Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2006, Article ID 75327, Pages 1–5
DOI 10.1155/JBB/2006/75327

Mini-Review Article

The Genomic Distribution of L1 Elements: The Role of Insertion Bias and Natural Selection

Todd Graham[1] and Stephane Boissinot[1, 2]

[1] *Department of Biology, Queens College, City University of New York, Flushing, NY 11367, USA*
[2] *Graduate School and University Center, City University of New York, New York, NY 10016, USA*

Received 5 August 2005; Revised 6 December 2005; Accepted 13 December 2005

LINE-1 (L1) retrotransposons constitute the most successful family of retroelements in mammals and account for as much as 20% of mammalian DNA. L1 elements can be found in all genomic regions but they are far more abundant in AT-rich, gene-poor, and low-recombining regions of the genome. In addition, the sex chromosomes and some genes seem disproportionately enriched in L1 elements. Insertion bias and selective processes can both account for this biased distribution of L1 elements. L1 elements do not appear to insert randomly in the genome and this insertion bias can at least partially explain the genomic distribution of L1. The contrasted distribution of L1 and Alu elements suggests that postinsertional processes play a major role in shaping L1 distribution. The most likely mechanism is the loss of recently integrated L1 elements that are deleterious (negative selection) either because of disruption of gene function or their ability to mediate ectopic recombination. By comparison, the retention of L1 elements because of some positive effect is limited to a small fraction of the genome. Understanding the respective importance of insertion bias and selection will require a better knowledge of insertion mechanisms and the dynamics of L1 inserts in populations.

INTRODUCTION

The sequencing of several mammalian genomes has revealed that all are littered with hundreds of thousands copies of LINE-1 (L1) retrotransposons that account for ~ 20% of their mass (Lander et al [1], Waterston et al [2]). The abundance of L1 elements in mammalian genomes is specific of this class of vertebrates and should be considered a diagnostic feature of mammals to the same extent the possession of hair and the production of milk by females are. As L1 elements are also responsible for the amplification of SINEs (eg, Alu in primates, B1 and B2 in mouse) and processed pseudogenes (Esnault et al [3], Dewannieux and Heidmann [4], Dewannieux et al [5]), it is believed that L1 activity may account for as much as 50% of mammalian DNA.

Although L1 elements can be found almost anywhere in the genome, their abundance varies considerably among genomic regions. In general, L1 elements are much more abundant in AT-rich, low-recombining, and gene-poor regions of the genome. In addition to this general trend, L1 elements can be locally very rare or extremely abundant. For instance, L1 elements constitute 89% of a 100 Kb region on chromosome X while they are virtually absent from the homeobox gene clusters (Lander et al [1]). They seem to be more abundant on the sex chromosomes (Korenberg and Rykowski [6], Boyle et al [7], Bailey et al [8], Boissinot et al [9], Parish et al [10]), in genes that are expressed at low level (Han et al [11]), and in monoallelically expressed genes (Allen et al [12]). Differences exist in the distribution of L1 elements with regard to their age and size. Younger L1 elements are located on average closer to genes than older elements (Medstrand et al [13]) and full-length elements are more abundant on the sex chromosomes than on autosomes (Boissinot et al [9]). Although most of the L1 elements found in the human and mouse genomes were inserted after the split between primates and rodents, their distributions are strikingly similar (Lander et al [1], Waterston et al [2]), suggesting that some common mechanisms have shaped L1 distribution in primates and rodents. Here we review the molecular mechanisms and evolutionary processes that might have played a role in shaping the genomic distribution of L1 elements and we evaluate their relative contribution to the biased distribution of L1 elements.

L1 ELEMENTS ARE NOT INSERTED RANDOMLY

The first possible source of bias comes from the retrotransposition process itself. The reaction of the retrotransposition

requires the target site to be cut by the L1-encoded endonuclease. As the consensus target site of L1 endonuclease is TT/AAAA (Jurka [14]), it is plausible that L1 inserts preferentially in AT-rich regions because this motif is overrepresented in these regions (Cost and Boeke [15]). It was even suggested that the preference of L1 elements for AT-rich regions could be an adaptation of L1 to its host because insertion of L1 in gene-poor regions limits the burden of L1 retrotransposition (Lander et al [1], Cost and Boeke [15]). However, the majority of L1 insertions sites differ from the insertion site consensus sequence (Jurka [14], Cost and Boeke [15]) and there is probably no shortage of insertion sites anywhere in the genome. In addition, Alu elements are more abundant in GC-rich regions of the genome despite the fact that they have the same consensus target site-as L1 elements (Jurka [14]). Although it is likely that the target-site preference of L1 endonuclease is, at least in part, responsible for the distribution bias of L1, this hypothesis has not been tested rigorously.

Beside the possible bias caused by the L1 endonuclease, the analysis of *de novo* insertions and of recently integrated elements revealed the presence of insertional hotspots in the human genome. Of 14 *de novo* disease-causing insertions listed in Ostertag and Kazazian [16], three were in the factor VIII gene, four in the dystrophin gene, and two in the CYBB gene. Another set of genes was the target of multiple L1 and L1-mediated (Alu, SVA) insertions: an L1 and two Alu elements inserted in the factor IX gene, and an Alu and an SVA inserted in the BTK gene. A novel L1 insertion in the factor IX of dog has recently been described (Brooks et al [17]) and two L1 insertions occurred recently and independently in human and gorilla within the same 1 Kb region (DeBerardinis and Kazazian [18]). In addition, a recent analysis of the currently amplifying Ta-1 family found that a number of recently integrated Ta-1 elements were clustered in the human genome more often than expected by chance suggesting the existence of insertional hotspots on several autosomes (Boissinot et al [19]). Together these observations indicate that some genomic regions are more likely to be the target of L1 retrotransposition events than others, and suggest that insertional hotspots may be conserved among mammalian species. It is unclear why some genomic regions are insertional hotspots but it is plausible that the transcriptional status of the target site region plays a role. If the structure of the DNA is modified during transcription in a way that makes it more hospitable for L1 retrotransposition, transcriptionally active regions would undergo a higher number of transposition events. This hypothesis requires further investigations with regard to some identified hotspots (Boissinot et al [19]). While some of these hotspots were in the vicinity of genes expressed in gonads and during early embryogenesis (Boissinot, Entezam, and Furano, unpublished observation), a genome-wide analysis of genes that are transcribed in testes failed to find a significant excess of L1 elements in those genes (Graham and Boissinot, unpublished observation). A recent analysis of L1 retrotransposition in neuronal precursor cells showed that a number of *de novo* L1 insertions were in neuronally expressed genes lending support for some relationship between the transcriptional activity of a gene and its hospitability to novel L1 insertions (Muotri et al [20]). Because insertional hotspots are not particularly enriched in old L1 elements, their contribution to the biased distribution of L1 remains unclear but they could very well explain the local abundance of elements in certain genes.

The abundance of recent L1 insertions varies significantly among chromosomes, with chromosomes 4 and X apparently being prone to L1 insertions. A significantly larger number of Ta-1 insertions were found on chromosome 4 than on other autosomes, not only because chromosome 4 is relatively gene-poor, but also because it contains several detectable insertional hotspots (Boissinot et al [19]). Eleven of the 14 disease-causing insertion sites mentioned above are on the X chromosome (Ostertag and Kazazian [16]). Although X-linked deleterious mutations are in general more likely to be apparent because of male hemizygosity, this bias is not sufficiently strong to account for the high frequency of disease-causing L1 insertions on the X. Therefore, it seems that the X chromosome is unusually prone to novel L1 insertions, although an analysis of the Ta-1 family did not reveal an excess of recent L1 insertions on the X (Boissinot et al [19]). Whatever the cause of the insertion bias, it is possible that insertion bias is, at least in part, responsible for the abundance of L1 elements on chromosomes X and 4.

NEGATIVE SELECTION ELIMINATES DELETERIOUS L1 ELEMENTS

In general, L1 insertions (like most genetic changes) are much more likely to be deleterious or neutral than favorable. An L1-containing allele is considered deleterious when it decreases the fitness of the individual that carries it either by reducing its survival or its reproductive success. As selection against deleterious allele will act as soon as the novel L1-containing allele is produced by retrotransposition, it is unlikely that such deleterious alleles reach high frequencies in populations. In most cases, they will be lost rapidly from populations and will never (or rarely) be observed.

L1 elements have the potential to disrupt the function of host genes in many ways. First, a novel L1 insertion in the coding sequence of a gene would most likely inactivate the protein-coding function of the gene, as exemplified by insertions in exons of the factor VIII gene and in the dystrophin gene (Kazazian et al [21], Narita et al [22]). L1 elements inserted in intronic sequences can also have a deleterious effect by introducing splice sites (Schwahn et al [23], Meischl et al [24]) and polyadenylation signals (Perepelitsa-Belancio and Deininger [25]), or by negatively affecting gene transcription (Han et al [11]). If inserted upstream of genes, L1 elements can also affect their regulation by disrupting regulatory sequences or by inserting their own regulatory sequence such as their sense or antisense internal promoters. Thus, L1 elements are significantly more abundant downstream of genes than upstream (Graham and Boissinot, unpublished observation). Finally, it has recently been demonstrated in a cell-culture assay that L1 retrotransposition can cause large (>3 Kb) genomic deletions (Gilbert et al [26],

Symer et al [27]). Such events would certainly be extremely deleterious if it occurred in a gene-rich region, but genomic deletions caused by L1 retrotransposition are, in general, small (< 500 bp) and relatively rare (Myers et al [28]) as they account for the total loss of only 18 Kb since the human-chimpanzee split (Han et al [29]). All the possible effects L1 elements can have on gene function would likely cause their selective loss from gene-rich regions.

The abundance of L1 sequences across the genome gives them the potential to be efficient mediator of ectopic (ie, nonallelic) recombination. Such events lead to chromosomal rearrangements that are, in general, very deleterious (Burwinkel and Kilimann [30], Segal et al [31]), although some have played an important role in genomic evolution (Fitch et al [32]). If we assume that the frequency of ectopic exchange correlates with the recombination rate, then we expect L1 elements to be more deleterious when they reside in highly recombining regions and therefore eliminated by negative selection. Because longer L1 elements are more likely to mediate ectopic recombination, this model of selection predicts a negative correlation between the length of L1 elements and the recombination rate of the genomic region where they reside. Indeed, long elements accumulate in low- and non-recombining regions of the genome (Boissinot et al [9]; Song and Boissinot, unpublished data) and are lacking from recombination hotspots (Myers et al [33]). Thus, the negative effect of ectopic recombination may cause the selective loss of L1 elements from highly-recombining regions and therefore their accumulation in low recombining regions, which are typically AT-rich and gene-poor.

POSITIVE SELECTION IN FAVOR OF L1 ELEMENTS

Since L1 elements have been described, scientists have wondered which benefit for its host L1 could have. So far, there is absolutely no evidence that L1 could have any useful function for its host. However, recent evidence suggests that in a few cases, L1 sequences may have been coopted by the host for its own benefit. Note that the occasional recruitment of L1 sequences does not imply a function for L1. In some rare cases, ready-to-use motifs contained within the L1 sequence seem to have been retained by the host (Makalowski [34], Kazazian [35]). For instance, the 5′ UTR of modern L1 elements contain sense and antisense promoters which have occasionally been recruited as regulators of the transcription of host genes (Yang et al [36], Speek [37], Nigumann et al [38]), and fragments of L1 sequences have been incorporated within protein-coding sequences (Nekrutenko and Li [39]). However, the number of described cases of cooptation is very small and this mechanism has no significant effect on the overall distribution of L1. In addition, one should always keep in mind that the retention of an L1 element affecting the expression or sequence of a gene does not imply that this element was positively selected (ie, improved the fitness of the host); it might as well have been neutral.

Although positive selection in favor of L1 inserts is unlikely to have affected the overall genomic distribution of L1 (ie, the bias toward AT-rich regions), it is possible that the recruitment of L1 sequences in some regions could result in a local enrichment of L1. It has been proposed that L1 may affect the expression pattern of entire genomic regions or chromosomes and that this effect could be sufficiently strong to positively affect the abundance of L1 in these regions. The idea is that L1 elements would act as "boosters" that promote the expansion of heterochromatin and consequently repress the transcription of genes. This hypothesis has been proposed to explain the spread of X-inactivation along the entire X chromosome (ie, the Lyon hypothesis) (Lyon [40]). Evidence for this role includes the strong enrichment for L1 elements near the X-inactivation center on the X chromosome (Bailey et al [8]) and the observation of X: autosome translocations, showing that the failure of the X-inactivation signal to spread is often correlated with the abundance of L1 elements. In addition, genes that escape X-inactivation are located in regions with a lower abundance of L1 (Bailey et al [8]). The Lyon hypothesis would explain the abundance of L1 elements on the X chromosomes in several mammalian species, although there are important variations in the abundance of L1 elements near the X-inactivation center (Chureau et al [41]) suggesting that the evolution of X-inactivation predates the recruitment of L1 elements as boosters. The hypothesis that L1 elements can promote the inactivation of one copy of a gene is also supported by the evidence that monoallelically expressed genes are located in regions of the genome that are enriched in L1 elements (Allen et al [12]). Another way L1 elements can affect the expression of genes comes from the ability of L1 elements to reduce the amount of transcript produced when they are inserted in an intron (Han et al [11]). This observation led to the suggestion that intronic L1 elements contribute to the fine tuning of gene expression (the Rheostat hypothesis) and may account for some of the differences in L1 abundance among genes (Han et al [11], Han and Boeke [42]). A negative correlation between the expression of genes and the abundance of L1 in their introns has recently been reported (Han et al [11]). Because the same observation could equally indicate that low-expressed genes are just more permissive to the presence of L1 in their introns than highly expressed genes, more data are needed to validate the rheostat hypothesis.

L1 elements may also be retained in the genome because they can reduce linkage between genes and therefore increase the efficiency of selection. In a region of low recombination, many weakly selected mutations can interfere with each other, therefore limiting the effect of selection due to tight linkage between loci. The insertion of L1 elements can mitigate this interference by simply increasing the distance between loci (Comeron [43]). Though this idea has not been tested so far, it has been proposed as a general mechanism to explain the length of introns and the amount of noncoding DNA in genomes (Comeron [43]). A prediction of this model is that longer introns and a higher proportion of noncoding DNA (including L1) will be favored in regions of low recombination.

CONCLUSION

L1 distribution is affected by a number of factors that act at the time of insertion or after the element is inserted. The main difficulty in determining the relative importance of insertion bias and selection is twofold. First, different mechanisms (ie, insertion bias and the different types of selection) can have the same effect on L1 distribution, and the same observation can be explained by radically different mechanisms. For instance, the abundance of L1 elements on the X chromosome can be explained by a bias of insertion, a reduced efficiency of negative selection, or the recruitment of L1 elements as mediator of X-inactivation. Second, genomic parameters such as GC content, gene richness, and recombination rate are not independent, and using correlations between any of these parameters and the abundance of L1 is unlikely to provide a clear explanation for the distribution bias of L1. Indeed, many of the mechanisms discussed in this review were inferred from the analysis of L1 distribution, that is, from the data they were trying to explain, and have not been tested rigorously. To fully understand the genomic distribution of L1 elements, a better knowledge of the molecular mechanism of insertion and the dynamics of L1 elements in natural populations will be necessary.

ACKNOWLEDGMENT

We thank Laurence Frabotta and two anonymous reviewers for their helpful comments on the manuscript.

REFERENCES

[1] Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature.* 2001;409(6822):860–921.

[2] Waterston RH, Lindblad-Toh K, Birney E, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature.* 2002;420(69 15):520–562.

[3] Esnault C, Maestre J, Heidmann T. Human LINE retrotransposons generate processed pseudogenes. *Nature Genetics.* 2000;24(4):363–367.

[4] Dewannieux M, Heidmann T. L1-mediated retrotransposition of murine B1 and B2 SINEs recapitulated in cultured cells. *Journal of Molecular Biology.* 2005;349(2):241–247.

[5] Dewannieux M, Esnault C, Heidmann T. LINE-mediated retrotransposition of marked Alu sequences. *Nature Genetics.* 2003;35(1):41–48.

[6] Korenberg JR, Rykowski MC. Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. *Cell.* 1988;53(3):391–400.

[7] Boyle AL, Ballard SG, Ward DC. Differential distribution of long and short interspersed element sequences in the mouse genome: chromosome karyotyping by fluorescence *in situ* hybridization. *Proceedings of the National Academy of Sciences of the United States of America.* 1990;87(19):7757–7761.

[8] Bailey JA, Carrel L, Chakravarti A, Eichler EE. Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. *Proceedings of the National Academy of Sciences of the United States of America.* 2000;97(12):6634–6639.

[9] Boissinot S, Entezam A, Furano AV. Selection against deleterious LINE-1-containing loci in the human lineage. *Molecular Biology and Evolution.* 2001;18(6):926–935.

[10] Parish DA, Vise P, Wichman HA, Bull JJ, Baker RJ. Distribution of LINEs and other repetitive elements in the karyotype of the bat *Carollia:* implications for X-chromosome inactivation. *Cytogenetic and Genome Research.* 2002;96(1–4):191–197.

[11] Han JS, Szak ST, Boeke JD. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature.* 2004;429(6989):268–274.

[12] Allen E, Horvath S, Tong F, et al. High concentrations of long interspersed nuclear element sequence distinguish monoallelically expressed genes. *Proceedings of the National Academy of Sciences of the United States of America.* 2003;100(17):9940–9945.

[13] Medstrand P, van de Lagemaat LN, Mager DL. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Research.* 2002;12(10):1483–1495.

[14] Jurka J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proceedings of the National Academy of Sciences of the United States of America.* 1997;94(5):1872–1877.

[15] Cost GJ, Boeke JD. Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. *Biochemistry.* 1998;37(51):18081–18093.

[16] Ostertag EM, Kazazian HH Jr. Biology of mammalian L1 retrotransposons. *Annual Review of Genetics.* 2001;35:501–538.

[17] Brooks MB, Gu W, Barnas JL, Ray J, Ray K. A Line 1 insertion in the Factor IX gene segregates with mild hemophilia B in dogs. *Mammalian Genome.* 2003;14(11):788–795.

[18] DeBerardinis RJ, Kazazian HH Jr. Full-length L1 elements have arisen recently in the same 1-kb region of the gorilla and human genomes. *Journal of Molecular Evolution.* 1998;47(3):292–301.

[19] Boissinot S, Entezam A, Young L, Munson PJ, Furano AV. The insertional history of an active family of L1 retrotransposons in humans. *Genome Research.* 2004;14(7):1221–1231.

[20] Muotri AR, Chu VT, Marchetto MCN, Deng W, Moran JV, Gage FH. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature.* 2005;435(7044):903–910.

[21] Kazazian HH Jr, Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE. Haemophilia A resulting from *de novo* insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature.* 1988;332(6160):164–166.

[22] Narita N, Nishio H, Kitoh Y, et al. Insertion of a 5′ truncated L1 element into the 3′ end of exon 44 of the dystrophin gene resulted in skipping of the exon during splicing in a case of Duchenne muscular dystrophy. *The Journal of Clinical Investigation.* 1993;91(5):1862–1867.

[23] Schwahn U, Lenzner S, Dong J, et al. Positional cloning of the gene for X-linked retinitis pigmentosa 2. *Nature Genetics.* 1998;19(4):327–332.

[24] Meischl C, de Boer M, Åhlin A, Roos D. A new exon created by intronic insertion of a rearranged LINE-1 element as the cause of chronic granulomatous disease. *European Journal of Human Genetics.* 2000;8(9):697–703.

[25] Perepelitsa-Belancio V, Deininger P. RNA truncation by premature polyadenylation attenuates human mobile element activity. *Nature Genetics.* 2003;35(4):363–366.

[26] Gilbert N, Lutz-Prigge S, Moran JV. Genomic deletions created upon LINE-1 retrotransposition. *Cell.* 2002;110(3):315–325.
[27] Symer DE, Connelly C, Szak ST, et al. Human L1 retrotransposition is associated with genetic instability in vivo. *Cell.* 2002;110(3):327–338.
[28] Myers JS, Vincent BJ, Udall H, et al. A comprehensive analysis of recently integrated human Ta L1 elements. *The American Journal of Human Genetics.* 2002;71(2):312–326.
[29] Han K, Sen SK, Wang J, et al. Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. *Nucleic Acids Research.* 2005;33(13):4040–4052.
[30] Burwinkel B, Kilimann MW. Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease. *Journal of Molecular Biology.* 1998;277(3):513–517.
[31] Segal Y, Peissel B, Renieri A, et al. LINE-1 elements at the sites of molecular rearrangements in Alport syndrome-diffuse leiomyomatosis. *The American Journal of Human Genetics.* 1999; 64(1):62–69.
[32] Fitch DH, Bailey WJ, Tagle DA, Goodman M, Sieu L, Slightom JL. Duplication of the gamma-globin gene mediated by L1 long interspersed repetitive elements in an early ancestor of simian primates. *Proceedings of the National Academy of Sciences of the United States of America.* 1991;88(16):7396–7400.
[33] Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. A fine-scale map of recombination rates and hotspots across the human genome. *Science.* 2005;310(5746):321–324.
[34] Makałowski W. Genomic scrap yard: how genomes utilize all that junk. *Gene.* 2000;259(1-2):61–67.
[35] Kazazian HH Jr. Mobile elements: drivers of genome evolution. *Science.* 2004;303(5664):1626–1632.
[36] Yang Z, Boffelli D, Boonmark N, Schwartz K, Lawn R. Apolipoprotein(a) gene enhancer resides within a LINE element. *The Journal of Biological Chemistry.* 1998;273(2):891–897.
[37] Speek M. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Molecular and Cellular Biology.* 2001;21(6):1973–1985.
[38] Nigumann P, Redik K, Mätlik K, Speek M. Many human genes are transcribed from the antisense promoter of L1 retrotransposon. *Genomics.* 2002;79(5):628–634.
[39] Nekrutenko A, Li W-H. Transposable elements are found in a large number of human protein-coding genes. *Trends in Genetics.* 2001;17(11):619–621.
[40] Lyon MF. X-chromosome inactivation: a repeat hypothesis. *Cytogenetics and Cell Genetics.* 1998;80(1–4):133–137.
[41] Chureau C, Prissette M, Bourdet A, et al. Comparative sequence analysis of the X-inactivation center region in mouse, human, and bovine. *Genome Research.* 2002;12(6):894–908.
[42] Han JS, Boeke JD. LINE-1 retrotransposons: modulators of quantity and quality of mammalian gene expression? *BioEssays.* 2005;27(8):775–784.
[43] Comeron JM. What controls the length of noncoding DNA? *Current Opinion in Genetics & Development.* 2001;11(6):652–659.