LETTER

doi:10.1038/nature19827

1970s and 'Patient 0' HIV-1 genomes illuminate early HIV/AIDS history in North America

Michael Worobey[1], Thomas D. Watts[1], Richard A. McKay[2], Marc A. Suchard[3], Timothy Granade[4], Dirk E. Teuwen[5], Beryl A. Koblin[6], Walid Heneine[4], Philippe Lemey[7] & Harold W. Jaffe[4]

The emergence of HIV-1 group M subtype B in North American men who have sex with men was a key turning point in the HIV/AIDS pandemic. Phylogenetic studies have suggested cryptic subtype B circulation in the United States (US) throughout the 1970s[1,2] and an even older presence in the Caribbean[2]. However, these temporal and geographical inferences, based upon partial HIV-1 genomes that postdate the recognition of AIDS in 1981, remain contentious[3,4] and the earliest movements of the virus within the US are unknown. We serologically screened >2,000 1970s serum samples and developed a highly sensitive approach for recovering viral RNA from degraded archival samples. Here, we report eight coding-complete genomes from US serum samples from 1978–1979—eight of the nine oldest HIV-1 group M genomes to date. This early, full-genome 'snapshot' reveals that the US HIV-1 epidemic exhibited extensive genetic diversity in the 1970s but also provides strong evidence for its emergence from a pre-existing Caribbean epidemic. Bayesian phylogenetic analyses estimate the jump to the US at around 1970 and place the ancestral US virus in New York City with 0.99 posterior probability support, strongly suggesting this was the crucial hub of early US HIV/AIDS diversification. Logistic growth coalescent models reveal epidemic doubling times of 0.86 and 1.12 years for the US and Caribbean, respectively, suggesting rapid early expansion in each location[3]. Comparisons with more recent data reveal many of these insights to be unattainable without archival, full-genome sequences. We also recovered the HIV-1 genome from the individual known as 'Patient 0' (ref. 5) and found neither biological nor historical evidence that he was the primary case in the US or for subtype B as a whole. We discuss the genesis and persistence of this belief in the light of these evolutionary insights.

No comprehensive genomic analysis of the emergence and early spread of HIV-1 in North America—where HIV/AIDS was first recognized—has been possible because the only pre-1980 HIV-1 group M genome currently available (strain Z321B) was sampled in Africa. To fill this gap, we performed serological screening and viral genome sequencing of archived serum samples dating back to 1978–1979, originally collected from men who have sex with men (MSM) cohort patients in New York City (NYC) and San Francisco (SF). NYC samples were from volunteers in a prospective study of AIDS established in 1984 (ref. 6), 378 of whom had been part of an earlier cohort of 8,906 men involved in hepatitis B virus (HBV) studies beginning in 1978 (ref. 7), and for which stored sera from 1978 and/or 1979 were available[8]. Previous work showed that 6.6% of these sera from NYC in 1978–1979 were HIV-1 seropositive[6]; 33 of these positive samples were chosen for attempted HIV-1 sequencing. The sera from SF originated from a study of approximately 6,875 patients enrolled in the late 1970s in HBV studies at the San Francisco City Clinic[9]. We tested 2,231 of these samples from 1978 and found, by western blot, that 83 (3.7%) were positive for HIV-1 antibodies; of these, 20 were randomly chosen for attempted HIV-1 sequencing.

Low template number and degradation arising from long-term storage were major challenges for genomic analysis, as encountered previously with similar samples[10]: recovered RNA was generally below the limits of quantification and initial attempts at amplification of reverse-transcribed viral RNA failed consistently and indicated that viral RNA survived in the 1970s samples only in short fragments. This led us to design an RNA 'jackhammering' approach to greatly increase both the ability to detect viral RNA-positive samples and to recover complete genomic HIV-1 sequences from them. Briefly, we used large panels of primers to amplify many short fragments in separate pools, such that amplicons overlapped between but not within each pool (Extended Data Fig. 1 and Supplementary Table 1). Each pool's amplicon set filled gaps between those of complementary pools, with the entire panel providing complete genomic coverage. Moreover, a preliminary, multiplex amplification step greatly concentrated target RNA before final amplification and sequencing.

Three samples from SF and five from NYC provided sufficient data to assemble coding-complete sequences. Bayesian phylogenetic analyses of these HIV-1 genomes (Fig. 1 and Extended Data Fig. 2) showed that although they were the oldest sampled outside Africa, they do not fall on the deepest branches even within subtype B. Instead, the 1970s genomes and the US epidemic as a whole were phylogenetically nested within the more genetically diverse, older subtype B epidemic in Caribbean countries. Separate analyses of *gag*, *pol* and *env* sequences also placed the US sequences in a strongly supported monophyletic clade nested within the paraphyletic Caribbean subtype B sequences from Haiti, Dominican Republic, Jamaica, Trinidad and Tobago and Haitian immigrants in the US (Extended Data Figs 3, 4). Molecular clock phylogeographic analysis of the complete genome data supported a subtype B ancestor in the Caribbean (posterior probability >0.99) dating to 1967 (95% credibility interval 1963–1970) (Extended Data Table 1). This provided genome-wide evidence that the epidemic moved from the Caribbean to the US rather than from the US to the Caribbean[2].

Location transition estimates recovered a relatively precise date (1971 (1969–1973), Extended Data Table 1) for the HIV-1 jump from the Caribbean, very shortly before the US most recent common ancestor (MRCA). This narrow timing is aided by the basal relationship of a very close relative from the Caribbean (sequence 'H6' from an individual who entered the US from Haiti in 1981)[2] (Extended Data Fig. 2). The probability density of the date of introduction to the US overlaps with the deep branching structure in Caribbean diversity (Fig. 1 and Extended Data Fig. 3), indicating that the US clade emerged from the Caribbean epidemic during its early growth phase. We estimated a relatively fast logistic growth rate of 0.62 (0.26–0.99) yr^{-1} within the Caribbean population (Fig. 2). That of the US population was even

[1]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA. [2]Department of History and Philosophy of Science, University of Cambridge, Cambridge CB2 3RH, UK. [3]Departments of Biomathematics, Biostatistics and Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, California 90095, USA. [4]Centers for Disease Control and Prevention, Atlanta, Georgia 30333, USA. [5]UCB, Brussels BE-1070, Belgium. [6]Laboratory of Infectious Disease Prevention, The New York Blood Center, New York, New York 10065, USA. [7]Department of Microbiology and Immunology, Rega Institute, KU Leuven—University of Leuven, Minderbroedersstaat 10, 3000 Leuven, Belgium.

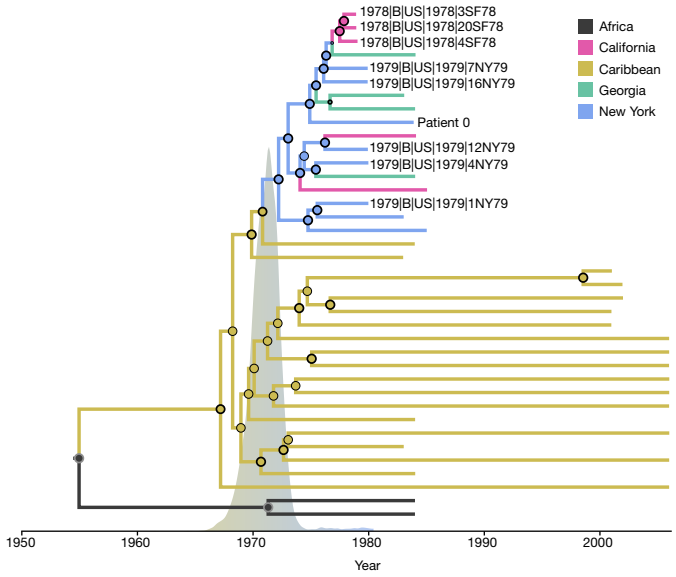


Figure 1 | Maximum clade credibility (MCC) tree summary of the Bayesian spatio-temporal reconstruction based on complete HIV-1 genome data. The tips of the tree correspond to year of sampling while branch and node colours reflect the sampling locations for the tip branches and the inferred locations for the internal branches. Tip labels are provided for the newly obtained archival HIV-1 genomes. Diameters of internal node circles reflect posterior location probability values. Thick outer circles represent internal nodes with posterior probability support >0.95. We also depict the posterior probability density (grey) for the time of the introduction event from the Caribbean into the US on the time scale of the tree. A fully annotated tree for this data set ('full genome 38', which includes only sequences sampled early in the US epidemic) is shown in Extended Data Fig. 2b; 'full genome 46' which includes all available complete genomes basal to the 'pandemic clade'[2] of subtype B, plus a similar number and date range of US pandemic clade sequences, is shown in Extended Data Fig. 2a. Separate analyses of *gag*, *pol*, *env*, and the coding-complete genomes (including also sequences sampled later in the US epidemic) provide consistent results (Extended Data Figs 3, 4).

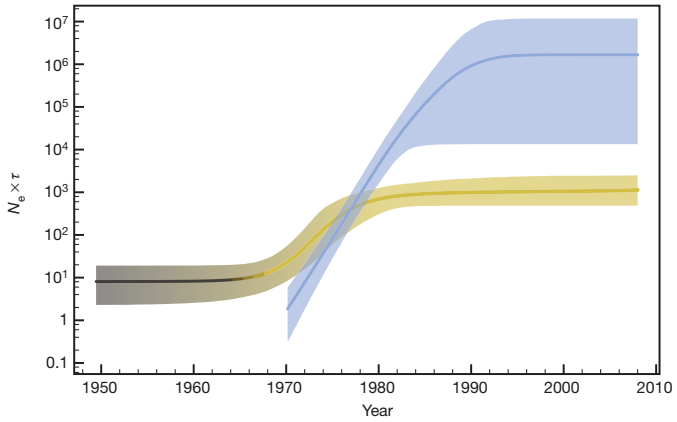


Figure 2 | Demographic reconstruction based on the nested coalescent model. The colour scheme is consistent with that of the phylogeographic analyses in Figs 1 and 3: the constant-logistic population size estimates (the 'effective number of infections', N_e, multiplied by the mean viral generation time, τ) through time are depicted in a black–yellow colour range (following the African and Caribbean locations in the phylogeographic analyses) while the logistic population size estimates for the nested US clade are shown in blue (as for the US/NYC location in the phylogeographic analyses).

higher, 0.81 (0.65–0.98) yr^{-1}, in line with a precipitous spread among existing high-risk sexual networks. These mean growth rate estimates corresponded to doubling times of 1.12 years and 0.86 years for the Caribbean and the US, respectively; both the more rapid and longer growth in the US appear to have contributed a higher number of 'effective infections' (Fig. 2), with the US overtaking the Caribbean by around 1977 despite the later HIV-1 emergence in the US.

Molecular clock analyses of larger numbers of *env* sequences revealed similar time of the most recent common ancestor (TMRCA) estimates for the key nodes (Fig. 3, Extended Data Table 1 and Extended Data Figs 5, 6). Interestingly, our modest snapshot of 1970s sequences from NYC and SF (Fig. 3, Extended Data Fig. 5b) encompassed the full diversity exhibited by HIV-1 sequences from later years (that is, it shares the same MRCA as larger sequence sets sampled in later years): all post-1985 US sequences are nested within the early diversity captured by the limited number of 1970s sequences we recovered (Extended Data Fig. 6).

A phylogeographic reconstruction including only those US sequences sampled from known locations between 1978–1984 (Fig. 1) demonstrated that the NYC epidemic was already relatively mature and genetically diverse by 1979, tracing back to an MRCA estimated at 1972 (1970–1974) and there is strong support for the idea that the US subtype B ancestor circulated in NYC (posterior probability = 0.99). Indeed, the extensive genetic diversity in the US (and in NYC in particular) in 1978–1979 can be explained only by several years of circulation of the virus before 1978–1979.

Using sequences sampled from NYC, North Carolina and California relatively late in the epidemic (comparable to the 1978–1984 East coast, West coast and Southern sampling), we still inferred a US ancestor in NYC, but with only modest support that prevents us from drawing firm conclusions (pp = 0.67, Extended Data Figure 6b and Extended Data Table 1). As a generality, early samples close to the deep branching structure are essential to confidently reconstruct the initial spatio-temporal expansion dynamics in exponentially growing populations.

Compared to NYC, the SF epidemic in 1978 appeared to have been established more recently (Figs 1, 3, and Extended Data Figs 2b, 5b). It is striking that all three independently detected complete HIV-1 genomes we found are so closely related; moreover, they form a cluster with three partial *env* sequences sampled in SF during the same period[10] (Extended Data Fig. 5b). This suggests that the bulk of the HIV-1 infections in SF in 1978 traced back to a single introduction from NYC in around 1976 (consistent with the lower HIV-1 seroprevalence in the SF cohort).

The sampled sequences thus reveal a series of key founder events in the genesis of subtype B (for example, Fig. 3 and Extended Data Table 1), with the epidemic spreading from the African HIV-1 group M epicentre to the Caribbean by about 1967, from the Caribbean to NYC by about 1971 and from NYC to SF by about 1976, quickly followed by extensive geographical mixing in the US and beyond.

Reports of one cluster of homosexual men with AIDS linked through sexual contact were important in suggesting the sexual transmission route of an infectious agent before the identification of HIV-1 (refs 5, 11). Beginning in California, CDC investigators eventually connected 40 men in ten American cities to this sexual network. Investigators placed one man with Kaposi's sarcoma near the centre of a sociogram representing this cluster and identified him as 'Patient 0'—a 'non-Californian AIDS patient' and a possible 'carrier' of an infectious agent (Extended Data Fig. 7). Before publication, Patient 'O' was the abbreviation used to indicate that this patient with Kaposi's sarcoma resided 'Out(side)-of-California.' As investigators numbered the cluster cases by date of symptom onset, the letter 'O' was misinterpreted as the number '0,' and the non-Californian AIDS patient entered the literature with that title[12]. Although the authors of the cluster study repeatedly maintained that Patient 0 was probably not the 'source' of AIDS for the cluster or the wider US epidemic, many people have subsequently employed the term 'patient zero' to denote an original or primary case, and many still believe the story today[13]. We therefore recovered the complete HIV-1 genome of Patient 0 and examined it against the backdrop provided by the 1970s sequences.

Although labelled as the cluster study's 'index patient', Patient 0 was neither the first AIDS case to come to CDC researchers'

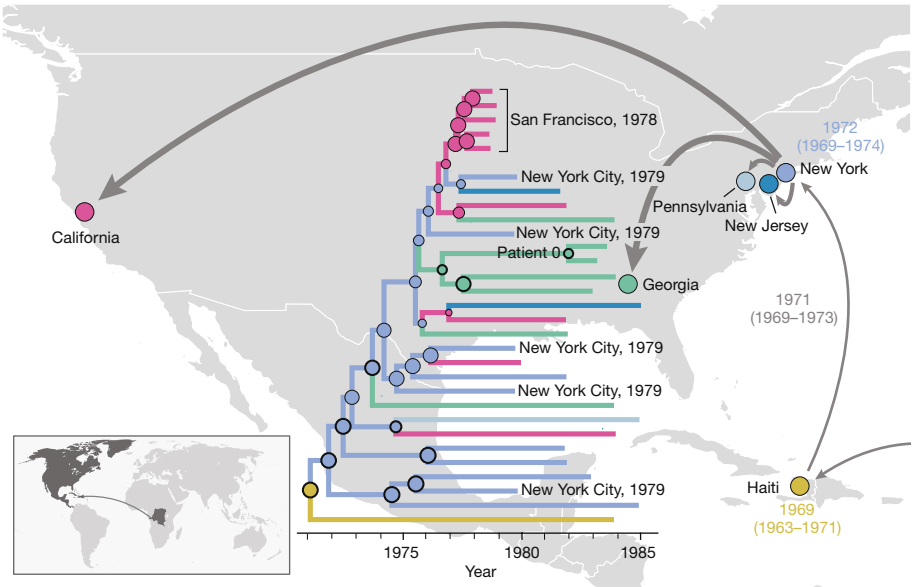


Figure 3 | The early patterns of HIV-1 subtype B spread in the Americas. The map summarizes the main patterns of spread inferred from the molecular clock phylogeographic analyses. The map inset shows the initial introduction of the subtype B lineage into the Caribbean from Africa. From there, the virus spreads first to NY and subsequently to different locations in the United States. The tree depicts the US clade, plus the most closely related basal Haiti strain, as inferred from the '*env* 74' analysis (Extended Data Fig. 5b). Tips of the clade correspond to the year of sampling. Tip branch colours reflect the actual sampling locations as indicated on the map; interior branches depict phylogenetically inferred locations using the same colour scheme. Diameters of internal node circles reflect posterior location probability values. Thick outer circles indicate internal nodes with posterior probability support >0.95. Thickness of the arrows reflects number of transitions inferred from this tree cluster. Mean dates and 95% credible intervals in yellow and blue represent the date estimates for the MRCA in the Caribbean and the US, respectively, based on the *env* 74 analysis. Date next to arrow between these locations represents the estimated timing of the corresponding jump. Patient 0 (represented by two sequences) and the earliest sequences from San Francisco (1978) and New York City (1979) are labelled. Maps made with Natural Earth.

attention, nor the first to display symptoms. In general, the CDC numbered cases in the order that the reports reached the agency from different cities and employed the terms 'cases' and 'patients' interchangeably. Patient 0, until he was linked to the cluster and took on his new name, was Case (or Patient) 057. The cluster study's LA 6 was the CDC's Case 032, and several cases in the New York section of the cluster[5] (Extended Data Fig. 7) were also reported before Patient 0 (and thus brought to investigators' attention first): NY 3 was Case 001, NY 2 was Case 002, NY 6 was Case 010 and NY 5 was Case 053 (ref. 14).

The information available for CDC investigators to establish symptom onset dates was often fragmentary and thus resisted uniform categorization. Sometimes onset was determined on the basis of lymphadenopathy, other times by the appearance or diagnosis of Kaposi's sarcoma or *Pneumocystis carinii* pneumonia. Investigators were unable to link to the cluster several NYC-based cases that had much earlier dates of symptom onset. For example, Case 154 was a middle-aged European man whose reported onset date for Kaposi's sarcoma was January 1975; and Case 153, when he was diagnosed with Kaposi's sarcoma in September 1981, recalled having swollen glands as early as June 1977 (ref. 15). Yet even within the cluster, Case 057's symptoms (lymphadenopathy in December 1979 and a Kaposi's sarcoma lesion diagnosed in May 1980, ref. 5) appeared considerably later than those of several other cases. LA 1 (Case 335) developed a lesion in February 1978 (ref. 16), whereas NY 1 (Case 152) experienced the onset of Kaposi's sarcoma in December 1978, NY 2 (Case 002) in May 1979 and NY 3 (Case 001) in August 1979 (ref. 14).

In his book *And the Band Played On*, Randy Shilts identified 'Patient Zero' by name as a highly sexually active French-Canadian flight attendant[17]. Unlike the initial reports of the cluster, media coverage of Shilts's book strongly insinuated that this individual was the source of the North American epidemic and an exemplar of dangerous disease transmission[18]—ideas which found a global audience (Supplementary Discussion). However, we found that the HIV-1 genome from this individual appeared typical of US strains of the time and was not basal to the US diversity, let alone to the deeper Caribbean subtype B diversity, in a manner that might be suggestive of a special role (Figs 1, 3). In short, we found no evidence that Patient 0 was the first person infected by this lineage of HIV-1.

In addition to donating plasma for analysis, Patient 0 provided investigators with the names of nearly 10% of his sexual partners over several years[5], while many other cluster patients were unable to share more than a handful of names[16]. This strongly suggests that ascertainment bias contributed to his central role in the cluster study and its diagrammatic representation. Later research would also call into question the cluster study's estimated average latency period of 10.5 months between sexual contact and symptom onset, with a revised average incubation period approaching 10 years for MSM. In retrospect, the study's sociogram (Extended Data Fig. 7) almost certainly depicted the sexual contacts of these men years after they had contracted HIV-1 (ref. 19) (Supplementary Discussion). Other East coast HIV-1 sequences fall much closer to the main early-California clade we identify than does that of Patient 0 (Fig. 3). Thus, while he did link AIDS cases in New York and Los Angeles through sexual contact, our results refute the widespread misinterpretation that he also infected them with HIV-1.

Much like historical reconstructions, phylogenetic inferences are often generated from data collected long after the critical events occurred. Our work highlights the importance of complete viral genomes from early archival specimens, carefully contextualized through historical analysis, without which this detailed picture of these early landmarks in the HIV/AIDS pandemic would not have been possible.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 5 April; accepted 7 September 2016.
Published online 26 October 2016.

1. Korber, B. *et al.* Timing the ancestor of the HIV-1 pandemic strains. *Science* **288,** 1789–1796 (2000).
2. Gilbert, M. T. *et al.* The emergence of HIV/AIDS in the Americas and beyond. *Proc. Natl Acad. Sci. USA* **104,** 18566–18570 (2007).
3. Holmes, E. C. When HIV spread afar. *Proc. Natl Acad. Sci. USA* **104,** 18351–18352 (2007).
4. Pape, J. W. *et al.* The epidemiology of AIDS in Haiti refutes the claims of Gilbert *et al*. *Proc. Natl Acad. Sci. USA* **105,** E13 (2008).
5. Auerbach, D. M., Darrow, W. W., Jaffe, H. W. & Curran, J. W. Cluster of cases of the acquired immune deficiency syndrome. Patients linked by sexual contact. *Am. J. Med.* **76,** 487–492 (1984).
6. Stevens, C. E. *et al.* Human T-cell lymphotropic virus type III infection in a cohort of homosexual men in New York City. *J. Am. Med. Assoc.* **255,** 2167–2172 (1986).
7. Szmuness, W., Stevens, C. E., Zang, E. A., Harley, E. J. & Kellner, A. A controlled clinical trial of the efficacy of the hepatitis B vaccine (Heptavax B): a final report. *Hepatology* **1,** 377–385 (1981).
8. Koblin, B. A., Morrison, J. M., Taylor, P. E., Stoneburner, R. L. & Stevens, C. E. Mortality trends in a cohort of homosexual men in New York City, 1978–1988. *Am. J. Epidemiol.* **136,** 646–656 (1992).
9. Jaffe, H. W. *et al.* The acquired immunodeficiency syndrome in a cohort of homosexual men. A six-year follow-up study. *Ann. Intern. Med*. **103,** 210–214 (1985).
10. Foley, B., Pan, H., Buchbinder, S. & Delwart, E. L. Apparent founder effect during the early years of the San Francisco HIV type 1 epidemic (1978–1979). *AIDS Res. Hum. Retroviruses* **16,** 1463–1469 (2000).
11. Centers for Disease Control (CDC) A cluster of Kaposi's sarcoma and *Pneumocystis carinii* pneumonia among homosexual male residents of Los Angeles and Orange Counties, California. *MMWR Morb. Mortal. Wkly. Rep.* **31,** 305–307 (1982).
12. McKay, R. A. *Imagining 'Patient Zero': Sexuality, Blame, and the Origins of the North American AIDS Epidemic*. Doctoral thesis, Univ. of Oxford (2011).
13. Harden, V. A. *AIDS at 30: A History* (Potomac Books, 2012).
14. Darrow, W. W. Trip report to New York City, July 12–16 and August 3–6, 1982. CDC Task Force on AIDS, internal communication (3 September 1982).
15. Darrow, W. W. Time–space clustering of KS cases in the City of New York: evidence for horizontal transmission of some mysterious microbe. CDC Task Force on Kaposi's Sarcoma and Opportunistic Infections, internal communication (3 March 1982).
16. Darrow, W. W. & Auerbach, D. M. Los Angeles cluster: background. CDC Task Force on Kaposi's Sarcoma and Opportunistic Infections, internal communication (12 May 1982).
17. Shilts, R. *And the Band Played On: Politics, People, and the AIDS Epidemic* (St. Martin's Press, 1987).
18. McKay, R. A. "Patient Zero": the absence of a patient's view of the early North American AIDS epidemic. *Bull. Hist. Med.* **88,** 161–194 (2014).
19. Moss, A. R. In response to: AIDS without end. *New York Rev. Books* **35,** 60 (1988).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank C. Stevens and D. Hemmerlein for facilitating access to archival sera; G.-Z. Han, A. Bjork, W. Switzer, V. Sullivan, R. Ruboyianes and P. Sprinkle for technical assistance; T. Spira and M. Owen for geographical data on some published sequences; and the NIH AIDS Reagent program for providing reference virus samples US657 and HT599. W. W. Darrow led the initial 1982 cluster investigation and provided R.A.M. with access to his copies of archival CDC documents. This work was supported by NIH/NIAID R01AI084691 and the David and Lucile Packard Foundation (M.W.); the Wellcome Trust (080651), the University of Oxford's Clarendon Fund, the Economic and Social Research Council (PTA-026-27-2838), and a J. Armand Bombardier Internationalist Fellowship (R.A.M.); the Research Fund KU Leuven (Onderzoeksfonds KU Leuven, Program Financing no. PF/10/018) and the 'Fonds voor Wetenschappelijk Onderzoek Vlaanderen' (FWO) (G066215N) (P.L); and NSF DMS 1264153, NIH R01 HG006139 and NIH R01 AI107034 (M.A.S.).

Author Contributions M.W., H.W.J., P.L. and R.A.M. conceived the study. T.D.W and M.W. designed the RNA jackhammering method. T.D.W. generated the sequences. B.A.K. provided serum samples from New York City. W.H. and T.G. acquired specimens and provided serological data. D.E.T. provided conceptual input. M.W., M.A.S. and P.L. prepared the data sets and performed the phylogenetic analyses. R.A.M. performed the historical analyses. M.W., H.W.J., P.L. and R.A.M. wrote the paper. All authors discussed the results and commented on the manuscript. The findings and conclusions in this report are those of the author(s) and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.W. (worobey@email.arizona.edu) or R.A.M. (ram78@cam.ac.uk).

Reviewer Information *Nature* thanks K. Andersen and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

HIV-1 serological screening of serum samples from San Francisco from 1978. We tested 2,231 samples collected from the MSM cohort in San Francisco in 1978 (ref. 9) and detected 83 HIV-1-positives by Western Blot (3.7% prevalence). Samples were first screened by GS HIV-1/HIV-2 Plus O EIA (Bio-Rad Laboratories) and reactive samples were further tested by WB Genetic Systems HIV-1 Western Blot (Bio-Rad Laboratories).

HIV-1 nucleic acid amplification. A total of 33 samples of frozen serum from New York City previously identified as positive for antibody to HIV-1[6–8] were assayed; and a total of 20 frozen serum samples from San Francisco[9], identified as part of the present study as positive for antibody to HIV-1, were assayed. The New York City samples were from 1978 and 1979 though no complete genomic sequences from 1978 were developed. The San Francisco samples were all from 1978. RNA recovered from samples from both NY and SF was generally undetectable when assaying 5-μl aliquots in a Qubit 2.0 fluorometer using the Qubit RNA HS reagents (detection limit, 250 pgμl^{-1}).

Additionally, a sample of peripheral blood mononuclear cells (PBMCs) and a sample of serum were both assayed; these had been collected from a single individual in 1983 (Patient 0), and the samples were stored at CDC Atlanta. Other than Patient 0, now deceased, the data recorded were unlinked to individual identifiers and the work was approved by the Human Subjects Protection Program at the University of Arizona.

Four panels of degenerate primers (Supplementary Table 1 and Extended Data Fig. 1) were designed using a suite of North American subtype B sequences. We aimed to design primers able to amplify both conserved regions and predictably variable sites. Primers within each panel were designed to generate sequence from the 5′ end of *gag* to the 3′ end of *nef* and were designed to amplify overlapping fragments. Two panels 'HIVL' ($n=25$) and 'HIVLb' ($n=22$) were designed to amplify fragments of approximately 500–650 bases in length. Two other panels 'HIVM' ($n=50$) and 'HIVR' ($n=46$) were designed to amplify fragments of approximately 200–320 bases in length.

Nucleic acids from 100-μl aliquots of serum (or PBMCs in the case of Patient 0) were isolated using the QIAamp viral RNA mini kit (Qiagen) with 5 mcg added carrier RNA. Serum samples were then treated with DNase I (Invitrogen, Life Technologies) before reverse transcription. PBMC nucleic acids were left untreated.

Proviral DNA from Patient 0's PBMCs was amplified with all four primer panels and from multiple separate isolations. Amplification was achieved using Invitrogen platinum Taq DNA polymerase high fidelity (Life Technologies) and run for 55 cycles at an annealing temperature of 52 °C. Additionally, attempts were made to amplify longer fragments using PCR supermix high fidelity (Life Technologies) and forward and reverse primers matched from the HIVLb primer panel for long fragment length followed by nesting with primers for slightly shorter fragment length. A single fragment of slightly more than 7,000 bases was generated after multiple attempts with multiple primer combinations and cloned using the Invitrogen TOPO XL PCR cloning kit (Life Technologies). Fragments of individual clones were then amplified using HIVLb forward and reverse primers matched to give approximately 1,000-base overlapping fragments and then sequenced.

RNA jackhammering. RNA jackhammering of the serum samples proceeded as follows: aliquots of RNA extract were reverse transcribed using the GoScript reverse transcription system (Promega) using a program of 4 cycles of 50 °C for 30 min followed by 55 °C for 30 min and a final incubation at 85 °C for 10 min. Primers used were pools of reverse primers from widely spaced amplicons (Supplementary Table 1, Extended Data Fig. 1), typically nine or ten primers per pool in a single reaction tube, with the wide spacing abrogating the possibility of incorporation of an internal primer into any given amplicon. Reverse transcription products were then briefly amplified in multiplex reactions in the pool-specific tube (denaturation for 3 min at 94 °C followed by 30 cycles of 94 °C for 30 s, 52 °C for 30 s, 68 °C for 30 s, and a final extension of 68 °C for 5 min) with matching forward primer pools (a 'preliminary amplification' step). Sequences were then amplified from individual aliquots taken from the pool-specific tubes, via single primer pairs (denaturation for 3 min at 94 °C followed by 40 cycles of 94 °C for 30 s, 52 °C for 30 s, 68 °C for 30 s, and a final extension of 68 °C for 5 min). Two separate isolates were amplified from each sample in this manner, with a minimum of one amplification with each primer panel per isolate. Five out of the 33 (15%) of the NY sera assayed yielded complete HIV-1 genomic data as did 3 out of the 20 (15%) SF sera, suggesting that levels of viral RNA preservation were very similar in each collection.

In Extended Data Fig. 1 we schematically illustrate the RNA jackhammering approach and its advantages over standard RT–PCR procedures for degraded, low input samples. For a conventional RT–PCR approach with a fairly long amplification product we would perform reverse transcription and obtain one potentially amplifiable cDNA product. We would then aliquot ~10% of the reverse transcription product for amplification in a PCR reaction with forward and reverse primers. Even if the single cDNA product made it into the PCR reaction, the desired amplification product would be too long and a PCR amplicon would therefore not be obtained. For RT–PCR with a shorter amplification product, more appropriately sized given the damaged RNA in the sample, there was still a 90% chance that it would be deemed a negative sample since most aliquots will not contain the rare cDNA product. Using multiple primer sets would increase the chance of a PCR-positive result, but most PCR reactions remained negative because most aliquots lack target cDNA. Even with a 10 primer-pair pool and 10 final PCR reactions, there may be no amplified product. The RNA jackhammering approach targets large panels of appropriately short amplicons, uses discrete pools of non-overlapping primers pairs for reverse transcription, and includes a crucial multiplex pre-amplification step to ensure that each aliquot contained ample template molecules for the final PCR amplification (a separate reaction for each primer pair in the entire panel).

Sequencing was performed at the University of Arizona Genetics Core using an ABI 3730XL. The Patient 0 sample contained considerable heterogeneity (mixed bases) both in proviral assembly and in viral RNA amplification. Heterogeneity in the NY and SF samples (all sequences derived from viral RNA) was low. In all cases consensus sequences were used in the phylogenetic analyses. Primer sequences were computationally removed from all sequence data before assembling genomic consensus sequences, which yielded coding-complete genomic data with exception of a few small gaps and the 3′ end of the *nef* gene (Supplementary Table 2).

Validation of the jackhammering approach. To validate this approach we obtained seed stock samples from the NIH AIDS Reagent program of subtype B viruses from the US (US657) and Haiti (HT599) and applied a jackhammering approach with independent runs of both the HIVM and HIVR primer panels (Extended Data Fig. 8).

For US657 we recovered, in total, from both runs combined, 8,194 nt of high quality data. HIVM and HIVR are independent runs with completely different primer sets, yet where the data overlapped, they were >99.9% similar. Moreover, the few heterogeneities did not line up with heterogeneous primers but fell in regions between primers, demonstrating that differences could not be attributed to the incorporation of primers into the recovered sequences. This was expected both because the wide spacing of amplicons within a single pool of primer pairs prevents incorporation of primers within amplified products and because all primer sequences from final amplification products were computationally removed from the sequences before assembly of genomic sequences. There are 3,354 bases in the published US657 sequence. Our data covered about 90% of the 3,354 bases of previously published US657 sequence (GenBank accession number U04908) and all of our individual amplicons in the region of overlap had US657 as the highest BLAST hit and were >99% similar to the published sequence.

For HT599 the HIVM and HIVR primer panels developed 8,545 nt of data, 99.6% of the target. The HIVM-derived sequence was >99.9% similar to the HIVR-derived sequence. We recovered 100% of the overlap with the previously published HT599 sequence (2,881 nt, GenBank accession number U08447) with 99.5% similarity.

To evaluate discrepancies between the jackhammering-recovered sequences and both US657 and HT599, we compared consensus sequences of combined HIVM and HIVR data with the respective published sequences by adding them to our complete genome alignment and reconstructing a maximum likelihood tree (Extended Data Fig. 8a). As expected, the independently generated sequences from each virus clustered very closely and only had short tips from their common ancestors, resulting from a very small number of substitutions in their overlapping regions. In a root-to-tip analysis (Extended Data Fig. 8b), our sequences (with a target symbol) were associated with somewhat smaller residuals than the published sequences (with a circle), indicating that our data are likely to be more accurate and, importantly, cannot contain primer remnants as this would result in much larger residuals.

Sequence data. To construct the data sets for the analyses shown in Fig. 1 and Extended Data Figs 2–4 we searched the Los Alamos National Laboratories (LANL) HIV database (http://hiv.lanl.gov/) for all available genome-length HIV-1 sequences from Caribbean countries, which had previously been shown to exhibit diverse subtype B lineages that fall basal to a monophyletic 'pandemic' clade of subtype B that accounts for most US and other non-Caribbean subtype infections[2]. These included sequences sampled in Haiti, Dominican Republic, Jamaica and from Haitians who had recently immigrated to the US from Haiti ('H3' and 'H5' from 1982, 'H6' and 'H7' from 1983, 'RF_HAT' from 1983)[2]. For sequences H3, H5, H6 and H7 *pol* sequences were not available, but partial *gag* and full-length *env* sequences were available. For the full-genome analyses the *pol* gene was treated as missing data. We then added a similar number of genomes from the US from

a similar time period (1982–2005), plus one each from France and the UK, as well as outgroup sequences of subtype D from the Democratic Republic of the Congo (D.R.C.). We called this the 'full genome 46' data set because it contained 46 genomes. The *gag*, *pol* and *env* data sets depicted in Extended Data Fig. 3 were each derived from the respective sub-genomic region of this same set of taxa. The subset of 'full genome 46' that contained only those US sequences sampled from 1978–1984 we called 'full genome 38'.

For the *env* analyses in Fig. 3 and Extended Data Fig. 5 the alignment from ref. 2 was used, with the addition of the sequences generated for the present study, additional Caribbean subtype B sequences from 2000 to 2005, and four early subtype B partial *env* sequences from San Francisco[10]. This alignment we called '*env* 105'. The subset that contained only those US sequences sampled from 1978–1984 we called '*env* 74'.

For Extended Data Fig. 6 we added to '*env* 105' a comparable number—relative to those sampled from 1978–1984 from known locations (New York, California, Georgia, Pennsylvania, New Jersey) (Extended Data Fig. 4b)–of randomly sampled sequences from 1997–2007 from NY, SF, and North Carolina (the closest available site with sufficient numbers to stand in for the Georgia ones from the 1978–1984 sample). We called this alignment '*env* 133'.

In all cases sequences were manually aligned using Se-Al (http://tree.bio.ed.ac.uk/software/seal/). All sequence alignments, input files, tree files and primer sequences are available at the Dryad Digital Repository (doi:10.5061/dryad.7mv7v).

Recombination analysis and maximum likelihood tree reconstruction. Maximum likelihood phylogenies were reconstructed using RAxML under on a general time-reversible model of substitution with gamma distributed rate variation among sites[20]. Bootstrap support values were calculated using 1,000 pseudo-replicates. To detect the presence of recombination, we first performed the Phi test[21] on every data set (Extended Data Table 1). When the null hypothesis of absence of recombination was rejected ($P < 0.05$), we subsequently analysed the data set using RDP4 (ref. 22) and produced new alignments in which the minor recombinant regions were deleted from putative recombinants. Re-analyses of these 'recombination-free' data sets using the Phi test confirmed the absence of detectable recombination signal ($P > 0.05$, Extended Data Table 1).

Bayesian phylogenetic inference. Time-measured phylogeographic histories were reconstructed using a Bayesian phylogenetic inference approach implemented in BEASTv1.8.2 (ref. 23). Our full probabilistic model combined sequence substitution over an unknown phylogeny calibrated in time units using a molecular clock process with dated tips[24], a coalescent tree prior and a discrete diffusion process among discrete location states[25]. For the sequence substitution process, we used the same model as for the maximum likelihood reconstructions. We accommodated rate variation among lineages using a lognormal distribution in an uncorrelated relaxed molecular clock model[26] and integrated out each sampling date over an uncertainty interval of one year. Visual inspections of root to tip divergence as a function of sampling time using TempEst[27] indicated a strong temporal signal with no clear outlier sequences (Extended Data Fig. 9).

For most analyses, we flexibly modelled changes in effective population size through time by specifying a Bayesian skygrid non-parametric tree prior with a grid of 50 years and yearly effective population size parameters[28]. (The notion of 'effective population size', or 'effective infections' in epidemiological applications, comes from population genetics, and is typically lower than the full (that is, census) population size, reflecting, for example, variance in reproductive success among individuals—transmissions to new hosts in this context). To estimate viral population growth rates in both the Caribbean and US populations, we fitted a 'nested' coalescent model to the data set with the largest taxon sampling (*env* 133). This model fits a constant-logistic demographic function[29] to the genealogy excluding the US clade. The initial constant phase was included in the model to accommodate the deep branching between the subtype B sequences and the African subtype D outgroup sequences. Nested within this model, a separate logistic growth model was fitted to the US clade in the genealogy.

The process of discrete diffusion among locations was modelled using a general non-reversible substitution model[30]. In our analyses including the African subtype D outgroup lineages, we set the root state frequency to one for the African state and zero for all other possible discrete states. We obtained estimates of the transitions among locations (Markov jumps) using a stochastic mapping implementation capable of inferring the complete Markov jump history[31,32]. We approximate the posterior distribution for our full probabilistic model using Markov chain Monte Carlo (MCMC) sampling. We use BEAGLE in conjunction with BEAST to improve the computational performance of our analyses[33]. MCMC chains were run for 50,000,000 generations, sampling every 5,000 generations. We diagnosed the runs by examining trace plots and effective samples sizes, and summarized continuous parameters (mean and 95% highest posterior density (HPD) intervals) using Tracer (http://tree.bio.ed.ac.uk/software/tracer/) after discarding a 10% burn-in. Trees were summarized as maximum clade credibility trees using TreeAnnotator and visualized in FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

In two specific phylogeographic analyses, we assessed (i) to what extent sequences sampled early in the US epidemic characterize the subtype B diversity in the US clade (Extended Data Fig. 6a) and (ii) to what extent the location state at the origin of the US clade can be estimated using sequences sampled later in the epidemic from three different US states (Extended Data Fig. 6b). For this purpose, we first reconstructed time-measured phylogenies for the *env* 133 data set using the substitution model, molecular clock model and coalescent model described above and subsequently reconstructed ancestral locations on the inferred posterior distribution of trees.

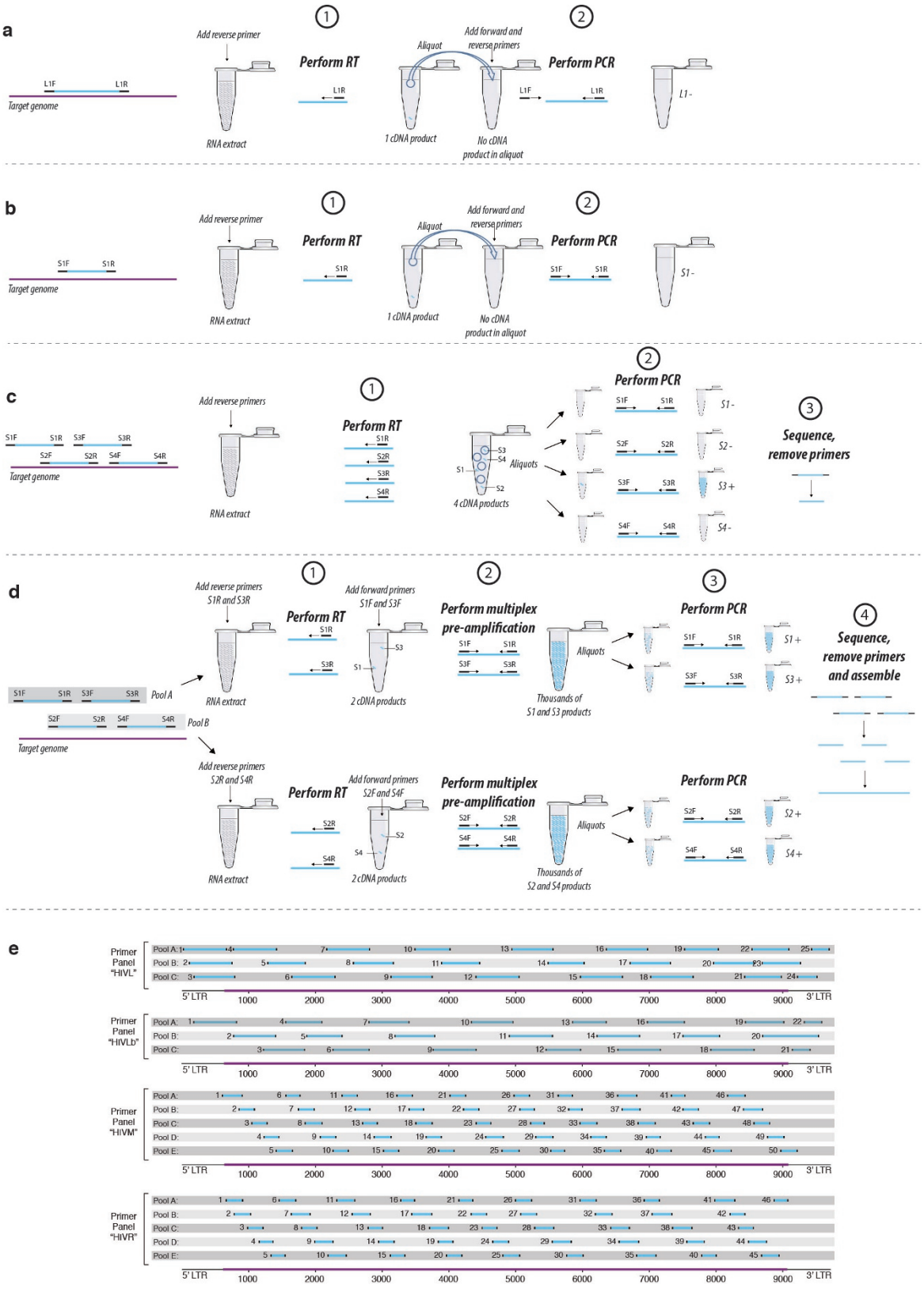
For Extended Data Fig. 6a, we classified US sequences as 'early' or 'late' depending on whether they were sampled before or after (and including) 1985. For Extended Data Fig. 6b, we first pruned the necessary US sequences from the posterior distributions in order to retain only 'late' sequences from New York, North Carolina and California (matching the sampling from New York, Georgia and California in Fig. 3 and Extended Data Fig. 5b). In this case, the support for a NYC ancestral state is likely upheld by the presence of two basal NYC representatives, but location estimates in a star-like tree structure with long tip branches will be critically dependent on how well the diversity of any location is represented in the contemporaneous sampling, as recently noted[34].

Comparison of phylogeographic estimates before and after deleting minor recombinant regions from putative recombinants (Extended Data Table 1) indicated highly consistent results.

Data availability. All sequence alignments, input files, tree files and primer sequences are available at the Dryad Digital Repository (doi:10.5061/dryad.7mv7v).

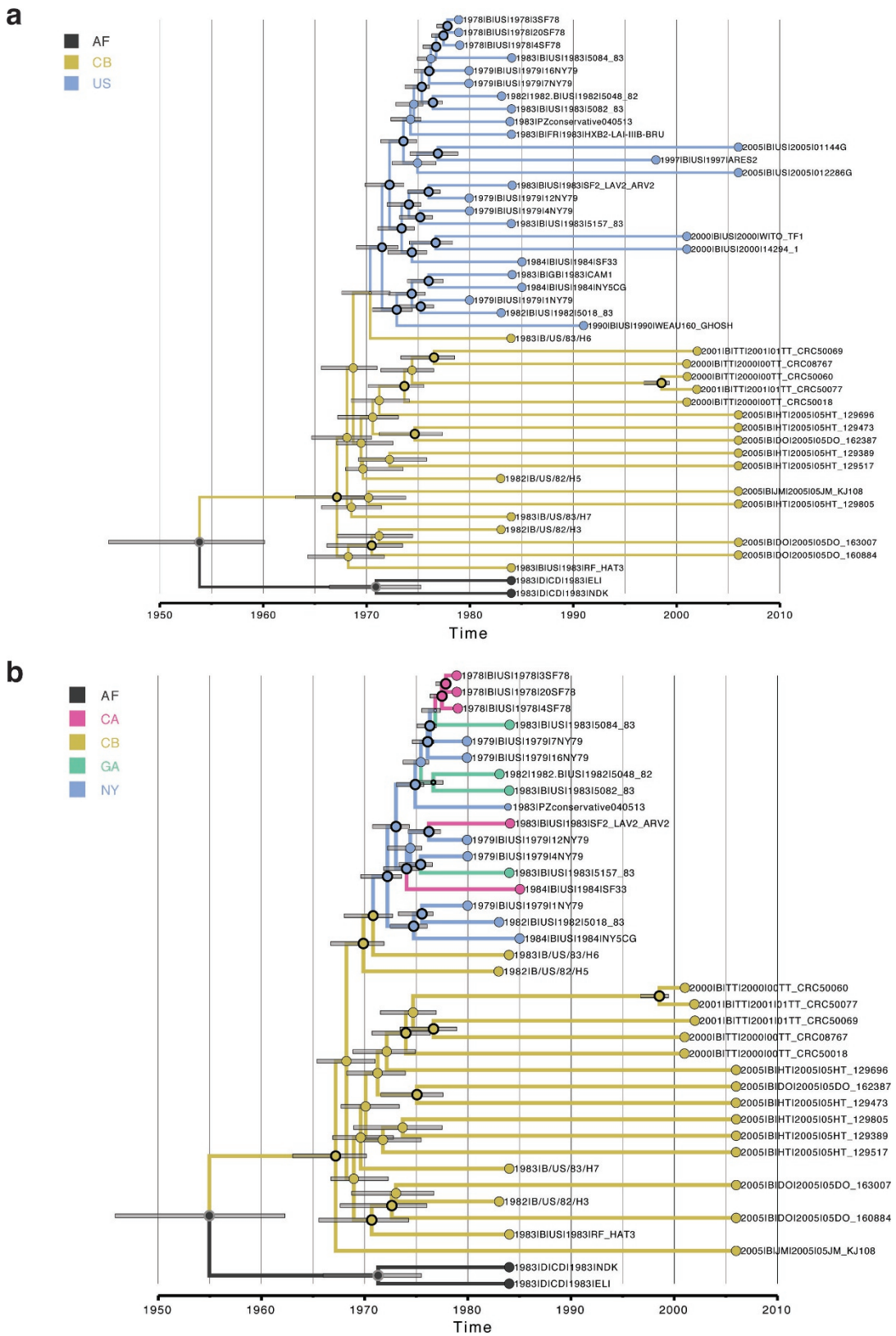
The HIV-1 sequences reported here have been deposited in GenBank under accession numbers KJ704787, KJ704788, KJ704789, KJ704790, KJ704791, KJ704792, KJ704793, KJ704794, KJ704795, KJ704796 and KJ704797.

20. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).
21. Bruen, T. C., Philippe, H. & Bryant, D. A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172,** 2665–2681 (2006).
22. Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: detection and analysis of recombination patterns in virus genomes. *Virus Evol.* **1,** vev003 (2015).
23. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29,** 1969–1973 (2012).
24. Rambaut, A. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* **16,** 395–399 (2000).
25. Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian phylogeography finds its roots. *PLOS Comput. Biol.* **5,** e1000520 (2009).
26. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4,** e88 (2006).
27. Rambaut, A., Lam, T. T., de Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst. *Virus Evol.* **2,** DOI: http://dx.doi.org/10.1093/ve/vew007 (2016).
28. Gill, M. S. *et al.* Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evol.* **30,** 713–724 (2013).
29. Faria, N. R. *et al.* HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. *Science* **346,** 56–61 (2014).
30. Edwards, C. J. *et al.* Ancient hybridization and an Irish origin for the modern polar bear matriline. *Curr. Biol.* **21,** 1251–1258 (2011).
31. Minin, V. N. & Suchard, M. A. Counting labeled transitions in continuous-time Markov models of evolution. *J. Math. Biol.* **56,** 391–412 (2008).
32. Lemey, P. *et al.* Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS Pathog.* **10,** e1003932 (2014).
33. Suchard, M. A. & Rambaut, A. Many-core algorithms for statistical phylogenetics. *Bioinformatics* **25,** 1370–1376 (2009).
34. Gräf, T. *et al.* Contribution of epidemiological predictors in unraveling the phylogeographic history of HIV-1 subtype C in Brazil. *J. Virol.* **89,** 12341–12348 (2015).

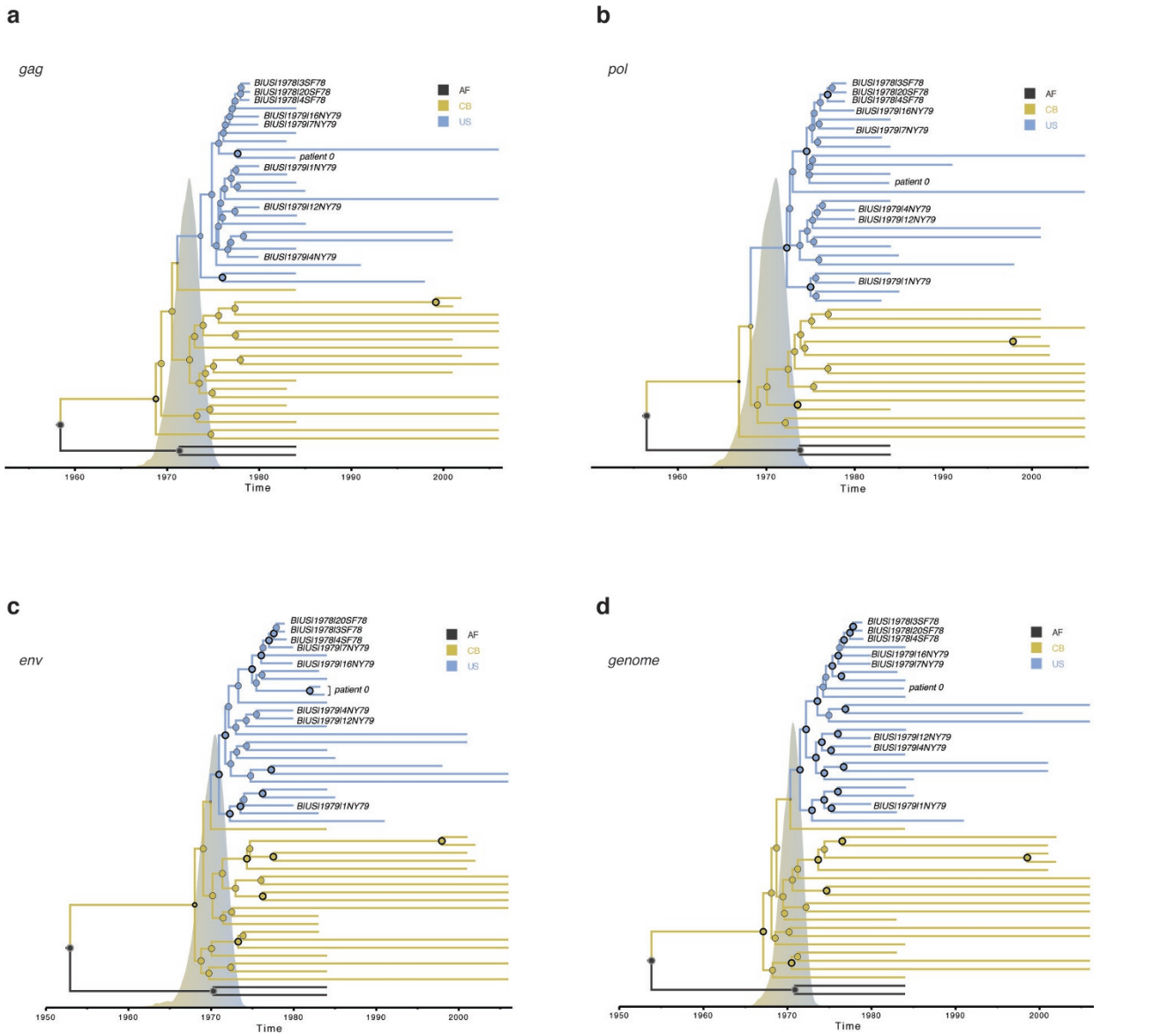


Extended Data Figure 1 | See next page for caption.

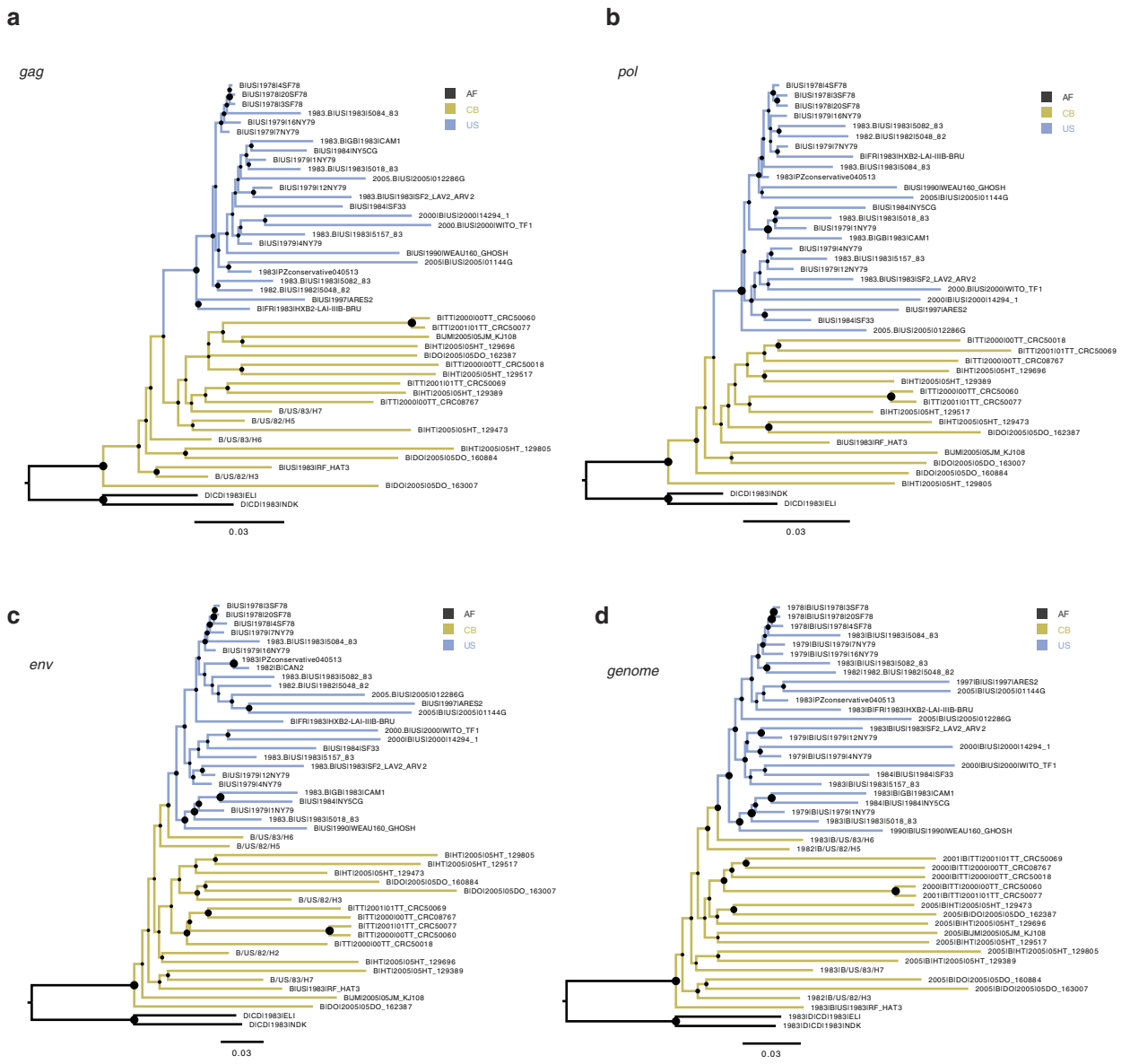
Extended Data Figure 1 | Jackhammering schematic and primer panels and pools. a–d, Detection and amplification of target RNA molecules in old, degraded, low-titre samples. For the purposes of illustration, consider a tube with 10^{13} RNA molecules, but (because of the low RNA quality) only one molecule that is (i) capable of being primed by the given reverse primer(s) and (ii) long enough to form a 200-bp product. **a**, Conventional RT–PCR with a long amplification product, oversized for a sample with RNA less than ~200 bases in length. **b**, RT–PCR with a shorter amplification product. **c**, Use of multiple primer pairs to increase the chance of at least one PCR-positive result. **d**, The jackhammering approach, which overcomes the problems encountered in **a–c** by (i) targeting an extensive panel of short amplicons appropriately sized to the level of RNA survival in the sample, (ii) conducting reverse transcription with pools of primer pairs that amplify discrete, non-overlapping genomic regions, and (iii) employing a multiplex pre-amplification step, in the tube with the reverse transcription product, to generate sufficient DNA to ensure that each aliquot from it contains numerous template molecules for final PCR amplification. In this schematic, we show just two primer pairs per pool, but we used pools of ten pairs with our largest primer panels (shown in **e**, HXB2 numbering along HIV-1 genome). With a 10 primer-pair pool, and 10 final reactions, one can reliably recover 10 bands for sequencing. Five such pools (one entire panel of 50 pairs), allows complete HIV-1 genome recovery even in heavily degraded samples.



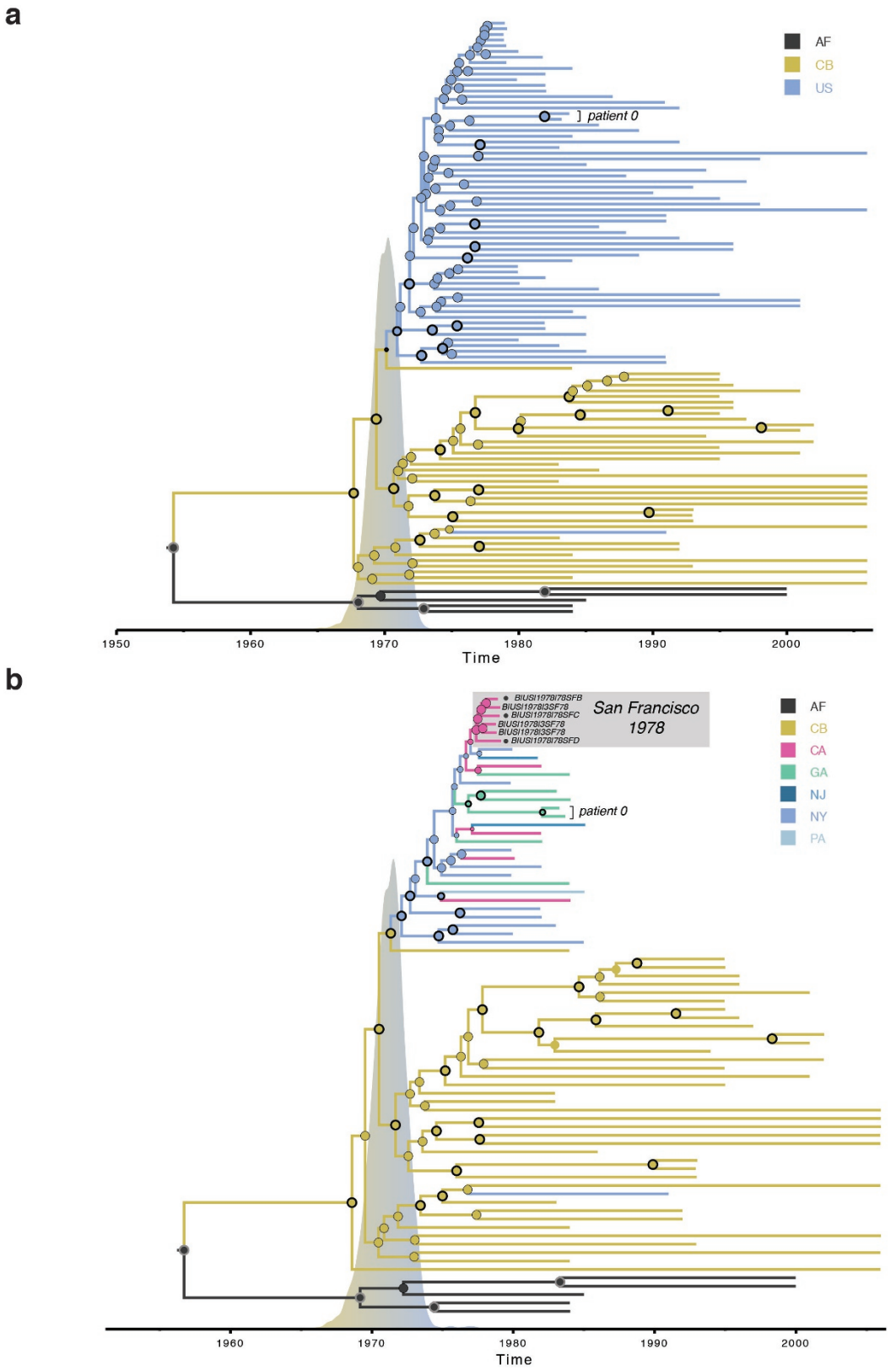
Extended Data Figure 2 | Maximum clade credibility (MCC) tree summaries of Bayesian spatio-temporal reconstructions based on complete HIV-1 genome data. a, 'full genome 46', **b**, 'full genome 38'. The tips of the trees correspond to the year of sampling while the branch (and node) colours reflect location: the sampling location for the tip branches and the inferred location for the internal branches. AF, Africa; CB, Caribbean; US, the United States; CA, California, GA, Georgia; NY, New York. The diameters of the internal node circles reflect posterior location probability values. Thick outer circles represent internal nodes with posterior probability support >0.95. Grey bars indicate the 95% credibility intervals for the internal node ages. The tree in **b** represents the fully annotated version of the tree in Fig. 1 in the main text.



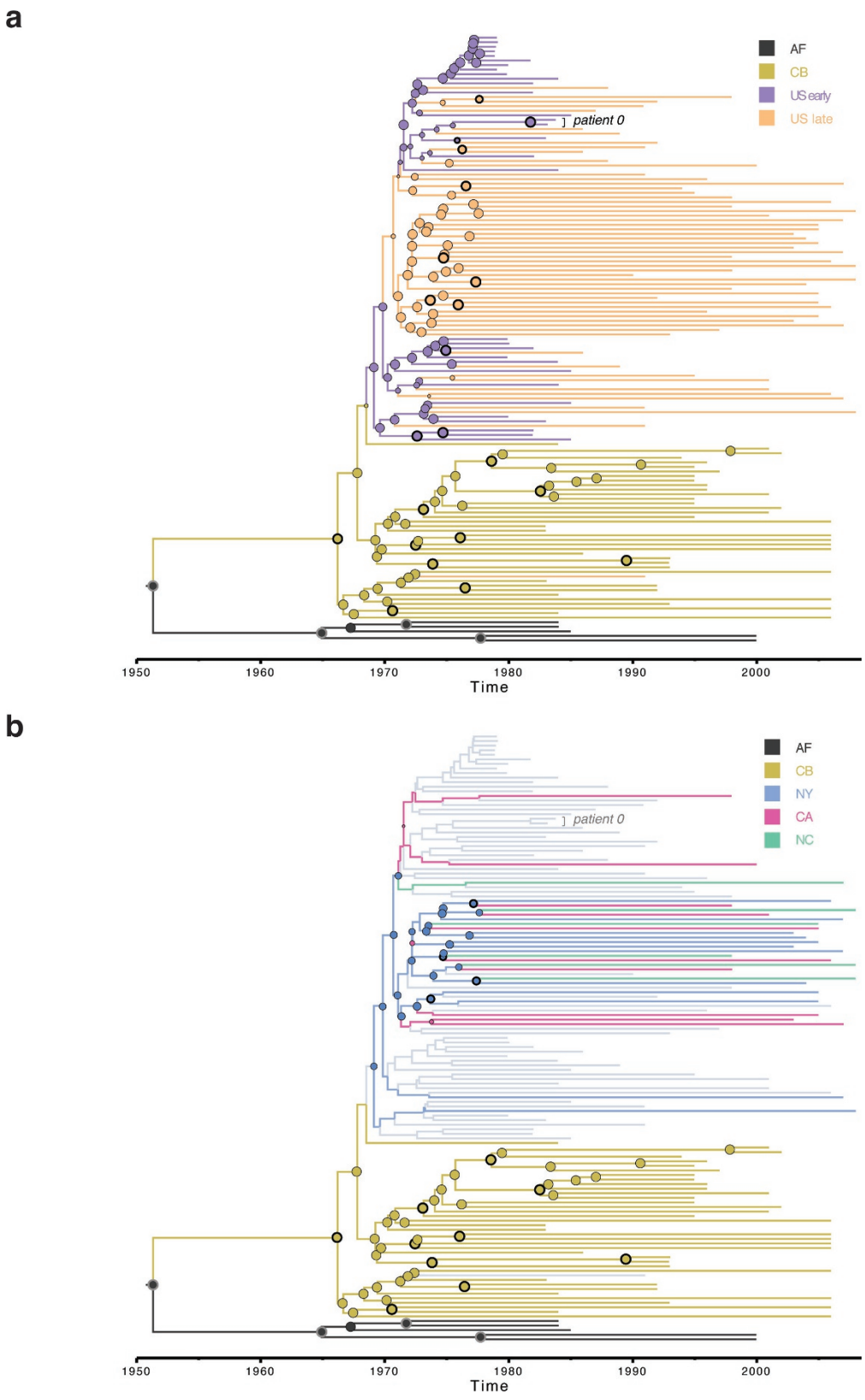
Extended Data Figure 3 | Maximum clade credibility (MCC) tree summaries of Bayesian spatio-temporal reconstructions based on different genome region data sets. MCC trees for the same strains are shown for **a**, *gag*, **b**, *pol*, **c**, *env* and **d**, the complete genome. The tips of the trees correspond to the year of sampling while the branch (and node) colours reflect location: the sampling location for the tip branches and the inferred location for the internal branches. AF, Africa; CB, Caribbean; US, the United States. Tip labels are provided for the newly obtained archival HIV-1 genomes. The diameters of the internal node circles reflect posterior location probability values. Thick outer circles represent internal nodes with posterior probability support >0.95. We also depict the posterior probability densities for the time of the introduction event from the Caribbean into the US on the time scale of the trees.



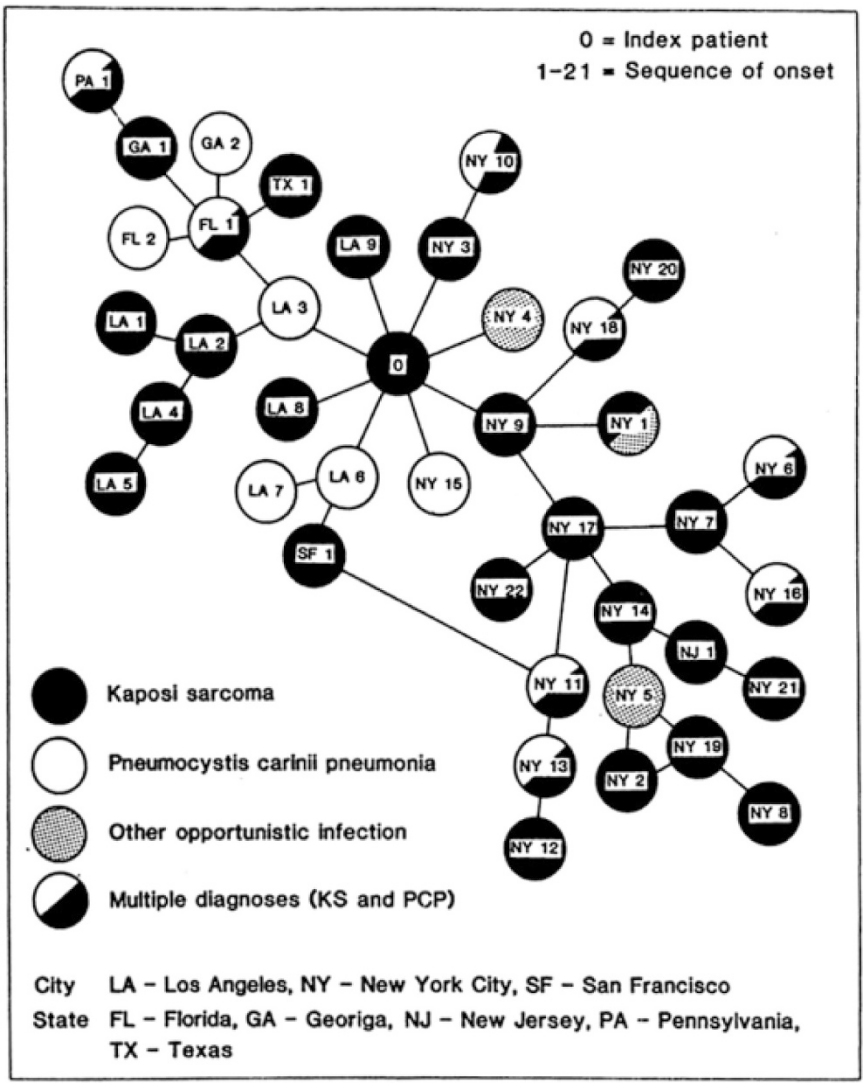
Extended Data Figure 4 | Maximum likelihood phylogenies for the different genome region data sets. a, *gag*, **b**, *pol*, **c**, *env* and **d**, the complete genome. We analysed the same data sets as in Extended Data Fig. 3. The diameters of the internal node circles reflect bootstrap support values. We manually coloured the branches in a similar way as for the Bayesian phylogeographic reconstructions.



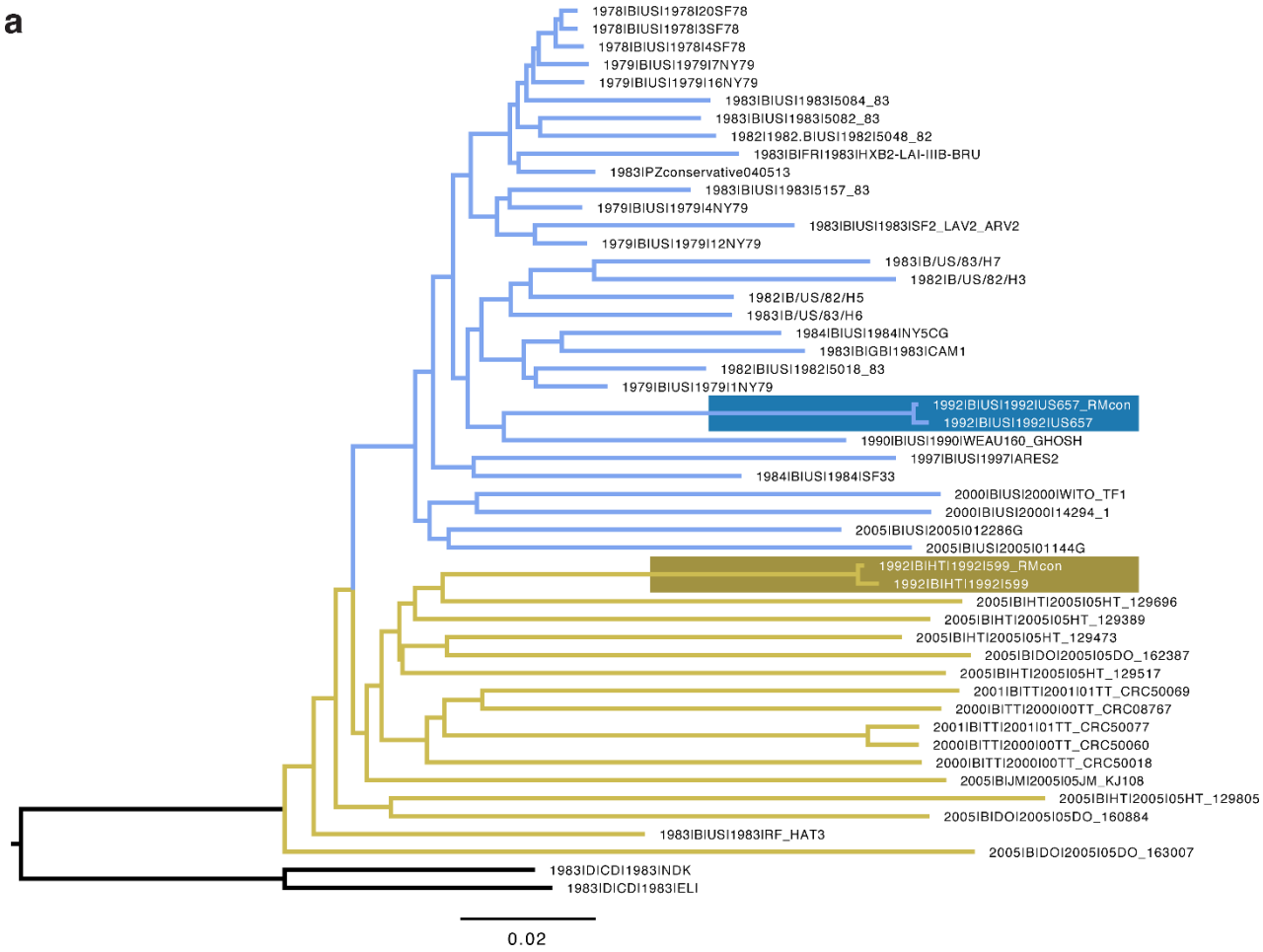
Extended Data Figure 5 | Maximum clade credibility (MCC) tree summaries of Bayesian spatio-temporal reconstructions based on different *env* data sets. a, '*env* 105', **b**, '*env* 74'. The tips of the trees correspond to the year of sampling while the branch (and node) colours reflect location: the sampling location for the tip branches and the inferred location for the internal branches. AF, Africa; CB, Caribbean; US, the United States, CA, California; GA, Georgia; NJ, New Jersey, NY, New York; PA, Pennsylvania. The diameters of the internal node circles reflect posterior location probability values. Thick outer circles represent internal nodes with posterior probability support >0.95. We also depict the posterior probability density for the time of the introduction event from the Caribbean into the U.S on the time scales of the trees. The three partial *env* sequences from SF in 1978 (ref. 10) are highlighted with bullets.



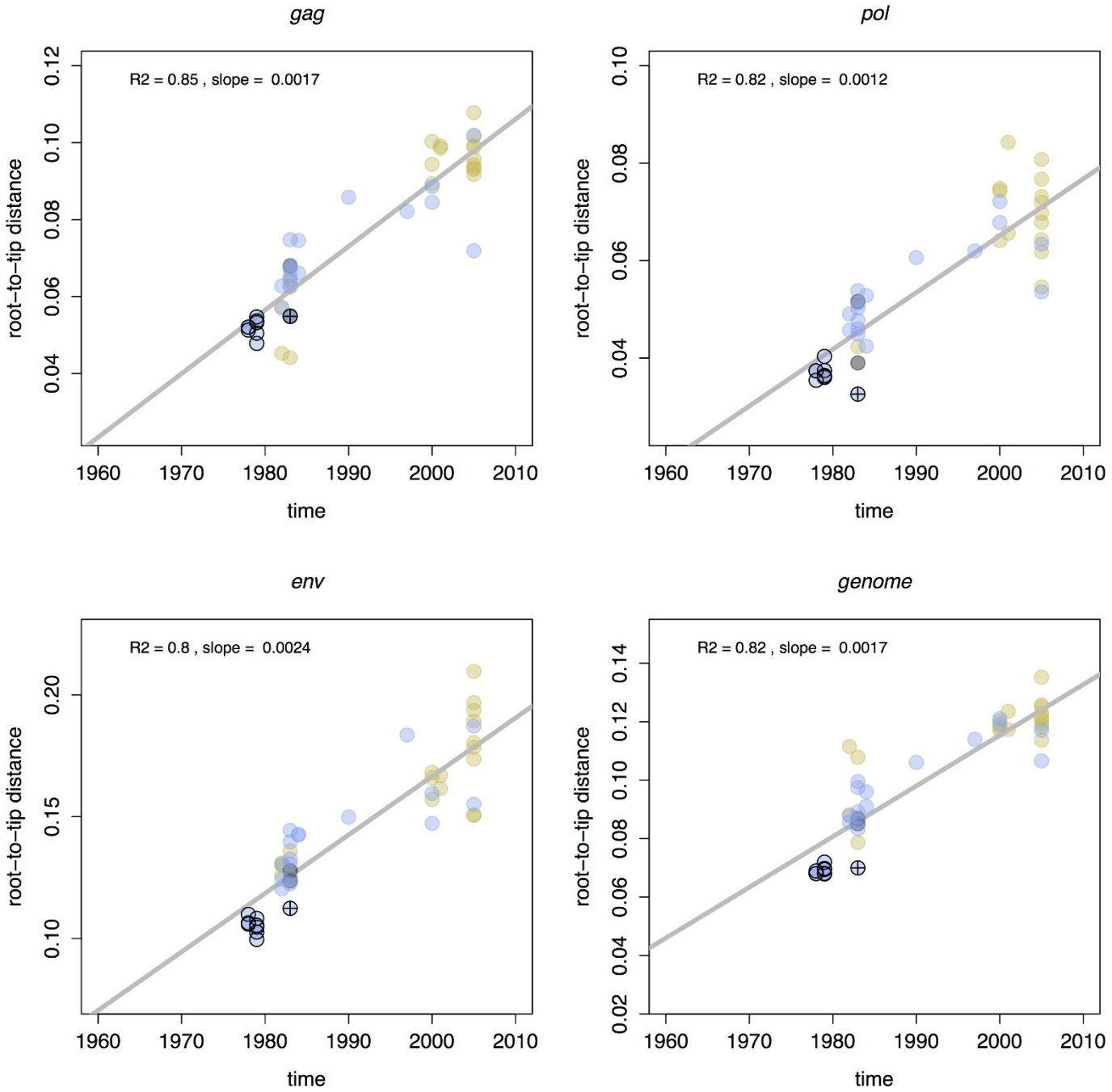
Extended Data Figure 6 | Maximum clade credibility (MCC) tree summaries of Bayesian spatio-temporal reconstruction comparing early and late strains. a, '*env* 133', **b**, only 'late' sequences from '*env* 133'. In **a**, we classified US sequences as 'early' or 'late' depending on whether they were sampled before or after (and including) 1985. In **b**, the analysis was conducted on an empirical tree distribution of '*env* 133' from which we pruned early US sequences (in grey), but we still annotate the reconstruction on the complete phylogenies for reference. The tips of the tree correspond to the year of sampling while the branch (and node) colours reflect location: the sampling location for the tip branches and the inferred location for the internal branches. AF, Africa; CB, Caribbean; US early, the United States sampled <1985; US late, the United States sampled in or after 1985; CA, California; GA, Georgia; NC, North Carolina, NY, New York. The diameters of the internal node circles reflect posterior location probability values. Thick outer circles represent internal nodes with posterior probability support >0.95.



Extended Data Figure 7 | A cluster of 40 early AIDS patients linked through sexual contact. Reprinted from figure 1 of ref. 5 with permission from Elsevier.



Extended Data Figure 8 | Jackhammering validation with reference viruses. a, The consensus sequences for primer panels HIVM and HIVR ('RMcon' suffix) were included, with previously published sequences for an US (US657) virus and a Haitian (HT599) virus, in a maximum likelihood tree. The two clusters of paired sequences are highlighted by coloured boxes. **b**, Plot of the root to tip genetic distance against sampling time for the tree in **a**. The colours for the data points are consistent with those used for sampling locations in the phylogenies (the two African outgroup tips are not shown for clarity). The data points with black circles represent the published sequences while the data points with a target symbol represent the newly obtained sequences.



Extended Data Figure 9 | Plots of the root-to-tip genetic distance against sampling time for different genome region data sets (*gag*, *pol*, *env* and the complete genome). We used TempEst[27] to obtain exploratory regressions based on the maximum likelihood trees (Extended Data Fig. 4). Each data point represents a tip; colours are consistent with those used for sampling locations in the phylogenies. The US data points with black circles represent the new genomes dating back to 1978–1979. The data point with the target symbol represents the Patient 0 genome. In each plot, we provide the R^2 for the regression and the slope, reflecting the evolutionary rate (in substitutions per site per year).

Extended Data Table 1 | Molecular clock, phylogeographic and recombination estimates for the different data sets

Data set	TMRCA (subtype B & D)	TMRCA (subtype B)	Location probability (subtype B)	Jump time (CB to US)	TMRCA (US subtype B)	Location probability (US subtype B)	Evolutionary rate	Rate, coefficient of variation	Phi test p-value
"full genome 46", ED Fig.2 & ED Fig. 3	1953 (1946,1961)	1967 (1963,1970)	CB: > 0.99	1970 (1968,1973)	1972 (1969,1973)	US: > 0.99	0.0027 (0.0024,0.0030)	0.25 (0.20,0.31)	0.99
"full genome 38", Fig. 1 & ED Fig. 2	1955 (1946,1962)	1967 (1963,1970)	CB: 0.99	1971 (1968.1973)	1972 (1970,1974)	NY: > 0.99	0.0024 (0.0021,0.0027)	0.26 (0.20,0.32)	0.99
"gag", ED Fig. 3	1958 (1950,1964)	1969 (1964,1972)	CB: > 0.99	1972 (1969,1974)	1974 (1971,1975)	US: > 0.99	0.0023 (0.0020,0.0026)	0.23 (0.14,0.33)	0.77
"pol", ED Fig. 3	1956 (1947,1965)	1967 (1961,1972)	CB: 0.92	1970 (1966,1973)	1973 (1969,1974)	US: > 0.99	0.0015 (0.0013,0.0017)	0.29 (0.20,0.37)	0.21
"env", ED Fig. 3	1953 (1943,1962)	1968 (1964,1972)	CB: > 0.99	1970 (1966,1973)	1971 (1968,1974)	US: 0.99	0.0037 (0.0032,0.0043)	0.25 (0.16,0.34)	< 0.01
*"env, recomb. free"**	1952 (1940,1961)	1968 (1964,1972)	CB: 0.99	1970 (1966,1973)	1971 (1967,1973)	US: 0.99	0.0039 (0.0031,0.0047)	0.26 (0.18,0.35)	0.59
"env 105", ED Fig. 5	1954 (1947,1961)	1968 (1964,1971)	CB: > 0.99	1970 (1968,1972)	1971 (1969,1973)	US: > 0.99	0.0047 (0.0042,0.0052)	0.23 (0.18,0.28)	0.01
*"env 105, recomb. free"**	1955 (1947,1961)	1968 (1974,1970)	CB: > 0.99	1970 (1968.1972)	1971 (1969,1972)	US: > 0.99	0.0047 (0.0041,0.0053)	0.23 (0.18,0.28)	0.26
"env 74", ED Fig. 5	1957 (1948,1963)	1969 (1963,1971)	CB: > 0.99	1971 (1969,1973)	1972 (1969,1974)	NY: 0.97	0.0044 (0.0038,0.0050)	0.28 (0.21,0.36)	< 0.01
*"env 74, recomb. free"**	1957 (1948,1964)	1969 (1964,1972)	CB: 0.99	1971 (1968,1973)	1972 (1970,1974)	NY: 0.97	0.0046 (0.0038,0.0054)	0.31 (0.23,0.39)	0.91
"env 133", ED Fig. 6†	1952 (1944,1958)	1966 (1963,1969)	CB: 0.99	1969 (1966,1971)	1969 (1967,1971)	NY: 0.67	0.0045 (0.0041,0.0048)	0.20 (0.16,0.23)	0.76

*The recombination free ('recomb. free') data sets were obtained by deleting the minor recombinant regions from the putative recombinants identified using RDP4.
†The empirical trees from the '*env* 133' analysis were used for two different ancestral reconstructions (Extended Data Fig. 6); here we list the location estimates for the analysis that considered different US states for the late samples (Extended Data Fig. 6b).