High-throughput single-cell sequencing for retroviral reservoir characterization

Lauren E Droske[a,b], Stephen D. Shank[c,d], Melanie N Cash[a,b], Sergei L Kosakovsky Pond[c,d], Marco Salemi[a,b,*], and Brittany Rife Magalis[a,b,*]

[a]University of Florida, Department of Pathology, Immunology, and Laboratory Medicine, Gainesville, 32610, USA
[b]University of Florida, Emerging Pathogens Institute, Gainesville, 32610, USA
[c]Temple University, Department of Biology, Philadelphia, 19122, USA
[d]Temple University, Institute for Genomics and Evolutionary Medicine, Philadelphia, 19122, USA
[*]To whom correspondence should be addressed. E-mail: brittany.rife@ufl.edu or salemi@pathology.ufl.edu

ABSTRACT

During the course of infection, human immunodeficiency virus (HIV) maintains a stably integrated reservoir of replication-competent proviruses within the host genome that are unaffected by antiretroviral therapy. Curative advancements rely heavily on targeting the reservoir, though determinants of its evolutionary origins remain ill-supported through current strategies and are often limited by sample variety. Here, we describe a single-cell deoxyribonucleic acid sequencing (scDNA-seq) method, optimized for sequencing of proviral and host DNA from a treatment-interrupted HIV animal model. We report its benefits for improving viral reservoir resolution to support critical evolutionary events otherwise considered unreliable using traditional viral envelope gene signal alone, as well as comparative advantages to existing near-full-length genome sequencing methods. Given the variety of proviral characteristics that may influence viral rebound, scDNA-seq holds immense value in its ability to streamline many of the present-day applications available in viral reservoir studies, such as integration status and putative replication competency.

Introduction

Though the advent of combined antiretroviral therapy (cART) has vastly improved survival rates of persons living with human immunodeficiency virus (HIV), complete clearance of the virus has yet to be attained[1,2]. Incomplete viral eradication is in large part due to the ability of HIV, as with other retroviruses, to maintain a form of latency through integration of the reverse-transcribed viral genome into the cellular genomic deoxyribonucleic acid (gDNA) as a provirus[3–5]. With the exception of rare elite or post-treatment controllers of viremia, viral rebound to levels approximating initial infection can result from the interruption of cART, indicating the production of new virions from the stably integrated and drug-impervious viral reservoir[6]. Persistent, low-level viral production can also be observed in anatomical sanctuaries, defined as anatomical locations with low drug penetration[7–9]. More recent investigation of the HIV+ DNA population within the infected host has demonstrated that the vast majority ($\geq 93\%$) actually consists of replication-incompetent, or defective, provirus generated through a variety of steps in the viral replication process[10–13], rendering intact virus characterization a "needles in a haystack" problem. While resting CD4+ T lymphocytes represent the largest and most well-characterized harbor for intact virus, there is mounting evidence from both animal models and human donors that other cell types may contribute to the total HIV reservoir[14]. Characterization of intact virus may also not be the only solution to the problem of HIV persistence. Whereas integration of viral DNA permits survival of the DNA molecule through long-lived cells or even cellular division, it is not the sole form of survival for individual variants within the DNA - genomic fragments (including drug-resistant mutations sites) can be inherited through the act of recombination, allowing for even defective virus to contribute to successful viral rebound[15,16].

With the ability to accurately measure features of reservoir populations and assess their contribution to viral rebound, more direct approaches to purge the collective reservoir can be established. This contribution can be qualified, as well as quantified, using information from the phylogenetic relationships among viral genetic sequence data extracted from the reservoir populations. Notably, shared common ancestry between samples taken at different time points and/or locations can provide insight into movement of viral variants through space and time[8,9,17]. Given known states (i.e., tissue sampled) supplied for the leaves of a phylogeny (and assumed trait neutrality)[18], statistical reconstruction of ancestral states at the internal nodes can be carried out, allowing for inference of the frequency of "transitions" between states along the branches of the tree, often interpreted as migration. As a systemically elusive virus, the task to fully characterize, and thereby reliably resolve HIV intra-host phylogenies, has remained an utmost challenge for researchers, consequently curbing curative efforts[19,20]. Studies focused on elucidating the phylogenetic component of HIV have primarily relied on sequence information derived from the

viral envelope gene region encoding the glycoprotein (*gp120*), owing to its abundant evolutionary signal necessary for reliable phylogenetic reconstruction[17,21,22]. Recently, reconstruction of discrete sampling states within the *env* phylogeny from infected tissue donors as part of the Last Gift cohort, Chaillon *et al.* (2020)[23] described a multitude of deep-tissue reservoirs and viral migration pathways following cessation of therapy. However, unreliable estimation of the breadth of defective provirus in the reservoir as a result of using a single gene lends potentially erroneous conclusions to both reconstruction of the phylogeny and ancestral states within the phylogeny. Improved sequence resolution and insight into the replication competency of viral genomes acts to increase reliability and interpretability of phylogenetic inference, fundamental to elucidating the origins of the HIV reservoir and viral rebound in space and time[24–26].

Although attempts have been made to improve HIV reservoir characterization through full-genome sequencing, they fall short of reliable completeness and/or depth, can be labor-intensive, and/or introduce several *a priori* assumptions[14,27,28]. Moreover, these approaches have not been used to discriminate fully integrated provirus from episomal DNA, which may act as a surrogate for active or recent infection[29,30] (though still debated[31]). Upon integration disruption, whether through antiretrovirals or stochastic processes, the viral DNA can circularize via ligation of the long terminal repeat (LTR) regions flanking the genome. Such a process results in LTR sequence junctions that can be detected using targeted polymerase chain reaction (PCR) amplification paired with full-genome sequencing to better understand their contribution to ongoing viral evolution. Coupled with limitations in patient sample procurement for a variety of tissues, methodological and animal model enhancements are critical to conclusively assess the range of viral quasispecies that act to incite reservoir formation, evolution, and viral rebound across anatomical sites[32].

In this study, we adapt a microfluidic, single-cell DNA sequencing (scDNA-seq) technology for use with a treatment-interrupted HIV animal model to more effectively capture reliable intra-host viral phylogenies and the contribution of reservoir cells to viral rebound at unprecedented depth. The scDNA-seq method described is scaled to improve both efficiency and capacity of proviral genome recovery for broader reservoir analyses, alleviating major burdens posed by currently available methods.

Results

Proviral scDNA-seq overview

The scDNA-seq approach described herein couples droplet microfluidic technology and next-generation sequencing to capture snapshots of near-full-length (NFL) viral gDNA within individual cells from a variety of tissues. While the Mission Bio Tapestri® microfluidic platform was originally developed for human cancer cell DNA profiling, adaptation of the method for proviral DNA targeting and sequence analysis pipeline proved successful. Briefly, individual cells within viable cell suspensions undergo encapsulation to form isolated droplets (Figure 1a, boxes 1-2) that contain protease enzymes to allow access to, and thereby release of, viral and host (macaque) gDNA. The lysate, including the now histone-free DNA, then undergoes a second encapsulation step, wherein the contents are merged with barcoding beads and PCR reagents for target amplification (Figure1a, boxes 3-5). Amplified regions, or amplicons, for this project consisted of approximately 250 base pairs (bp), collectively spanning nearly the entire S[imian]IV genome, the 2-LTR circularizing junction, and the SPRY domain of the $TRIM5\alpha$ macaque gene (primer details provided in Supplementary Table S1). Full genome coverage could not be attained owing to the repeat of the 5' LTR (also present as the 3' LTR) and because the forward and reverse primers for the first amplicon were designed so as to assess the integration status of reconstructed proviruses from their former host cells (Figure 1b, c). Primers designed to amplify the SPRY domain of the *Rhesus* macaque $TRIM5\alpha$ gene served as a control for measuring bioinformatic pipeline performance. See methods section for more details on the scDNA-seq workflow and strategies for enhancing sample integrity.

Performance of this sequencing approach was evaluated using peripheral blood mononuclear cell (PBMC) and lymph node (LN) samples taken from three monkeys (HG02, JA41, and JP70) either during therapy (119 days post-infection [dpi]) or post-therapy interruption at necropsy (197 dpi) (Supplementary Figure S1).

Depth in proviral genome sequencing

A total of $1,127$ SIV genomes were recovered with a mean coverage of 29% (Figure 2a). Average estimates of proviral copies per $1E6$ cells were $> 7,000$ copies for all necropsy samples across monkeys, with the lowest at $7,798$ copies/$1E6$ cells from HG02 PBMCs, and the highest at $48,043$ copies/$1E6$ cells from JA41 PBMCs (Figure 2b). JA41 consistently exhibited a higher ($> 2-fold$) estimated proportion of infected cells across tissues and time points, with the exception of PBMCs during cART, for which only animal HG02 recovered SIV+ cells and at a fairly low number (820 copies/$1E6$ cells) (Figures 2b, c). The number of total SIV genome copies for PBMCs during cART (again, only observed for HG02), relative to the total number of cells recovered from the same tissue and time point (Figure 2b, left), was low in comparison to post-cART and to that of LN, both during and following cART interruption. Earlier studies of proviral populations in LN and PBMC presented models supporting high similarity between the two populations during cART[33]. However, more recent studies have claimed to

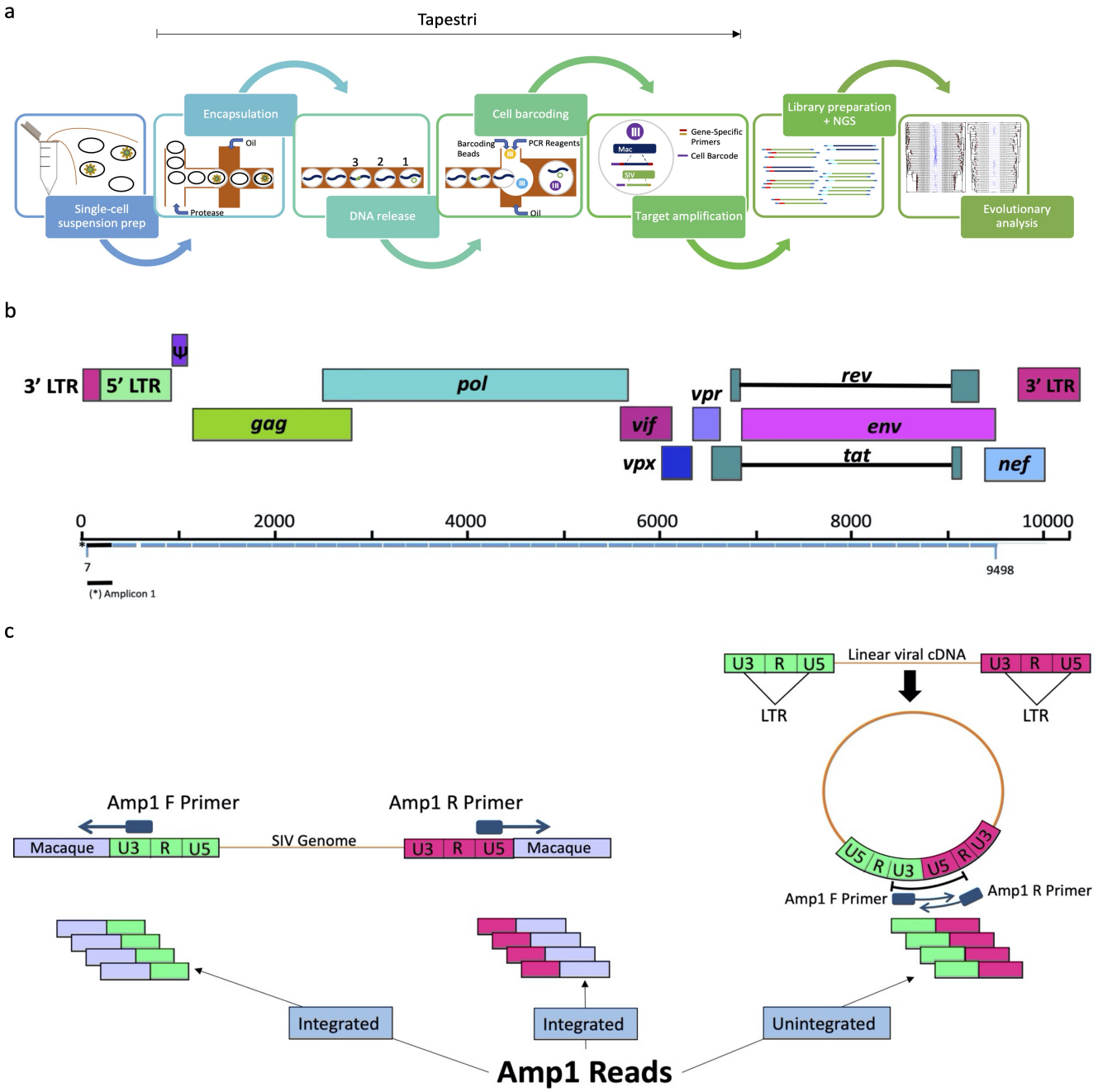

Figure 1. scDNA-seq increases viral genomic sequencing depth and breadth via a droplet microfluidics platform and enables determination of genome integration status. a Schematic overview of the scDNA-seq experimental workflow using the Tapestri® platform adapted for viruses. The encapsulated cell lysates numbered in the DNA release step exemplify the different ways in which scDNA-seq has been designed to capture genomic Macaque (Mac gDNA) and SIV (SIV gDNA) DNA for cells that are: 1. SIV-infected and unintegrated, 2. uninfected, and 3. SIV-infected and integrated. Following library prep and next-generation sequencing of NFL SIV genomes, scDNA-seq allows for enhanced phylogenetic analysis and in-depth study of the defective and intact provirus landscape on a per sample basis. **b** Amplicon coverage across SIVmacL28 reference genome (GenBank accession version KU892415.1). *Amplicon 1, covering U5U3 circularizing junction. **c** Visual representation of amplicon one construction and purposeful use for determining proviral integration status. Retrieval of complete amplicon one reads indicate a provirus present in 2-LTR form, whereas partial amplicon one reads are indicative of integrated provirus with reads that are a hybrid of SIV and Macaque.

detect viral evolution in tissues collected during cART that imply greater differential replication inhibition between anatomical sites[9]. Paired with scrutiny over whether cART can effectively penetrate LN follicles to prevent ongoing viral replication[34], the increased number of SIV+ LN cells identified during cART by comparison with PBMCs, may be a result of exploited putative sanctuary sites within LNs.

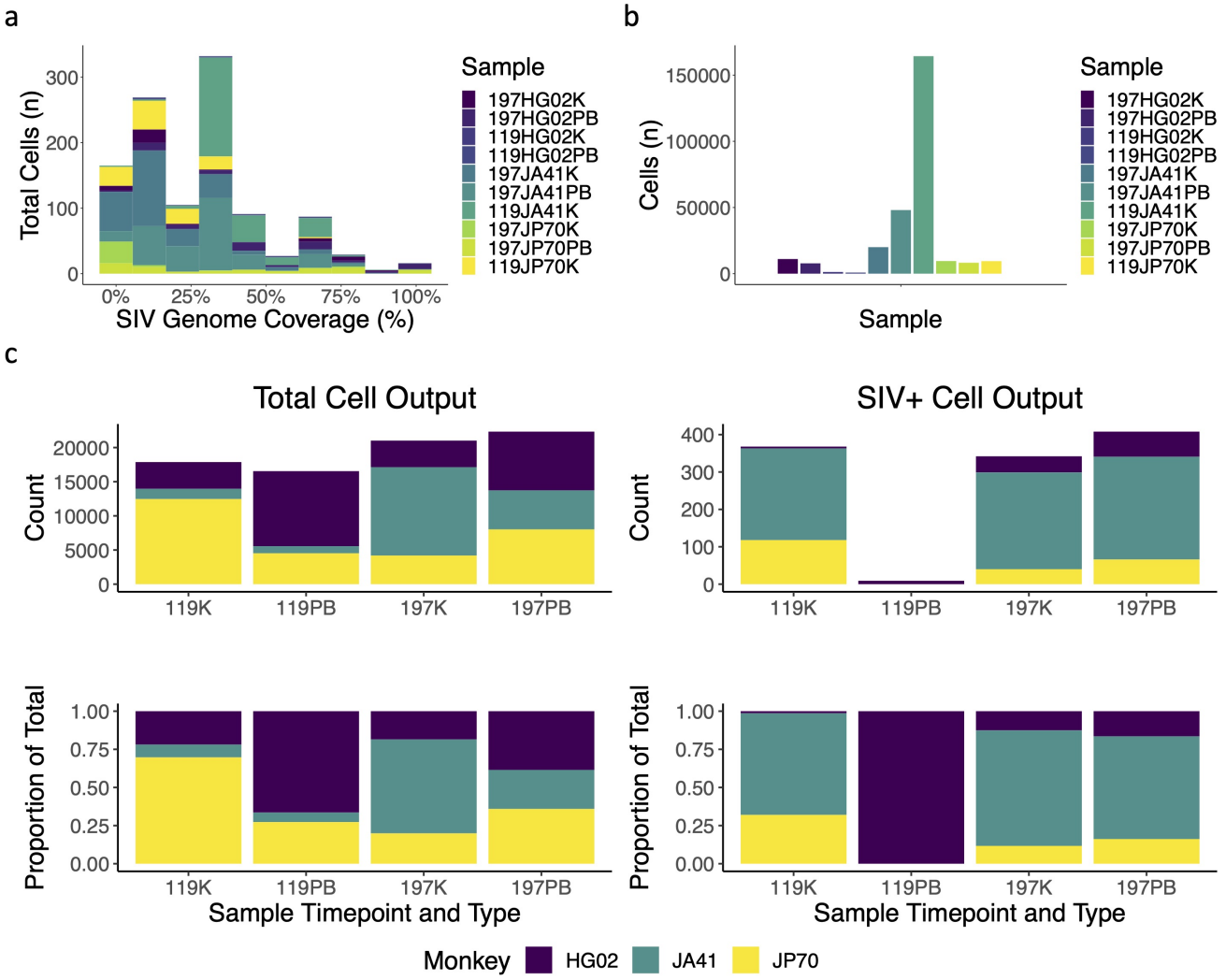

Figure 2. scDNA-seq proviral genome coverage and cellular output summaries. a Distribution of SIV genome coverage derived from 1,127 SIV+ cells. SIV+ cells were considered those in which SIV consensus genomes could be reconstructed. Corresponding samples from the three animals during cART and post-cART are indicated to the right. **b** Average estimation of proviral copies/1E6 cells. Each estimate is a reflection accounting for cumulative total and SIV+ cells per sample. **c** Stacked distributions of the cellular output counts (top) and proportion of cells contributed by individual animals (bottom) for overall cell output (left) and SIV+ only cell output (right). Each bar represents one of the four combined sample types and time points, filled according to animal contribution - 119K cART LN, 119PB cART PBMC, 197K post-cART LN, 197PB post-cART PBMC.

Intra-host genetic diversity estimates were consistent with previous studies, with mean pairwise genetic distance estimates across animals ranging from 2.89E-03 to 1.19E-02 nucleotide substitutions/site (subs/site)[35–38], though with a select minority of sequences (particularly in post-cART samples) $\geq$ 5.0E-02 subs/site, as would be expected with increasing depth of sequencing[39]. LN cART sequences exhibited a relatively low level of diversity, with a median of 2.193E-03 subs/site [IQR: 1.34E-03 − 4.15E-03]. Diversity for this tissue continued to remain low post-cART overall (median 3.06E-03), though covering a broader range [IQR: 1.25E-03 − 6.10E-03] consistent with reactivation of a subset of cells within this tissue in the absence of cART. Contrastingly, median diversity values for PBMC cART and post-cART differed by greater than 10-fold (PBMC cART 9.43E-02

[IQR: 7.24E-03 − 1.26E-02]; PBMC post-cART 4.81E-03 [IQR: 3.17E-03 − 7.33E-03]). The reduced diversity observed for PBMCs following therapy interruption may, however, be the result of animal-specific sampling bias, as the total PBMC cART SIV sequence collection consisted of only 9 sequences, all of which were derived from one animal (HG02) (Figure 3c).

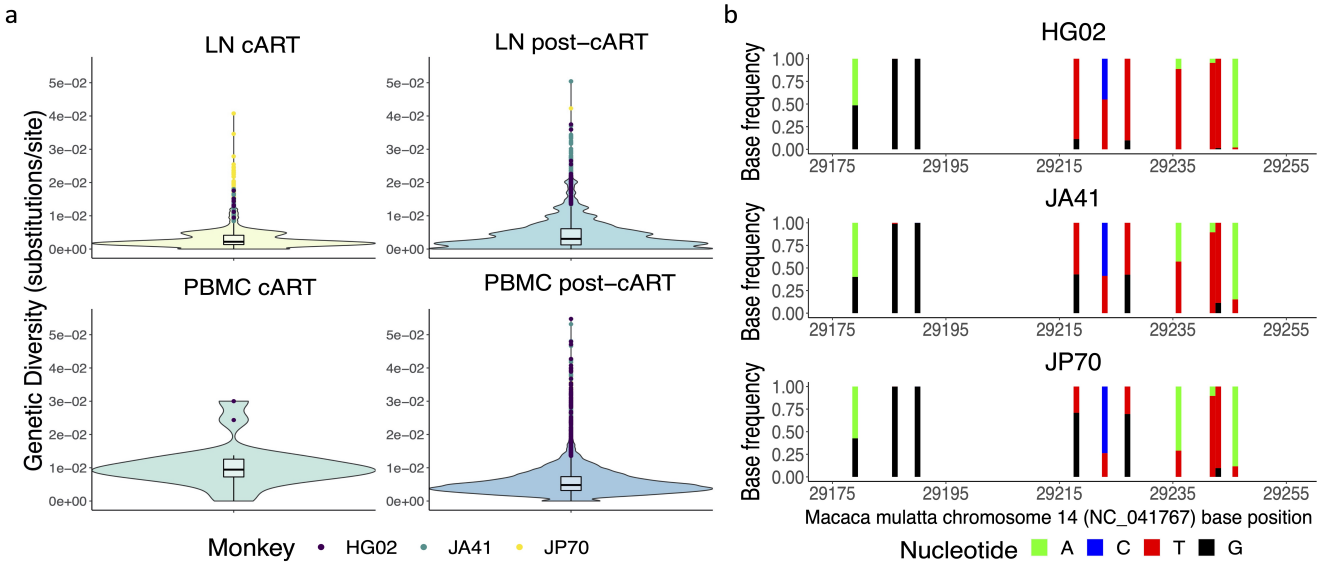

Figure 3. Genetic diversity for SIV proviral genome and *Rhesus* macaque $TRIM5\alpha$ exon 8 sequences. a Pairwise genetic distances for each tissue and time point collected among SIV DNA sequences using the TN93 model of evolution. **b** Frequency of individual alleles observed in $TRIM5\alpha$ sequences per macaque present in amplicon 42.

Reliability of produced sequence data

In order to evaluate the performance of sequencing and reliability of the respective bioinformatic pipeline, $TRIM5\alpha$ sequence variants (amplicon 42 in Table S1) were characterized with the expectation of rare multi-allelic sites at frequencies consistent with heterozygosity. A total of ten nucleotide sites were shared across all animals at which more than one allele was observed (Figure 3b). Only one instance of a third allele was observed and at a frequency of 0.006%. Two of the ten sites (positions 29190 and 29186) presented with minor allele frequencies of $< 1\%$, potentially a consequence of sequencing errors specific to the Illumina platform (e.g., certain trimer motifs, such as GGT and AGT, have been associated with Illumina machines[40]). Both trimers at these positions were represented by majority AGG; however, only one presented with a known problematic motif (AGT) in the reference sequence. The other minor allele trimer (reference motif AGC), while not established as a top ten error trimer for Illumina, remained suspect as a platform error, as it occurred in all three animals with a maximum incidence of 0.1%. Approximately 69% (18/26) of multi-allelic sites harbored minor alleles, though present at $< 1\%$ (Supplementary Figure S2), indicating a relatively low error rate of 0.17% (proportion of cells containing at least one dominant sequencing error) generated during sequencing and/or preparation of sequence library. Results thus demonstrate the reliability in calling variants for consensus sequence reconstruction for the vast majority of cells present.

Inference of proviral replication competency

Defective HIV provirus has previously been characterized by large deletions in the 5' and/or 3' end of the proviral genome. Additionally, the presence of APOBEC-induced hypermutation has been described in the form of significantly increased AG or AA di-nucleotide content relative to a closely related reference sequence (KU892415.1 used herein). However, in-depth characterization of these defects has not yet been provided owing to the limited sample size of previous analyses[10,41]. Approximately 0.80% (7/863) of SIV genomes in this study were considered to be hypermutated (Supplementary Table S2) according to a global search of hypermutated sites within the genome using the Hypermut2.0 algorithm[42]. An additional approach to hypermutation detection was described by Bruner *et al.* (2019)[11] in which modification of the di-nucleotide pattern GG (to AG or AA) at residues 4-5 and 33-34 of the rev response element (RRE) was deemed adequate to identify 95% of hypermutated sequences as defective, which was also used in this study. Proviral genomes were thus classified as "hypermutation-defective" according to this criterion, or "deletion-defective" if the first two amplicons within the 5' and/or 3' regions were missing. Two amplicons were used to define "deletion-defective" in order to reduce the false positive rate by accounting for potential variability in primer-binding sites that would result in loss of sequenced reads in one of the two

amplicons. Although hypermutation was detected according to Hypermut2.0, as described above, no RRE hypermutations were observed in the SIV data which may be attributable to a large majority of our sequences lacking the RRE probing region. Consistent with previous reports for HIV[11,43], our method classified 2.4% of genomes as intact. However, percentages of defective deletion patterns described in earlier studies for both SIV and HIV varied in comparison with our results - whereas Bender *et al.* (2019)[35] reported 3' deletions as most prevalent, our results indicate the vast majority (83.8%) of defective genomes are characterized by deletions in both the 5' and 3' ends (heretofore referred to as "double-deletions"), and 5' deletions (in the absence of 3' deletions) more prevalent (10.9%) than 3' deletions only (2.9%) (Figure 4a).

Only 0.23% NFL genomes (2/863) harbored intact 2-LTRs, one of which was a largely defective genome (119JA41K cell 199) and the other fully intact (197JP70PB cell 25). Although this percentage is lower than previous reports[44], the labile nature and scarcity of this particular form of unintegrated DNA in contrast with more common 1-LTR circles[45,46], which our current approach is not equipped to detect, may account for this disparity.

Characterization of defective provirus

Using the targeted scDNA-seq approach with relatively short (250-bp) amplicons, differences in 5' and 3' deletion length distributions could be constructed with more precision than previously[41] in order to characterize differences among these classified defective viruses. Among the double-deletion genomes, 3' deletions were 4-fold longer overall (median 4,821 bp) than 5' deletions (median 930 bp), though a non-Gaussian distribution of lengths was observed for both regions (Figure 4b). In contrast, defective genomes characterized as having only a 5' deletion or 3' deletion ranged widely in terms of the length of the deletion, with 5' deletions on average longer (mean 5,787) than those in the 3' end (mean 2,768). Short 5' deletions (mean 297) were also observed in genomes considered 3' deletion-defective only which may warrant further investigation into biological and/or sequencing origins. The distribution of 3' deletion lengths was consistent regardless of whether deletions in general were isolated to the 3' end or occurred in both ends (Figure 4b, top and bottom right). These consistent distributions might be expected in the absence of selection on deletion length or location. However, the opposite held true for 5' deletion lengths, for which the distributions were oppositely skewed (Figure 4b, top and middle left). Taken together, these results suggest the potential for a dependency of 5' genome fragment deletions on the absence of specific 3' fragments. Mapping of deletions at peak deletion lengths reported in Figure 4b top row revealed consistently intact protease (Pro) and reverse transcriptase (RT), but not the entire gag-pro-pol reading frame[47], as the common denominator among double-deletion defective genomes (Figure 4c). Though previous reports of defective proviruses have demonstrated the capability for protein production[43], the absence of the full open reading frame (ORF) within these defective DNA suggests that residual protein expression in the face of defective replication does not predominate when deletions exist at both ends. The source of integration in the absence of an intact viral integrase (the remainder of this ORF) is also unclear, though detection of LTR circles was only specific for the 2-LTR form, leaving 1-LTR circles as a possible mechanism for survival of SIV DNA without integrase function.

Phylogenetic resolution

To investigate improvements in the reliability of proviral evolution reconstruction using scDNA-seq, we compared the phylogenetic signal of NFL genomes with their respective envelope *gp120* fragments (i.e., NFL genomes obtained by scDNA-seq were disaggregated into subgenomic *gp120*) (Figure 5 and Supplementary Figure S4). Only those genomes with intact *gp120* were thus included for this analysis (HG02, n=62/124; JA41, n=155/779; JP70, n=47/224). Likelihood mapping was used to visualize phylogenetic signal, which provides a summary of the percentage of sampled sequence quartets that could be reliably resolved in the form of a bifurcating phylogeny (represented by the corners of the triangles in Figure 5a) and percentage unresolved (center of the triangle). Animals exhibited up to a nearly 3-fold reduction in the number of unresolved phylogenies for NFL genomes compared with their *gp120* counterparts.

Maximum likelihood trees for each animal were then reconstructed using all NFL genomes and pruned to exclude genomes with non-intact *gp120* for comparison with the *gp120* sequence region alone using a variety of quantitative and qualitative tree metrics (Figure 5b and Supplementary Figure S4b). Robinson-Foulds (RF) distances between the two trees for each animal were first calculated in order to describe quantitatively the differences in topology (branching patterns)[49]. Briefly, for two trees with identical sets of leaves, the RF describes the number of branches in the first tree that define bifurcations that are absent from the second tree, and vice versa. Final values were normalized by the maximum potential RF distance. RF distances ranged from 0.696 (JP70) to 0.942 (JA41) (Table S3), indicating the NFL and *gp120* trees were $< 31\%$ away from completely randomized (100%) tree topologies.

Differing topologies translate to the potential for differences in ancestral sequences and/or states (e.g., tissue) within the tree. As described above, ancestral state reconstruction within a phylogeny can be used to infer state transitions and thus origins but assumes that the individual phylogenetic relationships within the tree are known, or at least reliable; this is not always the case, as can be seen from the percentage of sequences considered unresolved in Figure 5 and differences in tree topology (S3). Increased resolution translates to increased reliability in the branching patterns observed, but it also has the potential to result in differing branching patterns altogether as compared with lower-resolution data. An overall summary of the frequency of

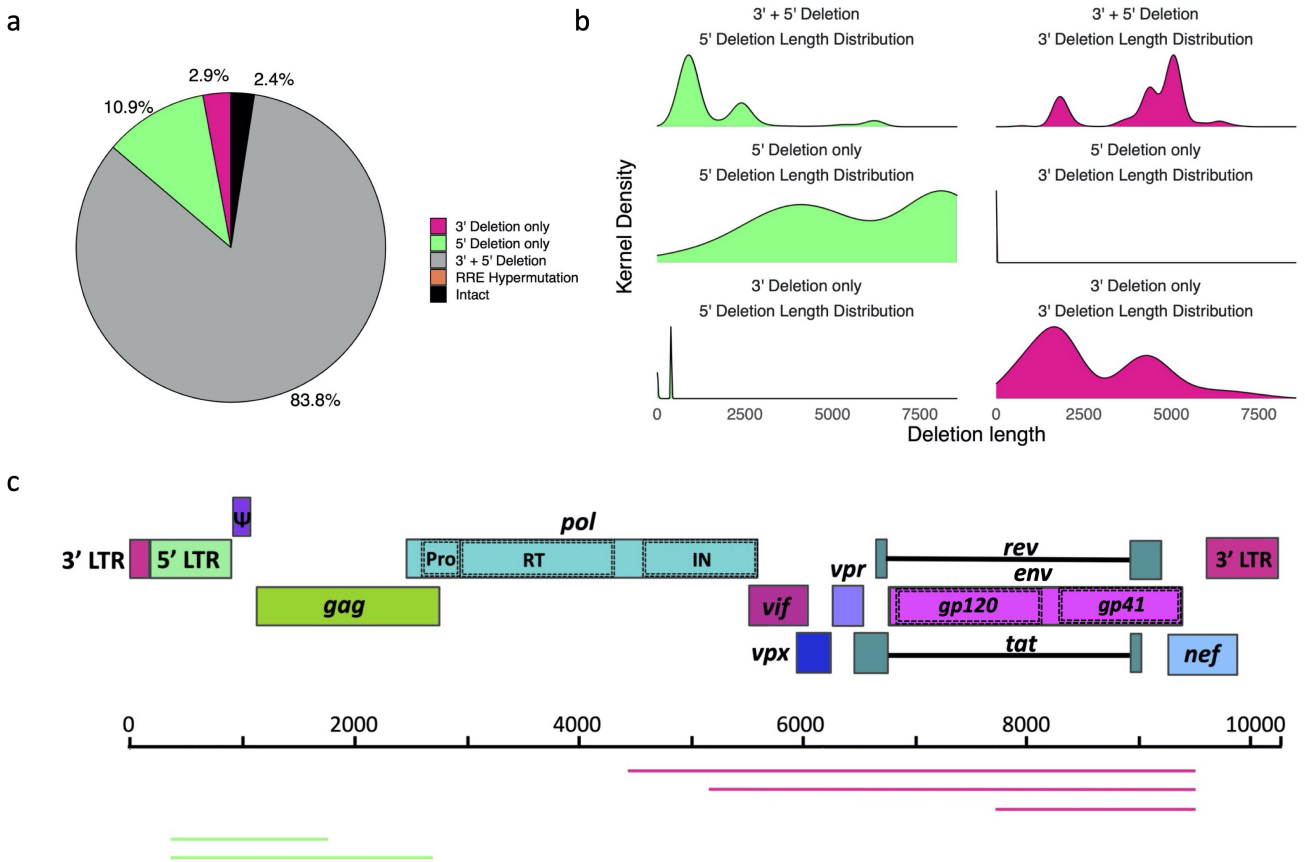

Figure 4. Defective profiles of SIV proviral genome sequences detected using scDNA-seq. a Proviruses persisting in macaque host cells as detected by scDNA-seq. Defect classifications included 3' only and 5' only deletions, 3' deletions subsisting with a 5' deletion, RRE hypermutation, and intact. **b** Deletion length distributions. **c** The *pol* portion of the gag-pro-pol open reading frame is segmented according to enzymes protease (Pro), reverse transcriptase (RT) and integrase (IN). Envelope gene delineations for *gp120* and *gp41* are outlined. Peak deletion lengths identified in (b) are mapped against the reference genome and colored according to 5' (green) or 3' (magenta) location.

transitions between sampling tissue and time points within the NFL and *gp120* trees revealed relatively consistent transition distributions across animals (Supplementary Figure S3), indicating the large RF distances above were largely the result of unresolved branches nearer the tips of the phylogeny, rather than the root. However, differences for individual clades of interest between trees were observed. In the HG02 NFL genome tree, a subtree of closely related post-cART LN sequences exhibited shared recent common ancestry with a LN sequence sampled during cART, which was supported in $\geq 90\%$ of bootstrap replicates (Figure 5b, clade "A"). Contrastingly, no such shared common ancestry was observed in the *gp120* tree (Figure 5b, "B"), suggesting differing origins could be inferred in the absence of the additional sequence information provided from the NFL genomes. The NFL genome tree for animal JA41 presented evidence of potential compartmentalization of cART LN sequences with bootstrap support (*BS*) $\geq 90\%$ (Supplementary Figure S4b, "A"), which was not supported in the *gp120* tree (Supplementary Figure S4b, "B"). In the JP70 trees, a supported group of post-cART PBMC sequences sharing common ancestry with a cART LN sequence in the NFL genome tree (Supplementary Figure S4b, "C") was also not observed in the respective *gp120* tree (Supplementary Figure S4b, "D"). Similarly, in JP70 a group of cART LN sequences which remained clustered together in both the NFL and *gp120* trees were only shown to exhibit supported shared ancestry with two post-cART LN and one post-cART PBMC sequence in the NFL tree (Supplementary Figure S4b, "E"), while identified in separate locations distanced from the cART LN sequences with BS $< 90\%$ in the *gp120* tree (Supplementary Figure S4b, "F"). Overall, HG02 and JA41 exhibited greater than two-fold increases in the number of clades with $BS \geq 90\%$ for the NFL tree (HG02, n=21; JA41, n=15) compared with that of *gp120* (HG02, n=8; JA41, n=6), whereas the difference for JP70 was four-fold (n=17 and n=4), further substantiating the need for greater genome coverage in phylogenetic inference.

Of the intact proviruses accounted for in reconstructed phylogenetic trees (i.e., with intact *gp120*), $> 95\%$ (20/21) were

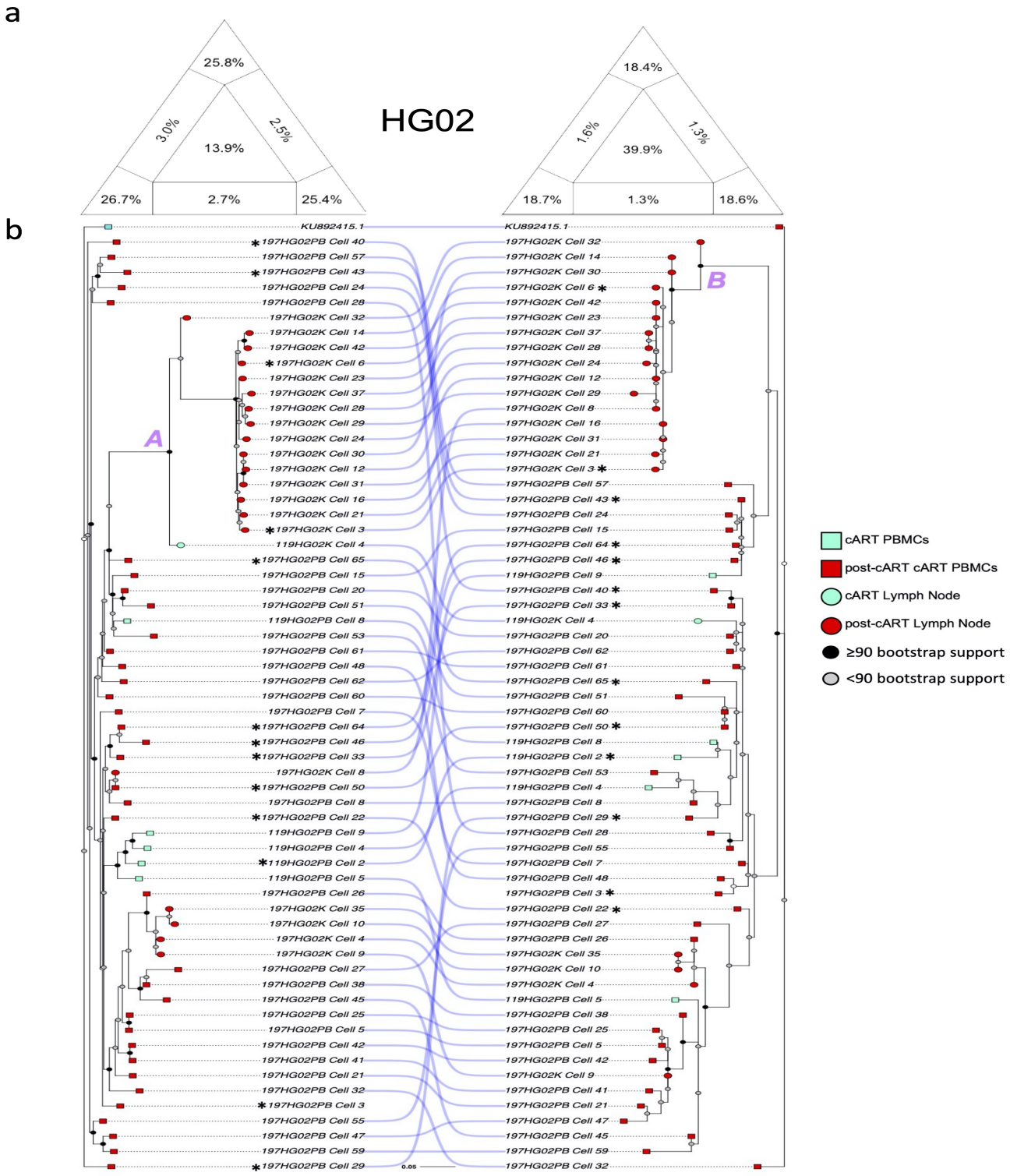

Figure 5. Comparison of phylogenetic resolution for near-full genomes and subgenomic *env gp120*. a. Phylogenetic likelihood map pairs for HG02 with NFL genome sequences (left) and *gp120* (right). **b.** Corresponding maximum likelihood phylogenetic tree (legend to the right). Shape indicates tissue type (square: PBMC, circle: LN), while color indicates time point (cyan: cART, red: post-cART). Support for branching events was provided through bootstrapping[48], with $\geq 90\%$ considered well-supported and represented by a black circle and $< 90\%$ represented by a gray circle. Asterisks indicate intact genomes. Significant differences in clade formation support between NFL genomes and *env gp120* alone are marked with letters A and B.

obtained from post-cART samples. These included intact genomes derived from both PBMC and LN, consistent with previous reports of multiple sources of viral rebound[50]. Within the NFL tree, the single identified intact cART sequence (119HG02PB Cell 2) did not share recent common ancestry with post-cART sequences, but rather belonged to a well-supported ($BS \geq 90\%$) clade consisting of the majority of remaining PBMC cART sequences. Similarly, the intact cART sequence in the *gp120* tree shared recent common ancestry with one other PBMC cART sequence (BS $< 90\%$), indicating not all intact proviruses persisting during therapy may contribute to viral rebound.

Prevalence of drug-resistant mutations in discrete proviral populations

Following screening of all sequences for major HIV/SIV drug-resistance mutations (DRMs) against those encompassed in our cART regimen (Raltegravir, Tenofovir and Emtricitabine) we recovered four DRMs within the RT and integrase (IN) regions (Table 1). RT DRM M184I was observed in only one defective sequence (PBMCs post-cART in animal HG02), whereas RT DRM S211N was observed in both intact and defective sequences from JP70 PBMCs following therapy interruption (Table 1). In contrast to RT mutations, IN DRMs were observed across all three animals, in both lymph node and PBMCs, and during both time points. Despite overlap in tissue and time point, Int and RT mutations were not observed together within the same sequence. IN DRM N155K was more frequent than IN DRM E138K and predominantly observed in defective provirus.

When traced over a phylogeny, the presence of a mutation across neighboring lineages in the context of genetic diversity is an indication of mutation persistence during ongoing replication and evolution. The extent of persistence of drug resistance, as opposed to independent evolutionary events, and the relationship between intact and defective virus were determined using the full NFL phylogenies for each animal. DRMs in this study could be observed among intact virus for two of the three animals (JP70 and HG02), for which phylogenetic relationships could be identified between intact and defective viral sequences. For animal JP70, two intact PBMC sequences harboring the S211N DRM following cART interruption shared recent common ancestry with one defective S211N-containing genome from the same sample (PBMCs post-cART), though this clade was not well supported ($BS \geq 90\%$) (Supplementary Figure S5, "A"). In addition to insufficient support, the contemporaneous, or homochronous, nature of these samples renders ancestral sequence reconstruction unreliable for identification of the source (i.e., defective or intact) of these two viral genomes. In animal HG02, however, defective provirus harboring the N155K DRM was recovered from the LN during cART and formed the base, and thus candidate source, of a clade containing two N155K DRM sequences from LN following therapy interruption (post-cART) (Supplementary Figure S6, "A"). Whereas relationships among heterochronous samples offer increased confidence, directionality is difficult to resolve when the relationship is not paraphyletic (e.g., an intact sequence nested within a clade of multiple defective genomes with the same DRM)[51]. Coupled with the low support given for the clade S6, "A"), we cannot conclude whether the N155K mutation could be revived from defective virus or if both intact and defective virus belonging to clade A were progeny of an earlier, unobserved viral genome. While replication competent, insufficient support for related intact sequences harboring these DRMs could also not exclude the independent emergence of resistance, alternative to persistence through replication, throughout each evolutionary history.

Table 1. Frequency of drug-resistant mutations in intact and defective proviruses. Four detected major DRMs targeting RT and IN were subdivided into prevalence among intact and defective proviruses. Frequencies are displayed as occurrences out of SIV genome sequence totals on a per sample basis.

Macaque	Time point/Tissue	RT Mutations				Integrase Mutations			
		M184I		S211N		E138K		N155K	
		(Freq Intact)	(Freq Defective)	(Freq Intact)	(Freq Defective)	(Freq Intact)	(Freq Defective)	(Freq Intact)	(Freq Defective)
JA41	119K	-	-	-	-	-	-	-	12/245
	197K	-	-	-	-	-	-	-	-
	197PB	-	-	-	-	-	-	-	1/275
JP70	119K	-	-	-	-	-	-	-	-
	197K	-	-	-	-	-	-	-	-
	197PB	-	-	2/66	5/66	-	-	1/66	3/66
HG02	119K	-	-	-	-	-	-	-	1/5
	119PB	-	-	-	-	-	-	-	-
	197K	-	-	-	-	-	-	1/43	1/43
	197PB	-	1/67	-	-	-	1/67	-	-

Discussion

Existence of the HIV reservoir, adept to replenishing viremia upon therapy interruption, has continued to pose a major barrier towards attaining HIV eradication within its host. Uncovering critical reservoir dynamics that facilitate the long-term maintenance of HIV are imperative to more precisely target the reservoir responsible for viral rebound[52]. Several sequence-based methods have been developed to further investigations of the reservoir, though often limited by the trade-off in depth and/or breadth, as well as the ability to correctly classify integration, of the proviral population[14,44,53–55]. Previous assays that

have successfully incorporated targeted, near-full-length (NFL) genome sequencing of HIV/SIV have been devised with the replicating virus in mind, incorporating a limited number of amplicons that ultimately 1) underestimate the size of the defective population and 2) are incapable of in-depth characterization of this population[12,41,56]. Other assays, such as the intact proviral DNA assay (IPDA), demonstrate the use of a duplexed droplet digital PCR (ddPCR) approach to estimate the frequency of intact HIV DNA[11]. However, without confirmed sequence data, IPDA may misidentify proviruses with defects elsewhere in the genome as intact (approximately 30%), or miss identifying genomes entirely due to HIV intra-host genetic variability restrictions with such limited targets for primer/probe binding[57,58]. Here, we present a novel method using single-cell DNA sequencing (scDNA-seq), that serves to overcome these limitations - by using a multi-amplicon approach that comprises the near-full SIV genome in short segments, we are able to both improve proviral genome recovery and discover with more certainty the lengths at which deletions exist and potentially contribute to reservoir dynamics. Our method not only recovers a wider range of proviruses, it does so at greater depth and with less labor intensity by circumventing the need for limiting dilutions. Increased depth has been previously viewed as a limitation, owing to the amount of input sample required[23]; however, the scDNA-seq method described requires only $100,000$ cells as minimum input with no trade-off in hands-on time[12,56,59]. Importantly, we demonstrate that this approach is not limited to peripheral blood samples[60], allowing for deep-tissue characterization of the reservoir from animal models.

Compared to other studies utilizing NFL genome sequencing with a similar number of samples, if not more, scDNA-seq was able to recover thousands of SIV+ NFL genomes for an average of 93 sequences per sample (maximum of nearly 300), contrasted with the typical total reported < 100 to approximately 300 genomes per study[11]. This is in stark comparison to per sample or participant averages of < 40 genomes using NFL sequencing with a Sanger approach[10,41] and doubles maximum output of that achieved with FLIPS (Full-Length Individual Proviral Sequencing, up to 144 sequences per participant)[12], all the while remaining largely consistent with detected percentage of infected cells[33,61]. For the vast majority of samples used in this study ($10/12$), total cell recovery based on a recommended input of between $1.0E5 - 1.5E5$ cells, met the minimum expected from frozen cells, at roughly $4,000$, consistent with human cancer cell study users[62,63]. The two samples with lower-than-expected yields were extracted from animal JA41 (group 2.1) during cART and were stored in a homemade freeze media as opposed to all other samples in this study, which were cryopreserved (see methods sample collection). Hence, the exceptionally poor quality of these two samples, despite efforts to enhance sample integrity, may have contributed to the significantly reduced total cell yields.

The ability to distinguish intact and defective genomes is critical to understanding the contribution of specific reservoir populations to viral rebound in infected hosts. Deletions encompassing either one or both ends of the SIV genome comprised the dominant defect classification present, corroborating existing evidence that deletions are the most common defect within the HIV DNA population[12,64]. Contrasted with the study performed by Bender *et al.* (2019)[35] in which 267 SIV sequences were recovered with 32% hypermutated and 28% intact, we recovered no RRE hypermutants and a percentage of intact genomes more consistent with other prior HIV reservoir studies of $< 8\%$[64]. These dissimilarities can likely be explained by a significant portion of our sequences harboring deletions in the region previously used to probe for hypermutations (RRE), and additionally, by way of their approach utilizing IPDA to classify intact proviruses, which as previously stated, lacks the ability to identify almost a third of defects by near-full genome sequencing. Although we were able to gain surmountable proviral genomes, the depth at which our data is provided to characterize deletion patterns lacks a formal comparison due to previous studies obtaining too few sequences.

The increased breadth of our deletion defective classification approach (using targeted, short amplicons) also allowed for enhanced insight into defective genomic profiles through population-level characterization of deletion lengths. The deletion length patterns described in this study were representative of the presence of two distinct "double-deletion" defective viral populations. The presence of the first population in which defective proviruses maintain intact *env* is consistent with earlier studies describing continued protein expression[43]. The ability to express proteins in the face of replication incompetence provides a mechanism of persistent immune activation, resulting in an increase in the susceptible cell population for infection and, ultimately, maintenance of the reservoir[65]. However, a second population with varying 5' deletion lengths (preserving RT) was identified in this study, which appear to be dependent on 3' deletions lengths (all resulting in removal of *env*). This finding poses the question as to the benefits of RT preservation in the absence of structural protein (*env* and *gag*) expression. Phylogenetic reconstruction of the virus evolutionary history within each animal revealed multiple instances of persistence of DRMs among intact viral sequences, but also the emergence of integrase DRM N155K in a defective virus during cART that was shared in related post-cART intact sequences. Though HIV DRMs have been hypothesized to persist through recombination of defective and intact viral genomes during co-infection[16], the potential for viral RNA transcript production required for this process could not be confirmed for this virus population. Alternatively, the maintenance of "double-deletion" defective genomes with intact RT may be the result of accumulation over time due to limited protein expression, thereby reducing their ability to be cleared. Additional investigation using RNA sequencing and cellular profiling is needed to determine the extent of transcription and the role of the individual defective populations as contenders for barriers to a cure, such as immune activation. Moreover,

despite improved overall phylogenetic resolution using NFL sequences, insufficient support in the form of bootstrapping for the events described above points to missing pieces of the puzzle (e.g., additional infected tissues) that may provide the resolution required for understanding the role of RT preservation in defective virus, drug resistance, and their relationship with 3' genomic variation.

As with the development of any new method, uncertainties exist throughout the original, unprecedented design that may serve as potential limitations when interpreting the data. In this study, SIV+ cells were defined as those possessing a minimum depth of ten reads specifically for amplicon 37. This amplicon was chosen primarily because of its proximity to the non-coding 3' end of the SIV genome and the current acceptance that HIV/SIV proviral deletions most often occur as internal deletions (in coding regions), with the exception of small psi/MSD deletions[66]. However, modifications may need to be made to this somewhat subjective definition, owing to the observed presence of additional amplicons in the absence of amplicon 37 ($n = 1,054/1,127$). Additional modifications to the cell-finder algorithm currently used in the post-processing pipeline, which serves to merge individual cell-specific barcodes into single cells, may also be necessary. Developed by the manufacturers to streamline data analysis for scDNA-seq of human cancer cells, the pipeline incorporates reference genome coverage and read depth when merging barcodes with or without potential sequencing errors. Whereas the default assumption is that every amplicon should ideally be present, SIV infection rate of a cell population is less than 100%, with proviral DNA missing anywhere from 0 to $> 75\%$ of the genome, as demonstrated here and by others[10]. Parameters relating to this function were thus modified to take this *a priori* assumption into account; however, additional optimization may be necessary to address dynamic, rather than uniform, coverage and its potential effects on false positive cell distinctions, though, as described in the above paragraph, cell output was as expected. Further information on pipeline modifications can be found on the scDNA-seq GitHub (https://github.com/salemilab/scDNA-seq).

We additionally recognize that, regardless of our increased resolution in defective classification and characterization owing to shorter amplicons, targeted sequencing has its own limitations. As is often observed for heterogeneous viruses, mutations in primer-binding sites can result in sub-optimal primer binding and even the complete absence of reads, referred to as "amplicon dropouts." This heterogeneity can exist in the form of divergence from the reference sequence or intra-host population heterogeneity. We recognize that, using this approach, we were unable to fully distinguish regional deletions from amplicon dropouts. A single reference sequence (KU892415.1) was used in the design of the primers, which were evaluated using preliminary PCR amplification, though this preliminary testing cannot account for within-host heterogeneity. Isolated, internal amplicon dropouts were not observed; however, single or even multiple amplicon dropouts within the flanking regions of the genome deserve additional attention and may have resulted in the underestimation of the number of intact genomes. Non-targeted, "shotgun," sequencing of samples used in this study is necessary (and underway) to measure the potential impact of divergence and diversity of individual samples on the results described herein.

In the face of the above described limitations, our scDNA-seq method permits unparalleled depth of NFL proviral genome sequencing that serves to: (1) broaden analyses of not only intact, but defective proviral populations that persist during cART, (2) determine integration status of proviral genomes, (3) provide a less labor-intensive NFL genome sequencing approach with minimal input, and (4) improve confidence in phylogenetic inference of viral evolution and phylodynamic patterns, such as migration. Now armored with multi-omic capabilities[67], this approach offers potentially expansive and informative insight into the maintenance and stability of the reservoir populations and underlying cellular compositions.

Acknowledgements

Research reported in this manuscript was supported by the National Institute of Neurological Disease and Stroke at the National Institutes of Health (NIH) [R01NS063897 to M.S.] and the Stephany W. Holloway University Chair in AIDS Research. The authors acknowledge University of Florida Research Computing for providing computational resources and support that have contributed to the research results reported in this publication. URL:http://researchcomputing.ufl.edu.

Author contributions statement

L.D. and M.C. contributed to the experimental aspects of the study, S.D.S. and B.R.M. to the development of analytical tools and analysis of data, and S.L.K.P., M.S., and B.R.M. to the supervision of experimental and analytical components. B.R.M. conceived of the presented idea and verified experimental and analytical methods. All authors discussed the results and contributed to the final manuscript.

Methods

Ethics statement. All animal procedures were performed by the Tulane National Primate Research Center (TNPRC) in accordance with Tulane University's Institutional Animal Care and Use Committee (IACUC) protocol P0376. The macaques

were housed in groups or pairs at Tulane animal facilities accredited by the Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC). Details surrounding animal welfare including environmental parameters and standard practices for treatment of non-human primates in research were followed as outlined in the *Guide for the Care and Use of Laboratory Animals*[68]. All possible measures were employed to minimize discomfort of the animals. Once IACUC defined endpoints were reached, macaques were humanely euthanized following the standard method of euthanasia for non-human primates at the TNPRC, consistent with the recommendations of the American Veterinary Medical Association Guidelines on Euthanasia.

Study population. Two Indian *Rhesus* macaque (*Macaca mulatta*) groups identified in the present study as 2.1 (n=4) and 2.2 (n=4), were staggered for use and represented a single cohort of SIV cART-interrupted animals. The study cohort was intravenously inoculated with the same SIVmac251 viral swarm at a dosage of 100 TCID50[69,70]. Following 35 days post-infection (dpi), at which median viral set point was established, cART treatment was initiated. The cART regimen comprised of two reverse transcriptase inhibitors: Tenofovir, 20 mg/kg (Gilead) and Emtricitabine, 30 mg/kg (Gilead) administered daily by subcutaneous injection, and one integrase inhibitor: Raltegravir, 22 mg/kg (Merck) twice daily by oral administration. Drug dosages were based on a protocol developed by Dr. Kenneth Williams' lab at Boston College, Boston, MA, and recommendations by the senior director of Biology at Gilead Sciences, San Francisco, CA. One macaque from group 2.2 was found deceased at 42 dpi and was therefore not included in further longitudinal sampling for this study. At 180 dpi (145 days of drug treatment), cART was suspended to allow viral rebound and disease reemergence. The animal cohort was followed for three additional weeks post-cART and subsequently euthanized for supplemental sample collection.

Sample collection. A baseline blood draw prior to SIVmac251 inoculation was taken from each animal. Routine weekly physical examinations over the course of pre-infection through necropsy, were accompanied by blood draws for the purpose of monitoring SIV viremic levels and viral reseeding into the peripheral blood. Peripheral lymph node collections were completed on the following days/day ranges post-infection: 14, once between 49-63, 120, and 180. Following cART interruption, macaques were euthanized and further sampling obtained at necropsy (days 196-202 post-cART) for the previously mentioned sample types. All group 2.1 sample time points preceding 180 dpi were frozen in a homemade freeze media containing 90% FBS : 10% DMSO using Mr. Frosties. At 180 dpi and all time points thereafter, samples were cryopreserved in cryovials to augment storage for downstream live cell recovery. Group 2.2 samples were frozen in the aforementioned live freeze media, and subsequently switched to cryopreservation beginning at 120 dpi.

Custom amplicon panel design. A custom reference genome was designed to obtain full length coverage of the SIVmacL28-C1 isolate (GenBank accession number KU892415.1), including one amplicon which targeted the U5U3 circularizing region (Figure 1b). Positions KU892415.1:10151-10276 were rearranged to the anterior end of the genome with resultant positions KU892415.1:1-126. Consequently, the overall coordinate-base map shifted, producing a slight deviation from the original SIVmacL28-C1 genome FASTA. Exon eight, encoding the SPRY domain of the macaque $TRIM5\alpha$ gene (GenBank accession number NC_041767.1), was added for amplicon coverage inclusive to all cells to aid in examination of pipeline performance for correctly calling variants. Virus amplicon coordinates of the custom reference genome can be located in Supplementary Table S1.

Single cell suspension preparation. Samples containing 1 mL cell suspensions were quickly thawed in a 37°C water bath and carefully aspirated and dispensed to a 1.5 mL Eppendorf tube. All cell suspension transfers and washes proceeded with the use of wide bore pipette tips to minimize cellular damage. The original storage vial was rinsed with 500 µl of room temperature 1xDPBS **(inlcude manufacturer and cat. no.)** to recover residual cells and gradually transferred to the 1.5 mL Eppendorf tube. The sample was then centrifuged at 300 *xg* for 5 minutes at 25°C to allow removal of storage media from the cells. The supernatant was discarded and cell pellet resuspended in 1 mL of DPBS to be washed using the aforementioned centrifuge parameters. Carefully, the supernatant was removed without disturbing the cell pellet and gently resuspended in 50 µl, or as low as 35 µl, of Cell Buffer (MissionBio) until no clumps were visible, and subsequently stored on ice. A cell count was performed on the suspension using trypan blue dye exclusion to determine percent viability and quantify live cells by visual inspection on a hemacytometer. In short, depending on concentration, cell suspensions were diluted with trypan blue either 1:4 (5 µl of cell suspension + 15 µl of trypan blue), or 1:10 (2 µl cell suspension + 18 µl trypan blue), and 18 µl of the diluted mixture pipetted into the loading chamber of a Nexcelom hemacytometer (order no. CP2-002). Total number of cells, live and dead, were counted in four corner quadrants using a brightfield microscope with 20X magnification. Counts were averaged and multiplied by dilution factor times 10,000 to obtain cells/mL. The cell suspension was then diluted, as necessary, using Cell Buffer to a concentration of approximately 3,500 cells/µl, and subsequently available for use as Tapestri® instrument input. To circumvent variation in sample integrity as a result of adjustments in storage media and preparation prior to thawing, a protocol was optimized to enhance viable cell recovery and executed when necessary.

Low-integrity sample enhancement workflow. In circumstances where viability fell below the 80% threshold established for single cell DNA sequencing, an optimized sample preparation protocol was performed. The workflow utilized the MACS Miltenyi Biotec Dead Cell Removal kit (cat. no. 130-090-101) and manufacturer's instructions adhered to with slight

modifications described as follows.

Peripheral blood mononuclear cells (PBMCs). PBMC samples characteristically possessed adequate initial cell concentrations and required MACS™ separators for smaller cell samples. Each sample was thus processed for dead cell removal with an MS column (order no. 130-042-201) inserted in a MiniMACS™ separator (cat. no. 130-042-102) on a MACS™ MultiStand (order no. 130-042-303). Following the initial cell count, samples requiring dead cell removal were washed a second time by centrifuging at 300 *xg* for 10 minutes at 25°C. The supernatant from wash two was discarded and the cell pellet was resuspended in 100 µl of Dead Cell Removal MicroBeads per 10^7 cells, and incubated for 15 minutes at room temperature to magnetically label dead and early apoptotic cells. During the incubation step, 20x Binding Buffer supplied in the dead cell removal kit was diluted with molecular grade water to make a 50 mL 1x stock solution. The 1x Binding Buffer was then degassed to remove any entrapped air that could result in suboptimal binding to the column. A pre-separation filter (70 µm, cat. no. 130-095-823) was fitted atop the column and 500 µl of degassed 1x Binding Buffer dispensed to moisten the filter and prime the column. Upon completed incubation, an additional 400 µl of Binding Buffer was added to the labeled cell suspension and mixed well. A 15mL conical tube was placed under the column and 500 µl of cell suspension applied to the moistened filter to limit passage of debris and cell clumps that would otherwise clog the column. The flow-through containing the unlabeled live cells was collected, and four column washes with 500 µl of Binding Buffer were subsequently recovered and combined for a final effluent volume of 2.5 mL.

Lymph node. Liquid lymph node (LN) biopsies were similarly processed as described above for PBMCs, with subtle differences attributed to cellular concentration and dramatically reduced viability. To account for a greater percentage of dead cells and lower initial cell concentrations in the LN samples, a minimum of four vials were combined for processing. The following details changes impacted by the aforementioned observations and difference in sample type.

A minimum of three LN vials per sample time point were thawed in a 37°C water bath and each carefully aspirated and passed through a moistened 70 µm pre-separation filter into a 15 mL conical tube. Each vial was rinsed with 1 mL of 1xDPBS to recover any residual cells, and subsequently combined to the 15 mL conical tube for a total of 6 mL. The 15 mL conical was centrifuged at 400 *xg* for 5 minutes at 4°C. The supernatant was discarded and cell pellet resuspended in 1 mL of 1xDPBS which was then transferred to a 1.5 mL Eppendorf tube. The cell suspension was washed in the 1.5 mL tube by centrifuging at 300 *xg* for 10 minutes at 25°C. The supernatant was discarded and cell pellet resuspended in 100 µl of Dead Cell Removal MicroBeads as previously stated, and incubated for 15 minutes at room temperature. To accommodate the higher volume of cells going into dead cell removal, an LS column (cat. no. 130-042-401) inserted in a MidiMACS™ separator (cat. no. 130-042-302) was alternatively used. A 1x Binding Buffer solution and pre-separation filter were prepared as described earlier for PBMCs. The LS column was primed by rinsing with 3 mL of Binding Buffer. Upon completed incubation, an additional 400 µl of 1x Binding Buffer was added to the labeled cell suspension and mixed well. A 15mL conical tube was placed under the column and 500 µl of cell suspension passed through the filter and LS column. The flow-through containing the unlabeled live cells was collected and four column washes with 3 mL of Binding Buffer were then collected and combined for a final flow-through volume of 12.5 mL.

Following completion of dead cell removal, cell counts were performed on both PBMC and LN collected effluents to determine total number of cells and ensure greater than 80% viability. The effluents of both sample types were mixed well with a wide bore pipette tip and their volumes alternatively used as dilution factors to trypan blue (i.e. 15 µl cell suspension + 2 µl trypan blue). Cell counts as well as all steps thereafter followed the methods described above in single cell suspension preparation.

Single-cell DNA sequencing. *Single-cell microfluidics*. The Tapestri® single-cell DNA sequencing V2 platform developed by MissionBio was performed as per manufacturer's instructions. Briefly, single-cell suspensions as processed above, were loaded onto the Tapestri® instrument and cells individually encapsulated via oil emulsion with a protease enzyme mix. The cells were then lysed and digested in a thermocycler to enable access to DNA. The encapsulated droplets containing histone-free DNA were then merged with a second droplet comprised of barcoding beads, PCR reagents and primers specific to the custom SIV amplicon panel (Supplementary Table S1). Regions of interest defined by amplicon design, were subjected to target amplification in a thermocycler and simultaneously tagged with a unique cell barcode. Emulsions were broken and samples stored at -20°C for less than six months before advancing to cleanup of PCR products.

Library cleanup proceeded with PCR product digestion in a thermocycler, followed by Ampure XP bead-based cleanup. In short, a 1.4X Ampure XP reagent (cat. no. A63881) ratio was initially incubated with the digested PCR products. The solution was placed on a magnet and beads allowed to separate. The beads containing adhered DNA were washed twice with fresh 80% ethanol, allowed to dry and all residual ethanol removed. The beads were then resuspended in nuclease-free water and further purified on the magnet. A fluorometer set to High Sensitivity dsDNA 1X (Qubit™ 4 Fluorometer, Invitrogen) was used to quantify 1 µl of the purified PCR product from library clean up one. Yields of DNA between the recommended 0.2 ng/µl to

4.0 ng/µl range were confirmed. A targeted library PCR step was then performed to add unique dual index sequences to the amplicons. Distinct pairings of index primers were used and recorded for separate samples to allow for multiplex sequencing. Upon completion of targeted library PCR, a second library cleanup was completed using a 0.69X Ampure XP bead ratio, and all other cleanup steps similarly executed as in cleanup one. The final DNA library concentration was again quantified by fluorometry and purity of on-target library analyzed with the Agilent 4150 TapeStation system (cat. no. G2992AA). For samples observed with >30% of smaller off-target product, further library purification using gel-size selection was carried out prior to sequencing. DNA samples containing unique dual index configurations were pooled and sequenced with an Illumina NovaSeq 6000 SP kit in 2 x 250 bp paired-end mode. The Tapestri® instrument protocol version referenced in this study was PN_3354H.

Gel library purification. Size selection of the library DNA was performed using an 11 cm by 14 cm 2% agarose gel (1.6g agarose + 1xTBE to 80 mL) with 8 µl of Syber safe gel dye. The samples for multiplexed sequencing that contained >30% off-target small fragments were pooled using equimolar quantities of on-target library prior to size selection. The gel was loaded with 25 µl of the pooled sample and 5 µl of 6x orange DNA loading dye-Fermentas into a single well. An O'Gene Ruler 1kb bp DNA 15 µl ladder-ready to go was added to an additional well to allow for identification of the band size of interest. The gel was then run at 64 V for 2 hours. The band corresponding to the size of interest at ~450bp was cut and QIAquick gel extraction performed as per manufacturer's instruction.

Bioinformatic analysis. *DNA pipeline virus optimization*. Fastq files output from sequencing were assessed for quality of read ends. Each file was run through FastQC to determine whether trimming of low quality 3' ends was necessary. For ends that required trimming, cutadapt[71] v3.4 was utilized. Modifications to pipeline uniformity and completeness thresholds were made to account for expected differences in largely "working amplicons" for human DNA, in which the pipeline was built to handle, as opposed to the breadth of variation experienced for viral DNA due to cumulative mutations and deletions. Scripts, as well as specifics on how and what pipeline parameters were changed, can be found at https://github.com/salemilab/scDNA-seq.

Near full-length proviral genome assembly. Amplicons from the scDNA-seq SIV panel were reviewed based on their targeted genome region to identify the best amplicon(s) to use for calling a cell SIV+. Based on previous NFL genome sequencing reports, the ends of HIV and SIV genomes were typically intact and so amplicon 37 was chosen. Cells were then considered SIV+ based on a criteria of possessing a minimum depth of 10 reads for amplicon 37. SIV+ barcodes were extracted and split from the rest of the cells, and individual bam files created for each barcode using samtools[72] v1.12. Consensus reconstruction using samtools[72] v1.12 and ivar[73] v1.3.1 resulted in fasta files for individual genomes that were then concatenated and aligned with viralMSA[74] for downstream analysis. All batch scripts to obtain the final consensus sequences can be found at https://github.com/salemilab/scDNA-seq.

Classification of proviral genomes. Proviral genome sequences were categorized as "deletion-defective", "hypermutation defective," or intact. To combat erroneous classifications as deletion defective owing to potential variability in primer-binding sites and therefore loss of sequenced reads, absence of two consecutive amplicons were used. Deletion-defective proviruses were further sub-classified by having either 3' only deletions, 5' only deletions, or double deletions comprising both ends. For any type of the above mentioned deletion-defects to be considered, the first two amplicons within the 5' and/or 3' regions were required to be missing. The extent of the deletion(s) from either or both ends were then measured until a nucleotide base was detected rather than a gap in the sequence. Hypermutated genomes were classified as previously described (Bruner *et al.*, 2019), by observing AG|AA at residues 4-5 and 33-34 of the RRE region. Resistance-associated mutations (RAMs) screened for in all sequences were those classified as major RAMs by Acharya *et al.* (2020)[75], Metzner *et al.* (2006)[76], and Stanford University HIV Drug resistance Database v.9.0, for Tenofovir, Emtricitabine, and Raltegravir.

Phylogenetic analyses. SIV NFL sequences obtained with scDNA-seq were first used to reconstruct phylogenetic trees for each animal and subsequently pruned to exclude all SIV+ cells missing *env gp120* entirely (*gp120* reference genome coordinates: $6832 - 8172$). Remaining SIV+ cells were run as both NFL genomes and *env gp120* alone, using IQ-tree v2.1.3 to construct likelihood maps and phylogenetic trees with 1,000 bootstrap replicates. Robinson-Foulds distances[49] were calculated and normalized in R[77] using the phangorn v2.8.1 package[78] RF.dist function in treedist. Resulting tree files are available in https://github.com/salemilab/scDNA-seq.

Diversity calculations. Hypermutated sequences identified using Hypermut2.0 available from the Los Alamos National Laboratory HIV sequence database (https://www.hiv.lanl.gov) were removed prior to diversity analysis. Calculations for population genetic diversity were performed in R[77] using the ape v5.6-2 package[79] function dist.dna to compute matrices of pairwise distances using the TN93 model of DNA evolution for SIV. $TRIM5\alpha$ diversity was assessed using mutational

frequencies at base positions of the targeted macaque $TRIM5\alpha$ chromosome 14 exon 8. Owing to poor amplification of amplicon 41 in certain samples resulting in large read number fluctuation, amplicon 42 was relied on.

Data and Code Availability

Raw SIV sequence data that support the findings of this study have been deposited in GenBank with accession codes (pending). Scripts concerning proviral genome and $TRIM5\alpha$ consensus reconstruction are available on the scDNA-seq GitHub along with scripts for generating the phylogenetic trees presented in the manuscript main text (https://github.com/salemilab/scDNA-seq). Guidance on setting pipeline parameters are additionally provided, as well as all recovered SIV sequences obtained in this study in fasta format.

References

1. Eron, J. J. HIV-1 protease inhibitors. *Clin Infect Dis* **30 Suppl 2**, S160–170 (2000).
2. Pomerantz, R. J. & Horn, D. L. Twenty years of therapy for HIV-1 infection. *Nat Med* **9**, 867–873 (2003).
3. Finzi, D. *et al.* Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* **278**, 1295–1300 (1997).
4. Wong, J. K. *et al.* Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. *Science* **278**, 1291–1295 (1997).
5. Siliciano, J. D. *et al.* Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4+ T cells. *Nat Med* **9**, 727–728 (2003).
6. Deeks, S. G. *et al.* International AIDS Society global scientific strategy: towards an HIV cure 2016. *Nat Med* **22**, 839–850 (2016).
7. Tyagi, M. & Bukrinsky, M. Human immunodeficiency virus (HIV) latency: the major hurdle in HIV eradication. *Mol Med* **18**, 1096–1108 (2012).
8. Salemi, M. & Rife, B. Phylogenetics and Phyloanatomy of HIV/SIV Intra-Host Compartments and Reservoirs: The Key Role of the Central Nervous System. *Curr HIV Res* **14**, 110–120 (2016).
9. Lorenzo-Redondo, R. *et al.* Persistent HIV-1 replication maintains the tissue reservoir during therapy. *Nature* **530**, 51–56 (2016).
10. Bruner, K. M. *et al.* Defective proviruses rapidly accumulate during acute HIV-1 infection. *Nat Med* **22**, 1043–1049 (2016).
11. Bruner, K. M. *et al.* A quantitative approach for measuring the reservoir of latent HIV-1 proviruses. *Nature* **566**, 120–125 (2019).
12. Hiener, B. *et al.* Identification of Genetically Intact HIV-1 Proviruses in specific CD4+T Cells from Effectively Treated Participants. *Cell Rep* **21**, 813–822 (2017).
13. Imamichi, H. *et al.* Defective HIV-1 proviruses produce viral proteins. *Proc Natl Acad Sci U S A* **117**, 3704–3710 (2020).
14. Falcinelli, S. D., Ceriani, C., Margolis, D. M. & Archin, N. M. New Frontiers in Measuring and Characterizing the HIV Reservoir. *Front Microbiol* **10**, 2878 (2019).
15. Pollack, R. A. *et al.* Defective HIV-1 Proviruses Are Expressed and Can Be Recognized by Cytotoxic T Lymphocytes, which Shape the Proviral Landscape. *Cell Host Microbe* **21**, 494–506 (2017).
16. Wei, H., Yu, D., Geng, X. & He, Y. Defective HIV-1 envelope gene promotes the evolution of the infectious strain through recombination in vitro. *BMC Infect Dis* **20**, 569 (2020).
17. RifeMagalis, B. *et al.* Phyloanatomic characterization of the distinct T cell and monocyte contributions to the peripheral blood HIV population within the host. *Virus Evol* **6**, veaa005 (2020).
18. Holland, B. R., Ketelaar-Jones, S., O'Mara, A. R., Woodhams, M. D. & Jordan, G. J. Accuracy of ancestral state reconstruction for non-neutral traits. *Sci Rep* **10**, 7644 (2020).
19. Richman, D. D. Introduction: challenges to finding a cure for HIV infection. *Curr Opin HIV AIDS* **6**, 1–3 (2011).

20. Sengupta, S. & Siliciano, R. F. Targeting the Latent Reservoir for HIV-1. *Immunity* **48**, 872–895 (2018).

21. Salazar-Gonzalez, J. F. *et al.* Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. *J Virol* **82**, 3952–3970 (2008).

22. Shankarappa, R. *et al.* Evolution of human immunodeficiency virus type 1 envelope sequences in infected individuals with differing disease progression profiles. *Virology* **241**, 251–259 (1998).

23. Chaillon, A. *et al.* HIV persists throughout deep tissues with repopulation from multiple anatomical sources. *J Clin Invest* **130**, 1699–1712 (2020).

24. Yebra, G. *et al.* Using nearly full-genome HIV sequence data improves phylogeny reconstruction in a simulated epidemic. *Sci Rep* **6**, 39489 (2016).

25. Rose, R. *et al.* HIV Maintains an Evolving and Dispersed Population in Multiple Tissues during Suppressive Combined Antiretroviral Therapy in Individuals with Cancer. *J Virol* **90**, 8984–8993 (2016).

26. Strickland, S. L. *et al.* Spatiotemporal dynamics of simian immunodeficiency virus brain infection in CD8+ lymphocyte-depleted rhesus macaques with neuroAIDS. *J Gen Virol* **95**, 2784–2795 (2014).

27. Gaebler, C. *et al.* Combination of quadruplex qPCR and next-generation sequencing for qualitative and quantitative analysis of the HIV-1 latent reservoir. *J Exp Med* **216**, 2253–2264 (2019).

28. Lee, G. Q. Chemistry and Bioinformatics Considerations in Using Next-Generation Sequencing Technologies to Inferring HIV Proviral DNA Genome-Intactness. *Viruses* **13** (2021).

29. Fox, J. M. *et al.* Long Terminal Repeat Circular DNA as Markers of Active Viral Replication of Human T Lymphotropic Virus-1 in Vivo. *Viruses* **8**, 80 (2016).

30. Richetta, C. *et al.* Two-long terminal repeat (LTR) DNA circles are a substrate for HIV-1 integrase. *J Biol Chem* **294**, 8286–8295 (2019).

31. White, J. A. *et al.* Complex decay dynamics of HIV virions, intact and defective proviruses, and 2LTR circles following initiation of antiretroviral therapy. *Proc Natl Acad Sci U S A* **119** (2022).

32. Rouzioux, C. & Richman, D. How to best measure HIV reservoirs? *Curr Opin HIV AIDS* **8**, 170–175 (2013).

33. McManus, W. R. *et al.* HIV-1 in lymph nodes is maintained by cellular proliferation during antiretroviral therapy. *J Clin Invest* **129**, 4629–4642 (2019).

34. Fletcher, C. V. *et al.* Persistent HIV-1 replication is associated with lower antiretroviral drug concentrations in lymphatic tissues. *Proc Natl Acad Sci U S A* **111**, 2307–2312 (2014).

35. Bender, A. M. *et al.* The Landscape of Persistent Viral Genomes in ART-Treated SIV, SHIV, and HIV-2 Infections. *Cell Host Microbe* **26**, 73–85 (2019).

36. Puller, V., Neher, R. & Albert, J. Estimating time of HIV-1 infection from next-generation sequence diversity. *PLoS Comput. Biol* **13**, e1005775 (2017).

37. Li, G. *et al.* An integrated map of HIV genome-wide variation from a population perspective. *Retrovirology* **12**, 18 (2015).

38. Kearney, M. *et al.* Human immunodeficiency virus type 1 population genetics and adaptation in newly infected individuals. *J Virol* **83**, 2715–2727 (2009).

39. Koboldt, D. C. Best practices for variant calling in clinical sequencing. *Genome Med* **12**, 91 (2020).

40. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genom Bioinform* **3**, lqab019 (2021).

41. Ho, Y. C. *et al.* Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. *Cell* **155**, 540–551 (2013).

42. Rose, P. P. & Korber, B. T. A hypermutation. *Bioinformatics* **16**, 400–401 (2000).

43. Imamichi, H. *et al.* Defective HIV-1 proviruses produce novel protein-coding RNA species in HIV-infected patients on combination antiretroviral therapy. *Proc Natl Acad Sci U S A* **113**, 8783–8788 (2016).

44. Gaebler, C. *et al.* Sequence Evaluation and Comparative Analysis of Novel Assays for Intact Proviral HIV-1 DNA. *J Virol* **95** (2021).

45. Orlandi, C. *et al.* A comparative analysis of unintegrated HIV-1 DNA measurement as a potential biomarker of the cellular reservoir in the blood of patients controlling and non-controlling viral replication. *J Transl Med* **18**, 204 (2020).

46. Hamid, F. B., Kim, J. & Shin, C. G. Distribution and fate of HIV-1 unintegrated DNA species: a comprehensive update. *AIDS Res Ther* **14**, 9 (2017).

47. Watts, J. M. *et al.* Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* **460**, 711–716 (2009).

48. Minh, B. Q., Nguyen, M. A. & von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol* **30**, 1188–1195 (2013).

49. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Math. Biosci.* **53**, 147–156 (1981).

50. De Scheerder, M. A. *et al.* HIV Rebound Is Predominantly Fueled by Genetically Identical Viral Expansions from Diverse Reservoirs. *Cell Host Microbe* **26**, 347–358 (2019).

51. Romero-Severson, E. O., Bulla, I. & Leitner, T. Phylogenetically resolving epidemiologic linkage. *Proc Natl Acad Sci U S A* **113**, 2690–2695 (2016).

52. Abdel-Mohsen, M. *et al.* Recommendations for measuring HIV reservoir size in cure-directed clinical trials. *Nat Med* **26**, 1339–1350 (2020).

53. Falcinelli, S. D. *et al.* Longitudinal Dynamics of Intact HIV Proviral DNA and Outgrowth Virus Frequencies in a Cohort of Individuals Receiving Antiretroviral Therapy. *J Infect Dis* **224**, 92–100 (2021).

54. Wang, X. Q. & Palmer, S. Single-molecule techniques to quantify and genetically characterise persistent HIV. *Retrovirology* **15**, 3 (2018).

55. Siliciano, J. D. & Siliciano, R. F. Enhanced culture assay for detection and quantitation of latently infected, resting CD4+ T-cells carrying replication-competent virus in HIV-1-infected individuals. *Methods Mol Biol* **304**, 3–15 (2005).

56. Einkauf, K. B. *et al.* Intact HIV-1 proviruses accumulate at distinct chromosomal positions during prolonged antiretroviral therapy. *J Clin Invest* **129**, 988–998 (2019).

57. Simonetti, F. R. *et al.* Intact proviral DNA assay analysis of large cohorts of people with HIV provides a benchmark for the frequency and composition of persistent proviral DNA. *Proc Natl Acad Sci U S A* **117**, 18692–18700 (2020).

58. Kinloch, N. N. *et al.* HIV-1 diversity considerations in the application of the Intact Proviral DNA Assay (IPDA). *Nat Commun* **12**, 165 (2021).

59. Patro, S. C. *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* **116**, 25891–25899 (2019).

60. Zhao, Y. *et al.* Diverse alterations associated with resistance to KRAS(G12C) inhibition. *Nature* **599**, 679–683 (2021).

61. Cole, B. *et al.* In-depth single-cell analysis of translation-competent HIV-1 reservoirs identifies cellular sources of plasma viremia. *Nat Commun* **12**, 3727 (2021).

62. Demaree, B. *et al.* Joint profiling of DNA and proteins in single cells to dissect genotype-phenotype associations in leukemia. *Nat Commun* **12**, 1583 (2021).

63. Miles, L. A. *et al.* Single-cell mutation analysis of clonal evolution in myeloid malignancies. *Nature* **587**, 477–482 (2020).

64. Lee, G. Q. *et al.* Clonal expansion of genome-intact HIV-1 in functionally polarized Th1 CD4+ T cells. *J Clin Invest* **127**, 2689–2696 (2017).

65. Klatt, N. R., Chomont, N., Douek, D. C. & Deeks, S. G. Immune activation and HIV persistence: implications for curative approaches to HIV infection. *Immunol Rev* **254**, 326–342 (2013).

66. Das, A. T., Pasternak, A. O. & Berkhout, B. On the generation of the MSD-ψ class of defective HIV proviruses. *Retrovirology* **16**, 19 (2019).

67. Morita, K. *et al.* Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. *Nat Commun* **11**, 5327 (2020).

68. National Research Council of the National Academies. *Guide for the Care and Use of Laboratory Animals* (National Academic Press, Washington, DC, 2011), 8 edn.

69. Rife, B. D. *et al.* Correction for Rife et al., Evolution of Neuroadaptation in the Periphery and Purifying Selection in the Brain Contribute to Compartmentalization of Simian Immunodeficiency Virus (SIV) in the Brains of Rhesus Macaques with SIV-Associated Encephalitis. *J Virol* **90**, 8947 (2016).

70. Strickland, S. L. *et al.* Significant genetic heterogeneity of the SIVmac251 viral swarm derived from different sources. *AIDS Res Hum Retroviruses* **27**, 1327–1332 (2011).

71. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12, DOI: 10.14806/ej.17.1.200 (2011).

72. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, DOI: 10.1093/gigascience/giab008 (2021). Giab008, https://academic.oup.com/gigascience/article-pdf/10/2/giab008/36332246/giab008.pdf.

73. Grubaugh, N. D. *et al.* An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol* **20**, 8 (2019).

74. Moshiri, N. ViralMSA: massively scalable reference-guided multiple sequence alignment of viral genomes. *Bioinformatics* **37**, 714–716 (2021).

75. Acharya, A. *et al.* Variability in HIV-1 Integrase Gene and 3'-Polypurine Tract Sequences in Cameroon Clinical Isolates, and Implications for Integrase Inhibitors Efficacy. *Int J Mol Sci* **21** (2020).

76. Metzner, K. J. *et al.* Tenofovir treatment augments anti-viral immunity against drug-resistant SIV challenge in chronically infected rhesus macaques. *Retrovirology* **3**, 97 (2006).

77. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2021).

78. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).

79. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289–290 (2004).

Additional information

Accession codes (where applicable);

Competing interests

The The author(s) declare no competing interests.

Supplementary

Table S1. Primer panel used in SIV scDNA-seq. Amplicon IDs are listed with their forward and reverse primers. Amplicons 1 – 39 start and end coordinates are with respect to our custom reference genome, while amplicon coordinates for 41 and 42 are with respect to the SPRY domain of the *Rhesus* macaque $TRIM5\alpha$ gene.

Tapestri® Single-Cell DNA SIV Panel					
Amplicon ID	**Amplicon Coordinate: Start**	**Amplicon coordinate: end**	**Primer (Fwd)**	**Primer (Rev)**	**Targets**
UF_AMP1	7	250	AGTGTGTGTTCCCATCTCTCC	GTCTTTGGGTATCTAATTCCTGGTC	SIVmacL28
UF_AMP2	251	512	ATTTGGCTGGCTATGGAAATTAGTC	AGCCATGTTAAGAAGGCCTCT	SIVmacL28
UF_AMP3	625	886	CTGCATTTTGCTCTGTATTCAGTC	TCTTCTTATTACCGAGTACCGAGT	SIVmacL28
UF_AMP4	918	1157	GAAACCGAAGCAGGAAAATCCC	CCTTCCTGGATAAAAGACAGCTATT	SIVmacL28
UF_AMP5	1177	1438	AGATGGGCGCGAGAAACT	CACCAGATGACGCAGACAGT	SIVmacL28
UF_AMP6	1439	1668	CATTCACGCAGAAGAGAAAGTGA	TTTCCTCTATCAATTTTACCCAGGC	SIVmacL28
UF_AMP7	1669	1936	AGAAATTTGGAGCAGAAGTAGTGC	TGTCTGTACATCCACTGGATTTGT	SIVmacL28
UF_AMP8	1937	2181	ACAGAACCCCATACCAGTAGG	GCACTAGCTTGCAATCTGGG	SIVmacL28
UF_AMP9	2182	2382	TGAAGGGGCTGGGTGTGAA	AGTGTCCTTCTTTCCCACAATTC	SIVmacL28
UF_AMP10	2384	2630	TGCAAGGCAATGCAGAGC	CTGCTTTTCTCTCTGCTGCTT	SIVmacL28
UF_AMP11	2631	2899	AGAGAAAGCAGAGAGAAGCCT	CCCTTTAATCCTTTTGCCTAAAACTT	SIVmacL28
UF_AMP12	2900	3158	ACAATCATGACAGGGGACACC	ATGTGGGGGTGTTGTATGGA	SIVmacL28
UF_AMP13	3175	3403	AGATAAGAACAAATGGAGAATGCTGA	TTTTCCTGGCTCTGCATTATTTACT	SIVmacL28
UF_AMP14	3404	3646	CGATACATTTATAAGGTTCTGCCTCAG	TTTTTGGAATTTCTCTTCTGGGGT	SIVmacL28
UF_AMP15	3656	3891	TTTCAATGGATGGGGTACGAATTG	TCCTCATATTCTGCTTCTGCCA	SIVmacL28
UF_AMP16	3897	4155	AAATAATTCTCAGTCAGGAACAAGAAGG	TCAACTGGTAAGTGGAATTTTGGAACC	SIVmacL28
UF_AMP17	4156	4424	GAGGGATGTATGGGAACAGTGG	TGAGAAATGCTTCCAATTCTGCTTG	SIVmacL28
UF_AMP18	4425	4692	TGGCATTGACAGACTCAGGG	TCATGTTCTTCTTGTGCTGGC	SIVmacL28
UF_AMP19	4709	4917	GTAAAAGAATTGGTATTCAAATTTGGATTGC	ACTTCTGCTTCTATGAATCCACTAGC	SIVmacL28
UF_AMP20	4918	5169	AATTCCACAAGAGACAGGAAGACA	ACTGAATTTGCTTGTTCCCTGA	SIVmacL28
UF_AMP21	5170	5436	AGAAACCATAGTATTAATGGCAGTTCA	GGTACTACCTTAATGTCTGTCCCT	SIVmacL28
UF_AMP22	5457	5702	TCAAAGATTATGGAGGAGGGAAAGA	TCTTTCTGGTGTCAAATGCCAA	SIVmacL28
UF_AMP23	5703	5961	GGGTGGCTCAGTACTTATGCA	GATTCTCTCCCTGGGATCTGAC	SIVmacL28
UF_AMP24	5967	6191	TGGAAACAGTGGAGAAGAGACC	TTCTTGCAATGCATAAATAAAGCCTT	SIVmacL28
UF_AMP25	6192	6446	AGGCTGTAGATGTCTAGGGGA	CCCTCAAGGGTGTCTCCAT	SIVmacL28
UF_AMP26	6447	6713	AGCAGGAGAACTCATTAGAATCCT	CAGAAGATGTATTAGCCTTAGCCTT	SIVmacL28
UF_AMP27	6717	6951	CAAACAAGTAAGTATGGGATGTCTTGG	GCTTTCTGTAACATTAAGGGCCA	SIVmacL28
UF_AMP28	6952	7218	TTTGATGCTTGGGAGAATACAGTC	TTGCTCTTGTTCCAAGCCTG	SIVmacL28
UF_AMP29	7219	7476	ATGATAGGCTGTAAATTCAACATGACA	GCCTGAATAATTTGTGTCATTACATCT	SIVmacL28
UF_AMP30	7477	7739	TTTATGCCTAAATGTTCTAAGGTGGT	CTATCATTGATTGGTTGTGAGTGG	SIVmacL28
UF_AMP31	7741	7993	CCAAAGCAGGCATGGTGTT	CCTTTGGCTTCTGGTTAGTTGT	SIVmacL28
UF_AMP32	7994	8238	AACAGCATAAAAGGAATTACGTGCC	TGGAGTGATCTCTACTAATTTATAATCTCC	SIVmacL28
UF_AMP33	8252	8517	CCACAGATGTGAAGAGGTACACT	CTCGATGGCAGTGACCCTA	SIVmacL28
UF_AMP34	8518	8786	AAGTACTTAAAGGACCAGGCGC	CAAGAAGCAAGGTCAAACCAATTG	SIVmacL28
UF_AMP35	8797	9025	ATACAATATGGAATTTATGTAGTTGTAGGAGT	CTATCTGCCAAGGCCAGGA	SIVmacL28
UF_AMP36	9031	9229	ATTCATTTCCTAATCCGCCAACTG	GGACCGCCTCCTGGAAATA	SIVmacL28
UF_AMP37	9230	9498	AAGTCGGCTGGAGATCTGC	ACCCCTACCAAGTCATCATCTT	SIVmacL28
UF_AMP38	9499	9768	ACCAGTGATGCCACGAGT	TCATCCTCCTGTGCCTCATC	SIVmacL28
UF_AMP39	9776	10025	TATTTAATGCATCCAGCTCAAACTTC	CCGTAACATCCCCTTGTGGA	SIVmacL28
UF_AMP40	28203	28432	CTCCTTTGCTTGTATTCAATATTATGCA	TGTTCAGAAGTTATAAGGAGGGGT	None
UF_AMP41	28433	28674	AGTTTCCTCCCAGTTTTCTCTCA	ATTGAAATTCGTGAGTGACTGAAAC	SPRY
UF_AMP42	28675	28935	TATTGTACTGGCGTCCTGGG	AGAGAGGGGCACAATGAAAGG	SPRY

Table S2. Hypermutated SIV genomes identified by Hypermut2.0. A total of 863 SIV sequences were assessed for APOBEC-induced hypermutation. Identified hypermutants are accounted for under Sequence ID and output values contributing to Fisher-Exact P-values displayed.

Hypermutants					
Sequence ID	**Muts (Match Sites)**	**Out of (Potential Mut Sites)**	**Controls (Control Muts)**	**Out of (Potential Controls)**	**Fisher Exact P-value**
119HG02PB Cell 7	31	1354	8	1231	0.000397436
197HG02PB Cell 51	40	1354	13	1231	0.000423786
197HG02PB Cell 10	14	1354	2	1231	0.00359902
197JA41PB Cell 67	6	1354	0	1231	0.0205425
197HG02K Cell 4	5	1354	0	1231	0.0392881
197HG02K Cell 10	5	1354	0	1231	0.0392881
197JA41K Cell 182	7	1354	1	1231	0.0466265

Table S3. Symmetry difference between phylogenetic trees. Robinson-Foulds distance (RF-distance) calculated between the NFL and *gp120* reconstructed phylogenetic trees for each animal.

Macaque	RF-distance
JA41	0.942
JP70	0.696
HG02	0.918

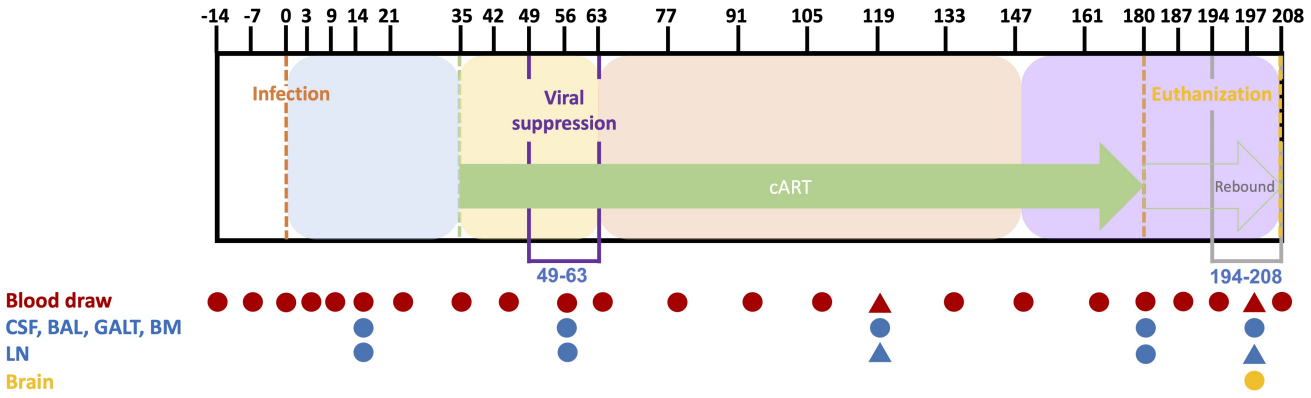

Figure S1. Therapy interrupted HIV animal model. *Rhesus* macaques were infected with a pathogenic strain of SIV. Samples of various types were taken at the labeled time points (longitudinal = blue, necropsy = yellow) throughout the therapy interruption timeline, as indicated by the legend on the left. Samples used in this study are identified by triangles, while circles represent other sampling events not included.

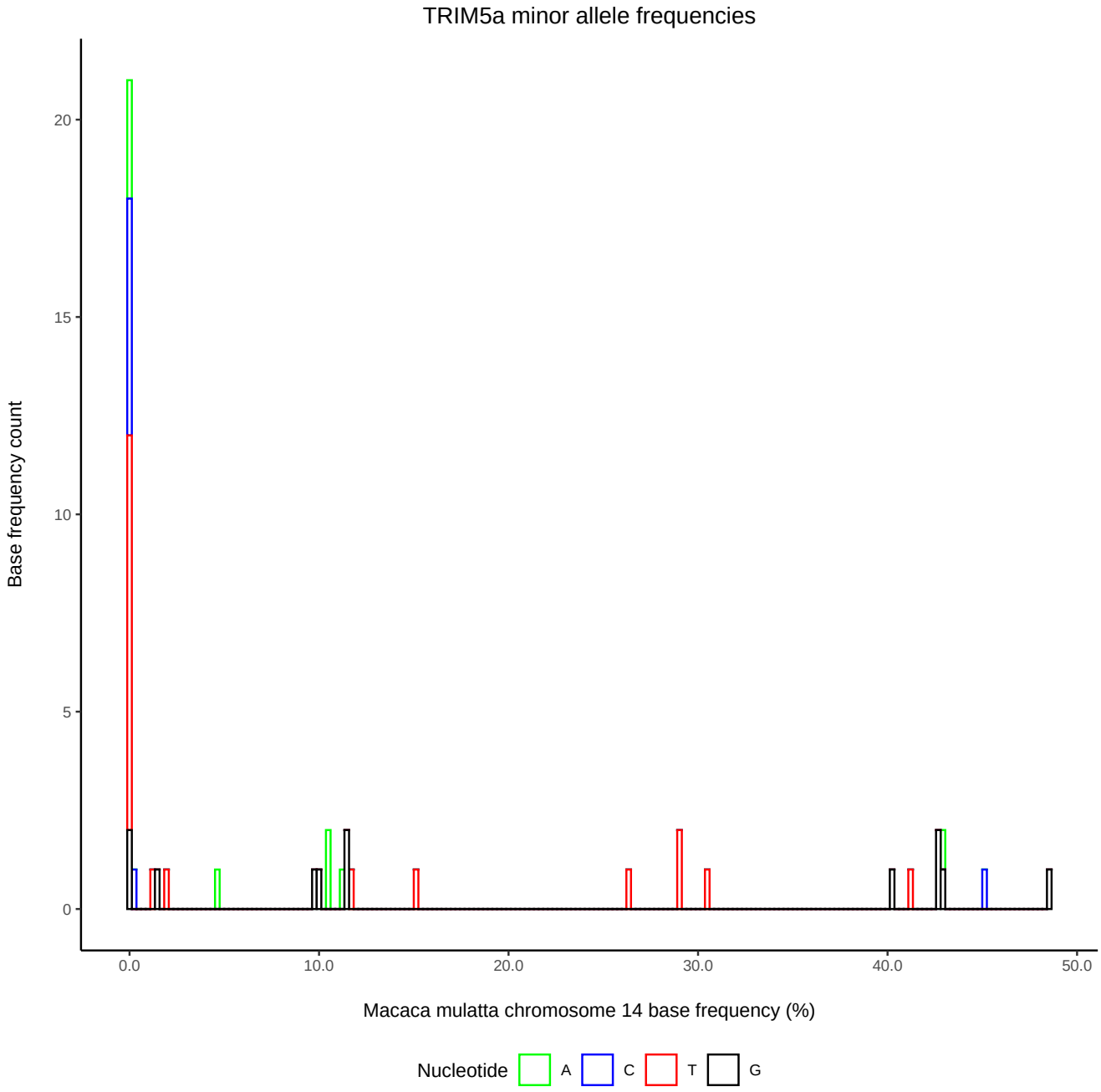

Figure S2. Distribution of TRIM5a minor allele frequencies. Minor allele nucleotide bases are distributed by percent frequency observed in amplicon 42 of the *Rhesus* macaque TRIM5a chromosome 14. Base counts are stacked to showcase nucleotide prevalence at all indicated frequencies.

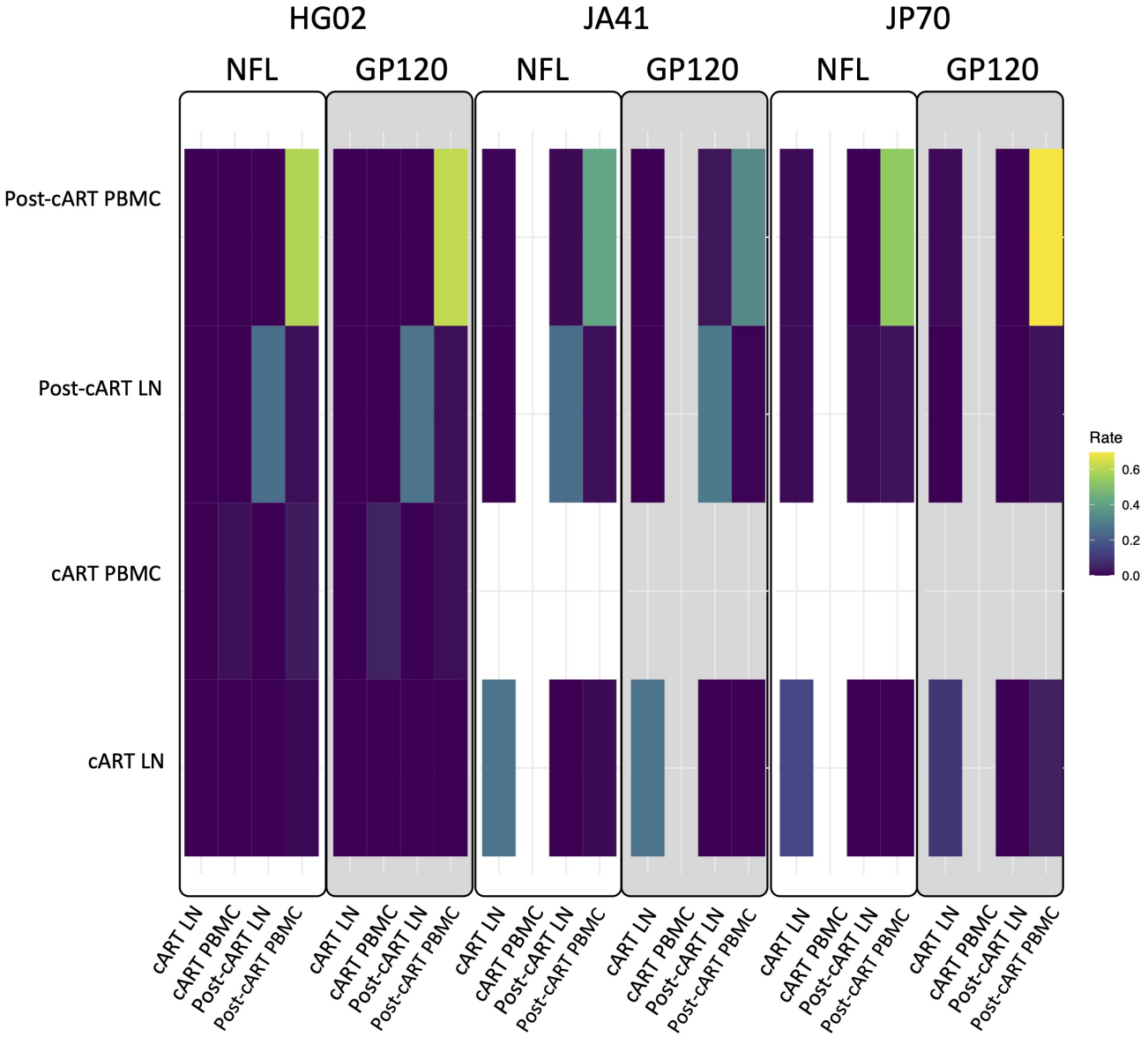

Figure S3. Density map of phylogeography tissue transitions. Animals are represented with two major columns: NFL and *gp120*. Each row is labeled with a time point and tissue that signifies the origin of the sequence, while each of the four smaller columns signify the transition "to" time point and tissue. Only sequences associated with branching events having > 90 bootstrap support were included.

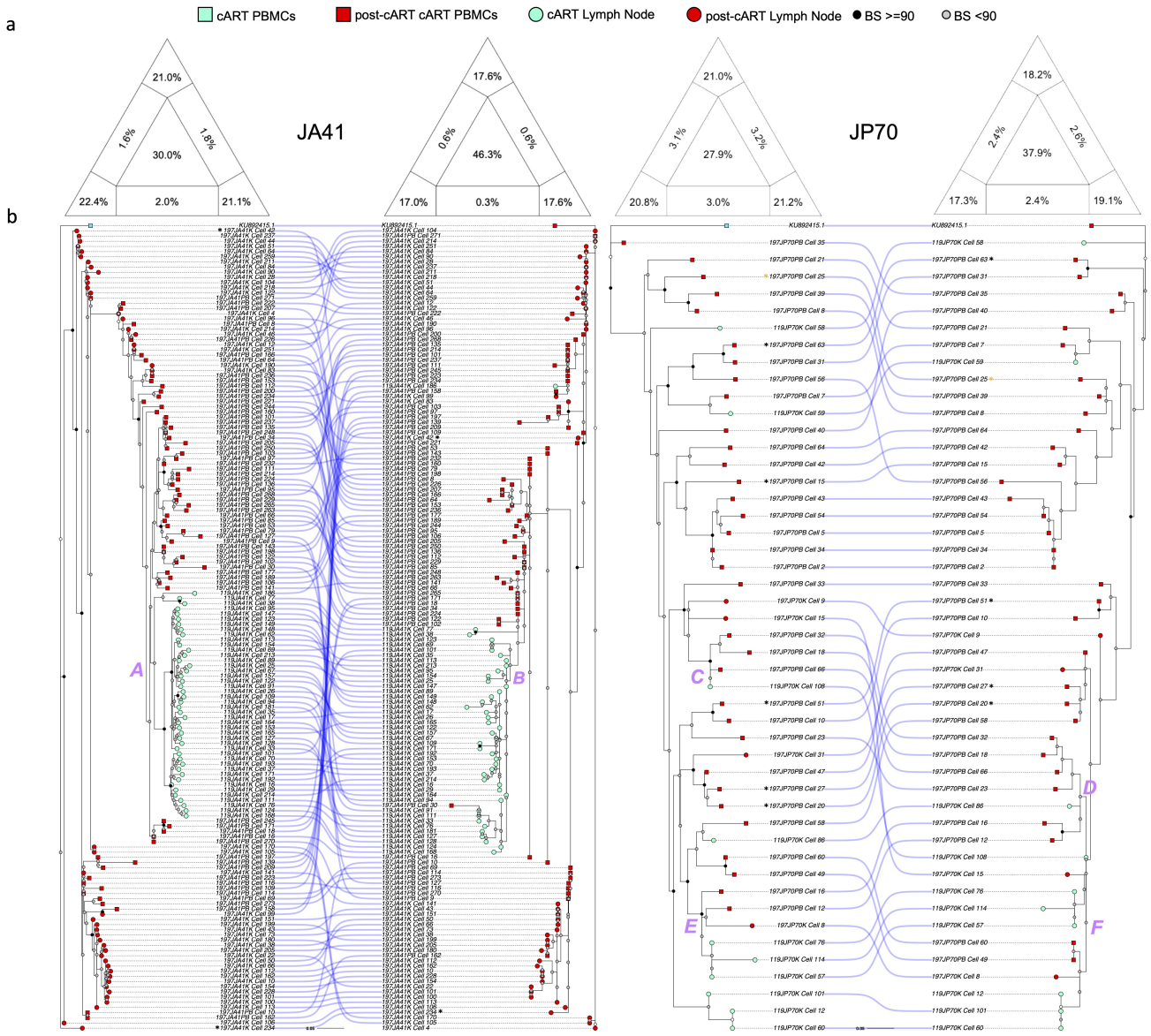

Figure S4. Comparison of phylogenetic resolution for near-full genomes and subgenomic *env gp120*. a. Phylogenetic likelihood map pairs for animals JA41 (left) and JP70 (right) with NFL genome sequences (left) and *gp120* (right). **b.** Corresponding maximum likelihood phylogenetic tree (legend at the top). Shape indicates tissue type (square: PBMC, circle: LN), while color indicates time point (cyan: cART, red: post-cART). Support for branching events was provided through bootstrapping[48], with $\geq 90\%$ considered well-supported and represented by a black circle and $< 90\%$ represented by a gray circle. Asterisks indicate intact genomes, and if colored yellow, further indicate those genomes as in the form of a 2-LTR circle. Significant differences in clade formation support between NFL genomes and *env gp120* alone are marked with letters A-F.

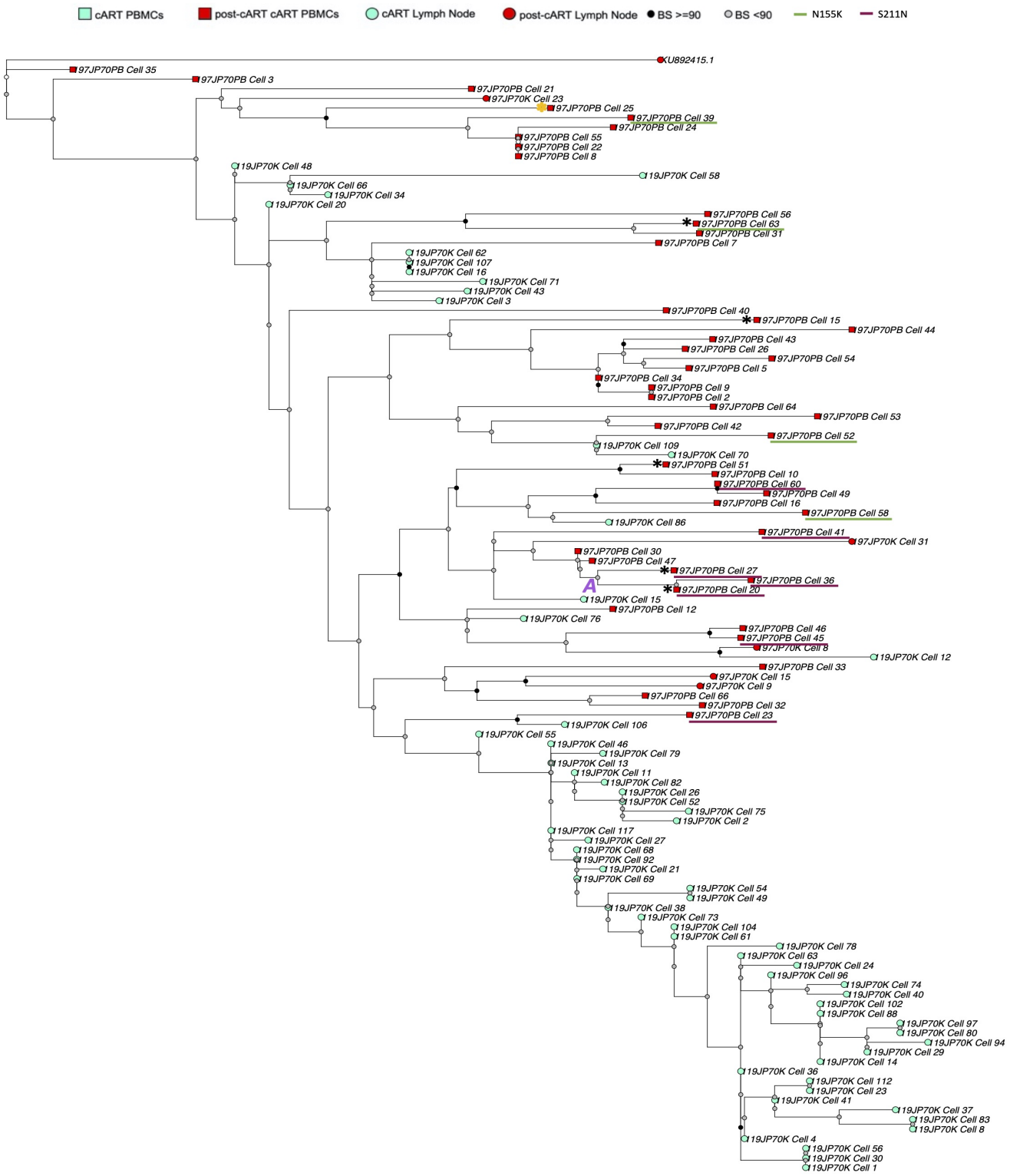

Figure S5. Phylogenetic reconstruction of JP70 NFL genomes. Generated maximum likelihood phylogenetic tree (legend at top). Shape indicates tissue type (square: PBMC, circle: LN), while color indicates time point (cyan: cART, red: post-cART). Support for branching events was provided through bootstrapping[48], with $\geq 90\%$ considered well-supported and represented by a black circle and $< 90\%$ represented by a gray circle. Asterisks indicate intact genomes, and if colored yellow, further indicate those genomes as in the form of a 2-LTR circle. Genomes harboring DRMs are underlined with corresponding color in legend (green: N155K, maroon: S211N). Significant clades containing DRMs are labeled with a letter.

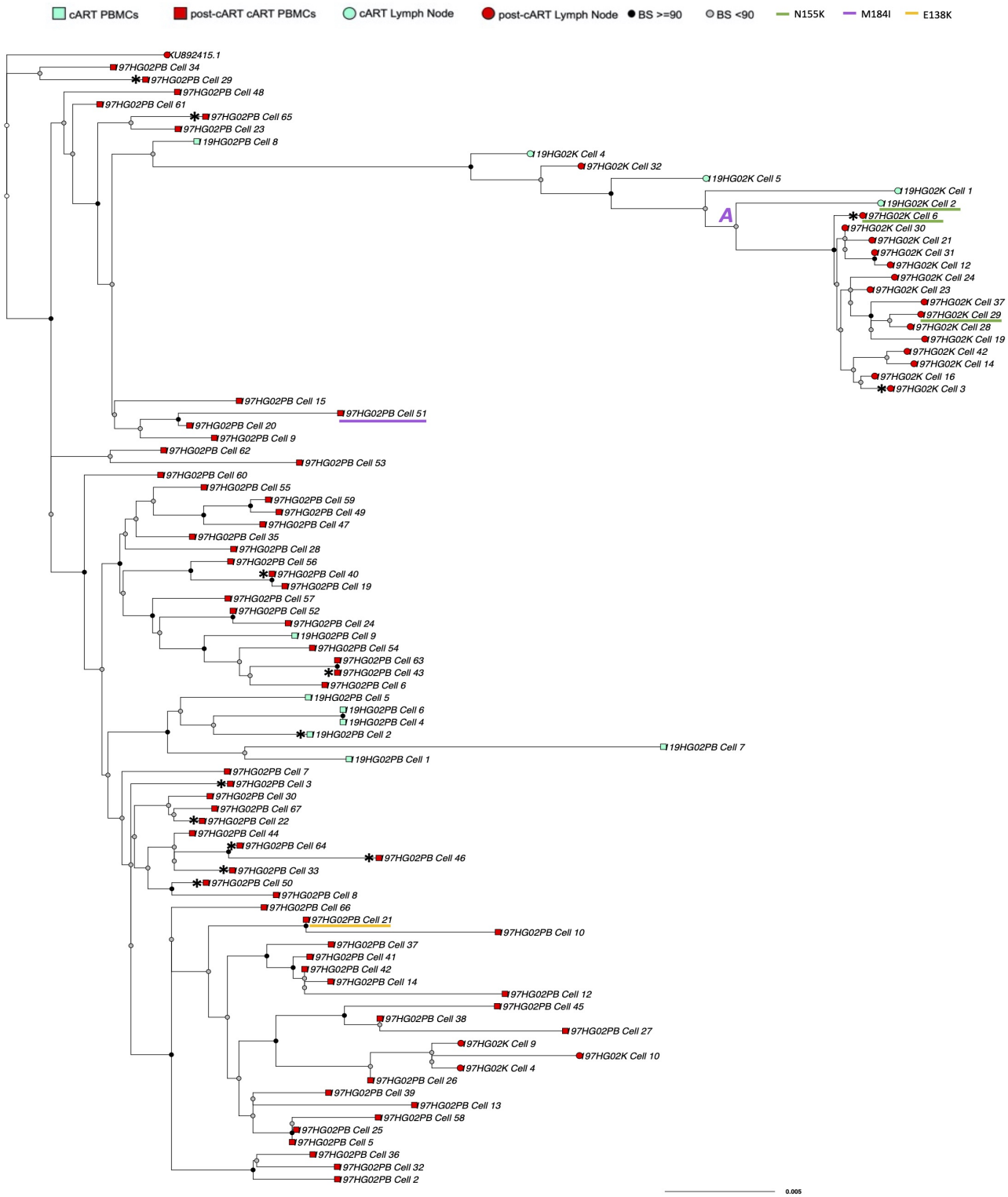

Figure S6. Phylogenetic reconstruction of HG02 NFL genomes. Maximum likelihood phylogenetic tree (legend at the top). Shape indicates tissue type (square: PBMC, circle: LN), while color indicates time point (cyan: cART, red: post-cART). Support for branching events was provided through bootstrapping[48], with $\geq 90\%$ considered well-supported and represented by a black circle and $< 90\%$ represented by a gray circle. Asterisks indicate intact genomes. Genomes harboring DRMs are underlined with corresponding color in legend (green: N155K, purple: M184I, yellow: E138K) Significant clades containing DRMs are labeled with a letter.