Scientific Research and Essays Vol. 7(33), pp. 2997-3009, 23 August, 2012
Available online at http://www.academicjournals.org/SRE
DOI: 10.5897/SRE12.242
ISSN 1992-2248 ©2012 Academic Journals

Full Length Research Paper

Characterization of *de novo* assemblies of quasispecies from next-generation sequencing via complex network modeling

Mattia C. F. Prosperi[1]*, Sandro Meloni[2,3], Iuri Fanti[4], Stefano Panzieri[3], Giovanni Ulivi[3] and Marco Salemi[1]

[1]Department of Pathology, Emerging Pathogens Institute, Immunology and Laboratory Medicine, University of Florida, Gainesville, Florida, USA.
[2]Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, Zaragoza, Spain.
[3]Department of Computer Science and Automation, Faculty of Computer Science Engineering, University of Roma TRE, Rome, Italy.
[4]Clinic of Infectious Diseases, Catholic University of the Sacred Heart, Rome, Italy.

Accepted 3 August, 2012

Several worldwide pandemics, such as influenza, human immunodeficiency virus, and coronavirus, are caused by viral quasispecies. Characterization of quasispecies harboring in a host is essential to unveil the mechanisms that are at the base of the pathogen evolution, infection and spread at the epidemic level. Next generation sequencing (NGS) produces many thousands of sequence fragments from a single sample, allowing the full genome sequencing at high resolution. In this work, an original approach for the *de novo* assembly (reconstruction of a full genome without the need of a reference genome) of NGS reads into the quasispecies present in the sample is introduced, using biased random walks over an overlap graph construction. The proposed framework is shown to be successful in reconstructing viral quasispecies at different diversities, using both simulated and empirical data. In addition, a broad set of measures describing topological properties of the overlap graphs is examined, in order to highlight differences in the data sets and therefore in the population structures.

Key words: Next-generation sequencing, genome assembly, quasispecies, complex network, random walk, *de novo* assembly.

INTRODUCTION

Several worldwide pandemics and chronic diseases are caused by viral quasispecies, such as influenza, human immunodeficiency virus (HIV), hepatitis C virus (HCV), and coronavirus. Quasispecies are characterized by a high genetic variability and can exhibit recombination, both due to error-prone viral replication mechanisms and host-virus interactions (Domingo et al., 1998). Knowledge and characterization of quasispecies harboring a host can be of dramatic importance, in order to unveil the mechanisms that are at the base of the pathogen evolution, infection and spread at the epidemic level.

Standard molecular sequencing, that is, Sanger's method (Sanger et al., 1977), coupled with shotgun techniques (Roach at al., 1995), produces random fragments of 500 to 1,000 base pairs from a deoxyribonucleic acid cloning. Shotgun Sanger's sequencing has a low overall base and base coverage throughput, although at high costs can be massively parallelized. Several whole genomes, including the first human genome, have been sequenced using this technique (Myers et al., 2000; Levy et al., 2007). Shotgun sequencing produces a set of sequence fragments that have to be merged together, or assembled, in order to reconstruct the original complete genome. During the past decade, a plethora of methodologies for *de novo*

*Corresponding author. E-mail: ahnven@yahoo.it. Tel: 001-352-213-7772, 001-352-273-9419. Facsimile: 001-352-273-9430.

whole-genome assembly, that is, reconstruction of a full genome without the need of any reference genome, have been introduced and implemented as software suites (Huang and Madan, 1999; Myers et al., 2000; Pevzner et al., 2001; Batzoglou et al., 2002; Mullikin and Ning, 2003).

Recent advancements on molecular sequencing (Nyrén, 2007) now allow to produce from thousands to billions of sequence fragments from a single sample, of variable length (from dozens to hundreds of bases) depending on the machinery (http://www.454.com/; http://www.illumina.com/; http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing.html; http://www.helicosbio.com/; http://www.polonator.org/; http://www.pacificbiosciences.com/; http://www.iontorrent.com/), and have been denominated by ultra-deep or next-generation sequencing (NGS). The times and costs required to sequence a human genome have been reduced radically, increasing also by several folds of the single base coverage (Wheeler et al., 2008; Wang et al., 2008; Kim et al., 2009). However, NGS techniques are usually more error-prone than Sanger sequencing. An issue with the NGS data analysis is the re-calibration of existing genome assembly algorithms, in order to deal with shorter sequence fragments, higher coverage, higher fragment number, and higher error rates (Dohm et al., 2007; Butler et al., 2008; Chaisson and Pevzner, 2008; Zerbino and Birney, 2008; Miller et al., 2008; Medvedev and Brudno, 2009).

Although the principal application of NGS is the (re-) sequencing of large and very-large genomes in short time (now approximately one month for a human genome), another important application is the sequencing of viral quasispecies for clinical and epidemiological purposes. For instance, recent applications of NGS produced the whole-genome assembly of H1N1 influenza A strain (Kuroda et al., 2010) and multiple type-1 HIV (HIV-1) strains (Henn et al., 2012). The case of HIV is of particular interest because the virus is characterized by a fast replication and a high mutation rate ($\approx 3\cdot 10^{-5}$ per nucleotide base per cycle of replication, producing $\approx 10^{10}$ virions per day), with a large number of different immune-escape and drug-resistance mutational pathways induced by host genetics and treatment experience (Wang et al., 2007; Archer et al., 2009). It has been shown that the HIV minority variants carrying drug resistance mutations (detectable by NGS, but not usually with Sanger sequencing) can impact the patient’s response to antiretroviral therapy (Simen et al., 2009).

The current software for *de novo* NGS assembly is not specifically designed to reconstruct all the coexisting variants within a quasispecies, but rather to infer a unique consensus genome and to map its allelic variations.

The *de novo* assemblers originally developed for Sanger’s shotgun sequencing are based on an overlap graph construction (Myers et al., 2000): each node of the graph represents a sequence fragment (read) and all pairwise alignments between them can be computed. An edge between two sequence fragments is then created when there is a consistent (under some criteria) overlap between the two reads (Figure 1). The edges can be bi-directed in order to account for forward- or reverse-strand sequenced deoxyribonucleic acid. Via some graph visit policies, such as maximum flow, a genome is progressively reconstructed, or in case of inconsistencies, a set of *contigs* (substrings of the genome to be reconstructed that could not be merged together) is reported. Some algorithms (Dohm et al., 2007) discard fragments to be added to a contig when they have same prefix and different suffixes. This might be a good way to handle repeats, but makes no sense when assembling a quasispecies. Other methods (Miller et al., 2008) account for base mismatches, but always output a consensus genome.

Other assemblers, especially those built for NGS, break the reads into smaller fragments of *k* length (*k*-mers) and compute a De Bruijn graph, reducing computational complexity (Medvedev and Brudno, 2009). The De Bruijn graph considers overlaps of length *k-1* between the *k*-mers; usually the value of *k* is optimized by testing some ranges in relation to the average read length and sequencing error rates. The De Brujin graph approach can be more efficient when the NGS technology has a very large overall throughput but short read length (like the Illumina technology, producing ~40 million reads, each of ~75 bases), and it is ideal for reconstructing long genomes. Since a viral quasispecies has usually a short- or medium-size genome (thousands of bases), but a high internal variability, it is preferable to use an NGS technique that gives less overall throughput but longer read length (like Roche 454, generating ~1 million reads, each of ~450 bases). The overlap graph paradigm can have some advantages, since the aim is to reconstruct a set of paths, explaining the coexisting variants.

Given this scenario, there is a need to develop and test *de novo* assemblers specifically tailored to quasispecies characterization. There are several studies that investigate the quasispecies reconstruction by means of re-sequencing using NGS data, but they are limited by the availability of an already sequenced genome (Westbrooks et al., 2008; Jojic et al., 2008; Eriksson et al., 2008; Zagordi et al., 2010a; Prosperi et al., 2011; Zagordi et al., 2011; Beerenwinkel and Zagordi, 2011). Along with reconstruction algorithms, procedures for error correction of NGS data outputs have been also introduced (Wang et al., 2007; Eriksson et al., 2008; Qu et al., 2009; Zagordi et al., 2010a, b; Skums et al., 2012).
In this work, an original approach for the *de novo* assembly of a quasispecies from NGS data is introduced. The approach relies on the overlap graph paradigm, but revises the edge construction policy, adopting a statistical criterion based on pairwise local alignment scores.

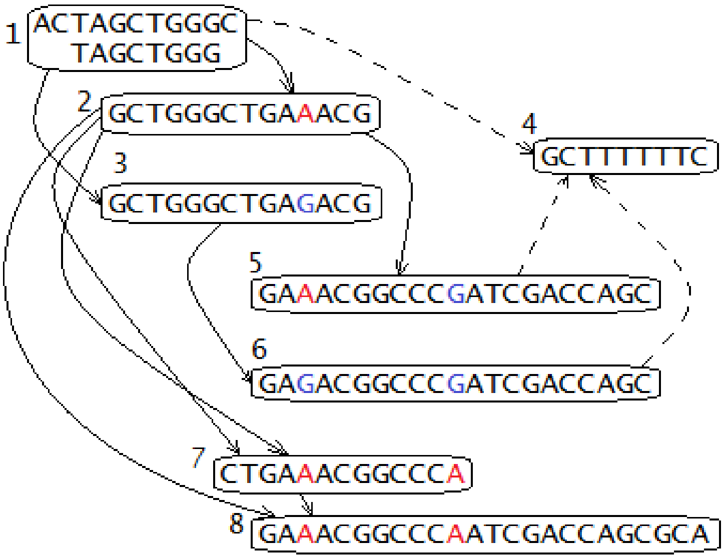

Figure 1. Overlap graph example. Sequence fragments (reads) are nodes and edges represent a consistent overlap (perfect local alignment match) between two reads. Identical reads or sub-strings of longer reads are merged into unique nodes. Directions are consistent with the ordering of the non-overlapping heads/tails of two aligned reads. The dashed edge represents a consistent overlap which might not be significant when comparing it against the null hypothesis of seeing a consistent overlap due to chance. Note that the local alignment algorithm automatically selects the longest substring match between two sequences, allowing a certain number of gaps/mismatches (given as input parameter).

Leveraging the theory of random walks on complex networks (Redner, 2001; Barrat et al., 2008; Boccaletti et al., 2006) a new algorithm is proposed here to produce a set of paths which are aimed at reconstructing the genomes of different variants in a quasispecies. This framework is applied to simulated viral quasispecies with different diversity, prevalence, repeats, super-infections and recombination. Also, empirical data publicly available are analyzed (Zagordi et al., 2010b). Finally, a broad set of topological indicators from graph theory is screened (Wasserman and Faust, 2004; Newman et al., 2006; Newman, 2010) in order to unveil differences among different experimental set ups and to characterize a quasispecies independently from the reconstruction phase.

MATERIALS AND METHODS

Overlap graph construction

Given a NGS data sample *N* composed by *n* reads, we assume that all reads are forward-stranded. A heuristic to induce a consistent orientation for all reads is given as supplementary material and can be computed at the same time with the overlap graph construction procedure, given as follows.

All collapse duplicate reads are stored in their relative frequency. If some reads appear as substring of one or more other different reads, it collapses them randomly into one of the superstrings. Substrings can contribute to the relative frequency of their superstring by a factor proportional to their length as compared to the superstring length. At this point, it is assumed that extremely short reads (± 2 standard deviations from the average read length) and reads with poor base quality (such as ambiguous base callings) have been filtered out. This is a pre-processing step that is often provided by the proprietary companion software of the NGS machine.

Computes all pairwise local alignments (Smith and Waterman, 1981; Gotoh, 1982) between reads and store them along with the corresponding alignment scores *s*. Estimate an *a priori* distribution of quasi-random local pairwise alignment scores *SR* by drawing randomly with replacement *m* ($m \approx n$) reads from *N*, shuffling the characters of each drawn sequence, each pairwise alignment, perform a statistical test by comparing each score *s* against the *SR* distribution (Bacro and Comet, 2000), adjusting p-values for multiple comparisons (Benjamini and Hochberg, 1995).

Retain all alignments with a consistent overlap. A consistent overlap is defined as (1) a perfect match in the local alignment and (2) a significant p-value (<0.01). Condition (1) is a strong assumption and is efficient only when the overall sequencing error rates are negligible. Indeed, error rates and types vary across different NGS technologies. For the Roche 454 technology, base insertions/deletions are more frequent than mismatches, and they depend on the neighboring region: in presence of homopolymers, that is, more than three consecutive bases of the same type, the error probability increases (Gilles et al., 2011). The Illumina technology instead is more prone to mismatches. By supposing an uniform error rate of 0.5% for any error type and any neighboring region, given an average read length of 250 bases, then ~29% of the reads are expected to contain at least one error. This can be a considerable problem when looking for a perfect local alignment match. However, either the local alignment can allow for a certain number of mismatches, insertions or deletions, or the reads can be pre-processed to correct errors, using *k*-mers-based algorithms that do not require any reference genome (Skums et al., 2012). Constraint (2), meaning an alignment p-value <0.01, allows dropping overlaps that might be obtained by chance.

The overlap graph is then constructed by setting a node for each read and an edge between two reads with a consistent and statistically significant overlap (Figure 1). The arc direction is induced by the local alignment; the edge direction goes from the read whose tail overlaps with the head of the other read.

Of note, cycles in the overlap graph should be virtually impossible, except for some degenerate cases such as for the two sequences AAAA...AAAAAAACCCCCC...CCCCC, and CCCCC...CCCCCAAAAAAA...AAAA.

The computational complexity for the overlap graph construction is $O(n^2)$, where n is the number of reads, since $n \cdot (n-1)/2$ pairwise comparisons are required. However, the complexity of a local pairwise alignment is $O(m_1 \cdot m_2)$, where m_1 and m_2 are the lengths of the two sequences to be aligned.

Topological analysis of overlap graph

Different topological indicators from graph theory and complex network analysis (Newman et al., 2006, Newman, 2010) have been evaluated as follows. A first insight on the possible number of quasispecies present in each experiment can be obtained studying the *ratio vertices/edges*, along with the percentage of *sources, sinks, isolated,* and *normal nodes*. The *number, diameter* and *size of connected components* (excluding isolated vertices) of the corresponding overlap graph were also evaluated. Then, for each component the *relative density* (ratio between the real number of edges and the maximum possible number of connections between the vertices) and the mean number of *in-going* and *out-going connections* can give a rough estimate of the number of possible different paths present in the network. Other structural measures examined were the *maximum degree* and *transitivity* (fraction of directed cliques, $i \rightarrow j$, $k \rightarrow j$, $i \rightarrow k$, in the graph). A force-based algorithm (Fruchterman and Reingold, 1991) was employed for visualising the network and its structure, implemented in the “igraph” package of the R suite (http://www.R-project.org).

Biased random walk and quasispecies reconstruction

Given the directed overlap graph and considering all its connected components, it is possible to determine sources (nodes with only incoming arcs) and sinks (nodes with only outgoing arcs). Some sources or sinks produced will be due to uncorrected sequencing errors, and others will correspond to the beginning or end of the genome(s) to be reconstructed. Except in the extreme case of a circular genome and no sequencing errors, there will always be sources and sinks.

In order to reconstruct a single genome or a quasispecies, we define a biased random walk (BRW) on the overlap graph as follows. Select at random a source node, weighting the selection probability by the relative frequency and number of outgoing arcs, wherein sources with higher frequency and a high number of connections are more likely to be selected. Move randomly from that node to another by giving higher probability to step into a node that also has higher frequency and higher number of outgoing nodes. Intuitively, nodes at low frequency with a low number of connections are likely to be reads with uncorrected sequencing error, or representatives of minority variants. The BRW stops when a sink is found or if the same node is visited. Since all connected nodes have a consistent overlap, it is possible to reconstruct uniquely a contig for each BRW. In the case where a consistent overlap was defined as a non-perfect match, with a fixed number of mismatches, insertions or deletions, then the reconstructed contig will contain the base from the node with the highest weight.

The BRW is executed for a large number of times and all different paths are counted along with their relative frequencies and their lengths. Paths with a length below the 25th percentile of the overall path length distribution are discarded. From the retained paths, all the different contigs are selected and counted.

Next-generation sequencing data simulation

A simulator of NGS was set up by considering the Roche 454 technology. The simulator takes as input one or more genomes (with associated prevalence) and draws a shotgun sample where the probability to generate a read from a genome is proportional to the genome prevalence. Average (st.dev.) length of reads was 250 (25) bases. Insertion and deletion error rate in homopolymeric regions (that is, more than three consecutive bases of the same type) was fixed to 0.005, while base mismatch rate was 0.001. Sequencing error probability in the head and tail of reads (10 bases) was higher, set to 0.005. The probability to generate a reverse-stranded read was set to 0.5.

Several data sets were generated, according to the following genome sets: (i) HIV-1, group M, subtype B polymerase gene (*pol,* n=2,844 bases considered); (ii) HIV-1, group M, subtypes A1, B, C, F1, and H *pol*, with an overall mean diversity of 11%; (iii) HIV-1, group M, subtypes B, F1, and one-point subtype B/J recombinant *pol*, with the form B/J/B; (iv) swarm of 10 variants of HIV-1, group M, subtype B *pol* genome at 4% diversity, mixing them at different prevalence increasing linearly from 2 to 18%; (v) modified HIV-1, group M, subtype B *pol* with two repeated regions according to the motif X/repeat/Y/repeat/Z.

According to Eriksson et al. (2008), the minimum number of reads n needed to detect a variant at frequency f, with probability p, is $n=-g \cdot \log_e (1-p^{1/n})/(f \cdot r)$, where g is the genome length and r is the read length. Therefore, at p=90%, given the HIV-1 *pol*, the minimum number of reads needed to detect a variant at 10% prevalence is n=1,160, while to detect a variant at 1% prevalence is n=11,607. For this reason, the number of reads simulated was 3,000, except for simulation (iv) for which 10,000 reads were used.

Empirical next-generation sequencing data set

The overlap graph construction and the BRW were applied to empirical HIV NGS data publicly available (Zagordi et al., 2010b), in which 10 HIV-1, group M, subtype B clones from different patients, encompassing a portion of the *pol* gene (1,245 base pairs, previously sequenced using Sanger technology), were pooled in a mixed sample, in different proportions, and re-sequenced using the Roche 454. The experiment was designed such that the variant proportions were halved progressively starting from 30%, to a minimum of 0.1%. The average population diversity was ~7%, with an estimated rate of heterogeneity of 0.35.

Model evaluation

The reconstructed genome(s) of each experiment were evaluated against the original strains, with some control sequences and outgroups, in terms of phylogenetic and recombination analyses, using the SplitsTree software suite (Huson and Bryant, 2006). For each experiment, reconstructed sequences, plus originals, controls and outgroups were fed to the MUSCLE multiple alignment software (Edgar, 2004). A thousand of bootstrap samples of each multiple alignment was used to assess node reliability for both

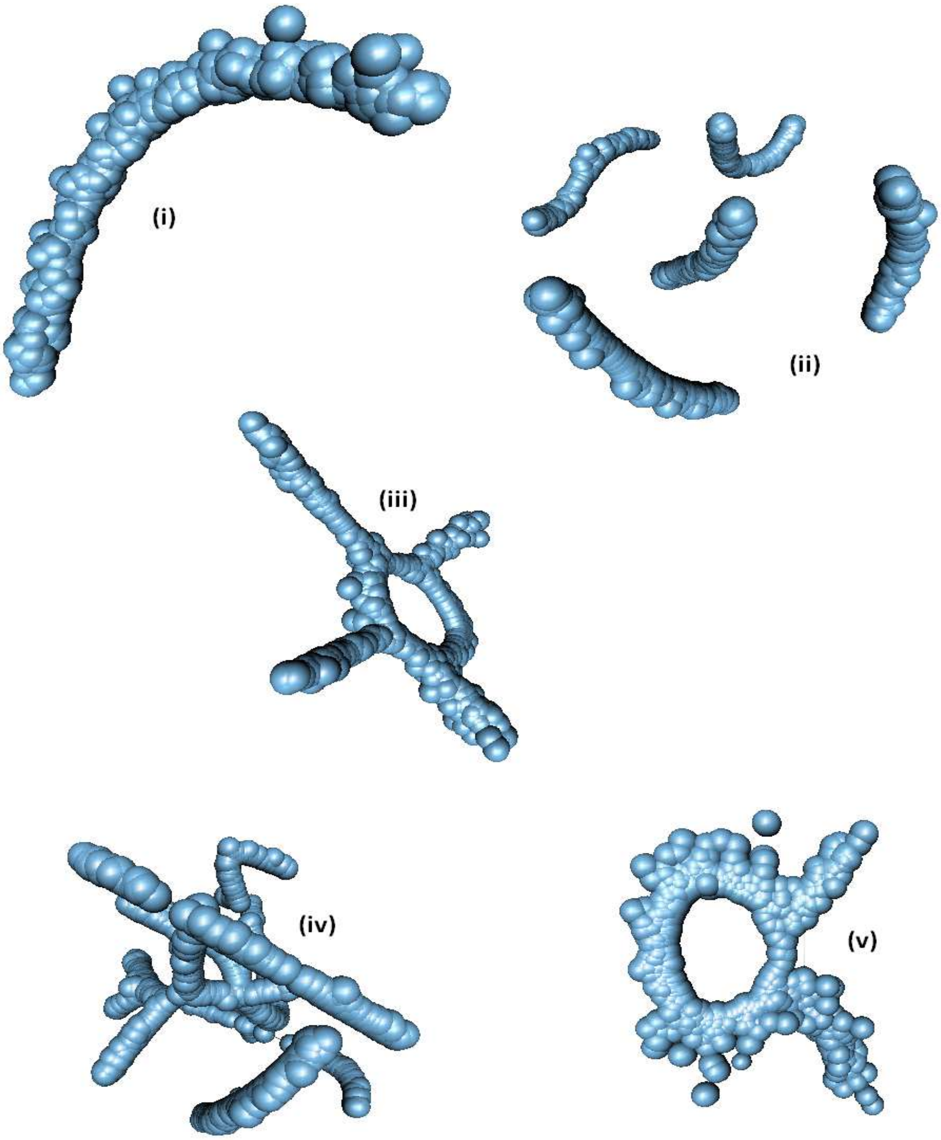

Figure 2. Three-dimensional visualization (Fruchterman-Reingold force-based algorithm) of overlap graphs for different NGS simulations of HIV-1 *pol*. Graph (i) represents a unique strain genome; (ii) five distinct strains at high diversity (11%); (iii) a super-infection of two pure strains and a recombinant form; (iv) a swarm of 10 variants at a low (4%) diversity; (v) a unique strain with a repeated region motif of the form X/repeat/Y/repeat/Z.

neighbor-joining and neighbor-net algorithms of phylogenetic inference, using the LogDet distance.

RESULTS

Simulations

For each of the generated data sets (n=3,000 reads, except for simulation (iv), with 10,000 shotgun read samples) an overlap graph was constructed. Figure 2 shows a three-dimensional plot of estimated networks by using a force-based layout algorithm. The visual inspection itself gives a good description of the different genome types: the unique variant (i) of HIV-1 induces a network with one (almost) fully connected component; for the set (ii) of five distinct HIV-1 subtypes (11% diversity) the network exhibits 5 independently connected components; the super-infection (iii) of two different viral subtypes that recombine within a host shows two single components (variants B and J) and the possibility to move from one to another (creating the actual recombinant B/J/B, but virtually also fake B/B/J and J/J/B); experiment (iv) has a similar graph layout as that of simulation (ii), but the components are more close to

Table 1. Topological measures and indices for overlap graphs inferred from the NGS HIV quasispecies simulations and empirical data.

Data set (HIV-1 group M *pol*)	Vertices / edges	Density	Connected components	Size of the biggest connected component	Diameter of the biggest connected component	Transitivity	Mean degree (in/out)	Max degree (in/out)	% of sources	% of sinks
Unique genome	663/4354	0.00992	1	591	82	0.334	6.6/6.5	22/23	29.1	26.9
Different subtypes at high diversity	1046/2449	0.0024	5	225	35	0.389	2.3/3.0	11/9	18.0	14.3
Super-infection of two variants with evidence of recombination	903/4266	0.00524	1	717	70	0.321	4.7/4.0	21/19	22.4	19.2
Quasispecies of 10 variants at low diversity (4%)	4828/22683	0.00097	10	1455	54	0.345	5.0/4.8	24/23	23.5	20.0
Unique variant with a repeated region	663/5623	0.0121	1	592	81	0.379	8.5/8.2	37/36	28.6	25.6
Empirical data	8671/26027	0.00035	14	1179	15	0.123	3.1/3.0	94/139	10.9	27.3
Empirical data (error-corrected)	527/2560	0.00924	5	18	10	0.4105	4.8/4.6	31/24	14.8	12.5

each other and there is the possibility to move from one to another at a certain point, since the diversity among variants is lower (4%) and in some regions one or more variants can be indistinguishable; finally, the genome with a repeated motif X/repeat/Y/repeat/Z is characterized by a main loop from the source to the sink, with the possibility to reconstruct both the genome X/repeat/Z and X/repeat/Y/repeat/Z, and eventually any other X/repeat/Y/repeat/Y/.../Y/repeat/Z.

Results of the topological analysis for the aforementioned graphs are given in Table 1. In all the cases, the networks show a peculiar structure characterized by a low number of edges with respect to the number of vertices, small differences between the average and the maximum degree, and also a small deviation between in- and out-degree. These indices suggest that vertices in the graphs are all characterized by almost the same number of connections and only few deviations are present. All the networks show a very low density denoting that the number of possible paths to be considered in the reconstruction is limited. More importantly, the number of connected components indicates the number of quasispecies present in the original genome.

Phylogenetic and recombination analysis

By applying the BRW described in the methods at each data set, we were able to reconstruct exactly the genomes of simulation (i) and (ii), with no errors. Concerning simulation (iii), both the pure B and J subtypes were reconstructed exactly, along with the recombinant form B/J/B. However, two *in silico* recombinants were also reported, namely the B/B/J and J/J/B variants. Nonetheless, it would be topologically impossible, due to the arc direction constraints, to reconstruct a J/B/J recombinant. For simulation (iv), the quasispecies at low diversity (4%), the BRW was able to reconstruct only 30% of the original population, capturing the variants with highest frequencies (>10%). The sample size (10.000) was allowing for a reconstruction up to 1% prevalence, but the induced error most likely caused the performance

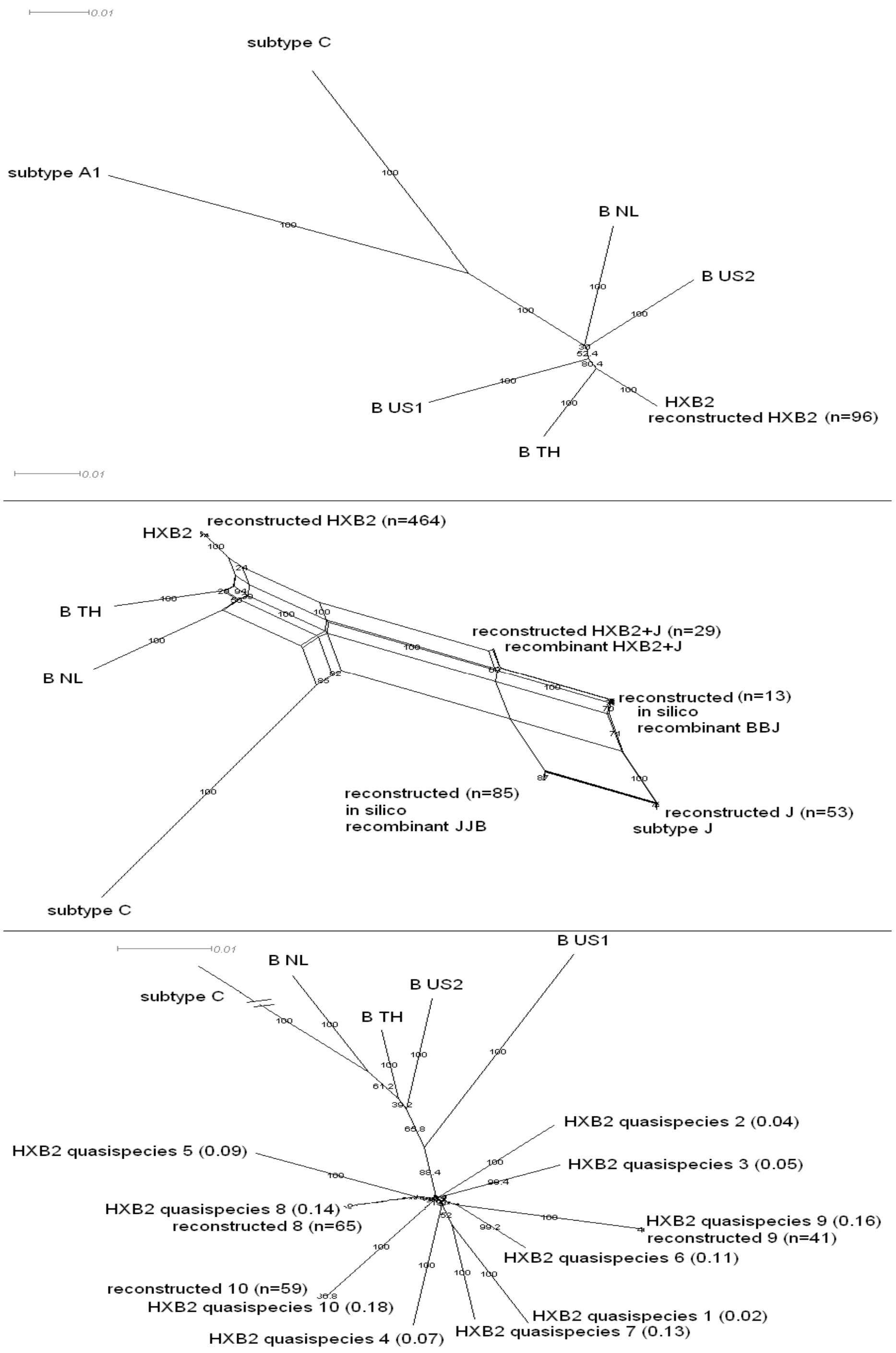
0.01
subtype C
subtype A1
B NL
B US2
B US1
HXB2
reconstructed HXB2 (n=96)
B TH
0.01
reconstructed HXB2 (n=464)
HXB2
B TH
B NL
reconstructed HXB2+J (n=29)
recombinant HXB2+J
reconstructed (n=13)
in silico
recombinant BBJ
reconstructed (n=85)
in silico
recombinant JJB
reconstructed J (n=53)
subtype J
subtype C
0.01
B US1
B NL
subtype C
B US2
B TH
HXB2 quasispecies 2 (0.04)
HXB2 quasispecies 5 (0.09)
HXB2 quasispecies 3 (0.05)
HXB2 quasispecies 8 (0.14)
reconstructed 8 (n=65)
HXB2 quasispecies 9 (0.16)
reconstructed 9 (n=41)
HXB2 quasispecies 6 (0.11)
reconstructed 10 (n=59)
HXB2 quasispecies 10 (0.18)
HXB2 quasispecies 1 (0.02)
HXB2 quasispecies 7 (0.13)
HXB2 quasispecies 4 (0.07)

Figure 3. Phylogenetic analysis of reconstructed genome variants (and prevalence) for selected simulation experiments and comparison with original genomes and other control genomes, plus some outgroup species. Upper panel is for a unique variant (case i), middle panel for recombinant superinfection (case iii), and lower panel for a quasispecies at a 4% diversity (case iv). Labels on network branches represent reliability (% of bootstrap replicates).

decrease. Finally, for BRW on simulation (v), that is, a genome with the large repeated motif, yielded genomes of considerable different lengths, with a variable number of repeats.

Figure 3 depicts bootstrapped phylogenetic Trees and recombination networks inferred after aligning the reconstructed genomes with the original strains, four control sequences (different isolates of the same subtype B) and two outgroup sequences (HIV-1 subtypes C and A1).

Empirical data analysis

The HIV-1 data set from Zagordi et al. (2010b) consisted of 16.540 reads. The average (st.dev.) read length was 305 (117) bases. The read set was preliminarily filtered by mapping and trimming all the reads against an HIV-1 group M subtype B *pol* gene reference sequence (Genbank identifier HXB2CG), in order to eliminate contaminants. This step was not necessary for the *de novo* assembly, but yielded a more compact data set (14.654 reads retained) and a less skewed distribution of quasi-random scores; in addition, we were able to calculate in advance the average (st.dev.) base coverage, which was 3.091 (1.405). The number of distinct reads was 8.672 and the network was created on this set. Over the total number of nodes, there were 948 (11%) sources, 2371 (27%) sinks, 4829 (56%) isolated, and 524 (6%) regular. The isolated nodes accounted for a high percentage of the overall number of connected components (4829/4903). Only 35 connected components had a size >3, and the first 10 had a size >10. By applying the read error correction method proposed by Zagordi et al. (2010a), before constructing the network, the number of distinct reads decreased to 527, yielding a network of 78 (15%) sources, 66 (12%) sinks, 104 (20%) isolated, and 279 (53%) regular nodes. Topological indices for the two overlap graphs constructed on the non-error-corrected and error-corrected set of reads are listed in Table 1, while Figure 4 depicts the two networks and phylogenetic trees including the variants estimated by the BRW plus the original Sanger clones. The BRW applied to both networks was able to reconstruct 5/6 variants (over a total of 10) unequivocally clustering with the original sequence clones, but the BRW on the error-corrected network yielded reconstructions much closer to the originals in terms of nucleotide differences. Of note, the 4 missed variants were those at the lowest frequency (<0.5%).

DISCUSSION

This study addresses the problem of *de novo* assembly of a quasispecies from NGS data outputs, investigating the properties of an overlap graph inferred from sequence fragments, and introducing an original method for full-genome reconstruction based on biased random walks. Different from previous *de novo* approaches (Miller et al., 2008), the presented method is able to reconstruct a whole population of genomes, that is, a viral quasispecies harboring a host, rather than a single genome, while at the moment only reference-based methods have been proposed (Beerenwinkel and Zagordi, 2011). The original overlap graph construction algorithm and the biased random walk permit also to partly correct for sequencing errors. With a set of simulated NGS experiments, and by means of phylogenetic analysis, we showed that the proposed framework can be successful in reconstructing viral quasispecies at different diversities, although performance can be affected by decreasing the variant prevalence and the overall diversity. It is also capable, to some extent, to function in the presence of super-infections, recombination, and problematic genomes with repeats. When applying our framework to empirical data (Zagordi et al., 2010b), it successfully reconstructed all the variants at a prevalence >0.5%, although a preliminary step of error correction becomes necessary.

The analysis of the topological properties of the overlap graph can provide interesting insights on the population structure of the sequenced quasispecies, independently from the reconstruction algorithm used. The number of connected components and the network density, for instance, give a direct indication of the number of coexisting variants.

Conclusions

The current available software for *de novo* NGS assembly is not specifically designed to reconstruct a quasispecies, but rather to infer a unique consensus genome and to map its allelic variations. In addition, all the quasispecies assembly methods developed so far are all based on re-sequencing, that is, reference alignment, thus the investigation of new pathogens is limited by the availability of a previously sequenced genome. Therefore, there is a need to develop and test *de novo* assemblers tailored to quasispecies characterization, given the potential impact in the translational science, especially in terms of clinical relevance, for chronic and life-threatening diseases such as HIV or HCV. This study was limited to the analysis of a few particular examples of a viral quasispecies, as a proof-of-concept, but the reported findings are promising in a broader perspective of NGS experiments for the characterization of any quasispecies.

LIMITATIONS

This work has some limitations. The first problem centers on the handling of errors. With high error rates (~0.5%) and long reads (>250 bases), a considerable proportion of the reads is expected to present errors (Gilles et al., 2011), affecting the whole overlap graph, therefore either a pre-processing step for error reduction is required (Skums et al., 2012), or a less strict definition of

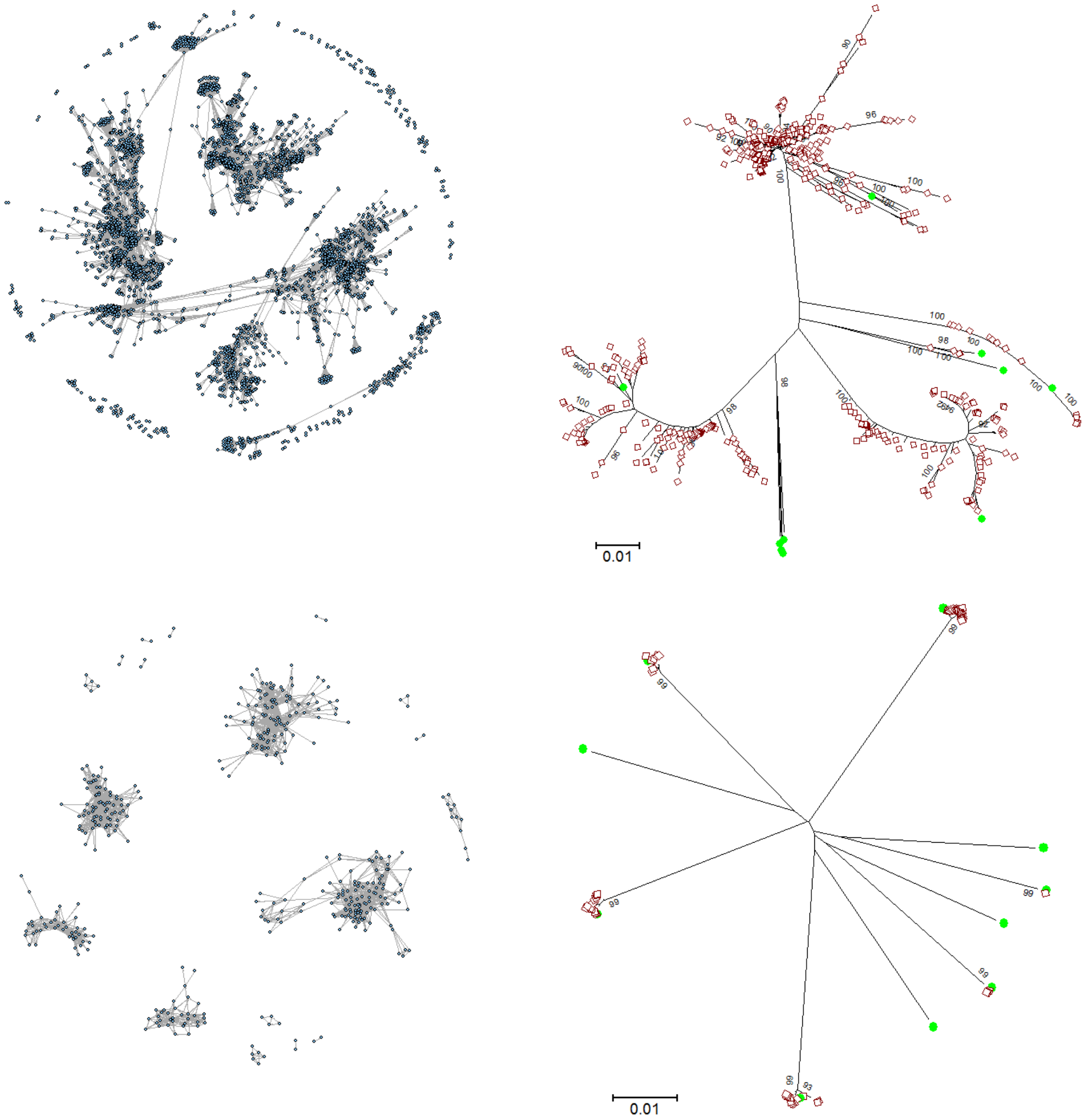

Figure 4. Overlap graph and phylogenetic analysis of reconstructed genome variants on the empirical HIV-1 NGS data set (Zagordi et al., 2010b), considering non-error-corrected (upper panel) and error-corrected (lower panel) reads. Green dots represent original (real) variants, whilst red diamonds represent inferred variants. Labels on tree branches represent reliability (% of bootstrap replicates).

consistent overlap, as explained in the methods. The second problem is the variant prevalence distribution and quasispecies diversity; as they decrease an increase in read number (to catch minority variants) and read length (to minimize the overlap ambiguities in conserved regions) is needed. The rate of heterogeneity also plays a role. However, this second problem is common to most existing methodologies. The problem of recombination and repeat handling deserves more investigation, especially when selecting paths from the BRW runs, along with other scenarios in which this approach should be tested. For instance, in the case of a circular genome, since the random walk usually starts usually from a source, the walk would start from fake sources probably caused by sequencing errors. However, additional rules can be introduced in these particular cases.

From an implementation point of view, probably this approach needs a software design tailored to multi-core parallel computation, given the complexity of the overlap graph inference and the huge data throughput of NGS. De Bruijn graph construction seems to be preferred to the overlap graph for NGS assembly as the core of software implementations, but in the case of long reads (as for Roche 454), possibly a hybrid approach based on long *k*-mers could be investigated.

ACKNOWLEDGMENTS

This work was supported in part by the University of Florida award UL1 RR02989, by the NIH/NINDS R01 grant NS063897-01A2, and by the EU FP7 FET “Dynanets” grant #233847. We thank David J. Nolan (University of Florida) for revising the English language throughout the manuscript.

REFERENCES

Archer J, Braverman MS, Taillon BE, Desany B, James I, Harrigan PR, Lewis M, Robertson DL (2009). Detection of low-frequency pretherapy chemokine (CXC motif) receptor 4 (CXCR4)-using HIV-1 with ultra-deep pyrosequencing. AIDS 23(10):1209-1218.

Bacro JN, Comet JP (2000). Sequence alignment: an approximation law for the Z-value with applications to databank scanning. Comput. Chem. 25(4):401-410.

Barrat A, Barthélemy M, Vespignani A (2008). Dynamical Processes on Complex Networks. Cambridge University Press, Cambridge, UK.

Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES (2002). ARACHNE: a whole-genome shotgun assembler. Genome Res. 12(1):177-189.

Beerenwinkel N, Zagordi O (2011). Ultra-deep sequencing for the analysis of viral populations. Curr. Opin. Virol. 1(5):413-418.

Benjamini Y, Hochberg Y (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Statist. Soc. B. 57(1):289-300.

Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006). Complex Networks: Structure and Dynamics. Phys. Reps. 424(4-5):175-308.

Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB (2008). ALLPATHS: *de novo* assembly of whole-genome shotgun microreads. Genome Res. 18(5):810-820.

Chaisson MJ, Pevzner PA (2008). Short read fragment assembly of bacterial genomes. Genome Res. 18(2):324-330.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2007). SHARCGS, a fast and highly accurate short-read assembly algorithm for *de novo* genomic sequencing. Genome Res. 17(11):1697-1706.

Domingo E, Baranowski E, Ruiz-Jarabo CM, Martin-Hernandez AM, Saiz JC, Escarmis C (1998). Quasispecies Structure and Persistence of RNA Viruses. Emerg. Infect. Dis. 4(4):521-527.

Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792-1797.

Eriksson N, Pachter L, Mitsuya Y, Rhee SY, Wang C, Gharizadeh B, Ronaghi M, Shafer RW, Beerenwinkel N (2008). Viral population estimation using pyrosequencing. PLoS Comput. Biol. 4(4):e1000074.

Fruchterman TMJ, Reingold EM (1991). Graph Drawing by Force-directed Placement. Softw. Pract. Exper. 21(11):1129-1164.

Gilles A, Meglécz E, Pech N, Ferreira S, Malausa T, Martin JF (2011). Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics 12:245.

Gotoh O (1982). An improved algorithm for matching biological sequences. J. Mol. Biol. 162(3):705-708.

Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, Macalalad AR, Berlin AM, Malboeuf CM, Ryan EM, Gnerre S, Zody MC, Erlich RL, Green LM, Berical A, Wang Y, Casali M, Streeck H, Bloom AK, Dudek T, Tully D, Newman R, Axten KL, Gladden AD, Battis L, Kemper M, Zeng Q, Shea TP, Gujja S, Zedlack C, Gasser O, Brander C, Hess C, Günthard HF, Brumme ZL, Brumme CJ, Bazner S, Rychert J, Tinsley JP, Mayer KH, Rosenberg E, Pereyra F, Levin JZ, Young SK, Jessen H, Altfeld M, Birren BW, Walker BD, Allen TM (2012). Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS Pathog. 8(3):e1002529.

Huang X, Madan A (1999). CAP3: A DNA sequence assembly program. Genome Res. 9(9):868-877.

Huson DH, Bryant D (2006). Application of Phylogenetic Networks in Evolutionary Studies. Mol. Biol. Evol. 23(2):254-2567.

Jojic V, Hertz T, Jojic N (2008). Population sequencing using short reads: HIV as a case study. Pac. Symp. Biocomput. 13:114-125.

Kim JI, Ju YS, Park H, Kim S, Lee S, Yi JH, Mudge J, Miller NA, Hong D, Bell CJ, Kim HS, Chung IS, Lee WC, Lee JS, Seo SH, Yun JY, Woo HN, Lee H, Suh D, Lee S, Kim HJ, Yavartanoo M, Kwak M, Zheng Y, Lee MK, Park H, Kim JY, Gokcumen O, Mills RE, Zaranek AW, Thakuria J, Wu X, Kim RW, Huntley JJ, Luo S, Schroth GP, Wu TD, Kim H, Yang KS, Park WY, Kim H, Church GM, Lee C, Kingsmore SF, Seo JS (2009). A highly annotated whole-genome sequence of a Korean individual. Nature 460(7258):1011-1015.

Kuroda M, Katano H, Nakajima N, Tobiume M, Ainai A, Sekizuka T, Hasegawa H, Tashiro M, Sasaki Y, Arakawa Y, Hata S, Watanabe M, Sata T (2010). Characterization of quasispecies of pandemic 2009 influenza A virus (A/H1N1/2009) by *de novo* sequencing using a next-generation DNA sequencer. PLoS One. 5(4):e10256.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, Venter JC (2007). The Diploid Genome Sequence of an Individual Human. PLoS Biol. 5(10): e254.

Medvedev P, Brudno M (2009). Maximum Likelihood Genome Assembly. J. Comput. Biol. 16(8):1-16.

Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, Johnson J, Li K, Mobarry C, Sutton G (2008). Aggressive assembly of pyrosequencing reads with mates. Bioinformatics 24(24):2818-2824.

Mullikin JC, Ning Z (2003). The phusion assembler. Genome Res. 13(1):81-90.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, Anson EL, Bolanos RA, Chou HH, Jordan CM, Halpern AL, Lonardi S, Beasley EM, Brandon RC, Chen L, Dunn PJ, Lai Z, Liang Y, Nusskern DR, Zhan M, Zhang Q, Zheng X, Rubin GM, Adams MD, Venter JC (2000). A whole-genome assembly of Drosophila. Science 287(5461):2196-2204.

Newman MEJ, Barabási AL, Watts DJ (2006). The Structure and Dynamics of Networks. Princeton University Press, Princeton, USA.

Newman MEJ (2010). Networks. An Introduction. Oxford University Press, Oxford, UK.

Nyrén P (2007). The History of Pyrosequencing. Methods Mol. Biol. 373:1-14.

Pevzner PA, Tang H, Waterman MS (2001). An Eulerian path approach to DNA fragment assembly. Proc. Natl. Acad. Sci. USA. 98(17):9748-9753.

Prosperi MC, Prosperi L, Bruselles A, Abbate I, Rozera G, Vincenti D, Solmone MC, Capobianchi MR, Ulivi G (2011). Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing. BMC Bioinformatics. 12:5.

Qu W, Hashimoto S, Morishita S (2009). Efficient frequency-based de novo short-read clustering for error trimming in next-generation sequencing. Genome Res. 19(7):1309-1315.

Redner S (2001). A guide to first-passage processes. Cambridge University Press, Cambridge, UK.

Roach JC, Boysen C, Wang K, Hood L (1995). Pairwise end sequencing: a unified approach to genomic mapping and sequencing. Genomics 26(2):345-353.

Sanger F, Nicklen S, Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA 74(12):5463-5467.

Simen BB, Simons JF, Hullsiek KH, Novak RM, Macarthur RD, Baxter JD, Huang C, Lubeski C, Turenchalk GS, Braverman MS, Desany B, Rothberg JM, Egholm M, Kozal MJ (2009). Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naive patients significantly impact treatment outcomes. J. Infect. Dis. 199(5):693-701.

Skums P, Dimitrova Z, Campo DS, Vaughan G, Rossi L, Forbi JC, Yokosawa J, Zelikovsky A, Khudyakov Y (2012). Efficient error correction for next-generation sequencing of viral amplicons. BMC Bioinforma. 13(10):S6.

Smith TF, Waterman MS (1981). Identification of common molecular subsequences. J. Mol. Biol. 147(1):195-197.

Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW (2007). Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. Genome Res. 17(8):1195-1201.

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, Guo Y, Feng B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, Ma L, Li G, Yang Z, Zhang G, Yang B, Yu C, Liang F, Li W, Li S, Li D, Ni P, Ruan J, Li Q, Zhu H, Liu D, Lu Z, Li N, Guo G, Zhang J, Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su Y, San A, Ping C, Yang S, Chen F, Li L, Zhou K, Zheng H, Ren Y, Yang L, Gao Y, Yang G, Li Z, Feng X, Kristiansen K, Wong GK, Nielsen R, Durbin R, Bolund L, Zhang X, Li S, Yang H, Wang J (2008). The diploid genome sequence of an Asian individual. Nature 456(7218):60-65.

Wasserman S, Faust K (2004). Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sci.) Cambridge University Press, Cambridge, UK.

Westbrooks K, Astrovskaya I, Campo D , Khudyakov Y , Berman P, Zelikovsky A (2008). HCV Quasispecies Assembly Using Network Flows. In: Proceedings of 4th International Symposium on Bioinformatics Research and Applications (ISBRA), pp. 159-170.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature 452(7189):872-876.

Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N (2011). ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinforma. 12:119.

Zagordi O, Geyrhofer L, Roth V, Beerenwinkel N (2010a). Deep Sequencing of a Genetically Heterogeneous Sample: Local Haplotype Reconstruction and Read Error Correction. J. Comput. Biol. 17(3):417-428.

Zagordi O, Klein R, Däumer M, Beerenwinkel N (2010). Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. Nucleic Acids Res. 38(21):7400-7409.

Zerbino DR, Birney E (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. 18(5):821-829.

Supplementary material

Induction of a whole forward-stranding given a sample of forward- and reverse-stranded reads

A NGS data output sample is usually formed both by forward- and reverse-stranded deoxyribonucleic acid sequence fragments. The overlap graph construction presented in the methods section requires a set of reads that are all forward-stranded. This can be achieved with a heuristic procedure, embeddable in the overlap graph construction. First, the local alignment score distribution or quasi-random reads *RS* accounts implicitly of read orientations, since it is calculated from character-shuffled original reads (that can be both forward and reverse-stranded). If the overlap graph construction is done in a depth-first fashion, given an arbitrary reference for the forward-strand, then it is possible to achieve an almost fully consistent transformation of reverse-stranded reads into forward-stranded ones, as follows.

0. Initialization)

```
INPUT: a set of forward- and reverse-stranded reads R, and a local alignment score distribution RS obtained from quasi-random reads.
Define an empty list C of “connected” reads (reads that have both a consistent and statistically significant overlap among each other), that is, $C=\emptyset$
Define a list V of “not yet connected and still to be visited” reads, that initially contains all the reads of R, that is, $V=R$.
```

1. Core algorithm

```
WHILE ( $C\neq\emptyset$ or $V\neq\emptyset$ )
    IF the list of connected reads is empty, that is, $C=\emptyset$
    THEN get and remove a read from V and put it into C, that is, C = {pop(V)}
    ELSE
        Get and remove the first read f from C, that is, f=pop(C)
        FOR EACH $c \in C$, compute all forward/forward local pairwise alignments between f and all the other reads $c \in C$.
            IF the overlap between f and c is consistent and statistically significant
            THEN set a link between f and c in the overlap graph (by setting the direction consistently with the heads and tails of the reads out of the local alignment)
        FOR EACH $v \in V$, compute all forward/forward and forward/reverse local pairwise alignments between f and all the reads $v \in V$
            IF the forward/forward alignment overlap between f and v is consistent and statistically significant
            THEN
                Remove v from V and add v to C, that is, remove (v,V) and push(v,C)
                Set a link between f and v in the overlap graph (by setting the direction consistently with the heads and tails of the reads out of the local alignment)
            ELSE
                IF the forward/reverse alignment overlap between f and v is consistent and statistically significant
                THEN
                    Remove v from V, reverse permanently v and add v to C, that is, remove(v,V), v=reverse(v) and push(v,C)
                    Set a link between f and v (now v is reversed) in the overlap graph (by setting the direction consistently with the heads and tails of the reads out of the local alignment)
```

2. Finalization

```
All the elements of C and V have been analyzed, that is, $C=\emptyset$ and $V=\emptyset$.
OUTPUT: the directed overlap graph of reads.
```

This algorithm can fail in some degenerate scenarios: for instance, when there is a read r_1 that has a consistent and significant overlap with another read r_a under a forward/forward alignment and with another read r_b under a forward/reverse alignment, and at the same time r_a and r_b have a consistent and significant overlap under a forward/forward alignment. However, such cases are likely to be rare, especially under NGS samples obtained by Roche 454 technology, which are characterized by a read length up to 450 bases.

In order to test the performance of this algorithm, a NGS sample of all forward-stranded reads, with the same settings as for simulation (i) described in the manuscript, was generated. By varying the initial *RS* distribution estimation via a selection of quasi-random sequences with different random seeds, we estimated for ten independent times the directed overlap graph. Then, 50% of the original reads was reverted and ten other overlap graphs were constructed by varying the initial conditions of the *RS* calculation and applying the above procedure whole forward-strand induction. Differences in the network edges were less than 5% across different runs, and probably more imputable to the *RS* distribution variation rather than to an algorithmic failure. In addition, no “clashes” were ever found when comparing two reads already put in the *C* list, that is, reads that had a consistent and significant overlap under a forward/reverse alignment, when instead both were supposed to be forward-stranded.