Supplemental Material for

L1 Integration in a Transgenic Mouse Model

Daria V. Babushok[1], Eric M. Ostertag[1,2,3], Christine E. Courtney[1], Janice M. Choi[1], and Haig H. Kazazian, Jr.[1,4]

[1]Department of Genetics, University of Pennsylvania, Philadelphia, PA 19104, USA; [2]Department of Pathology and Laboratory Medicine, Hospital of the University of Pennsylvania, Philadelphia, PA 19104, USA; [3]Transposagen Biopharmaceuticals, Inc., Philadelphia, PA 19104, USA.

[3]Corresponding author. Email: kazazian@mail.med.upenn.edu; Fax: (215) 573-7760.

Methods

Transgenic Mice

Purified transgenes were microinjected into fertilized mouse oocytes at the University of Pennsylvania School of Medicine Transgenic and Chimeric Mouse Facility to obtain founder mice. Founders were bred with wild type mice (129/SV) to establish nine stable transgenic lines. These studies were approved by the University of Pennsylvania Institutional Animal Care and Use Committee.

PCR Primers

Oligonucleotides used for genotyping are mlessfor2 (GAGAATTCCCACAACATCG) and L16045(30) (ATGCTAGATGACACATTAGTGGGTGCAGCG). Oligos used for transgene-specific PCR are GGAACACCACTCATTGTTCAAGGTC for transgene-11 and CACTAAACCACTATTCCTGCCCTTAG for transgene-18. Oligos used in TAIL-PCR are the SV40 pA-specific oligos SP1 (AACTTAAAGTATAATAAAGACGTCAGG), SP2 (AACTTAAAGTATAATAAAGACGTCAGGGTTCG); SP3 (GTCAGGGTTCGAAATCGATAAGC), and SP4 (CAAACCACAACTAGAATGCAGTG), and arbitrary degenerate oligos AD1 (NTCGANTWTSGWGTT), AD3 (WGTGNAGWANCANAGA), AD6 (AGWGNAGWANCTTAGG), and AD8 (TNSTICGNACWTWGGA). Oligo sequences used for characterization of the 5' ends of individual insertions were designed by Primer3 program (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi).

Genotyping PCRs

Screening for overall transgene presence was done by "genotyping PCR" amplifying across the intron region with primers L16045(30) and mlessfor2. The reactions were carried out on 160 ng of genomic DNA isolated from mouse tails with Taq DNA Polymerase (Promega) in a 25 μl reaction volume using 1.5 min extension time according to the manufacturer's protocol. The presence of individual transgenes (transgene-11 and transgene-18) was determined in a similar reaction, using SP4 and a 3' flanking primer specific to a particular transgene's genomic location as identified by TAIL-PCR. Genotyping F1 animals for presence of individual transgenes allowed us to calculate F0 to F1 transgene transmission rates of 39% for transgene-11, 22% for transgene-18, and 46% for transgene-Un, reflecting different degrees of mosaicism in the F0 germline.

Southern Blotting

Genomic DNA was isolated from mouse tails as described previously (Palmiter et al. 1985; Sambrook 2001). Ten micrograms of DNA were digested with AvrII (NEB), resolved on 0.7% agarose gel/1X TAE, and subjected to Southern Blot analysis using standard protocol (Sambrook 2001). The blots were probed with a 376 bp probe encompassing the whole region from the intron splice site junction to the polyadenylation cleavage site. The probe was ^{32}P-dCTP-labeled using the Random Primed DNA

Labeling Kit (Roche) according to the manufacturer's protocol. The blots were exposed on a phosphor screen and read on a phosphoimager (Molecular Dynamics).

Statistical Analysis

L1 integration preferences with respect to genomic landmarks were evaluated by normal approximation to binomial probability distribution.

Briefly, to test the hypothesis that L1 integrants are underrepresented in genes and CpG islands as is typical of endogenous L1s, a one-sided test was done with the following conditions: n = 48 uniquely identifiable insertions; two possible values for each insert: *in* or *out* of the genomic feature (GF) (e.g. a gene or a CpG island); π = the expected probability of being *in*; continuity correction, *cc* = 0.5; type 1 error, $\alpha = 0.05$. The binomial mean and variance are given by $\mu = n \cdot \pi$ and $\sigma^2 = n \cdot \pi \cdot (1 - \pi)$, respectively. H_0: π = the proportion of the mouse genome occupied by GF; H_1: $\pi <$ proportion of the genome occupied by GF. $prob_{bin}\{x \le in$, when π = proportion of GF in the genome$\} \approx prob_{nor}\{y \le (in + cc)$, where *y* has a normal distribution, with a mean, μ, and variance, σ^2) = $prob_{nor}\{\frac{(y-\mu)}{\sigma} \le \frac{(in+cc)-\mu}{\sigma}\} = prob_{nor}\{Z \le \frac{(in+cc)-\mu}{\sigma}\}$. For each *Z*, *p*-values were determined using standard normal distribution (μ=0, σ=1). H_0 was rejected for p-values ≤ 0.05.

To evaluate whether L1 insertions are distributed randomly with respect to genomic repeats and individual chromosomes, a two-sided test was done with the following conditions: n = 51 insertion sites amenable to repeat analysis or n = 49 insertions with assigned chromosome number (including Un_random); two possible values for each insert: *in* or *out* of the genomic feature (GF) (e.g. a repeat or a particular chromosome); π = the expected probability of being *in*; continuity correction, *cc* = 0.5; type 1 error, $\alpha = 0.05$. The binomial mean and variance are given by $\mu = n \cdot \pi$ and $\sigma^2 = n \cdot \pi \cdot (1 - \pi)$, respectively. H_0: π = the proportion of the mouse genome occupied by GF; H_1: $\pi \neq$ proportion of the genome occupied by GF. H_0 will be rejected if $prob_{bin}\{x \le in$, when π = proportion of GF in the genome$\} \le 0.025$ or $prob_{bin}\{x \ge in$, when π = proportion of GF in the genome$\} \le 0.025$. $prob_{bin}\{x \le in$, when π = proportion of GF in the genome$\} \approx prob_{nor}\{y \le (in + cc)$, where *y* has a normal distribution, with a mean, μ, and variance, σ^2) = $prob_{nor}\{\frac{(y-\mu)}{\sigma} \le \frac{(in+cc)-\mu}{\sigma}\} = prob_{nor}\{Z \le \frac{(in+cc)-\mu}{\sigma}\}$. $prob_{bin}\{x \ge in$, when π = proportion of GP in the genome$\} \approx prob_{nor}\{y \ge (in - cc)$, where *y* has a normal distribution, with a mean, μ, and variance, σ^2) = $prob_{nor}\{\frac{(y-\mu)}{\sigma} \ge \frac{(in-cc)-\mu}{\sigma}\} =$ $prob_{nor}\{Z \ge \frac{(in-cc)-\mu}{\sigma}\}$. For each *Z*, *p*-values were determined using standard normal distribution (μ=0, σ=1).

Finally, binomial probability distribution was used to conservatively estimate the frequency of small target site alterations (>30 nt TSDs and small target site deletions) accompanying L1 retrotransposition. Briefly, the binomial probability of having *x* target site alterations in *n* insertions is given by $\frac{n!}{x!(n-x)!}q^x(1-q)^{n-x}$, where *q* is a probability of having an alteration on any given insertion. Using this formula, the probability of seeing no target site alterations in 33 inserts with a q of 0.087 is 0.0496, strongly suggesting that target site alterations occur in less than 8.7% of L1 retrotransposition events *in vivo*.

Analysis of Microhomologies at 5' Boundaries and Inversion Points of *de novo* Integrants.

Distribution of 5' microhomologies (5'MHs) was determined separately for 5' truncated elements, 5' truncated elements containing an inversion, and full length elements. Elements containing extra 5' nts were excluded from these analyses. Only one characterized insert originated precisely at the L1 promoter, and, in agreement with lower rates of 5'MH in FL elements (Zingler et al. 2005), it lacked a 5'MH. The distribution of inversion point MHs was determined by analyzing the sequence at the junction of the inverted and direct fragments in 5' truncated elements containing an inversion. The MH distribution expected by chance was calculated as described previously (Roth et al. 1985; Symer et al. 2002; Zingler et al. 2005) for the random nucleotide distribution (p =0.25), as well as for the nucleotide distribution of L1-mless transgene construct and the 30 nucleotides downstream of the insertion site, where 2nd strand cleavage is thought to occur

($p = a_{host} \cdot a_{insert} + t_{host} \cdot t_{insert} + c_{host} \cdot c_{insert} + g_{host} \cdot g_{insert} = 0.27$, where *a*, *t*, *c*, and *g* are the average frequencies of the corresponding nucleotides in the target site sequences ("host") or in the insertion construct used in this study ("insert")).

Discussion

Supplemental Discussion 1

Another possible explanation for the lack of insertions with unusually long TSDs in our study is the possible growth disadvantage or selective loss of cells in chimeric animals. Long TSDs are likely to cause genomic instability in the host cell due to DNA polymerase slippage and subsequent expansion or contraction of direct repeats (Chen et al. 2005). Negative selective pressures are likely weaker in HeLa and HCT116 cancer cells allowing to recover insertions with long TSDs in these cell lines (Gilbert et al. 2002; Symer et al. 2002; Gilbert et al. 2005). If long TSDs are formed *in vivo* at high frequencies, they must be very detrimental and quickly lost from the population, explaining their extremely low occurrence among genomic L1s (Szak et al. 2002).

Supplemental Discussion 2

Frequencies of deletions accompanying endogenous L1s had been difficult to estimate because of the lack of knowledge of most preintegration sites. Recent studies employed conceptual reconstruction of preintegration sites of Ta and preTa L1 elements followed by empty site amplification by PCR in the orthologous loci in a panel of primates; 3-5 deletions/254 Ta elements (1.2-1.9%) and 1 deletion /254 preTa elements (0.4%) were reported (Vincent et al. 2003; Ho et al. 2005). Deletions ranged from 1 to 372 nts with a median size of 5.5 nts. As noted by the authors, small deletions of <20 nts may have been missed. 1.9%-4.7% of orthologous empty sites failed to amplify in any primate species, a result that could be caused by large deletions accompanying L1 integration. However, as noted by the authors, possible mutations in oligonucleotide sites used for PCR likely contributed to amplification failure. This explanation is supported by decreased successful amplifications with increasing evolutionary age (88.2% (224/254) successfully amplified sites for pigmy chimpanzee and gorilla, 16.5% (42/254)— for galago) (Ho et al. 2005). Additionally, unrelated genomic rearrangements, such as non L1-mediated deletions, repeat expansions, and independent retrotransposon insertions, made up the majority (87-89%) of all unexpected size alterations and likely contributed to amplification failure (Vincent et al. 2003; Ho et al. 2005). It is very likely that the frequency of large deletions in these studies is much less than 1.9% – 4.7% estimate derived from unamplified preintegration sites. Infrequent retrotransposition-mediated deletions were further confirmed in comparisons of lineage-specific L1 and Alu integrations in the human and chimpanzee genomes: 2.2% of L1 and 0.21% of Alu insertions were accompanied by deletions, with a median L1-mediated deletion size of 21 bp (Han et al. 2004; Callinan et al. 2005).

Supplemental Discussion 3

L1 RT template jumping in the course of twin priming likely explains the addition of 7 extra nucleotides in another insertion. In this case, 7 bp were added between the 5' flanking host DNA and the 5' end of an otherwise standard inverted integrant of 710 bp, containing an inversion of 30 bp and a direct fragment of 680 bp, with a concomitant deletion of the intervening 2645 nts. The insertion occurred at a typical 5'-TTTT'G-3' endonuclease cleavage site, and was flanked by a 14 bp TSD. Because the 7 unknown bases were inserted 5' to the unambiguous 30 bp inverted fragment, it is likely that they were formed in a preceding twin priming reaction. While we are unable to definitively establish the origin of these 7 bps because of insufficient length, there are three candidate templates for this complete 7 nucleotide stretch within the L1 element upstream of the direct fragment. It is possible that other cases of few extra nt additions could result from similar short inversion events.

References for Supplemental Material

Callinan, P.A., Wang, J., Herke, S.W., Garber, R.K., Liang, P., and Batzer, M.A. 2005. Alu retrotransposition-mediated deletion. *J Mol Biol* **348:** 791-800.

Chen, J.M., Chuzhanova, N., Stenson, P.D., Ferec, C., and Cooper, D.N. 2005. Meta-analysis of gross insertions causing human genetic disease: novel mutational mechanisms and the role of replication slippage. *Hum Mutat* **25:** 207-221.

Gilbert, N., Lutz, S., Morrish, T.A., and Moran, J.V. 2005. Multiple fates of L1 retrotransposition intermediates in cultured human cells. *Mol Cell Biol* **25:** 7780-7795.

Gilbert, N., Lutz-Prigge, S., and Moran, J.V. 2002. Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110:** 315-325.

Han, J.S., Szak, S.T., and Boeke, J.D. 2004. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* **429:** 268-274.

Ho, H.J., Ray, D.A., Salem, A.H., Myers, J.S., and Batzer, M.A. 2005. Straightening out the LINEs: LINE-1 orthologous loci. *Genomics* **85:** 201-207.

Palmiter, R.D., Chen, H.Y., Messing, A., and Brinster, R.L. 1985. SV40 enhancer and large-T antigen are instrumental in development of choroid plexus tumours in transgenic mice. *Nature* **316:** 457-460.

Roth, D.B., Porter, T.N., and Wilson, J.H. 1985. Mechanisms of nonhomologous recombination in mammalian cells. *Mol Cell Biol* **5:** 2599-2607.

Sambrook, J. 2001. *Molecular cloning: a laboratory manual/Joseph Sambrook, David W. Russell*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Symer, D.E., Connelly, C., Szak, S.T., Caputo, E.M., Cost, G.J., Parmigiani, G., and Boeke, J.D. 2002. Human l1 retrotransposition is associated with genetic instability in vivo. *Cell* **110:** 327-338.

Szak, S.T., Pickeral, O.K., Makalowski, W., Boguski, M.S., Landsman, D., and Boeke, J.D. 2002. Molecular archeology of L1 insertions in the human genome. *Genome Biol* **3:** research0052.

Vincent, B.J., Myers, J.S., Ho, H.J., Kilroy, G.E., Walker, J.A., Watkins, W.S., Jorde, L.B., and Batzer, M.A. 2003. Following the LINEs: an analysis of primate genomic variation at human-specific LINE-1 insertion sites. *Mol Biol Evol* **20:** 1338-1348.

Zingler, N., Willhoeft, U., Brose, H.P., Schoder, V., Jahns, T., Hanschmann, K.M., Morrish, T.A., Lower, J., and Schumann, G.G. 2005. Analysis of 5' junctions of human LINE-1 and Alu retrotransposons suggests an alternative model for 5'-end attachment requiring microhomology-mediated end-joining. *Genome Res* **15:** 780-789.

Web Site References for Supplemental Material

http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi; Primer3, the web based primer picking service at the Whitehead Institute for Biomedical Research.