HIV-1 evolutionary dynamics under non-suppressive, 2nd-line protease-inhibitor containing 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], Darren P. Martin[9], on behalf of the French National Agency for AIDS and Viral Hepatitis Research (ANRS) 12249 Treatment as Prevention (TasP) Study Group, Deenan Pillay[1], 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:
Steven A Kemp
Cambridge Institute for Therapeutic Immunology and Infectious Diseases
Jeffrey Cheah Biomedical Centre
Cambridge CB2 0AW, UK
sk2137@cam.ac.uk

or

Ravindra K Gupta
Cambridge Institute for Therapeutic Immunology and Infectious Diseases
Jeffrey Cheah Biomedical Centre
Cambridge CB2 0AW, UK

Rkg20@cam.ac.uk

Abstract

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.

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. Recombination and linkage disequilibrium between haplotypes and genes was also explored.

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, 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.

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

69 exposure. Drug resistance mutations in RT can be used as markers of viral populations in the
70 reservoir and we found evidence for loss of linkage disequilibrium for drug resistance mutations,
71 indicative of recombination. These data imply that even years of exposure to PIs, within the context
72 of large stable populations displaying ongoing selective competition, may not precipitate emergence
73 of major PI resistance mutations, indicating significant fitness costs for such mutations. Ongoing viral
74 diversification within reservoirs may compromise the goal of sustained viral suppression.
75

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 [7-10]. These works have shown non-linear diversification of virus away from the founder strain during chronic untreated infection.

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 1st-line ART[5,11]. As a result, ART-experienced participants failing 1st-line regimens for prolonged periods of time are characterised by high frequencies of common nucleoside reverse transcriptase (NRTI) and non-nucleoside reverse transcriptase (NNRTI) drug resistance mutations (DRMs) such as M184V, K65R and K103N [12]. Routinely, 2nd-line ART regimens consist of two NRTIs and and underpinned by a boosted protease inhibitor (PIs); PI DRMs are uncommonly reported [13] however, 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 impact PI susceptibility [14-20], though common pathways have been difficult to discern, likely reflecting plasticity to drug escape.

Prolonged virological failure on PI-based regimens, without 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 [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. HIV has been shown to exhibit significant genetic diversity within infected hosts, with different populations of virus accumulating beneficial mutations – these are referred to as 'quasispecies' [25,26]. 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 [27], 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 [28]. In this manuscript, we report details of evolutionary dynamics during non-suppressive 2nd-line ART, through investigation of individual individual quasispecies using a novel computational haplotype reconstruction tool, Haplotype Reconstruction of Longitidunal Deep sequencing data (HaROLD, [29].

Results

Participant Characteristics

Eight south African 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. Participant metadata collected included viral loads, regimens and time since ART initiation (**Table 1**). HIV RNA was isolated from venous blood samples and subject to whole-genome sequencing (WGS) using Illumina technology; from this whole-genome haplotypes were reconstructed. Prior to participation in the TasP trial, participants accessed 1st-line regimens for an average of 5.6yrs (±2.7yrs). At baseline enrolment 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 had DRMs associated with PI failure at minority frequencies (average 6.4%) and usually at single timepoints throughout the longitudinal sampling. Observed mutations included L23I, I47V, M46I/L, G73S, V82A, N83D and I85V (**Supplementary Tables 1a-3c**). Viral populations of four of the eight participants also carried major integrase stand inhibitor (INSTI) mutations, at minority frequencies (average 5.0%) and usually at single timepoints (T97A, E138K, Y143H, Q148K).

SNP frequencies and measures of diversity/divergence over time

WGS data was used to measure the changing frequencies of viral single nucleotide polymorphisms (SNPs) relative to a dual-tropic subtype C reference sequence (AF411967) within individuals over time (**Figure 1a-b**). The number of longitudinal synonymous SNPs approximately mirrored the number of non-synonymous SNPs, but the former were two-to-three-fold more common. In most participants, viral populations were homogenous. Diversification, by counting the number of SNPs

144 difference to the refence sequence was considered. There was largely idiosyncratic changes in the
145 number of SNPs over time, with both increases and decreases in number of SNPs, suggesting
146 different population competition, or selective sweeps occurring. From timepoint two onwards (all
147 participants now on 2nd-line, PI-containing regimens for >6 months), all participants (except 28545)
148 had increases in both synonymous and non-synonymous SNPs.
149
150 In previous literature, chronically infected, but untreated HIV-1 patients viral populations showed
151 some reversion to the founder or infecting virus[7]. We assessed this phenomenon in our chronically
152 infected, but treated HIV-1 population. HIV-1, subtype C and M group consensus sequences were
153 downloaded from the Los Alamos National Laboratory (LANL) database and an 'ancestral' consensus
154 sequence was constructed according to the Materials and Methods. Diversification towards or away
155 from the baseline (or infecting sequence) as well as both the subtype C and subtype M consensus
156 was measured.
157
158 When considering the whole genome, linear regression suggests that all sequences continued to
159 diversify away from the infecting/founder strain, though R^2 and p-values suggests the diversification
160 towards or away from the founder was not significant in most cases. Despite this, all participants
161 appeared to diversify away from the infecting strain (**Figure 1C**), all but two participants (15664 &
162 29447) diversified away from the ancestral subtype C (**Figure 1D**), and all but three participants
163 (15664, 22828 & 29447) diversified away from the ancestral Subtype M (**Figure 1E**). This is contrary
164 to current literature, and may be a novel feature for patients who fail 2nd-line regimens or
165 sporadically adhere to therapies. To determine whether specific genomic regions were responsible
166 for reversion or divergence, we re-examined the divergence from simulated ancestors across the
167 genome in a sliding window of 1000bp (**Supplementary Figures 1-3**). This revealed that whilst there
168 were heterogeneous patterns of divergence, when considering portions of each genome
169 independently there was a general trend towards divergence, although curiously the regions 1-
170 2000bp (covering the *gag* gene) appears to be show reversion (to the baseline strain). The highest
171 degree of diversification occurred in regions 2000-3000bp and 7000-8000bp, corresponding to the
172 *protease* and *reverse transcriptase* genes, and the *env* gene respectively. For any given patient their
173 HIV genomes are able to both revert in part, and diverge in others, likely enabled by recombination
174 which unlinks hyper-variable loci from strongly constrained neighboring sites.
175
176 To assess the relationship of the observed divergent patterns, we examined nucleotide diversity by
177 considering all pairwise nucleotide distances of each consensus sequence, by timepoint and

participant using a multidimensional scaling approach [30]. Intra-participant nucleotide diversity varied considerably between participants (**Figure 2a**). Some participants showed little diversity between timepoints (e.g. participant 16207), whereas others showed higher diversity between timepoints (e.g. participant 22763). Some participants were tightly clustered, suggesting little change over time (**Figure 3a,** participants 16207, 26892 & 47939), compared to others (participants 22828 & 28545). To corroborate the MDS approach, we used a novel method of examining nucleotide diversity of longitudinal timepoints using all positional information from BAM files (**Supplementary Figure 4**).

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 tool[29] to infer 289 unique haplotypes across all participants, with between 11 and 32 haplotypes (average 21) per participant. The number haplotypes changed dynamically between successive timepoints indicative of dynamically shifting populations (**Figure 2B**). To ensure that haplotypes were sensibly reconstructed, a phylogeny of all consensus sequences was also inferred (**Supplementary Figure 5**). Furthermore, to ensure that reconstructed haplotypes were sensible, a subsequent MDS plot of all viral haplotypes was constructed (**Supplementary Figure 6**).

Linkage Disequilibrium 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[31]. HIV is known to rapidly recombine such that sequences are not generaly in Linkage Disequilibrium (LD) beyond 400bp[7]. The significance of recombination in intra-host, single infection setting is less well understood[32]. To asses if intra-patient recombination was occuring between patient haploytypes for three most sampled participants, we determined LD decay patterns, assuming that if there was random recombination, this would equate to a smooth LD decay pattern. This was not observed, rather, each participant demonstrated a complex decay pattern, consistent with non-random recombination along the genome (**Figure 3A).** Given this, we charactersied recombination patterns (**Figure 3b**). Perceived recombination breakpoints were recurrent within participants and identifiable over successive timepoints. DRMs were gained over successive timepoints for time for participant 22763 whereas in participant 15664, the reverse was true, whereby the converse is seen for patient 15664, whereby there was a gradual loss of DRMs. This indicaties that the not all DRMs were required to overcome sporadic drug pressure and the original drug-resistant virus with lower fitness was selected preferentially by ART

drug pressure. Participant 16207 had recombinant breakpoints localised in the the *pol* gene in two timepoints, though it retained its majority DRM (K103N) across all haplotype populations.

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[33,34]. There are also data suggesting associations between *env* mutations and PI exposure [35,36]. **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[18,37]. 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 the absence of P225H (**Figure 4d**, dark grey 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**, light grey 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 three and five. Between timepoints 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 4d**, light blue circles). Several inferred haplotypes without DRMs was nonetheless phylogenetically distinct from the timepoint one minority haplotype (**Figure 4d**, compare small orange and pink circles in lowest clade).

Upon examining the phylogenetic relationships of the inferred haplotype sequences, there were several distinct clades with haplotypes from all timepoints interspersed throughout (except at timepoint 4, which remained phylogenetically distinct). This is indicative of ongoing viral population competition. DRMs showed some segregation by clade; viruses carrying a higher frequency of DRMs were observed in upper cladea (Clade A, **Figure 4d**), and those with either K103N alone, or no DRMs were preferentially located in the upper clade (Clade C, **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, Supplementary Figure 4**) 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 equally observed in the upper clade (A) and lower clade (C) (**Figure 5d**). K103N was the majority DRM at all timepoints, except for a minority haplotype at timepoint three, also carrying E157Q. Haplotypes did not cluster by time point. Significant diversity in haplotypes from this participant was confirmed by MDS (**Supplementary Figure 6**).

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 boosted lopinavir (LPV/r) and tenofovir (TDF) were measured in plasma and haplotypes clustered separately from the first two timepoints (**Figure 6d**, light and dark grey 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 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 observed the same groups of other non-synonymous or synonymous alleles exhibiting dramatic frequency shifts, but to a lesser degree than in the previous two participants i.e. 'chevron patterns' were less pronounced, outside of the *env* gene (**Figure 6b-c**). Variable drug pressures placed on the viral populations throughout the 2nd-line regimen appear to have played some role in limiting haplotype diversity. Timepoints 1-4 all formed distinct clades, without intermingling, indicating that competition between populations was not occurring to the same degree as in previous participants. 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[38-40]. One possible 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 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[41,42]. However, with the scale-up of ART, there has also been an increase in both pre-treatment drug resistance (PDR)[43,44] and acquired drug resistance[12,45] 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%[46]. 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

346 populations present in longitudinally collected plasma samples of chronically-infected HIV-1
347 participants under non-suppressive 2nd-line ART.
348
349 With the vast majority of PLWH treated in the post-ART era, virus dynamics during non-suppressive
350 ART is important to understand, as there may be implications for future therapeutic success. For
351 example broadly neutralising antibodies (bNab) are being tested not only for prevention, but also as
352 part of remission strategies in combination with latency reversal agents. We know that HIV
353 sensitivity to broadly neutralising antiboides (bNab) is dependent on *env* diversity[47,48], and therefore
354 prolonged ART failure with viral diversification could compromise sensitivity to these agents.
355
356 Our understanding of virus dynamics largely stems from studies that were limited to untreated
357 individuals[10], with largely subgenomic data analysed rather than whole genome[10]. Traditional
358 analysis of quasispecies distribution, for example as reported by Yu et al[49], suggests that the viral
359 diversity increases in longitudinal samples. However the findings of Yu et al were based entirely on
360 short-read NGS data without considering whole-genome haplotypes. The added benefit of
361 examining whole genome is that linked mutations can be identified statistically using an approach
362 that we recently developed[29]. Indeed haplotype reconstruction has proved beneficial in the analysis
363 of compartmentalisation and diversification of several RNA viruses, including HIV-1, CMV and SARS-
364 CoV-2[33,50,51].
365
366 Key findings of this study were: firstly that diversity increased over time with variable trajectory
367 away from the consensus baseline sequence and also the reconstructed ancestral subtype C and M
368 consensus. Approximately half of the participants appeared to diversify away from the
369 reconstructed ancestral subtype C and M sequence, wheras three participants showed possible
370 reversion back towards the ancestral consensus C and M (albeit with insufficient statistical support).
371
372 Secondly, and in contrast to the fractions of synonymous and non-synonymous mutations reported
373 by Zanini et al in a longitudinal untreated dataset[2], we show that the fractions of synonymous
374 mutations are generally two-to-three fold higher than non-synonymous mutations during non-
375 suppressive ART in chronic infection. This finding may reflect early versus chronic infection and
376 differing selective pressures. Haplotype reconstruction revealed evidence for competing haplotypes,
377 with evidence for numerous soft selective sweeps in phylogenies, evidenced by intermingling of
378 haplotypes during periods where there was low drug concentration measured in participant's blood
379 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[52]. 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[53]. Unlike the smooth LD decay curve as reported in the literature, we identified complex LD decay patterns within patients, indactive of non-random recombination. Recombination appears as the loss and gain of common genomic regions over successive tiempoints between each participant's haplotype populations (**Figure 3B**). Participant 15664 recombines between hapllotype populations in the *vif* and *vpr* genes in four of the six timepoints. In contrast, participant 22763 showed recombination in the *gag-pol* genes in three of the eight timepoints. We explain these recombination events in longitudinal sequences, as reflected in the previously discussed 'chevron' patterns whereby varianrts increase and susequenrly decrease between timepoints. HIV quasispecies allows the virus to increase fitness through recombination when selectively advantageous[26]. The relationship between recombination and aquisition of DRMs is unclear with each patient showing unique patterns; participant 16207 recombined in *pol* between haplotypes at timepoint two and six and maintained the major DRM K103N. Participant 22763 recombined in *pol*

at three timewpoints (two, four and six) resulting in no change of DRMS, gain of DRMs, and loss of DRMs respectively. This occurring at the same time as antagonism between TAMs and DRMs (K65R and D67N).Finally, participant 15664 steadily lost DRMs throughout longitudinal sampling, although we did not see evidence of recombination driving this. This suggests that in the absence of strong drug pressures, viral populations only mainted crucial DRMs which were useful to evading innate immunity.

This study had some limitations – we examined in-detail only 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 is unprecedented. We are confident that the combination of computational analyses has provided a detailed understanding of viral dynamics under non-suppressive ART may be 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 [50]. 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 analysed timepoints, nuances of the sequencing method led in some cases to suboptimal degrees of 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 >10 reads.

In summary we have found compelling evidence of HIV-1 within-host viral diversification, recombination and haplotype competition during non-suppressive ART. In future , 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[54,55], this approach needs assessment given data linking individuals with NNRTI resistance with poorer virological outcomes on Dolutegravir[56], coupled with a history of intermittent adherence. Having shown that long-time intra-host PI failure increases diversity of HIV viral populations, monitoring future drug-failure cases will be of interest due to their capacity to maintain a reservoir of drug-resistant and transmisible virus.

Methods

Study & Participant selection

This cohort was nested within the French ANRS 12249 Treatment as Prevention (TasP) trial [27]. 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 eight 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[57]. 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 [58] using a modified protocol previously described by Gall et al [59]. 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 [60] and mapped to a clade C South African reference genome (AF411967) with BWA-MEM [61]. The reference genome was manually annotated in Geneious Prime v2020.3 with

DRMs according to the Stanford HivDB [62]. Optical PCR duplicate reads were removed using Picard tools (http://broadinstitute.github.io/picard). Finally, QualiMap2 [63] 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 [64] 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 [65] 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 [66].

Haplotype Reconstruction & Phylogenetics

Whole-genome viral haplotypes were constructed for each participant timepoint using Haplotype Reconstruction for Longitudinal Samples (HaROLD) [67]. 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.

Maximum-likelihood phylogenetic trees and ancestral reconstruction were performed using IQTree2 v2.1.3[68] and a 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 AF411967.3 reference sequence, and nodes arranged in descending order. Phylogenies were manipulated and annotated using ggtrree v2.2.4.

Multi Dimension 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 v.5.4. Non-metric Multi-dimensional 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 whilst retaining most of the variation relationships between points. We find that like network trees, non-metric 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 dimensionalty 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 & Recombination

Starting wth a sequence alignment we determined the pairwise LD R^2 associations for all variable sites using WeightedLD[70]. This method allowed us to exclude sites with any insertions or ambiguous characters easily where we used the option --min-acgt 0.99 and --min-variability 0.05. The pairwise R^2 values were then binned per 200bp comparison distance blocks along the genome and the mean R^2 value were taken and represented graphically to assess LD decay. This analysis was run for the three participants taken forward for in-detail analysis, and run using an alignment of all their timepoint samples. Graphics were generated using Rv4.04.

We first performed an analysis for detecting individual recombination events in individual genome sequences using RDP, GENECONV, BOOTSCAN, MAXCHI, CHIMAERA, SISCAN, and 3SEQ methods implemented in RDP5[71] using default settings. Putative breakpoint hotspots 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 hotspots were confirmed if at least three or more methods supported the breakpoint.

Funding

SAK is supported by the Bill and Melinda Gates Foundation: OPP1175094. RKG is supported by Wellcome Trust Senior Fellowship in Clinical Science: WT108082AIA. OC is supported by a PhD studentship/UKRI MRC grant : MR/N013867/1.

Competing Interests

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

Author Contributions

Conceptualization of study: S.A.K, R.K.G, A.D, Bioinformatic processes: A.D, S.A.K, O.C, D.P.M, Writing and revising manuscript: S.A.K, O.C, A.D, D.P.M, D.P, R.A.G, R.K.G

Data Avilability Statement

All fasta files have been deposited on Genbank with the following accession numbers [pending approval].

Code Availability Statement

Custom code used to produce figures and graphs can be found at: https://github.com/Steven-Kemp/21-2_hiv_tasp/tree/main/scripts or within the references manuscripts.

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, 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 Zanini, F. *et al.* Population genomics of intrapatient HIV-1 evolution. *Elife* **4**, e11282, doi:10.7554/eLife.11282 (2015).

8 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, doi:10.1098/rspb.2012.0595 (2012).

9 Hedskog, C. *et al.* Dynamics of HIV-1 Quasispecies during Antiviral Treatment Dissected Using Ultra-Deep Pyrosequencing. *PloS one* **5**, e11345, doi:ARTN e11345 10.1371/journal.pone.0011345 (2010).

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

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, doi:10.1371/journal.pgen.1007859 (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. & Garcia-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, doi:10.1128/Jvi.02712-06 (2007).

25 Wilke, C. O. Quasispecies theory in the context of population genetics. *BMC Evol Biol* **5**, 44, doi:10.1186/1471-2148-5-44 (2005).

26 Lauring, A. S. & Andino, R. Quasispecies theory and the behavior of RNA viruses. *PLoS Pathog* **6**, e1001005, doi:10.1371/journal.ppat.1001005 (2010).

27 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).

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

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

31 Stephens, M. & Scheet, P. Accounting for Decay of Linkage Disequilibrium in Haplotype Inference and Missing-Data Imputation. *The American Journal of Human Genetics* **76**, 449-462, doi:https://doi.org/10.1086/428594 (2005).

32 Song, H. *et al.* Tracking HIV-1 recombination to resolve its contribution to HIV-1 evolution in natural infection. *Nature Communications* **9**, 1928, doi:10.1038/s41467-018-04217-5 (2018).

33 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).

34 Kletenkov, K. *et al.* Role of Gag mutations in PI resistance in the Swiss HIV cohort study: bystanders or contributors? *J Antimicrob Chemother* **72**, 866-875, doi:10.1093/jac/dkw493 (2017).

35 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).

36 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).

665 37 Datir, R., El Bouzidi, K., Dakum, P., Ndembi, N. & Gupta, R. K. Baseline PI susceptibility by
666 HIV-1 Gag-protease phenotyping and subsequent virological suppression with PI-based
667 second-line ART in Nigeria. *The Journal of antimicrobial chemotherapy* **74**, 1402-1407,
668 doi:10.1093/jac/dkz005 (2019).

669 38 Parikh, U. M., Zelina, S., Sluis-Cremer, N. & Mellors, J. W. Molecular mechanisms of
670 bidirectional antagonism between K65R and thymidine analog mutations in HIV-1 reverse
671 transcriptase. *Aids* **21**, 1405-1414 (2007).

672 39 Parikh, U. M., Bacheler, L., Koontz, D. & Mellors, J. W. The K65R mutation in human
673 immunodeficiency virus type 1 reverse transcriptase exhibits bidirectional phenotypic
674 antagonism with thymidine analog mutations. *Journal of virology* **80**, 4971-4977 (2006).

675 40 Parikh, U. M., Barnas, D. C., Faruki, H. & Mellors, J. W. Antagonism between the HIV-1
676 reverse-transcriptase mutation K65R and thymidine-analogue mutations at the genomic
677 level. *The Journal of infectious diseases* **194**, 651-660 (2006).

678 41 Department of Health. 2019 ART Clinical Guidelines for the Management of HIV in Adults,
679 Pregnancy, Adolescents, Children, Infants and Neonates. (Republic of South Africa National
680 Department of Health, 2019).

681 42 UNAIDS. *Global HIV & AIDS statistics — 2020 fact sheet*,
682 <https://www.unaids.org/en/resources/fact-sheet> (2020), Accessed 3rd March 2021.

683 43 Gupta, R. K. *et al.* HIV-1 drug resistance before initiation or re-initiation of first-line
684 antiretroviral therapy in low-income and middle-income countries: a systematic review and
685 meta-regression analysis. *Lancet Infect Dis* **18**, 346-355, doi:10.1016/S1473-3099(17)30702-
686 8 (2018).

687 44 Gupta, R. K. *et al.* Global trends in antiretroviral resistance in treatment-naive individuals
688 with HIV after rollout of antiretroviral treatment in resource-limited settings: a global
689 collaborative study and meta-regression analysis. *Lancet* **380**, 1250-1258,
690 doi:10.1016/S0140-6736(12)61038-1 (2012).

691 45 Gregson, J. *et al.* Occult HIV-1 drug resistance to thymidine analogues following failure of
692 first-line tenofovir combined with a cytosine analogue and nevirapine or efavirenz in sub
693 Saharan Africa: a retrospective multi-centre cohort study. *Lancet Infect Dis*,
694 doi:10.1016/S1473-3099(16)30469-8 (2017).

695 46 WHO, C. Global Fund. HIV drug resistance report. 2017. *World Health Organisation* (2017).

696 47 Stefic, K., Bouvin-Pley, M., Braibant, M. & Barin, F. Impact of HIV-1 Diversity on Its Sensitivity
697 to Neutralization. *Vaccines (Basel)* **7**, 74, doi:10.3390/vaccines7030074 (2019).

698 48 Pancera, M. *et al.* Structure and immune recognition of trimeric pre-fusion HIV-1 Env.
699 *Nature* **514**, 455-461, doi:10.1038/nature13808 (2014).

700 49 Yu, F. *et al.* The Transmission and Evolution of HIV-1 Quasispecies within One Couple: a
701 Follow-up Study based on Next-Generation Sequencing. *Scientific reports* **8**, 1404,
702 doi:10.1038/s41598-018-19783-3 (2018).

703 50 Pang, J. *et al.* Mixed cytomegalovirus genotypes in HIV-positive mothers show
704 compartmentalization and distinct patterns of transmission to infants. *Elife* **9**, e63199,
705 doi:10.7554/eLife.63199 (2020).

706 51 Boshier, F. A. T. *et al.* Remdesivir induced viral RNA and subgenomic RNA suppression, and
707 evolution of viral variants in SARS-CoV-2 infected patients. *medRxiv*,
708 2020.2011.2018.20230599, doi:10.1101/2020.11.18.20230599 (2020).

709 52 Parry, C. M. *et al.* Three residues in HIV-1 matrix contribute to protease inhibitor
710 susceptibility and replication capacity. *Antimicrobial agents and chemotherapy* **55**, 1106-
711 1113, doi:10.1128/AAC.01228-10 (2011).

712 53 Neher, R. A. & Leitner, T. Recombination rate and selection strength in HIV intra-patient
713 evolution. *PLoS Comput Biol* **6**, e1000660, doi:10.1371/journal.pcbi.1000660 (2010).

714 54 El Bouzidi, K. *et al.* High prevalence of integrase mutation L74I in West African HIV-1
715 subtypes prior to integrase inhibitor treatment. *J Antimicrob Chemother* **75**, 1575-1579,
716 doi:10.1093/jac/dkaa033 (2020).

717 55 Derache, A. *et al.* Predicted antiviral activity of tenofovir versus abacavir in combination with
718 a cytosine analogue and the integrase inhibitor dolutegravir in HIV-1-infected South African
719 patients initiating or failing first-line ART. *The Journal of antimicrobial chemotherapy*,
720 doi:10.1093/jac/dky428 (2018).

721 56 Siedner, M. J. *et al.* Reduced efficacy of HIV-1 integrase inhibitors in patients with drug
722 resistance mutations in reverse transcriptase. *Nat Commun* **11**, 5922, doi:10.1038/s41467-
723 020-19801-x (2020).

724 57 Iwuji, C. *et al.* Universal test and treat is not associated with sub-optimal antiretroviral
725 therapy adherence in rural South Africa: the ANRS 12249 TasP trial. *J Int AIDS Soc* **21**,
726 e25112, doi:10.1002/jia2.25112 (2018).

727 58 Derache, A. *et al.* Impact of Next-generation Sequencing Defined Human Immunodeficiency
728 Virus Pretreatment Drug Resistance on Virological Outcomes in the ANRS 12249 Treatment-
729 as-Prevention Trial. *Clinical infectious diseases : an official publication of the Infectious*
730 *Diseases Society of America* **69**, 207-214, doi:10.1093/cid/ciy881 (2019).

731 59 Gall, A. *et al.* Universal amplification, next-generation sequencing, and assembly of HIV-1
732 genomes. *Journal of clinical microbiology* **50**, 3838-3844, doi:10.1128/JCM.01516-12 (2012).

733 60 Martin, M. J. E. j. Cutadapt removes adapter sequences from high-throughput sequencing
734 reads. **17**, pp. 10-12 (2011).

735 61 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*
736 *preprint arXiv:1303.3997* (2013).

737 62 Shafer, R. W. Rationale and uses of a public HIV drug-resistance database. *The Journal of*
738 *infectious diseases* **194 Suppl 1**, S51-58, doi:10.1086/505356 (2006).

739 63 Okonechnikov, K., Conesa, A. & Garcia-Alcalde, F. Qualimap 2: advanced multi-sample
740 quality control for high-throughput sequencing data. *Bioinformatics (Oxford, England)* **32**,
741 292-294, doi:10.1093/bioinformatics/btv566 (2016).

742 64 Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in
743 cancer by exome sequencing. *Genome Res* **22**, 568-576, doi:10.1101/gr.129684.111 (2012).

744 65 Charles, O. J., Venturini, C. & Breuer, J. cmvdrg - An R package for Human Cytomegalovirus
745 antiviral Drug Resistance Genotyping. *bioRxiv*, 2020.2005.2015.097907,
746 doi:10.1101/2020.05.15.097907 (2020).

747 66 Perrier, M. *et al.* Evaluation of different analysis pipelines for the detection of HIV-1 minority
748 resistant variants. *PloS one* **13**, e0198334, doi:10.1371/journal.pone.0198334 (2018).

749 67 Goldstein, R. A., Tamuri, A. U., Roy, S. & Breuer, J. Haplotype assignment of virus NGS data
750 using co-variation of variant frequencies. *bioRxiv*, 444877 (2018).

68 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).

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 Charles, O. J., Roberts, J., Breuer, J. & Goldstein, R. A. WeightedLD: The Application of Sequence Weights to Linkage Disequilibrium. *bioRxiv*, 2021.2006.2004.447093, doi:10.1101/2021.06.04.447093 (2021).

71 Martin, D. P. *et al.* RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. *Virus Evol* **7**, veaa087, doi:10.1093/ve/veaa087 (2021).

Table 1. Regimens and viral load at final timepoint for all participants. Participants initiated and maintained 1st-line regimens for between 1-10 years before being switched to 2nd-line regimens as part of the TasP trial. Eight of the nine participants were failing 2nd-line regimens at the final timepoint.

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,	5.9	LPV/r, TDF, FTC	56660

		NVP			
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
28841	2	-	-	-	199011
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

779 **NRTI:** Stavudine, d4T; Lamivudine, 3TC; Tenofovir, TDF; Emtricitabine, FTC; Zidovudine, AZT. **NNRTI:**
780 Efavirenz, EFV; Nevirapine, NVP. **PI:** Lopinavir/ritonavir, LPV/r. "-", data missing.
781
782

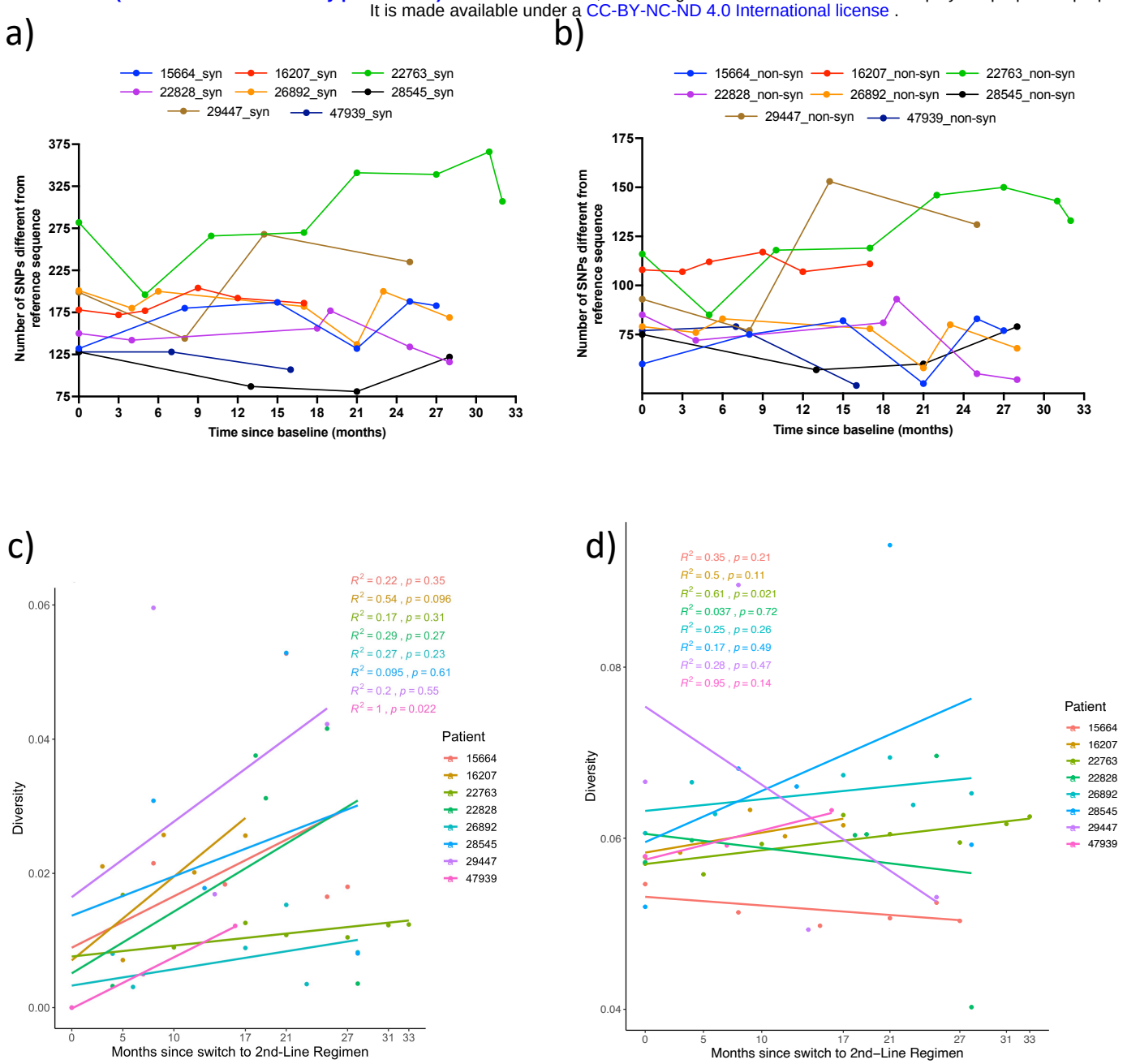


Figure 1. Measure of sequence divergence for eight participants under non-suppressive ART, relative to the subtype C reference strain at successive timepoints. These data were for SNPs detected by Illumina NGS at <2% abundance. Sites had coverage of at least 10 reads. In both a) synonymous and b) non-synonymous mutations, there was idiosyncratic change in number of SNPs relative to the reference strain over time. **1c-e**) **Linear regression of average pairwise distance relative to C) The baseline timepoint, D) a reconstructed subtype C consensus and E) a reconstructed subtype M consensus.** Average pairwise distances were estimated under a TN93 substitution model and reveal divergence from the initial samples. The Ancestral/Consensus HIV-1 subtype C and M was downloaded from the Los Alamos National Laboratory (https://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html). All ancestral HIV-1 subtype were downloaded from the same alignment and a consensus was created, as a proxy for an ancestral HIV-1 group M sequence. R^2 and p-values for logistic regression fits are indicated.

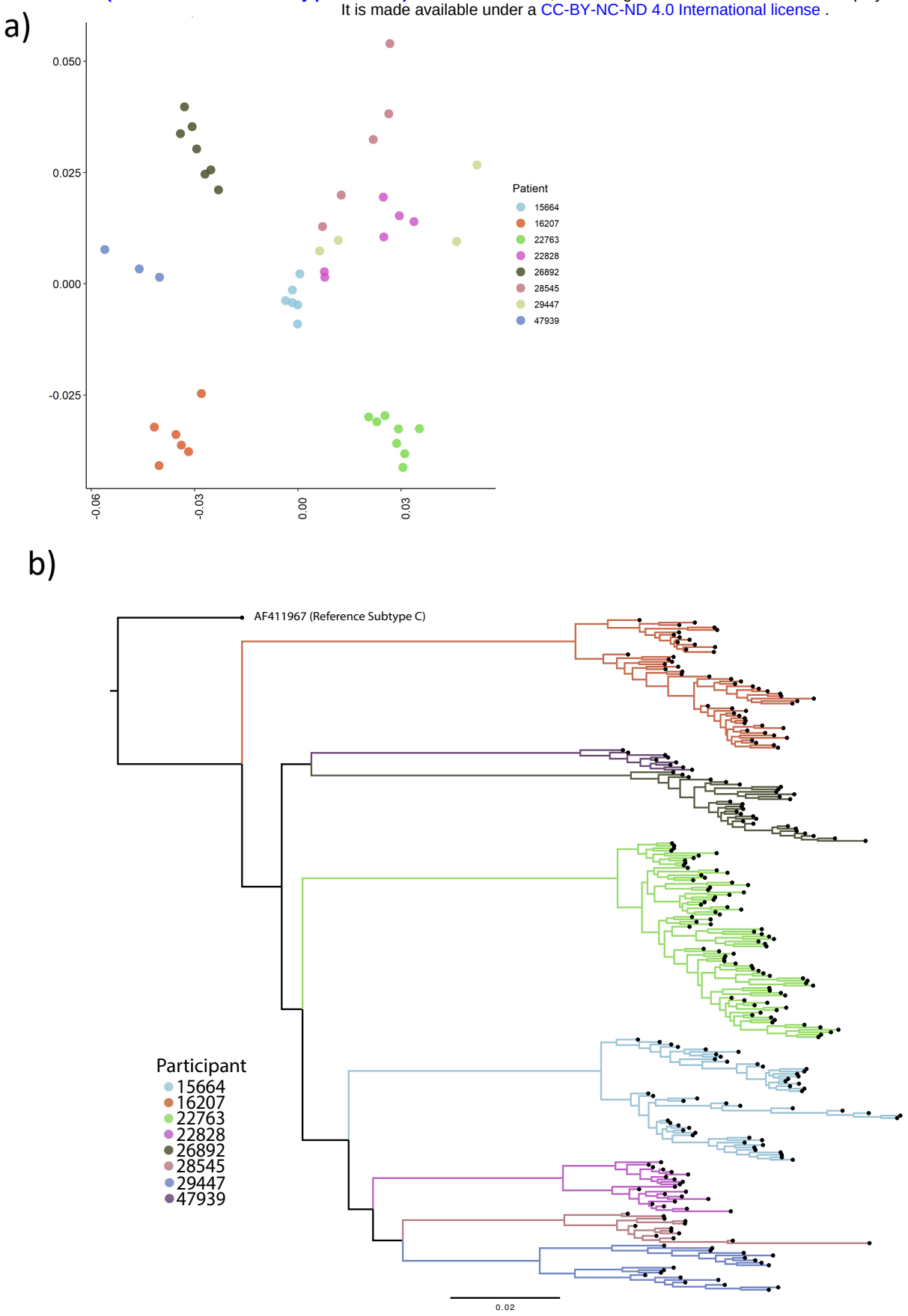
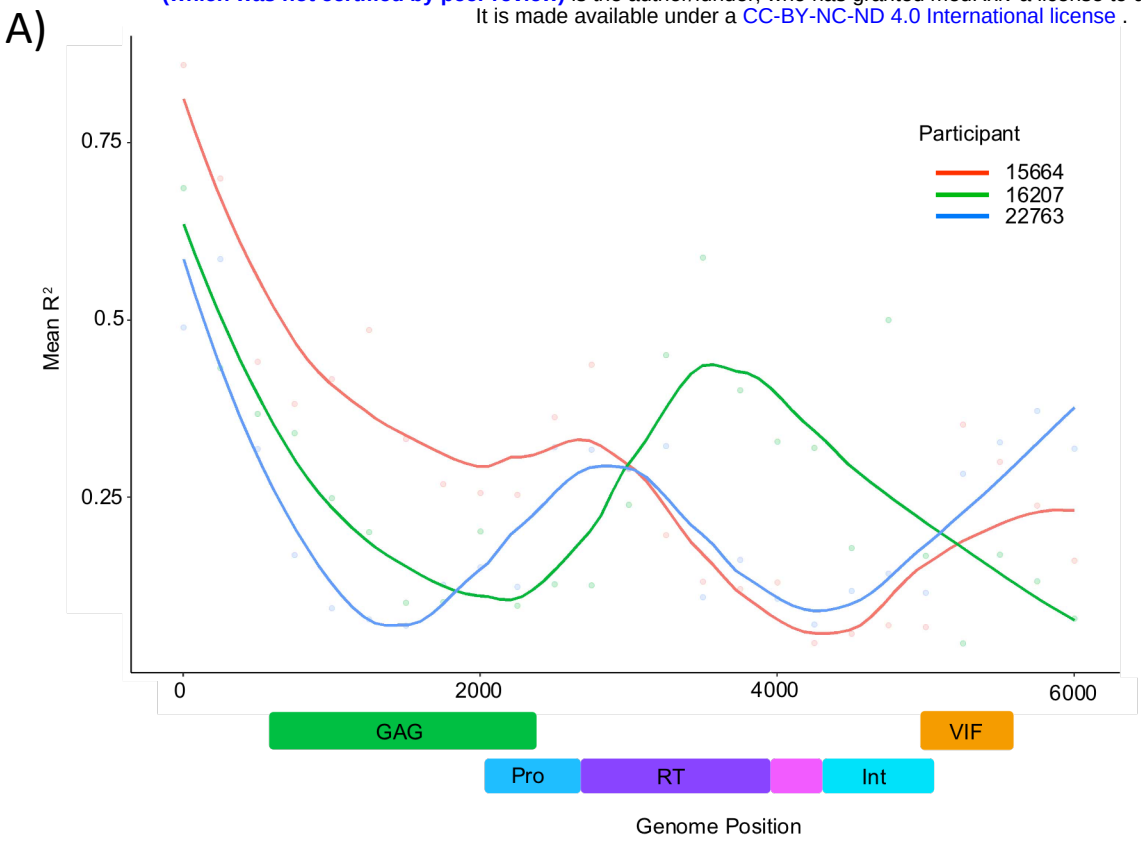


Figure 2. Multi-dimensional scaling showing A) clustering of HIV whole genomes from consensus sequences with high intra-participant diversity. Multi-dimensional scaling (MDS) were created by determining all pairwise distance comparisons under a TN93 substitution model, coloured by participant. Axis are MDS-1 and MDS-2. **B) 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.



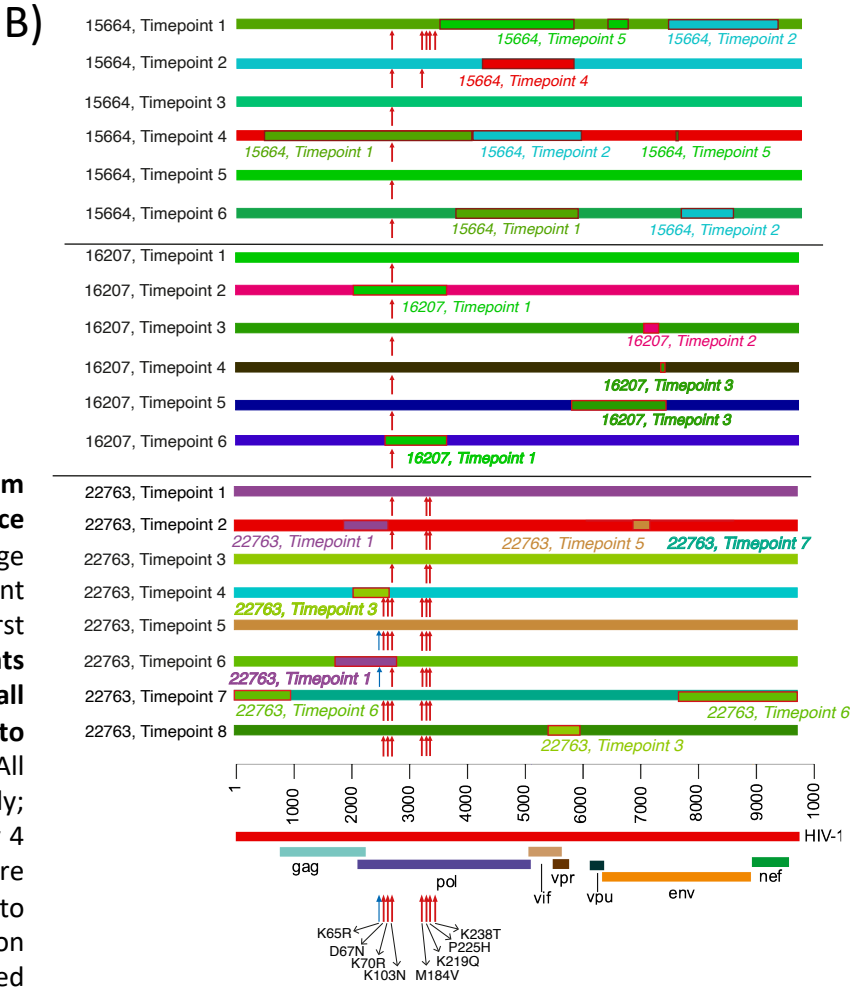


Figure 3A) Pairwise linkage disequilibrium decays rapidly with increasing distance between SNPs. The line indicates the average LD of all eight patients. There was a constant decrease in linkage disequilibrium over the first 800bp. **B) Perceived recombination breakpoints and drug-resistance associated mutations of all longitudinal consensus sequences belonging to three participants: 15664, 16207 and 22763**. All sequences were coloured uniquely 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 is indicated with a blue arrow.

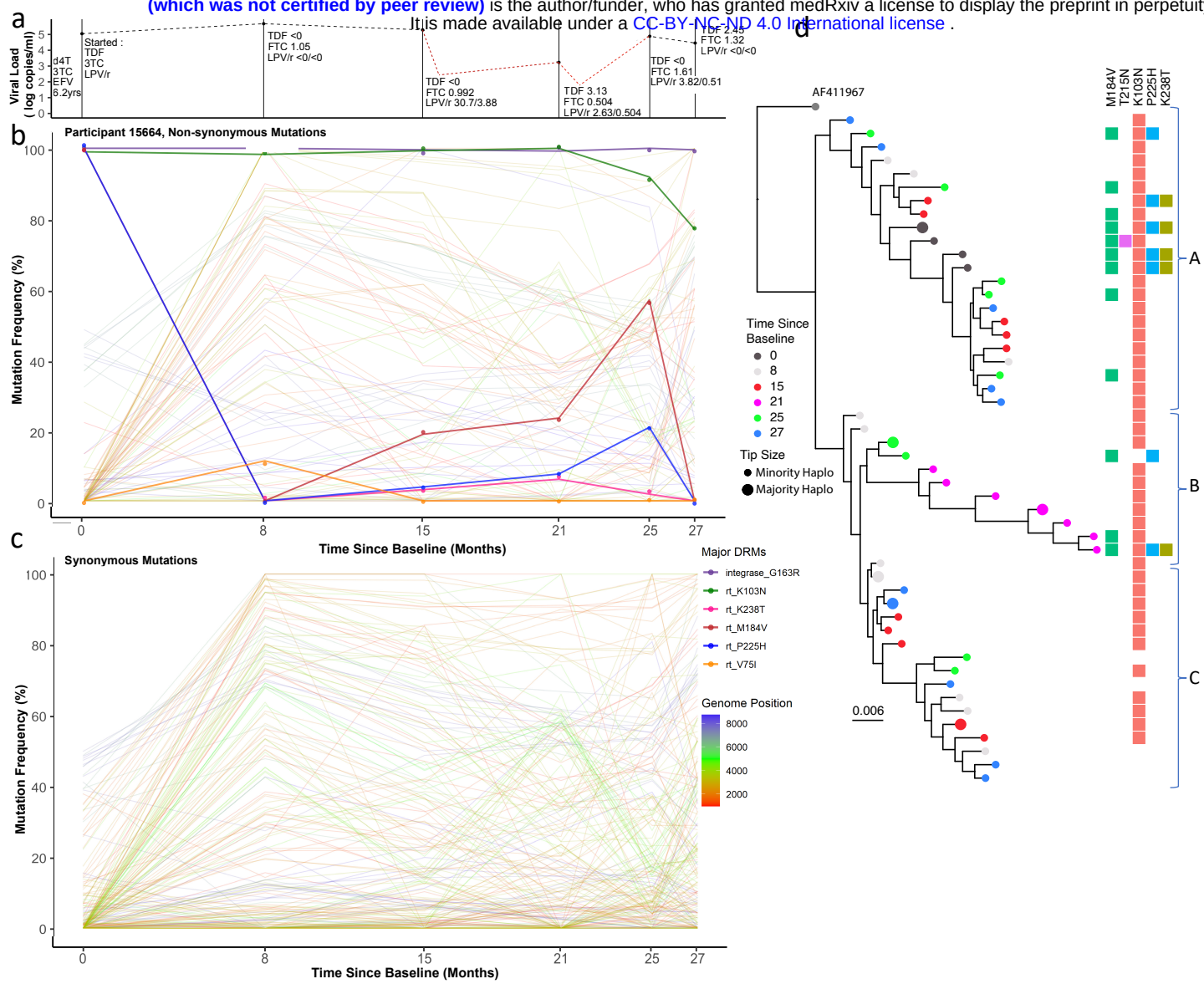


Figure 4. Drug regimen, adherence and viral dynamics within participant 15664. a) Viral load and drug levels. At successive timepoints drug regimen was noted and plasma drug concentration measured by HPLC (nmol/l). The participant was characterised by multiple partial suppression (<750 copies/ml, 16 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 three major clades (labelled A-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.

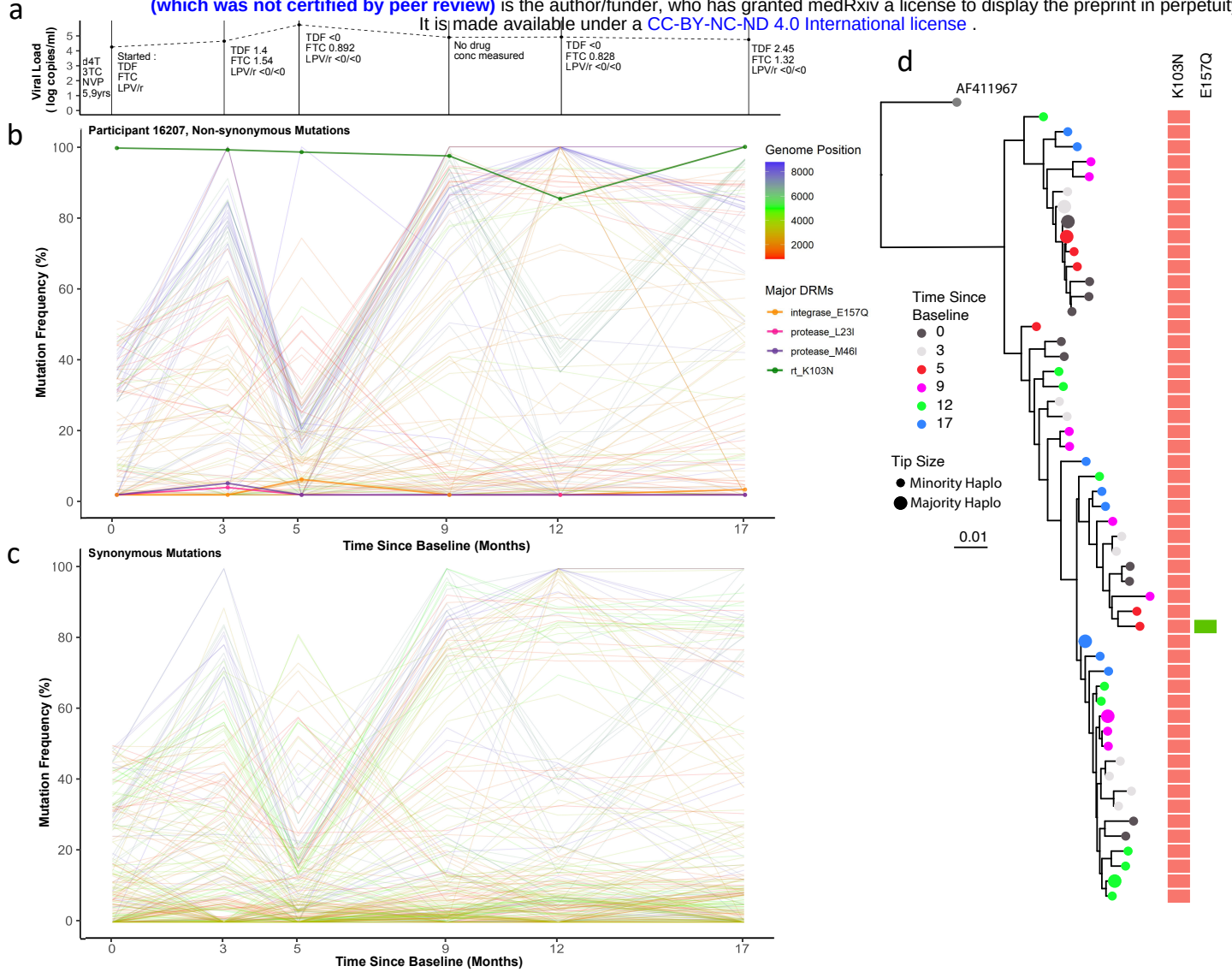


Figure 5. Drug regimen, adherence and viral dynamics within participant 16207. A) Viral load and drug levels. At successive timepoints regimen was noted and plasma drug concentration measured by HPLC (nmol/l). The participant displayed ongoing viraemia and poor adherence to the prescribed drug regimen. **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 again clearly divided intro three distinct clades; 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.

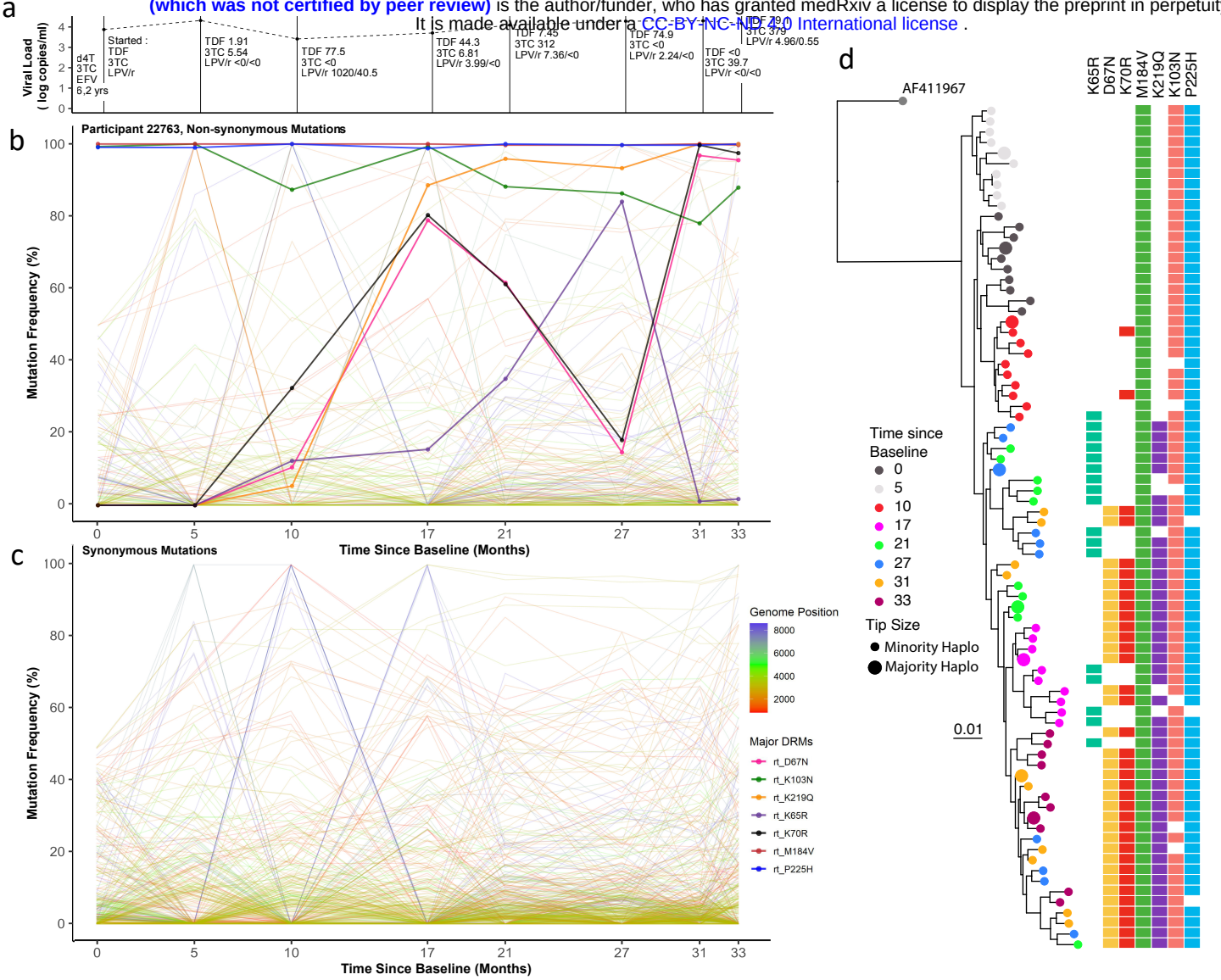


Figure 6. Drug regimen, adherence and viral dynamics of participant 22763. A) Viral load and regimen adherence. At successive timepoints the regimen was noted, and plasma drug concentration measured by HPLC (nmol/l). The participant had therapeutic levels of drug at several timepoints (3, 5 and 8), indicating 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 1-4 were found in distinct lineages. In later timepoints, from 5-8, haplotypes became more intermingled, whilst maintaining antagonism between K65R and K67N bearing viruses.