Published in final edited form as:
Gene. 2007 April 1; 390(1-2): 190–198.

Characterization of pre-insertion loci of *de novo* L1 insertions

Stephen L. Gasior[1], **Graeme Preston**[1], **Dale J. Hedges**[1], **Nicolas Gilbert**[2], **John V. Moran**[3], and **Prescott L. Deininger**[1,*]

1 *Tulane Cancer Center and Dept. of Epidemiology, Tulane University Health Sciences Center SL-66, 1430 Tulane Ave., New Orleans, LA 70112, Phone: (504) 988-6385, Fax: (504) 988-5516, pdeinin@tulane.edu*

2 *Institut de Génétique Humaine, CNRS, UPR 1142, 141 rue de la Cardonille, 34396 Montpellier cedex 5, France*

3 *Departments of Human Genetics and Internal Medicine, 1241 E. Catherine St., University of Michigan Medical School, Ann Arbor, Michigan 48109-0618*

Abstract

The human Long Interspersed Element-1 (LINE-1) and the Short Interspersed Element (SINE) Alu comprise 28% of the human genome. They share the same L1-encoded endonuclease for insertion, which recognizes an A+T-rich sequence. Under a simple model of insertion distribution, this nucleotide preference would lead to the prediction that the populations of both elements would be biased towards A+T-rich regions. Genomic L1 elements do show an A+T-rich bias. In contrast, Alu is biased towards G+C-rich regions when compared to the genome average. Several analyses have demonstrated that relatively recent insertions of both elements show less G+C content bias relative to older elements. We have analyzed the repetitive element and G+C composition of more than 100 pre-insertion loci derived from *de novo* L1 insertions in cultured human cancer cells, which should represent an evolutionarily unbiased set of insertions. An A+T-rich bias is observed in the 50 bp flanking the endonuclease target site, consistent with the known target site for the L1 endonuclease. The L1, Alu, and G+C content of 20 kb of the *de novo* pre-insertion loci show a different set of biases than those observed for fixed L1s in the human genome. In contrast to the insertion sites of genomic L1s, the *de novo* L1 pre-insertion loci are relatively L1-poor, Alu-rich and G+C-neutral. Finally, a statistically significant cluster of *de novo* L1 insertions was localized in the vicinity of the c-*myc* gene. These results suggest that the initial insertion preference of L1, while A+T-rich in the initial vicinity of the break site, can be influenced by the broader content of the flanking genomic region and have implications for understanding the dynamics of L1 and Alu distributions in the human genome.

Keywords

LINE; Retrotransposition; Alu; LINE; SINE

1. Introduction

Transposable elements (TEs) constitute substantial portions of all sequenced mammalian organisms (Chimpanzee Sequencing and Analysis Consortium. 2005; International Human

*Address for Correspondence: Tulane Cancer Center, SL66, Tulane University Health Sciences Center, 1430 Tulane Ave., New Orleans, LA 70112, 504-988-6385, pdeinin@tulane.edu.

Genome Sequence Consortium 2001; Lindblad-Toh et al 2005; Mouse Genome Sequencing Consortium 2002; Rat Genome Sequencing Project Consortium 2004). Within the human genome, the TE content is primarily composed of retrotransposons (Deininger et al 2003; International Human Genome Sequence Consortium 2001). Retrotransposons mobilize via an RNA intermediate which is reverse-transcribed and integrated into the genome (reviewed in Kazazian 2004). They can be divided into two general categories; autonomous retrotransposons encode the protein components necessary to initiate insertion into the genome, while non-autonomous elements parasitize the protein machinery of autonomous retrotransposons (Dewannieux et al 2003; Jurka 1997; Kajikawa and Okada 2002; Wei et al 2001). Nucleotide preferences exhibited by the endonucleases employed by retrotransposons can play a role in their genomic distribution. The human genome possesses an average G+C content of 41% but shows deviations when averaged over large 20–100 kb regions, called isochores (International Human Genome Sequence Consortium 2001; Macaya et al 1976; Soriano et al 1983). This uneven distribution of G+C content could be expected to bias the genomic distribution of non site-specific TEs.

The Long Interspersed Element-1 (LINE-1 or L1) is a ~6 kb autonomous non -LTR retrotransposon that comprises about 17% of the human genome (International Human Genome Sequence Consortium 2001; Moran and Gilbert 2002; Ostertag and Kazazian 2001). Alu is a ~300 bp non-autonomous, non-LTR retroposon that represents another 11% of the human genome (Deininger and Batzer 2002; International Human Genome Sequence Consortium 2001). L1 and Alu are both non-site-specific and demonstrate a strong bias for insertion into a short, A+T-rich L1 endonuclease target site, as has been demonstrated for genomic L1 and Alu insertions (Boissinot et al 2000; Boissinot et al 2004; Feng et al 1996; Jurka 1997; Jurka and Klonowski 1996; Morrish et al 2002; Myers et al 2002; Ovchinnikov et al 2001; Salem et al 2003; Szak et al 2002), for L1 and Alu insertions from molecular assays (Dewannieux et al 2003; Feng et al 1996; Gilbert et al 2002; Gilbert et al 2005; Morrish et al 2002; Symer et al 2002), and for the L1 endonuclease *in vitro* (Cost and Boeke 1998; Feng et al 1996). Even though they both utilize the same endonuclease, L1 and Alu are distributed differently in the human genome. L1 preferentially inhabits G+C-poor isochores, and Alu preferentially inhabits G+C-rich isochores (International Human Genome Sequence Consortium 2001; Soriano et al 1983).

Models have suggested various ways to explain how the differential distribution of Alu and L1 elements. These include positive selection for Alu in genic regions (Britten 1997; Kidwell and Lisch 1997), biased Alu loss in gene-poor regions via recombination (Batzer and Deininger 2002), or that G+C-rich or -poor regions strive to maintain their local G+C content in a process called compositional matching (Filipski et al 1989; Pavlicek et al 2001). It has also been noted that gene density positively correlates with G+C-rich regions of the genome (International Human Genome Sequence Consortium 2001). Thus, L1 could be excluded from genic (G+C-rich) regions relative to Alu due to negative selection from stronger gene disruptive effects (Boissinot et al 2001; Boissinot et al 2006; Han and Boeke 2005; Perepelitsa-Belancio and Deininger 2003; Roy-Engel et al 2005; Wheelan et al 2005).

To better understand the potential role of insertion-site preferences vs. post-insertion alterations of L1 insertions, we have characterized the pre-insertion loci of over one hundred *de novo* L1 insertions in HeLa cells generated from a tissue culture assay (Gilbert et al 2002; Gilbert et al 2005; Moran et al 1996; Symer et al 2002). HeLa cells (as compared to *in vivo*) provide a useful cell line for understanding insertions because their triploid content (Macville et al 1999) should generally prevent haplo-insufficiency from affecting growth rates. In addition, we have generated a computer simulation of random insertions specific to HeLa cells and used this to evaluate deviations from expected G+C content and repetitive element content. We have also

compiled a much smaller dataset of *de novo* SINE insertions in HeLa cells from the literature and characterized their pre-insertion loci.

2. Materials and Methods

2.1 Sources of pre-insertion loci for L1 insertions

Techniques for L1 insertions and their recovery have been previously described (Gilbert et al 2002; Gilbert et al 2005). The first data set consists of the published inserts generated from the L1.3, L1.2, LRE2, or $L1_{RP}$ based vectors in HeLa cells, comprehensively listed in (Gilbert et al 2002; Gilbert et al 2005; Moran et al 1996). The second source of data derives from L1 assays employing a derived vector that eliminated the 5' UTR of L1.3 (L1_CMV_rec) which would limit the ability of inserted L1s to remobilize (El Sawy et al 2005). During the analysis of those clones, more robust L1s were synthetically created. The third source was generated by swapping into L1_CMV_rec synthetically generated $L1_{RP}$ ORF2 sequence with several synonymous sequence changes to eliminate canonical polyA sites (synL1_neo) (El Sawy et al 2005). Full details of the previously unpublished insertions and differences in generation are presented in Supplemental Table 1, consistent with the format used previously (Gilbert et al 2005). L1_CMV_rec and syn_L1 transfected HeLa cells were grown under G418 selection for 2 weeks until colonies were visible, collected in batch and allowed to grow prior to genomic DNA isolation using a DNeasy kit (Qiagen). Genomic DNA was digested with *Apa*I (L1_CMV_rec) or *Hin*d III (syn_L1), ligated with T4 DNA ligase (New England Biolabs) at low concentration to circularize fragments, and then concentrated using microconcentrators (Microcon) prior to electroporation into competent *E. coli* (EP-Max 10B, Bio-Rad).

2.2 Analysis of L1 insertion clones

PCR primer pairs were used to determine the approximate insert length of recovered L1_CMV_rec clones with a primer in the neoR sequences which would be available in every insert (F1: GAATTCTACAAACTACCATCAGAGAATAC, F2: GAATTCTTCTTATACACCAACAACAG F3: GAATTCAGGACATGAACAGACACTTC, neo_3'_F: CCTTCTTGACGAGTTCTTC, underlined GAATTC sequences are non L1). For synL1_neo clones, restriction enzyme digests of DNA were used to determine the approximate insert length. *Xho*I-*Eco*RV-*Apa*I-*Eco*RI and *Eco*RI-*Bsr*GI-*Bgl*II multi digests yielded diagnostic fragments. In both cases DNA was then sequenced using several optimally placed primers based on the length of the insert (neo_mid_F: ATGATCTGGACGAAGAGCATCAGG, neo3'_F, L1_int_R: TTGGGAGAGTGTATGTGTCGAGGA, orlon_3788_R: TTTCTGAGGGCTCTGTTCTGTTCC, L1_int2_R: TGTAGTTGAGCGGCTTTGAGTGAG, orlon_3344_R: GAGTTCACCCATGATTTGGC, orlon_3788_R: TTTCTGAGGGCTCTGTTCTGTTCC, orlon_4349_R: CATGTGTTTTTTGGCTGCAT, orlon_1062_R: CTGGTGATTTTGCTCATTAG. Once a sequence read traversed the 5' L1-genomic junction, a BLAST query to the human genome was used to identify the genomic target site. Using this sequence, primers were designed to sequence the 3'L1-genomic junction for each clone. Primers were generally designed to be approximately400–600 bp 3' of the identified genomic sequence at the 5' L1 junction and sequence towards the neoR cassette of the L1 insertion. Pre-insertion loci were defined as the 20 kb (10 kb to each side) of genomic sequence flanking the utilized endonuclease incision site taken from the human genome in GenBank (accession numbers provided previously or in Supplemental Table 1).

2.3 Computation of random genomic sites

Simulation of random insertion of L1 sequences into a genome with a HeLa karyotype was conducted using local Perl scripts, which are available from the authors upon request. For the

purpose of the simulation, the sequenced nucleotides of the human genome (version hg17) were mapped to a corresponding set of unique consecutive integers. Using karyotypic data provided in (Macville et al 1999), the mapping process accounted for over and under-represented chromosomal regions of the HeLa karyotype by increasing or decreasing the amount of integer space allocated to the corresponding human regions. Insertion locations were chosen by randomly selecting an integer from the total mapped set using a uniform distribution. The sequence flanking the chosen location (10 kb for both flanks) was subsequently extracted from the human genome and analyzed for repeat content with a local installation of RepeatMasker (default settings).

For the purpose of evaluating the probability of insertions to cluster together within chromosomal regions of specified size, a separate Perl script was created which tracked insertion locations for 10000 replicates of N=104 insertions. The sampling process and sequence mapping for the Hela genome was identical to that described above. For each replicate, the number of inserts clustering within 1 Mb regions was recorded.

3. Results

3.1 Genomic characteristics of L1 pre- insertion loci

A tissue culture assay for L1 retrotransposition was used to obtain *de novo* L1 insertions in HeLa cells (Gilbert et al 2002; Gilbert et al 2005). This assay allows for the cloning of individual L1 insertions because the retrotransposition indicator cassette is equipped with a neo/kan^R gene under the control of both a viral and bacterial promoter and a plasmid origin of replication. Characterization of insertion clones with subsequent comparisons to Genbank allows for the reconstruction of the pre-insertion loci. In addition to the previously published L1 insertions, two additional unpublished L1 insertion sets were characterized from more recently developed vectors, L1_CMV_rec and synL1_neo (Materials and Methods).

RepeatMasker was used to annotate the repeat content of 108 *de novo*, endonuclease-dependent L1 pre-insertion loci including 10 kb to either side of the endonuclease cut site. To determine whether L1 insertion structures or different vectors demonstrated independent biases, the 108 pre-insertion loci were subdivided by L1 insertion structure and by the lab in which they were generated (Supplemental Table 2). The majority of events were simple 5' truncations (5' trunc) without any rearrangement of L1 sequence (n=80). These 80 *de novo* L1 pre-insertion loci are enriched for Alu (13%), MIR (3.2%), and L2 (3.7%) sequences, depleted for L1 (13%) sequences, and are neutral for MaLR content when compared to the overall genome. MaLR is a moderately interspersed element that uses a different insertion mechanism than L1 and demonstrates relatively weak G+C-poor and gene-poor biases (Medstrand et al 2002). Thus, MaLR should represent a suitable control for randomness assuming little or no genic bias (Medstrand et al 2002).

There was little difference between repetitive element content for the datasets generated with different L1 vectors. However, minor deviations relative to the 5' truncated set include sites were observed for full-length (FL) insertions (n=6) with Alu and L1 contents of 12.8% and 19.1%, respectively. A similar situation was observed for L1 insertions with internal inversions (with associated L1 duplication or internal L1 deletion, n=17) with an L1 content of 18.4%. Both classes of insertions approximate the genomic content of L1s (17%). Finally, sites associated with L1-L1 chimeras (n=5) showed dramatically low Alu content (2.8%) and higher than average L1 content (23.0%). L1-L1 chimeras are rarely detected in genomic DNA, either because they genuinely are infrequent or because systematic biases in assembly algorithms exclude their incorporation into assembled genome sequences. Therefore, we did not include L1-L1 chimeras in our overall analysis. We note that all three subclasses require a larger sample size to determine whether they truly deviate in their insertion characteristics. The repetitive

element content for all of the clearly endonuclease-dependent L1 pre-insertion loci (n=103) is summarized in Table 1 and this was used for subsequent comparative analyses.

To statistically determine whether *de novo* L1 pre-insertion loci deviate from genomic repetitive element content, we computationally generated a random site dataset (n=1042) with compensation for the HeLa karyotype (Macville et al 1999) and then used RepeatMasker to determine the repetitive element content of the 20 kb flanking regions. The random-site flanking regions demonstrated the expected genome repetitive element content and were then compared to the*de novo* L1 pre-insertion loci (Table 1). Deviation of Alu content between *de novo* L1 pre-insertion loci and random sites was significant (p=0.0039), whereas no statistical differences (p >0.05) were seen for the other assayed repetitive elements. In comparing the distributions of Alu content of *de novo* L1 pre-insertion loci with the random sites (Figure 1A), *de novo* L1 pre-insertion loci are particularly depleted flanking regions with low Alu content (<10% Alu). However, the L1 distributions in each set are largely similar (Figure 1B). A technical consideration that may influence these results is that several 5' truncations from the previously unpublished L1 insertions using L1_CMV_rec and synL1_neo had high repetitive element content (generally Alu) in the immediate predicted 3' genomic sequence. These flanking regions were not conducive to effective primer design and were thus excluded from this analysis. This bias would likely underestimate the Alu content and thus not adversely influence our interpretation. However, it does point out that our L1 insert population is derived from a insertion events that are amenable to complete analysis.

The G+C content of *de novo* L1 insertion sites was also characterized (Table 1). In general the full 20 kb was equivalent to the genome average of 41%. A small window (50 bp) centered around the endonuclease cut site was demonstrably G+C-poor (32% compared to 41%). The endonuclease 1st strand preferred cut sequence is 5'-TTTT^AA-3', although base substitutions are allowable and further flanking sequences may also have an influence (Cost and Boeke 1998;Feng et al 1996;Gilbert et al 2002;Gilbert et al 2005;Morrish et al 2002;Symer et al 2002). The second-strand target site shows a bias towards a consensus that includes an approximately 5-base long, A+T-rich sequence (Gentles et al 2005;Szak, et al 2002). Setting the 50 bp window to include 11 A and T bases would provide an expected G+C baseline of 32%. This value would then be very similar to that observed for *de novo* L1 pre-insertion loci (Table 1). To determine the statistical significance of these differences, the G+C content of the random sites was also characterized and compared to the *de novo* L1 pre-insertion loci (Table 1). The *de novo* L1 insertion 20 kb G+C content was not significantly different from the simulated insertions and showed a similar distribution (Table 1, Figure 1C) whereas the 50 bp region was significantly different as well as distributed differently (Table 1, Figure 1D). The overall results contrast with insertion sites of genomic L1s (G+C neutral versus G+C-poor, Alu-rich versus Alu-poor, respectively) suggesting that *de novo* L1s in a transformed cell line are distributed in a way significantly different than observed for the total population of genomic L1 elements.

Given the uneven distribution of G+C content of the human genome, the observation that the 50 bp flanking sequence of *de novo* L1 insertion sites was G+C-poor but the 20 kb flanking sequence was essentially G+C neutral was surprising. To determine whether 50 bp regions with this degree of A+T richness would be preferentially located in A+T-rich isochores, a subset of the random sites was generated in which the 50 bp region G+C content was restricted to 26–37% (mean= 32.6%, n=391). The average G+C content of the 20 kb regions decreased to 39% (Table 1). The average Alu and L1 contents decreased and increased, respectively (Table 1). The G+C, Alu, and L1 content of this restricted dataset was further differentiated from the G+C, Alu, and L1 content of the *de novo* L1 pre-insertion loci to enhance or create the statistically significant differences ($<10^{-8}$, $<10^{-6}$, and p=0.00051). These results provide data to support the argument that, while L1 prefers to utilize A+T rich endonuclease target

sites, the larger flanking content is biased towards G+C neutrality and higher Alu content. This is a significant deviation from expectations based on the total population of genomic L1s and increased likelihood of finding small regions of A+T-rich sequences within A+T-rich isochores.

3.2 Genomic characteristics of SINE pre- insertion loci

Pre-insertion loci derived from published *de novo* SINE insertions in HeLa cells (Dewannieux et al 2003; Dewannieux and Heidmann 2005) were also analyzed (Table 1). The SINE dataset was generated using a similar retrotransposition assay as used for L1 with the exception that genomic loci were cloned using inverse PCR. The SINEs characterized included five Alu, four mouse B1, and four mouse B2. Even though the B1 and B2 elements are mouse SINEs, they still demonstrate dependency on the cotransfected L1 expression vector for high levels of retrotransposition in HeLa cells (Dewannieux and Heidmann 2005). SINE pre-insertion loci showed a markedly stronger bias to be Alu-rich relative to the random insertion set (27% Alu relative to 11% for genomic, $p < 10^{-8}$) and no significant difference from the random insertions' L1 content (p= 0.84). MIR, L2, and MaLR sequences were also lower than observed in the genome (but was only marginally significant for MaLR (p=0.16, 0.11, 0.042 respectively). Consistent with *de novo* L1 insertions, the *de novo* SINE insertions showed no difference for G+C content in the 50 bp region compared to the A+T-rich subset of the random insertions (p=0.079). The small number of analyzed SINE insertions makes it difficult to make strong conclusions, and more data will be needed to strengthen these observations.

3.3 Chromosomal distribution of L1 and SINE pre- insertion loci

We also fine-mapped 104 L1 insertions and the 13 SINE insertions to chromosomes (Figure 2A). Insertion #53 was found in GenBank to be located to the X chromosome, but a full 20 kb contig could not be identified which is why it was excluded from the previous analyses. The proportion of L1 insertions on each chromosome generally correlated with the proportion of each chromosome to the total HeLa genome (Figure 2B) (modified HeLa chromosome content derived from (Narezkina et al 2004)). There were three notable exceptions: only 3 insertions were observed into the X chromosome, 12 were found in chromosome 12, and only 1 was found in chromosome 9. The chromosomal targeting of the *de novo* L1 insertions was also compared to the chromosomal locations of the computationally derived random insertion sites (Figure 2C). A similar degree of correlation was seen as with the chromosome proportion analysis. Finally, we compared the L1 insertion distribution to Avian sarcoma virus (ASV), an LTR retrovirus (Hindmarsh and Leis 1999). ASV demonstrates little or no specific gene or subgenic targeting biases (Mitchell et al 2004; Narezkina et al 2004). The chromosomal distribution of L1 insertions is more highly correlated to previously mapped ASV insertions (Narezkina et al 2004) than the analyses directly linked to chromosome size (Figure 2B, right Y axis).

The majority of the previously published L1 insertions were within 20 kb of known genes (Gilbert et al 2002; Gilbert et al 2005). The new pre-insertion loci introduced in this study are consistent with that observation (data not shown). One notable cluster of insertions occurred in the vicinity of the c-*myc* locus (Figure 3). Four of the described *de novo* insertions were into a 470 kb region with c-*myc* at the center. One inserted into a known breakpoint region 3 kb 3' of the last coding exon and another into the c-*myc* regulatory region. One insertion was into the nearby, oncogenic PVRT locus. The 4th insertion occurred into a pseudogene ~300 kb 5' of c-*myc*. A modification of the random insertion program was used to estimate the probability that a four-insert cluster could occur by chance (Materials and Methods). For a more conservative estimate we scored clusters in a 1 megabase region. Out of 1000 runs of 104 insertions, no cluster of 4 insertions within a 1 megabase region were found.

4. Discussion

4.1 Differences in G+C and TE content between de novo and genomic insertion loci

By examining the pre-insertion loci of *de novo* L1 insertions, we observed significant differences in their characteristics compared to insertion sites of the total population of genomic L1 elements (International Human Genome Sequence Consortium 2001; Medstrand et al 2002). These differences are notable even compared to analyses that specifically target young L1 insertions which demonstrate an A+T bias relative to the genome average (Boissinot et al 2004; Szak, et al 2002). The *de novo* L1 retrotransposition events (and to an even greater extent *de novo* Alu retroposition events) preferentially insert into Alu-enriched sequences and do not demonstrate a G+C bias in a 20 kb window. An A+T-rich bias is still reflected in the sequences immediately adjacent to the insertion site; however, our results suggest that the preference of the L1 endonuclease for A+T-rich sequences does not lead to an insertion bias over large genomic regions. This is a particularly striking deviation from expectation because our random site data distinctly demonstrate that A+T-rich 50 bp regions are preferentially located in A+T-rich isochores. A possible explanation for the *de novo* L1 insertion bias is that our insertions preferentially utilize A+T-rich endonuclease sites that are embedded within Alu-rich regions. The G+C-richness of Alu may then contribute unexpectedly higher G+C content to the flanking region.

These observations suggest that two common assumptions in explaining Alu and L1 dynamics in the human genome may require refinement. One is that both Alu and L1 preferentially insert into A+T-rich regions based on the L1 endonuclease preferred cleavage site. It appears that in this tissue culture assay G+C content (or sequence characteristics that correlate with G+C-rich regions) over a larger region plays a positive role in L1 insertion site selection, which counterbalances the preferential location of A+T rich local regions within A+T-rich isochores. The second is that L1 and Alu insertion preferences are similar because they both utilize the L1 endonuclease. Our data (notably based on a very small SINE data set) suggest that, although they are generally similar for G+C content, SINE elements have a more pronounced insertion bias for Alu-rich regions relative to L1 elements.

Our L1 and Alu insertion site data show a striking correlation to the insertion sites of young SVA elements (Wang et al 2005). The SVA insertions that postdate the human-chimpanzee divergence preferentially insert into Alu-rich, L1-poor, and G+C-rich sequences. Their low copy numbers would likely keep them from being eliminated due to recombination. The insertion preferences of SVA may best represent a genomic correlate of our tissue culture data that is obscured for Alu and L1 by their high copy numbers and long history in the human genome.

One caveat in our comparison of HeLa insertions vs. extant L1 elements in the human germline is that it is possible that there are different insertion preferences in cultured cells than in the germline. However, there is currently no mechanistic data that would support such an interpretation. It also is possible that ascertainment bias may influence the analysis of *de novo* L1 insertions because they enrich for insertions that can effectively express the neoR cassette or resist epigenetic silencing. Indeed, silencing of L1 selection cassettes has been observed *in vitro* (Muotri et al 2005; Ostertag et al 2000). This caveat would be interesting to revisit using L1 insertions in human cells without the use of a selectable marker. However, the observation that L1 insertions are similar to ASV insertions, which did not employ a selection cassette, suggests that silencing of L1 cassettes does not strongly affect our analyses.

4.2 Model for L1 and Alu distributions in the human genome

We propose a refined model to account for the observed genomic biases in L1 distribution. First, the insertion site preference is essentially neutral to G+C content (and thus A+T content) over a large region. However, a slight preference is seen for Alu-rich sequences. Initial negative selection would eliminate L1 from genic regions, thus imparting a weak A+T-rich bias. This selection would probably be stronger than expected solely from selection at the sexually mature organism level, as negative impacts on somatic germline precursors, meiotic cells, gametes, and the developing embryo (pre-zygotic selection) would lead to their under-representation the gene pool (Boissinot et al 2004; Hastings 1991). The observation that many L1 insertions in HeLa cells lead to L1-L1 chimeras with loss of intrachromosomal genomic regions, translocations, and endonuclease-independent target site deletions (Gilbert et al 2002; Gilbert et al 2005; Morrish et al 2002; Symer et al 2002) that are rarely observed in the human germline further supports this argument. We further postulate that a shift of L1 towards A+T-rich regions would occur over evolutionary time because negative selection acts to exclude full-length L1s from genic regions (Boissinot et al 2001; Boissinot et al 2006). Indeed, our data suggest that the forces affecting this shift are stronger than previous models may have taken into account.

Our data on SINE insertions in the Alu-rich regions are consistent with observations for the youngest SINE insertions in both the human (Jurka et al 2004) and mouse (Jurka et al 2005) genomes, although they cannot be compared quantitatively with those data. Thus, it seems likely that the HeLa system is recapitulating the general principles for SINE insertions in the genome. The insertion of SINEs into Alu-rich regions, may also help through proximity to contribute to the ability of Alu elements to cause genomic deletions during the insertion process (Callinan et al 2005) and through non-allelic recombination afterwards (Deininger et al 2003). Recent work has suggested that Alu insertions are predominantly neutral to gene function (Cordaux et al 2006) suggesting that Alus can insert into genes with little negative effect. However, their loss via non-allelic recombination would still be expected to be under negative selection in genic regions, contributing to their retention.

L1 insertions generally target chromosomes in proportion to size. This has also been observed for recent L1 insertions in the genome (Boissinot et al 2004). Our studies do show fewer *de novo* insertions than expected in the X chromosome relative to a random insertion model. The X chromosome may only represent a target a third its potential size if all but one copy is kept inactivated in a Barr-body state (HeLa generally is 3n (Macville et al 1999). However, it has not yet been demonstrated that compact chromosomes would necessarily impede L1 insertion. These data do highlight a difference compared to genomic L1 distributions in which the X and Y chromosome are enriched in L1 sequence (Boissinot et al 2001). It is hypothesized that elimination due to recombination on these chromosomes is reduced during meiosis allowing a biased retention of L1s with a reduced fitness cost (Boissinot et al 2006). These forces do not seem to play a role in HeLa cells under our growth conditions and underscores the utility of this approach for finding initial insertion preferences.

Possible reasons for the observed chromosome 9 depletion are difficult to envision, other than that a large majority of the chromosome is missing specifically from a large fraction of the two tissue culture samples even though different isolates of HeLa generally contain chromosome 9 in cytological evaluations (Macville et al 1999). We know of no other feature of chromosome 9 that would be expected to impact insertions. It also is possible that this underrepresentation simply represents a sampling bias.

The same sampling bias may explain the enrichment of *de novo* L1 insertions into chromosome 12. Although, the enrichment also could be explained by noncytologically observable chromosomal amplifications. Finally, it is also possible that chromosomes 9 and 12 have slight variations of chromatin, Alu clustering, gene density, etc, that affect L1 insertions in

conjunction with target size. Interestingly, chromosome 12 also demonstrated a specific enrichment for ASV insertions. An alternate explanation is that both estimates of HeLa chromosome content are not representative of HeLa cells used in the experiments.

4.3 Clustering of L1 and SINE insertions

The presence of insertion clustering is a potential indicator of specific chromosome regions that are particularly susceptible to the L1 endonuclease. There was generally no clustering of L1 insertions with the one potential exception of four L1 insertions near the c-*myc* locus. Interestingly, there is a previously published L1 insertion in the dog c-*myc* locus (Amariglio et al 1991). There was also a report of a L1-related rearrangement in the human c-*myc* locus (Morse et al 1988) but this event was not fully characterized and had several features suggesting that it was not a typical L1 insertion. The somatic insertion of endogenous L1 in the dog c-*myc* gene (Figure 3)(Amariglio et al 1991) may represent an ascertainment bias because of the influence c-*myc* has on cellular proliferation. However, it seems unlikely that the L1 insertions created in the HeLa assay could present a similar proliferation advantage. Thus, c-*myc* and its surrounding regions may represent a site susceptible to L1 insertion as has been proposed for other integrating elements (Dudley et al 2002). One underlying feature potentially related to insertion enhancement is that c-*myc* is highly expressed or amplified in cancer cells (Dudley et al 2002). Such c-*myc* amplifications have been reported as absent in HeLa cells (Macville et al 1999). Notably, the c-*myc* locus in HeLa cells harbors human papillomavirus virus (HPV) insertions leading to changes in c-myc expression (Couturier et al 1991; Macville et al 1999). The nearby presence of inserted viral DNA that is highly expressed could represent a chromatin context specifically altered by exogenously derived DNA.

Overall, our data suggest the possibility of higher order chromatin structure influencing L1 insertions, but greater numbers of insertion loci need to be characterized. It's nonetheless tempting to speculate that this retroposon insertion susceptibility may persist even after insertion. Thus, the insertion bias we've observed into Alu-containing regions actually reflects a bias into these regions that are "marked" with a prevalence of prior insertion activity. The nature of this insertion susceptibility is profoundly important for understanding genome evolution.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We would like to thank Melanie Palmisano for technical help with DNA preparation and tissue culture. Thanks to Richard Katz (University of Pennsylvania) for the raw ASV data set. S.L.G was supported by postdoctoral fellowship PF-01-077-01-LIB from the American Cancer Society and additionally from a grant from the Brown Foundation through the Tulane Cancer Center. G. P. was supported by the Louisiana Cancer Research Consortium Summer Research Internship program. The P.L.D. lab is supported by grants from the USPHS grant R01GM45668, National Science Foundation EPS-0346411 and the State of Louisiana Board of Regents Support Fund, as well as core resources supported by NIH P20RR020152. S.L.G. would also like to specifically thank the American Society of Human Genetics for post-Katrina funding as well as the Berry College Department of Communications (Rome, GA) and the Doug Bishop Lab (University of Chicago) for computer support during the writing of the manuscript. The databases created by N.G. and J.V.M. were supported by a grant from the NIH (GM60518).

References

Amariglio EN, Hakim I, Brok-Simoni F, Grossman Z, Katzir N, Harmelin A, Ramot B, Rechavi G. Identity of rearranged LINE/c-MYC junction sequences specific for the canine transmissible venereal tumor. Proc Natl Acad Sci U S A 1991;88:8136–8139. [PubMed: 1654559]

Batzer MA, Deininger PL. Alu repeats and human genomic diversity. Nat Rev Genet 2002;3:370–379. [PubMed: 11988762]

Boissinot S, Chevret P, Furano AV. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Mol Biol Evol 2000;17:915–928. [PubMed: 10833198]

Boissinot S, Davis J, Entezam A, Petrov D, Furano AV. Fitness cost of LINE-1 (L1) activity in humans. Proc Natl Acad Sci U S A 2006;103:9590–9594. [PubMed: 16766655]

Boissinot S, Entezam A, Furano AV. Selection against deleterious LINE-1-containing loci in the human lineage. Mol Biol Evol 2001;18:926–935. [PubMed: 11371580]

Boissinot S, Entezam A, Young L, Munson PJ, Furano AV. The insertional history of an active family of L1 retrotransposons in humans. Genome Res 2004;14:1221–1231. [PubMed: 15197167]

Britten RJ. Mobile elements inserted in the distant past have taken on important functions. Gene 1997;205:177–182. [PubMed: 9461392]

Callinan PA, Wang J, Herke SW, Garber RK, Liang P, Batzer MA. Alu retrotransposition-mediated deletion. J Mol Biol 2005;348:791–800. [PubMed: 15843013]

Chimpanzee, Sequencing; Analysis, Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 2005;437:69–87. [PubMed: 16136131]

Cordaux R, Lee J, Dinoso L, Batzer MA. Recently integrated Alu retrotransposons are essentially neutral residents of the human genome. Gene 2006;373:138–144. [PubMed: 16527433]

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;37:18081–18093. [PubMed: 9922177]

Couturier J, Sastre-Garau X, Schneider-Maunoury S, Labib A, Orth G. Integration of papillomavirus DNA near myc genes in genital carcinomas and its consequences for proto-oncogene expression. J Virol 1991;65:4534–4538. [PubMed: 1649348]

Deininger PL, Batzer MA. Mammalian retroelements. Genome Res 2002;12:1455–1465. [PubMed: 12368238]

Deininger PL, Moran JV, Batzer MA, Kazazian HH. Mobile elements and mammalian genome evolution. Curr Opin Genet Dev 2003;13:651–658. [PubMed: 14638329]

Dewannieux M, Esnault C, Heidmann T. LINE-mediated retrotransposition of marked Alu sequences. Nat Genet 2003;35:41–48. [PubMed: 12897783]

Dewannieux M, Heidmann T. L1-mediated retrotransposition of murine B1 and B2 SINEs recapitulated in cultured cells. J Mol Biol 2005;349:241–247. [PubMed: 15890192]

Dudley JP, Mertz JA, Rajan L, Lozano M, Broussard DR. What retroviruses teach us about the involvement of c-Myc in leukemias and lymphomas. Leukemia 2002;16:1086–1098. [PubMed: 12040439]

El Sawy M, Kale SP, Dugan C, Nguyen TQ, Belancio V, Bruch H, Roy-Engel AM, Deininger PL. Nickel stimulates L1 retrotransposition by a post-transcriptional mechanism. J Mol Biol 2005;354:246–257. [PubMed: 16249005]

Feng Q, Moran JV, Kazazian HH Jr, Boeke JD. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 1996;87:905–916. [PubMed: 8945517]

Filipski J, Salinas J, Rodier F. Chromosome localization-dependent compositional bias of point mutations in Alu repetitive sequences. J Mol Biol 1989;206:563–566. [PubMed: 2716062]

Gentles AJ, Kohany O, Jurka J. Evolutionary Diversity and Potential Recombinogenic Role of Integration Targets of Non-LTR Retrotransposons. Mol Biol Evol 2005;22:1983–1991. [PubMed: 15944437]

Gilbert N, Lutz S, Morrish TA, Moran JV. Multiple fates of L1 retrotransposition intermediates in cultured human cells. Mol Cell Biol 2005;25:7780–7795. [PubMed: 16107723]

Gilbert N, Lutz-Prigge S, Moran JV. Genomic deletions created upon LINE-1 retrotransposition. Cell 2002;110:315–325. [PubMed: 12176319]

Han JS, Boeke JD. LINE-1 retrotransposons: modulators of quantity and quality of mammalian gene expression? Bioessays 2005;27:775–784. [PubMed: 16015595]

Hastings IM. Germline selection: population genetic aspects of the sexual/asexual life cycle. Genetics 1991;129:1167–1176. [PubMed: 1783297]

Hindmarsh P, Leis J. Retroviral DNA integration. Microbiol Mol Biol Rev 1999;63:836–43. [PubMed: 10585967]table.

International Human Genome Sequence Consortium. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921. [PubMed: 11237011]

International Human Genome Sequence Constortium. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921. [PubMed: 11237011]

Jurka J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc Natl Acad Sci U S A 1997;94:1872–1877. [PubMed: 9050872]

Jurka J, Klonowski P. Integration of retroposable elements in mammals: selection of target sites. J Mol Evol 1996;43:685–689. [PubMed: 8995066]

Jurka J, Kohany O, Pavlicek A, Kapitonov VV, Jurka MV. Duplication, coclustering, and selection of human Alu retrotransposons. Proc Natl Acad Sci U S A 2004;101:1268–1272. [PubMed: 14736919]

Jurka J, Kohany O, Pavlicek A, Kapitonov VV, Jurka MV. Clustering, duplication and chromosomal distribution of mouse SINE retrotransposons. Cytogenet Genome Res 2005;110:117–123. [PubMed: 16093663]

Kajikawa M, Okada N. LINEs mobilize SINEs in the eel through a shared 3' sequence. Cell 2002;111:433–444. [PubMed: 12419252]

Kazazian HH Jr. Mobile elements: drivers of genome evolution. Science 2004;303:1626–1632. [PubMed: 15016989]

Kidwell MG, Lisch D. Transposable elements as sources of variation in animals and plants. Proc Natl Acad Sci U S A 1997;94:7704–7711. [PubMed: 9223252]

Lindblad-Toh K, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 2005;438:803–819. [PubMed: 16341006]

Macaya G, Thiery JP, Bernardi G. An approach to the organization of eukaryotic genomes at a macromolecular level. J Mol Biol 1976;108:237–254. [PubMed: 826644]

Macville M, Schrock E, Padilla-Nash H, Keck C, Ghadimi BM, Zimonjic D, Popescu N, Ried T. Comprehensive and definitive molecular cytogenetic characterization of HeLa cells by spectral karyotyping. Cancer Res 1999;59:141–150. [PubMed: 9892199]

Medstrand P, van de Lagemaat LN, Mager DL. Retroelement distributions in the human genome: variations associated with age and proximity to genes. Genome Res 2002;12:1483–1495. [PubMed: 12368240]

Mitchell RS, Beitzel BF, Schroder AR, Shinn P, Chen H, Berry CC, Ecker JR, Bushman FD. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. PLoS Biol 2004;2:E234. [PubMed: 15314653]

Moran, JV.; Gilbert, N. Mobile DNA II. Craig, N.; Craggie, R.; Gellert, M.; Lambowitz, A., editors. Am Soc Microbiol; Washington, DC: 2002. p. 836-869.

Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH Jr. High frequency retrotransposition in cultured mammalian cells. Cell 1996;87:917–927. [PubMed: 8945518]

Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD, Taccioli GE, Batzer MA, Moran JV. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat Genet 2002;31:159–165. [PubMed: 12006980]

Morse B, Rotherg PG, South VJ, Spandorfer JM, Astrin SM. Insertional mutagenesis of the myc locus by a LINE-1 sequence in a human breast carcinoma. Nature 1988;333:87–90. [PubMed: 2834650]

Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. Nature 2002;420:520–562. [PubMed: 12466850]

Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, Gage FH. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature 2005;435:903–910. [PubMed: 15959507]

Myers JS, Vincent BJ, Udall H, Watkins WS, Morrish TA, Kilroy GE, Swergold GD, Henke J, Henke L, Moran JV, Jorde LB, Batzer MA. A comprehensive analysis of recently integrated human Ta L1 elements. Am J Hum Genet 2002;71:312–326. [PubMed: 12070800]

Narezkina A, Taganov KD, Litwin S, Stoyanova R, Hayashi J, Seeger C, Skalka AM, Katz RA. Genome-wide analyses of avian sarcoma virus integration sites. J Virol 2004;78:11656–11663. [PubMed: 15479807]

Ostertag EM, Kazazian HH Jr. Biology of mammalian L1 retrotransposons. Annu Rev Genet 2001;35:501–538. [PubMed: 11700292]

Ostertag EM, Prak ET, DeBerardinis RJ, Moran JV, Kazazian HH Jr. Determination of L1 retrotransposition kinetics in cultured cells. Nucleic Acids Res 2000;28:1418–1423. [PubMed: 10684937]

Ovchinnikov I, Troxel AB, Swergold GD. Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. Genome Res 2001;11:2050–2058. [PubMed: 11731495]

Pavlicek A, Jabbari K, Paces J, Paces V, Hejnar JV, Bernardi G. Similar integration but different stability of Alus and LINEs in the human genome. Gene 2001;276:39–45. [PubMed: 11591470]

Perepelitsa-Belancio V, Deininger P. RNA truncation by premature polyadenylation attenuates human mobile element activity. Nat Genet 2003;35:363–366. [PubMed: 14625551]

Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 2004;428:493–521. [PubMed: 15057822]

Roy-Engel AM, El Sawy M, Farooq L, Odom GL, Perepelitsa-Belancio V, Bruch H, Oyeniran OO, Deininger PL. Human retroelements may introduce intragenic polyadenylation signals. Cytogenet Genome Res 2005;110:365–371. [PubMed: 16093688]

Salem AH, Myers JS, Otieno AC, Watkins WS, Jorde LB, Batzer MA. LINE-1 preTa elements in the human genome. J Mol Biol 2003;326:1127–1146. [PubMed: 12589758]

Soriano P, Meunier-Rotival M, Bernardi G. The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes. Proc Natl Acad Sci U S A 1983;80:1816–1820. [PubMed: 6572942]

Symer DE, Connelly C, Szak ST, Caputo EM, Cost GJ, Parmigiani G, Boeke JD. Human l1 retrotransposition is associated with genetic instability in vivo. Cell 2002;110:327–338. [PubMed: 12176320]

Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD. Molecular archeology of L1 insertions in the human genome. Genome Biol 2002;3:52.1–52.18.

Wang H, Xing J, Grover D, Hedges Kyudong Han DJ, Walker JA, Batzer MA. SVA elements: a hominid-specific retroposon family. J Mol Biol 2005;354:994–1007. [PubMed: 16288912]

Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, Boeke JD, Moran JV. Human L1 retrotransposition: cis preference versus trans complementation. Mol Cell Biol 2001;21:1429–1439. [PubMed: 11158327]

Wheelan SJ, Aizawa Y, Han JS, Boeke JD. Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution. Genome Res 2005;15:1073–1078. [PubMed: 16024818]

Abbreviations

ASV
avian sarcoma virus

CMV
Cytomegalovirus

HPV
human papillomavirus

L1
LINE-1 retrotransposon

LINE
long interspersed element

NeoR
neomycin resistance gene

ORF
open reading frame

RT

reverse transcriptase

SINE

short interspersed element

TPRT

target-primed reverse transcription

TSD

target site duplication

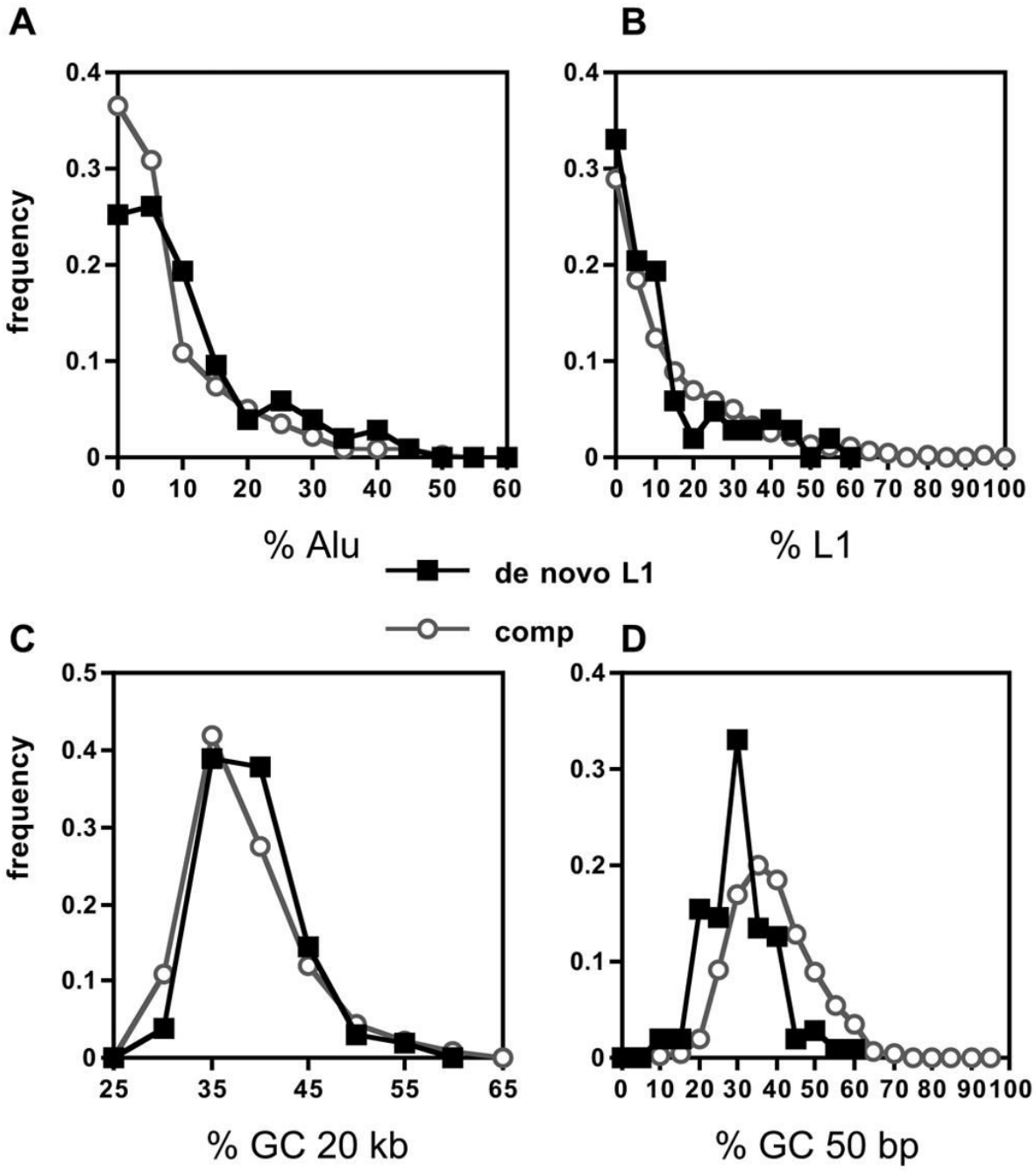

Fig 1.

Alu, L1, and G+C content of *de novo* HeLa L1 pre-insertion loci and random sites. The 10 kb flanks on each side of *de novo* L1 pre-insertion sites (closed squares) and simulated sites (open circles) were characterized for frequency containing A) %Alu content, B) %L1 content, C) % G+C content for 10 kb on either side of the endonuclease site/random site, and D) %G+C content for 25 bp on either side of the endonuclease site/random site. X-axis values are binned by 5.

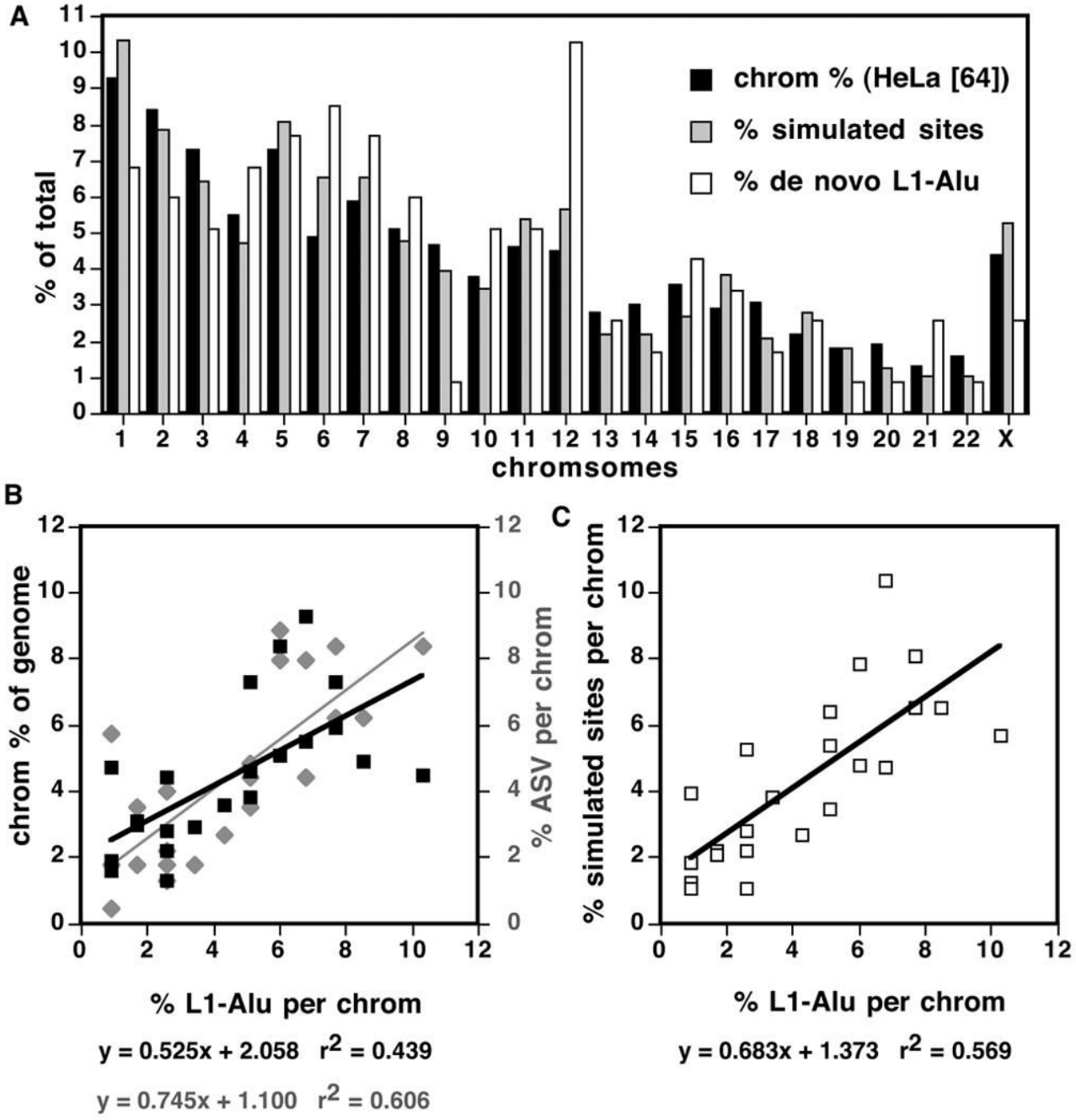

Fig 2.

Chromosome distribution of *de novo* L1 and SINE pre-insertion loci A) *de novo* L1 and SINE insertion sites were mapped to individual chromosomes according to the chromosome assignment of the highest contig BLAST hit and each chromosome is presented as a % of the total (white). This was compared to the percent chromosome content with corrections for the HeLa karyotype (black). The chromosome distributions of >1000 random insertion sites from a computer simulation with corrections for the HeLa karyotype are also shown (grey). B) Plot and curve fits of chromosome distributions of *de novo* L1 and SINE insertions (x-axis, as % of total) versus % chromosome content as calculated previously (squares's, black line) and versus % ASV insertions (grey diamonds, grey line, right Y-axis. C) Plot and curve fit of

chromosome distributions of *de novo* L1 and Alu insertions (x-axis, as % of total) versus a computer simulation of random sites in HeLa total (y- axis).

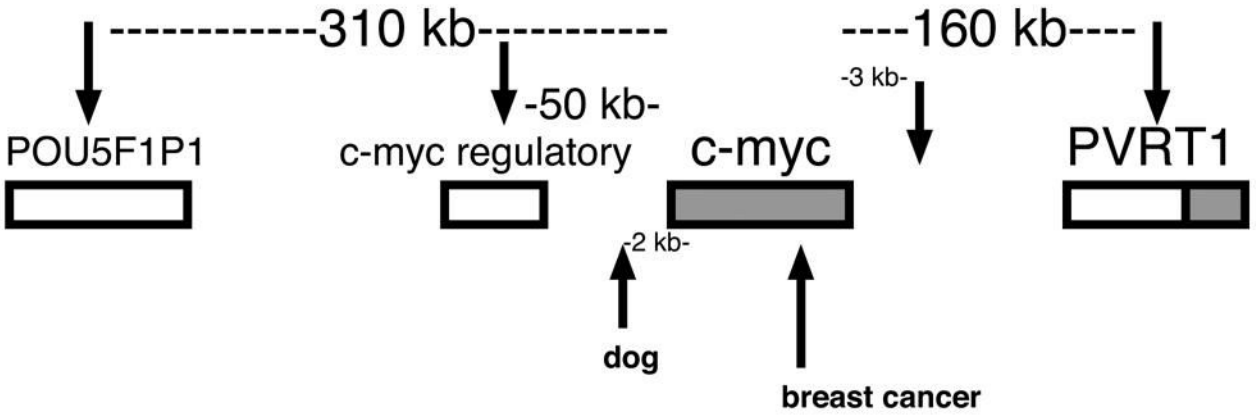

Fig 3.

Insertions near the c-*myc* locus. A schematic of the c-*myc* locus with 5' flanking pseudogene POU5F1P1 and 3' flanking PVRT gene is presented. The locations of 4 *de novo* L1 insertions are marked with arrows above the genes pointing down. The locations into c-*myc* of a somatic L1 insertion/rearrangement from a breast cancer and the site of a canine L1 insertion shown with arrows pointing up.

Table 1

Repetitive Element and GC content of L1 Pre-insertion Loci

	%Alu	%L1	% Alu +L1	%MIR	%L2	% MaLR	%GC	%GC 50bp[a]
L1[b]	**13.0**	**13.3**	**26.3**	**3.2**	**3.7**	**3.7**	**41**	**32**
Genome[c]	10.6	16.9	27.5	2.5	3.2	3.6	41	41 (32[d])
Sim. Random[e]	10.0	16.4	26.4	2.9	3.2	3.8	41	41
P-value[f]	0.0039	0.063		0.28	0.20	0.85	0.33	$<10^{-14}$
Sim. random (A +T-rich 50 bp[g])	8.3	19.5		2.5	3.2	3.8	39	33
P-value[f]	$<10^{-6}$	0.00051		0.011	0.18	0.81	$<10^{-8}$	0.67
SINES[h]	26.9	15.5	42.4	1.9	1.7	1.4	44	34
Sim. Random[e]	10.0	16.4	26.4	2.9	3.2	3.8	41	41
P-value[i]	$<10^{-9}$	0.84		0.16	0.11	0.042	0.034	0.029

[a] 25 bp to each side of endonuclease cut site

[b] All endonuclease-dependent and non L1-L1 chimeras (Gilbert et al 2005)

[c] (International Human Genome Sequence Constortium 2001)

[d] expected if 11 of 50 bases involved in the cleavage site are restricted to A+T

[e] random simulation of insertion sites in HeLa

[f] ANOVA comparing 103 *de novo* L1 PIL versus the simulated or subset thereof

[g] subset of random with restricted range of GC% in 50 bp region 26–37.9

[h] (Dewannieux et al 2003; Dewannieux and Heidmann 2005)

[i] ANOVA comparing 13 *de novo* SINE PIL versus the simulated