Sequence Quality Control

Carla Kuiken and Bette Korber

MS K710, Los Alamos National Laboratory, Los Alamos, NM 87545

Some HIV researchers are convinced that careful lab work is enough to prevent contamination. After seeing many examples of contamination pass by, we disagree. Contamination happens frequently, even in the best laboratories. It is not a thing of the past; cases of contamination can still be found in new publications. Screening for contamination should be done before the analysis of the sequences, and periodically during the course of large sequencing studies, so problems can be detected and corrected early.

To show what contamination looks like in practice, we have collected some illustrative examples of (mostly published) datasets where contamination is a problem, and included some references that discuss instances and consequences of contamination.

Following the steps below will help to check your sequences. They are no substitute for common sense and precautions, but they may help reveal contamination in your sequences. We have created interactive Web pages

☞ **http://hiv-web.lanl.gov/HTML/Contam/contam_main.html**

where you can build a tree and do a BLAST search with your sequences. We include some tips for identifying problem sequences in conserved regions of the HIV genome such as protease or reverse transcriptase (RT).

1. **Create a phylogenetic tree that includes all the sequences in the study.** Common signs of trouble are:

   - Extreme intrapatient divergence
   - Extreme interpatient similarity
   - Mixed clusters (sequences from patient A clustering with patient B)

   A phylogenetic tree can clarify the relations between the sequences. If you have lab strain contamination or sample mix-ups between two patients, a phylogenetic tree will likely show it. Once you have your sequences aligned, use our web site to generate a simple neighbor-joining tree (Saitou and Nei, 1987) to check for potential problems. Neighbor-joining is a computationally fast phylogenetic method that can easily handle hundreds of sequences in a single tree.

2. **Compare your sequences to all published sequences (BLAST search).**

   BLAST is a program that finds sequences with similarity to the query sequence (Altshul, et al., 1990). The output can be ordered by the degree of similarity. If your sequence is very similar to a published strain, especially a lab strain that is used for *in vitro* studies, it is likely that you have contamination. Even if your sequence is not identical to the lab strain, watch out for *in vitro* recombination, in which only part of the sequence matches the lab strain, and the other part is derived from your patient sample (see example 2 below). On our Web site you can compare your sequences to all Genbank entries, which contains the very latest sequences, or against the Los Alamos HIV database which can lag behind Genbank a bit, but contains more background information about the sequences.

What degree of similarity indicates contamination depends on the gene or region being analyzed. Thus, RT sequences are much less variable than V3 sequences. Figure 1 shows the frequency distribution of similarity scores for different genes. The population from which the samples have been obtained will also influence the degree of similarity to be expected. Sets of clonal sequences from different tissues in a single patient will tend to be more similar than sets from different persons in a clustered outbreak, which in turn will tend to be less similar than sequences from geographically disperse locations.

3. **Look carefully at the alignments, and pay attention to patient signature patterns.**

   Signature patterns (Korber and Myers, 1992) often help to show what is "typical" and "atypical" for a patient, and thus help to reveal sequences that don't seem to belong with a patient. The usefulness of signature patterns can be seen in the contamination example 3 below. You can use the Vespa program (http://hiv-web.lanl.gov/HTML/vespa.html) to find signature patterns, but often a simple alignment is sufficient to spot suspicious sequences.

4. **Keep a background set of sequences that are commonly used in your laboratory for comparison.**

   BLAST searches can detect contamination by common lab strains whose sequences are entered in Genbank, but contamination with other genetic material that was recently used in your lab may go undetected. Aligning sequences that look suspicious with other sequences that your lab has produced may bring this type of contamination to light.

Some examples of contamination.

We have selected a few datasets to illustrate the problems that can arise, and how they can be recognized. The sets are anonymous and unrecognizable; the purpose is to show real examples of contamination, not to cast blame on any particular person or group.

Example 1: A set of C1-C3 sequences containing LAI/HXB2 contamination and sample mix-ups (partial set, published). This partial dataset contains several examples of possible contamination. The tree is shown in Figure 2. Signs of trouble in this tree:

1. Sequences from patient F spread over three clusters. One cluster is very similar to HXB2/LAI and is probably a laboratory contaminant. The distinctness of the other two clusters suggests either dual infection, contamination with an isolate for which there is no sequence in Genbank, or mix-up with a patient that's not in the study.
2. Patients E and G both have a single sequence that clusters tightly with the other patient, suggesting a sample mix-up or mislabeling.

Example 2: This dataset contains three sequences, labeled 59, 77, and 65 in Figures 3 and 4, that are the result of *in vitro* recombination between the viral DNA from the patient and LAI/HXB2 DNA. In the tree, (Figure 3) three sequences clearly cluster with the LAI clone. That *in vitro* recombination has occurred can be seen very clearly in the alignment (Figure 4). The three recombinant sequences match the LAI sequence perfectly in the latter half of the alignment whereas the other sequences in the study do not.

Example 3: This set was generated to study CTL epitope variation, and consists of partially overlapping sequence fragments of variable length. Phylogenetic analysis was impossible because the sequences had too little overlap to create a tree, but a BLAST search and an alignment with the most similar Genbank sequence showed extensive contamination with pNL43. Figure 5 shows the sequences aligned to pNL43.

Yellow bars indicate sequences whose best BLAST match was to pNL43 and are thus considered contaminants. Signature patterns, shown as colored rectangles, are characteristic for each patient and are notably absent from the contaminant sequences. Although identity of one individual sequence to a lab strain would not conclusively prove contamination, all evidence taken together is very strong:

- the offending sequence in all cases was NL43
- NL43 was used in this laboratory
- all other patients' sequences had clear characteristic signatures that were not shared by NL43.

The special case of conserved genes.

Even if your region contains very little variation, there are ways to increase confidence in the validity of the sequences. Therefore, we will elaborate a bit on the problem of conserved genes.

The low variability of protease and RT and the occurrence of mutations associated with drug resistance make detecting problems more challenging, especially in short sequences. Phylogenetic analysis of the protease gene sometimes shows patient intermixing that does not necessarily indicate contamination. But even in these genes it is possible to get more information about the sequence quality. In addition to the methods suggested above, here are some other analyses that can help exclude contamination. If sequences from a patient cluster with sequences from another patient, contamination is likely. If they do not cluster with other sequences from the set, contamination is less plausible (but doesn’t exclude it). If patient sequences do not cluster together in a phylogenetic tree, check if the separation is associated with drug resistance; it can be the result of selection rather than contamination. Making a tree excluding the resistance-associated positions can sometimes resolve the problem. If a sequence from one patient clusters with sequences from another patient, check if they were PCR amplified on the same day. If so, this sequence is more likely to be suspect. Make a tree based on synonymous substitutions only. This can in some cases reunite all sequences from one patient into the same cluster and confirm their validity. Look for signature patterns that are characteristic of a patient and see if they are preserved in the outliers in question.

Analysis of synonymous substitutions may help to validate a sequence dataset (Figures 6 and 7). The dataset consists of 421 clonal culture-derived sequences from varying numbers of samples from pre- and post-treatment samples of 21 patients. A BLAST search revealed no indication of lab strain contamination (not shown). Only four single sequences clustered with other patients, and were probably contaminants or mix-ups, an excellent result for a study of this size. Here we will focus on samples from some patients that show unexpected clustering. A representative neighbor-joining tree containing sequences from three ‘well-behaved’ (N, P, U) and one ‘strange’ patient (K) is shown in Figure 2. This tree is based on all—both synonymous and nonsynonymous—nucleotide differences between sequences. Sequences from patient K form two separate clusters, separated by sequences from two other patients. It is possible that this behavior is caused by the emergence of a drug-resistant strain of virus. The virus isolated at week 60 was highly resistant to Indinavir, while the virus from week 0 and week 18 was not. To diminish the influence of drug-driven selection of amino acids at critical positions in the protein, we looked at only synonymous changes, using the program SNAP that was developed here at Los Alamos; the phylogenetic analysis programs MEGA (for PC) and Phylowin (for UNIX) can do the same thing. The tree in Figure7 is based on only synonymous changes. All sequences from patient K now cluster together, which makes it very plausible that they are indeed from the same patient, rather than a sample mix-up or cross-contaminant. Not all cases are as clear-cut. In this dataset, six patients showed unexpected clustering. In three of these, the sequences came together in the synonymous tree, and therefore most likely were legitimate. In one other case, the existence of a sequence from plasma from the same patient obtained and handled seperately from the PBMC sample indicated that the outlying cluster was valid. Two other cases couldn't be resolved; in view of the overall quality of the dataset and in the absence of strong evidence to the contrary, we considered them likely to be valid.

We thank Dr. Jon Condra and Dr. Andrew Leigh-Brown for help with the analysis and further consideration of the protease sequence set.

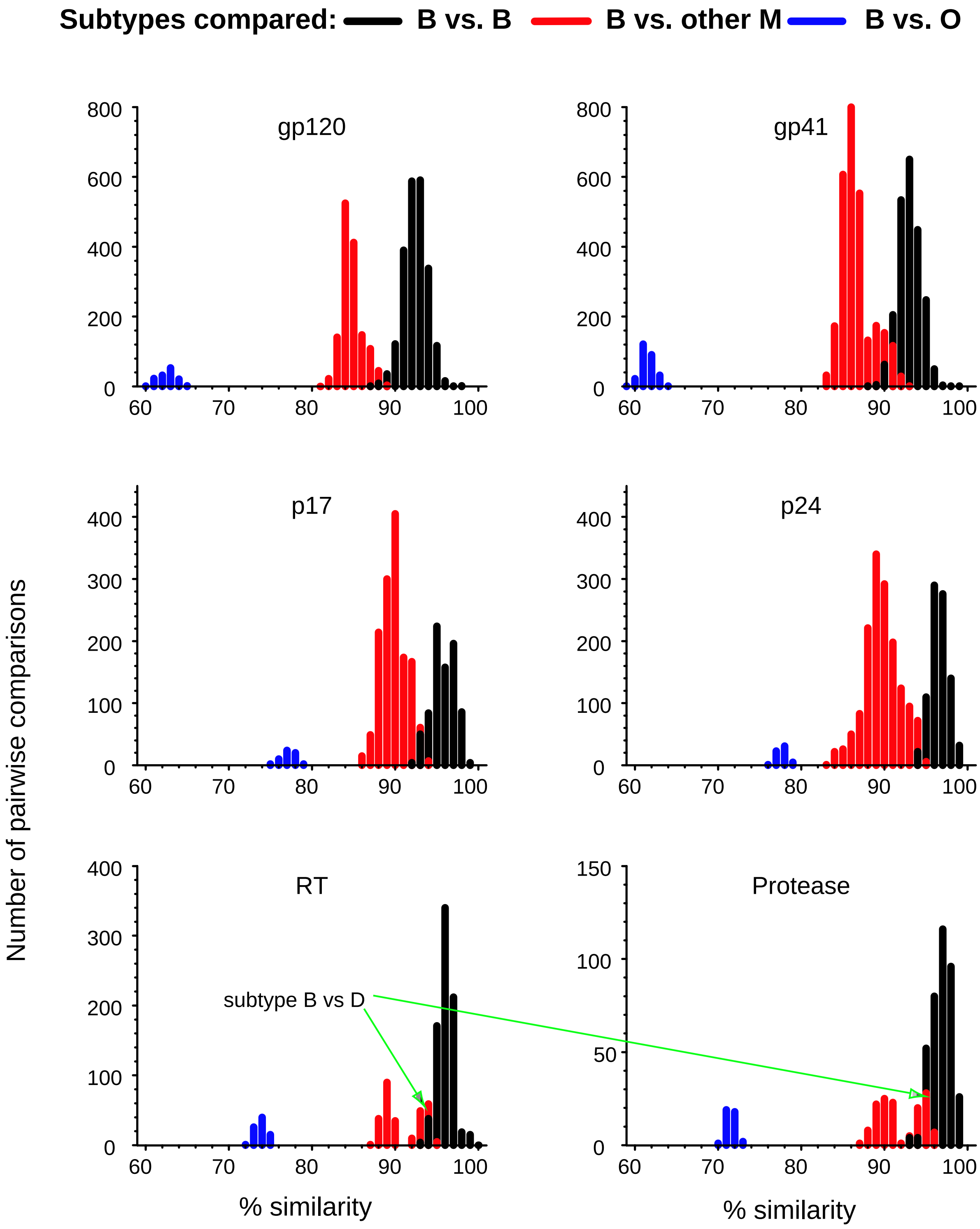

Figure 1. Frequency distribution of similarity scores for different genes.

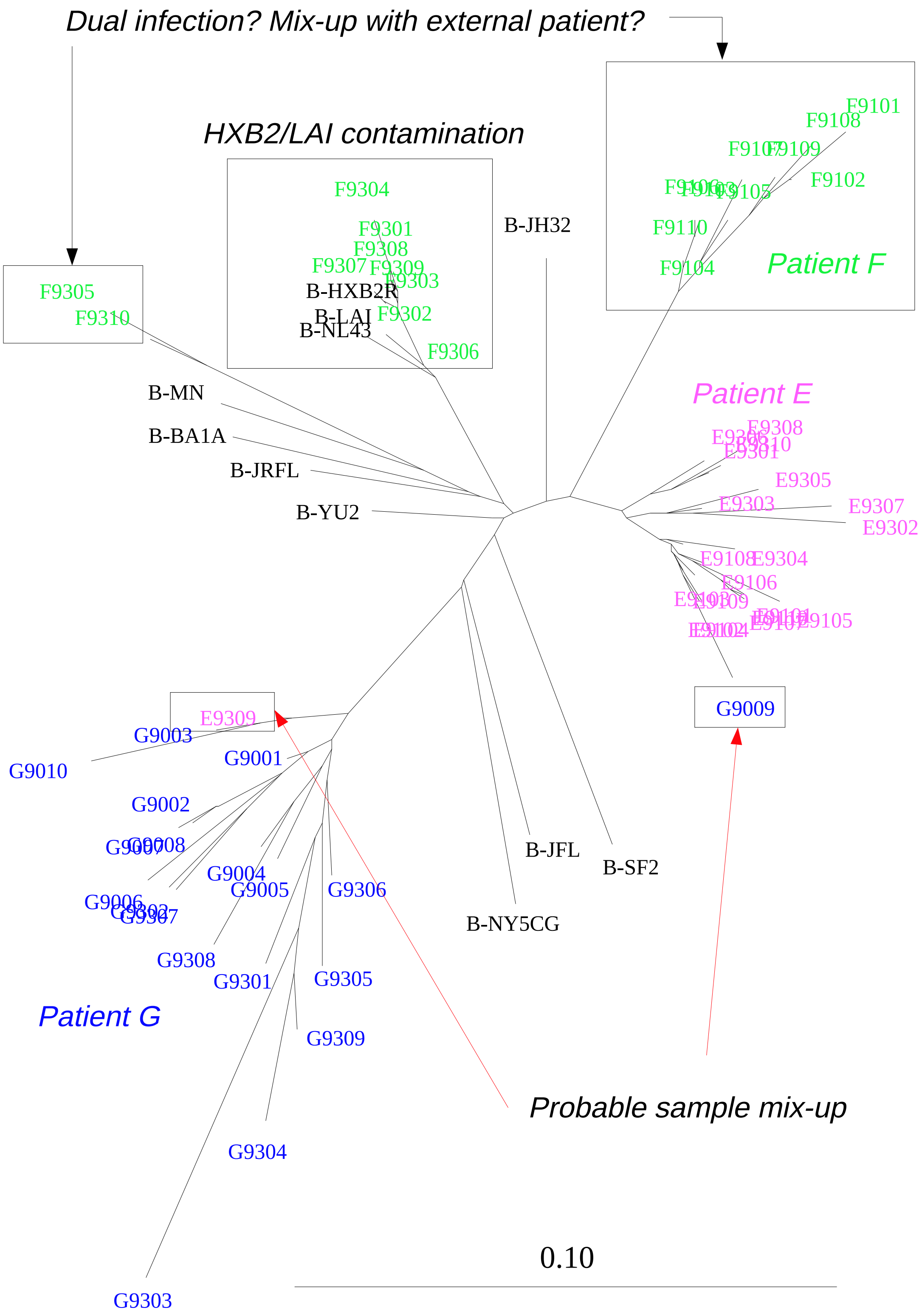

Figure 2. LAI/HXB2 contamination and sample mix-ups

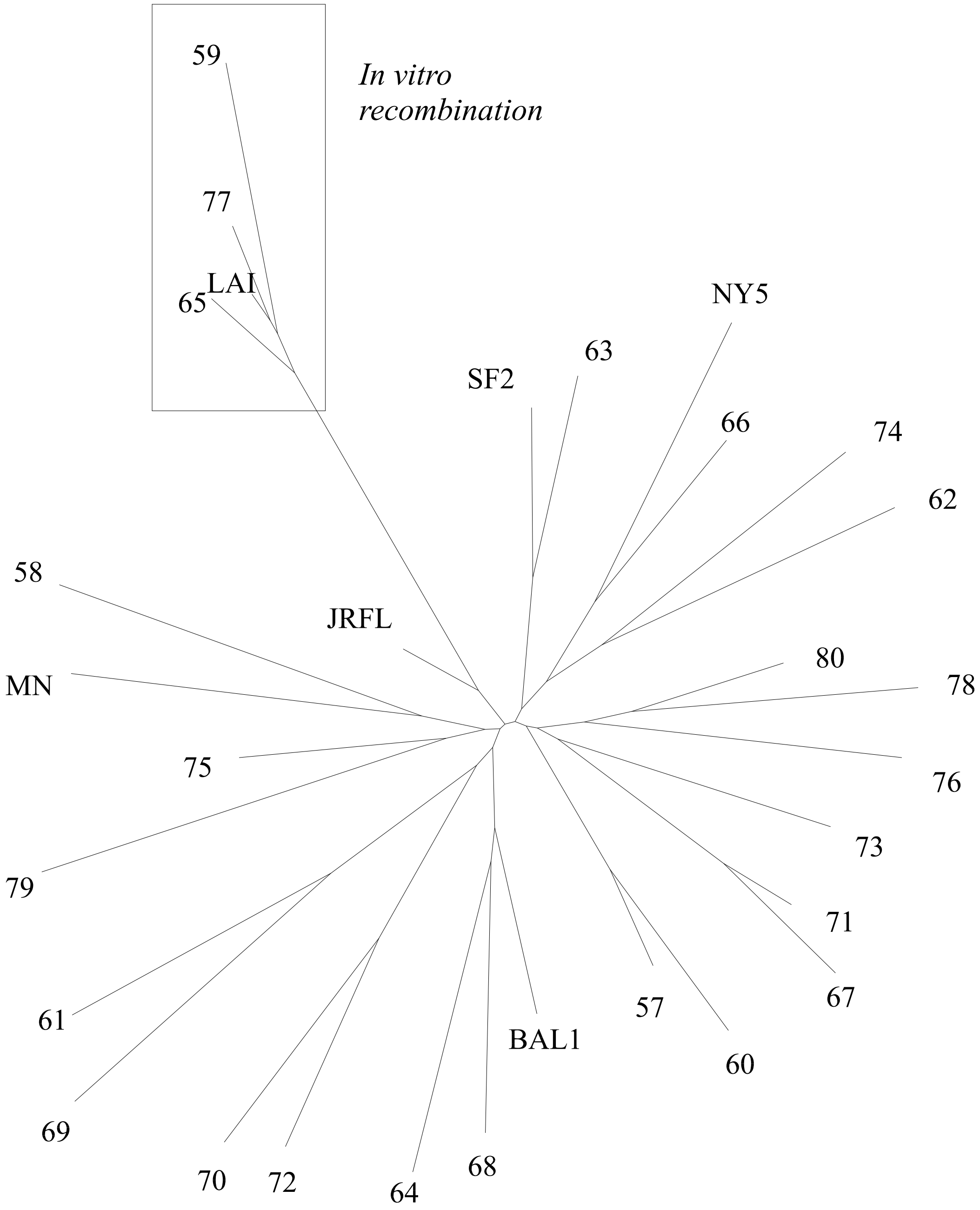

Figure 3. Three recombinant sequences clustering with LAI.

```
LAI   CTAGCAGAAGAAGAGGTAGTAATTAGATCTGCCAATTTCACAGACAATGCTAAAACCATAATAGTACAGCTGAACCAATCTGTAGAAATTAATTGTACAAGACCCAACAACAATACAAG

59    -------------------------------AA------G--------CG-------------CAG-T-----------------C---------------------------------
77    -------------------------------T---------C----T--------------------------------------C---------------------------------
65    -------------------------------T---------GA--------------------------------AC------------------------------------------
80    -------------G-----------------AA--A------------A-------------------------AG-------------------------------------------
79    ------------------A------------AA---C------------------------------GA-----TG--A----------------------------------------
78    -------------G-----------------AA--A--------------------------------------AG-----A--A---A---------T--------------------
76    -------------G-----------------A----A-----A-T-------G--------------------CAG-CA------T---------------------------------
75    ------------------A------------AA--------GA-------------------------------TG--A-----C----------------------------------
74    ------------A-------------G----AG------CGGA-------------------------------TG-------------------------------------------
73    -------------------------------AA--------C-----------------------G---------G-------------------------------------------
72    -------------------------------AAG--C--T-GA-------------------------------TG----A--------------------------------------
71    -------------G-----------------A---A-----GA-------------------------------TA-------------------------------------------
70    ------------A------------------A----C----G--------------------------------TG----A-------A------------------------------
69    -----------G-------------------AA---------A----------------------------A--TG-----A---C----G-G----T---------------------
68    -----------------------------A--A--------GA---------G--T------------------TG---------C---------------------T-----------
67    -------------------------------AA-CA-------------------------------------GGA-------------------------------------------
66    -------------G------------G----AA--------GA------T------------------------TG-------------------------------------------
64    -----------------------------C-AA----------------------T------------------AG--------C---------------------G------------
63    -------------G-----------------AA--------GA------T-------------------T-A--TG---------C------C--------------------------
62    --------------------------G----AA------CGGA----------------------G---------G--A-----C------------T--------G------------
61    -------------------------------AA--------GA------------T------------------TG-----A--A-----G-G--------------------------
60    ----T--------------------------AA------CGCA-------------------------------AG-CC------T---------------------------------
58    ------------A------------------AG----------------------------------------TTG-T--------------------------------T----A---
57    -------------------------------AA--------GA-------------------------------A---C------T---------------------------------
SF2   -------------------------------A---------GA-------------------------------TG---------C------C--------------------------
NY5CG ----------G---------------G----AA--------GA------T------------------------TG-------------------------------------------
BAL1  -----------------------------C---------G-G------------GT------------------TG-------------------------------------------
JRFL  -------------------------------A---------GA-------------------------------AG-------------------------------------------
MN    -------------------------------AG--------T--T--------------C--------T-----TG--------C-----------------------T------A---

LAI   AAAAAGTATCCGTATCCAGAGGGGACCAGGGAGAGCATTTGTTACAATAGGA...AAAATAGGAAATATGAGACAAGCACATTGTAACATTAGTAGAGCAAAATGGAATGCCACTTTAAA

59    ----------------------------------------------------...-----------------------------------------------------------------
77    ----------------------------------------------------...-----------------------------------------------------------------
65    ----------------------------------------------------...-----------------------------------------------------------------
80    ---------AAA---A......------------------TA-----C----GAC-T-------G----A-------------------------C-------------AA---C----T
79    ---------AAC---G......------------A-----TA-G---C----GAAGT------------C------------------C---A----------------AA---------
78    ---------A-A---A......------------------TA-G---C----AAC-T-------G----A-----------------TC-----------G-C------AA---C-----
76    ---------A-A---A......------------------TA-G---C----GAA-T-------G----A---------------------------------------AA---C----G
75    ---------AAA---A......------------------TA-----C----GAA-T-------G----A--------------------------------G---G-GAA---------
74    ---------A-A---A......------------------TA-----C----GAA-T-------G----A---------T-------TG------T-------------AGG--C----T
73    ---------A-A---A......-----------G------TA-G--TC----GAA-T-------G----A------------------T-A---GA-------------AA---------
72    --G------A-A---G......----------A----A--TA-----C----GCA-T-------G----A----G----T---------------T-A-G---------AA---------
71    ---------A-A---A......--------A---------TA-G---C-...GAT-T------------A------------------------G---T-G--------AA-------G-
70    ---------ATC---G......----------AG---A--TA-----C----GCA-T------------A---------T---------------T-A-----------AA--------T
69    ---------A-A---A......-----G----A------GTA-----C----...GT-------G----A---------T--------C---A----A------------A--------G
68    ---------A-C---A......--------C-A-------TA-----C----GAA-T-------G----A------------------C------A----G--------AA--------G
67    ---------A-A---A......--------A---------TA-G---C-...GAT-T------------A------------------------G---T-G--------AA-------G-
66    C---G----A-A---A......------------A-----TA-G---C----AGA-T-------G----A-------------------------T-A-----------AA---C----T
64    ---------A-C---A......--------C-A--TC--GTA-----C----GAA-T-------G----A-----G------------C--------------------AA--------C
63    ---------ATC---A......------------------TA-G---C----GAC-T-------G----A---A--------------C-----------C--------AA---------
62    ---------A-A---A......------------------TA-G---C----GAC-T-------G----A------------------------G-----C--------AAA--C-----
61    --G----G-A-A-G-A......-----G----A------GTA-----C-...CAT-T-------G-C--A--------------------------------C------AA--------G
60    ---------A-A---A......------------------TA-----C----GAA-T------G-----A------------------C--------T--G--------AA---------
58    ---------A-A---A......-C----------------TA-G---C----GTA-T-------G----A--------C---------C--------A--------G-A-A---------
57    ---------A-A---A......------------------TA-----C----GAA-T------------A------------------C--------T--GC-------AA--------C
SF2   ----------TA---A......------------------CA-----C----AGA-T-------G----A---A--------------------------C--------AA-------G-
NY5CG ----G----AGC---A......-----G------A--C-CTA-G---G--A-AAA-T-------G----A------------------C---------------------A---------
BAL1  ---------A-A---A......--------C---------TA-----C----GAA-T-------G----A------------------C---------------------A---------
JRFL  ---------A-A---A......------------------TA---T-C----GAA-T-------G----A----------------------------------------A---------
MN    ------G--A-A---A......------------------TA-----C-AA-AAT-T--------C---A----------------------------------------A--------G
```

Figure 4. Alignment of samples in Fig 3 showing perfect match with LAI in shaded region.

```
NL4-3        MGARASVLSGGELDKWEKIRLRPGGKKQYKLKHIVWASRELERFAVNPGLLETSEGCRQILGQLQPSLQTGSEELRSLYNTIAVLYCVHQRIDVKDTKEALDKIEEEQNK
CONSENSUS-B  --------------r------------k-----------------------------------------------------v-t--------e-----------------
pat01-B6                       ---------R-----L--------------------------------A--------K-------T-----
pat01-B2                       ---------R-----L-----------------------K--------A--------K-----V-T--------EI
pat01-B20                      --------------------------------------------------------------------------
pat01-C15                     ----------R-----L-----------------------K--------A--------K-----V-T--------E-
pat01-C14                     ----------R-----L-----------------------K--------A--------K-----V-T--------E-
pat01-C18                     ----------R-----LI----------------------K--------A--------K--F--V-T--------E-
pat01-C20                     ---------------------------------------------------------------------------E-
pat01-B22                        -------R-----L-----------L-----------K--------A--------K--F--V-T--------
pat01-A8                       ------------------------------------------------------------------------
pat01-A5                                         ---------L-----------K--------A--------K--F--V-T-----

pat02-D7                   -----------R-K-Q---L-------------------AD------E------K------------V-T--------EI-----------
pat02-D4                   -----------R-K-Q---L-------------------AD------E----A-K---------F--V-T--------EI-----------
pat02-D40                  -------------------------------------------------------------------V-T---------I------
pat02-D39                  ---R-------R-K-------------------------AD------E----A-K---------F-------------EI-----------
pat02-C6                   -----------R-K-Q---L-------------------AD------E----A-K---------F--V-T------Q-E------------
pat02-C17                  -----------R-K-------------------------AD------E------K---------F--V-T------Q-E------------
pat02-B17                  -----------R-K-Q---L-------------------AD------E----A-K---------F--V-T------Q-E------------
pat02-B15                  -----------R-K-------------------------AD------E------K---------F--V-T------Q-E--------
pat02-B14                  -----------R-K-------------------------AD------E----A-K---------F--V-T------Q-E--------
pat02-B13                  -----------R-K-Q---L-------------------AD------E----A-K---------F--V-T------Q-E-------
pat02-A9                   R----------R-K-------------------------AD------E------K---------F--V-T------Q-E--Y---------

pat03-C17                                               -------------------------A-------FK----LV---------
pat03-B2                                                            -------------A-------FK----LV---------
pat03-B18                                                  -------------------------------K-----V-T-------
pat03-B15                                                               ---------------------------------
pat03-A6                 --R---------S--S-----------------L--------------------A-------FK----LV----------EI----------------
pat03-A16                --R------K--S--------------------------------------------------------V----------EI----------------
pat03-A13                --R--R------S--R-----------------L------------------H----------K-----V-T--------EI--------R-----S-
pat03-C20                                                           -----------H----------K-----V-T------
pat03-A9                 --R--R------S--R-----------------L-----------------------------K-----V-T--------EI--------R-----S-

pat04-A9                                                -----------------------------------------------------
pat04-A7                                                -----------------------------------------------------
pat04-A5                                                -------------G---------------------------------------
pat04-A3                                                ----------------------------------K-------T------G-E-
pat04-A19                                               --------------------R-------------K-------T------G-E-
pat04-B2                                                ----------------------------------K-------T------G-E-R---
pat04-B18                                               ----------------------------------K-------T--------E-
pat04-C9                                                -----------------------------------------------------
pat04-C6                                                -----------------------------------------------------
pat04-C4                                                -----------------------------------------------------
pat04-C10                                               -----------------------------------------------------
pat04-C2                                                ----------------------------------K-------T------G-E-

pat05-9A                                               ----------------------------------------------------------------------
pat05-8A                                               ----------------------------------------------------------------------
pat05-7A                                               ------------------------------------------------------------
pat05-6A                                               ------------------------------------------------------
pat05-4A                                               ------------------------------------------------------
pat05-12A                                                       -------------------------------------------------------------
pat05-11A                                                       -----------------------D-------------------
pat05-10A                                              ----------------------------------------------------------------------

pat06-C5                                                         --------------------------------------------
pat06-C3                                                       ----------------------------------------
pat06-C                                               -----------------------------------------
pat06-B9                      ----------K----------K---------------------------A---R-------F--V-T----
pat06-B5               ----R--------------------------------------------------------------------------
pat06-B4                  -R----------------------------------------------------------------------------
pat06-B2                      ----------K--------------------------------------A-----------F--V-T---
pat06-B19                   --------------------------------------------------------------------------
pat06-B12                   --------------------------------------------------------------------
pat06-B10                   -----K--------------------------------------T-----------F--V-T------
pat06-A15              ----R---------------------------------------------------------------------------------
pat06-A14                 -R-----------------------------------------------------------------------------

pat07-9                       -------------------------------------------------------------------------
pat07-6                       --------R--------------------------------------A--------K-----V-T------K-E-
pat07-3                       -------------------------------------------------------------------------
pat07-                        -------------------------------------------------------------------------
pat07-16                      --------R--------------------------------------A--------K-----V-T------K-E-
pat07-15                      --------R--------------------------------------A--------K-----V-T------K-E
pat07-14                      -------------------------------------------------------------------------
pat07-13                      --------R--------------------------------------A--------K-----V-T------K-E--
pat07-12                      --------R--------------------------------------A--------K-----V-T------K-E
pat07-1                       --------R-----------------------------K--------A--------K-----V-T------K-E-

pat08-9A                       ---------R------------------------G--------------------Q--------------
pat08-8A                       ----------------------------------------------------------------------
pat08-7A                                           --------------------------------------------------
pat08-6A                                          -----------------------------------------------------
pat08-5A                                           --------------G--------------------------------
pat08-4A                       -------------------------------------------------------------------
pat08-3A                                          -----------------------------------------------------EI--
pat08-11A                      ---------R------------------------G---------H----------Q-----------
```

Figure 5. Alignment of samples to pNL43. Yellow bars: contaminant sequences, best BLAST match to pNL43. Other colors: patient signature patterns.

Analyses

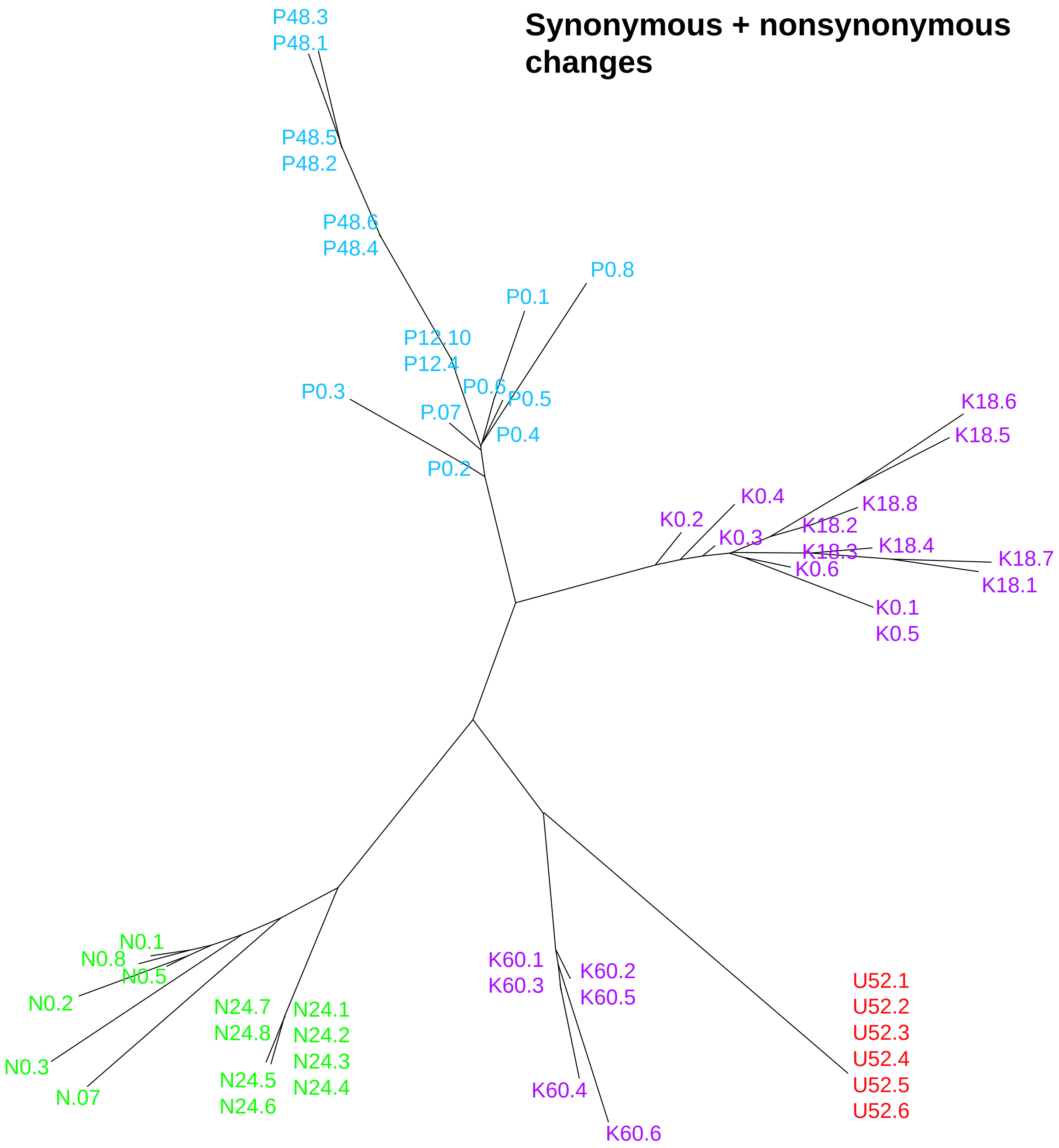

Figure 6. Neighbor-joining tree based on *all* changes.

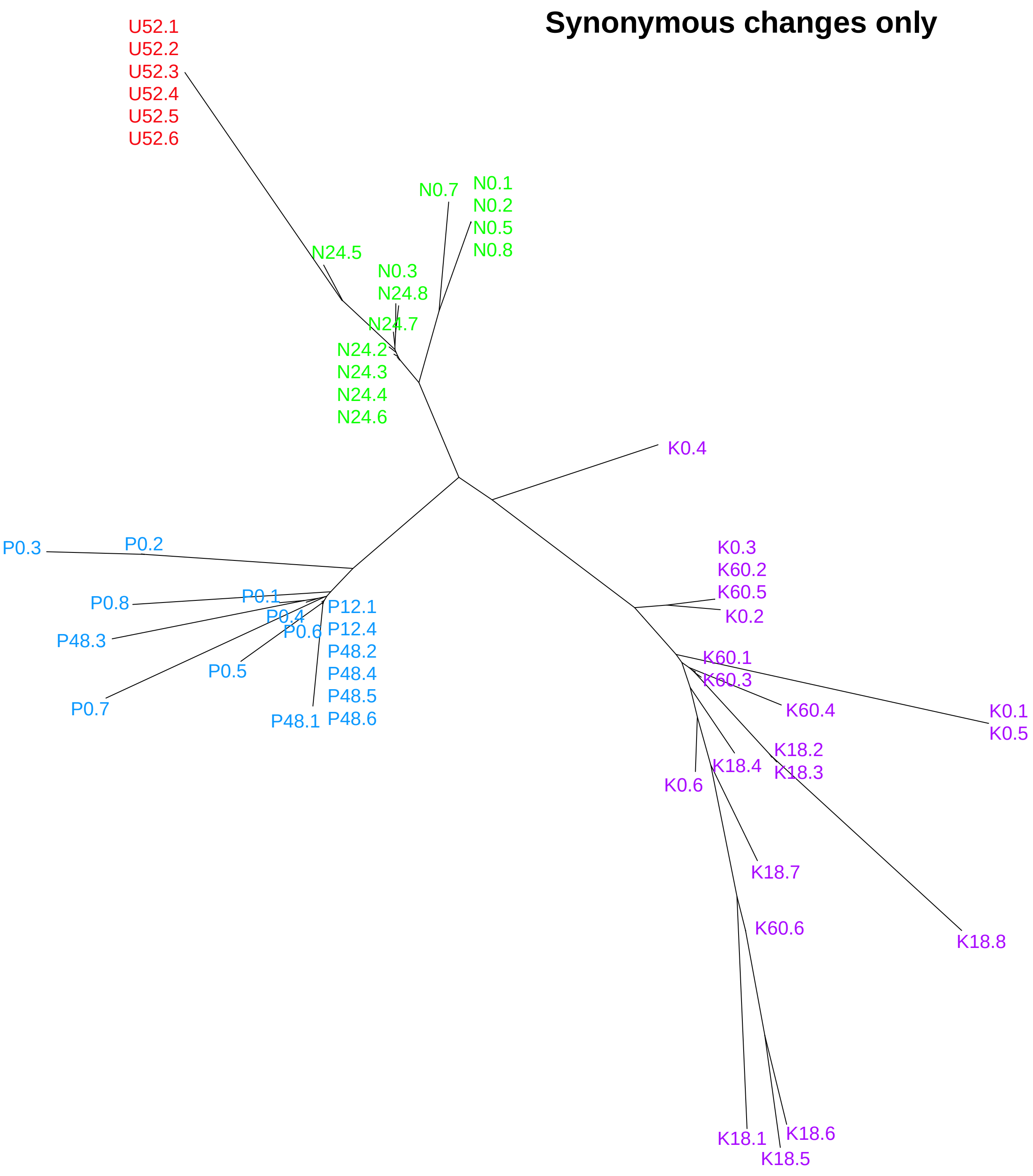

Figure 7. Neighbor-joining tree based on synonymous changes only.

Cited references:

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215**,403–10.

Korber, B., and G. Myers. 1992. Signature pattern analysis: a method for assessing viral sequence relatedness. *AIDS Res Hum Retroviruses* **8**,1549–60.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4**,**406–25.

Selected references on contamination and its consequences:

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 Aug;**70**(8):5720–5730.

Korber BT, Learn G, Mullins JI, Hahn BH, Wolinsky S., Protecting HIV databases, *Nature* 1995 Nov 16;**378**(6554):242–244.

Frenkel LM, Mullins JI, Learn GH et al., Genetic Evaluation of Suspected Cases of Transient HIV-1 Infection of Infants, *Science* 1998 May 15;**280**(5366):1073–1077.

McClure MO, Bieniasz PD, Weber JN, Tedder RS, O'Shea S, Banatvala JE, Tudor-Williams G, Simmonds P, Holmes EC., HIV clearance in an infant?, *Nature* 1995 Jun 22;**375**(6533):637-638

R. Schuurman, L. Demeter, P. Reichelderfer, J. Tijnagel, T. de Groot, C. Boucher on behalf of the ENVA laboratories, the Sequencing Working Group and participating laboratories, World-wide Evaluation of DNA Sequencing Approaches for the Identification Drug Resistance Mutations in the HIV-1 Reverse Transcriptase, *Proceedings of the 5th annual Conference on Retroviruses*, Abstract # 532.