The Microbial Forensic Use of HIV Sequences

Gerald H. Learn and James I. Mullins
Department of Microbiology, University of Washington School of Medicine, Box 358070, Seattle, WA 98195

Introduction

Analyses of microbial genetic sequences and phylogenies have become increasingly important in the tracking and investigation of events leading to infections. It has recently been proposed that these activities be termed microbial forensics [1,2]. Studies of HIV have frequently assumed a pivotal role in this developing discipline. Since the evolutionary rate of HIV is so great that it is unlikely to isolate viruses with identical genomes, their sequences provide an exquisite tool to investigate their evolution. The proper use of phylogenetic analyses to study the evolutionary patterns of variation in this retrovirus can enable investigators to discern events during its spread through a population.

Forensic studies of HIV sequence data first entered wide public awareness in the early 1990's in the case of the HIV-infected Florida dentist who was suspected of transmitting the virus to patients in his care [3]. Subsequent analyses of the HIV sequences obtained from the dentist and his infected patients strongly supported the finding that most of his patients who had no other risk factor for HIV infection had viruses that were closely related to those of the dentist [4]. Several subsequent analyses have reexamined these inferences and support the conclusions of Ou *et al*. [5,6] (but see [7]). Other studies of HIV transmissions have been conducted in several countries: in Sweden, a rape case [8] and two cases of knowingly exposing others to HIV infection [9,10] among others [10]; in the Netherlands, a deliberate injection [11]; in France, nosocomial transmissions from a surgeon to one of his patients [12] and a nurse to one of his patients [13]; in New Zealand, a case of deliberate transmission [14], and in the United States, deliberate transmissions to a cluster of sexual partners [15]. Investigations in several cases have exonerated suspected individuals [10,13,16,17]. Perhaps the most notorious case is that of the physician from Louisiana who was convicted of deliberately infecting his girlfriend with HIV and HCV that he obtained from two of his patients [18]. This widely covered case [19–21] established the legal precedent of the admissibility of HIV phylogenetic analysis in a criminal court in the United States. There is also the recent case in Libya of seven medical personnel convicted of and sentenced to death for the reckless nosocomial transmission of HIV [22–24], although it appears that the molecular sequence data presented in this case was either not considered or not properly applied [24,25].

It has been noted that phylogenetic patterns cannot act as the sole proof of a transmission event from a donor to recipient [18] and DNA forensic evidence needs to be considered in courts and by authorities in the context of the totality of the evidence available to them to adjudicate the guilt or innocence of a suspect in a case [10]. Even in cases in which patterns are consistent with a direction of transmission from the suspected donor to the alleged victim, it may be impossible to know with certainty that the transmission was directly from the donor to the victim without an intervening individual. The rapid evolution of HIV that gives each virion virtual uniqueness eliminates the possibility of establishing identity of viruses from one individual to the next and requires the use of phylogenetic inference to deduce the evolutionary relationships of viruses among individuals in a case.

There have been published guidelines for the use of human DNA identity profiling as evidence in court cases for more than a decade [26,27]. Leitner and Albert have provided an excellent review of 27 forensic investigations in Sweden and the methodologies followed in them [10]. In this methodological review, we add our recommendations on how these studies should be conducted. We first present a set of guidelines that we propose for use in HIV forensic studies and then present three recent case studies that illustrate patterns relevant to our discussion, including one that highlights the criticality of including local control specimens as part of the framework for analysis. We then close with some remarks that reflect our views on the potential benefits and drawbacks that forensic studies of HIV may present.

Requirements and Guidelines

Forensic specimen handling

Biological specimens used for forensic studies are generally governed by legal rules of evidence that require the maintenance of a so-called chain-of-custody. Once these specimens arrive in a laboratory, their handling needs to be closely recorded. Unlike human DNA identity profile work, these forensic studies are most often conducted in an academic research laboratory environment in which specimen handling and storage may not follow the rigorous protocols followed in forensic or clinical laboratories. However, it is essential that the handling and storage of specimens and products derived from them be closely monitored. Once the specimen is in the custody of the lab, it is the responsibility of principal investigators, laboratory supervisors, and the technicians that specimens and their derived products be properly handled. The technicians working with forensic specimens should be experienced in the methods and techniques that are used and any work that is done with the forensic specimen must be recorded in laboratory notebooks. Since these records may be called into court, we recommend that these notes be kept in special notebooks devoted solely to the work with the forensic specimens and their derived products. For further guidelines, see the Supporting Online Material from Budowle *et al*. [2] (http://www.sciencemag.org/cgi/data/301/5641/1852/DC1/1).

PCR methods and practices. Amplification of regions targeted for sequence analysis should be done in a laboratory designed in keeping with guidelines presented by Dieffenbach *et al*. [28] with standard techniques designed to minimize carryover contamination [29]. These considerations are good practice for all HIV-related PCR techniques and should be the absolute minimum standard for work with forensic specimens.

Amplification via PCR and the handling of specimens, reagents, and products should be done in dedicated laboratory areas specifically designed for such work. Diligence should be maintained to prevent the contamination of amplification reactions with either products of previous reactions or reagents from other reactions from the same time. Dieffenbach *et al*. have outlined procedures for setting up laboratory spaces to prevent contamination of reactions [28]. Practices of particular note include:

1. PCR templates should be prepared and reactions set up in an area or room dedicated for this purpose. PCR products or DNA clones should never be handled in this “clean” area.
2. Pipettes and other laboratory equipment used in the clean area should be designated for PCR and are not to be used for general laboratory work.
3. All reagents should be dispensed with positive-displacement pipettors or special barrier pipette tips.
4. In order to minimize aerosols and carryover, liquids and condensation should be briefly centrifuged before tubes are opened. Efforts should be made to avoid popping tubes open and touching the inside rim of the cap. Opening tubes with single use towel strips can prevent contamination between tube and gloves.
5. Gloves should be worn at all times and changed frequently. They should be tight fitting so that loose glove material does not get caught when closing PCR tubes.
6. Lab coats dedicated for use in the PCR set-up area should be worn, and washed frequently and always following a spill involving nucleic acids.

Dieffenbach *et al*. also discuss special handling procedures for solutions and reagents [28]. Below are our recommendations and requirements modified from their suggestions:

1. Solutions and reagents for PCR reactions should be designated for PCR, always handled with clean gloves, and stored in a clean area.
2. Solutions should be sterile, nucleic acid-free, and free of nucleases. To the extent possible, reagents should be purchased pre-made from an offsite company to completely eliminate the chance of in-house contamination.
3. All reagents should be dispensed from single-usage aliquots to minimize the chance of carryover contamination.

4. To avoid the possibility of carryover contamination, the PCR machines cannot be located in the PCR clean area and tubes that have been in a PCR machine should never be returned to the clean area.

5. A first-round reaction tube should never be opened once it leaves the clean area except for the addition of positive control DNA.

6. Second-round reaction mixes should be set-up in the clean area, but the addition of first-round material to the second-round reaction mix must be done outside of the clean area.

7. Negative control reactions that contain all reagents except the target template must be included in all PCR experiments.

Sequence assembly and error checking.

We recommend the use of automated sequencing machinery for generating HIV sequences for forensic studies, since they directly store results as computer files. All computer files generated in the work with the sequences must be kept and backups should be archived. The use of sequence assembly software, such as Sequencher (Gene Codes, Ann Arbor, MI), greatly facilitates the reconciliation of multiple sequencing reactions; sequencing of both template strands are strongly recommended. Sequencher software (among others) also helps maintain information that tracks the dates, times, sequencing reactions, and original automated base calls.

Any edits of the sequence data to reconcile multiple reactions must be recorded. In general, we recommend that, if multiple reactions differ in the base calls of a sequence, the assembled sequence record the difference as the most inclusive IUPAC ambiguity; for example, if the call of a forward reaction at a particular site in the sequence is a G while the call of the reverse reaction at the site is a T (complementary to A), the reconciled call at this site should be encoded as an R (indicating a purine, *i.e.,* G or A). An exception to this procedure would be the editing of base calls at the ends of the sequence reaction runs where the signal is poor and all base calls 3′ (or alternatively 5′) to it are removed.

Once the sequence data have been assembled and reconciled into contigs, they should be checked against both public (GenBank or Los Alamos HIV Sequence Database) and local laboratory databases [30]. The public databases may be queried using Worldwide Web interfaces (http://www.ncbi.nlm.nih.gov/BLAST/ or http://www.hiv.lanl.gov/content/hiv-db/BASIC_BLAST/basic_blast.html). Both FASTA [31] and BLAST [32] are suitable for the purpose of checking sequences against the local database.

Epidemiological issues and selection of local controls

It is essential that the degree of relationships between forensic specimens be examined in the context of the broader HIV epidemic. Wherever possible, a proper understanding of the molecular epidemiology of HIV is desirable; for example, local outbreaks with rapid spread may lead to clustering of very closely related variants among individuals (see for example, Liitsola *et al*. [33]). Therefore, it has been consistent practice to include so-called local controls in forensic studies [4,10]. The proper selection of these is crucial to deriving proper inferences. The local controls should be isolated at a time near the suspected date of transmission and reflect the risk group(s) and, where there are multiple circulating subtypes or recombinant forms, genetic clade or subtype of the forensic case [10].

The assertion of Hillis (in testimony [34]) and Metzker *et al*. (in the *Proceedings of the National Academy of Sciences, U.S.A*. [18]) that there is no evidence of correlation of risk group to genetic sequence variants is incorrect. Studies of subjects from Thailand [35], the Netherlands [36] (as well as elsewhere in Europe [37]), Brazil [38,39], Russia [33,40], Denmark [41], Argentina [42] and Australia [43] have demonstrated that individuals from different risk groups are infected by distinct populations of HIV variants. A recent study by Anderson *et al*. of a set of sequences drawn from clinics in five cities across the United States using both phylogenetic analysis and multidimensional scaling has demonstrated a degree of geographic and risk-group correlation with HIV genetic structure [44]. Furthermore, even if the correlation of genetic variability to risk group is weak or absent at a given time in the past, there is no way to predict that this would be the case at any given time in such a rapidly changing epidemic. Clearly, the prudent course is to make the best effort to have a properly designed population of local controls to the extent possible.

While it may be difficult to obtain an ideal set of local control individuals, a correctly conducted forensic study may depend on proper recruitment of this group. These individuals should be assured that their identities will be protected. If the local controls are not truly reflective of the populations from which the alleged donor and recipient pair are drawn, then the local control set is in effect a convenience sample. A convenience set of HIV reference sequences is better than no reference set at all. Such a set might be obtained from the Los Alamos National Laboratory HIV Sequence Database (http://www.hiv.lanl.gov/content/hiv-db/ALIGN_CURRENT/ALIGN-INDEX.html) or selected from GenBank either more or less at random, or guided by BLAST searches to derive a set of the most similar sequences. However, when using a convenience control set, conclusions made about the relative relatedness of the sequences of the alleged donor recipient may be incorrect (see Case Study 3 below).

Considerations of diversity and choice of genes for analysis

In all studies of HIV sequences it is wise to keep in mind the extent of genetic diversity seen in the retroviral genome. Levels of HIV-1 genetic diversity within an infected subject vary during the course of infection and these patterns have been well characterized (for an example, see [45]). Due to these considerations, the genetic diversity of the suspected source individual and the alleged recipient will likely differ. While it might seem desirable to examine patterns of diversity in the individuals concerned to make some inferences about the time of infection, such studies need to be carefully designed so that proper methods are used to gauge viral genetic diversity and conclusions from them interpreted with caution and only in the context of the other inferences derived from the sequence data.

Investigators who make use of PCR-generated HIV sequences should keep in mind that, while samples bearing low numbers of viral templates may be amplified, the resulting product will reflect this limited number of input copies and any diversity in the product may be the result of errors during PCR amplification [46]. Methods are available to assess this phenomenon and estimate the effective number of input templates [47].

There are several sequencing approaches that are intended to cope with viral genetic diversity. Leitner and Albert favor so-called population sequencing (also referred to as consensus sequencing) [10,48]. In their approach, they estimate the input copy number for PCR and adjust the reactions so that at least 10 gene copies are amplified. The products of these reactions are then sequenced and polymorphism at variable sites is noted using IUPAC ambiguity codes (*e.g.*, G and A nucleotides at a site are coded as R). While relatively fast and inexpensive, this methodology has several disadvantages. First, it does not quantitate the genetic diversity of a sample; second, the sequence estimated is reflective of the consensus of a sample and the genetic distance between two samples will be an underestimate of the average pairwise genetic distance; third, and perhaps most importantly for considerations in forensic investigations, it is not possible to reconstruct phylogenies in which the sequence lineages interdigitate (see *e.g.*, Figure 1, where clusters A and B contain both C and N sequences). We recommend instead that sequences be obtained from either a molecular clone (after appropriate quantitation of template input as described above) or, to avoid PCR-associated mutations, direct sequencing of PCR product derived from a single molecule template. Since regions of the HIV genome are prone to length variation, samples that are subjected to direct sequencing should contain relatively low variability and should be diluted to approximate a molecular end-point (one amplifiable copy) to obtain an unambiguous sequence.

A variety of HIV genes have been used to examine viral relationships in forensic studies. Portions of the gp120 region of *env* , including the V3 loop, contain ample genetic variability and have been most frequently used. Studies have also examined portions of *gag* (p17) [8,11,12,15] *pol* (RT [8,10,13,18,49], and IN [50]), and *tat* [51]. While it may useful to examine regions that are governed by different selective regimes[18], care should be exercised in interpreting results from genes that are less variable than *env*, particularly regions that are depauperate in variability and potentially subject to selective forces governed by resistance to antiviral drugs that may lead to potential parallel or convergent mutations. The p17-encoding region of *gag* frequently shows sufficient variability and is recommended if it is desirable to include an additional gene.

Phylogenetic and distance analysis of HIV sequences for forensic studies. Sequence sets used for genetic analyses should be aligned such that each position at every site in the alignment has positional

homology, *i.e.,* the nucleotides at any site are presumed to have descended from a common ancestor at that position in the alignment. Computational methods, such as CLUSTAL [52,53], using progressive clustering alignment algorithms may be used to facilitate the alignment process, but their results need to be visually inspected and perhaps adjusted. Furthermore, sequence regions containing sites that cannot be unambiguously aligned should be removed from the alignment (gap-stripped [30]). Forensic analyses of sequence data may also examine the shared presence of bases or segments in variable length regions that might have unclear positional homology. Researchers should be cautioned that the dynamics of sequence evolution in these regions is not clearly known and may be governed by parallel or convergent evolution; thus inferences made about such shared patterns may be particularly prone to error.

A pattern of frequent and repetitious G→A mutational bias, referred to as hypermutation [54], is often seen in retroviral sequences; these mutations are caused by a host enzyme, APOBEC3G, that apparently acts to protect against retroviral infection (see KewalRamani and Coffin [55] for a review). Such parallel mutations in different infected individuals can confound evolutionary inference, especially in the pol gene where genetic variability is relatively low. Therefore, sequences with regions of substantial G→A mutational bias should be removed from alignments intended for forensic study. HYPERMUT [56,57] (http://www.hiv.lanl.gov/content/hiv-db/HYPERMUT/hypermut.html) may be used to detect regions of hypermutation. Patterns of mutation at a site of greater than 5–10% of G's substituted by A's with G→A replacements more than twice that of A→G or a high level of unique nonsynonymous substitutions indicate probable hypermutation.

Standard methods of phylogenetic analysis should be used in forensic studies. In the field of HIV research, the current state of the art involves using evolutionary models of nucleotide replacements such as the Hasegawa-Kishino-Yano (HKY) [58] or the general time-reversible (REV or GTR) model [59,60], among others. Such models take into account the skewed base composition of retroviral sequences and are preferable to models that assume equal base frequencies (*e.g.*, Jukes-Cantor [61] or Kimura 2-parameter [62]). The computer program Modeltest [63], in conjunction with PAUP* [64], allows researchers to evaluate a set of evolutionary models with varying numbers of estimated parameters using either hierarchical likelihood-ratio testing [65] or the Akaike information criterion [66]. In addition to evaluating evolutionary rate parameters, these methods examine Γ- (gamma-) distributed site-to-site rate variability [67] and the impact of invariable sites. PAUP* runs on a variety of computer platforms (Macintosh, DOS/Windows, and UNIX) and can be used to implement models using these parameters to infer evolutionary trees using neighbor-joining [68], or maximum likelihood [69]. [It is also possible to perform maximum parsimony phylogenetic analyses using PAUP*, although the use of Fitch (unweighted) parsimony is not recommended due to the highly skewed base composition of retroviral genomes.] Strength of support can be evaluated using site-resampling bootstrapping [70] or parametric bootstrapping approaches [65]. Bayesian methods of phylogenetic inference [71] have also been suggested for evaluating reliability of evolutionary relationships [18].

Although phylogenetic analyses provide the most definitive information on the relatedness of sequences in a forensic case, measures of genetic distance might also be of value. Genetic distances can be estimated with PAUP* [64] using parameter settings estimated in a similar fashion to the methods mentioned above. While the inferences of relationships derived from phylogenetic methods are more robust, genetic distances for regions with well characterized evolutionary rates (*e.g.*, the C2-V5 region of *env* [45]) might be useful in estimating the time of transmission events. Researchers should keep in mind that the distance measured reflects divergence from the most recent common ancestor (MRCA) of the descendent lineages; the time of this MRCA would pre-date the transmission event. Again, conclusions based on genetic distances should be interpreted with great caution and considered in the context of the other inferences derived from the sequence data.

Other considerations

Despite efforts to prevent laboratory error, the possibility of human error remains. Errors in laboratory technique leading to detection of HIV in samples that are ultimately shown to come from uninfected

individuals (false positives) may be as great as 2-5% [72]. Errors resulting in misidentification can be caused by carryover contamination [29], by errors in labeling of specimens, or by attribution of laboratory results to the wrong individual [73]. While Bayesian approaches may be used to evaluate the impact of false positive results on the estimation of the posterior odds (and perhaps the odds of guilt) [73], the probability of laboratory error in a specific case is difficult to quantify. One solution suggested to reduce the impact of laboratory error is to split specimens and send them to separate, independent laboratories [74]. According to recommendations of the National Research Council Committee on DNA Forensic Science, specimens used for DNA identity profiling analysis should be subjected to repeat testing "whenever feasible" [27]. In instances of HIV forensic testing, it would seem that duplicate testing either in different laboratories or of specimens obtained at different times should be a minimum standard [18]. A further safeguard we have followed in our forensic analyses is to use different laboratories to analyze specimens from individuals who are suspected of epidemiologic linkage [51,72]; for example, specimens from the suspected donor would go to one laboratory, while those from the alleged recipient would go to another. The resultant sequences may then be compared with the virtual assurance that any linkage seen would not be due to inadvertent contamination or other laboratory errors.

Since HIV populations become more genetically distinct as the infection progresses, it is important in forensic cases that viral variants be sampled as close to the time of transmission as possible. While phylogenetic relationships will continue to be maintained, affinities among particular lineages may be lost due to the extinction of particular lineages during infection.

At times it may be desirable to conduct further analyses of specimens, for example, heteroduplex analyses (HMA or HTA) [75] to estimate genetic diversity or to investigate the presence of rare variants. Analyses such as these should be conducted with the greatest of care, since they involve simultaneous handling of specimens from multiple subjects and are thus prone to sample mix-up and carryover contamination. Such studies should only be done after preliminary genetic characterization of specimens has been completed.

Case Studies

In order to present some cases of forensic analyses, we include case studies of three data sets.

Case study 1 (see also de Oliveira *et al*. [51])

While attempting to draw a blood sample from a child with advanced AIDS, a health care worker was cut with glass from a broken blood collection tube that punctured her glove. Despite prompt washing of the cut and post-exposure prophylactic treatment with ZDV, the nurse presented symptoms compatible with acute HIV viremia 13 days following the accident. The nurse had never had a blood transfusion, denied the use of injection drugs, and reported celibacy in the 6 years before the accident. However, tests of blood samples obtained from her on the day of the accident reportedly tested positive for anti-HIV-1/2 antibodies, and it was alleged that she had been infected prior to the accident.

Blood samples from the nurse (taken 20 days after the accident) and the child (sampled after 10 days) were examined at three different sites for HIV-1 nucleic acids corresponding to a region in the first exon of *tat* and the V3-V5 region of *env* ; we present the results of the *env* analysis here (Figure 1). Phylogenetic reconstructions were done using maximum-likelihood estimation [69] under an HKY model of evolution [58] using the computer program PAUP* [64]. Bootstrap support [70] of phylogenetic relationships were calculating using the neighbor-joining method [68]. Two distinct viral populations (A and B) were detected in the phylogenetic analysis; these two form a monophyletic group to the exclusion of sequences from the reference data set with 99% bootstrap support. Both nurse (indicated as N) and child (C) sequences are found in each of the two clusters, indicating that the viral sequences from the two subjects of this study are very closely related. The likelihood that two disparate clusters would be found in two subjects also provides strong evidence of a shared transmission history. Since this was a civil case, the pattern of relatedness shown led to an out-of-court settlement before a local control population was included in the analysis.

Case study 2

This study involves sequences from a case of suspected deliberate transmission. The phylogram of the *env* C2-V5 region sequences from this study is shown in Figure 2. Four clusters of sequences (1–4) are indicated on the tree: 1 through 3 (dashed-line boxes) indicate the local controls from this study; cluster 4 (solid-line boxes) contains all of the sequences from the index case (A) and the alleged recipient (B). Average pairwise distance within cluster 4 is 1.31% (range: 0–3.39%), quite closely related for *env* sequences from two different individuals. Sequence B23 is most closely related to the sequences from the index case, as close as 0.82% distant (to A43). Also noteworthy and suggestive of direct transmission is that the *env* sequences from B form a monophyletic cluster within those from A.

Case study 3

This case involves a child (referred to here as subject V) who apparently became infected with HIV-1 while a patient at a hospital in City “Q,” a large city in the eastern United States. Initially a nosocomial source of the infection was suspected, but subject U, a known HIV-1-positive friend of the family, fell under suspicion. HIV-1 *env* C2-V5 gp120 region sequences from subject U and the child were obtained and examined. Figure 3a illustrates the results of a preliminary phylogenetic analysis of the *env* sequences: one from the suspected donor (U-1), three from the alleged recipient (V-1, V-2 and V-3), as well as a variety of sequences chosen as a convenience sample from the Los Alamos HIV sequence database. Note that this phylogram was constructed before local controls were available for analysis. The sequences from V cluster together with 100% bootstrap support. Note also that the U and V sequences cluster together with high bootstrap support (94%). While there is a high level of bootstrap support in this analysis, the average genetic distance between U1 and the V sequences is high, >10%, and does not appear to support the suspected time of transmission. However, based solely on the phylogenetic relationships evaluated, U-1 would appear to be related to the V sequences.

Figure 3b shows a maximum-likelihood phylogram of the data set constructed later after including 20 local controls (HIV sequences from infected children from the vicinity of City “Q” (reflecting the risk group of subject V). All of the sequences from the subject V samples formed a relatively closely related cluster (#1). A larger group of more distantly related sequences (#2) includes the V sequences, nearly all of the subject U sequences (excluding sequence C52), sequences from four of the local controls (LC-02, -09, -13, -20) and a sequence, AY139298, identified in the Anderson *et al*. study [44] as a sample from Los Angeles (no sequences in the Anderson study were sampled from City “Q”). Within this relatively large cluster of possibly epidemiologically linked viral sequences is a relatively tight cluster of seven sequences from subject U (indicated by #3). The J62 and Q31 sequences seem to be more divergent and, depending on the method of analysis, cluster variously with different local controls or AY139298. These analyses appear to have uncovered an epidemiologically linked cluster, and taken as a whole, the relationships seen do not indicate that subject U was the source of the infection of subject V. An alternative explanation of the results is that U and V were infected from either an unknown common source, or several individuals sharing a population of closely related viruses.

Case study 3 indicates the critical importance of including a proper set of local controls in a forensic study of HIV sequence data. When the earlier analysis was conducted using the convenience sample from subtype B sequences selected from a publicly available database, a phylogeny was reconstructed that was indicative of close relationship and the potential of a direct transmission. When controls from a relevant risk group were included, the genetic distance patterns seen for U-1 and V-1, V-2, and V-3 were, of course, maintained, but our perception of the relationships was changed. It is evident that there is a more complex epidemiologic pattern here, and the likelihood of direct transmission based on these data is significantly reduced.

Figure 1

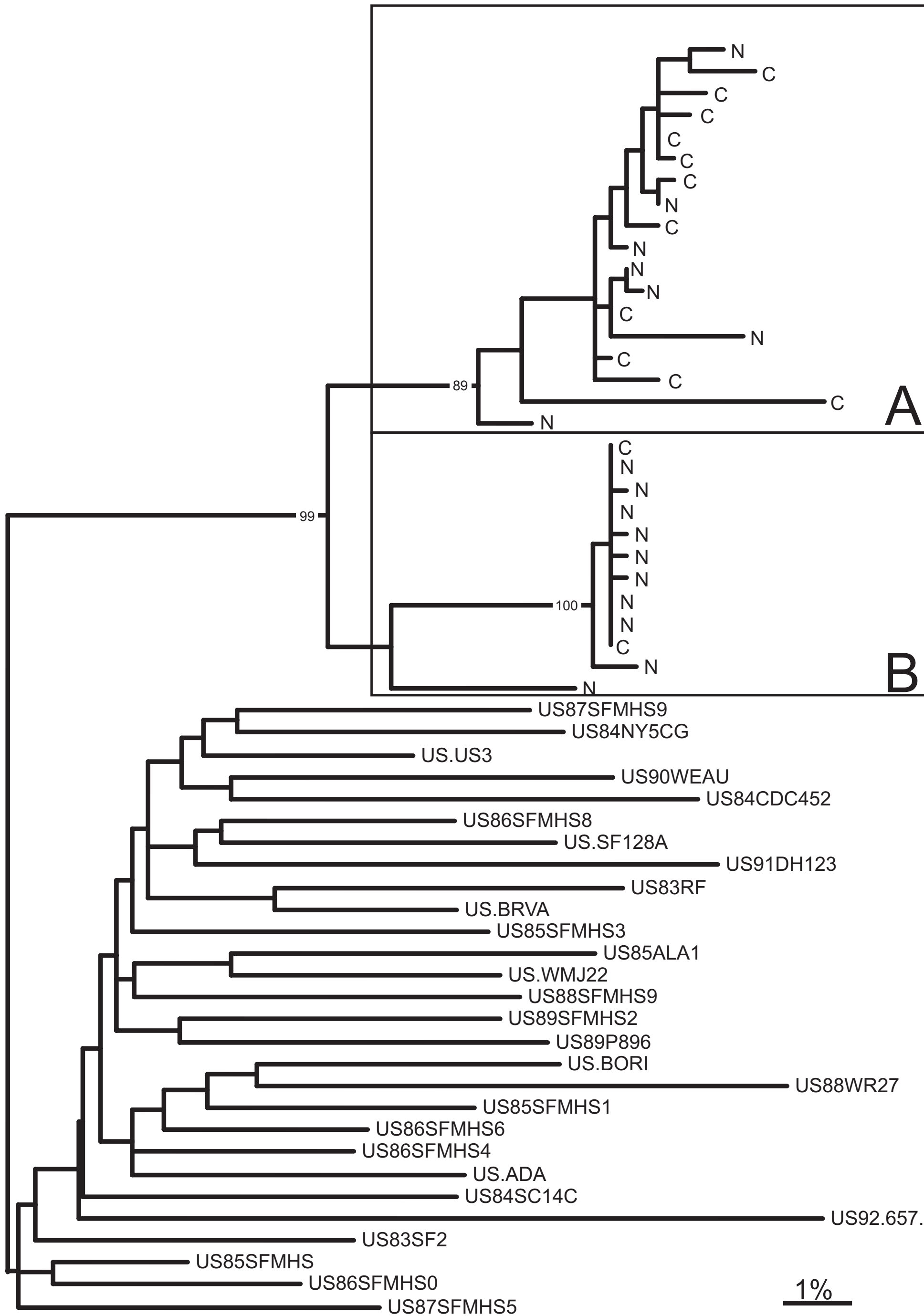


Figure 1. Maximum-likelihood phylogram (-LnL = 4221.0130) representing the reconstruction of the phylogenetic relationships between the env (V3-V5) sequences obtained from the nurse (N), the putative donor (C), and 28 sequences chosen from GenBank. Ten iterations with random sequence addition were used. Scale bar represents 1% genetic distance (0.01 substitutions/site). Bootstrap values are shown at nodes with greater than 70% support.

Figure 2

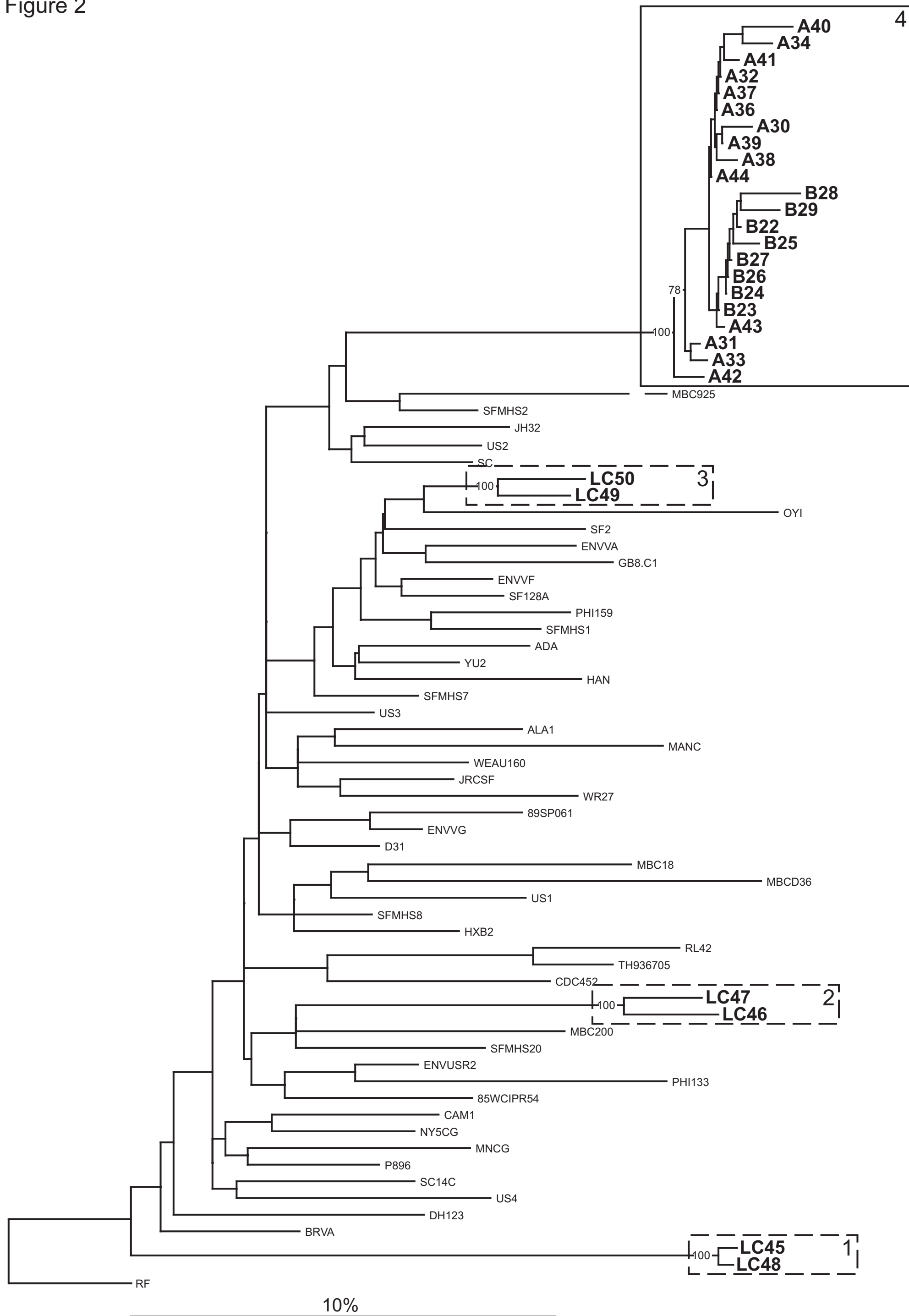


Figure 2. Neighbor-joining phylogram representing the reconstruction of the phylogenetic relationships between the env (C2-V5) sequences obtained from the index case (A31-44), the alleged recipient (B22-29), three local controls (LC45 and LC48; LC46 and LC47; and LC49 and LC50) and 48 sequences chosen from GenBank. Ten iterations of random sequence addition were used. Scale bar represents 10% genetic distance. Bootstrap values are shown at nodes with greater than 70% support.

Figure 3a

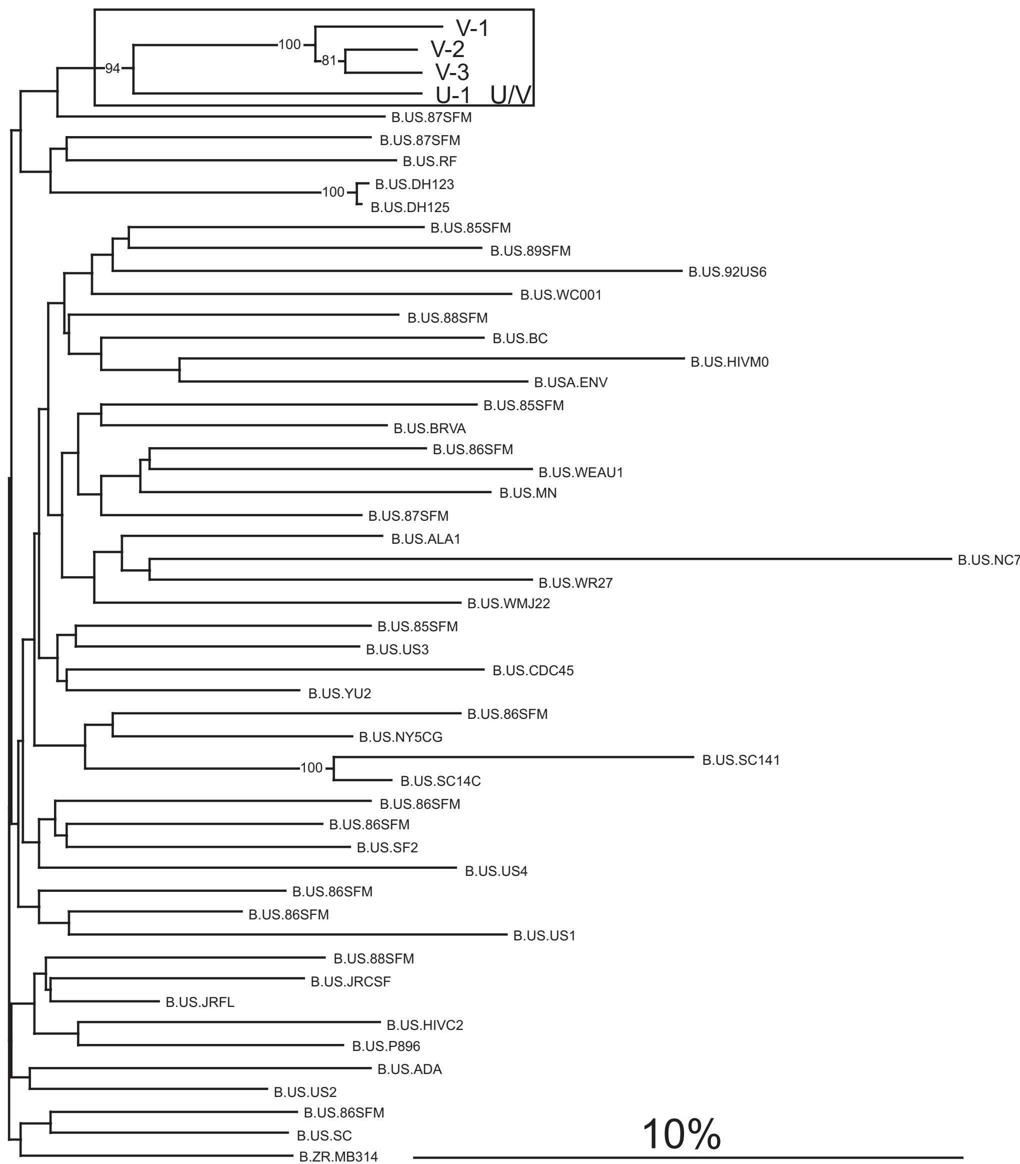
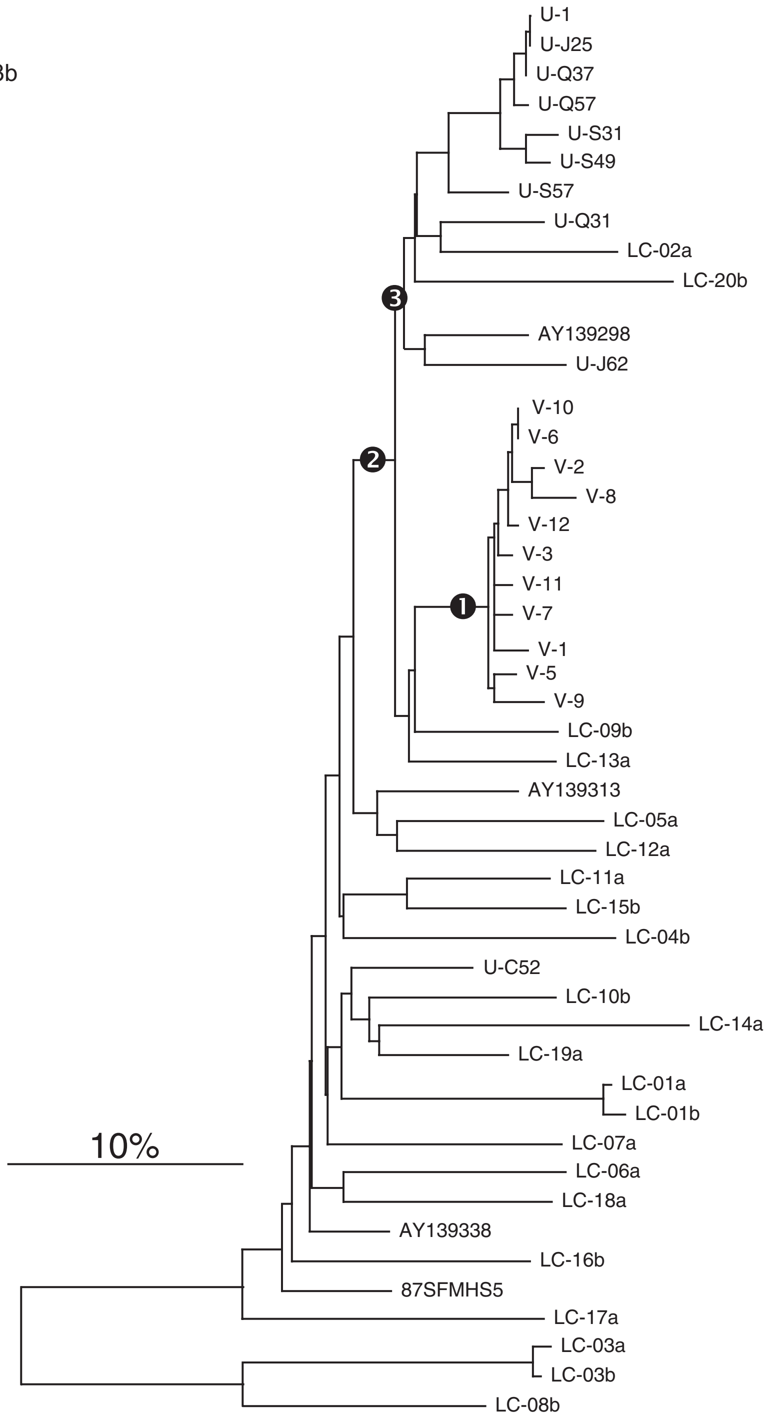


Figure 3. Phylograms of env (C2-V5) sequences from the U/V case. Suspected index case (subject U) sequences are labeled with U, while those from the alleged recipient are labeled V. a. Preliminary neighbor-joining tree including four sequences from the case (boxed) and 48 sequences obtained from the Los Alamos HIV Sequence Database. b. Maximum-likelihood phylogram (-LnL = 6024.5864) incorporating local controls. Local controls are labeled LC; AY139298, AY139313 and AY139338 are from studies by Anderson et al. *[44,76]; 87SFMHS5 is from the San Francisco Men's Health Study [77]. Starting trees for the ML searches were obtained either by stepwise addition of the 47 sequences (10 iterations) or random starting trees (50 iterations), and then swapped using the SPR (subtree-pruning and regrafting) method.*

Figure 3b



Concluding Remarks

The use of HIV sequences in forensic studies provides the opportunity to use evolutionary patterns to inform the courts and other authorities about the possibility of events leading to infections. Since HIV is both infectious and rapidly evolving, it provides an unusual opportunity for the application of microbial forensics to the investigation of so-called biocrimes [1,2].

We consider the impact of scientific analyses on the investigation of events leading to both accidental and deliberate infections as having two-fold importance. First, this information can guide the courts and juries in evaluating the totality of the evidence available to them to decide the guilt or innocence of suspects in a case. The impact of reckless or deliberate infection with HIV on a victim is catastrophic. It is fully understandable that the victim and society would want such wrongs to be redressed. On the other hand, HIV forensic evidence has an even greater potential to exclude suspects and clarify events during an investigation, perhaps contributing to the acquittal of the wrongly accused in court proceedings. Second, we wish to emphasize that care must be used in the examination of forensic data and in communicating the results of these examinations to the courts and to juries. These results are potentially very powerful, either in support of other evidence leading to conviction of the guilty or to exoneration of the falsely accused. However, for the value of the forensic evidence obtained from HIV sequence data to be maintained, the collection, analyses, and expert evaluation of results must remain of the highest quality. It would be of no value to the scientific community or to society as a whole for HIV sequence forensics to fall into the realm of "junk" science. Experts in this area, serving both as witnesses for the adversaries in a case and advisors to the court, need to be able to acknowledge the limitations of the inferences that might be made. We support the recommendations by Leitner and Albert that the language used by experts in written reports or testimony be appropriately conservative and acknowledge the limitations of inferences derived using microbial forensic methods [10]. Specifically, we recommend that, where applicable, statements used should be similar to "the viral sequences from the two subjects display a high level of similarity" or "the results from our analysis are compatible with the possibility that subject A infected subject B." Statements should include the possibility that an unknown third person might be involved in the transmission chain and that direction of transmission cannot be proven. Misleading testimony and other statements do not benefit the cause of justice, regardless of how convinced an expert may be of the guilt or innocence of the accused.

We hope that the guidelines and perspectives provided here will be helpful to researchers that conduct microbial forensic studies of HIV sequences. We anticipate that such studies in the future will remain of the highest quality.

Acknowledgements

We wish to thank Bette T. Korber, William Gallaher, Jan Albert, Thomas Leitner, Belinda Herring, David Nickle, Sherry McLaughlin and Daniel Shriner for valuable discussions and suggestions. This work was supported by U.S. Public Health Service grants, including to the University of Washington Center for AIDS Research.

References

[1] Larkin M: Microbial forensics aims to link pathogen, crime, and perpetrator. *Lancet Infect Dis* 2003, **3**:180.

[2] Budowle B, Schutzer SE, Einseln A, Kelley LC, Walsh AC, Smith JA, Marrone BL, Robertson J, Campos J: Public health. Building microbial forensics as a response to bioterrorism. *Science* 2003, **301**:1852–1853.

[3] Centers for Disease Control: Possible transmission of human immunodeficiency virus to a patient during an invasive dental procedure. *MMWR* 1990, **39**:489–493.

[4] Ou C-Y, Ciesielski CA, Myers G, Bandea CI, Luo C-C, Korber BTM, Mullins JI, Schochetman G, Berkelman R, Economou AN, *et al.*: Molecular epidemiology of HIV transmission in a dental practice. *Science* 1992, **256**:1165–1171.

[5] Hillis DM, Huelsenbeck JP: Support for dental HIV transmission. *Nature* 1994, **369**:24–25.

[6] Crandall KA: Intraspecific phylogenetics: support for dental transmission of human immunodeficiency virus. *J Virol* 1995, **69**:2351–2356.

[7] DeBry RW, Abele LG, Weiss SH, Hill MD, Bouzas M, Lorenzo E, Graebnitz F, Resnick L: Dental HIV transmission? *Nature* 1993, **361**:691.

[8] Albert J, Wahlberg J, Leitner T, Escanilla D, Uhl'en M: Analysis of a rape case by direct sequencing of the human immunodeficiency virus type 1 pol and gag genes. *J. Virol*. 1994, **68**:5918–5924.

[9] Leitner T, Escanilla D, Marquina S, Wahlberg J, Brostrom C, Hansson HB, Uhlen M, Albert J: Biological and molecular characterization of subtype D, G, and A/D recombinant HIV-1 transmissions in Sweden. *Virology* 1995, **209**:136–146.

[10] Leitner T, Albert J: Reconstruction of HIV-1 transmission chains for forensic purposes. *AIDS Rev* 2000, **2**:241–251.

[11] Cornelissen M, Mulder-Kampinga G, Veenstra J, Zorgdrager F, Kuiken C, Hartman S, Dekker J, van der Hoek L, Sol C, Coutinho R, et al.: Syncytium-inducing (SI) phenotype suppression at seroconversion after intramuscular inoculation of a non-syncytium-inducing/SI phenotypically mixed human immunodeficiency virus population. *Journal of Virology* 1995, **69**:1810–1818.

[12] 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**:4537–4540.

[13] Goujon CP, Schneider VM, Grofti J, Montigny J, Jeantils V, Astagneau P, Rozenbaum W, Lot F, Frocrain-Herchkovitch C, Delphin N, et al.: Phylogenetic analyses indicate an atypical nurse-to-patient transmission of human immunodeficiency virus type 1. *J Virol* 2000, **74**:2525–2532.

[14] Dwyer DE, Herring BL, Ge YC, Bolton WV, Ellis-Pegler RB, Thomas M, Schroeder BA, Croxson MC, Cunningham AL, Mullins JI, et al.: Human immunodeficiency virus type 1 subtypes B and C detected in New Zealand. *AIDS Res Hum Retroviruses* 1998, **14**:1105–1108.

[15] Robbins KE, Weidle PJ, Brown TM, Saekhou AM, Coles B, Holmberg SD, Folks TM, Kalish ML: Molecular Analysis in Support of an Investigation of a Cluster of HIV-1-Infected Women. *AIDS Res Hum Retroviruses* 2002, **18**:1157–1161.

[16] Holmes EC, Zhang LQ, Simmonds P, Rogers AS, Leigh Brown AJ: Molecular investigation of human immunodeficiency virus (HIV) infection in a patient of an HIV-infected surgeon. *J. Infect. Dis*. 1993, **167**:1411–1414.

[17] Jaffe HW, McCurdy JM, Kalish ML, Liberti T, Metellus G, Bowman BH, Neasman AR, Witte JJ: Lack of transmission of human immunodeficiency virus in the practice of a dentist with AIDS. *Ann. Intern. Med.* 1994, **121**:855–859.

[18] Metzker ML, Mindell DP, Liu XM, Ptak RG, Gibbs RA, Hillis DM: Molecular evidence of HIV-1 transmission in a criminal case. *Proc Natl Acad Sci U S A* 2002, **99**:14292–14297.

[19] Vogel G: Phylogenetic analysis: getting its day in court. *Science* 1997, **275**:1559–1560.

[20] King LS: Tracking the chameleon killer. *SFI Bulletin* 1999, 14.

[21] Katinas L, Crisci JV: La filogenia frente a la justicia: ¿el fin del caso? *Ciencia Hoy* 1999, 9.

[22] Visco-Comandini U, Cappiello G, Liuzzi G, Tozzi V, Anzidei G, Abbate I, Amendola A, Bordi L, Budabbus MA, Eljhawi OA, et al.: Monophyletic HIV type 1 CRF02-AG in a nosocomial outbreak in Benghazi, Libya. *AIDS Res Hum Retroviruses 2002*, **18**:727–732.

[23] Krosnar K: Libyan government lets AIDS experts comment on hospital deaths. *Bmj* 2003, **327**:360.

[24] Krosnar K: Six health workers sentenced to death in Libya. *Bmj* 2004, **328**:1153.

[25] A true test of leadership. *Nature* 2004, **430**:277.

[26] National Research Council Committee on DNA Technology in Forensic Science: DNA Technology in Forensic Science. Washington, D. C.: National Academy Press; 1992.

[27] National Research Council Committee on DNA Forensic Science: An Update: The evaluation of forensic DNA evidence. Washington, D. C.: National Academy Press; 1996.

[28] Dieffenbach CW, Dragon EA, Dveksler GS: Setting up a PCR laboratory. In PCR Primer: A Laboratory Manual. Edited by C. W. Dieffenback GSD: Cold Spring Harbor Labs Press; **1995**:7–16.

[29] Kwok S, Higuchi R: Avoiding false positives with PCR [published erratum appears in *Nature* 1989 Jun 8;**339**(6224):490]. *Nature* 1989, **339**:237–238.

[30] Learn GH, Jr., Korber BT, Foley B, Hahn BH, Wolinsky SM, Mullins JI: Maintaining the integrity of human immunodeficiency virus sequence databases. *J Virol* 1996, **70**:5720–5730.

[31] Pearson WR: Flexible similarity searching with the FASTA3 program package. In Bioinformatics methods and protocols. Edited by Misener S, Krawetz SA: Humana Press; **1999**:185–219.

[32] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997, **25**:3389–3402.

[33] Liitsola K, Tashkinova I, Laukkanen T, Korovina G, Smolskaja T, Momot O, Mashkilleyson N, Chaplinskas S, Brummer-Korvenkontio H, Vanhatalo J, et al.: HIV-1 genetic subtype A/B recombinant strain causing an explosive epidemic in injecting drug users in Kaliningrad. *Aids* 1998, **12**:1907–1919.

[34] State of Louisiana vs. Richard J. Schmidt. 15th Judicial District Court, Lafayette Parish, LA; 1998.

[35] Ou CY, Takebe Y, Weniger BG, Luo CC, Kalish ML, Auwanit W, Yamazaki S, Gayle HD, Young NL, Schochetman G: Independent introduction of two major HIV-1 genotypes into distinct high-risk populations in Thailand. *Lancet* 1993, **341**:1171–1174.

[36] Kuiken C, Zwart G, Baan E, Coutinho R, Anneke J, van den Hoek R, Goudsmit J: Increasing antigenic and genetic diversity of the HIV-1 V3 domain in the course of the AIDS epidemic. *Proc. Natl. Acad. Sci. USA* 1993, **90**:9061–9065.

[37] Kuiken CL, Goudsmit J: Silent mutation pattern in V3 sequences distinguishes virus according to risk group in Europe. *AIDS Res Hum Retroviruses* 1994, **10**:319–320.

[38] Sabino EC, Diaz RS, Brigido LF, Learn GH, Mullins JI, Reingold AL, Duarte AJS, Mayer A, Busch MP: Distribution of HIV-1 subtypes seen in an AIDS clinic in Sao Paulo City, Brazil. *AIDS* 1996, **10**:1579–1584.

[39] Couto-Fernandez JC, Morgado MG, Bongertz V, Tanuri A, Andrade T, Brites C, Galvao-Castro B: HIV-1 subtyping in Salvador, Bahia, Brazil: a city with African sociodemographic characteristics. *J Acquir Immune Defic Syndr* 1999, **22**:288–293.

[40] Bobkov A, Kazennova E, Selimova L, Bobkova M, Khanina T, Ladnaya N, Kravchenko A, Pokrovsky V, Cheingsong-Popov R, Weber J: A sudden epidemic of HIV type 1 among injecting drug users in the former Soviet Union: identification of subtype A, subtype B, and novel gagA/envB recombinants. *AIDS Res. Hum. Retroviruses* 1998, **14**:669–676.

[41] Iversen AK, Learn GH, Fugger L, Gerstoft J, Mullins JI, Skinhøj P: Presence of multiple HIV subtypes and a high frequency of subtype chimeric viruses in heterosexually infected women. *J. Acquir. Immune Defic. Syndr*. 1999, **22**:325–332.

[42] Avila MM, Pando MA, Carrion G, Peralta LM, Salomon H, Carrillo MG, Sanchez J, Maulen S, Hierholzer J, Marinello M, et al.: Two HIV-1 epidemics in Argentina: different genetic subtypes

Reviews

associated with different risk groups. *J Acquir Immune Defic Syndr* 2002, **29**:422–426.

[43] Herring BL, Ge YC, Wang B, Ratnamohan M, Zheng F, Cunningham AL, Saksena NK, Dwyer DE: Segregation of human immunodeficiency virus type 1 subtypes by risk factor in Australia. *J Clin Microbiol* 2003, **41**:4600–4604.

[44] Anderson JP, Learn GH, Rodrigo AG, He X, Wang Y, Weinstock H, Kalish ML, Robbins KE, Hood L, Mullins JI: Predicting demographic group structures based on DNA sequence data. *Mol Biol Evol* 2003, **20**:1168–1180.

[45] Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H, Gupta P, Rinaldo CR, Learn GH, He X, et al.: Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J. Virol*. 1999, **73**:10489–10502.

[46] Liu S-L, Rodrigo AG, Shankarappa R, Learn GH, Hsu L, Davidov O, Zhao LP, Mullins JI: HIV quasispecies and resampling. *Science* 1996, **273**:415–416.

[47] Rodrigo AG, Goracke PC, Rowhanian K, Mullins JI: Quantitation of target molecules from polymerase chain reaction-based limiting dilution assays. *AIDS Res. and Hum. Retrovir. 1997*, **13**:737–742.

[48] Leitner T, Escanilla D, Franzen C, Uhlen M, Albert J: Accurate reconstruction of a known HIV-1 transmission history for phylogenetic analysis. *Proc. Natl. Acad. Sci. USA* 1996, **93**:10864–10869.

[49] French MA, Herring BL, Kaldor JM, Sayer DC, Furner V, de Chaneet CC, Dwyer DE: Intrafamilial transmission of HIV-1 infection from individuals with unrecognized HIV-1 infection. *Aids* 2003, **17**:1977–1981.

[50] Herring BL: Molecular Investigations of Variation in HIV Genes. Sydney, NSW, Australia: University of Sydney: 2001.

[51] de Oliveira CF, Diaz RS, Harmache A, Frenkel LM, Gupta P, Learn GH, Mullins JI, Busch M: Passive transfer of HIV antibodies following a health care worker (HCW) accident, and transmission of HIV-1 antiretroviral resistant strain: implications for HCW post-exposure management. submitted.

[52] Higgins DG, Sharp PM: CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* 1988, **73**:237–244.

[53] Thompson JD, Higgins DG, Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994, **22**:4673–4680.

[54] Vartanian J-P, Meyerhans Åsjö B, Wain-Hobson S: Selection, recombination, and G→A hypermutation of human immunodeficiency virus type I genome. *J. Virol*. 1991, **65**:1779–1788.

[55] KewalRamani VN, Coffin JM: Weapons of mutational destruction. *Science* 2003, **301**:923–925.

[56] Korber BT, Rose P: HYPERMUT. Los Alamos, NM: Los Alamos National Laboratory; 1999.

[57] Rose PP, Korber BT: Detecting hypermutations in viral sequences with an emphasis on G→A hypermutation. *Bioinformatics* 2000, **16**:400–401.

[58] Hasegawa M, Kishino H, Yano T: Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 1985, **22**:160–174.

[59] Lanave C, Preparata G, Saccone C, Serio G: A new method for calculating evolutionary substitution rates. *J. Mol. Evol*. 1984, **20**:86–93.

[60] Yang Z: Estimating the pattern of nucleotide substitution. *J Mol Evol* 1994, **39**:105–111.

[61] Jukes TH, Cantor CR: Evolution of protein molecules. In Mammalian Protein Metabolism. Edited by Munro HN: Academic Press; **1969**:21–132.

[62] Kimura M: Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA* 1981, **78**:454–458. 63. Posada D, Crandall KA: MODELTEST: testing the model of DNA substitution. *Bioinformatics* 1998, 14:817-818.

[64] Swofford DL: PAUP* 4. 0: Phylogenetic Analysis Using Parsimony (*and Other Methods). version 4. 0b2a. Sunderland, MA: Sinauer Associates, Inc.; 1999.

[65] Huelsenbeck JP, Rannala B: Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* 1997, **276**:227–232.

[66] Akaike H: A new look at statistical model identification. *IEEE Trans Autom Contr* 1974, **19**:716–723.

[67] Yang Z: Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol*. 1994, **39**:306–314.

[68] Saitou N, Nei M: The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*. 1987, **4**:406–425.

[69] Felsenstein J: Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol*. 1981, **17**:368–376.

[70] Felsenstein J: Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 1985, **39**:783–791.

[71] Ronquist F, Huelsenbeck JP: MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 2003, **19**:1572–1574.

[72] Frenkel LM, Mullins JI, Learn GH, Manns-Arcuino L, Herring BL, Kalish ML, Steketee RW, Thea DM, Nichols JE, Liu SL, et al.: Genetic evaluation of suspected cases of transient HIV-1 infection of infants. *Science* 1998, **280**:1073–1077.

[73] Thompson WC, Taroni F, Aitken CG: How the probability of a false positive affects the value of DNA evidence. *J Forensic Sci* 2003, **48**:47–54.

[74] Hagerman PJ: DNA typing in the forensic arena. *Am J Hum Genet* 1990, **47**:876–877.

[75] Delwart EL, Shpaer EG, Louwagie J, McCutchan FE, Grez M, R¸bsamen-Waigmann H, Mullins JI: Genetic relationships determined by a DNA heteroduplex mobility assay: analysis of HIV-1 env genes. *Science* 1993, **262**:1257–1261.

[76] Anderson JP, Rodrigo AG, Learn GH, Wang Y, Weinstock H, Kalish ML, Robbins KE, Hood L, Mullins JI: Substitution model of sequence evolution for the human immunodeficiency virus type 1 subtype B gp120 gene over the C2-V5 region. *J Mol Evol* 2001, **53**:55–62.

[77] McCutchan FE, Sanders-Buell E, Salminen MO, Carr JK, Sheppard WH: Diversity of the human immunodeficiency virus type 1 envelope glycoprotein in San Francisco Men's Health Study participants. *AIDS Res Hum Retroviruses* 1998, **14**:329–337.