frontiers in Microbiology

METHODS
published: 12 November 2015
doi: 10.3389/fmicb.2015.01258

Quasispecies Analyses of the HIV-1 Near-full-length Genome With Illumina MiSeq

Hirotaka Ode[1‡], Masakazu Matsuda[1‡], Kazuhiro Matsuoka[1†‡], Atsuko Hachiya[1], Junko Hattori[1†], Yumiko Kito[1], Yoshiyuki Yokomaku[1], Yasumasa Iwatani[1,2] and Wataru Sugiura[1,2†]*

[1] Department of Infectious Diseases and Immunology, Clinical Research Center, National Hospital Organization Nagoya Medical Center, Nagoya, Japan, [2] Department of AIDS Research, Graduate School of Medicine, Nagoya University, Nagoya, Japan

OPEN ACCESS

Edited by:
Francois Villinger,
Emory University School of Medicine,
USA

Reviewed by:
Kazuhisa Yoshimura,
National Institute of Infectious Diseases, Japan
Siddappa Byrareddy,
Emory University, USA

***Correspondence:**
Hirotaka Ode
odehir@mail-nmc.jp

†Present Address:
Kazuhiro Matsuoka,
Proteo-Science Center, Ehime University, Ehime, Japan;
Junko Hattori,
HIV Dynamics and Replication Program, National Cancer Institute, National Institutes of Health, Frederick, USA;
Wataru Sugiura,
GlaxoSmithKline, Tokyo, Japan

‡These authors have contributed equally to this work.

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: 12 September 2015
Accepted: 29 October 2015
Published: 12 November 2015

Citation:
Ode H, Matsuda M, Matsuoka K, Hachiya A, Hattori J, Kito Y, Yokomaku Y, Iwatani Y and Sugiura W (2015) Quasispecies Analyses of the HIV-1 Near-full-length Genome With Illumina MiSeq. Front. Microbiol. 6:1258. doi: 10.3389/fmicb.2015.01258

Human immunodeficiency virus type-1 (HIV-1) exhibits high between-host genetic diversity and within-host heterogeneity, recognized as quasispecies. Because HIV-1 quasispecies fluctuate in terms of multiple factors, such as antiretroviral exposure and host immunity, analyzing the HIV-1 genome is critical for selecting effective antiretroviral therapy and understanding within-host viral coevolution mechanisms. Here, to obtain HIV-1 genome sequence information that includes minority variants, we sought to develop a method for evaluating quasispecies throughout the HIV-1 near-full-length genome using the Illumina MiSeq benchtop deep sequencer. To ensure the reliability of minority mutation detection, we applied an analysis method of sequence read mapping onto a consensus sequence derived from *de novo* assembly followed by iterative mapping and subsequent unique error correction. Deep sequencing analyses of aHIV-1 clone showed that the analysis method reduced erroneous base prevalence below 1% in each sequence position and discarded only <1% of all collected nucleotides, maximizing the usage of the collected genome sequences. Further, we designed primer sets to amplify the HIV-1 near-full-length genome from clinical plasma samples. Deep sequencing of 92 samples in combination with the primer sets and our analysis method provided sufficient coverage to identify >1%-frequency sequences throughout the genome. When we evaluated sequences of *pol* genes from 18 treatment-naïve patients' samples, the deep sequencing results were in agreement with Sanger sequencing and identified numerous additional minority mutations. The results suggest that our deep sequencing method would be suitable for identifying within-host viral population dynamics throughout the genome.

Keywords: HIV-1, deep sequencing, drug resistance, error correction, consensus sequence estimation, quasispecies

INTRODUCTION

Knowledge of the genome sequence of human immunodeficiency virus type-1 (HIV-1) is fundamental for improving the clinical outcome of patients infected with HIV-1 and for understanding viral co-evolution within hosts. However, not only between-host HIV-1 genetic diversity and within-host viral heterogeneous population make it difficult to determine the viral

sequences within host. HIV-1 is classified into four groups (M, N, O, and P), and the group that is most widespread globally, M, is further divided into nine subtypes (A, B, C, D, F, G, H, J, and K), with more than 70 circulating recombinant forms (CRFs), according to the Los Alamos HIV Sequence database (http://www.hiv.lanl.gov/), and numerous unique recombinant forms (URFs; Sharp, 2002; Taylor et al., 2008; Hemelaar et al., 2011; Sharp and Hahn, 2011). Genetic diversity between the subtypes ranges from 25 to 35% (Korber et al., 2001), which is extremely high compared to the human population, in which <1% of distinct DNA sequences are distinct (International HapMap, 2003, 2004). This diversity is considered a consequence of HIV-1's short replication period, lack of proofreading machinery, and recombination in viral replication (Robertson et al., 1995; Perelson et al., 1996; Blackard et al., 2002). The genetic diversity of HIV-1 likely influences the effectiveness of antiretroviral therapy (Wainberg and Brenner, 2012) and at least partially prevents the development of curable strategy against HIV-1 infection (Thomson et al., 2002). Moreover, the error-prone replication induces a within-host genetically diverse heterogeneous viral population, recognized as quasispecies (Ojosnegros et al., 2011). The quasispecies are considered a source of drug-resistant or immune escape variants. Within-host minority viruses likely influence clinical outcome (Johnson et al., 2008; Balduin et al., 2009; Geretti et al., 2009; Metzner et al., 2009; Simen et al., 2009; Paredes et al., 2010), although some reports have found no association between treatment failure and minority variants (Peuchant et al., 2008; Jakobsen et al., 2010; Metzner et al., 2010; Stekler et al., 2011).

To improve clinical outcomes and further understand viral co-evolution within-hosts, the HIV-1 RNA genome has been sequenced using the direct Sanger sequencing method. For example, before treatment against HIV-1 infection, the *pol* and *env* V3 regions are sequenced in genotyping resistance tests and tropism tests that predict viral susceptibility to antiretroviral drugs (Smit, 2014). However, analysis of viral polymorphic sequences is limited using Sanger sequencing method. For example, direct Sanger sequencing is inappropriate for analyzing regions containing heterogeneous insertions or deletions, such as *gag* and *env*. Within-host quasispecies population analyses using direct Sanger sequencing can detect low-frequency mutations in only up to 10–20% of the population. In addition, primer design may be troublesome when analyzing sequences of large or multiple segments.

Recently developed next-generation sequencing technologies that output unprecedented amounts of short sequence reads enable the analysis of viral quasispecies in further depth (Willerth et al., 2010; Dudley et al., 2012; Gall et al., 2012; Henn et al., 2012; Gibson et al., 2014; Park et al., 2014). Bench-top deep sequencers Roche/454 GS Junior and Ion PGM, both based on the pyrosequencing method, are applicable for analyses of limited regions of the HIV-1 genome (Dudley et al., 2012; Gibson et al., 2014; Park et al., 2014). Illumina has released another bench-top deep sequencer, MiSeq, based on bridge sequencing technology, which, compared with the aforementioned pyrosequencing platforms, can output large amounts of sequence reads with a lower intrinsic error rate, especially at homopolymeric regions, including the drug-resistance-related reverse transcriptase (RT) K65 codon (Varghese et al., 2010; Loman et al., 2012; Junemann et al., 2013). Here, we have proposed a practical method to analyze viral quasispecies of the HIV-1 near-full-length genome in clinical samples using the Illumina MiSeq deep sequencing method and especially evaluated nucleotide variations in viral sequences of the Pol and the Env V3 encoding regions.

MATERIALS AND METHODS

Plasmid Sample Preparation

To examine analytical biases that may produce misleading results and intrinsic errors in sequence reads from Illumina MiSeq, pNL4-3 (pNL4-3_{wt}) was used as a reference clone. Furthermore, to examine the threshold of deep sequencing in detecting minority mutations in clinical samples, artificially simulated samples were prepared by mixing multiple different clones. For this purpose, three pNL101-based recombinant infectious clones (Neuveut and Jeang, 1996) possessing drug-resistance mutations RT K103N ($pNL101_{rtK103N}$), RT M184V ($pNL101_{rtM184V}$), and integrase (IN) Q148H ($pNL101_{inQ148H}$) were constructed using standard site-directed mutagenesis protocols as described previously (Hachiya et al., 2008; Shimura et al., 2008). Seven ratios of pNL4-3_{wt}: $pNL101_{rtK103N}$: $pNL101_{rtM184V}$: $pNL101_{inQ148H}$ were mixed as follows: (a) 100:0:0:0, (b) 99.97:0.01:0.01:0.01, (c) 99.7:0.1:0.1:0.1, (d) 98.5:0.5:0.5:0.5, (e) 97:1:1:1, (f) 70:10:10:10, and (g) 40:20:20:20.

Clinical Sample Collection and Sanger Sequencing

Fifty-two plasma samples were collected from 33 HIV-1-infected patients who visited the Nagoya Medical Center in Japan from January 2009 to April 2014 (Supplementary Table S1). Forty-five plasma samples collected from 25 patients enrolled in the Japanese Drug Resistance HIV-1 Surveillance Network (Gatanaga et al., 2007; Hattori et al., 2010; Shiino et al., 2014) were also used in this study. Thus, a total of 97 plasma samples obtained from 58 patients were used. To monitor viral quasispecies chronologically, plasma samples were obtained at 10 time points from one patient failing protease inhibitors (PIs) containing regimens. The total RNA was extracted from 200- or 400-μL of the plasma sample using the MagNA Pure Compact Nucleic Acid Isolation Kit I (Roche Diagnostics K.K., Tokyo, Japan). Extracted RNA was eluted in a final volume of 50 μL of elution buffer and used for subsequent analyses. HIV-1 protease (PR) (297 bps; 2253–2549, positions based on HXB2 numbering), RT (720 bps; 2550–3269) and *env* V3 (108 bps; 7110–7217) sequences of each sample were analyzed using the bulk sequencing method as previously reported (Gatanaga et al., 2007; Hattori et al., 2010; Shiino et al., 2014). Drug-resistance mutations in Pol were determined according to the list reported by Shafer et al. (Bennett et al., 2009) and IAS-USA (Wensing et al., 2014). HIV-1 subtypes were determined using phylogenetic analysis with reference sequences recommended by the Los Alamos Database (http://www.hiv.lanl.gov/). Genotypic tropism

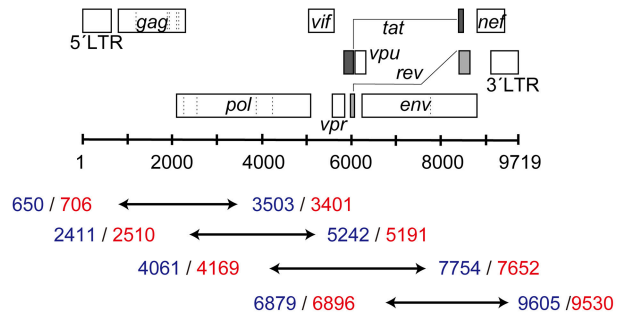

FIGURE 1 | Four overlapping segments covering the HIV-1 near-full-length genome. The blue numbers represent the 5′- and 3′-end positions (HXB2 numbering) for RT-PCR, whereas red numbers show those for nested PCR.

tests were performed using geno2pheno [coreceptor] (http://coreceptor.geno2pheno.org/) with a false-positive rate of 10%.

This study was conducted according to principles in the Declaration of Helsinki. The Ethical Committee at the National Institute of Infectious Diseases and Nagoya Medical Center in Japan approved the study. All patients provided written informed consent for the collection of samples and subsequent analyses.

Amplification of the HIV-1 Near-full-length Genome in Clinical Samples

To analyze the full-length HIV-1 genome (excluding LTR regions) using MiSeq, the *gag* to *nef* (8825 bps; 706–9530) region of the genome was amplified in four overlapping segments, as shown in **Figure 1**. The primer sequences used for the amplifications are listed in Supplementary Table S2. RT-PCR was performed using a PrimeScript II High Fidelity One Step RT-PCR Kit (Takara, Shiga, Japan), followed by nested PCR using PrimeSTAR GXL DNA Polymerase (Takara, Shiga, Japan). Finally, four amplified PCR products were combined into one sample with a MultiScreen HTS PCR96 filter plate according to the manufacturer's instructions (Merck Millipore, Billerica, Massachusetts, USA). The purified DNA was eluted into a final volume of 50 μL of distilled water.

Library Preparation and Deep Sequencing with Illumina MiSeq

Viral DNA libraries for deep sequencing were prepared using the Nextera DNA Sample Prep Kit (Illumina K.K.,Tokyo, Japan) according to the manufacturer's instructions. Unamplified DNA of the recombinant clones and amplified DNA obtained from clinical samples were used for the preparation. The prepared library was sequenced using Illumina MiSeq, generating paired-end 2 × 250-bps-long sequence reads. For each run, a maximum of 24 samples were concurrently examined.

Sequence Read Analysis for Deep Sequencing

The sequence reads obtained from Illumina MiSeq were analyzed following three procedures: (I) mapping of the sequence reads onto a reference sequence or a consensus sequence; (II) error correction; and (III) frequency estimation of bases, amino acids, or short (<250-bps-long) fragment sequences. The detailed methods at each step are described below.

(I) Consensus Sequence Estimation and Read Mapping. Sequence read mapping was conducted with the BWA-MEM algorithm implemented in the BWA-0.6.4 program (Li and Durbin, 2009, 2010) using the default setting. We performed four sequence read mapping methods (Supplementary Figure S1A) as follows:

(i) Simple mapping onto an infectious clone sequence. The sequence reads were mapped on a given infectious clone sequence, such as NL4-3, HXB2, JRCSF (GenBank accession no. M38429) in subtype B, and 93JP-NH1(NH1) in CRF01_AE (AB052995).

(ii) Mapping onto consensus sequence derived from *de novo* assembly. Long fragments (>1000 bps) of the HIV-1 near-full-length genome sequence were assembled *de novo* from sequence reads with the VICUNA program (Yang et al., 2012; Malboeuf et al., 2013). The estimated long fragment sequences were mapped onto the HXB2 sequence with the BWA program. Next, the long fragment sequences were connected to construct a near-full-length consensus sequence. Sequence reads were then mapped onto the constructed consensus sequence.

(iii) Mapping onto consensus sequence estimated from iterative mapping (McElroy et al., 2014; Verbist et al., 2015). The first cycle of iterative mapping was performed on a given reference sequence, such as NL4-3, HXB2, JRCSF, or NH1, with BWA using an option of "-B 1" to reduce mismatch penalty for mapping. The majority base at each position was accepted as the consensus base. In cases of insertions or deletions, the longer sequences were always chosen for consensus sequence estimation; i.e., deletions were ignored and insertions were accepted regardless of their prevalence. The consensus sequence estimated in the first cycle was used as a reference sequence of the second cycle of iterative mapping with the BWA program at the default setting. This procedure was repeated nine times to refine a consensus sequence, and, finally, the sequence reads were mapped onto the consensus sequence obtained in the 9th iterative mapping.

(iv) Combination of (ii) and (iii): Sequence reads were mapped onto the consensus sequence estimated from *de novo* assembly using the VICUNA program (Yang et al., 2012; Malboeuf et al., 2013) followed by the iterative mapping (McElroy et al., 2014; Verbist et al., 2015).

(II) Error Correction. We performed error correction using averaged quality score (QS)-values for each reference sequence position. The detailed method is described below.

(III) Frequency Estimation. We estimated the frequency of each base at a given position by counting the number of nucleotides for each base that remained after error correction. We also calculated the occupancy of sequences

within a <250-bps-long range because the maximum length of Illumina MiSeq sequence reads is 250 bp. First, we extracted short fragment sequences from the mapped sequence reads within the targeted range. Then, identical fragment sequences were grouped into a haplotype sequence. The number of the fragment sequences was counted for each haplotype sequence. Next, haplotype sequences including bases of averaged QSs below 20 were removed. The remaining haplotype sequences were used for frequency estimation.

RESULTS

Use of the Consensus Sequence Estimated from *de novo* Assembly or Iterative Mapping Improved the Sequence Read Mapping Results

To examine analytical bias that may produce misleading results and an intrinsic error rate of output sequence reads from Illumina MiSeq, pNL4-3_{wt} was deeply sequenced in quadruplicate. The obtained sequence reads from each of the quadruplicate runs were individually mapped onto the original pNL4-3_{wt} sequence [Schema (i) in Supplementary Figure S1A]. The mapping results demonstrated >5000-fold coverage throughout the HIV-1 genome (**Figure 2**, Supplementary Figure S1B and Supplementary Table S3). To investigate the effect of selected reference sequences on the mapping accuracy and coverage of the sequence reads, two different subtype B sequences, HXB2 (97.4% identical to NL4-3) and JRCSF (93.8%), and one CRF01_AE sequence, NH1 (85.0%), were selected as reference sequences for mapping [Schema (i) in Supplementary Figure S1A]. As shown in **Figure 2**, the mapping coverage by HXB2 (5722- to 23,058-fold) and JRCSF(3038- to 22,570-fold) were comparable to that of NL4-3 (5424- to 21,616-fold), whereas a considerable reduction in coverage was observed when NH1 was used as the reference sequence (0- to 21,064-fold). Thus, selection of the reference sequence is clearly a critical factor for accurate mapping and high coverage of deep sequencing reads. However, for clinical samples, selection of an appropriate reference sequence is problematic because the sample sequences are unknown at the time of deep sequencing. To overcome this problem, we estimated a consensus sequence from *de novo* assembly, iterative mapping, or *de novo* assembly followed by iterative mapping (Yang et al., 2012; Malboeuf et al., 2013; Gibson et al., 2014; McElroy et al., 2014; Verbist et al., 2015) [Schema (ii)–(iv) in Supplementary Figure S1A]. We used the resulting consensus sequence as the reference sequence for mapping, as previously proposed by others (Yang et al., 2012; Malboeuf et al., 2013; Gibson et al., 2014; McElroy et al., 2014; Verbist et al., 2015). The use of the *de novo* assembled consensus sequence provided comparable mapping results (5731- to 23,053-fold) to that of NL4-3. The assembled consensus sequence was analogous to the original NL4-3 reference sequence but included ∼10 *nef* mutations, suggesting that *de novo* assembly is likely insufficient to estimate the true consensus sequence. By contrast, iterative mapping or *de novo* assembly followed by iterative mapping estimated a consensus sequence that was identical to the NL4-3 sequence and resulted in the same mapping coverage as that achieved using NL4-3 as the reference sequence. These results suggest that iterative mapping and *de novo* assembly followed by iterative mapping can estimate the true consensus sequence and are most appropriate for sequence read mapping. Hence, in the following sections, we mapped sequence reads from deep sequencing using consensus sequence estimation by *de novo* assembly followed by iterative mapping (Yang et al., 2012; Malboeuf et al., 2013; Gibson et al., 2014; McElroy et al., 2014; Verbist et al., 2015).

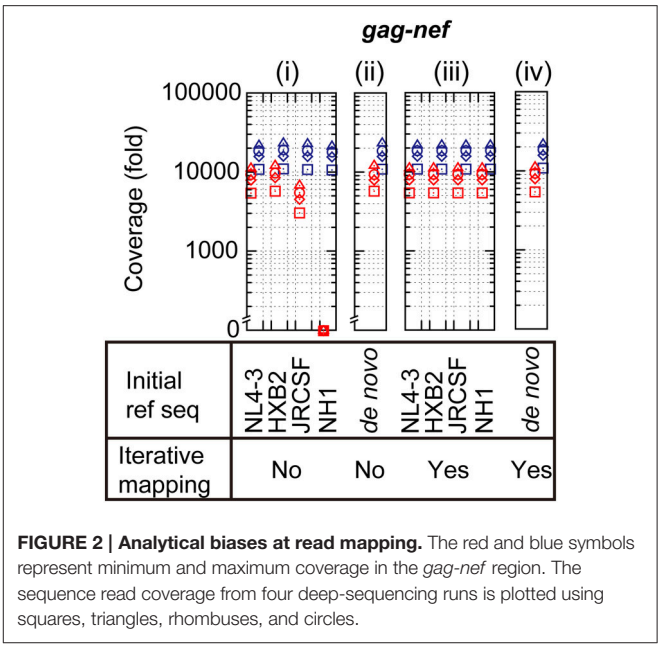

FIGURE 2 | Analytical biases at read mapping. The red and blue symbols represent minimum and maximum coverage in the *gag-nef* region. The sequence read coverage from four deep-sequencing runs is plotted using squares, triangles, rhombuses, and circles.

Our Unique Error-correction Method Reduced the Prevalence of Erroneous Bases Found in Sequence Reads for a Recombinant Clone Below 1% in Each Sequence Position

We also analyzed intrinsic errors in the sequence reads from deep sequencing for pNL4-3_{wt}. As shown in **Figure 3A**, without any error-correction handling, even with clonal pNL4-3_{wt} sequencing results, the erroneous bases occupied a maximum of 6.4% of the population at each reference sequence position. The erroneous bases induced drug-resistance-associated mutations such as IN T66A/I/K and Q148H/K/R at maximums of 1.7 and 1.6% of the population. Considering minority mutation detection by deep sequencing, this error rate is excessive, making minority-mutation determinations inaccurate. In analyzing the patterns of errors, a dominant pattern was substitution (∼99.6% in total errors), whereas insertions or deletions were not frequently observed, as previously reported by others (Loman et al., 2012; Junemann et al., 2013). Further examination of substitution patterns revealed that C>A and T>G transversion errors were most frequently observed in the pNL4-3_{wt} sequence

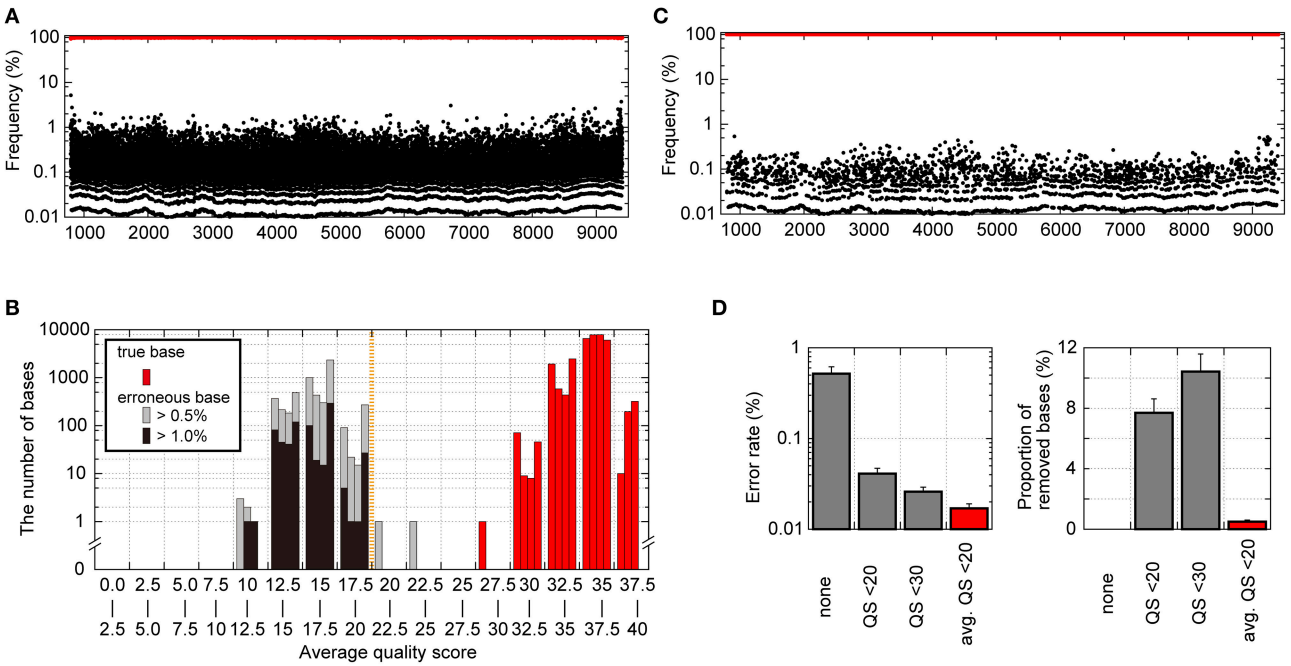

FIGURE 3 | Analytical bias in pNL4-3$_{wt}$ sequence reads from the Illumina MiSeq deep sequencer. (A) Frequencies of nucleotides with true bases (red dots) and with erroneous bases (black dots) throughout the pNL4-3$_{wt}$ genome except for the LTR regions prior to any error correction. **(B)** Histogram of averaged QSs for true bases and erroneous bases at each genome position. The averaged QSs for true bases are shown in red bars, whereas those for erroneous bases occupying >0.5 and >1.0% of the population are represented as gray and black bars, respectively. In each range (x-axis) of the histogram, the four bars representing results from each quadruplicate runs are shown. **(C)** Frequencies of nucleotides with true bases (red dots) and erroneous bases (black dots) throughout the pNL4-3$_{wt}$ genome after error correction removed each base with averaged QSs below 20 at the respective position. **(D)** Error rate after error correction and the proportion of removed nucleotides among all nucleotides using error correction. Results are shown for no error correction, quality-filtering error correction by removing nucleotides with QSs below 20 or 30, and our error correction removing bases of averaged QSs below 20.

reads (Supplementary Figure S2). The predominance of T>G transversion was previously reported for MiSeq (Schirmer et al., 2015) and Illumina Genome Analyzer (Nakamura et al., 2011; Flaherty et al., 2012), both of which are based on the bridge sequencing method, suggesting that the transversion errors might be intrinsic to the apparatuses and the technology.

To improve the accuracy of the sequencing results, we sought to establish a novel error correction method by distinguishing true minority bases from erroneous bases. To differentiate true and erroneous bases, we focused on Phred QS-values of nucleotides in pNL4-3$_{wt}$ sequence reads. QSs of 10, 20, and 30 indicate 90, 99, and 99.9% base-calling accuracy, respectively. Nucleotides with true bases tend to demonstrate high QS-values, whereas low QS-values are associated with erroneous bases (Supplementary Figure S3). When the QSs of nucleotides with true bases were averaged at the respective reference sequence positions, the averaged QS-values for true bases were >25, suggesting >99.7% base-calling accuracies (**Figure 3B**). By contrast, when we focused on bases occupying >1% of the population at the respective positions, the averaged QS-values for erroneous bases were below 20 and were clearly different from those for true bases at a threshold of 20 (**Figure 3B**). Thus, this result indicates that "an averaged QS-value of 20" is a reasonable cut-off threshold to distinguish true and erroneous bases.

Figure 3C shows the results of pNL4-3$_{wt}$ sequencing managed by our novel error-correction method that removed bases with averaged QS-values below 20 at each reference sequence position (Supplementary Figure S3), and the erroneous bases did not occupy >0.54% of the population. Especially, each population of the drug resistant mutations in Pol and erroneous sequences at Env V3 and PR cleavage sites in Gag, Pol, Nef (Shafer and Schapiro, 2008; Fun et al., 2012) was not exceeded 0.2%. The error rates were reduced to 0.017 ± 0.002 from 0.52 ± 0.098% of the raw sequence reads (**Figure 3D**). Of note, our error-correction method removed only <1% of all nucleotides (**Figure 3D**), suggesting selective removal of erroneous bases.

As a comparative method, we applied the simple quality-filtering correction method reported by others (Dudley et al., 2012; Gall et al., 2012; Pessoa et al., 2014) to the same pNL4-3$_{wt}$ sequence reads. The quality-filtering method simply discards any nucleotides with a QS below 20 or 30 as a cut-off value (Supplementary Figure S3). This alternative method was also successful in reducing error rates to 0.041 ± 0.006 and 0.026 ± 0.003% with a cut-off of QS < 20 and QS < 30, respectively (**Figure 3D**). However, the quality-filtering method discarded 7% and 11% of all nucleotides by the QS < 20 and QS < 30 cut-offs, respectively (**Figure 3D**).

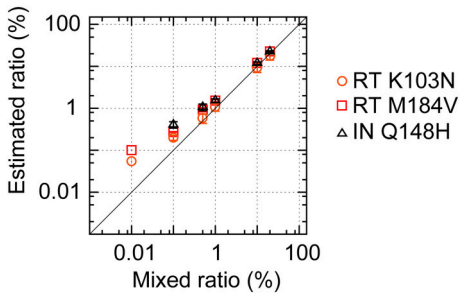

FIGURE 4 | Deep sequencing to detect minority viral populations. pNL4-3_{wt} and pNL101-based mutant clones ($pNL101_{rtK103N}$, $pNL101_{rtM184V}$, and $pNL101_{inQ148H}$) were mixed at seven different ratios: (a) 100:0:0:0, (b) 99.97:0.01:0.01:0.01, (c) 99.7:0.1:0.1:0.1, (d) 98.5:0.5:0.5:0.5, (e) 97:1:1:1, (f) 70:10:10:10, and (g) 40:20:20:20. The estimated frequencies of RT K103N, RT M184V, and IN Q148H are shown in orange circles, red squares, and black triangles, respectively.

Deep-sequencing Analyses Coupled with Our Analysis Method for Distinct Recombinant Clone Mixtures Successfully Detected Minority Mutations with a Prevalence of >1%

To confirm the potential of our mapping and error-correction methods, we performed deep sequencings of mixtures of four recombinant clones, pNL4-3_{wt}, $pNL101_{rtK103N}$, $pNL101_{rtM184V}$, and $pNL101_{inQ148H}$ in triplicate, at seven different ratios: (a) 100:0:0:0, (b) 99.97:0.01:0.01:0.01, (c) 99.7:0.1:0.1:0.1, (d) 98.5:0.5:0.5:0.5, (e) 97:1:1:1, (f) 70:10:10:10, and (g) 40:20:20:20 (**Figure 4**, Supplementary Table S4). We successfully detected three amino acid mutations stably in mixture samples when mutant clones were mixed at ≥0.5% prevalence [(d)–(g)]. For samples (b) and (c), where mutant clones were mixed at 0.01 and 0.1% prevalence, one and two in triplicate tests identified the three mutations, respectively. During the amino acid population analyses, our error-correction removed only ∼3% of all three-nucleotide sequences. By contrast, the simple quality-filtering correction method with a QS < 20 or QS < 30 cut-off, removed ∼25 or 27% of all three-nucleotide sequences, although the quality filtering correction method also allowed us to stably detect three mutations in samples (d)–(g), where mutant clones were mixed at ≥0.5% prevalence (Supplementary Table S4). Taken together with the aforementioned results on error prevalence, our analysis method enables us to detect amino acid mutations at >1% of the population reproducibly and semi-quantitatively while maximizing usage of the sequence read data.

Hence, in the following sections, errors in sequence reads from deep sequencing were corrected with our error-correction method based on averaged QS-values at a threshold of 20. Furthermore, we used error-corrected >1%-frequency bases, amino acids, or short <250-bps-long fragment sequences.

A Near-full-length Genome Amplification Protocol was Successfully Constructed

We designed primer sets to amplify near-full-length viral RNA genomes in four overlapping segments (**Figure 1**). A total of 97 clinical plasma samples were examined. Phylogeny analyses of *pol* sequences derived from Sanger sequencing indicated that 58 plasma samples from 22 patients and 39 samples from 36 patients contained subtype B and non-subtype B viral RNAs. The results showed that all of the four segments were successfully amplified for the subtype B samples with >200 copies/mL (Supplementary Table S5). By contrast, two subtype B samples with 50 and 65.7 copies/mL were incompletely amplified with missing segments. The same primer sets were tested for the remaining 39 non-subtype B viruses, including 10 subtype C, 10 CRF01_AE, 9 subtype F, and 10 CRF02_AG HIV-1. We successfully amplified all four segments for subtype C, CRF01_AE, subtype F, and CRF02_AG viral genomes from samples with up to 432, 700, 2980, and 1600 copies/mL, respectively (Supplementary Table S5). However, we failed to amplify three subtype F samples below 1240 copies/mL, suggesting lower amplification efficacy for non-subtype B viral genomes than subtype B genomes. In particular, amplification of the *in-env v5* regions in non-subtype B genomes was relatively unsuccessful. This is likely due to more nucleotide mismatches between the primer sequences and non-subtype B viral genome sequences (Supplementary Table S6). Further adjustment of these primers, especially for the *in-env v5* region, is required to improve amplification for non-subtype B viral genomes and to effectively amplify HIV-1 genomes regardless of their subtypes. Consequently, all four segments were successfully amplified from 92 plasma samples.

Subsequently, we deeply sequenced 92 amplified samples. When we analyzed the obtained sequence reads, >95% of the sequence reads were mapped onto the estimated consensus sequence for each sample. Further, as shown in Supplementary Figure S4, >1000-fold coverage of sequence reads were obtained at each position throughout *gag-nef*, except for the 5′-end of *matrix* in six samples, *env* in 1 sample, and the 3′-end of *nef* in 1 sample (Supplementary Table S7). Therefore, each minority nucleotide mutation occupying >1% of the population was confirmed from at least 10 sequence reads. Taken together, the results highlight that amplification with our primer sets followed by deep sequencing enabled us to analyze low-frequency mutations with sufficient sequence read coverage.

Our Deep Sequencing Method Detected more Minority Variants than the Direct Sanger Sequencing Method

To evaluate the potential of detecting quasispecies and minority population using our near-full-length deep sequencing method, we compared the results obtained with the proposed method and the direct Sanger sequencing method. We focused on sequences at PR-RT encoding regions (1017 bps; 2253–3269) from 18 treatment-naïve patients, including 13 subtype B, 4 CRF01_AE, and 1 CRF02_AG viruses (Supplementary Table S1). We achieved both deep and Sanger sequencings from the same amplicons. Only one mismatch was found in one sample (TN07) between the consensus sequences derived using the deep and Sanger sequencing methods (**Figure 5A**), except 216 positions in 18 samples where Sanger sequencing detected

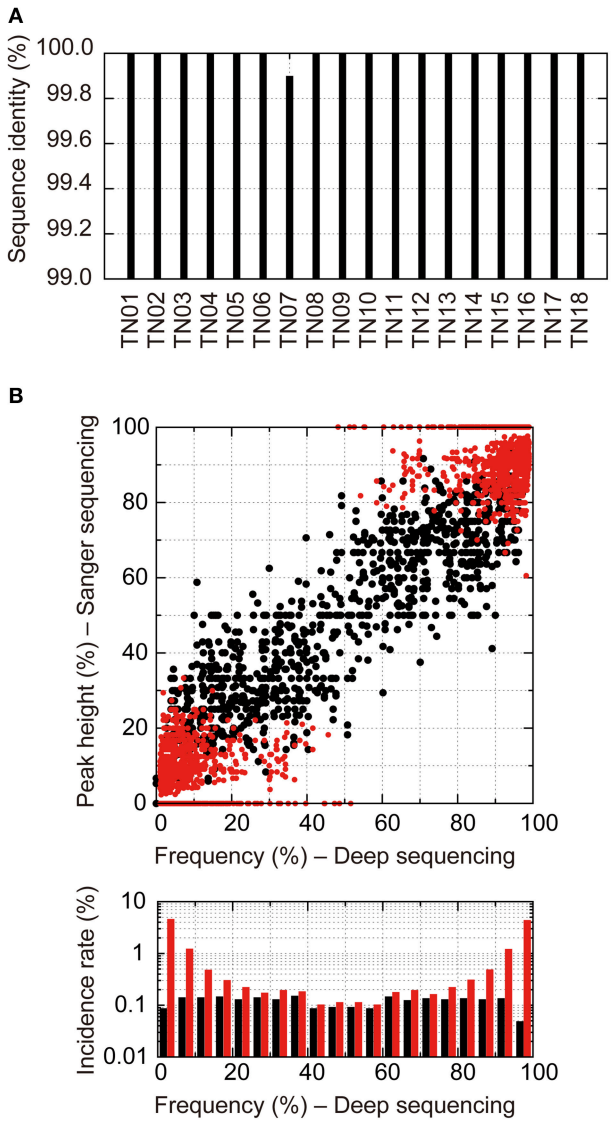

FIGURE 5 | Direct Sanger sequencing and deep sequencing comparison of 18 treatment-naïve samples. (A) Identity of consensus nucleotide sequences at positions where A, T, G, or C was detected by Sanger sequencings between the two methods. **(B)** Correlation of occupancies of bases at the positions where multiple bases at >1% prevalence were detected by deep sequencings between the two methods. For Sanger sequencing, occupancies of bases were estimated from their electropherogram peak heights. Black dots show mixture bases assigned from Sanger sequencing electropherograms. Red dots show mixture bases that were unassigned by Sanger sequencing but were identified by the deep sequencing method. Bottom graphs represent histograms of occupancies detected by the deep sequencing method. The black and red bars show incident rates detected by Sanger sequencings and those detected only by deep sequencing methods, respectively.

mixture bases. Thus, the concordance rate of the two methods was 99.994% [one mismatch in 18,090 (1017 × 18 − 216) positions]. Furthermore, we evaluated the sensitivity of the deep and Sanger sequencing methods in detecting minority populations. For Sanger sequencing, base occupancies were calculated from the electropherogram peak height ratios. As shown in **Figure 5B**, when base occupancies were analyzed at the 216 positions where Sanger sequencing detected mixture bases, a good correlation was observed between the Sanger sequencing method and deep sequencing method (1143 peaks, $R^2 = 0.76$, $P < 0.0001$, single regression analysis; black dots in **Figure 5B**), and all mixture bases detected with Sanger sequencings were identified with the deep sequencing method. However, using the deep sequencing method, we detected an additional 1069 minority bases in all 18 samples that Sanger sequencing failed to recognize (red dots in **Figure 5B**). These results suggest that deep sequencing is more sensitive in detecting minority variants than Sanger sequencing and enables us to analyze quasispecies in clinical samples semi-quantitatively. These additional minority bases contained substitutions conferring drug resistance, and PR M46I (1.1% in TN03, 3.8% in TN04), RT T215S (1.0% in TN01), and RT K219R (1.2% in TN04) were identified. Thus, deep sequencing is likely useful in determining effective treatment regimen in clinical settings.

Our Deep Sequencing Method is Applicable for Detecting Minority X4-tropic Viruses and Examining the Chronological Population Dynamics of Quasispecies

To confirm the clinical advantages of deep sequencing, we applied our deep sequencing method in genotypic tropism testing to determine the co-receptor usage of 18 treatment-naïve patients' samples (Supplementary Figure S5). The deep sequencing method identified heterogeneous V3 sequences in each sample (Supplementary Table S8) and identified nine samples possessing X4-tropic viruses, whereas only 5 samples were identified using Sanger sequencing. As shown in Supplementary Figure S5, all X4 tropic viral sequences in four samples diagnosed using the deep sequencing had less than a 20% minority population (Supplementary Table S8). Thus, the deep sequencing is more sensitive in detecting minority X4-tropic viruses than Sanger sequencings.

To confirm whether our deep sequencing method allows us to dissect quasispecies population dynamics and identify minority drug-resistance mutations relevant to treatment failure, we retrospectively monitored changes in drug-resistance-related mutations in one multi-drug-resistant case (PI-resistant patient #1 in Supplementary Table S1) with M41L/D67N/T69D/M184V/L210W/T215Y in the RT region and M46IL/G73S/I84V/L90M in the PR region (**Figure 6**). We analyzed 10 time points and found that under an EFV-based regimen (time points 3–6), the prevalence of non-nucleoside RT inhibitor (NNRTI)-resistant mutations L100I and K103N increased from <1 to 89.2% and 17.7 to 95.2%, respectively, with viral load relapse from 500 (time point 3) to 5400 copies/mL (time point 5). At time point 6 with 7200 copies/mL, the population of the other NNRTI-resistant mutations, Y181C (33.3%) and G190S (36.4%) increased, whereas the prevalence of L100I and K103N decreased (58.3 and 59.5%). These NNRTI-mutations became undetectable after the regimen was switched

to LPV/r-based therapy (time points 7–10). Emergence of PR I54L followed by I54V was also correlated with relapse under LPV/r-based therapy. Thus, the dynamic population movements of drug-resistance mutations were successfully monitored in detail using our deep sequencing method. In addition to drug-resistance mutations, we also found a population possessing the NC/p1 cleavage-site mutation AP2V was fluctuating with the regimen switches, when we analyzed the sequences at all 11 PR cleavage sites in Gag, Pol, and Nef (Shafer and Schapiro, 2008; Fun et al., 2012). The mutation was first identified as the major population at time point 1, when SQV-based therapy was in progress. Subsequently, the mutation became a minority with EFV-based therapy, and revived as the majority with LPV/r-based therapy.

Although clinical significance was not confirmed in these two analyses of tropism testing and drug resistance mutation monitoring, our results suggest that our deep sequencing method for clinical sample analysis generates more data for understanding within-host viral co-evolutions such as tropism drifting and selection of antiretroviral resistances.

DISCUSSION

In this study, we have proposed a practical method to analyze viral quasispecies of the HIV-1 near-full-length genome in clinical samples using the Illumina MiSeq deep sequencing method (Supplementary Figure S6). Sequence data with low error rates are crucial for accurately analyzing minority populations and genetic diversity of HIV-1. We applied a unique error correction to minimize the effect of artificial errors and facilitate HIV-1 genome analysis using the Illumina bridge sequencing technology. Of note, Illumina bridge sequencing produces $0.52 \pm 0.098\%$ reading errors, which is significantly greater than sequencing platforms using high-fidelity polymerases (Cline et al., 1996; Palmer et al., 2005), but lower than pyrosequencing methods exhibiting high error rates at homopolymeric regions (Varghese et al., 2010; Dudley et al., 2012; Loman et al., 2012; Di Giallonardo et al., 2013; Junemann et al., 2013; Gibson et al., 2014). Our sequencing analysis of the infectious clone pNL4-3_{wt} indicated that the averaged QS-value is a reasonable guide to distinguish true and erroneous bases. One advantage of our error-correction method based on the averaged QS-value is that it removes significantly fewer nucleotides than quality-filtering error correction methods previously reported (Dudley et al., 2012; Gall et al., 2012; Pessoa et al., 2014), which increases the opportunity for detecting minority mutations.

In addition, sequencing analysis of the pNL4-3_{wt} clone suggested that reference sequence choice is critical for accurate and efficient sequence read mapping. To select an appropriate reference sequence in clinical sample analyses, we found that consensus sequence estimated from sequence reads is applicable as the reference sequence for the mapping as reported previously by others (Yang et al., 2012; Malboeuf et al., 2013; McElroy et al., 2014; Verbist et al., 2015), and that *de novo* assembly followed by iterative mapping [Schema (iv) in Supplementary Figure S1A] precisely estimates consensus sequence. During sequence analysis at PR-RT encoding regions of treatment-naïve patients' samples, *de novo* assembly followed by iterative mapping estimated the same consensus sequence as Sanger sequencing, except for one mutation in TN07 (**Figure 5**). By contrast, another estimation method, *de novo* assembly [Schema (ii) in Supplementary Figure S1A] inferred consensus sequences with two mutations for TN11 and another mutation for TN07.

When we performed phylogeny analyses of these estimated near-full-length consensus sequences for each clinical sample, their phylogenetic tree showed concordant subtypes to those based on their *pol* sequences by Sanger method (Supplementary Figure S7), except two samples (Non-subtype B samples #23 and #26 in Supplementary Table S1). Although these two samples were classified into CRF02_AG from the *pol* and of subtype A or G from the near-full-length sequences, this is likely due to analyzed sequence lengths and/or recombination breakpoint positions within CRF02_AG. Furthermore, phylogenetic clusters were found among samples from each drug-resistant patient. Samples from a partner pair (TN12 and TN13) diagnosed in our hospital were also phylogenetically close to each other. Taken together with high sequence identities of the Pol and Env V3 encoding regions between Sanger and our methods, these results suggest that our method may estimate consensus sequences throughout near-full-length regions accurately.

With our error-correction and mapping methods, we can obtain benefit of large genome information from the bridge sequencing. Our method enables both in-depth and semi-quantitative quasispecies analyses (**Figure 5** and Supplementary Figure S5). Although we especially evaluated sequences at the Pol and Env V3 encoding regions in this study, our method would be applicable for quasispecies analyses at the other regions such at PR cleavage sites as shown in **Figure 6**. This is an advantage of our method that is applicable to analyze sequences throughout near-full-length genomes in depth at a run, unlike Sanger or allele specific sequencing methods. Of note, we successfully amplified genomes from low viral load samples using our designed primer sets. Therefore, when patient's viral load increases above the detection limit, our method might be helpful for early detection of drug resistant mutations. Collecting and analyzing genome date using our methods will lead to a comprehensive understanding of unknown mechanisms of resistance acquisition and treatment failure, such as the recent finding demonstrating the importance of the Env cytoplasmic tail mutation in PI resistance (Rabi et al., 2013). Moreover, our method would also be applicable to examine whether drug resistant variants are persistent as proviral DNA, although further assessment is required. Combination of *in vitro* resistance induction experiments or *in vivo* infection of HIV-1 relatives to animal models with our method would help recognize drug resistant machinery or viral evolution. However, there are at least two limitations with our analysis method. The first is due to short sequence reads. Because sequence reads in our study were up to 250-bps long, it was difficult to evaluate interferences of two or more mutations that locate more than 250-bps distant positions. Despite the limitation, our deep sequencing method could help

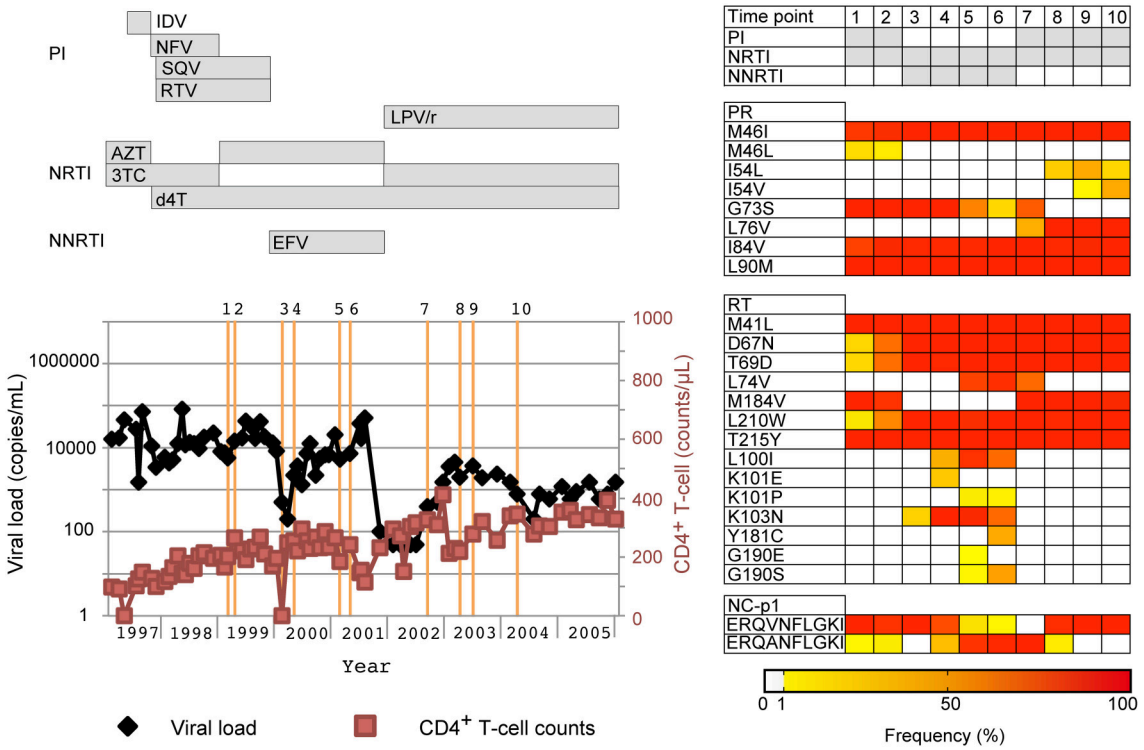

FIGURE 6 | Application of our deep sequencing method in examination of chronological population changes of drug-resistance mutations. Chronological changes in viral loads, CD4$^+$ T-cell counts, and occupancies of drug-resistance mutations for a patient who failed a PI-containing regime. In the left graph, viral loads and CD4$^+$ T-cell counts are shown in black and light red lines, respectively. Collection dates of analyzed samples are highlighted in orange perpendicular lines. Over the orange line, the identification numbers for the collection dates are also shown. At the top of the graph, periods when drugs were used are shown in horizontal bars. Furthermore, population dynamics of resistance related mutations are shown on the right of the graph. The occupancies of the mutations are highlighted with colors according to the bottom bar. PI, PR inhibitor; NRTI, nucleoside/nucleotide RT inhibitor; NNRTI, non-nucleoside RT inhibitor. IDV, indinavir; NFV, nelfinavir; SQV, saquinavir; RTV, ritonavir; LPV/r, lopinavir boosted with ritonavir; AZT, azidothymidine; 3TC, lamivudine; d4T, stavudine; EFV, efavirenz.

obtain hints to know co-evolution within the genome, like mutations in PR and its cleavage sites, in combination with clonal sequencings or haplotype inference by recently proposed some bioinformatics algorithms (Beerenwinkel and Zagordi, 2011; Beerenwinkel et al., 2012; Prosperi et al., 2013; Giallonardo et al., 2014; Schirmer et al., 2014; Jayasundara et al., 2015). The second limitation is attributable to limited stocks of plasma viral RNAs. In several clinical samples, cDNA was amplified from 2.5 μL of extracted viral RNA for each of the four segments, which theoretically contained less than 100 copies of viral RNA, as the original 200 μL plasma had a viral load of <5000 copies/mL. This limitation alerts that the results might be less heterogeneous population than in reality, although we attempted to reduce this risk by triplicate genome amplifications for deep sequencing of each sample. To address this second limitation, in addition to triplicate genome amplification, we must consider increasing several parameters, including the amount of templates, viral RNA, plasma, PCR volume, interestingly retrogressing direction of downsizing sequence technology progress in the last decade.

In conclusion, we devised a data management method and library preparation protocol to analyze quasispecies throughout the HIV-1 near-full-length genome using Illumina MiSeq bentchtop deep sequencing technology. Using deep-sequencing technology with larger genome datasets to precisely analyze minority drug-resistance mutations may improve the efficacy of antiretroviral therapy management in clinical settings.

ACCESSION NUMBER

The data sets analyzed in this study have been deposited in the DNA Data Bank of Japan (DDBJ) under Bioproject accession number PRJDB3502.

AUTHOR CONTRIBUTION

Conceived and designed the experiments: HO, MM, KM, WS. Performed the experiments: HO, MM, KM, AH, JH, YK, YY, YI, WS. Analyzed the data: HO, MM, KM, WS. Contributed reagents/materials/analysis tools: YY, YI, WS. Wrote the paper: HO, MM, KM, WS.

FUNDING

This study was supported by a Grant-in-Aid for AIDS research from the Ministry of Health, Labour, and Welfare of Japan

(H25-AIDS-004) and the Research Program on HIV/AIDS from the Japan Agency for Medical Research and Development, AMED.

ACKNOWLEDGMENTS

We are grateful to all the patients who participated in this study. We thank Ms. Reiko Okazaki, Urara Shigemi and Masumi Hosaka for sample preparation. pNL101 was kindly provided by K.-T. Jeang (National Institutes of Health, Bethesda, MD).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2015.01258

REFERENCES

Balduin, M., Oette, M., Daumer, M. P., Hoffmann, D., Pfister, H. J., and Kaiser, R. (2009). Prevalence of minor variants of HIV strains at reverse transcriptase position 103 in therapy-naive patients and their impact on the virological failure. *J. Clin. Virol.* 45, 34–38. doi: 10.1016/j.jcv.2009.03.002

Beerenwinkel, N., Gunthard, H. F., Roth, V., and Metzner, K. J. (2012). Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Front. Microbiol.* 3:329. doi: 10.3389/fmicb.2012.00329

Beerenwinkel, N., and Zagordi, O. (2011). Ultra-deep sequencing for the analysis of viral populations. *Curr. Opin. Virol.* 1, 413–418. doi: 10.1016/j.coviro.2011.07.008

Bennett, D. E., Camacho, R. J., Otelea, D., Kuritzkes, D. R., Fleury, H., Kiuchi, M., et al. (2009). Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. *PLoS ONE* 4:e4724. doi: 10.1371/journal.pone.0004724

Blackard, J. T., Cohen, D. E., and Mayer, K. H. (2002). Human immunodeficiency virus superinfection and recombination: current state of knowledge and potential clinical consequences. *Clin. Infect. Dis.* 34, 1108–1114. doi: 10.1086/339547

Cline, J., Braman, J. C., and Hogrefe, H. H. (1996). PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Res.* 24, 3546–3551. doi: 10.1093/nar/24.18.3546

Di Giallonardo, F., Zagordi, O., Duport, Y., Leemann, C., Joos, B., Kunzli-Gontarczyk, M., et al. (2013). Next-generation sequencing of HIV-1 RNA genomes: determination of error rates and minimizing artificial recombination. *PLoS ONE* 8:e74249. doi: 10.1371/journal.pone.0074249

Dudley, D. M., Chin, E. N., Bimber, B. N., Sanabani, S. S., Tarosso, L. F., Costa, P. R., et al. (2012). Low-cost ultra-wide genotyping using Roche/454 pyrosequencing for surveillance of HIV drug resistance. *PLoS ONE* 7:e36494. doi: 10.1371/journal.pone.0036494

Flaherty, P., Natsoulis, G., Muralidharan, O., Winters, M., Buenrostro, J., Bell, J., et al. (2012). Ultrasensitive detection of rare mutations using next-generation targeted resequencing. *Nucleic Acids Res.* 40:e2. doi: 10.1093/nar/gkr861

Fun, A., Wensing, A. M., Verheyen, J., and Nijhuis, M. (2012). Human Immunodeficiency Virus Gag and protease: partners in resistance. *Retrovirology* 9:63. doi: 10.1186/1742-4690-9-63

Gall, A., Ferns, B., Morris, C., Watson, S., Cotten, M., Robinson, M., et al. (2012). Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. *J. Clin. Microbiol.* 50, 3838–3844. doi: 10.1128/JCM.01516-12

Gatanaga, H., Ibe, S., Matsuda, M., Yoshida, S., Asagi, T., Kondo, M., et al. (2007). Drug-resistant HIV-1 prevalence in patients newly diagnosed with HIV/AIDS in Japan. *Antiviral Res.* 75, 75–82. doi: 10.1016/j.antiviral.2006.11.012

Geretti, A. M., Fox, Z. V., Booth, C. L., Smith, C. J., Phillips, A. N., Johnson, M., et al. (2009). Low-frequency K103N strengthens the impact of transmitted drug resistance on virologic responses to first-line efavirenz or nevirapine-based highly active antiretroviral therapy. *J. Acquir. Immune Defic. Syndr.* 52, 569–573. doi: 10.1097/QAI.0b013e3181ba11e8

Giallonardo, F. D., Topfer, A., Rey, M., Prabhakaran, S., Duport, Y., Leemann, C., et al. (2014). Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. *Nucleic Acids Res.* 42:e115. doi: 10.1093/nar/gku537

Gibson, R. M., Meyer, A. M., Winner, D., Archer, J., Feyertag, F., Ruiz-Mateos, E., et al. (2014). 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.* 58, 2167–2185. doi: 10.1128/AAC.02710-13

Hachiya, A., Kodama, E. N., Sarafianos, S. G., Schuckmann, M. M., Sakagami, Y., Matsuoka, M., et al. (2008). Amino acid mutation N348I in the connection subdomain of human immunodeficiency virus type 1 reverse transcriptase confers multiclass resistance to nucleoside and nonnucleoside reverse transcriptase inhibitors. *J. Virol.* 82, 3261–3270. doi: 10.1128/JVI.01154-07

Hattori, J., Shiino, T., Gatanaga, H., Yoshida, S., Watanabe, D., Minami, R., et al. (2010). Trends in transmitted drug-resistant HIV-1 and demographic characteristics of newly diagnosed patients: nationwide surveillance from 2003 to 2008 in Japan. *Antiviral Res.* 88, 72–79. doi: 10.1016/j.antiviral.2010.07.008

Hemelaar, J., Gouws, E., Ghys, P. D., Osmanov, S., and WHO-UNAIDS Network for HIV Isolation and Characterisation (2011). Global trends in molecular epidemiology of HIV-1 during 2000-2007. *AIDS* 25, 679–689. doi: 10.1097/QAD.0b013e328342ff93

Henn, M. R., Boutwell, C. L., Charlebois, P., Lennon, N. J., Power, K. A., Macalalad, A. R., et al. (2012). Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog.* 8:e1002529. doi: 10.1371/journal.ppat.1002529

International HapMap, C. (2003). The international hapmap project. *Nature* 426, 789–796. doi: 10.1038/nature02168

International HapMap, C. (2004). Integrating ethics and science in the international hapmap project. *Nat. Rev. Genet.* 5, 467–475. doi: 10.1038/nrg1351

Jakobsen, M. R., Tolstrup, M., Sogaard, O. S., Jorgensen, L. B., Gorry, P. R., Laursen, A., et al. (2010). Transmission of HIV-1 drug-resistant variants: prevalence and effect on treatment outcome. *Clin. Infect. Dis.* 50, 566–573. doi: 10.1086/650001

Jayasundara, D., Saeed, I., Maheswararajah, S., Chang, B. C., Tang, S. L., and Halgamuge, S. K. (2015). ViQuaS: an improved reconstruction pipeline for viral quasispecies spectra generated by next-generation sequencing. *Bioinformatics* 31, 886–896. doi: 10.1093/bioinformatics/btu754

Johnson, J. A., Li, J. F., Wei, X., Lipscomb, J., Irlbeck, D., Craig, C., 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. doi: 10.1371/journal.pmed.0050158

Junemann, S., Sedlazeck, F. J., Prior, K., Albersmeier, A., John, U., Kalinowski, J., et al. (2013). Updating benchtop sequencing performance comparison. *Nat. Biotechnol.* 31, 294–296. doi: 10.1038/nbt.2522

Korber, B., Gaschen, B., Yusim, K., Thakallapally, R., Kesmir, C., and Detours, V. (2001). Evolutionary and immunological implications of contemporary HIV-1 variation. *Br. Med. Bull.* 58, 19–42. doi: 10.1093/bmb/58.1.19

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26, 589–595. doi: 10.1093/bioinformatics/btp698

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., et al. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* 30, 434–439. doi: 10.1038/nbt.2198

Malboeuf, C. M., Yang, X., Charlebois, P., Qu, J., Berlin, A. M., Casali, M., et al. (2013). Complete viral RNA genome sequencing of ultra-low copy

samples by sequence-independent amplification. *Nucleic Acids Res.* 41:e13. doi: 10.1093/nar/gks794

McElroy, K., Thomas, T., and Luciani, F. (2014). Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. *Microb. Inform. Exp.* 4:1. doi: 10.1186/2042-5783-4-1

Metzner, K. J., Giulieri, S. G., Knoepfel, S. A., Rauch, P., Burgisser, P., Yerly, S., 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. doi: 10.1086/595703

Metzner, K. J., Rauch, P., von Wyl, V., Leemann, C., Grube, C., Kuster, H., et al. (2010). Efficient suppression of minority drug-resistant HIV type 1 (HIV-1) variants present at primary HIV-1 infection by ritonavir-boosted protease inhibitor-containing antiretroviral therapy. *J. Infect. Dis.* 201, 1063–1071. doi: 10.1086/651136

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., et al. (2011). Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* 39:e90. doi: 10.1093/nar/gkr344

Neuveut, C., and Jeang, K. T. (1996). Recombinant human immunodeficiency virus type 1 genomes with tat unconstrained by overlapping reading frames reveal residues in Tat important for replication in tissue culture. *J. Virol.* 70, 5572–5581.

Ojosnegros, S., Perales, C., Mas, A., and Domingo, E. (2011). Quasispecies as a matter of fact: viruses and beyond. *Virus Res.* 162, 203–215. doi: 10.1016/j.virusres.2011.09.018

Palmer, S., Kearney, M., Maldarelli, F., Halvas, E. K., Bixby, C. J., Bazmi, H., 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. doi: 10.1128/JCM.43.1.406-413.2005

Paredes, R., Lalama, C. M., Ribaudo, H. J., Schackman, B. R., Shikuma, C., Giguel, F., et al. (2010). Pre-existing minority drug-resistant HIV-1 variants, adherence, and risk of antiretroviral treatment failure. *J. Infect. Dis.* 201, 662–671. doi: 10.1086/650543

Park, S. Y., Goeken, N., Lee, H. J., Bolan, R., Dube, M. P., and Lee, H. Y. (2014). Developing high-throughput HIV incidence assay with pyrosequencing platform. *J. Virol.* 88, 2977–2990. doi: 10.1128/JVI.03128-13

Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M., and Ho, D. D. (1996). HIV-1 dynamics *in vivo*: virion clearance rate, infected cell life-span, and viral generation time. *Science* 271, 1582–1586. doi: 10.1126/science.271.5255.1582

Pessoa, R., Watanabe, J. T., Calabria, P., Felix, A. C., Loureiro, P., Sabino, E. C., et al. (2014). Deep sequencing of HIV-1 near full-length proviral genomes identifies high rates of BF1 recombinants including two novel circulating recombinant forms (CRF) 70_BF1 and a disseminating 71_BF1 among blood donors in Pernambuco, Brazil. *PLoS ONE* 9:e112674. doi: 10.1371/journal.pone.0112674

Peuchant, O., Thiebaut, R., Capdepont, S., Lavignolle-Aurillac, V., Neau, D., Morlat, P., et al. (2008). Transmission of HIV-1 minority-resistant variants and response to first-line antiretroviral therapy. *AIDS* 22, 1417–1423. doi: 10.1097/QAD.0b013e3283034953

Prosperi, M. C., Yin, L., Nolan, D. J., Lowe, A. D., Goodenow, M. M., and Salemi, M. (2013). Empirical validation of viral quasispecies assembly algorithms: state-of-the-art and challenges. *Sci. Rep.* 3:2837. doi: 10.1038/srep02837

Rabi, S. A., Laird, G. M., Durand, C. M., Laskey, S., Shan, L., Bailey, J. R., et al. (2013). Multi-step inhibition explains HIV-1 protease inhibitor pharmacodynamics and resistance. *J. Clin. Invest.* 123, 3848–3860. doi: 10.1172/JCI67399

Robertson, D. L., Sharp, P. M., McCutchan, F. E., and Hahn, B. H. (1995). Recombination in HIV-1. *Nature* 374, 124–126. doi: 10.1038/374124b0

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., and Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* 43:e37. doi: 10.1093/nar/gku1341

Schirmer, M., Sloan, W. T., and Quince, C. (2014). Benchmarking of viral haplotype reconstruction programmes: an overview of the capacities and limitations of currently available programmes. *Brief. Bioinform.* 15, 431–442. doi: 10.1093/bib/bbs081

Shafer, R. W., and Schapiro, J. M. (2008). HIV-1 drug resistance mutations: an updated framework for the second decade of HAART. *AIDS Rev.* 10, 67–84.

Sharp, P. M. (2002). Origins of human virus diversity. *Cell* 108, 305–312. doi: 10.1016/S0092-8674(02)00639-6

Sharp, P. M., and Hahn, B. H. (2011). Origins of HIV and the AIDS pandemic. *Cold Spring Harb. Perspect. Med.* 1:a006841. doi: 10.1101/cshperspect.a006841

Shiino, T., Hattori, J., Yokomaku, Y., Iwatani, Y., Sugiura, W., and Japanese Drug Resistance HIV-1 Surveillance Network. (2014). Phylodynamic analysis reveals CRF01_AE dissemination between Japan and neighboring Asian countries and the role of intravenous drug use in transmission. *PLoS ONE* 9:e102633. doi: 10.1371/journal.pone.0102633

Shimura, K., Kodama, E., Sakagami, Y., Matsuzaki, Y., Watanabe, W., Yamataka, K., et al. (2008). Broad antiretroviral activity and resistance profile of the novel human immunodeficiency virus integrase inhibitor elvitegravir (JTK-303/GS-9137). *J. Virol.* 82, 764–774. doi: 10.1128/JVI.01534-07

Simen, B. B., Simons, J. F., Hullsiek, K. H., Novak, R. M., Macarthur, R. D., Baxter, J. D., et al. (2009). Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naive patients significantly impact treatment outcomes. *J. Infect. Dis.* 199, 693–701. doi: 10.1086/596736

Smit, E. (2014). Antiviral resistance testing. *Curr. Opin. Infect. Dis.* 27, 566–572. doi: 10.1097/QCO.0000000000000108

Stekler, J. D., Ellis, G. M., Carlsson, J., Eilers, B., Holte, S., Maenza, J., et al. (2011). Prevalence and impact of minority variant drug resistance mutations in primary HIV-1 infection. *PLoS ONE* 6:e28952. doi: 10.1371/journal.pone.0028952

Taylor, B. S., Sobieszczyk, M. E., McCutchan, F. E., and Hammer, S. M. (2008). The challenge of HIV-1 subtype diversity. *N. Engl. J. Med.* 358, 1590–1602. doi: 10.1056/NEJMra0706737

Thomson, M. M., Perez-Alvarez, L., and Najera, R. (2002). Molecular epidemiology of HIV-1 genetic forms and its significance for vaccine development and therapy. *Lancet Infect. Dis.* 2, 461–471. doi: 10.1016/S1473-3099(02)00343-2

Varghese, V., Wang, E., Babrzadeh, F., Bachmann, M. H., Shahriar, R., Liu, T., et al. (2010). Nucleic acid template and the risk of a PCR-Induced HIV-1 drug resistance mutation. *PLoS ONE* 5:e10992. doi: 10.1371/journal.pone.0010992

Verbist, B. M., Thys, K., Reumers, J., Wetzels, Y., Van der Borght, K., Talloen, W., et al. (2015). VirVarSeq: a low-frequency virus variant detection pipeline for Illumina sequencing using adaptive base-calling accuracy filtering. *Bioinformatics* 31, 94–101. doi: 10.1093/bioinformatics/btu587

Wainberg, M. A., and Brenner, B. G. (2012). The impact of HIV genetic polymorphisms and subtype differences on the occurrence of resistance to antiretroviral drugs. *Mol. Biol. Int.* 2012:256982. doi: 10.1155/2012/256982

Wensing, A. M., Calvez, V., Gunthard, H. F., Johnson, V. A., Paredes, R., Pillay, D., et al. (2014). 2014 update of the drug resistance mutations in HIV-1. *Top. Antivir. Med.* 22, 642–650.

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

Yang, X., Charlebois, P., Gnerre, S., Coole, M. G., Lennon, N. J., Levin, J. Z., et al. (2012). *De novo* assembly of highly diverse viral populations. *BMC Genomics* 13:475. doi: 10.1186/1471-2164-13-475

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.