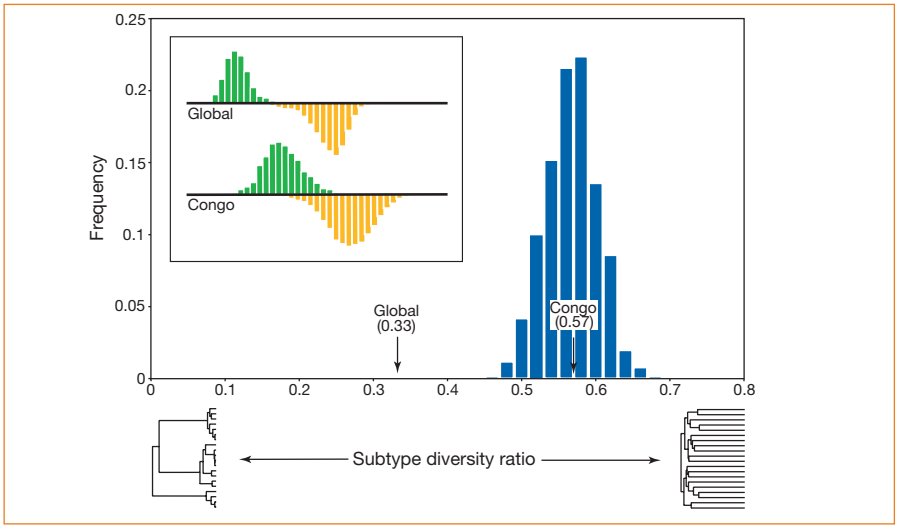

Figure 1 Maximum-likelihood phylogenies were estimated for three V3–V5 data sets: HIV-1 sequences from the Democratic Republic of Congo (423 base pairs), global isolates (426 base pairs), and Congo and global isolates combined (396 base pairs). Given a phylogeny with tips labelled according to subtype, the subtype diversity ratio (SDR) was calculated as the mean path length between tips of the same subtype divided by the mean path length between tips of different subtypes. For the phylogeny of the global isolates, 11 subtypes were allocated according to standard HIV-1 nomenclature[7]. For the Congo phylogeny, 11 subtypes were allocated so as to minimize the SDR score, using a heuristic optimization algorithm. This assignment is that which gives the maximum possible subtype structure for the Congo phylogeny. The global phylogeny gave an SDR of 0.33 and the Congo a value of 0.57. The analysis was repeated after removal of the Congo and global sequences previously identified as intersubtype recombinants[4,5]. Our analysis will only be affected if recombination breakpoints fall within the V3–V5 region, so excluding recombinants changes the SDR only marginally (0.35 for the global phylogeny; 0.58 for the Congo). SDR values were similar when Congo isolates were assigned to different numbers of subtypes (for example, 0.59 and 0.55 in the case of 8 and 14 subtypes, respectively). To assess the significance of the difference between the global and Congo SDRs, we obtained a null distribution by simulating phylogenies under an exponential growth coalescent process inferred from *env* gene sequences of subtype A (ref. 6), which is common in Africa. The frequency distribution of minimum SDR values for these simulated phylogenies is shown in blue. Inset: normalized frequency distributions of intrasubtype path lengths (above the line) and intersubtype path lengths (below the line), plotted on the same horizontal scale (0.0–0.8 substitutions per site), for the global and Congo phylogenies. See supplementary information for details of trees and phylogenetic methods.

Human immunodeficiency virus

Phylogeny and the origin of HIV-1

The origin of human immunodeficiency virus type 1 (HIV-1) is controversial. We show here that viruses obtained from the Democratic Republic of Congo in Africa have a quantitatively different phylogenetic tree structure from those sampled in other parts of the world. This indicates that the structure of HIV-1 phylogenies is the result of epidemiological processes acting within human populations alone, and is not due to multiple cross-species transmission initiated by oral polio vaccination.

According to the oral polio vaccination (OPV) hypothesis, the main (M) group of HIV-1 (the viruses responsible for the majority of global AIDS cases) emerged as a result of the vaccination of about one million people, who were largely living in the Congo from 1957–60, with an oral vaccine against polio virus that had allegedly been cultured in chimpanzee kidneys[1]. This is claimed to have enabled the transfer to humans of chimpanzee simian immunodeficiency virus, the closest relative of HIV-1.

Conversely, phylogenetic analysis of HIV-1 sequences indicates that group M originated before the vaccination campaign[2], supporting a model of 'natural transfer' from chimpanzees to humans[3]. If this timescale is correct, then the OPV theory remains a viable hypothesis of HIV-1 origins only if the subtypes of group M differentiated in chimpanzees before their transmission to humans.

It has been suggested that the distinctive structure of the global group-M tree, which has been called a 'starburst' because of the apparently simultaneous appearance of viral subtypes, is consistent with the transfer of multiple viral lineages from chimpanzees to humans[1]. To test this, we analysed partial *env* sequences (V3–V5) of 197 HIV-1 isolates sampled in 1997 from the Congo[4], a likely location for the origin of HIV-1 group M under both hypotheses.

A phylogeny comprising the Congo data, plus 223 sequences representing the global diversity of HIV-1 (including all known subtypes), reveals comparable genetic diversity in the Congo strains to that among global strains, with many Congo lineages falling basal to the origin of each subtype as currently defined by the phylogeny of global strains[5]. We tested whether the structure of the Congo phylogeny differed from that of the global HIV-1 M group by comparing the subtype diversity ratio (SDR) of the two phylogenies (Fig.1). This is defined as the ratio of the mean within-subtype pairwise distance to the mean between-subtype pairwise distance.

Rather than assigning the Congo isolates to subtypes by their phylogenetic relationship to global strains, we used a heuristic algorithm to assign subtypes such that the subtype diversity ratio was minimized. The Congo and global phylogenies differ significantly in the SDR statistic, with the former showing no more subtype structure than phylogenetic trees simulated under a model of exponential population growth[6] (Fig. 1; see supplementary information). This result is conservative because the minimum possible ratio value (representing maximum subtype structure) was used in the Congo analysis. Furthermore, although subtypes can be clearly identified in the distribution of pairwise distances for the global sequences (Fig. 1, inset), there is much less distinction between intra- and intersubtype comparisons for the Congo sequences. Hence, for any two randomly chosen Congo sequences, it is difficult to determine unambiguously whether they belong to the same or to different subtypes.

Our results indicate that the Congo and global phylogenies probably result from different epidemiological histories. As many Congo strains appear to be basal, we propose that each global subtype is the result of the chance exportation of some Congo strains to other geographical regions, thus producing an apparent starburst. Such founder effects have

been proposed to explain the phylogenetically distinct subtypes B and E of HIV-1 group M (ref. 2). The observation that many Congo strains fall basal to the global subtypes also suggests that previous phylogenetic analysis has underestimated the number of lineages that pre-date 1957–60, and hence underestimated the minimum number of cross-species transmissions necessary to reconcile the OPV hypothesis with phylogenetic data.

In conclusion, the HIV-1 sequences from the Congo are evidence that the claim of the OPV theory[1] that it is "probably the only hypothesis of origin that can readily explain the starburst phenomenon" is incorrect. Our results give us no reason to doubt that the last common ancestor of HIV-1 group M was present in a human host.

Andrew Rambaut*, David L. Robertson*, Oliver G. Pybus*, Martine Peeters†, Edward C. Holmes*

**Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK*
e-mail: edward.holmes@zoo.ox.ac.uk
†Laboratoire Retrovirus, IRD, BP 5045, 34032 Montpellier Cedex1, France

1. Hooper, E. *The River: A Journey to the Source of HIV and AIDS* (Penguin, London, 1999).
2. Korber, B. *et al. Science* **288,** 1789–1796 (2000).
3. Hahn, B., Shaw, G., de Cock, K. M. & Sharp, P. M. *Science* **287,** 607–614 (2000).
4. Vidal, N. *et al. J. Virol.* **74,** 10498–10507 (2000).
5. Los Alamos National Laboratory HIV Sequence Database, http://hiv-web.lanl.gov.
6. Pybus, O. G., Rambaut, A. & Harvey, P. H. *Genetics* **155,** 1429–1437 (2000).
7. Robertson, D. L. *et al. Science* **288,** 55–57 (2000).

Supplementary information is available on *Nature*'s website at www.nature.com or as paper copy from the London editorial office of *Nature.*