Molecular evolution of selected families of human endogenous retroviruses

Palle Villesen, Hugo Martins & Jotun Hein 5.9.11

Aim: Collect genomic data on genetically conserved endogenous retroviruses. Compare the integration time estimates on endogenous retroviral loci under neutral or purifying selection. Integration time estimates should be done using different methods, on the basis of human homologue and orthologue ERV loci in mammalian genomes.

Background:

Endogenous retroviruses (ERVs) are genetic fossils of ancient retroviral integrations that remain in the genome of many organisms. Most loci are rendered non-functional by mutations, but several intact retroviral genes are known in mammalian genomes and may play important roles in disease or physiological functions [1,2]. The determination of the integration time of these viruses has been based upon the assumption that both 5' and 3' Long Terminal Repeats (LTRs) sequences are identical at the time of integration, but evolve separately afterwards, thus behaving as paralogue genes. By studying the similarity at the genetic level between both LTRs of an ERV and using species-related mutation rates, one can estimate at what time that ERV might have inserted itself in the host genome[3]. Such approaches are not faithful representations of genetic reality, however, since ERVs, once inserted, are usually submitted to selective pressures. Because they are viral remnants, and thus harmful in principle to the host organism, many ERVs are indeed deleterious genes and suffer extreme genetic degradation. Degraded ERVs might reach present day as nothing more that solo LTRs. However, there are exceptions. Some ERVs remain relatively safe in the genome under the umbrella of neutral selection (genetic drift) and a few are even co-opted by the organism as fully functional genes.

An ERV has a typical structure of a retrovirus. Besides the two LTRs, one on each extremity of the sequence, the ERV possesses at least three major genes that encode for the virus' functional proteins[4]. These genes are the *gag* gene (encodes for several structural proteins), the *pol* gene (encodes for the polyprotein that includes the protease, integrase and reverse transcriptase) and the *env* gene (encodes for surface envelope proteins). Sometimes, *gag*, *pol* and the protease gene and considered separate. Some minor regulatory genes may or may not be present, depending on the viral family.

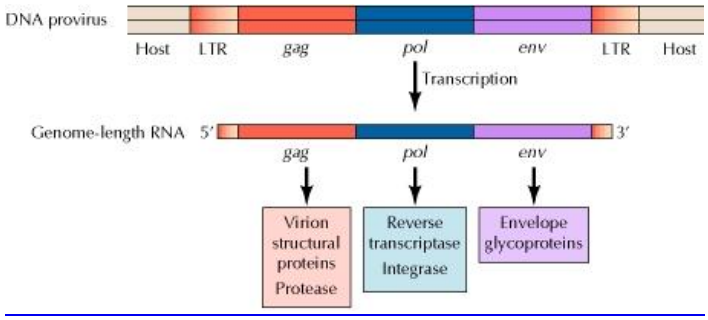

Fig.1 – Integrated proviral sequence (top) and exogenous retroviral sequence (bottom). Picture from [5].

ERVs have been detected and characterized in all studied vertebrate genomes so far, and 8-9% of the human genome itself is composed of degenerate ERVs. Data on repetitive elements on genomes can be visualized via online databases that use the Repeatmasker tool such as the UCSC Genome Database[6]. The detected ERVs sequences and classifications can be consulted via the online database RepBase [7].

In order to perform insertion date estimations, it is necessary to reconstruct the genetic history of each ERV locus. For that, multiple orthologue copies of the same ERV locus are need across as much species as possible. Once the original ERV locus is taken from the bibliography, its copies and paralogues can be acquired by BLASTing the sequence on other species' genomes or, in some cases, by direct sequence retrieval on RepBase. The workflow will consist in the following steps: 1) data retrieval, 2) sequence alignment, 3) phylogeny building, 4) selection events study, 5) integration date estimation.

After relevant data on conserved ERV families is collected, a dedicated alignment tool such as CLUSTAL or MUSCLE should be used to produce an alignment for each family of ERVs. These ERVs should align rather well in the LTR regions, but the alignments in the intra-LTR region may vary in quality depending on the degree of conservation of the specific ERV copy. A correct alignment should reproduce phylogenetically the paralogue behavior of the ERV, with a first separation between 3' and 5' LTRs in the original host and only latter reproducing the vertical transmission along the evolutionary tree. This phylogeny can be build using specialized tools, such as PhyML or MEGA.

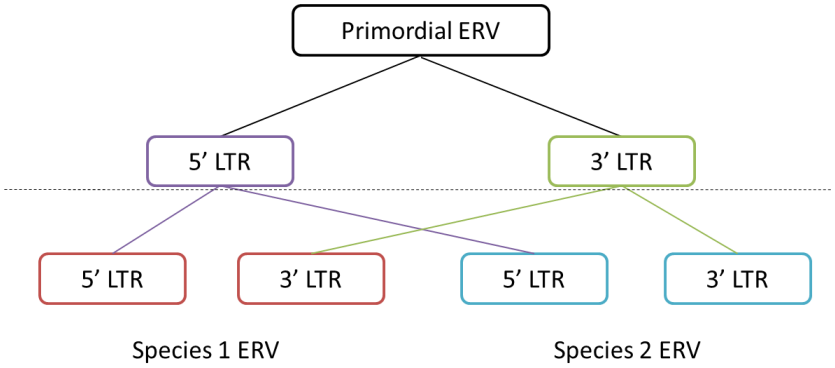

Fig. 2 – Paralogue behavior of ERVs. The first genetic separation is done between 5' and 3' LTRs still in the original host. After the first vertical transmission, both LTRs, once identical, evolve separately, however belonging to the same genetic locus.

In case of severely degenerated sequences that affect the alignment of LTRs, these terminal sequences can be cropped out of the larger sequence and aligned in separate. LTR boundaries can be detected by a simple DNA dot-plot that will reveal inverted and repeated sequences in the genetic sequence. If annotation information on the LTR boundaries (in RepeatMasker tables, for instance) is available, cropping out the LTRs becomes trivial.

Selection studies in ERVs can be performed using the PAML[8] software. Testing for selection usually implies treating the DNA sequence as coding, so degeneracy in the intra-LTR region is crucial. More recent integrations, theoretically more conserved, will yield better results due to sequence integrity and will also diminish the noise from cumulative, untraceable point mutations at single sites. Selection tests in PAML involve the building and use of several comparative models for negative (purifying) or positive selection against a null hypothesis of neutrality. The Ka/Ks ratio (also referred to as dN/dS or ω) measures the rate of substitutions on non-synonymous sites versus the rate on synonymous sites, acting as indicator of selective pressure on protein-coding genes[9]. The methodology will be applied on the selected datasets to identify selective pressures on the ERVs at the site and sequence levels. For a more detailed information on implementation, refer to the PAML manual.

In order to estimate the dates of insertion, various methods can be used. The most basic is to apply a standard mutation rate for the studied organism of ERV family and calculate the divergence between the 5' and 3' LTRs of the species. This should yield a point-estimate of the insertion date assuming a steady mutation rate across time. A second method that corrects for variable mutation rates along the phylogeny is given in [10], and becomes more accurate with increasing available data. Finally, a more complex methodology involves performing a Montecarlo phylogenetic modeling of the data, which will yield tree node ages, in time intervals. This method is more computationally demanding and laborious, but it provides the best estimate, again depending on data quality. Such methodology can be implemented through the mcmctree application in PAML.

Future perspectives: The expected outcome of this project will be a characterization of selected ERV families at the phylogenetic, evolutionary and insertion time level. Problems may arise in several points of the data collection and treatment that can be circumvented either by the suggested pathways, by novel

initiatives, or tackled altogether as spinoff projects. These problems will stem from the degenerate nature of the ERV sequence which results in difficult alignments and selective pressure studies on non-conserved genes/pseudogenes. An efficient dating method is also dependent on good LTR data quality. Identification of previously unknown ERV loci under selective pressure may lead to the discovery of potential sites of interest for molecular characterization and expression (Protein/RNA) studies.

Requirements: BLAST, alignment, phylogeny building, PAML.

References:

[1] Nexø BA et al, *The etiology of multiple sclerosis: genetic evidence for the involvmente of the human endogenous retrovirus HERV-Fc1*. PLoS One 2001 Feb 2; 6(2).
[2] Kämmerer U et al, *Human endogenous retrovirus K (HERV-K) is expressed in villous and extravillous cytotrophoblast cells of the human placenta*. J Reprod Immunol 2011 Aug 12.
[3] Bowen NJ, McDonald JF, *Drosophila euchromatic LTR retrotransposons are much younger than the host species in which they reside*. Gen Res 2001 Sep; 11(9):1527-40.
[4] Coffin JM, *Structure and Classification of Retroviruses*. In Levy JA *The Retroviridae* (1st ed), New York Plenum Press. 20-34.
[5] Geoffrey M. Cooper; *The Cell – A Molecular Approach*, 2nd edition, Boston University.
[6] Kent WJ et al, *The human genome browser at UCSC*. Genome Res 2002 Jun; 12(6):996-1006
[8] Yang Z, *PAML 4: phylogenetic analysis by maximum likelihood*. Mol Biol Evol 24(8): 1586-91 (2007)
[7] Jurka J et al, *Repbase Update, a database of eukaryotic repetitive elements*. Cyt Gen Research (2005) 110:462-7
[9] Massingham T, Goldman N, *Detecting Amino Acid Sites Under Positive Selection and Purifying Selection*. Genetics 169: 1753-62 (March 2005)
[10] Martins H, Villesen P (2011) *Improved Integration Time Estimation of Endogenous Retroviruses with Phylogenetic Data*. PLoS ONE 6(3): e14745. doi:10.1371/journal.pone.0014745

Gifford R, Tristem M. *The evolution, distribution and diversity of endogenous retroviruses. Virus Genes*. 2003 May;26(3):291-315.

Links:

PAML - http://abacus.gene.ucl.ac.uk/software/paml.html
UCSC Genome Browser - http://genome.ucsc.edu/
BLAST - http://blast.ncbi.nlm.nih.gov/Blast.cgi
RepBase - http://www.girinst.org/repbase/update/browse.php
PhyML – http://www.atgc-montpellier.fr/phyml/binaries.php
MEGA – http://www.megasoftware.net/
CLUSTAL – http://www.clustal.org/
MUSCLE – http://www.drive5.com/muscle/