1-1-2004

Retrotransposition of Alu elements: How many sources?

Richard Cordaux
Louisiana State University

Dale J. Hedges
Louisiana State University

Mark A. Batzer
Louisiana State University

Recommended Citation

Cordaux, R., Hedges, D., & Batzer, M. (2004). Retrotransposition of Alu elements: How many sources?. *Trends in Genetics, 20* (10), 464-467. https://doi.org/10.1016/j.tig.2004.07.012

6 Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151–1155
7 The Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815
8 Simillion, C. *et al*. (2002) The hidden duplication past of *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U. S. A.* 99, 13627–13632
9 Blanc, G. *et al*. (2003) A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res.* 13, 137–144
10 Ermolaeva, M.D. *et al*. (2003) The age of the *Arabidopsis thaliana* genome duplication. *Plant Mol. Biol.* 51, 859–866
11 Bowers, J.E. *et al*. (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422, 433–438
12 Ku, H.M. *et al*. (2000) Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proc. Natl. Acad. Sci. U. S. A.* 97, 9121–9126
13 Yang, J. *et al*. (2003) Organismal complexity, protein complexity, and gene duplicability. *Proc. Natl. Acad. Sci. U. S. A.* 100, 15661–15665
14 Ashburner, M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29
15 Riddle, N.C. and Birchler, J.A. (2003) Effects of reunited diverged regulatory hierarchies in allopolyploids and species hybrids. *Trends Genet.* 19, 597–600
16 Birchler, J.A. *et al*. (2001) Dosage-dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploid syndromes, and quantitative traits. *Dev. Biol.* 234, 275–288
17 Osborn, T.C. *et al*. (2003) Understanding mechanisms of novel gene expression in polyploids. *Trends Genet.* 19, 141–147
18 Adams, K.L. *et al*. (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl. Acad. Sci. U. S. A.* 100, 4649–4654
19 Kondrashov, F.A. *et al*. (2002) Selection in the evolution of gene duplications. *Genome Biol.* 3. RESEARCH0008
20 Hughes, A.L. and Friedman, R. (2003) Parallel evolution by gene duplication in the genomes of two unicellular fungi. *Genome Res.* 13, 794–799
21 Wolfe, K.H. (2001) Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.* 2, 333–341

0168-9525/$ - see front matter
doi:10.1016/j.tig.2004.07.008

Genome Analysis

Retrotransposition of *Alu* elements: how many sources?

Richard Cordaux, Dale J. Hedges and Mark A. Batzer

Department of Biological Sciences, Biological Computation and Visualization Center, Louisiana State University, Baton Rouge, LA 70803, USA

It is generally thought that only a few *Alu* elements are capable of retrotransposition and that these 'master' sources produce inactive copies. Here, we use a network phylogenetic approach to demonstrate that recently integrated human-specific *Alu* subfamilies typically contain 10–20% of secondary source elements that contributed 20–40% of all subfamily members. This multiplicity of source elements provides new insight into the remarkably successful amplification strategy of the *Alu* family.

Alu inserts are short interspersed elements of ~300 base-pairs that have inserted in primate genomes within the last 65 million years through a mechanism termed retrotransposition [1]. They are the most abundant class of all mobile elements in the human genome, with >1 000 000 copies and making up >10% of the human genome by mass [1,2]. *Alu* elements have been reported to contribute to genetic disorders through insertional mutagenesis and postintegration recombination [3], to shape the architecture of the genome through segmental duplication and retrotransposition-mediated genomic deletion [4–6], and to affect proteome diversity through alternative splicing [7]. As such, their impact on the human genome and proteome has been substantial and it is therefore important to understand how these elements spread within their host genomes. One popular model of amplification of *Alu* elements is termed the 'master gene' model, in which only a few *Alu* elements are capable of retrotransposition and produce inactive copies [1,8,9] (Box 1). A strong argument for the master gene model is the hierarchical subfamily structure that typically characterizes *Alu* element sequence diversity [1,8,9]. Indeed, *Alu* subfamilies are collections of closely related *Alu* elements that share diagnostic nucleotide substitutions that are thought to arise in the master or source gene(s) and, subsequently, to represent a signature of close affinity to this master gene. However, alternative models in which many subfamily members are capable of generating new copies are also possible [9–11] (Box 1), and the actual number of truly retrotransposition-competent *Alu* elements in the human genome remains unresolved.

Benefits of networks over traditional phylogenetic methods

Phylogenetic methods have been widely used to study the relationships and evolution of mobile elements, including *Alu* elements. However, traditional phylogenetic methods used thus far assume bifurcating relationships and do not allow for persistent ancestral nodes (Box 2). Therefore, they might be inappropriate for reconstructing the genealogy of closely related sequences [12], such as those

Corresponding author: Mark A. Batzer (mbatzer@lsu.edu).
Available online 12 August 2004

Box 1. Models of *Alu* subfamily expansion

The popular 'master gene' model of *Alu* subfamily expansion posits that a single element α generated all other subfamily members, which are themselves inactive (Figure Ia). According to this model, the relationships between subfamily members will be star-like, with all inactive copies derived from the α element. By contrast, the extreme opposite model (termed the transposon model) posits that all subfamily members are capable of producing new elements (Figure Ic). In terms of relationships between subfamily members, this model is distinguished from the master gene model by its absence of radiating structure from a central node. In between these two extreme scenarios are intermediate models suggesting that several or many subfamily members are active and contribute to *Alu* subfamily expansion (Figure Ib). With this intermediate model, relationships among elements are expected to be at least partly star-like, but also to show varying proportions of elements that are not directly connected to the center of radiation.

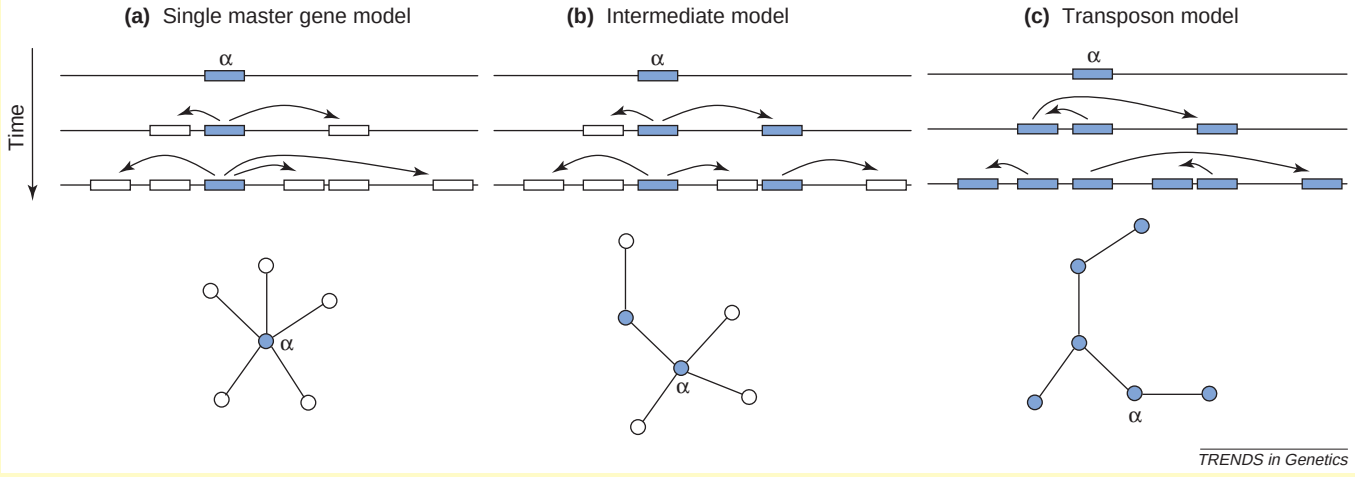

Figure I. Three models of *Alu* expansion. The active elements are represented in blue and the inactive elements in white. Branch lengths and circle size in the genealogies are arbitrary.

of *Alu* subfamilies (Box 2). These properties of *Alu* subfamilies are taken into account by network phylogenetic approaches (Box 2), making such methods better suited for studying evolutionary relationships of *Alu* elements. Here, we use networks [13] to investigate the relationships and expansion patterns of *Alu* subfamilies that have recently expanded in the human genome. What mobility model best fits the patterns of *Alu* subfamilies' sequence diversity? Are *Alu* subfamily members other than the 'master' gene capable of producing new copies? If they are, what is the proportion of these 'secondary' master genes and how large is their contribution to *Alu* subfamily expansions?

How many *Alu* element sources?

We analyzed 706 *Alu* elements belonging to all of the human-specific *Alu* subfamilies reported to date that have <310 members, which is the maximum number of sequences handled by the software NETWORK version 3.1 [13]. We used *Alu* subfamily sequence alignments

Box 2. Network versus traditional phylogenetic methods

Traditional phylogenetic methods have been designed to investigate interspecific relationships. Interspecific relationships are hierarchical because they are the product of reproductive isolation over long periods of time, leading to high divergence and non-overlapping gene pools. Thus, interspecific genealogies, as estimated by traditional phylogenetic methods, can typically be represented by strictly bifurcating trees (in which each ancestral branch splits into two descendant branches). In this case, all sampled units occupy terminal branches whereas all internal nodes (representing ancestors) are unsampled and therefore reconstructed.

By contrast, intraspecific relationships, or in a broader sense, relationships among closely related samples (such as *Alu* subfamily members), can be characterized by low divergence, multifurcating relationships and persistence of ancestral nodes. Datasets showing reduced variation will have fewer characters for analysis, which can result in poor resolution or incorrect inferences if traditional phylogenetic methods designed for highly divergent datasets are used. In addition, *Alu* subfamily members might be derived from a single source or master gene (see Box 1), which means that one element could have generated more than two descendant elements. This would yield *Alu* subfamily genealogies with true multifurcations (as illustrated in Box 1, Figure Ia), which violate the principle of bifurcating relationships assumed by traditional phylogenetic methods. Finally, *Alu* master or source genes might persist in their host genome and coexist with their descendants. This means that both ancestral and descendant *Alu* subfamily members can be sampled, which violates the principle of traditional phylogenetic methods according to which ancestral types are unsampled and have to be reconstructed.

Contrary to traditional methods, network phylogenetic approaches have been designed for investigating the relationships of closely related samples, and they allow for persistent ancestral nodes and multifurcations. The network approach is based on the parsimony principle and connects datasets in the way that requires the smallest number of evolutionary steps. However, contrary to traditional phylogenetic trees, networks depict alternative evolutionary pathways that require the same minimum number of steps, which create reticulations or loops in the network. In the absence of recombination, reticulations result from homoplasy (when two characters are identical by state – because of parallel or reverse mutations – and not by descent). The absence of homoplasy in the dataset results in networks without reticulations.

Table 1. Information on seven human-specific *Alu* subfamilies analyzed in this study

Alu subfamily	Sample size	Proportion of secondary source genes (%)		Contribution of secondary source genes (%)		IPL[a] (%)	Refs
		No correction	After correction	No correction	After correction		
Ya5a2	33	3	6	3	6	80	This study; [17]
Ya8	36	14	17	22	28	50	This study; [18]
Yb9	58	15	20	33	44	36	This study; [19]
Yc1	232	8	12	12	19	21	[20]
Yd6	96	10	15	14	21	12	[21]
Yg6	150	9	13	26	37	11	[6]
Yi6	101	10	16	30	42	10	[6]
All families[b]		9	14	20	28		
All except Ya5a2[b]		11	16	23	32		

[a]Insertion polymorphism level of *Alu* subfamilies.
[b]Average values.

published in the original papers characterizing these subfamilies (Table 1), except for Ya5a2, Ya8 and Yb9 subfamilies, whose elements were extracted from the July 2003 assembly of the human genome sequence, through a Basic Local Alignment Tool (BLAT) screening [14] of the human genome database. Sequence alignments are available on the authors' website (http://batzerlab.lsu.edu). These *Alu* subfamilies encompass a wide range of insertion polymorphism levels (IPL) per subfamily (80–10%, Table 1). The IPL is the proportion of *Alu* subfamily members that are polymorphic for presence/absence in the human population. As the IPL decreases with the time since subfamily expansion, it can be used as a proxy for estimating relative expansion times of the different *Alu* subfamilies that are independent of DNA sequence data.

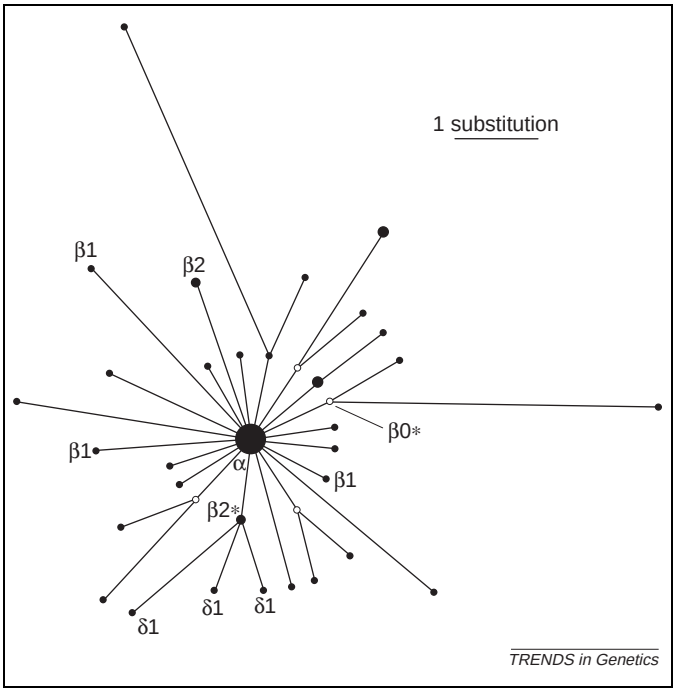

Figure 1. Median-joining network of the human-specific Yb9 *Alu* subfamily. Circles denote sequence types. The size of circles is proportional to the number of *Alu* loci with this sequence type. Lines denote substitution steps, with a one-step distance being indicated in the upper-right corner. Reconstructed nodes are identified as empty circles. The labels close to some nodes illustrate the nomenclature used to refer to the different components of the genealogy: the original master sequence of each *Alu* subfamily is called the α type, whereas derived types are called βx (where x is the number of loci matching this sequence type; if a type is reconstructed, then $x=0$) if they are directly derived from the α type or δx (where x is the number of loci matching this sequence type) if not so derived. When a βx type links α and δx types, it is identified by a star (e.g. βx*). $\beta 0$ types in the genealogies are reconstructed nodes absent from the dataset because: (i) they were not sampled; or (ii) they correspond to *Alu* loci that retrotransposed while polymorphic and that were eventually lost from the genome. The 3′ poly-A tails of all *Alu* elements were excluded from the analyses.

The networks of older subfamilies (IPL $<25\%$) showed multidimensional reticulations, which is suggestive of homoplasy in the data (Box 2). This is not surprising given the high number of CpG dinucleotides contained in *Alu* elements, which mutate at least at a sixfold higher rate compared with non-CpG dinucleotides [15,16]. When CpG dinucleotides were excluded, most reticulations disappeared from the networks. These results suggest that homoplasy is primarily attributable to CpG dinucleotides and that non-CpG sites are more stable and thus informative for reconstructing *Alu* subfamily genealogies. Therefore, to ensure that homoplasy would not affect the results, further analyses were performed using complete *Alu* sequences for the youngest *Alu* subfamilies (IPL $>35\%$), whereas CpG sites were disregarded for older subfamilies (IPL $<25\%$).

The Yb9 subfamily network displays a star-like phylogeny in which 36% of the elements fall in the central node α (Figure 1). This is typically expected under the 'master' gene model, where one *Alu* locus (the master gene) generated the other members of the subfamily. The node α can be inferred to be the ancestral node (and thus correspond to the original master or source gene sequence) of the Yb9 subfamily because: (i) it is the most frequent sequence type found in the subfamily; and (ii) it occupies a central position in the network [12]. Thus, overall, the spread of the Yb9 subfamily in the human genome is consistent with the master gene model of a single driver. Similar results were obtained for all other *Alu* subfamilies, in that the networks also displayed starlike topologies with the central nodes (α) corresponding to the most frequent sequence types found in each subfamily. However, within all *Alu* subfamilies, there were some sequence types that were not directly connected to the master sequence α (δx types; Figure 1) and others that were directly connected to the master sequence α, but were encompassing several *Alu* loci (βx and βx* types; Figure 1).

Because hypervariable sites were removed from the analyses when appropriate and the networks do not show any excess of multidimensional reticulations (Figure 1, Box 2), homoplasy can be considered as negligible. Thus, it is

unlikely that a δ*x* type and its most closely related β*x** type could be both independently and directly derived from the α type and would have accumulated substitutions independently at the same positions. If so, the most parsimonious explanation for these patterns is that δ*x* sequences are derived from β*x** types rather than directly from the α type. The same reasoning leads us to conclude that for β*x* types with $x \geq 2$, the two or more sequences of any given node are not independently and directly derived from the α type, but rather that one of the loci gave rise to the other members of the given β*x* node. We conclude that β*x* with $x \geq 2$ and β*x** types correspond to secondary source genes capable of amplification within the *Alu* subfamilies. The network approach shows that *Alu* subfamilies examined all contain between 3% and 15% of such secondary source or submaster genes (Table 1, Figure 1). We also estimate that these secondary source genes generated between 3% and 33% of the total members of each *Alu* subfamily (Table 1, Figure 1). On average, we find that *Alu* subfamilies comprise ~9% of secondary source genes that contributed ~20% of subfamily copies. These values could underestimate the true proportions because it is likely that the α type of each *Alu* subfamily might encompass several loci that could contribute new subfamily members. To correct for this possibility, we assumed that α nodes contained the same proportion of secondary source genes as estimated for each subfamily without correction. The corresponding number of elements was added to the uncorrected number of secondary source elements initially estimated in each subfamily, allowing estimation of a corrected proportion of secondary sources in the subfamilies (Table 1). This conservative correction suggests that *Alu* subfamilies show on average ~15% of secondary source genes contributing ~30% of subfamily members.

Multiple sources and the evolutionary success of *Alu* elements

In summary, we confirm here that human *Alu* subfamilies do not follow a single 'master' gene model of expansion. Indeed, the 'sprout' or multiple source model [9–11] best explains the observed patterns of *Alu* subfamily sequence variation, in which *Alu* subfamilies contain secondary source genes that can contribute a substantial portion of subfamily members. It is noteworthy that the *Alu* subfamilies examined here consistently show 10–20% of secondary sources contributing 20–40% of the subfamily members, regardless of the IPL or subfamily copy number. Although these estimates might vary for the oldest *Alu* subfamilies (with IPL <10%), this considerably strengthens the credibility of the sprout model of human *Alu* subfamily expansion over other previously proposed models, because it spans multiple *Alu* subfamilies that have amplified at different times throughout human evolution. The existence of a considerable number of active elements with lower levels of amplification instead of a few hyperactive 'master' genes might have been the evolutionary strategy that enabled *Alu* elements to bypass mutational inactivation, negative selection and/or putative host defense mechanisms that could have limited their expansion. This ultimately contributed to make *Alu* elements the most successful class of mobile elements (in terms of copy number) in the human genome.

Acknowledgements

We thank David Ray, Jerilyn Walker and Jinchuan Xing for comments on an earlier version of the manuscript. This research was supported by Louisiana Board of Regents Millennium Trust Health Excellence Fund HEF (2000-05)-05 and (2000-05)-01 (M.A.B), and by National Science Foundation grant BCS-0218338 (M.A.B).

References

1 Batzer, M.A. and Deininger, P.L. (2002) *Alu* repeats and human genomic diversity. *Nat. Rev. Genet.* 3, 370–379
2 Lander, E.S. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
3 Deininger, P.L. and Batzer, M.A. (1999) *Alu* repeats and human disease. *Mol. Genet. Metab.* 67, 183–193
4 Bailey, J.A. *et al*. (2003) An *Alu* transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* 73, 823–834
5 Hayakawa, T. *et al*. (2001) *Alu*-mediated inactivation of the human CMP-*N*-acetylneuraminic acid hydroxylase gene. *Proc. Natl. Acad. Sci. U. S. A.* 98, 11399–11404
6 Salem, A.H. *et al*. (2003) Recently integrated *Alu* elements and human genomic diversity. *Mol. Biol. Evol.* 20, 1349–1361
7 Lev-Maor, G. *et al*. (2003) The birth of an alternatively spliced exon: 3′ splice-site selection in *Alu* exons. *Science* 300, 1288–1291
8 Deininger, P.L. *et al*. (1992) Master genes in mammalian repetitive DNA amplification. *Trends Genet.* 8, 307–311
9 Deininger, P.L. and Batzer, M.A. (1993) Evolution of retroposons. In *Evolutionary Biology* (Vol. 27) (Hecht, M.K. *et al.*, eds), pp. 157–196, Plenum Press
10 Brookfield, J.F.Y. (1993) The generation of sequence similarity in SINEs and LINEs. *Trends Genet.* 9, 38–39
11 Matera, A.G. *et al*. (1990) Recently transposed *Alu* repeats result from multiple source genes. *Nucleic Acids Res.* 18, 6019–6023
12 Posada, D. and Crandall, K.A. (2001) Intraspecific gene genealogies: trees grafting into networks. *Trends Ecol. Evol.* 16, 37–45
13 Bandelt, H.J. *et al*. (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* 16, 37–48
14 Kent, W.J. (2002) BLAT – the BLAST-like alignment tool. *Genome Res.* 12, 656–664
15 Labuda, D. and Striker, G. (1989) Sequence conservation in *Alu* evolution. *Nucleic Acids Res.* 17, 2477–2491
16 Batzer, M.A. *et al*. (1990) Structure and variability of recently inserted *Alu* family members. *Nucleic Acids Res.* 18, 6793–6798
17 Roy, A.M. *et al*. (2000) Potential gene conversion and source genes for recently integrated *Alu* elements. *Genome Res.* 10, 1485–1495
18 Roy, A.M. *et al*. (1999) Recently integrated human *Alu* repeats: finding needles in the haystack. *Genetica* 107, 149–161
19 Roy-Engel, A.M. *et al*. (2001) *Alu* insertion polymorphisms for the study of human genomic diversity. *Genetics* 159, 279–290
20 Garber, R.K. *et al.* The *Alu* Yc1 subfamily: sorting the wheat from the chaff. *Cytogenet. Genome Res.* (in press)
21 Xing, J. *et al*. (2003) Comprehensive analysis of two *Alu* Yd subfamilies. *J. Mol. Evol.* 57, S76–S89

0168-9525/$ - see front matter
doi:10.1016/j.tig.2004.07.012