQuaS outperformed ShoRAH, QuRe and PredictHaplo in terms of total strain reconstruction (Supplementary Table S4 of Supplementary File S5).

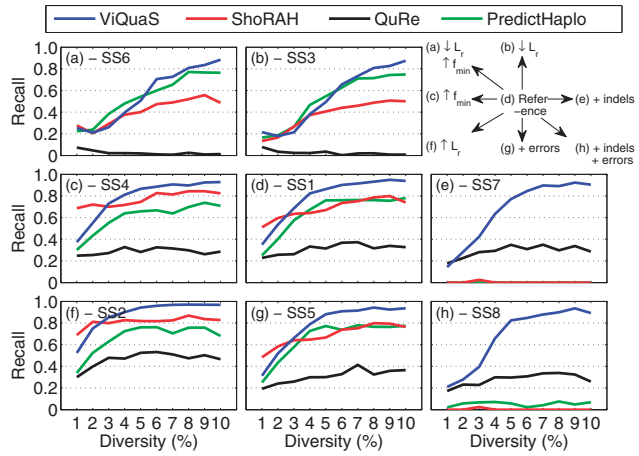
3.3 Performance comparison on real HIV-1 strains

Supplementary Table S5 of Supplementary File S5 summarizes the performance comparison between the four methods on *SS10* and *SS11*. All four methods were able to reconstruct a higher number of true strains under a longer read length, whereas ViQuaS showed the best performance among the four methods under both *SS10* and *SS11*.

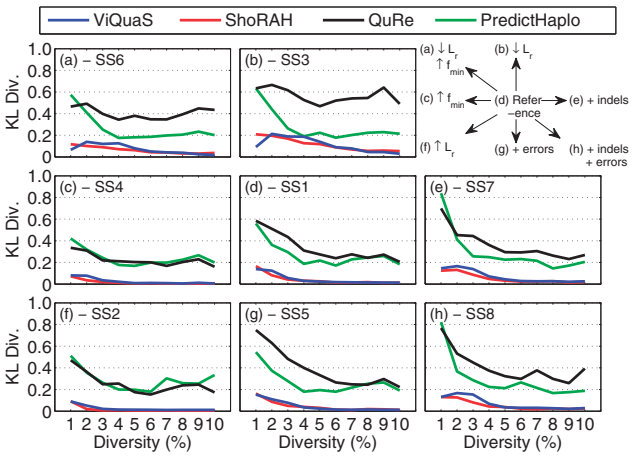
3.4 Performance comparison on in vitro population real Illumina HIV-1 data

Given the 376 545 number of quality filtered reads derived from the *5-Virus-Mix in vitro* population as input, none of the methods used in our benchmark study were able to successfully reconstruct any of the five known strains in the population. This confirmed our observation on simulated data that the global reconstruction of haplotypes becomes harder when the ratio between the genomic region of interest and the read length grows high.

ViQuaS reconstructed 15 strains where the relative frequencies of the five most abundant strains were estimated to be 20.4%, 14.2%, 10.7%, 8.3% and 8.2%, whereas PredictHaplo reconstructed 23 strains out of which the 5 most abundant strains were estimated to be 18.4%, 14.0%, 10.1%, 8.4% and 8.4%. Even though the estimated frequencies show similar profiles, the reconstructed strains did not show such agreement among the two



I - Variations in performance with L_r, e, indels and f_{min} in terms of *Recall*. Each point indicates mean value of the measure.



II - Variations in performance with L_r, e, indels and f_{min} in terms of *KL-Divergence*. Each point indicates mean value of the measure.

Fig. 2. Variations in (**I**)-total strain sequence reconstruction and (**II**)-strain frequency estimation performance of ViQuaS, ShoRAH, QuRe and PredictHaplo with L_r, e, presence of indels in the strains and f_{min}. L_r is the mean read length (read length $\sim \mathcal{N}(L_r, 25)$), e is the substitutional error probability of the reads and f_{min} is the theoretically reconstructible minimum relative frequency

Table 2. Performance variation in terms of *F-score* with technical sequencing errors

Population parameters	Method	$e = 0\%$	$e = 0.1\%$ (1:0)	$e = 0.3\%$ (4:1)
Diversity = 1%, $N_s = 3$	ViQuaS	0.083	0.218	0.174
	QuRe	0.280	0.450	0.267
	ShoRAH	0.000	0.000	0.000
	PredictHaplo	0.000	0.000	0.000
Diversity = 7%, $N_s = 75$	ViQuaS	0.702	0.682	0.413
	QuRe	0.421	0.542	0.364
	ShoRAH	0.000	0.000	0.000
	PredictHaplo	0.000	0.000	0.000

$e = x\%(y:z)$ denotes a technical sequencing error probability of $x\%$ with a $y:z$ ratio between the mismatch errors and indel errors. Each value indicates mean value of the measure.

Table 3. Running time comparison: *SS1–SS8*

Sample set name	ViQuaS	ShoRAH	QuRe	PredictHaplo
SS1	**100.16**	311.61	1640.74	484.94
SS2	**224.60**	257.49	1745.03	490.51
SS3	**124.23**	564.13	1827.87	390.07
SS4	**35.39**	167.68	298.73	269.06
SS5	**101.55**	308.20	1715.03	549.54
SS6	**60.29**	320.07	713.37	275.43
SS7	**101.23**	343.35	2009.11	566.42
SS8	**102.26**	326.06	2017.94	602.75

All reported running time measures are in seconds and indicate the average over 10 samples. From each dataset *SS1–SS8*, the 10 most complex (i.e. *Diversity* = 10% and $N_s = 100$) samples were chosen for this analysis. All measurements were done on an Ubuntu 12.04 (64-bit) operating system running on a Intel® Core™ i5-4200U CPU at 1.60 GHz × 4 with 5.6-GB random access memory. For ViQuaS and QuRe, input reads were given in FASTA format, whereas ShoRAH was provided with BAM alignment format, and PredictHaplo was provided with SAM alignment format. Indicated running times of ShoRAH and PredictHaplo do not include the time taken to perform sequence alignment. Bold face figures show the best performance measure in each row.

methods. On the other hand, ShoRAH could not produce a proper output on the dataset as it experienced a runtime error and QuRe could not handle the volume of input data and resulted in a memory handling error without producing an output. A comparison between the methods in terms of average pair-wise alignment cost between the known and reconstructed haplotypes is presented in Table 4. We observed that both ViQuaS and PredictHaplo have reconstructed strains with similar pair-wise alignment distances from the true strains with the exception of ViQuaS becoming very close to reconstructing HIV $- 1_{HXB2}$ (pair-wise alignment cost=197).

3.5 Performance on real Illumina FMDV data

Ideally, the reconstructed populations should be identical for the two input datasets ERR180978 and ERR180979 as they were derived from the same biological sample. Indicators of similarity between the pairs of populations reconstructed using the ViQuaS, ShoRAH and QuRe are summarized in Supplementary Table S6 of Supplementary File S5.

On ERR180978, ViQuaS reconstructed 153 strains out of which 62 strains (all 62 were <0.2% relative frequency) had internal stop codons. On ERR180979, ViQuaS reconstructed 52 strains out of which 35 strains (all 35 were <0.2% relative frequency) had internal stop codons. Both spectra contained the same dominant strain with relative frequencies of 97.29% (ERR180978) and 99.08% (ERR180979), respectively.

On ERR180978, ShoRAH reconstructed 114 strains out of which 14 had internal stop codons and on ERR180979, 195 strains were reconstructed out of which 3 had internal stop codons. On the other hand, ShoRAH reconstructed two dominant strains of 40.73% and 31.90% for ERR180978 and a single dominant strain of 96.53% for ERR180979. Furthermore, both dominant strains in ERR180978 were different from the dominant strain of ERR180979 indicating a clear deviation from identical spectra.

On ERR180978, QuRe reconstructed four strains out of which none had internal stop codons and on ERR180979, two strains were reconstructed, none of which had internal stop codons. Similar to ShoRAH, QuRe reconstructed two dominant strains of 50.32% and 48.99% for ERR180978 and a single dominant strain of 96.30% for ERR180979. Also, both dominant strains in ERR180978 were different from the dominant strain of ERR180979 indicating a clear deviation from identical spectra.

On ERR180978, PredictHaplo reconstructed two strains out of which none had internal stop codons and on ERR180979, one strain was reconstructed, which did not have internal stop codons. Similar to ViQuaS, both spectra reconstructed by PredictHaplo contained a single dominant strain with relative frequencies of 98.43% (ERR180978) and 100.00% (ERR180979), respectively. Even though PredictHaplo reconstructed similar spectra on ERR180978 and ERR180979, it is evident that it significantly underestimates the richness of the quasispecies populations confirming the observation under simulated data.

As second criteria of justifying the better performance of ViQuaS over ShoRAH and QuRe on sequence reads with Illumina platform-specific technical sequencing errors, we simulated 10 quasispecies populations each with a sequencing error profile that mimics Illumina platform, modelled in Korbel *et al.* (2009) and implemented in the read simulator software named Grinder (Angly *et al.*, 2012), 60-bp read length and $n_{total} = 100\,000$. Having observed the performance of the three methods over different *Diversity* and N_s values (Fig. 2), we set the *Diversity* of these 10 samples to be around 7% to minimize performance degradation due to a low *Diversity*. Also, we selected $N_s = 75$ to eliminate undue bias towards any particular method by using a very low or very high value for N_s. Strains were derived from the same 339-nt-long region analysed using ERR180978 and ERR180979. The overall error rate was calculated to be $\sim 0.11\%$ for each sample. It should be noted that ShoRAH and QuRe were designed for Roche 454 reads. The comparison of performance of ViQuaS, ShoRAH, QuRe and PredictHaplo on this dataset is presented under Supplementary Table S7 of Supplementary File S5, which highlights the better performance of ViQuaS under Illumina error profile. Especially, the clear improvement in *Recall* and *KL-divergence* provides evidence to conclude that the populations reconstructed by ViQuaS on ERR180978 and ERR180979 have a much higher likelihood of representing the actual but unknown FMDV populations in the biological sample.

3.6 Performance on real Roche 454 HIV-1 data

Analysis of the dataset V11909 (Wang *et al.*, 2007) using ViQuaS resulted 24 different strains out of which 6 strains had internal stop codons in the *protease* and *reverse transcriptase* genes due to the presence of indels. We used two criteria to evaluate the biological validity of the remaining 18 strains. First criterion was to check the agreement between the mutations present in the reconstructed strains and that are published in Rhee *et al.* (2003). All 18 strains under consideration produce amino acid sequences in the *protease*

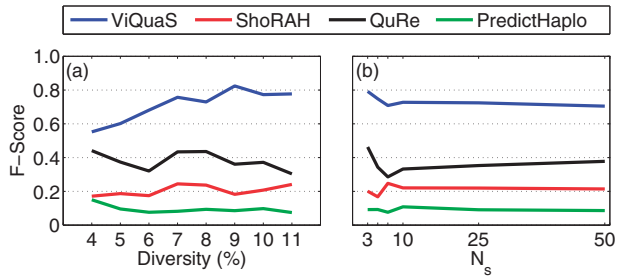


Fig. 3. Performance comparison with ShoRAH and QuRe on *SS9* ($L_r = 100$ bp, $e = 0\%$, both SNPs and indels are present in strains ($T = P, I, D$), $n_{total} = 30\,000$ and $f_{min} = 1.4\%$). L_r is the mean read length (read length $\sim \mathcal{N}(L_r, 25)$), e is the substitutional error probability of the reads, n_{total} is the total number of reads and f_{min} is the theoretically reconstructible minimum relative frequency. Each point indicates mean value of the measure

Table 4. Average pair-wise alignment cost of strains reconstructed by ViQuaS, ShoRAH, QuRe and PredictHaplo on *5-Virus-Mix* HIV-1 dataset with respect to the five known input strains

Input strain	ViQuaS	PredictHaplo	ShoRAH	QuRe
$HIV-1_{89.6}$	673	554	NA	NA
$HIV-1_{HXB2}$	**197**	515	NA	NA
$HIV-1_{JR-CSF}$	612	530	NA	NA
$HIV-1_{NL4-3}$	591	527	NA	NA
$HIV-1_{YU2}$	554	544	NA	NA

Pair-wise alignment was performed using the Smith–Watermann algorithm with the following parameters. Match score = 1, mismatch penalty = −3, gap opening penalty = −5 and gap extension penalty = −2. NA indicates that a result is "Not Available". Bold face figure show the minimum average pair-wise alignment cost.

and *reverse transcriptase* regions with mutations agreeing with Rhee *et al.* (2003) except for one strain which had the amino acid E at the 65th amino acid position of *reverse transcriptase* gene that does not agree with Rhee *et al.* (2003). As a second criterion, we checked these 18 strains for the presence of 40 highly correlated and 10 least correlated mutation pairs in the *protease* and *reverse transcriptase* regions published in Rhee *et al.* (2007). None of the 10 least correlated mutation pairs are present in these strains, and 14 and 2 out of the 40 highly correlated mutation pairs in the *protease* and *reverse transcriptase* regions, respectively, are present in these 23 strains indicating the biological validity of the strains.

Results of the two validation criteria on strains reconstructed by ViQuaS, ShoRAH, QuRe and PredictHaplo are summarized in Table 5.

Analysis of the dataset V11909 using ShoRAH resulted 184 different strains out of which 131 strains had internal stop codons in the *protease* and *reverse transcriptase* genes due to the presence of indels. Out of the remaining 53 strains without internal stop codons, 3 strains including the highest frequency strain had the amino acid K at the 89th amino acid position of the *protease* gene and 3 strains had Y at the 48th, 5 strains had E at the 65th, 4 strains had N at the 73rd and 1 strain had V at the 214th amino acid positions of the *reverse transcriptase* gene, all of which do not agree with Rhee *et al.* (2003). Under the second validation criteria, these 53 strains contained zero and 2 out of the 10 least correlated mutation pairs in the *protease* and *reverse transcriptase* regions, respectively, and 14 and 9 out of the 40 highly correlated mutation pairs in the *protease* and *reverse transcriptase* regions, respectively.

Analysis of the dataset V11909 using QuRe resulted 15 different strains out of which 13 strains had internal stop codons in the *protease* and *reverse transcriptase* genes due to the presence of indels. Out of the remaining two strains without internal stop codons, all mutations agree with Rhee *et al.* (2003). Under the second validation criteria, these two strains contained zero and 2 out of the 10 least correlated mutation pairs in the *protease* and *reverse transcriptase* regions, respectively, and 11 and 2 out of the 40 highly correlated mutation pairs in the *protease* and *reverse transcriptase* regions, respectively.

Analysis of the dataset V11909 using PredictHaplo resulted in only 1 strain, which had no internal stop codons in the *protease* and *reverse transcriptase* genes. This only strain had I at the 73rd position that does not agree with Rhee *et al.* (2003). Under the second validation criteria, this strain contained none of the 10 least correlated mutation pairs in the *protease* or the *reverse transcriptase* regions, respectively, and 9 and 7 out of the 40 highly correlated mutation pairs in the *protease* and *reverse transcriptase* regions, respectively.

Table 5. Biological validation of strains reconstructed by ViQuaS, ShoRAH, QuRe and PredictHaplo on V11909 dataset

Method	1st criterion			2nd criterion			
	S_T	S_V (S_V/S_T)	N_{mu}	$N_{p,l}$	$N_{r,l}$	$N_{p,h}$	$N_{r,h}$
ViQuaS	24	18 (75.0%)	1	0/10	0/10	14/40	2/40
ShoRAH	184	53 (28.8%)	16	0/10	2/10	14/40	9/40
QuRe	15	2 (13.3%)	0	0/10	2/10	11/40	2/40
PredictHaplo	1	1 (100.0%)	1	0/10	0/10	9/40	7/40

S_T, total number of strains reconstructed; S_V, no. of strains without internal stop codons; N_{mu}, no. of mutations contained in the stains without internal stop codons not agreeing with Rhee *et al.* (2003); $N_{p,l}$, no. of least correlated mutation pairs present in *protease*; $N_{r,l}$, no. of least correlated mutation pairs present in *reverse transcriptase*; $N_{p,h}$, no. of highly correlated mutation pairs present in *protease*; $N_{r,h}$, no. of highly correlated mutation pairs present in *reverse transcriptase*.

Analysis of the dataset V11909 using PredictHaplo resulted in only 1 strain, which had no internal stop codons in the *protease* and *reverse transcriptase* genes. This only strain had I at the 73rd position that does not agree with Rhee *et al.* (2003). Under the second validation criteria, this strain contained none of the 10 least correlated mutation pairs in the *protease* or the *reverse transcriptase* regions, respectively, and 9 and 7 out of the 40 highly correlated mutation pairs in the *protease* and *reverse transcriptase* regions, respectively.

The results of the two validation criteria used here (Table 5) provide clear evidence to conclude that ViQuaS reconstructed a population with a higher degree of biological validity compared with ShoRAH, QuRe and PredictHaplo.

4 Discussion

We present in this article, a novel method for QSR named ViQuaS. The main novelty is the proposed reference-assisted *de novo* assembly algorithm constituted by the three consecutive steps *de novo assembly*, *Mutation calling* and *Chimeric error correction*. The greedy graph-based *de novo* assembly algorithm and the subsequent mutation loci-guided chimeric error correction algorithm provide ViQuaS with a significant advantage over ShoRAH, QuRe and PredictHaplo when it is used to analyse short NGS reads.

Our method clearly outperformed ShoRAH, QuRe and PredictHaplo on the 6810 simulated datasets used in the benchmark study, demonstrating the capability of reconstructing quasispecies spectra using short NGS read samples (100 bp, 200 bp and 300 bp) with comparatively less susceptibility to the presence of sequencing errors and indels. Given a viral genomic region of interest, the average read length of the NGS read sample and the *Diversity* of the quasispecies population were identified as the two most critical factors affecting the performance of ViQuaS, provided that the technical sequencing error rate is kept at a minimum level.

It is evident that strain sequence assembly gets easier when the *Diversity* of a quasispecies population becomes higher, provided that read data are available covering the whole genomic region of

interest. This pattern of performance was common to all four methods considered in our study. Under all eight different simulation conditions, ViQuaS gave the highest average *Recall* values demonstrating the best total strain reconstruction capability. ViQuaS and ShoRAH performed at comparable levels in terms of strain frequency estimation, while clearly outperforming QuRe and PredictHaplo.

Due to the unavailability of real datasets with known quasispecies populations, the best alternative for validating these methods under real world conditions was to use simulated datasets such as *SS9–SS11* that can be assumed to resemble such conditions. Agreement between the performance patterns of entirely simulated data (*SS1–SS8*) and data resembling real conditions (*SS9–SS11*) further confirmed the enhanced performance of ViQuaS over ShoRAH, QuRe and PredictHaplo.

Even though the correctness of the reconstructed strains cannot be evaluated due to the unavailability of exact nucleotide content of the strains in the real FMDV samples (ERR180978 and ERR180979) and real HIV-1 dataset V11909 (Wang *et al.*, 2007), the consistency of results produced by ViQuaS on real FMDV samples and the higher degree of biological validity of the strains reconstructed from real HIV-1 data by ViQuaS indicate the applicability of ViQuaS on real sequence data and better performance over ShoRAH, QuRe and PredictHaplo. In addition, simulated samples with Illumina sequencing parameters further confirmed the superior performance of ViQuaS.

In this study, we performed a comprehensive comparison of performance between ViQuaS, ShoRAH, QuRe and PredictHaplo in a broad range of *Diversity* and N_s values. Because of the requirement of ViSpA (Astrovskaya *et al.*, 2011) to have the *Diversity* of a sample as an input parameter, which is impractical for real data, and its poor performance under read lengths in the range of 100 bp and 200 bp, we did not use it in this study. Another method found in literature named QuasiRecomb (Töpfer *et al.*, 2013) was also omitted from further analysis due to its poor performance even on *SS2* (on which the other three method used in the analysis showed their best performance). Two of the most significant observations of the comparison study are the drastic drop of performance of ShoRAH and PredictHaplo under the presence of indels [refer Fig. 2I(e) and Fig. 2I(h)] and of QuRe under a shorter (100 bp) read length [Fig. 2I(b)]. On the contrary, ViQuaS was much less susceptible against both indels and reduced read lengths. All four methods handled the sequencing error rate of 0.1% without a significant drop of performance. In particular, we observed a better performance of ShoRAH and PredictHaplo on *SS9* in the presence of indels compared with their poor performance on *SS7*. This observation is due to the fact that some strains in *SS9* do not have insertions (but SNPs and deletions), which are correctly reconstructed, while every strain in *SS7* has both insertions and deletions. It was not evident that ShoRAH and PredictHaplo eliminates all insertion considering them as technical sequencing errors because some reconstructed strains in *SS7* contain a few (but not all) correct insertions. This also explains the improved performance of ShoRAH on *SS10* and *SS11* compared with QuRe (Supplementary Table S5), as the 10 real HIV strains used in *SS10* and *SS11* does not have any mutations in the form of insertions. On the other hand, QuRe performed better than ShoRAH and PredictHaplo on *SS9* (Supplementary Table S4) where some of the simulated HBV strains contain insertions.

The second significant factor is that ShoRAH, QuRe and PredictHaplo define local analysis windows *a priori*. These window boundaries restrict the growth of the read clusters along the nucleotide space of the region of interest. In contrast, ViQuaS allows the contigs to extend as far as the reads provide information supporting the co-occurrence of mutations. This unrestricted growth of contigs becomes advantageous in the total strain reconstruction stage. In ShoRAH, each of the paths in the *cover* of the read graph contains a number of nodes equal to the number of windows (Eriksson *et al.*, 2008). In QuRe, the complexity of the global reconstruction algorithm is proportional to the number of windows (Prosperi *et al.*, 2011). The window length is typically set equal to the average read length in both ShoRAH and QuRe. Hence, the algorithmic complexity of the global reconstruction methods of both ShoRAH and QuRe increases in proportionate to the ratio between the genomic region of interest and the read length. These factors significantly affect the performance of the respective methods as been discussed in Eriksson *et al.* (2008) and Prosperi *et al.* (2011). Similar to ShoRAH, PredictHaplo also uses a Dirichlet process mixture model to analyse reads within fixed local windows and uses the local analysis results as prior knowledge in its propagating probabilistic global strain inference method (Prabhakaran *et al.*, 2014). In contrast, the *Global Spectrum Reconstruction* algorithm in ViQuaS, which was extended from Prosperi *et al.* (2011), takes advantage of the unrestricted *de novo* assembly algorithm and dynamically defines the local haplotype boundaries for each strain, at which the decisions for connecting local haplotypes are made. This dynamic definition of local haplotype boundaries significantly reduces the number of heuristic decisions to be made when connecting local haplotypes providing a clear advantage over QuRe. This advantage becomes prominent especially under shorter read lengths.

The third source of performance enhancement in ViQuaS is the improvement of estimated strain frequency values (demonstrated in terms of *KL-divergence* and *root mean-squared error*) resulted from the second and third major extensions to the algorithm in Prosperi *et al.* (2011) described under the subsection 2.2.6.

Limitations of ViQuaS include the requirement of higher computational resources when analysing longer read lengths owing to the use of Smith–Waterman algorithm in the pipeline. However, this computational complexity does not degrade the performance of the pipeline. A shorter read length compared with the genomic region of interest reduced coverage depth of the sequences and a higher rate of technical sequencing errors were identified as the factors influencing negatively on the performance of ViQuaS. In fact, these factors were common to all four methods in our study and ViQuaS was observed to be the least susceptible to them. Also, all four methods failed to reconstruct strains with a relative frequency less than $f_{\min,}$ and the correctly reconstructible minimum relative frequency increased with the presence of technical sequencing errors.

We notice that the performance of the methods analyzed vary with different population characteristics and sequencing parameters. Therefore, it is a difficult task to define parameters within which a certain method performs at an acceptable level. In fact, the acceptable level of performance varies according to the needs of the individual user. Hence, the decision on choosing the best method of analysis and the appropriate sequencing parameters is left to the end user. Furthermore, the longest region that can be successfully reconstructed depends on both the read length and *Diversity*. For instance, when the region of interest and the *Diversity* are given, we observe that the performance increases with increasing read length. This does not imply that we can expect the same levels of performance with the same *Diversity* and read length over a larger genomic region. A thorough analysis of this fact is presented in Zagordi *et al.* (2012b).

We anticipate the use of ViQuaS to extend beyond quasispecies spectra profiling with applications in other *omics* fields including

cancer genomics, metagenomics and immunogenomics. For instance, applications such as estimating genetic diversity of tumour cells and understanding gene heterogeneity and diversity of microbiomes in various environments such as human gut or waste treatment plants may benefit from the underlying methodology.

Acknowledgements

We wish to acknowledge the contributions of Rene Warren for providing valuable feedback on SSAKE algorithm and Osvaldo Zagordi, Irina Astrovskaya and Mattia Prosperi for their support during the comparison study.

Funding

This work was partially supported by Australian Research Council [grant number LP140100670] and by MIFRS and MIRS scholarships of The University of Melbourne (to D.J.).

Conflict of interest: none declared.

References

Angly,F.E. *et al.* (2012) Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res.,* **40**, e94.

Astrovskaya,I. *et al.* (2011) Inferring viral quasispecies spectra from 454 pyrosequencing reads. *BMC Bioinformatics,* **12**, 1–10.

Baldick,C. *et al.* (2008) Hepatitis b virus quasispecies susceptibility to entecavir confirms the relationship between genotypic resistance and patient virologic response. *J. Hepatol.,* **48**, 895–902.

Beerenwinkel,N. and Zagordi,O. (2011) Ultra-deep sequencing for the analysis of viral populations. *Curr. Opin. Virol.,* **1**, 413–418.

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

Brunetto,M. *et al.* (1999) Hepatitis b virus mutants. *Intervirology,* **42**, 69–80.

Carter,J.B. and Saunders,V.A. (2007) *Virology: Principles and Applications*. John Wiley, Chichester, UK.

Dimmock,N.J. *et al.* (2007) *Introduction to Modern Virology*. Blackwell Pub., Malden, MA.

Eriksson,N. *et al.* (2008) Viral population estimation using pyrosequencing. *PLoS Comput. Biol.,* **4**, 1–13.

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

Huang,A. *et al.* (2011) Qcolors: An algorithm for conservative viral quasispecies reconstruction from short and non-contiguous next generation sequencing reads. In: *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, IEEE, Atlanta, GA, pp. 130–136.

Korbel,J. *et al.* (2009) Pemer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol.,* **10**, R23.

Kullback,S. and Leibler,R. (1951) On information and sufficiency. *Ann. Math. Stat.*, **22**, 7986.

Kunin,V. *et al.* (2008) A bioinformatician's guide to metagenomics. *Microbiol. Mol. Biol. Rev.,* **72**, 557–578.

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

Mancuso,N. *et al.* (2011) Viral quasispecies reconstruction from amplicon 454 pyrosequencing reads. In: *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, IEEE, Atlanta, GA, pp. 94–101.

Miller,J.R. *et al.* (2010) Assembly algorithms for next-generation sequencing data. *Genomics,* **95**, 315–327.

Morelli,M. *et al.* (2013) Evolution of foot-and-mouth disease virus intra-sample sequence diversity during serial transmission in bovine hosts. *Vet. Res.,* **44**, 12.

Nishijima,N. *et al.* (2012) Dynamics of hepatitis b virus quasispecies in association with nucleos(t)ide analogue treatment determined by ultra-deep sequencing. *PLoS One,* **7**, 1–10.

O'Neil,S. and Emrich,S. (2012) Haplotype and minimum-chimerism consensus determination using short sequence data. *BMC Genomics,* **13**(Suppl. 2), S4.

Prabhakaran,S. *et al.* (2010) HIV-haplotype inference using a constraint-based dirichlet process mixture model. *Machine Learn. Comput. Biol. NIPS Workshop*.

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

Prosperi,M. *et al.* (2011) Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing. *BMC Bioinformatics,* **12**, 5.

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

Quince,C. *et al.* (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics,* **12**, 38.

Ramakrishnan,M.A. *et al.* (2009) The feasibility of using high resolution genome sequencing of influenza a viruses to detect mixed infections and quasispecies. *PLoS One,* **4**, e7105.

Rhee,S.-Y. *et al.* (2003) Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.,* **31**, 298–303.

Rhee,S.-Y. *et al.* (2007) HIV-1 subtype b protease and reverse transcriptase amino acid covariation. *PLoS Comput. Biol.,* **3**, e87.

Schirmer,M. *et al.* (2012) Benchmarking of viral haplotype reconstruction programmes: an overview of the capacities and limitations of currently available programmes. *Brief. Bioinform.,* **15**, 431–442.

Smith,T. and Waterman,M. (1981) Identification of common molecular subsequences. *J. Mol. Biol.,* **147**, 195–197.

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

Vignuzzi,M. *et al.* (2006) Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature,* **439**, 344–348.

Wang,C. *et al.* (2007) Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res.,* **17**, 1195–1201.

Warren,R.L. *et al.* (2007) Assembling millions of short DNA sequences using ssake. *Bioinformatics,* **23**, 500–501.

Zagordi,O. *et al.* (2010) Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Res.,* **38**, 7400–7409.

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

Zagordi,O. *et al.* (2012a) Probabilistic inference of viral quasispecies subject to recombination. In: B. Chor (ed.) *Research in Computational Molecular Biology,* volume 7262 of *Lecture Notes in Computer Science*. Springer, Berlin, Germany, pp. 342–354.

Zagordi,O. *et al.* (2012b) Read length versus depth of coverage for viral quasispecies reconstruction. *PLoS One,* **7**, e47046.