HIV-1 evolutionary dynamics under non-suppressive antiretroviral therapy.

Steven A Kemp[1,3], Oscar Charles[1], Anne Derache[2], Collins Iwuji[2,5], John Adamson[2], Katya Govender[2], Tulio de Oliveira[2,6], , Nonhlanhla Okesola[2], Francois Dabis[7,8], on behalf of the French National Agency for AIDS and Viral Hepatitis Research (ANRS) 12249 Treatment as Prevention (TasP) Study Group, Deenan Pillay[1], Darren P Martin[9], Richard A. Goldstein[1] & Ravindra K Gupta[2,3]

1. Division of Infection & Immunity, University College London, London, UK
2. Africa Health Research Institute, Durban, South Africa
3. Cambridge Institute of Therapeutic Immunology & Infectious Disease (CITIID), Cambridge, UK
4. Department of Medicine, University of Cambridge, Cambridge, UK
5. Research Department of Infection and Population Health, University College London, United Kingdom.
6. KRISP - KwaZulu-Natal Research and Innovation Sequencing Platform, UKZN, Durban, South Africa.
7. INSERM U1219-Centre Inserm Bordeaux Population Health, Université de Bordeaux, France.
8. Université de Bordeaux, ISPED, Centre INSERM U1219-Bordeaux Population Health, France.
9. Department of Integrative Biomedical Sciences, University of Cape Town, South Africa

Address for correspondence:
Ravindra K. Gupta
Cambridge Institute for Therapeutic Immunology and Infectious Diseases
Jeffrey Cheah Biomedical Centre
Puddicombe Way
Cambridge CB2 0AW, UK
Rkg20@cam.ac.uk

Or

Richard Goldstein
Division of Infection and Immunity
UCL
London WC1E 6BT

Abstract

Background: Viral population dynamics in long term viraemic antiretroviral therapy (ART) treated individuals have not been well characterised. Prolonged virologic failure on 2nd-line protease inhibitor (PI) based ART without emergence of major protease mutations is well recognised, providing an opportunity to study within-host evolution.

Methods: Using next-generation Illumina short read sequencing and in silico haplotype reconstruction we analysed whole genome sequences from longitudinal plasma samples of eight chronically infected HIV-1 individuals failing 2nd-line regimens from the ANRS 12249 TasP trial, in the absence of high frequency major PI resistance mutations. Plasma drug levels were measured by HPLC. Three participants were selective for in-depth variant and haplotype analyses, each with five or more timepoints spanning at least 16 months.

Results: During PI failure synonymous mutations were around twice as frequent as non-synonymous mutations across participants. Prior to or during exposure to PI, we observed several polymorphic amino acids in *gag* (e.g. T81A, T375N) which are have also been previously associated with exposure to protease inhibitor exposure. Although overall SNP frequency at abundance above 2% appeared stable across time in each individual, divergence from the consensus baseline sequence did increase over time. Non-synonymous changes were enriched in known polymorphic regions such as *env* whereas synonymous changes were more often observed to fluctuate in the conserved *pol* gene. Phylogenetic analyses of whole genome viral haplotypes demonstrated two common features: Firstly, evidence for selective sweeps following therapy switches or large changes in plasma drug concentrations, with hitchhiking of synonymous and non-synonymous mutations. Secondly, we observed competition between multiple viral haplotypes that intermingled phylogenetically alongside soft selective sweeps. The diversity of viral populations was maintained between successive timepoints 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.

Conclusions: Large fluctuations in variant frequencies with diversification occur during apparently 'stable' viremia on non-suppressive ART. Reconstructed haplotypes provided further evidence for sweeps during periods of partial adherence, and competition between haplotypes during periods of low drug exposure. Drug resistance mutations in RT can be used as markers of viral populations in the reservoir and we found evidence for loss of linkage disequilibrium for drug resistance mutations,

73 indicative of recombination. These data imply that even years of exposure to PIs, within the context
74 of large stable populations displaying ongoing selective competition, may not precipitate emergence
75 of major PI resistance mutations, indicating significant fitness costs for such mutations. Ongoing viral
76 diversification within reservoirs may compromise the goal of sustained viral suppression.
77

Introduction

Even though HIV-1 infections are most commonly initiated with a single founder virus[1], acute and chronic disease are characterised by extensive inter- and intra-participant 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, and penetration 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 co-exist as quasispecies. Existing literature on HIV-1 intrahost population dynamics is largely limited to untreated infections, predominantly in subtype B infected individuals[2,7-9]. These works have shown non-linear diversification of virus away from the founder strain during chronic untreated infection. Synonymous mutations can be the product of neutral evolution, but they are also expected to occur more commonly in established chronic infection due to higher fitness costs associated with non-synonymous mutations[10], with the exception of immune escape during early infection[2].

Viral population dynamics in long-term viraemic antiretroviral therapy (ART) treated individuals have not been well characterised. HIV rapidly accumulates drug-resistance associated mutations (DRMs), particularly during non-suppressive first line ART[5,11]. As a result, ART-experienced participants failing 1st-line regimens for prolonged periods of time particularly in low- and middle-income countries (LMICs), are characterised by high frequencies of common DRMs such as M184V, K65R and K103N[12]. By contrast, PI resistance mutations following failure of current boosted PI treatments are uncommon in LMIC[13], a situation that differs for less potent drugs used in the early PI era[5]. A number of studies have indicated that less well characterised mutations accumulating in the *gag* gene during PI failure might also impact PI susceptibility[14-20], though common pathways have been difficult to discern, likely reflecting plasticity to drug escape.

Prolonged virologic failure on PI regimens without emergence of protease mutations 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[21,22] and it was reported that major PI DRMs and other non-synonymous mutations in regulatory regions such as *pol,* significantly lower fitness[2,23,24]. However, this has been typically shown outside of the context of longitudinal sampling. By sampling participants consistently over several years, we propose that ongoing evolution is driven by the dynamic flux between selection, recombination, and genetic drift.

We have deployed next-generation sequencing of stored blood plasma specimens from participants in the Treatment as Prevention (TasP) ANRS 12249 study[25], conducted in Kwazulu-Natal, South Africa. All participants were infected with HIV-1 subtype C and characterised as failing 2nd-line regimens containing Lopinavir and Ritonavir (LPV/r), with prolonged virological failure in the absence of major PI mutations[26]. In this paper, we report details of evolutionary dynamics during non-suppressive ART.

Results

Participant Characteristics

Eight participants with virological failure of 2nd-line PI based ART and at least two timepoints with viraemia above 1000 copies/ml were selected from the French ANRS TasP trial for viral dynamic analysis, including whole genome haplotye reconstruction (**Table 1**). Prior to participation in the TasP trial, participants were accessing 1st-line regimens for an average of 5.6yrs (±2.7yrs). At baseline enrolment visit into TasP (whilst failing 1st-line regimens), median viral load was $4.96x10^{10}$ (IQR: $4.17x10^{10}$ – $5.15x10^{10}$); 12 DRMs were found at a threshold of >2%; the most common of which were RT mutations K103N, M184V and P225H, consistent with previous use of d4T, NVP, EFV and FTC/3TC. Six of the eight participants carried drug resistance mutations (DRMs) associated with PI failure at minority frequencies (average 6.4%) and at usually at only single timepoints throughout the longitudinal sampling. Observed mutations included L23I, I47V, M46I/L, G73S, V82A, N83D and I85V (**Supplementary Tables 1a-3c**). Four of the six participants also carried major integrase stand inhibitor (INSTI) mutations, again at minority frequencies (average 5.0%) and again normally only at single timepoints (T97A, E138K, Y143H, Q148K).

SNP frequencies and measures of diversity/divergence over time

We used whole-genome sequencing (WGS) data to measure the frequencies of viral single nucleotide polymorphisms (SNPs) relative to a dual-tropic subtype C reference sequence (AF411967) within individuals over time using short-read deep-sequence data (**Figure 1**). We observed that SNPs resulting from synonymous changes roughly mirrored those from non-synonymous changes, but the former were two-to-three-fold more common. Furthermore, when we considered diversity relative to the consensus sequence at baseline (switch from 1st- to 2nd-line regimens), we observed a moderate decrease in diversity in four of the eight participants at the second timepoint, moderate increase in diversity in one participant and relatively little diversity in the remaining three (**Figure 1, Supplementary Figure 1**). From timepoint two onwards (all participants now on 2nd-line regimens for >6 months), three participants saw increases in both synonymous and non-synonymous changes, two participants saw decreases in those changes and three participants did not differ significantly. Finally,

HIV-1 subtype M group consensus sequences (Alignment ID: 102CG1) were downloaded from the Los Alamos National Laboratory (LANL) and a merged consensus sequence was used as a proxy of a distant ancestor for all participants. Using this ancestral sequence we measured sequence divergence over sequential timepoints (**Figure 2**) to determine whether under non-suppressive ART, viral populations were reverting towards an ancestral state, or diverging away from the M-group consensus. All but two participants (15664 and 29447) showed a trend towards divergence away from the ancestor. This is indicative of potential ongoing selection or could be due to random genetic drift. This finding contrasts with studies in transmission where groups have demonstrated that transmitted viruses in the 'recipient' are closer to consensus as compared to the majority virus in the transmitting 'donor'[27].

To further investigate nucleotide diversity, we first considered all pairwise nucleotide distances of each consensus WGS by timepoint and participant using a multidimensional scaling approach[28]. Intra-participant nucleotide diversity varied considerably between participants (**Figure 3A-B**). Some participants showed little diversity between timepoints (for example participant 16207), whereas others showed higher diversity between timepoints (e.g. participant 22763). A similar range of diversity profiles between participants were inferred by multidimensional scaling. Some participants were tightly clustered, suggesting little change over time (**Figure 3a** participants 26892, 47939 & 16207), compared to others (participants 22828 & 28545).

When examining the *gag*, *pol* and *env* genes independently, the average diversity over all timepoints (for participants 15664, 16207 and 22763) suggested that the highest diversity (π) was in the *env* gene. Watterson's genetic diversity (θ) was highest in the *env* gene of participants 15664 and 16207. By contrast, θ was highest in the *gag* gene for participant 22763 (**Table 2**).

Phylogenetic analysis of inferred haplotypes

The preceding diversity assessments suggested the existence of distinct viral haplotypes within each participant. We therefore used a recently reported computational method[29] to infer 166 unique haplotypes across all participants, with between 11 and 32 haplotypes (average 21) per participant (**Figure 3C**). The number haplotypes changed dynamically between successive timepoints indicative of dynamically shifting populations. To ensure that haplotypes were sensibly reconstructed, a phylogeny of all consensus sequences was also inferred (**Supplementary Figure 2**). Three participants were selected for deeper analysis based on these participants having the highest number of longitudinal plasma samples. These included participants 15664, 16207 and 22763 who each had whole-genome short-read deep-sequence data at more than five timepoints.

Intra-host evolutionary dynamics and relation to drug levels

Changing landscapes of non-synonymous and synonymous mutations

In the absence of major PI mutations, we first examined non-synonymous mutations across the whole genome (**Figures 4-6**), with a specific focus on *pol* (to observe first and second line NRTI-associated mutations) and *gag* (given its involvement in PI susceptibility). We and others have previously shown that *gag* mutations accumulate during non-suppressive PI therapy[30,31]. There are also data suggesting associations between *env* mutations and PI exposure[32,33]. **Supplementary Tables 1-3** summarise the changes in variant frequencies of *gag*, *pol* and *env* mutations in participants over time. We found between two and four mutations at sites previously associated with PI resistance in each participant, 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[18,34]. 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 timepoints (**Figures 4c & 5b**), indicative of linked alleles dynamically shifting and suggestive of competition between viral haplotypes.

Participant 15664 had consistently low drug plasma concentration of all drugs at each measured timepoint, with detectable levels measured only at month 15 and beyond (**Figure 4a**). At baseline, whilst 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 whilst 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 none of the RT mutations (**Figure 4d**, orange circles). Following the switch to a 2nd-line regimen, variant frequencies of M184V and P225H dropped below detection limits (<2% of reads), whilst K103N remained at high frequency (**Figure 4b**). Haplotype analysis was concordant, revealing that viruses with K103N, M184V and P225H were replaced by haplotypes with only K103N (**Figure 4d**, large brown circles), with minority haplotypes that had no RT mutations (**Figure 4d**, small brown circles). At timepoint two (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 timepoint three (15 months post-switch to 2nd-line regimen) drug concentrations were highest, though still low in absolute terms, indicating partial adherence. Between timepoints three and four we observed a two-log reduction in viral load, with 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 (**Figure 4b**). However many of the changes were between intermediate frequencies, (e.g. between 20% and 60%), which differed from changes between time points one and two where multiple variants changed more dramatically in frequency from <5% to more than 80%, indicating harder selective sweep. These data are in keeping with a soft selective sweep between time points 3 and 5. Between time points five and six, the final two samples, there was another population shift - M184V and P225H frequencies fell below the detection limit at timepoint six, whereas the frequency of K103N dropped from almost 100% to around 75% (**Figure 4b**). This was consistent with haplotype reconstruction, which inferred a dominant viral haplotype at timepoint six bearing only K103N, as well as a minor haplotype with no DRMs at all (**Figure 2d**, pink circles). The inferred haplotype without DRMs was nonetheless phylogenetically distinct from the timepoint one minority haplotype (**Figure 4d**, compare small orange and pink circles in upper clade).

Upon examining the phylogenetic relationships of the inferred haplotype sequences, there were two distinct clades with members of both clades present at all time points apart from the first. This suggests an intermingling of viral haplotypes and competition. DRMs showed some segregation by clade; viruses carrying a higher frequency of DRMs were observed in the lower clade (Clade B, **Figure 4d**), and those with either K103N alone, or no DRMs were preferentially located in the upper clade (Clade A, **Figure 4d**). However, this relationship was not clear cut, and therefore consistent with competition between haplotypes during low drug exposure. Soft sweeps were evident, given the increasing diversity (**Figure 1**) of this participant, as well as constrained variant frequencies between 20-80% (**Figure 4b,c**).

Participant 16207. Viral load in this participant were consistently elevated >10,000 copies/ml (**Figure 5a)**. As with participant 15664, drug concentrations in blood plasma remained extremely low or absent at each measured timepoint, consistent with non-adherence to the prescribed regimen. There was almost no 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[24] and can therefore persist in the absence of NNRTI pressure. Throughout treatment the participant maintained K103N at a frequency of >95% but also carried several integrase strand transfer inhibitor (INSTI) associated changes (E157Q) and PI-exposure associated amino acid replacements (L23I and M46I) at low frequencies at timepoints two and three. Despite little change in DRM site frequencies, very significant viral population shifts were observed at the whole genome level, again

indicative of selective sweeps (**Figures 5b-c**). Between timepoints one and four, several linked mutations changed abundance contemporaneously, generating chevron-like patterns of non-synonymous changes in *env* specifically (blue lines). A large number of alleles increased in frequency from <20% to >80% at the same time as numerous others decreased in frequency from above 80% to below 20%. Whereas large shifts in gag and pol alleles also occurred, the mutations involved were almost exclusively synonymous (red and green lines). These putative selective sweeps in *env* were evident in the phylogenetic analysis (**Figure 5d**, see long branch lengths between timepoints one and four, and cladal structure) possibly driven by neutralising antibodies and/or T-cell immune pressures.

Phylogenetic analysis of inferred whole genome haplotypes overall showed a distinct cladal structure as observed in 15664 (**Figure 5d**), although the dominant haplotypes were mostly observed in the lower clade (Clade B, **Figure 5d**). K103N, the only major DRM, was inferred in both clades A and B. Haplotypes did not cluster by time point. Significant diversity in haplotypes from this participant was confirmed by MDS (**Supplementary Figure 3**).

Participant 22763 was notable for a number of large shifts in variant frequencies across multiple drug resistance associated residues and synonymous sites. Drug plasma concentration for different drugs was variable yet detectable at most measured timepoints reflecting changing levels of adherence across the treatment period (**Figure 6a).** 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 one and two. Most of the highly variable synonymous changes in this participant were found in the *gag* and *pol* genes (as in participant 16207) (**Figure 6c**), but in this case *env* displayed large fluctuations in synonymous and non-synonymous allelic frequencies over time. At timepoint three therapeutic concentrations of LPV/r and TDF were measured in plasma and haplotypes clustered separately from the first two timepoints (**Figure 6d**, green circles). NGS confirmed that the D67N, K219Q, K65R, L70R, M184V DRMs and NNRTI-resistance mutations were present at low frequencies from timepoint three onwards. Of note, between timepoints three and six, therapeutic concentration of tenofovir (TDF) was 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, whilst the lamivudine (3TC) associated resistance mutation, M184V, persisted throughout. In the final three timepoints M46I emerged in *protease*, but never increased in frequency above <6%. At timepoint seven, populations shifted again with some haplotypes resembling those previously timepoint four, with D67N and K70R again being predominant over K65R in *reverse transcriptase* (**Figure 6d**, green and blue circles). At the

final timepoint (eight) the frequency of K103N was approximately 85% and the TAM-bearing populations continued to dominate over the K65R population, which at this timepoint had a low frequency.

Although the DRM profile suggested the possibility of a selective sweep, we did not observe groups of other non-synonymous or synonymous alleles exhibiting dramatic frequency shifts, i.e. we did not observe the same 'chevron patterns' in synonymous- and non-synonymous variant plots (**Figure 6b-c**) as those seen in participants 15664 and 16207. Indeed, haplotypes were spread throughout the phylogenetic tree, consistent with low drug pressure. 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[35-37]. One possible explanation for the disconnect between the trajectories of DRM frequencies over time and haplotype phylogeny is recombination. Alternatively, emergence of haplotypes from previously unsampled reservoir with different DRM profile is possible, but one might have expected other mutations to characterise such haplotypes that would manifest as change in frequencies of large numbers of other mutations.

Discussion

The proportion of people living with HIV (PLWH) accessing ART has increased from 24% in 2010, to 68% in 2020[38,39]. However, with the scale-up of ART, there has also been an increase in both pre-treatment drug resistance (PDR)[40,41] and acquired drug resistance[12,42] 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%[43]. Boosted PI-containing regimens remain second line drugs following first 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 PI, have not emerged. Here we have comprehensively analysed viral populations present in longitudinally collected plasma samples of chronically-infected HIV-1 participants under non-suppressive 2nd-line ART.

With the vast majority of PLWH treated in the post-ART era, virus dynamics during non-suppressive ART is important to understand, as there may be implications for future therapeutic success. For example broadly neutralising 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 bNab is dependent on *env* diversity[44,45], 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[46], with largely subgenomic data analysed rather than whole genome[9]. Traditional analysis of quasispecies distribution, for example as reported by Yu et al[47], suggests that the 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 genome is that linked mutations can be identified statistically using an approach that we gave recently developed[29]. Indeed haplotype reconstruction has proved beneficial in the analysis of compartmentalisation and diversification of several RNA viruses, including HIV-1, CMV and SARS-CoV-2[30,48,49].

Key findings of this study were: firstly that diversity increased over time with variable trajectory away from the consensus baseline sequence and also the reconstructed ancestral subtype M consensus. Approximately half of the participants appeared to diversify away from the reconstructed ancestral subtype M sequence, but interestingly three participants showed possible diversification back towards the ancestral consensus (albeit with insufficient statistical support).

Secondly, and in contrast to the fractions of synonymous and non-synonymous mutations reported by Zanini et al in a longitudinal untreated dataset[2], we show that the fractions of synonymous mutations are generally two-to-three fold higher than non-synonymous mutations during non-suppressive ART in chronic infection. This finding may reflect early versus chronic infection and differing selective pressures. Haplotype reconstruction revealed evidence for competing haplotypes, with evidence for numerous soft selective sweeps in phylogenies, evidenced by intermingling of haplotypes during periods where there was low drug concentration measured in participant's blood plasma.

Individuals in the present study were treated with Ritonavir boosted Lopinavir along with two NRTIs (typically Tenofovir + Emtricitabine). We observed significant change in the frequencies of NRTI mutations in two of the three participants studied in-depth. These fluctuations likely reflected adherence to the 2nd-line regimen though we saw evidence for possible archived virus populations with DRMs emerging during follow-up because large changes in DRM frequency were not always accompanied by changes at other sites. This is consistent with soft sweeps occurring and that non-

DRMs do not necessarily drift with other mutations to fixation[21] and that the same mutations are occurring on different backgrounds. As frequencies of RT DRMs did not always segregate with haplotype frequencies, we suggest that a high number of recombination events, known to be common in HIV infections, was responsible for the haplotypic diversity.

Although no participant developed major DRMS at consistently high frequencies to PIs (https://hivdb.stanford.edu/dr-summary/resistance-notes/PI/), we did observe non-synonymous mutations associated with PI exposure that are also known to be polymorphic; however, there was no temporal evidence of specific changes being associated with selective sweeps. For example PI exposure associated residues in matrix (positions 76 and 81) were observed in participant 16207 prior to PI initiation[50]. Furthermore, participant 16207 was one of few participants who achieved two partial suppressions (<750 copies/ml). After both of these partial suppressions, the rebound populations appeared to be less diverse, consistent with drug-resistant virus re-emerging.

Mutations in all genes that are further apart than 100bp are subject to shuffling via recombination[51]. Indeed, particularly in participant 22763, there is evidence of recombination between timepoint corresponding to months 10 and 27. Whilst we did not undertake a formal recombination analysis at this time, in the future insights into the relationship between DRMs and the dissociation from synonymous changes, relative to genomic regions that have been transferred by recombination, may lead to additional insights of viral populations pressures under non-suppressive ART.

This study had some limitations – we examined in-detail three participants with ongoing viraemia and variable adherence to 2nd-line drug regimens. Despite the small sample size, this type of longitudinal sampling of ART-experienced participants in unprecedented. We are confident that the combination of computational analyses has provided a detailed understanding of viral dynamics under non-suppressive ART and is directly applicable to wider datasets. The method used to reconstruct viral haplotypes *in silico* is novel and has previously been validated in HIV-positive participants with CMV[48]. Validation of haplotypes is based on inputting a known mix of artificial viruses and comparing the output to the known input, as is common with such software. We are confident that 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 analysed timepoints, nuances of the sequencing method led in some cases to suboptimal degrees of gene coverage (**Supplementary Figure 4**). Mapping metrics showed that gaps in sequence coverage were random, which excludes the

possibility of systematic bias in the participant data. To ensure that uneven sequencing coverage did not bias our analyses, we ensured that variant analysis was only performed where coverage was >10 reads.

In summary we have found compelling evidence of HIV-1 within host viral diversification and haplotype competition during non-suppressive ART, in addition to recombination. Going forward, participants 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[52,53], this approach needs assessment given data linking individuals with NNRTI resistance with poorer virological outcomes on Dolutegravir[54], coupled with a history of intermittent adherence. The increases in diversity generated during long term PI failure may have biological effects that renders durable suppression less likely.

Methods

Study & Participant selection

This cohort was nested within the French ANRS 12249 Treatment as Prevention (TasP) trial[25]. TasP was a cluster-randomised trial comparing an intervention arm who offered ART after HIV diagnosis irrespective of participant CD4 + count, to a control arm which offered ART according to prevailing South African guidelines. A subset of 44 longitudinal samples from 8 chronically infected participants. Participants were selected for examination if there were >3 timepoint samples available. All samples were collected from blood plasma. The Illumina MiSeq platform was used and an adapted protocol for sequencing[55]. Adherence to 2nd-line regimens was measured by HPLC using plasma concentration of drug levels as a proxy. Drug levels were measured at each timepoint with detectable viral loads, post-PI initiation.

Ethical approval was originally grant 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 participants. 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[56] using a modified protocol previously described by Gall et al[57]. Briefly, RNA was extracted from 1ml of plasma with detectable viral load of >1000 copies/ml, using QIAamp Viral RNA mini kits (Qiagen, Hilden, Germany), and eluted in 60µl of elution buffer. The near-full HIV genome was amplified with 4 subtype C primers pairs, generating 4 overlapping amplicons of between 2100 and 3900kb.

DNA concentrations of amplicons were quantified with the Qubit dsDNA HS Assay kit (Invitrogen, Carlsbad, CA). Diluted amplicons were pooled equimolarly and prepared for 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 & Bioinformatics

Poor quality reads (with Phred score <30) and adapter sequences were trimmed from FastQ files with TrimGalore! v0.6.519[58] and mapped to a clade C South African reference genome (AF411967.1) with BWA-MEM[59]. The reference genome was manually annotated in Geneious Prime v2020.3 with DRMs according to the Stanford HivDB[60]. Optical PCR duplicate reads were removed using Picard tools (http://broadinstitute.github.io/picard). Finally, QualiMap2[61] 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 nucleotides polymorphisms (SNPs) were called using VarScan2[62] with a minimum average quality of 20, minimum variant frequency of 2% and in at least 10 reads. These were then annotated by gene, codon and amino acid alterations using an in-house script[63] modified to utilise HIV genomes.

All synonymous variants and DRMs were examined, and their frequency compared across successive timepoints. Synonymous variants were excluded from analysis if their prevalence remained at ≤10% or ≥90% across all timepoints. 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 cut-off[64].

Haplotype Reconstruction & Phylogenetics

Whole-genome viral haplotypes were constructed for each participant timepoint using Haplotype Reconstruction for Longitudinal Samples (HaROLD)[65]. Briefly, SNPs were assigned to each haplotype such that the frequency of variants was equal to the sum of the frequencies of haplotypes containing a specific variant. Maximal log likelihood was used to optimise time-dependent frequencies for

longitudinal haplotypes which was calculated by summing over all possible assignment of haplotype variants. Haplotypes were then constructed based on posterior probabilities. After constructing haplotypes, a refinement process remapped reads from BAM files to those constructed haplotypes. The number of haplotypes either increased or decreased as a result of combination or division according to AIC scores, in order to present the most accurate representation of viral populations at each timepoint.

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 optimised by maximum log likelihood. Individual gene nucleotide diversity (π), number of segregating sites (S) and Watterson's genetic diversity (θ) were calculated separately using all reconstructed haplotypes per participant, using DnaSP v6.12.03[66].

Maximum-likelihood phylogenetic trees and ancestral reconstruction were performed using IQTree2 v2.1.2[67]. Evolutionary model selection for trees were inferred using ModelFinder[68] and trees were estimated using the GTR+F+I model with 1000 ultrafast bootstrap replicates[69]. All trees were visualised with Figtree v.1.4.4 (http://tree. bio.ed.ac.uk/software/figtree/), rooted on the AF411967.1 reference sequence, and nodes arranged in descending order. Phylogenies were manipulated and annotated using ggtrree v2.24.

Multi Dimension Scaling (MDS) Plots

Whole genomes were obtained as previously described. Pairwise distances between these consensus sequences were calculated using the dist.dna() package in R, with a TN93 nucleotide-nucleotide substitution matrix[70] and with pairwise deletion by way of the R package Ape v.5.4[71]. Multi-dimensional scaling (MDS) was implemented using the cmdscale() function with pairwise deletion in R v4.0.4. Much like PCA is a method to attempt to simplify complex data into a more interpretable format, by reducing dimensionality of data whilst retaining most of the variation. In a genomics context we can use this 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.

Funding

SAK is supported by the Bill and Melinda Gates Foundation: OPP1175094. RKG is supported by Wellcome Trust Senior Fellowship in Clinical Science: WT108082AIA.

Transparency declarations

RKG has received ad hoc consulting fees from Gilead, ViiV and UMOVIS Lab.

References

1 Abrahams, M. R. *et al.* 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, doi:10.1128/JVI.02132-08 (2009).

2 Zanini, F., Puller, V., Brodin, J., Albert, J. & Neher, R. A. In vivo mutation rates and the landscape of fitness costs of HIV-1. *Virus Evol* **3**, vex003, doi:10.1093/ve/vex003 (2017).

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

4 Lemey, P., Rambaut, A. & Pybus, O. G. HIV evolutionary dynamics within and among hosts. *Aids Reviews* **8**, 125-140 (2006).

5 Collier, D. A., Monit, C. & Gupta, R. K. The Impact of HIV-1 Drug Escape on the Global Treatment Landscape. *Cell host & microbe* **26**, 48-60, doi:10.1016/j.chom.2019.06.010 (2019).

6 Biebricher, C. K. & Eigen, M. What is a quasispecies? *Curr Top Microbiol Immunol* **299**, 1-31, doi:10.1007/3-540-26397-7_1 (2006).

7 Lythgoe, K. A. & Fraser, C. New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels. *Proceedings of the Royal Society B: Biological Sciences* **279**, 3367-3375 (2012).

8 Hedskog, C. *et al.* Dynamics of HIV-1 quasispecies during antiviral treatment dissected using ultra-deep pyrosequencing. *PloS one* **5**, e11345 (2010).

9 Shankarappa, R. *et al.* Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *Journal of virology* **73**, 10489-10502 (1999).

10 Theys, K. *et al.* Within-patient mutation frequencies reveal fitness costs of CpG dinucleotides and drastic amino acid changes in HIV. *PLoS genetics* **14**, e1007420, doi:10.1371/journal.pgen.1007420 (2018).

11 Masikini, P. & Mpondo, B. C. HIV drug resistance mutations following poor adherence in HIV-infected patient: a case report. *Clin Case Rep* **3**, 353-356, doi:10.1002/ccr3.254 (2015).

12 TenoRes Study, 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, doi:10.1016/S1473-3099(15)00536-8 (2016).

13 Collier, D. *et al.* 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. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **64**, 1006-1016, doi:10.1093/cid/cix015 (2017).

14 Giandhari, J. *et al.* Genetic Changes in HIV-1 Gag-Protease Associated with Protease Inhibitor-Based Therapy Failure in Pediatric Patients. *AIDS Res Hum Retroviruses* **31**, 776-782, doi:10.1089/AID.2014.0349 (2015).

15 Kelly Pillay, S., Singh, U., Singh, A., Gordon, M. & Ndungu, T. Gag drug resistance mutations in HIV-1 subtype C patients, failing a protease inhibitor inclusive treatment regimen, with detectable lopinavir levels. *Journal of the International AIDS Society* **17**, 19784 (2014).

16 Sutherland, K. A. *et al.* Evidence for Reduced Drug Susceptibility without Emergence of Major Protease Mutations following Protease Inhibitor Monotherapy Failure in the SARA Trial. *PloS one* **10**, e0137834, doi:10.1371/journal.pone.0137834 (2015).

17 Sutherland, K. A. *et al.* Phenotypic characterization of virological failure following lopinavir/ritonavir monotherapy using full-length Gag-protease genes. *The Journal of antimicrobial chemotherapy* **69**, 3340-3348, doi:10.1093/jac/dku296 (2014).

18 Sutherland, K. A. *et al.* Gag-Protease Sequence Evolution Following Protease Inhibitor Monotherapy Treatment Failure in HIV-1 Viruses Circulating in East Africa. *AIDS research and human retroviruses* **31**, 1032-1037, doi:10.1089/aid.2015.0138 (2015).

19 Day, C. L. *et al.* Proliferative capacity of epitope-specific CD8 T-cell responses is inversely related to viral load in chronic human immunodeficiency virus type 1 infection. *Journal of virology* **81**, 434-438, doi:10.1128/JVI.01754-06 (2007).

20 Blanch-Lombarte, O. *et al.* HIV-1 Gag mutations alone are sufficient to reduce darunavir susceptibility during virological failure to boosted PI therapy. *The Journal of antimicrobial chemotherapy* **75**, 2535-2546, doi:10.1093/jac/dkaa228 (2020).

21 Feder, A. F. *et al.* More effective drugs lead to harder selective sweeps in the evolution of drug resistance in HIV-1. *Elife* **5**, e10670, doi:10.7554/eLife.10670 (2016).

22 Harris, R. B., Sackman, A. & Jensen, J. D. On the unfounded enthusiasm for soft selective sweeps II: Examining recent evidence from humans, flies, and viruses. *PLoS genetics* **14**, e1007859 (2018).

23 Dam, E. *et al.* Gag mutations strongly contribute to HIV-1 resistance to protease inhibitors in highly drug-experienced patients besides compensating for fitness loss. *PLoS pathogens* **5**, e1000345 (2009).

24 Cong, M.-e., Heneine, W. & García-Lerma, J. G. The fitness cost of mutations associated with human immunodeficiency virus type 1 drug resistance is modulated by mutational interactions. *Journal of virology* **81**, 3037-3041 (2007).

25 Iwuji, C. C. *et al.* 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, doi:10.1186/1745-6215-14-230 (2013).

26 Organization, W. H. *Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: recommendations for a public health approach*. (World Health Organization, 2016).

27 Carlson, J. M. *et al.* Selection bias at the heterosexual HIV-1 transmission bottleneck. *Science* **345**, 1254031, doi:10.1126/science.1254031 (2014).

28 Cox, M. A. & Cox, T. F. in *Handbook of data visualization* 315-347 (Springer, 2008).

29 Pang, J. *et al.* Haplotype assignment of longitudinal viral deep-sequencing data using co-variation of variant frequencies. *bioRxiv*, 444877, doi:10.1101/444877 (2020).

30 Datir, R. *et al.* In Vivo Emergence of a Novel Protease Inhibitor Resistance Signature in HIV-1 Matrix. *mBio* **11**, e02036-02020, doi:10.1128/mBio.02036-20 (2020).

31 Kletenkov, K. *et al.* Role of Gag mutations in PI resistance in the Swiss HIV cohort study: bystanders or contributors? *Journal of Antimicrobial Chemotherapy* **72**, 866-875, doi:10.1093/jac/dkw493 (2016).

32 Rabi, S. A. *et al.* Multi-step inhibition explains HIV-1 protease inhibitor pharmacodynamics and resistance. *The Journal of clinical investigation* **123**, 3848-3860, doi:10.1172/JCI67399 (2013).

33 Manasa, J. *et al.* Evolution of gag and gp41 in Patients Receiving Ritonavir-Boosted Protease Inhibitors. *Sci Rep* **7**, 11559, doi:10.1038/s41598-017-11893-8 (2017).

34 Datir, R., El Bouzidi, K., Dakum, P., Ndembi, N. & Gupta, R. K. Baseline PI susceptibility by HIV-1 Gag-protease phenotyping and subsequent virological suppression with PI-based

second-line ART in Nigeria. *The Journal of antimicrobial chemotherapy* **74**, 1402-1407, doi:10.1093/jac/dkz005 (2019).

35 Parikh, U. M., Zelina, S., Sluis-Cremer, N. & Mellors, J. W. Molecular mechanisms of bidirectional antagonism between K65R and thymidine analog mutations in HIV-1 reverse transcriptase. *Aids* **21**, 1405-1414 (2007).

36 Parikh, U. M., Bacheler, L., Koontz, D. & Mellors, J. W. The K65R mutation in human immunodeficiency virus type 1 reverse transcriptase exhibits bidirectional phenotypic antagonism with thymidine analog mutations. *Journal of virology* **80**, 4971-4977 (2006).

37 Parikh, U. M., Barnas, D. C., Faruki, H. & Mellors, J. W. Antagonism between the HIV-1 reverse-transcriptase mutation K65R and thymidine-analogue mutations at the genomic level. *The Journal of infectious diseases* **194**, 651-660 (2006).

38 Department of Health. 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, 2019).

39 UNAIDS. *Global HIV & AIDS statistics — 2020 fact sheet*, <https://www.unaids.org/en/resources/fact-sheet> (2020), Accessed 3rd March 2021.

40 Gupta, R. K. *et al.* 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, doi:10.1016/S1473-3099(17)30702-8 (2018).

41 Gupta, R. K. *et al.* 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, doi:10.1016/S0140-6736(12)61038-1 (2012).

42 Gregson, J. *et al.* 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 multi-centre cohort study. *Lancet Infect Dis*, doi:10.1016/S1473-3099(16)30469-8 (2017).

43 WHO, C. Global Fund. HIV drug resistance report. 2017. *World Health Organisation* (2017).

44 Stefic, K., Bouvin-Pley, M., Braibant, M. & Barin, F. Impact of HIV-1 diversity on its sensitivity to neutralization. *Vaccines* **7**, 74 (2019).

45 Pancera, M. *et al.* Structure and immune recognition of trimeric pre-fusion HIV-1 Env. *Nature* **514**, 455-461 (2014).

46 Shankarappa, R. *et al.* Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *Journal of virology* **73**, 10489-10502, doi:10.1128/JVI.73.12.10489-10502.1999 (1999).

47 Yu, F. J. *et al.* The Transmission and Evolution of HIV-1 Quasispecies within One Couple: a Follow-up Study based on Next-Generation Sequencing. *Scientific reports* **8**, 1-8, doi:ARTN 140410.1038/s41598-018-19783-3 (2018).

48 Pang, J. *et al.* Mixed cytomegalovirus genotypes in HIV-positive mothers show compartmentalization and distinct patterns of transmission to infants. *Elife* **9**, e63199, doi:10.7554/eLife.63199 (2020).

49 Boshier, F. A. T. *et al.* Remdesivir induced viral RNA and subgenomic RNA suppression, and evolution of viral variants in SARS-CoV-2 infected patients. *medRxiv*, 2020.2011.2018.20230599, doi:10.1101/2020.11.18.20230599 (2020).

50 Parry, C. M. *et al.* Three residues in HIV-1 matrix contribute to protease inhibitor susceptibility and replication capacity. *Antimicrobial agents and chemotherapy* **55**, 1106-1113, doi:10.1128/AAC.01228-10 (2011).

51 Neher, R. A. & Leitner, T. Recombination rate and selection strength in HIV intra-patient evolution. *PLoS Comput Biol* **6**, e1000660 (2010).

52 El Bouzidi, K. *et al.* High prevalence of integrase mutation L74I in West African HIV-1 subtypes prior to integrase inhibitor treatment. *J Antimicrob Chemother* **75**, 1575-1579, doi:10.1093/jac/dkaa033 (2020).

53 Derache, A. *et al.* 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. *The Journal of antimicrobial chemotherapy*, doi:10.1093/jac/dky428 (2018).

54 Siedner, M. J. *et al.* Reduced efficacy of HIV-1 integrase inhibitors in patients with drug resistance mutations in reverse transcriptase. *Nat Commun* **11**, 5922, doi:10.1038/s41467-020-19801-x (2020).

55 Iwuji, C. *et al.* 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, doi:10.1002/jia2.25112 (2018).

56 Derache, A. *et al.* Impact of Next-generation Sequencing Defined Human Immunodeficiency Virus Pretreatment Drug Resistance on Virological Outcomes in the ANRS 12249 Treatment-as-Prevention Trial. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **69**, 207-214, doi:10.1093/cid/ciy881 (2019).

57 Gall, A. *et al.* Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. *Journal of clinical microbiology* **50**, 3838-3844, doi:10.1128/JCM.01516-12 (2012).

58 Martin, M. J. E. j. Cutadapt removes adapter sequences from high-throughput sequencing reads. **17**, pp. 10-12 (2011).

59 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).

60 Shafer, R. W. Rationale and uses of a public HIV drug-resistance database. *The Journal of infectious diseases* **194 Suppl 1**, S51-58, doi:10.1086/505356 (2006).

61 Okonechnikov, K., Conesa, A. & Garcia-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics (Oxford, England)* **32**, 292-294, doi:10.1093/bioinformatics/btv566 (2016).

62 Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568-576, doi:10.1101/gr.129684.111 (2012).

63 Charles, O. J., Venturini, C. & Breuer, J. cmvdrg - An R package for Human Cytomegalovirus antiviral Drug Resistance Genotyping. *bioRxiv*, 2020.2005.2015.097907, doi:10.1101/2020.05.15.097907 (2020).

64 Perrier, M. *et al.* Evaluation of different analysis pipelines for the detection of HIV-1 minority resistant variants. *PloS one* **13**, e0198334, doi:10.1371/journal.pone.0198334 (2018).

65 Goldstein, R. A., Tamuri, A. U., Roy, S. & Breuer, J. Haplotype assignment of virus NGS data using co-variation of variant frequencies. *bioRxiv*, 444877 (2018).

66 Rozas, J. *et al.* DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Mol Biol Evol* **34**, 3299-3302, doi:10.1093/molbev/msx248 (2017).

67 Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *bioRxiv*, 849372, doi:10.1101/849372 (2019).

68 Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**, 587-589, doi:10.1038/nmeth.4285 (2017).

69 Minh, B. Q., Nguyen, M. A. & von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol* **30**, 1188-1195, doi:10.1093/molbev/mst024 (2013).

70 Tamura, K. & Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* **10**, 512-526, doi:10.1093/oxfordjournals.molbev.a040023 (1993).

71 Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics (Oxford, England)* **20**, 289-290, doi:10.1093/bioinformatics/btg412 (2004).

689 **Table 1.** Regimens and viral load at final timepoint for all participants. Participants initiated and
690 maintained 1st-line regimens for between 1-10 years before being switched to 2nd-line regimens as
691 part of the TasP trial. Seven of the eight participants were failing 2nd-line regimens at the final
692 timepoint.
693

Participant	No. of timepoints	1st-line regimen	Time since initiation of 1st-line treatment (yrs.)	2nd-line regimen	Viral Load at final timepoint (copies/ml)
15664	6	d4T, 3TC, FTC	6.2	LPV/r, TDF, FTC	28655
16207	5	d4T, 3TC, NVP	5.9	LPV/r, TDF, FTC	56660
22763	8	d4T, 3TC, EFV	6.2	LPV/r, TDF, 3TC	15017
22828	6	d4T, 3TC, NVP	6.4	LPV/r, TDF, 3TC/FTC	947
26892	7	d4T, 3TC, EFV	6	LPV/r, TDF, FTC	12221
28545	5	TDF, FTC, EFV	1.3	LPV/r, AZT, 3TC	12964
29447	4	TDF, FTC, EFV	2.8	LPV/r, TDF, FTC	64362
47939	3	d4T, 3TC, EFV	10.1	LPV/r, AZT, 3TC/FTC	6328

694 **NRTI:** Stavudine, d4T; Lamivudine, 3TC; Tenofovir, TDF; Emtricitabine, FTC; Zidovudine, AZT. **NNRTI:**
695 Efavirenz, EFV; Nevirapine, NVP. **PI:** Lopinavir/ritonavir, LPV/r
696
697
698

699 **Table 2.** Nucleotide diversity between *gag*, *pol* and *env* reconstructed haplotype sequences from
700 three participants.
701

Participant	Diversity			
	Gene	S	π	θ
15664	*gag*	72	0.01721	0.01321
15664	*pol*	150	0.01489	0.01349
15664	env	190	0.02840	0.02241
16207	*gag*	86	0.02018	0.01543
16207	*pol*	192	0.01575	0.01675
16207	env	246	0.03173	0.02733
22763	*gag*	160	0.02265	0.02608
22763	*pol*	292	0.01724	0.02323
22763	env	251	0.02498	0.02558

702 Diversity (**S** = number of segregating sites, **π** = nucleotide diversity, **θ** = Watterson genetic diversity)
703 rates
704
705
706

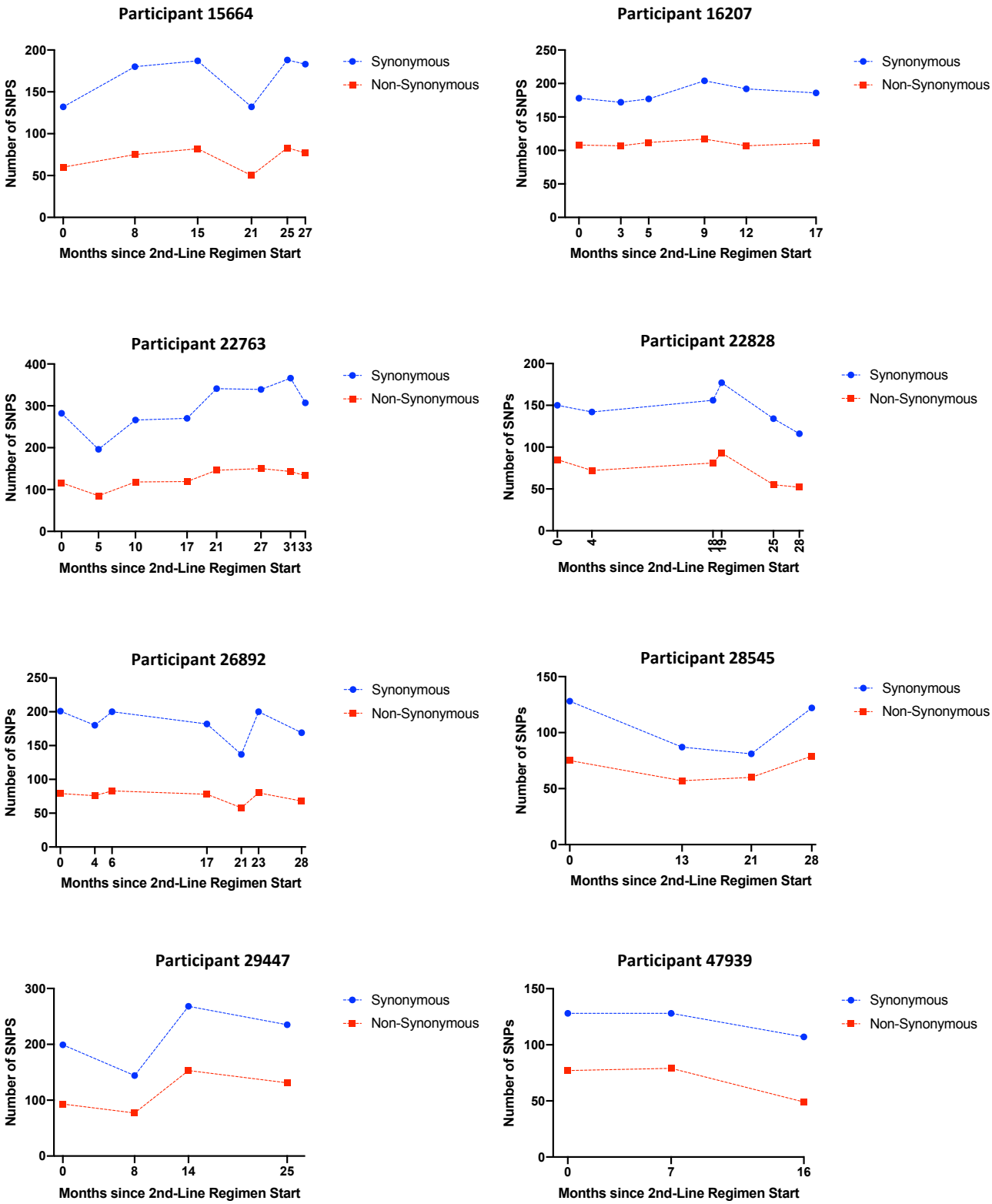

Figure 1. Measure of divergence from the SNPs relative to subtype C reference strain at successive timepoints in eight individuals under non-suppressive ART. These data were for SNPs detected by Illumina NGS at a threshold of 2% abundance. Sites in this analysis had coverage of at least 10 reads.

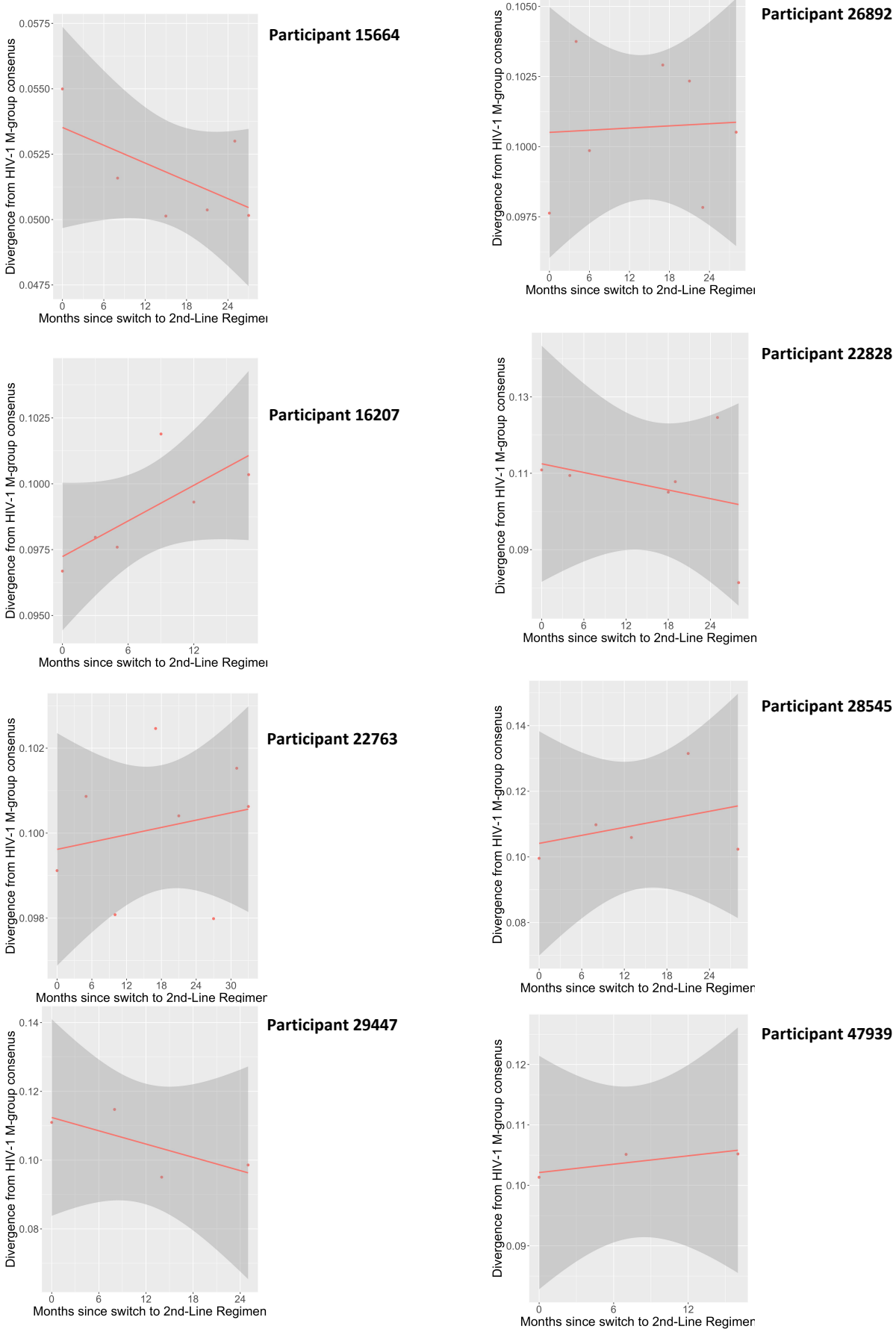

Figure 2. Divergence under non suppressive ART: Linear regression of average pairwise distance relative to a reconstructed subtype M consensus. Average pairwise distances were estimated under a TN93 substitution model and reveal increase divergence from from the initial samples. The 2002 HIV-1 subtype M group consensus sequences (Alignment ID: 102CG1) was downloaded and an overall ancestral consensus created from this using Geneious Prime v2021.1. This was used as a proxy for an ancestral HIV-1 group M sequence, which allowed us to determine if throughout ongoing treatment, the virus was reverting to an ancestral/founder state or diverging further. Shaded regions represented 95% CI of the linear regression fit.

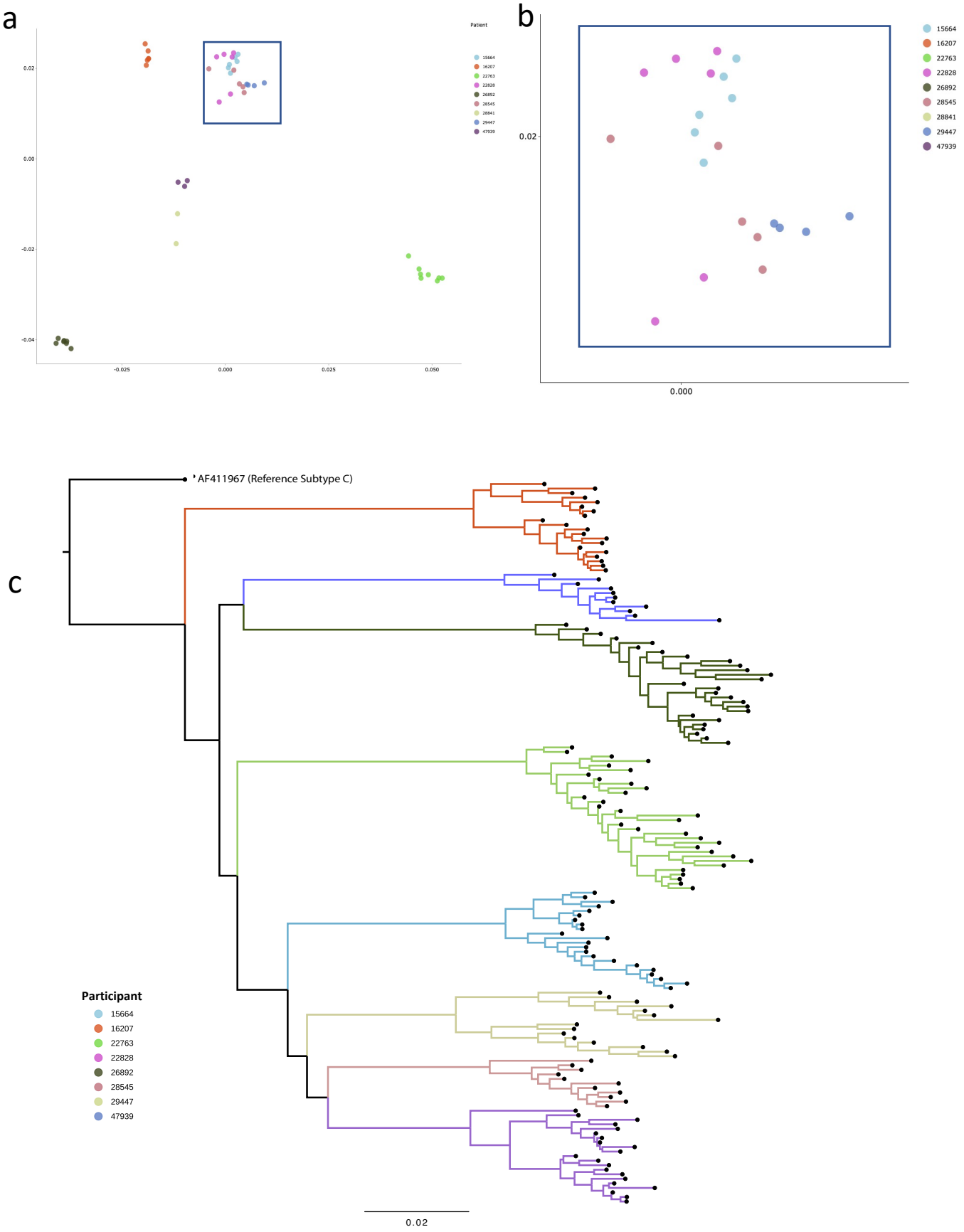

Figure 3. Multi-dimensional scaling showing A) clustering of HIV whole genomes from consensus sequences and **B) zoomed-in view of four participants with high intra-participant diversity.** Multi-dimensional scaling (MDS) were created by determining all pairwise distance comparisons under a TN93 substitution model, coloured by participant. **C) Maximum likelihood phylogeny of constructed viral haplotypes for all participants**. The phylogeny was rooted on the AF411967 clade C reference genome. Reconstructed haplotypes were genetically diverse and did not typically cluster by timepoint.

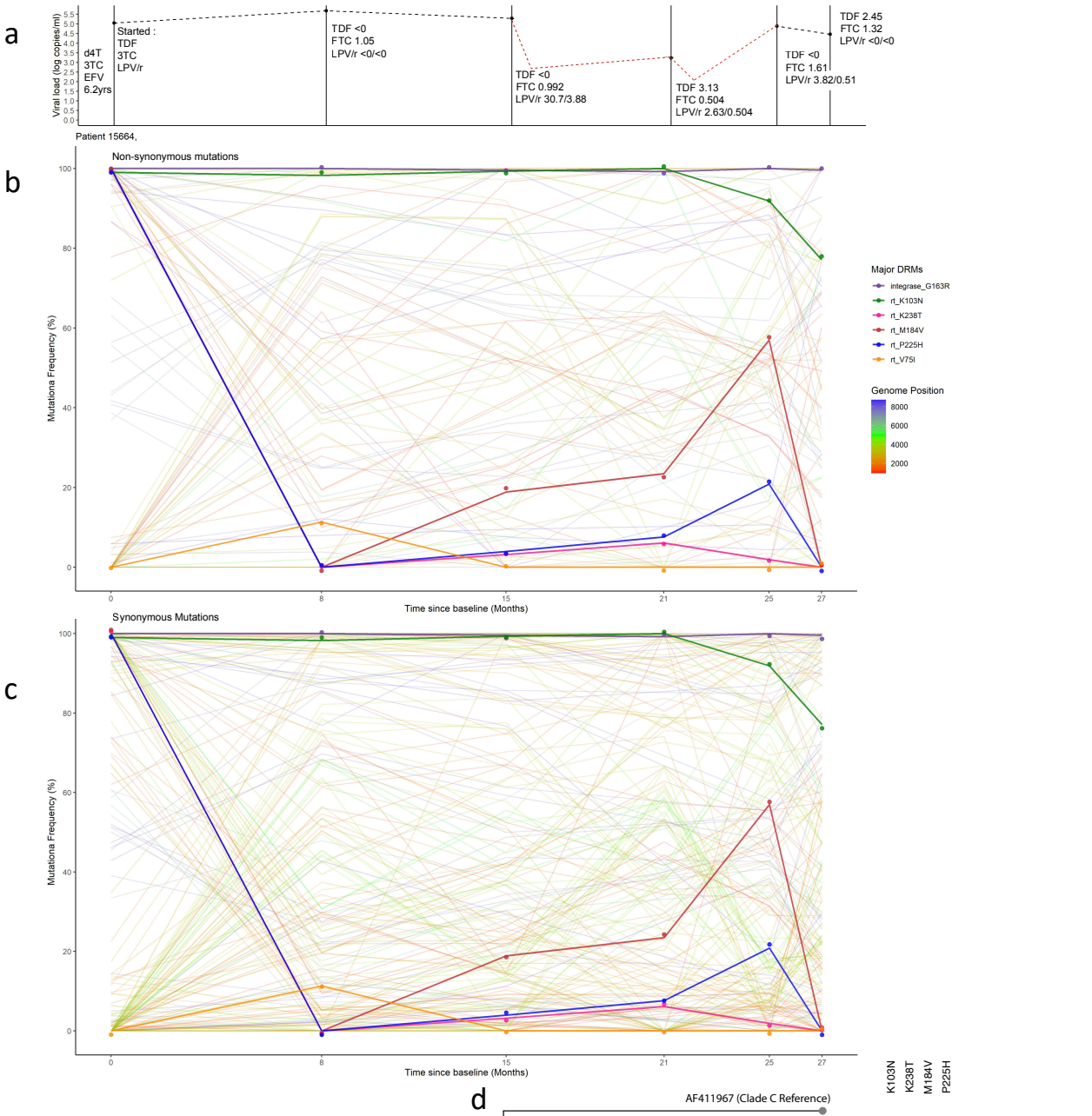

Figure 4. Drug regimen, adherence and viral dynamics within participant 15664. a) Viral load and drug levels. At successive timepoints drug regimen was noted and plasma drug concentration measured by HPLC (nmol/l). The participant was characterised by multiple partial suppression (<750 copies/ml, 16 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 non-synonymous mutation frequencies by Illumina NGS.** The participant had large population shifts between timepoints 1-2, consistent with a hard selective sweep, coincident with the shift from 1st-line regimen to 2nd-line. **c) Synonymous mutation frequencies.** All mutations with a frequency of <10% or >90% at two or more timepoints were tracked over successive timepoints. Most changes were restricted to *gag* and *pol* regions and had limited shifts in frequency i.e. between 20-60%. **d) Maximum-likelihood phylogeny of reconstructed haplotypes.** Haplotypes largely segregated into two major clades, those with the NNRTI mutation K103N and those without. 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.

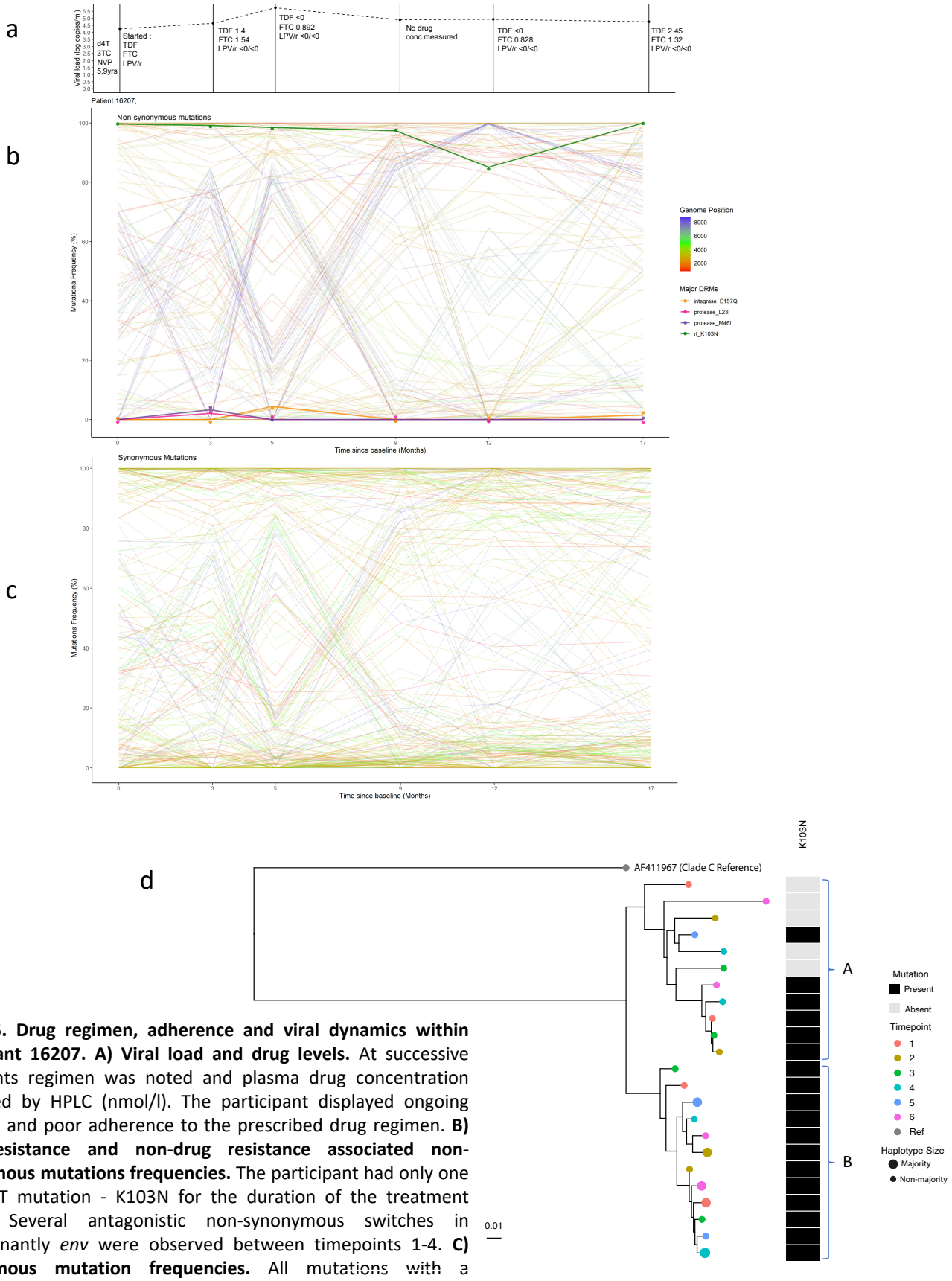

Figure 5. Drug regimen, adherence and viral dynamics within participant 16207. A) Viral load and drug levels. At successive timepoints regimen was noted and plasma drug concentration measured by HPLC (nmol/l). The participant displayed ongoing viraemia and poor adherence to the prescribed drug regimen. **B) Drug resistance and non-drug resistance associated non-synonymous mutations frequencies.** The participant had only one major RT mutation - K103N for the duration of the treatment period. Several antagonistic non-synonymous switches in predominantly *env* were observed between timepoints 1-4. **C) Synonymous mutation frequencies.** All mutations with a frequency of <10% or >90% at two or more timepoints were followed over successive timepoints. In contrast to non-synonymous mutations, most synonymous changes were in *pol*, indicative of linkage to the env coding changes. **D) Maximum-likelihood phylogeny of reconstructed haplotypes.** Haplotypes were clearly divided by a bifurcation; each clade contained haplotypes from all timepoints, suggesting lack of hard selective sweeps and intermingling of viral haplotypes with softer sweeps. that most viral competition occurred outside of drug pressure.

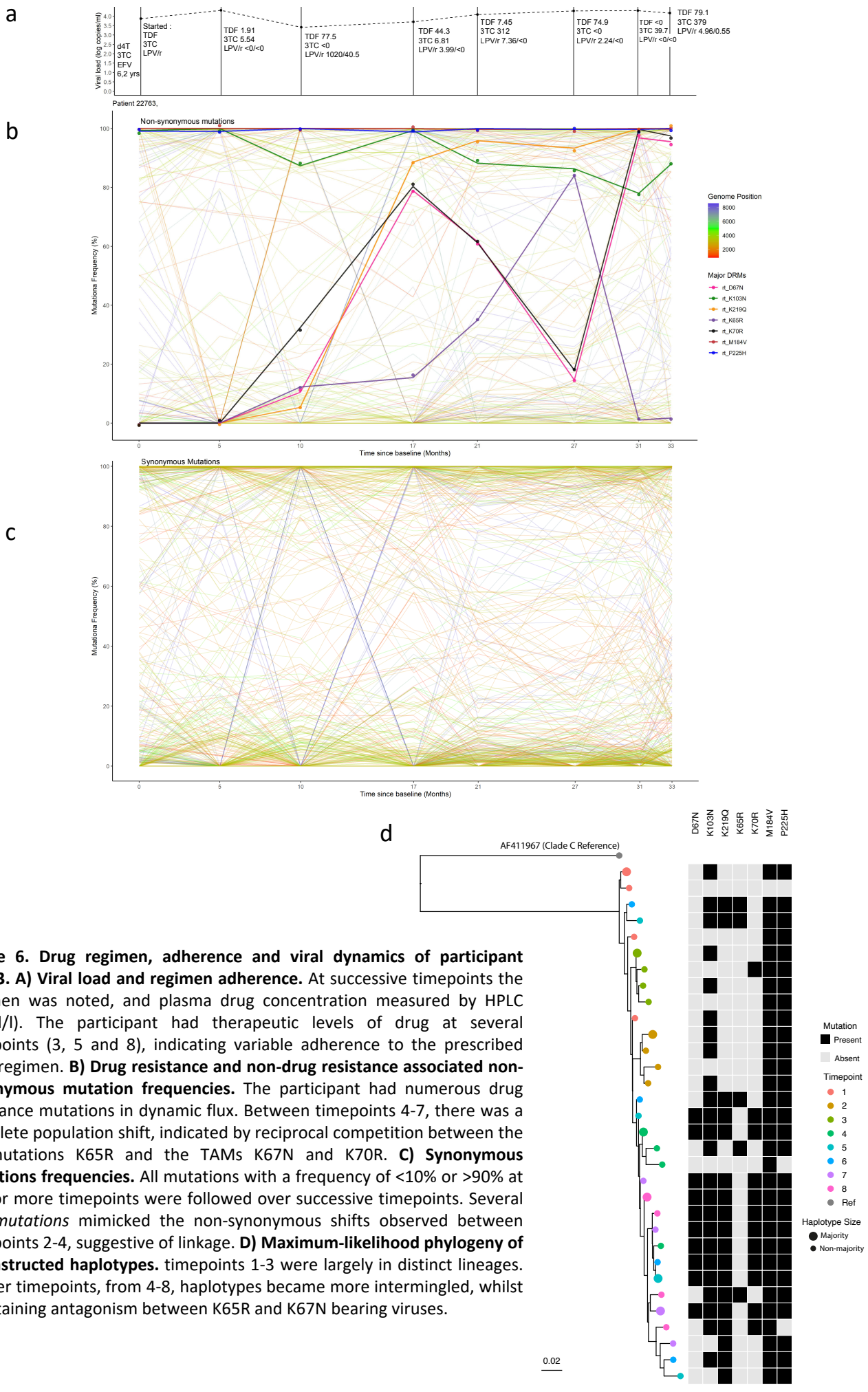

Figure 6. Drug regimen, adherence and viral dynamics of participant 22763. A) Viral load and regimen adherence. At successive timepoints the regimen was noted, and plasma drug concentration measured by HPLC (nmol/l). The participant had therapeutic levels of drug at several timepoints (3, 5 and 8), indicating variable adherence to the prescribed drug regimen. **B) Drug resistance and non-drug resistance associated non-synonymous mutation frequencies.** The participant had numerous drug resistance mutations in dynamic flux. Between timepoints 4-7, there was a complete population shift, indicated by reciprocal competition between the RT mutations K65R and the TAMs K67N and K70R. **C) Synonymous mutations frequencies.** All mutations with a frequency of <10% or >90% at two or more timepoints were followed over successive timepoints. Several *env mutations* mimicked the non-synonymous shifts observed between timepoints 2-4, suggestive of linkage. **D) Maximum-likelihood phylogeny of reconstructed haplotypes.** timepoints 1-3 were largely in distinct lineages. In later timepoints, from 4-8, haplotypes became more intermingled, whilst maintaining antagonism between K65R and K67N bearing viruses.