Tracing the Proliferation Process of Class II Transposon Families Throughout Mammalian Host Species

(Hellen & Brookfield, 2012)

Elizabeth HB Hellen & John FY Brookfield

Introduction

Three percent of the human genome consists of class II (DNA) transposable elements (Lander *et al.*, 2001), although these sequences are found far less abundantly than class I (RNA) transposable elements, such as Alu (de Koning *et al.*, 2011). However, when we consider that protein-coding gene sequences make up roughly 1.5% of the human genome, the importance of class II elements' contribution to the make-up of mammalian genomes can be put into perspective.

A number of interesting questions can be asked of these transposon families. We are particularly interested in the process of transposons originating in a genome, proliferating and becoming inactive. In particular, we can ask whether the inactivation of elements is somehow linked to their initial spread, such that the nature of their proliferation through genomes and populations has the effect that inactivation follows inevitably, or whether the elements proliferate to create a population of elements, residing stably in the chromosomes and undergoing a turnover process, with extinction being a subsequent, random event, unlinked to this initial proliferation.

In any given genome the collection of transposable elements of a given family will be connected to a most recent common ancestor element (element MRCA) by a phylogenetic tree, and the changes in elements' sequence that have happened since the element MRCA can be used to estimate the shape of the tree. Thus, for an element family a time to the element MRCA can be estimated (Hellen & Brookfield, 2011). As an alternative to the MRCA of transposable elements, the family can be dated by looking at the presence of elements in modern organisms and calculating the time to a host MRCA using predicted divergence dates between organisms.

What is the significance of any differences between these times to MRCA? Is this the dating of the first invasion of the genome with that family of transposable element? Or did the element family exist in the genome for millions (or tens of millions) of years prior to the element MRCA? For transposable elements, a theoretical possibility is that there is an ongoing process of turnover, such that the time to element MRCA for a family is much more recent than is the time of origin of the family.

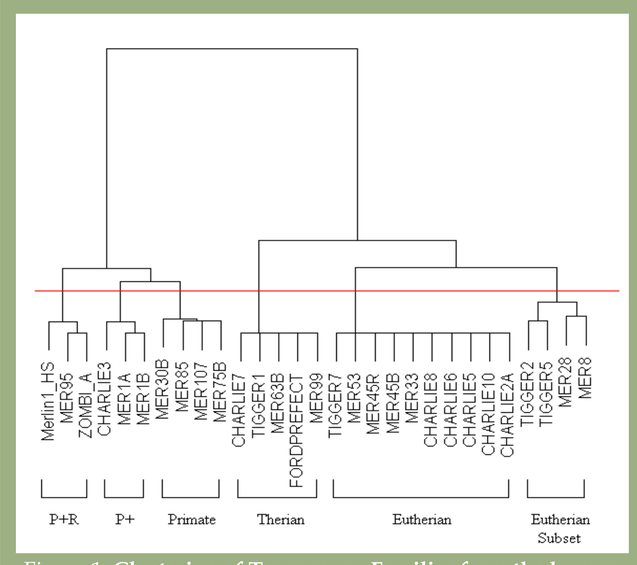

Figure 1. **Clustering of Transposon Families from the human genome**

Transposon families have been divided into 6 groups using a hierarchical clustering algorithm implemented in R.

- P+R - present in Primates and Rodentia
- P+ - present in Primate as well as some other species
- Primate - present in Primates
- Therian - found in all eutherian and marsupial species
- Eutherian - found in all eutherian species
- Eutherian subset - found in a large number of eutherian species

Dating Method 1 – MRCA of Host Species

Consensus sequences for human class II transposable elements were retrieved from the Repbase database (Jurka *et al.*, 2005). The consensus sequences were used to carry out Ensembl BLAT (Flicek *et al.*, 2012) searches. The presence or absence of matches to the consensus sequences allowed a rough origin date to be assigned to each sequence. Transposons were clustered into groups with similar predicted origin times using hierarchical clustering, implemented through the statistical language R (version 2.11.1) (http://www.R-project.org.).

Dating Method 2 – MRCA of Transposons

Multiple sequence alignments, for each transposon family, created from pairs of orthologous transposon sequences were dated using BEAST (Drummond & Rambaut, 2007), to predict an origin date for the family (the MRCA of all modern elements). Five different analyses were carried out for each family using different pairs of orthologs; Human-Chimp, Human-Orangutan, Dog-Cat, Dog-Panda, Cow-Pig. Intermediate dates were assigned to the phylogeny at the divergence points between orthologs for each of the transposon elements using the mean divergence times found in Timetree.org (Kumar & Hedges, 2011).

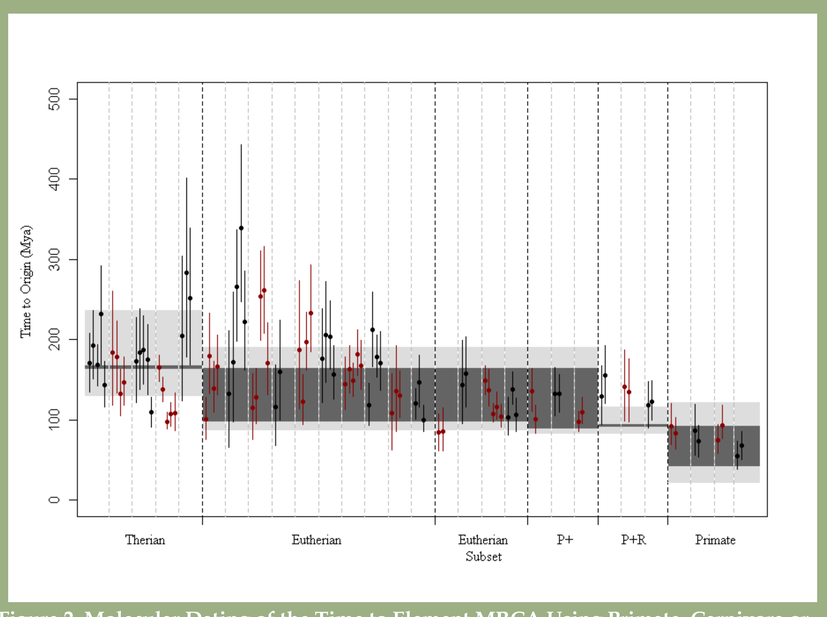

Figure 2. Molecular Dating of the Time to Element MRCA Using Primate, Carnivora or Artiodactyla Orthologous Pairs.
Mean Predicted time to element MRCA of each transposon family, using human-chimp, human-orangutan, dog-panda, dog-cat and cow-pig divergence dates as constraints. Error bars show the highest posterior density interval. Dark grey - time predicted for the species MRCA using timetree.org mean values, light grey - time predicted using highest/lowest published values.

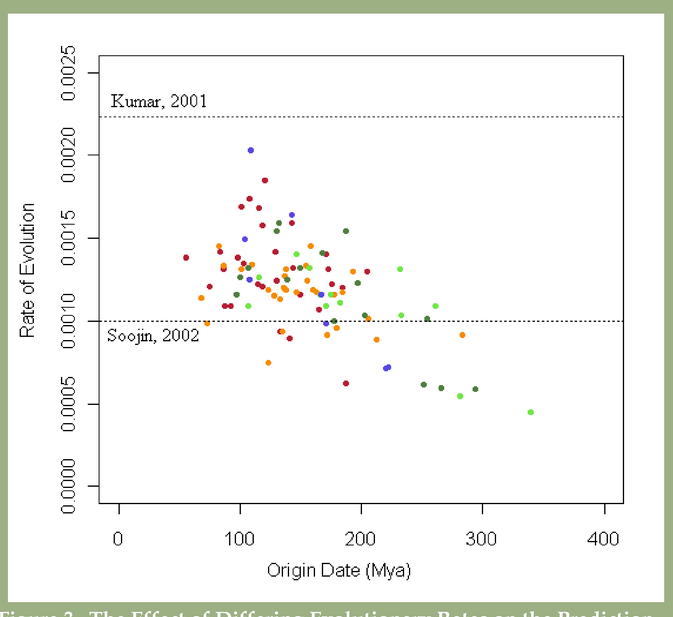

Figure 3. The Effect of Differing Evolutionary Rates on the Prediction of the Element MRCA.
Plot showing the mean rate of evolution (clock.rate) and the mean time to element MRCA (treemodel.rootHeight) for each of the BEAST analyses; data from different families and using different orthologous pairs has been pooled. human-chimp, human-orangutan, dog-panda, dog-cat, cow-pig.

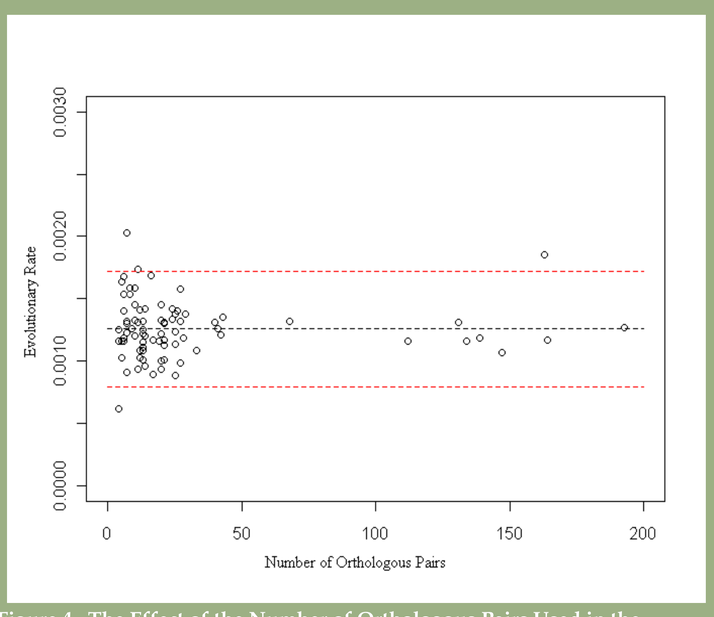

Figure 4. The Effect of the Number of Orthologous Pairs Used in the Prediction of Evolutionary Rate.
Each point represents one BEAST analysis, comparing the predicted evolutionary rate against # orthologous pairs used in the analysis; data from different families and different orthologous pairs has been pooled. Black dotted line = mean rate, red dotted lines = mean ± 2 standard deviations.

Results

The BEAST analyses, using human-chimp orthologs, predicts ancestral element dates which mostly fall within the bounds of the species MCRA predictions made using the presence of the family in modern genomes and timetree.org mean divergence estimates (Figure 2). The analyses carried out with different species pairs largely agree in their predictions.

The Primate+ group is predicted to have occurred at a similar time to the Therian and Eutherian subset groups. This would imply that the occurrence of the elements in certain organisms, but not others which are closely related, is due to the loss of elements from certain lineages rather than horizontal transfer of the element into these species. The earlier than expected dates found for the 'Primate + Rodent' group are confirmed when using the human-orangutan orthologs, but cannot be assessed through a different lineage due to the lack of examples in modern genomes. These early origin dates, coupled with a lack of examples in most modern eutherian species, may imply that the latter is the result of a large scale deletion of these elements in species not on the primate-rodentia lineage.

A strong negative correlation can be seen between the evolution rate and the predicted time to element MRCA (Figure 3), particularly for the earliest predicted dates (Pearson; $r = -0.64$, $p<0.05$). The correlation is not an effect we would expect to occur naturally, instead it can be assumed that the numbers of observed changes between orthologous elements in the species chosen for comparison, which may be higher, or lower, than expected from the long-term evolutionary rate, are pushing the estimate of the origin date from its true position.

Most variation in evolutionary rates can be seen in the analyses carried out using a smaller number of orthologs (Figure 4). In larger analyses, where the clock is averaged over a greater number of branches, this effect is muted.

Discussion

Our predictions have shown that, for the majority of cases, the prediction of the time to element MRCA is similar to the range of dates predicted for the species MRCA of extant elements.

This observation is consistent with the concept of a life cycle of the proliferation of the elements followed by inactivation, rather than an ongoing process of turnover, extending many tens of millions of years after the elements' origin. If the latter were to have happened, the time to element MCRA in a given genome would be expected to be much more recent than the time to the species MRCA of the host organisms that now contain the elements.

Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology 7:214.
Flicek P, Amode MR, Barrell D *et al.* (2012.) Ensembl 2012. Nucleic Acids Res. 40:D84-90.
Hellen EHB, Brookfield JFY (2011) Investigation of the Origin and Spread of a Mammalian Transposable Element Based on Current Sequence Diversity. J. Mol. Ev. 73(5-6):287-296
Hellen EHB, Brookfield JFY (2012) The diversity of Class II transposable elements in mammalian genomes has arisen from ancestral phylogenetic splits during ancient waves of proliferation through the genome. Mol Biol Ev.
Jurka J, Kapitonov VV, Pavlicek A *et al.* (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogentic and Genome Research 110:462-467.
de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD (2011) Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. PLoS Genet. 7(12):e1002384
Kumar S, Hedges SB (2011) TimeTree2: species divergence times on the iPhone. Bioinformatics 27:2023-2024.
Lander ES, Linton LM, Birren B *et al.* (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921.