

1 **Persistent HIV-1 replication maintains the tissue reservoir**  
2 **during therapy**

3

4

5 **Ramon Lorenzo-Redondo\*, Helen R. Fryer\*, Trevor Bedford, Eun-Young Kim,**  
6 **John Archer, Sergei L. Kosakovsky Pond, Yoon-Seok Chung, Sudhir Penugonda,**  
7 **Jeffrey Chipman, Courtney V. Fletcher, Timothy W. Schacker, Michael H. Malim,**  
8 **Andrew Rambaut, Ashley T. Haase, Angela R. McLean & Steven M. Wolinsky**

9

10 \*These authors contributed equally to this work.

11

12

13 Running Title: Inferring viral evolutionary processes in well-suppressed patients using  
14 Bayesian phyloanalysis and models of viral dynamics

15

16

17 Abstract word count: 150 words

18 Text word count: 3,029 words (with section headings)

19 Figure legend word count: 411 words (with titles)

20 Methods word count: 2221 words (with section headings)

21

22

**Lymphoid tissue is a key reservoir established by HIV-1 during acute infection. It is a site of viral production, storage of viral particles in immune complexes, and viral persistence. Whilst combinations of antiretroviral drugs usually suppress viral replication and reduce viral RNA to undetectable levels in blood, it is unclear whether treatment fully suppresses viral replication in lymphoid tissue reservoirs. Here we show that virus evolution and trafficking between tissue compartments continues in patients with undetectable levels of virus in their bloodstream. A spatial dynamic model of persistent viral replication and spread explains why the development of drug resistance is not a foregone conclusion under conditions where drug concentrations are insufficient to completely block virus replication. These data provide fresh insights into the evolutionary and infection dynamics of the virus population within the host, revealing that HIV-1 can continue to replicate and refill the viral reservoir despite potent antiretroviral therapy.**

Combinations of antiretroviral drugs routinely cripple HIV-1 production and replication to levels undetectable in the blood within weeks of starting treatment<sup>1</sup>. None of the current treatments, however, are capable of eradicating the virus from a long-lived reservoir in resting memory CD4<sup>+</sup> T cells and other potential cell types that insulate the virus from antiretroviral drugs or immune surveillance<sup>2-5</sup>. Intermittent virus production from reactivation of a small proportion of latently infected CD4<sup>+</sup> T cells (rather than low levels of ongoing replication) is thought to drive viral rebound detected in blood of well-suppressed patients on treatment<sup>6-8</sup>. Ongoing replication is considered unlikely because neither viral genetic divergence over time, nor the emergence of drug resistance

46 mutations have been convincingly documented<sup>9,10</sup>. As earlier studies only examined viral  
47 sequences derived from the blood of patients who continued to suppress viral replication  
48 in that anatomic compartment<sup>11</sup>, the conclusions are not necessarily generalizable to other  
49 compartments in the body, particularly to lymphoid tissue where the frequency of  
50 infection per cell is mostly higher<sup>12</sup> and the intracellular drug concentrations are much  
51 lower than in blood<sup>13</sup>. Under low drug concentrations, the virus may continue to replicate  
52 and evolve in sanctuary sites within the reservoir of cells in lymphoid tissue, and remain  
53 undetectable in the bloodstream for a time depending on viral population migration  
54 dynamics between the two compartments. Here we use a multi-pronged strategy of deep-  
55 sequencing, time-calibrated phylogenetic analysis, and mathematical modeling to  
56 characterize the distinct temporal structure and divergence of compartmentally sampled  
57 viral sequences. We discover ongoing replication in lymphoid tissue sanctuaries of  
58 patients despite undetectable blood levels of virus. Our sampling approach differs  
59 fundamentally from those of previous studies<sup>14-16</sup>, which do not address evolutionary  
60 dynamics within lymphoid tissue, and better suits investigation of the dynamic nature of  
61 the viral reservoir during treatment with potent antiretroviral drugs.

### 63 **HIV-1 sequence determination**

64 To investigate the evolution and spatial dispersion of virus with high accuracy, we deep  
65 sequenced (using the Roche 454 Sciences' GS-FLX sequencing platform) HIV-1 DNA in  
66 cells from blood and inguinal lymph nodes collected from three subjects at three separate  
67 times (at day 0, and after 3 and 6 months of treatment) described elsewhere<sup>13</sup>. Previous  
68 work established that viral sequences contemporaneously sampled from lymphoid tissue

in different locations are genetically homogeneous<sup>17</sup>, consistent with CD4+ T cell homing and trafficking<sup>18</sup>. Consequently, detailed assessment of a portion of a solitary lymph node is no more susceptible to bias than wider anatomical sampling. We also sequenced viral RNA in the plasma (day 0) from these three study subjects. Two subjects (1727 and 1679) had well-suppressed infections (< 48 copies/mL of plasma); and the third subject (1774) continued to have measureable amounts of viral RNA in plasma after 3, but not 6, months of treatment (Extended Data Fig. 1). Subjects 1727 and 1679 were each infected with HIV-1 for approximately 3 to 4 months and were antiretroviral drug naïve before the study. Subject 1774 was infected with HIV-1 for approximately 17 years and was antiretroviral-therapy experienced, but had not received any treatment for at least 1 year prior to the study.

We aligned individual reads with an average length of 548bp to a consensus sequence using reference-guided assembly, and corrected sequencing errors for potentially inflated estimates of genetic diversity<sup>19</sup>. We then used a previously described approach<sup>20</sup> to reconstruct the minimum number of viral haplotypes needed to adequately explain the observed reads. We calculated the sequencing error rate and set the cut-off for the subsequent analyses using a known internal control sequence. We found no significant evidence for recombinant sequences that could bias the analysis. High coverage enabled us to correct PCR and sequencing errors (leaving an average 25,000 final long-reads per sample) to detect variants present in at least 0.04% of the virus population (see Methods). The sequences from the Pol region of HIV-1 that spanned the genomic region encoding

the viral enzymes protease or reverse transcriptase showed no evidence of new mutations that confer resistance to the particular antiretroviral drug used (data not shown).

To avoid uncertainties in the haplotype detection due to sparse sampling and prevent systematically biased evolutionary analyses, we established a higher number of template molecules than the depth of the sequence data by a median 9.6 fold (range 3.2 to 362 fold). The high coverage of ultra deep sequencing ensured reliable detection of low-frequency viral variants (see Methods). In support of this conclusion, we found limited variability across inferred haplotypes in two completely independent technical replicates made from the same RNA or DNA sample at each time point from each subject using the same procedures (average Spearman rank correlations coefficient between single nucleotide polymorphism frequencies across replicates of  $\rho = 0.832$ , interquartile range across samples,  $\rho = 0.820$  to  $0.851$ ; 93.7% of haplotypes above 1% frequency appeared above 1% frequency in the replicate; 97.7% of haplotypes below 1% frequency appeared below 1% frequency in the replicate), indicating that the haplotype representation is not notably biased from random amplification of some sequences and not others. The high degree of concordance between technical replicates validated our approach for the computational characterization of the viral populations, which is robust with respect to experimental error and stochastic effect<sup>20</sup>.

### **Phylogenies show temporal structure**

The inferred haplotypes corresponding to the Gag or Pol regions of HIV-1 were subjected to maximum-likelihood methods of phylogeny estimation (Extended Data Figs. 2 and 3).

We masked out the guanosines in the APOBEC3 trinucleotide contexts of the edited sites from the alignments to avoid distortion and retain the phylogenetic information<sup>20</sup> (Extended Data Figs. 2 and 3). Phylogenetic relationships between the distinct haplotypes showed a temporal structure consistent with the molecular clock (continuing nucleotide substitutions occurring at a constant rate), as evidenced by strong correlation between root-to-tip distance and sampling date in the regression analyses (despite the short branches). Branch support computed using an approximate likelihood ratio test and the proportion of sites that are different (p-distance) verify the divergence of haplotypes between day 0 and after 6 months of treatment in most of the Gag and Pol regions of HIV-1 analyzed (Extended Data Table 1). Consistent with ongoing replication rather than sampling of different virus populations in lymph nodes, viral sequences contemporaneously sampled from lymphoid tissue and blood showed a similar degree of divergence. With removal of the haplotypes found to harbor repetitive inactivating base substitutions of guanosine-to-adenosine (G-to-A) the evolutionary lineage emerged that lead up to the APOBEC3-mediated hypermutation event, and the evolutionary rate estimates (range,  $6.24 \times 10^{-4}$  to  $1.02 \times 10^{-3}$  substitutions per site per month; Extended Data Table 2) are consistent with those of intra-host virus estimations (range,  $5.22 \times 10^{-4}$  to  $8.42 \times 10^{-4}$  substitutions per site per month)<sup>21</sup>. We therefore conclude that continued virus replication contributes to the viral reservoir.

### **Genetic differentiation due to migration**

The pairwise fixation index ( $F_{ST}$ ), a standard measure of genetic differentiation between populations, confirmed significant genetic variation between lymph node and blood at

each time point (Extended Data Table 3)<sup>22,23</sup>. Because  $F_{ST}$  measures of population structure among sampled haplotypes in spatially distinct compartments can be affected by selection and migration, we used an unrestricted branch-site random effects model to test whether the proportion of sites along the branches of the phylogeny significantly differ among lineages, indicating episodic diversifying selection<sup>24</sup> (see Methods and Extended Data Table 4). We restricted the test to internal branches, which capture at least one and likely multiple rounds of virus replication, to mitigate the biasing effects of neutral or deleterious mutations on the ratio of nonsynonymous-to-synonymous substitution rates ( $\omega$ ) estimates where selection has not yet fully filtered such population level variation<sup>25,26</sup>. Except for one study subject (1679), where a small proportion of sites (0.3%) were under strong diversifying positive selection along internal tree branches (likelihood ratio test,  $P \sim 10^{-6}$ ), we found scarce evidence to suggest the virus is evolving in response to strong selective forces. We concluded that the  $F_{ST}$  values are more likely due to migration of haplotypes from one compartment to another.

## **The phyloanatomic history of HIV-1**

To infer evolutionary patterns and population dynamic processes from the time-structured sequence data, we used a Bayesian statistical framework that estimates the substitution rate, divergence time, and demographic history of the sampled viral lineages to structure-rooted, time-resolved phylogenies<sup>27,28</sup> (see Methods). This approach resolves evolutionary patterns to infer the timing and direction of the key migrations of the virus within hosts. Figure 1 shows the phyloanatomic history of HIV-1 within the study subjects inferred from the ancestral and descendent haplotypes on a Bayesian maximum

clade credibility (MCC) phylogenetic tree. The branching patterns in the trees, which reconstruct the origins and trace the flow of HIV-1 within hosts, and the tissue of origin of the internal nodes, show strong statistical support (highlighted by their posterior probabilities). Branch lengths in these time-structured trees (left panels with all haplotypes; right panels with the putative G-to-A hypermutant sequences removed to avoid distortion) represent posterior median estimates of calendar time. The temporal structure in the trees shows a strong clock-like signal and rates consistent with HIV-1 within-host evolution. The best-fit model included a strict molecular clock and assumed a constant population size, though a model with a relaxed molecular clock gave qualitatively similar results (data not shown).

Based on inferred MCC tree topologies and the compartment assignments to unobserved internal nodes, an underlying conformity and strong correlation exist between genetic and anatomic locations and the direction of the virus's spread in the body. A particular pattern recurs: the haplotypes in lymph nodes are the source of viral lineages that migrate from lymph node to blood (Fig. 1). We deduce that viral lineages in blood are derived from replicating virus in lymph nodes with little or no evidence of an additional source in blood. The time-scaled trees show a strong and significant result and are robust to different substitution models (see Methods). These data, only revealed through temporarily and spatially resolved sampling, further support the conclusion that the pattern does not result from distinct populations of haplotypes being sampled from different compartments, but rather migration and colonization of haplotypes between lymphoid tissue and blood. A structured coalescent model<sup>29</sup>, less prone to potential bias



in spatial inference estimates, shows higher migration rates from lymph node to blood (Extended Data Table 5), confirming that the direction of flow is not due to oversampling of a particular anatomic location that would have increased estimates of traffic into that location<sup>30</sup>.

Our results, which reconstruct the dynamics of HIV-1 spread within the body, imply that in patients with no detectable viral RNA in plasma, the virus reservoir is constantly replenished by low-level virus replication in lymphoid tissue. Distinguishing between low amounts of viral replication and pools of latently infected cells that may persist and rekindle HIV-1 infection is methodologically difficult. A small number of HIV-1 sequences isolated at consecutive time points that persisted without evidence of genetic change might be due to long-lived central memory cells, a fraction of which may have reverted to a resting state, or latently infected transitional memory cells that persist by clonal expansions (driven by homeostatic proliferation) or survivorship for long-lived infected CD4<sup>+</sup> T cells that contain replication-competent virus<sup>8,15,22,31-34</sup>. Two of the subjects (1774 and 1679) showed that some haplotypes persist as a single tree branch through time (Fig. 1d and f), consistent with proliferation of HIV-1 infected cells or long-term cell survival<sup>33,34</sup>. Regardless of the different mechanisms for self-renewal and/or persistence by which some of these quite similar latent or defective viral lineages may have persisted, these quiescent viruses differ from others that have evolved and trafficked between compartments. The temporally and compartmentally sampled data show that viral lineages continue to diverge in well-suppressed patients and help explain the persistence of infectious viral reservoirs with scant decay of the virus pool<sup>35</sup>. The

dynamic nature of the viral population in lymphoid tissue sanctuaries, where infected cells can still produce new viruses, infect new target cells and replenish the pool, undermines previous estimates of the time necessary to purge the reservoir of latently infected cells and achieve virus eradication<sup>3</sup>.

## **A spatial dynamic model**

Even though viral genetic diversity accumulated over time, the infected cells did not produce new virus with drug-resistance mutations, conferring a putative fitness advantage that might have led to a systemic viral rebound. To provide a mechanistic explanation of this scenario, we developed a drug concentration-dependent mathematical model of virus replication and spread between spatially distinct compartments to explore the deterministic components of viral dynamics<sup>36</sup>. In this model (see Fig. 2 and Supplementary Information), the virus can occupy two spatially distinct compartments that have limited traffic of virus particles or cells. The main compartment has a larger volume and high effective drug concentration. The other, smaller volume compartment (<0.01% of the size of the main compartment) has a lower drug concentration and represents a sanctuary site within the lymphoid tissue reservoir<sup>37</sup>. The model includes competition between two viral strains, one that is sensitive to a potent multidrug regimen and one that is partially resistant to the full complement of antiretroviral drugs, but less fit than the drug-sensitive strain in the absence of treatment. Within each compartment the balance between the strains is determined by the difference in their replicative fitness without treatment, the effectiveness of drug therapy to curb new rounds of infection in

that region (determined by drug concentration) and the susceptibility of each strain to antiretroviral therapy.

Figure 3 illustrates a hypothetical fitness landscape, which portrays the relationship between the fitness of each strain and the evolutionary adaptation across a range of drug concentrations. The fitness landscape illuminates the persistence of the pool of virus that is unaffected by antiretroviral drugs. The effective reproductive number  $R$  (the average number of secondary *de novo* infections of cells produced by one infected cell) is a function of the basic reproductive number  $R_0$  and the effectiveness of treatment ( $R = R_0$  in the absence of treatment). The fitness cost for drug resistance determines the difference between the effective reproductive numbers for drug-sensitive and drug-resistant strains ( $R_S$  and  $R_R$ , respectively) at zero effective drug concentration. In a competitive system where both strains are present, the maximum of the two effective reproductive numbers  $R_S$  and  $R_R$  equals the effective reproductive number  $R$  for the system as a whole at that effective drug concentration.

With two spatial compartments, heterogeneity in the distribution of the drug can lead to heterogeneity in  $R$ . In the sanctuary site, where drug-penetration is low, the drug-selective pressure on the replicating virus population is too low to compensate for the fitness costs associated with resistance; thus, the effective reproductive number for a drug-sensitive strain is greater than that of the partially drug-resistant strain ( $R_S > R_R$ ), enabling *de novo* infection dominated by the drug-sensitive strain<sup>38</sup>. As drug effectiveness increases, the partially drug-resistant strain gradually becomes more fit relative to the drug-sensitive

strain such that at some intermediate drug concentrations, characterized by threshold levels, the effective reproductive number for the resistant strain can be greater than that for the drug-sensitive strain ( $R_R > R_S$ ). This would allow for *de novo* infection dominated by the partially drug-resistant strain. In our model, however, we assume that drug concentration in the main compartment exceeds this threshold. At this high drug concentration, infection is no longer sustainable because the effective reproductive numbers are less than unity for both strains ( $R_R$  and  $R_S$ ).

By assuming a relatively simple two-compartment model with drug concentration differences between them, the model predicts that virus, dominated by the drug-sensitive strain, can proliferate in a small sanctuary site where the drug concentration is low (Fig. 4a). Although all single point mutations will be generated sufficiently often to prompt partially drug-resistant strains within individuals, their numbers will remain low through competition. The likelihood is exceedingly low (see Supplementary Information and Supplementary Table 2) of stepwise accumulation of mutations from a partially drug-resistant strain to one that confers resistance to all drugs, each of which would have to come to fixation in the absence of drug selection (Extended Data Fig. 4), or the presence or absence of recombination. The model calculations show that increasing drug effectiveness or penetration across spatial regions can affect evolutionary dynamics, and lead either to the emergence of drug-resistant strains (Fig. 4b) or the elimination of ongoing replication (Fig. 4c). Our model predictions fit the data well (Extended Data Fig. 5) and confirm that both competition between strains and regional spatial heterogeneity in

antiretroviral concentrations help capture the observed dynamics of the viral reservoir in these well-suppressed patients.

Though probabilistic models suggest (and this model allows) production of partially drug-resistant strains, any which arise cannot populate the sanctuary site because of low drug penetration and competition from drug-sensitive strains. Equally, they cannot repopulate the larger, main compartment where drug concentrations preclude any ongoing replication. In agreement with our phylogenetic inferences, these results suggest that the low-level viral replication in lymphoid tissues where antiretroviral drugs concentration are low could allow drug-sensitive strains to grow and spillover to the blood<sup>39</sup>.

## **Conclusion**

This study reveals how dynamic and spatial processes work together to permit HIV-1 to persist within the infected host and avoid development of resistance despite antiretroviral therapy. From these temporally and compartmentally structured sequence data, we conclude that continued virus production from infected cells in lymphoid tissue sanctuary sites, where drug concentrations are not fully suppressive, can continue to refill the viral reservoir and traffic to blood or lymphoid tissue<sup>18</sup>. We further show that the virus does not inexorably develop resistance to antiretroviral drugs because the lower concentrations of drug in the sanctuaries are not sufficient to confer a competitive advantage upon drug-resistant strains. Our findings explain the failure of treatment intensification to fully suppress *de novo* infection and highlight issues surrounding the barriers to delivering

antiretroviral drugs at clinically effective concentrations in the infectious viral reservoir. The state-of-the-art sequencing approach, innovative time-calibrated phyloanatomic tree construction, and a novel model of compartmentalized intra-host population dynamics provide a new perspective on HIV-1's seemingly untouchable strongholds in the body. Achieving optimal cellular pharmacokinetics and spatial distribution of antiretroviral drugs in lymphoid tissue to fully suppress viral replication and preserve immune function is, therefore, a prerequisite to the elimination of the viral reservoir and ultimately a cure for HIV-1 infection.

## References

- 1 Perelson, A. S. *et al.* Decay characteristics of HIV-1-infected compartments during combination therapy. *Nature* **387**, 188-191, doi:10.1038/387188a0 (1997).
- 2 Chun, T. W. *et al.* Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature* **387**, 183-188, doi:10.1038/387183a0 (1997).
- 3 Finzi, D. *et al.* Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nature medicine* **5**, 512-517, doi:10.1038/8394 (1999).
- 4 Finzi, D. *et al.* Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* **278**, 1295-1300 (1997).
- 5 Wong, J. K. *et al.* Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. *Science* **278**, 1291-1295 (1997).

318 6 Brenchley, J. M. *et al.* T-cell subsets that harbor human immunodeficiency virus  
319 (HIV) in vivo: implications for HIV pathogenesis. *Journal of virology* **78**, 1160-  
320 1168 (2004).

321 7 Chomont, N. *et al.* HIV reservoir size and persistence are driven by T cell  
322 survival and homeostatic proliferation. *Nature medicine* **15**, 893-900,  
323 doi:10.1038/nm.1972 (2009).

324 8 Zhu, T. *et al.* Evidence for human immunodeficiency virus type 1 replication in  
325 vivo in CD14(+) monocytes and its potential role as a source of virus in patients  
326 on highly active antiretroviral therapy. *Journal of virology* **76**, 707-716 (2002).

327 9 Persaud, D. *et al.* Continued production of drug-sensitive human  
328 immunodeficiency virus type 1 in children on combination antiretroviral therapy  
329 who have undetectable viral loads. *Journal of virology* **78**, 968-979 (2004).

330 10 Shen, L. & Siliciano, R. F. Viral reservoirs, residual viremia, and the potential of  
331 highly active antiretroviral therapy to eradicate HIV infection. *The Journal of*  
332 *allergy and clinical immunology* **122**, 22-28, doi:10.1016/j.jaci.2008.05.033  
333 (2008).

334 11 Persaud, D. *et al.* A stable latent reservoir for HIV-1 in resting CD4(+) T  
335 lymphocytes in infected children. *The Journal of clinical investigation* **105**, 995-  
336 1003, doi:10.1172/JCI9006 (2000).

337 12 Yukl, S. A. *et al.* The distribution of HIV DNA and RNA in cell subsets differs in  
338 gut and blood of HIV-positive patients on ART: implications for viral persistence.  
339 *J Infect Dis* **208**, 1212-1220, doi:10.1093/infdis/jit308 (2013).

340 13 Fletcher, C. V. *et al.* Persistent HIV-1 replication is associated with lower  
341 antiretroviral drug concentrations in lymphatic tissues. *Proceedings of the*

342 *National Academy of Sciences of the United States of America* **111**, 2307-2312,  
343 doi:10.1073/pnas.1318249111 (2014).

344 14 Gunthard, H. F. *et al.* Evolution of envelope sequences of human  
345 immunodeficiency virus type 1 in cellular reservoirs in the setting of potent  
346 antiviral therapy. *Journal of virology* **73**, 9404-9412 (1999).

347 15 Kearney, M. F. *et al.* Lack of detectable HIV-1 molecular evolution during  
348 suppressive antiretroviral therapy. *PLoS pathogens* **10**, e1004010,  
349 doi:10.1371/journal.ppat.1004010 (2014).

350 16 Josefsson, L. *et al.* Majority of CD4+ T cells from peripheral blood of HIV-1-  
351 infected individuals contain only one HIV DNA molecule. *Proceedings of the*  
352 *National Academy of Sciences of the United States of America* **108**, 11199-11204,  
353 doi:10.1073/pnas.1107729108 (2011).

354 17 Wong, J. K. *et al.* In vivo compartmentalization of human immunodeficiency  
355 virus: evidence from the examination of pol sequences from autopsy tissues.  
356 *Journal of virology* **71**, 2059-2071 (1997).

357 18 von Andrian, U. H. & Mempel, T. R. Homing and cellular traffic in lymph nodes.  
358 *Nature reviews. Immunology* **3**, 867-878, doi:10.1038/nri1222 (2003).

359 19 Archer, J. *et al.* Analysis of high-depth sequence data for studying viral diversity:  
360 a comparison of next generation sequencing platforms using Segminator II. *BMC*  
361 *bioinformatics* **13**, 47, doi:10.1186/1471-2105-13-47 (2012).

362 20 Kim, E. Y. *et al.* Human APOBEC3 induced mutation of human  
363 immunodeficiency virus type-1 contributes to adaptation and evolution in natural  
364 infection. *PLoS pathogens* **10**, e1004281, doi:10.1371/journal.ppat.1004281  
365 (2014).



- 366 21 Lemey, P., Rambaut, A. & Pybus, O. G. HIV evolutionary dynamics within and  
367 among hosts. *AIDS reviews* **8**, 125-140 (2006).
- 368 22 Frenkel, L. M. *et al.* Multiple viral genetic analyses detect low-level human  
369 immunodeficiency virus type 1 replication during effective highly active  
370 antiretroviral therapy. *Journal of virology* **77**, 5721-5730 (2003).
- 371 23 Nickle, D. C. *et al.* Evolutionary indicators of human immunodeficiency virus  
372 type 1 reservoirs and compartments. *Journal of virology* **77**, 5540-5546 (2003).
- 373 24 Murrell, B. *et al.* Gene-wide identification of episodic selection. *Molecular*  
374 *biology and evolution*, doi:10.1093/molbev/msv035 (2015).
- 375 25 Kryazhimskiy, S. & Plotkin, J. B. The population genetics of dN/dS. *PLoS*  
376 *genetics* **4**, e1000304, doi:10.1371/journal.pgen.1000304 (2008).
- 377 26 Mugal, C. F., Wolf, J. B. & Kaj, I. Why time matters: codon evolution and the  
378 temporal dynamics of dN/dS. *Molecular biology and evolution* **31**, 212-231,  
379 doi:10.1093/molbev/mst192 (2014).
- 380 27 Bouckaert, R. *et al.* BEAST 2: a software platform for Bayesian evolutionary  
381 analysis. *PLoS computational biology* **10**, e1003537,  
382 doi:10.1371/journal.pcbi.1003537 (2014).
- 383 28 Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian  
384 phylogeography finds its roots. *PLoS computational biology* **5**, e1000520,  
385 doi:10.1371/journal.pcbi.1000520 (2009).
- 386 29 Vaughan, T. G., Kuhnert, D., Poppinga, A., Welch, D. & Drummond, A. J.  
387 Efficient Bayesian inference under the structured coalescent. *Bioinformatics* **30**,  
388 2272-2279, doi:10.1093/bioinformatics/btu201 (2014).

389 30 Frost, S. D. W. *et al.* Eight challenges in phylodynamic inference. *Epidemics*,  
390 doi:http://dx.doi.org/10.1016/j.epidem.2014.09.001.

391 31 Tobin, N. H. *et al.* Evidence that low-level viremias during effective highly active  
392 antiretroviral therapy result from two processes: expression of archival virus and  
393 replication of virus. *Journal of virology* **79**, 9625-9634,  
394 doi:10.1128/JVI.79.15.9625-9634.2005 (2005).

395 32 Anderson, J. A. *et al.* Clonal sequences recovered from plasma from patients with  
396 residual HIV-1 viremia and on intensified antiretroviral therapy are identical to  
397 replicating viral RNAs recovered from circulating resting CD4+ T cells. *Journal*  
398 *of virology* **85**, 5220-5223, doi:10.1128/JVI.00284-11 (2011).

399 33 Maldarelli, F. *et al.* HIV latency. Specific HIV integration sites are linked to  
400 clonal expansion and persistence of infected cells. *Science* **345**, 179-183,  
401 doi:10.1126/science.1254194 (2014).

402 34 Wagner, T. A. *et al.* Proliferation of cells with HIV integrated into cancer genes  
403 contributes to persistent infection. *Science*, doi:10.1126/science.1256304 (2014).

404 35 Althaus, C. L., Joos, B., Perelson, A. S. & Gunthard, H. F. Quantifying the  
405 turnover of transcriptional subclasses of HIV-1-infected cells. *PLoS*  
406 *computational biology* **10**, e1003871, doi:10.1371/journal.pcbi.1003871 (2014).

407 36 Kepler, T. B. & Perelson, A. S. Drug concentration heterogeneity facilitates the  
408 evolution of drug resistance. *Proceedings of the National Academy of Sciences of*  
409 *the United States of America* **95**, 11514-11519 (1998).

410 37 Rong, L., Dahari, H., Ribeiro, R. M. & Perelson, A. S. Rapid emergence of  
411 protease inhibitor resistance in hepatitis C virus. *Science translational medicine* **2**,  
412 30ra32, doi:10.1126/scitranslmed.3000544 (2010).

- 413 38 McLean, A. R. & Nowak, M. A. Competition between zidovudine-sensitive and  
414 zidovudine-resistant strains of HIV. *Aids* **6**, 71-79 (1992).
- 415 39 Furtado, M. R. *et al.* Persistence of HIV-1 transcription in peripheral-blood  
416 mononuclear cells in patients receiving potent antiretroviral therapy. *The New*  
417 *England journal of medicine* **340**, 1614-1622,  
418 doi:10.1056/NEJM199905273402102 (1999).

## 419 **Acknowledgements**

420 We thank Gregory J. Beilman, Ann Thorkelson, Peter Swantek and Kristin Mars for their  
421 technical assistance. We thank Esteban Domingo and Tanmoy Bhattacharya for their  
422 constructive and informed review. We are indebted to the patients who participated in  
423 this study. This work was supported by the National Institutes of Health (DA033773 to  
424 S.M.W., AI1074340 to T.W.S., and GM110749 to S.L.K.P.), the Medical Research  
425 Council (G1000196 to M.H.M.), the Framework Programme for Research and  
426 Technological Development (278433-PREDEMICS to A.R.) and the European Research  
427 Council (260864 to A.R.). The Oxford Martin School supports H.R.F. All Souls College  
428 supports A.R.M. where S.M.W. held a Visiting Fellowship. A Newton International  
429 Fellowship from the Royal Society supported T.B. The funders had no role in study  
430 design, data collection and analysis, decision to publish, or preparation of the manuscript.  
431 The Institutional Review Board of the University of Minnesota approved the study. All  
432 subjects provided written informed consent.

433

## 434 **Author Contributions**

435 T.W.S., A.T.H., and S.M.W. conceived the experiments and designed the study. T.W.S.  
436 and J.C. acquired the patient tissue samples. A.T.H. performed the *in situ* hybridization

experiments. C.F. measured the intracellular drug concentrations. R.L.-R., E.-Y.K., and Y.-S.C. generated the viral sequences. R.L.-R., T.B., E.-Y.K., S.P., M.H.M., S.L.K.P., A.R., and S.M.W. analyzed the data. R.L.-R., T.B., J.A., and A.R. conducted the Bayesian inference analyses. H.R.F, A.R.M, and S.M.W. developed the evolutionary model. R.L.-R., T.B., H.R.F., A.T.H., S.L.K.P., A.R.M., A.R., and S.M.W. wrote the paper, with extensive input from all authors. All authors discussed the results and commented on the manuscript.

#### **Affiliations**

Division of Infectious Diseases, Northwestern University Feinberg School of Medicine, Chicago, IL 60011, USA.

Ramon Lorenzo-Redondo, Eun-Young Kim, Sudhir Penugonda, & Steven M. Wolinsky

Institute for Emerging Infections, Department of Zoology, University of Oxford, Oxford, OX1 3PS, UK.

Helen R. Fryer & Angela R. McLean

Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA.

Trevor Bedford

Centro de Investigação em Biodiversidade e Recursos Genéticos Universidade do Porto, Vairão, Portugal.

460 John Archer  
461  
462 Department of Medicine, University of California, San Diego, CA 92093, USA.  
463 Sergei L. Kosakovsky Pond  
464  
465 Center for Infectious Disease Research, Korean National Institutes of Health, Osong,  
466 Korea.  
467 Yoon-Seok Chung  
468  
469 Antiviral Pharmacology Laboratory, University of Nebraska Medical Center, College of  
470 Pharmacy, Omaha, NE 68198, USA.  
471 Courtney V. Fletcher  
472  
473 Department of Infectious Diseases, King's College London, Guy's Hospital, London, UK.  
474 Michael H. Malim  
475  
476 Centre for Immunology, Infection and Evolution, University of Edinburgh, Edinburgh,  
477 UK.  
478 Andrew Rambaut  
479  
480 Department of Surgery, University of Minnesota, Minneapolis, MN, 55455 USA.  
481 Jeffrey Chipman  
482

Division of Infectious Diseases, University of Minnesota, Minneapolis, MN 55455, USA.

Timothy W. Schacker

Department of Microbiology, University of Minnesota, Minneapolis, MN 55455, USA.

Ashley T. Haase

Nucleotide sequence alignments were deposited in GenBank with the accession numbers (KT829617 - KT831260).

Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints).

The authors declare no competing financial interests.

Correspondence and requests for materials should be addressed to Steven Wolinsky

**Figure 1. Time-structured phyloanatomic history of haplotypes in lymph nodes and blood.** MCC phylogenetic trees constructed from the complete alignments of the haplotypes from the Gag region of HIV-1 for subjects 1774, 1727 and 1679 with all haplotypes (**a**, **b** and **c**, respectively) and with the haplotypes containing G-to-A hypermutations removed (**d**, **e** and **f**, respectively). Branch colors represent the most probable (modal) anatomic location of their descendent node inferred through Bayesian reconstruction of the ancestral state, along with the posterior probabilities for the location of their parental nodes.

506

507 **Figure 2. Cartoon illustration of the drug concentration-dependent spatial model.** In  
508 the main compartment ( $i=0$ ; the majority of lymphoid tissue and the blood) drug  
509 concentration is high (grey and red). In the sanctuary site ( $i=1$ ; a small fraction of the  
510 lymphoid tissue and localized extracellular fluid) drug concentration is low (pink). There  
511 are uninfected cells ( $X_i$ ), long-lived infected cells ( $Y_i$ ), and short-lived infected cells ( $Q_i$ ),  
512 as well as virus particles ( $V_i$ ) that can be bound by few ( $F_i$ ) or many ( $G_i$ ) receptors on the  
513 follicular dendritic cell network. The dashed lines represent the effect of treatment in  
514 blocking infection and production of infectious virus particles. For graphical simplicity,  
515 we do not show the emergence of drug resistance, the production of noninfectious virus  
516 particles, virus clearance, nor cell death.

517

518 **Figure 3. Drug-dependent fitness landscape.** Effective reproductive numbers for drug-  
519 sensitive ( $R_S$ , orange line) and partially drug-resistant ( $R_R$ , blue line) strains are driven by  
520 the effective drug concentration of a single drug in the relevant compartment. Grey lines  
521 mark thresholds separating three possible outcomes. When effective drug concentrations  
522 are low, the benefit of drug resistance does not overcome the fitness cost of mutations  
523 and drug-sensitive strains dominate. This ceases to be true at intermediate effective drug  
524 concentrations and drug-resistant strains dominate. At high concentrations, both  $R_S$  and  
525  $R_R$  fall below one (red line), neither strain can grow and virus replication is halted.

526

527 **Figure 4. Modelling replication dynamics and treatment effectiveness in the viral**  
528 **reservoir.** In the main compartment, antiretroviral therapy is wholly effective against

drug-sensitive virus ( $z_0 = \tilde{z}_0 = 1$ ). In the sanctuary site, **a**, low treatment effectiveness ( $z_1 = \tilde{z}_1 = 0.3$ ) favors drug-sensitive strains. Partially drug-resistant strains will exist, but at levels below those favoring stepwise evolution towards a fully drug-resistant strain. **b**, Intermediate treatment effectiveness ( $z_1 = \tilde{z}_1 = 0.6$ ), favors partially drug-resistant virus at levels that may suffice for evolution towards a fully drug-resistant strain. **c**, high treatment effectiveness ( $z_1 = \tilde{z}_1 = 1$ ) favors the decline of all strains and the cessation of virus replication.

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

## **Methods**

### **Study subjects**

The three study subjects were enrolled into a clinical protocol where treatment was started after we obtained peripheral blood by venipuncture, inguinal lymph nodes by excisional biopsy, and ileum and rectal biopsies through colonoscopy<sup>13</sup>. Two subjects were antiretroviral drug naïve (1727 and 1679) and one subject (1774) had not received any drugs for at least 1 year before enrollment. Subjects 1679 and 1774 received emtricitabine (FTC), tenofovir (TFV; as the disoproxil fumarate [TDF]), and atazanavir with ritonavir (ATV/R). Subject 1727 received FTC, TDF and efavirenz (EFV). In all three subjects, conventional typing methods confirmed that the plasma virus was sensitive to the potent antiretroviral regimen that they received. Subjects 1727 and 1679 were well-suppressed patients. Subject 1774 continued to have measureable amounts of



HIV-1 RNA in plasma after 3, but not 6 months of treatment. We obtained peripheral blood, inguinal lymph node, and terminal ileum and rectum biopsies once again at 3 and 6 months after starting treatment. Laboratory procedures for tissue management, *in situ* hybridization, and analytical pharmacology are described elsewhere<sup>13</sup>.

#### **Extraction and quantification of viral nucleic acids**

Nucleic acids were extracted from frozen cells obtained from blood or lymphoid tissue using the MasterPure Complete DNA and RNA Purification Kit (Epicentre, Madison, WI). Viral RNA was isolated from plasma using the PureLink Viral RNA/DNA Mini Kit (Life Technologies, Carlsbad, CA). HIV-1 was quantified using a quantitative reverse transcription PCR assay. The relative amount of HIV-1 target DNA was normalized to the quantification cycle for a concentration calibrator by using an external standard curve of serial 10-fold dilutions of reference DNA for the Gag region of HIV-1 derived from the plasmid pNL-43. All reactions were performed in triplicate on the ABI 7900HT sequence detector (Applied Biosystems, Foster City, CA).

#### **Library preparation for deep sequencing**

As the number of viral templates in the sampled material is low, we used an amplicon-based deep sequencing strategy. To minimize biased amplification of the target sequence, primer locations in the Gag and Pol regions of HIV-1 were selected on the basis of the alignment positional entropy in the multiple sequences aligned from the Los Alamos National Laboratory HIV-1 sequence database (<http://www.hiv.lanl.gov/>). The primers were computationally screened for cross-dimer interactions and the concentration of each

primer was optimized for amplification. For each sample, blanks were included to screen for contamination. We used designated, physically separated areas within the laboratory to set-up PCR, which avoids contact with potentially contaminating amplicons.

To generate the long read-length PCR amplicons sequenced in this study (range, 509bp to 587bp read-lengths per run depending on the gene region being analyzed), we amplified the Gag and Pol regions of HIV-1. For the Gag region of HIV-1, we used forward primer gag\_632F\_EK (5'-GCAGTGGCGCCCGAAC-3' (corresponding to HXB2 nucleic acid sequence numbering positions 632 → 647) and reverse primer gag\_1788R\_EK (5'-AATAGTCTTACAATCTGGGTTCGC -3' (1788 → 1765). For the Pol region of HIV-1 that spanned the genomic region encoding the viral enzyme protease and reverse transcriptase, we used forward primer HIV-Pro1\_2137F(5'-CAGAGCAGACCAGAGCCAAC-3', corresponding to positions 2137 → 2156) and reverse primer HIV-RT1\_3531R (5'-CTGCTATTAAGTCTTTTGATGGGTC-3' (3531 → 3507). PCR amplification was performed using the High Fidelity Platinum Taq DNA Polymerase (Invitrogen, Carlsbad, CA) with thermal cycling conditions of 94°C for 2 mins, followed by 35 cycles of 94°C for 15s, 54°C for 15s, 68°C for 1 mins, with a final extension step at 68°C for 5 mins.

We used an integrated sequencing pipeline for library construction, template amplification, and DNA sequencing as in<sup>20</sup>. Multiplex Identifiers were included during library preparation for sample barcoding. For the Gag region of HIV-1, we used the forward primer A-Gag\_977F\_degEK 5'-primer A-

598 GCTACAACCAKCCCTYCAGACAG-3' (977 → 1000) and the reverse primer B-  
 599 Gag\_1564R\_degEK 5'-primer B-CTACTGGGATAGGTGGATTAYKTG-3' (1564 →  
 600 1541) to generate a 587bp amplicon (977 → 1564). For the Pol region of HIV-1 that  
 601 spanned the genomic region encoding the viral enzyme protease, we used forward primer  
 602 A-Pol1\_2235F 5'-primer A-ACTGTATCCTTTAGCTTCCCTCA-3' (2235 → 2262) and  
 603 the reverse primer B-Pol1\_2744R 5'-primer B-TTCTTTATGGCAAATACTGGAG-3'  
 604 (2744 → 2721) to generate a 509bp amplicon (2235 → 2744). For the Pol region of HIV-  
 605 1 that spanned the genomic region encoding the viral enzyme reverse transcriptase, we  
 606 used forward primer A-Pol2\_2700F 5'-primer A- GGCCTGAAAATCCATACAAT-3'  
 607 (2700 → 2721) and the reverse primer B-Pol2\_3265R 5'-primer B-  
 608 CATTTATCAGGATGGAGTTCATA-3' (3265 → 3242) to generate a 565bp amplicon  
 609 (2700 → 3265). PCR was performed using the High Fidelity Platinum Taq DNA  
 610 Polymerase (Invitrogen, Carlsbad, CA) with thermal cycling conditions of 94°C for 2  
 611 mins, followed by 35 cycles of 94°C for 15s, 54°C for 15s, 68°C for 1 mins, with a final  
 612 extension step at 68°C for 5 mins. DNA amplicon libraries were resolved on a pre-cast  
 613 2% agarose gel and purified with QIAquick Gel Extraction kit (Qiagen, Hilden,  
 614 Germany) and AMPure XP SPRI beads (Beckman Coulter, Indianapolis, IN).  
 615  
 616 Libraries were quantified with KAPA Library Quant Kit (Kapa Biosystems, Wilmington,  
 617 MA) on Agilent 2100 Bioanalyzer High Sensitivity DNA chip (Agilent, Santa Clara, CA)  
 618 for concentration and size distribution. The concentration of the product DNA was  
 619 normalized before pooling to achieve sequence uniformity across amplicons. Controls  
 620 were spiked into the reaction to monitor the library construction process and potential

index cross-contamination. The known internal control sequence (clonal sequence of 456 bp) introduced into the reaction was used to calculate the single nucleotide error rate and set the cut-off for the sequence analysis where no control errors could be detected. PCR errors are more common in later cycles of amplification, and being limited to small copy numbers they have little impact on the haplotype distribution. Bidirectional sequencing using the 454 Life Sciences' GS-FLX sequencing platform (Roche, Basel, Switzerland) provided independent confirmation of sequence information. The long read-length sequences were sorted based on index sequences, trimmed to remove residual adapter bases from the ends of the reads, and filtered for length and duplicates prior to alignment. Read depth and coverage estimation that met predetermined coverage thresholds were performed as in<sup>20</sup>. Sequencing quality metrics were calculated for all samples using FastQC and only high-quality sequencing libraries were used in the ensuing analyses.

#### **Sequence clean up and assembly**

Experimental precision along with deep coverage allows for accurate estimation of the underpinning diversity of the virus population. We binned the sequence reads by multiplex identifier barcodes, and identified and excluded sequencing errors and misaligned regions from the analysis by computational methods. A small number of reads had a disproportionate number of errors that accounted for most of the inaccuracy in the full dataset. After quality filtering and trimming to a uniform length, we proceeded to build the different haplotypes present in each of the samples. We began by collapsing the identical sequences into haplotypes using reference-guided assembly to avoid the use of uninformative sequence repeats. Haplotypes at prevalence above the error threshold

(defined by the internal control spiked in the sequencing runs), which corresponded to variants present above 0.04% of the total existing variants in the collapsed alignments, were used in the analysis.

For the processing of the temporally and spatially linked deep-sequencing data for studying viral diversity, the viral sequences were first aligned against the HXB2 reference sequence (GenBank accession number K03455) using Segminator II (version 0.1.1). We then generated a consensus viral sequence for each patient as a reference for assembly to improve the alignment quality. A statistical model that utilizes platform error rates in conjunction with patterns within nucleotide frequencies derived from data obtained from related samples, that is, different compartments within the host or temporally linked samples, was used to separate low frequency platform error from true variation. This model is termed “probabilistic read error detection across temporally obtained reads” (PREDATOR) and is implemented within Segminator II. This statistical framework maintains the reading frame and corrects for deep sequencing errors<sup>19</sup>. We corroborated the number of haplotypes and the frequency of haplotypes that explain the data using a second reconstruction algorithm based on combinations of multinomial distributions to analyze the k-mer frequency spectrum of the sequencing data implemented with QuRe<sup>40</sup>. We screened the sequence alignments for recombinant sequences using the GARD algorithm implemented in HyPhy<sup>41,42</sup>.

## **Sampling variance**

The high coverage of massively parallel sequencing is necessary to ensure reliable detection of low-frequency viral variants. A simple calculation based on the geometric distribution shows that in order to guarantee that a viral template occurring at frequency  $f$  is detected with probability  $p$  or better, it is necessary to sequence at least  $\log(1-p) / \log(1-f) - 1$  templates. A variant of frequency 0.01 (that is, 1%), for example, would require 450 sequences to ensure its detection at probability 0.99 (that is, 99%) or better. Conversely, a study using 100 single-genome sequences would detect a variant of frequency 0.01 with probability 0.64. A low number of input DNA templates derived from one compartment that catches the spillover from another does not account for the complexities of partial observation and spatial heterogeneity that could lead to measurement error elsewhere<sup>16</sup>. This complication emphasizes the challenge in trying to extrapolate from single template sequencing the magnitude and character of the virus population that comprises the viral reservoir.

### **Maximum-likelihood tree construction**

Maximum-likelihood phylogenies were created with PhyML using the general time-reversible model with the proportion of invariant sites and gamma distribution of among-site rate variation (GTR+I+ $\Gamma_4$ ) nucleotide substitution model<sup>43</sup> applying an approximate likelihood ratio test for branch support<sup>44</sup>. We estimated trees on viral sequence sets from which gaps in the alignment were removed and considered as missing data for the reason that maximum-likelihood tree error may increase with inclusion of unreliable sites. We assessed the temporal structure of the trees by performing linear regression on the root-to-tip distances of samples versus the time of sampling and tested the validity of the time-

dependency of the evolution rate estimates with the assumption of a strict molecular clock using the program Path-O-Gen v1.4 (<http://tree.bio.ed.ac.uk/software/pathogen/>). We used the Highlighter sequence visualization tool ([www.HIV.lanl.gov](http://www.HIV.lanl.gov)) to trace commonality between sequences in an alignment based on individual nucleotide changes.

### **Compartmentalization**

We tested for subdivision of viral sequences into sub-populations in the different compartments at each time point. We calculated genetic distances using the Wright's  $F_{ST}$  and  $S_{nn}$  test statistics<sup>45,46</sup>. We used a bootstrap test to determine the confidence of the estimates and performed a permutation test (1000 iterations) to assess the significance levels of the obtained scores.

### **Identifying selection**

We used a modification of a random effects branch-site model to detect positive selection and test whether the phylogeny diverged over time<sup>47</sup>. A likelihood ratio hypothesis test compared the fit of the model using a 3-bin  $\omega$  distribution ( $\omega_1$  and  $\omega_2$  in  $[0,1]$ ,  $\omega_3$  unrestricted) to describe the evolution of all branches in the tree, to the fit of the model where all  $\omega_3$  is restricted to be in  $[0,1]$ . We tested whether or not a proportion of sites along internal branches of the intra-host viral phylogeny have been subject to episodic selection ( $\omega > 1$ )<sup>24</sup>, restricting the test to internal branches to lessen the biasing effects of neutral or deleterious mutations on  $\omega$  estimates<sup>48,49</sup> and serve as a proxy for population level selection<sup>25,26</sup>.

## **Time-calibrated phylogenetic tree construction**

To resolve the phyloanatomy, we reconstructed the temporal and spatial dynamics of the viral haplotype lineages with a Bayesian statistical framework using Markov chain Monte Carlo (MCMC) sampling for evolutionary hypothesis testing, as implemented in BEAST version 2.1.2<sup>27</sup>. This approach was used to sample phylogenies from their joint posterior distribution, in which the viral haplotypes are restricted by their known date of sampling, using a simple substitution model described by Hasegawa–Kishino–Yano (HKY) to avoid over-parameterization<sup>50</sup>. Models differing in assumptions on mutation rate and effective population size were run for 100 million generations each and compared using the Bayes factor as implemented in Tracer version 1.6. We determined that the best-fit model included a strict molecular clock and assumed a constant population size. We used a symmetric transition model with constant rates over time that considered a discretized diffusion process among the different compartments. This was formalized as a continuous time Markov chain model to reconstruct the spatial dynamics between compartments. All chains were run for sufficient length and convergence of the relevant parameters was assessed using Tracer version 1.6, ignoring 10% of the chain as burn-in.

We summarized the connections between virus evolution and anatomical compartment history using an annotated MCC phylogenetic tree estimated with BEAST. The model and its parameters were chosen after computing the posterior probability of several models to obtain the discriminatory Bayes factors. Because population structure, whether due to spatial segregation or limitations to gene flow, may affect evolutionary dynamics, we confirmed that the direction of flow was not due to oversampling of a particular



environment, by running a two-deme Bayesian inference under a structured coalescent model with a HKY substitution model assuming a strict molecular clock<sup>29</sup>, which is less susceptible to sampling issues than our trait-based analysis. For completeness, we conducted a search for topologies and divergence times assuming a relaxed molecular clock as well. In the analyses performed, HIV-1 showed a high degree of clock-like evolution and a mean nucleotide substitution rate expected to be within the bounds necessary to obtain meaningful phylogenetic information from sequence data. Using the location of each of the haplotypes, a discrete trait was included in the inference. We used BEAST to estimate the probabilities of each of the possible states.

#### Statistical analysis

Standard descriptive statistics were performed with the use of the STATA, GraphPad or R packages.

40 Prosperi, M. C. & Salemi, M. QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. *Bioinformatics* **28**, 132-133, doi:10.1093/bioinformatics/btr627 (2012).

41 Pond, S. L., Frost, S. D. & Muse, S. V. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**, 676-679, doi:10.1093/bioinformatics/bti079 (2005).

42 Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H. & Frost, S. D. Automated phylogenetic detection of recombination using a genetic algorithm. *Molecular biology and evolution* **23**, 1891-1901, doi:10.1093/molbev/msl051 (2006).

759 43 Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate  
760 large phylogenies by maximum likelihood. *Systematic biology* **52**, 696-704  
761 (2003).

762 44 Anisimova, M. & Gascuel, O. Approximate likelihood-ratio test for branches: A  
763 fast, accurate, and powerful alternative. *Systematic biology* **55**, 539-552,  
764 doi:10.1080/10635150600755453 (2006).

765 45 Kosakovsky Pond, S. L. & Frost, S. D. Not so different after all: a comparison of  
766 methods for detecting amino acid sites under selection. *Molecular biology and*  
767 *evolution* **22**, 1208-1222, doi:10.1093/molbev/msi105 (2005).

768 46 Pond, S. L. & Frost, S. D. Datamonkey: rapid detection of selective pressure on  
769 individual sites of codon alignments. *Bioinformatics* **21**, 2531-2533,  
770 doi:10.1093/bioinformatics/bti320 (2005).

771 47 Kosakovsky Pond, S. L. *et al.* A random effects branch-site model for detecting  
772 episodic diversifying selection. *Molecular biology and evolution* **28**, 3033-3043,  
773 doi:10.1093/molbev/msr125 (2011).

774 48 Pond, S. L. *et al.* Adaptation to different human populations by HIV-1 revealed  
775 by codon-based analyses. *PLoS computational biology* **2**, e62,  
776 doi:10.1371/journal.pcbi.0020062 (2006).

777 49 Pybus, O. G. *et al.* Phylogenetic evidence for deleterious mutation load in RNA  
778 viruses and its contribution to viral evolution. *Molecular biology and evolution*  
779 **24**, 845-852, doi:10.1093/molbev/msm001 (2007).

780 50 Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a  
781 molecular clock of mitochondrial DNA. *Journal of molecular evolution* **22**, 160-  
782 174 (1985).  
783

**Extended Data Figure 1. The amounts of virus and the concentrations of drugs measured during antiretroviral therapy.** Panels **a-c** show how the number of copies HIV-1 RNA per ml of blood, the number of the HIV-1 RNA particles bound to the follicular dendritic cell network per gram of lymphoid tissue, and the number HIV-1 RNA positive cells per gram of lymphoid tissue change over the first 6 months of treatment in subjects 1774, 1727 and 1679 (**a**, **b** and **c**, respectively). Filled circles represent detectable measures. Unfilled circles represent undetectable measures and are plotted at the limit of detection. Panels **d-f** show antiretroviral concentrations in cells from lymph node (dashed line) or blood (solid line) in subjects 1774, 1727 and 1679 (**d**, **e** and **f**, respectively) (see Methods). Intracellular TFV-diphosphate concentrations (fmol/10<sup>6</sup> cells) are shown in orange, FTC-triphosphate (fmol/10<sup>6</sup> cells) in green, ATV (ng/mL) in purple, and EFV (ng/mL) in blue. Samples with concentrations that were below the limits of quantification (2.5 fmol/10<sup>6</sup> cells, 2.5 fmol/10<sup>6</sup> cells, 0.014 ng/mL and 0.063 ng/mL, respectively) were assigned a value of 1 for graphical illustration purposes.

**Extended Data Figure 2. Phylogenies and Highlighter plots for the Gag region of HIV-1.** Maximum-likelihood trees were constructed using gene sequences from the Gag region of HIV-1 from lymph node and blood before and after the guanosines within all possible APOBEC3 trinucleotide sequence context of edited sites were masked in the alignments, regardless of their presence in hypermutant or non-hypermutant sequences, to avoid their distortion in the phylogenetic reconstructions. Branch tips are colored according to compartment sampled: red for plasma; gold for lymph node; and blue for

blood. The progressive shading of the colors of the branch tips indicate the points in time sampled. Phylogenetic trees reconstructed from the haplotypes in which the guanosines in the APOBEC3 trinucleotide context of the edited sites are masked in the alignments correct the skewing effect caused by clustering of shared haplotypes that harbor repetitive G-to-A substitutions and longer branch lengths caused by a larger number of these mutations in the hypermutated sequences whilst retaining the phylogenetic information. The horizontal scale indicates the expected number of substitutions per nucleotide site per unit time with haplotypes from later time points having diverged more. The Highlighter plots show the haplotypes from the lymphoid tissue and blood time point clusters aligned to the plasma virus sequence from day 0. The particular nucleotide changes are color coded in the alignment (thymidine, red; adenosine, green; cytosine, blue; and guanosine, orange). Magenta circles represent APOBEC3-induced G-to-A change in a trinucleotide context of the edited sites, which are distinguishable from the more random error-prone viral reverse transcriptase and RNA polymerase II replicating enzyme induced mutations<sup>6</sup>. Gene sequences from the Gag region of HIV-1 from Subject 1774, who continued to have measureable amounts of HIV-1 RNA in plasma on treatment, and Subjects 1727 and 1679 who were well-suppressed on treatment (**a**, **b** and **c**, respectively) before and after the guanosines within the particular APOBEC3 trinucleotide sequence context of edited sites were masked in the entire sequence alignment (left and right panels, respectively).

**Extended Data Figure 3. Phylogenies and Highlighter plots for the Pol region of HIV-1.** Maximum-likelihood trees were constructed using gene sequences from the Pol

region (retrotranscriptase [pol2]) of HIV-1 from lymph node and blood before and after the guanosines within all possible APOBEC3 trinucleotide sequence context of edited sites were masked in the alignments, regardless of their presence in hypermutant or non-hypermutant sequences, to avoid their distortion in the phylogenetic reconstructions. Branch tips are colored according to compartment sampled: red for plasma; gold for lymph node; and blue for blood. The progressive shading of the colors of the branch tips indicate the points in time sampled. The horizontal scale indicates the expected number of substitutions per nucleotide site per unit time with haplotypes from later time points having diverged more. The Highlighter plots show the haplotypes from the lymphoid tissue and blood time point clusters aligned to the plasma virus sequence from day 0. The particular nucleotide changes are color coded in the alignment (thymidine, red; adenosine, green; cytosine, blue; and guanosine, orange). Magenta circles represent APOBEC3-induced G-to-A change in a trinucleotide context of the edited sites. Gene sequences from the Pol region of HIV-1 that spanned the genomic region encoding the viral enzyme reverse transcriptase from Subjects 1774, 1727, and 1679 (**a**, **b** and **c**, respectively) before and after the guanosines within the particular APOBEC3 trinucleotide sequence context of edited sites were masked in the entire sequence alignment (left and right panels, respectively).

**Extended Data Figure 4. Alternative drug-dependent fitness landscape plots. a,** Fitness landscape plot for a partially drug-resistant strain that confers relatively low-level resistance to drugs as compared with the fitness costs imposed by the drug-resistant mutations. The drug-resistant strain (blue line) does not outcompete the drug-sensitive

strain (orange line) at any effective treatment concentration where it can grow. There are two phases to the dynamics. At lower effective drug concentrations (left of grey line) the drug-sensitive strain thrives. Beyond this threshold, neither strain can continuously replicate. **b**, Fitness landscape plot for a highly drug-resistant strain. This strain confers a high-level of drug resistance relative to the replicative fitness cost imposed by the resistance mutations. At low effective drug concentrations (left of grey line), the drug-sensitive strain outcompetes the drug-resistant strain. At high effective drug concentrations, the drug-resistant strain outcompetes the drug-sensitive strain and can continuously replicate. We argue that, typically, fully drug resistant mutants of this sort neither exist in the viral population of patients before treatment, nor arise through random mutation during the course of antiretroviral therapy (see Supplementary Information and Supplementary Table 2). Drug-resistant strains, which are capable of ongoing replication at high effective drug concentrations are not typically generated in individuals because: they are generated in a single step very rarely; and stepwise generation from partially resistant strains is also rare because partially resistant strains are outcompeted in the sanctuary site which constantly refills the pool. The strain specific effective reproductive numbers for the drug-sensitive  $R_S$  (orange) and drug-resistant  $R_R$  (blue) strains are shown. For simplicity, only the impact of changes to the effectiveness of a single drug in a single compartment is shown.

**Extended Data Figure 5. Model of replication dynamics and treatment effectiveness in the viral reservoir fitted to the data.** The model is fitted to the total inferred average body counts of free virus particles (green line), infected CD4<sup>+</sup> T cells (orange line) and

virus bound to the follicular dendritic cell network of B cell follicles (grey line). **a**, Demonstrates the dynamics over the first 200 days of treatment. Note that early on during antiretroviral therapy, HIV-1 RNA in plasma declines more rapidly than virus bound to the follicular dendritic cell network of B cell follicles. Circles demonstrate average data from the 3 patients discussed in detail in this study and an additional 9 patients presented elsewhere<sup>13</sup>. Where the average value was indeterminate because of test sensitivity, the data are fitted below the upper limit of the average log<sub>10</sub> infectious units. **b**, Demonstrates the dynamics over a longer period. The model predicts the persistent low-level viral RNA in plasma. The diamond symbol represents data relating to the long-term persistent virus, as measured using quantitative reverse transcription PCR (see Methods). The optimal model fit parameters are presented in Supplementary Table 1.

**Extended Data Table 1. The genetic distance measured between the haplotypes in the Gag or Pol regions of HIV-1.** We used sequence data from the Gag or Pol regions (protease [pol1] and retrotranscriptase [pol2]) regions of HIV-1 (range, 315bp to 487bp) to perform a genetic distance analysis. We masked the guanosines in APOBEC3 trinucleotide contexts of the edited sites from the alignments. We compared the proportion of substitutions per site (p-distance) for haplotypes within lymph node or blood from each subject at 6 months to the most common haplotype present at day 0. Most comparisons were statistically significant (Mann-Whitney U test, *P*-values < 0.1), indicating ongoing diversification of the viral populations.

**Extended Data Table 2. HIV-1 evolutionary rate calculated after removing hypermutated haplotypes.** We used a linear regression model to estimate the evolutionary rate for the Gag or Pol regions (protease [pol1] and retrotranscriptase [pol2]) of HIV-1. We calculated the slope ( $\mu$ ) of the linear regression between time and the direct pairwise genetic distances (number of substitutions per site per month) from the most common haplotype at day 0 for each subject. Haplotypes found to harbor G-to-A hypermutant sequences were removed from the analysis to limit the effect of inactivating mutations on the estimates. There is very strong evidence, now presented in numerous studies<sup>17</sup>, that acute HIV-1 infections are by and large founded by a single (or at best a few) viral strains; this initial bottleneck is followed by rapid population diversification in the absence of treatment. The structures of our intra-host phylogenies at day 0 support this pattern. Even if multiple populations are transmitted and are able to establish infection and co-circulate, Bayesian phylodynamics models can properly account for it, and estimate evolutionary rates accurately.

**Extended Data Table 3. Patterns of genetic divergence measured between viral haplotypes in lymphoid tissue and blood.**  $S_{nn}$  and  $F_{ST}$  pairwise genetic differentiation (**a** and **b**, respectively) between peripheral blood and lymph samples for the Gag and Pol regions of HIV-1 from subjects 1774, 1727, and 1679 at day 0 and after 6 months of antiretroviral therapy.

**Extended Data Table 4. Episodic selection estimated across sites along internal branches of the phylogeny.** A branch-site evolutionary model provided a statistical



framework to search for evidence of episodic selection on internal branches in the tree.

The  $\omega$  distribution is given for both internal and external branches.

**Extended Data Table 5. Estimated migration rates between lymph nodes and blood**

**using Bayesian inference under the structured coalescence.** Migration rates (in

fraction of emigrants per month) were estimated in the Gag region of HIV-1 for the three

study subjects using Bayesian inference under the structured coalescence model,

assuming a constant population size and a strict molecular clock. Migration rates from

lymph node to blood and vice versa are shown with their standard error of the mean

(SEM) after running at least 50 million MCMC steps and reaching high values of

estimated sample sizes (ESS) for all parameters.







