Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID

Cassandra B. Jabara[a,b,c], Corbin D. Jones[a,d], Jeffrey Roach[e], Jeffrey A. Anderson[b,c,f,1], and Ronald Swanstrom[b,c,g,2]

[a]Department of Biology, [b]Lineberger Comprehensive Cancer Center, [c]University of North Carolina Center for AIDS Research, [d]Carolina Center for Genome Sciences, [e]Research Computing Center, [f]Division of Infectious Diseases, and [g]Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC 27599

Edited by John M. Coffin, Tufts University School of Medicine, Boston, MA, and approved November 8, 2011 (received for review June 24, 2011)

Viruses can create complex genetic populations within a host, and deep sequencing technologies allow extensive sampling of these populations. Limitations of these technologies, however, potentially bias this sampling, particularly when a PCR step precedes the sequencing protocol. Typically, an unknown number of templates are used in initiating the PCR amplification, and this can lead to unrecognized sequence resampling creating apparent homogeneity; also, PCR-mediated recombination can disrupt linkage, and differential amplification can skew allele frequency. Finally, misincorporation of nucleotides during PCR and errors during the sequencing protocol can inflate diversity. We have solved these problems by including a random sequence tag in the initial primer such that each template receives a unique Primer ID. After sequencing, repeated identification of a Primer ID reveals sequence resampling. These resampled sequences are then used to create an accurate consensus sequence for each template, correcting for recombination, allelic skewing, and misincorporation/sequencing errors. The resulting population of consensus sequences directly represents the initial sampled templates. We applied this approach to the HIV-1 protease (*pro*) gene to view the distribution of sequence variation of a complex viral population within a host. We identified major and minor polymorphisms at coding and noncoding positions. In addition, we observed dynamic genetic changes within the population during intermittent drug exposure, including the emergence of multiple resistant alleles. These results provide an unprecedented view of a complex viral population in the absence of PCR resampling.

drug resistance | genetic diversity | high throughput sequencing | HIV | population dynamics

High throughput sequencing allows the acquisition of large amounts of sequence data that can encompass entire genomes (1–4). With sufficient amounts of starting DNA, PCR is not needed before the library preparation step of the sequencing protocol. Sequencing miscalls inherent in high throughput sequencing approaches are resolved using multiple reads over a given base.

Deep sequencing can also capture the genetic diversity of viral populations (5–10), including intrahost populations derived from clinical samples. This approach offers the opportunity to view population diversity and dynamics and viral evolution in unprecedented detail. One place where the presence of minor variants is of immediate practical importance is in the detection of drug-resistant variants. Standard bulk sequencing methods typically miss allelic variants below 20% in frequency within a population (11, 12). Alternative assays can detect less abundant variants that confer drug resistance, but require a priori selection of sites and variants (13–23). Thus, deep sequencing approaches offer the opportunity to identify minor variants associated with resistance de novo with the goal of understanding their role in therapy failure.

Although screening for drug-resistant variants is a practical application of the deep sequencing technology, this technology also addresses broader questions of sequence diversity and structure for a complex population like HIV-1. However, the relatively high sequencing error rates of these technologies artificially increase genetic diversity, which confounds the detection of natural genetic variation especially when sequencing a highly heterogeneous viral population. Moreover, the use of PCR to amplify the amount of material before starting the sequencing protocol adds the potential for several serious artifacts (24–27): First, nucleotide misincorporation by the polymerase during many rounds of amplification artificially increases sequence diversity; second, artifactual recombination during amplification occurs when premature termination products prime a subsequent round of synthesis, which can obscure the linkage of two sequence polymorphisms (28, 29); third, differential amplification can skew allelic frequencies; and fourth, PCR amplification can create a significant mass of DNA from a small number of starting templates, which obscures the true sampling of the original population as these few starting templates/genomes get resampled in the PCR product, creating sequence resampling rather than the observation of independent genomes (30). Overall, these biases artificially decrease true diversity while introducing artifactual diversity and also skew allelic frequencies, which can lead to incongruence between the real and observed viral populations. Most investigators use statistical tools to attempt to control for the types of sequencing errors that are associated with each sequencing platform.

To make deep sequencing useful for complex populations, it is necessary to overcome PCR resampling, which is mistaken for sampling of the original population, and PCR and sequencing errors, which can be mistaken for diversity. As nucleotide misincorporation is largely random across sites and template switching/recombination is more likely to occur in the later cycles of a PCR (31), strategies that create a bulk or consensus sequence for each sampled template will call the correct base at each position. One approach to sampling highly heterogeneous populations, such as the HIV-1 *env* gene, is through endpoint dilution titration of the template before nested PCR, such that a single template is present in each PCR amplification (32–35). In addition to masking the misincorporations, PCR-mediated recombination produces recombinant templates identical to the parental sequence. Although highly accurate, this technique is labor-intensive and, as population sampling is dependent on the number of templates sequenced, this methodology does not lend itself to the identification of minor variants or to understanding the structure of a complex population, nor is it easily adaptable to a high throughput approach.

We have developed a high throughput technique for directly resolving the genetic diversity of a viral population. This technique avoids the recording of PCR and sequencing errors that

Author contributions: C.B.J., C.D.J., J.A.A., and R.S. designed research; C.B.J. and J.A.A. performed research; C.B.J., C.D.J., and J.R. contributed new reagents/analytic tools; C.B.J., C.D.J., J.R., J.A.A., and R.S. analyzed data; and C.B.J., C.D.J., and R.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]Present address: Discovery Medicine-Virology, Discovery Medicine and Clinical Pharmacology, 311 Pennington-Rocky Hill Rd., 8A-1.14, Bristol-Myers Squibb, Pennington, NJ 08543.

[2]To whom correspondence should be addressed. E-mail: risunc@med.unc.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1110064108/-/DCSupplemental.

create artificial diversity, and corrects for artificial allelic skewing and PCR resampling, revealing the original genomes in the population. This is accomplished by embedding a degenerate block of nucleotides within the primer used in the first round of cDNA synthesis. This creates a random library of sequences within the primer population. As primers are individually used out of this library, each viral template is copied such that the complement (cDNA) now includes a unique sequence tag, or Primer ID. This Primer ID is carried through all of the subsequent manipulations to mark all sequences that derive from each independent templating event, and PCR resampling then becomes over-coverage for each template to create a consensus sequence of that template. Using this approach, we were able to directly remove error, correct for PCR resampling, and capture the fluctuation of minor variants in the viral population within a host. We also resolved minor drug-resistant variants below 1% in frequency before the initiation of antiretroviral therapy, and were able to correlate these variants with the emergence of drug resistance. The value of this strategy and its applicability to other deep sequencing protocols has been further emphasized through a recent parallel effort by Kinde et al. in short read sequencing of human genomic material (36).

Results

A cDNA Synthesis Primer Containing a Primer ID Can Be Used to Track Individual Viral Templates. A population of cDNA synthesis primers was designed to prime DNA synthesis downstream of the HIV-1 protease (*pro*) gene, with the primer containing two additional blocks of identifying information (Fig. 1*A*). The first block was a string of eight degenerate nucleotides that created 65,536 distinct sequence combinations (4^8), or Primer IDs. This region was flanked by an a priori selected three nucleotide barcode, creating a sample identification block so that multiple samples could be pooled together in a sequencing run (7). A designed sequence at the 5′ end of the cDNA primer was used for subsequent amplification of the cDNA sequences by nested PCR.

Viral RNA was extracted from three longitudinal blood plasma samples from an individual infected with subtype B HIV-1 who was participating in a protease inhibitor efficacy trial (M94-247) (ref. 37; Fig. S1). Approximately 10,000 copies of viral RNA from each sample were used in a reverse transcription reaction for cDNA synthesis and tagging using the Primer ID. The cDNA product was separated from the unused cDNA primers, and then the viral sequences were amplified by nested PCR and sequenced on the 454 GS FLX Titanium. Our data were distilled from total reads of 20,429, 24,658, and 27,075 for the three time points (T1, T2, and T3, respectively). Raw sequence reads were assessed for the cDNA tagging primer and a full length *pro* gene sequence (297 nucleotides long representing 99 codons), and when three or more sequences within a sample contained an identical Primer ID, a consensus sequence was formed to represent one sequence/genome in the population (Fig. 1 *B* and *C* and Fig. S2).

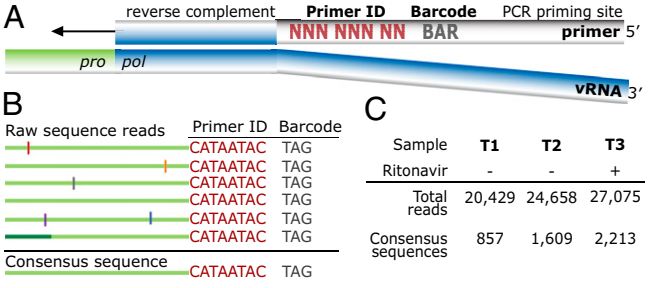

Fig. 1. Tagging viral RNA templates with a Primer ID before PCR amplification and sequencing allows for direct removal of artifactual errors and identifies resampling. (*A*) A primer was designed to bind downstream of the protease coding domain. In the 5′ tail of the primer, a degenerate string of eight nucleotides created a Primer ID, allowing for 65,536 unique combinations. An a priori selected three nucleotide barcode was designed for the sample ID. Finally, a heterologous string of nucleotides with low affinity to the HIV-1 genome was included in the far 5′ end for use as the priming site in the PCR amplification. (*B*) PCR biases and sequencing error are introduced during amplification and sequencing of viral templates. Repetitive identification of the barcode and Primer ID allow for tracking of each templating event from a single tagged cDNA. As errors are minor components within the Primer ID population, forming a consensus sequence directly removes them, and corrects for PCR resampling. (*C*) HIV-1 RNA templates isolated from plasma samples from two pre- and one postintermittent ritonavir drug therapy were tagged, amplified, and deep sequenced. Tagged sequences containing full-length protease were used to create a population of consensus sequences when at least three sequences contained an identical barcode and Primer ID.

With these manipulations we generated 857, 1,609, and 2,213 consensus sequences, respectively, for the three time points (Fig. 1*C*). The median number of reads per Primer ID was 6, ranging from 1 to 96 (Fig. S3*A*). The distribution of identical Primer IDs did not form a normal distribution as would be expected if all templates were amplified equally. We saw a higher than expected number of single reads of Primer IDs; although we do not know the reason for this, such a result is consistent with different cDNA templates entering the PCR at different cycles. Because each template is individually tagged the different number of reads is an indication of allelic skewing, as noted this can be nearly 100-fold. In an analysis of a number of low abundant variants we saw a 20-fold range of representation through allelic skewing, with half of the variants up to 2- to 3-fold more abundant than the mean, and the other half up to 5- to 10-fold less abundant (Fig. S4).

We conservatively estimate the combined in vitro error rate of the cDNA synthesis step by reverse transcriptase (RT) and the first strand synthesis by the Taq polymerase to be on the order of 1 mutation in 10,000 bases, or approximately one mutation per 33 *pro* gene sequences, based on an RT error rate of 1 in 22,000 nucleotides (38) and a Taq polymerase error rate of 1.1 in 10,000 nucleotides (39) but reduced by half because only the first round of synthesis is relevant and a misincorporation at this step gives a mixture. Later rounds of Taq polymerase errors should be largely lost through the creation of the consensus sequence. Thus, we would expect 139 sequence misincorporations to be present in the data set of 4,679 total sequences representing T1+T2+T3, and with an excess of transitions. These would be expected to occur as 113 single copy single-nucleotide polymorphisms (SNPs) and 13 SNPs that appeared twice. We observed 98 single copy SNPs in the data set with a threefold excess of transitions, and with three-fourths of them being coding changes, which is consistent with random mutations. We expect there to be low frequency SNPs in the viral population from rare but persistent variants that are fortuitously sampled, and from the intrinsic error rate of viral replication (the error rate during one round of viral replication would represent approximately one mutation per 150 *pro* gene sequences; ref. 24). However, we cannot distinguish real polymorphisms from the inferred background error rate associated with the first and second rounds of in vitro DNA synthesis. Thus, we have limited the analysis of population diversity to SNPs that appeared at least twice in the data set (i.e., linked to at least two separate Primer IDs), either at the same time point or at multiple time points in the overall data set (Table S1). We have not corrected the data set for the presumed 13 SNPs that appeared twice that are expected to be present due to error even though this represents 33% of all of the SNPs that appeared twice (13 of 39). Overall, 80% of the SNPs (i.e., any sequence change from the consensus that appeared at least once) in the total data set of 72,162 sequence reads were removed as error. Also, 60–65% of the sequence reads were revealed as resampling. Finally, allelic skewing of up to nearly 100 fold was corrected (Fig. S4).

Longitudinal Sequencing of the HIV-1 Protease (*pro*) Gene in an Untreated Individual Reveals Dynamic Changes in Genetic Variation. We analyzed the sequences of the *pro* gene populations to assess allelic frequency at the two sampled time points, separated by 6 mo

MICROBIOLOGY

and before ritonavir (37) drug selection (Fig. S1). The combined sequence population from the two time points (T1 and T2) before therapy consisted of 492 unique *pro* gene sequences with 155 SNPs. About 4% (i.e., 21) of these unique gene sequences were above 0.5% abundance, and these 21 unique gene sequences represented 67% of all sampled genomes, with the genome representing the overall consensus sequence comprising 21% of the total population (Fig. S5 *A* and *B*). The relatively small number of unique gene sequences above 0.5% frequency in the population contained only 7% of the 155 detected SNPs. Thus, a large proportion of the viral population's diversity was associated with a large number of *pro* gene sequences that were present at low abundance (Fig. S5 *A* and *C*); conversely, the majority of the population consisted of a small number of SNPs. Similarly, Tajima's *D* statistic for T1 and T2 in this individual were −2.35 and −2.31, respectively (Table S2), indicative of a population structure that has an excess of low frequency polymorphisms. This pattern is consistent with but more extreme than that observed in a prior shallow intrahost survey in which a metapopulation model was proposed to explain the pattern of Tajima's *D* statistic (40). Fig. 2 shows the encoded amino acid variability and synonymous nucleotide variability present in two or more individual genomes across the 99 codons in the *pro* gene for these samples.

Synonymous variability. There were 57 codons (with 63 variants/SNPs) that contained synonymous diversity that appeared in both pretherapy time points, and 30 codons (with 31 variants) that appeared in only one time point. Taken together, 75 of the 99 codons contained some level of synonymous diversity (Fig. 2 and Table S1). Of the 63 variants that were present in both untreated time points, 92% were transitions. Of the 31 variants that appeared in only one of the time points, 71% were transitions, representing a significantly smaller fraction of transitions than among the synonymous variants that appeared at both time points ($P = 0.012$; Fisher's exact test). This suggests that synonymous transversions are selected against over time.

Nonsynonymous variability. There were 26 codons (28 variants) that contained coding variability that appeared in both pretherapy time points, and an additional 28 codons (33 variants) with nonsynonymous changes found in only one of the time points. Taken together, 49 of the 99 codons contained some level of nonsynonymous diversity (Fig. 2 and Table S1). For the 28 nonsynonymous variants detected at both time points, 22 were transitions, and these mostly represented conservative amino acid changes. In the case of synonymous mutations two-thirds of the variants were present at both time points, whereas in the case of nonsynonymous mutations, less than half were present at both time points ($P = 0.012$; Fisher's exact test). This observation suggests that, at this level of sequence sampling, we are able to see a difference in stability within the population in comparing synonymous and nonsynonymous substitutions.

Genetic fluctuation. We compared the stability of minor SNPs present at both T1 and T2. A total of 14 of the 91 SNPs (synonymous and nonsynonymous that appeared at both time points) had significant changes in abundance between the two time points (χ^2 test with a false discovery rate of 0.05). Of the 14 SNPs with significant changes in abundance, 11 had a decrease in the abundance, with an average decrease around 7.5-fold. There were three SNPs that had a significant increase in abundance, all of which were synonymous, ranging from a 4- to 47-fold increase. Although a majority of SNPs that changed in abundance had a decrease in the frequency between T1 and T2, on a population level, there was not a large change in diversity between the two time points (T1 $\pi = 0.0080$, T2 $\pi = 0.0079$; Table S2). However, the trend of increased abundance at the three sites may be driven by selection of cryptic epitopes in an alternative reading frame (see *Discussion*).

Significance of rare variants. We observed two extremes in terms of biological relevance in the untreated population among variants detected as at least two independent sequences across the three time points. At one extreme was the detection of nonviable genomes in the form of a coding variant at position 25, which mutates the active site of the protease, and the detection of termination codons at positions 42 and 61 (Table S1). At the other extreme was the detection of the L90M and V82A variants (at time points 1 and 2, respectively) that became the major resistance populations after ritonavir therapy was initiated (see below, Fig. 3); in addition, V82I and V82L were detected at T2. We found two more examples of primary resistance mutations at low abundance, K20R at all three time points and M46I at two time points, but these did not grow out in the presence of ritonavir (Fig. 3 and Table S1). Similarly, fitness compensatory mutations were also detected at low abundance (L10F, M36I, L63P, A71T, and V77I), all below 1%, and only L63P increased (modestly) in abundance after exposure to ritonavir. More generally, of the 28 substitutions most closely associated with protease inhibitor drug resistance (41, 42), we found 10 such variants, half of which were detected at both pretherapy time points (Table S1).

Assessment of Linkage Disequilibrium (LD) Within the HIV-1 *pro* Gene Population. We measured LD for the sequences in the T1 and T2 populations. We identified very few examples of LD at these two time points using the Fisher's exact test with a Bonferroni correction. Of the 103 polymorphic sites in T1, only three pairs were in significant LD. Similarly, in T2 with 118 polymorphic sites, only

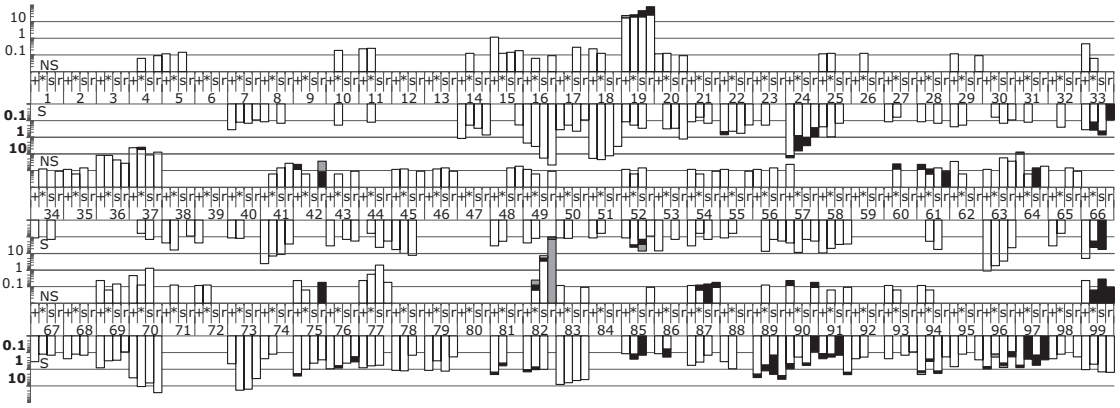

Fig. 2. Frequency of codon variation across all 99 positions in protease over three time points. Within a codon position, the first two bars represent untreated time points 1 and 2, respectively. Bars 3 and 4 are the third time point split based on the presence or absence of the resistance mutations to ritonavir. Bar 3 is the population of susceptible genotypes (defined as not V82A, I84V, or L90M), and bar 4 is the major resistant variant, V82A, population. Upward facing bars are nonsynonymous changes (scale in regular typeface), and downward facing bars are synonymous changes (scale in bolded typeface). Within a codon position, different shading represents different SNPs.

four pairs displayed significant LD. A positive D (i.e., linkage) was found for six of the seven pairs in the untreated populations, with one pair associating at a lower than expected frequency. Overall, LD did not appear to play a significant role in defining the *pro* gene population in this late stage individual, with only a single pair of SNPs showing linkage in both of the time points.

Detection of Multiple Drug-Resistant Alleles After Exposure to Selection by a Protease Inhibitor. The third plasma sample we examined from this subject was from a time point (T3) after the initiation of therapy with the protease inhibitor ritonavir. It is apparent from the cyclical pattern of viral load and self-report that this person had incomplete adherence to the drug regimen (Fig. S1). Thus, we expected selective pressure from the drug to disrupt the viral population but not to select for the more homogeneous populations that are associated with virologic failure solely due to the appearance of drug resistance. The choice of this sample allowed us to look at the evolution of resistance and the persistence of polymorphisms in both the resistant and nonresistant portions of the population. Over two-thirds of the sequences from T3 carried a resistance mutation, with ~50% of the sequences carrying the V82A allele, the most common resistance mutation associated with resistance to ritonavir (43).

There were two divergent paths for population diversity at the third time point. For the large V82A-containing population there was a general trend of decreased diversity ($\pi = 0.0069$), consistent with the expected bottleneck associated with fixing a drug resistance mutation. In contrast, the diversity in the coexisting drug sensitive population was higher than the drug-resistant population and comparable to the earlier time points ($\pi = 0.0082$; Table S2).

Although V82A is the most common resistance mutation associated with ritonavir resistance, the I84V allele and L90M allele can also be selected and in combination with V82A can confer a higher level of resistance (44). We detected all three of these distinct drug resistance alleles in the T3 sequence population, collectively representing 69% of the total T3 population: V82A (50% of the population), I84V (5%), and L90M (14%). These three resistance mutations appeared on different genomes, with only a single example of a sequence with two of these resistance mutations (V82A/L90M). In total, there were 136 unique sequences carrying the V82A mutation (all with the GCC Ala codon), 29 unique sequences carrying the I84V mutation (all with the GTA Val codon), and 36 unique sequences carrying the L90M mutation.

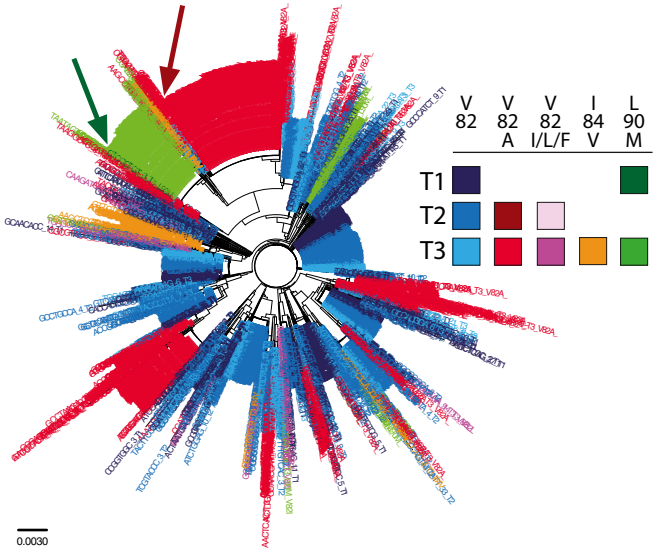

Fig. 3. Phylogenetic representation of protease population derived from deep sequencing with a Primer ID. A Neighbor-Joining tree was constructed from sequences derived from all three time points and colored based on susceptibility to ritonavir. Blue colored taxa represent susceptible variants (defined as not V82A/I/L/F, I84V, or L90M). Red colored taxa represent variants containing the major ritonavir resistant variant, V82A. Pink colored taxa represent the minor resistant variants V82I/L/F. Green and orange colored taxa represent the minor resistant alleles L90M and I84V, respectively. Within a color, color brightness is correlated with sample time. Dark green and red arrows point to pre-RTV low-abundance sequences that clonally amplified to their respective clades.

There were also small groups of *pro* gene sequences in T3 that appear to be the result of selection by ritonavir. Two other substitutions at position 82, V82I and V82L, were detected at a low level at T2 and also seen at T3, but now representing 1.3% and 1.1% of the population. V82F was also detected as 0.14% of the population at T3. Finally, the compensatory mutation L63P was detected at T1 and modestly expanded at T3, with half of the sequences in the V82A background (Table S2).

An important issue is the number of times each of the resistance mutations evolved in the presence of drug selection. The data are consistent with the major V82A variant (42% of the V82 sequences) growing out from the preexisting variant detected at T2. For the six genomic variants of V82A that each accounted for greater than 2.5% of the V82A population, all were on the background of the consensus except for the three different polymorphisms at positions 19 and 70 (Fig. S6*B*). In total, these represented ~71% of the V82A population and presumably arose via recombination with the founding sequence (Fig. S6*A*). The remaining 29% of the V82A-containing genomes vary in relative abundance from 2.3% to 0.1%, including over 100 unique sequences that each appeared once but to a large extent represent the variation seen at T1 and T2 added on to the predominant V82A genotypes.

The composition of the I84V and L90M populations were similar to the V82A population. In each case there was a predominant population defined by a 5′ polymorphism: the major L90M lineage (69% of the L90M sequences) was on the G16G/L19V background (Fig. S6 *C* and *D*), whereas the major I84V lineage (35% of the I84V sequences) was on the consensus sequence background for the 5′ polymorphisms (G16/L19) (Fig. S6 *E* and *F*). The next three most abundant I84V lineages, representing 28% of the I84V sequences, differed from the most abundant sequence by other 5′ polymorphisms (Fig. S6*F*). Similarly, the next three most abundant L90M lineages, representing 14% of the L90M sequences, differed from the most abundant L90M sequence by 5′ polymorphisms (Fig. S6*D*). With the exception of the 5′ polymorphisms and the resistance mutations, all eight of these lineages were in the consensus sequence background. The remaining sequences are accounted for by the low level variability added onto these major lineages.

As noted above, the major V82A lineage was detected at T2 (as a single genome), and this population was likely clonally amplified to form the large proportion of the drug-resistant population seen at T3 (Fig. 3). L90M was also detected on the same *pro* gene background in the therapy-naïve environment at T1, and was likely also clonally amplified to form the large proportion of the L90M sequences (Fig. 3 and Fig. S6*D*). In contrast, V82I and V82L were detected in the pretherapy time points on background sequences that did not become the predominant sequence when these mutation modestly expanded at T3, although these two populations have complex mixtures of the 5′ polymorphisms, which may indicate low level persistence and recombination during the period of drug exposure. Finally, I84V and V82F were not detected in either pretherapy population (Table S1).

Discussion

Complex viral populations can form within a host (45–47). High throughput sequencing technologies allow for extensive sampling of these populations (1–3, 5, 22, 48). However, these technologies are severely limited when a PCR amplification precedes the sequencing protocol, as each sequence read has the potential to be reported as an independent observation without properly controlling for PCR resampling, PCR-mediated recombination, allelic skewing, PCR-introduced misincorporations, and sequencing errors. When working with pathogenic agents in clinical samples,

the number of pathogen genomes in the sample is limited, and the use of PCR can obscure the quality of the sampling by creating a large amount of DNA from a relatively small number of starting templates. This can create artificial homogeneity, inflate estimates of segregating genetic variation, skew the distribution of alleles in the population, and introduce artificial diversity.

We have developed a strategy that allows each sampled template to be tagged with a unique ID by a primer that has a degenerate sequence tag incorporated during the primer oligonucleotide synthesis (Fig. S7). This tag can then be followed through the PCR and the deep sequencing protocol to identify sequencing over-coverage (resampling) of the individual viral templates. Because the Primer ID allows for the identification of over-coverage, this can then be used to create a consensus sequence for each template, avoiding both PCR-related errors and sequencing errors (Fig. S8). In addition, the number of different Primer IDs reflects the number of templates that were actually sampled. This allows a realistic assessment of the depth of population sampling and makes it possible to apply a more rigorous analysis of minor variants by correcting the allelic skewing during the PCR.

We tested the Primer ID approach by sequencing the HIV-1 protease coding domain at three time points in a subject who was intermittently exposed to a protease inhibitor between the second and third time points. A key feature of our approach is the removal of fortuitous errors and accounting for resampling, which results in a dramatic reshaping of the original data set of 72,162 reads. Other approaches that rely on statistical modeling have been developed to deal with the problem of high sequencing error rates associated with deep sequencing technologies (49–51). The use of the Primer ID to create consensus sequences resulted in the removal of 80% of the unique sequence polymorphisms (defined as a change in the consensus without regard to frequency of appearance) in the data set. Similarly, allelic skewing was dramatic among the sampled sequences, in most cases ranging from 2- to 15-fold but going up to nearly 100-fold. Although the Primer ID reveals such skewing and helps correct it, this is clearly a poorly controlled feature of PCR amplifications that can dramatically affect the observed abundance of complex populations, especially the minor variants. Allelic skewing may still persist if the cDNA primer or the upstream PCR primer binds differentially among the templates, or if cDNAs enter the PCR amplification in later rounds and are discarded because they do not result in at least three reads to allow a consensus sequence to be formed. Also, residual misincorporation errors by RT and in the first round of PCR synthesis still limit the interpretation of mutations that occur in the range of 0.01–0.1%. This problem is not overcome with larger numbers of sequences. Given the low diversity in these samples, we removed all substitutions that appeared once because their number approximated the expected number of residual sequence errors, and this resulted in a sensitivity of detection in the range of 0.1% for SNPs that appeared above the frequency of the residual sequence error rate.

Using the Primer ID approach, we were able to describe a number of features of the protease sequence population, however our results are from a single individual and therefore cannot be generalized. First, a pooled analysis of two time points six months apart showed that the variants present at greater than 0.5% in abundance made up two-thirds of the total population but represented only 4% of unique genome sequences and contained only 7% of the total unique sequence polymorphisms. About 60% of the diversity was stable over both time points, with synonymous SNPs maintained at a significantly higher proportion in the two time points than nonsynonymous SNPs. Only 18% of the total diversity represented nonsynonymous SNPs that were present at both time points. However, our ability to assess persistence of these sequences is limited by the depth of sampling, although we feel we are approaching the practical limit of sampling with this technology. We observed nonviable substitutions and estimate that most of the SNPs that appeared once were the result of remaining method error. We found no pattern of conserved linkage among these SNPs, consistent with high levels of recombination across the population.

Although the overall measurement of diversity (π) was similar between the first two time points, we noted that the biggest changes in SNP abundance between the two time points were in three synonymous codon positions (L24L, K70K, and G73G). These dynamic increases made these SNPs part of a larger group of SNPs that accounted for 51% of the total sequences that were otherwise identical to the consensus sequence (Q18Q, L19I, L24L, K70K, G73G, and Q18Q/L19I/L24L'). These SNPs also overlapped the major SNPs that defined subgroups of the resistant variants (L19I; L19V; G16G/L19V). We considered the possibility that there was a unifying feature of these SNPs. We found such a feature in that all of these SNPs, both coding and noncoding, result in changes in two relatively large alternative ORFs that lie at the 5′ and 3′ ends of the *pro* gene. Alternative reading frames have been suggested to generate cryptic CTL epitopes (52–54). In this scenario, these abundant SNPs would represent various escape mutants. Such selective pressures could explain the dynamic behavior of several of these SNPs between the first two time points.

After intermittent exposure to the protease inhibitor ritonavir, we were able to identify six independent lineages of drug resistance mutations. With the intermittent exposure in this particular subject, it was possible to see the major V82A lineage most often seen with ritonavir resistance, but also significant populations of I84V and L90M. We also saw minor populations of V82I, V82L, and V82F. This mixed population of resistant lineages likely represents the early stages of the evolution of resistance, a conclusion supported by the minor appearance of the L63P compensatory mutation and the complete absence of I54V, which is an often seen compensatory mutation for V82A. We saw few examples of genomes with multiple resistance mutations, although these would be expected after more extensive selection (55, 56). We and others have previously examined viral sequences that have been collected in large databases. Typically, these sequences represent the single predominant sequence within an individual, and the use of these sequences allows for assessment of interperson diversity. In the future, it will be an interesting exercise to compare the conclusions reached by examining viral diversity within a person to viral diversity between people; however more intraperson diversity needs to be measured at this level of detail to allow comparison of inter- versus intraperson diversity.

The presence of preexisting drug-resistant variants and their role in therapy failure is of great interest, and accurate, deep sampling of a viral population can add significantly to our understanding of this question. We were able to detect several examples of drug-resistance mutations but only at a very low level. Our ability to reliably detect these mutations is limited to those that appear at a frequency of 0.1–0.2%, limited in part by the low overall diversity in the population. We were able to see examples of mutations that are typically seen only in the presence of drug selection. However, the detection was usually as one genome at two time points or two genomes at one time point. This was also the level of detection of active site mutations in the protease and of termination codons, which must represent either transient viral genomes or residual misincorporation errors. In two cases, we were able to observe the resistance mutation (V82A and L90M) at pretherapy time points linked to the same polymorphisms that were present on the variant that grew out during drug exposure. Thus, although it is likely that we are detecting relevant preexisting drug-resistant variants, these are at the limit of detection and, if they are maintained at a steady-state level, it is well under 0.5% abundance.

Most protocols of high throughput sequencing technologies still require an initial quantity of DNA that necessitates an upfront PCR step for many applications. The use of a Primer ID will help clarify the sequencing products in any strategy that uses an initial PCR step with its attendant error rate, recombination, and resampling. In an independent effort Kinde et al. have described an analogous approach in another deep sequencing

setting (36). We believe a strategy that allows an initial tagging of individual templates before PCR and subsequent sequence analysis will be essential for understanding the true complexity and diversity of genetically dynamic populations.

Materials and Methods

Viral RNA was isolated from blood plasma using the QIAmp Viral RNA kit (Qiagen). cDNA was generated using SuperScript III Reverse Transcriptase (Invitrogen) using the primer (with Primer ID) as described. Following the reaction, RNA in hybrid was removed by RNaseH treatment (Invitrogen). Unincorporated cDNA primer was removed, and the cDNA product amplified by PCR. Sequencing was done using the 454 platform (Roche). Detailed methods for cDNA tagging, amplification, and analysis are presented in the *SI Materials and Methods*.

ACKNOWLEDGMENTS. C.B.J. thanks Jesse Walsh for training on the GS FLX platform. We also thank Dr. Dale Kempf of Abbott Laboratories for making clinical samples available. This work was supported by National Institutes of Health (NIH) Awards GM P01 GM066524 (with subcontract to R.S.) and R37 AI44667 (to R.S.). In addition, we received support from University of North Carolina (UNC) Center For AIDS Research NIH Award P30 AI50410 and UNC Lineberger Comprehensive Cancer Center NIH Award P30 CA16086.

1. Margulies M, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380.
2. Eid J, et al. (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323:133–138.
3. Bentley DR, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53–59.
4. Metzker ML (2010) Sequencing technologies - the next generation. *Nat Rev Genet* 11: 31–46.
5. Fischer W, et al. (2010) Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PLoS ONE* 5:e12303.
6. Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW (2007) Characterization of mutation spectra with ultra-deep pyrosequencing: Application to HIV-1 drug resistance. *Genome Res* 17:1195–1201.
7. Hoffmann C, et al. (2007) DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Res* 35:e91.
8. Bushman FD, et al. (2008) Massively parallel pyrosequencing in HIV research. *AIDS* 22: 1411–1415.
9. Varghese V, et al. (2009) Minority variants associated with transmitted and acquired HIV-1 nonnucleoside reverse transcriptase inhibitor resistance: Implications for the use of second-generation nonnucleoside reverse transcriptase inhibitors. *J Acquir Immune Defic Syndr* 52:309–315.
10. Mitsuya Y, et al. (2008) Minority human immunodeficiency virus type 1 variants in antiretroviral-naive persons with reverse transcriptase codon 215 revertant mutations. *J Virol* 82:10747–10755.
11. Gunthard HF, Wong JK, Ignacio CC, Havlir DV, Richman DD (1998) Comparative performance of high-density oligonucleotide sequencing and dideoxynucleotide sequencing of HIV type 1 pol from clinical samples. *AIDS Res Hum Retroviruses* 14: 869–876.
12. Van Laethem K, et al. (1999) Phenotypic assays and sequencing are less sensitive than point mutation assays for detection of resistance in mixed HIV-1 genotypic populations. *J Acquir Immune Defic Syndr* 22:107–118.
13. Palmer S, et al. (2006) Persistence of nevirapine-resistant HIV-1 in women after single-dose nevirapine therapy for prevention of maternal-to-fetal HIV-1 transmission. *Proc Natl Acad Sci USA* 103:7094–7099.
14. Flys TS, et al. (2006) Quantitative analysis of HIV-1 variants with the K103N resistance mutation after single-dose nevirapine in women with HIV-1 subtypes A, C, and D. *J Acquir Immune Defic Syndr* 42:610–613.
15. Cai F, et al. (2007) Detection of minor drug-resistant populations by parallel allele-specific sequencing. *Nat Methods* 4:123–125.
16. Beck IA, et al. (2008) Optimization of the oligonucleotide ligation assay, a rapid and inexpensive test for detection of HIV-1 drug resistance mutations, for non-North American variants. *J Acquir Immune Defic Syndr* 48:418–427.
17. Johnson JA, et al. (2008) Minority HIV-1 drug resistance mutations are present in antiretroviral treatment-naive populations and associate with reduced treatment efficacy. *PLoS Med* 5:e158.
18. Johnson JA, et al. (2007) Simple PCR assays improve the sensitivity of HIV-1 subtype B drug resistance testing and allow linking of resistance mutations. *PLoS ONE* 2:e638.
19. Metzner KJ, et al. (2003) Emergence of minor populations of human immunodeficiency virus type 1 carrying the M184V and L90M mutations in subjects undergoing structured treatment interruptions. *J Infect Dis* 188:1433–1443.
20. Paredes R, Marconi VC, Campbell TB, Kuritzkes DR (2007) Systematic evaluation of allele-specific real-time PCR for the detection of minor HIV-1 variants with pol and env resistance mutations. *J Virol Methods* 146:136–146.
21. Li JZ, et al. (2011) Low-frequency HIV-1 drug resistance mutations and risk of NNRTI-based antiretroviral treatment failure: A systematic review and pooled analysis. *JAMA* 305:1327–1335.
22. Metzner KJ, et al. (2009) Minority quasispecies of drug-resistant HIV-1 that lead to early therapy failure in treatment-naive and -adherent patients. *Clin Infect Dis* 48: 239–247.
23. Halvas EK, et al. (2006) Blinded, multicenter comparison of methods to detect a drug-resistant mutant of human immunodeficiency virus type 1 at low frequency. *J Clin Microbiol* 44:2612–2614.
24. Mansky LM, Temin HM (1995) Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J Virol* 69:5087–5094.
25. Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD (1996) HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time. *Science* 271:1582–1586.
26. Hughes JP, Totten P (2003) Estimating the accuracy of polymerase chain reaction-based tests using endpoint dilution. *Biometrics* 59:505–511.
27. Kanagawa T (2003) Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J Biosci Bioeng* 96:317–323.
28. Meyerhans A, Vartanian JP, Wain-Hobson S (1990) DNA recombination during PCR. *Nucleic Acids Res* 18:1687–1691.
29. Yang YL, Wang G, Dorman K, Kaplan AH (1996) Long polymerase chain reaction amplification of heterogeneous HIV type 1 templates produces recombination at a relatively high frequency. *AIDS Res Hum Retroviruses* 12:303–306.
30. Liu SL, et al. (1996) HIV quasispecies and resampling. *Science* 273:415–416.
31. Judo MS, Wedel AB, Wilson C (1998) Stimulation and suppression of PCR-mediated recombination. *Nucleic Acids Res* 26:1819–1825.
32. Salazar-Gonzalez JF, et al. (2008) Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. *J Virol* 82:3952–3970.
33. Palmer S, et al. (2005) Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. *J Clin Microbiol* 43:406–413.
34. Simmonds P, Balfe P, Ludlam CA, Bishop JO, Brown AJ (1990) Analysis of sequence diversity in hypervariable regions of the external glycoprotein of human immunodeficiency virus type 1. *J Virol* 64:5840–5850.
35. Edmonson PF, Mullins JI (1992) Efficient amplification of HIV half-genomes from tissue DNA. *Nucleic Acids Res* 20:4933.
36. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B (2011) Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci USA* 108:9530–9535.
37. Cameron DW, et al. (1998) Randomised placebo-controlled trial of ritonavir in advanced HIV-1 disease. The Advanced HIV Disease Ritonavir Study Group. *Lancet* 351: 543–549.
38. Potter J, Zheng W, Lee J (2003) Thermal stability and cDNA synthesis capability of SuperScript III reverse transcriptase. *Focus* (Invitrogen, Carlsbad, CA), pp 19–24.
39. Barnes WM (1992) The fidelity of Taq polymerase catalyzing PCR is improved by an N-terminal deletion. *Gene* 112:29–35.
40. Shriner D, Liu Y, Nickle DC, Mullins JI (2006) Evolution of intrahost HIV-1 genetic diversity during chronic infection. *Evolution* 60:1165–1176.
41. Johnson VA, et al. (2005) Update of the drug resistance mutations in HIV-1: Fall 2005. *Top HIV Med* 13:125–131.
42. Shafer RW, Jung DR, Betts BJ (2000) Human immunodeficiency virus type 1 reverse transcriptase and protease mutation search engine for queries. *Nat Med* 6:1290–1292.
43. Carrillo A, et al. (1998) In vitro selection and characterization of human immunodeficiency virus type 1 variants with increased resistance to ABT-378, a novel protease inhibitor. *J Virol* 72:7532–7541.
44. Eastman PS, et al. (1998) Genotypic changes in human immunodeficiency virus type 1 associated with loss of suppression of plasma viral RNA levels in subjects treated with ritonavir (Norvir) monotherapy. *J Virol* 72:5154–5164.
45. Drake JW, Holland JJ (1999) Mutation rates among RNA viruses. *Proc Natl Acad Sci USA* 96:13910–13913.
46. Duffy S, Shackelton LA, Holmes EC (2008) Rates of evolutionary change in viruses: Patterns and determinants. *Nat Rev Genet* 9:267–276.
47. Onafuwa-Nuga A, Telesnitsky A (2009) The remarkable frequency of human immunodeficiency virus type 1 genetic recombination. *Microbiol Mol Biol Rev* 73:451–480.
48. Shafer RW (2009) Low-abundance drug-resistant HIV-1 variants: Finding significance in an era of abundant diagnostic and therapeutic options. *J Infect Dis* 199:610–612.
49. Eriksson N, et al. (2008) Viral population estimation using pyrosequencing. *PLOS Comput Biol* 4:e1000074.
50. Zagordi O, Klein R, Daumer M, Beerenwinkel N (2010) Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Res* 38:7400–7409.
51. Zagordi O, Geyrhofer L, Roth V, Beerenwinkel N (2010) Deep sequencing of a genetically heterogeneous sample: Local haplotype reconstruction and read error correction. *J Comput Biol* 17:417–428.
52. Cardinaud S, et al. (2004) Identification of cryptic MHC I-restricted epitopes encoded by HIV-1 alternative reading frames. *J Exp Med* 199:1053–1063.
53. Bansal A, et al. (2010) CD8 T cell response and evolutionary pressure to HIV-1 cryptic epitopes derived from antisense transcription. *J Exp Med* 207:51–59.
54. Berger CT, et al. (2010) Viral adaptation to immune selection pressure by HLA class I-restricted CTL responses targeting epitopes in HIV frameshift sequences. *J Exp Med* 207:61–75.
55. Resch W, Parkin N, Watkins T, Harris J, Swanstrom R (2005) Evolution of human immunodeficiency virus type 1 protease genotypes and phenotypes in vivo under selective pressure of the protease inhibitor ritonavir. *J Virol* 79:10638–10649.
56. Hance AJ, et al. (2001) Changes in human immunodeficiency virus type 1 populations after treatment interruption in patients failing antiretroviral therapy. *J Virol* 75:6410–6417.