5 Intra-host Evolution of Simian Immunodeficiency Virus

5.1 Introduction and aims

As a retrovirus, HIV/SIV is present as both DNA and RNA within an infected cell. After the virus binds to the cell surface and gains entry to the cytoplasm, it undergoes capsid un-coating, reverse transcription of the RNA genome and insertion of the newly formed proviral DNA into the cellular genome. Most HIV/SIV infected cells contain a single proviral copy of viral DNA, indicating a single round of infection[36,382]. Recombination (also called superinfection) occurs when two genetically distinct viruses infect a single cell, and is estimated to occur at a rate of 1% per genome per generation[383]. Genomic integration does not guarantee expression, which requires cellular activation and the HIV/SIV DNA inserted into a favourably transcribed region for viral genes to be expressed. Cellular restriction factors, including the APOBEC family of mRNA editing enzymes, can also limit the replication of HIV/SIV by introducing lethal G-to-A mutations throughout the viral genome. Thus, the pool of DNA-infected cells contains only a fraction of cells that produce viral RNA and new, infectious virus.

Upon initiation of cART, viral RNA levels in plasma decline significantly, dropping by 99% within two weeks due to the elimination of free virus and productively infected cells [384]. The decay in cells containing HIV DNA is less dramatic, with an estimated 8 fold to 30 fold decline immediately following therapy and plateauing after 4 years of cART[385]. In spite of this, the number of cells with replication-competent HIV is low (estimated at 1 in a million PBCMs from patients on long-term ART). This gap in frequencies, between the total DNA-infected cells and the smaller subset that produce virus, is also reflected in the viral sequences derived from the two populations. DNA-infected cells include so-called “graveyard” sequences, and recent studies suggest that the integration site of proviral DNA [69,386,387] may drive persistence, clonal expansion, and RNA expression. In active untreated infection, turnover of different T cell subsets, as well as surveillance of CD8 T cells to different regions of lymph nodes, may affect which T cells harbor graveyard sequences and produce the most viral RNA.

To investigate the evolution of SIV in lymph nodes, we adapted a protocol from Gall et all[327] to amplify and sequence nearly full-length SIV genomes from plasma, cell-associated RNA, and proviral DNA. We sequenced SIV genomes from proviral DNA and cell-associated RNA in lymph node T cell subpopulations to examine the relationship between integrated genomes and transcribed viral RNA. We also sequenced plasma virus from the same time points to examine the contributions of lymph node cells to plasma sequences throughout infection.

The aims of this chapter are to:

- Develop a protocol to sequence SIV cell-associated RNA and proviral DNA near-full-length genomes
- Evaluate the relationship between circulating plasma virus and cell-associated viral sequences in lymph node T cell subsets.
- Measure the compartmentalization of virus in lymph node T cell subsets, and the relative mutation rates and diversity of proviral DNA and cell-associated RNA
- Compare the intra-host evolution and compartmentalization of virus between individual animals.

5.2 Results

5.2.1 A near-full-length genome sequencing protocol for SIV DNA and RNA

While sequencing a single gene or region of the HIV/SIV genome can identify viral subtypes and provide insight into viral adaptations, full genomes offer a more complete picture of virus evolution. Given the short time frame of this study (twenty four weeks) and the relatively small number of viral mutations expected within a single host in that time period, we wanted to sequence near full length SIV genomes to be better equipped to detect viral intra-host evolution. We adapted the protocol used by Gall et al for HIV sequencing, using four overlapping amplicons spanning the entire coding region of the genome. We designed new primers targeting semi-conserved regions to amplify 2-3 kb overlapping regions of the SIV genome with help from the O'Connor laboratory at the University of Wisconsin, using SIVmac239 as a reference sequence. After a single round of RT-PCR (RNA samples) or PCR (DNA samples), we visualized amplicons on an agarose gel to verify amplification (Fig. 5.1) before pooling amplicons for Illumina HiSeq sequencing.

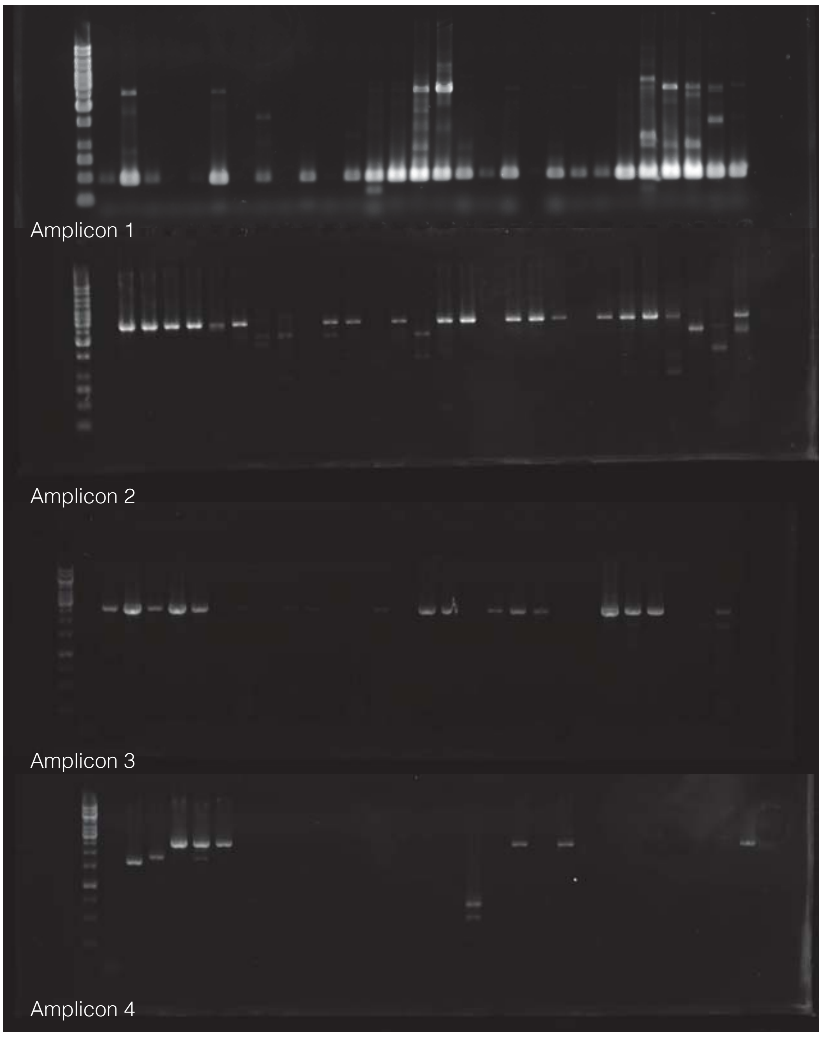

Figure 5.1 Verification of amplicons from SIV RNA (amplicon 1) and DNA (amplicons 2-4). Samples were run on an agarose gel with a 1kb ladder (left) to confirm amplification of the four PCR products (2-3 kb in length) from SIV RNA and DNA. Shown are (top to bottom) representative gels of amplicons 1, 2, 3, and 4. Amplicons 2 and 4 had the highest success rate (indicated by multiple lanes with bands in the 2-3kb range) while amplicon 1 had the highest failure rate.

5.2.1.1 Plasma virus sequencing

Plasma virus was sequenced from three animals (ZF61, ZF76, ZG13) at four time points, at 2, 4, 10, and 24 weeks post infection. Full-genome sequences, with all four amplicons sequenced and assembled into a single contiguous sequence (contig), were generated from six of the twelve samples (Table 5.1). Five of the samples had coverage of 75% of the genome, with three amplicons successfully sequenced and assembled, and the remaining sample had two amplicons sequenced and assembled. Amplicon 1 (covering the first 3kb of the genome) failed to assemble in five of the samples, while amplicon 4 failed to assemble in two of the samples. Amplicons 2 and 3 assembled in all samples.

Table 5.1 SIV PCR Amplicons sequenced from plasma RNA, proviral DNA, and cell-associated RNA

	Plasma RNA	Proviral DNA	Cell-associated RNA
Total Sequences	**12**	**62**	**65**
4 amplicons	6	0	30
3 amplicons	5	42	16
2 amplicons	1	15	8
1 amplicon	0	5	10
0 amplicons	0	0	1
Amplicon Failure			
Amplicon 1	5	62	27
Amplicon 2	0	4	5
Amplicon 3	0	3	19
Amplicon 4	2	18	15

5.2.2 Sequencing and assembly of SIV DNA and cell-associated RNA

To measure how lymph node SIV RNA sequences reflected circulating plasma virus and compare viral sequences present in different T cell subpopulations, we sorted six subsets of central memory CD4 T cells from iliac lymph nodes (at 2, 4, 10, and 24 weeks post infection) and mesenteric lymph nodes (at 24 weeks post infection). From CD3+CD8- CD4- cells, we sorted CXCR5- and CXCR5- cells, and from CD4 bright and dim central memory we sorted CXCR5++PD1++, CXCR5+PD1+, CXCR5+PD1-, and CXCR5- cells (for sample gating criteria, see chapter 4). We extracted DNA and RNA from the samples and split each nucleic acid sample four ways for amplification (Fig. 5.1) Sixty-two DNA samples that had sufficient material for sequencing library preparation and were sequenced on the Illumina HiSeq platform, although none had full genomes assembled (Table 5.2). However, forty-two had three of four amplicons sequenced and assembled, fifteen had two amplicons, and five had a single amplicon assemble. Amplicon 1 failed to assemble in all of the DNA samples sequenced, and amplicons 4 assembled in only eighteen of the sixty two DNA sequences. Amplicons 2 and 3 were assembled in nearly all the samples, failing only in 4 and 3 cases, respectively. Sixty-five cell-associated RNA sequences had sufficient material for sequencing, and thirty had full genomes assembled from the four PCR amplicons. Sixteen had three amplicons successfully amplified, eight had two amplicons, and ten had a single amplicons assemble. Amplicon 1 failed in twenty-seven of the samples, amplicons three failed in nineteen samples, amplicons 4 failed in fifteen samples, and amplicon 2 failed in five samples.

Of the six memory CD4 T cell subsets sorted for viral sequencing, CXCR5++PD1++ and CXCR5+PD1+ CD4 central memory T samples had the highest recovery of DNA amplicons, with 86% and 79% of samples containing 3 amplicons. CXCR5+PD1- and CXCR5- samples had 67% and 50% of samples with three amplicons sequenced and assembled, while CD4 null samples had 43% and 67% of samples assembled (for CXCR5+ and CXCR5-, respectively). While amplicon 1 failed to assemble in all the DNA samples, amplicons 2 and 3 failed in an average of 18% of DNA samples, while amplicon 4 failed in 41% of samples. In the RNA samples, three of the four CD4 high populations had 3 or 4 amplicons assemble (79% of CXCR5++PD1++, 92% of CXCR5+PD1+, and 82% of CXCR5-). Null CXCR5+ samples also had good recovery of assembled amplicons, with 75% of samples with 3 or 4 amplicons. CD4 bright CXCR5+PD1- cells had only 33% of samples with 3 or 4 amplicons, and CD4 null CXCR5- cells had 40% of samples with 3 or 4 amplicons. Amplicon 1 failed to assemble in 25%-83% of samples, amplicon 2 failed to assemble in 8%-40% of samples, amplicon 3 failed in 18%-60%, and amplicon 4 failed in 9%-58% of samples. Overall, amplicons 2 and 3 were most likely to assemble, while amplicon 4 and particularly amplicon 1, did not assemble in most of the samples. The different CD4 T cell subsets had similar efficiencies of sequence recovery from DNA, while in RNA samples, CD4 bright CXCR5+PD1- and CD4 null CXCR5- samples did not assemble and recover as many full genomes as the other populations.

Table 5.2 Amplicons sequenced per sample and amplicon failure rate.

	CD4 Bright/Dim				CD4 Null	
Proviral DNA	**CXCR5++ PD1++**	**CXCR5+ PD1+**	**CXCR5+ PD1-**	**CXCR5-**	**CXCR5+**	**CXCR5-**
Total Sequences	14	14	12	12	7	3
4 amplicons	0	0	0	0	0	0
3 amplicons	12	11	8	6	3	2
2 amplicons	1	3	4	0	1	0
1 amplicon	1	0	0	6	3	1
0 amplicons		0	0	0	0	0
Amplicon Failure						
Amplicon 1	14	14	12	12	7	3
Amplicon 2	4	3	1	1	2	1
Amplicon 3	3	3	1	1	2	1
Amplicon 4	5	6	5	7	3	0

	CD4 Bright/Dim				CD4 Null	
Cell-associated RNA	**CXCR5++ PD1++**	**CXCR5+ PD1+**	**CXCR5+ PD1-**	**CXCR5-**	**CXCR5+**	**CXCR5-**
Total Sequences	14	12	12	11	8	5
4 amplicons	8	9	1	7	4	0
3 amplicons	3	2	3	2	2	2
2 amplicons	0	0	5	0	0	1
1 amplicon	3	0	3	2	2	2
0 amplicons		1	0	0	0	0
Amplicon Failure						
Amplicon 1	6	3	10	4	3	3
Amplicon 2	3	2	1	1	1	2
Amplicon 3	6	2	6	2	3	3
Amplicon 4	5	2	7	1	1	2

5.2.3 Hypermutation in assembled DNA contigs

For each sample, a consensus genome was created by merging overlapping contigs generated in IVA[333]. In some samples, multiple contigs covering the same genomic region were generated, indicating the presence of multiple distinct viral genomes within the sample. All contigs mapping to the SIV genome underwent hypermutation analysis using Hypermut [334], which detects APOBEC-driven G-to-A mutations in HIV and SIV genomes. Hypermutated sequences are replication-incompetent but if included in reconstructions of viral phylogenies will skew the trees by exaggerating the apparent rate of viral evolution. We used this program to detect and remove hypermutated contigs from the sequence assembly. Hypermut calculates the number of APOBEC G-to-A mutations (with a downstream RD (A/G, A/G/T) relative to control G-to-A mutations (with downstream YN or RC), and calculates a Fisher exact P value that evaluates the probability of APOBEC induced mutations relative to a predetermined reference sequence. The sixty-two DNA samples contained a total of one hundred forty eight contigs that mapped to the SIV reference genome. Seventy contigs had significantly more APOBEC-induced G-to-A mutations than random mutations ($p<0.00001$), indicating that a significant proportion of the proviral genomes from those samples contained hypermutated regions so that they were reflected in the sequence assembly and were separately mapped into contigs. Hypermutated contigs were assembled in forty one of sixty two samples (66%), but did not differ by animal (69% of samples in ZF61, 64% in ZF76, and 66% in ZG13 contained hypermutated contigs) (Table 5.3). Early sequences (2 and 4 weeks) did not contain more hypermutated sequences (68%, 15/22) than 10 week and 24 week samples (65%, 26/40). The percentage of samples with hypermutated contigs varied widely by population, with 93% (13/14) of CD4 bright CXCR5++PD1++ samples containing hypermutated contigs, 71% of CD4 bright CXCR5+PD1+ samples, 75% of CD4 bright CXCR5+PD1-, and 58% (7/12) of CXCR5- samples. Fewer hypermutated contigs were detected in CD4 null samples (29% of CXCR5+ and 0% of CXCR-), however, there were fewer CD4 null samples overall that had sufficient material for amplification and analysis (7 samples and 3 samples, respectively). Hypermutated contigs were discarded from the assembly of consensus genomes for each sample, and subsequent Hypermut analysis of the resulting genomes confirmed that there was no evidence of APOBEC induced G-to-A mutations in the consensus proviral genomes.

Table 5.3 Hypermutated contigs in SIV DNA samples

ZF61	CD4 bright/dim				CD4 null	
	CXCR5++PD1++	CXCR5+PD1+	CXCR5+PD1-	CXCR5-	CXCR5+	CXCR5-
W4	■	■	□	■	□	□
W10	■	■	■	■	□	□
W24 ILN	▨	▨	□	□	□	□
W24 MLN	■	■	▨	□	▨	□

ZF76	CD4 bright/dim				CD4 null	
	CXCR5++PD1++	CXCR5+PD1+	CXCR5+PD1-	CXCR5-	CXCR5+	CXCR5-
W2	■	▨	■	▨	□	▨
W4	■	■	■	■	□	□
W10	■	■	■	▨	■	□
W24 ILN	■	■	▨	▨	▨	▨
W24 MLN	■	■	■	■	□	▨

ZG13	CD4 bright/dim				CD4 null	
	CXCR5++PD1++	CXCR5+PD1+	CXCR5+PD1-	CXCR5-	CXCR5+	CXCR5-
W2	■	■	■	■	■	□
W4	■	▨	▨	▨	▨	□
W10	■	■	■	■	▨	□
W24 ILN	■	▨	■	▨	□	□
W24 MLN	■	■	■	■	▨	□

▨	only non-hypermutated contigs
■	both hypermutated and non-hypermutated contigs
□	no contigs

5.2.4 Diversity of sequences within and between animals

HIV/SIV undergoes a complete replication cycle in approximately 48 hours, and mutated sequences arise rapidly within the viral swarm. To measure the evolutionary distance between sequences in different compartments and over time, we calculated p-distances using the proportion of nucleotide sites that differed between each pair of aligned sequences. While this method does not correct for multiple substitutions (e.g. A to T to A) or rate biases between different regions of the genome, it nonetheless offers a metric for the mean diversity within a group of sequences. A large mean distance indicates that many of the sequences in that group contain different mutations, while a small distance indicates that the sequences are similar.

There was no difference in the mean pairwise distance of plasma samples between animals, with an average distance between all plasma sequences of 0.00198 for ZF61, 0.00192 for ZF76, and 0.00180 for ZG13 (Fig. 5.2). However, the mean distance between all samples (plasma, proviral DNA and cell-associated RNA) was significantly lower in ZG13 (0.00195) than in ZF61 (0.00255, $p<0.00001$) and ZF76 (0.00248, $p<0.00001$). There was no significant difference between the mean distances of ZF61 and ZF76 total sequences (p=0.38) or DNA sequences (p=0.33) but their mean distance between RNA sequences was slightly higher in ZF61 (0.00297 vs. 0.00258, p=0.036). ZG13 had significantly lower mean distances between its DNA (0.00193 vs. 0.00235 and 0.00250, both p<0.00001) and RNA (0.00196 vs. 0.00297 and 0.00258, both p<0.00001) sequences than in ZF61 or ZF76. This indicates that the sequences from animal ZG13 were generally closer to each other (i.e. less divergent) than the sequences from animals ZF76 or ZF61. The diversity in the RNA sequences in ZF61 was higher than the DNA.

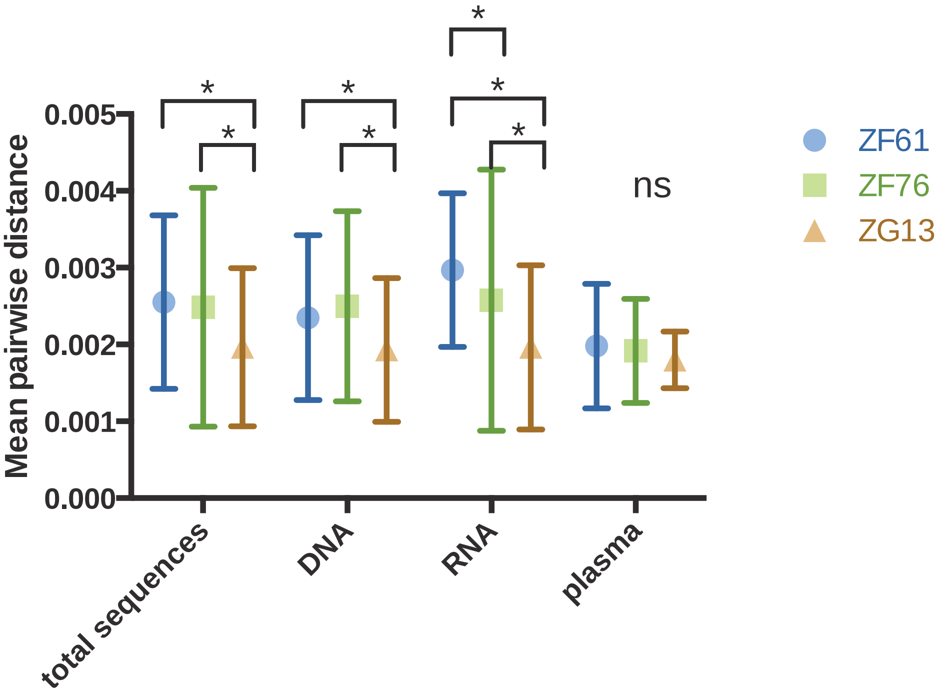

Fig. 5.2 Mean pairwise distance between sequences in individual animals. Mean pairwise distance was calculated for all sequences, proviral DNA only, cell-associated RNA only, and plasma from all time points for animals ZF61, ZF76, and ZG13. Significant differences were determined using a one way ANOVA test.

Next, we examined whether different CD4 T cell subsets contained different diversities of sequences (Fig 5.3). The mean distance between DNA sequences was not significantly different in CXCR5++PD1++, CXCR5+PD1+, and CXCR5+PD1- cells (means of 0.0033-0.0036), but was significantly lower in CXCR5- cells (0.0025, p<0.0003). However, in RNA from the same samples, CXCR5++PD1++ samples had the lowest mean distance (0.0027) and were significantly lower than the distances between CXCR5+PD1+, CXCR5+PD1-, and CXCR5- samples (p<0.01). CXCR5+PD1+ samples, with a mean distance of 0.0039, were significantly higher (p<0.006) than CXCR5+PD1- (0.0031) and CXCR5- (0.0034). The differences in the mean distance between the DNA and RNA samples varied, with CXCR5++PD1++ and CXCR5+PD1- samples having a greater genetic distance in DNA than in RNA (p<0.00001 and p<0.02), but CXCR5+PD1+ and CXCR5- samples with a greater genetic distance in RNA than in DNA (p<0.004 and p<0.0001). Thus, in DNA sequences, CXCR5- cells were less diverse than other CD4 T cell subsets. For RNA sequences, CXCR5++PD1++ sequences were most similar, while CXCR5+PD1+ sequences differed the most.

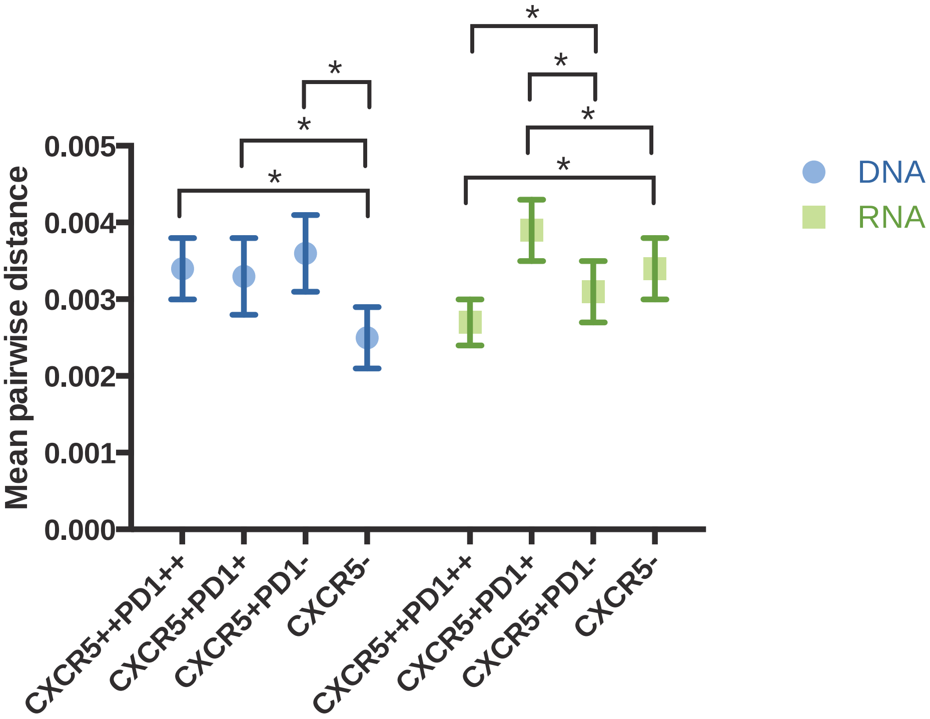

Fig. 5.3 Mean pairwise distance between T cell subpopulations. Mean pairwise distance was calculated between all DNA and RNA sequences isolated from populations of CXCR5++PD1++, CXCR5+PD1+, CXCR5+PD1- and CXCR5- CD4 T cells. Significant differences were determined using a one way ANOVA test.

Next, we examined the distances between TFH (CXCR5++PD1++ and CXCR5+PD1+) and non TFH (CXCR5+PD1- and CXCR5-) samples in early infection (2 and 4 weeks) and late infection (ten weeks and twenty-four weeks). There was no significant difference between the mean distances in DNA and RNA samples from early TFH and early non-TFH, or between late TFH and late non-TFH (Fig. 5.4). However, the mean distance between early and late samples was significant (p<0.0001) in both DNA and RNA TFH, with higher mean distance in the late samples (0.0038 in DNA and RNA) than in the early samples (0.0021 and 0.0023). There was also a significant difference (p<0.00001) between the mean distance in early vs. late samples in non-TFH, with an increase from 0.0017 and 0.0018 to 0.0038 and 0.0039 in DNA and RNA, respectively. Over the time frame in this study, later sequences show a wider range of mutations and diversity, while there is not a difference in the mean pairwise distances within TFH and non-TFH.

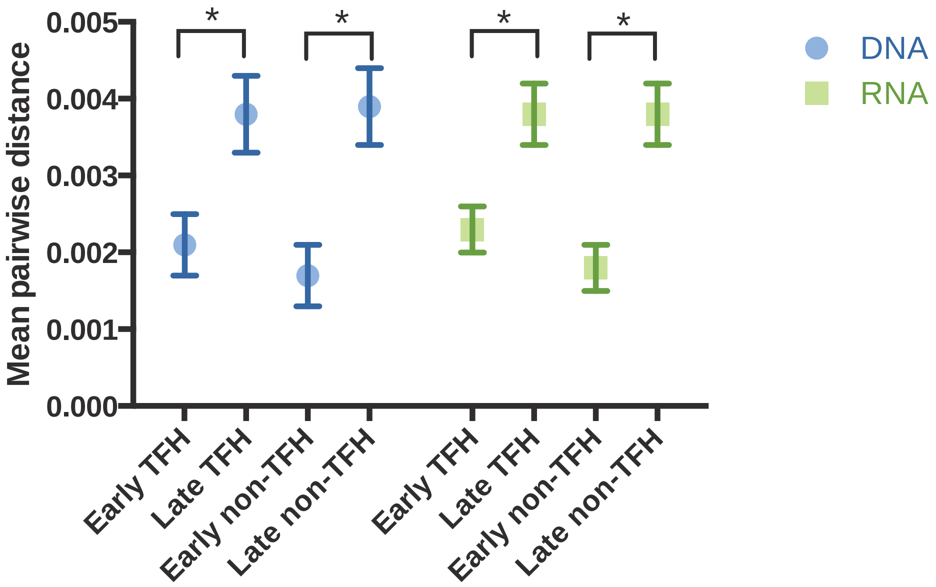

Fig. 5.4 Mean pairwise distance between early and late T cell subpopulations. Mean pairwise distance was calculated for all early (2 and 4 weeks post infection) and late (10 and 24 weeks post infection) DNA and RNA sequences from TFH (CXCR5++PD1++ and CXCR5+PD1+) and non-TFH (CXCR5+PD1- and CXCR5-) CD4 T cells. Significant differences were determined using a one way ANOVA test.

5.2.5 Phylogeny of SIV early infection

To explore the phylogenetic relationships between viruses sampled in this study, we constructed four phylogenetic trees: one with plasma sequences, one tree for each animal containing all the sequences from that animal, and one tree of sequences from all three animals. all samples from individual animals, and all viruses in the study. Sequences were aligned using MAFFT in Aliview and trees were constructed using FastTree with bootstrapping. All alignments included reference sequence KC52232, the consensus genome of the infectious stock used to infect the animals in this study[388]. All trees were rooted on the reference genome and are displayed with nodes in decreasing order.

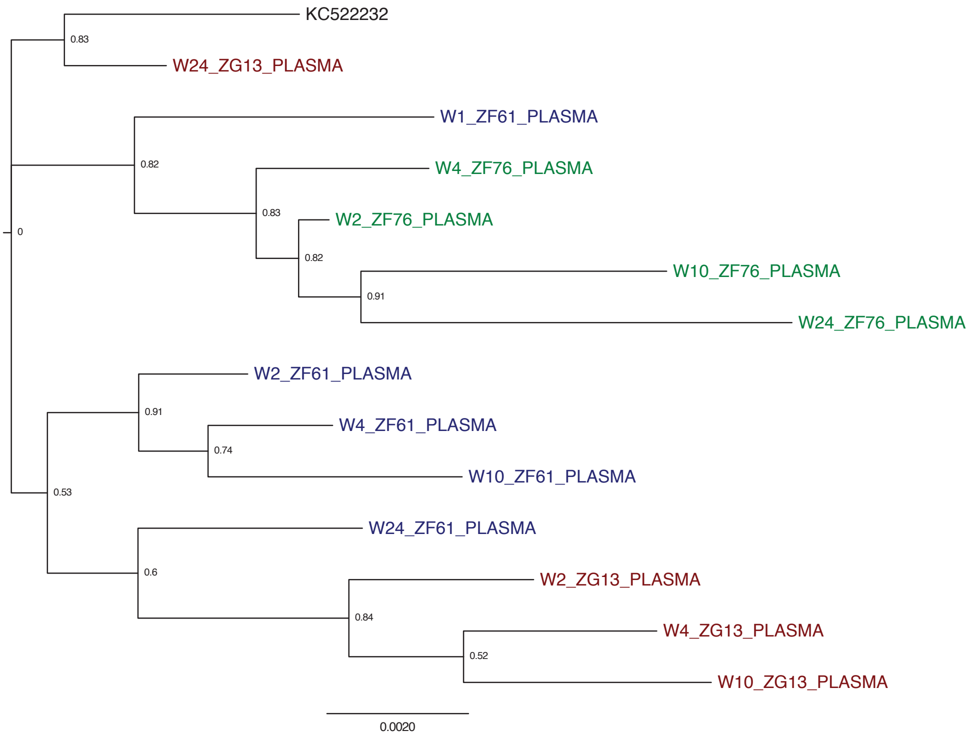

Fig. 5.5 Maximum likelihood phylogenetic tree of near-full-length genome plasma sequences. The maximum likelihood tree rooted on reference sequence KC52232. Bootstrap values on each node indicate the reliability of the split and were calculated using 1,000 resamplings.

5.2.5.1 SIV plasma phylogeny

We first constructed a phylogenetic tree of all the plasma consensus sequences, consisting of samples from infection weeks 2, 4, 10, and 24 in each of three animals (one animal, ZF61, had an additional sample from 1 week post infection) (Fig. 5.5). All four sequences from animal ZF76 (in green) clustered together, with the earlier sequences (weeks 2 and 4) close to the root and the week 10 and 24 sequences successively farther away. The four ZF76 sequences and the earliest ZF61 sequence (in blue) made up a larger cluster, distinct from the remaining ZF61 sequences and three of the four ZG13 sequences (in red). Three ZF61 sequences (from weeks 2, 4, and 10) formed a small cluster, with the week 24 sequences in a nearby branch. The week 2, 4, and 10 ZF61 sequences had increasing branch length with time of sampling. Animal ZG13 had three of four sequences (from 2, 4, and 10 weeks) in a cluster farthest away from the root, also with branch length increasing with time post infection. However, the latest ZG13 plasma sequence (week 24) was closest to the root but far from the earlier ZG13 sequences.

5.2.5.2 Interhost phylogeny

Next, we wanted to look at the relationship between all the sequences in this study. The first phylogenetic tree containing plasma, RNA, and DNA sequences from all animals is divided into three well-supported clades which largely cluster by animal (Fig. 5.6). The first, which shows the least divergence from the reference sequence, contained most of the sequences from animal ZF61 (blue) and two sequences from ZF76 (green). The second major clade is almost entirely made up of sequences from animal ZF76, with one sequence from ZG13 (red). The third clade contains all but one of the ZG13 sequences, but also several ZF61 and ZF76 sequences. The first two clades, from ZF61 and ZF76, generally show a classic model of on-going viral escape from immune pressure. The virus in each of these two animals is close to the reference infection sequence at early time points, and later (4 weeks to 24 weeks) sequences are derived from the earlier circulating viruses. Some early viruses explore alternative sequence space but were ultimately unsuccessful, and most of the samples from ZF61 and ZF76 belong to major lineages. In contrast, sequences from animal ZG13 do not form a single lineage, but instead show more extensive sampling of the sequence space and evolution of several minor lineages over time. The sequences from ZG13 make up several sub-linages that temporally co-circulate within the animal. Overall, these data suggest two models of within-host viral evolution. In the first, one lineage dominates the sequences sampled with directed evolution over time, likely due to constant immune pressure. The other shows several smaller lineages arising throughout the sampling period, but without any single lineage dominating the samples over the course of the infection. The most significant branching in the tree was due to animal, not time point, nucleic acid, or T cell population.

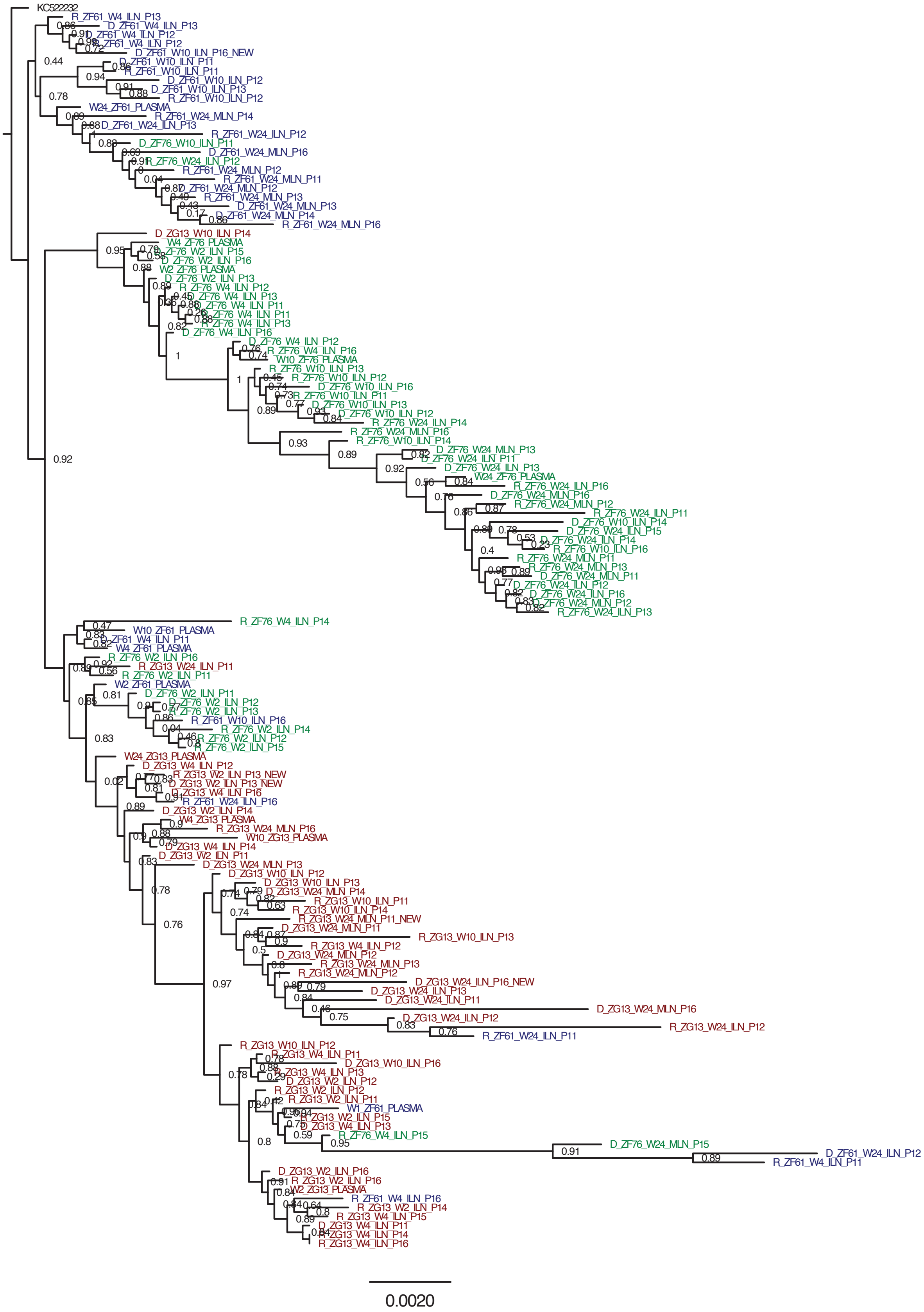

Fig. 5.6 Maximum likelihood phylogenetic tree of near-full-length genome sequences from all animals. An alignment of plasma, cell-associated RNA, and proviral DNA sequences from animal ZF76 were used to construct a maximum likelihood tree rooted on reference sequence KC52232. Bootstrap values on each node indicate the reliability of the split and were calculated using 1,000 resamplings. Sequences are named by nucleic acid (D-DNA, R-RNA), animal, time point (W2, W4, W10, W24), tissue (ILN-iliac, MLN-mesenteric lymph node), and sample population (P11- CXCR5-, P12- CXCR5++PD1++, P13 – CXCR5+PD1+, P14- CD4 null CXCR5+, P15- CD4 null CXCR5-, P16 – CXCR5+PD1-)

5.2.5.3 Intrahost phylogeny

Next, we looked at the evolution and phylogenetic relationship within each of the three animals. In ZF61, the sequences formed four distinct clusters, broadly clustering by time points (Fig. 5.7). The cluster closest to the root sequence contained most of the week 4 sequences, while the next-nearest cluster contained the early plasma sequences. The most distant cluster contained most of the week 24 sequences, although the week 1 plasma sequence was in the middle of this cluster. The week 10 sequences lay between the early and late sequences. At 24 weeks post infection, where sequences from two distinct lymph nodes (mesenteric and iliac) were obtained, the mesenteric sequences clustered together far from the root, while some of the iliac sequences clustered with week 10 sequences. The branching structure of the ZF61 tree suggests rapid exploration of the sequence space by the virus (although mostly unsampled in this study) with no single lineage predominant, until week 24 when the bottom clade becomes the most prevalent. RNA and DNA sequences were evenly dispersed throughout the tree and did not cluster together. Although we did not recover matched DNA and RNA sequences from all cell populations, in those that we did we were able to investigate how closely the RNA sequences reflected the DNA. In some cases, the DNA and RNA sequences were nearly identical branches on the same node, but in others they were separated by several branches and sequences.

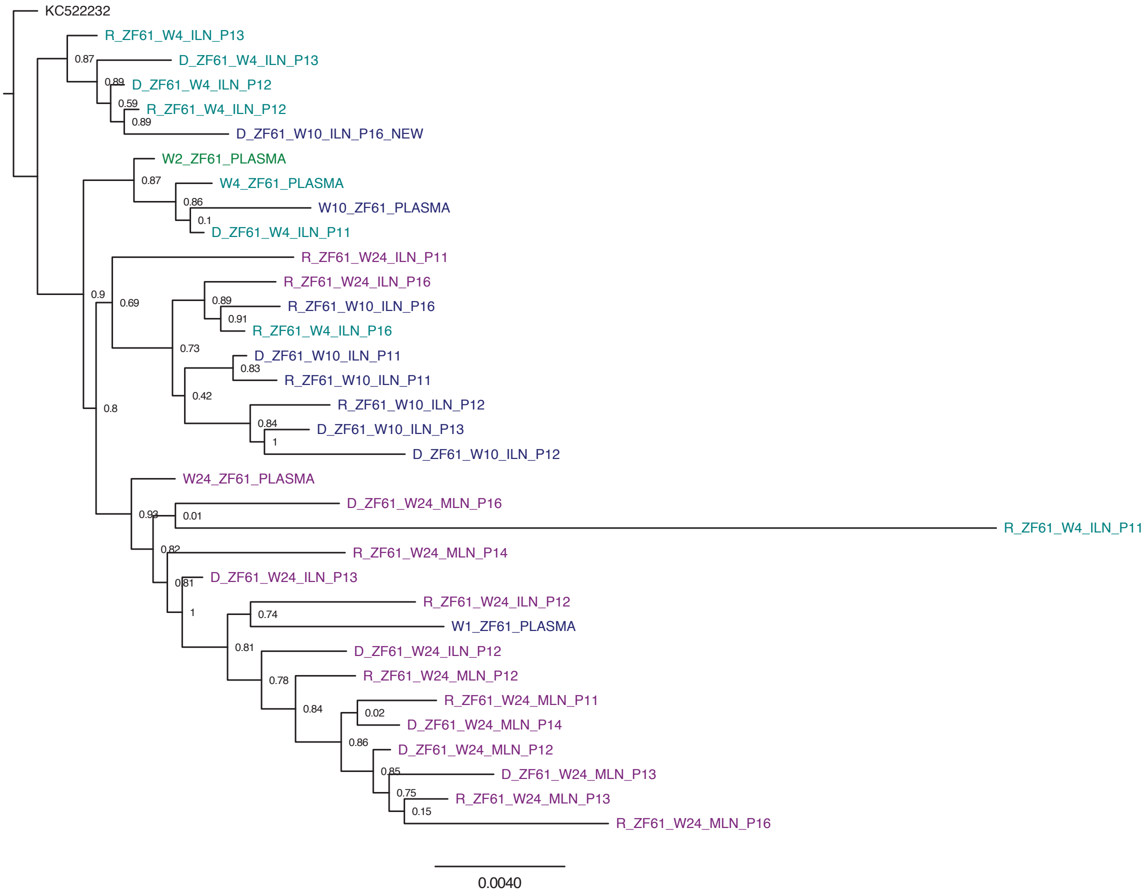

Fig. 5.7 Maximum likelihood phylogenetic tree of near-full-length genome sequences from animal ZF61. An alignment of plasma, cell-associated RNA, and proviral DNA sequences from animal ZF76 were used to construct a maximum likelihood tree rooted on reference sequence KC52232. Bootstrap values on each node indicate the reliability of the split and were calculated using 1,000 resamplings. Sequences are named by nucleic acid (D-DNA, R-RNA), animal, time point (W2, W4, W10, W24), tissue (ILN-iliac, MLN-mesenteric lymph node), and sample population (P11- CXCR5-, P12- CXCR5++PD1++, P13 – CXCR5+PD1+, P14- CD4 null CXCR5+, P15- CD4 null CXCR5-, P16 – CXCR5+PD1-)

ZF76 sequences formed two major clusters (Fig. 5.8). The first and smaller cluster contained most of the week 2 sequences and was close to the root, while the remaining week 2 sequences were part of the large branch that made up the majority of the tree. This large branch had a very strong time resolution, with the week 2 sequences closest to the root, followed by the week 4 sequences and week 10 sequences, with the week 24 sequences farthest away from the root. There were two sequences, from weeks 10 and 4, were on a separate branch close to the reference. DNA and RNA sequences were evenly distributed throughout the tree. The mesenteric and iliac sequences from week 24 were not clustered by lymph node, but were evenly intermingled. While there were small differences in the DNA and RNA sequences obtained from the same time point and cell subset, they were generally close to their counterparts on the tree, indicating that the proviral DNA did not contain a majority of substantially different genomes than the expressed RNA. Plasma sequences clustered with the DNA/RNA from the same time point. Like the ZF61 tree, the ZF76 samples showed evidence of a single lineage gradually mutating over time in response to host immune pressure, with early sequences closest to the root reference sequences, and later sequences on the most distant nodes.

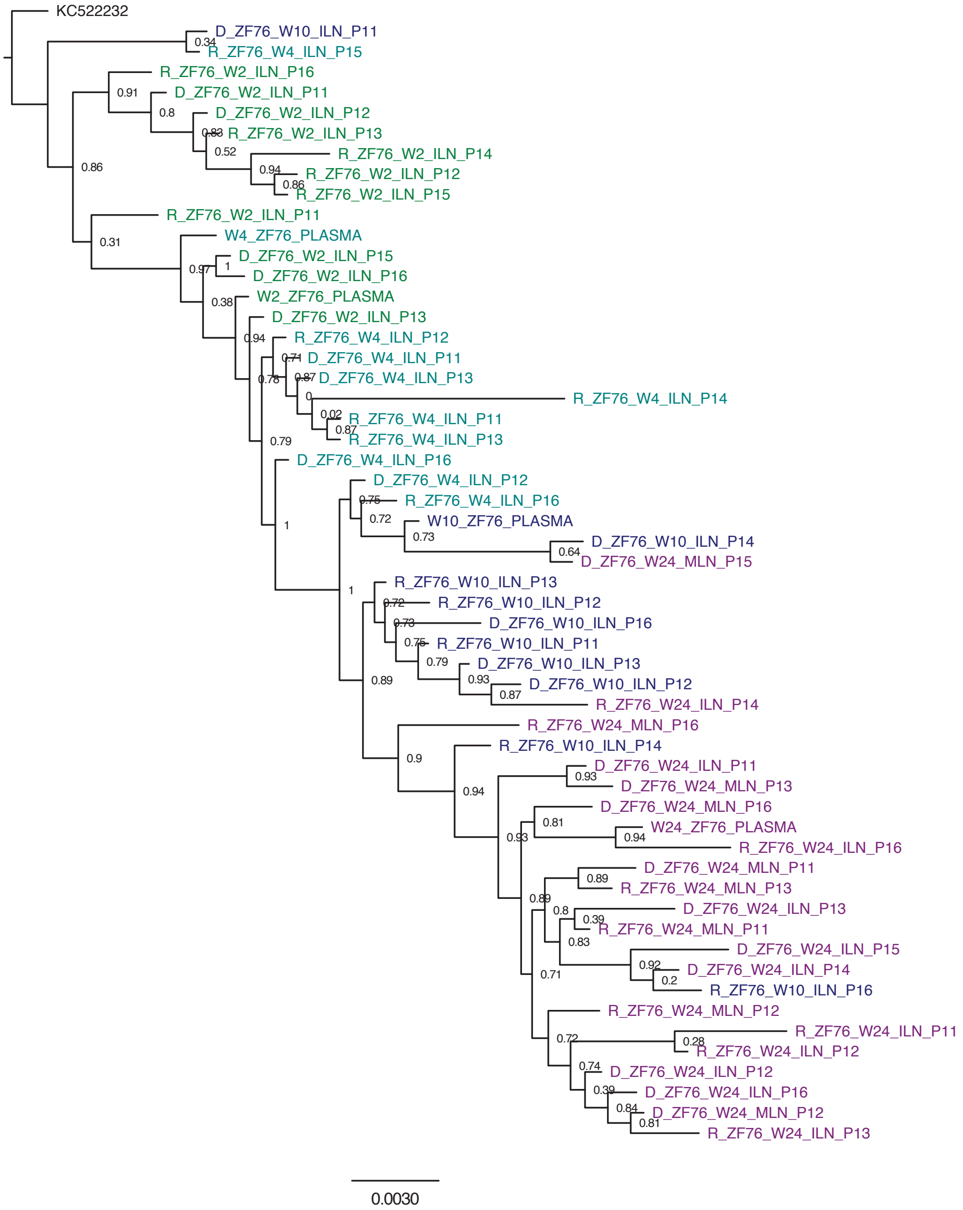

Fig. 5.8 Maximum likelihood phylogenetic tree of near-full-length genome sequences from animal ZF76. An alignment of plasma, cell-associated RNA, and proviral DNA sequences from animal ZF76 were used to construct a maximum likelihood tree rooted on reference sequence KC52232. Bootstrap values on each node indicate the reliability of the split and were calculated using 1,000 resamplings. Sequences are named by nucleic acid (D-DNA, R-RNA), animal, time point (W2, W4, W10, W24), tissue (ILN-iliac, MLN-mesenteric lymph node), and sample population (P11- CXCR5-, P12- CXCR5++PD1++, P13 – CXCR5+PD1+, P14- CD4 null CXCR5+, P15- CD4 null CXCR5-, P16 – CXCR5+PD1-)

Unlike the other two animals, ZG13 did not have a single distinct lineage with early sequences near the root and later sequences at the greatest distance (Fig. 5.9). Instead, ZG13 sequences formed several smaller clusters with sequences from multiple time points in each cluster. Most week 2 and week 4 sequences were in two clusters: a small cluster near the root, as well as a larger cluster that was farthest from the root. The week 4 plasma sequence was in a separate cluster with week 10 and week 24 sequences. A separate cluster had both DNA and RNA sequences from weeks 4, 10, and 24, from multiple T cell populations. As in the other trees, DNA and RNA sequences were distributed throughout the tree, and many of the DNA and RNA sequences from the same subset and time period were clustered closely. Overall, several distinct lineages were present in the early sampling but not in later samples, while others emerged later and were not closely related to the early samples. Some of the later sequences more closely resemble the root sequence, and may be present at low frequencies in early samplings but were not represented in this tree since we used only majority consensus sequences.

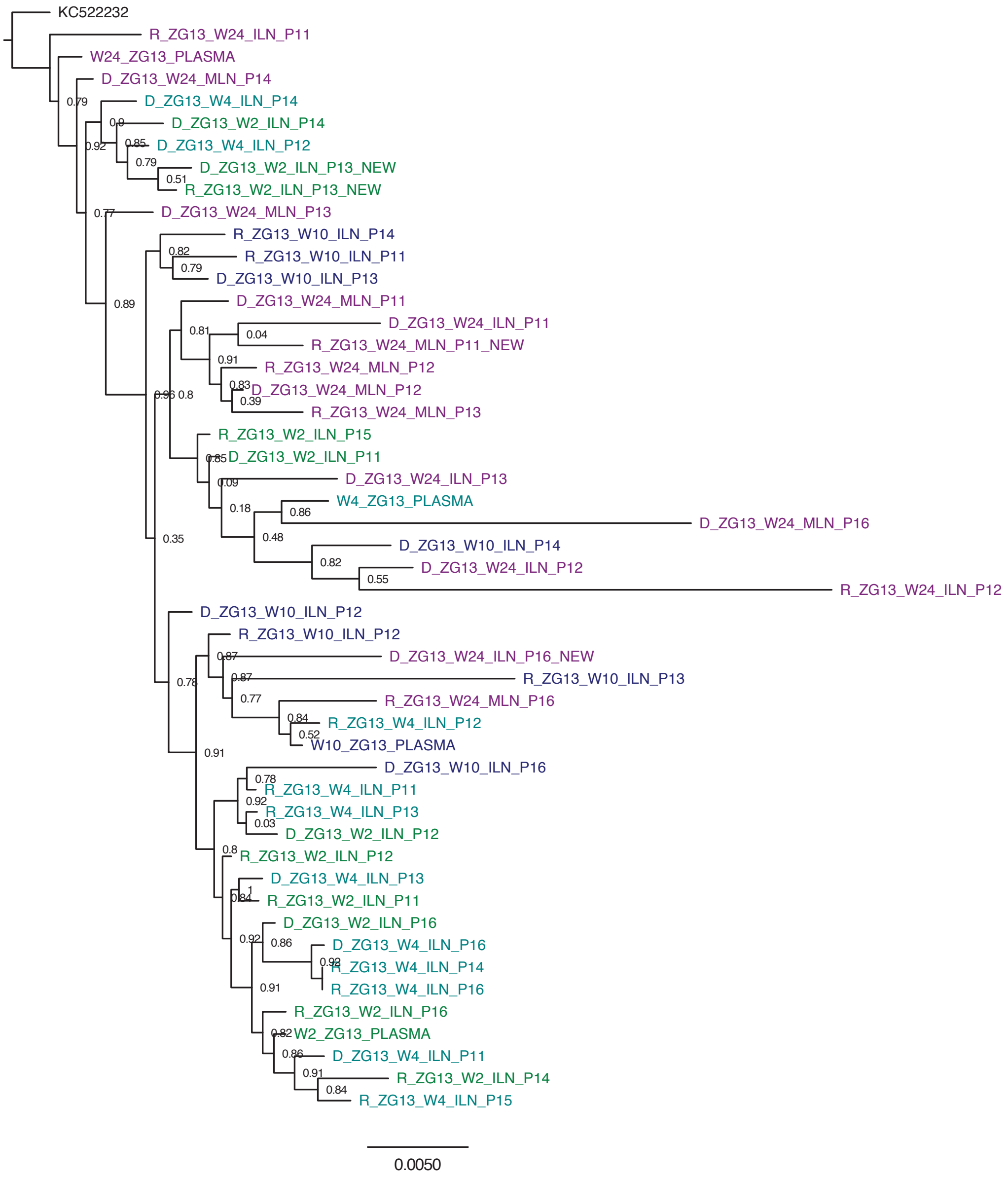

Fig. 5.9 Maximum likelihood phylogenetic tree of near-full-length genome sequences from animal ZG13. An alignment of plasma, cell-associated RNA, and proviral DNA sequences from animal ZF76 were used to construct a maximum likelihood tree rooted on reference sequence KC52232. Bootstrap values on each node indicate the reliability of the split and were calculated using 1,000 resamplings. Sequences are named by nucleic acid (D-DNA, R-RNA), animal, time point (W2, W4, W10, W24), tissue (ILN-iliac, MLN-mesenteric lymph node), and sample population (P11- CXCR5-, P12- CXCR5++PD1++, P13 – CXCR5+PD1+, P14- CD4 null CXCR5+, P15- CD4 null CXCR5-, P16 – CXCR5+PD1-)

5.3 Conclusions

Sequencing SIV genomes throughout infection offers a perspective on the evolution of the virus within a single host. As the host immune system begins to bring viremia under control, the virus adapts to the humoral and cell-mediated defences and escapes. Here, we infected three animals with the same infectious stock of SIVmac251, and tracked the evolution of virus in plasma and in lymph nodes. We adapted a protocol for sequencing near-full length genomes from HIV plasma RNA for our SIV cell-associated RNA and proviral DNA samples. The protocol for sequencing proviral DNA alongside cell-associated RNA was relatively robust, although dependent on sample input, with additional efficiencies to be gain from addition optimization of primer design. While we obtained at least partial genomes for all of the samples sent for sequencing (after screening for amplification using visualization on an agarose gel), there were variable efficiencies between the amplicons. The sensitivity of the assay was also a factor in the recovery of SIV sequences, as many of the samples had only a few thousand cells, of which only a fraction were expected to contain proviral DNA, and a subset of those containing RNA. An RNA producing cell can produce hundreds of thousands of viruses each with two copies of the RNA genome[389], whereas each infected cell contains an average of 1.5 copies of SIV DNA, which contributes to the lower recovery of DNA genomes. The presence of hypermutated sequences in the DNA samples indicates that a significant portion of the proviral genomes in lymph node CD4 T cells contained APOBEC-mutated genomes, with the highest frequencies of hypermutated sequences detected in CXCR5++PD1++ cells.

Plasma virus sequences were closely related to lymph node sequences from the same time point, but were not clustered with a specific T cell subpopulation. CXCR5- DNA samples generally had the lowest diversity (as measured by mean pairwise distance between sequences), while CXCR5+PD1+ RNA samples had the highest diversity. Both TFH and non-TFH sequences 2 and 4 weeks post infection, had lower diversity than later (10 and 24 week) samples.

A maximum likelihood phylogenetic tree with all the sequences showed that the genomes clustered largely by animal, and each animal explored a distinct sequence space. In two of the animals, nearly all the sequences in each animal were in a single lineage that grew from the founder sequence and accumulated mutations over the course of infection. In the third animal, several smaller lineages were present and there was not a clear linear evolution from early to late infection. We used consensus genomes for each sample, and do not in this study investigate minority variants. In the two animals where a single lineage appears to persist over the course of infection, we may be missing low frequency viral mutations and haplotypes. In the third animal, where we see several linages appear and disappear, we may be missing the low-level persistence of some of those lineages.