Journal of Clinical Virology 57 (2013) 249–253

Contents lists available at SciVerse ScienceDirect

Journal of Clinical Virology

journal homepage: www.elsevier.com/locate/jcv

Application of a new informatics tool for contamination screening in the HIV sequencing laboratory

Mark T.W. Ebbert[a,b,*,1], Melanie A. Mallory[a,1,2], Andrew R. Wilson[a,3], Shane K. Dooley[b,4], David R. Hillyard[a,c,5]

[a] ARUP Institute for Clinical and Experimental Pathology, ARUP Laboratories, 500 Chipeta Way, Salt Lake City, UT 84108, USA
[b] Department of Biology, Brigham Young University, Provo, UT 84602, USA
[c] Department of Pathology, University of Utah, 15 North Medical Drive East, Salt Lake City, UT 84112, USA

ARTICLE INFO

Article history:
Received 19 December 2012
Received in revised form 14 March 2013
Accepted 16 March 2013

Keywords:
HIV
HIVCD
Contamination
Informatics
Sequencing

ABSTRACT

Background: Current HIV-1 sequencing-based methods for detecting drug resistance-associated mutations are open and susceptible to contamination. Informatic identification of clinical sequences that are nearly identical to one another may indicate specimen-to-specimen contamination or another laboratory-associated issue.
Objectives: To design an informatic tool to rapidly identify potential contamination in the clinical laboratory using sequence analysis and to establish reference ranges for sequence variation in the HIV-1 protease and reverse transcriptase regions among a U.S. patient population.
Study design: We developed an open-source tool named HIV Contamination Detection (HIVCD). HIVCD was utilized to make pairwise comparisons of nearly 8000 partial HIV-1 pol gene sequences from patients across the United States and to calculate percent identities (PIDs) for each pair. ROC analysis and standard deviations of PID data were used to determine reference ranges for between-patient and within-patient comparisons and to guide selection of a threshold for identifying abnormally high PID between two unrelated sequences.
Results: The PID reference range for between-patient comparisons ranged from 83.8 to 95.7% while within-patient comparisons ranged from 96 to 100%. Interestingly, 48% of between-patient sequence pairs with a PID > 96.5 were geographically related. The selected threshold for abnormally high PIDs was 96 (AUC = 0.993, sensitivity = 0.980, specificity = 0.999). During routine use, HIVCD identified a specimen mix-up and the source of contamination of a negative control.
Conclusions: In our experience, HIVCD is easily incorporated into laboratory workflow, useful for identifying potential laboratory errors, and contributes to quality testing. This type of analysis should be incorporated into routine laboratory practice.

1. Background

Identifying drug resistance-associated mutations in the human immunodeficiency virus type 1 (HIV-1) protease and reverse transcriptase regions through sequence analysis is critical for disease management.[1,2] Many well-validated software are available for mutation identification and interpretation.[3–8] These programs also assess sequence quality to assure accurate base calling. Beyond resistance calling, other less recognized tools can have an important role in assuring test quality. Current sequencing-based methods require multiple capping and uncapping steps, introducing risk for specimen-to-specimen contamination. Furthermore, some specimens submitted for testing may be below the limit of sequence determination. Thus, there is a real risk for contamination of negative or low-titer specimens by high-titer specimens tested simultaneously or on previous runs. If such a contamination event

Abbreviations: HIV, human immunodeficiency virus; HIVCD, HIV Contamination Detection; PID, percent identity; ROC, Receiver Operator Characteristic; AUC, area under the curve; FDA, Food and Drug Administration; ABI, Applied Biosystems.

* Corresponding author at: 500 Chipeta Way, Salt Lake City, UT 84108, USA. Tel.: +1 801 583 2787x2289; fax: +1 801 584 5207.
E-mail addresses: mark.ebbert@aruplab.com (M.T.W. Ebbert), melanie.mallory@aruplab.com (M.A. Mallory), andrew.wilson@aruplab.com (A.R. Wilson), dooley.shanek@byu.edu (S.K. Dooley), hillyadr@aruplab.com (D.R. Hillyard).
[1] These authors contributed equally to this work.
[2] Tel.: +1 801 583 2787x2266; fax: +1 801 584 5207.
[3] Tel.: +1 801 583 2787x3359; fax: +1 801 584 5207.
[4] 673 WIDB, Brigham Young University, Provo, UT 84602, USA.
[5] Tel.: +1 801 583 2787x2202; fax: +1 801 584 5207.

1386-6532/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jcv.2013.03.013

goes undetected, a false-positive result may be reported with adverse consequences for clinical care.

The standard approach for contamination control in the sequencing laboratory includes spatial separation, unidirectional workflow,[9,10] and use of uracil N-glycosylase[11] for amplicon inactivation. Informatics analysis provides an additional level of quality control to detect contamination by identifying highly similar patient sequences. This type of analysis is possible because of the extreme variability of HIV-1 sequence,[12–19] which serves as a unique barcode for each patient. Indeed, previous studies have demonstrated the need for careful screening of HIV-1 sequence data to rule out contamination[20] and suggest that identifying a clinical sequence that is nearly identical to another clinical sequence is a marker for contamination.[21]

The two principal informatics approaches for contamination screening are phylogenetic analysis and sequence comparison based on percent identity or genetic distance.[20–23] Tools for monitoring HIV-1 inter-specimen sequence similarity include the "genetic fingerprint" tool associated with the FDA-approved TruGene HIV-1 analysis system (Siemens),[5,24,25] DBCollHIV,[22] SQUAT,[23] ViroBLAST,[26,27] BioAfrica's HIV-1 Sequence Quality Analysis Tool (http://bioafrica.mrc.ac.za/tools/pppweb.html) and the Los Alamos National Laboratory's Quality Control tool (http://www.hiv.lanl.gov/content/sequence/QC/index.html). Although useful, currently available tools are slow, closed source, incapable of batch analysis, or only available for use with a particular commercial kit.[7] Tools employing phylogenetic analysis can be complex and time consuming and may not be well suited for routine use in the clinical laboratory.[28] Furthermore, a large-scale study exploring and validating such informatics tools for contamination screening has not been previously undertaken.

2. Objectives

The aims of this study were to (i) design an open-source tool to rapidly detect potential contamination events for routine use in the clinical laboratory; and (ii) establish reference ranges for sequence variation in the HIV-1 protease and reverse transcriptase regions based on a large, unselected patient population from across the United States. In addition to contamination screening, other potential applications of the tool include detecting specimen mix-ups, verifying results for patients tested repeatedly over time, and monitoring transmission.

3. Study design

3.1. Sequencing of HIV-1 pol region

A total of 7942 partial HIV-1 pol gene sequences were generated from 6842 patients in the infectious disease sequencing laboratory at ARUP Laboratories from 2004 to 2008 using the Abbott ViroSeq HIV-1 Genotyping System (Alameda, CA). The ViroSeq HIV-1 Genotyping System generates a 1302 bp sequence of the pol gene (protease amino acids 1–99 and reverse transcriptase amino acids 1–324). Sequences from patient specimens were generated according to the manufacturer's recommendation with one modification: the specimen input volume for nucleic acid extraction was increased from 500 μL to 1 mL. Cycle sequencing was carried out using the ABI BigDye Terminator v1.1 Cycle Sequence Kit (Foster City, CA) and sequencing products were electrophoresed on an ABI Prism 3730 DNA analyzer (Foster City, CA).

3.2. Sequence data

Patient sequences included in this study represent a broad geographic area of the United States; specimens used for sequencing were collected from 360 unique ARUP client collection sites from 244 cities in 41 of the 50 United States. Patient ages ranged from 0 to 77 years with a mean of 41 years. Poor quality sequences were excluded based on the presence of: (i) >50 ambiguous or mixed bases and/or (ii) insertions or deletions.

Patient sequences were subdivided into two groups: sequences from different patients (interpatient) and sequences obtained from the same patient at various time points (intrapatient). The interpatient dataset included 6842 sequences comprised of comparisons using one sequence from each patient. The intrapatient dataset included 1960 sequences comprised of comparisons among multiple sequences obtained from the same patient. If multiple specimens from a single patient were collected on the same day, only one sequence was included by random selection in the intrapatient analysis to prevent sequence duplication. Of the 7942 total sequences, 860 were used in both interpatient and intrapatient sequence datasets.

3.3. HIVCD

To automate the process of contamination screening using sequence data, a tool named HIV Contamination Detection (HIVCD) was developed. HIVCD is a web-based application written mainly in the Perl programming language as a Common Gateway Interface (CGI) application. HIVCD consists of the multiple sequence alignment program MAFFT and a percent identity (PID) calculator,[29–36] which are written in C and Java, respectively. HIVCD uses MAFFT's default settings to create a multiple sequence alignment by invoking the '–auto' command. HIVCD subsequently performs pairwise comparisons of all sequences in the alignment and calculates a PID for each pair. PID is defined as the total number of matches divided by the total alignment length. Matches are determined strictly based on the character of the nucleotides at a given location even for mixed bases. For example, HIVCD's PID calculator considers the bases a match if both sequences exhibit 'M', but does not make special consideration for mixed bases that share a common nucleotide such as 'M' (A or C) and 'R' (A or G). The total alignment length is decreased by one if a gap at a homologous location is observed in both sequences.

3.4. Sequence and statistical analysis

To characterize inter- and intrapatient sequence variation, PIDs for all sequence pairs in both datasets were generated using HIVCD. Sequences were subtyped using the REGA HIV-1 Subtyping Tool version 2.0.[37,38] Receiver Operator Characteristic (ROC) curve analysis was used to determine whether the bimodal interpatient PID distribution was associated with subtype and to guide selection of an appropriate threshold for detecting contamination. All statistical analyses were performed in R.[39] Geographic linkage was determined by comparing the collection site zip code for each de-identified patient in a sequence pair.

4. Results

4.1. Interpatient analysis

The interpatient dataset consisted of sequences from 6842 patients, where each patient was represented by only one sequence. A total of 23,403,061 PIDs were generated from pairwise comparisons of all interpatient sequences. No two sequences were identical to one another. PIDs were distributed bimodally (Fig. 1).

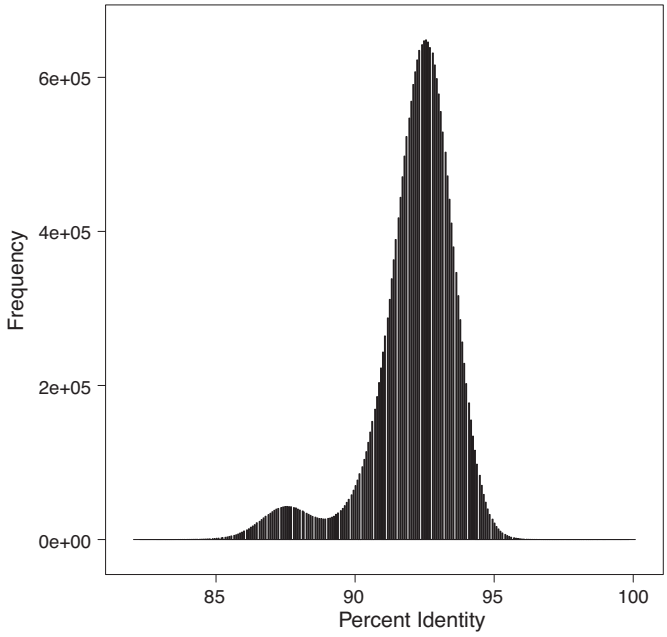

Fig. 1. Distribution of 23,403,061 PIDs from pairwise comparisons of interpatient sequences. The distribution is bimodal with no two patients having identical sequences. ROC analysis suggests the bimodal distribution is a result of within-subtype comparisons (higher distribution) and across-subtype comparisons (lower distribution) (AUC = 0.986, sensitivity = 0.979, specificity = 0.947).

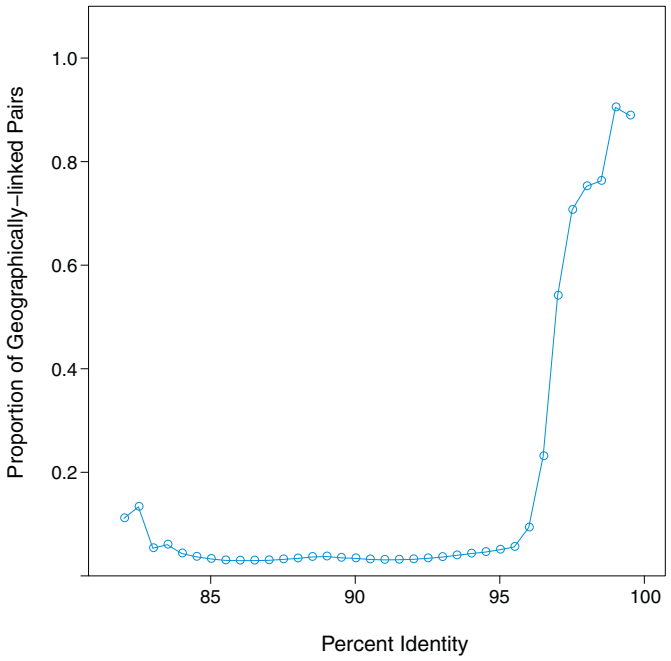

Fig. 2. Proportion of geographically-linked sequence pairs increases with PID. Geographic linkage was determined using the zip code for each de-identified patient specimen's collection site. Proportions of geographically-linked patient specimens were calculated by binning within 0.5 increments. For example, all linked patient pairs with a PID ≥ 90 and < 90.5 were binned at 90.

Subtyping suggested the smaller distribution consisted of comparisons between sequences of different subtypes and the larger distribution consisted of comparisons between sequences of the same subtype (AUC = 0.986). The mean and standard deviation of across-subtype comparisons were 87.7% and 1.3, respectively. The mean and standard deviation of within-subtype comparisons were 92.4% and 1.1, respectively. Based on three standard deviations, the normal PID range for across-subtype comparisons and within-subtype comparisons was 83.8–91.6 and 89.1–95.7, respectively.

4.2. Geographic linkage

Although no two sequences originating from different patients were identical, the maximum PID observed for any pair was 99.85%. Furthermore, the proportion of geographically-linked pairs increased with PID (Fig. 2). Approximately 48% (382 of 789) of pairs with a PID > 96.5 consisted of sequence data from specimens collected at the same geographic location based on zip code. For pairs with a PID > 99, 93% (56 of 60) were geographically linked. The geographic linkage among pairs with high identity implies an epidemiological relationship and further suggests HIVCD has potential applications in transmission surveillance. For the remaining four pairs at the 99% identity level, no clear geographic relationship could be established. These four specimens may be epidemiologically related to one another although no obvious geographic relationship exists or their high identity could indicate other issues such as specimen-to-specimen contamination. In a previous study, Learn and colleagues[20] indicated that although unexpectedly high levels of similarity may indicate an epidemiological relationship, sequence analysis alone cannot prove epidemiological linkage over a laboratory-associated problem. Such cases merit further investigation.

4.3. Intrapatient analysis

A total of 1417 PIDs were generated from the data set of 1960 sequences originating from 860 patients with sequence data available from multiple time points (Fig. 3). Interestingly, only 0.7% (10 of 1417) of sequence pairs were identical (i.e. PID = 100). The mean PID was 98.4% with a standard deviation of 1.2. Variation in the PID among sequences generated from the same patient may reflect biological adaptations of the virus over time, sampling error of quasispecies, therapy changes, and/or inherent inaccuracies of the sequencing process itself.[40] Indeed, when retesting replicates

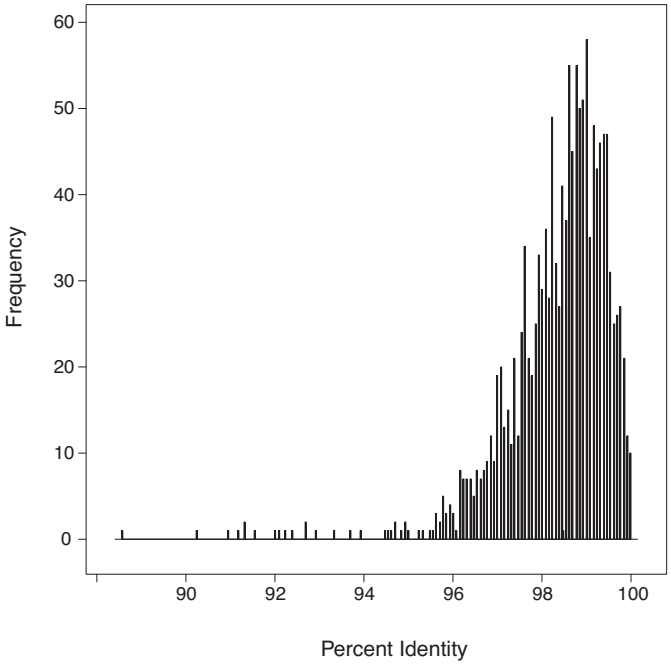

Fig. 3. Distribution of 1417 PIDs from pairwise comparisons of intrapatient sequences. The distribution is left skewed. Approximately 0.7% (10 of 1417) of sequence pairs were identical (i.e. PID = 100), 0.8% (12 of 1417) had one mismatch (PID = 99.92), and 1.4% (21 of 1417) had two mismatches (PID = 99.85).

in a dilution series of the same specimen using the ViroSeq assay, PID ranged from 98.4 to 99.4 for replicates with titers ≥250 IU/mL. Other laboratories have reported similar levels of identity when retesting the same specimen.[40–42]

The difference in days between specimen collections among specimens from the same patient varied significantly, ranging from 1 to 1526 days with a mean and standard deviation of 251 and 220 days, respectively. The difference in days between collections for the 10 identical sequences also varied significantly, ranging from 1 to 237 days with a mean and standard deviation of 58 and 75.5 days, respectively. A comparison of PID with respect to difference in collection date shows a weak downward trend over time with a correlation coefficient of −0.33, implying that intrapatient PIDs vary significantly but generally decrease over time.

4.4. Threshold selection

Based on the observation that >99% of sequences originating from different patients were 83.8–95.7% identical when compared to one another, a PID threshold of 96 was designated. Approximately 3.3% of intrapatient PIDs were <96, while 0.01% of interpatient PIDs were >96. These data support 96% as a reasonable threshold above which abnormally high interpatient PIDs can be identified without excessive false alarms (AUC = 0.993, sensitivity = 0.980, specificity = 0.999). Results from geographic linkage also support this threshold selection considering that the proportion of geographically-linked pairs remained consistently low at <0.05 as PID increased until approximately 96% identity where linkage increased dramatically.

4.5. Implementation in clinical laboratory

The HIVCD tool is currently used in ARUP's diagnostic laboratory to compare 24 sequences from new runs to each other and all sequences generated during the previous six months (∼4500 sequences). HIVCD then displays comparison results in a simple table, flagging any sequences with PIDs ≥96. This process requires minimal user interaction and takes about six minutes. Laboratory technologists then inspect patient names, specimen collection site, run date and mutation pattern of flagged sequence pairs. If the sequence pair represents the same patient or is an example of two patients with a geographic relationship, results are reported in the usual way. Again, although a geographic relationship strongly implies an epidemiological relationship, specimen-to-specimen contamination in the laboratory or at the collection site cannot be completely ruled out in such cases. If the sequences in the pair are identified in the same run or a recent run, contamination or a specimen mix-up is suspected and repeat testing is performed.

The utility of HIVCD has been demonstrated in various situations. For example, a specimen mix-up was detected during repeat testing. On a separate run, HIVCD identified the specimen that was the source of contamination of a negative control, demonstrating the tool's ability to flag the worst-case scenario of reporting a false positive. In our experience, HIVCD also accurately identifies multiple sequence examples from the same patient even when typographical errors in demographic data prevent other informatics tools from doing so.

5. Discussion

Although HIV resistance testing is widely performed by an increasing number of laboratories, the implementation of informatics quality checks beyond sequence accuracy (e.g. accuracy of base calling) has received relatively little attention. Current sequencing methods are open and vulnerable to contamination potentially leading to inaccurate reporting. Compounding the problem is the high frequency of submitted specimens with RNA levels near the limit of sequence determination.

Given this vulnerability to contamination, a specialized screening tool will help laboratories prevent misreporting. HIVCD expands on the capabilities of currently available tools used for contamination screening. For example, some tools may process large batches of sequences slowly because they were designed primarily for other uses such as virus classification[26] or assessment of base calling and alignment accuracy.[23] HIVCD, however, was designed specifically for contamination screening and can compare 24 sequences from a current run to each other and several thousand previous sequences in just a few minutes. With the advantage of this rapid analysis, HIVCD may provide additional utility in establishing epidemiological relationships among patients and transmission surveillance.[23,43,44] The source code for HIVCD is freely available at www.sourceforge.net/projects/hivcd/ and includes a user-friendly graphical interface. Because HIVCD is web-based, the tool and all necessary components including a database can be installed on a single computer, eliminating the need for a network or external Internet connection, a useful feature in resource-limited settings.

Using HIVCD, we identified 96 as a reasonable threshold above which interpatient sequence identity is considered abnormally high. For the region interrogated by the commonly-used ViroSeq HIV-1 Genotyping System, 99% of sequences originating from different patients from across the United States are 83.8–95.7% identical. Additionally, 96.7% of intrapatient sequences were at least 96% identical to one another. While other studies have suggested PID thresholds as high as 98 or 99.5,[20,23,45] we chose a threshold of 96 not only based on empirical data, but also to avoid situations involving a high PID that may lead to misreporting if not thoroughly investigated. The exact PID threshold, however, is selected arbitrarily and may vary based on the particular population analyzed or the needs of the researcher. For this reason, HIVCD's PID threshold can be modified by the user.

Although HIVCD is effective, identifying errors in the laboratory process based solely on sequence information remains challenging. Laboratory-specific errors resulting in highly similar sequences include specimen mix-ups and contamination; however, sequential testing of the same patient as well as transmission between two patients can also produce sequences with high identity. Thus, sequence analysis alone cannot differentiate between laboratory-associated problems and naturally occurring similarity.[20] Patient history and clinical findings should also be considered. Despite limitations of sequence analysis, by flagging highly similar sequences HIVCD allows laboratories to make all reasonable efforts to correct a laboratory error.

In our experience HIVCD is easily incorporated into laboratory workflow and has proved useful for identifying situations that may compromise patient care during HIV-1 resistance testing. This type of analysis contributes to quality testing and should be incorporated into routine laboratory practice.

Funding

ARUP Institute for Clinical and Experimental Pathology.

Competing interests

None declared.

Ethical approval

All patient sequences were de-identified in accordance with regulations set by the University of Utah Institutional Review Board (IRB no. 7275).

Acknowledgements

This research was funded by the ARUP Institute for Clinical and Experimental Pathology. We gratefully acknowledge Maria Erali for assistance with manuscript preparation and Mitchell T. Sears for technical assistance.

References

1. Hirsch MS, Günthard HF, Schapiro JM, Brun-Vézinet F, Clotet B, Hammer SM, et al. Antiretroviral drug resistance testing in adult HIV-1 infection: 2008 recommendations of an International AIDS Society-USA panel. *Top HIV Med* 2008;**16**:266–85.
2. Vandamme AM, Camacho RJ, Ceccherini-Silberstein F, de Luca A, Palmisano L, Paraskevis D, et al. European recommendations for the clinical use of HIV drug resistance testing: 2011 update. *AIDS Rev* 2011;**13**:77–108.
3. De Luca A, Perno CF. Impact of different HIV resistance interpretation by distinct systems on clinical utility of resistance testing. *Curr Opin Infect Dis* 2003;**16**:573–80.
4. Eshleman SH, Crutcher G, Petrauskene O, Kunstman K, Cunningham SP, Trevino C, et al. Sensitivity and specificity of the ViroSeq human immunodeficiency virus type 1 (HIV-1) genotyping system for detection of HIV-1 drug resistance mutations by use of an ABI PRISM 3100 genetic analyzer. *J Clin Microbiol* 2005;**43**:813–7.
5. Grant RM, Kuritzkes DR, Johnson VA, Mellors JW, Sullivan JL, Swanstrom R, et al. Accuracy of the TRUGENE HIV-1 genotyping kit. *J Clin Microbiol* 2003;**41**:1586–93.
6. Lengauer T, Sander O, Sierra S, Thielen A, Kaiser R. Bioinformatics prediction of HIV coreceptor usage. *Nat Biotechnol* 2007;**25**:1407–10.
7. Liu TF, Shafer RW. Web resources for HIV type 1 genotypic-resistance test interpretation. *Clin Infect Dis* 2006;**42**:1608–18.
8. Zazzi M, Prosperi M, Vicenti I, Di Giambenedetto S, Callegaro A, Bruzzone B, et al. Rules-based HIV-1 genotypic resistance interpretation systems predict 8 week and 24 week virological antiretroviral treatment outcome and benefit from drug potency weighting. *J Antimicrob Chemother* 2009;**64**:616–24.
9. The Clinical Laboratory Improvement Amendments (CLIA) regulations; laboratory requirements, 42 C.F.R. Part 493; 2004.
10. Chen B, Gagnon M, Shahangian S, Anderson NL, Howerton DA, Boone DJ. *Good laboratory practices for molecular genetic testing for heritable diseases and conditions*; 2009.
11. Longo MC, Berninger MS, Hartley JL. Use of uracil DNA glycosylase to control carry-over contamination in polymerase chain reactions. *Gene* 1990;**93**: 125–8.
12. Kijak GH, McCutchan FE. HIV diversity, molecular epidemiology, and the role of recombination. *Curr Infect Dis Rep* 2005;**7**:480–8.
13. Ramirez BC, Simon-Loriere E, Galetto R, Negroni M. Implications of recombination for HIV diversity. *Virus Res* 2008;**134**:64–73.
14. Pandrea I, Descamps D, Collin G, Robertson DL, Damond F, Dimitrienco V, et al. HIV type 1 genetic diversity and genotypic drug susceptibility in the Republic of Moldova. *AIDS Res Hum Retroviruses* 2001;**17**:1297–304.
15. Makuwa M, Souquiere S, Apetrei C, Tevi-Benissan C, Bedjabaga I, Simon F. HIV prevalence and strain diversity in Gabon: the end of a paradox. *AIDS* 2000;**14**:1275–6.
16. Gaschen B, Taylor J, Yusim K, Foley B, Gao F, Lang D, et al. Diversity considerations in HIV-1 vaccine selection. *Science* 2002;**296**:2354–60.
17. Korber B, Gnanakaran S. The implications of patterns in HIV diversity for neutralizing antibody induction and susceptibility. *Curr Opin HIV AIDS* 2009;**4**: 408–17.
18. Coffin JM. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science* 1995;**267**:483–9.
19. Wain-Hobson S. The fastest genome evolution ever described: HIV variation in situ. *Curr Opin Genet Dev* 1993;**3**:878–83.
20. Learn Jr GH, Korber BT, Foley B, Hahn BH, Wolinsky SM, Mullins JI. Maintaining the integrity of human immunodeficiency virus sequence databases. *J Virol* 1996;**70**:5720–30.
21. Korber BT, Learn G, Mullins JI, Hahn BH, Wolinsky S. Protecting HIV databases. *Nature* 1995;**378**:242–4.
22. Araújo LV, Soares MA, Oliveira SM, Chequer P, Tanuri A, Sabino EC, et al. DBColl-HIV: a database system for collaborative HIV analysis in Brazil. *Genet Mol Res* 2006;**5**:203–15.
23. Delong AK, Wu M, Bennett D, Parkin N, Wu Z, Hogan JW, et al. Sequence quality analysis tool for HIV type 1 protease and reverse transcriptase. *AIDS Res Hum Retroviruses* 2012;**28**:894–901.
24. Gale HB, Kan VL, Shinol RC. Performance of the TruGene human immunodeficiency virus type 1 genotyping kit and OpenGene DNA sequencing system on clinical samples diluted to approximately 100 copies per milliliter. *Clin Vaccine Immunol* 2006;**13**:235–8.
25. Perandin F, Pollara PC, Gargiulo F, Bonfanti C, Manca N. Performance evaluation of the automated NucliSens easyMAG nucleic acid extraction platform in comparison with QIAamp Mini kit from clinical specimens. *Diagn Microbiol Infect Dis* 2009;**64**:158–65.
26. Deng W, Nickle DC, Learn GH, Maust B, Mullins JI. ViroBLAST: a stand-alone BLAST web server for flexible queries of multiple databases and user's datasets. *Bioinformatics* 2007;**23**:2334–6.
27. Heath L, Conway S, Jones L, Semrau K, Nakamura K, Walter J, et al. Restriction of HIV-1 genotypes in breast milk does not account for the population transmission genetic bottleneck that occurs following transmission. *PLoS ONE* 2010;**5**:e10213.
28. Bernard EJ, Azad Y, Vandamme AM, Weait M, Geretti AM. HIV forensics: pitfalls and acceptable standards in the use of phylogenetic analysis as evidence in criminal investigations of HIV transmission. *HIV Med* 2007;**8**:382–7.
29. Katoh K, Asimenos G, Toh H. Multiple alignment of DNA sequences with MAFFT. *Methods Mol Biol* 2009;**537**:39–64.
30. Katoh K, Frith MC. Adding unaligned sequences into an existing alignment using MAFFT and LAST. *Bioinformatics* 2012.
31. Katoh K, Kuma K, Miyata T, Toh H. Improvement in the accuracy of multiple sequence alignment program MAFFT. *Genome Inform* 2005;**16**:22–33.
32. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 2005;**33**:511–8.
33. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 2002;**30**:3059–66.
34. Katoh K, Toh H. Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. *BMC Bioinformatics* 2008;**9**:212.
35. Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* 2008;**9**:286–98.
36. Katoh K, Toh H. Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics* 2010;**26**:1899–900.
37. Alcantara LC, Cassol S, Libin P, Deforche K, Pybus OG, Van Ranst M, et al. A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. *Nucleic Acids Res* 2009;**37**:W634–42.
38. de Oliveira T, Deforche K, Cassol S, Salminen M, Paraskevis D, Seebregts C, et al. An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics* 2005;**21**:3797–800.
39. R Core Team. *R: a language and environment for statistical computing*. 2.15.1 ed. Vienna, Austria: R Foundation for Statistical Computing; 2012.
40. Galli RA, Sattha B, Wynhoven B, O'Shaughnessy MV, Harrigan PR. Sources and magnitude of intralaboratory variability in a sequence-based genotypic assay for human immunodeficiency virus type 1 drug resistance. *J Clin Microbiol* 2003;**41**:2900–7.
41. Shafer RW, Hertogs K, Zolopa AR, Warford A, Bloor S, Betts BJ, et al. High degree of interlaboratory reproducibility of human immunodeficiency virus type 1 protease and reverse transcriptase sequencing of plasma samples from heavily treated patients. *J Clin Microbiol* 2001;**39**:1522–9.
42. Eshleman SH, Hackett Jr J, Swanson P, Cunningham SP, Drews B, Brennan C, et al. Performance of the Celera Diagnostics ViroSeq HIV-1 Genotyping System for sequence-based analysis of diverse human immunodeficiency virus type 1 strains. *J Clin Microbiol* 2004;**42**:2711–7.
43. Paraskevis D, Pybus O, Magiorkinis G, Hatzakis A, Wensing AM, van de Vijver DA, et al. Tracing the HIV-1 subtype B mobility in Europe: a phylogeographic approach. *Retrovirology* 2009;**6**:49.
44. Hue S, Clewley JP, Cane PA, Pillay D. HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *AIDS* 2004;**18**:719–28.
45. Shafer RW, Hsu P, Patick AK, Craig C, Brendel V. Identification of biased amino acid substitution patterns in human immunodeficiency virus type 1 isolates from patients treated with protease inhibitors. *J Virol* 1999;**73**:6197–202.