HIV-1 Evolutionary Dynamics under Nonsuppressive Antiretroviral Therapy

Steven A. Kemp,[a] Oscar J. Charles,[b] Anne Derache,[c] Werner Smidt,[c] Darren P. Martin,[d] Collins Iwuji,[c,e] John Adamson,[c] Katya Govender,[c] Tulio de Oliveira,[c,f] Francois Dabis,[g,h] on behalf of the ANRS 12249 TasP Study Group, Deenan Pillay,[b] Richard A. Goldstein,[b] Ravindra K. Gupta[a,c]

[a]Cambridge Institute of Therapeutic Immunology & Infectious Disease (CITIID), University of Cambridge, Cambridge, United Kingdom
[b]Division of Infection & Immunity, University College London, London, United Kingdom
[c]Africa Health Research Institute, Durban, South Africa
[d]Department of Integrative Biomedical Sciences, University of Cape Town, Cape Town, South Africa
[e]Research Department of Infection and Population Health, University College London, United Kingdom
[f]KRISP - KwaZulu-Natal Research and Innovation Sequencing Platform, UKZN, Durban, South Africa
[g]INSERM U1219-Centre Inserm Bordeaux Population Health, Université de Bordeaux, France
[h]Université de Bordeaux, ISPED, Centre INSERM U1219-Bordeaux Population Health, France

Steven A. Kemp and Oscar J. Charles contributed equally. Author order was determined in order of increasing seniority.

ABSTRACT Prolonged virologic failure on 2nd-line protease inhibitor (PI)-based antiretroviral therapy (ART) without emergence of major protease mutations is well recognized and provides an opportunity to study within-host evolution in long-term viremic individuals. Using next-generation sequencing and *in silico* haplotype reconstruction, we analyzed whole-genome sequences from longitudinal plasma samples of eight chronically infected HIV-1-positive individuals failing 2nd-line regimens from the French National Agency for AIDS and Viral Hepatitis Research (ANRS) 12249 Treatment as Prevention (TasP) trial. On nonsuppressive ART, there were large fluctuations in synonymous and nonsynonymous variant frequencies despite stable viremia. Reconstructed haplotypes provided evidence for selective sweeps during periods of partial adherence, and viral haplotype competition, during periods of low drug exposure. Drug resistance mutations in reverse transcriptase (RT) were used as markers of viral haplotypes in the reservoir, and their distribution over time indicated recombination. We independently observed linkage disequilibrium decay, indicative of recombination. These data highlight dramatic changes in virus population structure that occur during stable viremia under nonsuppressive ART.

IMPORTANCE HIV-1 infections are most commonly initiated with a single founder virus and are characterized by extensive inter- and intraparticipant genetic diversity. However, existing literature on HIV-1 intrahost population dynamics is largely limited to untreated infections, predominantly in subtype B-infected individuals. The manuscript characterizes viral population dynamics in long-term viremic treatment-experienced individuals, which has not been previously characterized. These data are particularly relevant for understanding HIV dynamics but can also be applied to other RNA viruses. With this unique data set we propose that the virus is highly unstable, and we have found compelling evidence of HIV-1 within-host viral diversification, recombination, and haplotype competition during nonsuppressive ART.

KEYWORDS antiretroviral resistance, clinical failure, drug resistance evolution, human immunodeficiency virus

Even though HIV-1 infections are most commonly initiated with a single founder virus (1), acute and chronic disease are characterized by extensive inter- and intrapatient genetic diversity (2, 3). The rate and degree of diversification is influenced by multiple factors, including selection pressures imposed by the adaptive immune system, exposure of the virus to drugs, and tropism/fitness constraints relating to replication and cell-to-cell transmission in different tissue compartments (4, 5). During HIV-1 infection, high rates of reverse transcriptase (RT)-related mutation and high viral turnover during replication result in swarms of genetically diverse variants (6) which coexist as quasispecies (7, 8). The existing literature on HIV-1 intrahost population dynamics is largely limited to untreated infection, in subtype B-infected individuals (9–12). These works have shown nonlinear diversification of virus both toward and away from the founder strain during chronic untreated infection.

Viral population dynamics in long-term viremic antiretroviral therapy (ART)-treated individuals have not been characterized. HIV-1 rapidly accumulates drug resistance-associated mutations (DRMs), particularly during nonsuppressive 1st-line ART (5, 13). As a result, ART-experienced patients failing 1st-line regimens for prolonged periods of time are characterized by high frequencies of common nucleoside reverse transcriptase (NRTI) and nonnucleoside reverse transcriptase (NNRTI) DRMs such as M184V, K65R, and K103N (14). Routinely, 2nd-line ART regimens consist of two NRTIs in conjunction with a boosted protease inhibitor (PI). Although PI DRMs are uncommonly reported (15), a situation that differs for less potent drugs used in the early PI era (5), multiple studies have indicated that diverse mutations accumulating in the *gag* gene during PI failure might impact PI susceptibility (16–22). Common pathways for these diverse mutations have, however, been difficult to discern, likely reflecting multiple routes to drug escape.

Prolonged virological failure on PI-based regimens without the emergence of PI DRMs provides an opportunity to study evolution under partially suppressive ART. The process of selective sweeps in the context of HIV-1 infection has previously been described (23, 24). Although major PI DRMs and other nonsynonymous mutations in regulatory regions such as *pol* can significantly lower fitness (2, 25, 26), these studies typically are oblivious to temporal sequencing.

We have deployed next-generation sequencing of stored blood plasma specimens from patients in the Treatment as Prevention (TasP) French National Agency for AIDS and Viral Hepatitis Research (ANRS) 12249 study (27), conducted in Kwazulu-Natal, South Africa. All patients were infected with HIV-1 subtype C and characterized as failing 2nd-line regimens containing lopinavir and ritonavir (LPV/r), with prolonged virological failure in the absence of major known PI mutations (28). In the manuscript, we report the details of evolutionary dynamics during nonsuppressive 2nd-line ART. By sampling patients consistently over 2 or more years, we propose that ongoing evolution is driven by the dynamic flux between genetic drift, fitness-driven selection, and recombination, exemplified by resistance mutations that have undergone reassortment across haplotypes through recombination.

RESULTS

Patient characteristics. Eight south African patients with virological failure of 2nd-line PI-based ART, with between three and eight time points and viremia of >1,000 copies/mL, were selected from the French ANRS TasP trial for viral dynamic analysis. Collected patient metadata included viral loads, regimens, and time since ART initiation (Table 1). HIV RNA was isolated from venous blood samples and subjected to whole-genome sequencing (WGS) using Illumina technology; from this, whole-genome haplotypes were reconstructed using sites with a depth of ≥100 reads (Fig. S1). Prior to participation in the TasP trial, patients accessed 1st-line regimens for an average of 5.6 years (±2.7 years). At baseline enrollment into TasP (while failing 1st-line regimens), the median patient viral load was 4.96×10^{10} copies/mL (interquartile range [IQR], 4.17×10^{10} to 5.15×10^{10}); 12 DRMs were found at a threshold of >2%, the most

TABLE 1 Regimens and viral load at the final time point for all patients[a,b]

Patient	No. of time points	1st-line regimen	Time since initiation of 1st-line treatment (yrs)	2nd-line regimen	Viral load at final time point (copies/mL)
15664	6	d4T, 3TC, FTC	6.2	LPV/r, TDF, FTC	28,655
16207	5	d4T, 3TC, NVP	5.9	LPV/r, TDF, FTC	56,660
22763	8	d4T, 3TC, EFV	6.2	LPV/r, TDF, 3TC	15,017
22828	6	d4T, 3TC, NVP	6.4	LPV/r, TDF, 3TC/FTC	947
26892	7	d4T, 3TC, EFV	6	LPV/r, TDF, FTC	12,221
28545	5	TDF, FTC, EFV	1.3	LPV/r, AZT, 3TC	12,964
29447	4	TDF, FTC, EFV	2.8	LPV/r, TDF, FTC	64,362
47939	3	d4T, 3TC, EFV	10.1	LPV/r, AZT, 3TC/FTC	6,328

[a]Patients initiated and maintained 1st-line regimens for between 1 and 10 years before being switched to 2nd-line regimens as part of the TasP trial. Eight of the nine patients were failing 2nd-line regimens at the final time point.

[b]NRTI: d4T, stavudine; 3TC, lamivudine; TDF, tenofovir; FTC, emtricitabine; AZT, zidovudine. NNRTI: EFV, efavirenz; NVP, nevirapine. PI: LPV/r, lopinavir/ritonavir.

common of which were the RT mutations, K103N, M184V, and P225H, which are consistent with previous use of stavudine (d4T), nevirapine (NVP), efavirenz (EFV), and emtricitabine/lamivudine (FTC/3TC). Six of the eight patients had minority frequency DRMs associated with PI failure (average, 6.4%) which were usually seen only in one sample per patient throughout the longitudinal sampling. The observed mutations included L23I, I47V, M46I/L, G73S, V82A, N83D, and I85V (Tables S1a to 3c). The viral populations of four of the eight patients also carried major integrase strand inhibitor (INSTI) mutations, also at minority frequencies (average, 5.0%) and also usually at single time points (T97A, E138K, Y143H, Q148K). Of note, patients were maintained on protease inhibitors during viremia, as poor adherence was suspected as the reason for ongoing failure. Sanger sequencing of all subtype C viruses was undertaken during routine clinical monitoring and was consistent with next-generation sequencing (NGS) data (Tables S1a to 3c) regarding the absence of PI DRMS.

SNP frequencies and measures of diversity/divergence over time. WGS data were used to measure the changing frequencies of viral single nucleotide polymorphisms (SNPs) relative to a dual-tropic subtype C reference sequence (GenBank accession number AF411967) within individuals over time (Fig. 1a and b). The number of longitudinal synonymous SNPs mirrored the number of nonsynonymous SNPs, but the former were 2- to 3-fold more common. Diversification was considered by counting the number of SNPs relative to the reference sequence. There were dynamic changes in the numbers of SNPs over time, with both increases and decreases in the numbers of SNPs, suggesting population competition, and/or the occurrence of selective sweeps. From time point 2 onward (all patients now on 2nd-line, PI-containing regimens for >6 months), all patients (except 28545) had increases in both synonymous and nonsynonymous SNPs.

In previous literature, viral populations within untreated, chronically infected HIV-1 patients have been shown to revert toward the founder or infecting virus states (9). We repeated this analysis with our chronically infected, but treated, HIV-1 population, considering separately the earliest consensus sequence, HIV-1 subtype C consensus, and M group consensus sequences as founder strains. Divergence from the founder strain per patient time point was measured by calculating the genetic distance between patient and founder for each longitudinal sample.

To assess if (i) there was a general trend of reversion to founder and (ii) time was an explanatory variable to that trend, we utilized a linear mixed-effects model (LMEM). Divergence from the founder was modeled as the response, each patient was treated as a random effect, and the time from first patient sample was treated as a fixed effect, "time." Modeling the whole-genome sequences indicated that there was no significant effect of time (in months) on viral diversification or reversion to the infecting/baseline strain or on ancestral C state (Fig. 1c and d, Table S4).

When assessing the constituent 1,000-bp genomic regions of each alignment, four genomic regions were significant for divergence from the ancestral C state, indicating that time in months impacted viral divergence. This revealed that in portions of the
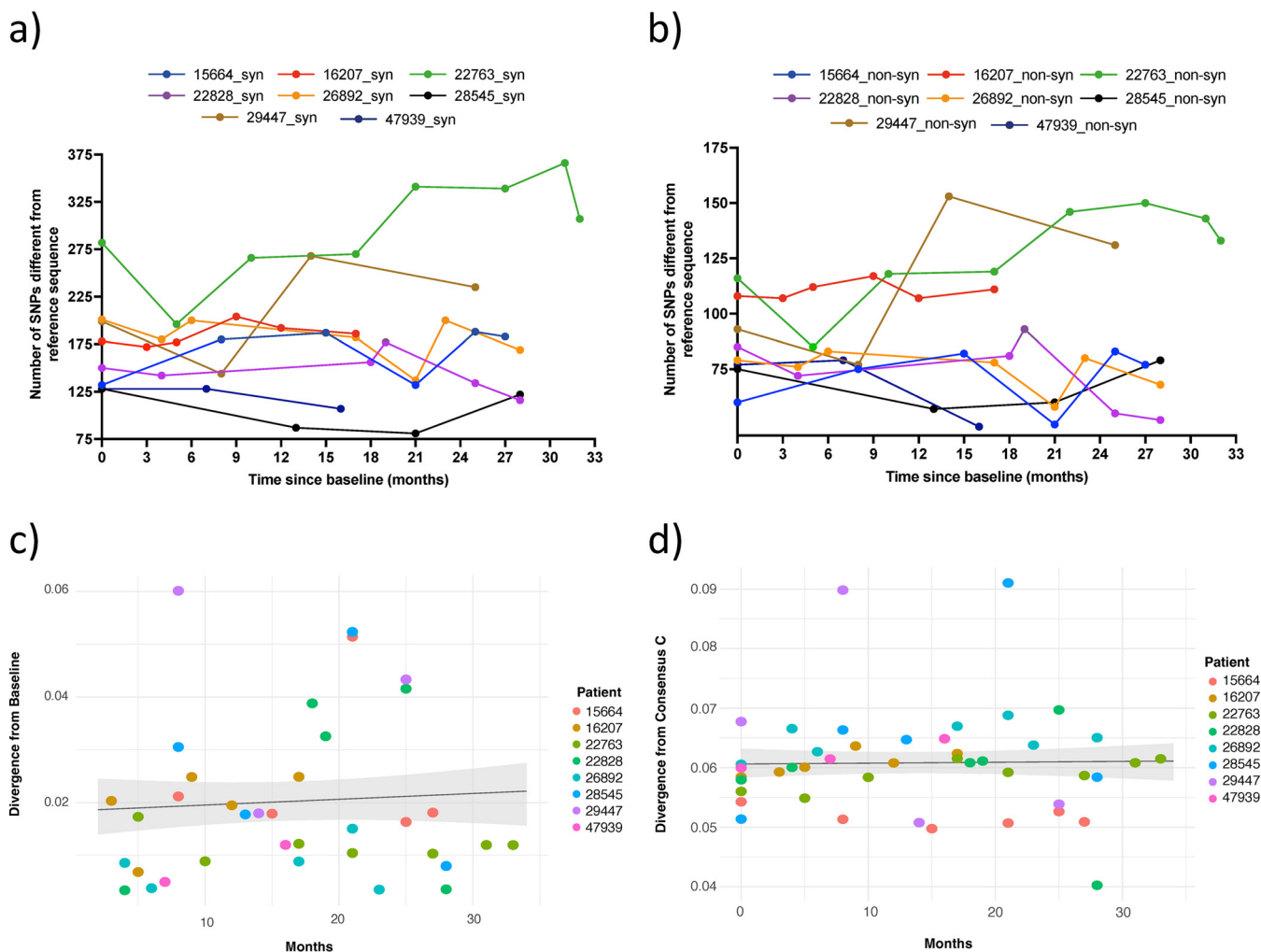
FIG 1 Sequence divergence for eight patients under nonsuppressive ART. These data were for SNPs detected by Illumina NGS at <2% abundance. Sites had coverage of at least 10 reads. (a and b) In both (a) synonymous and (b) nonsynonymous mutations, there was an idiosyncratic change in the number of SNPs relative to the reference strain over time. (c) Mixed-effects linear model of divergence from the baseline time point and (d) consensus C subtype. The trend and 95% confidence interval (CI) (gray shadows) suggest that there is no strong positive or negative linear relationship between divergence and time.

genome (*pol*, *vpu*, and *env*) there was sufficient statistical support to confirm that there was ongoing divergence from the subtype C consensus. However, correction for false-discovery rate (FDR) with a Benjamini-Hochberg correction revealed that this divergence was not significant. Furthermore, analysis of amino acids on a site-by-site basis showed that AA mutations almost always resulted in a divergence, rather than a reversion toward baseline/ancestral sequences. Divergence from these ancestral sequences is likely enabled by recombination, which unlinks hyper-variable loci from strongly constrained neighboring sites. Collectively, we found no statistically significant evidence for reversions and were therefore unable to conclude that these patients are reverting to founder as described in previous literature (9, 29).

To assess the relationship of the observed divergence patterns, we examined nucleotide diversity by considering all pairwise nucleotide distances of each consensus sequence, by time point and patient utilizing multidimensional scaling (30). Intrapatient nucleotide diversity varied considerably between patients (Fig. 2a). Viruses from some patients showed little diversity between time points (e.g., patient 16207), whereas those from others showed higher diversity between time points (e.g., patient 22763). In some instances, a patient's viruses were tightly clustered, suggesting little change over time (Fig. 3a, patients 16207, 26892, and 47939) compared to others (patients 22828 and
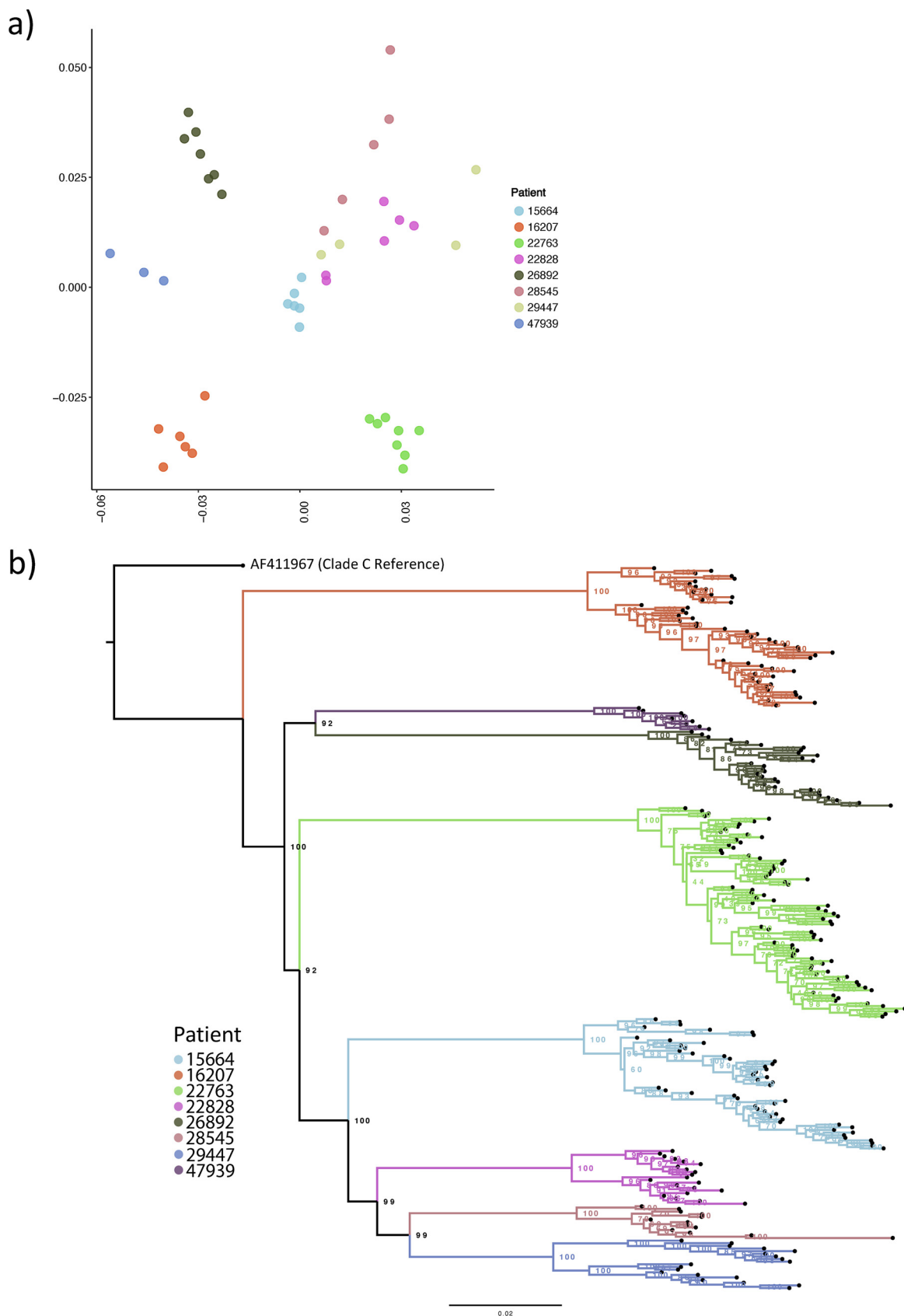
FIG 2 (a) Multidimensional scaling showing clustering of HIV whole genomes from consensus sequences with high intrapatient diversity. Multidimensional scaling (MDS) was created by determining all pairwise distance comparisons under a TN93 substitution

28545). To corroborate the multidimensional scaling (MDS) approach, we used an alternative novel method of examining nucleotide diversity of longitudinal time points using all positional information from BAM files (Fig. S2).

Phylogenetic analysis of inferred haplotypes. The preceding diversity assessments suggested the existence of distinct viral haplotypes within each patient. We therefore used a recently reported computational tool, Haplotype Reconstruction for Longitudinal Samples (HaROLD) (31), to infer 289 unique haplotypes across all patients, with between 11 and 32 haplotypes (average, 21) per patient. The number of haplotypes changed dynamically between successive time points, indicative of dynamically shifting populations (Fig. 2b). To confirm the plausibility of haplotypes, a phylogeny of all consensus sequences was inferred (Fig. S3), and an MDS plot of all viral haplotypes was constructed (Fig. S4).

Linkage disequilibrium (LD) and recombination. LD between two pairwise loci is reduced by recombination, such that LD tends to be higher for loci that are close and lower for more distant loci (32). HIV-1 is known to recombine such that sequences are not generally in linkage disequilibrium beyond 400 bp (9). The significance of recombination in an intrahost, chronic-infection setting is less well understood (33). To assess whether intrapatient recombination was occurring between the haplotypes observed in each of the three most sampled patients, we determined LD decay patterns. We assumed that if there was random recombination, this would equate to smooth LD decay patterns. This was not observed. Rather, each patient demonstrated a complex decay pattern, consistent with nonrandom recombination along the genome (Fig. 3a). Given this, we characterized recombination patterns (Fig. 3b). Inferred recombination breakpoints were identified within patients over successive time points (Fig. S5). DRMs accumulated over successive time points for patient 22763, whereas in patient 15664 the reverse was true. Patient 16207 had recombinant breakpoints localized in the *pol* gene in two time points, though it retained its majority DRM (K103N) across all haplotype populations, possibly as a result of K103N being acquired as a transmitted DRM or as all variants were under the same selective pressure.

Changing landscapes of nonsynonymous and synonymous mutations. In the absence of major PI mutations, we first examined nonsynonymous mutations across the whole genome (Fig. 4 and 5), with a specific focus on *pol* (to identify known first- and second-line NRTI-associated mutations) and *gag* (given its known involvement in PI susceptibility). We and others have previously shown that *gag* mutations accumulate during nonsuppressive PI therapy (34, 35). There are also data suggesting associations between *env* mutations and PI exposure (36, 37). Tables S1 to S3 summarize the changes in variant frequencies of *gag*, *pol*, and *env* mutations in patients over time. We found between two and four mutations at sites previously associated with PI resistance in each patient, all at persistently high frequencies (>90%) even in the absence of presumed drug pressure. This is explained by the fact that a significant proportion of sites associated with PI exposure are also polymorphic across HIV-1 subtypes (20, 38). To complement this analysis, we examined underlying synonymous mutations across the genome. This revealed complex changes in the frequencies of multiple nucleotide residues across all genes. These changes often formed distinct chevron-like pattens between time points (Fig. 4c and 6b), indicative of linked alleles dynamically shifting, which is in turn suggestive of competition between viral haplotypes.

Three patients (15664, 16207, and 22763), which had the greatest number of time points for ongoing comparison and had the highest read coverage, were selected for in-depth viral dynamics analysis as discussed below.

Patient 15664. This patient had consistently low plasma concentrations of all drugs at each measured time point, with detectable levels measured only at month 15 and

FIG 2 Legend (Continued)

model, color-coded by patient. Axes are MDS-1 and MDS-2. (b) Maximum likelihood phylogeny of constructed viral haplotypes for all patients. The phylogeny was rooted on the clade C reference genome (GenBank accession number AF411967). Reconstructed haplotypes were genetically diverse and did not typically cluster by time point.
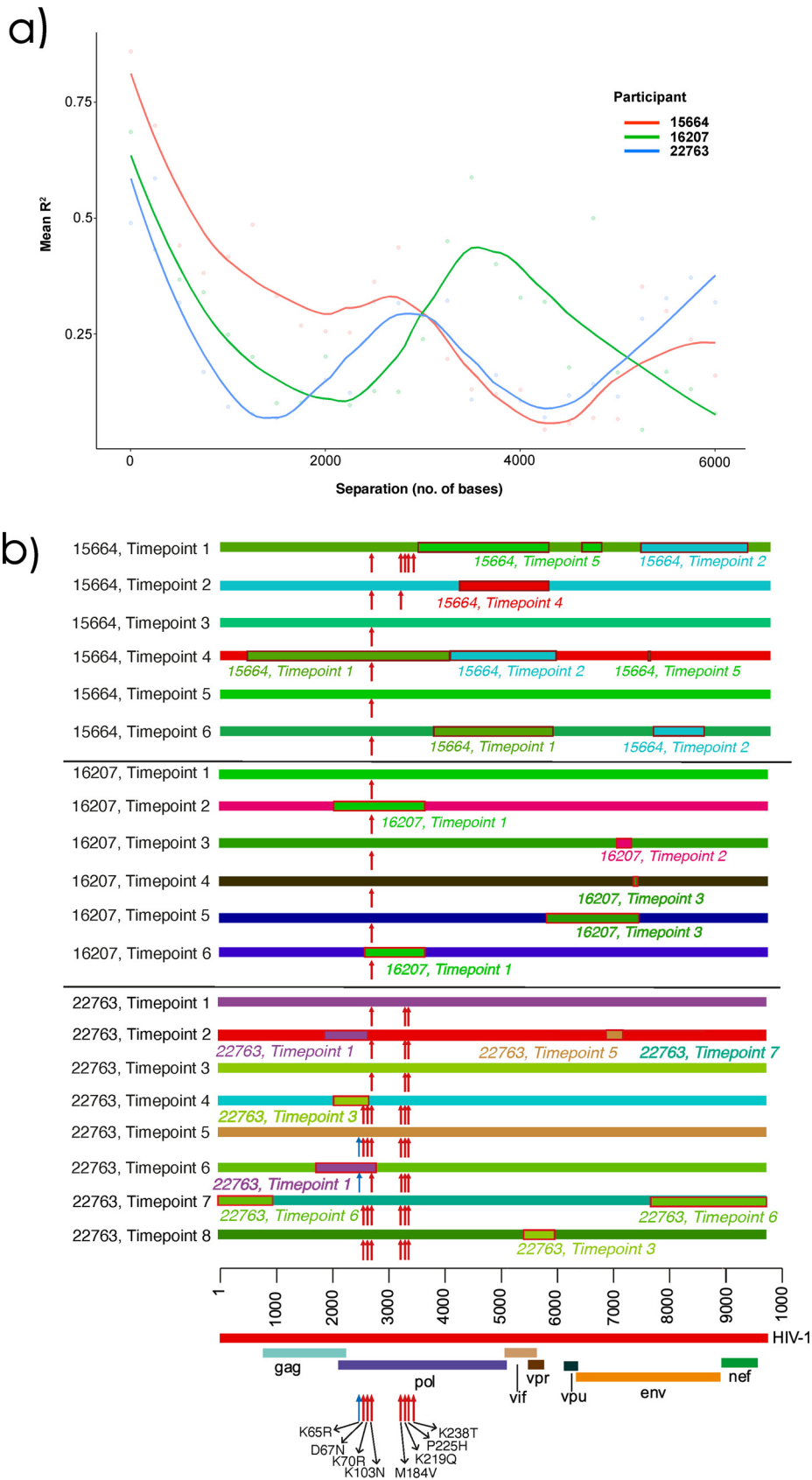
FIG 3 (a) Pairwise linkage disequilibrium decays rapidly with increasing distance between SNPs. Lines represent patterns of LD for each patient examined in-depth. There was a constant decrease in linkage disequilibrium

(Continued on next page)

beyond (Fig. 4a). At baseline, while on NNRTI-based 1st-line ART, known NRTI (M184V) and NNRTI (K103N and P225H) DRMs (5) were at high prevalence in the virus populations, which is as expected while adhering to 1st-line treatments. Haplotype reconstruction and subsequent analysis inferred the presence of a majority haplotype carrying all three of these mutations at baseline, as well as a minority haplotype with the absence of P225H (Fig. 4d, dark gray circles). Following the switch to a 2nd-line regimen, variant frequencies of M184V and P225H dropped below detection limits (<2% of reads), while K103N remained at high frequency (Fig. 4b). Haplotype analysis was concordant, revealing that viruses with K103N, M184V, and P225H were replaced by haplotypes with only K103N (Fig. 4d, light gray circles). At time point 2 (month 8), there were also numerous synonymous mutations observed at high frequency in both *gag* and *pol* genes, corresponding with the switch to a 2nd-line regimen. At time point 3 (15 months post-switch to 2nd-line regimen) drug concentrations were highest, though still low in absolute terms, indicating poor adherence. Between time points 3 and 4 we observed a 2-log reduction in viral load, with a modest change in frequency of RT DRMs. However, we observed synonymous variant frequency shifts predominantly in both *gag* and *pol* genes, as indicated by multiple variants increasing and decreasing contemporaneously, creating characteristic chevron patterning (Fig. 4b). However, many of the changes were between intermediate frequencies, (e.g., between 20% and 60%), which differed from changes between time points 1 and 2, where multiple variants changed more dramatically in frequency from <5% to more than 80%, indicating harder selective sweeps. These data are in keeping with a soft selective sweep between time points 3 and 5. Between time points 5 and 6, the final two samples, there was another population shift; M184V and P225H frequencies fell below the detection limit at time point 6, whereas the frequency of K103N dropped from almost 100% to around 80% (Fig. 4b). This was consistent with the haplotype reconstruction, which inferred a dominant viral haplotype at time point 6 bearing only K103N, as well as three minor haplotypes with no DRMs at all (Fig. 4d, light blue circles).

The phylogeny of inferred haplotype sequences showed that haplotypes from all time points were interspersed throughout the tree (except at time point 4, which remained phylogenetically distinct). DRMs showed some segregation by clade; viruses carrying a higher frequency of DRMs (M184V, P225H, and K238T) were observed in clade A (Fig. 4d), and those with either K103N alone or no DRMs were preferentially located in clade C (Fig. 4d). However, this relationship was not clear-cut and therefore was consistent with competition between haplotypes during low drug exposure. Soft sweeps were evident, given the increasing diversity (Fig. 1, Fig. S4) of this patient.

Patient 16207. Viral loads in this patient were consistently above 10,000 copies/mL (Fig. 6a). As with patient 15664, detectable drug concentrations in blood plasma were either extremely low or absent at each measured time point, consistent with nonadherence to the prescribed regimen. There was little change in the frequency of DRMs throughout the follow-up period, even when making the switch to the 2nd-line regimen. NNRTI resistance mutations such as K103N are known to have minimal fitness costs (26) and can therefore persist in the absence of NNRTI pressure. Throughout treatment, the viruses from this patient maintained K103N at a frequency of >85% but also carried an integrase strand transfer inhibitor (INSTI)-associated mutation (E157Q) and PI-exposure-associated amino acid replacements (L23I and M46I) at low frequencies at time points 2 and 3. Despite little change in DRM site frequencies, very significant viral population shifts were observed at the whole-genome level, again indicative of selective sweeps (Fig. 6b and c). Between time points 1 and 4, several linked muta-

FIG 3 Legend (Continued)

over the first 800 bp. (b) Putative recombination breakpoints and drug resistance-associated mutations of all longitudinal consensus sequences belonging to three patients, 15664, 16207, and 22763. All sequences were colored uniquely; perceived recombination events supported by 4 or more methods implemented in RDP5 are highlighted with a red border and italic text to show the major parent and recombinant portion of the sequence. Drug resistance-associated mutations are indicated with a red arrow, relative to the key at the bottom of the image. For ease of distinguishment, the K65R mutations are indicated with a blue arrow.
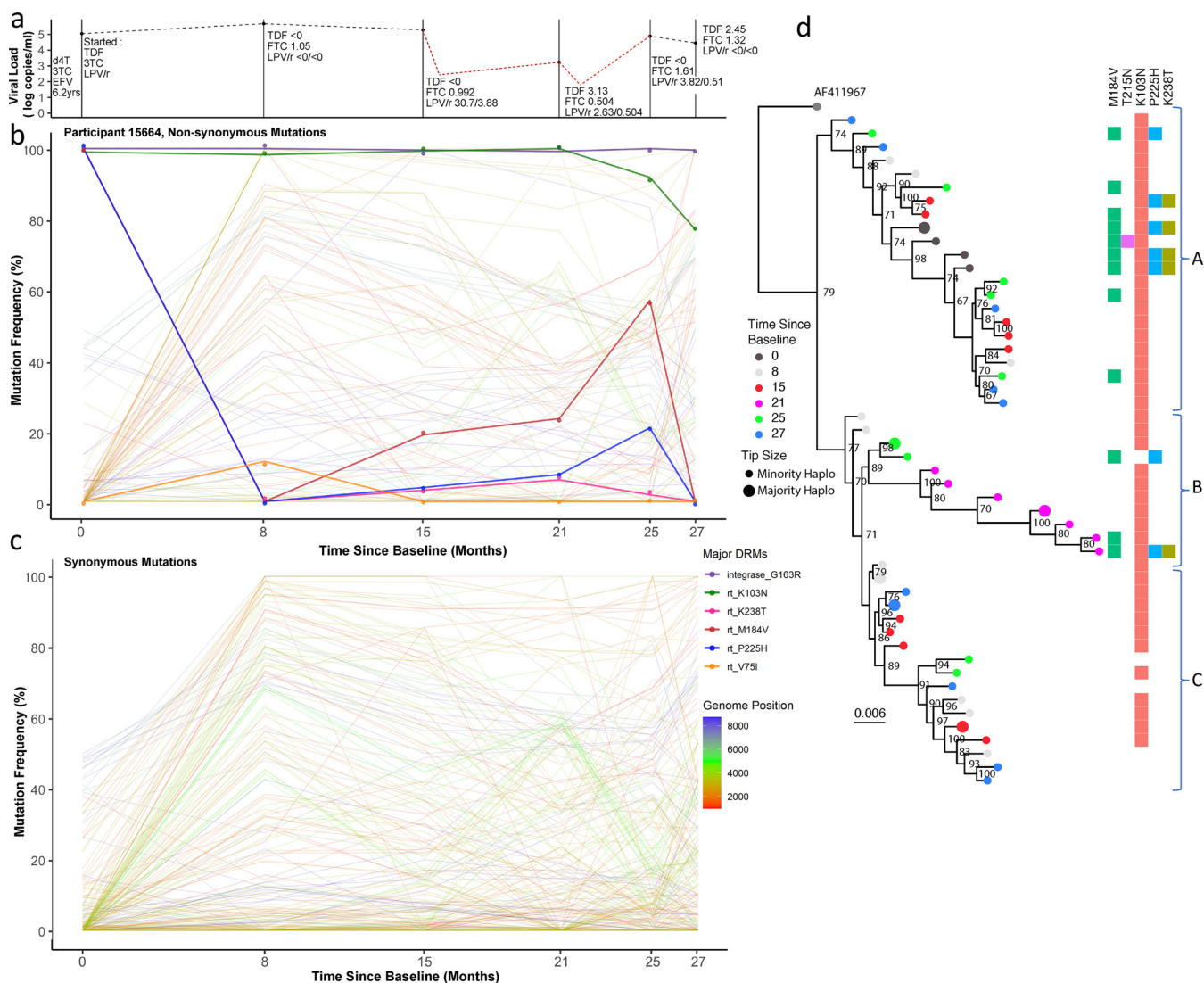
FIG 4 Drug regimen, adherence, and viral dynamics within patient 15664. (a) Viral load and drug levels. At successive time points the drug regimen was noted and the plasma drug concentration measured by HPLC (nmol/L). The patient was characterized by multiple partial suppression (<750 copies/mL, 16 months; <250 copies/mL, 22 months) and rebound events (red dotted line) and poor adherence to the drug regimen. (b) Drug resistance- and non-drug resistance-associated nonsynonymous mutation frequencies determined by Illumina NGS. The Patient had large population shifts between time points 1 and 2, consistent with a hard selective sweep, coincident with the shift from the 1st-line regimen to 2nd-line. (c) Synonymous mutation frequencies. All mutations with a frequency of <10% or >90% at two or more time points were tracked over successive time points. Most changes were restricted to *gag* and *pol* regions and had limited shifts in frequency, i.e., between 20 and 60%. (d) Maximum likelihood phylogeny of reconstructed haplotypes. Haplotypes largely segregated into three major clades (labeled A to C). Majority and minority haplotypes, some carrying lamivudine resistance mutation M184V. Clades referred to in the text body are shown to the right of the heatmap.

tions changed abundance contemporaneously, generating chevron-like patterns of nonsynonymous changes in *env* specifically (blue lines, Fig. 6b). A large number of alleles increased in frequency from <40% to >80% at time point 1, followed by decreases in frequency from >70% to <30% at time point 3. Whereas large shifts in *gag* and *pol* alleles also occurred, the mutations involved were almost exclusively synonymous (red and green lines).

Phylogenetic analysis of inferred whole-genome haplotypes again showed a distinct cladal structure as observed in patient 15664 (Fig. 6d), although the dominant haplotypes were equally observed in the upper clade (A) and lower clade (C) (Fig. 6d). K103N was the majority DRM at all time points, except for a minority haplotype at time point 3, also carrying E157Q. Haplotypes did not cluster by time point. Significant diversity in haplotypes from this patient was confirmed by MDS (Fig. S4).

Patient 22763. This patient was notable for a number of large shifts in variant
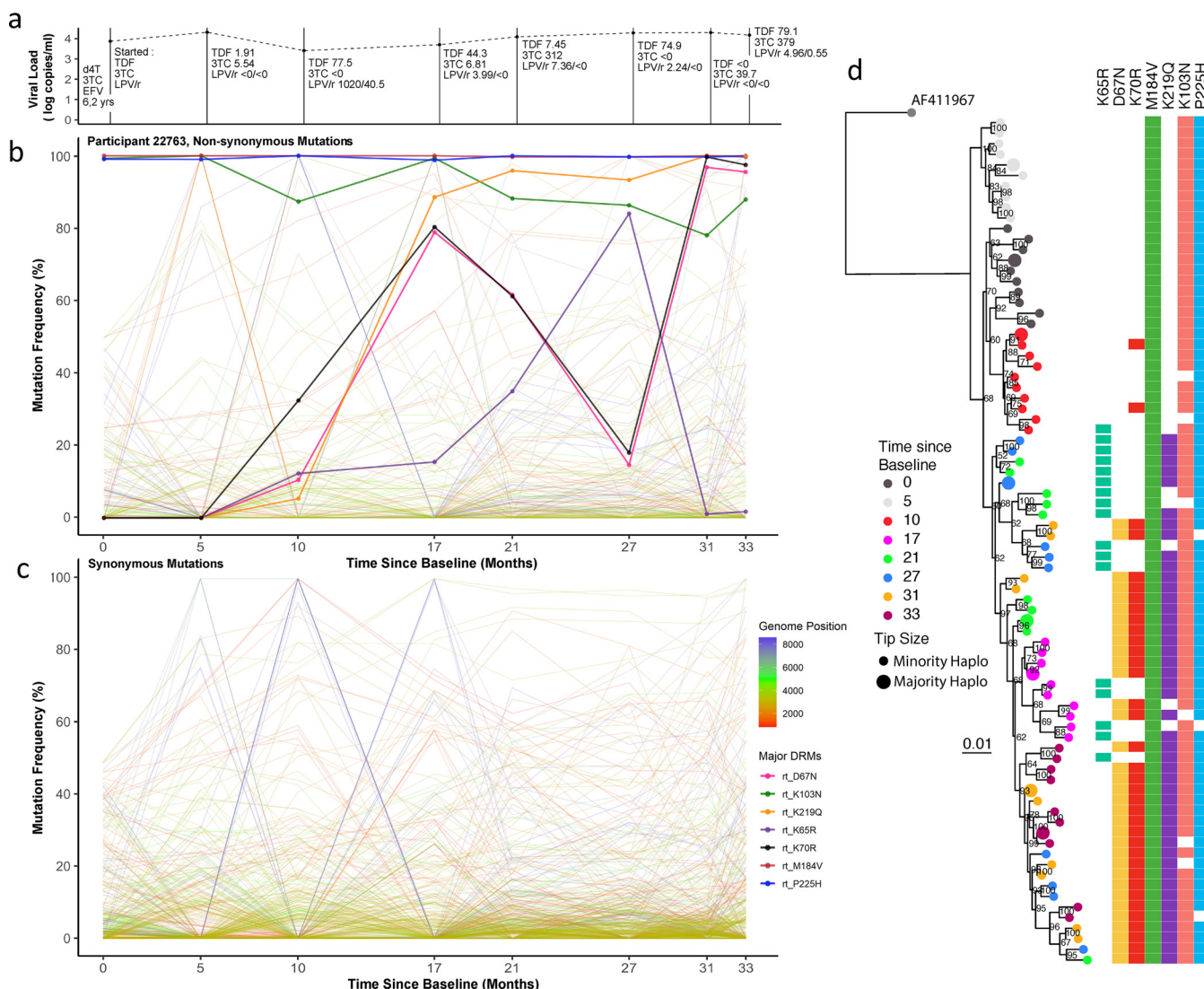
FIG 5 Drug regimen, adherence, and viral dynamics of patient 22763. (a) Viral load and regimen adherence. At successive time points the regimen was noted and plasma drug concentration was measured by HPLC (nmol/L). The patient had therapeutic levels of drug at several time points (3, 5, and 8), indicating variable adherence to the prescribed drug regimen. (b) Drug resistance- and non-drug resistance-associated nonsynonymous mutation frequencies. The patient had numerous drug resistance mutations in dynamic flux. Between time points 4 and 7, there was a complete population shift, indicated by reciprocal competition between the RT mutation K65R and the TAMs K67N and K70R. (c) Synonymous mutations frequencies. All mutations with a frequency of <10% or >90% at two or more time points were followed over successive time points. Several *env* mutations mimicked the nonsynonymous shifts observed between time points 2 and 4, suggestive of linkage. (d) Maximum likelihood phylogeny of reconstructed haplotypes. Time points 1 to 4 were found in distinct lineages. In later time points, from 5 to 8, haplotypes became more intermingled while maintaining antagonism between K65R- and K67N-bearing viruses.

frequencies across multiple drug resistance-associated residues and synonymous sites. The drug plasma concentration for different drugs was variable yet detectable at most measured time points. This suggests that the patient took some of their prescribed drugs throughout the follow-up period (Fig. 5a). Non-PI DRMs such as M184V, P225H, and K103N were present at baseline (time of switch from first- to second-line treatments). These mutations persisted despite synonymous changes between time points 1 and 2. Most of the highly variable synonymous changes in this patient were found in the *gag* and *pol* genes (as in patient 16207) (Fig. 5c), but in this case *env* displayed large fluctuations in synonymous and nonsynonymous allelic frequencies over time. At time point 3, therapeutic concentrations of boosted lopinavir (LPV/r) and tenofovir (TDF) were measured in plasma and haplotypes clustered separately from the first two time points (Fig. 5d, light and dark gray circles). NGS confirmed that the D67N, K219Q, K65R, L70R, and M184V DRMs and NNRTI resistance mutations were present at low
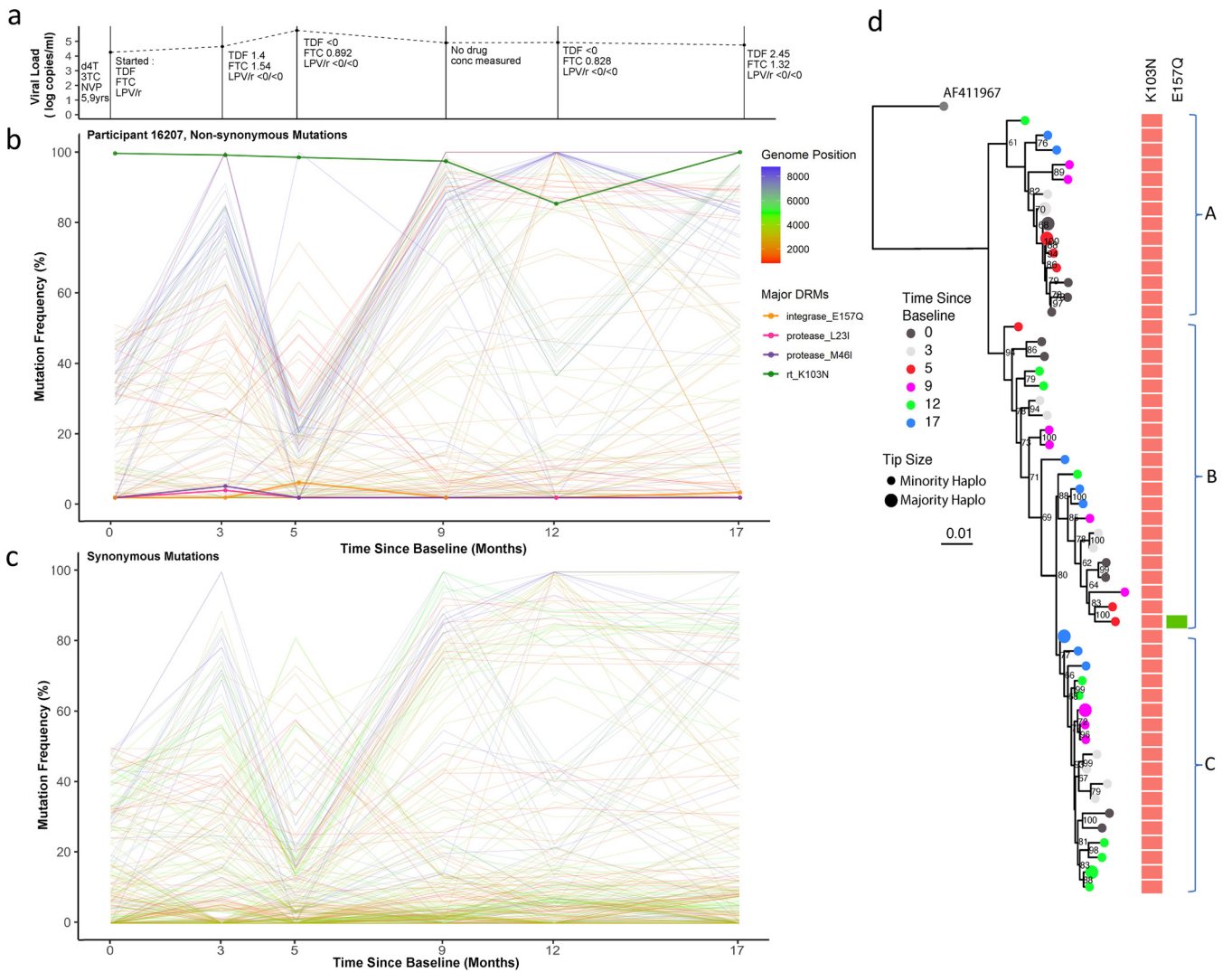
FIG 6 Drug regimen, adherence, and viral dynamics within patient 16207. (a) Viral load and drug levels. At successive time points the regimen was noted and the plasma drug concentration measured by HPLC (nmol/L). The patient displayed ongoing viremia and poor adherence to the prescribed drug regimen. (b) Drug resistance- and non-drug resistance-associated nonsynonymous mutation frequencies. The patient had only one major RT mutation, K103N, for the duration of the treatment period. Several antagonistic nonsynonymous switches predominantly in *env* were observed between time points 1 and 4. (c) Synonymous mutation frequencies. All mutations with a frequency of <10% or >90% at two or more time points were followed over successive time points. In contrast to nonsynonymous mutations, most synonymous changes were in *pol*, indicative of linkage to the *env* coding changes. (d) Maximum likelihood phylogeny of reconstructed haplotypes. Haplotypes were again clearly divided intro three distinct clades; each clade contained haplotypes from all time points, suggesting a lack of hard selective sweeps and intermingling of viral haplotypes with softer sweeps. Most viral competition occurred outside drug pressure.

frequencies from time point 3 onward. Of note, between time points 3 and 6, therapeutic concentrations of TDF were detectable and coincided with increased frequencies of the canonical TDF DRM, K65R (5). The viruses carrying K65R outcompeted those carrying the thymidine analogue mutants (TAMs) D67N and K70R, while the lamivudine (3TC)-associated resistance mutation, M184V, persisted throughout. In the final three time points M46I emerged in protease but never increased in frequency above 6%. At time point 7, populations shifted again, with some haplotypes resembling those previously seen in time point 4, with D67N and K70R again being predominant over K65R in reverse transcriptase (Fig. 5d, green and blue circles). At the final time point (8) the frequency of K103N was approximately 85%, and the TAM-bearing populations continued to dominate over the K65R population, which at this time point had a low frequency.

Although the DRM profile suggested the possibility of a selective sweep, we observed the same groups of other nonsynonymous or synonymous alleles exhibiting

dramatic frequency shifts, but to a lesser degree than in patients 16207 and 15664; i.e., chevron patterns were less pronounced, outside the *env* gene (Fig. 5b and c). Variable drug pressures placed on the viral populations throughout the 2nd-line regimen appear to have played some role in limiting haplotype diversity. Time points 1 to 4 all formed distinct clades, without intermingling, indicating that competition between populations was not occurring to the same degree as in previous patients. Some inferred haplotypes had K65R and others, the TAMs D67N and K70R. K65R was not observed in combination with D67N or K70R, consistent with previously reported antagonism between K65R and TAMs, whereby these mutations are not commonly found together within a single genome (39–41). One explanation for the disconnect between the trajectories of DRM frequencies over time and haplotype phylogeny is competition between different viral populations. Alternatively, emergence of haplotypes from previously unsampled reservoirs with different DRM profiles is possible, but one might have expected other mutations to characterize such haplotypes that would manifest as changes in the frequencies of large numbers of other mutations.

DISCUSSION

The proportion of people living with HIV (PLWH) who are accessing ART has increased from 24% in 2010, to 68% in 2020 (42, 43). However, with the scale-up of ART, there has also been an increase in both pretreatment drug resistance (PDR) (44, 45) and acquired drug resistance (14, 46) to 1st-line ART regimens containing NNRTIs. Integrase inhibitors (specifically dolutegravir) are now recommended for first-line regimens by the WHO in regions where PDR exceeds 10% (47). Boosted PI-containing regimens remain 2nd-line drugs following 1st-line failure, though one unanswered question relates to the nature of viral populations during failure on PI-based ART, where major mutations in protease, described largely for less potent PIs, have not emerged. Here, we have comprehensively analyzed viral populations present in longitudinally collected plasma samples of chronically infected HIV-1 patients under nonsuppressive 2nd-line ART.

With the vast majority of PLWH who have been treated in the post-ART era, virus dynamics during nonsuppressive ART are important to understand, as there may be implications for future therapeutic success. For example, broadly neutralizing antibodies (bNab) are being tested not only for prevention, but also as part of remission strategies in combination with latency reversal agents. We know that HIV sensitivity to bNabs is dependent on *env* diversity (48, 49), and therefore prolonged ART failure with viral diversification could compromise sensitivity to these agents.

Our understanding of virus dynamics largely stems from studies that were limited to untreated individuals (12), with mostly subgenomic data analyzed rather than whole genomes (12). Traditional analyses of quasispecies distributions, for example, as reported by Yu et al. (50), suggest that viral diversity increases in longitudinal samples. However, the findings of Yu et al. were based entirely on short-read NGS data without considering whole-genome haplotypes. The added benefit of examining whole genomes is that linked mutations can be identified statistically using an approach that we recently developed (31). Indeed, haplotype reconstruction has proved beneficial in the analysis of compartmentalization and diversification of several RNA and DNA viruses, including HIV-1, cytomegalovirus (CMV), and SARS-CoV-2 (34, 51, 52).

Key findings of this study were, first, that diversity as defined by the number of quasispecies in each sample typically increased over time. Considering divergence, (a measure at the consensus level of how many mutations have accumulated in a current sequence, from the founder infection) in contrast to previous literature which showed that there was a degree of reversion to the founder strain (9), we show that there was no significant reversion in our study population. There was also no significant divergence from baseline or ancestral C consensus sequences when considering the whole genome. However, when considering 1,000-bp fragments of the genome in a sliding window, several regions in *pol*, *vpu*, and *env* significantly diverged from the consensus C sequence. Using 1,000-bp windows, we were unable to identify any large-scale

reversions. This is consistent with the divergence hypothesis, as if large-scale reversion was an accepted phenomenon, then HIV-1 would eventually converge to a homogeneous sequence, rather than what we have presented in the manuscript.

A second key finding in our study was that synonymous mutations were generally 2- to 3-fold more frequent than nonsynonymous mutations during nonsuppressive ART during chronic infection—a finding in contrast to that seen previously in a longitudinal study of untreated individuals (2, 12, 50). Nonsynonymous changes were enriched in known polymorphic regions such as *env*, whereas synonymous changes were more often observed to fluctuate in the conserved *pol* gene. This finding may reflect early versus chronic infection and differing selective pressures. Haplotype reconstruction revealed evidence for competing haplotypes, with phylogenetic evidence for numerous soft selective sweeps in that haplotypes intermingled during periods when there were low drug concentrations measured in the blood plasma samples of patients. Nonadherence to drug regimens therefore offers opportunity for the HIV-1 reservoir to increase in size and is associated with higher levels of residual viremia (53), preventing future viral suppression due to accumulation and maintenance of beneficial mutations.

Individuals in the present study were treated with ritonavir-boosted lopinavir along with two NRTIs (typically tenofovir plus emtricitabine). We observed significant changes in the frequencies of NRTI mutations in two of the three patients studied indepth. We saw evidence for possible archived virus populations with DRMs emerging during follow-up, in that large changes in DRM frequency were not always accompanied by changes at other sites. This is consistent both with the occurrence of soft selective sweeps and with previous observations that non-DRMs do not necessarily drift with other mutations to fixation (23). As frequencies of RT DRMs did not always segregate with haplotype frequencies (i.e., the same mutations were repeatedly observed on different genetic backgrounds), we suggest that a high number of recombination events, known to be common in HIV infections, were likely contributing to the observed haplotypic diversity.

Although no patient developed major resistance mutations to PIs at consistently high frequencies (https://hivdb.stanford.edu/dr-summary/resistance-notes/PI/), we did observe nonsynonymous mutations in *gag* which have been previously associated with mediating resistance to PI. There was, however, no temporal evidence of specific mutations being associated with selective sweeps. For example, PI exposure-associated residues in matrix (positions 76 and 81) were observed in patient 16207 prior to PI initiation (54). Furthermore, patient 16207 was one of two patients who achieved low-level viremic suppression (45 to 999 copies/mL) of viral replication at one or two time points. After both of these partial suppressions, the rebound populations appeared to be less diverse, consistent with drug-resistant viruses reemerging.

Mutations at sites in the HIV genome that are further apart than 100 bp are subject to frequent shuffling via recombination (55). Unlike the smooth LD decay curves for pairs of HIV mutations reported in the literature, we identified complex LD decay patterns within the genomes of viruses from individual patients—patterns indicative of nonrandom recombination. Recombination appears as the loss and gain of common genomic regions over successive time points between each patient's haplotype populations (Fig. 3b). Viruses from patient 15664 with interhaplotype recombination events detectable in the *vif* and *vpr* genes were present at four of the six analyzed time points. In contrast, viruses in patient 22763 that had evidence of interhaplotype recombination events in the *gag-pol* genes were present at three of the eight analyzed time points. We explain these recombination events detectable in longitudinally sampled sequences, as reflected in the previously discussed chevron patterns whereby variants increase and subsequently decrease between time points. HIV quasispecies foster a degree of genetic diversity that facilitates rapid adaptive evolution through recombination whenever there exists within the quasispecies combinations of mutations that provide fitness advantages (8). The relationship between recombination and the

accumulation of multiple DRMs within individual genomes is not clearly evident within the analyzed sequence data sets, with viruses sampled from each patient showing unique patterns of recombination. Interhaplotype recombinants detected at time points 2 and 6 in patient 16207 had recombination events in *pol* that involved the transfer of the major DRM, K103N. Three independent inter-haplotype recombination events detected in the *pol* gene from patient 22763 at time points 2, 4, and 6 resulted in no change in DRMS at time point 2, gain of DRMS at time point 4, or loss of DRMs at time point 6. The recombination dynamics in this patient were occurring against a backdrop of apparent antagonism between TAMs and DRMs (K65R and D67N). Finally, patient 15664 steadily lost DRMs throughout the longitudinal sampling period, although we found no evidence of recombination being implicated in this loss. This suggests that, in the absence of strong drug pressures, viral populations only maintained DRMs which were crucial for providing resistance to drugs that the patient was variably adhering to at the time.

Phylogenetic analyses of whole-genome viral haplotypes demonstrated two common features: (i) evidence for selective sweeps following therapy switches or large changes in plasma drug concentrations, with hitchhiking of synonymous and nonsynonymous mutations, and (ii) competition between multiple viral haplotypes that intermingled phylogenetically alongside soft selective sweeps. The diversity of viral populations was maintained between successive time points with ongoing viremia, particularly in *env*. Changes in haplotype dominance were often distinct from the dynamics of drug resistance mutations in reverse transcriptase (RT), indicating the presence of softer selective sweeps and/or recombination.

This study had some limitations; we examined eight patients with ongoing viremia and variable adherence to 2nd-line drug regimens, with three of these being examined in-depth. Despite the small sample size, this type of longitudinal sampling of ART-experienced patients is unprecedented. We are confident that the combination of computational analyses has provided a detailed understanding of viral dynamics under nonsuppressive ART that will be applicable to wider data sets. The method used to reconstruct viral haplotypes *in silico* is novel and has previously been validated in HIV-1-positive patients coinfected with CMV (51). We are confident that the approach implemented by HaROLD has accurately, if conservatively, estimated haplotype frequencies, and future studies should look to validate these frequencies using an *in vitro* method such as single-genome amplification.

Despite there being high viral loads present at each of the analyzed time points, nuances of the sequencing method resulted in suboptimal gene coverage, particularly in the *env* gene. To ensure that uneven sequencing coverage did not bias our analyses, we ensured that variant analysis was only performed where coverage was >100 reads. We also utilized a second method of haplotype reconstruction, in order to determine concordance of DRM calls between the two methods used. We find that there was good concordance between the two methods, specifically highlighted by the antagonism between TAMs (D67N and K70R) and NRTI mutations (K65R) in patient 22763 (Fig. S6).

In summary, we have found compelling evidence of HIV-1 within-host viral diversification, recombination, and haplotype competition during nonsuppressive ART. In the future, patients failing PI-based regimens are likely to be switched to INSTI-based ART (specifically dolutegravir in South Africa) prior to genotypic typing or resistance analysis. Although the prevalence of underlying major INSTI resistance mutations is low in sub-Saharan Africa (56, 57), data linking individuals with NNRTI resistance with poorer virological outcomes on dolutegravir (58), coupled with a history of intermittent adherence, warrant further investigation. Having shown that long-time intrahost PI failure increases the intrapatient diversity of HIV populations, monitoring future drug-failure cases will be of interest due to their capacity to maintain a reservoir of transmissible drug-resistant viruses, as well as impacting responses to future therapies.

MATERIALS AND METHODS

Study and patient selection. This cohort was nested within the French ANRS 12249 Treatment as Prevention (TasP) trial (27). TasP was a cluster-randomized trial comparing an intervention arm which offered ART after HIV diagnosis irrespective of patient CD4$^+$ count to a control arm offering ART according to prevailing South African guidelines. In total, a subset of 44 longitudinal samples from eight chronically infected patients with virological failure of 2nd-line PI-based ART, with viremia above 1,000 copies/mL, were analyzed. From these eight patients, three patients with a mean coverage of >2,000 reads across the whole genome were selected from in-depth viral dynamic analysis. All samples were collected from blood plasma. The Illumina MiSeq platform and an adapted protocol for sequencing were used (59). Adherence to 2nd-line regimens was measured by high-performance liquid chromatography (HPLC) using plasma concentration of drug levels as a proxy. Drug levels were measured at each time point with detectable viral loads, post-PI initiation. Cutoffs for assessment of adherence were selected from published literature.

Ethical approval was originally granted by the Biomedical Research Ethics Committee (BFC 104/11) at the University of KwaZulu-Natal and the Medicines Control Council of South Africa for the TasP trial (Clinicaltrials.gov: NCT01509508; South African Trial Register: DOH-27-0512-3974). The study was also authorized by the KwaZulu-Natal Department of Health in South Africa. Written informed consent was obtained from all patients. Original ethical approval also included downstream sequencing of blood plasma samples and analysis of those sequences to better understand drug resistance. No additional ethical approval was required for this.

Illumina sequencing. Sequencing of viral RNA was performed as previously described by Derache et al. (60) using a modified protocol previously described by Gall et al. (61). Briefly, RNA was extracted from 1 mL of plasma with a detectable viral load of >1,000 copies/mL, using QIAamp viral RNA minikits (Qiagen, Hilden, Germany) and eluted in 60 μL of elution buffer. The near-full HIV genome was amplified with four HIV-1 subtype C primer pairs, generating 4 overlapping amplicons between 2,100 and 3,900 kb.

DNA concentrations of amplicons were quantified with the Qubit double-stranded DNA (dsDNA) high-sensitivity (HS) assay kit (Invitrogen, Carlsbad, CA). Diluted amplicons were pooled equimolarly and prepared for the library using the Nextera XT DNA library preparation and the Nextera XT DNA sample preparation index kits (Illumina, San Diego, CA), following the manufacturer's protocol.

Genomics and bioinformatics. Poor-quality reads (with a Phred score of <30) and adapter sequences were trimmed from FastQ files with TrimGalore! v0.6.519 (62) and mapped to a dual-tropic, clade C, South African reference genome (GenBank accession number AF411967) with Minimap2 (63). The reference genome was manually annotated in Geneious Prime v2020.3 with DRMs according to the Stanford HIV Drug Resistance Database (HIVdb) (64). Optical PCR duplicate reads were removed using Picard tools (http://broadinstitute.github.io/picard). Finally, Qualimap 2 (65) was used to assess the mean mapping quality scores and coverage in relation to the reference genome for the purpose of excluding poorly mapped sequences from further analysis. Single nucleotide polymorphisms (SNPs) were called using VarScan 2 (66) with a minimum average quality of 20, minimum variant frequency of 2%, and in at least 100 reads. These were then annotated by gene, codon, and amino acid alterations using an in-house script (67) modified to utilize HIV genomes.

All synonymous and nonsynonymous variants (including DRMs) were examined, and their frequencies were compared across successive time points. Synonymous variants were excluded from analysis if their prevalence remained at ≤10% or ≥90% across all time points. DRMs were retained for analysis if they were present at over 2% frequency and on at least two reads. A threshold of 2% is supported by a study evaluating different analysis pipelines, which reported fewer discordances over this cutoff (68).

Measuring divergence or reversion to baseline and consensus C ancestor. For each patient, divergence over time from inferred founder state was measured for (i) the baseline sequence for each patient and (ii) a reconstructed subtype C consensus. The full-length HIV-1 subtype C consensus was downloaded from the LANL HIV database, and annotations from the subtype C reference sequence (GenBank accession number AF411967) used for haplotype reconstruction were transferred to this genome using Geneious Prime v2021.1.0 to ensure that positions remained consistent throughout.

Divergence was measured as the pairwise distance between time point consensus and founder, calculated using the dist.dna() package with a TN93 nucleotide-nucleotide substitution matrix and with pairwise deletion as implemented in the R package Ape v5.4.

As a validation, an in-house script was used to examine all amino acids on a site-per-site basis. The initial AA at time point 1 (baseline or ancestral C) is recorded. Where there is a mutation at any subsequent time point except the last time point, we measure if the AA at the final time point is the same or different from the first. If there has been a reversion mutation, this will be the same as the first time point. If there has been a diverging mutation, this will be different from the first.

Linear mixed-effects models. To investigate the general relationship of time in months to divergence, incorporating all 8 patients, we built a series of linear mixed-effects models implemented in the lmer R package. Divergence was treated as the response, time was treated as a fixed effect, and patient was treated as a random effect. We built similar models for the whole-genome and discrete genomic portions analysis for each founder strain. We tested if time had a non-0 effect on divergence by calculating the P value using Satterthwaite's method as implemented in the lmerTEst package. For the 1,000-bp analyses, a Benjamini-Hochberg correction adjustment was undertaken to account for 9 tests within the same sample.

Haplotype reconstruction and phylogenetics. Whole-genome viral haplotypes were constructed for each patient time point using HaROLD (Haplotype Reconstruction for Longitudinal Samples) (31). The first stage consists of SNPs being assigned to each haplotype such that the frequency of variants is

equal to the sum of the frequencies of haplotypes containing a specific variant. This considers the frequency of haplotypes in each sample, the base found at each position in each haplotype, and the probability of erroneous measurements at that site. Maximal log likelihood was used to optimize time-dependent frequencies for longitudinal haplotypes, which was calculated by summing over all possible assignments of haplotype variants. Haplotypes were then constructed based on posterior probabilities.

After constructing haplotypes, a 2nd stage or refinement process remaps reads from BAM files to constructed haplotypes. This begins with the *a posteriori* probability of each base occurring at each site in each haplotype from the first stage but relaxes the assumption that haplotypes are identical at each sample time point and instead uses variant colocalization to refine haplotype predictions. Starting with the estimated frequency of each haplotype in a sample, haplotypes are optimized by probabilistically assigning reads to the various haplotypes. Reads are then reassigned iteratively until haplotype frequencies converge. The number of haplotypes either increases or decreases as a result of combination or division according to Akaike information criterion (AIC) scores, in order to present the most accurate representation of viral populations at each time point.

Whole-genome nucleotide diversity was calculated from BAM files using an in-house script (https://github.com/ucl-pathgenomics/NucleotideDiversity). Briefly diversity is calculated by fitting all observed variant frequencies to either a beta distribution or four-dimensional Dirichlet distribution plus delta function (representing invariant sites). These parameters were optimized by maximum log likelihood.

Maximum-likelihood phylogenetic trees and ancestral reconstruction were performed using IQ-TREE 2 v2.1.3 (69) and a GTR+F+I model with 1,000 ultrafast bootstrap replicates (70). All trees were visualized with Figtree v1.4.4 (https://github.com/rambaut/figtree/releases), rooted on the AF411967 reference sequence, and nodes were arranged in descending order. Phylogenies were manipulated and annotated using ggtrree v2.2.4.

Additionally, as a sensitivity analysis, Cliquesnv (71) was used to infer a second set of haplotypes using the following flags: -m snv-illumina -fdf extended4 -threads 20 -cm accurate. This was to determine concordance of drug resistance mutation calls within haplotypes.

Multidimensional scaling (MDS) plots. Pairwise distances between these consensus sequences were calculated using the dist.dna() package, with a TN93 nucleotide-nucleotide substitution matrix and with pairwise deletion implemented in the R package Ape v5.4. Nonmetric multidimensional scaling (MDS) was implemented using the metaMDS() function in the R package vegan v2.5.7. MDS is a method to attempt to simplify high-dimensional data into a simpler representation of reducing dimensionality while retaining most of the variation relationships between points. We find that like network trees, nonmetric MDS better represents the true relative distances between sequences, whereas eigenvector methods are less reliable in this sense. In a genomics context we can apply dimensionality reduction on pairwise distance matrices, where each dimension is a sequence with data points of $n - 1$ sequences pairwise distance. The process was repeated with whole-genome haplotype sequences.

Linkage disequilibrium and recombination. Starting with a sequence alignment, we determined the pairwise LD r^2 associations for all variable sites using WeightedLD (72) without weighting. This method allowed us to easily exclude sites with any insertions or ambiguous characters, where we used the options –min-acgt 0.99 and –min-variability 0.05. The pairwise r^2 values (the square of the correlation coefficient between two indicator variables) were then binned per 200-bp comparison distance blocks along the genome, and the mean r^2 values were taken and represented graphically to assess LD decay. This analysis was run for the three patients taken forward for in-depth analysis and run using an alignment of all their time point samples. Graphics were generated using R v4.04.

We first performed an analysis for detecting individual recombination events in individual genome sequences using the RDP, GENECONV, BOOTSCAN, MAXCHI, CHIMAERA, SISCAN, and 3SEQ methods implemented in RDP5 (73) with default settings. Putative breakpoint sites were identified and manually checked and adjusted if necessary using the BURT method with the MAXCHI matrix and LARD two-breakpoint scan methods. Final recombination breakpoint sites were confirmed if at least three or more methods supported the existence of the recombination breakpoint.

Data and code availability. All BAM files used to undertake analyses have been deposited in the SRA database with the accession numbers SRR15510046 to SRR15510072. Custom code used to produce figures and graphs can be found at https://github.com/ojcharles/HIV1-evolutionary-dynamics.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, PDF file, 2.1 MB.
FIG S2, PDF file, 0.1 MB.
FIG S3, PDF file, 0.1 MB.
FIG S4, PDF file, 0.3 MB.
FIG S5, PDF file, 0.1 MB.
FIG S6, PDF file, 0.2 MB.
TABLE S1, DOCX file, 0.1 MB.
TABLE S2, DOCX file, 0.1 MB.
TABLE S3, DOCX file, 0.1 MB.
TABLE S4, DOCX file, 0.01 MB.

REFERENCES

1. Abrahams M-R, Anderson JA, Giorgi EE, Seoighe C, Mlisana K, Ping L-H, Athreya GS, Treurnicht FK, Keele BF, Wood N, Salazar-Gonzalez JF, Bhattacharya T, Chu H, Hoffman I, Galvin S, Mapanje C, Kazembe P, Thebus R, Fiscus S, Hide W, Cohen MS, Karim SA, Haynes BF, Shaw GM, Hahn BH, Korber BT, Swanstrom R, Williamson C, Center for HIV-AIDS Vaccine Immunology Consortium. 2009. Quantitating the multiplicity of infection with human immunodeficiency virus type 1 subtype C reveals a non-Poisson distribution of transmitted variants. J Virol 83:3556–3567. https://doi.org/10.1128/JVI.02132-08.

2. Zanini F, Puller V, Brodin J, Albert J, Neher RA. 2017. In vivo mutation rates and the landscape of fitness costs of HIV-1. Virus Evol 3:vex003. https://doi.org/10.1093/ve/vex003.

3. Salemi M. 2013. The intra-host evolutionary and population dynamics of human immunodeficiency virus type 1: a phylogenetic perspective. Infect Dis Rep 5:e3. https://doi.org/10.4081/idr.2013.s1.e3.

4. Lemey P, Rambaut A, Pybus OG. 2006. HIV evolutionary dynamics within and among hosts. Aids Rev 8:125–140.

5. Collier DA, Monit C, Gupta RK. 2019. The impact of HIV-1 drug escape on the global treatment landscape. Cell Host Microbe 26:48–60. https://doi.org/10.1016/j.chom.2019.06.010.

6. Biebricher CK, Eigen M. 2006. What is a quasispecies? Curr Top Microbiol Immunol 299:1–31. https://doi.org/10.1007/3-540-26397-7_1.

7. Wilke CO. 2005. Quasispecies theory in the context of population genetics. BMC Evol Biol 5:44. https://doi.org/10.1186/1471-2148-5-44.

8. Lauring AS, Andino R. 2010. Quasispecies theory and the behavior of RNA viruses. PLoS Pathog 6:e1001005. https://doi.org/10.1371/journal.ppat.1001005.

9. Zanini F, Brodin J, Thebo L, Lanz C, Bratt G, Albert J, Neher RA. 2015. Population genomics of intrapatient HIV-1 evolution. Elife 4:e11282. https://doi.org/10.7554/eLife.11282.

10. Lythgoe KA, Fraser C. 2012. New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels. . Proc Biol Sci 279:3367–3375. https://doi.org/10.1098/rspb.2012.0595.

11. Hedskog C, Mild M, Jernberg J, Sherwood E, Bratt G, Leitner T, Lundeberg J, Andersson B, Albert J. 2010. Dynamics of HIV-1 quasispecies during antiviral treatment dissected using ultra-deep pyrosequencing. PLoS One 5:e11345. https://doi.org/10.1371/journal.pone.0011345.

12. Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H, Gupta P, Rinaldo CR, Learn GH, He X, Huang X-L, Mullins JI. 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J Virol 73:10489–10502. https://doi.org/10.1128/JVI.73.12.10489-10502.1999.

13. Masikini P, Mpondo BC. 2015. HIV drug resistance mutations following poor adherence in HIV-infected patient: a case report. Clin Case Rep 3:353–356. https://doi.org/10.1002/ccr3.254.

14. TenoRes Study Group. 2016. G. Global epidemiology of drug resistance after failure of WHO recommended first-line regimens for adult HIV-1 infection: a multicentre retrospective cohort study. Lancet Infect Dis 16:565–575. https://doi.org/10.1016/S1473-3099(15)00536-8.

15. Collier D, Iwuji C, Derache A, de Oliveira T, Okesola N, Calmy A, Dabis F, Pillay D, Gupta RK. 2017. Virological outcomes of second-line protease inhibitor-based treatment for human immunodeficiency virus type 1 in a high-prevalence rural South African setting: a competing-risks prospective cohort analysis. Clin Infect Dis 64:1006–1016. https://doi.org/10.1093/cid/cix015.

16. Giandhari J, Basson AE, Coovadia A, Kuhn L, Abrams EJ, Strehlau R, Morris L, Hunt GM. 2015. Genetic changes in HIV-1 Gag-protease associated with protease inhibitor-based therapy failure in pediatric patients. AIDS Res Hum Retroviruses 31:776–782. https://doi.org/10.1089/AID.2014.0349.

17. Kelly Pillay S, Singh U, Singh A, Gordon M, Ndungu T. 2014. Gag drug resistance mutations in HIV-1 subtype C patients, failing a protease inhibitor inclusive treatment regimen, with detectable lopinavir levels. J Int Aids Soc 17:19784. https://doi.org/10.7448/IAS.17.4.19784.

18. Sutherland KA, Parry CM, McCormick A, Kapaata A, Lyagoba F, Kaleebu P, Gilks CF, Goodall R, Spyer M, Kityo C, Pillay D, Gupta RK, DART Virology Group. 2015. Evidence for reduced drug susceptibility without emergence of major protease mutations following protease inhibitor monotherapy failure in the SARA Trial. PLoS One 10:e0137834. https://doi.org/10.1371/journal.pone.0137834.

19. Sutherland KA, Mbisa JL, Ghosn J, Chaix M-L, Cohen-Codar I, Hue S, Delfraissy J-F, Delaugerre C, Gupta RK. 2014. Phenotypic characterization of virological failure following lopinavir/ritonavir monotherapy using full-length Gag-protease genes. J Antimicrob Chemother 69:3340–3348. https://doi.org/10.1093/jac/dku296.

20. Sutherland KA, Goodall RL, McCormick A, Kapaata A, Lyagoba F, Kaleebu P, Thiltgen G, Gilks CF, Spyer M, Kityo C, Pillay D, Dunn D, Gupta RK, DART Trial Team. 2015. Gag-protease sequence evolution following protease inhibitor monotherapy treatment failure in HIV-1 viruses circulating in East Africa. AIDS Res Hum Retroviruses 31:1032–1037. https://doi.org/10.1089/aid.2015.0138.

21. Day CL, Kiepiela P, Leslie AJ, van der Stok M, Nair K, Ismail N, Honeyborne I, Crawford H, Coovadia HM, Goulder PJR, Walker BD, Klenerman P. 2007. Proliferative capacity of epitope-specific CD8 T-cell responses is inversely related to viral load in chronic human immunodeficiency virus type 1 infection. J Virol 81:434–438. https://doi.org/10.1128/JVI.01754-06.

22. Blanch-Lombarte O, Santos JR, Peña R, Jiménez-Moyano E, Clotet B, Paredes R, Prado JG. 2020. HIV-1 Gag mutations alone are sufficient to reduce darunavir susceptibility during virological failure to boosted PI therapy. J Antimicrob Chemother 75:2535–2546. https://doi.org/10.1093/jac/dkaa228.

23. Feder AF, Rhee S-Y, Holmes SP, Shafer RW, Petrov DA, Pennings PS. 2016. More effective drugs lead to harder selective sweeps in the evolution of drug resistance in HIV-1. Elife 5:e10670. https://doi.org/10.7554/eLife.10670.

24. Harris RB, Sackman A, Jensen JD. 2018. On the unfounded enthusiasm for soft selective sweeps II: examining recent evidence from humans, flies, and viruses. PLoS Genet 14:e1007859. https://doi.org/10.1371/journal.pgen.1007859.

25. Dam E, Quercia R, Glass B, Descamps D, Launay O, Duval X, Kräusslich H-G, Hance AJ, Clavel F, ANRS 109 Study Group. 2009. Gag mutations strongly contribute to HIV-1 resistance to protease inhibitors in highly drug-experienced patients besides compensating for fitness loss. PLoS Pathog 5:e1000345. https://doi.org/10.1371/journal.ppat.1000345.

26. Cong ME, Heneine W, Garcia-Lerma JG. 2007. The fitness cost of mutations associated with human immunodeficiency virus type 1 drug resistance is modulated by mutational interactions. J Virol 81:3037–3041. https://doi.org/10.1128/JVI.02712-06.

27. Iwuji CC, Orne-Gliemann J, Tanser F, Boyer S, Lessells RJ, Lert F, Imrie J, Bärnighausen T, Rekacewicz C, Bazin B, Newell M-L, Dabis F, ANRS 12249 TasP Study Group. 2013. Evaluation of the impact of immediate versus WHO recommendations-guided antiretroviral therapy initiation on HIV incidence: the ANRS 12249 TasP (Treatment as Prevention) trial in Hlabisa sub-district, KwaZulu-Natal, South Africa: study protocol for a cluster randomised controlled trial. Trials 14:230. https://doi.org/10.1186/1745-6215-14-230.

28. World Health Organization. 2016. Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: recommendations for a public health approach. World Health Organization, Geneva, Switzerland.

29. Carlson JM, Schaefer M, Monaco DC, Batorsky R, Claiborne DT, Prince J, Deymier MJ, Ende ZS, Klatt NR, DeZiel CE, Lin T-H, Peng J, Seese AM, Shapiro R, Frater J, Ndung'u T, Tang J, Goepfert P, Gilmour J, Price MA, Kilembe W, Heckerman D, Goulder PJR, Allen TM, Allen S, Hunter E. 2014. HIV transmission. Selection bias at the heterosexual HIV-1 transmission bottleneck. Science 345:1254031–1254031. https://doi.org/10.1126/science.1254031.

30. Cox MA, Cox TF. 2008. Multidimensional scaling, 315–347. In Chen C-H, Härdle W, Unwin A (ed), Handbook of data visualization. Springer, Berlin, Germany.

31. Pang J, Venturini C, Tamuri AU, Roy S, Breuer J, Goldstein RA. 2020. Haplotype assignment of longitudinal viral deep-sequencing data using co-variation of variant frequencies. bioRxiv https://doi.org/10.1101/444877.

32. Stephens M, Scheet P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 76:449–462. https://doi.org/10.1086/428594.

33. Song H, Giorgi EE, Ganusov VV, Cai F, Athreya G, Yoon H, Carja O, Hora B, Hraber P, Romero-Severson E, Jiang C, Li X, Wang S, Li H, Salazar-Gonzalez JF, Salazar MG, Goonetilleke N, Keele BF, Montefiori DC, Cohen MS, Shaw GM, Hahn BH, McMichael AJ, Haynes BF, Korber B, Bhattacharya T, Gao F. 2018. Tracking HIV-1 recombination to resolve its contribution to HIV-1 evolution in natural infection. Nat Commun 9:1928. https://doi.org/10.1038/s41467-018-04217-5.

34. Datir R, Kemp S, El Bouzidi K, Mlchocova P, Goldstein R, Breuer J, Towers GJ, Jolly C, Quiñones-Mateu ME, Dakum PS, Ndembi N, Gupta RK. 2020. In vivo emergence of a novel protease inhibitor resistance signature in HIV-1 matrix. mBio 11:e02036-20. https://doi.org/10.1128/mBio.02036-20.

35. Kletenkov K, Hoffmann D, Böni J, Yerly S, Aubert V, Schöni-Affolter F, Struck D, Verheyen J, Klimkait T, Swiss HIV Cohort Study. 2017. Role of Gag mutations in PI resistance in the Swiss HIV cohort study: bystanders or contributors? J Antimicrob Chemother 72:866–875. https://doi.org/10.1093/jac/dkw493.

36. Rabi SA, Laird GM, Durand CM, Laskey S, Shan L, Bailey JR, Chioma S, Moore RD, Siliciano RF. 2013. Multi-step inhibition explains HIV-1 protease inhibitor pharmacodynamics and resistance. J Clin Invest 123:3848–3860. https://doi.org/10.1172/JCl67399.

37. Manasa J, Varghese V, Pond SLK, Rhee S-Y, Tzou PL, Fessel WJ, Jang KS, White E, Rögnvaldsson T, Katzenstein DA, Shafer RW. 2017. Evolution of gag and gp41 in patients receiving ritonavir-boosted protease inhibitors. Sci Rep 7:11559. https://doi.org/10.1038/s41598-017-11893-8.

38. Datir R, El Bouzidi K, Dakum P, Ndembi N, Gupta RK. 2019. Baseline PI susceptibility by HIV-1 Gag-protease phenotyping and subsequent virological suppression with PI-based second-line ART in Nigeria. J Antimicrob Chemother 74:1402–1407. https://doi.org/10.1093/jac/dkz005.

39. Parikh UM, Zelina S, Sluis-Cremer N, Mellors JW. 2007. Molecular mechanisms of bidirectional antagonism between K65R and thymidine analog mutations in HIV-1 reverse transcriptase. AIDS 21:1405–1414. https://doi.org/10.1097/QAD.0b013e3281ac229b.

40. Parikh UM, Bacheler L, Koontz D, Mellors JW. 2006. The K65R mutation in human immunodeficiency virus type 1 reverse transcriptase exhibits bidirectional phenotypic antagonism with thymidine analog mutations. J Virol 80:4971–4977. https://doi.org/10.1128/JVI.80.10.4971-4977.2006.

41. Parikh UM, Barnas DC, Faruki H, Mellors JW. 2006. Antagonism between the HIV-1 reverse-transcriptase mutation K65R and thymidine-analogue mutations at the genomic level. J Infect Dis 194:651–660. https://doi.org/10.1086/505711.

42. Department of Health. 2019. 2019 ART clinical guidelines for the management of HIV in adults, pregnancy, adolescents, children, infants and neonates. Republic of South Africa National Department of Health, Pretoria, South Africa.

43. UNAIDS. Global HIV & AIDS statistics: 2020 fact sheet. https://www.unaids.org/en/resources/fact-sheet. Accessed 3 March 2021.

44. Gupta RK, Gregson J, Parkin N, Haile-Selassie H, Tanuri A, Andrade Forero L, Kaleebu P, Watera C, Aghokeng A, Mutenda N, Dzangare J, Hone S, Hang ZZ, Garcia J, Garcia Z, Marchorro P, Beteta E, Giron A, Hamers R, Inzaule S, Frenkel LM, Chung MH, de Oliveira T, Pillay D, Naidoo K, Kharsany A, Kugathasan R, Cutino T, Hunt G, Avila Rios S, Doherty M, Jordan MR, Bertagnolio S. 2018. HIV-1 drug resistance before initiation or re-initiation of first-line antiretroviral therapy in low-income and middle-income countries: a systematic review and meta-regression analysis. Lancet Infect Dis 18:346–355. https://doi.org/10.1016/S1473-3099(17)30702-8.

45. Gupta RK, Jordan MR, Sultan BJ, Hill A, Davis DH, Gregson J, Sawyer AW, Hamers RL, Ndembi N, Pillay D, Bertagnolio S. 2012. Global trends in antiretroviral resistance in treatment-naive individuals with HIV after rollout of antiretroviral treatment in resource-limited settings: a global collaborative study and meta-regression analysis. Lancet 380:1250–1258. https://doi.org/10.1016/S0140-6736(12)61038-1.

46. Gregson J, Kaleebu VC, Marconi VC, van Vuuren C, Ndembi N, Hamers RL, Kanki P, Hoffmann CJ, Lockman S, Pillay D, de Oliveira T, Clumeck N, Hunt G, Kerschberger B, Shafer RW, Yang C, Raizes E, Kantor R, Gupta RK. 2017. Occult HIV-1 drug resistance to thymidine analogues following failure of first-line tenofovir combined with a cytosine analogue and nevirapine or efavirenz in sub Saharan Africa: a retrospective multicentre cohort study. Lancet Infect Dis 17:296–304. https://doi.org/10.1016/S1473-3099(16)30469-8.

47. WHO. 2017. HIV drug resistance report. 2017. World Health Organization, Geneva, Switzerland.

48. Stefic K, Bouvin-Pley M, Braibant M, Barin F. 2019. Impact of HIV-1 diversity on its sensitivity to neutralization. Vaccines (Basel) 7:74. https://doi.org/10.3390/vaccines7030074.

49. Pancera M, Zhou T, Druz A, Georgiev IS, Soto C, Gorman J, Huang J, Acharya P, Chuang G-Y, Ofek G, Stewart-Jones GBE, Stuckey J, Bailer RT, Joyce MG, Louder MK, Tumba N, Yang Y, Zhang B, Cohen MS, Haynes BF, Mascola JR, Morris L, Munro JB, Blanchard SC, Mothes W, Connors M, Kwong PD. 2014. Structure and immune recognition of trimeric pre-fusion HIV-1 Env. Nature 514:455–461. https://doi.org/10.1038/nature13808.

50. Yu F, Wen Y, Wang J, Gong Y, Feng K, Ye R, Jiang Y, Zhao Q, Pan P, Wu H, Duan S, Su B, Qiu M. 2018. The transmission and evolution of HIV-1

quasispecies within one couple: a follow-up study based on next-generation sequencing. Sci Rep 8:1404. https://doi.org/10.1038/s41598-018-19783-3.

51. Pang J, Slyker JA, Roy S, Bryant J, Atkinson C, Cudini J, Farquhar C, Griffiths P, Kiarie J, Morfopoulou S, Roxby AC, Tutil H, Williams R, Gantt S, Goldstein RA, Breuer J. 2020. Mixed cytomegalovirus genotypes in HIV-positive mothers show compartmentalization and distinct patterns of transmission to infants. Elife 9:e63199. https://doi.org/10.7554/eLife.63199.

52. Boshier FAT, Pang J, Penner J, Hughes J, Parker M, Shepherd J, Alders N, Bamford A, Grandjean L, Grunewald S, Hatcher J, Best T, Dalton C, Bynoe PD, Frauenfelder C, Köeglmeier J, Myerson P, Roy S, Williams R, Thomson EC, de Silva TI, Goldstein RA, Breuer J, The COVID-19 Genomics UK (COG-UK) Consortium. 2020. Remdesivir induced viral RNA and subgenomic RNA suppression, and evolution of viral variants in SARS-CoV-2 infected patients. medRxiv https://doi.org/10.1101/2020.11.18.20230599.

53. Li JZ, Gallien S, Ribaudo H, Heisey A, Bangsberg DR, Kuritzkes DR. 2014. Incomplete adherence to antiretroviral therapy is associated with higher levels of residual HIV-1 viremia. AIDS 28:181–186. https://doi.org/10.1097/QAD.0000000000000123.

54. Parry CM, Kolli M, Myers RE, Cane PA, Schiffer C, Pillay D. 2011. Three residues in HIV-1 matrix contribute to protease inhibitor susceptibility and replication capacity. Antimicrob Agents Chemother 55:1106–1113. https://doi.org/10.1128/AAC.01228-10.

55. Neher RA, Leitner T. 2010. Recombination rate and selection strength in HIV intra-patient evolution. PLoS Comput Biol 6:e1000660. https://doi.org/10.1371/journal.pcbi.1000660.

56. El Bouzidi K, Kemp SA, Datir RP, Murtala-Ibrahim F, Aliyu A, Kwaghe V, Frampton D, Roy S, Breuer J, Sabin CA, Ogbanufe O, Charurat ME, Bonsall D, Golubchik T, Fraser C, Dakum P, Ndembi N, Gupta RK. 2020. High prevalence of integrase mutation L74I in West African HIV-1 subtypes prior to integrase inhibitor treatment. J Antimicrob Chemother 75:1575–1579. https://doi.org/10.1093/jac/dkaa033.

57. Derache A, Iwuji CC, Danaviah S, Giandhari J, Marcelin A-G, Calvez V, de Oliveira T, Dabis F, Pillay D, Gupta RK. 2018. Predicted antiviral activity of tenofovir versus abacavir in combination with a cytosine analogue and the integrase inhibitor dolutegravir in HIV-1-infected South African patients initiating or failing first-line ART. J Antimicrob Chemother 74:473–479. https://doi.org/10.1093/jac/dky428.

58. Siedner MJ, Moorhouse MA, Simmons B, de Oliveira T, Lessells R, Giandhari J, Kemp SA, Chimukangara B, Akpomiemie G, Serenata CM, Venter WDF, Hill A, Gupta RK. 2020. Reduced efficacy of HIV-1 integrase inhibitors in patients with drug resistance mutations in reverse transcriptase. Nat Commun 11:5922. https://doi.org/10.1038/s41467-020-19801-x.

59. Iwuji C, McGrath N, Calmy A, Dabis F, Pillay D, Newell M-L, Baisley K, Porter K. 2018. Universal test and treat is not associated with sub-optimal antiretroviral therapy adherence in rural South Africa: the ANRS 12249 TasP trial. J Int AIDS Soc 21:e25112. https://doi.org/10.1002/jia2.25112.

60. Derache A, Iwuji CC, Baisley K, Danaviah S, Marcelin A-G, Calvez V, de Oliveira T, Dabis F, Porter K, Pillay D. 2019. Impact of next-generation sequencing defined human immunodeficiency virus pretreatment drug resistance on virological outcomes in the ANRS 12249 Treatment-as-Prevention Trial. Clin Infect Dis 69:207–214. https://doi.org/10.1093/cid/ciy881.

61. Gall A, Ferns B, Morris C, Watson S, Cotten M, Robinson M, Berry N, Pillay D, Kellam P. 2012. Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. J Clin Microbiol 50:3838–3844. https://doi.org/10.1128/JCM.01516-12.

62. Martin MJE. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17:10–12. https://doi.org/10.14806/ej.17.1.200.

63. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

64. Shafer RW. 2006. Rationale and uses of a public HIV drug-resistance database. J Infect Dis 194:S51–58. https://doi.org/10.1086/505356.

65. Okonechnikov K, Conesa A, Garcia-Alcalde F. 2016. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics 32:292–294. https://doi.org/10.1093/bioinformatics/btv566.

66. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22:568–576. https://doi.org/10.1101/gr.129684.111.

67. Charles OJ, Venturini C, Breuer J. 2020. cmvdrg: an R package for human cytomegalovirus antiviral drug resistance genotyping. bioRxiv https://doi.org/10.1101/2020.05.15.097907.

68. Perrier M, Désiré N, Storto A, Todesco E, Rodriguez C, Bertine M, Le Hingrat Q, Visseaux B, Calvez V, Descamps D, Marcelin A-G, Charpentier C. 2018. Evaluation of different analysis pipelines for the detection of HIV-1 minority resistant variants. PLoS One 13:e0198334. https://doi.org/10.1371/journal.pone.0198334.

69. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2019. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530–1534. https://doi.org/10.1093/molbev/msaa015.

70. Minh BQ, Nguyen MA, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol 30:1188–1195. https://doi.org/10.1093/molbev/mst024.

71. Knyazev S, Tsyvina V, Melnyk A, Artyomenko A, Malygina T, Porozov YP, Campbell E, Switzer WM, Skums P, Zelikovsky A. 2018. Cliquesnv: scalable reconstruction of intra-host viral populations from NGS reads. bioRxiv https://doi.org/10.1093/nar/gkab576.

72. Charles OJ, Roberts J, Breuer J, Goldstein RA. 2021. WeightedLD: the application of sequence weights to linkage disequilibrium. bioRxiv https://doi.org/10.1101/2021.06.04.447093.

73. Martin DP, Varsani A, Roumagnac P, Botha G, Maslamoney S, Schwab T, Kelz Z, Kumar V, Murrell B. 2021. RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. Virus Evol 7:veaa087. https://doi.org/10.1093/ve/veaa087.