Phylogenetic Analysis of the Origin of the HIV virus

Jian Xu

Acquired immunodeficiency syndrome (AIDS) is one of the most devastating epidemics. According to the report from the Centers for Disease Control and Prevention (CDC) at June 30, 1997, there were 612,078 cases of AIDS in the United States, and 379,258 deaths. Today, AIDS is the second leading killer of people aged 25 to 44 in the United States. The cause of AIDS is human immunodeficiency virus (HIV), which belongs to the lentivirus subgroup of retrovirus. Like other retroviruses, the genes of HIV are composed of RNA, while the genes of human and most other organisms are made of DNA. Once HIV virus invades a cell, it uses an enzyme called reverse transcriptase to convert their RNA into DNA, which can be incorporated into the host cell's genes. Then most notably, the "T-helper cells", which plays a central role in the immune response, are disabled and killed during the typical course of infection. Therefore, HIV progressively destroys the body's ability to fight infections and certain cancers. People diagnosed with AIDS may get life-threatening diseases called opportunistic infections, which are caused by microorganisms that are normally not dangerous.

The first HIV identification was in 1983, and the studies of previously stored blood samples indicate that the HIV entered the U.S. population sometime in the late 1970s. The earliest HIV cases links to blood sample from central Africa around the end of 1950s. Therefore, humans are not the nature hosts of either HIV-1 or HIV-2. Instead, according to the host-dependent evolution [1], the closest relative of HIV, simian immunodeficiency virus (SIV) has been confirmed to host in primates for hundreds of thousands or even millions of years. Therefore, it is believed that HIV have entered the human population as a result of cross-species transmission. But when and how does HIV entered human population?

According to the oral polio vaccination (OPV) hypothesis by E. Hooper [2], the main (M) group of HIV-1 virus (the viruses responsible for the majority of global ADIS cases) emerged as a result of the vaccination of about one million people, who were largely living in the Congo from 1957-1960, with an oral vaccine against polio virus that had allegedly been cultured in chimpanzee kidneys. This is claimed to enable the transfer to humans of chimpanzee simian immunodeficiency virus (SIVcpz), the closest relative of HIV-1.

Lacking solid clues, the OPV hypothesis is not conclusive. In order to closely examine the origin of HIV, phylogenetic methods to are used to estimate the date of the last common ancestor of the main group of HIV-1. Many previous attempts have been hampered by the very limited HIV-1 sequence data [3,4] that antedate the discovery of HIV-

1 in 1983 [5]. Further more, the computations are so intensive that it was difficult to derive a reliable time scale for the molecular evolution of HIV-1 at the population level. At 2000, Korber et al.[6] successfully use phylogenetic calculations to estimate the date of the last common ancester of the main group of HIV-1 to be 1931 (1915-1941). Their phylogenetic methodology relies on the assumption of a molecular clock, that is, the molecular change is a linear function of time and the substitutions accumulate according to a Poisson distribution. Despite the limitations, such as the evolutionary rates may vary, the average rate of evolution of viral sequences within individuals is still relatively constant [7]. However, evolutionary models incorporate different rates of substitution between bases, different site-specific rates of evolution, and different base frequencies [8,9]. In order to consider all these factors to give a realistic estimation, they also used an adaptation of a method that relaxed the assumption of a strict molecular clock.

In a rooted phylogenetic tree, time is assigned to the distance from the root to the leaf. Therefore, how to choose the root, the branching point at which divergence from the common ancestor of the lineage first occurs, is very important in timing. Traditionally, the root is defined as the branch position of an "outgroup", where an outgroup is specified as a sequence or sequence set that is known to be external to the lineage under consideration. Thus the SIVcpz sequences are best suited as an outgroup. But Korber et. al.'s study found that, when using SIVcpz as the outgroup, the subtype can differ depending on the phylogenetic methodology, the region of the gene considered, or whether different combinations of single and multiple outgroup sequences were used. Therefore, they constructed a different outgroup, which is composed of the consensus sequences from each subtype. Compare to the SIVcpz outgroup, this construction avoids bias resulting from over-sampling of some clades relative to others, and the branch lengths are very stable. Finally, they considered a third outgroup strategy, which held the branching order obtained with the consensus-outgroup fixed, then replaced the consensus outgroup by the SIVcpzUS sequence, and then reoptimized the branch lengths. This produced very similar likelihood scores to the trees constructed with SIVcpzUS as the outgroup from the start, indicating the root positions determined by the consensus sequences were, within statistical fluctuations, compatible with those obtained by the conventional method.

After constructing a root, one can plot the total branch length against the year of sampling, base on the phylogenetic tree and assuming a uniform rate of evolution (Fig.1). From the inferred linear relation, one can project back to estimate the time associated with zero branch length, the time of the ancestral sequence. Two factors are known to affect the uncertainty of the estimation. First, the sampling time is generally only recorded to a precision of 1 year; second, and more importantly, an HIV-1 provius can be harbored, not evolving for an extended period of time in persistently infected cells, so viral DNA sampled in a given year may actually have an origin some years earlier [6]. Considering these effects, they construct a bootstrap method to provide a 95% confidence intervals (CIs) on the timing and rate-of-evolution estimates, and thus yields the date of the last common ancestor of the M group of HIV-1 to be 1931 (1915-1941).

To test the validity, they used two documented dates. The first control case was the viral sequence ZR.59, obtained from a blood sample colleted in 1959 in the Democratic Republic of the Congo. Since the tree is constructed by the samples only from the 1980-2000, the ZR.59 is decades before these data and can serve as a good control. Their calculation shows estimate the origin was 1957 (95%CI 1934-1962), well matches the documented date 1959 (Fig.2). Another control is Thai subtype E, which was first appeared in the northern part of Thailand. Related documentations suggest a single founder subtype E virus some time near 1986-1987. And their calculation estimate the date to be 1987 (95% CI 1978-1989), also close to documentation.

These analyses do not specifically address the question of when the simian progenitor lentivirus trafficked between species, since they do not provide information as to which species (human or chimpanzee) was infected at the time of the M group expansion. Nevertheless, they give satisfied estimation of when the HIV virus began to evolve from SIVcpz. Because the estimation suggests that the HIV-1 M group ancestral sequence occurred decades before the vaccination programs and that the diverse subtypes were well established by 1957. Therefore, for the OPV hypothesis to be consistent with their analyses, at least nine genetically distinct viruses would have had to differentiated in chimpanzees before their transmission to humans and then enter the human population through the vaccine, which is implausible. The fact that later PCR experiments [10] cannot detect the presence of HIV-1 related nucleic acids or chimpanzee mitochondrial DNA in suspected OPV samples, also suggests the OPV hypothesis may be false.

In conclusion, phylogenetic analyses shows that HIV-1 M group ancestral sequence occurs around 1931 (95% CI 1915-1941), and it is more likely to be a nature transmission event, rather than the human fault in OPV hypothesis.

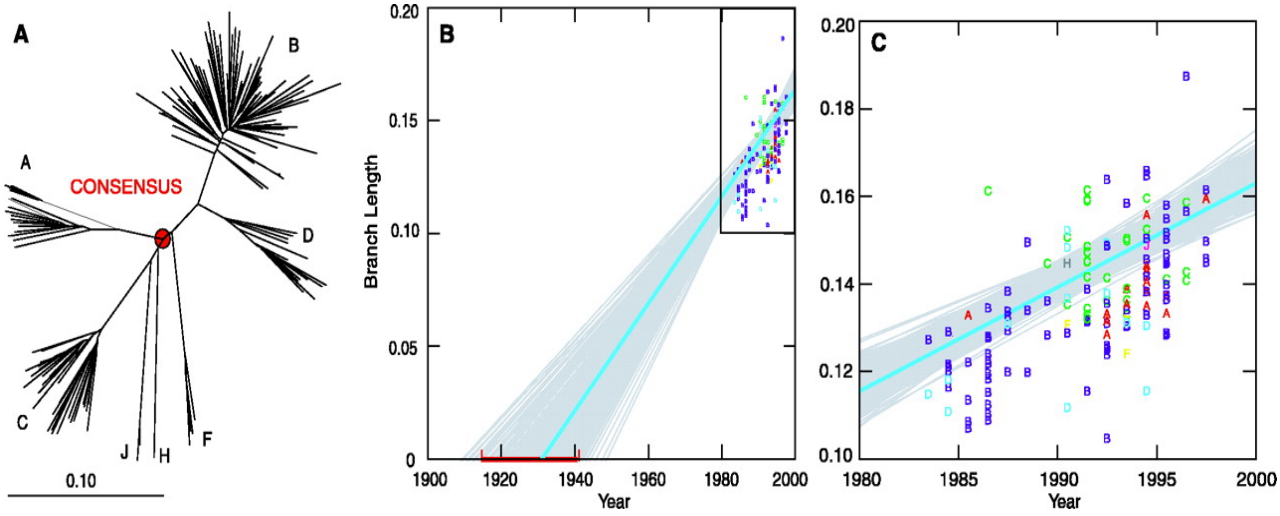

Fig.1. Estimating the last common ancestor of the HIV-1 M group on the basis of data collected over the last two decades. (**A**) The gp160 phylogenetic tree used for this calculation. (**B**) The branch lengths from each leaf to the root of the tree are plotted

against time. The subtype of the sequence is indicated by colored letters. Four hundred eighty bootstrap fits to data points were used to calculate 95% CIs, shown as a red line along the horizontal axis. (**C**) A magnified view of the boxed region in (B), showing that the points derived from different subtypes tend to be reasonably well distributed about the line, a consequence of the approximate equality of the intraclade evolutionary rates.

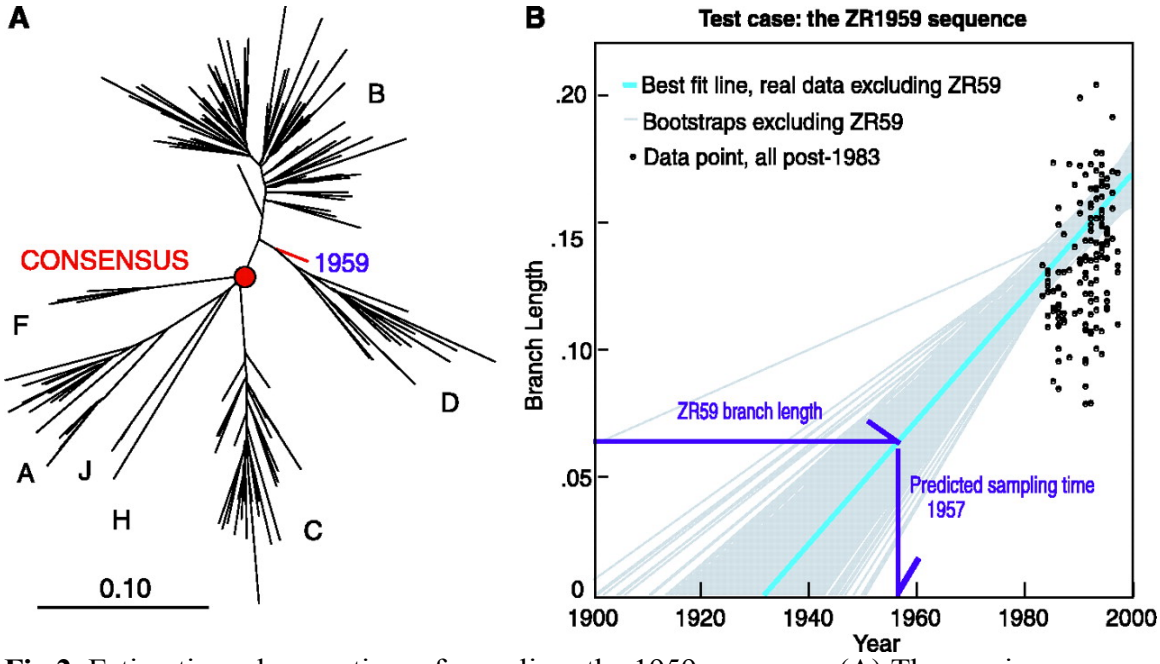

Fig.2. Estimating a known time of sampling: the 1959 sequence. (**A**) The maximum-likelihood tree. (**B**) Estimate of the time of origin of the 1959 sequence based on its branch length. A linear fit to these data points, excluding the 1959 sequence, was made with the strategy described in the text. The 480 bootstrap replicates were generated by random-with-replacement resampling of the data points to estimate the 95% CI (light gray lines). On the basis of the branch length of the 1959 sequence (horizontal purple arrow), its time of origin was estimated (vertical purple arrow) to be 1957 (95% CI 1934-62).

---

[1] B. H. Hahn, et. al., Science 287, 607 (2000).
[2] E. Hooper, *The River: A Journey to the Source of HIV and AIDS* (Penguin, London, 1999).
[3] T. Zhu et al., Nature 391, 594 (1998).
[4] T. Jonassen et al., Virology 231, 43 (1997).
[5] F. Barre-Sinoussi et al., Science 220, 865 (1983).
[6] B. Korber, et. al., Science 288, 1789 (2000).
[7] R. Shankarappa et al., J. Virol. 73, 10489 (1999).
[8] T. Leitner et. al., Proc. Natl. Acad. Sci. U.S.A. 93, 10864 (1996).
[9] D. L. Swofford, et. al., in Molecular Systematics, E. M. Hillis, C. Moritz, B. K. Mable, Eds. (Sinauer, Sunderland, MA, 1996), pp.407-514.
[10] P. Blancou, et. al., Nature 410, 1045 (2001).