Alu elements in Primates are preferentially lost from areas of high GC content

Hellen & Brookfield (2013)

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

EHB Hellen & JFY Brookfield

BBSRC bioscience for the future

Introduction

The human genome project (Lander et al., 2001) argued strongly that the Alu sequences in our genome are mostly functional. This argument was based on the mean GC-content of DNA flanking Alu subfamilies of different ages. Alu elements from subfamilies with peak transpositional activity more than 5 million years ago tend to be located in high GC, gene-rich, areas of the genome. In contrast, elements belonging to "human-specific" subfamilies, peaking in activity within the last five million years, tend to be seen in low GC, gene-poor regions. In this analysis, the age of the subfamily was taken as a proxy for the age of insertion of the element.

They suggested 3 possible hypotheses to explain this:

1. **A higher rate of random loss of Alus in GC-poor regions**
2. **Negative selection iacting against Alu elements in GC-poor areas**
3. **Positive selection is acting in favour of Alu elements in GC-rich areas.**

The final hypothesis is deemed the most likely to be correct (Lander et al., 2001).

To explore these hypotheses further we have individually dated each element by looking for orthologous and syntenous sequences in other primate species. This will allow for the age/GC region correlation to be calculated for individual elements rather than assuming that all elements in a subfamily have a similar age and behaviour and are found in a similar GC regions.

Methods (See Hellen & Brookfield 2013 for more details)

Each element and its 500bp flanking region was separately BLASTed to find orthologous and syntenous sequences in other primate species. Age was assigned to each element depending on the primates the elements could be found in e.g. an element found in Human, Chimpanzee and Gorilla was assumed to be older than one found in only Human and Chimpanzee. Elements where the presence of orthologous elements was patchy, were removed from the analysis.

Elements were assigned to a subfamily by using the Repeat masker track on the UCSC genome browser. The % GC content of the flanking regions were calculated using bespoke perl scripts. Mean flanking region % GC values were calculated for elements divided by subfamily, by presence in primate orthologs (i.e. Age) and by both of these factors.

References

Belle EMS, Webster MT & Eyre-Walker A (2005) Why are young and old repetitive elements distributed differently in the human genome? *Journal of Molecular Evolution* 60:290–296 DOI 10.1007/s00239-004-0020-0.

Hellen EHB & Brookfield JFY (2013), Alu elements in primates are preferentially lost from areas of high GC content. *PeerJ* 1:e78; DOI 10.7717/peerj.78

Kapitonov V & Jurka J (1996) The age of Alu subfamilies. *Journal of Molecular Evolution* **42**:59–65 DOI 10.1007/BF00163212.

Lander et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921 DOI 10.1038/35057062.

Analysis by Subfamily

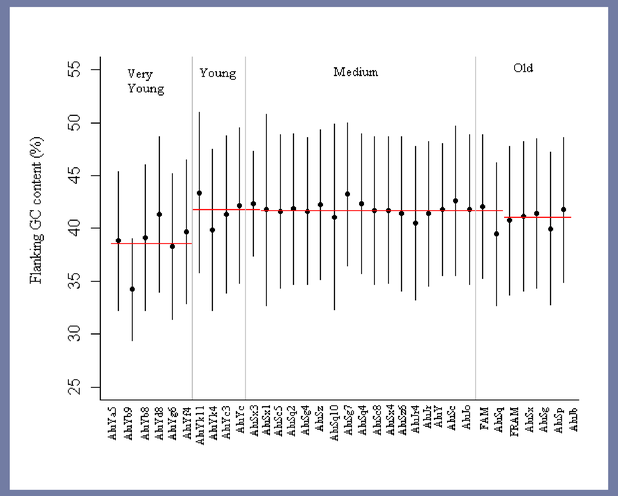

Figure 1: The GC flanking regions of Alu subfamilies with subfamily as a proxy for age.
34 Alu subfamilies are grouped into 4 ages: very young, young, medium, old. The mean GC flanking content of each family is plotted. Horizontal red lines represent the mean GC content of the age group, with bars showing standard deviations.

Conclusions

For all but the youngest subfamilies, the GC content flanking human-specific elements are found in areas of lower GC content than those of older families. If the mean GC flanking content of the most recent insertion events is the same as that for earlier insertion events for these subfamilies, then this must represent preferential loss from high GC regions. This is in direct contrast to the hypotheses from Lander et al. (2001) which suggests these regions are protected from loss by positive selection.

This hypothesis of loss from high GC regions is sufficient to explain the pattern of GC bias shown within the subfamilies and to explain the patterns shown (Figure 3) for all but the very young Alu subfamilies. The differences in GC flanking content shown by very young AluY elements appears to be due to a difference in insertion bias. As suggested by Belle, Webster & Eyre-Walker (2005), it appears as if these subfamilies have a different pattern of insertion from the old elements. However, we could find no systematic DNA sequence differences between the 'low GC' and 'high GC' subfamily groups which would account for this differing behaviour.

Analysis by Ortholog Detection

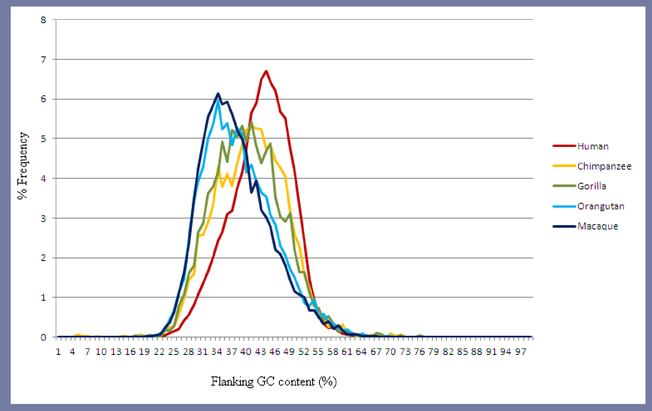

Figure 2: The frequency distribution of Alus with presence of primate orthologs as a proxy for age.
Frequency distribution of GC content of human Alu flanking regions, categorised by the most distant primate species in which orthologs could be found. Elements are found in the labelled species and all primate species with a smaller divergence time from human, but not in the species with a larger divergence time from human. GC content is calculated in 500bp up- and down-stream of the Alu element sequence in the human genome.

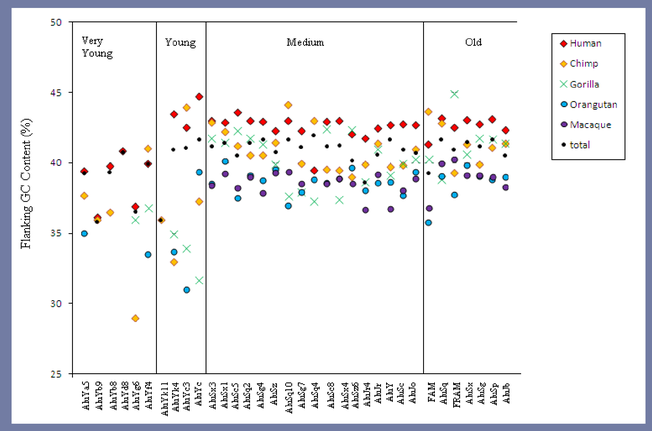

Figure 3: The GC flanking regions of Alu elements by both primate ortholog presence and subfamily.
Mean GC content of Alu flanking regions in each subfamily, classified by presence in each of the primate species in a cumulative manner.

elizabeth.hellen@nottingham.ac.uk

@Phdgeek

http://nottingham.academia.edu/ElizabethHellen