7400–7409 *Nucleic Acids Research, 2010, Vol. 38, No. 21* *Published online 29 July 2010*
doi:10.1093/nar/gkq655

Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies

Osvaldo Zagordi[1,2,*], Rolf Klein[3], Martin Däumer[3] and Niko Beerenwinkel[1,2]

[1]Department of Biosystems Sciences and Engineering, ETH Zurich, Mattenstrasse 26, 4058 Basel, [2]SIB – Swiss Institute of Bioinformatics, Switzerland and [3]Institut für Immunologie und Genetik, Pfaffplatz 10, 67655 Kaiserslautern, Germany

Received April 29, 2010; Revised July 2, 2010; Accepted July 9, 2010

ABSTRACT

Next-generation sequencing technologies can be used to analyse genetically heterogeneous samples at unprecedented detail. The high coverage achievable with these methods enables the detection of many low-frequency variants. However, sequencing errors complicate the analysis of mixed populations and result in inflated estimates of genetic diversity. We developed a probabilistic Bayesian approach to minimize the effect of errors on the detection of minority variants. We applied it to pyrosequencing data obtained from a 1.5-kb-fragment of the HIV-1 *gag/pol* gene in two control and two clinical samples. The effect of PCR amplification was analysed. Error correction resulted in a two- and five-fold decrease of the pyrosequencing base substitution rate, from 0.05% to 0.03% and from 0.25% to 0.05% in the non-PCR and PCR-amplified samples, respectively. We were able to detect viral clones as rare as 0.1% with perfect sequence reconstruction. Probabilistic haplotype inference outperforms the counting-based calling method in both precision and recall. Genetic diversity observed within and between two clinical samples resulted in various patterns of phenotypic drug resistance and suggests a close epidemiological link. We conclude that pyrosequencing can be used to investigate genetically diverse samples with high accuracy if technical errors are properly treated.

INTRODUCTION

Recent technological advances have drastically decreased the time and the cost required to obtain DNA sequences (1). Several next-generation sequencing (NGS), or deep sequencing, platforms are available now that can read millions of base pairs in a more cost-effective and faster way than traditional Sanger sequencing (2). NGS is applied in *de novo* genome sequencing projects (3), as well as in targeted resequencing studies, for example, of tumours (4), in epigenetic studies (5) and in transcriptome analysis (6). In this article, we demonstrate how NGS can be used to detect low-frequency variants in genetically heterogeneous samples such as those obtained from HIV-infected patients.

In general, the population structure of infectious pathogens is highly relevant because genetic pathogen diversity is often associated with disease progression, poor prognosis and treatment failure. RNA viruses, such as HIV or influenza, are prominent examples. The low fidelity of their viral polymerases, which lack classic proofreading mechanisms, is responsible for the continuous production of mutated viral copies. Many mutations are maintained on clones in proportion related to their fitness. This mutant spectrum, referred to as a viral quasispecies (7), allows the virus to rapidly adapt to fluctuating environments (8,9).

Without additional experimental effort, such as individual cloning, Sanger capillary sequencing can only determine the consensus sequence of a mixed sample, and mutations can only be detected if their frequency exceeds a threshold of ~20%. If two or more loci display variation, information on whether and how often these variants occur on the same DNA molecule is lost. NGS can overcome these limitations by directly sequencing the mixed sample at high coverage. Every read obtained in this manner represents a contiguous fragment of DNA from a single molecule in the DNA library of the sample. Therefore, the set of reads provides a statistical sample of the DNA library and it can be used to make inference about the genetic structure of the population.

The potential for NGS to detect low-frequency variants has been noted early on in the context of viral infections. Pyrosequencing detected low-frequency mutations conferring resistance to antiviral drugs in hepatitis B virus (10,11) and in HIV (12–16), and found mixed infections with different influenza strains (17,18).

*To whom correspondence should be addressed. Tel: +41 61 387 3177; Fax:+41 61 387 3994; Email: osvaldo.zagordi@bsse.ethz.ch

Despite these early successful case studies, application of NGS to resolve the population structure of pathogens remains challenging, chiefly because of the sequencing error rate. A consequence of the high error rate is the risk of considering a technical error as a low-frequency variant. For example, with a uniform error rate of 0.25% per base pair and an average read length of 350 bp, ~58% of the reads obtained in an experiment are expected to be contaminated with at least one sequencing error. Thus, the genetic diversity of the sample would be vastly overestimated based on the raw data alone. To avoid these huge numbers of false positives, some *ad hoc* strategies have been proposed to discard reads that are much shorter than average, contain ambiguous characters, or induce frameshifts in the translation (19,20).

Statistically, measurement noise can be distinguished from real variation in a deep sequencing experiment, if variants that are sequenced more than once can be identified as groups (clusters) of reads that are more similar to each other than to reads in other groups (Supplementary Figure S1). Here, we use a Bayesian statistical modelling approach to read error correction that can be regarded as a probabilistic clustering method (21). The main advantages of our approach are: (i) that an independent estimate of the error rate is not necessary; (ii) that the number of haplotypes does not need to be specified in advance; and (iii) that we generate a full probabilistic clustering solution which contains all information about the uncertainty associated with it, including the estimated error rate, the estimated number of haplotypes, their DNA sequence composition and the assignment of reads to haplotypes. This algorithm, together with other tools, is implemented in the open source software package *ShoRAH* (www.cbg.ethz.ch/software/shorah).

We present a comprehensive analysis of this method and assess its performance in error correction, haplotype reconstruction and haplotype frequency estimation. We applied the *ShoRAH* algorithm to read data from four HIV samples obtained with the 454/Roche GS FLX Titanium platform (Table 1). Two samples define control experiments and consist of 10 different clonal isolates of the HIV *pol* gene, mixed in different proportions. To assess the effect of PCR amplification, another confounding factor, one of the control samples was PCR-amplified from an aliquot of the other before pyrosequencing. The remaining two samples were derived from HIV patients suspected to be part of the same infection chain.

Table 1. Next-generation sequencing experiments

Experiment	Sample	Coverage	Indels (%)	Mismatches (%)	ε (%)	nt hap
Control	Non-PCR	2110	1.0	0.05	0.35–0.60	10
Control	PCR	6030	1.0	0.25	0.40–0.45	10
Clinical	Patient 1	2100	0.44	–	1.1	13
Clinical	Patient 2	6240	0.42	–	0.8	15

Type of experiment, coverage, error rates, estimates for the parameter ϵ and number of haplotypes (nt hap) are reported for all experiments. ϵ represents the probability that a position in one of the reads in the multiple sequence alignment is wrong. The number of haplotypes in the clinical samples is the number of distinct haplotypes that were detected.

The control experiments allow for a hard assessment of the performance of the computational haplotype inference method, because the reads can be mapped directly to the original clones, which are the only true haplotypes in the mixtures. We demonstrate that haplotypes at frequencies as low as 0.1% can be detected reliably and that their estimated frequencies show high agreement with the expected frequencies. Haplotype reconstruction based on probabilistic clustering is shown to outperform, in both precision and recall, *ad hoc* methods based on a minimal number of required observations.

In the clinical samples obtained from two infected patients, we found several low-frequency mutations, which were invisible to Sanger sequencing. The pattern of observed mutations was used to analyse the level of resistance of individual variants to several protease inhibitors (PIs). Both viral populations display intra-host genetic heterogeneity that resulted in predicted phenotypic diversity of drug resistance. In particular, we found minority drug resistant variants that, as it has been repeatedly observed, can affect treatment outcome (11,14–16,22).

It seems obvious that high-coverage NGS can detect low-frequency variants in the pathogen populations. In practice, however, the feasibility of this approach depends critically on correcting sequencing errors. Probabilistic read clustering as implemented in *ShoRAH* provides a robust error correction method and it can be used to estimate the structure of pathogen populations in short genomic regions.

MATERIALS AND METHODS

Experiments

Sample preparation. For the generation of a defined clone mixture, amplicons of the partial *gag*/*pol*-gene were produced in the context of routine genotypic HIV drug resistance testing. Briefly, viral RNA was isolated from plasma of HIV infected patients using QIAamp viral RNA mini kit (Qiagen, Hilden, Germany) according to the manufacturers protocol. Reverse transcription and polymerase chain reaction were carried out using OneStep RT–PCR kit (Qiagen, Hilden, Germany) and primers

1RES, 5′-GAAGAAATGATGACAGCATGTCAGGG-3′ (nt 1819–1844 numbered according to HXB2 reference genome) and

2RES, 5′-TAATTTATCTACTTGTTCATTTCCTCCAAT-3′ (nt 4173–4202).

Nested PCR was carried out with HotStarTaq (Qiagen, Hilden, Germany) and the following inner primer pair:

RES3, 5′-AGACAGGCTAATTTTTTAGGGA-3′ (nt 2074–2095) and

RES4, 5′-ATGGYTCTTGATAAATTTGATATGTCC-3′ (nt 3559–3585).

The 1.5-kb PCR product was purified by using the QIAquick spin PCR purification kit (Qiagen, Hilden, Germany). Standard Sanger sequencing of the HIV-1 *pol* region was done by using ABI Prism 3730 capillary

sequencer (Applied Biosystems, Foster City, CA, USA). PCR-products from 10 different subtype-B clinical isolates were cloned into pCRII-TOPO (Invitrogen, Carlsbad, CA, USA). After control sequencing and propagation, the inserts were excised from the vector and restriction enzymes SpeI and NotI. To analyse the impact of the PCR on the error rate an aliquot containing 100 000 copies of the fragment was used as the template in a single round PCR reaction using primers RES3 and RES4 (PCR-amplified sample). Samples from Patients 1 and 2 were generated by PCR from clinical specimen as described above. All PCR product were purified using QIAamp PCR purification kit (Qiagen, Hilden, Germany) until further processing.

Massively parallel sequencing. For the library preparation all four samples were nebulized according to 454 shotgun protocol (Roche/454-Life sciences, Branford, CT, USA). Fragmented DNA was purified using AMPure SPRI beads (Agencourt) to remove fragments <400 bp. The purified fragmented DNA was further processed according to the 454 FLX Titanium Library construction kit and protocol (Roche/454-Life Sciences, Branford, CT, USA) to ligate multiple identifier (MID) adaptors specific to the Titanium sequencing chemistry. The resulting single-stranded DNA library was assessed for size distribution using the RNA 6000 Pico chip on the 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) and quantified using the Ribogreen RNA Quantitation Kit (Invitrogen) on a Fluorometer (FLUO-star Omega, BMG Labtech Offenburg, Germany). Emulsion PCR (emPCR) and titration by enrichment at 0.5, 1, 2 and 4 copies per bead (cpb) was carried out according to the 454 Titanium emPCR protocol to determine the optimal ratio of the library DNA fragments to emPCR beads. Large volume emulsions were set up at the optimal ratio of 2 DNA cpb. The sequencing run was carried out on a 2/4 picotiter plate using the Roche/454 Genome Sequencer FLX with the 454 Titanium update.

Data analysis

Sample selection. In the first experiment, we mixed 10 different 1.5-kbp HIV-1 clonal fragments of the *gag*/*pol* gene at different proportions between 0.3% and 30%. In the analysis, we focused on a 1245-bp long region including the complete viral protease and part of the reverse transcriptase from position 2253 to 3497. The clones were chosen such that their mutual distance be in a narrow range ($6.8 \pm 0.5\%$). Nevertheless, some regions are more conserved than others, so when the haplotypes are compared on localized windows, the diversity varies between 2% and 12%, see Supplementary Figures 4 and 5. The original and the PCR-amplified mixtures were ligated with MID barcodes and sequenced. Among the 63 019 reads obtained, the MIDs assigned 16 540 reads to Sample 1 (non-PCR-amplified) and 45 973 to Sample 2 (PCR-amplified). We decided not to assign the remaining reads because they presented two or more mismatches in the barcode, and we regarded this as an indication of a low-quality read, although we cannot exclude that the mismatches were due to mistakes in the production of MIDs. In the second experiment, plasma samples from two HIV infected patients were analysed. The same genomic region as in the control experiment was isolated, ligated to barcodes and sequenced. In this experiment, 48 285 reads were obtained, 12 626 assigned to Patient 1 and 35 088 to Patient 2.

Preprocessing. For all four samples, we removed low-quality reads defined as having a phred score below 10 at one or more sequence positions. This filtering step selected 10 907 and 26 814 reads for the non-PCR and PCR control samples, and 7985 and 23 284 reads for Patients 1 and 2, respectively. The average read length was 340 bp for the control samples and 380 bp for the clinical samples (Table 1). After the filtering step, we built a multiple sequence alignment (MSA) of the reads by padding the gaps obtained in pairwise alignments of the reads to the reference genome, because off-the-shelf MSA software is impractical in this setting (23). The MSA was subsequently split into regions of width approximately equal to the read length, and the correction algorithm was applied to each region.

Haplotype reconstruction and error correction. Every time a variation is observed in the reads, the possibility that this is a technical error rather than a true biological variation must be considered. Due to the high coverage of NGS, we exploited the power gained from multiple independent observations. We used clustering to group reads by their similarity and interpreted an optimal clustering as follows. Each cluster consists of exactly those reads that originate from the same haplotype. The cluster centre corresponds to the haplotype sequence and the cluster size is proportional to its frequency. With this interpretation, sequencing errors can be corrected by removing any variation within read clusters and haplotypes can be reconstructed as cluster centres.

The most difficult task in clustering is usually to find the number of clusters that best explain the data (24). In our context, given a set of reads, this number depends on the error rate, that should be measured in a different experiment (25). To circumvent this difficulty, we developed a probabilistic generative model for the sequencing of error-prone reads from a mixed sample. In our approach, the number of haplotypes and the technical error rate are not fixed *a priori*, rather they are parameters to be estimated together with the biological diversity. The statistical model used to explore different numbers of clusters employs a non-parametric prior distribution called Dirichlet process mixture (26). In a Bayesian fashion, we compute the posterior distribution of the assignment of reads to haplotypes, the identity of these haplotypes, and the parameters controlling the error rate and the biological diversity, given the observed reads. Due to the complicated form of this probability distribution, these estimates cannot be derived analytically. We used Gibbs sampling, a Markov chain Monte Carlo (MCMC) algorithm, to sample from the posterior distribution of these quantities.

This model, together with other tools, is implemented in the software *ShoRAH* (21). A detailed mathematical description of the method and an assessment of its performance on simulated data has been presented elsewhere (26).

Confidence levels. The probabilistic nature of the clustering algorithm allows for estimating the reliability of predictions. The sampling method we devised explores different configurations of the model parameters, including the haplotype sequences. The fraction of iterations a haplotype is reported estimates the posterior probability of the existence of that haplotype. This posterior provides a confidence level for the haplotype, and in general, we report only those haplotypes with confidence value (posterior probability) greater than 0.9.

Control experiments. For the control experiments, we aligned the reads to all original sequences and assigned them to the best matched if there was no ambiguity in the assignment (the difference between the identities with the best match and the second best match had to be $>1.25\%$). We could then estimate the frequencies of the haplotypes, as the fraction of reads assigned to each one, and the sequencing errors from the number of discrepant bases between each read and their closest haplotype. The haplotype frequencies of the two control samples with and without PCR amplification are reported in Supplementary Figure 2. Some deviations are observed, especially for low frequencies, which indicate a selective amplification bias during PCR. Thus, PCR amplification is critical to estimating haplotype frequencies in mixed samples. Haplotype frequencies are also affected by stochastic sampling effects, if the coverage is non-uniform and frequencies are estimated at a local level. Therefore, we locally re-estimated the expected frequencies by re-aligning the read segments in every window of the MSA to the original clone segments of the respective genomic region. These estimates were considered the ground truth and compared to the output of the *ShoRAH* algorithm.

In the precision–recall analysis, we compared the performance of the cut-off method (in which a minimum number of reads must support a haplotype to call it true) with the probabilistic clustering results. In the cut-off method, the minimum threshold is varied between one and the maximum number of reads in order to draw the precision–recall graph. Similarly, a threshold on the confidence value defines the positive and negative haplotypes reconstructed by the clustering algorithm. The inferred haplotypes were aligned to all sequences, and they were considered correctly inferred if they matched one of the sequences with at most one mismatch. Precision was defined as the ratio of the number of matched sequences (out of the 10 present) over the total number of reconstructed haplotypes. Recall was the ratio of the number of matched sequences over the total number of real sequences (10).

Clinical samples. Phenotypic drug resistance for the reconstructed haplotypes has been predicted with a support vector machine model with linear kernel trained on the data in the Stanford HIV Drug Resistance Database (27) as described in (28) and (29). The coordinates of the points on the plot representing the inferred haplotypes have been determined with a multi-dimensional-scaling technique such that their distance reflects the distance between amino acid sequences.

Software and computational details. Sequence manipulation was performed using Biopython (30) and EMBOSS (31). Prediction of the phenotypic drug resistance and multi-dimensional scaling for the clinical samples were computed using the statistical language R (32).

RESULTS

We used deep sequencing to infer low-frequency variants in a mixed population with high precision, despite the error rate that is typically associated with this technology. To achieve this result, we used *ShoRAH*, a Bayesian probabilistic clustering method to identify biological variation and to filter out technical errors. The output of the algorithm is a list of haplotypes with a confidence value and an estimate of its frequency (Supplementary Figure S1). In Figure 1, one example of the distribution of these posteriors is shown for a region of the control sample with 10 haplotypes. We observe that about 10 haplotypes have a high posterior probability, while for the 5 additional haplotypes the confidence values quickly drop to negligible levels. For the haplotypes reported we also estimated the frequencies. In particular, low-frequency variants can have high confidence levels, emphasising the benefit of a full probabilistic model over cut-off-based haplotype calling.

PCR errors

Error rate of pyrosequencing. We estimated the error rate of the NGS procedure from the number of discrepant bases in the alignment of each read with its best match among the original clones (see 'Materials and Methods' section). Excluding indels, the base substitution error rate for the non-PCR sample was 0.05%, while for the PCR-amplified sample it was 0.25%. The PCR-associated increase of the error rate was highly significant ($P<10^{-6}$, Wilcoxon rank-sum test).

Recombination. PCR can not only introduce base substitution errors, but when used to amplify a mixture of heterogeneous templates, it may also produce cross-overs among templates (33). To estimate the amount of this artificial recombination, we analysed the read data from the two control samples with the software *Recco* (34). For each read, we determined whether it is better explained as resulting from mutations and indels in a single haplotype, or from recombination between two haplotypes and fewer mutations and indels.

The amount of mutations that can be saved by invoking recombination to explain the observed read is a measure of the likelihood of a recombination event. The result of this analysis is reported in Supplementary Figure S3. While for the non-PCR-amplified sample, the number of

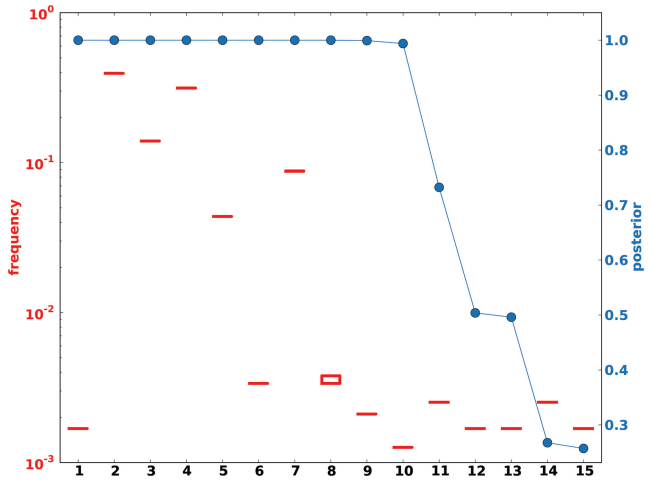

Figure 1. Posterior probability of reconstructed haplotypes. The algorithm computes posterior probabilities for the inferred haplotypes and their frequency given the observed reads. The figure shows (for window 3 in the PCR-amplified control experiment) the posterior of the haplotype frequencies in a box-plot (red box-plots) and the posterior of the reconstructed haplotype sequences (blue circles). In most cases, the box-plot height (lower-upper quartile) is invisible on this scale, because the clustering assignment is stable and the number of reads assigned to the cluster in the sampling does not change. The posterior distribution for the frequency is then very peaked. The figure reports estimates for the haplotypes reported by the algorithm without further processing. With additional analysis one finds that haplotypes 12 and 13 differ by one gap only in a homopolymeric region, and that their posterior probabilities sum up to one. Moreover, haplotype 1 consists of four reads and is the result of a recombination event between two of the original haplotypes.

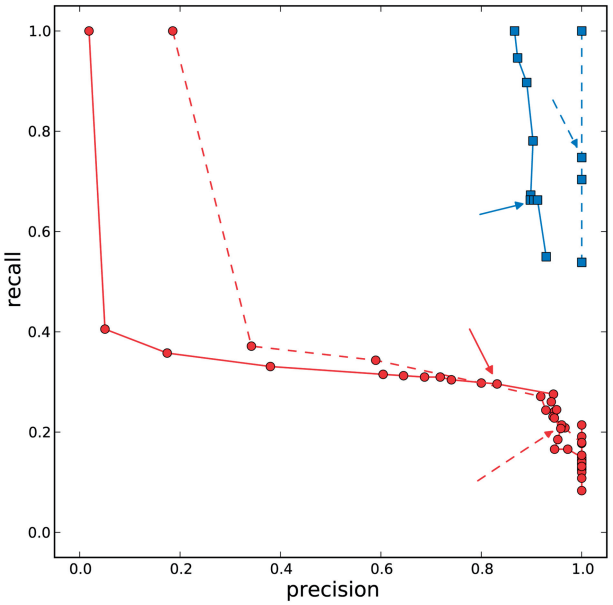

Figure 2. Precision–recall analysis. We considered the haplotypes inferred in all windows by the clustering algorithm and by the cut-off method based on the minimum number of reads supporting the variant. Red circles represent precision and recall for a set of threshold values chosen in the cut-off method (values from 1 to the number of reads in the most-supported haplotype), red arrows annotate points for a cut-off equal to 50. We performed a similar analysis on the output of the clustering algorithm, considering haplotypes whose confidence value (posterior probability) was greater or equal than a given threshold. Blue squares represent threshold values from 0.01 to 1, with blue arrows annotating the values obtained when the threshold is 0.9. Dashed lines and arrows are used for points obtained in the non-PCR-ampified sample, solid lines and arrows for points in the PCR-amplified one. In the non-PCR-amplified sample, we have a perfect precision (no false positives), and very good results for the recall. In the PCR-amplified sample, some false positives are found. In both cases, the performance of the clustering method is superior to the cut-off method. Results for individual windows can be found in Supplementary Data.

mutations that could be saved is always lower than five, in the PCR-amplified sample, up to 16 possible savings are observed. In 1.9% of reads from the PCR-amplified sample, we observed more than five savings, indicating PCR-associated cross-overs.

Error correction and haplotype reconstruction

Detection of haplotypes. The high coverage of NGS promises increased sensitivity to detect haplotypes in a mixture, but PCR and sequencing errors can induce large false positive rates without careful analysis of the read data. Using the clonal mixture samples, we compared the ability to detect the 10 real haplotypes between our probabilistic clustering method and the baseline cut-off method. A low cut-off results in the detection of many haplotypes at the cost of many false positives, while a large cut-off avoids false alarms at the cost of missing true variants. For example, requiring only at least two observations of each haplotype (i.e. cut-off 2), we found between 7 and 37 haplotypes in the non-PCR sample and between 80 and 235 haplotypes in the PCR-amplified sample. For a cut-off of 20 observations, between 0 and 7 and between 5 and 16 haplotypes were called for the non-PCR and PCR sample, respectively. Some manufacturers recommend a cut-off of 50 observations (35). For this cut-off, between 0 and 5 haplotypes were found in the non-PCR-amplified sample and between 3 and 10 were found in the PCR-amplified one.

The choice of an optimal cut-off does not only depend on the trade-off between precision and recall, but also on specific parameters such as coverage, error rate and genetic diversity of the sample, some of which are unknown in any real application. The Bayesian clustering approach avoids this problem and instead estimates these parameters directly from the data and reports the confidence level for each reported haplotype.

We assessed haplotype reconstruction performance in a precision–recall analysis over the entire range of possible cut-offs and posterior probabilities. Precision is the fraction of true haplotypes among all called haplotypes and recall is the fraction of called haplotypes among all true haplotypes. Figure 2 shows the aggregate result of this analysis for all windows of the MSA. The cut-off method performs poorly for all choices of the cut-off value and can achieve high recall ($\geq$80%) only at the expense of very low precision ($\leq$20%) and vice versa. For example, a cut-off of 50 observations, results in 80% precision, but only <40% recall (Figure 2, red arrows). The probabilistic clustering method outperforms the cut-off method for any choice of the haplotype

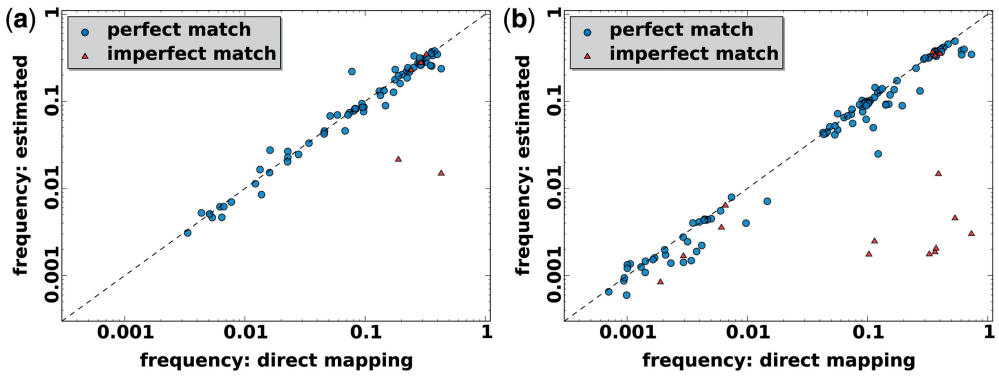

Figure 3. Frequency estimation with the clustering method. In each window true frequency of the haplotypes was estimated by aligning the raw reads to the original sequences (direct mapping) for non-PCR-amplified sample (**a**) and PCR-amplified sample (**b**). Then, haplotypes were reconstructed and it was checked whether they matched the originals in identity and frequency. Circles represent perfect matches with one of the original haplotypes, triangles indicate imperfect match. Except for a few spurious cases at low frequencies, there is good agreement both in identity and frequency between inferred and actual haplotypes.

confidence level. Overall, the performance is less sensitive to this value and many choices result in haplotype reconstruction with both precision and recall $\geq 80\%$. In general, this method is conservative in the sense that for all confidence thresholds a called haplotype is likely to be a true haplotype, especially for thresholds of 0.9 and higher. For both methods, performance is slightly better on the non-PCR sample than on the PCR sample.

These results are confirmed in each individual window of the MSA (Supplementary Figures 7 and 8). The local analysis also highlights the fact that for the cut-off method, there is no single optimal cut-off value. In contrast, the posterior haplotype probability is interpretable across experiments and the default choice of 0.9, which we fixed for all further analyses, results in good performance throughout (Figure 2, blue arrows).

Error correction. The non-PCR-amplified sample showed a per-base error rate of 0.05%, while the amplified sample had a higher error rate of 0.25%. As a consequence, ~11% of the reads in the non-PCR sample and 43% in the PCR sample have one or more sequencing errors. After the error correction procedure, the error rate drops to 0.03% and 0.05% for the non-PCR and PCR sample, respectively, which results in ~92% of the reads being error free in both samples.

Frequency estimation. We assessed the ability of the *ShoRAH* algorithm to estimate the frequency of individual clones in the population. For the non-PCR-amplified sample, almost all haplotypes were reconstructed correctly and their frequency estimates were highly correlated with the ground truth (Pearson's correlation coefficient $r = 0.88$ for all haplotypes and $r = 0.96$ if outliers, i.e. haplotypes with ≥ 1 mismatches, are excluded; Figure 3a). Both haplotype reconstruction and frequency estimation are more difficult for the PCR-amplified sample, as shown by an increased number of imperfect haplotype matches and more discrepant frequency estimates ($r = 0.78$ for all and $r = 0.95$ for all perfect matching haplotypes; Figure 3b).

Perfect reconstruction and frequency estimation was possible for many haplotypes with frequencies as low as 1% for the non-PCR sample and 0.1% for the PCR sample. This difference in resolution can be explained by the different average coverage of ~2100 and 6000 base pairs per sequence position for the non-PCR and PCR sample, respectively. In each window, the sum of the frequencies of all true haplotypes that we were able to detect was always >96.5% for the non-PCR sample and >99.0% for the PCR sample.

Clinical samples

We applied error correction, haplotype reconstruction and frequency estimation to NGS data obtained from two clinical samples derived from two HIV-infected patients suspected to be in the same infection chain. The *ShoRAH* algorithm was run on the aligned reads and the output was used to make inference about the genetic diversity of the two viral quasispecies and about the distribution of drug resistance among individual clones of the populations. We focused on the 99 codons of the HIV-1 protease and considered all reads that covered this 297 bp region completely: 891 reads for Patient 1 and 3272 reads for Patient 2.

In Figure 4, the allele frequency spectrum is shown after translation into amino acids. The consensus sequences for Patients 1 and 2 share the following mutations relative to the HXB2 reference strain: V3I, T12S, L19I, S37N, I54V, D60E, L63P, V77I and I93L, some of which are associated with resistance to PIs (16). They differ only at Position 10, where for Patient 1 an isoleucine is observed and a valine for Patient 2. However, we identified several additional mutations at low frequencies and the pattern of this variation was very different between the two patients. Sanger sequencing would not be able to detect this variation, because the frequency of most mutations is far below it detection limit of ~20% (Figure 4).

The allele frequency spectrum provides only a summary of the underlying population structure that ignores covariation, or phasing, of mutations at different sites of the genome. The real strength of NGS is to locally resolve

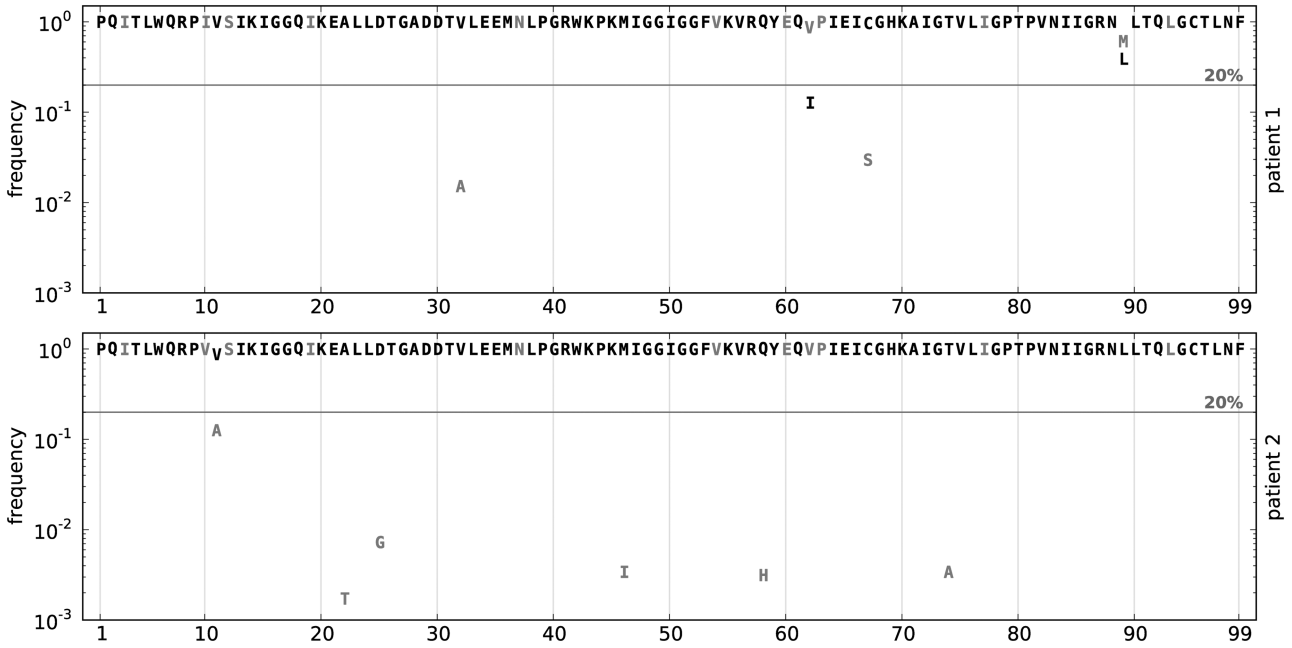

Figure 4. HIV protease amino acid allele frequency spectra of two patient samples. We analysed the frequency of amino acid substitution in the protease for two patients suspected to be in the same infection chain. They both present the drug resistance mutation I54V. Their consensus sequences differ at Position 10, Patient 1 showing an isoleucine and Patient 2 a valine. The horizontal line shows the 20% threshold typical of Sanger sequencing.

the population structure. On the DNA level, we identified a total of 13 different haplotypes for Patient 1, and 15 different haplotypes for Patient 2, all with posterior probabilities close to one, which gave rise to five and six different protein sequences for Patients 1 and 2, respectively, (Supplementary Table S1). Using the cut-off method and a threshold of 50 observations to support a variant, we would call only 2 haplotypes for Patient 1 and 12 for Patient 2.

To compare the diversity of the two protease quasispecies, we employed two statistical tests on the estimated virus populations, i.e. on the inferred haplotype sequences and frequencies. A general non-parametric procedure for comparing vectors with categorical components was used to detect differences between the two haplotype distributions (36). The permutation test indicated that the observed difference between the quasispecies could not be explained by the sampling variance alone ($P < 5 \times 10^{-4}$) and hence is more likely to be the result of evolutionary change. We also quantified the amount of genetic diversity found in either population using the Simpson's index (37). The Simpson's, index is defined as $D = \sum_{i=1}^{n} p_i^2$, where p_i is the frequency of the i-th haplotype. It is the probability that two sequences that are randomly drawn from the population are identical. Thus, a low Simpson's, index indicates a high level of diversity. We found significantly higher viral genetic diversity in Patient 1 ($D = 0.115 \pm 0.005$) than in Patient 2 ($D = 0.138 \pm 0.004$).

The reconstructed virus populations show a diverse pattern of drug resistance mutations. PI resistance can be caused by many different protease mutations and usually it is mediated by combinations of these (27). To account for the effect of mutational patterns, we

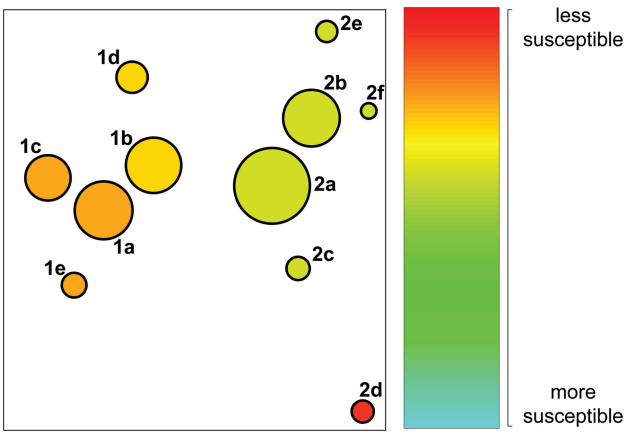

Figure 5. Structure of the viral quasispecies and predicted resistance to lopinavir. Circles represent detected haplotypes translated into amino acid sequences. The size reflects the frequency of the amino acid sequences, while the fill colour indicates the predicted resistance to the PI lopinavir. Green indicates higher and red lower levels of predicted drug susceptibility. The circles are positioned in the plot such that their distance approximately preserves the Hamming distance of the amino acid sequences. The number and the letter next to each circle denote, respectively, the patient and the protease sequence reported in the Supplementary Data.

predicted, for each inferred haplotype, the level of phenotypic drug resistance to the PIs atazanavir, lopinavir, amprenavir, saquinavir and tipranavir (see 'Materials and Methods' section). The results are reported in Supplementary Table S1 and visualized for lopinavir in Figure 5 and for all five drugs in Supplementary Figure S9. Haplotype sequences are shown as discs in the plane in such a way that their pairwise distances

reflect the Hamming distances in sequence space. The size of the discs is proportional to the haplotype frequency and the colour indicates the level of resistance. This analysis revealed variation in the predicted levels of PI resistance not only between the two patient samples, but also within individual quasispecies. For example, the viral quasispecies of Patient 1 displays differential levels of resistance to four of the five PIs, and the virus population of Patient 2 contains a low-frequency clone with reduced susceptibility to lopinavir and tipranavir. The latter haplotype would almost certainly go undetected with Sanger sequencing.

DISCUSSION

We have shown how NGS can detect low-frequency variants in a mixed sample, provided that the technical sequencing errors are properly treated. Without this step, technical artefacts can easily be mistaken for real variants, and the amount of variation detected would necessarily overestimate the real one. In the case of HIV infection, low-frequency drug resistance mutations have been correlated with treatment failure (14,15). Thus, reliable estimates of these minority variants are crucial in effectively tailoring anti-viral therapy.

Different error sources can affect the reliability of these estimates. Reverse transcription and PCR amplification can introduce mismatches in the DNA sequences. Recombination between different templates in PCR can introduce haplotypes not existing in the original sample, and haplotypes can be amplified with different efficiency in the PCR reaction, altering the frequency spectrum of the original mixture. Finally, the sequencing process is far from being error-free.

We designed control experiments in order to investigate several error sources in a quantitative manner. In particular, we assessed two types of errors introduced by PCR amplification: recombination and amplification bias. We found that a fraction of $<2\%$ of the reads most likely originated from cross-over events during PCR amplification. Selective amplification (a change in the spectrum of frequencies after PCR) was also observed (see Control experiments in 'Materials and Methods' section). We conclude that PCR reactions must be designed and performed with great care in order to avoid artificial recombinants and to decrease the amount of bias in the investigated sample (33).

However, in our analysis we could separate these error sources from those due to pyrosequencing, and we could assess the performance of the *ShoRAH* algorithm used to correct sequencing errors. We obtained up to a five-fold decrease in the error rate, from 0.25% to 0.05% per base.

The cut-off method often used to detect minority variants requires previous knowledge of the sequencing error rate in order to set an optimal threshold for the minimum number of reads to call a variant. In contrast, the Bayesian method employed here does not require the error rate as input, but estimates it from the data. Using data from the control experiments we showed that, in fact, any choice of such a cut-off value results in poor haplotype reconstructions suffering from either low precision or low recall, and often both.

We stress that, unlike conventional Sanger sequencing, deep sequencing can detect the co-occurrence of mutations at least at a distance of the read length. Full HIV protease haplotypes were reconstructed from samples obtained from infected patients and showed how in a single host a diverse pattern of drug resistance can be observed. In HIV as well as in other viral infections, it will be of interest to investigate whether therapeutic interventions will benefit from the higher level of detail at which the viral population can be studied.

The data present in drug resistance databases, such as those used for the predictions shown in Figure 5, report phenotypic resistance to antiviral drugs of the entire population, but the matched DNA sequences are obtained by traditional Sanger sequencing. These data cannot inform about the contribution of individual clones to the level of resistance of the whole population. Therefore, the predictions shown in Figure 5 should be interpreted as the level of phenotypic drug resistance that can be expected in case the respective clone becomes dominant in the population. From a therapeutic point of view, the structure of the whole population is likely to have prognostic value, because pre-existing resistant minority clones will be selected rapidly as soon as the corresponding selective drug pressure is applied. In the future, models for predicting phenotypic drug resistance might include information on the viral population structure rather than the consensus sequence alone, eventually leading to a better understanding of resistance mechanisms.

Reliable reconstruction of the virus population is a prerequisite for making inference about the population structure, and it can serve as a starting point for epidemiological and clinical investigations. For example, the pattern of diversity observed in Figure 5 shows that the intra-patient diversity is larger than the inter-patient diversity, suggesting a close epidemiological link. Although it is in general difficult to establish the occurrence and the direction of transmission events with certainty, reliable high-resolution estimates of the population structure can be used in phylogenetic analyses to support or reject transmission hypotheses (38).

The knowledge of the population structure at the local level (on the order of the read length) can be used to reconstruct the haplotypes at a global level. Combinatorial approaches that assemble the reads after error correction into longer haplotypes have been proposed (25). Together with information obtained from paired-end reads, local structure estimates impose constraints on the global structure of haplotypes. We envisage extending our model of read generation from diverse haplotypes to the reconstruction of larger regions, and even the entire viral genome in the future.

An important step will be extending the *ShoRAH* algorithm to read data obtained from other NGS technologies, including the Illumina and SOLiD platforms, and highlighting the advantages and disadvantages of each in this specific application. Moreover, we foresee a better usage of base-specific quality scores, which we used here only to discard low-quality reads. The simple error model used in

our approach does not distinguish between different substitutions as it considers the probability of a technical error independent of both the real and the erroneous base. It has been observed, for example for the Illumina platform (39), that some substitutions occur more frequently than others, and a model with additional parameters might take this into account. Such a model might be more accurate, but would not be efficient in correcting technical insertions and deletions. For this reason, we believe that another important contribution might be to perform alignment and error correction within the same probabilistic framework. A promising effort in this direction, though directed to metagenomics, has been recently presented (40).

Genetic diversity is important not only in retrovirus infections, but also in cancer and bacterial communities (41). We imagine the application of error correction to NGS data to all cases where sequencing is targeted towards a genetically heterogeneous sample and the goal is to estimate the structure of the population. Understanding the relation between phenotypic features of a pathogen population and its genetic structure will help in understanding the pathogenicity of parasite populations and possibly lead to the discovery of novel therapeutic options.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Patrick Knupfer is kindly acknowledged for calculating the predicted drug resistance to antiviral drugs of the clinical samples.

FUNDING

Swiss National Science Foundation (grant number CR32I2_127017) (in part).

Conflict of interest statement. None declared.

REFERENCES

1. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.-J., Chen,Z. *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
2. Metzker,M.L. (2010) Sequencing technologies—the next generation. *Nat. Rev. Genet*., **11**, 31–46.
3. Miller,W., Drautz,D.I., Ratan,A., Pusey,B., Qi,J., Lesk,A.M., Tomsho,L.P., Packard,M.D., Zhao,F., Sher,A. *et al*. (2008) Sequencing the nuclear genome of the extinct woolly mammoth. *Nature*, **456**, 387–390.
4. Shah,S.P., Morin,R.D., Khattra,J., Prentice,L., Pugh,T., Burleigh,A., Delaney,A., Gelmon,K., Guliany,R., Senz,J. *et al*. (2009) Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, **461**, 809–813.
5. Varley,K.E., Mutch,D.G., Edmonston,T.B., Goodfellow,P.J. and Mitra,R.D. (2009) Intra-tumor heterogeneity of MLH1 promoter methylation revealed by deep single molecule bisulfite sequencing. *Nucleic Acids Res*., **37**, 4603–4612.
6. Albrecht,M., Sharma,C., Reinhardt,R., Vogel,J. and Rudel,T. (2010) Deep sequencing-based discovery of the Chlamydia trachomatis transcriptome. *Nucleic Acids Res*., **38**, 868.
7. Domingo,E., Martin,V., Perales,C., Grande-Pérez,A., García-Arriaza,J. and Arias,A. (2006) Viruses as quasispecies: biological implications. *Curr. Top. Microbiol. Immunol*., **299**, 51–82.
8. Vignuzzi,M., Stone,J., Arnold,J., Cameron,C. and Andino,R. (2006) Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature*, **439**, 344–348.
9. Korber,B., Gaschen,B., Yusim,K., Thakallapally,R., Kesmir,C. and Detours,V. (2001) Evolutionary and immunological implications of contemporary HIV-1 variation. *Bri. Med. Bull*., **58**, 19–42.
10. Solmone,M., Vincenti,D., Prosperi,M.C.F., Bruselles,A., Ippolito,G. and Capobianchi,M.R. (2009) Use of massively parallel ultradeep pyrosequencing to characterize the genetic diversity of hepatitis B virus in drug-resistant and drug-naive patients and to detect minor variants in reverse transcriptase and hepatitis B S antigen. *J. Virol*., **83**, 1718–1726.
11. Margeridon-Thermet,S., Shulman,N.S., Ahmed,A., Shahriar,R., Liu,T., Wang,C., Holmes,S.P., Babrzadeh,F., Gharizadeh,B., Hanczaruk,B. *et al*. (2009) Ultra-deep pyrosequencing of hepatitis B virus quasispecies from nucleoside and nucleotide reverse-transcriptase inhibitor NRTI-treated patients and NRTI-naive patients. *J. Infect. Dis*, **199**, 1275–1285.
12. Wang,C., Mitsuya,Y., Gharizadeh,B., Ronaghi,M. and Shafer,R.W. (2007) Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res*., **17**, 1195–1201.
13. Hoffmann,C., Minkah,N., Leipzig,J., Wang,G., Arens,M.Q., Tebas,P. and Bushman,F.D. (2007) DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Res*., **35**, e91.
14. Simen,B., Simons,J., Hullsiek,K., Novak,R., Macarthur,R., Baxter,J., Huang,C., Lubeski,C., Turenchalk,G., Braverman,M. *et al*. (2009) Low-Abundance Drug-Resistant Viral Variants in Chronically HIV-Infected, Antiretroviral Treatment-Naive Patients Significantly Impact Treatment Outcomes. *J. Infect. Dis*., **199**, 693–701.
15. Le,T., Chiarella,J., Simen,B.B., Hanczaruk,B., Egholm,M., Landry,M.L., Dieckhaus,K., Rosen,M.I. and Kozal,M.J. (2009) Low-abundance HIV drug-resistant viral variants in treatment-experienced persons correlate with historical antiretroviral use. *PLoS ONE*, **4**, e6079.
16. Johnson,J.A., Li,J.-F., Wei,X., Lipscomb,J., Irlbeck,D., Craig,C., Smith,A., Bennett,D.E., Monsour,M., Sandstrom,P. *et al*. (2008) Minority HIV-1 drug resistance mutations are present in antiretroviral treatment-naïve populations and associate with reduced treatment efficacy. *PLoS Med*., **5**, e158.
17. Ghedin,E., Fitch,A., Boyne,A., Griesemer,S., Depasse,J., Bera,J., Zhang,X., Halpin,R.A., Smit,M., Jennings,L. *et al*. (2009) Mixed infection and the genesis of influenza virus diversity. *J. Virol*., **83**, 8832–8841.
18. Ramakrishnan,M.A., Tu,Z.J., Singh,S., Chockalingam,A.K., Gramer,M.R., Wang,P., Goyal,S.M., Yang,M., Halvorson,D.A. and Sreevatsan,S. (2009) The feasibility of using high resolution genome sequencing of influenza a viruses to detect mixed infections and quasispecies. *PLoS ONE*, **4**, e7105.
19. Huse,S.M., Huber,J.A., Morrison,H.G., Sogin,M.L. and Welch,D.M. (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol*., **8**, R143.
20. Tsibris,A.M.N., Korber,B., Arnaout,R., Russ,C., Lo,C.-C., Leitner,T., Gaschen,B., Theiler,J., Paredes,R., Su,Z. *et al*. (2009) Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo. *PLoS ONE*, **4**, e5683.
21. Zagordi,O., Geyrhofer,L., Roth,V. and Beerenwinkel,N. (2009) Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. In *Proceedings of the 13th Annual International Conference on Research in Computational Molecular Biology*. Springer, Berlin, p. 284.

22. Metzner,K.J., Bonhoeffer,S., Fischer,M., Karanicolas,R., Allers,K., Joos,B., Weber,R., Hirschel,B., Kostrikis,L.G., Günthard,H.F. *et al*. (2003) Emergence of minor populations of human immunodeficiency virus type 1 carrying the M184V and L90M mutations in subjects undergoing structured treatment interruptions. *J. Infect. Dis.*, **188**, 1433–1443.
23. Saeed,F., Khokhar,A., Zagordi,O. and Beerenwinkel,N. (2009) Multiple sequence alignment system for pyrosequencing reads. *LNBI*, **5462**, 362–375.
24. Hastie,T., Tibshirani,R. and Friedman,J. (2009) *The elements of statistical learning: data mining, inference, and prediction*. Springer Science+Business Media LLC, New York.
25. Eriksson,N., Pachter,L., Mitsuya,Y., Rhee,S.-Y., Wang,C., Gharizadeh,B., Ronaghi,M., Shafer,R.W. and Beerenwinkel,N. (2008) Viral population estimation using pyrosequencing. *PLoS Comput. Biol.*, **4**, e1000074.
26. Zagordi,O., Geyrhofer,L., Roth,V. and Beerenwinkel,N. (2010) Deep Sequencing of a Genetically Heterogeneous Sample: local Haplotype Reconstruction and Read Error Correction. *J. Comput. Biol.*, **17**, 417–428.
27. Shafer,R. and Schapiro,J. (2008) HIV-1 drug resistance mutations: an updated framework for the second decade of HAART. *AIDS Rev.*, **10**, 67.
28. Beerenwinkel,N., Schmidt,B., Walter,H., Kaiser,R., Lengauer,T., Hoffmann,D., Korn,K. and Selbig,J. (2001) Geno2pheno: interpreting genotypic HIV drug resistance tests. *IEEE Intell. Syst.*, **16**, 35–41.
29. Beerenwinkel,N., Däeumer,M., Oette,M., Korn,K., Hoffmann,D., Kaiser,R., Lengauer,T., Selbig,J. and Walter,H. (2003) Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids Res.*, **31**, 3850–3855.
30. Cock,P.J.A., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B. *et al*. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
31. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
32. R Development Core Team R. (2009) *A Language and Environment for Statistical Computing R Foundation for Statistical Computing Vienna*, Austria.
33. Kanagawa,T. (2003) Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J. Biosci. Bioeng.*, **96**, 317–323.
34. Maydt,J. and Lengauer,T. (2006) Recco: recombination analysis using cost optimization. *Bioinformatics*, **22**, 1064–1071.
35. Roche Applied Science Genome Sequencer System. Amplicon Sequencing. Application Note No. 5 (Feb, 2007).
36. Nettleton,D. and Banerjee,T. (2001) Testing the equality of distributions of random vectors with categorical components. *Comput. Stat. Data Anal.*, **37**, 195–208.
37. Simpson,E. (1949) Measurement of diversity. *Nature*, **163**, 688.
38. Metzker,M.L., Mindell,D.P., Liu,X.-M., Ptak,R.G., Gibbs,R.A. and Hillis,D.M. (2002) Molecular evidence of HIV-1 transmission in a criminal case. *Proc. Natl Acad. Sci. USA*, **99**, 14292–14297.
39. Dohm,J.C., Lottaz,C., Borodina,T. and Himmelbauer,H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, **36**, e105.
40. Laserson,J., Jojic,V. and Koller,D. (2010) Genovo: De Novo Assembly for Metagenomes. *In Research In Computational Molecular Biology*. Springer, Berlin, pp. 341–356.
41. Barrick,J.E. and Lenski,R.E. (2009) Genome-wide mutational diversity in an evolving population of *Escherichia coli*. *Cold Spring Harb. Symp. Quant. Biol.*, **74**, 119–129.