Reconstruction of HIV-1 transmission chains for forensic purposes

Thomas Leitner[1] and Jan Albert[2]

[1]Department of Virology, Swedish Institute for Infectious Disease Control, Solna, Sweden
[2]Department of Clinical Virology, Karolinska Institute, Huddinge University Hospital, Huddinge/Stockholm, Sweden

Abstract

Phylogenetic reconstruction has become a standard way to investigate and reconstruct transmission histories. The accuracy and reliability of the methods becomes a particularly critical issue when the reconstruction is performed as part of a criminal investigation. Here we report on and discuss experiences gained from phylogenetic reconstructions of 27 Swedish HIV-1 transmission chains. The police or public health authorities had requested the investigations and in many cases the results were used as evidence in court. We have established a relatively simple procedure for reliable reconstruction of transmission chains based on double sampling, direct population sequencing, and maximum-likelihood tree analysis. In 19 cases we found support for an epidemiological link between the index case and the recipient(s), whereas in 4 cases we found no such support. In the remaining cases the epidemiological questions or the results were more complex. An important limitation is that it is usually impossible to determine the direction of HIV-1 transfer or formally exclude that the subjects under investigation are indirectly, rather than directly, linked, i.e. that the virus has been transmitted via a third person. Given these limitations, we conclude that phylogenetic analysis is a very powerful tool for reconstruction of HIV-1 transmission chains for forensic as well as other purposes.

Key words

Phylogenetic analysis. Molecular epidemiology. Disease control. Molecular clock. Criminal investigation

Introduction

HIV-1 infection has been, and still is, a very serious disease. Before the introduction of highly active antiretroviral therapy (HAART) the long-term mortality of HIV-1 infection was close to 100%. It is possible that HAART may reduce long-term mortality, but if so, at the expense of life-long treatment with drugs that have considerable side effects. Therefore, some countries, including Sweden, have passed laws which dictate that it is criminal to deliberately transmit HIV-1 and other severe infections. In Sweden, it is also illegal to expose someone to these severe infections without informing the potential recipient about the risk. For HIV-1, such exposure can occur during unprotected sex and sharing needles. The Swedish communicable disease act also states that contact tracing must be performed for sexually transmitted diseases (STD) and that an infected person must inform about his or her contacts. If it is discovered that an HIV-1 transmission may have been unlawful, the case can be transferred to the

Correspondence to:
Thomas Leitner
Department of Virology
Swedish Institute for Infectious Disease Control
SE-171 82 Solna, Sweden

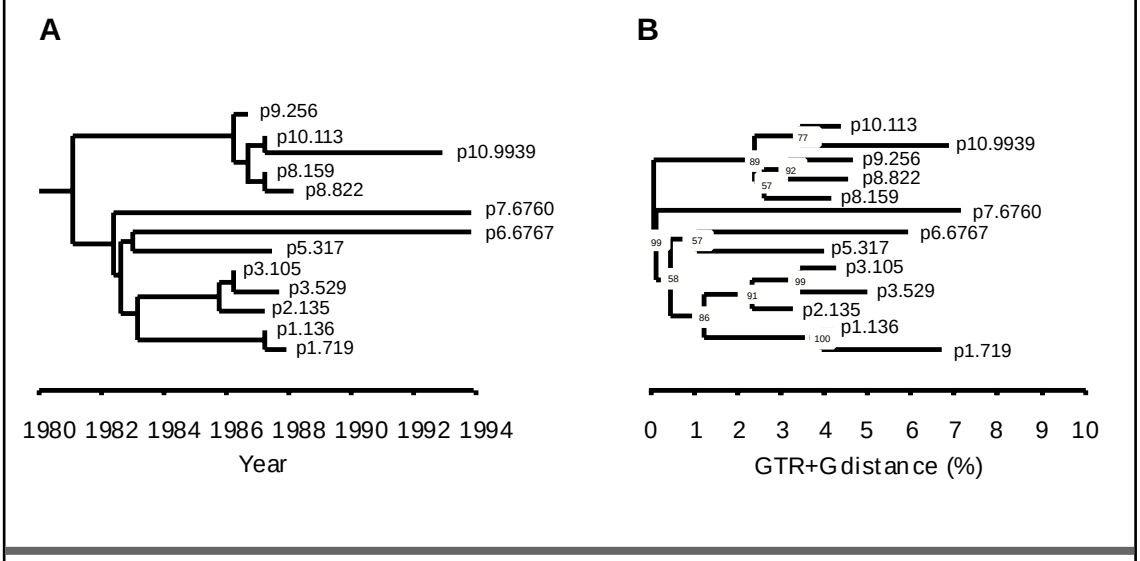

Fig. 1 (A). *The true transmission history of the Swedish Transmission Chain. Each tip in the tree represents a sample from an infected individual. From some individuals more than one sample was available, for instance, p1.136 and p1.719 are two samples from patient 1 (index case in the study). All internal nodes indicate transmission events where the donor lineage continues below the split and the recipient lineage above. All events are according to the indicated time scale.* **(B)** *A reconstructed maximum likelihood tree based on a combined env-env gene fragment. Bootstrap values are indicated for all clusters as percent out of 1000 resamplings. Branch lengths were optimized under the GTR + G substitution model. The reconstructed tree resembled the true transmission history to 94%, and all epidemiological linkages were accurate.*

public health authorities or to the police. Note that HIV-1 transmission can never be unlawful if the transmitter was not aware of his or her infection at the time of transmission. This part can often be resolved by prior HIV test results and interviews with the transmitter's doctor, etc. An important part of the forensic investigation of deliberate HIV-1 transmission is a reconstruction of the transmission chain using DNA sequencing and phylogenetic inference. In this review we report on and discuss experiences gained from phylogenetic reconstructions of 27 Swedish HIV-1 transmission chains. The police or public health authorities had requested the investigations and in many cases the results were used as evidence in court.

The Swedish transmission chain

In order to allow results from phylogenetic reconstructions of HIV-1 transmissions to be used as evidence in court, it is imperative that the conclusions drawn from the investigations are accurate and reliable. What evidence do we have that our reconstructions of HIV-1 transmission chains reflect reality? Simulation experiments have shown that phylogenetic methods can reconstruct evolutionary histories. Such simulations usually consist of three steps. First, a tree is constructed by the investigator. Second, sequences are generated using the constructed tree and a suitable model of evolution. Finally, attempts are made to reconstruct the tree from the sequences. The results from such simulations have provided important information about the strengths and weaknesses of different phylogenetic methods[1-5], but they do not fully reveal how accurate the methods are on real data. There are at least two reasons for this. First, all models used are oversimplifications of real evolution. Second, the evolutionary process is often incompletely understood and thus impossible to accurately model. For these reasons it is important to document whether real evolutionary histories can be reconstructed using phylogenetic methods. However, in most biological systems evolution is too slow to follow in a lifetime. RNA viruses are an important exception because they display very rapid genetic evolution, and thus it is possible to follow their evolution in real time.

One of the most carefully studied examples of a known phylogenetic tree involves a Swedish HIV-1 transmission chain[6]. This transmission chain consisted of 9 individuals from whom 13 samples were obtained. The index case, a Swedish male (p1) became heterosexually HIV-1 infected in Haiti in 1980. During the next five years he transmitted the virus to six women (p2, p4, p5, p7, p8, and p11). Three of these women later infected two male sexual partners (p6 and p10) and two children (p3 and p9). Detailed knowledge about the transmission history was obtained through in-depth interviews that were done by doctors or nurses with special training in contact tracing. For all subjects it was possible to define a narrow time-interval of a few months during which the transmission had occurred. Records of probable symptomatic primary HIV infection further narrowed this interval down for some individuals. DNA population sequences[7] from the *env* V3 and p17 *gag* regions of the HIV-1 genome were determined and several common tree-reconstruction methods were tested for their ability to reconstruct the true transmission history. Reassuringly, the reconstructed trees were almost identical to the true tree (Fig. 1). Thus, 94% of the quartets agreed between the reconstructed tree and the true transmission history[6].

Table 1. *Swedish transmission investigations*

Case	Year	Subjects	Route	Index[2]	Recipient[2]	Linkage	Police[3]	Subtype
1[8]	1992	2	HE	M	F	Y	Y	B
2[9]	1993	3	HE	M	2F	Y	Y	G
3[9]	1993	2	HE	M	F	Y	Y	D
4[9]	1994	4	HE/VE	M	1F1M1C	Y	N	A/D
5	1994	3	HE	M	2F	Y	Y	B
6	1994	2	HE	F	M	Y	Y	A
7	1994/1995	3	HE	M	2F	Y	Y	A
8	1994	2	HE	M	F	Y	Y	B
9	1995	2	HO	M	M	Y	N	B
10	1995	2	HO	M	M	Y	N	B
11	1995	2	HE	M	F	4	N	
12	1996	2	HE	M	F	Y	Y	A
13	1996	2	HE/IVD	M	F	Y	N	B
14	1996	2	HE	M	F	N	Y	B
15	1995	4	HE/IVD	F	2M1F	Y	N	B
16	1996	2	HE	M	F	Y	Y	B
17	1996	6	IVD			Y/N5	N	B
18	1996	2	HO	M	M	Y	Y	B
19	1997	3	HE	F	2M	Y	Y	B
20	1997/2000	2	HE	M	F	Y	Y	D
21	1998	2	HE	M	F	Y	Y	A
22	1998	1	IVD?		6	N	N	B
23	1998	2	HO	M	M	N	Y	B
24	1999	3	HE	M	2F	Y	Y	D
25	1999	17	IVD/HE			7	N	B
26	1999	1	IVD		6	Y	N	B
27	2000	2	HE	M	F	N	Y	A

[1]HE, heterosexual; HO, homosexual; VE, mother-child contact; IVD, intravenous drug use. [2]M, male; F, female; C, child. [3] Y, the case was investigated on the commission of the police or the court; N, the case did not become part of a police investigation. [4]This case was closed after failure to amplify DNA from one of the subjects. [5]Samples from two individuals were not possible to analyze. Of the remaining four, three were linked and one was unlinked. [6]This investigation was whether an IVD user was linked to a previously established group of IVD users. [7]This was an epidemiological investigation of IVD users in southern Sweden. Molecular as well as traditional methods were used. [8]Ref.no. [9]Ref. no.

The main conclusion from the studies of the Swedish transmission chain was that complex HIV-1 transmission chains could be accurately reconstructed by the methods used, i.e. direct population sequences, realistic substitution models, and maximum-likelihood calculations. However, the reconstructed tree had one error that involved a mother-to-child transmission. The reconstructed tree suggested that the child had infected the mother, which was obviously wrong. As we will discuss further under *Limitations* and in the section on genetic distance and time, this is not a surprising result because a viral phylogeny is not necessarily identical to the transmission history of the virus[8]. However, it is important to stress that every epidemiological linkage was accurately reconstructed, even if the suggested direction of the transmission in one case was incorrect.

Forensic investigation

Phylogenetic reconstruction based on genetic information in epidemiological investigations is known as molecular epidemiology. On a large scale, molecular epidemiology has been used to follow the pandemic by determining HIV-1 subtype distributions and to investigate the origin of the virus. Molecular epidemiology has been used on a smaller scale to study HIV-1 transmission chains and local epidemics. One special example of the latter is the use of molecular epidemiology for forensic purposes. The most well known example is the Florida dentist case from 1992[9]. The severe consequence of the use of phylogenetic reconstruction in a court case raised a large amount of controversy[10-16]. The molecular evidence indicated that the Florida dentist had transmitted HIV-1 to several of his patients, but the case was finally settled out-of-court[17]. The first case in which evidence based on molecular epidemiology was used in court involved a rape in Stockholm, Sweden, in 1992[18]. Based partly on molecular evidence, the rapist was found guilty of deliberate transmission of HIV-1 and imprisoned. Since then several other forensic investigations of HIV-1 transmission have been performed in Sweden (Table 1). Epidemiological investigations of possible HIV-1 transmissions have also been conducted in several other countries, but it is not clear if some of them were part of criminal investigations. Possible transmissions from health care workers to their patients have frequently been investigated and obviously this is an important issue. After the Florida dentist case, investigations of other possible dentist-to-patient transmissions have found no evidence of epidemiological linkage[19,20]. Furthermore, several investigations have evaluated possible transmissions from HIV-infected surgeons to their patients. Some studies have supported sur-

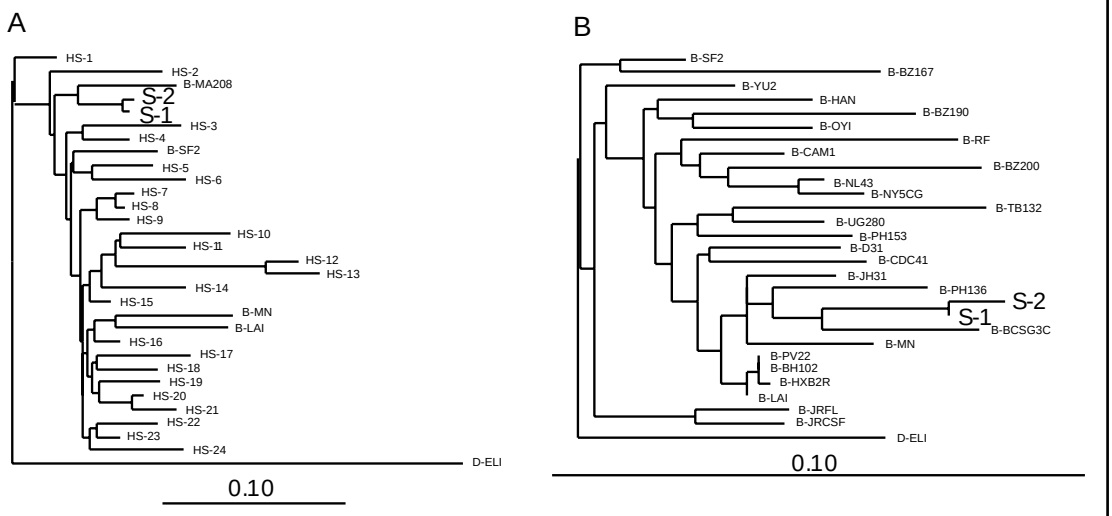

Fig. 2. *Reconstruction of transmission case 18. (A) Maximum likelihood tree based on the env V3 sequences. (B) Maximum likelihood tree based on the p17 gag sequences. S-1, subject 1, virus population sequence from the accused person. S-2, subject 2, virus population sequences from the person aggrieved. HS-#, virus population sequence from controls from the same risk group as the accused person (homosexual). Other sequences are representatives for database sequences of HIV-1 subtype B, and sequence D-ELI (subtype D) is used to root the tree.*

geon-to-patient transmission[21], while others have not[22]. Recently, a nurse-to-patient transmission in France was supported by molecular evidence[23]. In addition, other investigations of epidemics in social contexts have included outbreaks in a prison[24] and in a paediatric hospital[25,26]. A criminal investigation was recently reported from Australia where an HIV-positive man had infected at least one other person, who subsequently had infected yet another person[27].

In Sweden, where the rape case was the first investigation, we have, up to now, completed 27 investigations of HIV-1 transmission requested by the police or the public health authorities (Table 1). Heterosexual contact has been the most common mode of transmission (63%). In the majority of the heterosexual cases a man was suspected to have transmitted to one or more females, but three cases involved a female perpetrator. The remaining cases involved homosexual contact (four cases), intravenous drug use (three cases), or a combination of heterosexual contact and intravenous drug use (two cases). One case, case 25, involved a group of 17 intravenous drug (IVD) users in southern Sweden. In this case molecular epidemiology was used together with traditional epidemiology to disentangle the transmission pathways between the subjects. The investigation has revealed a complex transmission pattern and it is not fully clear if there have been one or several introductions of HIV-1 into the group[28]. At least two other cases (cases 1 and 12) have involved rape, but the reconstruction of HIV-1 transfer can obviously not prove rape. The investigations only test whether the virus variants carried by subjects involved are epidemiologically linked or not. In 19 cases a clear epidemiological linkage could be reconstructed, and in 4 cases the reconstruction suggested no direct linkage.

Examples

Case 18. This investigation involved two homosexual men. The index case was accused of having had unprotected sex with, and thereby infecting, the recipient without informing the recipient about the fact that he was HIV-1 infected. It was clear that the index case was aware of his infection, since, among other things, he had received antiretroviral treatment. Figure 2 shows the reconstructed evolutionary relationship of the *env* V3 and p17 *gag* sequence fragments. The virus was of subtype B, as expected in this risk group in Sweden. A control group consisting of viral sequences from the same risk group and geographical area was used together with sequences from the international HIV database in Los Alamos, NM[29]. As seen in figure 2, the viruses carried by the two subjects were clearly linked, a finding that was supported by a bootstrap value of 97% for the V3 analysis. Based on the molecular as well as other evidence, the accused man was found guilty.

Case 27. This investigation involved a man and a woman where the man was accused of having infected the woman through unprotected sex. Also, in this case it was clear that the man was aware of his HIV-1 infection before they had met. The female claimed that the accused man had infected her, while he stated that she had told him that she thought that she had been infected at some prior time point. Both individuals were found to carry subtype A virus. The viruses were, however, not closely related to each other. Figure 3 shows the env V3 tree in which the sequences from the man and the female are separated by several local and database control sequences. The man, who was held in custody during the investigation, was immediately released and the charges were dropped when the results from the molecular investigation were presented to the police and the district attorney.

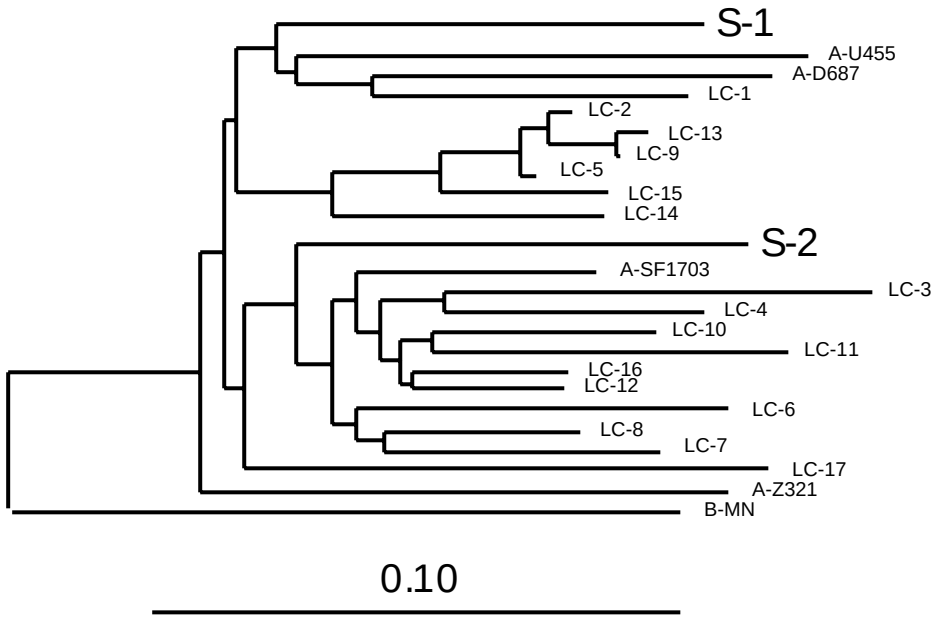

Fig. 3. *Reconstruction of transmission case 27. Maximum likelihood tree based on the env V3 sequences. S-1, subject 1, virus population sequence from the accused person. S-2, subject 2, virus population sequence from the person aggrieved. LC-#, virus population sequences from controls from the same risk group (subtype A infected person sampled in Sweden). Other sequences are representatives for database sequences of HIV-1 subtype A, and sequence B-MN (subtype B) is used to root the tree.*

Investigation procedure

In the following sections we will present and discuss the methods that were used in the forensic investigations of HIV-1 transmissions in Sweden. As outlined in the flowchart (Table 1), the investigation consists of several steps that are consecutively dependent on each other. We will look in detail at each step. Naturally, there are many other valid ways to perform these investigations. Nevertheless, we feel that our investigation method provides an attractive balance between accuracy and reliability on the one hand and cost in time and labour on the other. It is an absolute requirement that the methods used are accurate and reliable, but it is also important that the investigation can be completed within a reasonable time and at a reasonable cost. Often the accused person has been held in custody while we have performed the investigation, and therefore it has been our goal to complete the investigation in less than four weeks.

Samples

Two samples of whole blood, drawn on different days, are obtained from each subject. The two samples are treated individually throughout the complete investigation to minimise the risk of sample mix-up and contamination. If the analysis does not show the same result on both samples it should be considered invalid. The identity of each individual is protected by laboratory codes, and all information on the samples is kept in an access-restricted database and fireproof safes. On a few occasions we have, in addition, included earlier samples from some of the individuals involved in the investigation[18]. Such prior samples, however, must be used with caution since they probably were collected for other than forensic reasons. This brings up ethical considerations as to which stored samples can and cannot be used, and consent from the subject may have to be acquired.

In most investigations we have used sequenced proviral DNA from peripheral blood mononuclear

Table 2. *Flowchart for forensie HIV-1 investigation*

1. **Two separate samples are drawn at separate days from each subject.**
2. **Viral load is determined and samples are made sure to contain at least ten copies.**
3. **Two gene fragments, env V3 and p17 gag, are sequenced.**
4. **The subtype (or group) of the sequences is determined by neighbour joining.**
5. **A final tree is reconstructed using maximum-likelihood with local controls and database sequences matched to subtype and risk group.**
6. **The final tree topology is tested with bootstrap analysis.**
7. **A statement is written where the result is presented together the limitations of the method. Oral testimony is given if required by the court.**

Step 2. At least ten copies should be present in the PCR tube used for sequencing. Steps 4, 5, and 6 are performed on both V3 and p17 sequence data. Steps 5 and 6 may also be performed on a combined V3+p17 fragment. See text for further details.

Table 3. *PCR and sequencing primers for forensic investigations*

Fragment	Primer	Use[1]	Position[2]	Sequence[3]
env V3	JA167	outer sense	6876	TAT CYT TTG AGC CAA TTC CYA TAC A
	JA168	inner sense	6988	ACA ATG YAC ACA TGG AAT TAR GCC A
	JA169	inner antisense	7373	AGA AAA ATT CYC CTC YAC AAT TAA A
	JA170	outer antisense	7388	GTG ATG TAT TRC ART AGA AAA ATT C
p17 gag	JA152	outer sense	624	ATC TCT AGC AGT GGC GCC CGA ACA G
	JA153	inner sense	679	CTC TCG ACG CAG GAC TCG GCT TGC T
	JA154	inner antisense	1237	CCC ATG CAT TCA AAG TTC TAG GTG A
	JA155	outer antisense	1300	CTG ATA ATG CTG AAA ACA TGG GTA T

[1]Inner primers also serve as sequencing primers.
[2]Position according to HIV-1 strain MN.
[3]Degenerate nucleotides are indicated by the IUPAC-IUB codes.

cells (PBMC), but in some cases RNA from plasma virions has been used instead. The PBMC virus population often is slightly more heterogeneous, but our studies on the Swedish transmission chain show that both DNA and RNA can be used for accurate reconstruction of transmission chains. However, DNA may be the only option if a subject is on successful HAART. Before sequencing, the number of viral DNA copies is determined in each sample. This is done to ensure that the virus population is correctly sampled. Hence, it is important that each direct population sequence is based on a minimum of ten DNA (or RNA) molecules (see further below under *Genetic information*). If the viral copy numbers are low, we run several PCR reactions from the original sample and pool the amplicons to ensure that the *in vivo* virus population is adequately sampled. The DNA load in the PCR tubes is determined by the limiting dilution procedure[30]. Briefly, each sample is diluted in sequential steps and several parallel PCR reactions are run on each dilution. The number of viral DNA copies per PCR reaction is then calculated according to the Poisson distribution formula[30]. In this context it is important to stress that commercial or in-house HIV-1 quantification methods based on competitive PCR are often not adequate for this specific purpose. These assays are designed to measure virus load in vivo, while we want to know how many viral DNA or RNA molecules have been successfully extracted and amplified in each PCR reaction tube.

Genetic information

One important conclusion from the studies of the Swedish Transmission Chain was that accuracy of the reconstructed tree topology (branching order) was more dependent on the amount of genetic information than the phylogenetic reconstruction method[6]. As discussed further below, the choice of phylogenetic reconstruction method is not irrelevant, however. Some methods clearly produced inferior topologies and, even more importantly, branch lengths and molecular clock estimates are critically dependent on the choice of evolutionary model[31]. It is important to stress that the amount of genetic information is not the same as the sequence length. Thus, the more variable *env* V3 fragment contained more genetic information than the longer p17 *gag* fragment. The best results, however, were obtained with a combined *env* V3 + p17 *gag* fragment. In all investigations performed after case 1, where *a pol* fragment was used together with the p17 *gag* fragment, *env* V3 and p17 *gag* were used. The primer sequences are given in Table 3. This choice was based on our studies of the Swedish Transmission Chain that show that complex HIV-1 transmission histories can be accurately reconstructed using these relatively short sequence fragments. Furthermore, the same studies have identified models of evolution and tree reconstruction methods that allow accurate reconstructions (see further below).

We have used direct population sequencing[7] rather than sequencing of individual clones. There are several reasons for this choice. First, we avoid the problem of PCR-induced substitutions, because each direct sequence is generated from a minimum of 10 individual HIV-1 dsDNA templates. Even if the Taq polymerase misincorporates a nucleotide in the first PCR cycle, a specific error will only be present in 1 out of ≥ 40 DNA strands (≥ 20 dsDNA molecules) present after the first PCR cycle (i.e. < 2.5%). The direct population sequencing method has a detection limit of 10-20% for polymorphic nucleotide positions. Thus, PCR-induced errors will not affect the deduced nucleotide sequence. Second, by direct population sequencing many more samples can be analysed in a given time. One sequence is sufficient to describe the population with a resolution of 10-20% for polymorphic positions, while 5-10 clonal sequences would normally be required for the same resolution. As mentioned previously, time is important when a person may be held in custody until the investigation is completed. Third, direct population sequencing involves less laboratory manipulations on the virus population. Traditional cloning procedures may introduce unwanted selection, thereby biasing the population structure. Finally, the tree reconstruction will involve 5-10 times fewer sequences. This makes it more feasible to use better, but computer-intense, models of evolution and reconstruction methods. One drawback of direct population sequencing is that the

linkage between the nucleotides in mixed positions is usually impossible to decipher. This is less of a problem in studies on transmission pathways between individuals than in studies on virus population dynamics within a single individual.

There are two ways of describing a population sequence, either by ambiguity codes or by a majority rule. In our investigations we have used the IUPAC-IUB ambiguity codes, where R, for instance, indicates the nucleotides A and G. Originally R was defined as A *or* G (i.e. an unresolved nucleotide position). However, we use ambiguity codes to indicate polymorphic nucleotide positions (i.e. when some virus variants have an A and others have a G in a certain position). Therefore, a population sequence describes many sequences at the same time and, thus, the ambiguity codes should be interpreted as multi-state characters rather than ambiguities. In contrast, a majority rule population sequence contains the most abundant nucleotide at every position. Note that this still is a population sequence although its sequence suggests a clone. There is a problem with this, namely that the sequence may not exist as a real variant in the population. The use of multi-state characters avoids this problem since it includes all variants down to a certain resolution, but unfortunately some phylogenetic software does not accept sequences that contain ambiguity codes.

Phylogenetic reconstruction

Inference of evolutionary history from DNA sequences is a very powerful tool. Many methods of tree reconstruction have been proposed and used in biological contexts. The efficacy and accuracy of phylogenetic reconstruction methods have been studied in many different ways including computer simulations[1-5], experimental phylogenetics[32], and real evolutionary histories[6]. For HIV-1 and many other systems, the best reconstructions have been derived using maximum-likelihood (ML) methods[6,31,33]. The ML method has the advantage of allowing the investigator to choose an explicit model of evolution but, on the other hand, requires that a reasonably realistic model be chosen. The use of an inadequate model may give the wrong answer. HIV-1 evolution has been carefully investigated[31] and shown to be best described for the *env* V3 and p17 *gag* fragments by the general time-reversible (GTR or REV) model including Gamma (G) distributed substitution rates among sites (GTR + G)[31,34]. Another important advantage of the ML method is that it has a sound statistical framework. This allows hypothesis testing and estimation of confidence intervals for all inferred parameters. For instance, it is easy to test the statistical support for alternative transmission pathways. Using ML hypothesis testing, Holmes *et al*[22]*.* investigated whether a patient was more likely to have been infected by an HIV-1 positive surgeon or through an HIV-1 contaminated batch of blood. It was found that the tree linking the patient to the blood batch was significantly better ($p < 0.01$) than the tree linking the patient to the surgeon.

As mentioned above, given sufficient and consistent phylogenetic information, the ability of the ML method to find the correct topology is relatively insensitive to violations of the substitution model[6]. This is fortunate because many phylogenetic reconstruction programs do not include the GTR + G model[35], while others do not accept, or correctly handle, multi-state characters[36]. Phylogenetic software is constantly being improved and refined, and thus our ability to accurately reconstruct HIV-1 transmission histories also improves over time. Currently, we carry out the reconstruction of a transmission chain in several steps,

1) The sequences from the subjects are aligned to a subtype reference set[37]. The genetic subtype of the virus carried by the subjects is determined by a quick neighbour-joining (NJ) tree.

2) Based on step 1, a set of control sequences from the same subtype is selected. After realignment, a ML tree is calculated using the F84 substitution model. This tree represents the final tree if the result is clear, i.e. if the sequences from the subjects under study are clearly not linked, or if they are clearly directly linked and separated only by very short distances.

3) If the ML tree is not completely clear, the branch lengths are recalculated under the GTR + G model. Steps 1-3 are performed on both the *env* V3 and p17 *gag* sequences.

4) A combined V3 + p17 tree is calculated using ML and either the F84 or the GTR + G model.

5) Bootstrap values are calculated under the F84 model using NJ on 1000 permutations of the alignment in step 2.

All ML trees are calculated on a parallel computer using a parallelised version of the program DNAML[35,38,39,] and the GTR + G corrections are calculated using BASEML[36]. NJ trees and bootstrap analysis are calculated using the PHYLIP package[35].

Controls

The HIV-1 genome is amazingly plastic, and this is why we are able to perform reconstructions of transmission chains at all. We cannot, however, say much about the possible epidemiological linkage between two sequences without control sequences. Phylogenetic reconstruction of a tree with only two sequences is trivial and uninformative. Thus, we need to include control or background material. As already mentioned above, HIV-1 is divided into subtypes, or more correctly, group M (main) of HIV-1 is divided into subtypes[40]. If we investigate the possible linkage between two sequences of subtype A, and used control sequences from subtype B, then obviously our two subtype A sequences would be more closely related to each other than to any of the controls, irrespective of whether they were linked or not. Because of the poor choice of control sequences, we have essentially performed the trivial phylogenetic

reconstruction of two sequences. Scientists may arrive at completely wrong conclusions if the control sequences are not correctly chosen. Consequently, relevant control sequences need to be carefully selected, but unfortunately such sequences are not always available.

In our analyses we always try to use local control sequences of the same subtype as the virus carried by the subjects under study. If possible, the local control sequences are also matched for risk group and geographical origin. For each case we have investigated, our set of local control sequences has increased. In addition, other studies of circulating HIV-1 genotypes in Sweden have provided a rich set of additional control sequences[41]. Besides local controls, we always include all available sequences from the same subtype found in the HIV database[42]. When a large amount of sequences are available, like for subtype B in the *env* V3 region (currently >1900 sequences)[44], a BLAST pre-screening is advisable. Finally, if the set of control group sequences still is very large, we first perform an analysis of all sequences and then recalculate a representative tree that includes all local controls plus the most relevant database sequences (Figs. 2 & 3). In addition to serving as epidemiological controls, sequences from local controls as well as sequences of commonly used laboratory strains serve as controls of laboratory contamination[43].

Genetic distance in relation to time since transmission

It has been shown that HIV-1 evolution involves a relatively constant rate of nucleotide substitutions[45]. The number of nucleotide substitutions per unit of time is referred to as the molecular clock[46,47]. In many other biological systems researchers have found that molecular clocks often are over-dispersed[48-50], i.e. there is more variation around the expected rate than predicted by the Poisson error. One reason for over-dispersion is that different lineages in a tree may evolve at different rates, i.e. local clocks may diverge significantly from the overall clock[51]. In HIV, it has been reported that the rate of evolution may vary in different stages of the disease[52,53]. Nevertheless, several studies indicate that the rate of evolution is relatively constant, at least when averaged out over several individuals and longer time periods[45,54,55]. In most forensic investigations of HIV-1 transmissions the samples are collected within one or few years after the time of the alleged transmission. It is interesting to discuss if it is possible to estimate the time of transmission based on the genetic distance between the sequences. At least two aspects of this problem require comments.

First, it is obvious that detailed knowledge about the rate and dispersion of the molecular clock is required to estimate the time for transmission. It is important to stress that both of these parameters have to be determined for the sequence fragment under investigation, as well as for the specific reconstruction method and evolutionary model that is used. Using the Swedish transmission chain, we have determined the rate and dispersion of the molecular clocks for the p17 *gag* and *env* V3 fragments as determined by ML and the GTR + G model[45]. For most other sequence fragments and phylogenetic methods, the knowledge about the characteristics of the molecular clock is much more incomplete. Given that the rate and dispersion of the molecular clock is known, it is relatively simple to estimate the time for transmission and also provide confidence limits on this estimate. From these estimates we can determine if the molecular data fits with the proposed time for transmission. This may sound very attractive, but unfortunately the method is severely hampered by the fact that the confidence limits usually are quite wide, especially if the time since transmission is short. The main reason for this, and the second aspect to consider is that we reconstruct the viral evolutionary history, not the transmission history. The viral evolutionary history is closely dependent on the transmission history, but, because the virus is a heterogeneous population, the genealogy of the transmitted variant may not precisely describe the transmission history[8,45]. Intuitively, it is easy to understand that a virus variant that is transmitted must already have existed for a while in the donor before it is transmitted to the recipient. We have named this time interval «the pre-transmission interval»[8,45] and more precisely it describes the (average) distance in time from the most recent common ancestor of the virus in the donor and the recipient to the time point of transmission. The pre-transmission interval will depend on the effective population size of the virus in the donor. In addition, it will be influenced by the transmission bottleneck, regardless of whether this selection process is purely stochastic or if it also has elements of positive or purifying selection. The pre-transmission interval may cause as much genetic distance between two samples as a few years of intra-patient evolution. Faster evolving gene fragments will be less affected by the pre-transmission interval because of the steeper slope of the substitution rate[45].

In conclusion, estimation of time for transmission using the molecular clock has limited use in most forensic investigations because the confidence limits are wide. Furthermore, the molecular clock is poorly characterised for most HIV-1 sequence fragments. Nevertheless, it is often useful to be able to state that a transmission was unlikely to have occurred more than 4-5 years ago (something which is often feasible). The method also has one important advantage: namely that the measure is not dependent on the choice of controls. Thus, it can be the only valid method if no suitable control sequences are available.

Limitations

The reconstruction of a transmission chain for forensic use requires the highest standards of laboratory and analysis procedures. Good laboratory practice, including detailed records of sample handling, is important to control sample mix-ups and

contamination risks. Limitations due to use of an inappropriate control group have already been discussed above. When it comes to the analysis part, the reconstruction method is better suited to provide evidence against than for linkage. This is because we cannot formally exclude the possibility that there exists an un-sampled link between the investigated sequences. In other words, if we find that A is epidemiologically linked to B, we cannot exclude that there may exist a third subject, C, who was infected by A and then infected B. It is also not possible to exclude the possibility that C, who has not been sampled, infected both A and B. If, on the other hand, the sequences from the index case and the recipient are separated by one or several control sequences, we can with much higher confidence state that it is highly unlikely that the proposed transmission has occurred. Note that if more than two persons are involved in the transmission chain the situation becomes more complex. Some investigators have, therefore, suggested that each transmission pair should be investigated separately[14].

From a tree reconstruction alone it is usually not possible to determine the direction of a transmission. Said differently, we cannot with certainty say whether A infected B or B infected A. This is surely the case if we are not able to root the tree, but may also be the case for a rooted tree. As previously stated, this is due to the fact that we reconstruct the evolutionary history of the virus, not the transmission history[8,45]. It may very well be that a variant that has not produced progeny in A was transmitted to B. In a later reconstruction it would then seem as if the population in B was ancestral to the population in A which would be correct, but it would mislead us about the direction of the transmission. This is one possible explanation for the only «mistake» in the Swedish Transmission Chain, where it seemed as if a child had infected its mother[6,8]. Limited or non-representative sampling can induce the same pattern. Note, however, that the linkage between two subjects is not disrupted by this limitation; it is only the direction of transfer that may be obscured. In some special cases it may be possible to determine the direction. If the sequences of B are significantly more closely related to some subset of sequences in A than many sequences of A are to each other, the transmission was surely from A to B. This investigation would, however, require detailed analysis of multiple clones from each subject. Also, we warn against using the degree of heterogeneity (i.e. intra-sample distances or number of polymorphic nucleotides as indicators of time since transmission). Intra-sample distances tend to vary during disease progression[54] and are also very sensitive to sampling errors. In our forensic cases, information about the probable direction of the virus transmission has often been obtained through other evidence, i.e. prior HIV antibody tests.

A third limitation of the tree reconstructions is that the resolution may sometimes be too limited. In situations where there is explosive spread of HIV-1 from a point source, such as when hundreds of intravenous drug users became infected within a few months in Kaliningrad, we will be unable to reconstruct the exact transmission pathways.

In the written statement of the investigation we always notify about the limitations of the methods. It is specifically stated that we cannot formally exclude the existence of «person C» by tree analysis. Thus, if the sequences from the subjects under study are linked we write that the result from the investigation agrees very well with the possibility that the proposed index case infected the proposed recipient.

Perspective

Reconstruction of transmission histories in criminal investigations will, unfortunately, not be obsolete in the future. Information campaigns and better social care may reduce their number, but we have not yet seen such trends. It is, once again, very important that the investigation is made according to the best method available since the judicial system relies strongly on fair and correct information from expert witnesses. We must, therefore, continuously follow the scientific field and improve our methods. In Sweden, and in other countries as well, studies are made of the epidemiological situation. For example, in many European countries immigration has brought new subtypes into the circulating HIV population[41,56-59,60]. It is naturally very important that the local controls reflect the epidemiological situation of the region where the alleged crime took place. We must, therefore, be up to date with the current epidemiological situation and keep representative sets of reference sequences at hand. The ideal situation would be to have HIV sequence material from all new cases, but in some regions this might not be practical for various reasons. It is also important to follow development in the methodology for phylogenetic reconstruction. Improvements of established methods and introductions of new techniques are frequently reported. Indeed, phylogenetic reconstruction is a scientific field in itself, with several specialised journals reporting on the progress.

In conclusion, we have shown that HIV-1 transmission chains can be accurately reconstructed by phylogenetic analysis of viral sequences. Sample handling, sequencing techniques, gene regions, and phylogenetic reconstruction methods are all important for an accurate result. We have presented an investigation method, which we feel provides an attractive balance between relative simplicity and high accuracy. However, we need to follow epidemiological changes as well as developments in laboratory and computer analysis. Furthermore, scientific studies of virus evolution will contribute to a better understanding of genetic variation and transmission. Thus, in the future the method of investigation is likely to become even better.

Acknowledgements

This study was supported by the Swedish Medical Research Council.

References

1. Sourdis J, Nei M. Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree. Mol Biol Evol 1988; 5: 298-311.
2. Saitou N, Imanishi T. Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. Mol Biol Evol 1989; 6: 514-25.
3. Nei M. Relative efficiencies of different tree-making methods for molecular data. Phylogenetic analysis of DNA sequences. MM Miyamoto and J Cracraft. 1991. New York, Oxford Univ. Press: 90-128.
4. Hillis D, Huelsenbeck J, Cunningham C. Application and accuracy of molecular phylogenies. Science 1994; 264: 671-7.
5. Hillis D, Moritz C, Mable B. Molecular Systematics. 1996. Sunderland, MA, Sinauer.
6. Leitner T, Escanilla D, Franzén C, Uhlén M, Albert J. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proc Natl Acad Sci USA 1996; 93: 10864-9.
7. Leitner T, Halapi E, Scarlatti G *et al*. Analysis of heterogeneous viral populations by direct DNA sequencing. BioTechniques 1993; 15: 120-6.
8. Leitner T, Fitch W. The phylogenetics of known transmission histories. Molecular Evolution of HIV. KA Crandall. 1999. Baltimore, MD, Johns Hopkins.
9. Ou C, Ciesielski C, Myers G *et al*. Molecular epidemiology of HIV transmission in a dental practice. Science 1992; 256: 1165-71.
10. Abele L, DeBry R. Florida dentist case: research affiliation and ethics. Science 1992; 255: 903.
11. Smith T, Waterman M. The continuing case of the Florida dentist. Science 1992; 256: 1155-6.
12. DeBry R, Abele L, Welss S *et al*. Dental HIV transmission? Nature 1993; 361: 691.
13. Holmes E, Leigh Brown A, Simmonds P. Sequence data as evidence. Nature 1993; 364: 766.
14. Hillis D, Huelsenbeck J. Support for dental HIV transmission. Nature 1994; 369: 24-5.
15. Crandall K. Intraspecific phylogenetics: support for dental transmission of human immunodeficiency virus. J Virol 1995; 69: 2351-6.
16. Korber B, Learn G, Mullins J, Hahn B, Wolinsky S. Protecting HIV databases. Nature 1995; 378: 242-4.
17. Anonymous. No trial to come in Florida dentist case. Science 1992; 255: 787.
18. Albert J, Wahlberg J, Leitner T, Escanilla D, Uhlén M. Analysis of a rape case by direct sequencing of the HIV-1 pol and gag genes. J Virol 1994; 68: 5918-24.
19. Dickinson G, Morhart R, Klimas N *et al*. Absence of HIV transmission from an infected dentist to his patients. An epidemiologic and DNA sequence analysis. JAMA 1993; 269: 1802-6.
20. Jaffe H, McCurdy J, Kalish M *et al*. Lack of HIV transmission in the practice of a dentist with AIDS. Ann Intern Med 1994; 121: 855-9.
21. Blanchard A, Ferris S, Chamaret S, Guetard D, Montagnier L. Molecular evidence for nosocomial transmission of human immunodeficiency virus from a surgeon to one of his patients. J Virol 1998; 72: 4537-40.
22. Holmes E, Zhang L, Simmonds P, Rogers A, Brown A. Molecular investigation of human immunodeficiency virus (HIV) infection in a patient of an HIV-infected surgeon. J Infect Dis 1993; 167: 1411-4.
23. Goujon C, Schneider V, Grofti J *et al*. Phylogenetic Analyses Indicate an Atypical Nurse-to-Patient Transmission of Human Immunodeficiency Virus Type 1. J Virol 2000; 74: 2525-32.
24. Yirrel D, Robertson P, Goldberg D *et al*. Molecular investigation into outbreak of HIV in a Scottish prison. Brit Med J 1997; 314: 1446-50.
25. Bobkov A, Cheingsong-Popov R, Garaev M *et al*. Identification of an *env* G subtype and heterogeneity of HIV-1 strains in the Russian Federation and Belarus. AIDS 1994; 8: 1649-55.
26. Bobkov A, Cheingsong-Popov R, Garaev M, Weber J. Glycoprotein 120 polymorphism in an HIV type 1 epidemic originating from a point source: nucleotide sequence analysis of variants with conserved V3 loop sequences. AIDS Res Hum Retroviruses 1996; 12: 251-3.
27. Birch C, McCaw R, Bulach D *et al*. Molecular analysis of human immunodeficiency virus strains associated with a case of criminal transmission of the virus. J Infect Dis 2000; 182: 941-4.
28. Leitner T, Hansson H. Unpublished results. 2000.
29. Myers G, Korber B, Hahn B *et al*. Human retroviruses and AIDS 1995: a compilation and analysis of nucleic acid and amino acid sequences. 1996. Los Alamos, NM, Los Alamos National Laboratory.
30. Brinchman J, Albert J, Vartdal F. Few infected CD4+ T cells but a high proportion of replication-competent provirus copies in asymptomatic human immunodeficiency virus type 1 infection. J Virol 1991; 65: 2019-23.
31. Leitner T, Kumar S, Albert J. Tempo and mode of nucleotide substitutions in gag and env gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. J Virol 1997; 71: 4761-70 (see also correction 1998: 72; 2565).
32. Hillis D, Bull J, White M, Badgett M, Molineux I. Experimental phylogenetics: generation of a known phylogeny. Science 1992; 255: 589-92.
33. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 1981; 17: 368-76.
34. Yang Z. Estimating the pattern of nucleotide substitution. J Mol Evol 1994; 39: 105-11.
35. Felsenstein J. PHYLIP: Phylogeny Inference Package. Seattle, WA, University of Washington. 1993.
36. Yang Z. PAML: Phylogenetic Analysis by Maximum Likelihood. The Pennsylvania State University, Institute of Molecular Evolutionary Genetics. 1995.
37. Carr J, Foley B, Leitner T *et al*. Reference sequences representing the principal genetic diversity of HIV-1 in the pandemic. Human retroviruses and AIDS 1998: a compilation and analysis of nucleic acid and amino acid sequences. B Korber and *et al*. 1999. Los Alamos, NM, Los Alamos National Laboratory: III19-III26.
38. Trelles O, Ceron C, Wang HC, Dopazo J, Carazo J. New phylogenetic venues opened by a novel implementation of the DNAml algorithm. Bioinformatics 1998; 14: 544-5.
39. Leitner T, Buresund R, Holmberg A. Optimized and parallel maximum likelihood inference of phylogenetic trees. 1999; unpublished work.
40. Robertson D, Anderson J, Bradac J *et al*. HIV-1 nomenclature proposal. Science 2000; 288: 55.
41. Alaeus A, Leitner T, Lidman K, Albert J. Most genetic subtypes of HIV-1 have entered Sweden. AIDS 1997; 11: 199-202.
42. Kuiken C, Foley B, Hahn B *et al*., (eds). Human Retroviruses and AIDS 1999: A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences. 2000. Los Alamos, NM, Theoretical Biology and Biophysics Group, Los Alamos National Laboratory.
43. Learn G, Korber B, Foley B *et al*. Maintaining the integrity of human immunodeficiency virus sequence databases. J Virol 1996; 70: 5720-30.
44. Gaschen B, Korber B, Foley B. Global variation in the HIV-1 V3 region. Human Retroviruses and AIDS 1999: A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences. CL Kuiken, B Foley, B Hahn *et al*. 2000. Los Alamos, NM, Theoretical Biology and Biophysics Group, Los Alamos National Laboratory: Part VII. 594-789.
45. Leitner T, Albert J. The molecular clock of HIV-1 unveiled through analysis of a known transmission history. Proc Natl Acad Sci USA 1999; 96: 10752-7.
46. Zuckerkandl E, Pauling L. Evolutionary divergence and convergence in proteins. Evolving genes and proteins. V Bryson and HJ Vogel. New York: Academic Press: 1965; 97-166.
47. Kimura M. The neutral theory of molecular evolution. Cambridge: 1983. Cambridge University Press.
48. Otha T, Kimura M. On the constancy of the evolutionary rate of cistrons. J Mol Evol 1971; 1: 18-25.
49. Langley C, Fitch W. An examination of the constancy of the rate of molecular evolution. J Mol Evol 1974; 3: 161-77.
50. Takahata N. On the overdispersed molecular clock. Genetics 1987; 116: 169-79.

51. Gillespie JH. The molecular clock may be an episodic clock. Proc Natl Acad Sci USA 1984; 81: 8009-13.
52. Wolinsky S, Korber B, Neumann A *et al*. Adaptive evolution of human immunodeficiency virus type1 during the natural course of infection. Science 1996; 272: 537-42.
53. Halapi E, Leitner T, Jansson M *et al*. Correlation between HIV sequence evolution, specific immune response and clinical outcome in vertically infected infants. AIDS 1997; 11: 1709-17.
54. Shankarappa R, Margolick J, Gange S *et al*. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J Virol 1999; 73: 10489-502.
55. Korber B, Muldoon M, Theiler J *et al*. Timing the ancestor of the HIV-1 pandemic strains. Science 2000; 288: 1789-96.
56. Salminen M, Nykänen A, Brummer-Korvenkontio H *et al*. Molecular epidemiology of HIV-1 based on phylogenetic analysis of in vivo gag p7/p9 direct sequences. Virology 1993; 195: 185-94.
57. Arnold C, Barlow K, Parry J, Clewley J. At least five HIV-1 subtypes (A, B, C, D, A/E) occur in England. AIDS Res Hum Retroviruses 1995; 11: 427-9.
58. Leitner T, Escanilla D, Marquina S *et al*. Biological and molecular characterization of subtype D, G and A/D recombinant HIV-1 transmissions in Sweden. Virology 1995; 209: 136-46.
59. Simon F, Loussert-Ajaka I, Damond F *et al*. HIV type 1 diversity in northern Paris, France. AIDS Res Hum Retroviruses 1996; 12: 1427-33.
60. Lasky M, Perret J, Peeters M *et al*. Presence of multiple non-B subtypes and divergent subtype B strains of HIV-1 in individuals infected after overseas deployment. AIDS 1997; 11: 43-51.