



ELSEVIER

Available online at www.sciencedirect.com

Ultra-deep sequencing for the analysis of viral populations

Niko Beerenwinkel^{1,2} and Osvaldo Zagordi^{1,2}

Next-generation sequencing allows for cost-effective probing of virus populations at an unprecedented level of detail. The massively parallel sequencing approach can detect low-frequency mutations and it provides a snapshot of the entire virus population. However, analyzing ultra-deep sequencing data obtained from diverse virus populations is challenging because of PCR and sequencing errors and short read lengths, such that the experiment provides only indirect evidence of the underlying viral population structure. Recent computational and statistical advances allow for accommodating some of the confounding factors, including methods for read error correction, haplotype reconstruction, and haplotype frequency estimation. With these methods ultra-deep sequencing can be more reliably used to analyze, in a quantitative manner, the genetic diversity of virus populations.

Addresses

¹ Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland

² Swiss Institute of Bioinformatics, Basel, Switzerland

Corresponding author: Beerenwinkel, Niko
(niko.beerenwinkel@bsse.ethz.ch)

Current Opinion in Virology 2011, 1:1–6

This review comes from a themed issue on
Virus evolution
Edited by Peter Simmonds and Esteban Domingo

1879-6257/\$ – see front matter
© 2011 Published by Elsevier B.V.

DOI [10.1016/j.coviro.2011.07.008](https://doi.org/10.1016/j.coviro.2011.07.008)

Introduction

Viruses exist in their hosts as dynamic ensembles of individual viral particles and integrated proviruses. The evolutionary dynamics of many viruses, including most RNA viruses such as HIV and HCV, is characterized by high turnover rates, large population sizes, and high mutation rates. Under these conditions, a large number of viral mutants are constantly produced, creating a large genetic diversity on which natural selection operates. Heterogeneous virus populations are often referred to as viral quasispecies [1–3] (Figure 1a). Genetic diversity plays a key role in the biology and medical treatment of viruses. In HIV, for example, it has been identified as an important factor of disease progression [4], pathogenesis [5], immune escape [6], vaccine design [7], and drug resistance development [8,9].

Until recently, the genetic diversity of a virus population, including the co-occurrence of mutations, could only be assessed by cloning individual viruses and applying Sanger capillary sequencing to each of them. Because this approach is time-consuming and involves a lot of labor, only few studies have surveyed virus populations in appreciable detail, never analyzing more than a few hundred viruses per sample. The situation has changed dramatically with the introduction of ultra-deep sequencing (UDS) based on next-generation sequencing (NGS) technologies [10]. This massively parallel sequencing approach can overcome the limitation of clonal Sanger sequencing by directly sequencing the mixed sample at high coverage of 10,000 or more reads per base pair.

UDS can detect low-frequency mutations and it provides substantial information on the structure of the population, that is, the set of different variants and their relative frequencies. The power of this approach for estimating the diversity of within-host virus populations has been recognized soon after the introduction of pyrosequencing [11], which was initially used for detecting low-frequency drug resistance mutations in HIV [12,13,14,15,16] and for analyzing HIV tropism and coreceptor usage [9,17–20]. The number and variety of viral UDS applications is increasing rapidly, including HCV transmission bottlenecks [21,22], HBV diversity and low-frequency drug resistance mutations [23,24], mixed influenza infections [25,26], and foot-and-mouth disease virus diversity [27].

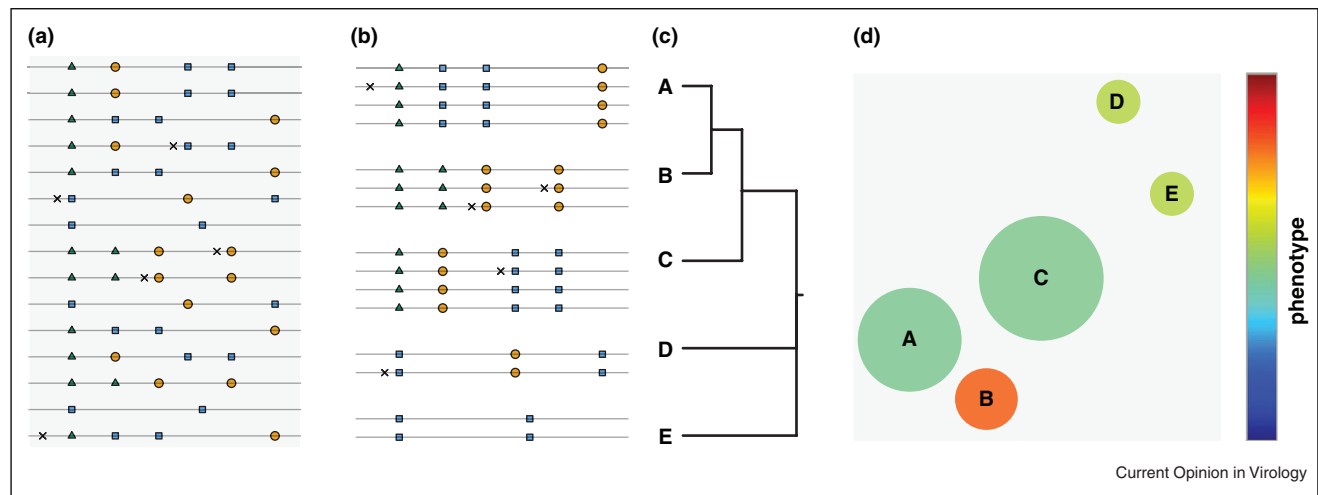
To date, pyrosequencing as commercialized by 454/RocheTM is the most frequently applied sequencing technology for these tasks, but other NGS platforms, including IlluminaTM [27,28] and ABI SOLiDTM, might also be used. All NGS approaches have in common that they produce short DNA segments, called reads, which provide only imperfect and incomplete information on the viral population structure. Sequencing errors and short read lengths complicate the analysis of UDS data obtained from viral quasispecies. In this review, we highlight some of the challenges associated with estimating a viral population from UDS data and we discuss computational and statistical methods for error correction, reconstruction of viral haplotypes (strains), and haplotype frequency estimation.

Challenges in the analysis of viral UDS data

While NGS platforms can be used to detect unknown viruses *de novo* [29], we focus here on the resequencing of known viruses. In the case of RNA viruses, reverse transcription is applied first. Then, PCR is used to amplify defined genomic regions, and the diverse PCR

2 Virus evolution

Figure 1



Analysis of viral quasispecies. A schematic representation of an intra-host virus population is displayed in (a), where each line corresponds to a haplotype and symbols indicate differences relative to a reference strain. In local haplotype reconstruction, all reads cover the sequence window. Once reads are clustered, it is easier to distinguish technical errors (crosses) from true biological variations (all other symbols). Haplotypes are identified as the consensus sequences of each group (denoted A, . . . , E) and their frequencies as the cluster sizes (b). The reconstructed viral quasispecies can then be further analyzed, for example, by phylogenetic methods (c). In (d), the virus population is displayed such that the size of each disc corresponds to the proportion of the respective haplotype in the population and the distances between them reflect evolutionary distance. The color coding suggests that haplotypes may be annotated with phenotypic properties, for example the level of drug resistance [32**].

products, or amplicons, are sequenced. The 454/RocheTM pyrosequencing platform is the most popular choice due to the longest reads of about 400 bp. Depending on the length of the amplicon, sequencing can be preceded by DNA fragmentation.

Although NGS technologies differ in several respects, all platforms currently in use involve steps of DNA library preparation, amplification, and sequencing by synthesis or ligation [10]. Fragment amplification is performed for each individual molecule either on beads captured in water-in-oil microreactors (emulsion PCR) or attached to a surface by bridge amplification. The amplified molecules are then sequenced in a massively parallel fashion, with 454/RocheTM pyrosequencing producing up to one million reads per run. An important consequence of single-molecule amplification is that each read obtained from the sequencing experiment originates from exactly one DNA molecule in the DNA library. Therefore, the read data can be regarded as a statistical sample of the original virus population.

However, reverse transcription, PCR amplification, and NGS are all error-prone processes. PCR can introduce point mutations and indels and it can also generate recombinant sequences, or chimeras, that are composed of two or more true template sequences. In addition, the relative frequencies of genetic variants can be perturbed due to selective amplification bias during PCR. Additional single-base errors can occur during emulsion

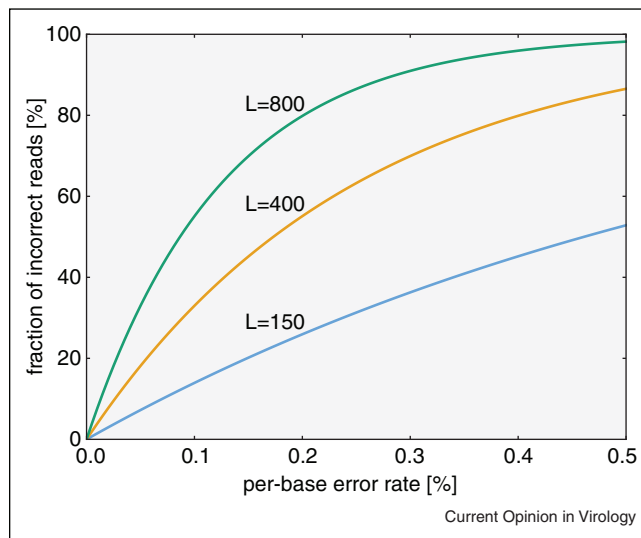
PCR. Finally, sequencing itself will introduce base substitution errors and indels. The impact of these errors on viral diversity studies can be enormous. To illustrate the effect, let us consider an error rate of 0.2% per base pair and a read length of 400 bp. Then the proportion of reads with at least one error is $1 - (1 - 0.002)^{400} = 0.551$. In other words, over 55% of the reads are incorrect. Thus, any diversity estimate based on the raw sequence data will be vastly inflated [30,31,32**]. In general, the fraction of erroneous reads increases with error rate and read length (Figure 2).

UDS data are not only noisy but also censored if the genomic region of interest is larger than the read length. Making inference about the genetic structure of a virus population based on such confounded and incomplete data is challenging. It involves several steps including filtering, alignment, and error correction of reads, and haplotype inference and frequency estimation. These tasks are not independent and some methods address only one step while others cover several.

Filtering and alignment

Filtering refers to removing reads of low quality from the data set. The read quality can be assessed from base quality scores provided by the NGS platform. Read quality tends to drop toward the end of the read, such that low-quality reads can be either truncated or discarded entirely. The original measurement from which the bases were called can also be used to filter reads, for

Figure 2



The number of incorrect reads in an UDS experiment. Displayed is $1 - (1 - \epsilon)^L$, the expected fraction of reads with at least one sequencing error, as a function of the per-base error rate ϵ , for three different read lengths L typical of Sanger sequencing ($L = 800$), 454/Roche™ ($L = 400$), and Illumina™ ($L = 150$). The 454/Roche error rate is around 0.1–0.5%. The graph illustrates that the number of reads with one or more errors increases with both the error rate and the read length. It is not meant as a comprehensive error analysis as it ignores other sources of variation (e.g., PCR chimeras) and does not distinguish different types of errors (e.g., base substitutions from indels).

example, the light signal intensity of pyrosequencing flowgrams [33*,34]. In resequencing studies, the remaining reads are aligned, or mapped, to a reference or consensus genome. This task can be accomplished by short read mappers or semiglobal alignment algorithms based on dynamic programming. The accuracy can further be improved by accounting for specific error patterns of the sequencer. For example, most errors in pyrosequencing reads are indels in homopolymeric regions [35] and their alignment can be optimized by using reduced gap costs in such regions [12*,33*]. A multiple sequence alignment of all reads can further enhance alignment quality, but is computationally expensive. It can be constructed from the pairwise read alignments, or by using customized algorithms [36].

Error correction

Separating true genetic variation from measurement noise is important for estimating the diversity of mixed samples. Viral quasispecies consist of many, evolutionarily related variants resulting from mutation and selection. Furthermore, many important viruses recombine within their hosts, including, for example, HIV, HBV, and HCV. The different haplotypes of quasispecies are typically very similar to each other and often form a connected mutant cloud. Because PCR errors arise by essentially the

same mechanisms *in vitro* as viral quasispecies *in vivo*, it is impossible to separate them from true mutations without additional assumptions or experiments. For example, PCR chimeras can only be detected in populations that are known not to recombine [33*]. In general, single-base substitution and recombination occur frequently during PCR and they represent the most severe limitation of UDS-based analyses of viral populations. The critical PCR step requires careful experimental design, including choice of polymerase, PCR conditions, and parallel runs.

The basic idea of error correction is that technical errors are randomly distributed and rare, whereas true mutations are sampled in proportion to their frequency in the population. Thus, a group of reads that are more similar to each other than to any other read is likely to represent a true haplotype (Figure 1b). Finding such groups is a clustering problem. For each cluster, the haplotype sequence is the consensus sequence (centroid), the haplotype frequency is the cluster size, and the technical noise is the within-cluster variation. Clustering is based on pairwise distances that can be computed from aligned reads or aligned flowgrams. Thus, it is applied locally to reads covering a common window of the genome. The result of this step is the local population structure and the corrected reads [32**,33*,37**,38,39*].

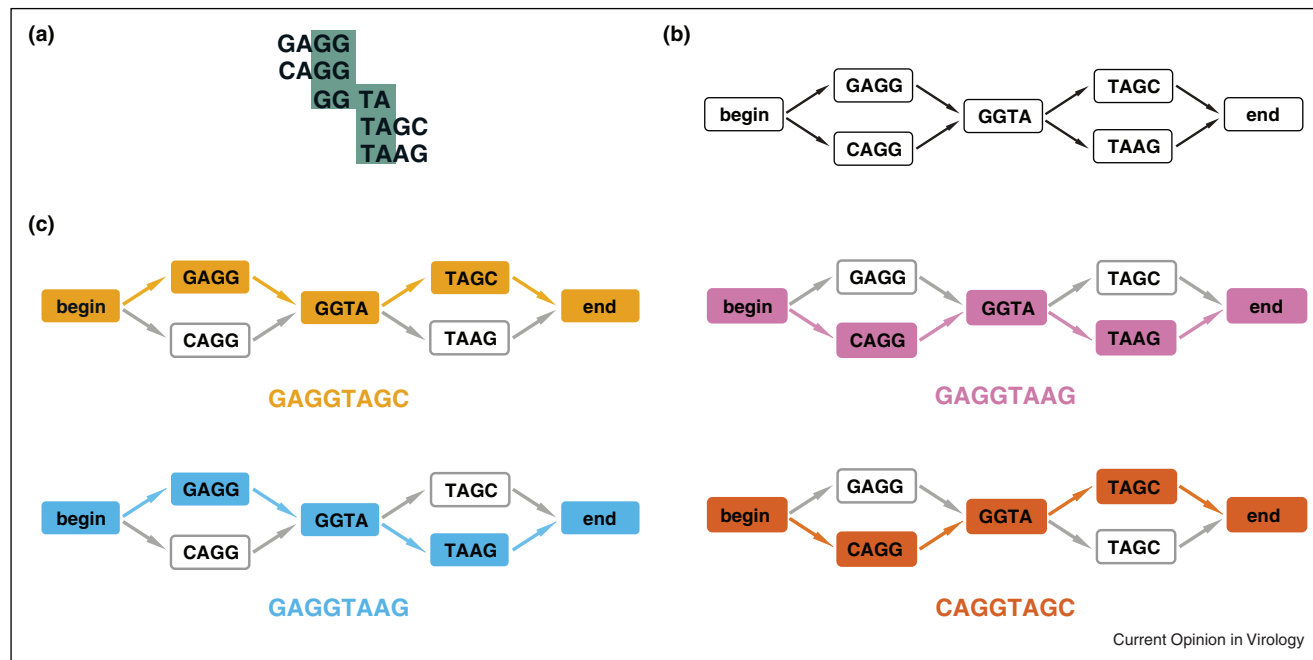
Different clustering algorithms have been proposed, some operating on flowgrams and some on the DNA sequences of the reads. A common difficulty is to infer the correct number of clusters (haplotypes) from the data. This problem has been addressed in a Bayesian fashion, which also yields the posterior probability of each haplotype, that is, the confidence in its prediction [32**,38]. The performance of error correction methods is difficult to assess, but initial control experiments suggest that clustering can reduce the overall error rate by a factor of at least 2–20. The performance increases with read length and, in particular, clustering is superior to calling mutations individually for each site, because it considers covariation. Finally, clustering has also been shown to outperform read calling based simply on read abundance [32**].

Global haplotype assembly

For global haplotype reconstruction, the short reads need to be assembled into longer contiguous strains. This problem is much harder than local haplotype reconstruction and it does not have a unique solution if intermediate genomic regions display lower genetic diversity (Figure 3). The clustering approach can be extended to this situation to obtain a hidden Markov model with unobserved global haplotypes [40], which can also be treated in a Bayesian fashion [41]. The latter approach makes explicit use of local clustering solutions and tries to extend them to larger regions subject to the local diversity constraints.

4 Virus evolution

Figure 3



Global haplotype assembly. Reads covering different regions of the genome can be assembled into longer haplotypes. In this example, five reads, each of lengths four, cover a genomic region of length 8 bp (a). After alignment, the reads GAGG and CAGG both overlap with GGTA and they all agree on the overlap (GG, shaded). Similarly, GGTA agrees on its overlap with both TAGC and TAAG (TA, shaded). In the read graph (b), reads agreeing on their non-empty overlaps are connected by directed edges. A possible haplotype is constructed by traversing the read graph from the 'begin' node to the 'end' node. In this example, four different haplotypes are compatible with the observed reads (c) and at least two are necessary to explain all of them. There are two different minimal haplotype sets explaining all reads, namely GAGGTAGC and GAGGTAAG (orange and purple), and GAGGTAAG and CAGGTAGC (blue and red).

Alternatively, global haplotype assembly methods have been proposed based on the read graph. The nodes of this graph correspond to unique, error-corrected reads, and reads are connected by a directed edge if they agree on their nonempty overlap as illustrated in Figure 3. Each maximal path in the read graph corresponds to a possible haplotype for which local evidence exists. Employing a

parsimony principle, several combinatorial algorithms have been proposed to extract a minimal subset of haplotypes that explain all observed reads [37,42,43,44]. This assembly into candidate haplotypes is then followed by estimation of their frequencies [37]. Because error correction and haplotype inference are separated, this approach is computationally more efficient than global

Table 1

Computational methods for viral quasispecies reconstruction.

| Method | Error correction | Global assembly | Confidence values | Software available | Applications | References |
|------------------------|------------------|-----------------|-------------------|--------------------|---------------|---|
| ShoRAH | Yes | Yes | Yes | Yes ^a | HIV, HCV | [37 ^{**} ,38,39 [*]] |
| ViSpA | Yes | Yes | No | Yes ^b | HIV, HCV | [43 [*] ,44] |
| Jojic <i>et al.</i> | Yes | Yes | No | No | HIV | [40] |
| AmpliconNoise | Yes | No | No | Yes ^c | 16S rRNA, HCV | [31,33 [*]] |
| PredictHaplo | Yes | Yes | Yes | Yes ^d | HIV | [41] |
| Prosperi <i>et al.</i> | No | Yes | No | No | HBV | [42] |

Listed are all published methods to date that have been applied to at least one real virus population. Each method is characterized by its ability to correct read errors, to assemble global haplotypes, and to provide confidence values with its predictions. Software is available for four of the methods. Most applications concern HIV, HBV, and HCV.

^a <http://www.cbg.ethz.ch/software/shorah>.

^b <http://alla.cs.gsu.edu/~software/VISPA/vispa.html>.

^c <http://code.google.com/p/ampliconnoise/>.

^d http://www.cs.unibas.ch/personen/roth_volker/HivHaploTyper.

clustering, but more sensitive to miscorrections. For the analysis of highly diverse populations, for example, those resulting from multiple infections with different subtypes, *de novo* sequence assembly programs can also be useful [26]. The result of haplotype assembly is the predicted set of DNA sequences in the sample and their relative frequencies (Figure 1c,d).

Conclusions

UDS can be used to assess the diversity of virus populations, but several pitfalls can confound the analysis, including PCR chimeras. PCR and sequencing errors can be corrected to some extent by local haplotype inference, while global haplotype assembly is more challenging and limited by the structure of the underlying population. Some of the described methods have been implemented in software packages (Table 1) and tested on simulated and on experimental control data. However, all programs are relatively new and more extensive validation and comparison is necessary to better understand their behavior and performance. It is also likely that new software will be developed in the near future.

UDS-based viral quasispecies reconstruction offers a simple and economic way to obtain a snapshot of the entire virus population. This snapshot is imperfect, but it can be improved substantially by careful data analysis. Further advancements can be expected from improved and new NGS technologies, such as single-molecule sequencing, and from novel computational and statistical methods. Therefore, UDS will play an increasingly important role in the analysis of virus populations with a wide range of applications including viral evolution, epidemiology, antiviral therapy, and forensics.

Acknowledgement

Part of this research has been funded by the Swiss National Science Foundation under grant number CR32I2_127017.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- of outstanding interest

1. Eigen M: **Selforganization of matter and the evolution of biological macromolecules.** *Naturwissenschaften* 1971, **58**:465-523.
 2. Domingo E, Holland JJ: **RNA virus mutations and fitness for survival.** *Annu Rev Microbiol* 1997, **51**:151-178.
 3. Nowak MA, May RM: *Virus Dynamics.* Oxford University Press; 2000
 4. Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H, Gupta P, Rinaldo CR, Learn GH, He X *et al.*: **Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection.** *J Virol* 1999, **73**:10489-10502.
 5. Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R: **Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population.** *Nature* 2006, **439**:344-348.
 6. Nowak MA, Anderson RM, McLean AR, Wolfs TF, Goudsmit J, May RM: **Antigenic diversity thresholds and the development of AIDS.** *Science* 1991, **254**:963-969.
 7. Gaschen B, Taylor J, Yusim K, Foley B, Gao F, Lang D, Novitsky V, Haynes B, Hahn BH, Bhattacharya T *et al.*: **Diversity considerations in HIV-1 vaccine selection.** *Science* 2002, **296**:2354-2360.
 8. Johnson JA, Li J-F, Wei X, Lipscomb J, Irlbeck D, Craig C, Smith A, Bennett DE, Monsour M, Sandstrom P *et al.*: **Minority HIV-1 drug resistance mutations are present in antiretroviral treatment-naïve populations and associate with reduced treatment efficacy.** *PLoS Med* 2008, **5**:e158.
 9. Tsbiris AMN, Korber B, Arnaout R, Russ C, Lo C-C, Leitner T, Gaschen B, Theiler J, Paredes R, Su Z *et al.*: **Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo.** *PLoS ONE* 2009, **4**:e5683.
 10. Mardis ER: **Next-generation DNA sequencing methods.** *Annu Rev Genomics Hum Genet* 2008, **9**:387-402.
 11. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z *et al.*: **Genome sequencing in microfabricated high-density picoliter reactors.** *Nature* 2005, **437**:376-380.
 12. Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW: **Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance.** *Genome Res* 2007, **17**:1195-1201.
- First application of pyrosequencing to estimate the diversity of a virus population.
13. Hoffmann C, Minkah N, Leipzig J, Wang G, Arens MQ, Tebas P, Bushman FD: **DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations.** *Nucleic Acids Res* 2007, **35**:e91.
 14. Le T, Chiarella J, Simen BB, Hanczaruk B, Egholm M, Landry ML, Dieckhaus K, Rosen MI, Kozal MJ: **Low-abundance HIV drug-resistant viral variants in treatment-experienced persons correlate with historical antiretroviral use.** *PLoS ONE* 2009, **4**:e6079.
 15. Simen BB, Simons JF, Hullsiek KH, Novak RM, Macarthur RD, Baxter JD, Huang C, Lubeski C, Turenchalk GS, Braverman MS *et al.*: **Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naïve patients significantly impact treatment outcomes.** *J Infect Dis* 2009, **199**:693-701.
 16. Codoñer FM, Pou C, Thielen A, García F, Delgado R, Dalmau D, Alvarez-Tejado M, Ruiz L, Clotet B, Paredes R: **Added value of deep sequencing relative to population sequencing in heavily pre-treated HIV-1-infected subjects.** *PLoS ONE* 2011, **6**:e19461.
 17. Knoepfel SA, Giallonardo FD, Däumer M, Thielen A, Metzner KJ: **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* 2011, **171**:329-338.
 18. Swenson LC, Mo T, Dong WWY, Zhong X, Woods CK, Jensen MA, Thielen A, Chapman D, Lewis M, James I *et al.*: **Deep sequencing to infer HIV-1 co-receptor usage: application to three clinical trials of maraviroc in treatment-experienced patients.** *J Infect Dis* 2011, **203**:237-245.
 19. Däumer M, Kaiser R, Klein R, Lengauer T, Thiele B, Thielen A: **Genotypic tropism testing by massively parallel sequencing: qualitative and quantitative analysis.** *BMC Med Inform Decis Mak* 2011, **11**:30.
 20. Dybowski JN, Heider D, Hoffmann D: **Structure of HIV-1 quasi-species as early indicator for switches of co-receptor tropism.** *AIDS Res Ther* 2010, **7**:41.
 21. Wang GP, Sherrill-Mix SA, Chang K-M, Quince C, Bushman FD: **Hepatitis C virus transmission bottlenecks analyzed by deep sequencing.** *J Virol* 2010, **84**:6218-6228.
 22. Bull RA, Luciani F, McElroy K, Gaudieri S, Pham ST, Chopra A, Cameron B, Maher L, Dore GJ, White PA, *et al.* **Sequential**

6 Virus evolution

- bottlenecks drive viral evolution in early acute hepatitis C virus infection.** *PLoS Pathogens*, in press.
23. Solmone M, Vincenti D, Prosperi MCF, Bruselles A, Ippolito G, Capobianchi MR: **Use of massively parallel ultradeep pyrosequencing to characterize the genetic diversity of hepatitis B virus in drug-resistant and drug-naïve patients and to detect minor variants in reverse transcriptase and hepatitis B S antigen.** *J Virol* 2009, **83**:1718-1726.
 24. Margeridon-Thermet S, Shulman NS, Ahmed A, Shahriar R, Liu T, Wang C, Holmes SP, Babrzadeh F, Gharizadeh B, Hanczaruk B *et al.*: **Ultra-deep pyrosequencing of hepatitis B virus quasispecies from nucleoside and nucleotide reverse-transcriptase inhibitor (NRTI)-treated patients and NRTI-naïve patients.** *J Infect Dis* 2009, **199**:1275-1285.
 25. Ghedin E, Fitch A, Boyne A, Griesemer S, DePasse J, Bera J, Zhang X, Halpin RA, Smit M, Jennings L *et al.*: **Mixed infection and the genesis of influenza virus diversity.** *J Virol* 2009, **83**:8832-8841.
 26. Ramakrishnan MA, Tu ZJ, Singh S, Chockalingam AK, Gramer MR, Wang P, Goyal SM, Yang M, Halvorson DA, Sreevatsan S: **The feasibility of using high resolution genome sequencing of influenza A viruses to detect mixed infections and quasispecies.** *PLoS ONE* 2009, **4**:e7105.
 27. Wright CF, Morelli MJ, Thébaud G, Knowles NJ, Herzyk P, Paton DJ, Haydon DT, King DP: **Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing.** *J Virol* 2011, **85**:2266-2275.
 28. Renzette N, Bhattacharjee B, Jensen JD, Gibson L, Kowalik TF: **Extensive genome-wide variability of human cytomegalovirus in congenitally infected infants.** *PLoS Pathog* 2011, **7**:e1001344.
 29. Palacios G, Druce J, Du L, Tran T, Birch C, Briese T, Conlan S, Quan P-L, Hui J, Marshall J *et al.*: **A new arenavirus in a cluster of fatal transplant-associated diseases.** *N Engl J Med* 2008, **358**:991-998.
 30. Kunin V, Engelbrektson A, Ochman H, Hugenholtz P: **Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates.** *Environ Microbiol* 2010, **12**:118-123.
 31. Quince C, Lanzén A, Curtis TP, Davenport RJ, Hall N, Head IM, Read LF, Sloan WT: **Accurate determination of microbial diversity from 454 pyrosequencing data.** *Nat Methods* 2009, **6**:639-641.
 32. Zagordi O, Klein R, Däumer M, Beerenwinkel N: **Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies.** *Nucleic Acids Res* 2010, **38**:7400-7409.
Detailed comparison of probabilistic reads clustering to *ad hoc* read counting methods on UDS data from control experiments and clinical HIV samples.
 33. Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ: **Removing noise from pyrosequenced amplicons.** *BMC Bioinformatics* 2011, **12**:38.
- Efficient local haplotype clustering based on flowgrams. Describes the AmpliconNoise software.
34. Reeder J, Knight R: **Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions.** *Nat Methods* 2010, **7**:668-669.
 35. Huse S, Huber J, Morrison H, Sogin M, Welch D: **Accuracy and quality of massively parallel DNA pyrosequencing.** *Genome Biol* 2007, **8**:R143.
 36. Saeed F, Khokhar A, Zagordi O, Beerenwinkel N: **Multiple sequence alignment system for pyrosequencing reads.** In *BICoB 2009 – Bioinformatics and Computational Biology*. Edited by Rajasekaran S. Springer; 2009:362-375.
 37. Eriksson N, Pachter L, Mitsuya Y, Rhee S-Y, Wang C, Gharizadeh B, Ronaghi M, Shafer RW, Beerenwinkel N: **Viral population estimation using pyrosequencing.** *PLoS Comput Biol* 2008, **4**:e1000074.
Introduces error correction and local haplotype inference via read clustering, parsimonious global haplotype reconstruction, and haplotype frequency estimation.
 38. Zagordi O, Geyrhofer L, Roth V, Beerenwinkel N: **Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction.** *J Comput Biol* 2010, **17**:417-428.
 39. Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N: **ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data.** *BMC Bioinformatics* 2011, **12**:119.
Describes the ShoRAH software.
 40. Jojic V, Hertz T, Jojic N: **Population sequencing using short reads: HIV as a case study.** In *Pacific Symposium on Biocomputing*. Edited by Altman RB, Dunker AK, Hunter L, Murray T, Klein TE. World Scientific; 2008:114-125.
 41. Prabhakaran S, Rey M, Zagordi O, Beerenwinkel N, Roth V: **HIV haplotype inference using a constraint-based Dirichlet process mixture model.** *NIPS Workshop on Machine Learning in Computational Biology*. 2010.
 42. Prosperi MCF, Prosperi L, Bruselles A, Abbate I, Rozera G, Vincenti D, Solmone MC, Capobianchi MR, Ulivi G: **Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing.** *BMC Bioinformatics* 2011, **12**:5.
 43. Astrovskaya I, Tork B, Mangul S, Westbrooks K, Mandoiu I, Balfe P, Zelikovsky A: **Inferring viral quasispecies spectra from 454 pyro-sequencing reads.** *BMC Bioinformatics* 2011, **12**:S1.
Improved haplotype assembly by accounting for sequencing errors at various steps. Describes the ViSpA software.
 44. Westbrooks K, Astrovskaya I, Campo D, Khudyakov Y, Berman P, Zelikovsky A: **HCV quasispecies assembly using network flows.** In *Proceedings of the Bioinformatics Research and Applications, Fourth International Symposium, ISBRA 2008*. Edited by Mandoiu II, Sunderraman R, Zelikovsky A. *Proceedings of the Bioinformatics Research and Applications, Fourth International Symposium, ISBRA 2008* Atlanta, GA, USA, May 6-9 2008: Springer; 2008:159-170.