Available online at www.sciencedirect.com
SCIENCE DIRECT
Gene 373 (2006) 134–137
GENE
SECTION
EVOLUTIONARY GENOMICS
www.elsevier.com/locate/gene

Estimating the retrotransposition rate of human *Alu* elements

Richard Cordaux, Dale J. Hedges, Scott W. Herke, Mark A. Batzer *

Department of Biological Sciences, Biological Computation and Visualization Center, Center for BioModular Multi-scale Systems, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803, USA

Received 29 November 2005; received in revised form 18 January 2006; accepted 21 January 2006
Available online 7 March 2006
Received by W. Makalowski

Abstract

Mobile elements such as *Alu* repeats have substantially altered the architecture of the human genome, and de novo mobile element insertions sometimes cause genetic disorders. Previous estimates for the retrotransposition rate (RR) of *Alu* elements in humans of one new insertion every ~100–125 births were developed prior to the sequencing of the human and chimpanzee genomes. Here, we used two independent methods (based on the new genomic data and on disease-causing de novo *Alu* insertions) to generate refined *Alu* RR estimates in humans. Both methods consistently yielded RR on the order of one new *Alu* insertion every ~20 births, despite the fact that the evolutionary-based method represents an average RR over the past ~6 million years while the mutation-based method better reflects the current-day RR. These results suggest that *Alu* elements retrotranspose at a faster rate in humans than previously thought, and support the potential of *Alu* elements as mutagenic factors in the human genome.

Keywords: SINEs; Retrotransposons; Mobile elements

1. Introduction

About 30% of the human genome is comprised of *Alu* and LINE-1 (L1) retrotransposons (Lander et al., 2001), which are the two major families of mobile elements that are currently expanding in humans (Smit, 1999; Ostertag and Kazazian, 2001; Batzer and Deininger, 2002; Deininger et al., 2003; Cordaux et al., 2004; Hedges and Batzer, 2005). New *Alu* and L1 copies are generated via a "copy and paste" mechanism, in which the RNA transcript of an active element is reverse transcribed as cDNA and the duplicate element is inserted at a new genomic location (Rogers, 1985; Ostertag and Kazazian, 2001; Batzer and Deininger, 2002). While L1 elements are autonomous because they encode the proteins required for retrotransposition, *Alu* elements are non-autonomous and are *trans*-mobilized by the L1 retrotransposition machinery (Sakaki et al., 1986; Skowronski et al., 1988; Feng et al., 1996; Ostertag and Kazazian, 2001; Wei et al., 2001; Dewannieux et al., 2003).

Alu and L1 elements have substantially impacted the architecture of the genome and their insertions have caused a number of human genetic disorders (Deininger and Batzer, 1999; Ostertag and Kazazian, 2001; Batzer and Deininger, 2002; Deininger et al., 2003; Chen et al., 2005). Because of these extensive interactions, it is important to understand the dynamics of expansion of these mobile element families. For instance, how often are new retrotransposon copies generated in the human genome? Retrotransposition rates (RR) for all classes of retrotransposons combined or L1 elements only have been previously estimated to be on the order of one new insertion every 2–30 births, based on the frequency of disease-causing de novo retrotransposon events compared to nucleotide mutations and cell-culture based retrotransposition assays (Kazazian, 1999; Li et al., 2001; Brouha et al., 2003). By contrast, *Alu* RR estimated using evolutionary approaches are on the order of one insertion every 100–125 births (Deininger and Batzer, 1993; Deininger and Batzer, 1995). Because these *Alu* RR were obtained prior to the sequencing of the human and chimpanzee genomes (which allow higher resolution estimates of the number of human and chimpanzee lineage-specific *Alu*

Abbreviations: RR, retrotransposition rate; L1, LINE-1; HGMD, Human Gene Mutation Database.

* Corresponding author. Tel.: +1 225 578 7102; fax: +1 225 578 7113.
E-mail address: mbatzer@lsu.edu (M.A. Batzer).

0378-1119/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2006.01.019

elements), we have re-evaluated the evolutionary-based *Alu* RR estimates. In addition, we have used a second independent approach to estimate the *Alu* RR, based on the frequency of disease-causing de novo *Alu* insertions as compared to nucleotide mutations (Kazazian, 1999). Strikingly, we found that, although the two methods use different assumptions, they both produce similar *Alu* RR estimates of one insertion every ~20 births.

2. Materials and methods

2.1. Evolutionary-based method

First, using an evolutionary framework, we estimated the average human *Alu* RR over the past ~6 million years. Assuming a neutral model of evolution, we inferred the RR that explains the copy number of fixed human-specific *Alu* elements if genetic drift is the sole driving evolutionary force. Under this model, the frequency of each element in the population fluctuates randomly and this force drives the loss or fixation of the element. Population genetics theory predicts that the probability of fixation of any element under neutral evolution is $1/2N_e$, where N_e is the human effective population size (Graur and Li, 2000). Thus, the expected total number of *Alu* insertions (i.e. encompassing insertions that ultimately became fixed or lost, as well as currently polymorphic elements) that must have occurred in humans to produce n fixed human-specific elements at the time of observation (present-day) is $n2N_e$. For a human–chimpanzee divergence time t and a generation time g, the RR expected under a neutral model of evolution is $n2N_e/(t/g)/N_e$ new insertion events per birth. As this expression results in a fractional insert value, it can be more intuitively expressed as one new insertion event every $(2ng/t)^{-1}$ births. We emphasize here that, because N_e is factored out of the equation, the number or severity of population bottlenecks having occurred in recent human evolution is irrelevant to this calculation.

2.2. Mutation-based method

The second approach we used estimates the current-day human *Alu* RR by comparing the frequency of de novo *Alu* insertions involved in genetic disease in humans to that of de novo nucleotide mutations causing disease in the same set of genes (Kazazian, 1999), as reported in the Human Gene Mutation Database (HGMD) (Stenson et al., 2003; Chen et al., 2005). The HGMD has not been specifically designed to calculate RR. Therefore, the accuracy of this approach is limited by the data present in the database and the results can potentially be influenced by several sources of bias, such as unreported mutations at non-disease-causing sites and parallel mutations. Using estimates of the extent of bias that these two major sources (see descriptions below) may exert on the HGMD data, we corrected the observed *Alu* RR to obtain more reliable estimates.

2.2.1. Unreported mutations at non-disease-causing sites

While most (if not all) *Alu* insertions in the protein encoding portion of a gene are likely to have detectable phenotypic effects (e.g., by truncating the protein encoded by the gene), many nucleotide substitutions do not modify the encoded amino acid (i.e., synonymous substitutions) or result in a conservative amino acid substitution with no detectable phenotypic effect. Such single base substitutions are not listed in the HGMD, which therefore leads to an overestimate of the RR because it decreases the frequency of nucleotide mutations relative to *Alu* insertion events. To quantify this effect, we estimated the proportion of synonymous nucleotide substitutions and non-synonymous nucleotide substitutions resulting in conservative amino acid changes out of all possible nucleotide substitutions. The proportions of synonymous and non-synonymous nucleotide substitutions were estimated according to the codon base positions of the human genetic code (Strachan and Read, 2004). Conservative and non-conservative amino acid substitutions were defined according to the BLOSUM62 matrix (Henikoff and Henikoff, 1992); changes with a positive or neutral sign in the matrix were considered conservative changes while those with negative signs were considered non-conservative (Cargill et al., 1999).

2.2.2. Parallel mutations

Alu elements are essentially homoplasy-free characters, particularly at the scale of human evolution (Ray et al., 2005; Salem et al., 2005; Xing et al., 2005), indicating that *Alu* parallel insertions are not a concern for estimating RR. By contrast, nucleotides are more prone to parallel or homoplasic mutations and nucleotide mutations at any site are counted only once in the HGMD (Stenson et al., 2003), even though multiple independent nucleotide mutations might have occurred at the same site (Li et al., 2001; Stenson et al., 2003). This source of bias leads to an overestimate of the RR. To quantify this effect, we estimated the proportion of parallel nucleotide mutations from ~5.5 kb of pseudogene sequence determined in 100 human chromosomes (Martinez-Arias et al., 2001). As the neutral substitution rate equals the spontaneous mutation rate (Graur and Li, 2000; Nachman and Crowell, 2000), the proportion of parallel substitutions observed in neutrally evolving pseudogene sequences equals the proportion of spontaneous parallel mutations in the HGMD.

3. Results

3.1. Evolutionary-based method

The comparison of the human and chimpanzee genomes yielded an estimate of the total number of human-specific *Alu* elements of 7082 (Chimpanzee Sequencing and Analysis Consortium, 2005). This figure includes both polymorphic and fixed elements. Since the vast majority of human-specific *Alu* elements belong to the Ya and Yb lineages in which ~80% of the elements are fixed in the population (Carter et al., 2004; Otieno et al., 2004), we deduce that there are ~5700 fixed human-specific *Alu* elements. Using this number of fixed human-specific *Alu* copies n, a human–chimpanzee divergence time t of 6 million years (Goodman et al., 1998), and a generation time g of 20–30 years, we estimate that the average *Alu* RR has been one insertion every 18–26 individuals over the past ~6 million years.

Table 1
Human *Alu* retrotransposition rate (births/insertion) based on disease-causing mutations

Number of de novo mutations	*Alu* elements	Nucleotide mutations
Reported in the HGMD as of November 2004	26	29,505
Missed due to unreported mutations at non-disease-causing sites	–	21,495 [a]
Missed due to parallel mutations	–	0–6000 [b]
Corrected estimates	26	51,000–57,000
Alu retrotransposition rate	14–16 [c]	

[a] Total non-synonymous nucleotide substitutions (i.e., encompassing non-conservative and conservative amino acid changes) = 38,930 (29,505/75.8%). The value of 75.8% is derived from the BLOSSUM62 amino acid substitution matrix (Henikoff and Henikoff, 1992), in which 144/190 (75.8%) of all possible amino acid substitutions are predicted to be non-conservative. Total nucleotide mutations (i.e., encompassing non-synonymous and synonymous substitutions) = 51,000 (38,930/76.3%). The value of 76.3% is derived as follows: ~96%, 100% and ~33% of all substitutions occurring at first, second and third codon positions, respectively, are non-synonymous (Strachan and Read, 2004), and codon usage barely affects these percentages (e.g., accounting for codon usage shifts the proportion of non-synonymous substitutions at first codon position from 95.6% to 96.1%). Thus, assuming that each codon position has the same probability of being hit by a spontaneous mutation, 76.3% (96+100+33 out of 300 spontaneous mutations) are expected to be non-synonymous.

[b] = (29,505+21,495)×11.8%. The value of 11.8% is derived from the analysis of ~500 kb of human pseudogene sequence that contained 17 polymorphic sites involving nucleotide substitutions where parallel mutations were inferred to have occurred at no more than two sites (Martinez-Arias et al., 2001). Thus, the proportion of polymorphic sites having experienced parallel mutations ranges from 0% (0/17) to 11.8% (2/17).

[c] 51,000/26/140 and 57,000/26/140 (where 140 is the number of spontaneous nucleotide mutations per diploid genome per generation; see Section 3.2).

3.2. Mutation-based method

As of November 2004, the HGMD contained 29,505 independent de novo nucleotide mutations as well as 26 de novo *Alu* insertions located in 1690 genes (Stenson et al., 2003; Chen et al., 2005). Correcting for unreported nucleotide mutations in the HGMD occurring at sites where no disease-causing mutation is known and for parallel nucleotide mutations results in a total of 51,000–57,000 actual nucleotide mutations (see Table 1 for details). Thus, the frequency of *Alu* insertions relative to nucleotide mutations is one in ~2000–2200. Assuming a spontaneous nucleotide mutation rate of 2.3×10^{-8} per site per generation (Nachman and Crowell, 2000) and 6.2×10^{9} nucleotides per diploid genome, there are on average ~140 spontaneous nucleotide mutations per diploid genome per generation. This translates into a current-day *Alu* RR of one new insertion every ~15 individuals (Table 1).

4. Discussion

The two methods we used consistently yielded *Alu* RR on the order of one insertion every ~20 births. The fact that these methods are independent in that they are based on different assumptions and use different datasets adds to the credibility of our *Alu* RR estimate, even though each method has its own inherent uncertainties. For example, for the method using disease-causing *Alu* elements and nucleotide mutations, we used a database that was not originally designed for the purpose of calculating RR (Stenson et al., 2003). Therefore, although the HGMD is the best suited database currently available for the type of analysis we performed, it has inherent biases that need to be corrected. We speculate that if more accurate *Alu* RR based on the disease-causing elements method can be estimated in the future, it may become possible to address questions related to selective forces acting on *Alu* elements, by comparing the improved *Alu* RR estimate to that expected under neutral evolution. Also, it is noteworthy that the evolutionary RR is an average RR over the past few million years and it is possible that the *Alu* RR has varied during this time period. The fact that different recent *Alu* subfamilies have different estimated ages and copy numbers (Xing et al., 2004; Hedges et al., 2005) is consistent with the idea of temporal *Alu* RR variation, although its magnitude is difficult to estimate. The observation that the disease-based *Alu* RR (which better reflects the present-day *Alu* RR) is slightly faster than the evolutionary *Alu* RR may suggest that, within the limits of accuracy of the methods used, *Alu* elements are currently in a phase of slightly higher retrotranspositional activity than average.

The *Alu* RR we estimated here is about five times faster than the *Alu* RR previously estimated by evolutionary methods (Deininger and Batzer, 1993, 1995). This is not completely unexpected given that the recent availability of sequence data from the human and chimpanzee genome projects has allowed refined estimates of the number of human-specific *Alu* elements (Lander et al., 2001; Chimpanzee Sequencing and Analysis Consortium, 2005). Our results confirm that *Alu* elements continue to vigorously proliferate in the human lineage (Hedges et al., 2004; Chen et al., 2005; Chimpanzee Sequencing and Analysis Consortium, 2005), although the current *Alu* RR is much lower than it was ~40 million years ago when the majority of *Alu* elements were generated (Batzer and Deininger, 2002).

Our results also suggest that *Alu* and L1 RR are fairly similar, both being on the order of one insertion every ~20 births (this study; Kazazian, 1999; Li et al., 2001; Brouha et al., 2003). This suggests that *Alu* elements might have a similar retrotransposition efficiency compared to L1 elements vis-à-vis the molecular retrotransposition machinery. This result is somewhat unexpected, given that L1 proteins are more efficient at retrotransposing the mRNA that encodes them than they are for any other RNA, including *Alu* RNA (Dewannieux et al., 2003). Nevertheless, *Alu* elements can be retrotransposed by L1 elements in which only ORF2 is functional; by contrast, L1 retrotransposition requires both ORF1 and ORF2 to be functional (Ostertag and Kazazian, 2001; Dewannieux et al., 2003).

In conclusion, previous estimates of one new *Alu* insertion every ~100–125 births were developed prior to the sequencing of the human and chimpanzee genomes. Here, we have used two independent methods (based on the new genomic data and on disease-causing de novo *Alu* insertions) to provide refined estimates for the RR of *Alu* elements in humans. Both methods consistently yielded RR on the order of one new *Alu* insertion every ~20 births, despite the fact that the evolutionary-based method represents an average RR over ~6 million years while

the mutation method represents current day RR. These results suggest that *Alu* elements retrotranspose in humans at a rate about five times faster than previously thought, and they testify to the high potential of *Alu* elements as mutagenic factors in the human genome.

Acknowledgments

We thank J.M. Chen, P.D. Stenson, D.N. Cooper and C. Férec for sharing data prior to publication. We also thank D. Grover, D.A. Ray and M. Konkel for useful discussions. This research was supported by Louisiana Board of Regents Millennium Trust Health Excellence Fund HEF (2000–05)-05, (2000–05)-01 and (2001–06)-02; National Institutes of Health GM 59290, National Science Foundation grants BCS-0218338 and EPS-0346411, and the State of Louisiana Board of Regents Support Fund.

References

Batzer, M.A., Deininger, P.L., 2002. Alu repeats and human genomic diversity. Nat. Rev. Genet. 3, 370–379.

Brouha, B., et al., 2003. Hot L1s account for the bulk of retrotransposition in the human population. Proc. Natl. Acad. Sci. U. S. A. 100, 5280–5285.

Cargill, M., et al., 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat. Genet. 22, 231–238.

Carter, A.B., et al., 2004. Genome-wide analysis of the human Alu Yb-lineage. Hum. Genomics 1, 167–178.

Chen, J.M., Stenson, P.D., Cooper, D.N., Ferec, C., 2005. A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. Hum. Genet. 117, 411–427.

Chimpanzee Sequencing and Analysis Consortium, 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437, 69–87.

Cordaux, R., Hedges, D.J., Batzer, M.A., 2004. Retrotransposition of Alu elements: how many sources? Trends Genet. 20, 464–467.

Deininger, P.L., Batzer, M.A., 1993. Evolution of retroposons. Evol. Biol. 27, 157–196.

Deininger, P.L., Batzer, M.A., 1995. SINE master genes and population biology. In: Maraia, R.J. (Ed.), The Impact of Short Interspersed Elements (SINEs) on the Host Genome. RG Landes Publishers, Georgetown, pp. 43–60.

Deininger, P.L., Batzer, M.A., 1999. Alu repeats and human disease. Mol. Genet. Metab. 67, 183–193.

Deininger, P.L., Moran, J.V., Batzer, M.A., Kazazian Jr., H.H., 2003. Mobile elements and mammalian genome evolution. Curr. Opin. Genet. Dev. 13, 651–658.

Dewannieux, M., Esnault, C., Heidmann, T., 2003. LINE-mediated retrotransposition of marked Alu sequences. Nat. Genet. 35, 41–48.

Feng, Q., Moran, J.V., Kazazian Jr., H.H., Boeke, J.D., 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87, 905–916.

Goodman, M., et al., 1998. Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. Mol. Phylogenet. Evol. 9, 585–598.

Graur, D., Li, W.H., 2000. Fundamentals of Molecular Evolution, 2 ed. Sinauer Associates, Sunderland.

Hedges, D.J., Batzer, M.A., 2005. From the margins of the genome: mobile elements shape primate evolution. BioEssays 27, 785–794.

Hedges, D.J., Callinan, P.A., Cordaux, R., Xing, J., Barnes, E., Batzer, M.A., 2004. Differential Alu mobilization and polymorphism among the human and chimpanzee lineages. Genome Res. 14, 1068–1075.

Hedges, D.J., et al., 2005. Modeling the amplification dynamics of human Alu retrotransposons. PLoS Comput. Biol. 1, e44.

Henikoff, S., Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U. S. A. 89, 10915–10919.

Kazazian Jr., H.H., 1999. An estimated frequency of endogenous insertional mutations in humans. Nat. Genet. 22, 130.

Lander, E.S., et al., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Li, X., et al., 2001. Frequency of recent retrotransposition events in the human factor IX gene. Hum. Mutat. 17, 511–519.

Martinez-Arias, R., Calafell, F., Mateu, E., Comas, D., Andres, A., Bertranpetit, J., 2001. Sequence variability of a human pseudogene. Genome Res. 11, 1071–1085.

Nachman, M.W., Crowell, S.L., 2000. Estimate of the mutation rate per nucleotide in humans. Genetics 156, 297–304.

Ostertag, E.M., Kazazian Jr., H.H., 2001. Biology of mammalian L1 retrotransposons. Annu. Rev. Genet. 35, 501–538.

Otieno, A.C., et al., 2004. Analysis of the human Alu Ya-lineage. J. Mol. Biol. 342, 109–118.

Ray, D.A., et al., 2005. Alu insertion loci and platyrrhine primate phylogeny. Mol. Phylogenet. Evol. 35, 117–126.

Rogers, J.H., 1985. The origin and evolution of retroposons. Int. Rev. Cytol. 93, 187–279.

Sakaki, Y., Hattori, M., Fujita, A., Yoshioka, K., Kuhara, S., Takenaka, O., 1986. The LINE-1 family of primates may encode a reverse transcriptase-like protein. Cold Spring Harbor Symp. Quant. Biol. 51 (Pt 1), 465–469.

Salem, A.H., Ray, D.A., Batzer, M.A., 2005. Identity by descent and DNA sequence variation of human SINE and LINE elements. Cytogenet. Genome Res. 108, 63–72.

Skowronski, J., Fanning, T.G., Singer, M.F., 1988. Unit-length line-1 transcripts in human teratocarcinoma cells. Mol. Cell. Biol. 8, 1385–1397.

Smit, A.F., 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. 9, 657–663.

Stenson, P.D., et al., 2003. Human gene mutation database (HGMD): 2003 update. Hum. Mutat. 21, 577–581.

Strachan, T., Read, A.P., 2004. Instability of the human genome: mutation and DNA repair. In: Strachan, T., Read, A.P. (Eds.), Human Molecular Genetics. Garland Science, New York, pp. 315–349.

Wei, W., et al., 2001. Human L1 retrotransposition: *cis* preference versus *trans* complementation. Mol. Cell. Biol. 21, 1429–1439.

Xing, J., Hedges, D.J., Han, K., Wang, H., Cordaux, R., Batzer, M.A., 2004. Alu element mutation spectra: molecular clocks and the effect of DNA methylation. J. Mol. Biol. 344, 675–682.

Xing, J., et al., 2005. A mobile element based phylogeny of Old World monkeys. Mol. Phylogenet. Evol. 37, 872–880.