Published in final edited form as:

J Virol Methods. 2014 September 1; 205: 7–16. doi:10.1016/j.jviromet.2014.04.017.

Targeted deep sequencing of HIV-1 using the IonTorrentPGM platform

Gustavo H Kijak^{1,*}, Eric Sanders-Buell¹, Elizabeth A Harbolick¹, Phuc Pham¹, Agnes L Chenine¹, Leigh Anne Eller¹, Kathleen Rono², Merlin L Robb¹, Nelson L Michael³, Jerome H Kim³, and Sodsai Tovanabutra¹

¹U.S. Military HIV Research Program, Henry M. Jackson Foundation for the Advancement of Military Medicine, Bethesda, Maryland, USA

²Kenya Medical Research Institute/Walter Reed Project, Kericho, Kenya

³U.S. Military HIV Research Program, Walter Reed Army Institute of Research, Silver Spring, Maryland, USA

Summary

The characterization of mixed HIV-1 populations is a key question in clinical and basic research settings. This can be achieved through targeted deep sequencing (TDS), where next-generation sequencing is used to examine in depth a sub-genomic region of interest. This study explores the suitability of IonTorrent PGM(LifeTechnologies) for the TDS-based analysis of HIV-1 evolution. Using laboratory reagents and primary specimens sampled at pre-peak viremia the error rates from misincorporation and in vitro recombination were <0.5%. The sequencing error rate was 2- to 3-fold higher in/around homopolymeric tracts, and could be discerned from true polymorphism using bidirectional sequencing. The limit of detection of complex variants was further lowered by using haplotyping. The application of this system was illustrated on primary samples from an individual infected with HIV-1 followed from pre-peak viremia through six months post-acquisition. TDS provided an augmented view of the extent of genetic diversity, the covariation among polymorphisms, the evolutionary pathways, and the boundaries of the mutational space explored by the viral swarm. Based on its performance, the system can be applied for the characterization of minor viral variants in support of studies of viral evolution, which can inform the rational design of the next generation of vaccines and therapeutics.

Keywords

targeted deep sequencing; next-generation sequencing; HIV-1; molecular evolution; IonTorrent

Corresponding author: Gustavo Hernan Kijak, PharmD PhD, US Military HIV Research Program/Henry M Jackson Foundation, Walter Reed Army Institute of Research, 503 Robert Grant Avenue, Rm 2N27, Silver Spring, MD 20910, USA, Phone: 1-301-319-9746, gkijak@hivresearch.org.

Conflict of interest: The authors declare no conflict of interest

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. Introduction

As of the end of 2012, more than 35 million people were living worldwide with HIV/AIDS, and hundreds of millions were at risk of acquiring the infection (UNAIDS, 2013). Despite colossal research efforts in the past three decades, the quest for a safe and effective preventive vaccine and a cure for the disease are still ongoing and constitute main public health priorities.

The extensive genetic variation of the virus, which results from the combination of high rates of mutation, recombination and replication, is responsible for the fast emergence of viral strains that can evade host immune responses and resist the effects of antiretroviral agents (Coffin and Swanstrom, 2013). During most of the course of the infection, viral populations within an individual are heterogeneous, with varying levels of genetic diversity (Salazar-Gonzalez et al., 2009). Being able to characterize mixed viral populations is a key question both in the clinical and in the basic research settings. For instance, the study of the ontogeny of viral escape from the host adaptive immune responses, such as antibodies and cytotoxic T lymphocytes (CTLs), can support the rational design of preventive vaccine (Goonetilleke et al., 2009). Similarly, the detection of viral quasiespecies that are resistant to antiretroviral drugs can inform the selection of regimens with better prospects of success; the risk of treatment failure may be significantly increased even when the drug-resistant viruses are present at 1% (Li et al., 2011).

Traditionally, viral populations have been sampled using cloning in bacterial systems or single genome amplification (SGA) (Palmer et al., 2005), followed by Sanger sequencing; these methods tend to be costly, tedious, low-throughput, and are not prone to extensive automation (McCutchan et al., 2005; Shankarappa et al., 1999). Furthermore, power calculations show that the detection of viral subpopulations carrying multiple mutations and present at frequencies below 5% requires acquiring >100 sequences (Keele et al., 2008), which renders this approach impractical and cost-prohibitive for routine detection of such minor viral variants. Real-time PCR-based methods represent cost-effective methods for the sensitive and selective detection of single nucleotide polymorphisms (SNPs) at frequencies <1% (Wei et al., 2011) but they are not designed to capture linkage among multiple polymorphisms. The parallel allele-specific sequencing (PASS) assay can detect low-frequency variants and can determine linkage among polymorphisms (Cai et al., 2007) but its use is not widespread because no commercial kit is available.

The advent of next-generation sequencing (NGS) technologies brings the promise of simplicity, increased throughput, low cost, and automation, which can translate into affordable higher sequencing depth. At the same time, the error rate associated with NGS is recognized to be higher than that of Sanger-based sequencing (Chang, 2009). When applied to HIV-1, the measured noise of NGS results from the composite of: 1) reverse transcription (when using viral RNA (vRNA) as initial template), 2) PCR for library preparation, 3) emulsion PCR (ePCR) or cluster generation, and 4) sequencing. The introduction of noise in the first two processes can be limited using high-fidelity enzymes (Di Giallonardo et al., 2013; Larsen et al., 2013), whereas the end-user has little capacity to influence the conditions of processes 3 and 4. In the study of HIV-1, the capacity to accurately discern

true genetic signals from background noise is paramount; especially considering that SNPs can drastically impact susceptibility to antiretroviral drugs (Richman et al., 1994) or immune effectors (Leslie et al., 2004).

The main experimental approaches of HIV-1 NGS have been whole genome sequencing, whole gene sequencing, and targeted deep sequencing (TDS). The first two approaches involve the amplification, fragmentation, sequencing, and alignment of partially overlapping reads. TDS examines a defined sub-genomic region of interest at great sampling (i.e., number of input genomes) and sequencing (i.e., number of reads) depth to determine the frequency of the different variants. One of the main advantages of TDS compared to the other approaches derives from the fact that the capacity to obtain longer reads has increased substantially over the recent years. As in TDS the obtained reads are, for the most part, fully overlapping, it is now possible to directly determine, rather than just infer (Zagordi et al., 2011), the linkage among measured polymorphisms (i.e., haplotyping). This is particularly important in HIV-1, where the presence of multiple mutations in the same genome can result in viruses with increased levels of antiretroviral drug resistance (Larder and Kemp, 1989) or can compensate for the loss of fitness that may have resulted from particular CTL-escape mutations (Brockman et al., 2007). Additionally, as PCR efficiency is inversely associated with amplicon length (Piyamongkol et al., 2003), targeting shorter regions results in more efficient library preparation, larger pools of sequenceable genomes, and lower chances of allelic dropout. Targeting shorter regions also results in lower in vitro recombination rates, and this is particularly important in HIV-1, where linkage can be created or disrupted by recombination events occurring during HIV-1 reverse transcription (Clavel et al., 1989), but it can also be an in vitro artifact generated in NGS library preparation (Shao et al., 2013).

The vast majority of the HIV-1 TDS work published to date has been produced using the 454 pyrosequencing platform (Armenia et al., 2012; Di Giallonardo et al., 2013; Fischer et al., 2010; Hedskog et al., 2013; Hedskog et al., 2010; Messiaen et al., 2012; Mild et al., 2011; Mukherjee et al., 2011; Poon et al., 2010; Redd et al., 2011; Redd et al., 2012; Rozera et al., 2009; Shao et al., 2013; Simen et al., 2009; Swenson et al., 2011; Swenson et al., 2010; Tsibris et al., 2009; Vandenbroucke et al., 2010) which was the first commercial NGS system, and there is also one report using the Pacific Bioscience platform (Archer et al., 2012). The use of the Illumina and Ion Torrent PGM systems for HIV-1 genotyping has been reported for whole-gene sequencing (Archer et al., 2012; Chang et al., 2013). Most recently, Gibson et al. (2014) have reported on DEEPGENTMHIV, a genotyping platform that applies Ion Torrent PGM whole-gene sequencing and TDS, but is tailored towards the clinical management of patients. While other platforms offer higher yields, at this time, the IonTorrent PGM system offers one of the lowest costs per experiment (<http://www.molecularecologist.com/next-gen-table-2b/>), making this a preferred choice for small laboratories. The aim of the current study was to determine the suitability of TDS implemented in the IonTorrent PGM platform for the analysis of HIV-1 evolution. Using a combination of laboratory reagents and primary clinical specimens the system performed with a lower limit of detection of 0.5% for single nucleotide variants, and with an even lower limit for more complex haplotypes. Based on these specifications, the system can be applied for the characterization of minor viral variants in support of studies of viral evolution, as

illustrated here for the analysis of an HIV-1 individual infected followed longitudinally from pre-peak viremia through six months post-infection.

2. Materials and Methods

2.1. Samples under study

Plasmid pBCKS(+) plasmid (Agilent Technologies, Santa Clara, CA), encoding the chloramphenicol acetyltransferase (CAT) gene, was expanded in Stbl2 bacterial system (Invitrogen/LifeTechnologies, Carlsbad, CA) in the presence of chloramphenicol, and was extracted using the PureYield Plasmid Maxiprep System (Promega, Madison, WI), according to manufacturer's instructions. The concentration of the extracted plasmid DNA was determined by OD260 spectrophotometry. Three samples from a female commercial sex worker from Kenya enrolled in the RV217/ECHO cohort (henceforth referred to as "participant 20225") were used to assess the assay's performance on primary specimens. The individual was enrolled before HIV-1 infection. HIV-1 acquisition was documented by HIV-1 nucleic acid testing (NAT)(APTIMA HIV-1 RNA Qualitative Assay, GenProbe, San Diego, CA), plasma viral load (pVL) (Abbott RealTime HIV-1, Abbott Laboratories, Abbott Park, IL, lower limit of detection: 40 copies/ml) and serology(Robb et al., manuscript in preparation). Specimens were sampled at day 5 (pVL=17,782,794 copies/ml), day 36(pVL=131,826 copies/ml), and day 178 (pVL=17783) (considering day -4 as the date of the last NAT negative determination, day 0 as the date of first positive HIV-1 NAT, and day 9 as peak viremia) (Supplemental figure 1). vRNA was extracted using the QIAamp Viral RNA Mini Kit (QIAGEN, Valencia, CA) (see Supplemental Table 1 for utilized plasma volumes).

2.2. Library preparation

2.2.1. Reverse transcription—For HIV-1 subgenomic regions gp120-V2 and gp120-C3, vRNA was retrotranscribed using the SuperScript III First Strand Synthesis System for RT-PCR (Invitrogen/LifeTechnologies, Carlsbad, CA). All primer sequences are shown in Supplemental Table 2. Master mix 1 contained 2 mM dNTPs, 5 uM of outer reverse primer, extracted vRNA, and diethylpyrocarbonate (DEPC)-treated water in a final volume of 10 ul. The mix was incubated at 65°C for 5 min then immediately transferred to ice for 4 min. Master mix 2 (10x RT buffer, 0.02 M DTT, 10 mM MgCl₂, RnaseOUT, and SuperScript III RT in a final volume of 10 ul) was then added on top of master mix 1. The thermocycler routine was: 50°C for 3 hrs, 85°C for 5 min. Then 1 ul of E. coli Rnase H was added and incubated at 37°C for 20 min.

2.2.2. Nested PCR—For all assessed regions, nested PCR used the Titanium Taq kit (Clontech, Mountain View, CA). The first round PCR mix contained 0.2 mM dNTPs, 0.2 uM forward and reverse primers, 10x Titanium Buffer, 50x Titanium Taq, and cDNA, in a final volume of 25 ul. In order to avoid PCR saturation and to minimize the chances of in vitro recombination, 2,000 copies of cDNA (as determined by titration, see below) were distributed into separate first round PCRs (<250 copies/reaction) (Supplemental Table 1). The thermocycler routine was: hold at 95°C for 1 min, then 30 cycles of 95°C 30s, 55°C 30s, 68 1 min, and a final extension at 68°C for 3 min. Each first round PCR product was

then distributed into two sets of second round PCR, in order to allow for bidirectional sequencing. Based on the fusion method, forward primers containing the A adaptor were paired with reverse primers containing the P1 adaptor, while reverse primers containing the A adaptor were paired with forward primers containing the P1 adaptor (Ion Amplicon Library Preparation (Fusion Method); Publication Number 4468326; Revision C; LifeTechnologies; 2012). The master mix contained 0.2 mM dNTPs, 0.2 uM forward and reverse primers, 10x Titanium Buffer, 50x Titanium Taq, and 1 ul of first round PCR product in a final volume of 25 ul. The thermocycler routine was: hold at 95°C for 1 min, then 27 cycles of 95°C 30s, 55°C 30s, 68°C 1 min, and a final extension at 68°C for 3 min. PCR products were visualized by electrophoresis on a 1.5% agarose gel, and were purified using the NucleoSpin Extract II kit (Machery-Nagel, Düren, Germany) following the manufacturer's instructions. PCR products were quantified using a 2100 Bioanalyzer (DNA 1000 kit, Agilent Technologies, Sunnyvale CA).

2.3 Sequencing

ePCR and enrichment were carried out using the Ion OneTouch 100 bp Template kit (LifeTechnologies, Carlsbad, CA) on OneTouch andES instruments. Sequencing was carried out on PGM instrument using Ion 316 Chips, the Ion PGM sequencing 100 bp kit (LifeTechnologies, Carlsbad, CA), following the manufacturer's instructions.

2.4 cDNA Titration

The number of amplifiable copies present in the cDNA was determined by endpoint dilution. Serial dilutions of the cDNA were used as template for nested PCR in the same conditions that are used for library preparation. Reactions were run in triplicate and were visualized by electrophoresis on a 1.5% agarose gel.

2.5 Determination of in vitro recombination rates

Day 5 and day 178 plasma samples from participant 20225 were the source of near-full genome HIV-1 SGAmplicons, which were amplified and sequenced as previously described (Salazar-Gonzalez et al., 2009). The dsDNA concentration was quantified using the Qubit dsDNA HS Assay Kit (LifeTechnologies, Carlsbad, CA). The amplicons were mixed in equimolar concentrations, the mixture was diluted, and 2,000 total copies of DNA template were subject to library preparation and TDS as described above. Due to intrinsic errors in measuring the DNA concentration and multiple serial dilutions, the final mixture was actually 60:40.

The same reagents, principles, and protocols were applied for the study of the effect of homopolymeric tracts on sequencing error rates. For those experiments, the resulting amplicon mixtures were 60:40, 25:75, 7:93, and 1:99.

2.6 NGS data analysis

Fastq files were exported from the PGM using the Torrent Suite 3.6.2 software, and were imported into CLC Genomics Workbench version 6.0 (CLC bio, Cambridge, MA) where sequencing adapters were removed and sequences were trimmed based on quality (limit=0.05; maximal 2 ambiguous nucleotides) and length (minimum length=45

nucleotides). The resulting fastq files were then aligned to reference using tmap (by Nils Homer, distributed through <https://github.com/iontorrent/TMAP>) using the following parameters: match score= 1; mismatch penalty= 3; gap open penalty= 5; gap extension penalty= 2; and soft-clip only the right portion of the read. Employed references were a representative day 5 SGA-derived HIV-1 sequence from participant 20225 (regions gp120-V2 and gp120-C3), and the pBCKS(+) vector sequence (<http://www.genomics.agilent.com/>). SAM files were analyzed using Nautilus (Kijak et al., 2013), which tallied the frequency of each base at each sequence position and determined the frequency of haplotypes. Alignments were also visually inspected using Tablet (Milne et al., 2013). The study of sequencing error rate was focused on substitutions, and excluded insertions/deletions that were not multiple of three (i.e., leading to frameshift), following the model of previous publications (Gibson et al., 2014; Tsibris et al., 2009). All references to HIV-1 codon are based on the HXB2 coordinate system (Korber et al., 1998).

3. Results

3.1 Targeted Deep sequencing approach

The experimental design of TDS can be schematically divided into five steps: i) selection of the target sequence; ii) template quantitation; iii) library preparation; iv) sequencing; and v) data analysis. The used protocol is detailed in the Materials and Methods section, and below are distinguished aspects of each step.

3.1.1 Selection of the target sequence—In order to determine the level of background sequencing noise a section of the CAT gene encoded in the pBCKS(+) plasmid was targeted. This gene is used for in vitro antibiotic selection, and thus very limited genetic variation was expected. To assess the suitability of the TDS strategy for the study of viral evolution two HIV-1 subgenomic areas were selected: the env gene -gp120-V2 (HXB2: 6700-6838) and gp120-C3 (HXB2:7229-7357). These areas were selected because they were some of the regions where the traditional SGA-based longitudinal analysis of samples from participant 20225 -an individual infected with HIV-1 followed from pre-peak viremia through six months post-acquisition-had shown concentrated genetic diversity. The SGA-based analysis had also indicated that in this individual the infection had been established by a single transmitted/founder (T/F) virus (Kijak et al., 2012), which made the pre-peak specimen -with its expected highly homogenous viral population (Salazar-Gonzalez et al., 2009)- an ideal setting to assess the level of background sequencing noise in primary samples. In order to prevent primer-bias, primers were tailor-designed for the individual under study, and were located in genetically invariant sequences.

3.1.2 Template quantitation—Following the principles stated by Larsen et al. (2013) endpoint dilution employing the same amplification system was used for library preparation in order to estimate the number of amplifiable template copies present in the cDNA sample. This method was preferred rather than directly using pVL values from clinical testing, as the latter can underestimate the presence of amplifiable templates in the sample.

3.1.3 Library preparation—The input for library preparation was set at 2,000 genomic templates, as power calculations indicated that a smaller input would have <90% probability of sampling variants present at 0.1% (Supplemental figure 2). To allow for optimal reverse transcription, amplification, and representation of all templates, the input was distributed into separate reverse transcription reactions (input<1,000 RNA copies/reaction), which were then distributed into multiple nested-PCRs (input<250 cDNA copies/reaction) (see supplemental Table 1). In order to allow for bidirectional sequencing, two sets of second round nested-PCRs were prepared per targeted region, alternating the location of the P1 and A IonTorrent adapters.

3.1.4 Sequencing—The intrinsic error rate of NGS calls for oversequencing each template (see discussion by Larsen et al. (2013)). In the current study, 50-100K reads were obtained for an input of 2,000 genome templates, which equated to 25-50x coverage per template (Supplemental Table 3).

3.1.5 Data analysis—Ion sequencing uses terminator-free chemistry, which leads to higher error rates in homopolymeric tracts (Rothberg et al., 2011). Due to its dependence on sequence context, this kind of error can be studied using bidirectional sequencing. In the current study, noise was distinguished from true genetic variation by comparing the frequency of each variant derived from forward and reverse reads. In addition, as the vast majority of the reads spanned the whole length of the target, the information derived from haplotyping was used to decrease the lower limit of detection of complex variants. When evidencing the co-circulation of two or more variants that can be distinguished by multiple polymorphisms, under the assumption of sequencing error being independent at each position, the probability of one of them deriving from the other solely by the occurrence of multiple sequencing errors is defined by:

$$P(A) = P(E_1) \times P(E_2) \times P(E_i) \times \dots \times P(E_n) \quad \text{Eq.(1)}$$

where:

$P(A)$ = probability of variant #1 being an artifact originating exclusively by sequencing errors in variant #2

$P(E_i)$ = sequencing error at position i

n = number of mutations distinguishing variant #1 from variant #2.

A conservative estimate would be:

$$P(A) \leq \{ \max[P(E_1), \dots, P(E_n)] \}^n \quad \text{Eq.(2)}$$

For instance, considering a maximum per-base error rate of 0.5%, in the case of the detection of two variants differing in four polymorphisms, the probability of this being a

spurious result would be 6.3×10^{-9} . If they were to differ by ten polymorphisms, the probability would be 6.3×10^{-24} .

3.2 Level and context of background noise

The level of background sequencing noise was determined based on a plasmid sequence (CAT gene from *E. coli*) and two subgenomic areas from a primary HIV-1 specimen sampled at pre-peak viremia (*gp120-V2* and *gp120-C3* regions). Figure 1a shows that the error rate (i.e., the frequency of nucleotide bases discrepant from the consensus sequence) was uniform throughout the CAT sequence. The 95th percentile of the error rate was 0.30%, the 97.5th percentile was 0.35%, and the 99th percentile was 0.42%. High concordance was observed among nucleotide base frequencies derived from forward and reverse reads (Figure 1b) (slope=0.9993; $r^2=1.000$; $p<0.0001$). In only one position the error rate was greater than 0.5% (i.e., 0.51%), and the frequency of the C nucleotide computed based on the forward reads was two-fold higher than in the reverse reads (Figure 1c).

Figure 2 shows the distribution of error rate in the *gp120-V2* region of the pre-peak viremia sample. The 95th percentile was 0.27%, the 97.5th percentile was 0.31%, and the 99th percentile was 0.73% (Figure 2a). When comparing the base frequencies at each site measured from forward and reverse reads, there was a great agreement (slope=0.9990; $r^2=1.000$; $p<0.0001$; Figure 2b). The error rate was evenly distributed throughout the sequence, and it was greater than 0.5% in only in three positions, which clustered in an area with two neighboring homopolymeric tracts (A_6GA_3). The observed high error rates derived from the miscalling of the number of As in the homopolymers: position #1 was affected by miscalling the sequence as A_5GA_4 (and mostly on the forward reads) while position #3 was affected by miscalling the sequence as A_7GA_2 (and mostly on the reverse reads) (Figure 2c-e, and Supplemental Figure 3 a and b); the higher frequency of A noted in position #2 resulted from the composite of miscalling errors at positions #1 and #3. Thus, all of the bases present at frequencies $>0.5\%$ in the *gp120-V2* region could be ascribed to homopolymer sequencing error. When excluding these positions, the distribution of error rate was: 95th percentile: 0.25%; 97.5th percentile: 0.28%; 99th percentile: 0.31%.

Figure 3 shows the distribution of error rate in the *gp210-C3* region of the pre-peak viremia sample. The 95th percentile was 0.27%, the 97.5th percentile was 0.32%, and the 99th percentile was 0.51%. When comparing the base frequencies at each site measured from forward and reverse reads, there was a great agreement (slope=0.9999; $r^2=1.000$; $p<0.0001$; Figure 3b). The error rate was evenly distributed throughout the sequence, with three notable exceptions, where error was greater than 0.45%, (Figure 3a); a detailed analysis of these positions revealed two distinct patterns. The first pattern was represented by positions #1 and #2, where the frequencies for the minor bases were concordant between forward and reverse reads (Figure 3c and d). Longitudinal data strongly supports the interpretation that the G detected at position #1 and the A detected in position #2 were true polymorphisms present at low level in the viral population (see section on longitudinal analysis). However, for position #3, there was an extreme discordance between forward and reverse reads (Figure 3e). A detailed inspection revealed that the high frequency of A at position #3 derived from miscalling the number of Gs in the homopolymer tract two bases upstream from position #3

(i.e., G₃ was recorded instead of G₄), combined with the miscalling of the CC dinucleotide (i.e., recording only one C instead of two Cs) and mapping errors (i.e., as the penalty for opening gaps was higher than the one for extending gaps or mismatch (see Materials and Methods section) the resulting read was aligned as “GGG--ACT” instead as “GGG-ACT”) (Supplemental Figure 2c). This was further confirmed when the original sequence and another variant, from a later time-point -where GGGG had mutated into GGAG, thus disturbing the homopolymeric tract-were cloned (Supplemental Figure 4). Mixing both variants at different proportions did not affect the frequency of A at position #3 measured in the reverse reads (range 0.000-0.012%) but substantially diminished the frequency of A based on the forward reads. Thus, in the gp210-C3 region, the higher error rate at one position neighboring a homopolymeric tract could be undoubtedly ascribed to homopolymer sequencing and mapping errors, while the detection of minor variants at >0.5% in two other positions could be ascribed to the presence of low-level polymorphisms. When excluding these three positions, the distribution of error rate was: 95th percentile: 0.25%; 97.5th percentile: 0.28%; 99th percentile: 0.29%.

Next the impact of in vitro recombination on the system was assessed. In this experiment two highly related molecular clones, which derived from actual genetic variants circulating in primary samples, and which could be distinguished based on six markers, were combined. To maximize the detectability of recombinants, the clones were mixed in equimolar proportions, and were subjected to library preparation. Recombinants were detected at 0.2-0.8% (Supplemental Figure 5); all of them could be ascribed to single crossover events. Most of the recombinants were one mutation removed from the parental strains, so some of the putative recombinants may in fact reflect some level of sequencing error rather than recombination; based on this, the estimated in vitro recombination rates would be slightly lower than the sequencing error rates.

In summary, these data indicate that for the vast majority of the sequence positions the expected sequencing error rate was <0.5%, which is at the same level as the in vitro recombination rate. On the other hand, the error rate in or around some homopolymeric tracts was 2- to 3-fold higher, and could be distinguished from actual signals by bidirectional sequencing.

3.3 Longitudinal TDS analysis

After determining that the system's noise level was <0.5% and showing that true polymorphisms could be distinguish from spurious signals by bidirectional sequencing and contextual sequence analysis, the system's suitability for the study of HIV-1 evolution was determined using longitudinal samples. The analysis was carried out with specimens sampled at pre-peak viremia (day 5 after the first detection of viral nucleic acids), immediate post-peak viremia (days 36), and 6-months post infection (day 178). For each region and each time-point, positions with SNPs at frequencies >0.5% (i.e., above the system's noise level) that could be validated by bidirectional sequencing were identified. As the length of the NGS reads covered the whole span of the region of interest it was possible to determine the physical linkage among polymorphism.

3.3.1 Longitudinal TDS analysis of gp120-V2—The analysis of the day5 sample had indicated the presence of a single variant (#1), and the same profile was observed at day 36 (Figure 4). At day 178, the original variant was a minor strain (2.8%) that was virtually eclipsed by a collection of seven other variants present at frequencies ranging 0.7%-51.4%. The majority of the positions did not show appreciable variation above 0.5%; polymorphic sites were limited to two areas (i.e., codon 171 and codons 185-190), which were located in different domains: an anti-parallel beta-strand and a loop, respectively. The V2 region in gp120 is a known target for host antibodies and it can tolerate extensive residue substitution and length variation, the latter concentrated between residues 187 and 190. All of the observed nucleotide substitutions resulted in non-synonymous amino acid changes, most of which were non-conservative substitutions. In addition, a single-codon in-frame deletion was detected. Thus, the observed variation likely represents escape from humoral immune responses. Interestingly, variation followed a pattern in which only one of the non-synonymous changes in codons 185-188 -or the deletion at codon 189 combined with the non-synonymous change at codon 190- was present, either with or without the non-synonymous change at codon 171. This indicates constraints to variation in codons 185-190 and also suggests separate selective pressures in the two domains. It should be noted that it is possible that variants #7 and #8 might have emerged not solely by mutation but also by in vivo recombination between variants #2 and #3. Considering that the variants #2 and #3 constituted the major strains at that time-point, and given the low frequencies of variants #7 and #8 (i.e., 0.7%-0.8%) it might be possible that the latter variants had resulted from in vitro recombination.

3.3.2 Longitudinal TDS analysis of gp120-C3—At day 5, variant #1 was present at 97.2% and was accompanied by two minor variants (variant #2: 2.1%, and variant #3: 0.5%), which differed from it in two different non-synonymous conservative substitutions (Figure 5). This pattern is consistent with acquisition of a single HIV-1 T/F virus with mutations arising in early replication cycles (Salazar-Gonzalez et al., 2009). At day 36, variant #1 still dominated (94.4%) and variants #2 and #3 were also observed, but at increased frequencies. In addition, a new variant emerged (variant #4; frequency: 0.6%), which presented a non-conservative amino acid change at position 347. By day 178, variants #1 and #4 were detected as minor strains (frequencies: 0.4%-0.7%). Additionally, eight new variants (variants #5 through #12) were observed at frequencies ranging 0.5%-64.3%, all of which differed from variant #1 in two to four non-synonymous substitutions -most of which were non-conservative- and some variants also incorporated synonymous changes. Most of the observed mutations occurred inside the alpha2-helix in the C3 region (defined by codons 335- 352), where alternations between positively- and negatively-charged amino acids are very frequent and they have been ascribed to immune escape (Gnanakaran et al., 2007; Moore et al., 2009). Figure 5c shows a likely reconstruction of the evolution of the different variants. Variant #1 gave rise independently to variants #2, #3, and #4 by single-base substitutions. Variant #4 gave rise to variant #6, which in turn gave rise to variants #7, #10, #11, and #12, also by independent single-base substitutions. Variant #8 seems to have emerged either by single-base substitutions from variant #2 or by recombination between variants #2 and #7. On the other hand, variants #5 and #9 seem to have emerged from variant #6 through multiple G-to-A substitutions in GpG or GpA dinucleotide context. The high

number of accumulated changes, the failure to detect the intermediate variants (i.e., “missing links”), and the context of the G-to-A mutations convergently support the hypothesis that variants #5 and #9 derived from variant #6 by action of host cytidine deaminases (e.g., APOBEC3F and APOBEC3G) rather than by mutations caused by the viral reverse transcriptase (RT) (Kijak et al., 2008). It should be noted that it is possible that variants #9 and #11 might have emerged not solely by mutation but also by in vivo recombination between variants #5 and #6. Considering that the variants #5 and #6 constituted the major strains at that time-point, and given the low frequencies of variants #9 and #11 (i.e., 0.6%-0.8%) it might be possible that the latter variants had resulted from in vitro recombination.

In summary, the longitudinal analysis of the evolution of the gp120-V2 region indicated the presence of a single variant at pre-peak viremia and immediate post-peak viremia. By six months post-infection the original strain was virtually replaced by a constellation of complex variants. The emerging strains presented non-synonymous changes and deletions concentrated in two sub-regions, which probably evolved under separate selective pressures. On the other hand, gp120-C3 showed earlier signs of genetic variation, with minor variants bearing non-synonymous substitution already present at pre-peak viremia and immediate post-peak viremia. One of these variants gave subsequently rise to a new strain which further diversified by mutations mediated by HIV-1 RT and hosts APOBECs. In both regions the low-level persistence of the original variants was noticed at six months post infection, which may reflect the reactivation of latent cells that had been infected earlier in the infection.

4. Discussion

The present work provides the first systematic assessment of suitability of the IonTorrent PGM platform for the TDS-based study of HIV-1 evolution. A combination of laboratory reagents and primary HIV-1 specimens was employed to determine the rates of misincorporation and recombination error.

The TDS strategy presented in the current study was based on defining the number of input template molecules for library preparation and oversequencing each template, following the approach by Larsen et al (Larsen et al., 2013). Pelak et al. have reported that optimal coverage to determine variants through whole human genome sequencing was 30-35x (i.e., 15-18x per genetic complement) (Pelak et al., 2010). In the current experiments, the coverage was 25-50x per template. Direct comparison with the existing HIV-1 TDS literature is difficult because most publications tend only to report the overall number of obtained reads, but do not set this in the context of the number of input genome templates. This also impedes direct comparisons on misincorporation error rates. Still operational cut-off of 0.5% used in the current study is in agreement with the ones used in most of the literature (Archer et al., 2012; Gibson et al., 2014; Shao et al., 2013; Tsibris et al., 2009).

In agreement with experiments performed on the 454 platform (Shao et al., 2013), the results from the current study show that the capacity to detect minor variants in or around homopolymeric tracks is subjected to a higher cutoff value. In this study it was shown that the noise level in these cases was 2- to 3-fold higher than in other contexts, and that, assisted

by bidirectional sequencing, it was possible to distinguish homopolymeric sequencing error from true polymorphisms.

The estimates of in vitro recombination from the current study are in agreement with previous reports (Hedskog et al., 2010; Mild et al., 2011; Shao et al., 2013). Another indication that in vitro recombination levels were very low came from primary clinical samples from day-178. When multiple variants co-circulated and the two major ones were at 51.4%/27.0% (gp120-V2) and 64.3%/30.0% (gp120-C3), variants that could potentially have derived from in vitro recombination among them (i.e., gp120-V2: #7 and #8; gp120-C3: #9 and #11) were present at very low frequencies (0.6%-0.8%). These results indicate that the current platform will be suitable for the study of the rates of *in vivo* recombination.

The experiments from the current study have not only shown the suitability of the current platform for the study of HIV-1 evolution but also demonstrate that TDS can provide an augmented view of the level and mechanisms of viral genetic variation. The coupling of several factors was critical to the modeling of viral evolution: 1) availability of longitudinal samples, 2) the deep sampling and sequencing, and 3) the ability to link polymorphisms. The first factor was necessary to determine the temporal precedence of some variants over others and provided directionality to the model. The second and the third factors were lent by the TDS approach and were essential to describe the level of genetic diversity of the viral swarms. Deep sampling and sequencing revealed the presence of a multitude of minor variants, some of which were intermediate forms in the chain linking early and late strains. The exclusive use of traditional methods would have missed these minor variants and would have provided an incomplete picture of the timing and pathway of HIV-1 evolution. Finally, determining the linkage among polymorphisms permitted to evidence the covariation patterns, including the action of cytidine deaminases, and helped define the extent and boundaries of the mutational space explored by the virus.

NGS constitutes one of the fastest evolving technologies to date, with improved reagents rapidly replacing older-generation kits. The results reported here thus constitute the present state of the art. The future will likely bring technologies with increased read length and higher accuracy; nevertheless the current findings demonstrate the contribution that the current technology can make to expand the knowledge of the evolution of viral swarms.

A major advantage of TDS compared to other approaches is that it not only determines the frequency of each base at each position but also informs about the linkage among polymorphisms. This can help trace the ontogeny of HIV-1 variants, allows for the study of covariation, and can assist in the development of structural models and antigen design.

4. Conclusion

The performance of the tested system supports the conclusion that it can be applied for the characterization of minor viral variants in support of studies of viral evolution. This knowledge will be essential to inform the rational design of the next generation of vaccines and therapeutics.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are thankful to the subjects participating in the RV217/ECHO study and to the personnel at the Kenya Medical Research Institute/Walter Reed Project, Kericho, Kenya for their help collecting specimens, epidemiological data, and laboratory technical assistance. The authors would also like to thank Mr. Adam Bates for technical assistance, and Dr. Morgane Rolland for critical reading of the manuscript. This work was supported in part by an Interagency Agreement (Y1-AI-2642-12) between the U.S. Army Medical Research and Materiel Command and the National Institute of Allergy and Infectious Diseases. This work was also supported by a cooperative agreement (W81XWH-07-2-0067) between the Henry M. Jackson Foundation for the Advancement of Military Medicine and the U.S. Department of Defense. The opinions expressed in this article are those of the authors and do not represent the official views of the U.S. Department of Health and Human Services, the National Institute of Allergy and Infectious Diseases, the U.S. Department of Defense, or the Department of the Army.

References

- Archer J, Weber J, Henry K, Winner D, Gibson R, Lee L, Paxinos E, Arts EJ, Robertson DL, Mimms L, Quinones-Mateu ME. Use of four next-generation sequencing platforms to determine HIV-1 coreceptor tropism. *PLoS One*. 2012; 7:e49602. [PubMed: 23166726]
- 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 CF, Ceccherini-Silberstein F. 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*. 2012; 205:557–67. [PubMed: 22238474]
- Brockman MA, Schneidewind A, Lahaie M, Schmidt A, Miura T, Desouza I, Ryvkin F, Derdeyn CA, Allen S, Hunter E, Mulenga J, Goepfert PA, Walker BD, Allen TM. Escape and compensation from early HLA-B57-mediated cytotoxic T-lymphocyte pressure on human immunodeficiency virus type 1 Gag alter capsid interactions with cyclophilin A. *J Virol*. 2007; 81:12608–18. [PubMed: 17728232]
- Cai F, Chen H, Hicks CB, Bartlett JA, Zhu J, Gao F. Detection of minor drug-resistant populations by parallel allele-specific sequencing. *Nat Methods*. 2007; 4:123–5. [PubMed: 17206150]
- Chang, E. Next-Generation Sequencing Methods: Impact of Sequencing Accuracy on SNP Discovery. In: Komar, A., editor. *Single Nucleotide Polymorphisms*. Humana Press; New York, NY: 2009.
- Chang MW, Oliveira G, Yuan J, Okulicz JF, Levy S, Torbett BE. Rapid deep sequencing of patient-derived HIV with ion semiconductor technology. *J Virol Methods*. 2013; 189:232–4. [PubMed: 23384677]
- Clavel F, Hoggan MD, Willey RL, Strebel K, Martin MA, Repaske R. Genetic recombination of human immunodeficiency virus. *J Virol*. 1989; 63:1455–9. [PubMed: 2915387]
- Coffin J, Swanstrom R. HIV pathogenesis: dynamics and genetics of viral populations and infected cells. *Cold Spring Harb Perspect Med*. 2013; 3:a012526. [PubMed: 23284080]
- Di Giallonardo F, Zagordi O, Duport Y, Leemann C, Joos B, Kunzli-Gontarczyk M, Bruggmann R, Beerenwinkel N, Gunthard HF, Metzner KJ. Next-generation sequencing of HIV-1 RNA genomes: determination of error rates and minimizing artificial recombination. *PLoS One*. 2013; 8:e74249. [PubMed: 24058534]
- Fischer W, Ganosov VV, Giorgi EE, Hraber PT, Keele BF, Leitner T, Han CS, Gleasner CD, Green L, Lo CC, Nag A, Wallstrom TC, Wang S, McMichael AJ, Haynes BF, Hahn BH, Perelson AS, Borrow P, Shaw GM, Bhattacharya T, Korber BT. Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PLoS One*. 2010; 5:e12303. [PubMed: 20808830]
- Gibson RM, Meyer AM, Winner D, Archer J, Feyertag F, Ruiz-Mateos E, Leal M, Robertson DL, Schmotzer CL, Quinones-Mateu ME. Sensitive Deep-Sequencing-Based HIV-1 Genotyping Assay To Simultaneously Determine Susceptibility to Protease, Reverse Transcriptase, Integrase, and

Maturation Inhibitors, as Well as HIV-1 Coreceptor Tropism. *Antimicrob Agents Chemother.* 2014; 58:2167–85. [PubMed: 24468782]

- Gnanakaran S, Lang D, Daniels M, Bhattacharya T, Derdeyn CA, Korber B. Clade-specific differences between human immunodeficiency virus type 1 clades B and C: diversity and correlations in C3-V4 regions of gp120. *J Virol.* 2007; 81:4886–91. [PubMed: 17166900]
- Goonetilleke N, Liu MK, Salazar-Gonzalez JF, Ferrari G, Giorgi E, Gnanusov VV, Keele BF, Learn GH, Turnbull EL, Salazar MG, Weinhold KJ, Moore S, Letvin N, Haynes BF, Cohen MS, Hraber P, Bhattacharya T, Borrow P, Perelson AS, Hahn BH, Shaw GM, Korber BT, McMichael AJ. The first T cell response to transmitted/founder virus contributes to the control of acute viremia in HIV-1 infection. *J Exp Med.* 2009; 206:1253–72. [PubMed: 19487423]
- Hedskog C, Mild M, Jernberg J, Sherwood E, Bratt G, Leitner T, Lundeberg J, Andersson B, Albert J. Dynamics of HIV-1 quasispecies during antiviral treatment dissected using ultra-deep pyrosequencing. *PLoS One.* 2010; 5:e11345. [PubMed: 20628644]
- Hedskog C, Brodin J, Heddini A, Bratt G, Albert J, Mild M. Longitudinal ultra-deep characterization of HIV type 1 R5 and X4 subpopulations in patients followed from primary infection to coreceptor switch. *AIDS Res Hum Retroviruses.* 2013; 29:1237–44. [PubMed: 23745959]
- 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. [PubMed: 18490657]
- Kijak GH, Janini LM, Tovanabutra S, Sanders-Buell E, Arroyo MA, Robb ML, Michael NL, Birx DL, McCutchan FE. Variable contexts and levels of hypermutation in HIV-1 proviral genomes recovered from primary peripheral blood mononuclear cells. *Virology.* 2008; 376:101–11. [PubMed: 18436274]
- Kijak, G.; Sanders-Buell, W.; Rolland, M.; Li, H.; Bates, A.; Bose, M.; O'Sullivan, A.; Eller, L.; O'Connell, R.; Shaw, G.; Michael, N.; Kim, J.; Robb, M.; Tovanabutra, S. *AIDS Vaccine.* Boston, MA: 2012. Incident cases characterization and deep sequencing provide new insight into multiplicity of infection and HIV evolution in very early acute infection.
- Kijak GH, Pham P, Sanders-Buell E, Harbolick EA, Eller LA, Robb ML, Michael NL, Kim JH, Tovanabutra S. Nautilus: a bioinformatics package for the analysis of HIV type 1 targeted deep sequencing data. *AIDS Res Hum Retroviruses.* 2013; 29:1361–4. [PubMed: 23809062]
- Korber, B.; Foley, B.; Kuiken, C.; Pillai, S.; Sodroski, J. Numbering Positions in HIV Relative to HXB2CG. In: Korber, B.; Kuiken, C.; Foley, B.; Hahn, B.; McCutchan, F.; Mellors, J.; Sodroski, J., editors. *Human Retroviruses and AIDS Theoretical Biology and Biophysics Group.* Los Alamos National Laboratory; Los Alamos, NM: 1998. p. III-102-111.
- Larder BA, Kemp SD. Multiple mutations in HIV-1 reverse transcriptase confer high-level resistance to zidovudine (AZT). *Science.* 1989; 246:1155–8. [PubMed: 2479983]
- Larsen BB, Chen L, Maust BS, Kim M, Zhao H, Deng W, Westfall D, Beck I, Frenkel LM, Mullins JI. Improved Detection of Rare HIV-1 Variants using 454 Pyrosequencing. *PLoS One.* 2013; 8:e76502. [PubMed: 24098517]
- Leslie AJ, Pfafferoth KJ, Chetty P, Draenert R, Addo MM, Feeney M, Tang Y, Holmes EC, Allen T, Prado JG, Altfeld M, Brander C, Dixon C, Ramduth D, Jeena P, Thomas SA, St John A, Roach TA, Kupfer B, Luzzi G, Edwards A, Taylor G, Lyall H, Tudor-Williams G, Novelli V, Martinez-Picado J, Kiepiela P, Walker BD, Goulder PJ. HIV evolution: CTL escape mutation and reversion after transmission. *Nat Med.* 2004; 10:282–9. [PubMed: 14770175]
- Li JZ, Paredes R, Ribaudo HJ, Svarovskaia ES, Metzner KJ, Kozal MJ, Hullsiek KH, Balduin M, Jakobsen MR, Geretti AM, Thiebaut R, Ostergaard L, Masquelier B, Johnson JA, Miller MD, Kuritzkes DR. Low-frequency HIV-1 drug resistance mutations and risk of NNRTI-based antiretroviral treatment failure: a systematic review and pooled analysis. *J Amer Med Assoc.* 2011; 305:1327–35.
- McCutchan FE, Hoelscher M, Tovanabutra S, Piyasirisilp S, Sanders-Buell E, Ramos G, Jagodzinski L, Polonis V, Maboko L, Mmbando D, Hoffmann O, Riedner G, von Sonnenburg F, Robb M, Birx

DL. In-depth analysis of a heterosexually acquired human immunodeficiency virus type 1 superinfection: evolution, temporal fluctuation, and intercompartment dynamics from the seronegative window period through 30 months postinfection. *J Virol*. 2005; 79:11693–704. [PubMed: 16140747]

Messiaen P, Verhofstede C, Vandenbroucke I, Dinakis S, Van Eygen V, Thys K, Winters B, Aerssens J, Vogelaers D, Stuyver LJ, Vandekerckhove L. Ultra-deep sequencing of HIV-1 reverse transcriptase before start of an NNRTI-based regimen in treatment-naïve patients. *Virology*. 2012; 426:7–11. [PubMed: 22305619]

Mild M, Hedskog C, Jernberg J, Albert J. Performance of ultra-deep pyrosequencing in analysis of HIV-1 pol gene variation. *PLoS One*. 2011; 6:e22741. [PubMed: 21799940]

Milne I, Stephen G, Bayer M, Cock PJ, Pritchard L, Cardle L, Shaw PD, Marshall D. Using Tablet for visual exploration of second-generation sequencing data. *Brief Bioinform*. 2013; 14:193–202. [PubMed: 22445902]

Moore PL, Ranchobe N, Lambson BE, Gray ES, Cave E, Abrahams MR, Bandawe G, Mlisana K, Abdool Karim SS, Williamson C, Morris L. Limited neutralizing antibody specificities drive neutralization escape in early HIV-1 subtype C infection. *PLoS Pathog*. 2009; 5:e1000598. [PubMed: 19763271]

Mukherjee R, Jensen ST, Male F, Bittinger K, Hodinka RL, Miller MD, Bushman FD. Switching between raltegravir resistance pathways analyzed by deep sequencing. *Aids*. 2011; 25:1951–9. [PubMed: 21832937]

Palmer S, Kearney M, Maldarelli F, Halvas EK, Bixby CJ, Bazmi H, Rock D, Falloon J, Davey RT Jr, Dewar RL, Metcalf JA, Hammer S, Mellors JW, Coffin JM. Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. *J Clin Microbiol*. 2005; 43:406–13. [PubMed: 15635002]

Pelak K, Shianna KV, Ge D, Maia JM, Zhu M, Smith JP, Cirulli ET, Fellay J, Dickson SP, Gumbs CE, Heinzen EL, Need AC, Ruzzo EK, Singh A, Campbell CR, Hong LK, Lornsen KA, McKenzie AM, Sobreira NL, Hoover-Fong JE, Milner JD, Ottman R, Haynes BF, Goedert JJ, Goldstein DB. The characterization of twenty sequenced human genomes. *PLoS Genet*. 2010; 6:e1001111. [PubMed: 20838461]

Piyamongkol W, Bermudez MG, Harper JC, Wells D. Detailed investigation of factors influencing amplification efficiency and allele drop-out in single cell PCR: implications for preimplantation genetic diagnosis. *Mol Hum Reprod*. 2003; 9:411–20. [PubMed: 12802048]

Poon AF, Swenson LC, Dong WW, Deng W, Kosakovsky Pond SL, Brumme ZL, Mullins JI, Richman DD, Harrigan PR, Frost SD. Phylogenetic analysis of population-based and deep sequencing data to identify coevolving sites in the nef gene of HIV-1. *Mol Biol Evol*. 2010; 27:819–32. [PubMed: 19955476]

Redd AD, Collinson-Streng A, Martens C, Ricklefs S, Mullis CE, Manucci J, Tobian AA, Selig EJ, Laeyendecker O, Sewankambo N, Gray RH, Serwadda D, Wawer MJ, Porcella SF, Quinn TC. Identification of HIV superinfection in seroconcordant couples in Rakai, Uganda, by use of next-generation deep sequencing. *J Clin Microbiol*. 2011; 49:2859–67. [PubMed: 21697329]

Redd AD, Collinson-Streng AN, Chatziandreou N, Mullis CE, Laeyendecker O, Martens C, Ricklefs S, Kiwanuka N, Nyein PH, Lutalo T, Grabowski MK, Kong X, Manucci J, Sewankambo N, Wawer MJ, Gray RH, Porcella SF, Fauci AS, Sagar M, Serwadda D, Quinn TC. Previously transmitted HIV-1 strains are preferentially selected during subsequent sexual transmissions. *J Infect Dis*. 2012; 206:1433–42. [PubMed: 22997233]

Richman DD, Havlir D, Corbeil J, Looney D, Ignacio C, Spector SA, Sullivan J, Cheeseman S, Barringer K, Pauletti D, et al. Nevirapine resistance mutations of human immunodeficiency virus type 1 selected during therapy. *J Virol*. 1994; 68:1660–6. [PubMed: 7509000]

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, Hoon J, Simons JF, Marran D, Myers JW, Davidson JF, Branting A, Nobile JR, Puc BP, Light D, Clark TA, Huber M, Branciforte JT, Stoner IB, Cawley SE, Lyons M, Fu Y, Homer N, Sedova M, Miao X, Reed B, Sabina J, Feierstein E, Schorn M, Alanjary M, Dimalanta E, Dressman D, Kasinskas R, Sokolsky T, Fidanza JA, Namsaraev E, McKernan KJ, Williams A, Roth GT, Bustillo J. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*. 2011; 475:348–52. [PubMed: 21776081]

- Rozera G, Abbate I, Bruxelles 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 quasiespecies deriving from lymphomonocyte sub-populations. *Retrovirology*. 2009; 6:15. [PubMed: 19216757]
- Salazar-Gonzalez JF, Salazar MG, Keele BF, Learn GH, Giorgi EE, Li H, Decker JM, Wang S, Baalwa J, Kraus MH, Parrish NF, Shaw KS, Guffey MB, Bar KJ, Davis KL, Ochsenbauer-Jambor C, Kappes JC, Saag MS, Cohen MS, Mulenga J, Derdeyn CA, Allen S, Hunter E, Markowitz M, Hraber P, Perelson AS, Bhattacharya T, Haynes BF, Korber BT, Hahn BH, Shaw GM. Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. *J Exp Med*. 2009; 206:1273–89. [PubMed: 19487424]
- Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H, Gupta P, Rinaldo CR, Learn GH, He X, Huang XL, Mullins JI. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J Virol*. 1999; 73:10489–502. [PubMed: 10559367]
- 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. [PubMed: 23402264]
- Simen BB, Simons JF, Hullsiek KH, Novak RM, Macarthur RD, Baxter JD, Huang C, Lubeski C, Turenchalk GS, Braverman MS, Desany B, Rothberg JM, Egholm M, Kozal MJ. 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. [PubMed: 19210162]
- Swenson LC, Moores A, Low AJ, Thielen A, Dong W, Woods C, Jensen MA, Wynhoven B, Chan D, Glascock C, Harrigan PR. 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*. 2010; 54:506–10. [PubMed: 20512044]
- Swenson LC, Mo T, Dong WW, Zhong X, Woods CK, Jensen MA, Thielen A, Chapman D, Lewis M, James I, Heera J, Valdez H, Harrigan PR. Deep sequencing to infer HIV-1 co-receptor usage: application to three clinical trials of maraviroc in treatment-experienced patients. *J Infect Dis*. 2011; 203:237–45. [PubMed: 21288824]
- Tsibris AM, Korber B, Arnaout R, Russ C, Lo CC, Leitner T, Gaschen B, Theiler J, Paredes R, Su Z, Hughes MD, Gulick RM, Greaves W, Coakley E, Flexner C, Nusbaum C, Kuritzkes DR. Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo. *PLoS One*. 2009; 4:e5683. [PubMed: 19479085]
- UNAIDS. Global Report: UNAIDS report on the global AIDS epidemic 2013. Geneva, Switzerland: 2013.
- 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, Stuyver LJ. HIV-1 V3 envelope deep sequencing for clinical plasma specimens failing in phenotypic tropism assays. *AIDS Res Ther*. 2010; 7:4. [PubMed: 20804564]
- Wei X, Youngpairaj AS, Garrido C, Zahonero N, Corral A, de Mendoza C, Heneine W, Johnson JA, Garcia-Lerma JG. Minority HIV mutation detection in dried blood spots indicates high specimen integrity and reveals hidden archived drug resistance. *J Clin Virol*. 2011; 50:148–52. [PubMed: 21130027]
- 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. [PubMed: 21521499]

Abbreviations

CAT	chloramphenicol acetyltransferase
CTL	cytotoxic T lymphocyte
DEPC	diethylpyrocarbonate

ePCR	emulsion polymerase chain reaction
NAT	nucleic acid testing
NGS	next-generation sequencing
PASS	parallel allele-specific sequencing
pVL	plasma viral load
RT	reverse transcriptase
SGA	single genome amplification
SNP	single nucleotide polymorphism
T/F	transmitted/founder
TDS	targeted deep sequencing
vRNA	viral RNA

Highlights

- Exploration of suitability of IonTorrent PGM for targeted deep sequencing of HIV-1
- The error rates from misincorporation and *in vitro* recombination were <0.5%
- TDS provided an augmented view of HIV-1 evolution during acute infection

Figure 1.

Sequencing error rate based on *E. coli* CAT. a) At each sequenced position, the frequency of each base differing from the reference sequence (shown in the abscissa) is depicted. The inset shows the distribution of the error rate, and the single position where the error rate was greater than 0.5% (#1) is highlighted. b) For each base sequenced at each position, the frequency derived from forward and reverse reads is shown. c) Detail of the frequency of C at position #1 computed based on forward and reverse reads. See text for details.

Figure 2.

Sequencing error rate based on the *gp120*-V2 region of a pre- peak viremia sample. a) At each sequenced position, the frequency of each base differing from the reference sequence (shown in the abscissa) is depicted. The inset shows the distribution of the error rate, and the three positions where the error rate was greater than 0.5% (#1, #2, and #3) are highlighted. b) For each base sequenced at each position, the frequency derived from forward and reverse reads is shown. c-e) Details of the frequency of G, A and G at positions #1, #2, and #3, respectively, computed based on forward and reverse reads. See text for details.

Figure 3.

Sequencing error rate based on the *gp120*-C3 region of a pre- peak viremia sample. a) At each sequenced position, the frequency of each base differing from the reference sequence (shown in the abscissa) is depicted. A G₄ homopolymeric tract responsible for sequencing error at position #3 is underlined. The inset shows the distribution of the error rate, and the three positions where the error rate was greater than 0.45% (#1, #2, and #3) are highlighted. b) For each base sequenced at each position, the frequency derived from forward and reverse reads is shown. c-e) Details of the frequency of G, A and A at positions #1, #2, and #3, respectively, computed based on forward and reverse reads. See text for details.

Figure 4.
TDS-based longitudinal analysis of the HIV-1 *gp120*-V2 region from pre-peak viremia through six months post-infection. a) The frequency of viral variants detected at each time-point are shown. b) Sequences of detected viral variants are depicted, and differences from the T/F virus (variant #1) are highlighted. Codon numbers in the HIV-1 *env* gene, relative to the HXB2 coordinate nomenclature, are indicated. See text for details.

Figure 5.

TDS-based longitudinal analysis of the HIV-1 *gp120*-C3 region from pre-peak viremia through six months post-infection. a) The frequency of viral variants detected at each time-point is shown. b) Sequences of detected viral variants are depicted, and differences from the T/F virus (variant #1) are highlighted. Codon numbers in the HIV-1 *env* gene, relative to the HXB2 coordinate nomenclature, are indicated. c) The figure shows a likely reconstruction of the evolution of the different variants. Color-coded diamonds and square boxes represent the corresponding variants. Large diamonds represent major variants, detectable by traditional SGA-based analyses, and small squares represent minor variants (frequency <5%) that would have been missed by traditional SGA-based analyses. Solid blue arrows depict single-base substitution events, red arrows depict multiple G-to-A substitution events, and dotted lines represent alternative single-base substitution or recombination events. The date of first detection of the variants is shown. See text for details.