Chapter 5
HIV-1 Sequencing

Shelby L. O'Connor

Introduction

There is a great interest in deep sequencing RNA virus populations, including HIV /SIV, influenza, hepatitis, and Ebola. The RNA polymerases that are critical for replication of many RNA viruses have much lower fidelity than DNA polymerases employed during genome replication [1]. This lower fidelity facilitates the accumulation of nucleotide variants into a progeny viral genome during each replication. This can have enormous consequences on immune evasion, zoonotic events, drug resistance, and pathogenesis [2–7]. With such an enormity of hypotheses to test, there has been an increasing interest in deep sequencing virus populations, and numerous methods have been developed alongside this interest. From a proteomics perspective, these accumulated viral polymorphisms can impact the degree of coverage one can obtain from a proteomics experiment. This chapter focuses exclusively on the evolution of deep sequencing HIV and SIV populations as a means to understand the genetic makeup of these virus populations that contributes to their behavior in vivo. With this understanding the reader should be able to learn how to obtain detailed sequence information to build custom databases that will inform their subsequent proteomics experiments.

S.L. O'Connor, Ph.D. (✉)
Department of Pathology and Laboratory Medicine, University of Wisconsin-Madison,
555 Science Drive, Madison, WI 53711, USA
e-mail: slfeinberg@wisc.edu

D.R.M. Graham, D.E. Ott (eds.), *HIV-1 Proteomics*,
DOI 10.1007/978-1-4939-6542-7_5

Reasons Why Understanding HIV/SIV Variation Is Important from a Biological and Proteomics Perspective

HIV/SIV can evade host immune responses. The first data providing evidence that the accumulation of mutations in HIV can lead to escape from CD8 T cells was presented in 1991 [8] and then shown more clearly a few years later [9–11]. This was followed by data in SIV-infected nonhuman primates to demonstrate that immune escape from CTL occurs during acute and chronic infection [12, 13]. Evidence of HIV/SIV escape from antibodies was also shown on several occasions [14–17]. These important findings required the collection of sequence data from virus populations replicating within individuals. This evidence provided an explanation for why host immune responses are ultimately unable to contain replication of HIV in most infected individuals. Even though immune escape has been well documented, there is an ongoing need to assess the evolutionary path of immune escape in the context of different host genetic backgrounds and in the face of other immune modulatory factors. Glycosylation plays an important role in HIV/SIV host immune evasion [18]. Since many mass spectrometry algorithms reject information lacking consensus glycosylation sequences, for these studies, high sequence coverage is necessary to properly interpret mass spectrometry data.

In addition to evasion of natural immune responses, there are concerns over evasion of HIV from vaccine-elicited immune responses. From past HIV vaccine trials, we have learned that the presence of vaccine-elicited immune responses can select for transmission of virus sequences that are less likely to be neutralized by host immune responses [19–21]. Understanding whether vaccine-elicited immune responses can select for transmission of viruses with specific sequences or affect the emergence of immune escape variants after HIV transmission is essential for evaluating vaccine efficacy. While newer algorithms are much better at assigning polymorphisms from mass spectrometry data alone, having extensive information on polymorphisms present in the virus population assists in the proper assignment of spectra to peptides.

HIV viruses can accumulate mutations such that viral proteins become resistant to antiretroviral medications. Antiretroviral treatment is the most commonly used approach to treat people with HIV. It is well established that strict adherence to antiretroviral drug regimens is essential to prevent the emergence of drug-resistant mutations [22, 23]. To prevent widespread circulation of drug-resistant variants, there is ongoing global surveillance of HIV drug-resistant viruses, and there are efforts to implement low-cost sequencing protocols in these countries to expand these programs [24–27]. By improving our tools to track HIV drug-resistant mutations, there will be better global surveillance of transmitted HIV drug-resistant strains so that numerous individuals are not vulnerable to HIV infection without the option for using antiretrovirals. In this case obtaining sequence information about known mutations related to drug resistance might inform an experiment so that viral enzymes are enriched or targeted in the analysis. Also obtaining detailed sequence information could assist in better quantitation of the relative abundance of the mutation using targeted proteomics.

Globally, circulating strains of HIV are diverse and still evolving. There are at least nine genetic subtypes (clades) and numerous circulating recombinant forms (CRFs) of HIV worldwide [28, 29]. These different subtypes can differ by approximately 30 % in *env*, thus making them quite distinct virus groups [30]. Although it has been speculated that HIV transmission varies for the different subtypes, there is no significant in vivo evidence for this [28]. Still, there is evidence that there are differences in disease progression of antiretroviral-naïve individuals infected with different HIV subtypes and CRFs [29, 31–33]. There is no ecological reason why these different subtypes of HIV will remain localized to their current location; the opposite is true, as there have been several introductions of HIV subtypes into new regions throughout the epidemic [34]. Thus, ongoing monitoring of the prevalence of the different subtypes of HIV will help predict how the epidemic may change at an epidemiological level.

In addition to surveillance of HIV in humans, monitoring viruses in wild nonhuman primates will help either prevent the next zoonotic event or track the origins of the zoonotic event that may occur. There have been at least four instances of zoonotic SIV transmission events leading to the groups of HIV circulating globally [35]. Naturally, this raises concerns that there will be future zoonotic transmission events that will introduce another SIV into the human population or, perhaps, a different RNA virus. Tools to monitor viruses circulating in wild nonhuman primates thus provide a view on the underlying world of existing viruses that may have the potential to jump species. These sorts of surveillance studies are needed to prevent future outbreaks. Thus, we posit that sequencing is the backbone of a successful viral proteomics experiment.

The Evolution of HIV/SIV Sequencing Technologies

The description of the first full-length sequence of HIV (called HTLV-III) in 1985 was performed using Maxam-Gilbert and Sanger techniques [36]. This was followed by Sanger sequencing of the first full-length clone of SIV [37]. These revolutionary studies paved the way for our ability to understand the complex nature of HIV/SIV sequence evolution and diversity.

Over time, Sanger sequencing methods improved [38]. The development of slab gel sequencers that took advantage of fluorescently labeled dideoxynucleotides allowed for higher-throughput DNA sequencing. In many cases, plasmids containing virus genes were sequenced, with the assumption that a single clone was derived from a single virus. Sequence lengths of about 700 bp could be obtained, which took longer, but this approach generated information about linked mutations. At its maximum capacity, the ABI 377 sequencers could sequence 96 samples per 10 h run. This translated into the interrogation of 192 viruses in a 24-h period. The subsequent introduction of capillary DNA sequencing further increased throughput. With an ABI 3730 capillary sequencer, it was possible to sequence 48 plasmids per 2 h run. This translated into the interrogation of 576 viruses in a 24-h period. At that time, these were astounding numbers for sequencing numerous HIV/SIV genomes in a single day.

While sequencing plasmids was informative, it was time consuming and expensive. A transition was then made from sequencing plasmids to sequencing bulk PCR products generated from SIV cDNA. With this approach, it was possible to explore the bulk virus population [39]. Bulk sequencing was limited, however, because it was only possible to identify evidence that a specific nucleotide position was accumulating variants in the total virus population. This approach gave a better perspective of the evolving virus population, but it could not be used to quantify the frequency of a given nucleotide variant in a population.

A benefit of Sanger sequencing is the relatively long length of sequencing reads (approximately 400–700 bp), facilitating analyses of strings of nucleotide sequences. This key advantage has kept the use of Sanger sequencing in favor for single genome amplification (SGA) sequencing of HIV/SIV [40]. SGA approaches have been instrumental in defining transmitted/founder viruses that initiate an HIV or SIV infection [41, 42]. Unfortunately, the cost and time required to sequence large virus populations using this technology has made it somewhat obsolete for characterizing total virus population.

In 2005, a sequencing revolution began with the description of pyrosequencing technology [43]. This discovery opened doors to increase the throughput of sequencing HIV/SIV. With this technology came a transition to pyrosequence HIV/SIV. The GS 20 and GS FLX were early instruments typically found at core facilities, and they were used for initial virus pyrosequencing experiments [44–47]. These were expensive and difficult to use, so Roche/454 developed the benchtop GS Jr. sequencer in 2009 that was adopted by many labs. Using this technology, a single piece of PCR-amplified DNA was attached to a bead, clonally amplified, and then pyrosequenced. This approach yielded about 100,000 sequences on a GS Jr. in a single run that took about 24 h to process. Once the GS Jr. became commonplace, it was used routinely for sequencing virus populations for the next few years [48–52]. Each sequence generated by the Roche/454 pyrosequencing technology was thought to represent a single piece of DNA that was part of a PCR product that was derived from an original virus particle. Of course, PCR amplification of the same virus sequence was still expected. Yet, if all viruses were equally amplified, then the distribution of virus sequences would be a reasonable reflection of the relative viruses in the overall population. Overall, this meant that 100,000 viruses could be interrogated in 24 h.

The length of reads that can be obtained on the Roche/454 platform can vary from 400 to 1000 bp. This is dependent upon the sample preparation and the sequencer being used. This length is on par with that obtained by Sanger technology, so linkage and haplotype information can be obtained from any given read. In contrast, a major problem associated with pyrosequencing was a high error rate frequently associated with the difficulties defining the number of specific nucleotides present in homopolymers [47, 53]. This is especially a major concern when sequencing drug resistance mutations in HIV because many drug resistance mutations occur in homopolymer regions. To accurately characterize mutations in HIV, an assortment of data filtering steps and analysis tools were developed to ensure that a mutation in a homopolymer region was authentic [52, 54, 55]. Consequently, this technique is useable, but not ideal, for widespread screening of drug-resistant HIV.

Sequencing by synthesis using the Illumina sequencing platform is the current workhorse for deep sequencing HIV/SIV. Solexa, the original company to develop sequencing-by-synthesis technology, was formed in 1998. It was acquired by Illumina in 2007, and since then, the use of this technology has increased dramatically [56]. When originally implemented, Solexa Genome Analyzers were used for sequencing but were replaced with Illumina HiSeq machines in 2010. Since then, these machines have been available in core facilities, but are expensive to operate. In 2011, Illumina developed the benchtop MiSeq that was more amenable for use in smaller labs. This machine has greatly improved the accessibility for larger numbers of labs to incorporate deep sequencing in their research portfolios [57]. One drawback to this technology is that it splits DNA into a lot of small pieces, and then the small pieces are sequenced. This makes it difficult to generate long contiguous sequences, and it can be difficult to create analysis pipelines to process the data. Currently, the length of reads ranges from 125 to 300 bp. When merging paired end reads, it is possible to sequence a single DNA fragment of about 500 bp. As these lengths have approached the capability of Sanger sequencing, it has made deep sequencing of HIV/SIV substantially more practical. In addition, sequences generated by this technology are not subject to the same homopolymer errors that plague pyrosequencing. Further, a single MiSeq run can sequence ten million pieces of DNA in a run that lasts 3 days. If each sequence came from a single piece of DNA generated in a PCR amplification, then this means that more than three million viruses can be interrogated in a 24-h period using the Illumina MiSeq.

Given the advantages of the Illumina MiSeq (high fidelity, low cost, increasingly longer read length), there has been a movement to focus development efforts to deep sequence HIV/SIV on the Illumina platform. For this reason, the remainder of this chapter will focus on the methodology and analysis approaches used to sequence HIV/SIV on the Illumina MiSeq.

HIV/SIV Deep Sequencing Methodologies

Pre-experimental Questions

Before initiating any experiment designed to deep sequence HIV or SIV, it is necessary to carefully identify the goals for the experiment and the subsequent goals one wishes to obtain with a complementary mass spectrometry experiment. The following section outlines some key questions to address before undertaking the task of deep sequencing.

Question 1: Am I interested in characterizing SNPs across the entire coding sequence of the genome?

With current technologies, it is difficult to sequence the complete coding region of a single virus without cloning it into a plasmid. It is possible, however, to obtain the frequency of each independent nucleotide at every position throughout the HIV/

SIV coding sequence. For this approach, long amplicons are generated and then fragmented into libraries. The pieces in the libraries are tagged and sequenced. The sequenced pieces can then be mapped back to a reference sequence and the frequency of each nucleotide at each position can be measured. Even though this approach provides little information about linkage, it can provide useful information about ongoing evolution of sequence variants in the virus population that may confer resistance to host immune responses or antiretroviral drugs [26, 45, 58].

Question 2: Do I want to determine linked sequences across a small section of the genome?

It is entirely feasible to obtain linkage information about short stretches of virus sequences, as long as the maximum nucleotide distance is realistically within the limitations of the sequencing methodology. For this approach, relatively small amplicons (~300–500 bp) are generated, tagged, and sequenced from both ends of the DNA in individual clusters. The paired sequences from the same cluster are merged, and then each merged piece of DNA is treated as a single DNA sequence. Sequence information is obtained across the entire piece of DNA and for each piece of DNA from the PCR product that was sequenced. The frequency that a certain stretch of nucleotides appears in the total population can then be calculated. Even though this approach is limited to a small region of the genome, it is an effective way to quantify variation within an entire T cell epitope, investigate linked variants within the envelope gene, or track individual virus populations replicating in an animal.

Question 3: Do I know the sequence of my SIV inoculum or the sequence of the virus population soon after HIV infection?

It is important to know the sequences of HIV/SIV that are replicating during acute infection for two reasons. First, it is important to know the sequences of the "baseline" virus population so that sequences present at later time points can be compared back to early virus sequences. Second, it is important to design primers in conserved virus regions that will amplify virus sequences in an unbiased manner. If the sequences of the early virus populations or the SIV inoculum are not known, these can be obtained through unbiased sequencing of viral cDNA. This approach has been used to discover new strains of SIV and other RNA viruses from the plasma of wild nonhuman primates [3, 48, 59].

Question 4: What is the virus titer?

Knowing the virus titer is important for experimental setup. To sequence SIV and HIV, there obviously needs to be detectable virus so that there is material to sequence. When titers are low, concentrating the virus becomes essential. In addition, long stretches of RNA are difficult to amplify by RT-PCR when there is a low amount of starting material, and thus amplifying shorter segments by RT-PCR is more practical. When titers are high ($>10^4$ copies/ml of plasma), then RT-PCR occurs readily and yields large amounts of viral cDNA for sequencing. Given the need to amplify the viral cDNA so there is product available for sequencing, it is expected that templates will be resampled. Although idealistic, it is reasonable to expect that the PCR products generated by amplification of templates using primers located in conserved regions are a good representation of the starting virus population [48].

It is useful to consider the following hypothetical example when assessing how virus titer will influence the results. Imagine that there are 20 vRNA templates present at a 50/50 ratio of the variant to the wild type. If only 50 % of these templates are amplified by RT-PCR, then amplification of five wild-type and five variant templates are needed to observe a 50/50 ratio in the final data set. If, by chance, this changes by one template, then four wild-type and six variant templates are amplified, so the ratio of wild type to variant is 40/60 in the final data set. In contrast, imagine that there are 2000 vRNA templates present at a 50/50 ratio of the variant to the wild type. If only 50 % of these templates are amplified by RT-PCR, then amplification of 500 wild-type and 500 variant templates are needed to observe a 50/50 ratio of the variant to the wild type. If, by chance, this changes by one template, then 499 wild-type and 501 variant templates are amplified, so the ratio of wild type to variant will still be approximately 50/50 in the final data set. Thus, slight perturbations in template amplification are less apparent when the titers are high. When analyzing low titer samples, more accurate information about the sequence of the virus population can be obtained by sequencing multiple independent samples.

Sequencing Methodology

Once the above questions have been carefully considered, there are some standard HIV/SIV sequencing pipelines that can be easily modified for a given experimental question. In the following sections, three methodologies will be outlined: (a) unbiased whole genome, (b) PCR-amplified whole genome, and (c) PCR-amplified short segments.

a. Unbiased whole genome

   This approach is typically used to sequence an entire viral genome when very little is known about the virus used in the study and there are no primers available to amplify the virus segments. Initially, viral RNA is isolated from a sample. Random hexamers are used to prime the RNA to initiate synthesis of the first strand of cDNA. The RNA is degraded and then the second strand of cDNA synthesis is completed. The double-stranded cDNA segments are fragmented into a library using the Nextera tagmentation kits (Illumina). Library ID tags are added and the fragments are sequenced on an instrument, such as the MiSeq. Analysis of the data set will be described below.

   This unbiased whole genome approach has been used extensively to discover RNA viruses present in wild nonhuman primates [3, 4, 60–63]. We have also used this approach routinely to sequence virus stocks for clients who have little a priori knowledge of the inoculum sequence. In an experiment comparing data obtained by this unbiased approach vs. amplicon-based approaches (below), the data appears to be quite similar [48]. Notably, the unbiased approach requires a high titer, or there will otherwise be insufficient starting material to yield adequate cDNA of the entire viral genome.

b. Amplicon-based whole genome

This approach is typically used to sequence an entire viral genome when there is an extensive amount of information known about the inocula. Primers specific for conserved sites in the virus genome are designed and then used to amplify viral gene segments by RT-PCR. Similar to the double-stranded cDNA created using the unbiased approach, the double-stranded PCR products are fragmented into libraries with the Nextera technology, tags are added, and then sequenced on an instrument, such as the MiSeq.

This biased sequencing approach is advantageous for many reasons. Unlike the unbiased approach, samples with low viral titers can be used as starting material, a critical need when trying to sequence from individuals who have low viral loads attributed to either natural or drug-mediated control. In addition, there is an extensive amount of flexibility inherent to this approach. As long as a PCR product can be generated from nucleic acid starting material, then it can be fragmented into a library and sequenced. This approach has been used to interrogate the frequencies of nucleotides across the viral genome to quantify variation in T cell epitopes [58, 64]. Besides analyses of entire viral genomes, it is possible to amplify a single gene, such as HIV polymerase, to characterize drug-resistant mutations [26].

One limitation to fragmenting DNA into a library is that sequences of variable length are created that have diverse start and stop positions. These inconsistencies do not affect the mapping of reads to a reference sequence, but they limit the information that can be gained about linkage between sites. In sum, both amplicon-based and unbiased whole genome sequencing are flexible and can yield information about variant frequencies across a wide data set, but they are limited in the information that can be obtained about linkage across a specific area of a genome.

c. Amplicon-based sequencing of a small segment of the viral genome

This approach is typically used to sequence a small segment of a genome when the start and end positions are known and the entire string of nucleic acids across each piece of DNA are of interest. Primers specific for the region of interest are used to amplify the virus segment by RT-PCR. Tags are added that are both unique for the specific sample and help initiate sequencing on the instrument being used. These tags can be added using the TruSeq kits (Illumina) with or without PCR. The addition of tags by PCR leaves open the possibility that bias or errors may be incorporated into the DNA with the additional amplification steps. The addition of tags without PCR leaves open the possibility of greater sample loss. Either way, the tagged samples can be pooled and loaded on an instrument, such as an Illumina MiSeq.

This sequencing approach is advantageous for sequencing entire short segments of DNA that are the same length. On the Illumina MiSeq platform, sequences of up to 300 bp from each end of a single DNA molecule can be generated. After trimming and merging of paired sequences, it is entirely feasible to generate high-quality sequence information for an amplicon of about 400–500 bp long. Information about identical sequences or linked variants can be determined. Ultimately, this data covers a relatively small length of sequence, but a great depth of coverage can be obtained for a segment of DNA that is up to 500 bp.

Sequence Analysis

Gathering sequence data has become relatively easy, but analyzing it can be difficult and time consuming. There are many tools available to help investigators analyze sequence data. Many of these are complex and some require a basic knowledge of the command line. Many biologists who are interested in sequencing, however, are not familiar with the sort of bioinformatics needed to transform sequence information into something that is meaningful to an end user. Fortunately, there is software available for non-bioinformaticians to interact with their data. This chapter will emphasize the use of the software product called Geneious (Biomatters, Ltd). This software has a graphical user interface to visualize aligned reads, making sequence analyses more accessible to non-bioinformaticians. The next section will walk you through a general outline of how data that is obtained from an Illumina MiSeq can be explored in Geneious to obtain a set of information that can be handled in a viewable format.

Prior to sequencing samples, information linking each barcode tag to a sample is entered on the instrument. This is an essential step so that sequence information can be deconvoluted after the run. Once the sample information has been entered, then a DNA pool containing the denatured samples is loaded onto a flow cell, and the single-stranded DNA molecules hybridize to specific adapter oligos on the surface of the flow cell. The hybridized DNA molecules are amplified on the flow cell to form a cluster of single-stranded DNA molecules located at a single coordinate. DNA in each cluster is sequenced so that a single consensus forward and a single consensus reverse sequence read are obtained for the DNA located at each cluster [56].

Trimming, Merging, and Mapping FASTQ Reads

After the sequencing run is complete, the instrument deconvolutes the data by barcode. Two FASTQ files are generated for each barcode: one file containing all the forward reads and one file containing all the reverse reads. FASTQ reads are FASTA reads (strings of nucleotide sequences) with quality information attached to each nucleotide. Each FASTQ read is labeled with a header containing information about the machine, the run number, the coordinate position of the cluster, and the direction of the sequence (http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf).

FASTQ reads can be imported into Geneious and then quality trimmed. The trimming stringency may depend on your specific experiment. Once trimmed, the coordinate position associated with each read can be used so that the forward and reverse reads from the same cluster can be interrogated as a single unit. Paired reads can then be merged using a tool called FLASH (Fast Length Adjustment of SHort reads) [65]. There is an available plug-in that can be used for this purpose in Geneious (http://www.geneious.com/plugins/flash).

After trimming and merging, the reads can be mapped to a reference sequence, if available. The appropriate reference sequence should be imported into Geneious. Both the reference and the reads are selected and then the "Map to reference" function can be applied. The mapping settings are specific to the experiment being performed, but there is no harm using one of the default settings to explore the data. There is no perfect group of settings for mapping reads to a reference. By mapping the reads, all the sequences will be arranged in the same direction as the virus sequence. Once reads are mapped, then there is typically interest in either quantifying the frequency of individual SNPs or linked segments in the genome. These two methods are outlined below.

Quantify the Frequency of Individual SNPs Across the Virus Genome

Under the "Annotate & Predict" menu in Geneious, there is an option to "Find Variations/SNPs." After selecting that option, parameters are chosen to detect variants in the mapped sequences, relative to the reference. Some parameters should be considered carefully:

a. Depth of coverage: It is important to determine how much coverage at a given position is required to determine if a variant is authentic. Oftentimes, there is low coverage at the ends of the amplicons that are fragmented and then sequenced. Requiring a minimum amount of coverage is important so detection of false positives is minimized.
b. Minimum variant frequency: This is a highly subjective number. Control experiments have determined that 2 % is a reliable threshold to use to consider a variant authentic or not [26]. This threshold, however, will be somewhat dependent on the technical procedures used upstream of the sequencing reactions. Another consideration is the amount of computer memory required to characterize variants. Choosing a threshold that will detect an informative number of variants without being overwhelming is key.
c. *P*-value: This is a number that takes into account the authenticity of the SNP of interest in the context of the surrounding sequences. An alternative to setting a *p*-value threshold is to have Geneious calculate the *p*-value for the characterized variants. With this approach, the user can consider the *p*-value when determining the authenticity of each variant.

Quantify the Frequency of Linked Segments

Once the alignment is made, portions of sequence reads spanning a specific region of interest can be extracted from the alignment. The region in the reference sequence of interest can be highlighted, and then all the sequences spanning that region can

be selected, vertically. These sequence portions can be extracted into a new sequence list. When sequences are extracted from an alignment created from a Nextera library, the reads will be of variable length. When sequences are extracted from an amplicon of a single size, the reads will be of a similar length. Either way, after sequences are extracted, they should be filtered by size, and then reads that are the exact length of the sequence of interest should be extracted. For instance, to examine variation in a nine-amino-acid CD8 T cell epitope, reads that are 27 nucleotides in length should be examined. Once reads of the appropriate length have been extracted, then there are two major analysis routes to choose:

a. Assess the frequency of a nucleotide sequence in the data set: This is a simple counting exercise. In Geneious, there is an option under the Edit menu to "Find duplicates." With this tool, Geneious will identify sequences that are perfectly identical. It will report the number of times a given nucleotide sequence is detected in the total nucleotide sequence list.
b. Assess the frequency of an amino acid sequence in the data set: For this analysis, the reads need to be translated. Under the Sequence menu, there is an option to "Translate." A key point is to ensure the reads are translated in the correct reading frame. Once they are translated, then "Find Duplicates" can be used to determine the number of times a given amino acid sequence is present in the total amino acid sequence list.

In both cases, after the duplicates have been identified, the data can be exported as a FASTA file. The FASTA file can be converted to a tab-delimited file in a text editor. At this point, the frequencies of each nucleotide or amino acid sequence in the population can be calculated.

When measuring variant frequencies, it is important to determine the threshold frequency of the variants to consider as authentic. This is a common concern among researchers and is dependent on a given type of experiment. To determine a minimum threshold for variation, it is best to sequence a sample that should be clonal and has been prepared in the same conditions as the experimental samples. In one example, an HIV clonal stock from a plasmid was created and used to quantify the variant threshold when deep sequencing by the Illumina platform. In this study, a variant frequency of 2 % was found to be a conservative threshold for determining the authenticity of a variant [26]. The threshold to use, however, is dependent upon the preparation protocol, such that empirically measuring the error for an experimental technique is useful.

If possible, detecting a variant in multiple samples from the same individual further increases confidence that the variant is authentic. Longitudinal samples are ideal to assess whether a variant present at a low frequency at an early time point will then expand at a later time point. This type of observation increases the likelihood that the initial variant present at 2 % (or less) is authentic.

De Novo Assembly

There are times when information about the reference sequences is unknown, so it is not possible to map reads to a reference. This is common when using the unbiased sequencing approach described above. In these cases, a de novo assembly of the reads may be appropriate. When performing a de novo assembly, reads will be arranged so that overlapping segments will be linked together to form a single contig. Parameters need to be selected for running a de novo assembly. These parameters are likely going to be specific to a given experiment, but an investigator should not be afraid to use default parameters to explore the data set. Geneious will generate several de novo assembled contigs. For each contig, the closest identity of the consensus sequence can be determined using BLAST. There is an option to perform a "Sequence Search" in Geneious. With this tool, the consensus can be used in a BLAST search against a number of databases. The end user can then explore the hits to determine the identity of the sequence contig.

Discussion and Final Thoughts

Throughout this chapter, a variety of points to consider when deep sequencing HIV/SIV have been raised. Simply saying "I want to deep sequence SIV or HIV" is not enough when designing one of these experiments. Different deep sequencing approaches will yield different data sets that will inform subsequent mass spectrometry experiments in a different contextual manner. While collecting the data may be relatively easy, it is useful to consider how data will be analyzed and stored before beginning.

It is also important to consider whether deep sequencing is appropriate for the specific experimental needs. For instance, should single genome amplification (SGA) be used rather than deep sequencing? The differences between these techniques are substantial. For SGA, vRNA is diluted so that a single amplified product will have been generated from a single virus [66]. This amplified product is sequenced with Sanger technologies. While expensive and time consuming, the sequence data obtained can be attributed to a single template, which may be important for the experiment. In contrast, generating an amplicon and then deep sequencing the amplicon will generate a lot of data about SNP frequencies, but information about linkage between distant sites will be lost.

There is a lot of time and expense associated with deep sequencing. A single run on an Illumina instrument is expensive. Multiplexing samples significantly reduces the cost per sample, but, oftentimes, methods need to be tested on a few samples before expanding to a larger cohort. Testing these sequencing methodologies is made easier when there are colleagues with whom samples can be tested on other ongoing runs. Nonetheless, planning for method development is key, since most SIV/HIV deep sequencing experiments require a custom approach.

Analysis of data must also be considered. Each MiSeq run will generate multiple gigabytes of data. After performing an analysis of the data in Geneious, project sizes can be tens of gigabytes in size. The processing time of large data sets gets longer and becomes overwhelming. Managing, storing, and backing up this data with appropriate hardware needs to be considered in any grant budget.

Besides storing and analyzing data, it can be difficult to display the data. Variation at 10,000 nucleotides across a genome is difficult to display on a single screen. Often, key pieces of information are extracted and then put into a graph or a table to present to an audience. To explore the entire genome, there are some tools available. At UW-Madison, a program called LayerCake was developed to explore and compare variation in multiple genomes [63, 67]. Alternately the V-phaser and V-profiler programs that were developed by the Broad Institute can be used to generate heat maps of the data sets [58, 68].

Predicting the future of HIV/SIV deep sequencing is impossible. Major improvements will come as read lengths increase. While it would be ideal to deep sequence entire virus genomes from the beginning of the virus transcript to the end, this is unlikely to happen in the near future. In addition to read length limitations, the amplification of entire viral genomes without recombination is technically difficult. Single virus template sequencing will likely require new technologies that have not yet been developed.

Still, the technologies available to deep sequences SIV/HIV are useful and can be used to sequence other RNA viruses. The exponential improvements in sequencing technologies are continuing to diversify the hypotheses that are being tested with deep sequencing experiments, and thus our understanding of virus populations will continue to evolve. What is unlikely to change in the near future is the need for better sequence databases to maximize the return on investment for mass spectrometry experiments.

References

1. te Velthuis AJ. Common and unique features of viral RNA-dependent polymerases. Cell Mol Life Sci. 2014;71:4403–20.
2. Gire SK, Goba A, Andersen KG, Sealfon RS, Park DJ, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G, Wohl S, Moses LM, Yozwiak NL, Winnicki S, Matranga CB, Malboeuf CM, Qu J, Gladden AD, Schaffner SF, Yang X, Jiang PP, Nekoui M, Colubri A, Coomber MR, Fonnie M, Moigboi A, Gbakie M, Kamara FK, Tucker V, Konuwa E, Saffa S, Sellu J, Jalloh AA, Kovoma A, Koninga J, Mustapha I, Kargbo K, Foday M, Yillah M, Kanneh F, Robert W, Massally JL, Chapman SB, Bochicchio J, Murphy C, Nusbaum C, Young S, Birren BW, Grant DS, Scheiffelin JS, Lander ES, Happi C, Gevao SM, Gnirke A, Rambaut A, Garry RF, Khan SH, Sabeti PC. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 2014;345:1369–72.
3. Lauck M, Switzer WM, Sibley SD, Hyeroba D, Tumukunde A, Weny G, Taylor B, Shankar A, Ting N, Chapman CA, Friedrich TC, Goldberg TL, O'Connor DH. Discovery and full genome characterization of two highly divergent simian immunodeficiency viruses infecting black-and-white colobus monkeys (*Colobus guereza*) in Kibale National Park, Uganda. Retrovirology. 2013;10:107.

4. Lauck M, Switzer WM, Sibley SD, Hyeroba D, Tumukunde A, Weny G, Shankar A, Greene JM, Ericsen AJ, Zheng H, Ting N, Chapman CA, Friedrich TC, Goldberg TL, O'Connor DH. Discovery and full genome characterization of a new SIV lineage infecting red-tailed guenons (*Cercopithecus ascanius schmidti*) in Kibale National Park, Uganda. Retrovirology. 2014;11:55.
5. Wilker PR, Dinis JM, Starrett G, Imai M, Hatta M, Nelson CW, O'Connor DH, Hughes AL, Neumann G, Kawaoka Y, Friedrich TC. Selection on haemagglutinin imposes a bottleneck during mammalian transmission of reassortant H5N1 influenza viruses. Nat Commun. 2013;4:2636.
6. Newman RM, Kuntzen T, Weiner B, Berical A, Charlebois P, Kuiken C, Murphy DG, Simmonds P, Bennett P, Lennon NJ, Birren BW, Zody MC, Allen TM, Henn MR. Whole genome pyrosequencing of rare hepatitis C virus genotypes enhances subtype classification and identification of naturally occurring drug resistance variants. J Infect Dis. 2013;208:17–31.
7. Ram D, Leshkowitz D, Gonzalez D, Forer R, Levy I, Chowers M, Lorber M, Hindiyeh M, Mendelson E, Mor O. Evaluation of GS Junior and MiSeq next-generation sequencing technologies as an alternative to Trugene population sequencing in the clinical HIV laboratory. J Virol Methods. 2015;212:12–6.
8. Phillips RE, Rowland-Jones S, Nixon DF, Gotch FM, Edwards JP, Ogunlesi AO, Elvin JG, Rothbard JA, Bangham CR, Rizza CR, et al. Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. Nature. 1991;354:453–9.
9. Goulder PJ, Phillips RE, Colbert RA, McAdam S, Ogg G, Nowak MA, Giangrande P, Luzzi G, Morgan B, Edwards A, McMichael AJ, Rowland-Jones S. Late escape from an immunodominant cytotoxic T-lymphocyte response associated with progression to AIDS. Nat Med. 1997;3:212–7.
10. Borrow P, Lewicki H, Wei X, Horwitz MS, Peffer N, Meyers H, Nelson JA, Gairin JE, Hahn BH, Oldstone MB, Shaw GM. Antiviral pressure exerted by HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus [see comments]. Nat Med. 1997;3:205–11.
11. Price DA, Goulder PJ, Klenerman P, Sewell AK, Easterbrook PJ, Troop M, Bangham CR, Phillips RE. Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection. Proc Natl Acad Sci U S A. 1997;94:1890–5.
12. Evans DT, O'Connor DH, Jing P, Dzuris JL, Sidney J, da Silva J, Allen TM, Horton H, Venham JE, Rudersdorf RA, Vogel T, Pauza CD, Bontrop RE, DeMars R, Sette A, Hughes AL, Watkins DI. Virus-specific cytotoxic T-lymphocyte responses select for amino-acid variation in simian immunodeficiency virus Env and Nef. Nat Med. 1999;5:1270–6.
13. Allen TM, O'Connor DH, Jing P, Dzuris JL, Mothe BR, Vogel TU, Dunphy E, Liebl ME, Emerson C, Wilson N, Kunstman KJ, Wang X, Allison DB, Hughes AL, Desrosiers RC, Altman JD, Wolinsky SM, Sette A, Watkins DI. Tat-specific cytotoxic T lymphocytes select for SIV escape variants during resolution of primary viraemia. Nature. 2000;407:386–90.
14. Richman DD, Wrin T, Little SJ, Petropoulos CJ. Rapid evolution of the neutralizing antibody response to HIV type 1 infection. Proc Natl Acad Sci U S A. 2003;100:4144–9.
15. Wei X, Decker JM, Wang S, Hui H, Kappes JC, Wu X, Salazar-Gonzalez JF, Salazar MG, Kilby JM, Saag MS, Komarova NL, Nowak MA, Hahn BH, Kwong PD, Shaw GM. Antibody neutralization and escape by HIV-1. Nature. 2003;422:307–12.
16. Roederer M, Keele BF, Schmidt SD, Mason RD, Welles HC, Fischer W, Labranche C, Foulds KE, Louder MK, Yang ZY, Todd JP, Buzby AP, Mach LV, Shen L, Seaton KE, Ward BM, Bailer RT, Gottardo R, Gu W, Ferrari G, Alam SM, Denny TN, Montefiori DC, Tomaras GD, Korber BT, Nason MC, Seder RA, Koup RA, Letvin NL, Rao SS, Nabel GJ, Mascola JR. Immunological and virological mechanisms of vaccine-mediated protection against SIV and HIV. Nature. 2014;505:502–8.
17. Wu F, Ourmanov I, Kuwata T, Goeken R, Brown CR, Buckler-White A, Iyengar R, Plishka R, Aoki ST, Hirsch VM. Sequential evolution and escape from neutralization of simian immunodeficiency virus SIVsmE660 clones in rhesus macaques. J Virol. 2012;86:8835–47.

18. Vigerust DJ, Shepherd VL. Virus glycosylation: role in virulence and immune interactions. Trends Microbiol. 2007;15:211–8.
19. Edlefsen PT, Rolland M, Hertz T, Tovanabutra S, Gartland AJ, deCamp AC, Magaret CA, Ahmed H, Gottardo R, Juraska M, McCoy C, Larsen BB, Sanders-Buell E, Carrico C, Menis S, Bose M, Arroyo MA, O'Connell RJ, Nitayaphan S, Pitisuttithum P, Kaewkungwal J, Rerks-Ngarm S, Robb ML, Kirys T, Georgiev IS, Kwong PD, Scheffler K, Pond SL, Carlson JM, Michael NL, Schief WR, Mullins JI, Kim JH, Gilbert PB. Comprehensive sieve analysis of breakthrough HIV-1 sequences in the RV144 vaccine efficacy trial. PLoS Comput Biol. 2015;11:e1003973.
20. Rolland M, Tovanabutra S, deCamp AC, Frahm N, Gilbert PB, Sanders-Buell E, Heath L, Magaret CA, Bose M, Bradfield A, O'Sullivan A, Crossler J, Jones T, Nau M, Wong K, Zhao H, Raugi DN, Sorensen S, Stoddard JN, Maust BS, Deng W, Hural J, Dubey S, Michael NL, Shiver J, Corey L, Li F, Self SG, Kim J, Buchbinder S, Casimiro DR, Robertson MN, Duerr A, McElrath MJ, McCutchan FE, Mullins JI. Genetic impact of vaccination on breakthrough HIV-1 sequences from the STEP trial. Nat Med. 2011;17:366–71.
21. Rolland M, Edlefsen PT, Larsen BB, Tovanabutra S, Sanders-Buell E, Hertz T, deCamp AC, Carrico C, Menis S, Magaret CA, Ahmed H, Juraska M, Chen L, Konopa P, Nariya S, Stoddard JN, Wong K, Zhao H, Deng W, Maust BS, Bose M, Howell S, Bates A, Lazzaro M, O'Sullivan A, Lei E, Bradfield A, Ibitamuno G, Assawadarachai V, O'Connell RJ, deSouza MS, Nitayaphan S, Rerks-Ngarm S, Robb ML, McLellan JS, Georgiev I, Kwong PD, Carlson JM, Michael NL, Schief WR, Gilbert PB, Mullins JI, Kim JH. Increased HIV-1 vaccine efficacy against viruses with genetic signatures in Env V2. Nature. 2012;490:417–20.
22. Hosseinipour MC, Gupta RK, Van Zyl G, Eron JJ, Nachega JB. Emergence of HIV drug resistance during first- and second-line antiretroviral therapy in resource-limited settings. J Infect Dis. 2013;207 Suppl 2:S49–56.
23. Gross R, Yip B, Lo Re V, Wood E, Alexander CS, Harrigan PR, Bangsberg DR, Montaner JS, Hogg RS. A simple, dynamic measure of antiretroviral therapy adherence predicts failure to maintain HIV-1 suppression. J Infect Dis. 2006;194:1108–14.
24. Bennett DE, Bertagnolio S, Sutherland D, Gilks CF. The World Health Organization's global strategy for prevention and assessment of HIV drug resistance. Antivir Ther. 2008;13 Suppl 2:1–13.
25. Jordan MR, Bennett DE, Wainberg MA, Havlir D, Hammer S, Yang C, Morris L, Peeters M, Wensing AM, Parkin N, Nachega JB, Phillips A, De Luca A, Geng E, Calmy A, Raizes E, Sandstrom P, Archibald CP, Perriens J, McClure CM, Hong SY, McMahon JH, Dedes N, Sutherland D, Bertagnolio S. Update on World Health Organization HIV drug resistance prevention and assessment strategy: 2004–2011. Clin Infect Dis. 2012;54 Suppl 4:S245–9.
26. Dudley DM, Bailey AL, Mehta SH, Hughes AL, Kirk GD, Westergaard RP, O'Connor DH. Cross-clade simultaneous HIV drug resistance genotyping for reverse transcriptase, protease, and integrase inhibitor mutations by Illumina MiSeq. Retrovirology. 2014;11:122.
27. Estill J, Salazar-Vizcaya L, Blaser N, Egger M, Keiser O. The cost-effectiveness of monitoring strategies for antiretroviral therapy of HIV infected patients in resource-limited settings: software tool. PLoS One. 2015;10:e0119299.
28. Taylor BS, Sobieszczyk ME, McCutchan FE, Hammer SM. The challenge of HIV-1 subtype diversity. N Engl J Med. 2008;358:1590–602.
29. Hemelaar J. The origin and diversity of the HIV-1 pandemic. Trends Mol Med. 2012;18:182–92.
30. Shaw, GM, Hunter, E. 2012. HIV transmission. Cold Spring Harb Perspect Med 2
31. Easterbrook PJ, Smith M, Mullen J, O'Shea S, Chrystie I, de Ruiter A, Tatt ID, Geretti AM, Zuckerman M. Impact of HIV-1 viral subtype on disease progression and response to antiretroviral therapy. J Int AIDS Soc. 2010;13:4.
32. Pant Pai N, Shivkumar S, Cajas JM. Does genetic diversity of HIV-1 non-B subtypes differentially impact disease progression in treatment-naive HIV-1-infected individuals? A systematic review of evidence: 1996–2010. J Acquir Immune Defic Syndr. 2012;59:382–8.

33. Tarosso LF, Sanabani SS, Ribeiro SP, Sauer MM, Tomiyama HI, Sucupira MC, Diaz RS, Sabino EC, Kalil J, Kallas EG. Short communication: HIV type 1 subtype BF leads to faster CD4+ T cell loss compared to subtype B. AIDS Res Hum Retroviruses. 2014;30:190–4.
34. Tebit DM, Arts EJ. Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease. Lancet Infect Dis. 2011;11:45–56.
35. Hahn BH, Shaw GM, De Cock KM, Sharp PM. AIDS as a zoonosis: scientific and public health implications. Science. 2000;287:607–14.
36. Ratner L, Haseltine W, Patarca R, Livak KJ, Starcich B, Josephs SF, Doran ER, Rafalski JA, Whitehorn EA, Baumeister K, et al. Complete nucleotide sequence of the AIDS virus, HTLV-III. Nature. 1985;313:277–84.
37. Regier DA, Desrosiers RC. The complete nucleotide sequence of a pathogenic molecular clone of simian immunodeficiency virus. AIDS Res Hum Retroviruses. 1990;6:1221–31.
38. Springer M. Applied biosystems: celebrating 25 years of advancing science. Am Lab. 2006;38(11):4–8.
39. O'Connor DH, Allen TM, Vogel TU, Jing P, DeSouza IP, Dodds E, Dunphy EJ, Melsaether C, Mothe B, Yamamoto H, Horton H, Wilson N, Hughes AL, Watkins DI. Acute phase cytotoxic T lymphocyte escape is a hallmark of simian immunodeficiency virus infection. Nat Med. 2002;8:493–9.
40. Keele BF. Identifying and characterizing recently transmitted viruses. Curr Opin HIV AIDS. 2010;5:327–34.
41. Keele BF, Li H, Learn GH, Hraber P, Giorgi EE, Grayson T, Sun C, Chen Y, Yeh WW, Letvin NL, Mascola JR, Nabel GJ, Haynes BF, Bhattacharya T, Perelson AS, Korber BT, Hahn BH, Shaw GM. Low-dose rectal inoculation of rhesus macaques by SIVsmE660 or SIVmac251 recapitulates human mucosal infection by HIV-1. J Exp Med. 2009;206:1117–34.
42. Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, Salazar MG, Sun C, Grayson T, Wang S, Li H, Wei X, Jiang C, Kirchherr JL, Gao F, Anderson JA, Ping LH, Swanstrom R, Tomaras GD, Blattner WA, Goepfert PA, Kilby JM, Saag MS, Delwart EL, Busch MP, Cohen MS, Montefiori DC, Haynes BF, Gaschen B, Athreya GS, Lee HY, Wood N, Seoighe C, Perelson AS, Bhattacharya T, Korber BT, Hahn BH, Shaw GM. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proc Natl Acad Sci U S A. 2008;105:7552–7.
43. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437:376–80.
44. Bimber BN, Chugh P, Giorgi EE, Kim B, Almudevar AL, Dewhurst S, O'Connor DH, Lee HY. Nef gene evolution from a single transmitted strain in acute SIV infection. Retrovirology. 2009;6:57.
45. Bimber BN, Dudley DM, Lauck M, Becker EA, Chin EN, Lank SM, Grunenwald HL, Caruccio NC, Maffitt M, Wilson NA, Reed JS, Sosman JM, Tarosso LF, Sanabani S, Kallas EG, Hughes AL, O'Connor DH. Whole-genome characterization of human and simian immunodeficiency virus intrahost diversity by ultradeep pyrosequencing. J Virol. 2010;84:12087–92.
46. Cale EM, Hraber P, Giorgi EE, Fischer W, Bhattacharya T, Leitner T, Yeh WW, Gleasner C, Green LD, Han CS, Korber B, Letvin NL. Epitope-specific CD8+ T lymphocytes cross-recognize mutant simian immunodeficiency virus (SIV) sequences but fail to contain very early evolution and eventual fixation of epitope escape mutations during SIV infection. J Virol. 2011;85:3746–57.

47. Rozera G, Abbate I, Bruselles A, Vlassi C, D'Offizi G, Narciso P, Chillemi G, Prosperi M, Ippolito G, Capobianchi MR. Massively parallel pyrosequencing highlights minority variants in the HIV-1 env quasispecies deriving from lymphomonocyte sub-populations. Retrovirology. 2009;6:15.
48. Hughes AL, Becker EA, Lauck M, Karl JA, Braasch AT, O'Connor DH, O'Connor SL. SIV genome-wide pyrosequencing provides a comprehensive and unbiased view of variation within and outside CD8 T lymphocyte epitopes. PLoS One. 2012;7:e47818.
49. O'Connor SL, Becker EA, Weinfurter JT, Chin EN, Budde ML, Gostick E, Correll M, Gleicher M, Hughes AL, Price DA, Friedrich TC, O'Connor DH. Conditional CD8+ T cell escape during acute simian immunodeficiency virus infection. J Virol. 2012;86:605–9.
50. Brumme CJ, Huber KD, Dong W, Poon AF, Harrigan PR, Sluis-Cremer N. Replication fitness of multiple nonnucleoside reverse transcriptase-resistant HIV-1 variants in the presence of etravirine measured by 454 deep sequencing. J Virol. 2013;87:8805–7.
51. Avidor B, Girshengorn S, Matus N, Talio H, Achsanov S, Zeldis I, Fratty IS, Katchman E, Brosh-Nissimov T, Hassin D, Alon D, Bentwich Z, Yust I, Amit S, Forer R, Vulih Shultsman I, Turner D. Evaluation of a benchtop HIV ultradeep pyrosequencing drug resistance assay in the clinical laboratory. J Clin Microbiol. 2013;51:880–6.
52. Dudley DM, Chin EN, Bimber BN, Sanabani SS, Tarosso LF, Costa PR, Sauer MM, Kallas EG, O'Connor DH. Low-cost ultra-wide genotyping using Roche/454 pyrosequencing for surveillance of HIV drug resistance. PLoS One. 2012;7:e36494.
53. Shao W, Boltz VF, Spindler JE, Kearney MF, Maldarelli F, Mellors JW, Stewart C, Volfovsky N, Levitsky A, Stephens RM, Coffin JM. Analysis of 454 sequencing error rate, error sources, and artifact recombination for detection of low-frequency drug resistance mutations in HIV-1 DNA. Retrovirology. 2013;10:18.
54. Iyer S, Bouzek H, Deng W, Larsen B, Casey E, Mullins JI. Quality score based identification and correction of pyrosequencing errors. PLoS One. 2013;8:e73015.
55. Macalalad AR, Zody MC, Charlebois P, Lennon NJ, Newman RM, Malboeuf CM, Ryan EM, Boutwell CL, Power KA, Brackney DE, Pesko KN, Levin JZ, Ebel GD, Allen TM, Birren BW, Henn MR. Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data. PLoS Comput Biol. 2012;8:e1002417.
56. Balasubramanian S. Sequencing nucleic acids: from chemistry to medicine. Chem Commun (Camb). 2011;47:7281–6.
57. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012;2012:251364.
58. Adnan S, Colantonio AD, Yu Y, Gillis J, Wong FE, Becker EA, Piatak MJ, Reeves RK, Lifson JD, O'Connor SL, Johnson RP. CD8 T Cell Response Maturation Defined by Anentropic Specificity and Repertoire Depth Correlates with SIVDeltanef-induced Protection. PLoS Pathog. 2015;11:e1004633.
59. Lauck M, Sibley SD, Hyeroba D, Tumukunde A, Weny G, Chapman CA, Ting N, Switzer WM, Kuhn JH, Friedrich TC, O'Connor DH, Goldberg TL. Exceptional simian hemorrhagic fever virus diversity in a wild African primate community. J Virol. 2013;87:688–91.
60. Sibley SD, Lauck M, Bailey AL, Hyeroba D, Tumukunde A, Weny G, Chapman CA, O'Connor DH, Goldberg TL, Friedrich TC. Discovery and characterization of distinct simian pegiviruses in three wild African Old World monkey species. PLoS One. 2014;9:e98569.
61. Goldberg TL, Gendron-Fitzpatrick A, Deering KM, Wallace RS, Clyde VL, Lauck M, Rosen GE, Bennett AJ, Greiner EC, O'Connor DH. Fatal metacestode infection in Bornean orangutan caused by unknown Versteria species. Emerg Infect Dis. 2014;20:109–13.
62. Lauck M, Palacios G, Wiley MR, Li Y, Fang Y, Lackemeyer MG, Cai Y, Bailey AL, Postnikova E, Radoshitzky SR, Johnson RF, Alkhovsky SV, Deriabin PG, Friedrich TC, Goldberg TL, Jahrling PB, O'Connor DH, Kuhn JH. Genome sequences of simian hemorrhagic fever virus variant NIH LVR42-0/M6941 isolates (Arteriviridae: Arterivirus). Genome Announc. 2014;2(5):e00978–14.

63. Bailey AL, Lauck M, Weiler A, Sibley SD, Dinis JM, Bergman Z, Nelson CW, Correll M, Gleicher M, Hyeroba D, Tumukunde A, Weny G, Chapman C, Kuhn JH, Hughes AL, Friedrich TC, Goldberg TL, O'Connor DH. High genetic diversity and adaptive potential of two simian hemorrhagic fever viruses in a wild primate population. PLoS One. 2014;9:e90714.
64. Harris M, Burns CM, Becker EA, Braasch AT, Gostick E, Johnson RC, Broman KW, Price DA, Friedrich TC, O'Connor SL. Acute-phase CD8 T cell responses that select for escape variants are needed to control live attenuated simian immunodeficiency virus. J Virol. 2013;87:9353–64.
65. Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011;27:2957–63.
66. Salazar-Gonzalez JF, Bailes E, Pham KT, Salazar MG, Guffey MB, Keele BF, Derdeyn CA, Farmer P, Hunter E, Allen S, Manigart O, Mulenga J, Anderson JA, Swanstrom R, Haynes BF, Athreya GS, Korber BT, Sharp PM, Shaw GM, Hahn BH. Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. J Virol. 2008;82:3952–70.
67. Correll M, Ghosh S, O'Connor D, Gleicher M. Visualizing virus population variability from next generation sequencing data. In: Proceedings of BioVis; 2011
68. Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, Macalalad AR, Berlin AM, Malboeuf CM, Ryan EM, Gnerre S, Zody MC, Erlich RL, Green LM, Berical A, Wang Y, Casali M, Streeck H, Bloom AK, Dudek T, Tully D, Newman R, Axten KL, Gladden AD, Battis L, Kemper M, Zeng Q, Shea TP, Gujja S, Zedlack C, Gasser O, Brander C, Hess C, Gunthard HF, Brumme ZL, Brumme CJ, Bazner S, Rychert J, Tinsley JP, Mayer KH, Rosenberg E, Pereyra F, Levin JZ, Young SK, Jessen H, Altfeld M, Birren BW, Walker BD, Allen TM. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS Pathog. 2012;8:e1002529.