frontiers in
MICROBIOLOGY

REVIEW ARTICLE
published: 11 September 2012
doi: 10.3389/fmicb.2012.00329

Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data

***Niko Beerenwinkel*[1,2]*, *Huldrych F. Günthard*[3], *Volker Roth*[4] and *Karin J. Metzner*[3]**

[1] *Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland*
[2] *Swiss Institute of Bioinformatics, Basel, Switzerland*
[3] *Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, University of Zurich, Zurich, Switzerland*
[4] *Department of Mathematics and Computer Science, University of Basel, Basel, Switzerland*

Edited by:
Masaru Yokoyama, National Institute of Infectious Diseases, Japan

Reviewed by:
Masaru Yokoyama, National Institute of Infectious Diseases, Japan
Fabio Luciani, University of New South Wales, Australia

***Correspondence:**
Niko Beerenwinkel, Department of Biosystems Science and Engineering, ETH Zurich, WRO-1058 8.40, Mattenstrasse 26, 4058 Basel, Switzerland.
e-mail: niko.beerenwinkel@bsse.ethz.ch

Many viruses, including the clinically relevant RNA viruses HIV (human immunodeficiency virus) and HCV (hepatitis C virus), exist in large populations and display high genetic heterogeneity within and between infected hosts. Assessing intra-patient viral genetic diversity is essential for understanding the evolutionary dynamics of viruses, for designing effective vaccines, and for the success of antiviral therapy. Next-generation sequencing (NGS) technologies allow the rapid and cost-effective acquisition of thousands to millions of short DNA sequences from a single sample. However, this approach entails several challenges in experimental design and computational data analysis. Here, we review the entire process of inferring viral diversity from sample collection to computing measures of genetic diversity. We discuss sample preparation, including reverse transcription and amplification, and the effect of experimental conditions on diversity estimates due to *in vitro* base substitutions, insertions, deletions, and recombination. The use of different NGS platforms and their sequencing error profiles are compared in the context of various applications of diversity estimation, ranging from the detection of single nucleotide variants (SNVs) to the reconstruction of whole-genome haplotypes. We describe the statistical and computational challenges arising from these technical artifacts, and we review existing approaches, including available software, for their solution. Finally, we discuss open problems, and highlight successful biomedical applications and potential future clinical use of NGS to estimate viral diversity.

Keywords: next-generation sequencing, viral diversity, viral quasispecies, statistics, bioinformatics, haplotype inference, error correction, quasispecies assembly

INTRODUCTION

Many viruses, in particular RNA or single-stranded DNA viruses, exhibit extreme evolutionary dynamics. They have very high mutation rates, up to six orders of magnitude higher than in humans, short generation times, and large population sizes (Duffy et al., 2008). Under these conditions, genetic variants are produced constantly, and in each infected host, the virus population displays a high degree of genetic diversity. Rapidly evolving viruses are not only ideal systems for studying evolutionary mechanisms (Drummond et al., 2003), but many of them are significant pathogens of vital medical interest, including HIV, HCV, and Influenza (WHO, 2012).

Because of their diversity, intra-host virus populations are often referred to as mutant clouds, swarms, or viral quasispecies. The latter terms were originally introduced in the context of self-replicating macromolecules (Eigen, 1971; Eigen and Schuster, 1977) and have a precise mathematical meaning. A quasispecies is the equilibrium distribution of mutants in a mathematical model that accounts for mutation and selection (Eigen et al., 1988, 1989). In the framework of classical population genetics, it can be regarded as a coupled mutation-selection balance (Wilke, 2005). The main prediction of the quasispecies model is that selection acts on the population as a whole and hence the population dynamics cannot be understood from the fittest strain alone (Van Nimwegen et al., 1999; Wilke et al., 2001). The quasispecies model has later been applied to RNA viruses (Nowak, 1992; Domingo and Holland, 1997), hence the term viral quasispecies. The impact of the quasispecies model is not only due to its mathematical feasibility, but also its conceptual focus on the population as the target of natural selection (Burch and Chao, 2000).

The diversity of virus populations has repeatedly been shown to provide a selective advantage. For example, decreasing the mutation rate of poliovirus artificially, while maintaining its replication rate, resulted in reduced genomic diversity and in failure to adapt to adverse growth conditions (Vignuzzi et al., 2006). Similarly, pre-existing minority drug-resistant variants of HIV-1 have been shown to facilitate rapid viral adaptation leading to failure of antiretroviral therapy (Metzner et al., 2009; Li et al., 2011). In general, viral diversity is advantageous when the virus faces different selection pressures that need to be overcome by evolutionary escape (Iwasa et al., 2003, 2004). Changing selection pressures are common in the life of viruses, for example, after infecting a new host with a different immune response (Pybus and Rambaut, 2009), when infecting different cell types, while being exposed to different chemical agents, or due to changing multiplicity of infection (Ojosnegros et al., 2010). Understanding

and modeling the escape dynamics of these processes is of direct relevance for clinical and public health decisions.

With the introduction of next-generation sequencing (NGS) technologies, the experimental analysis of viral genetic diversity has changed dramatically. Rather than using labor-intensive limiting dilution and individual cloning of viruses followed by traditional Sanger sequencing, NGS now allows for sampling the virus population in a highly parallel fashion in a single experiment. However, the novel high-throughput approach has several pitfalls associated with both the experimental protocol and the statistical analysis of the data. We address both aspects in this review and discuss several successful applications of NGS to viral diversity studies, including drug resistance, immune escape, and epidemiology.

SAMPLE PREPARATION

The usefulness of NGS for viral diversity estimation depends crucially on the quality of the sample and on the procedure to prepare the sample. NGS sequence reads mirror the accumulation of errors, some of them preventable others unavoidable. To minimize the error rate, each step requires careful handling, starting with biological sample retrieval and storage up to the last steps of the NGS procedure itself (**Figure 1**).

Viral genomes are usually protected by the viral capsid and some of them additionally by an envelope, for instance, HIV and HCV. However, retrieval and storage conditions of biological specimens are especially important when studying RNA viruses due to the fragility of RNA (Holodniy et al., 1995; Jose et al., 2005), because degraded RNA will jeopardize all further steps of the analysis. Before starting the extraction of viral genomes, the viral load of the specimen should be considered. The final number of genome copies sequenced provides the basis for assessing viral diversity from the sequence reads (Metzner et al., 2003; Casbon et al., 2011). Low amounts might require a concentrating step, for instance, ultracentrifugation of plasma.

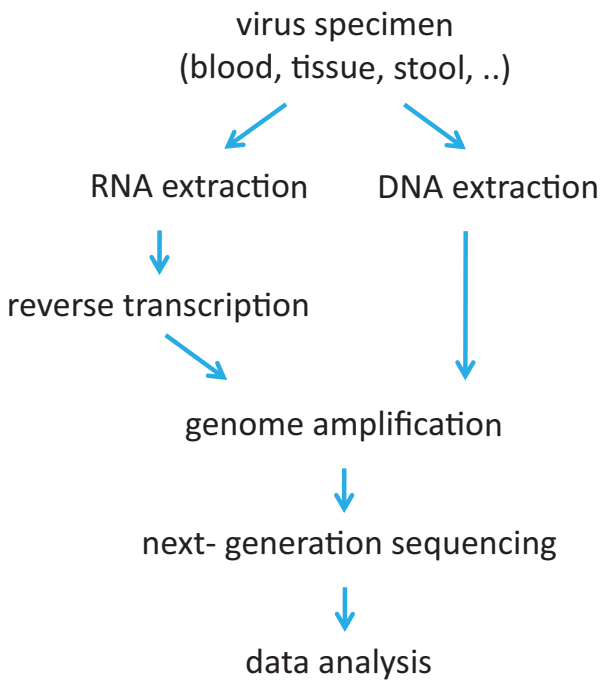

FIGURE 1 | Flow chart of sample processing for next-generation sequencing (NGS) of virus samples.

The choice of protocols used for genome extraction and elimination of contaminating RNA and DNA from other sources like host cells depends on the intended downstream procedures. Numerous kits are offered to extract viral DNA or RNA whose pros and cons will not be discussed here. A more critical point is the enrichment of viral genomes in the context of sample complexity. Three scenarios can be envisioned. (1) The virus is known and an amplicon approach is chosen for NGS. Here, the specificity of the primers might allow for amplifying the viral genome without any upstream enrichment. Nevertheless, it is often beneficial to eliminate contaminating DNA or RNA by DNase or RNase treatment. For instance, investigating HIV RNA genomes requires the elimination of proviral DNA genomes (Fischer et al., 2002). (2) The virus is known, but a random approach is chosen for NGS. Due to the high heterogeneity of some viruses, it might be disadvantageous to use virus-specific primers for amplification due to potential primer bias or even complete failure of amplification (Metzner et al., 2003). In contrast, any random approach, including amplification using degenerated or random primers as well as non-specific adaptor ligation and subsequent amplification using adaptor-specific primers, cannot differentiate between the viral genome and any other nucleic acid (Reyes and Kim, 1991; Chang et al., 1992). Thus, the elimination of contaminating nucleic acids is mandatory when a high coverage of viral genomes is required, as for studying diversity, since the viral genomes represent only a low-abundant fraction in almost all biological specimens (Daly et al., 2011). DNase and RNase treatment, filtration, density gradient centrifugation, and their combinations are commonly used procedures. Enrichment strategies based on hybridization capture might also be suitable (Turner et al., 2009; Althaus et al., 2012) and, potentially, freeze thaw nuclease digestion protocols may also be beneficial to minimize contaminating RNA or DNA (Fischer et al., 2002) (3). The virus in unknown, therefore, random approaches have to be applied. The enrichment of viral genomes is an even greater challenge in this set-up. In this review, we focus on estimating viral diversity from NGS data, a second step after virus discovery (Lipkin, 2010).

After viral genome extraction, an amplification procedure has to be performed, because the current NGS technologies require a high input DNA amount and the viral genome amount is several orders of magnitude lower. Furthermore, RNA genomes have to be reverse transcribed prior to PCR. Every amplification process introduces errors. Reverse transcriptases (RTs) are error-prone enzymes, because of the lack of any proof-reading activity (Preston et al., 1988; Roberts et al., 1988). Some RTs are less error-prone than others, but, in general, RT errors are unavoidable and very difficult to distinguish from real mutations since they are introduced in the first step of amplification. Another important but often ignored problem with reverse transcription is that short, incomplete cDNA fragments can act as primers in subsequent

PCRs and lead to *in vitro* recombination. This phenomenon has been considered only for RT-PCRs amplifying several kilobases (kb) long fragments (Fang et al., 1998). We have recently shown that this effect also occurs very frequently when amplifying short cDNA fragments of a size of only 0.6 kb and can be minimized by using an RNaseH-negative RT (Di Giallonardo et al., submitted).

Four main types of errors can occur during PCR and are relevant for NGS data: (i) biased amplification due to primer mismatches, (ii) *in vitro* recombination due to premature termination of strand elongation and subsequent false hybridization of short DNA fragments acting as primers or, less frequently, due to template switching, (iii) nucleotide misincorporation due to the inaccuracy of DNA polymerases, and (iv) resampling due to, for instance, too low amounts of input DNA copies (Eckert and Kunkel, 1991; Liu et al., 1996; Kanagawa, 2003). Several precautions can be taken to minimize these errors. Primer mismatches can be diminished by choosing primer binding sites in conserved regions of the viral genome or by using degenerated primers. Chimera formation can be reduced by several improvements of PCR conditions such as increasing the elongation time, decreasing the number of cycles, and deleting the final extension step (Meyerhans et al., 1990; Judo et al., 1998). Nucleotide misincorporation can be lowered by using high-fidelity DNA polymerases, and resampling can be reduced, for instance, by optimizing the input copy number. Even when applying all these precautions, it is currently not possible to completely avoid these PCR errors. Furthermore, the discrimination between artificial and real viral variants can be very difficult if not impossible. One possibility is to perform several independent PCRs assuming that most of the errors occur randomly with regard to the sequence position and the timing of the error, i.e., in which PCR cycle the error occurs, resulting in different variants of different frequencies in the replicates. A recently described method uses primer identifiers (IDs) to uniquely label each cDNA molecule (Jabara et al., 2011). This is an elegant procedure to reduce or even eliminate PCR errors, although errors induced during the reverse transcription cannot be addressed in this manner. In addition, the method is only applicable to amplicon-based approaches and a high number of sequence reads are required to obtain a sufficient number of consensus sequences, each of which has to be derived from at least three reads with the same primer ID. Thus, all unique or twice occurring reads, which represent the majority of sequence reads, cannot be considered in the analysis.

Overall, sample preparation is a critical issue in the process of NGS. If unrecognized, errors during sample preparation can lead to an artificially increased diversity of the investigated virus population. To avoid such misinterpretation, the pitfalls of sample preparation need to be identified and properly addressed.

NEXT-GENERATION SEQUENCING

In the last decade, many NGS technologies have been developed and several are commercially available today or about to become available in the near future (Mardis, 2008b; Metzker, 2010). Due to its massively parallel approach, NGS allows for generating much larger volumes of sequencing data in a cost-effective manner as compared to conventional sequencing methods. The increase in throughput has been so far-reaching that NGS is considered revolutionary, because it facilitates many new sequencing applications that had been out of reach (Mardis, 2008a; Schuster, 2008). One of these novel applications is the inference of viral genetic diversity from a single deep-coverage NGS experiment.

All NGS technologies involve the steps of template preparation, sequencing, and imaging, followed by data analysis, but they differ in the realization of each step. 454/Roche pyrosequencing has been the first NGS method commercially available and until today it is the most commonly used technology for the analysis of viruses (Margulies et al., 2005). For pyrosequencing, DNA is isolated, amplified and/or fragmented, adaptor-annealed, and amplified on beads in a micro-droplet emulsion PCR. DNA and beads have to be used in a ratio allowing the hybridization of only one DNA molecule to one bead, i.e., the majority of beads do not contain any DNA molecule. Thus, on each DNA-hybridized bead, a single template gives rise to several thousand copies. These beads are separated from the empty beads and loaded into 1.6 million wells of a picotiter plate, one bead per well, and enzymes for pyrophosphate sequencing are added. Sequencing by synthesis proceeds by adding the four bases in a cyclic order. In each cycle, the light emission associated with base incorporation is detected and remaining chemicals are washed out. The intensity of the light signal is approximately proportional to the number of nucleotides that have been incorporated. All generated signals are recorded as a series of peaks, called a flowgram, from which DNA bases are eventually called (Margulies et al., 2005).

The Illumina Genome Analyzer and HiSeq systems are currently dominating the NGS market (Bentley et al., 2008). Rather than emulsion PCR, Illumina relies on solid-phase amplification, which consists of initial priming and extending of single-stranded templates, followed by bridge amplification of each immobilized template with adjacent primers. In multiple cycles of annealing, extension, and denaturation, around 200 million molecular clusters are formed. For sequencing, all four nucleotides are added simultaneously. Each nucleotide is labeled with a different dye and they are modified to terminate DNA synthesis after incorporation. Color imaging is used to detect the incorporated nucleotide. In a cleavage step, the fluorescent dye is removed and termination is reversed by regenerating the 3′-OH group. Bases are called from the resulting four-color images.

We focus here on the 454/Roche and Illumina platforms, because the vast majority of reported virus sequencing applications have used these systems, but several other technologies can, and are likely to, be used as well, including ABI SOLiD, Ion Torrent, PacBio RS, and Polonator. The technical details in which platforms differ can have important consequences for their applicability to viral sequencing studies. Among other aspects, NGS platforms differ in throughput, runtime, costs, read lengths, and error patterns (Metzker, 2010). The currently most powerful 454/Roche sequencer GS FLX Titanium XL+ can produce up to 1 million reads per run of 700 bp average length, while Illumina's largest machine, HiSeq 2500, can generate up to 1.2 billion paired-end reads of 2 × 150 bp length. Both companies also offer smaller benchtop devices of their platforms that may be preferable in certain diagnostic and clinical settings. The Roche/454 Junior produces up to 100,000 reads of 400 bp average length in a single

10-h run, and the Illumina MiSeq generates up to 30 million paired-end reads of 2 × 150 bp length in 24 h. Thus, longer reads can be produced with the 454/Roche technology, but ultra-deep coverage is easier to obtain with Illumina (Loman et al., 2012).

In addition to the various errors that can occur during sample preparation, as discussed in "Sample Preparation", all NGS platforms introduce sequencing errors. With 454/Roche pyrosequencing, insertions and deletions (indels) are the most common type of errors. They occur predominantly in homopolymeric regions of the target sequence, where the linear relationship between signal intensity and number of incorporated nucleotides starts to fail. Remaining nucleotides after washing can give rise to insertions or carry forward errors, while deletion errors can result from incomplete extension (Margulies et al., 2005; Balzer et al., 2011). The error rate has been shown to increase with read length and to depend on several other biological and technical factors, including the organism and genomic region to be analyzed and the position on the picotiter plate with respect to the flow of chemicals and the position of the camera (Gilles et al., 2011).

Illumina reads are not as susceptible to indel errors in homopolymeric regions, but artificial indels outside these regions and substitutions have similar frequencies (Archer et al., 2012). The Illumina mismatch rate also increases with read length and it further depends on the sequence context and the substitution type (Dohm et al., 2008; Kircher et al., 2009; Nakamura et al., 2011). Illumina reads are generated in forward and reverse direction, and errors predominantly occur on one of the two strands (Chapman et al., 2011; Varela et al., 2011). All NGS platforms report quality scores, defined as $Q = -10 \log_{10} p$, where p is the error probability (Ewing and Green, 1998), together with the called bases, but the calibration of these scores is challenging (Brockman et al., 2008; Kircher et al., 2009) and there is no consensus on how to compare scores across platforms.

Besides errors, the distribution of reads along the genome is critical for diversity estimation, especially if phasing of genetic variants is the goal. However, uniform coverage is difficult to achieve and, in practice, the read coverage often varies by orders of magnitude. The reasons for this variation are poorly understood, but for Illumina, the GC content of the target sequence is an important factor (Dohm et al., 2008). Uniform coverage is feasible within short segments by using a single amplicon. However, increasing the number of amplicons to cover longer segments can impair this uniformity, and shot-gun approaches introduce even more variation. For 454/Roche, Illumina, and ABI SOLiD, correlation of coverage and errors is fairly weak among the three different NGS platforms (Harismendy et al., 2009). Thus, for viral diversity estimation, where uniform coverage and error correction are critical, complementary sequencing strategies involving more than one platform may be more efficient than increasing the coverage on a single platform.

The large amounts of viral sequencing data obtained by NGS place substantial demands on information technology and computational data analysis in terms of storage, quality control, mapping, error correction, single nucleotide variant (SNV) calling, haplotype reconstruction, diversity estimation, and data integration (Pop and Salzberg, 2008; Vrancken et al., 2010; Barzon et al., 2011; Beerenwinkel and Zagordi, 2011). Data analysis usually starts by removing reads of exceptionally low quality. The rationale for this initial filtering step is that low-quality reads contribute disproportionally to the overall error rate, i.e., most errors occur on a few reads (Huse et al., 2007). Filtering can be based on quality scores or on properties of the read or the target sequence known to affect error rates, as discussed above. Optimized filtering has been shown to reduce the error rate in detecting genomic variation up to 300-fold (Reumers et al., 2011).

After filtering, the next step is to align the remaining reads. In re-sequencing studies of known viruses, this is typically done by mapping reads individually to a reference sequence and then aggregating the pairwise alignments into a multiple sequence alignment (MSA). For read mapping, local alignment using dynamic programming may be applied (Wang et al., 2007; Zagordi et al., 2011), but for larger data sets, efficient short read mappers are required. Several efficient mapping algorithms based on indexing techniques are available. Some of them can handle gaps, account for quality scores, and have a paired ends option (Trapnell and Salzberg, 2009; Wikipedia, 2012). In coding regions, a major goal of the alignment step is to identify indels that cause frameshifts. These alterations are likely to be sequencing errors, which are frequently observed using the 454/Roche platform. Hence, they are usually removed, but this bears the risk of losing virus variants harboring real indels. For correcting indel errors, a high-quality alignment is necessary, but in mixed samples, the use of a reference sequence can be suboptimal if reads originating from some subpopulations align only poorly to the reference sequence. To address this concern, a MSA may be computed directly, for example, by using a progressive MSA strategy that takes into account the approximate location of reads on the genome (Saeed et al., 2009). Similarly, for the HIV *env* gene, a multi-step procedure has been proposed, in which reads are located efficiently on a reference sequence by k-mer matching and MSAs are built locally in windows of width 70 nucleotides along the genome. From all local MSAs, in-frame consensus sequences are generated and concatenated. Finally, the reads are re-aligned to the global consensus sequence and all indels causing frameshifts are removed. Using the consensus rather than a reference sequence was shown to improve the alignment quality, especially if their divergence is high (Archer et al., 2010).

LOCAL DIVERSITY ESTIMATION

From the aligned reads, one wants to reconstruct the original virus population in the sample, meaning the composition and relative frequencies of all individual viral genomes, also referred to as strains or haplotypes. Even after filtering and removal of frameshift-causing indels, many reads are still erroneous. Therefore, in mixed samples, error correction and haplotype inference are intrinsically tied to each other and, in fact, addressed jointly by most computational methods. This is in contrast to the simpler task of error correction in clonal samples, where implausible variants can easily be discarded using either k-mers, suffix trees/arrays, or MSA (Yang et al., 2012).

The haplotype inference problem occurs at different spatial scales depending on the length of the genomic region to be analyzed for diversity (**Figure 2**). When only a single genomic site

is considered, diversity estimation means detecting SNVs. Local haplotype inference refers to analyzing windows in the MSA that are covered entirely by reads. Finally, global haplotype inference, also called quasispecies assembly, involves a jigsaw puzzling step of assembling local fragments into multiple haplotype sequences that span the entire genomic region of interest.

SNV calling is based on the observed nucleotide counts at a single sequence position. The simplest statistical model for separating errors from true variations is to assume that, at each genomic site, the number of errors follows the same Poisson distribution and to call SNVs that occur more often than expected by chance for a given error rate (Wang et al., 2007). This approach has been extended to account for site-specific error rates (Macalalad et al., 2012). The power and accuracy of SNV calling can be increased substantially by a control experiment, in which the same genomic region is sequenced from a clonal sample under conditions as similar as possible to those for the mixed sample. The rationale for this comparative sequencing approach is that the control experiment allows for estimating the specific error patterns of the experiment and hence for improved separation of biological signal from technical noise. In this setting, SNV detection is based on comparing nucleotide counts between two experiments, for example, using Fisher's exact test (Koboldt et al., 2012). Assuming independent Poisson distributions, another test is based on the difference of the number of observed nucleotides (Altmann et al., 2011). Count data from NGS experiments have repeatedly been shown to display more variation across sites than is captured by a binomial distribution, and the beta-binomial distribution is a popular choice for such overdispersed data (Flaherty et al., 2012; Gerstung et al., 2012). Based on this model and accounting for the strand-bias of sequencing errors, a sensitivity of up to 1/10,000 has been achieved for SNV calling at a coverage of around 10^5 (Gerstung et al., 2012).

By dropping the assumption of independence among sites, SNV calling can be further improved. Considering the number of joint sequencing errors at two positions has been shown to significantly decrease the minimal frequency at which a variant is detectable (Macalalad et al., 2012). This phasing of two SNVs is possible only at a distance smaller than the maximal read length. For small distances, the SNV pair will be covered by many reads, but for larger distances the benefit of phasing will be undone by the loss of joint coverage. In fact, for deep coverage, pairs are more informative than single sites only if their distance is not larger than the average read length (Macalalad et al., 2012).

The idea of phasing SNVs is further extended by comparing entire reads within a sequence window they overlap. The size of the window is subject to the same trade-off as the distance between two SNVs discussed above: Small windows contain many reads but few SNVs for robust pairwise comparisons of reads, while large windows contain less reads but more segregating sites. Local haplotype inference is based on clustering reads within a given window (**Figure 3**). The rationale for clustering is that reads originating from the same haplotype should be more similar to each other than to reads from other haplotypes. This assumption is only valid if the error rate is low relative to the diversity of the population, and the ability to identify haplotype clusters increases with coverage (Eriksson et al., 2008).

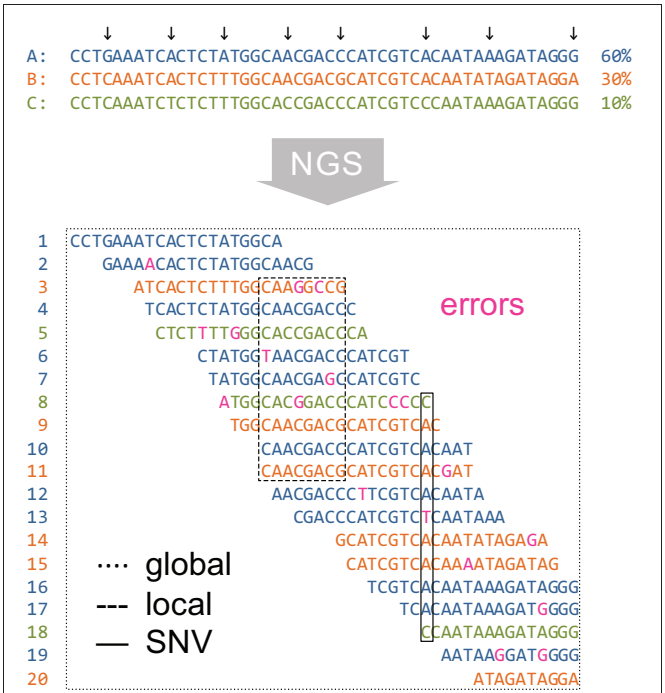

FIGURE 2 | Spatial scales of diversity estimation from NGS data. In this example, it is assumed that the true virus population (top of figure) consists of three haplotypes of relative frequencies 60% (A, blue), 30% (B, orange), and 10% (C, green). Segregating sites are indicated by arrows. Twenty short reads (labeled 1 through 20) are generated by NGS from the virus population subject to sequencing errors (indicated in magenta). Reads are displayed in a MSA and in the color of their corresponding parental haplotype. Diversity estimation can be approached at single sites (SNV detection, solid-line rectangle), in windows of the MSA (local haplotype inference, dashed-line rectangle), or over the entire genomic region (global haplotype reconstruction, dotted-line rectangle).

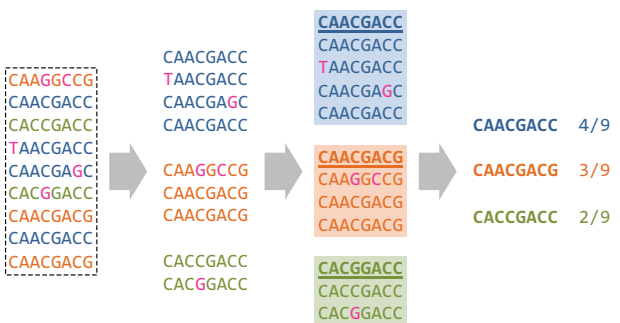

FIGURE 3 | Local read clustering. The local window of the MSA displayed in **Figure 2** is considered (dashed-line rectangle), with colors defined as in **Figure 2**. Reads that are more similar to each other than to other reads are grouped together which recovers the three original haplotypes A, B, and C of **Figure 2** as indicated by the three different colors. Each cluster center sequence is a predicted haplotype (bold, underlined) and the size of its corresponding cluster is an estimate of the frequency of the haplotype (here, 4/f/9, and 2/9, respectively).

Clustering was initially performed using the classical k-means algorithm (Jain and Dubes, 1981) and later formulated probabilistically and solved in a Bayesian fashion (Eriksson et al., 2008; Zagordi et al., 2010a). In particular, the latter approach allows for estimating the error rate and the number of clusters from the data—a notoriously difficult problem with any clustering method. The cluster centers are the predicted haplotypes and the cluster sizes are interpreted as the haplotype frequencies in the population. Error correction is based on a local read clustering solution by replacing all read bases with those of its cluster center (**Figure 3**). This method has been shown to reduce the per-base error rate after correction, to increase the sensitivity and specificity of local haplotype calling, and to improve the estimation of haplotype frequencies as compared to simple read counting or k-means clustering (Zagordi et al., 2010b). For the 454/Roche platform, a similar clustering approach called AmpliconNoise can be applied before base calling on the flowgrams (Quince et al., 2009, 2011). Here, the observed flowgrams are obtained from ideal flowgrams corresponding to read sequences subject to measurement noise. Whether clustering is based on sequences or on flowgrams, the distance measure between reads should reflect the pattern of experimental noise.

As an alternative to clustering, k-mer-based error correction, implemented in the program KEC, has been proposed for viral amplicon sequencing (Skums et al., 2012). This approach extends the EDAR error correction algorithm (Zhao et al., 2010) and initially does not require a read alignment. It consists of a number of heuristic steps with the goal of locating error regions in reads by considering rare k-mers and removing errors in these regions. In a final step, which eventually involves MSAs of the corrected reads, local haplotypes are reconstructed.

GLOBAL DIVERSITY ESTIMATION

The local methods discussed in the previous section focus on reconstructing haplotypes in a local window, the maximum size of which is effectively restricted to the average length of the reads. The global reconstruction problem, on the other hand, is defined as the genome-wide assembly of quasispecies, irrespective of machine-specific parameters like the average read length. The various approaches to solving this jigsaw puzzle described in the literature can be roughly divided into three groups: (1) graph-based methods that first aggregate the reads in a read graph and then search for a minimum set of paths through this graph, (2) probabilistic clustering models based on mixture models, and (3) *de novo* assembly methods which do not rely on the availability of a reference sequence.

Read graph-based global haplotype reconstruction consists in aggregating the reads in a read graph and subsequently identifying haplotypes as paths in this graph. The concept of a read graph has been independently introduced by Eriksson et al. (2008) and Westbrooks et al. (2008). The read graph contains the possibly pre-processed, for instance, locally error-corrected, reads as nodes. Directed edges connect two nodes when the reads agree on their non-empty overlap (**Figure 4**). The direction of the edge reflects the order of the starting positions on the reference sequence. The set of nodes is restricted to all irredundant reads, where a read is considered redundant if there is another read that overlaps completely and if both reads agree on this overlap. In a similar manner, the set of edges is restricted to include only those edges for which there would be no path between the corresponding nodes without this edge. The latter restriction is called transductive reduction in (Westbrooks et al., 2008), and it has been shown that this reduction can be computed efficiently. Finally, a source and a sink node are added to the graph, along with edges connecting all reads starting at the first position to the source and all reads ending at the last position to the sink (**Figure 4**).

Every path in the read graph connecting source and sink is a potential haplotype, and the problem of estimating the haplotypes present in a certain sample might be restated as finding a set of such source-sink paths that explains the reads well. Different formalizations of this problem lead to different optimization problems. One example is the search for the minimum set of paths that covers all reads implemented in ShoRAH (Eriksson et al., 2008; Zagordi et al., 2011). The same problem has been studied in a different way as a network flow problem (Westbrooks et al., 2008). A variant of the network flow formulation is the search for a set of haplotypes covering all reads with minimum costs (Westbrooks et al., 2008) and, in a slightly different

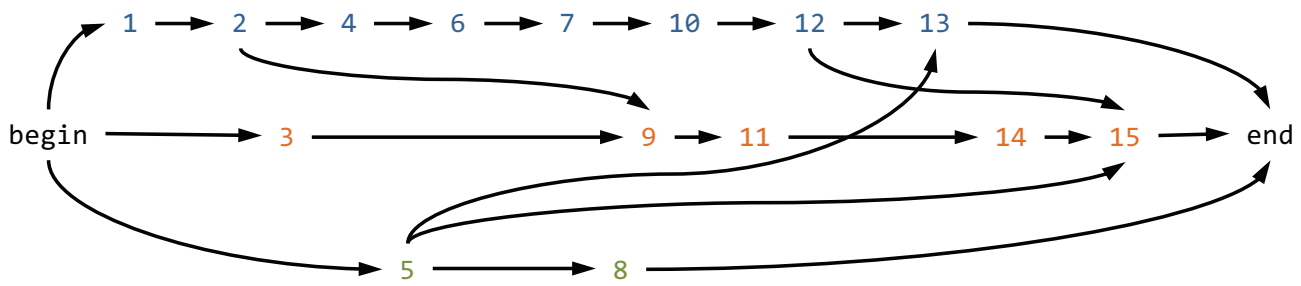

FIGURE 4 | Read graph-based global haplotype reconstruction. Shown is the read graph for the first 15 reads of the MSA shown in **Figure 2**. Each read is represented by its index and colored according to its parental haplotype (A, blue, first row; B, orange, second row; and C, green, third row). Reads are connected by a direct edge if they agree on their non-empty overlap. Each path from the begin node to the end node represents a potential global haplotype, but there are more paths in the graph than the original three haplotypes the reads have been derived from.

fashion relaxing the requirement of a complete read cover, implemented in ViSpA (Astrovskaya et al., 2011). The combinatorial reconstruction is followed by frequency estimation using an Expectation Maximization (EM) algorithm (Eriksson et al., 2008; Westbrooks et al., 2008; Astrovskaya et al., 2011).

In a related approach termed QuRe, the same read graph idea is used to find a set of consistent quasispecies explaining the reads (Prosperi et al., 2011; Prosperi and Salemi, 2012). It differs from the methods above in the optimization procedure for finding the quasispecies. This is formalized as minimizing the number of *in silico* recombinants instead of finding a path cover explaining the reads. However, both optimization strategies are similar in nature, since avoiding *in silico* recombinants can be regarded as avoiding redundant paths in the read graph. Another advantage of QuRe is that it explicitly addresses the blockwise structure of the reads due to amplicon-based sequencing in the statistical analysis (Prosperi et al., 2011; Prosperi and Salemi, 2012).

Haplotype assembly based on amplicon sequencing is also addressed by the BIOA software (Mancuso et al., 2011). Here, a read graph-based framework is proposed that includes balancing of haplotype frequencies between neighboring amplicons followed by quasispecies reconstruction using a maximum bandwidth approach or a greedy algorithm. In the assembly step, the parsimony criterion of explaining the observed reads with a minimal number of haplotypes is relaxed to finding a quasispecies of minimal entropy explaining the reads. This strategy was shown to outperform shotgun-based quasispecies assembly using ViSpA.

QColors is another method that relies on the read graph as the main source of information for assembling reads into haplotypes, but it uses in addition a conflict graph consisting of edges between reads that overlap but disagree on the overlap (Huang et al., 2011). The reconstruction problem is then to find a partition of the reads into a minimal number of non-conflicting subsets, which defines a vertex graph coloring problem, hence the name QColors. A potential problem with this approach might be the sensitivity of the conflict graph to sequencing errors and the uncertainty in placing alignment gaps, which are not explicitly dealt with.

Another method that uses the read graph approach is called Hapler (O'Neil and Emrich, 2012). This method is specifically designed for situations characterized by low haplotype diversity and low read coverage ($<25\times$), which, for instance, occur in the context of population-level *de novo* transcriptome assemblies or ecological studies. The minimum path cover problem is generalized and reformulated as a weighted bipartite graph matching problem, such that erroneous reads can be identified. Since, in general, the resulting path covers are again not unique, the analysis is equipped with a randomization step in which samples are drawn from the set of path covers, although this process seems to lack a clear probabilistic interpretation. Experiments under low-coverage conditions indicate that this method is successful in reconstructing local haplotypes over a region that is roughly determined by the average read length, which in our terminology would be classified as local reconstruction. Nevertheless, longer haplotype assemblies are possible with Hapler and specific care is taken in reconstructing consensus sequences with a minimal number of chimeric points.

A common property of all read graph-based approaches is that the haplotype reconstruction problem itself becomes deterministic in nature, while the unavoidable noise component present in observed reads is dealt with in a pre-processing error correction step—if at all.

Removing all the stochasticity in the observed reads by way of local error correction prior to global haplotype reconstruction has the limitation that corrections cannot be revised in the global context and miscorrections are propagated through subsequent steps. A probabilistic hierarchical model that circumvents this problem has been introduced (Jojic et al., 2008). The main idea is to model the generative stochastic process of read generation. Parameters and hidden variables in this method include the parental haplotype, the starting position, and the parameters related to the error transformation. Inference is carried out by maximizing the likelihood using the EM algorithm. A potential drawback of this approach is that the user has to fix the number of haplotypes to be reconstructed in advance, and no well-defined estimation process for this number is provided.

Probabilistic approaches are a second methodology for global haplotype reconstruction. PredictHaplo is one of these approaches which also automatically adjusts the number of haplotypes (Prabhakaran et al., 2010). In this model, a haplotype is represented as a set of position-specific probability tables over the four nucleotides, which can be augmented to include a fifth character representing alignment gaps (**Figure 5**). The underlying generative model assumes that reads are sampled from a mixture model, where each mixture component is interpreted as a haplotype, and the associated mixing proportion estimates the haplotype frequency. In order to avoid a priori specification of the number of mixture components, an infinite mixture model is employed (Ewens, 1972; Ferguson, 1973; Rasmussen, 2000), and for computational reasons, a truncated approximation of this stochastic process is used.

A further refinement of probabilistic haplotype reconstruction has been implemented in the program QuasiRecomb (Zagordi et al., 2012). Here, haplotypes are not reconstructed individually, but rather their distribution is estimated by a hidden Markov model. The model assumes that all haplotypes are generated from a small set of sequences by mutation and recombination. This model is taking into account that in some RNA viruses, such as HIV, recombination is very frequent and hence an important factor generating genetic diversity.

All approaches described so far make use of a known reference genome that serves as a fixed spatial coordinate system after read alignment. By contrast, *de novo* assembly methods are more general in nature since they do not require such reference genomes. Several assemblers specifically designed for certain NGS platforms like 454/Roche have been proposed in recent years (Finotello et al., 2012). The original goal of *de novo* assembly is reconstructing a single target genome sequence, rather than an ensemble of different genomes. Hence, the currently available genome assemblers are not designed to solve the whole-genome quasispecies assembly problem, but the different contigs they reconstruct may

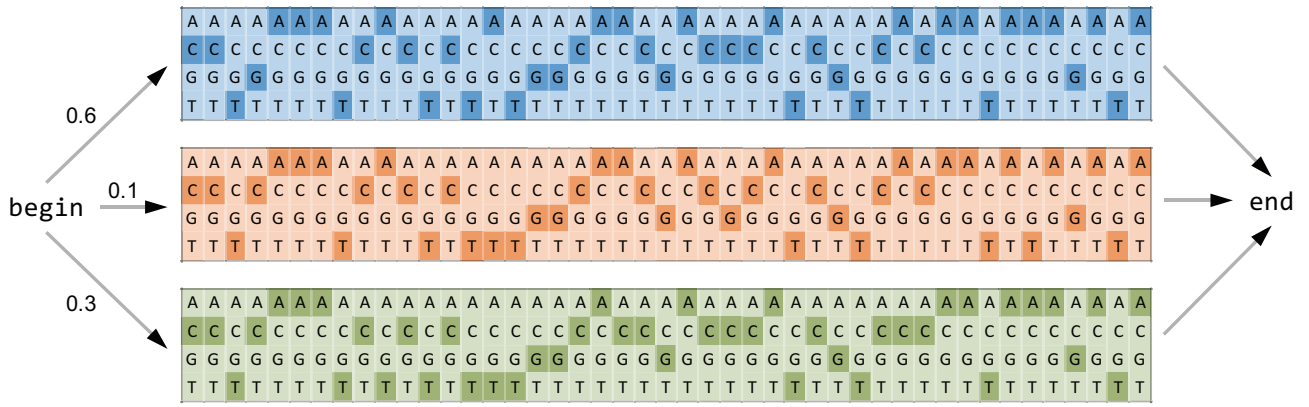

FIGURE 5 | Probabilistic global haplotype reconstruction using a generative mixture model. Each of the three haplotypes colored as in **Figure 2** (A, blue; B, orange; and C, green) is represented as a chain of probability tables over the four nucleotides, where darker shading of a base indicates higher probability. The probabilities of traversing from the begin node to one of the haplotypes serve as an estimate for the haplotype frequencies. Each read is regarded as an independent observation from this statistical model.

serve as a starting point for this jigsaw puzzle (Ramakrishnan et al., 2009).

Large-scale simulation studies show that all global reconstruction methods rely on the availability of relatively long reads. Coverage is also important when it comes to detecting low-abundant mutants, but even an arbitrarily high coverage cannot compensate for insufficient overlaps due to short reads. Given the typical diversity of virus populations, it appears that global haplotype reconstruction is currently only realistic for sequencing platforms producing long reads on the order of at least 300–500 bp. Accordingly, successful reconstructions have been reported predominantly for the 454/Roche sequencing platform.

Regarding the different computational approaches described above, it is generally difficult to conduct informative comparative simulation experiments, but two general trends have become evident. First, local read error correction has the tendency to under-correct the reads, which can lead to a large number of false positive global haplotypes, in particular, when combined with read graph approaches requiring a complete coverage of all reads. Quasispecies assembly methods that relax this coverage requirement (Astrovskaya et al., 2011; O'Neil and Emrich, 2012) or probabilistic approaches avoiding the read-graph construction (Jojic et al., 2008; Prabhakaran et al., 2010) are successful in decreasing the false positive rate. Second, the most problematic step in genome-wide reconstruction is the usually unavoidable (RT-)PCR pre-processing which can introduce significant artifacts. These artifacts might have a much stronger effect on the final quality of the haplotype reconstruction than the actual choice of the computational reconstruction method.

Computational methods for local and global haplotype reconstruction are summarized in **Table 1**. All of these tools have been developed in research environments and most are subject to continuous enhancements. Their usability and performance also depends on the quickly changing characteristics of the sequencing machines. In the future, comparative studies using simulated data, mixed control samples, or Sanger-sequenced gold standard samples are required to assess the performance of these tools under different conditions. In addition, software tools are available for NGS read data management and visualization. For example, Segminator II has been specifically designed to display sequence variability of temporally sampled virus populations (Archer et al., 2012).

Table 1 | Available software tools for viral quasispecies inference.

Program	Method	URL	References
QuRe	Read graph	https://sourceforge.net/projects/qure/	Prosperi and Salemi, 2012
ShoRAH	Read graph	http://www.cbg.ethz.ch/software/shorah	Zagordi et al., 2011
ViSpA	Read graph	http://alla.cs.gsu.edu/~software/VISPA/vispa.html	Astrovskaya et al., 2011
BIOA	Read graph	https://bitbucket.org/nmancuso/bioa/	Mancuso et al., 2011
Hapler	Read graph	http://nd.edu/~biocmp/hapler/	O'Neil and Emrich, 2012
AmpliconNoise	Probabilistic	http://code.google.com/p/ampliconnoise	Quince et al., 2011
PredictHaplo	Probabilistic	http://www.cs.unibas.ch/personen/roth_volker/HivHaploTyper	Prabhakaran et al., 2010
QuasiRecomb	Probabilistic	http://www.cbg.ethz.ch/software/quasirecomb	Zagordi et al., 2012

Table 2 | Applications of 454/Roche pyrosequencing and Illumina NGS technologies in clinical virology.

Virus	Study	NGS platform	NGS approach	Basis of analysis	References
CMV	Epidemiology	454/Roche	Amplicon-based	Reads	Gorzer et al., 2010
CMV	Epidemiology	454/Roche	Shotgun	Consensus sequence	Jung et al., 2011
EBV	Epidemiology	Illumina	Shotgun	SNV, consensus sequence	Liu et al., 2011
EBV	Epidemiology	Illumina	Shotgun (amplicons)	SNV	Kwok et al., 2012
HBV	Drug resistance	454/Roche	Amplicon-based	Reads, SNV	Solmone et al., 2009; Homs et al., 2011; Rodriguez-Frías et al., 2012
HBV	Drug resistance	454/Roche	Amplicon-based	SNV	Margeridon-Thermet et al., 2009; Ko et al., 2012; Sede et al., 2012
HBV	Drug resistance	Illumina	Shotgun	SNV	Nishijima et al., 2012
HCV	Drug resistance	454/Roche	Amplicon-based	Reads	Bolcic et al., 2012; Fonseca-Coronado et al., 2012
HCV	Drug resistance	Illumina	Shotgun (cDNA)	SNV	Hiraga et al., 2011
HCV	Drug resistance	454/Roche	Shotgun (amplicons)	SNV, consensus sequences	Lauck et al., 2012
HCV	Drug resistance	Illumina	Paired-end (amplicons)	SNV	Nasu et al., 2011
HCV	Drug resistance	454/Roche	Amplicon-based	SNV	Powdrill et al., 2011
HCV	Epidemiology	454/Roche	Amplicon-based	Reads	Escobar-Gutiérrez et al., 2012; Forbi et al., 2012
HCV	Epidemiology	Illumina	Shotgun (cDNA)	SNV, consensus sequences	Ninomiya et al., 2012
HIV	Drug resistance	454/Roche	Amplicon-based	SNV	Hoffmann et al., 2007; Wang et al., 2007; Mitsuya et al., 2008; Le et al., 2009; Simen et al., 2009; Varghese et al., 2009; Lataillade et al., 2010, 2012; Alteri et al., 2011; D'Aquila et al., 2011; Delobel et al., 2011; Gianella et al., 2011; Ji et al., 2011; Kozal et al., 2011; Moorthy et al., 2011; Stelzl et al., 2011; Fisher et al., 2012; Messiaen et al., 2012
HIV	Drug resistance	454/Roche	Amplicon-based	Reads, SNV	Hedskog et al., 2010; Ji et al., 2010; Mild et al., 2011; Mukherjee et al., 2011; Armenia et al., 2012
HIV	Epidemiology	454/Roche	Shotgun (amplicons)	Consensus sequence	Bruselles et al., 2009
HIV	Epidemiology	454/Roche	Amplicon-based	Consensus sequence	Eshleman et al., 2011
HIV	Epidemiology	454/Roche	Amplicon-based	Reads	Redd et al., 2012
HIV	Tropism	454/Roche	Amplicon-based	Reads	Archer et al., 2009; Rozera et al., 2009; Abbate et al., 2011; Swenson et al., 2010; Vandenbroucke et al., 2010; Baatz et al., 2011; Bunnik et al., 2011; Raymond et al., 2011; Saliou et al., 2011; Svicher et al., 2011; Swenson et al., 2011a,b; Vandekerckhove et al., 2011
Influenza A virus	Epidemiology	Illumina	Shotgun (amplicons)	SNV	Kuroda et al., 2010; Kampmann et al., 2011
Influenza A virus	Epidemiology	454/Roche	Shotgun (amplicons)	SNV	Bartolini et al., 2011
Influenza A virus	Epidemiology	454/Roche	Shotgun	Reads	Lorusso et al., 2011
norovirus	Epidemiology	454/Roche	Shotgun (amplicons)	SNV, haplotype recon-struction	Bull et al., 2012
rhinovirus	Epidemiology	Illumina	Shotgun (amplicons)	SNV, consensus sequences	Tapparel et al., 2011
rotavirus	Epidemiology	454/Roche	Shotgun (cDNA)	Consensus sequences	Jere et al., 2011
VZV	Epidemiology	454/Roche	Shotgun (amplicons)	Consensus sequences	Zell et al., 2012

BAL, bronchoalveolar lavage; CMV, cytomegalovirus; EBV, Epstein Barr virus; HBV, hepatitis B virus; HCV, hepatitis C virus; HIV, human immunodeficiency virus; SNV, single nucleotide variant; VZV, varicella zoster virus.

APPLICATIONS

NGS is widely applied to study viral diversity mainly in the context of drug resistance of clinically relevant viruses such as HIV, HCV, and HBV (**Table 2**). Most studies focus on pre-existing minority drug-resistant virus variants in treatment-naïve individuals and their impact on the success of antiviral therapy, epidemiological surveillance, and virus population dynamics during virological failure. The pathways of drug resistance development are of particular clinical importance, since they can lead to new drug design or new therapeutic strategies, for instance, avoiding cross resistance or rapid selection of resistant viruses (Beerenwinkel et al., 2003). Furthermore, epidemiological studies for a huge variety of human pathogenic viruses were performed using NGS technologies, including cytomegalovirus (CMV), Epstein Barr virus (EBV), HCV, influenza virus, norovirus, rhinovirus, rotavirus, and varicella zoster virus (VZV) (**Table 2**).

NGS is also increasingly used in more basic research areas, such as characterization of transmitted HIV (Fischer et al., 2010) and HCV (Wang et al., 2010; Bull et al., 2011), estimation of infection dates (Poon et al., 2011), evolution during the course of infection with HIV (Rozera et al., 2009; Poon et al., 2010; Wu et al., 2011), HCV (Bull et al., 2011), and rhinovirus (Cordey et al., 2010), and hypermutation patterns (Reuman et al., 2010; Knoepfel et al., 2011). Recently, NGS technologies have been applied to obtain the whole genome of HIV using a coverage allowing quasispecies analysis beyond the generation of consensus sequences to study, for instance, patterns of immune escape (Bimber et al., 2010; Willerth et al., 2010; Henn et al., 2012).

All these applications demonstrate the growing importance of NGS in studying viral diversity. With this technology, we will gain further insights into transmission traits, viral evolution, and its association with pathogenesis. World-wide viral diversity surveillance will be important for vaccine design and vaccination strategies. Currently, genetic diversity is mainly studied based on the detection and analyses of SNVs, rather than the reconstruction of linked mutations, due to the challenges in local and global haplotype reconstruction discussed above. It will be a huge step forward when haplotype reconstruction in heterogeneous viruses matures into a routine procedure based on standardized experimental protocols and validated, automatic data analysis pipelines.

OUTLOOK AND CONCLUSIONS

NGS opens up new roads to study viral diversity. It will tremendously increase our knowledge in virus evolution, fitness, selection pathways, and pathogenesis. Together with host genomics, viral diversity will allow insights into complex virus-host interactions. Full-length viral sequences may ultimately define truly conserved regions in viral genomes which might also be of relevance for vaccine and drug design. Clinically, the first application we can foresee is that in a single assay all drug targets relevant for antiviral treatment can be sequenced including information on minority drug-resistant variants. For all applications, sample procedures have to be chosen that minimize errors during sample preparation and sequencing. Several challenges in data analysis remain, especially in regard to alignments and global diversity estimation. In the future, some of these challenges might be diminished by upcoming third- and fourth-generation sequencing technologies, like single molecule or direct RNA sequencing.

Another not yet addressed future challenge will be making sense of the large amounts of genome data generated by NGS. For instance, clinical cut-offs need to be defined for minority drug-resistant virus variants, the clinical importance of new virus subtypes or even new viruses needs to be determined, and pathogenesis factors need to be confirmed in clinical settings. Thus, downstream analyses have to include large sets of well-documented patients, results from other experimental setups, etc. These are challenges as well as opportunities to answer important research questions which could not be addressed with conventional sequencing techniques.

ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation under grant number CR32I2_127017.

REFERENCES

Abbate, I., Vlassi, C., Rozera, G., Bruselles, A., Bartolini, B., Giombini, E., Corpolongo, A., D'Offizi, G., Narciso, P., Desideri, A., Ippolito, G., and Capobianchi, M. R. (2011). Detection of quasispecies variants predicted to use CXCR4 by ultra-deep pyrosequencing during early HIV infection. *AIDS* 25, 611–617.

Alteri, C., Santoro, M. M., Abbate, I., Rozera, G., Bruselles, A., Bartolini, B., Gori, C., Forbici, F., Orchi, N., Tozzi, V., Palamara, G., Antinori, A., Narciso, P., Girardi, E., Svicher, V., Ceccherini-Silberstein, F., Capobianchi, M. R., and Perno, C. F. (2011). 'Sentinel' mutations in standard population sequencing can predict the presence of HIV-1 reverse transcriptase major mutations detectable only by ultra-deep pyrosequencing. *J. Antimicrob. Chemother.* 66, 2615–2623.

Althaus, C. F., Vongrad, V., Niederost, B., Joos, B., Di Giallonardo, F., Rieder, P., Pavlovic, J., Trkola, A., Gunthard, H. F., Metzner, K. J., and Fischer, M. (2012). Tailored enrichment strategy detects low abundant small noncoding RNAs in HIV-1 infected cells. *Retrovirology* 9, 27.

Altmann, A., Weber, P., Quast, C., Rex-Haffner, M., Binder, E. B., and Müller-Myhsok, B. (2011). vipR: variant identification in pooled DNA using R. *Bioinformatics* 27, i77–i84.

Archer, J., Baillie, G., Watson, S. J., Kellam, P., Rambaut, A., and Robertson, D. L. (2012). Analysis of high-depth sequence data for studying viral diversity: a comparison of next generation sequencing platforms using Segminator, I. I. *BMC Bioinformatics* 13, 47.

Archer, J., Braverman, M. S., Taillon, B. E., Desany, B., James, I., Harrigan, P. R., Lewis, M., and Robertson, D. L. (2009). Detection of low-frequency pretherapy chemokine (CXC motif) receptor 4 (CXCR4)-using HIV-1 with ultra-deep pyrosequencing. *AIDS* 23, 1209–1218.

Archer, J., Rambaut, A., Taillon, B. E., Harrigan, P. R., Lewis, M., and Robertson, D. L. (2010). The evolutionary analysis of emerging low frequency HIV-1 CXCR4 using variants through time–an ultra-deep approach. *PLoS Comput. Biol.* 6:e1001022. doi: 10.1371/journal.pcbi.1001022

Armenia, D., Vandenbroucke, I., Fabeni, L., Van Marck, H., Cento, V., D'Arrigo, R., Van Wesenbeeck, L., Scopelliti, F., Micheli, V., Bruzzone, B., Lo Caputo, S., Aerssens, J., Rizzardini, G., Tozzi, V., Narciso, P., Antinori, A., Stuyver, L., Perno, C. F., and Ceccherini-Silberstein, F. (2012). Study of genotypic and phenotypic HIV-1 dynamics of integrase mutations during raltegravir treatment: a refined analysis by ultra-deep 454 pyrosequencing. *J. Infect. Dis.* 205, 557–567.

Astrovskaya, I., Tork, B., Mangul, S., Westbrooks, K., Măndoiu, I., Balfe, P., and Zelikovsky, A. (2011). Inferring viral quasispecies spectra from 454 pyrosequencing reads. *BMC Bioinformatics* 12(Suppl. 6), S1.

Baatz, F., Struck, D., Lemaire, M., De Landtsheer, S., Servais, J. Y., Arendt, V., Schmit, J. C., and Perez Bercoff, D. (2011). Rescue of HIV-1 long-time archived X4 strains to escape maraviroc. *Antiviral Res.* 92, 488–492.

Balzer, S., Malde, K., and Jonassen, I. (2011). Systematic exploration of error sources in pyrosequencing flowgram data. *Bioinformatics* 27, i304–i309.

Bartolini, B., Chillemi, G., Abbate, I., Bruselles, A., Rozera, G., Castrignano, T., Paoletti, D., Picardi, E., Desideri, A., Pesole, G., and Capobianchi, M. R. (2011). Assembly and characterization of pandemic influenza A H1N1 genome in nasopharyngeal swabs using high-throughput pyrosequencing. *New Microbiol.* 34, 391–397.

Barzon, L., Lavezzo, E., Militello, V., Toppo, S., and Palù, G. (2011). Applications of next-generation sequencing technologies to diagnostic virology. *Int. J. Mol. Sci.* 12, 7861–7884.

Beerenwinkel, N., Lengauer, T., Däumer, M., Kaiser, R., Walter, H., Korn, K., Hoffmann, D., and Selbig, J. (2003). Methods for optimizing antiviral combination therapies. *Bioinformatics* 19(Suppl. 1), i16–i25.

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

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Cheetham, R. K., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Catenazzi, M. C. E., Chang, S., Cooley, R. N., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fajardo, K. V. F., Furey, W. S., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G. D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R., and Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59.

Bimber, B. N., Dudley, D. M., Lauck, M., Becker, E. A., Chin, E. N., Lank, S. M., Grunenwald, H. L., Caruccio, N. C., Maffitt, M., Wilson, N. A., Reed, J. S., Sosman, J. M., Tarosso, L. F., Sanabani, S., Kallas, E. G., Hughes, A. L., and O'Connor, D. H. (2010). Whole-genome characterization of human and simian immunodeficiency virus intrahost diversity by ultra-deep pyrosequencing. *J. Virol.* 84, 12087–12092.

Bolcic, F., Sede, M., Moretti, F., Westergaard, G., Vazquez, M., Laufer, N., and Quarleri, J. (2012). Analysis of the PKR-eIF2alpha phosphorylation homology domain (PePHD) of hepatitis C virus genotype 1 in HIV-coinfected patients by ultra-deep pyrosequencing and its relationship to responses to pegylated interferon-ribavirin treatment. *Arch. Virol.* 157, 703–711.

Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W. L., Russ, C., Lander, E. S., Nusbaum, C., and Jaffe, D. B. (2008). Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res.* 18, 763–770.

Bruselles, A., Rozera, G., Bartolini, B., Prosperi, M., Del Nonno, F., Narciso, P., Capobianchi, M. R., and Abbate, I. (2009). Use of massive parallel pyrosequencing for near full-length characterization of a unique HIV Type 1 BF recombinant associated with a fatal primary infection. *AIDS Res. Hum. Retroviruses* 25, 937–942.

Bull, R. A., Eden, J.-S., Luciani, F., McElroy, K., Rawlinson, W. D., and White, P. A. (2012). Contribution of intra- and interhost dynamics to norovirus evolution. *J. Virol.* 86, 3219–3229.

Bull, R. A., Luciani, F., McElroy, K., Gaudieri, S., Pham, S. T., Chopra, A., Cameron, B., Maher, L., Dore, G. J., White, P. A., and Lloyd, A. R. (2011). Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection. *PLoS Pathog.* 7:e1002243. doi: 10.1371/journal.ppat.1002243

Bunnik, E. M., Swenson, L. C., Edo-Matas, D., Huang, W., Dong, W., Frantzell, A., Petropoulos, C. J., Coakley, E., Schuitemaker, H., Harrigan, P. R., and van 't Wout, A. B. (2011). Detection of inferred CCR5- and CXCR4-using HIV-1 variants and evolutionary intermediates using ultra-deep pyrosequencing. *PLoS Pathog.* 7:e1002106. doi: 10.1371/journal.ppat.1002106

Burch, C. L., and Chao, L. (2000). Evolvability of an RNA virus is determined by its mutational neighbourhood. *Nature* 406, 625–628.

Casbon, J. A., Osborne, R. J., Brenner, S., and Lichtenstein, C. P. (2011). A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Res.* 39, e81.

Chang, K. S., Vyas, R. C., Deaven, L. L., Trujillo, J. M., Stass, S. A., and Hittelman, W. N. (1992). PCR amplification of chromosome-specific DNA isolated from flow cytometry-sorted chromosomes. *Genomics* 12, 307–312.

Chapman, M. A., Lawrence, M. S., Keats, J. J., Cibulskis, K., Sougnez, C., Schinzel, A. C., Harview, C. L., Brunet, J.-P., Ahmann, G. J., Adli, M., Anderson, K. C., Ardlie, K. G., Auclair, D., Baker, A., Bergsagel, P. L., Bernstein, B. E., Drier, Y., Fonseca, R., Gabriel, S. B., Hofmeister, C. C., Jagannath, S., Jakubowiak, A. J., Krishnan, A., Levy, J., Liefeld, T., Lonial, S., Mahan, S., Mfuko, B., Monti, S., Perkins, L. M., Onofrio, R., Pugh, T. J., Rajkumar, S. V., Ramos, A. H., Siegel, D. S., Sivachenko, A., Stewart, A. K., Trudel, S., Vij, R., Voet, D., Winckler, W., Zimmerman, T., Carpten, J., Trent, J., Hahn, W. C., Garraway, L. A., Meyerson, M., Lander, E. S., Getz, G., and Golub, T. R. (2011). Initial genome sequencing and analysis of multiple myeloma. *Nature* 471, 467–472.

Cordey, S., Junier, T., Gerlach, D., Gobbini, F., Farinelli, L., Zdobnov, E. M., Winther, B., Tapparel, C., and Kaiser, L. (2010). Rhinovirus genome evolution during experimental human infection. *PLoS ONE* 5:e10588. doi: 10.1371/journal.pone.0010588

D'Aquila, R. T., Geretti, A. M., Horton, J. H., Rouse, E., Kheshti, A., Raffanti, S., Oie, K., Pappa, K., and Ross, L. L. (2011). Tenofovir (TDF)-selected or abacavir (ABC)-selected low-frequency HIV type 1 subpopulations during failure with persistent viremia as detected by ultradeep pyrosequencing. *AIDS Res. Hum. Retroviruses* 27, 201–209.

Daly, G. M., Bexfield, N., Heaney, J., Stubbs, S., Mayer, A. P., Palser, A., Kellam, P., Drou, N., Caccamo, M., Tiley, L., Alexander, G. J., Bernal, W., and Heeney, J. L. (2011). A viral discovery methodology for clinical biopsy samples utilising massively parallel next generation sequencing. *PLoS ONE* 6:e28879. doi: 10.1371/journal.pone.0028879

Delobel, P., Saliou, A., Nicot, F., Dubois, M., Trancart, S., Tangre, P., Aboulker, J. P., Taburet, A. M., Molina, J. M., Massip, P., Marchou, B., and Izopet, J. (2011). Minor HIV-1 variants with the K103N resistance mutation during intermittent Efavirenz-containing antiretroviral therapy and virological failure. *PLoS ONE* 6:e21655. doi: 10.1371/journal.pone.0021655

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.

Domingo, E., and Holland, J. J. (1997). RNA virus mutations and fitness for survival. *Annu. Rev. Microbiol.* 51, 151–178.

Drummond, A. J., Pybus, O. G., Rambaut, A., Forsberg, R., and Rodrigo, A. G. (2003). Measurably evolving populations. *Trends Ecol. Evol.* 18, 481–488.

Duffy, S., Shackelton, L. A., and Holmes, E. C. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.* 9, 267–276.

Eckert, K. A., and Kunkel, T. A. (1991). DNA polymerase fidelity and the polymerase chain reaction. *PCR Methods Appl.* 1, 17–24.

Eigen, M. (1971). Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* 58, 465–523.

Eigen, M., McCaskill, J., and Schuster, P. (1988). Molecular quasi-species. *J. Phys. Chem.* 92, 6881–6891.

Eigen, M., McCaskill, J., and Schuster, P. (1989). The molecular quasi-species. *Adv. Chem. Phys.* 75, 149–263.

Eigen, M., and Schuster, P. (1977). The hypercycle. A principle of natural self-organization. Part A: emergence of the hypercycle. *Naturwissenschaften* 64, 541–565.

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. doi: 10.1371/journal.pcbi.1000074

Escobar-Gutiérrez, A., Vazquez-Pichardo, M., Cruz-Rivera, M., Rivera-Osorio, P., Carpio-Pedroza, J. C., Ruíz-Pacheco, J. A., Ruiz-Tovar, K., and Vaughan, G. (2012). Identification of hepatitis C virus transmission using a next-generation sequencing approach. *J. Clin. Microbiol.* 50, 1461–1463.

Eshleman, S. H., Hudelson, S. E., Redd, A. D., Wang, L., Debes, R., Chen, Y. Q., Martens, C. A., Ricklefs, S. M., Selig, E. J., Porcella, S. F., Munshaw, S., Ray, S. C., Piwowar-Manning, E., McCauley, M., Hosseinipour, M. C., Kumwenda, J., Hakim, J. G., Chariyalertsak, S., De Bruyn, G., Grinsztejn, B., Kumarasamy, N., Makhema, J., Mayer, K. H., Pilotto, J., Santos, B. R., Quinn, T. C., Cohen, M. S., and Hughes, J. P. (2011). Analysis of genetic linkage of HIV from couples enrolled in the HIV prevention trials network 052 trial. *J. Infect. Dis.* 204, 1918–1926.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* 3, 87–112.

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 8, 186–194.

Fang, G., Zhu, G., Burger, H., Keithly, J. S., and Weiser, B. (1998). Minimizing DNA recombination during long RT-PCR. *J. Virol. Methods* 76, 139–148.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *Ann. Stat.* 1, 209–230.

Finotello, F., Lavezzo, E., Fontana, P., Peruzzo, D., Albiero, A., Barzon, L., Falda, M., Camillo, B. D., and Toppo, S. (2012). Comparative analysis of algorithms for whole-genome assembly of pyrosequencing data. *Brief Bioinform.* 13, 269–280.

Fischer, M., Wong, J. K., Russenberger, D., Joos, B., Opravil, M., Hirschel, B., Trkola, A., Kuster, H., Weber, R., and Gunthard, H. F. (2002). Residual cell-associated unspliced HIV-1 RNA in peripheral blood of patients on potent antiretroviral therapy represents intracellular transcripts. *Antivir. Ther.* 7, 91–103.

Fischer, W., Ganusov, V. V., Giorgi, E. E., Hraber, P. T., Keele, B. F., Leitner, T., Han, C. S., Gleasner, C. D., Green, L., Lo, C. C., Nag, A., Wallstrom, T. C., Wang, S., McMichael, A. J., Haynes, B. F., Hahn, B. H., Perelson, A. S., Borrow, P., Shaw, G. M., Bhattacharya, T., and Korber, B. T. (2010). Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PLoS ONE* 5:e12303. doi: 10.1371/journal.pone.0012303

Fisher, R., Van Zyl, G. U., Travers, S. A., Pond, S. L. K., Engelbrech, S., Murrell, B., Scheffler, K., and Smith, D. (2012). Deep sequencing reveals minor protease resistance mutations in patients failing a protease inhibitor regimen. *J. Virol.* 86, 6231–6237.

Flaherty, P., Natsoulis, G., Muralidharan, O., Winters, M., Buenrostro, J., Bell, J., Brown, S., Holodniy, M., Zhang, N., and Ji, H. P. (2012). Ultrasensitive detection of rare mutations using next-generation targeted resequencing. *Nucleic Acids Res.* 40, e2.

Fonseca-Coronado, S., Escobar-Gutierrez, A., Ruiz-Tovar, K., Cruz-Rivera, M. Y., Rivera-Osorio, P., Vazquez-Pichardo, M., Carpio-Pedroza, J. C., Ruiz-Pacheco, J. A., Cazares, F., and Vaughan, G. (2012). Specific detection of naturally occurring hepatitis C virus mutants with resistance to telaprevir and boceprevir (protease inhibitors) among treatment-naive infected individuals. *J. Clin. Microbiol.* 50, 281–287.

Forbi, J. C., Purdy, M. A., Campo, D. S., Vaughan, G., Dimitrova, Z. E., Ganova-Raeva, L. M., Xia, G. L., and Khudyakov, Y. E. (2012). Epidemic history of hepatitis C virus infection in two remote communities in Nigeria, West Africa. *J. Gen. Virol.* 93, 1410–1421.

Gerstung, M., Beisel, C., Rechsteiner, M., Wild, P., Schraml, P., Moch, H., and Beerenwinkel, N. (2012). Reliable detection of subclonal single-nucleotide variants in tumor cell populations. *Nat. Commun.* 3, 811.

Gianella, S., Delport, W., Pacold, M. E., Young, J. A., Choi, J. Y., Little, S. J., Richman, D. D., Pond, S. L. K., and Smith, D. M. (2011). Detection of minority resistance during early HIV-1 infection: natural variation and spurious detection rather than transmission and evolution of multiple viral variants. *J. Virol.* 85, 8359–8367.

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

Gorzer, I., Guelly, C., Trajanoski, S., and Puchhammer-Stockl, E. (2010). Deep sequencing reveals highly complex dynamics of human cytomegalovirus genotypes in transplant patients over time. *J. Virol.* 84, 7195–7203.

Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., Schork, N. J., Murray, S. S., Topol, E. J., Levy, S., and Frazer, K. A. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* 10, R32.

Hedskog, C., Mild, M., Jernberg, J., Sherwood, E., Bratt, G., Leitner, T., Lundeberg, J., Andersson, B., and Albert, J. (2010). Dynamics of HIV-1 quasispecies during antiviral treatment dissected using ultra-deep pyrosequencing. *PLoS ONE* 5:e11345. doi: 10.1371/journal.pone.0011345

Henn, M. R., Boutwell, C. L., Charlebois, P., Lennon, N. J., Power, K. A., Macalalad, A. R., Berlin, A. M., Malboeuf, C. M., Ryan, E. M., Gnerre, S., Zody, M. C., Erlich, R. L., Green, L. M., Berical, A., Wang, Y., Casali, M., Streeck, H., Bloom, A. K., Dudek, T., Tully, D., Newman, R., Axten, K. L., Gladden, A. D., Battis, L., Kemper, M., Zeng, Q., Shea, T. P., Gujja, S., Zedlack, C., Gasser, O., Brander, C., Hess, C., Gunthard, H. F., Brumme, Z. L., Brumme, C. J., Bazner, S., Rychert, J., Tinsley, J. P., Mayer, K. H., Rosenberg, E., Pereyra, F., Levin, J. Z., Young, S. K., Jessen, H., Altfeld, M., Birren, B. W., Walker, B. D., and Allen, T. M. (2012). Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog.* 8:e1002529. doi: 10.1371/journal.ppat.1002529

Hiraga, N., Imamura, M., Abe, H., Hayes, C. N., Kono, T., Onishi, M., Tsuge, M., Takahashi, S., Ochi, H., Iwao, E., Kamiya, N., Yamada, I., Tateno, C., Yoshizato, K., Matsui, H., Kanai, A., Inaba, T., Tanaka, S., and Chayama, K. (2011). Rapid emergence of telaprevir resistant hepatitis C virus strain from wild-type clone *in vivo*. *Hepatology* 54, 781–788.

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.

Holodniy, M., Mole, L., Yen-Lieberman, B., Margolis, D., Starkey, C., Carroll, R., Spahlinger, T., Todd, J., and Jackson, J. B. (1995). Comparative stabilities of quantitative human immunodeficiency virus RNA in plasma from samples collected in VACUTAINER CPT, VACUTAINER PPT, and standard VACUTAINER tubes. *J. Clin. Microbiol.* 33, 1562–1566.

Homs, M., Buti, M., Quer, J., Jardi, R., Schaper, M., Tabernero, D., Ortega, I., Sanchez, A., Esteban, R., and Rodriguez-Frias, F. (2011). Ultra-deep pyrosequencing analysis of the hepatitis B virus preCore region and main catalytic motif of the viral polymerase in the same viral genome. *Nucleic Acids Res.* 39, 8457–8471.

Huang, A., Kantor, R., Delong, A., Schreier, L., and Istrail, S. (2011). "QColors: An algorithm for conservative viral quasispecies reconstruction from short and non-contiguous

next generation sequencing reads," in *IEEE International Conference on Bioinformatics and Biomedicine Workshops*. Publisher is Institute of Electrical and Electronics Engineers (IEEE), 130–136.

Huse, S., Huber, J., Morrison, H., Sogin, M., and Welch, D. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* 8, R143.

Iwasa, Y., Michor, F., and Nowak, M. A. (2003). Evolutionary dynamics of escape from biomedical intervention. *Proc. Biol. Sci.* 270, 2573–2578.

Iwasa, Y., Michor, F., and Nowak, M. A. (2004). Evolutionary dynamics of invasion and escape. *J. Theor. Biol.* 226, 205–214.

Jabara, C. B., Jones, C. D., Roach, J., Anderson, J. A., and Swanstrom, R. (2011). Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer, I. D. *Proc. Natl. Acad. Sci. U.S.A.* 108, 20166–20171.

Jain, A. K., and Dubes, R. C. (1981). *Algorithms for Clustering Data*. Upper Saddle River, NJ: Prentice-Hall.

Jere, K. C., Mlera, L., Page, N. A., Van Dijk, A. A., and O'neill, H. G. (2011). Whole genome analysis of multiple rotavirus strains from a single stool specimen using sequence-independent amplification and 454(R) pyrosequencing reveals evidence of intergenotype genome segment recombination. *Infect. Genet. Evol.* 11, 2072–2082.

Ji, H., Li, Y., Graham, M., Liang, B. B., Pilon, R., Tyson, S., Peters, G., Tyler, S., Merks, H., Bertagnolio, S., Soto-Ramirez, L., Sandstrom, P., and Brooks, J. (2011). Next-generation sequencing of dried blood spot specimens: a novel approach to HIV drug-resistance surveillance. *Antivir. Ther.* 16, 871–878.

Ji, H., Masse, N., Tyler, S., Liang, B., Li, Y., Merks, H., Graham, M., Sandstrom, P., and Brooks, J. (2010). HIV drug resistance surveillance using pooled pyrosequencing. *PLoS ONE* 5:e9263. doi: 10.1371/journal.pone.0009263

Jojic, V., Hertz, T., and Jojic, N. (2008). "Population sequencing using short reads: HIV as a case study," in *Pacific Symposium on Biocomputing*, eds R. B. Altman, A. K. Dunker, L. Hunter, T. Murray, and T. E. Klein (World Scientific), 114–125. ISBN 978-981-277-608-2.

Jose, M., Gajardo, R., and Jorquera, J. I. (2005). Stability of HCV, HIV-1 and HBV nucleic acids in plasma samples under long-term storage. *Biologicals* 33, 9–16.

Judo, M. S., Wedel, A. B., and Wilson, C. (1998). Stimulation and suppression of PCR-mediated recombination. *Nucleic Acids Res.* 26, 1819–1825.

Jung, G. S., Kim, Y. Y., Kim, J. I., Ji, G. Y., Jeon, J. S., Yoon, H. W., Lee, G. C., Ahn, J. H., Lee, K. M., and Lee, C. H. (2011). Full genome sequencing and analysis of human cytomegalovirus strain JHC isolated from a Korean patient. *Virus Res.* 156, 113–120.

Kampmann, M. L., Fordyce, S. L., Avila-Arcos, M. C., Rasmussen, M., Willerslev, E., Nielsen, L. P., and Gilbert, M. T. (2011). A simple method for the parallel deep sequencing of full influenza A genomes. *J. Virol. Methods* 178, 243–248.

Kanagawa, T. (2003). Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J. Biosci. Bioeng.* 96, 317–323.

Kircher, M., Stenzel, U., and Kelso, J. (2009). Improved base calling for the illumina genome analyzer using machine learning strategies. *Genome Biol.* 10, R83.

Knoepfel, S. A., Di Giallonardo, F., Daumer, M., Thielen, A., and Metzner, K. J. (2011). In-depth analysis of G-to-A hypermutation rate in HIV-1 env DNA induced by endogenous APOBEC3 proteins using massively parallel sequencing. *J. Virol. Methods* 171, 329–338.

Ko, S.-Y., Oh, H.-B., Park, C.-W., Lee, H. C., and Lee, J.-E. (2012). Analysis of hepatitis B virus drug-resistant mutant haplotypes by ultra-deep pyrosequencing. *Clin. Microbiol. Infect.* doi: 10.1111/j.1469-0691.2012.03951.x. [Epub ahead of print].

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., and Wilson, R. K. (2012). VarScan 2, Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 22, 568–576.

Kozal, M. J., Chiarella, J., St. John, E. P., Moreno, E. A., Simen, B. B., Arnold, T. E., and Lataillade, M. (2011). Prevalence of low-level HIV-1 variants with reverse transcriptase mutation K65R and the effect of antiretroviral drug exposure on variant levels. *Antivir. Ther.* 16, 925–929.

Kuroda, M., Katano, H., Nakajima, N., Tobiume, M., Ainai, A., Sekizuka, T., Hasegawa, H., Tashiro, M., Sasaki, Y., Arakawa, Y., Hata, S., Watanabe, M., and 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:e10256. doi: 10.1371/journal.pone.0010256

Kwok, H., Tong, A. H. Y., Lin, C. H., Lok, S., Farrell, P. J., Kwong, D. L. W., and Chiang, A. K. S. (2012). Genomic sequencing and comparative analysis of Epstein-Barr virus genome isolated from primary nasopharyngeal carcinoma biopsy. *PLoS ONE* 7:e36939. doi: 10.1371/journal.pone.0036939

Lataillade, M., Chiarella, J., Yang, R., Degrosky, M., Uy, J., Seekins, D., Simen, B., John, E. S., Moreno, E., and Kozal, M. (2012). Virologic failures on initial boosted-PI regimen infrequently possess low-level variants with major PI resistance mutations by ultra-deep sequencing. *PLoS ONE* 7:e30118. doi: 10.1371/journal.pone.0030118

Lataillade, M., Chiarella, J., Yang, R., Schnittman, S., Wirtz, V., Uy, J., Seekins, D., Krystal, M., Mancini, M., McGrath, D., Simen, B., Egholm, M., and Kozal, M. (2010). Prevalence and clinical significance of HIV drug resistance mutations by ultra-deep sequencing in antiretroviral-naive subjects in the CASTLE study. *PLoS ONE* 5:e10952. doi: 10.1371/journal.pone.0010952

Lauck, M., Alvarado-Mora, M. V., Becker, E. A., Bhattacharya, D., Striker, R., Hughes, A. L., Carrilho, F. J., O'connor, D. H., and Pinho, J. R. (2012). Analysis of hepatitis C virus intrahost diversity across the coding region by ultradeep pyrosequencing. *J. Virol.* 86, 3952–3960.

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. doi: 10.1371/journal.pone.0006079

Li, J. Z., Paredes, R., Ribaudo, H. J., Svarovskaia, E. S., Metzner, K. J., Kozal, M. J., Hullsiek, K. H., Balduin, M., Jakobsen, M. R., Geretti, A. M., Thiebaut, R., Ostergaard, L., Masquelier, B., Johnson, J. A., Miller, M. D., and Kuritzkes, D. R. (2011). Low-frequency HIV-1 drug resistance mutations and risk of NNRTI-based antiretroviral treatment failure: a systematic review and pooled analysis. *JAMA* 305, 1327–1335.

Lipkin, W. I. (2010). Microbe hunting. *Microbiol. Mol. Biol. Rev.* 74, 363–377.

Liu, P., Fang, X., Feng, Z., Guo, Y.-M., Peng, R.-J., Liu, T., Huang, Z., Feng, Y., Sun, X., Xiong, Z., Guo, X., Pang, S.-S., Wang, B., Lv, X., Feng, F.-T., Li, D.-J., Chen, L.-Z., Feng, Q.-S., Huang, W.-L., Zeng, M.-S., Bei, J.-X., Zhang, Y., and Zeng, Y.-X. (2011). Direct sequencing and characterization of a clinical isolate of Epstein-Barr virus from nasopharyngeal carcinoma tissue by using next-generation sequencing technology. *J. Virol.* 85, 11291–11299.

Liu, S. L., Rodrigo, A. G., Shankarappa, R., Learn, G. H., Hsu, L., Davidov, O., Zhao, L. P., and Mullins, J. I. (1996). HIV quasispecies and resampling. *Science* 273, 415–416.

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., and Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* 30, 434–439.

Lorusso, A., Vincent, A. L., Harland, M. L., Alt, D., Bayles, D. O., Swenson, S. L., Gramer, M. R., Russell, C. A., Smith, D. J., Lager, K. M., and Lewis, N. S. (2011). Genetic and antigenic characterization of H1 influenza viruses from United States swine from 2008. *J. Gen. Virol.* 92, 919–930.

Macalalad, A. R., Zody, M. C., Charlebois, P., Lennon, N. J., Newman, R. M., Malboeuf, C. M., Ryan, E. M., Boutwell, C. L., Power, K. A., Brackney, D. E., Pesko, K. N., Levin, J. Z., Ebel, G. D., Allen, T. M., Birren, B. W., and Henn, M. R. (2012). Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data. *PLoS Comput. Biol.* 8:e1002417. doi: 10.1371/journal.pcbi.1002417

Mancuso, N., Tork, B., Mandoiu, I. I., Skums, P., and Zelikovsky, A. (2011). "Viral quasispecies reconstruction from amplicon 454 pyrosequencing reads," in *Proceedings of the 1st Workshop on Computational Advances in Molecular Epidemiology,* (IEEE), 94–101. ISBN: 978-1-4577-1612-6.

Mardis, E. R. (2008a). The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24, 133–141.

Mardis, E. R. (2008b). Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* 9, 387–402.

Margeridon-Thermet, S., Shulman, N. S., Ahmed, A., Shahriar, R., Liu, T.,

Wang, C., Holmes, S. P., Babrzadeh, F., Gharizadeh, B., Hanczaruk, B., Simen, B. B., Egholm, M., and Shafer, R. W. (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.

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., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380.

Messiaen, P., Verhofstede, C., Vandenbroucke, I., Dinakis, S., Van Eygen, V., Thys, K., Winters, B., Aerssens, J., Vogelaers, D., Stuyver, L. J., and Vandekerckhove, L. (2012). Ultra-deep sequencing of HIV-1 reverse transcriptase before start of an NNRTI-based regimen in treatment-naive patients. *Virology* 426, 7–11.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat. Rev. Genet.* 11, 31–46.

Metzner, K. J., Bonhoeffer, S., Fischer, M., Karanicolas, R., Allers, K., Joos, B., Weber, R., Hirschel, B., Kostrikis, L. G., Günthard, H. F., and Study, T. S. H. C. (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.

Metzner, K. J., Giulieri, S. G., Knoepfel, S. A., Rauch, P., Burgisser, P., Yerly, S., Gunthard, H. F., and Cavassini, M. (2009). Minority quasispecies of drug-resistant HIV-1 that lead to early therapy failure in treatment-naive and -adherent patients. *Clin. Infect. Dis.* 48, 239–247.

Meyerhans, A., Vartanian, J. P., and Wain-Hobson, S. (1990). DNA recombination during PCR. *Nucleic Acids Res.* 18, 1687–1691.

Mild, M., Hedskog, C., Jernberg, J., and Albert, J. (2011). Performance of ultra-deep pyrosequencing in analysis of HIV-1 pol gene variation. *PLoS ONE* 6:e22741. doi: 10.1371/journal.pone.0022741

Mitsuya, Y., Varghese, V., Wang, C., Liu, T. F., Holmes, S. P., Jayakumar, P., Gharizadeh, B., Ronaghi, M., Klein, D., Fessel, W. J., and Shafer, R. W. (2008). Minority human immunodeficiency virus type 1 variants in antiretroviral-naive persons with reverse transcriptase codon 215 revertant mutations. *J. Virol.* 82, 10747–10755.

Moorthy, A., Kuhn, L., Coovadia, A., Meyers, T., Strehlau, R., Sherman, G., Tsai, W. Y., Chen, Y. H., Abrams, E. J., and Persaud, D. (2011). Induction therapy with protease-inhibitors modifies the effect of nevirapine resistance on virologic response to nevirapine-based HAART in children. *Clin. Infect. Dis.* 52, 514–521.

Mukherjee, R., Jensen, S. T., Male, F., Bittinger, K., Hodinka, R. L., Miller, M. D., and Bushman, F. D. (2011). Switching between raltegravir resistance pathways analyzed by deep sequencing. *AIDS* 25, 1951–1959.

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M. C., Hirai, A., Takahashi, H., Ul-Amin, M. A., Ogasawara, N., and Kanaya, S. (2011). Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* 39, e90.

Nasu, A., Marusawa, H., Ueda, Y., Nishijima, N., Takahashi, K., Osaki, Y., Yamashita, Y., Inokuma, T., Tamada, T., Fujiwara, T., Sato, F., Shimizu, K., and Chiba, T. (2011). Genetic heterogeneity of hepatitis C virus in association with antiviral therapy determined by ultra-deep sequencing. *PLoS ONE* 6:e24907. doi: 10.1371/journal.pone.0024907

Ninomiya, M., Ueno, Y., Funayama, R., Nagashima, T., Nishida, Y., Kondo, Y., Inoue, J., Kakazu, E., Kimura, O., Nakayama, K., and Shimosegawa, T. (2012). Use of illumina deep sequencing technology to differentiate hepatitis C virus variants. *J. Clin. Microbiol.* 50, 857–866.

Nishijima, N., Marusawa, H., Ueda, Y., Takahashi, K., Nasu, A., Osaki, Y., Kou, T., Yazumi, S., Fujiwara, T., Tsuchiya, S., Shimizu, K., Uemoto, S., and Chiba, T. (2012). Dynamics of hepatitis B virus quasispecies in association with nucleos(t)ide analogue treatment determined by ultra-deep sequencing. *PLoS ONE* 7:e35052. doi: 10.1371/journal.pone.0035052

Nowak, M. A. (1992). What is a quasispecies? *Trends Ecol. Evol.* 7, 118–121.

O'Neil, S. T., and Emrich, S. J. (2012). Haplotype and minimum-chimerism consesus determination using short sequence data. *BMC Genomics* 13, S4.

Ojosnegros, S., Beerenwinkel, N., Antal, T., Nowak, M. A., Escarmís, C., and Domingo, E. (2010). Competition-colonization dynamics in an RNA virus. *Proc. Natl. Acad. Sci. U.S.A.* 107, 2108–2112.

Poon, A. F., McGovern, R. A., Mo, T., Knapp, D. J., Brenner, B., Routy, J. P., Wainberg, M. A., and Harrigan, P. R. (2011). Dates of HIV infection can be estimated for seroprevalent patients by coalescent analysis of serial next-generation sequencing data. *AIDS* 25, 2019–2026.

Poon, A. F., Swenson, L. C., Dong, W. W., Deng, W., Kosakovsky Pond, S. L., Brumme, Z. L., Mullins, J. I., Richman, D. D., Harrigan, P. R., and Frost, S. D. (2010). Phylogenetic analysis of population-based and deep sequencing data to identify coevolving sites in the nef gene of HIV-1. *Mol. Biol. Evol.* 27, 819–832.

Pop, M., and Salzberg, S. L. (2008). Bioinformatics challenges of new sequencing technology. *Trends Genet.* 24, 142–149.

Powdrill, M. H., Tchesnokov, E. P., Kozak, R. A., Russell, R. S., Martin, R., Svarovskaia, E. S., Mo, H., Kouyos, R. D., and Gotte, M. (2011). Contribution of a mutational bias in hepatitis C virus replication to the genetic barrier in the development of drug resistance. *Proc. Natl. Acad. Sci. U.S.A.* 108, 20509–20513.

Prabhakaran, S., Rey, M., Zagordi, O., Beerenwinkel, N., and Roth, V. (2010). "HIV haplotype inference using a constraint-based Dirichlet process mixture model," in *NIPS Workshop on Machine Learning in Computational Biology*.

Preston, B. D., Poiesz, B. J., and Loeb, L. A. (1988). Fidelity of HIV-1 reverse transcriptase. *Science* 242, 1168–1171.

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

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

Pybus, O. G., and Rambaut, A. (2009). Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.* 10, 540–550.

Quince, C., Lanzén, A., Curtis, T. P., Davenport, R. J., Hall, N., Head, I. M., Read, L. F., and Sloan, W. T. (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat. Methods* 6, 639–641.

Quince, C., Lanzen, A., Davenport, R. J., and Turnbaugh, P. J. (2011). Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* 12, 38.

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. doi: 10.1371/journal.pone.0007105

Rasmussen, C. E. (2000). "The infinite gaussian mixture model," in *NIPS*, eds S. A. Solla, T. K. Leen, and K.-R. Müller (The MIT Press), 554–560.

Raymond, S., Saliou, A., Nicot, F., Delobel, P., Dubois, M., Cazabat, M., Sandres-Saune, K., Marchou, B., Massip, P., and Izopet, J. (2011). Frequency of CXCR4-using viruses in primary HIV-1 infections using ultra-deep pyrosequencing. *AIDS* 25, 1668–1670.

Redd, A. D., Mullis, C. E., Serwadda, D., Kong, X., Martens, C., Ricklefs, S. M., Tobian, A. A., Xiao, C., Grabowski, M. K., Nalugoda, F., Kigozi, G., Laeyendecker, O., Kagaayi, J., Sewankambo, N., Gray, R. H., Porcella, S. F., Wawer, M. J., and Quinn, T. C. (2012). The rates of HIV superinfection and primary HIV incidence in a general population in Rakai, Uganda. *J. Infect. Dis.* 206, 267–274.

Reuman, E. C., Margeridon-Thermet, S., Caudill, H. B., Liu, T., Borroto-Esoda, K., Svarovskaia, E. S., Holmes, S. P., and Shafer, R. W. (2010). A classification model for G-to-A hypermutation in hepatitis B virus ultra-deep pyrosequencing reads. *Bioinformatics* 26, 2929–2932.

Reumers, J., Rijk, P. D., Zhao, H., Liekens, A., Smeets, D., Cleary, J., Loo, P. V., Bossche, M. V. D., Catthoor, K., Sabbe, B., Despierre, E., Vergote, I., Hilbush, B., Lambrechts, D., and Del-Favero,

J. (2011). Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. *Nat. Biotechnol.* 30, 61–68.

Reyes, G. R., and Kim, J. P. (1991). Sequence-independent, single-primer amplification (SISPA) of complex DNA populations. *Mol. Cell. Probes* 5, 473–481.

Roberts, J. D., Bebenek, K., and Kunkel, T. A. (1988). The accuracy of reverse transcriptase from HIV-1. *Science* 242, 1171–1173.

Rodriguez-Frías, F., Tabernero, D., Quer, J., Esteban, J. I., Ortega, I., Domingo, E., Cubero, M., Camós, S., Ferrer-Costa, C., Sánchez, A., Jardí, R., Schaper, M., Homs, M., Garcia-Cehic, D., Guardia, J., Esteban, R., and Buti, M. (2012). Ultra-deep pyrosequencing detects conserved genomic sites and quantifies linkage of drug-resistant amino acid changes in the hepatitis B virus genome. *PLoS ONE* 7:e37874. doi: 10.1371/journal.pone.0037874

Rozera, G., Abbate, I., Bruselles, A., Vlassi, C., D'offizi, G., Narciso, P., Chillemi, G., Prosperi, M., Ippolito, G., and Capobianchi, M. R. (2009). Massively parallel pyrosequencing highlights minority variants in the HIV-1 *env* quasispecies deriving from lymphomonocyte sub-populations. *Retrovirology* 6, 15.

Saeed, F., Khokhar, A., Zagordi, O., and Beerenwinkel, N. (2009). "Multiple sequence alignment system for pyrosequencing reads," in *BICoB 2009, LNBI 5462*, ed S. Rajasekaran (Berlin Heidelberg: Springer-Verlag), 362–375.

Saliou, A., Delobel, P., Dubois, M., Nicot, F., Raymond, S., Calvez, V., Masquelier, B., and Izopet, J. (2011). Concordance between two phenotypic assays and ultradeep pyrosequencing for determining HIV-1 tropism. *Antimicrob. Agents Chemother.* 55, 2831–2836.

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nat. Methods* 5, 16–18.

Sede, M., Ojeda, D., Cassino, L., Westergaard, G., Vazquez, M., Benetti, S., Fay, F., Tanno, H., and Quarleri, J. (2012). Long-term monitoring drug resistance by ultra-deep pyrosequencing in a chronic hepatitis B virus (HBV)-infected patient exposed to several unsuccessful therapy schemes. *Antiviral Res*. 94, 184–187.

Simen, B. B., Simons, J. F., Hullsiek, K. H., Novak, R. M., MacArthur, R. D., Baxter, J. D., Huang, C., Lubeski, C., Turenchalk, G. S., Braverman, M. S., Desany, B., Rothberg, J. M., Egholm, M., and Kozal, M. J. (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.

Skums, P., Dimitrova, Z., Campo, D. S., Vaughan, G., Rossi, L., Forbi, J. C., Yokosawa, J., Zelikovsky, A., and Khudyakov, Y. (2012). Efficient error correction for next-generation sequencing of viral amplicons. *BMC Bioinformatics* 13(Suppl. 10), S6.

Solmone, M., Vincenti, D., Prosperi, M. C., 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.

Stelzl, E., Proll, J., Bizon, B., Niklas, N., Danzer, M., Hackl, C., Stabentheiner, S., Gabriel, C., and Kessler, H. H. (2011). Human immunodeficiency virus type 1 drug resistance testing: evaluation of a new ultra-deep sequencing-based protocol and comparison with the TRUGENE HIV-1 genotyping kit. *J. Virol. Methods* 178, 94–97.

Svicher, V., Balestra, E., Cento, V., Sarmati, L., Dori, L., Vandenbroucke, I., D'Arrigo, R., Buonomini, A. R., Marck, H. V., Surdo, M., Saccomandi, P., Mostmans, W., Aerssens, J., Aquaro, S., Stuyver, L. J., Andreoni, M., Ceccherini-Silberstein, F., and Perno, C. F. (2011). HIV-1 dual/mixed tropic isolates show different genetic and phenotypic characteristics and response to maraviroc *in vitro*. *Antiviral Res.* 90, 42–53.

Swenson, L. C., Mo, T., Dong, W. W., Zhong, X., Woods, C. K., Jensen, M. A., Thielen, A., Chapman, D., Lewis, M., James, I., Heera, J., Valdez, H., and Harrigan, P. R. (2011a). Deep sequencing to infer HIV-1 co-receptor usage: application to three clinical trials of maraviroc in treatment-experienced patients. *J. Infect. Dis.* 203, 237–245.

Swenson, L. C., Mo, T., Dong, W. W. Y., Zhong, X., Woods, C. K., Thielen, A., Jensen, M. A., Knapp, D. J. H. F., Chapman, D., Portsmouth, S., Lewis, M., James, I., Heera, J., Valdez, H., and Harrigan, P. R. (2011b). Deep V3 sequencing for HIV type 1 tropism in treatment-naive patients: a reanalysis of the MERIT trial of maraviroc. *Clin. Infect. Dis.* 53, 732–742.

Swenson, L. C., Moores, A., Low, A. J., Thielen, A., Dong, W., Woods, C., Jensen, M. A., Wynhoven, B., Chan, D., Glascock, C., and Harrigan, P. R. (2010). Improved detection of CXCR4-using HIV by V3 genotyping: application of population-based and "deep" sequencing to plasma RNA and proviral DNA. *J. Acquir. Immune Defic. Syndr.* 54, 506–510.

Tapparel, C., Cordey, S., Junier, T., Farinelli, L., Van Belle, S., Soccal, P. M., Aubert, J. D., Zdobnov, E., and Kaiser, L. (2011). Rhinovirus genome variation during chronic upper and lower respiratory tract infections. *PLoS ONE* 6:e21163. doi: 10.1371/journal.pone.0021163

Trapnell, C., and Salzberg, S. L. (2009). How to map billions of short reads onto genomes. *Nat. Biotechnol.* 27, 455–457.

Turner, E. H., Ng, S. B., Nickerson, D. A., and Shendure, J. (2009). Methods for genomic partitioning. *Annu. Rev. Genomics Hum. Genet.* 10, 263–284.

Van Nimwegen, E., Crutchfield, J. P., and Huynen, M. (1999). Neutral evolution of mutational robustness. *Proc. Natl. Acad. Sci. U.S.A.* 96, 9716–9720.

Vandekerckhove, L., Verhofstede, C., Demecheleer, E., De Wit, S., Florence, E., Fransen, K., Moutschen, M., Mostmans, W., Kabeya, K., Mackie, N., Plum, J., Vaira, D., Van Baelen, K., Vandenbroucke, I., Van Eygen, V., Van Marck, H., Vogelaers, D., Geretti, A. M., and Stuyver, L. J. (2011). Comparison of phenotypic and genotypic tropism determination in triple-class-experienced HIV patients eligible for maraviroc treatment. *J. Antimicrob. Chemother.* 66, 265–272.

Vandenbroucke, I., Van Marck, H., Mostmans, W., Van Eygen, V., Rondelez, E., Thys, K., Van Baelen, K., Fransen, K., Vaira, D., Kabeya, K., De Wit, S., Florence, E., Moutschen, M., Vandekerckhove, L., Verhofstede, C., and Stuyver, L. J. (2010). HIV-1 V3 envelope deep sequencing for clinical plasma specimens failing in phenotypic tropism assays. *AIDS Res. Ther.* 7, 4.

Varela, I., Tarpey, P., Raine, K., Huang, D., Ong, C. K., Stephens, P., Davies, H., Jones, D., Lin, M.-L., Teague, J., Bignell, G., Butler, A., Cho, J., Dalgliesh, G. L., Galappaththige, D., Greenman, C., Hardy, C., Jia, M., Latimer, C., Lau, K. W., Marshall, J., Mclaren, S., Menzies, A., Mudie, L., Stebbings, L., Largaespada, D. A., Wessels, L. F. A., Richard, S., Kahnoski, R. J., Anema, J., Tuveson, D. A., Perez-Mancera, P. A., Mustonen, V., Fischer, A., Adams, D. J., Rust, A., On, W. C., Subimerb, C., Dykema, K., Furge, K., Campbell, P. J., Teh, B. T., Stratton, M. R., and Futreal, P. A. (2011). Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* 469, 539–542.

Varghese, V., Shahriar, R., Rhee, S. Y., Liu, T., Simen, B. B., Egholm, M., Hanczaruk, B., Blake, L. A., Gharizadeh, B., Babrzadeh, F., Bachmann, M. H., Fessel, W. J., and Shafer, R. W. (2009). Minority variants associated with transmitted and acquired HIV-1 nonnucleoside reverse transcriptase inhibitor resistance: implications for the use of second-generation nonnucleoside reverse transcriptase inhibitors. *J. Acquir. Immune Defic. Syndr.* 52, 309–315.

Vignuzzi, M., Stone, J. K., Arnold, J. J., Cameron, C. E., and Andino, R. (2006). Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* 439, 344–348.

Vrancken, B., Lequime, S., Theys, K., and Lemey, P. (2010). Covering all bases in HIV research: unveiling a hidden world of viral evolution. *AIDS Rev.* 12, 89–102.

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.

Wang, G. P., Sherrill-Mix, S. A., Chang, K. M., Quince, C., and Bushman, F. D. (2010). Hepatitis C virus transmission bottlenecks analyzed by deep sequencing. *J. Virol.* 84, 6218–6228.

Westbrooks, K., Astrovskaya, I., Campo, D., Khudyakov, Y., Berman, P., and Zelikovsky, A. (2008). "HCV quasispecies assembly using network flows," in *ISBRA 2008, LNBI 4983*, eds I. Măndoiu, R. Sunderraman, and A. Zelikovsky (Berlin Heidelberg: Springer-Verlag), 159–170.

WHO. (2012). *World Health Organization* [Online]. Available online at: www.who.int [Accessed 1 May 2012].

Wikipedia (2012). List of sequence alignment software [Online]. Available: http://en.wikipedia.org/wiki/List_of_sequence_alignment_software # Short - Read _ Sequence _ alignment [Accessed 1 May 2012].

Wilke, C. O. (2005). Quasispecies theory in the context of population genetics. *BMC Evol. Biol.* 5, 44.

Wilke, C. O., Wang, J. L., Ofria, C., Lenski, R. E., and Adami, C. (2001). Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature* 412, 331–333.

Willerth, S. M., Pedro, H. A., Pachter, L., Humeau, L. M., Arkin, A. P., and Schaffer, D. V. (2010). Development of a low bias method for characterizing viral populations using next generation sequencing technology. *PLoS ONE* 5:e13564. doi: 10.1371/journal.pone.0013564

Wu, X., Zhou, T., Zhu, J., Zhang, B., Georgiev, I., Wang, C., Chen, X., Longo, N. S., Louder, M., McKee, K., O'Dell, S., Perfetto, S., Schmidt, S. D., Shi, W., Wu, L., Yang, Y., Yang, Z. Y., Yang, Z., Zhang, Z., Bonsignori, M., Crump, J. A., Kapiga, S. H., Sam, N. E., Haynes, B. F., Simek, M., Burton, D. R., Koff, W. C., Doria-Rose, N. A., Connors, M., Mullikin, J. C., Nabel, G. J., Roederer, M., Shapiro, L., Kwong, P. D., and Mascola, J. R. (2011). Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* 333, 1593–1602.

Yang, X., Chockalingam, S. P., and Aluru, S. (2012). A survey of error-correction methods for next-generation sequencing. *Brief Bioinform*. doi: 10.1093/bib/bbs015. [Epub ahead of print]

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

Zagordi, O., Geyrhofer, L., Roth, V., and Beerenwinkel, N. (2010a). Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. *J. Comput. Biol.* 17, 417–428.

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

Zagordi, O., Töpfer, A., Prabhakaran, S., Roth, V., Halperin, E., and Beerenwinkel, N. (2012). "Probabilistic inference of viral quasispecies subject to recombination," in *RECOMB 2012, LNBI 7262*, ed B. Chor (Berlin Heidelberg: Springer-Verlag), 342–354.

Zell, R., Taudien, S., Pfaff, F., Wutzler, P., Platzer, M., and Sauerbrei, A. (2012). Sequencing of 21 varicella-zoster virus genomes reveals two novel genotypes and evidence of recombination. *J. Virol.* 86, 1608–1622.

Zhao, X., Palmer, L. E., Bolanos, R., Mircean, C., Fasulo, D., and Wittenberg, G. M. (2010). EDAR: an efficient error detection and removal algorithm for next generation sequencing data. *J. Comput. Biol.* 17, 1549–1560.

Conflict of Interest Statement: Karin J. Metzner has received travel grants and honoraria from Gilead, Roche Diagnostics, GlaxoSmithKline, Bristol-Myers Squibb, Tibotec, and Abbott, and has received research grants from Abbott, Gilead, and Roche Diagnostics. Huldrych F. Günthard has been an adviser and/or consultant for the following companies: GlaxoSmithKline, Abbott, Novartis, Gilead, Boehringer Ingelheim, Roche, Tibotec and Bristol-Myers Squibb, and has received unrestricted research and educational grants from Roche, Abbott, Bristol-Myers Squibb, GlaxoSmithKline, Gilead, Tibotec and Merck Sharp & Dohme (all money went to institution). The other authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 June 2012; paper pending published: 06 July 2012; accepted: 24 August 2012; published online: 11 September 2012.

*Citation: Beerenwinkel N, Günthard HF, Roth V and Metzner KJ (2012) Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front. Microbio. **3**:329. doi: 10.3389/fmicb.2012.00329*

This article was submitted to Frontiers in Virology, a specialty of Frontiers in Microbiology.