Gene 448 (2009) 134–138

Contents lists available at ScienceDirect

Gene

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

The phylogeny of orthoretroviral long terminal repeats (LTRs)

Farid Benachenhou, Vidar Blikstad, Jonas Blomberg *

Section of Virology, Dept. of Medical Sciences, Uppsala University, Dag Hammarskjolds v. 17, SE-75185 Uppsala, Sweden

ARTICLE INFO

Article history:
Received 29 April 2009
Received in revised form 24 June 2009
Accepted 2 July 2009
Available online 9 July 2009

Received by N. Okada

Keywords:
Endogenous retrovirus
LTR
Phylogeny
Hidden Markov Model

ABSTRACT

LTRs are sequence elements in retroviruses and retrotransposons which are difficult to align due to their variability. One way of handling such cases is to use Hidden Markov Models (HMMs). In this work HMMs of LTRs were constructed for three groups of orthoretroviruses: the betaretroviruslike human MMTV-like (HML) endogenous retroviruses, the lentiviruses, including HIV, and gammaretroviruslike human endogenous retroviruses (HERVs). The HMM-generated LTR alignments and the phylogenetic trees constructed from them were compared with trees based on alignments of the *pol* gene at the nucleic acid level. The majority of branches in the LTR and *pol* based trees had the same order for the three retroviral genera, showing that HMM methods are successful in aligning and constructing phylogenies of LTRs. The HML LTR tree deviated somewhat from the *pol* tree for the groups HML3, HML7 and HML6. Among the gammaretroviruslike proviruses, the exogenous Mouse Leukemia Virus (MLV) was highly related to HERV-T in the *pol* based tree, but not in the LTR based tree. Aside from these differences, the similarity between the trees indicates that LTRs and *pol* coevolved in a largely monophyletic way.

1. Introduction

Endogenous retroviruses and LTR retrotransposons are present in a wide variety of organisms ranging from plants and insects to humans. For a review, see for example Blikstad et al. (2008) and Jern and Coffin (2008). One of their most characteristic features is two identical long terminal repeats (LTRs) flanking the protein coding genes. The LTRs vary considerably in length and internal structure but do have a few conserved motifs such as target site duplications (TG-CA), three A-rich regions, which occasionally encompass a TATA signal and always a polyadenylation signal (AATAAA box), see Benachenhou et al. (2009). Due to this diversity LTRs cannot be aligned with the commonly used alignment algorithms such as ClustalW (Thompson et al., 1994) and consequently are not used in e.g. phylogenetic analyses. This is despite the fact that the majority of endogenous retroviruses occur as single LTRs in many genomes (Mager and Medstrand, 2003) and completely lack the other protein coding genes such as the *pol* gene. Thus an important source of information is ignored.

In Benachenhou et al. (2009) several groups of LTRs from vertebrate exogenous and endogenous retroviruses were aligned by means of Hidden Markov Models (HMMs). The two most conserved groups were LTRs from the human MMTV-like (HML) endogenous retroviruses (Blikstad et al., 2008) and from the exogenous lentiviral retroviruses (including HIV-1 and HIV-2). The gammaretroviral HERV LTRs were on the other hand more variable. Here we explore whether phylogenies can be reconstructed from LTR Viterbi alignments for the three groups and compare them with trees obtained from *pol* gene alignments.

2. Results

Profile Hidden Markov Models were built according to the methodology of Benachenhou et al. (2009). The most important issue in the model building is to avoid overfitting by regularising the HMMs. The regularisation method in Benachenhou et al. (2009) was taken from Brand (1999). It has a parameter z that can be thought of as introducing disorder in the training set if negative.

The scoring of the sequences was performed using reverse-sequence null models (Karplus et al., 2005). This scoring method has the virtue of being insensitive to the composition bias of the sequence since it is the difference between the logarithm of the raw score of the sequence and the logarithm of the same sequence in reverse order.

For the three retroviral groups many HMMs were built with increasing number of match states (M) and with different z-values.

Abbreviations: LTR, long terminal repeat; HMM, Hidden Markov Model; MMTV, Mouse Mammary Tumor Virus; HML, human MMTV-like; HERV, human endogenous retrovirus; *pol*, polymerase gene nucleotide sequence; Pol, polymerase gene amino-acid sequence; Gag, group antigen amino-acid sequence; ERV, human endogenous retrovirus; PBS, primer binding site; HIV, human immunodeficiency virus; SIV, simian immunodeficiency virus; LST, lhoest's monkey; MND, mandrill; SUN, sun-tailed macaque; DRL, drill monkey; RCM, red-capped mangabey; CPZ, chimpanzee; O.BE, O. CM, HIV-1 type O; SAB, African green monkey, *sabaeus* subspecies; SIV-VER, vervet monkey; SYK, Sykes' monkey; MON, Mona's monkey; MUS, moustached monkey; GSN, greater spot-nosed monkey; DEB, DeBrazza's monkey; DEN, Dent's Mona monkey; COL, *guereza colobus*; BIV, bovine immunodeficiency virus; Visna, ovine maedi-visna virus; FIV, feline immunodeficiency virus; EIAV, equine infectious anemia virus; MLV, murine leukemia virus; GaLV, gibbon ape leukemia virus; FLV, feline leukemia virus.

* Corresponding author. Fax: +46 18 55 10 12.
E-mail address: Jonas.Blomberg@medsci.uu.se (J. Blomberg).

0378-1119/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2009.07.002

The score of the training set plotted against the number of match states showed a characteristic linear rise followed by a plateau where the score stayed constant (see Supplementary materials 1, 2 and 3).

For lentiviral LTRs it was found necessary to remove the part of the LTR which codes for the *nef* protein in order to obtain good alignments. Otherwise, this part interfered with the non-coding part of the LTR during HMM training. For the HML LTRs the long insertion mentioned in Benachenhou et al. (2009) (which sometimes also contained an open reading frame) had a similar effect and was therefore also removed.

Each HMM yielded a Viterbi alignment (Rabiner, 1989) of the training set. The Viterbi alignment with its insert states removed was used to construct a phylogenetic tree. Individual trees from different models varied to some extent. Ten to fifteen trees from models with *M*-values in the plateau and a fixed *z* were therefore combined to yield a 50% majority rule consensus tree. This proved especially useful for the broader gammaretroviral group because some groupings appeared consistently but not in the same individual tree. Negative *z* yielded consensus trees with somewhat higher bootstrap support. In Benachenhou et al. (2009) it was found that the HMMs trained with negative *z*-parameters were the most sensitively detecting ones and this is in line with other approaches to regularisation such as simulated annealing (see Eddy, 1995).

The resulting LTR trees were compared with trees based on ClustalX (Thompson et al., 1997) alignments of the *pol* gene at the nucleic acid level (see Figs. 1, 2 and 3).

For HML endogenous retroviruses the LTR tree did not group hml-7 with hml-8 which seems to be the correct grouping according to both ClustalW (Thompson et al., 1994) alignments of LTRs (data not shown) and the *pol* tree (see Fig. 1). However it is well known that HMM methods perform less efficiently than ClustalW when aligning closely related sequences within subgroups (Edgar and Sjolander, 2003). HMMs are on the other hand superior in aligning the different subgroups. The other difference between the LTR and the *pol* tree is the branching order of HML3 and HML6. If the LTR coevolved with the *pol* gene, this discrepancy could be explained as a long branch attraction between HML5 and HML6 in the *pol* tree, since they are both relatively distant from the other HML groups (see Fig. 1). However, there may also be other explanations (see below). In the LTR tree the bootstrap support for MER9 (HML3) is admittedly weaker but this could be due to the misplacement of MER11D (HML7). In Benachenhou et al. (2009) HML6 was detected in human chromosome 19 even though it was absent from the LTR training set. This was not the case for HML3 when it was absent from the training set, confirming the branching order of the LTR tree, i.e. that the HML6 LTR is closer to the HML1-2/4/8-10 LTRs than is the HML3 LTR. In addition, as described in Lavie et al. (2004), both HML-5 and HML-3 use different primer binding sites in comparison to the other HMLs. HML-5 uses methionine or isoleucine tRNA while HML-3 uses arginine or asparagine tRNA instead of lysine tRNA (which gave the alternative name HERV-K). This opens the possibility that the true LTR and *pol* trees are not identical, i.e. the evolution of the LTR and the *pol* gene may not have been monophyletic for all HML groups.

For lentiviruses the LTR and *pol* trees (Fig. 2) can be compared to the robust phylogenetic tree in Gifford et al. (2008). This tree was based on the Gag and Pol proteins at the amino-acid level. Both trees correctly group the non-primate lentiviruses BIV, Visna, FIV and EIAV outside the primate lentiviruses but their branching orders do not completely agree with Gifford et al. (2008). In the *pol* tree the SAB lentiviral sequence (African green monkey, *sabaeus* subspecies) branches differently. On the other hand the LTR tree has generally lower bootstrap support than the *pol* tree.

The gammaretroviral LTR tree (Fig. 3) has as expected (because this group of LTRs is more variable) lower support and more unresolved nodes than the HML and lentiviral LTR trees. Nevertheless it generally follows the gammaretroviral *pol* tree (Fig. 3). The *pol* tree

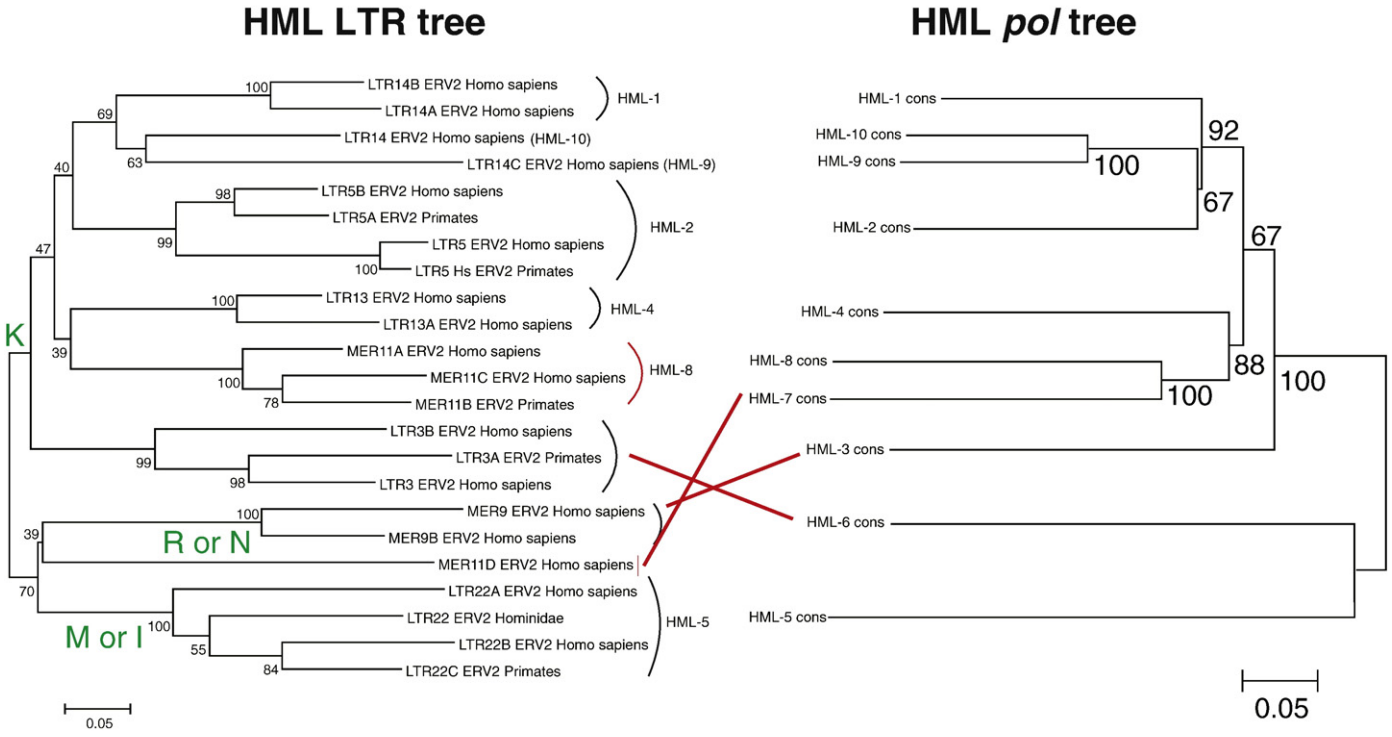

Fig. 1. HML LTR and *pol* trees. Comparison between neighbour-joining trees of HML LTR sequences and HML *pol* nucleic acid sequences. The two trees are aligned when possible; the non-congruent branches are connected with red lines. The amino acid corresponding to the primer binding site (PBS) as described in Lavie et al. (2004) is shown in the LTR tree. K: lysine, M: methionine, R: arginine, N: asparagine. The correspondence between the RepBase nomenclature and the HML names follows (Mager and Medstrand, 2003; Blikstad et al., 2008). Mega 4.1 was used with default parameters except for the pairwise deletion option. LTRs have RepBase names and *pol* sequences ERV names from the literature.

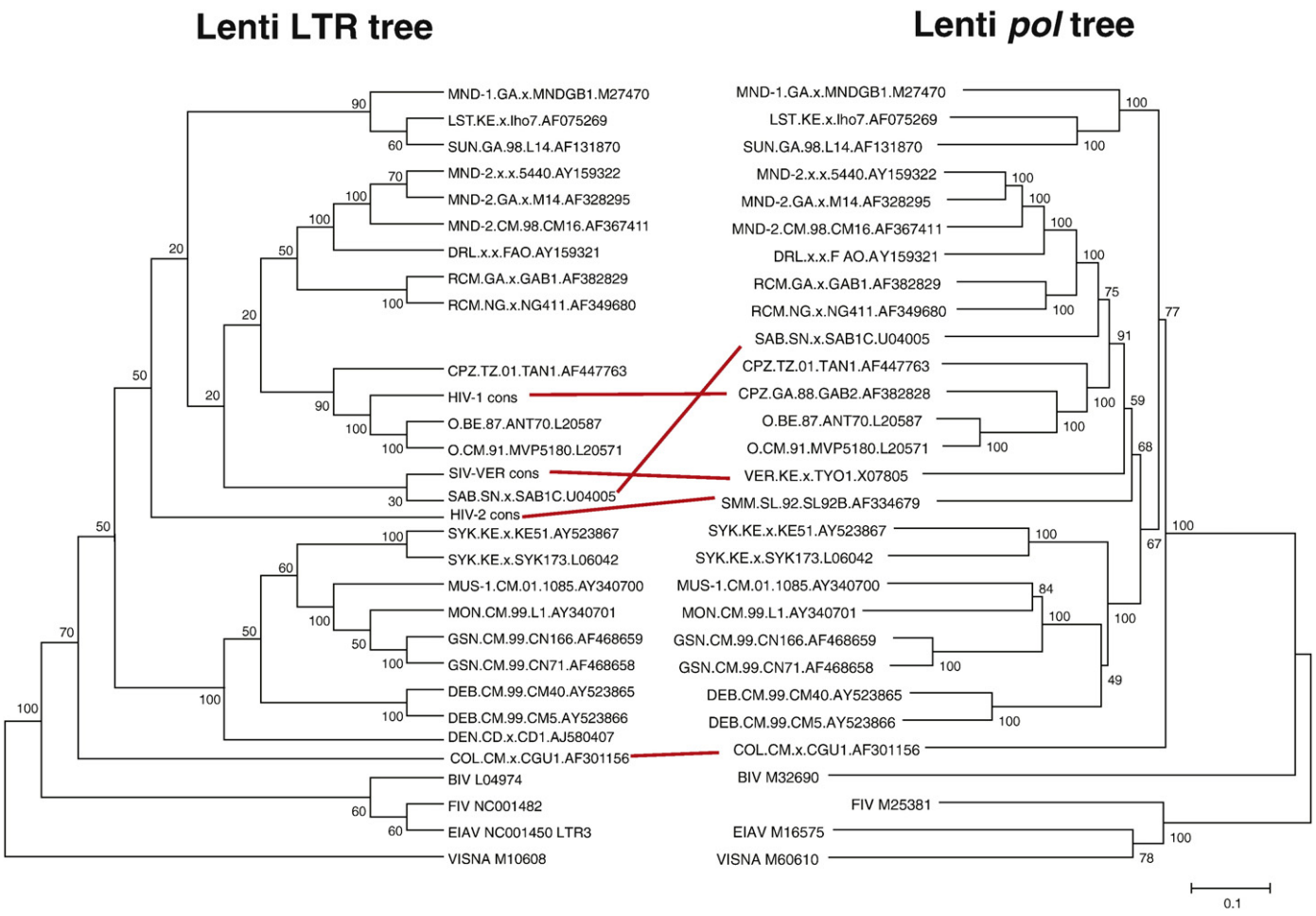

Fig. 2. Lenti LTR and *pol* trees. Comparison between neighbour-joining trees of lentiviral LTR sequences and lentiviral *pol* nucleic acid sequences. Each of the three consensuses in the LTR tree is for simplicity represented in the *pol* tree by one of the sequences that were replaced by the consensus. LST: lhoest's monkey; MND: mandrill; SUN: sun-tailed macaque; DRL: drill monkey; RCM: red-capped mangabey; CPZ: chimpanzee; O.BE, O.CM: HIV-1 type O; SAB: African green monkey, *sabaeus* subspecies; SIV-VER: vervet monkey; SYK: Sykes' monkey; MON: Mona's monkey; MUS: moustached monkey; GSN: greater spot-nosed monkey; DEB: DeBrazza's monkey; DEN: Dent's Mona monkey; COL: *guereza colobus*; BIV: bovine immunodeficiency virus; Visna: ovine maedi-visna virus; FIV: feline immunodeficiency virus; EIAV: equine infectious anemia virus. Methods and conventions were as in Fig. 1.

has both higher bootstrap support and larger groups: for example it groups MLV with HERV-T, ERV3, HERV15 and HERV-E, whereas the LTR tree only clusters HERV-T, ERV3, HERV15 and HERV-E.

The results show that the HMM-based approach makes it possible to align relatively unrelated LTRs. It was successful in aligning subsets of orthoretroviral LTRs and can be used for phylogenetic analyses at least for the relatively homogenous groups studied here.

3. Discussion

LTRs are partly very variable and partly highly conserved. The conservation is observable only after advanced bioinformatic treatment. In this paper we explored whether HMM-generated (Viterbi) LTR alignments would give the same, similar, or dissimilar phylogenetic trees as trees derived from the conserved core of retroviruses, the *pol* gene. Three major genera of orthoretroviruses, lentiviruses, the human betaretroviruslike sequences (HML), and human endogenous gammaretroviruslike sequences, were analyzed. Overall, the Viterbi alignments lead to trees which were largely congruent with *pol* trees for all three groups studied here. The HMM-based methods generated Viterbi alignments which allowed reconstruction of HML, lentiviral and gammaretroviral LTR phylogeny. The bootstrap support was high for many of the branches. This represents a step forward in understanding the phylogenetic relationships of retroviruses. LTR alignments are particularly useful for phylogenetic inference if no other gene is available, like for single LTRs, but can as well be combined with alignments of other genes and structural taxonomic markers to yield stronger phylogenetic hypotheses (Jern et al., 2005; Blomberg et al., 2009). Recombination is frequent in retroviruses. Despite this, the trees derived from one of the most variable retroviral subsequences, the LTRs, had a similar branching order as *pol* trees. Thus, recombination probably did not frequently separate *pol* from LTR of the three clades during many millions of years. However, the deviations in branching pattern may or may not indicate that occasionally a separation between LTR and *pol* occurred, possibly due to recombination. The two discrepancies between the HML LTR and *pol* trees both came from HMLs with an aberrant PBS, not using lysine tRNA. The PBS is situated just outside of the 5′LTR. A recombination involving the LTR thus could have involved also the PBS. Another notable difference was seen in the LTR and *pol* based trees of the diverse gammaretroviruslike HERVs. The LTRs of MLV and its close relatives exogenous relative GaLV were separated from the LTRs of HERV-T, HERV-E, ERV-3 and RRHERVI (HERV15). In the *pol* tree, MLV came out together with these HERVs. Both LTR trees have some uncertain branches, which makes it hard to reach a definite conclusion.

On the technical side, a subjective feature of the method is the manual removal of the long insertion in HML LTRs. By incorporating it in the training process better alignments and phylogenetic trees would possibly be achieved. We are currently working on this issue.

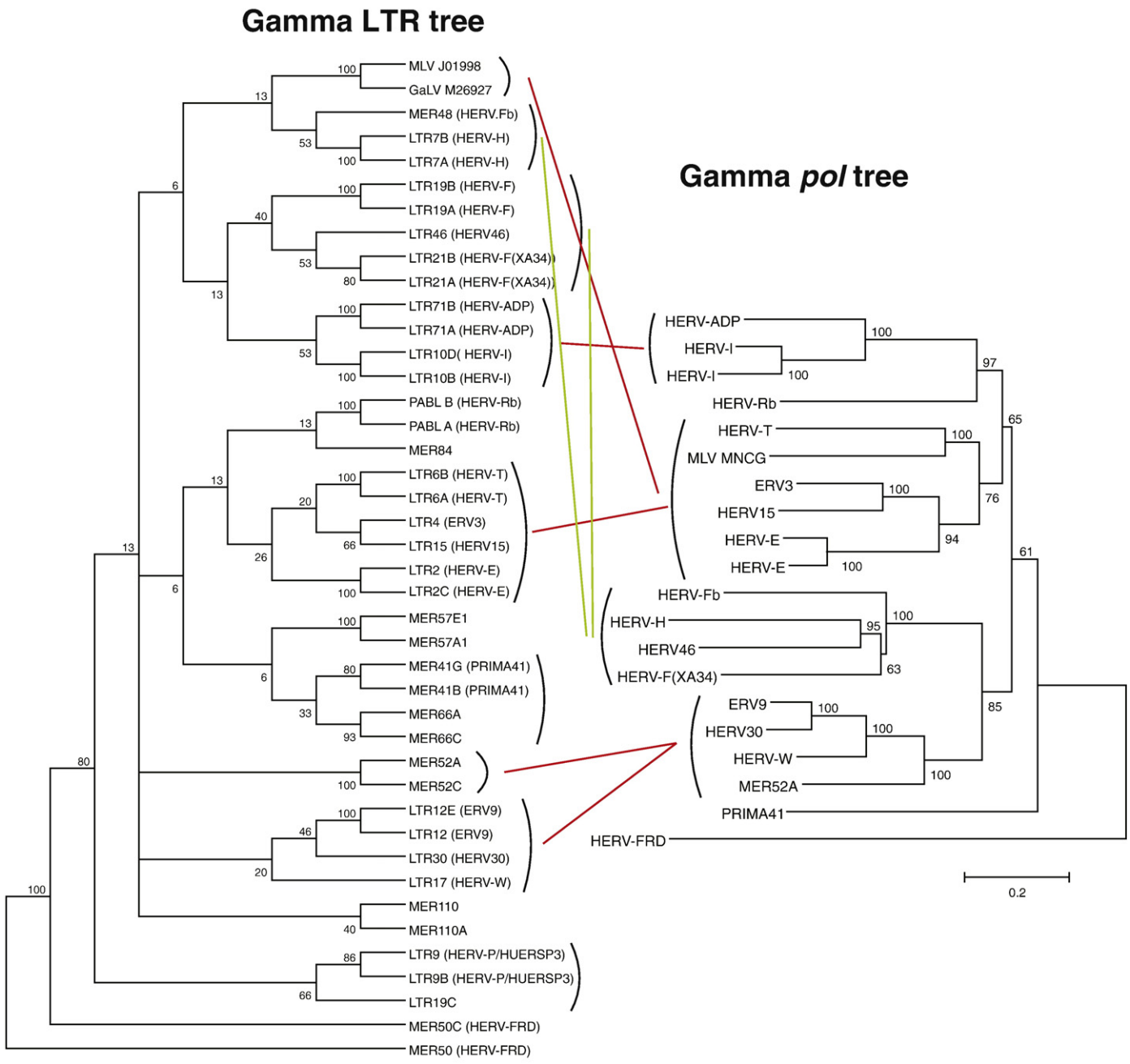

Fig. 3. Gamma LTR and *pol* trees. Comparison between neighbour-joining trees of human gammaretroviruslike (ERV1 in RepBase) LTR sequences (a consensus cladogram) and the corresponding *pol* sequences (a phylogram). The latter were extracted from internal retroviral RepBase entries using RetroTector Online (Benachenhou et al., 2009; Sperber et al., 2009). Not all internal RepBase consensus sequences yielded a *pol* sequence with ROL. They are therefore fewer. Methods and conventions were as in Fig. 1. Lines join groups which largely contain the same members. Red lines join groups which occur in similar positions in the two trees. Green lines join the HERV-H like groups, which occur in different positions in the two trees. A notable exception is MLV, whose *pol* clusters with HERV-T *pol*, but whose LTR does not cluster with HERV-T LTR. However, the bootstrap support in the LTR tree is weak for branches which do not fit into the *pol* tree order. The LTR consensus tree was derived from multiple alternative trees using the program Consense in the PHYLIP program package.

Furthermore, the standard profile HMMs used here may not be the optimal choice due to the repetitive nature of LTRs. An architecture more adapted to LTRs is perhaps warranted.

In conclusion, LTRs are adapted to a particular cellular, tissue and organismic environment, but still retain basic functions in mRNA transcriptional start and polyadenylation, as well as during integration. Much of the variability is due to interaction with a variable set of host proteins, like transcription factor binding sites in the cellular microenvironment. This generates LTR variation, even between the members of a viral quasispecies in an infected organism. It is therefore reassuring that despite this high variability, there exists a backbone of conserved nucleotides which to a large extent coevolved with the most conserved retroviral gene, the *pol* gene.

4. Materials and methods

The primate lentiviral LTRs and *pol* genes were mainly obtained from the "HIV Databases", http://www.hiv.lanl.gov/ (see Fig. 2 for accession numbers). The lentiviral training set also contained four non-primate LTRs retrieved from GenBank: BIV, Visna, FIV and EIAV, see Fig. 2 for accession numbers. Three LTR subgroupings were replaced by majority consensuses based on ClustalX alignments (Thompson et al., 1997) (see

Fig. 2 and Supplementary materials 4–6). The *pol* nucleotide sequences were aligned by ClustalX with alignment length equal to 3720 base pairs.

The HML LTRs consisted of 23 RepBase (Jurka et al., 2005) consensus sequences. The HML *pol* sequences were 10 nucleic acid consensuses (HML-1 to HML-10) constructed by Blikstad et al. (in preparation). The correspondence between the RepBase nomenclature and the HML names (see Fig. 1) follows (Mager and Medstrand, 2003). The ClustalX alignment obtained from the *pol* sequences was 2926 base pairs long.

The gammaretroviral LTRs were 69 RepBase (Jurka et al., 2005) consensus sequences. The 19 endogenous *pol* sequences were reconstructed by RetroTector© online (Sperber et al., in press) from cognate RepBase internal consensus sequences. Three exogenous LTRs were added to the LTR set: The Murine Leukemia Virus MLV LTR, accession number J01998, the Gibbon ape Leukemia Virus GaLV LTR, accession number M26927 and the Feline Leukemia Virus FLV LTR, accession number M18247. The *pol* set was supplemented with a sequence from the MNCG MLV.

The software used was programs written in C by FB, implementing the HMMs. The phylogenetic trees were made with Mega version 4.1 (Tamura et al., 2007). The trees were neighbour-joining trees with the pairwise deletion option but otherwise default parameters. The 50% majority rule consensus trees were constructed with PHYLIP version 3.68 (http://evolution.gs.washington.edu/phylip.html) (Felsenstein, 1988). In many cases minimum evolution trees were also computed and were found to give the same topology and similar bootstrap values as the neighbour-joining trees.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2009.07.002.

References

Benachenhou, F., et al., 2009. Evolutionary conservation of orthoretroviral long terminal repeats (LTRs) and ab initio detection of single LTRs in genomic data. PLoS ONE 4, e5179.

Blikstad, V., Benachenhou, F., Sperber, G.O., Blomberg, J., 2008. Endogenous retroviruses: Evolution of human endogenous retroviral sequences: a conceptual account. Cell. Mol. Life. Sci. 65, 3348–3365.

Blomberg, J., Benachenhou, F., Blikstad, V., Sperber, G., Mayer, J, 2009. Classification and nomenclature of endogenous retroviral sequences (ERVs), problems and recommendations. Gene, PMID 19540319.

Brand, M, 1999. Structure learning in conditional probability models via an entropic prior and parameter extinction. Neural. Comp. 11, 1155–1182.

Eddy, S.R, 1995. Multiple alignment using hidden Markov models. Proc. Int. Conf. Intell. Syst. Mol. Biol. 3, 114–120.

Edgar, R.C., Sjolander, K., 2003. Simultaneous sequence alignment and tree construction using hidden Markov models. Pac. Symp. Biocomput. 180–191.

Felsenstein, J, 1988. Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. 22, 521–565.

Gifford, R.J., Katzourakis, A., Tristem, M., Pybus, O.G., Winters, M., Shafer, R.W, 2008. A transitional endogenous lentivirus from the genome of a basal primate and implications for lentivirus evolution. Proc. Natl. Acad. Sci. U. S. A. 105, 20362–20367.

Jern, P., Coffin, J.M, 2008. Effects of retroviruses on host genome function. Annu. Rev. Genet. 42, 709–732.

Jern, P., Sperber, G.O., Blomberg, J, 2005. Use of endogenous retroviral sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy. Retrovirology 2, 50.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J, 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. 110, 462–467.

Karplus, K., Karchin, R., Shackelford, G., Hughey, R, 2005. Calibrating *E*-values for hidden Markov models using reverse-sequence null models. Bioinformatics 21, 4107–4115.

Lavie, L., Medstrand, P., Schempp, W., Meese, E., Mayer, J, 2004. Human endogenous retrovirus family HERV-K(HML-5): status, evolution, and reconstruction of an ancient betaretrovirus in the human genome. J. Virol. 78, 8788–8798.

Mager, D.L., Medstrand, P., 2003. Retroviral repeat sequences. D, C. (D, C.)D, C.s), Nature Encyclopedia of the Human Genome. InNature publishing group, London, UK.

Rabiner, L.A, 1989. tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77, 257–286.

Sperber, G., Lövgren, A., Eriksson, N.-E., Benachenhou, F., Blomberg, J, 2009. RetroTector online, a rational tool for analysis of retroviral elements in small and medium size vertebrate genomic sequences. BMC Bioinformatics 10, Suppl 6:S4. PMID 19534753.

Tamura, K., Dudley, J., Nei, M., Kumar, S., 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24, 1596–1599.

Thompson, J.D., Higgins, D.G., Gibson, T.J, 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic. Acids Res. 22, 4673–4680.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G, 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic. Acids Res. 25, 4876–4882.