BRIEFINGS IN BIOINFORMATICS. VOL 15. NO 3. 431–442
Advance Access published on 19 December 2012
doi:10.1093/bib/bbs081

Benchmarking of viral haplotype reconstruction programmes: an overview of the capacities and limitations of currently available programmes

Melanie Schirmer, William T. Sloan and Christopher Quince

Submitted: 16th August 2012; Received (in revised form): 12th November 2012

Abstract

Viral haplotype reconstruction from a set of observed reads is one of the most challenging problems in bioinformatics today. Next-generation sequencing technologies enable us to detect single-nucleotide polymorphisms (SNPs) of haplotypes—even if the haplotypes appear at low frequencies. However, there are two major problems. First, we need to distinguish real SNPs from sequencing errors. Second, we need to determine which SNPs occur on the same haplotype, which cannot be inferred from the reads if the distance between SNPs on a haplotype exceeds the read length. We conducted an independent benchmarking study that directly compares the currently available viral haplotype reconstruction programmes. We also present nine *in silico* data sets that we generated to reflect biologically plausible populations. For these data sets, we simulated 454 and Illumina reads and applied the programmes to test their capacity to reconstruct whole genomes and individual genes. We developed a novel statistical framework to demonstrate the strengths and limitations of the programmes. Our benchmarking demonstrated that all the programmes we tested performed poorly when sequence divergence was low and failed to recover haplotype populations with rare haplotypes.

Keywords: *benchmarking; viral haplotype reconstruction programmes; quasispecies; in silico data sets; statistics for validation*

INTRODUCTION

RNA viruses are the most common pathogens for humans and animals. The human immunodeficiency virus (HIV), the hepatitis C virus (HCV) and the foot-and-mouth virus (FMV) are just some examples of RNA viruses that pose major health threats. There is no general treatment available and for many viruses we have not been able to develop effective vaccines. RNA viruses are able to mutate quickly and cause acute epidemics with novel strains. To develop successful and preventive treatments and to act quickly when a new viral strain occurs, we need more detailed and accurate information about the population structure, the mutations and the viral haplotypes for the specific infection. The development of next-generation sequencing (NGS) technologies opens up the opportunity to respond quickly to outbreaks and gain a better understanding of viral populations. However, there are significant challenges that we need to overcome to reconstruct viral populations from NGS data.

The lack of proof checking during replication causes RNA viruses to have mutation rates about a million times larger than within human cells [1].

Corresponding author. Melanie Schirmer, University of Glasgow, Rankine Building, Oakfield Avenue, Glasgow G12 8LT, UK. Tel.: +44-141-330-6311. E-mail: m.schirmer.1@research.gla.ac.uk

Melanie Schirmer is a PhD student of Bioinformatics in the Computational Microbial Genomics Group at the University of Glasgow. She is working on algorithms for viral haplotype reconstruction and metagenomics.

William T. Sloan is a professor of Environmental Engineering at the University of Glasgow. His research centres on developing mathematical models of microbial communities in natural and engineered systems.

Christopher Quince is a reader at the University of Glasgow working on microbial diversity and population genetics.

This results in a population of closely related genomes, a so-called quasispecies. The high mutation rate is dangerous for the virus, since it results in many non-viable clones but it also provides the virus with a large number of potentially beneficial mutations allowing it to adapt quickly to changing environments during infection. It is likely that the haplotypes that enable the virus to survive selective pressure pre-exist in the population [2]. Thus, it is essential to determine all haplotypes to develop effective treatments and vaccines.

With NGS technologies, we are now able to detect SNPs in a viral population—even for low abundance haplotypes. However, reads from any sequencing technology contain noise from polymerase chain reaction (PCR) amplification and platform specific noise, which we need to distinguish from real diversity to be able to reconstruct the haplotypes accurately. Another major challenge is that, because read lengths are short, it can be difficult, or sometimes impossible, to determine which SNPs reside on the same haplotype. The three columns in Figure 1 give a schematic of the different steps in the process of reconstructing a population from NGS data. In the first column, we can see two haplotypes occurring at different abundances. They have one SNP in common. The next column displays a set of observed reads obtained from NGS technologies including sequencing noise. The third presents different scenarios that can occur during reconstruction. In the first scenario, the reconstruction is successful. We encounter two reads that contain SNPs and have a sufficient overlap to be assembled correctly into a contig of the first haplotype. In the second, the distance between SNPs exceeds the read length which means we cannot map the reads to a haplotype based on read overlap. In the third, noise is mistaken for diversity. And in the forth, we cannot infer the origin of the read as the SNP occurs on both haplotypes.

Here, we study the currently available haplotype reconstruction programmes and benchmark their performance across various *in silico* data sets. It is important to know their capabilities and highlight scenarios that might expose their limitations. Our test data sets were deliberately selected to challenge the reconstruction programmes with different sequence divergence and haplotype abundance distributions.

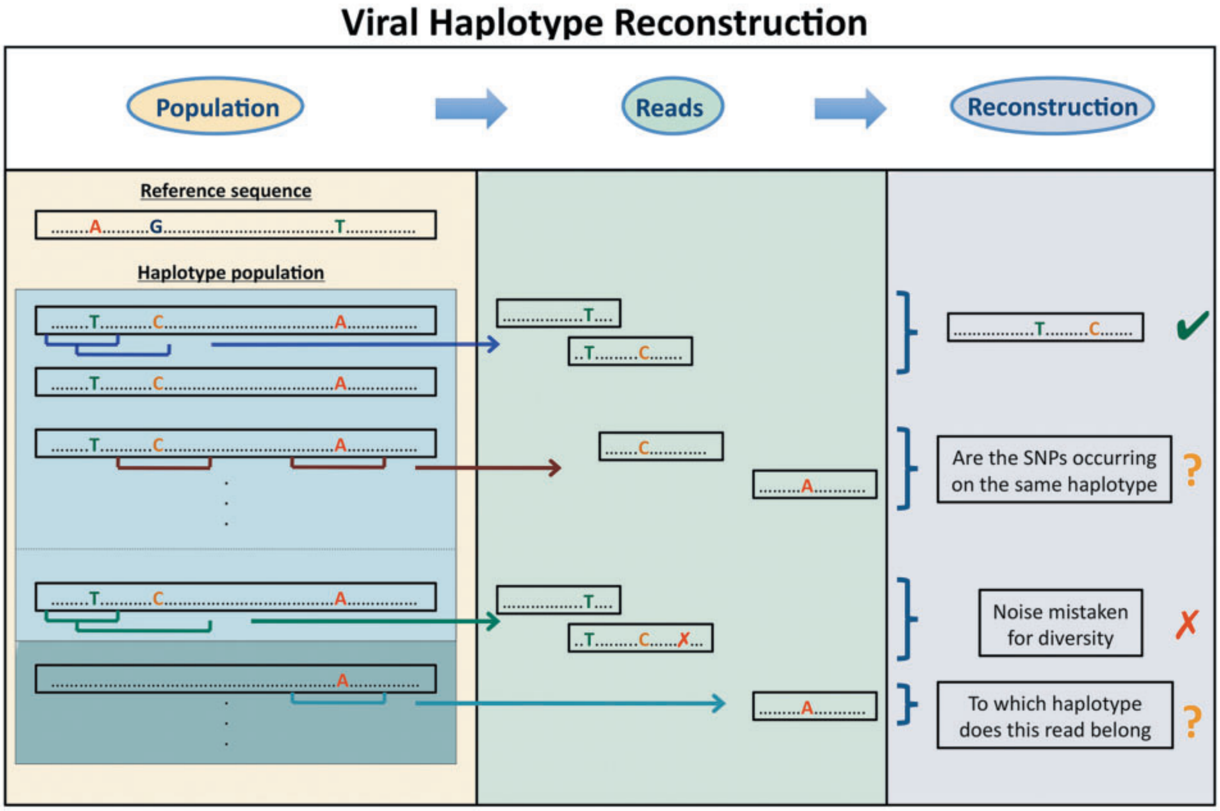
Viral Haplotype Reconstruction
Population
Reads
Reconstruction
Reference sequence
........A.........G..................................T..............
Haplotype population
........T..........C.....................................A..............
........T..........C.....................................A..............
........T..........C.....................................A..............
........T..........C.....................................A..............
....A..............
................T....
..T.........C.......
.......C...........
.........A..........
................T....
..T.........C......X...
.........A..........
................T.........C.......
Are the SNPs occurring on the same haplotype
?
Noise mistaken for diversity
To which haplotype does this read belong
?

Figure 1: Schematic diagram representing the process of reconstructing viral haplotypes from next-generation sequencing reads.

Thus, we could assess the programmes' abilities to reproduce the underlying population structure. We also evaluated the accuracy of the reconstructed haplotypes by outlining how many haplotypes are reconstructed with zero, one and two mismatches. It is important to identify the number of false positives in the reconstructed population to assess the overall accuracy of the reconstruction.

We simulated 454 and Illumina reads for all the test data sets. Table 1 compares the run time (for maximum read length), number of reads per run, read length, yield per run and reagent cost for Illumina and 454 sequencers. The great advantage of 454 reads is that they are longer than Illumina reads which could be important (as seen in Figure 1). On the other hand, Illumina reads provide much higher coverage at much lower cost. None of the programmes that we tested was specifically designed for Illumina reads, but all of them take Illumina fasta files or fastq files. We tested the programmes on the Illumina reads to examine their potential for viral haplotype reconstruction.

THE TEST DATA SETS

We generated nine different *in silico* haplotype populations that vary in sequence diversity and number of haplotypes and we considered different frequency distributions for these populations. These data sets are based on observations from real data sets for FMV and HIV. For FMV we targeted whole genome reconstruction and for HIV we considered single gene reconstruction. The details of the data sets, including the sequence divergence and frequency distributions of the haplotypes, can be found in Table 2.

Simulating the evolution of a single FMV sequence (DSIa–DSIf)

We created six data sets based on an experimental foot-and-mouth virus data set (FMVD). Nucleotide substitution rates are organism specific, gene specific, vary between the four nucleotides (e.g. purine–purine substitutions are more likely than purine–pyrimidine substitutions) and depend on the codon

Table 1: Comparison of different 454 and Illumina sequencing instruments (taken from [3, 4])

Instrument	Run time	Millions of reads/run	Read length	Yield Mb/run	Reagent cost/run ($)	Reagent cost/Mb ($)
454 FLX Titanium	10 h	1	400[a]	500	6200	12.40
454 FLX +	18–20 h	1	700[a]	900	6200	7.00
Illumina HiSeq 2000	8 days	1000	36–100[b]	200 000	20 120	0.10
Illumina GAIIx	14 days	320	35–150[c]	96 000	11 524	0.12
454 GS Junior	10 h	0.10	400[a]	50	1100	22
Illumina MiSeq	26 h	3.4	25–150[d]	1200	750	0.74

[a] Average read length. [b] Possible read lengths: 36 bp, 50 bp, 100 bp. [c] Possible read lengths: 35 bp, 50 bp, 75 bp, 100 bp, 150 bp. [d] Possible read lengths: 25 bp, 35 bp, 100 bp, 150 bp.

Table 2: Overview of all test data sets, including the number of haplotypes in the population, the number of mutations on each haplotype and the frequency distribution of the haplotypes

Data sets	No. haplotypes	Genome size	No. mutations per haplotype	Divergence (%)	Frequency distribution
DSIa	10	8162 bp	10 bp	0.23	Uniform
DSIb	10	8162 bp	10 bp	0.23	Log-normal
DSIc	10	8162 bp	50 bp	1.12	Uniform
DSId	10	8162 bp	50 bp	1.12	Log-normal
DSIe	10	8162 bp	200 bp	3.98	Uniform
DSIf	10	8162 bp	200 bp	3.98	Log-normal
DSIIa	44	2256–2581 bp	2–328 bp	0.08–12.71	Uniform
DSIIb	44	2256–2581 bp	2–328 bp	0.08–12.71	Log-normal
DSIII	4359	8162 bp	1–41 bp	0.01–0.50	Empirical

We used the Levenshtein distance to evaluate the pairwise sequence divergence between the haplotypes. The Levenshtein distance is the minimum number of substitutions and indels to turn one sequence into another. Note that the same mutation can occur on more than one haplotype.

position [5, 6]. The situation is even more complicated in the case of viruses as genes can overlap. Many models work with a substitution matrix where changing probabilities between nucleotides can vary depending on the initial base and the mutated base. With the FMVD, we went one step further and inferred position-specific substitution rates for the entire genome. The sample was sequenced with high coverage (4873×) on the Illumina Genome Analyzer II [7]. The nucleotide frequencies for every position in the genome were inferred from the sequenced reads.

We started with the consensus sequence of FMV and simulated its evolution into 10 haplotypes by introducing mutations at 10, 50 and 200 positions in the genome to create data sets of varying diversity. The position-specific nucleotide frequencies from the experimental data set can be interpreted as a discrete probability distribution giving the probability for each nucleotide (A, C, T and G) to occur at a specific position in the genome. For each mutation, we chose a position in the genome at random with probabilities proportional to the number of observed SNPs. A second random number specified the mutation according to the probability distribution for this position. For each of the three sets of haplotypes, we simulated a population with uniformly distributed haplotype frequencies and log-normal distributed frequencies (Figure 2).

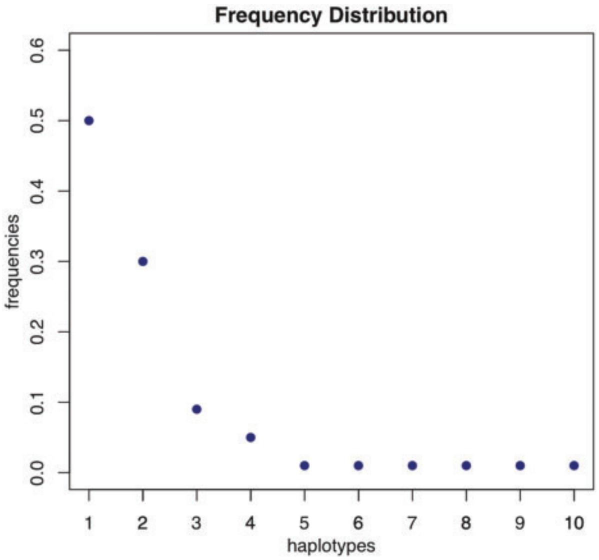

Figure 2: Frequency distribution of the haplotypes for the data sets DSIb, DSId and DSIf.

The HIV-1 envelope gene (DSIIa and DSIIb)

We generated two *in silico* populations of the envelope glycoprotein (env) gene of HIV-1 from data used to assess whether an HIV-1 infection was initiated by a single or multiple viral strains [8]. For one patient, 44 sequences were isolated (GenBank accession number EU577344–EU57787). We used the first sequence as a reference sequence. In comparison to this reference, 40 of the sequences had between 2 and 4 mutations, one had 6, another had 18 and one had major deletions that resulted in a Levenshtein distance of 328. We constructed populations from these sequences where the abundances were distributed uniformly and log-normally ($\mu = 1$, $\sigma = 2$).

An *in silico* FMV population (DSIII)

We generated a more complex *in silico* viral population from the experimental FMVD described above. The SNPs and their frequencies in this *in silico* data set are consistent with the experimental data. Taking the coverage into account, we filtered out mutations that were only observed once. The mutation rate after filtering is $\approx 5.6 \times 10^{-4}$ and conforms to the upper threshold of the mutation rate specified by Nowak [9]. We used our algorithm to create a haplotype population that is based on the SNP frequencies of the filtered FMVD which contains 5479 SNPs occurring at $\approx$48.42% of all positions in the genome. (Note that up to three different SNPs can occur at the same position in the genome.) The algorithm yielded a population of 4359 haplotypes with an average of 6 polymorphisms per haplotype and a maximum is 22 polymorphisms within a single haplotype where the same SNP can occur on multiple haplotypes. The wildtype is dominant in the population and appears with a frequency of $\approx$0.2796 and the majority of the haplotypes are present at very low frequencies with <0.01. This data set is very challenging for a haplotype reconstruction programme, but highly complex haplotype populations occur in nature.

454 read data sets

For each of the data sets, DSIa–DSIf FlowSim (V 0.3) was used to generate 120 000 reads with an average read length of 492 bp that included 454 sequencing noise and PCR noise. For DSIIa, 40 000 reads were generated.

We used our own metagenomic read simulation programme to generate 454 reads for the more complex non-uniform distributions. The read lengths were normally distributed ($\mu = 500$ bp, $\sigma = 80$ bp) which reflects an experimental distribution. We simulated 40 000 reads for DSIIb and 120 000 reads for DSIII, again with 454 and PCR noise.

Illumina read data sets

We used our programme to simulate 75 bp Illumina reads and used the corresponding Illumina error profile from [18]. We simulated 1 million reads for the whole-genome test data sets (DSIa–DSIf and DSIII) and 400 000 reads for the single-gene test data sets (DSIIa and DSIIb).

THE PROGRAMMES

Table 3 gives an overview of the currently available viral haplotype reconstruction programmes. Four of these programmes attempt whole gene/genome reconstruction: ShoRAH, PredictHaplo, QuRe and ViSpA. Two of them, V-Phaser/V-Profiler and QuasiRecomb, attempt reconstruction over local windows that are smaller than the read length. Here, we only benchmarked whole gene/genome programmes. Unfortunately, ViSpA could not be installed and is unsupported, hence three programmes were benchmarked. A more detailed description of the programmes can be found in the Supplementary Material.

MEASURES FOR THE EVALUATION: SIMILARITY AND COMPLETENESS

Two novel measurements were developed to evaluate the accuracy of the haplotype reconstruction programmes. These measures not only take the number of successfully reconstructed haplotypes into account but also their frequencies, the number of mismatches,

Table 3: Overview of the currently available haplotype reconstruction programmes

Programme	Whole gene/ genome reconstruction	Method
ShoRAH [10]	✓	– read-graph approach in combination with solving the maximum weight matching problem for path selection – reconstructs a minimal set of haplotypes – Dirichlet process mixture model for error correction
PredictHaplo [11, 12]	✓	– non-standard clustering problem in combination with a Dirichlet process mixture model – reconstructs the most likely set of haplotypes
QuRe [13]	✓	– error correction according to a Poisson distribution with different parameters for homopolymeric and non-homopolymeric regions – infers multinomial distributions for locally reconstructed haplotypes and uses those for global reconstruction
ViSpA[a] [14]	✓	– read-graph approach where the most probable paths in the graph are selected – reconstructs the most likely set of haplotypes
V-Phaser [15] & V-Profiler [16]	[b]	– V-Phaser uses covariation and an EM algorithm to recalibrate quality scores to detect the SNPs for every position – V-Profiler calculates the frequency of each triplet codon of the accepted nucleotides and constructs the haplotypes
QuasiRecomb [17]	[b]	– jumping hidden Markov model taking recombination events and SNPs into account – samples from the inferred distribution of viral haplotypes

The first three programmes (ShoRAH, PredictHaplo and QuRe) were included in our benchmarking.
[a]Could not be installed and is unsupported. [b]Only attempt reconstruction over a small local window (window size < read length).

the number of false positives and the reconstructed length. We first introduce a measurement for the similarity of the reconstructed haplotypes and the true haplotypes, where only the fraction of the genome is taken into consideration that is covered by the reconstruction; the second measurement reflects the completeness of the reconstructed haplotypes.

Similarity measure

We need a measurement that takes the distance of each reconstructed haplotype to its closest true haplotype into account and at the same time penalises the reconstruction of too many or too few haplotypes. We achieved this by defining a probability distribution based on the reconstructed population and the 'true' population respectively. We then use the Hellinger distance to quantify the similarity between the two distributions. Here, we only take the reconstructed part of the true haplotype sequence into account and measure the distance between two sequences by computing the Levenshtein distance. By allowing *i* mismatches for a reconstructed haplotype to 'equal' a true haplotype, we can see how close the reconstructed population is to the underlying 'true' population.

Let P_1 denote the set of true haplotypes. Then the frequencies of the true haplotypes represent a discrete probability distribution over P_1, which we denote with *f*:

$$f : P_1 \rightarrow]0,1]$$

The set of reconstructed haplotypes is denoted with $\mathbf{P_2}$ and analogously the frequencies of the reconstructed haplotypes represent a discrete probability distribution *g*:

$$g : P_2 \rightarrow]0,1]$$

We now take the union of the two sets:

$$P := P_1 \cup P_2$$

which is the set of all true haplotypes combined with the reconstructed haplotypes that do not match any of the true haplotypes.

We can now extend the probability distribution $f(x)$ to the set $P = p_1, p_2, \ldots, p_{|P|}$ by setting the frequency of any haplotype $p_k \in P \setminus P_1$ to zero. We denote this distribution with $\tilde{f}(x) : P \rightarrow [0,1]$. Analogously, we can define the extension of the probability distribution $g(x)$ to P and denote it with $\tilde{g}(x) : P \rightarrow [0,1]$. So, the two distributions $\tilde{f}$ and $\tilde{g}$ overlap on the set of correctly reconstructed haplotypes.

Let $i \in \mathbb{N}$ denote the number of allowed mismatches. Then a reconstructed haplotype matches a true haplotype if the Levenshtein distance is $\leq i$. So for $i > 0$, an element $p_k \in P_1$ and an element $p_j \in P_2$ can be mapped to the same element in P if the Levenshtein distance $Ldist(p_k, p_j) \leq i$. When allowing mismatches, more than one reconstructed haplotype can match the same true haplotype. In that case, we add up the frequencies of all matching reconstructed haplotypes.

We define the similarity measure (SiM) in terms of the Hellinger distance of the probability distributions $\tilde{f}$ and $\tilde{g}$ as follows:

$$\mathrm{SiM}_i := 1 - H(\tilde{f}, \tilde{g})$$

with

$$H^2(\tilde{f}, \tilde{g}) = \frac{1}{2} \sum_{x = p_0}^{p_{|P|}} \left(\sqrt{\tilde{f}(x)} - \sqrt{\tilde{g}(x)} \right)^2$$

So SiM_0 enforces strict similarity where a reconstructed haplotype must match the 'true' haplotype exactly. Also note that we have $0 \leq \mathrm{SiM_i}$ where zero corresponds to the maximal distance between two distributions (if none of the reconstructed haplotypes matches a true haplotype) and a similarity of one corresponds to the minimal distance (if the true population was exactly reconstructed).

Completeness measure

The completeness measure (CoM) returns the average percentage of the sequence length that a reconstructed haplotype recovered of a true haplotype:

$$\mathrm{CoM} = \frac{1}{n_{\mathrm{rec}}} \sum_{i=0}^{n_{\mathrm{rec}}-1} \frac{\text{length of reconstructed haplotype}}{\text{length of closest true haplotype}}$$

where n_{rec} denotes the total number of reconstructed haplotypes.

RESULTS

We benchmarked PredictHaplo, ShoRAH and QuRe on 454 and Illumina reads. In general, PredictHaplo tended to underestimate the number of haplotypes for the larger populations, whereas

ShoRAH greatly overestimated the population size (e.g. for DSI: 25–55× the real population size). The same was true for QuRe where the population size was up to 25 × the size of the real population. Also, for ShoRAH different sets of parameters yielded very different results on our data sets. On data set DSIc, for example, we ran ShoRAH with 10 different sets of parameters. A Levenshtein distance of 1 was achieved for the best reconstructed haplotype on the 'best run' but for the 'worst run' the Levenshtein distance was 156. PredictHaplo requires fewer parameters and the effect of different choices on the Levenshtein distance is in the range of 1–2. QuRe takes three different parameters. Two of them are specific for the sequencing technology and default values are provided for 454.

The similarity measure for the 454 read data sets

We summarised the results in Table 4. The first two columns show our *in silico* data sets and the average mutual sequence divergence of the haplotypes in the respective population. It is important to interpret the results (especially the Levenshtein distance) with regards to the sequence divergence within the population. For each programme, we stated the results for the best run (after testing the programme with various parameters) including the Levenshtein distance, the number of reconstructed haplotypes and their length. ShoRAH has an additional column that specifies the number of analyses we ran with different parameters followed by the number of analyses that did not complete in brackets (i.e. where ShoRAH aborted computations at some point during the analysis). We ran PredictHaplo with a value of 2000, 3000 and 4000 for the number of reads that are considered over a local window and used an entropy threshold of 0.005. PredictHaplo completed all of the analyses. For QuRe, we used the default settings for the homopolymeric and non-homopolymeric 454 error rates (0.0044 and 0.0007, respectively) and the default number of iterations (10 000 iterations). An increase to 30 000 and 50 000 iterations resulted in much longer running times but no improvement. The results for the similarity measure are visualised in Figures 3 and 4.

Figure 3 includes the SiM_i results ($i = 0$, $i = 1$ and $i = 2$) on the 454 reads for PredictHaplo, QuRe and ShoRAH and the results for PredictHaplo on Illumina reads for each data set. The uniformly distributed data sets (e.g. DSIa) are plotted next to their corresponding log-normally distributed data sets (e.g. DSIb). Overall, PredictHaplo showed the best performance and was able to increasingly reconstruct the population as sequence divergence increased.

In case of the low sequence divergence in DSIa and DSIb, none of the programmes were able to reconstruct any of the haplotypes with up to two mismatches and the majority of the returned

Table 4: 454 reads: overview of the performance of the haplotype reconstruction programmes across all test data sets

		ShoRAH				PredictHaplo			QuRe		
	Diversity	No. analysis (failed)	Best Ldist	No. rec.hap	Length	Best Ldist	No. rec.hap	Length	Best Ldist	No. rec.hap	Length
DSIa	19 bp	14 (5)	7–172	328	8040	9–267	8	8162	7–382	110	7590
DSIb	19 bp	10 (4)	17–160	295	8040	16–374	6	8162	17–216	256	7595
DSIc	91 bp	10 (4)	24–161	259	8040	1, 3–25	10	8162	25–326	114	7582
DSId	91 bp	10 (4)	21–98	551	8107	(3 ×)0, 2–34	10	8162	7–96	210	7589
DSIe	325 bp	10 (4)	1–209	300	8040	(3 ×)0, 1–2	10	8162	a	a	a
DSIf	325 bp	10 (3)	3–231	252	8040	(6 ×)0, 1–3	10	8162	a	a	a
DSIIa	2–328 bp	14 (0)	(2×)1, 2–82	586	2546	0, 8–9	5	2580	1, 4–173	84	2280
DSIIb	2–328 bp	13 (1)	0, 26–302	41	2541	0, 14–422	19	2580	0, 8–989[b]	272	2254
DSIII	1–41 bp	11 (0)	0–672[b]	53	8085	1, 195–1591	7	8162	a	a	a

The first two columns specify the data set and the mutual sequence divergences of the haplotypes. For each programme we specify the results of the best run in three columns. That includes the range of Levenshtein distances between each of the reconstructed haplotypes and the corresponding closest "true" haplotype, the number of reconstructed haplotypes in the population and their lengths. ShoRAH has an additional column where we specify the number of runs (with different parameters) and in brackets the number of runs that did not complete.

[a] Analysis did not complete. [b] The reference sequence was reconstructed.

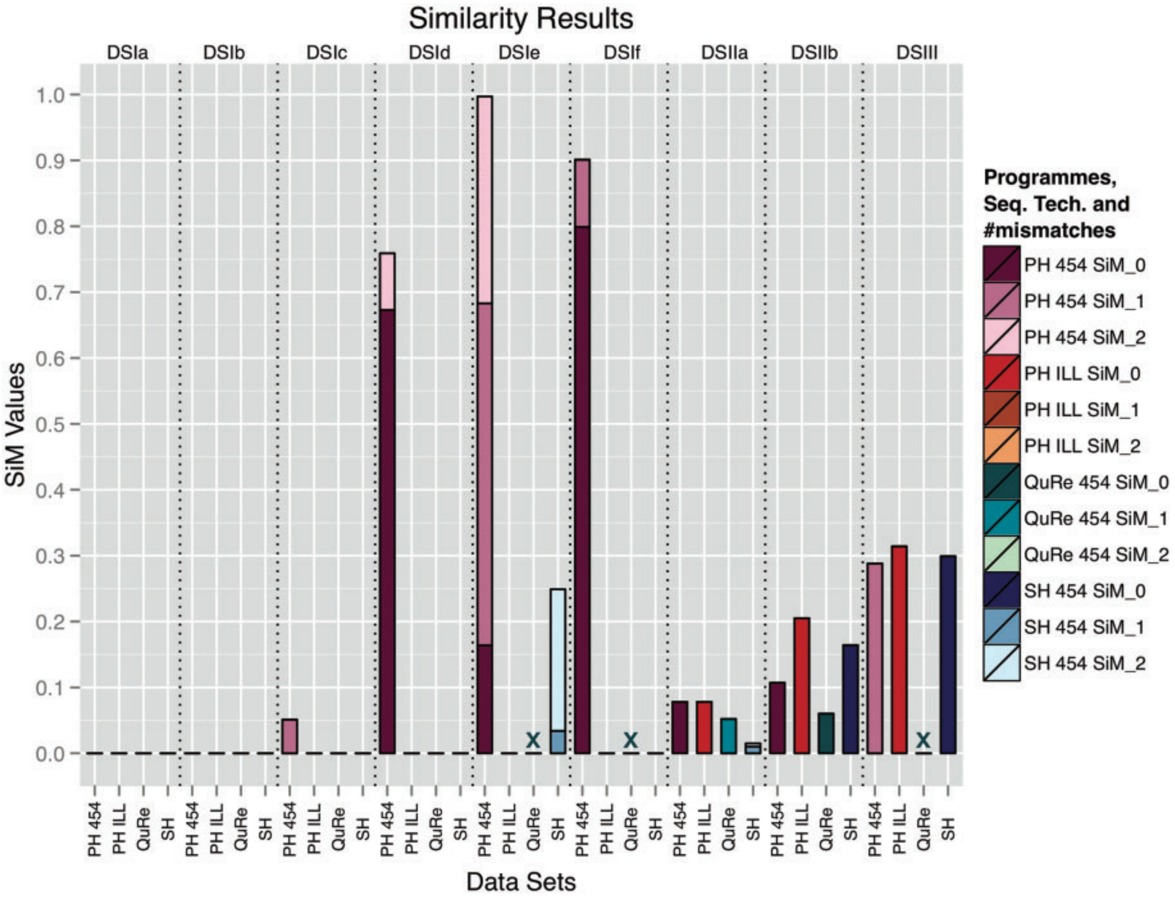

Figure 3: The plot displays the similarity results. The test data sets are indicated on the upper *x*-axis. For each data set, we have four bars and the respective programmes and sequencing technologies are indicated on the lower *x*-axis. The bars correspond to the results for PredictHaplo on 454 reads, PredictHaplo on Illumina reads, QuRe on 454 reads and ShoRAH on 454 reads. The different shades indicate how much of the population was reconstructed with zero, one or two mismatches in the sequence. The higher the value the better the reconstruction, where a perfect reconstruction corresponds to a value of one.

haplotypes show a Levenshtein distance that exceeds the mutual sequence divergence in the population. For the uniformly distributed data set DSIc (with 50 mutations on each haplotype), PredictHaplo reconstructed one haplotype with one mismatch and for the corresponding log-normally distributed data set DSId three haplotypes were reconstructed exactly and one haplotype with two mismatches (Table 4). But the reconstructed population also contains haplotypes with a Levenshtein distance of up to 34 where the mutual distance between the "true" haplotypes is on average 91 bp. None of the other programmes were able to reconstruct any haplotypes for these populations. For the much more diverse data sets DSIe and DSIf (≈ 4 %), PredictHaplo achieved good results: three haplotypes were exactly reconstructed for the uniformly distributed data set DSIe and six haplotypes were exactly reconstructed for the log-normally distributed data set DSIf. The rest of the reconstructed haplotypes have ≤ 3 mismatches. ShoRAH was able to reconstruct one haplotype with one mismatch and seven haplotypes with two mismatches, but the population also contains haplotypes with up to 209 mismatches (compared with their closest true haplotype). QuRe aborted calculations for DSIe and DSIf with an error message.

For data sets DSIIa and DSIIb, we concentrated on the reconstruction of a single gene. PredictHaplo was able to reconstruct one haplotype exactly for each of the two data sets. But in the case of DSIIb, we also find haplotypes with up to 422 mismatches in the reconstructed population. ShoRAH reconstructed two haplotypes with one mismatch and

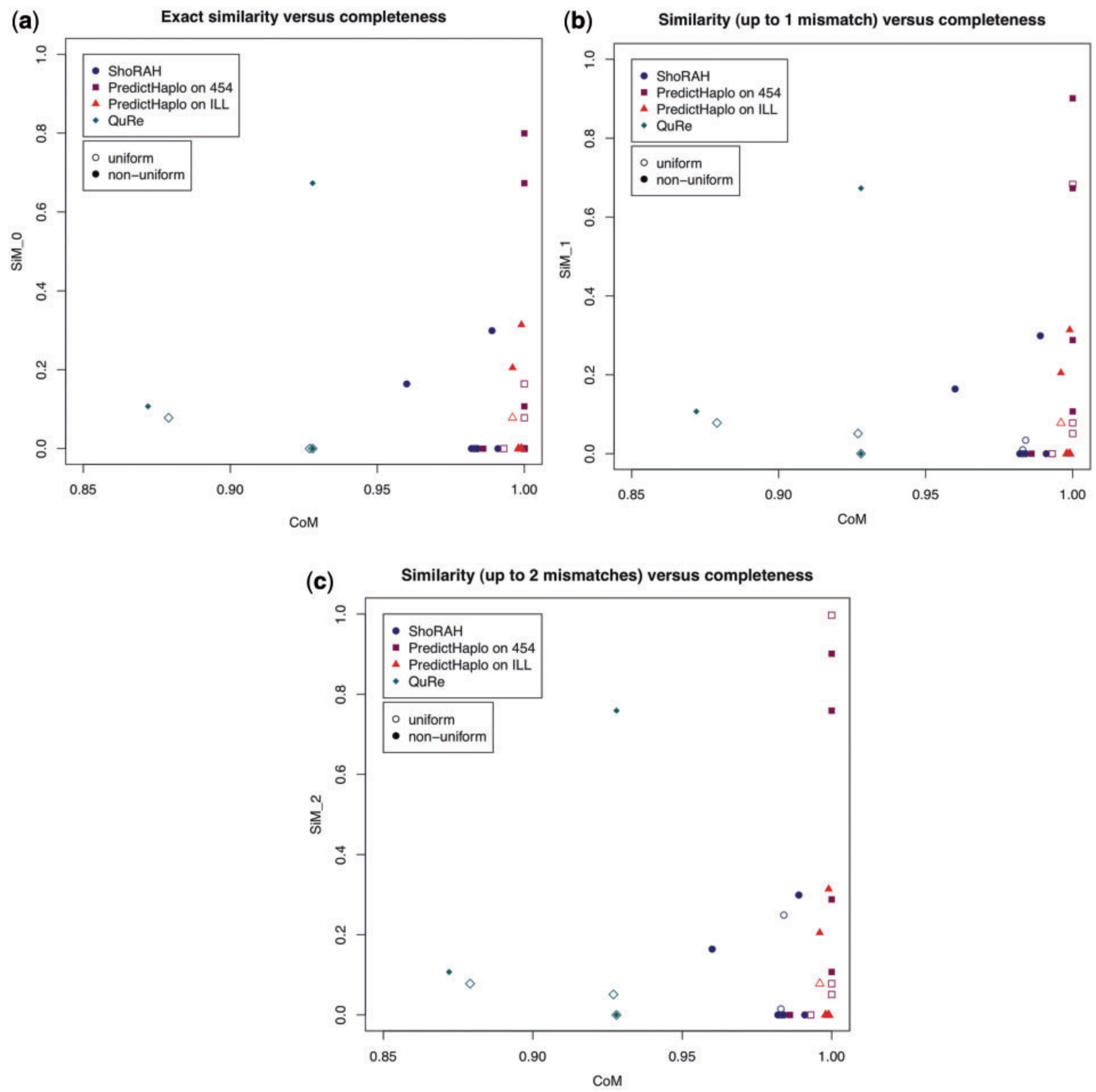

Figure 4: Similarity versus completeness for ShoRAH, PredictHaplo and QuRe: The similarity results with zero, one and two mismatches are displayed on the *y*-axis of the respective graph. The completeness (which is independent of the number of mismatches) is displayed on the *x*-axis. All completeness values were between 85% and 100%. (a) SiM_0 versus CoM, (b) SiM_I versus CoM and (c) SiM_2 versus CoM.

one haplotype with two mismatches. But the reconstructed population contains a total of 586 haplotypes compared with 44 ‘true’ haplotypes. In the case of DSIIb, ShoRAH reconstructed one of the haplotypes exactly. QuRe achieved its best result across all data sets for DSIIb where the dominant haplotype was reconstructed exactly. But the reconstructed population also contains haplotypes with up to 989 mismatches. For DSIIa, QuRe reconstructed one of the haplotypes with one mismatch.

DSIII was the most complex *in silico* population with a large number of haplotypes. ShoRAH was able to reconstruct the reference haplotype exactly but the rest of the population shows between 208 and 672 mismatches compared with their closest true haplotype. PredictHaplo found one ‘true’ haplotype with one mismatch and the Levenshtein distance for the rest of the population is between 195 and 1591. PredictHaplo seems to have problems in the presence of many low abundant haplotypes and seems to incorporate SNPs from many different haplotypes into very few reconstructed haplotypes. QuRe aborted the calculations for this data set with an error message.

The similarity measure for the Illumina read data sets

In the documentation of ShoRAH, the usage of Illumina reads is only described for the reconstruction over a local window. We tested ShoRAH on Illumina reads as a matter of completeness on data sets where ShoRAH has been able to reconstruct haplotypes with 454 reads. We did two runs with different parameters for each of those data sets. ShoRAH only produced results for one run on DSIIa where 102 haplotypes were reconstructed with 424–467 mismatches and one run for DSIIb where 987 haplotypes were reconstructed with 588–681 mismatches. We ran QuRe on all data sets with Illumina-specific error rates of 0.0012 [19]. Unlike 454, Illumina has the same error probability in homopolymeric and non-homopolymeric regions. QuRe aborted computations with an error message for all Illumina data sets. We ran PredictHaplo three times for all of the data sets with the same parameters as for the 454 data sets. PredictHaplo returned only one or two haplotypes in each case. For DSIa – DSIf, none of the 'true' haplotypes were found and the reconstructed haplotype(s) showed an increasing number of mismatches compared with their closest true haplotype as the sequence divergence in the data sets increased. It seems that PredictHaplo recognised the SNPs occurring on various haplotypes but was not able to assign them to different reconstructed haplotypes. As Illumina reads are very short compared to 454 reads, there seems to be insufficient information to link the SNPs to their respective haplotype. In the case of DSIIa and DSIIb, one haplotypes was successfully reconstructed. Only one haplotypes was returned for DSIIa, thus PredictHaplo assessed all the SNPs occurring on other haplotypes as errors. In the case of DSIII, the wildtype was successfully reconstructed. The results for PredictHaplo and ShoRAH are summarised in Table 5 and the PredictHaplo results are included in Figure 3.

Completeness of the reconstruction for the 454 and Illumina read data sets

In Figure 4, we plotted the similarity versus the completeness measure. The completeness measure tells us how much of the haplotype was reconstructed on average (in percent). For those parts of the haplotypes that were reconstructed, the similarity tells us about the quality; a value of zero means that none of the 'true' haplotypes was reconstructed and a value of one corresponds to a perfect reconstruction. Across all the data sets, the reconstructed haplotypes covered between 87% and 100% of the length of the 'true' haplotypes. PredictHaplo produced the best results in terms of completeness with values of $\geq$99%. ShoRAH produced very good results as well with values above 98% for all data sets besides DSIIb (96%). The QuRe results were $\approx$93% for DSIa, DSIb, DSIc and DSId and $\approx$87% for DSIIa and DSIIb.

Table 5: Illumina reads: the results for PredictHaplo and ShoRAH for the Illumina read data sets

		PredictHaplo			**ShoRAH**		
	Diversity	**Best Ldist**	**No. rec.hap**	**Length**	**Best Ldist**	**No. rec.hap**	**Length**
DSIa	19 bp	9, 134	2	8152	[a]	[a]	[a]
DSIb	19 bp	8	1	8152	[a]	[a]	[a]
DSIc	91 bp	45	1	8151	[a]	[a]	[a]
DSId	91 bp	45	1	8152	[a]	[a]	[a]
DSIe	325 bp	185	1	8149	[a]	[a]	[a]
DSIf	325 bp	184	1	8147	[a]	[a]	[a]
DSIIa	2–328 bp	0	1	2570	424–467	102	2540
DSIIb	2–328 bp	0, 57	2	2569	588–681	987	2580
DSIII	1–41 bp	0[b]	1	8152	[c]	[c]	[c]

Analogous to Table 4, the Levenshtein distance, the number of reconstructed haplotypes and the length of the reconstructed haplotypes are displayed. [a]We did not run any analyses as the reconstruction was not successful with the much longer 454 reads. [b]The reference sequence was reconstructed. [c]Analysis did not complete.

CONCLUSION

The programmes that we have tested were unable to cope with populations with low sequence divergences and low abundance levels. There are two problems. First, sequencing errors can be mistaken for sequence divergence. Second, if SNPs occur with a distance exceeding the read length we cannot infer from the reads if these SNPs occurred on the same haplotype. The reconstructed populations contained many false-positive SNPs and false-positive haplotypes and many true haplotypes were not recovered. This has serious consequences, for example in the development of drug therapies and vaccine design because the haplotypes that were not recovered cannot be targeted by the treatment. Haplotypes in a viral population can be resistant to a drug treatment due to mutations and, even if they occur only at low frequencies, their presence has been shown to correlate with treatment failure [20, 21]. Thus it is particularly important that reconstruction programmes are also able to recover low-frequency haplotypes in the population. We also need to bear in mind that for real data sets (in contrast to mock communities) we have no possibility to distinguish the successfully reconstructed haplotypes from the false positives. In many of the reconstructed populations, the number of false positives far exceeds the number of successfully reconstructed haplotypes and false positives often showed a substantial number of mismatches compared with the 'true' haplotypes. A high number of false positives needlessly complicate the development of an effective treatment.

In general, the read-graph approach seems computationally expensive and not the optimal approach for the whole gene/genome haplotype reconstruction; the inferred set of haplotypes was, in most cases, much larger than the actual 'true' population. The approach with a Dirichlet process mixture model (DPMM), where errors are not corrected but incorporated as prior information, is computationally less expensive and seems to be more powerful. ShoRAH has been previously tested on single gene reconstruction rather than whole genome reconstruction. On our data sets, ShoRAH achieved better results on data set DSIIb where the reconstruction concentrates on a single gene and haplotypes are distributed according to a log-normal distribution. For the 454 reads, ShoRAH outperformed PredictHaplo on this data set. But the DPMM approach of PredictHaplo seems to scale better to whole genome reconstruction than the read-graph approach. Also, as the read-graph approach reconstructs the haplotypes based on the overlap of the reads, the much shorter Illumina reads are not suitable for ShoRAH and QuRe.

None of the currently available programmes for whole gene/genome reconstruction are designed to detect recombination. The programme QuasiRecomb accounts for recombination events but the currently published version only attempts local reconstruction. We created an *in silico* data set of 10 haplotypes including two recombinants to test the ability of PredictHaplo, ShoRAH and QuRe to detect recombination events. The data set consists of the first eight haplotypes from data set DSIe (≈4% sequence divergence). We created two recombinants with breaking points at position 2685 and 4450, respectively, and added them to the population. The 10 haplotypes were mixed according to a uniform distribution and we simulated 120 000 454 reads with FlowSim. QuRe did not yield any results on this data set. For PredictHaplo, we ran three analyses (with parameters 2000, 3000 and 4000). On the best run, PredictHaplo returned all eight non-recombinants with 0–2 mismatches but was not able to identify any of the recombinants. For ShoRAH, we used the parameters of its three best runs for data set DSIe. The best reconstructed population consisted of 271 haplotypes with a Levenshtein distance between 1 and 157. The closest reconstruct haplotype to the first recombinant showed 58 mismatches and the closest haplotype to the second recombinant showed 4 mismatches. This shows that the approach that ShoRAH takes is in principle able to identify recombination events but the high number of mismatches and false positives remains a problem.

All the haplotype reconstruction programmes have been tested previously but mostly on data sets with a much larger sequence divergence. The highest sequence divergence in our data sets is ≈4%. For bacteria, the threshold to distinguish between different species is 3%. Though there is no similar method for comparing viruses (as there is no gene or region that all viruses have in common), we expect that a large fraction of the haplotypes in a viral population show very low sequence divergence. Thus, programmes need to be able to find haplotypes that only differ by a few nucleotides.

Further advances in sequencing technologies are just a matter of time. This will be accompanied by reduced error rates, increased read length and higher coverage. Those advances will greatly improve our ability to reconstruct viral haplotype population from a set of observed reads. However, reducing the rate of false positives in a reconstructed population remains an imperative if we are to obtain reliable results from the reconstruction that will facilitate the development of effective vaccines and treatments.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- None of the programmes were able to cope with low sequence divergences (≪4%) and did not identify a large fraction of the population.
- For populations with a high sequence divergence, reconstruction is feasible with 454 reads but not with Illumina reads.
- Many of the programmes reconstructed a high number of false positives, which can contain a large number of incorrect SNPs.
- On 'real' data sets it is not possible to distinguish the successfully reconstructed sequences from the false positives with a high number of incorrect SNPs.
- To evaluate how good the reconstruction is, we need to take the accuracy of the reconstructed population into account as well as the frequencies of the reconstructed haplotypes and the number of false positives.

ACKNOWLEDGEMENTS

We thank Daniel Haydon and Marco Morelli for providing the experimental foot-and-mouth virus data set.

FUNDING

M.S. is supported by Unilever R&D Port Sunlight, Bebington, United Kingdom. C.Q. is supported by an Engineering and Physical Sciences Research Council Career Acceleration Fellowship (EP/H003851/1).

References

1. Ojosnegros O, Beerenwinkel N. Models of RNA virus evolution and their roles in vaccine design. *Immunome Res* 2010;**6(Suppl 2)**:S5.
2. Bonhoeffer S, Nowak MA. Pre-existence and emergence of drug resistance in HIV–1 infection. *Proc R Soc Lond B Biol Sci* 1997;**264**(1382):631–7.
3. Glenn TC. Field guide to next-generation DNA sequencers. *Mol Ecol Resour* 2011;**11**(5):759–69.
4. Illumina. http://www.illumina.com/systems.ilmn (December 2012, date last accessed).
5. Yang Z, Goldman N, Friday A. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol Biol Evol* 1994;**11**(2):316–24.
6. Shimizu N, Okamoto T, Moriyama EN, *et al*. Patterns of nucleotide substitutions and implications for the immunological diversity of human immunodeficiency virus. *FEBS Lett* 1989;**250**(2):591–5.
7. Wright CF, Morelli MJ, Thebaud G, *et al*. Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. *J Virol* 2011;**85**(5):2266–75.
8. Lee HY, Giorgi EE, Keele BF, *et al*. Modeling sequence evolution in acute HIV-1 infection. *J Theor Biol* 2009; **261**(2):341–60.
9. Nowak MA. What is a quasispecies? *Trends Ecol Evol* 1992; **7**(4):118–21.
10. Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics* 2011;**12**(1):119.
11. Roth V. *PredictHaplo*. http://cs-wwwarchiv.cs.unibas.ch/personen/roth_volker/HivHaploTyper/index.html (December 2012, date last accessed).
12. Prabhakaran S, Rey M, Zagordi O, *et al*. HIV-haplotype inference using a constraint-based Dirichlet process mixture model. In: *Machine Learning in Computational Biology (MLCB) NIPS Workshop* 2010;1–4.
13. Prosperi MCF, Salemi M. QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. *Bioinformatics* 2012;**28**(1):132–3.
14. Astrovskaya I, Tork B, Mangul S, *et al*. Inferring viral quasispecies spectra from 454 pyrosequencing reads. *BMC Bioinformatics* 2011;**12(Suppl 6)**:S1.
15. Macalalad AR, Zody MC, Charlebois P, *et al*. Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data. *PLoS Comput Biol* 2012;**8**(3):e1002417.
16. Henn MR, Boutwell CL, Charlebois P, *et al*. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog* 2012;**8**(3):e1002529.
17. Zagordi O, Töpfer A, Prabhakaran S, *et al*. Probabilistic inference of viral quasispecies subject to recombination. In: RECOMB'12: Proceedings of the 16th Annual international conference on Research in Computational Molecular Biology. BerlinSpringer, 2012;342–54.
18. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. *Bioinformatics* 2012; **28**(4):593–4.
19. Archer J, Baillie G, Watson SJ, *et al*. Analysis of high-depth sequence data for studying viral diversity: a comparison of next generation sequencing platforms using Segminator II. *BMC Bioinformatics* 2012;**13**(1):47.
20. Simen BB, Simons JF, Hullsiek KH, *et al*. Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment—naive patients significantly impact treatment outcomes. *J Infect Dis* 2009;**199**(5):693–701.
21. Le T, Chiarella J, Simen BB, *et al*. Low-abundance HIV drug-resistant viral variants in treatment-experienced persons correlate with historical antiretroviral use. *PLoS ONE* 2009;**4**(6):e6079.
22. Eriksson N, Pachter L, Mitsuya Y, *et al*. Viral population estimation using pyrosequencing. *PLoS Comput Biol* 2008; **4**(5):e1000074.