In-depth characterization of HIV-1 reservoirs reveals links to viral rebound during treatment interruption

Basiel Cole[1,#], Laurens Lambrechts[1,2,#], Zoe Boyer[3], Ytse Noppe[1], Marie-Angélique De Scheerder[4], John-Sebastian Eden[3], Bram Vrancken[5], Timothy E. Schlub[6], Sherry McLaughlin[7], Lisa M. Frenkel[7,8], Sarah Palmer[3,#] , Linos Vandekerckhove[1,4,#,*]

Affiliations

[1]HIV Cure Research Center, Department of Internal Medicine and Pediatrics, Ghent University Hospital, Ghent University, Ghent 9000, Belgium.
[2]BioBix, Department of Data Analysis and Mathematical Modelling, Faculty of Bioscience Engineering, Ghent University, Ghent 9000, Belgium
[3]Centre for Virus Research, The Westmead Institute for Medical Research, The University of Sydney, Sydney 2145, NSW, Australia.
[4]Department of General Internal Medicine and Infectious Diseases, Ghent University Hospital, Corneel Heymanslaan 10, Ghent 9000, Belgium.
[5]Department of Microbiology, Immunology and transplantation, Rega Institute, Laboratory of Evolutionary and Computational Virology, KU Leuven-University of Leuven, Leuven 3000, Belgium.
[6]University of Sydney, Faculty of Medicine and Health, Sydney School of Public Health, Sydney 2000, NSW, Australia.
[7]Center for Global Infectious Disease Research, Seattle Children's Research Institute, Seattle, Washington, United States of America.
[8]Departments of Global Health, Laboratory Medicine, Medicine, and Pediatrics, University of Washington, Seattle, Washington, United States of America.

[#]These authors contributed equally

[*]Correspondence to:

Prof. Dr. Linos Vandekerckhove
Department of Internal Medicine and Pediatrics
Corneel Heymanslaan 10, 9000 Ghent, Belgium
Ghent University
Linos.Vandekerckhove@UGent.be

Abstract

The HIV-1 reservoir is composed of cells harboring latent proviruses that are capable of refuelling viremia upon antiretroviral treatment interruption. This reservoir is in part maintained by clonal expansion of infected cells. However, the contribution of large, infected cell clones to rebound remains underexplored. Here, we performed an in-depth study on four chronically treated HIV-1 infected individuals that underwent an analytical treatment interruption (ATI). A combination of single-genome sequencing, integration site analysis, near-full length proviral sequencing and multiple displacement amplification was used to identify infected cell clones and link these to plasma viruses before and during an ATI. A total of six proviruses could be linked to plasma sequences recovered during ATI. Interestingly, only two of six proviruses were genome intact, one of which is integrated in the *ZNF141* gene. To our knowledge, this is the first instance of an intact provirus with its matched IS being matched to plasma virus during an ATI.

These findings demonstrate that with in-depth reservoir characterization, clones of infected cells harboring genome-intact proviruses can be linked to rebound viremia, confirming the previously proposed notion that infected clonal cell populations play an important role in the long-term maintenance of the replication-competent HIV-1 reservoir.

Introduction

HIV-1 infection remains incurable due to the presence of a persistent viral reservoir, capable of rebounding upon treatment interruption (TI) (1–4). Despite efforts to better understand the dynamics and maintenance of the HIV-1 viral reservoir, pinpointing the origins of viruses that rebound remains elusive (5). Previously, it was shown that CD4+ T cells carrying an HIV-1 provirus in their genome can undergo clonal expansion, contributing to the long-term persistence of the HIV-1 viral reservoir during antiretroviral therapy (ART) (6–14). The observation that low level viremias (LLV) under ART (15–20) and rebound viremia upon TI (5,19,21,22) often consist of monotypic populations of viruses, suggest that HIV-1 infected cell clones are key contributors to refueling viremia during TI. Clonality of infected cells has historically been demonstrated by recovering identical proviral sequences or identical integration sites (IS) in multiple cells (8,9,23–27). While the former method allows for qualitative assessment of the proviral genome, it is often not adequate to confidently predict clonal expansion of HIV-1 infected cells, especially when evaluating a short subgenomic region (28,29). On the other hand, integration site analysis (ISA) provides direct proof of clonal expansion, though it typically leaves the proviral sequence uncharacterized. Recently, two techniques to link near full-length (NFL) proviral sequences to IS were developed by Einkauf *et al.* (14) and Patro *et al.* (30), respectively called *Matched Integration site and Proviral sequencing* (MIP-Seq) and *Multiple Displacement Amplification Single Genome Sequencing* (MDA-SGS). These assays combine the qualitative strength of NFL HIV-1 sequencing with ISA, shedding light on the integration profile of intact versus defective proviruses.

Analytical treatment interruption (ATI) studies allow for the investigation of the dynamics and genetic makeup of rebounding viruses (21,31,32). To identify the source of rebounding viruses, we previously conducted the HIV-STAR (HIV-1 sequencing before analytical treatment interruption to identify the anatomically relevant HIV reservoir) study (5). During this study, in-depth sampling was performed on 11 chronically treated HIV-1 infected participants prior to ATI. Cells were isolated from different anatomical compartments and sorted into several CD4+ T cell subsets. Subgenomic proviral sequences (V1-V3 region of *env*) were recovered and phylogenetically linked to sequences from rebounding plasma virus collected during different stages of the ATI. This study suggested that HIV-1 rebound is predominantly fueled by genetically identical viral expansions, highlighting the potentially important role of clonal expansion in the maintenance of the HIV-1 reservoir. While this set-up allowed for the generation of a very broad and comprehensive dataset, it left some questions unanswered. Most importantly, the evaluation of a short subgenomic region (V1-V3 *env*) to link proviral sequences to rebounding plasma virus made it impossible to investigate the entire genome structure of proviruses linked to rebound. Furthermore, the lack of ISA did not allow for the study of the chromosomal location of the rebounding versus non-rebounding proviruses.

To address these points, we performed a combination of multiple displacement amplification (MDA), ISA and NFL proviral sequencing on four participants that were enrolled in the HIV-STAR study, with special attention on clonally expanded HIV-1 infected cells. We demonstrate that HIV-1 proviral sequences and corresponding IS of clonally expanded infected cells could be retrieved, and in rare cases these could be linked to rebounding plasma viruses. To our knowledge, we report the first instance of an intact proviral sequence with its associated IS being linked to plasma virus identified during an ATI. This provirus is integrated in a gene of the Krüppel-associated box domain (KRAB) containing zinc finger nuclease (ZNF) family, which adds to the growing body of evidence that this class of genes is a hotspot for genetically intact proviruses in patients on long-term ART.

Results

Experimental set-up

To investigate the genetic composition and chromosomal location of proviruses stemming from clonally expanded cells, and their relationship to rebound viremia, several qualitative assays were performed on samples from chronically treated HIV-1 infected individuals undergoing an ATI (Supplemental Table 1). These individuals were sampled longitudinally before and during the ATI, as summarized in Figure 1A, B.

First, the overall landscape of HIV-1 infected cell clones prior to ATI (Timepoint 1 (T1), Figure 1A) was determined by subgenomic single-genome sequencing (SGS) and Full-length Individual Proviral Sequencing (FLIPS) at the proviral level, and with Integration Site Loop Amplification (ISLA) at the integration site level (Figure 1A, Supplemental Table 1). This yielded three datasets that were used independently as a reference to identify potential clonally expanded infected cell populations.

In order to find links between the different datasets, multiple displacement amplification (MDA) was performed on sorted cell lysates from peripheral blood obtained during the pre-ATI timepoint (T1). MDA wells were subjected to V1-V3 *env* SGS and ISLA, and MDA reactions that yielded a V1-V3 *env* sequence and/or an IS corresponding to a suspected cellular clone, were further investigated. This was determined by an exact link to ISLA/FLIPS/SGS data generated in the first step, or by identical V1-V3 *env* sequences and/or IS shared between MDA wells. The proviruses in these selected MDA wells were sequenced using either a one-amplicon, four-amplicon, or five-amplicon approach, or a combination thereof (see methods). These MDA sequences were subsequently mapped back to proviral FLIPS sequences and historic V1-V3 *env* proviral sequences from PBMCs, gut-associated lymphoid tissue (GALT) and lymph node (LN) subsets prior to ATI (T1, Figure 1B), as well as V1-V3 *env* plasma sequences retrieved during the ATI (Timepoints 2-4 (T2-T4), Figure 1B).

This set-up allowed for the assessment of the genetic structure of proviruses in clonally expanded infected cells, their placement across cellular subsets and anatomical compartments, and their contribution to refuelling viremia during an ATI.

Integration site analysis and full-length proviral sequencing

To gain insight into the composition of the viral reservoirs of the four STAR participants, especially in terms of clonal expansion of infected cells, we initially performed bulk NFL proviral sequencing and ISA.

ISLA was performed on bulk cell lysate and on MDA-amplified cell lysate of TCM and TEM subsets from peripheral blood for three of the four study participants: STAR 9, STAR 10 and STAR 11 (Figure 2, Supplemental Table 2, Supplemental Table 4). Analysis of IS revealed a significantly higher degree of clonally expanded HIV-1 infected cells in the TEM proportion (mean 55%) compared to the TCM proportion (mean 16%) of the peripheral blood (P < 0.001 for STAR 9 and STAR 11; P = 0.036 for STAR 10), as previously reported (24). Identical IS between subsets, indicative of linear differentiation from an originally infected TCM into a TEM, was observed in rare instances, with 5 shared IS between subsets out of 284 distinct IS recovered (178 in TCM and 106 in TEM).

Near full-length HIV genomes (spanning 92% of the proviral genome) were recovered from TCM and TEM subsets in the peripheral blood and from CD45+ cells in the GALT for all four ART-treated participants before ATI (T1, Figure 1B). In addition, based on sample availability per participant, other cell subsets from the peripheral blood and LN were assayed with FLIPS as listed in Supplemental Table 1. This yielded a total number of 536 individual proviral genomes with a mean of 134 genomes per participant (Figure 2, Supplemental Table 3). Across all participants, only 30 (6%) intact proviral genomes were retrieved, with a majority of proviral sequences (68%, n=365) displaying large internal deletions (Supplemental Figure 1). In addition, the HIV-1 infection frequency differed significantly across cell subsets from the peripheral blood ($P < 0.001$), with the TEM subset having the highest infection frequency, except for participant STAR 4 (Supplemental Figure 2). Across the four participants, some deviations to the overall proportions of sequence types were observed, such as a higher fraction of hypermutated sequences (23%) in STAR 9 as compared to the overall proportion (15%) and a higher frequency of intact sequences (19%) in STAR 11 versus the overall frequency of 6%. These observations can be explained by cellular proliferation of HIV-1 infected cells as identified by expansions of identical sequences (EIS) in the FLIPS data. If proviruses belonging to such EIS are counted only once, these divergent proportions of hypermutated proviruses in STAR 9 and intact proviruses in STAR 11 disappear since they are driven by an EIS for that sequence type (Supplemental Figure 3). Furthermore, in each participant we observed more proviral genomes from the peripheral blood belong to an EIS in the TEM subset (mean average of 70%) than in the TCM subset (mean average of 34%), confirming the ISLA findings (Figure 2).

Multiple displacement amplification-mediated characterization of near full-length proviruses

MDA-mediated HIV-1 provirus sequencing and ISA offers the unique opportunity of linking NFL proviral sequences to their precise chromosomal location. Applying this technique to three of the four study participants, we could identify several expanded clones from which NFL sequences and matched IS could be retrieved, as shown in Figure 3 (Supplemental Figure 4, Supplemental Table 4).

STAR 9 displayed one major hypermutated clone (14% of all retrieved proviral genomes), predominantly present in the peripheral blood TEM fraction and integrated at an intergenic location on chromosome 11 (Figure 3). Interestingly, this clone could also be retrieved in the peripheral blood in the TCM subset by ISLA, and in the TCM/TTM/TEM subsets by FLIPS, which is indicative of differentiation of a clonally expanded cell population harboring a defective, hypermutated provirus (Supplemental Figure 4).

For STAR 10, one major clonally expanded cell population was found in the peripheral blood TCM fraction, with a provirus integrated in the *STAT5B* gene (Figure 3). This gene has previously been found to be significantly overrepresented in HIV-1 IS datasets (8,9,33,34). In most of these cases, the integration took place in the first intron in the same orientation as the gene, which can lead to aberrant transcription and production of the STAT5B protein (27,34). In this case however, the provirus was integrated against the orientation of the gene, in the first intron. Also, the provirus was shown to be defective, with a packaging signal defect in the form of a 25-bp deletion in stem loop 2 at the 5’ end of the genome. Three more clonal NFL genomes were retrieved in STAR 10, all with packaging signal or multiple splice donor site (MSD) defects: one in an intergenic region on chromosome 8, one in the long non-coding RNA gene *LINC00649*, and the third in the *CASC5* gene. One clonal intact NFL provirus was detected in an intronic region of the *CIT* gene, in the same orientation as the gene (Figure 3, Supplemental Figure 4).

With FLIPS, three different EIS containing genetically intact sequences were identified in the peripheral blood TEM fraction of STAR 11, which represent 10%, 5% and 3% of all NFL sequences retrieved in that subset. These EIS were also detected in several MDA wells, which enabled the identification of their corresponding IS. Looking at these clones at the IS level, they represent 7%, 19% and 2% of IS retrieved by ISLA in the peripheral blood TEM fraction, and are integrated in the *GGNBP2* gene, *ZNF274* gene and the *ZNF141* gene respectively. The provirus in the *GGNBP2* gene was integrated in the sixth intron, in the same orientation as the gene. Of note, this clone was not only observed in the TEM fraction but was also retrieved in the TCM fraction of the peripheral blood by both ISLA and FLIPS. The proviruses in the *ZNF141* gene and the *ZNF274* gene were integrated in the reverse orientation with respect to the gene. Interestingly, these genes belong to categories that have recently been described as harboring proviruses responsible for non-suppressed viremia, and 'deep latency' respectively (20,35). Finally, two proviruses with small deletions in the packaging signal and MSD were found in intergenic regions of chromosome 17 (Figure 3).

We conclude that a large fraction of the clonally expanded infected cell populations we identified harbor defective proviruses that would not be able to rebound during an ATI, however, in participant STAR 11, three clonal cell populations were identified that harbor a genetically intact provirus.

Large discrepancies between suspected clonal HIV-1 infected cell populations identified with ISLA, SGS and FLIPS

ISLA, SGS and FLIPS can independently be used to assess clonality of infected cells, the former based on the integration site and the two latter on the (subgenomic) proviral sequence of the provirus. To investigate whether the methods appear biased in their ability to detect specific clones, we used V1-V3 *env* or NFL sequences to assess overlap between assays (Figure 4).

Matches between MDA-ISLA data and FLIPS data were based on NFL sequences, whereas other links were based on V1-V3 *env* sequences. The Elimdupes tool (LANL) was used to identify EIS, which were validated by construction of maximum-likelihood (ML) trees using PHYML. For NFL matches, a total of 3-bp differences were allowed, to account for PCR-induced errors and sequencing errors, where for V1-V3 *env* matches, 100% accordance was required. In the case of IS data, only those IS that were associated with a corresponding V1-V3 *env* sequence (as found with MDA) could be linked. For FLIPS sequences, proviruses that have an internal deletion covering the V1-V3 *env* region could not be linked to SGS data.

Upon comparison of EIS present in SGS data and FLIPS data from participant STAR 4, one clear overlap could be found, in the peripheral blood TCM fraction. All other proviral sequences retrieved with SGS could not be linked unequivocally to sequences derived by using FLIPS, indicating a significant primer bias. However, one V1-V3 *env* sequence found with SGS in the TEM and the TCM fractions perfectly matched two distinct FLIPS sequences (Figure 4, green arrow). This is an example of a presumed 'clonal' EIS detected with SGS that consists of two or more proviruses sharing the same V1-V3 *env* region, although differing elsewhere in their genome.

A similar picture was observed for STAR 9, with only limited overlaps between assays. One major clone, integrated in an intergenic region on chromosome 11, was detected with both MDA-ISLA and FLIPS. In both assays, this clone was predominantly found in the peripheral blood TEM fraction (23% and 29% respectively), but also appeared in the peripheral blood TCM fraction. Strikingly, this provirus was never amplified with SGS, which can be explained by the fact that V1-V3 *env* primers did not anneal to this hypermutated sequence. This is another example of primer bias, which in this case can be explained by the hypermutated nature of the provirus.

Participant STAR 10 displays several instances of clear discrepancies between the assays. One large suspected EIS, based on V1-V3 *env* SGS, could be linked to four different IS (Figure 4, blue arrow). This most likely is the result of multiple distinct proviruses sharing a similar V1-V3 *env* sequence but integrated at different sites. Alternatively, this observation could result from MDA reactions containing more than one provirus, however, there was no evidence of mixed sequences observed from these wells. In addition, similar to the STAR 4 observation, one V1-V3 *env* sequence from the SGS data could be linked to two different NFL sequences, again indicating that in some cases subgenomic regions can be linked to different full-length sequences (Figure 4, red arrow).

Remarkable consistency between assays was observed for STAR 11, with all the clonal NFL sequences being linked to both SGS and MDA-ISLA data (Figure 4). However, the largest clone based on ISLA data, integrated in the *ZFC3H1* gene (Figure 2), could not be linked to SGS and FLIPS data, which was probably the result of large internal deletions spanning the entire length of the genome. In fact, out of ten MDA wells that yielded this integration site, the proviral sequence could never be amplified by V1-V3 *env* SGS, or by one-, four- or a five-amplicon approach NFL sequencing (Supplemental Table 4).

To quantify the discrepancies between assays, the clonal prediction score (CPS, described by Laskey *et al.* (28)) for the V1-V3 *env* region was calculated for all participants individually, based on available FLIPS data (Supplemental Table 5). The CPS for STAR 4 and STAR 10 were 96% and 95% respectively, while the CPS was 100% for both STAR 9 and STAR 11. This is consistent with the aberrant results described above in STAR 4 and STAR 10, where identical V1-V3 *env* sequences could be linked to distinct IS and/or distinct NFL sequences. To investigate whether this is the result of limited genomic variability, the average nucleotide distances of all participants were calculated based on V1-V3 SGS *env* data. This revealed that indeed, participants with a lower CPS displayed a lower nucleotide diversity (Supplemental Table 5).

Overall, we demonstrate that for two out of four participants, the CPS is lower than 100%, leading to inaccuracies when using the V1-V3 *env* to predict clonality of infected cells. Furthermore, we show compartmentalization between the viral populations identified by the V1-V3 *env* SGS method versus the FLIPS method. This could either result from primer bias or from limited sampling depth, leading us to miss a large proportion of viral strains with intact V1-V3 *env* when using FLIPS, when the frequency of the former are less and thus are obscured by *env*-deleted strains.

Rebounding sequences match intact proviruses and proviruses with major deletions or defects in the packaging signal

In our previously conducted HIV-STAR study, proviral V1-V3 *env* SGS sequences from several subsets and anatomical compartments were linked to rebounding plasma sequences (5). Yet, no conclusions about the genomic structure of the NFL proviruses and their associated IS could be inferred, since these subgenomic sequences did not allow for such analysis. The FLIPS and MDA-ISLA data generated in the present study allowed for a deeper characterization of the proviral landscape through linkage of NFL proviral sequences to rebounding plasma sequences.

To determine if the FLIPS- and MDA-derived NFL sequences matched rebound plasma sequences, phylogenetic trees were constructed. In conducting this comparison, all sequences which belonged to an EIS were only included once. All sequences were then trimmed to the V1-V3 *env* region and aligned with the plasma-derived V1-V3 *env* sequences from several timepoints during rebound. Phylogenetic analysis was performed using ML trees constructed via PHYML v3.0 with 1000 bootstraps (Figure 5).

For participants STAR 10 and 4, one or more identical matches between rebound plasma sequences and defective proviral genomes could be observed (Figure 5). In fact, STAR 10 had three matches between rebounding V1-V3 *env* sequences and largely deleted proviruses: one match to a provirus that was sampled only once with FLIPS, hence no IS recovered, and two to proviruses located in the *ZBTB20* gene and in an intergenic region on chromosome 8. STAR 4's plasma V1-V3 *env* sequences from all three timepoints during the ATI matched an NFL provirus with a PSI/MSD deletion. These observations further suggest that subgenomic SGS is unable to distinguish between distinct proviruses, which is reflected by a CPS smaller than 100% in these two participants (Supplemental Table 5).

For participants STAR 9 and STAR 11, a match was found between intact proviruses and rebounding plasma sequences (Figure 5). For STAR 9, a provirus found only once using FLIPS matched plasma sequences found at T2 (3/4 plasma sequences from that timepoint) and T4. For STAR 11, an intact provirus that was found using both FLIPS and MDA-assisted NFL proviral sequencing, could be linked to a plasma virus at T2 (1 out of 3 plasma sequences from that timepoint). This provirus was found to be integrated in the *ZNF141* gene, which belongs to the Krüppel-associated box domain (KRAB) containing zinc finger nuclease family. Interestingly, the same viral sequences were not identified in the plasma from rebounding timepoints T3 and T4.

To investigate how the proviruses that could be linked to rebounding viruses compare to the historic plasma and proviral V1-V3 *env* sequences generated during the original HIV-STAR study, including sequences stemming from different anatomical compartments, the trimmed V1-V3 *env* region from the MDA- and FLIPS- derived NFL sequences were aligned with SGS-derived and MDA-derived V1-V3 *env* sequences. Subsequently, phylogenetic trees were constructed for each participant, where sequences belonging to an EIS including one or more MDA or FLIPS derived V1-V3 *env* sequences were highlighted (Figure 6, Supplemental Figure 5). For STAR 9, the unique intact FLIPS provirus matching T2 and T4 plasma sequences falls within a cluster of proviral peripheral blood and GALT SGS V1-V3 *env* sequences (Figure 6, indicated by black arc). For STAR 10, a match between the *STAT5B* clone and proviral sequences from LN and peripheral blood could be observed, suggesting intermingling between these two compartments (Figure 6, indicated by black arc). For STAR 11, the cluster containing *ZNF141*, which could be linked to potential residual viremia, also matches T0 plasma sequences, suggesting a phylogenetic relationship to the founder virus (Figure 6, indicated by black arc).

In conclusion, by performing MDA-mediated NFL and ISA, we identified several proviruses with matched IS that linked to sequences from plasma before and/or during an ATI. Multiple of these proviruses displayed major defects of the packaging signal, raising the question whether these are still capable of producing viremia. Furthermore, some intact proviral sequences could be linked to multiple anatomical compartments, suggesting that certain clones harboring genome-intact proviruses can traffic between different compartments.

Discussion

Stable integration of HIV-1 genomes into the DNA of host cells leads to the establishment of a persistent HIV-1 latent reservoir. While most of these integrated proviruses are defective, a small proportion are genetically intact and fully capable of producing infectious virions upon latency reversal (7,24,26,36–41). The proportion of genetically intact HIV-1 proviruses, as measured by Intact Proviral DNA Assay (IPDA), has been shown to decay slowly, with an estimated average half-life of 4 years during the first 7 years of suppression, and 18.7 years thereafter (42). This long half-life can in part be explained by continuous clonal expansion of infected cells harboring these genetically intact HIV-1 proviruses (30,43). While this phenomenon is well-established, the contribution of clonally expanded HIV-1 infected cells to refueling viremia upon treatment interruption remains underexplored. Previously, others have tried to characterize rebounding viruses by phylogenetically linking these to proviral sequences and viral sequences obtained by viral outgrowth assays (VOA), with limited success. While two studies were unable to find links between rebounding sequences and viral sequences recovered by VOA (44,45), two other groups did find several links using similar techniques (13,46). However, these latter studies were performed in the context of interventional clinical trials and the IS of these viruses remained unknown. In addition, two groups were able to link proviral sequences to rebound sequences, though only a small part of the proviral genome was queried (21,47). We previously conducted the HIV-STAR clinical study, where SGS on the V1-V3 *env* region was used to link proviral sequences to rebounding plasma sequences (5). We found multiple links between proviral sequences and rebounding plasma sequences, however, this study was limited by the sequencing of a small subgenomic region of the proviruses. In the current study, we used a combination of NFL sequencing, ISA and MDA-mediated IS/NFL sequencing to more accurately define the source of rebounding virus detected during ATI in a subset of HIV STAR participants.

We first showed that large discrepancies exist between different techniques to assess clonal expansion of HIV-1 infected cells. These discrepancies are often the result of primer biases, which dictate which proviruses are amplified. This has important implications for HIV-1 reservoir research, as some assays will be unable to detect potentially relevant proviruses. In addition, we demonstrated that the use of a short subgenomic region of the HIV-1 genome (V1-V3 *env*) to assess clonality of infected cells can lead to inaccurate results. This was shown by the recovery of distinct NFL proviruses, integrated at different sites, displaying identical V1-V3 *env* sequences. Similar observations were made in a recently published study, where P6-PR-RT sequences were compared to matched NFL/IS sequences (30). They found multiple instances of unique proviral P6-PR-RT sequences, with distinct IS. Taken together, we conclude that evaluating clonality of HIV-1 infected cells based on the assessment of a subgenomic region should be done with caution.

We next set out to find links between NFL proviral sequences and sequences found in the plasma during different stages of an ATI. First, we identified several links between defective proviruses and rebounding plasma viruses. Interestingly, for participant STAR 4, a link was found with a provirus containing a small packaging signal deletion. It has been shown previously that proviruses with such defects are still capable of producing infectious virions, though with significantly lower efficiency (48). Therefore, we cannot exclude the possibility that the detected sequences in the plasma at rebound originate from such proviruses. Three other defective proviruses linked to rebound viruses, all in participant STAR 10, contain large internal deletions, making it unlikely that these are the real source of the virus rebounding during ATI. Rather, these are probably related proviruses, as they share an identical V1-V3 *env* sequence. Two previous studies that tried to link proviral sequences to rebound sequences, based on full *env* sequences, concluded that while they were not able to directly link the proviral sequences to the rebounding ones, the rebounding sequences could often be accounted for by recombination (45,46). Because we assessed only a small portion of the *env* gene (V1-V3 region), we were not able to comprehensively study recombination events, though we hypothesize that recombination may be a probable cause of identical overlap between defective proviral sequences and rebounding virus sequences.

We further identified two links between genetically intact NFL proviruses and plasma viruses emerging upon treatment interruption. The first link was found in participant STAR 9, where an intact provirus obtained with FLIPS could be linked to plasma virus at T2 and T4. Because this provirus was not retrieved in an MDA reaction, the IS remains unknown. Interestingly, this virus was first sampled at T2 and persisted into T4, which suggests that this virus emerged during the phase of an ATI when the viral load was still undetectable. In participant STAR 11, an intact provirus integrated in the *ZNF141* gene could be linked to plasma virus at T2 during an ATI. Another recent publication found a clonal infected cell population with IS in the *ZNF721/ABCA11P* gene, that contributed to persistent residual viremia which was not suppressed by ART (20).

This gene is located at the extreme end of chromosome 4, and belongs to the KRAB-containing zinc finger nuclease family. This integration event shows great similarities with the provirus we identified in the *ZNF141* gene, which also belongs to the KRAB-containing zinc finger nuclease family and which is located on chromosome 4, just upstream of the *ZNF721/ABCA11P* gene. Interestingly, three other studies also described infected cell clones harboring a genetically intact provirus integrated in the *ZNF721/ABCA11P* gene, suggesting that this region is a particular hotspot for the persistence of genetically intact proviruses (14, 20, 27). Because the plasma virus that was linked to our ZNF141 clone stems from T2, the latest timepoint with undetectable viral load during the ATI, but did not persist in the later timepoints (T3 and T4), we cannot exclude that the virus we sampled emerged as a result of continuous virus shedding, as described by Halvas *et al.* (20), rather than 'true' rebounding virus. Previously, it was suggested that the origin of rebounding plasma viruses includes clonally expanded infected cells that are transcriptionally active before TI (21). Similarly, a recent study found several overlaps between monotypic low-level residual viremia sequences, which persisted for years, and rebound plasma sequences (19). These two findings, together with the observations by Halvas *et al.* (20), leads to the expectation that the provirus integrated in the *ZNF141* gene is a prime candidate to contribute to viral rebound, however, our current data does not support this. Off course, we cannot exclude that this viral strain was not identified at T3 and T4 because it was obscured by other rebound viruses, causing us to miss it.

In a recent study it was observed that 'elite controllers' (EC), individuals that control HIV-1 infection spontaneously, often carry genetically intact proviral sequences integrated at spots associated with 'deep latency', which persist over time and are not cleared by the immune system (35). In one EC, they described a persistently infected cell population with an intact provirus integrated in the *ZNF274* gene, which is associated with highly condensed chromatin. Interestingly, we also observed a clonally expanded infected cell population in the peripheral blood TEM fraction from STAR11, with a genetically intact provirus integrated in the *ZNF274* gene. Despite the rather large size of the clone, we did not observe the emergence of the corresponding viral sequence in the plasma during the ATI, which is in agreement with its presumed ‘deep latent’ state. In fact, it is possible that because of the heterochromatin state of the DNA at this spot, this provirus would tend to remain latent. Alternatively, we cannot exclude that this virus was not identified during the ATI due to timing of our specimen collection. Indeed, it is possible that this virus would be detected if the treatment interruption would have been prolonged and if the participant was sampled at later time-points, especially knowing that transcription at this specific IS could be diminished and, if possible at all, would need more time to complete. These findings add to the current understanding that not all genetically intact proviral sequences contribute to the ‘replication competent HIV-1 viral reservoir’, as some are unlikely to rebound due to an unfavorable IS, though they may possess all the necessary attributes to rebound under specific conditions.

We acknowledge several limitations in this study. The first one is the limited sampling from tissue compartments, possibly causing us to miss important rebound lineages. Indeed, it has been shown that tissues, including lymph nodes and GALT, harbor most of the HIV-1 latent reservoir, orders of magnitude higher than the peripheral blood compartment (49). Whether there is compartmentalization between different anatomical compartments is under debate. Several studies, including our previously conducted HIV-STAR study, have suggested that there is limited compartmentalization between the HIV-1 proviral sequences recovered from lymph nodes and from peripheral blood (5,23,45,50,51), based on identical proviral sequences and/or IS shared between both compartments. In contrast, another recently published study reports partial compartmentalization between lymph nodes and peripheral blood when specifically enriching for tissue resident CD4+ T cells, based on IS sequencing results (52). In addition, our previous HIV-STAR study did not show evidence of any enrichment of rebounding sequences stemming from specific anatomical compartments (5), justifying our decision to focus the current study primarily on the peripheral blood compartment. The second limitation of the current study is that the link to plasma rebounding sequences is based on the V1-V3 *env* region, rather than on plasma NFL sequences. This means that we cannot exclude the possibility that links between proviral sequences and rebounding plasma sequences are the result of false V1-V3 *env* matches, however the CPS for the V1-V3 *env* region for participants STAR 9 and STAR 11, which display links between intact proviral sequences and plasma rebound sequences, was calculated at 100%.

In conclusion, our data show that reservoir characterization using multiple methods, including ISA, NFL proviral sequencing and a combination of both, one can identify matches between proviral sequences and plasma sequences emerging during an ATI, however these matches are rare. We report a link between a genome-intact provirus integrated in the *ZNF141* gene and a plasma sequence recovered during an ATI. This finding further adds to the body of evidence that genes of the KRAB-containing zinc finger nucleases are a particular hotspot for persistence of genetically intact proviruses (20,35).

Special focus on this class of genes, and the proviruses integrated within, will be needed in future studies to elucidate their role in reservoir persistence.

Methods

Samples

A total of four HIV-1 infected, ART treated participants were included in this study. All had an undetectable viral load (<20 copies/ml) for at least 1 year prior to treatment interruption, and all initiated ART during the chronic phase of infection. The participants characteristics are summarized in Supplemental Table 6. Participants were sampled longitudinally, prior to and during an ATI (Figure 1B). Anatomical compartments that were sampled, and corresponding cell subsets sorted from these, are summarized in Supplemental Table 1.

CD4+ T cell subset sorting

Cryopreserved PBMCs were thawed and CD4+ T cell enrichment was carried out with negative magnet-activated cell sorting (Beckton Dickinson, BD IMag™, Cat. No. 557939). CD4+ T cells were stained with the following monoclonal antibodies: CD3 (Becton Dickinson, Cat. No. 564465), CD8 (Becton Dickinson, Cat. No. 557746), CD45RO (Becton Dickinson, Cat. No. 555493), CD27 (Becton Dickinson, Cat. No. 561400), CCR7 (Becton Dickinson, Cat. No. 560765) and a fixable viability stain (Becton Dickinson, Cat. No. 565388). Fluorescence-activated cell sorting was used to sort stained peripheral blood-derived CD4+ T cells into naïve CD4+ T cells (CD45RO-, CD45RA+), central memory CD4+ T cells (CD3+ CD8- CD45RO+ CD27+), transitional memory CD4+ T cells (CD3+ CD8- CD45RO+, CD27+ CCR7-) and effector memory CD4+ T cells (CD3+ CD8- CD45RO+ CD27-), GALT cells into CD45+ cells and cells from lymph nodes into central memory CD4+ T cells (CD3+ CD8- CD45RO+ CD27+) and effector memory CD4+ T cells (CD3+ CD8- CD45RO+ CD27-), using a BD FACSJazz cell sorter machine, as previously described (5). A small fraction of each sorted cell population was analyzed by flow cytometry to check for purity, which was over 95% on average. Flow cytometry data was analyzed using FlowJo software (Tree-Star).

Droplet digital PCR (ddPCR)

Sorted cells were pelleted and lysed in 100µL lysis buffer (10mM TRisHCl, 0.5% NP-40, 0.5% Tween-20 and proteinase K at 20mg/ml) by incubating for 1 hour at 55°C and 15 min at 85°C. HIV-1 copy number was determined by a total HIV-1 DNA assay on droplet digital PCR (Bio-Rad, QX200 system), as described previously (53). PCR amplification was carried out with the following cycling program: 10 min at 98°C; 45 cycles (30 sec at 95°C, 1 min at 58°C); 10 min at 98°C. Droplets were read on a QX200 droplet reader (Bio-Rad). Analysis was performed using ddpcRquant software (54).

Whole genome amplification (WGA)

Cell lysates were diluted according to ddPCR HIV-1 copy quantification, so that less than 30% of reactions contained a single proviral genome. Whole genome amplification was performed by multiple displacement amplification with the REPLI-g single cell kit (Qiagen, Cat. No. 150345), according to manufacturer's instructions. The resulting amplification product was split for downstream ISA, single genome/proviral sequencing, and, for selected reactions, near full-length HIV-1 sequencing.

Single genome/proviral sequencing

Single genome/proviral sequencing (SGS) of the V1-V3 region of *env* was performed as described before (55,56), with a few adaptations. The amplification consists of a nested PCR with the following primers: Round 1 forward (E20) 5’-GGGCCACACATGCCTGTGTACCCACAG-3’ and reverse (E115) 5’-AGAAAAATTCCCCTCCACAATTAA-3’; round 2, forward (E30) 5’-GTGTACCCACAGACCCCAGCCCACAAG-3’ and reverse (E125) 5’-CAATTTCTGGGTCCCCTCCTGAGG-3’. The 25 µL PCR mix for the first round is composed of: 5 µL 5X Mytaq buffer, 0.375 µL Mytaq polymerase (Bioline, Cat. No. BIO-21105), 400 nM forward primer, 400 nM reverse primer and 1 µL REPLI-g product. The mix for the second round has the same composition and takes 1 µL of the first-round product as an input. Thermocycling conditions for first and second PCR rounds are as follows: 2 min at 94°C; 35 cycles (30 sec at 94°C, 30 sec at 60°C, 1 min at 72°C); 5 min at 72°C. Resulting amplicons were visualized on a 1% agarose gel and Sanger sequenced (Eurofins Genomics, Ebersberg, Germany) from both ends, using second round PCR primers.

Integration site loop amplification (ISLA)

Integration site sequencing was carried out by integration site loop amplification (ISLA), as described by Wagner *et al.* (8), but with a few modifications. Firstly, the *env* primer used during the linear amplification step was omitted, as it was not necessary to recover the *env* portion of the provirus at a later stage. Therefore, the reaction was not split after the linear amplification, and the entire reaction was used as an input into subsequent decamer binding and loop formation. For some proviruses, an alternative set of primers were used to retrieve the IS from the 5' end (Supplemental Table 7). Resulting amplicons were visualized on a 1% agarose gel and positives were sequenced by Sanger sequencing. Analysis of the generated sequences was performed using the 'Integration Sites' webtool developed by the Mullins lab; https://indra.mullins.microbiol.washington.edu/integrationsites/.

Full-length individual proviral sequencing assay

Proviral sequences from the genomic DNA of sorted subsets were recovered by the Full-length Individual Proviral Sequencing (FLIPS) assay as first described by Hiener *et al.* (28) with some minor alterations. Briefly, the assay consists of two rounds of nested PCR at an end-point dilution where 30% of the wells are positive. This yields proviral fragments of up to 9 kb using the following primers for the first round BLOuterF (5'-AAATCTCTAGCAGTGGCGCCCGAACAG-3') and BLOuterR (5'-TGAGGGATCTCTAGTTACCAGAGTC-3') followed by a second round using primers 275F (5'-ACAGGGACCTGAAAGCGAAAG-3') and 280R (5'-CTAGTTACCAGAGTCACACAACAGACG-3'). The cycling conditions are 94°C for 2 m; then 94°C for 30 s, 64°C for 30 s, 68°C for 10 m for 3 cycles; 94°C for 30 s, 61°C for 30 s, 68°C for 10 m for 3 cycle; 94°C for 30 s, 58°C for 30 s, 68°C for 10 m for 3 cycle; 94°C for 30 s, 55°C for 30 s, 68°C for 10 m for 21 cycle; then 68°C for 10 m. For the second round, 10 extra cycles at 55°C are included. The PCR products were visualized using agarose gel electrophoresis. Amplified proviruses from positive wells were cleaned using AMPure XP beads (Beckman Coulter), followed by a quantification of each cleaned provirus with Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen). Next, an NGS library preparation using the Nextera XT DNA Library Preparation Kit (Illumina) with indexing of 96-samples per run was used according to the manufacturer's instructions, except that input and reagents volumes were halved and libraries were normalized manually. The pooled library was sequenced on a MiSeq Illumina platform via 2x150 nt paired-end sequencing using the 300 cycle v2 kit.

Near full-length provirus amplification from MDA reactions

MDA reactions containing a potentially clonal proviral sequence were subjected to near full-length proviral sequencing, using either a single-amplicon approach (24), a four-amplicon approach (30), or a five-amplicon approach (14), as previously described. In case of the multiple amplicon approaches, amplicons were pooled equimolarly and sequenced as described above.

De Novo assembly of HIV-1 proviruses and analysis

The generated sequencing data from either FLIPS or multiple amplicon approaches was demultiplexed and used to *de novo* assemble individual proviruses using a custom inhouse pipeline. In short, the workflow consists of following steps: (i) check of sequencing quality for each library using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and removal of Illumina adaptor sequences and trimming of 5' and 3' terminal ends. (ii) The trimmed reads are fed to the MEGAHIT (57) *de novo*-assembler generating multiple contigs for each library. (iii) Per library, all *de novo* contigs were checked using blastn against the HXB2 reference virus as a filter to exclude non-HIV-1 contigs in the following analysis steps. (iv)

Subsequently, the trimmed reads were mapped against the *de novo* assembled HIV-1 contigs to enable the calling of the final majority consensus sequence of each provirus. Alignments of proviral sequences for each participant were made via MAFFT (58) and manually inspected via MEGA7 (59). The generated HIV-1 proviruses were categorized as intact or defective as described previously (24). Phylogenetic trees were constructed using PhyML v3.0 (60) (best of NNI and SPR rearrangements) and 1000 bootstraps. MEGA7 (59) and iTOL v5 (61) were used to visualise phylogenetic trees.

Statistical analysis

P-values in figure 2A test for a difference in the proportion of unique IS between TCM and TEM. P-values were calculated using “prop.test” command in R versions 3.6.2 (62). Infection frequencies for FLIPS data were calculated by expressing the total number of identified HIV positive cells as a proportion of all cells analysed. The infection frequency was compared across cellular subsets using a logistic regression on the number of cells positive for HIV and total number of cells using “glm” function in R. Interaction between participant and cellular subset was detected (P < 0.001) and included in the logistic regression. P-values were calculated using the “Anova” function from the “car” package in R (63).

Data availability statement

Data will be uploaded to public repositories upon acceptance of the manuscript.

Study approval

This study was approved by the Ethics Committee of the Ghent University Hospital (Belgian registration number: B670201525474). Written informed consent was obtained from all study participants.

Author contributions

BC, LL, LF, SP and LV conceptualized the experiments. MADS processed the samples from the initial HIV STAR study, including cell isolation from peripheral blood and tissue, and she performed cell sorting and single-genome sequencing. BC and YN performed experiments involving cell sorting, multiple displacement amplification, single-genome sequencing and integration site sequencing. LL and ZB performed experiments involving near full-length proviral sequencing. BC, LL, BV, JSE and TS analyzed data and performed associated analyses. BC, LL, TS and BV made figures and tables. BC and LL wrote the manuscript. All co-authors edited and approved the manuscript.

Acknowledgements and funding sources

We would like to acknowledge and thank all participants who donated samples to the HIV-STAR study, and all the MDs and study nurses that assisted with the sample collection. We would also like to thank Marion Pardons, Tine Struyve and Sofie Rutsaert for providing guidance during initial data analyses, for the constructive discussions and for critically reading the manuscript. We are grateful for the discussions with and input from James Mullins, Rafick Sékaly, Susan Pereira Ribeiro, Hadega Aamer, Sam Kint, Oleg Denisenko, Katie Fisher and Bethany Horsburgh. In addition, we would like to thank Kim De Leeneer, Céline Helsmoortel and Bram Parton for their assistance in performing MiSeq sequencing at UZ Ghent. This current research work was supported by the NIH (R01-AI134419, MPI: LV and LF) and the Research Foundation Flanders (S000319N and G0B3820N). LV was supported by the Research Foundation Flanders (1.8.020.09.N.00) and the Collen-Francqui Research Professor Mandate. SP was supported by the Delaney AIDS Research Enterprise (DARE) to Find a Cure (1U19AI096109 and 1UM1AI126611-01) and the Australian National Health and Medical Research Council (APP1061681 and APP1149990). The sample collection at UZ Ghent was supported by an MSD investigator grant (ISS 52777). BC and LL were supported by FWO Vlaanderen (1S28918N, 1S29220N). BV was supported by a postdoctoral grant (12U7121N) of the Research Foundation - Flanders (Fonds voor Wetenschappelijk Onderzoek).

Competing interests

The authors declare that no conflict of interest exists.

References

1. Chun TW, et al. Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. *Proc Natl Acad Sci U S A*. 1997;94(24):13193-13197.
2. Chun T, Fauci AS. Latent reservoirs of HIV : Obstacles to the eradication of virus. *Proc Natl Acad Sci U S A.* 1999;96(20):10958-10961.
3. Finzi D, et al. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science*. 1997;278(5341):1295-1300.
4. Chun TW, et al. Early establishment of a pool of latently infected, resting CD4(+) T cells during primary HIV-1 infection. *Proc Natl Acad Sci U S A*. 1998;95(15):8869-8873.
5. De Scheerder M-A, et al. HIV Rebound Is Predominantly Fueled by Genetically Identical Viral Expansions from Diverse Reservoirs. *Cell Host Microbe*. 2019;26(3):347-358.
6. Wang Z, et al. Expanded cellular clones carrying replication-competent HIV-1 persist, wax, and wane. *Proc Natl Acad Sci U S A*. 2018;115(11):E2575-E2584.
7. Simonetti FR, et al. Clonally expanded CD4+ T cells can produce infectious HIV-1 in vivo. *Proc Natl Acad Sci U S A*. 2016;113(7):1883-1888.
8. Wagner TA, et al. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science*. 2014;345(6196):570-573.
9. Maldarelli F, et al. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science*. 2014;345(6193):179-183.
10. Cohn LB, et al. HIV-1 integration landscape during latent and active infection. *Cell*. 2015;160(3):420-432.

11. Boritz EA, et al. Multiple Origins of Virus Persistence during Natural Control of HIV Infection. *Cell*. 2016;166(4):1004-1015.

12. Hosmane NN, et al. Proliferation of latently infected CD4 $^{+}$ T cells carrying replication-competent HIV-1: Potential role in latent reservoir dynamics. *J Exp Med*. 2017;214(4):959-972.

13. Salantes DB, et al. HIV-1 latent reservoir size and diversity are stable following brief treatment interruption. *J Clin Invest*. 2018;128(7):3102-3115.

14. Einkauf K, et al. Distinct chromosomal positions of intact HIV-1 proviruses. *J Clin Invest*. 2018;129(3):988-998.

15. Brennan TP, et al. Analysis of Human Immunodeficiency Virus Type 1 Viremia and Provirus in Resting CD4+ T Cells Reveals a Novel Source of Residual Viremia in Patients on Antiretroviral Therapy. *J Virol*. 2009;83(17):8470-8481.

16. Bailey JR, et al. Residual Human Immunodeficiency Virus Type 1 Viremia in Some Patients on Antiretroviral Therapy Is Dominated by a Small Number of Invariant Clones Rarely Found in Circulating CD4+ T Cells. *J Virol*. 2006;80(13):6441-6457.

17. Wagner TA, et al. An increasing proportion of monotypic HIV-1 DNA sequences during antiretroviral treatment suggests proliferation of HIV-infected cells. *J Virol*. 2013;87(3):1770-1778.

18. Tobin NH, et al. Evidence that low-level viremias during effective highly active antiretroviral therapy result from two processes: expression of archival virus and replication of virus. *J Virol*. 2005;79(15):9625-9634.

19. Aamer HA, et al. Cells producing residual viremia during antiretroviral treatment appear to contribute to rebound viremia following interruption of treatment. 2020:1-

24.

20. Halvas EK, et al. HIV-1 viremia not suppressible by antiretroviral therapy can originate from large T cell clones producing infectious virus. *J Clin Invest*. 2020;130(11) 5847-5857.

21. Kearney MF, et al. Origin of Rebound Plasma HIV Includes Cells with Identical Proviruses That Are Transcriptionally Active before Stopping of Antiretroviral Therapy. *J Virol*. 2016;90(3):1369-1376.

22. Lu C, et al. Relationship between intact HIV-1 proviruses in circulating CD4 + T cells and rebound viruses emerging during treatment interruption. *Proc Natl Acad Sci U S A*. 2018;115(48):11341-11348.

23. Von Stockenstrom S, et al. Longitudinal Genetic Characterization Reveals That Cell Proliferation Maintains a Persistent HIV Type 1 DNA Pool During Effective HIV Therapy. *J Infect Dis*. 2015;212(4):596-607.

24. Hiener B, et al. Identification of Genetically Intact HIV-1 Proviruses in Specific CD4+T Cells from Effectively Treated Participants. *Cell Rep*. 2017;21(3):813-822.

25. Cohn LB, et al. HIV-1 integration landscape during latent and active infection. *Cell*. 2015;160(3):420-432.

26. Lee GQ, et al. Clonal expansion of genome-intact HIV-1 in functionally polarized Th1 CD4+ T cells. *J Clin Invest*. 2017;127(7):2689-2696.

27. Pinzone MR, et al. Longitudinal HIV sequencing reveals reservoir expression leading to decay which is obscured by clonal expansion. *Nat Commun*. 2019;10(1): 728.

28. Laskey SB, et al. Evaluating Clonal Expansion of HIV-Infected Cells: Optimization

of PCR Strategies to Predict Clonality. Douek DC, ed. *PLOS Pathog*. 2016;12(8):e1005689.

29. Lambrechts L, et al. Emerging PCR-Based Techniques to Study HIV-1 Reservoir Persistence. *Viruses*. 2020;12(2):1-12.

30. Patro SC, et al. Combined HIV-1 sequence and integration site analysis informs viral dynamics and allows reconstruction of replicating viral ancestors. *Proc Natl Acad Sci U S A*. 2019;116(51):25891-25899.

31. Clarridge KE, et al. Effect of analytical treatment interruption and reinitiation of antiretroviral therapy on HIV reservoirs and immunologic parameters in infected individuals. *PLoS Pathog*. 2018;14(1):e1006792.

32. Garner SA, et al. Interrupting antiretroviral treatment in HIV cure research : scientific and ethical considerations. *J Virus Erad*. 2017;3(2):82-84.

33. Ikeda T, et al. Recurrent HIV-1 integration at the BACH2 locus in resting CD4+ T cell populations during effective highly active antiretroviral therapy. *J Infect Dis*. 2007;195(5):716-725.

34. Cesana D, et al. HIV-1-mediated insertional activation of STAT5B and BACH2 trigger viral reservoir in T regulatory cells. *Nat Commun*. 2017;8(1):498.

35. Jiang C, et al. Distinct viral reservoirs in individuals with spontaneous control of HIV-1. *Nature*. 2020;585(7824):261–267.

36. Ho Y-C, et al. Replication-Competent Noninduced Proviruses in the Latent Reservoir Increase Barrier to HIV-1 Cure. *Cell*. 2013;155(3):540-551.

37. Bruner KM, et al. Defective proviruses rapidly accumulate during acute HIV-1 infection. *Nat Med*. 2016;22(9):1043-1049.

38. Pollack RA, et al. Defective HIV-1 Proviruses Are Expressed and Can Be Recognized by Cytotoxic T Lymphocytes, which Shape the Proviral Landscape. *Cell Host Microbe*. 2017;21(4):494-506.

39. Bui JK, et al. Proviruses with identical sequences comprise a large fraction of the replication-competent HIV reservoir. Ross SR, ed. *PLOS Pathog*. 2017;13(3):e1006283.

40. Bruner KM, et al. A quantitative approach for measuring the reservoir of latent HIV-1 proviruses. *Nature*. 2019;566(7742):120-125.

41. Coffin JM, et al. Clones of infected cells arise early in HIV-infected individuals. *JCI Insight*. 2019;4(12):e128432.

42. Peluso MJ, et al. Differential decay of intact and defective proviral DNA in HIV-1-infected individuals on suppressive antiretroviral therapy. *JCI Insight*. 2020;5(4):e132997.

43. Liu R, et al. The forces driving clonal expansion of the HIV-1 latent reservoir. *Virol J*. 2020;17(1):4.

44. Lu C-L, et al. Relationship between intact HIV-1 proviruses in circulating CD4 + T cells and rebound viruses emerging during treatment interruption. *Proc Natl Acad Sci*. 2018;115(48):E11341-E11348.

45. Vibholm LK, et al. Characterization of Intact Proviruses in Blood and Lymph Node from HIV-Infected Individuals Undergoing Analytical Treatment Interruption. *J Virol*. 2019;93(8):e01920-18.

46. Cohen YZ, et al. Relationship between latent and rebound viruses in a clinical trial of anti – HIV-1 antibody 3BNC117. *J Exp Med*. 2018;215(9):2311-2324.

47. Barton K, et al. Broad activation of latent HIV-1 in vivo. 2016;7(12731):1-8.

48. Pollack RA, et al. Defective HIV-1 Proviruses Are Expressed and Can Be Recognized by Cytotoxic T Lymphocytes, which Shape the Proviral Landscape. *Cell Host Microbe*. 2017;21(4):494-506.e4.

49. Estes JD, et al. Defining total-body AIDS-virus burden with implications for curative strategies. *Nat Med*. 2017;23(11):1271-1276.

50. Mcmanus WR, et al. HIV-1 in lymph nodes is maintained by cellular proliferation during antiretroviral therapy Graphical abstract. *J Clin Invest*. 2019;129(11):4629-4642.

51. Josefsson L, et al. The HIV-1 reservoir in eight patients on long-term suppressive antiretroviral therapy is stable with few genetic changes over time. *Proc Natl Acad Sci U S A*. 2013;110(51):E4987-96.

52. Wu VH, et al. Assessment of HIV-1 integration in tissues and subsets across infection stages Find the latest version : *JCI Insight*. 2020;5(20):139783.

53. Rutsaert S, et al. Evaluation of HIV-1 reservoir levels as possible markers for virological failure during boosted darunavir monotherapy. *J Antimicrob Chemother*. 2019;74(10):3030-3034.

54. Trypsteen W, et al. ddpcRquant: threshold determination for single channel droplet digital PCR experiments. *Anal Bioanal Chem*. 2015;407(19):5827-5834.

55. Josefsson L, et al. Hematopoietic Precursor Cells Isolated From Patients on Long-term Suppressive HIV Therapy Did Not Contain HIV-1 DNA. *J Infect Dis*. 2012;206(1):28-34.

56. Von Stockenstrom S, et al. Longitudinal Genetic Characterization Reveals That Cell

Proliferation Maintains a Persistent HIV Type 1 DNA Pool during Effective HIV Therapy. *J Infect Dis*. 2015;212(4):596-607.

57. Li D, et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*. 2015;31(10):1674-1676.

58. Katoh K, et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 2002;30(14):3059-3066.

59. Kumar S, et al. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol*. 2016;33(7):1870-1874.

60. Guindon S, et al. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst Biol*. 2010;59(3):307-321.

61. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res*. 2019;47(W1):W256-W259.

62. “R Core Team.” R: A language and environment for statistical computing. 2020. https://www.r-project.org/.

63. John F, Sanford W. *An R Companion to Applied Regression*. Third edit. Thousand Oaks CA: Sage; 2019. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

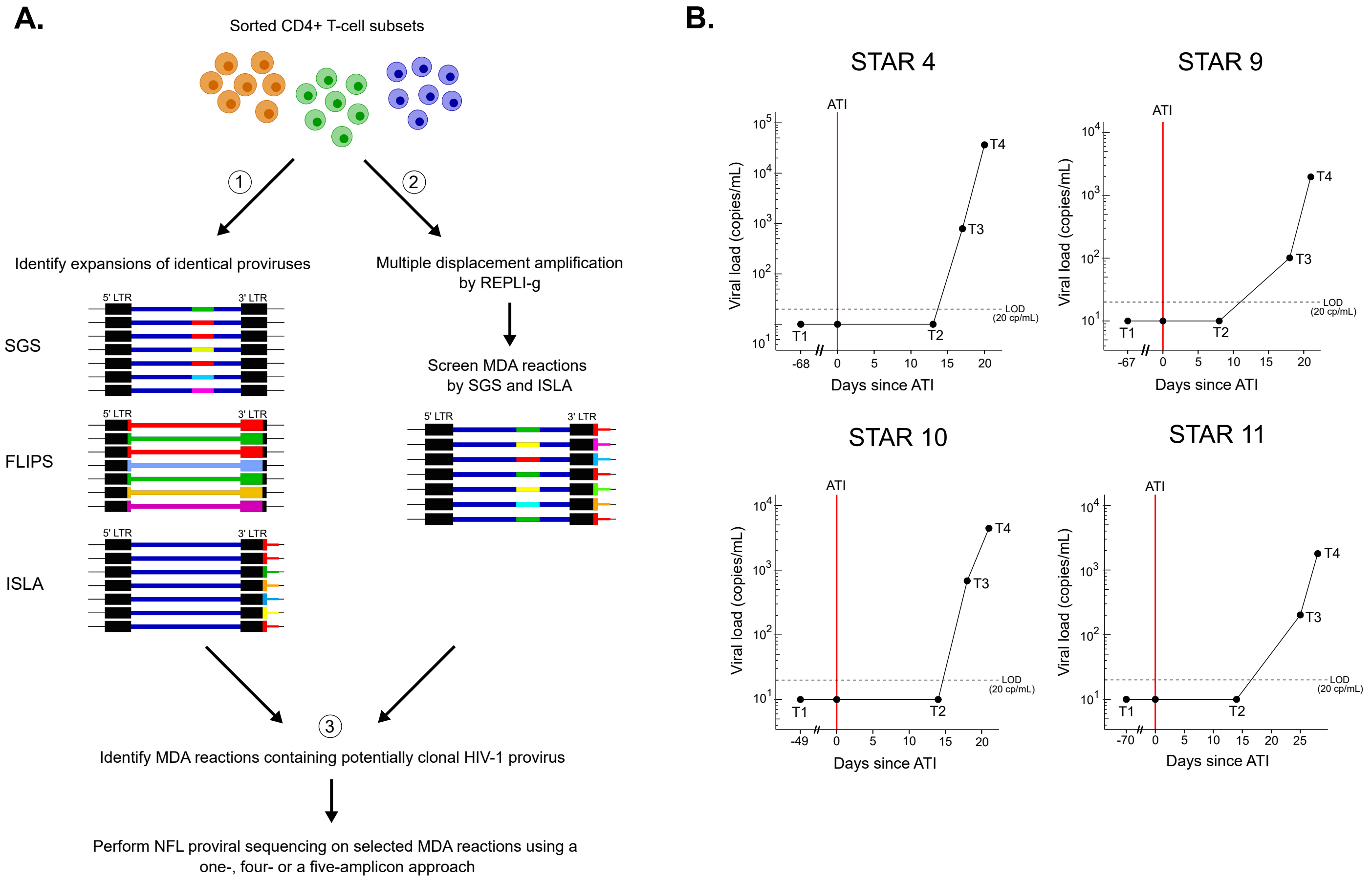
A.
Sorted CD4+ T-cell subsets
1
2
Identify expansions of identical proviruses
Multiple displacement amplification by REPLI-g
SGS
FLIPS
ISLA
5' LTR
3' LTR
Screen MDA reactions by SGS and ISLA
3
Identify MDA reactions containing potentially clonal HIV-1 provirus
Perform NFL proviral sequencing on selected MDA reactions using a one-, four- or a five-amplicon approach
B.
STAR 4
STAR 9
STAR 10
STAR 11
ATI
T1
T2
T3
T4
LOD (20 cp/mL)
Viral load (copies/mL)
Days since ATI
-68
-67
-49
-70
0
5
10
15
20
25
10^1
10^2
10^3
10^4
10^5

Figure 1: Overview of the workflow for deep HIV-1 reservoir characterization and viral loads at each timepoint of sample collection for all participants. (A) Workflow of deep HIV-1 reservoir characterization by single genome sequencing (SGS), full-length individual proviral sequencing (FLIPS), integration site loop amplification (ISLA) and multiple displacement amplification (MDA). In a first step, potentially clonal HIV-1 infected cells were identified by SGS, FLIPS and ISLA at the bulk level, on lysed sorted CD4+ T-cell subsets. In a second step, MDA with subsequent SGS and ISLA was performed on selected sorted cell lysates. In the final step, MDA reactions containing a potentially clonal provirus were identified and the NFL genome of the according provirus was amplified and sequenced. (B) Viral load (copies/mL) at each time of sample collection for all participants. The day of ATI initiation is indicated with a vertical red line. The plasma was sampled during ART (time point 1, T1), 8 to 14 days after ATI (time point 2, T2), at the first detectable viral load (time point 3, T3), and at rebound (time point 4, T4). Note that T1 is not shown to scale. The horizontal dashed lines indicate the limit of detection at 20 copies/mL. ATI = analytical treatment interruption.

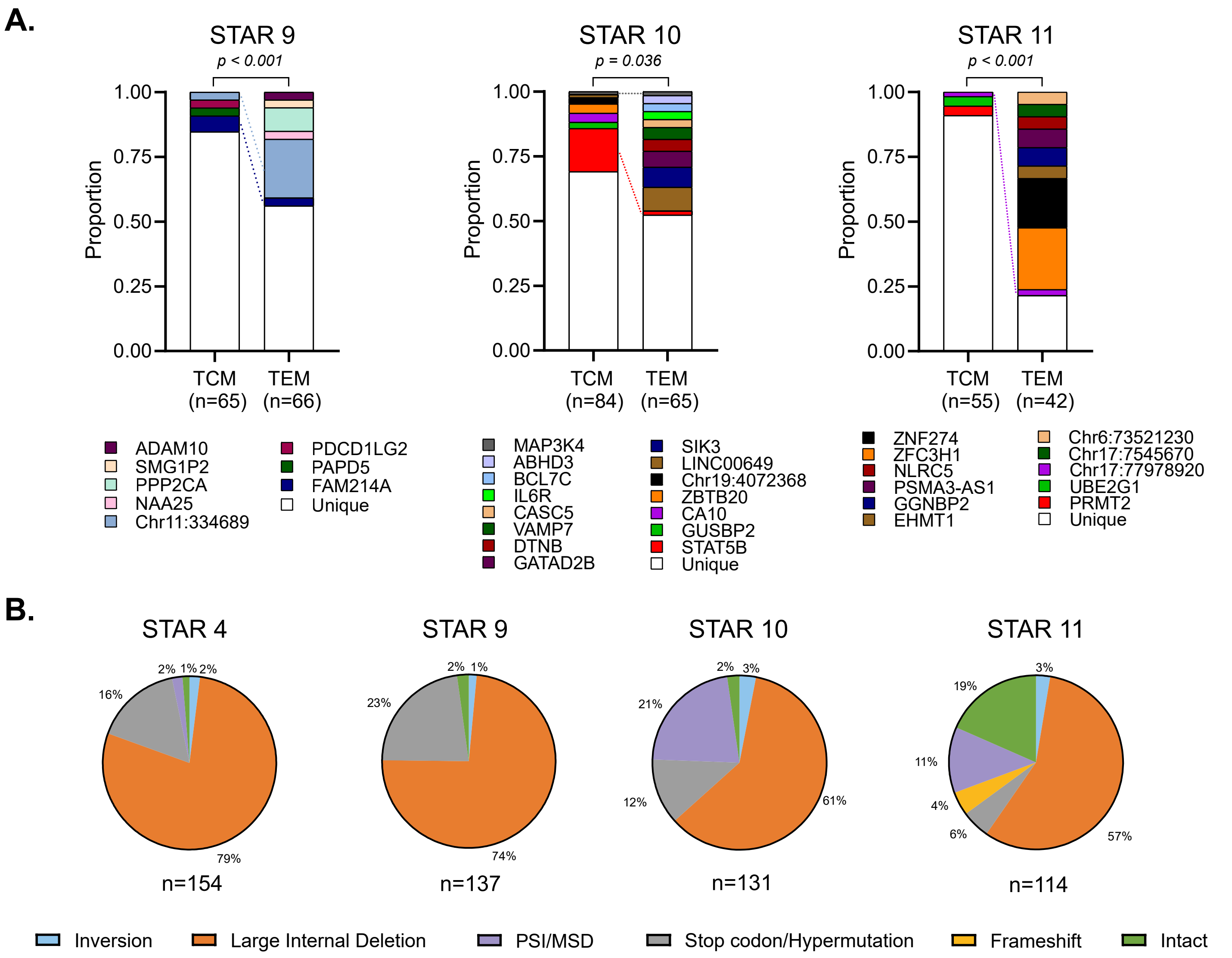

Figure 2: Clonal HIV-1 integration sites and proviral near full-length genome sequences per category from different participants across different cell subsets before ATI. (A) Proportions of retrieved integration sites (IS) by ISLA for participants STAR 9, STAR 10 and STAR 11 from TCM and TEM subsets from peripheral blood. IS found more than once are shown as colored proportions and represent clonally expanded HIV-1 infected cells. Identical IS found in both subsets are linked with dashed lines. P-values test was used for a difference in the proportion of unique IS between TCM and TEM. ISLA = integration site loop amplification, TCM = central memory T cell, TEM = effector memory T cell. (B) Proportions of intact and defective near full-length sequences from FLIPS within all sequenced proviruses from peripheral blood, GALT and lymph nodes for each participant. FLIPS = Full-Length Individual Provirus sequencing, GALT = gut-associated lymphoid tissue, PSI = packaging signal, MSD = major splice donor.

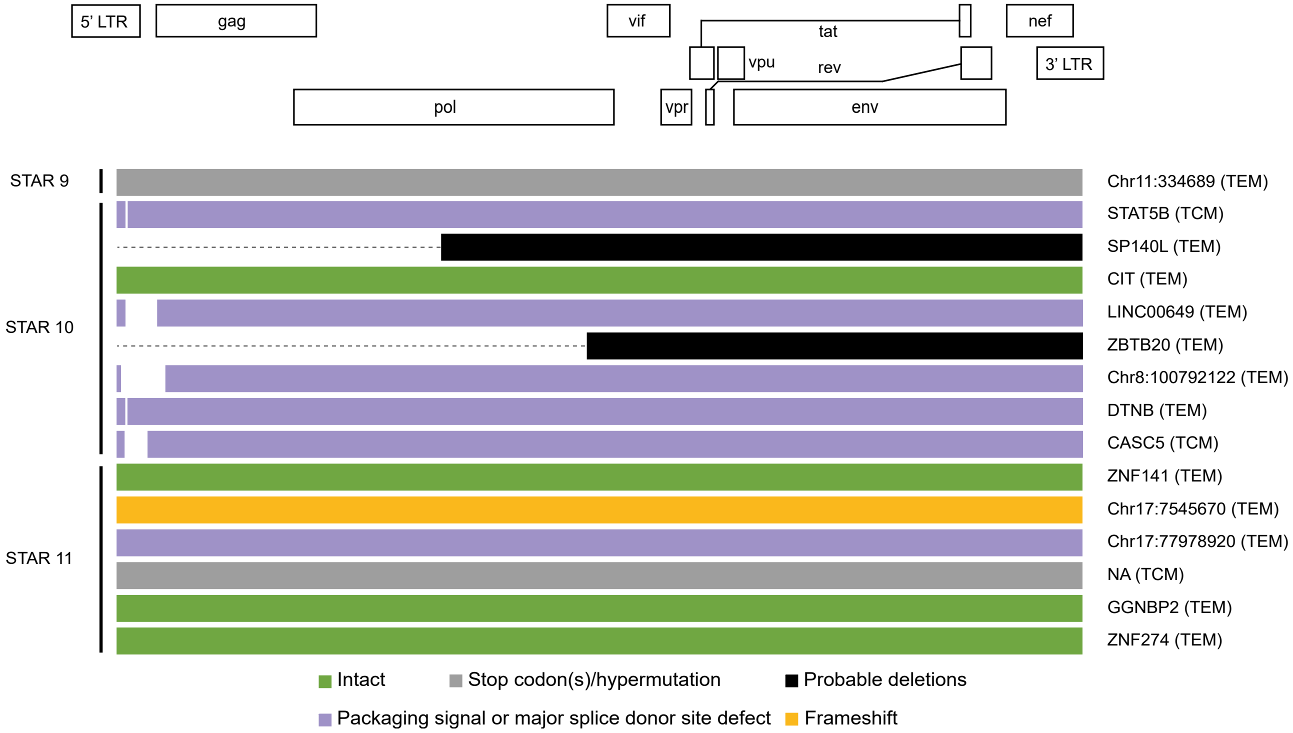

Figure 3: Near full-length proviral HIV-1 genomes and associated integration sites recovered from the peripheral blood by MDA. For each participant, the recovered proviral genome structures are shown aligned to the HXB2 reference sequence and corresponding integration sites, if available, are listed on the right hand side, together with the memory subset between brackets. For two proviruses (*SP140L* and *ZBTB20*) no near full-length genomes could be retrieved despite multiple attempts (Supplemental Table 4). The regions that could not be recovered are indicated by a dashed line. MDA = multiple displacement amplifcation, TCM = central memory T cell, TEM = effector memory T cell, NA = not available.

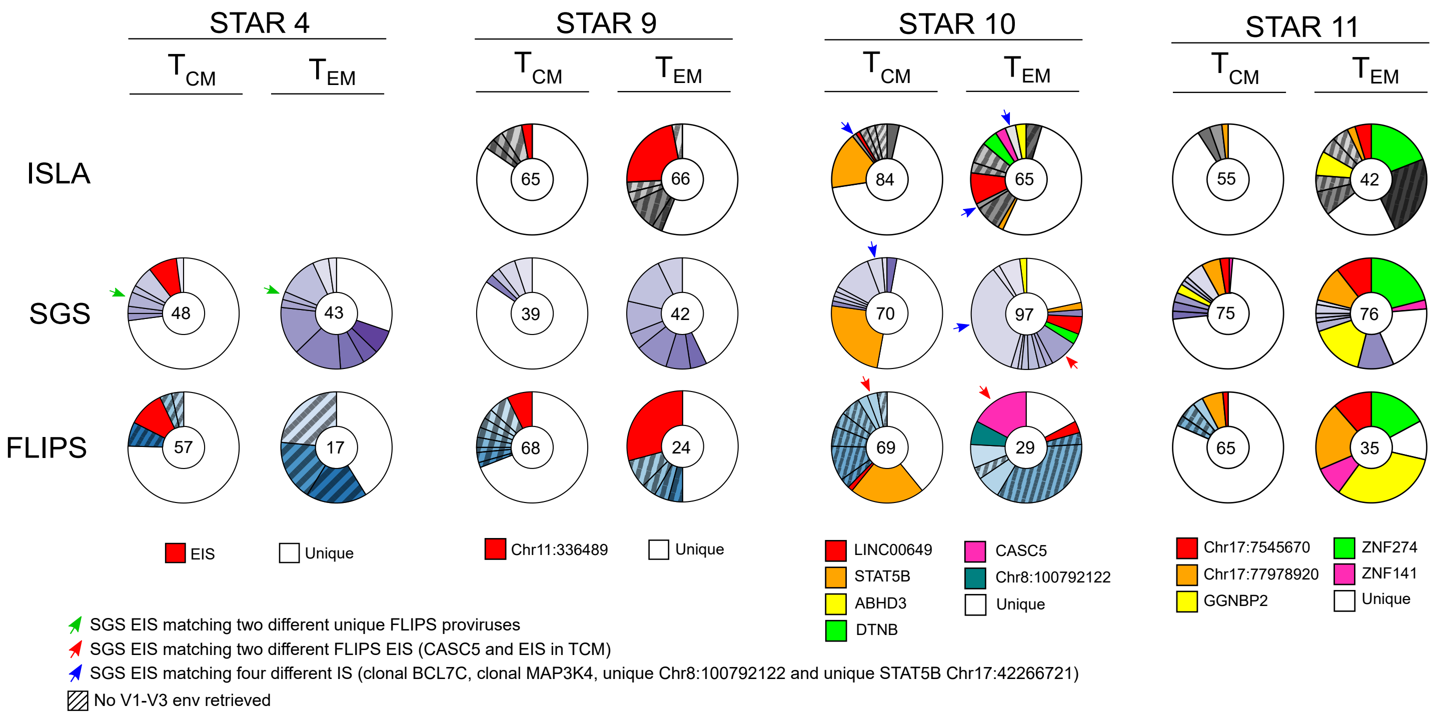

Figure 4: Comparison of assays to identify potentially clonal HIV-1 infected cell populations. The total number of examined integration sites (IS), V1-V3 *env* sequences and near full-length proviral (NFL) sequences is noted in the middle of each donut plot. Sequences found multiple times within the same assay are colored by a shade of grey, purple or blue (for ISLA, SGS and FLIPS respectively). When NFL or V1-V3 *env* sequences overlapped between assays, they were given a distinct standout color, and these are named in the legend. Populations of identical FLIPS or ISLA sequences that are not associated with a V1-V3 *env* sequence (due to deletions and/or primer mismatches) are shaded. Arrows are used to indicate discrepancies between the different assays. ISLA = integration site loop amplification, SGS = single-genome sequencing, FLIPS = Full-Length Individual Provirus sequencing, EIS = expansion of identical sequences.

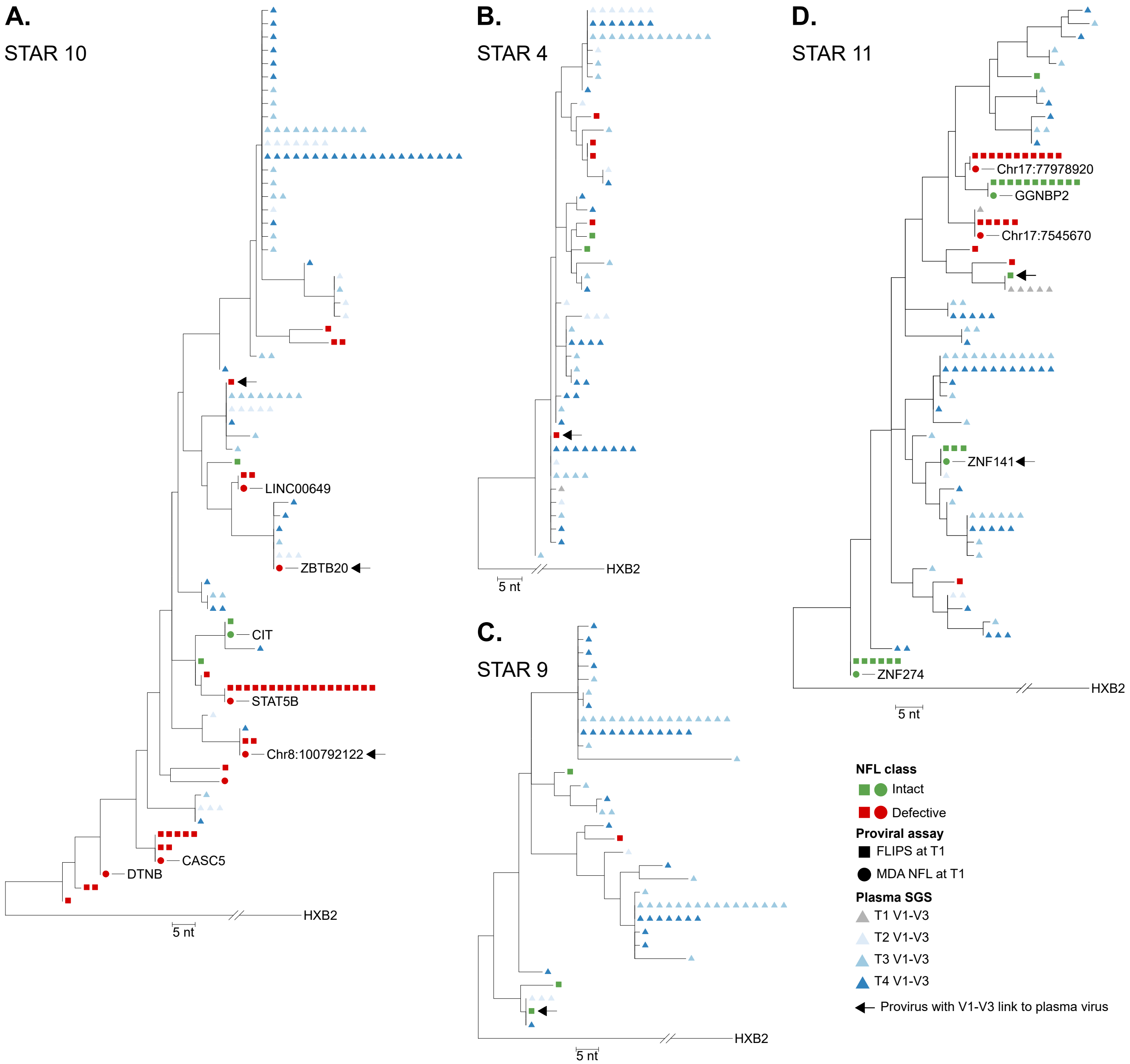

Figure 5: Maximum-likelihood phylogenetic trees of V1-V3 *env* sequences derived from FLIPS- and MDA-derived intact and defective proviral sequences before ATI and rebounding plasma viruses during different stages of ATI. Proviral sequences derived from FLIPS and MDA are shown as squares and circles respectively. The integration sites associated with MDA-derived proviruses are noted if available. Plasma sequences are shown as triangles where the colour indicates the timepoint during ATI. Arrows indicate identical matches between proviral and plasma V1-V3 *env* sequences. All trees are rooted to the HXB2 reference sequence. (A) In participant STAR 10, three identical matches between defective proviral and plasma rebound sequences were found. For two, the corresponding IS *ZBTB20* and *Chr8:100792122* could be recovered. (B) In participant STAR 4, only one match between a unique MSD deleted provirus and plasma sequences was observed. (C) In STAR 9, a match between a unique intact provirus and multiple plasma sequences from different timepoints were found. (D) In STAR 11, a rebounding plasma sequence could be linked to an expansion of identical intact NFL genomes located in the *ZNF141* gene. One unique intact provirus can be linked to a residual plasma sequence from T1. FLIPS = Full-Length Individual Provirus sequencing, MDA = multiple displacement amplifcation, ATI = analytical treatment interruption, IS = integration site, MSD = major splice donor, SGS = single-genome sequencing, NFL = near full-length.

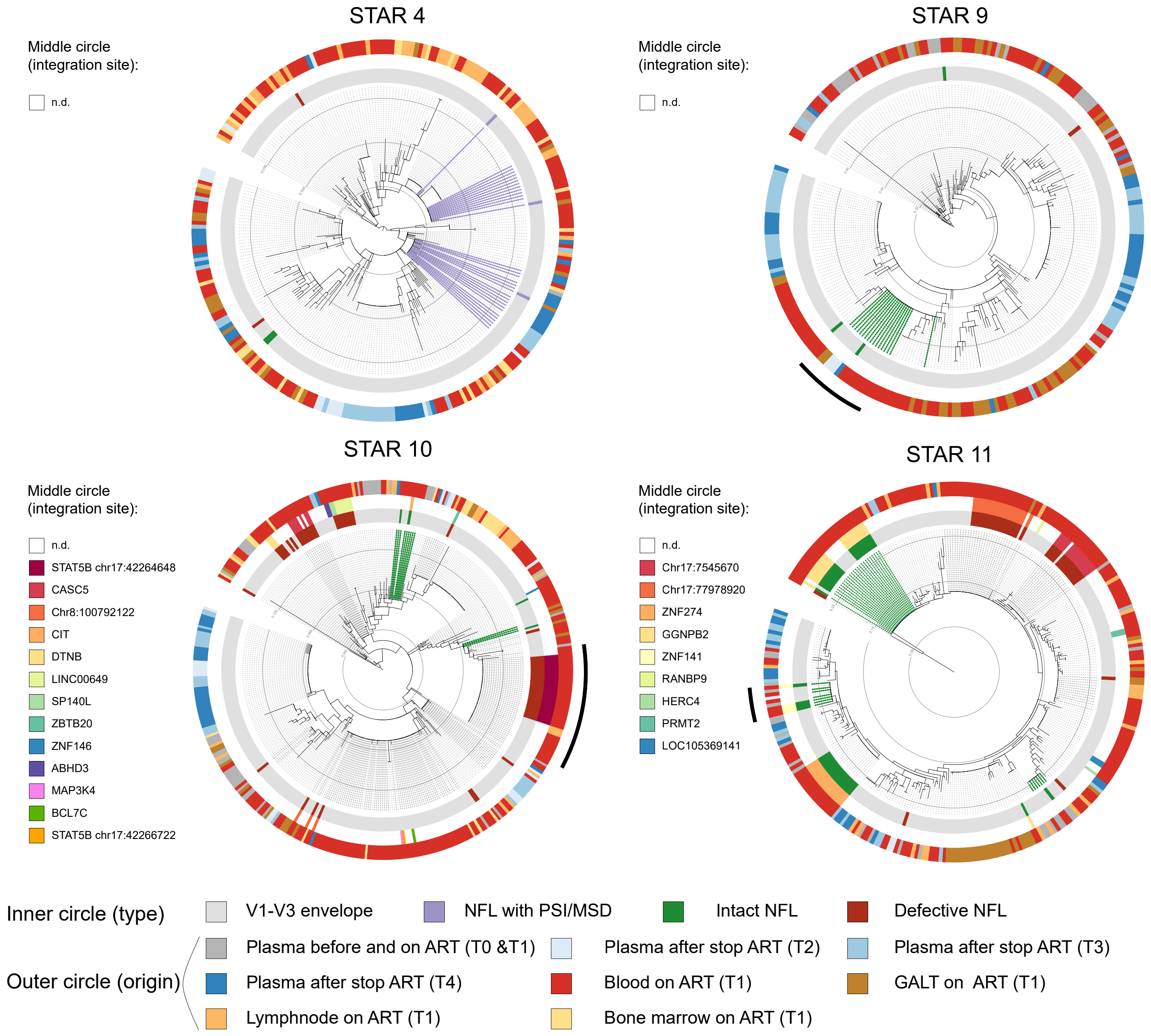

Figure 6: Circular maximum likelihood phylogenetic trees for each participant using all generated proviral and plasma V1-V3 *env* sequences before and during different stages of the ATI. The inner circle represents the sequence type, either obtained through single-genome sequencing (SGS) of the V1-V3 *env* region shown in grey and V1-V3 *env* trimmed near full-length (NFL) genomes in colors, respective of their intactness category. Clusters of identical sequences containing both subgenomic SGS and NFL are highlighted in bold dashed lines. The middle circle shows the integration site associated with MDA-derived proviruses (multiple displacement amplification) if available. The anatomical compartment origin of each plasma and proviral sequence is shown on the outer circle. The black arcs around the outer circles of STAR 9, STAR 10 and STAR 11 denote the discussed clusters of identical V1-V3 *env* sequences. ATI = analytical treatment interruption, PSI = packaging signal, MSD = major splice donor, n.d. = not determined.