Available online at www.sciencedirect.com
ScienceDirect
Gene xx (2006) xxx – xxx
GENE
SECTION
EVOLUTIONARY GENOMICS
www.elsevier.com/locate/gene

Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination

Mingzhou Song [a], Stéphane Boissinot [b,c,*]

[a] *Department of Computer Science, New Mexico State University, Las Cruces, New Mexico 88003, USA*
[b] *Department of Biology, Queens College, The City University of New York, Flushing, New York 11367, USA*
[c] *Graduate School and University Center, The City University of New York, New York, New York 10016, USA*

Received 1 August 2006; received in revised form 27 September 2006; accepted 28 September 2006

Received by M. Batzer

Abstract

LINE-1 (L1) retrotransposons constitute the most successful family of autonomous retroelements in mammals and they represent at least 17% of the size of the human genome. L1 insertions have occasionally been recruited to perform a beneficial function but the vast majority of L1 inserts are either neutral or deleterious. The basis for the deleterious effect of L1 remains a matter of debate and three possible mechanisms have been suggested: the direct effect of L1 inserts on gene activity, genetic rearrangements caused by L1-mediated ectopic recombination, or the retrotransposition process *per se*. We performed a genome-wide analysis of the distribution of L1 retrotransposons relative to the local recombination rate and the age and length of the elements. The proportion of L1 elements that are longer than 1.2 Kb is higher in low-recombining regions of the genome than in regions with a high recombination rate, but the genomic distributions of full-length elements (i.e. elements capable of retrotransposition) and long truncated elements were indistinguishable. We also found that the intensity of selection against long elements is proportional to the replicative success of L1 families. This suggests that the deleterious effect of L1 elements results principally from their ability to mediate ectopic recombination.
© 2006 Elsevier B.V. All rights reserved.

Keywords: L1/LINE-1; Retrotransposon; Human; Recombination; Genome evolution

1. Introduction

The abundance of LINE-1 (L1) retrotransposons constitutes one of the most puzzling features of mammalian genomes and it is now clear that they have profoundly affected the structure and function of these genomes (Lander et al., 2001; Waterston et al., 2002; Kazazian, 2004). However, the evolutionary forces affecting their genomic distribution and dynamics in natural populations remain incompletely understood. Although L1 sequences have occasionally been recruited to perform a function beneficial to the host (Kazazian, 2004; Han and Boeke, 2005), the vast majority of new insertions are more likely to be either neutral or detrimental. Therefore, the extent of L1 amplification will depend on two opposing factors: the retrotransposition rate and the intensity of selection against the deleterious effect of L1 activity. The basis for selection against retrotransposon insertions could be either the direct effect of where elements insert (e.g., gene inactivation) (Charlesworth and Charlesworth, 1983; Finnegan, 1992), the effect of genetic rearrangements caused by ectopic recombination (Langley et al., 1988), or the retrotransposition process *per se* (e.g., a deleterious effect of L1 gene products) (Nuzhdin, 1999; Boissinot et al., 2001). Although these three mechanisms could all affect the fitness of individuals, their relative importance remains a matter of debate (Biemont et al., 1997; Charlesworth et al., 1997; Boissinot et al., 2001; Furano et al., 2004; Neafsey et al., 2004).

Abbreviations: LINE-1 or L1, Long interspersed element 1; Kb, Kilo base pair; bp, Base pair; Mb, Mega base pair; TR, truncated; FL, full-length; Myr, Million of year; UTR, Untranslated region; Ta, Transcribed subset a; NRR, non-recombining region; PAR, pseudo-autosomal region; ANOVA, analysis of variance; HSD, Honest Significant Difference; RR, recombination rate; UCSC, University of California-Santa Cruz.

* Corresponding author. Department of Biology, Queens College, CUNY, 65-30 Kissena Boulevard, Flushing, NY 11367-1597, USA. Tel.: +1 718 997 3437; fax: +1 718 997 3321.
E-mail address: stephane.boissinot@qc.cuny.edu (S. Boissinot).

0378-1119/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2006.09.033

Table 1
Proportion of full-length (FL) L1 elements on autosomes, the X and the Y chromosomes (excluding the pseudo-autosomal region)

	Autosomes			X chromosome			Y chromosome			Ratios of FL abundance		
	Total L1	FL L1	% FL	Total L1	FL L1	% FL	Total L1	FL L1	% FL	Y to Aut	X to Aut	Y to X
L1PA2	3377	965	28.6	348	96	27.6	76	27	35.5	1.24	0.97	1.29
L1PA3	7305	1318	18.0	852	189	22.2	161	41	25.5	1.41	1.23	1.15
L1PA4	8575	1193	13.9	1001	211	21.1	180	48	26.7	1.92	1.52	1.27
L1PA5	8275	967	11.7	772	135	17.5	108	34	31.5	2.69	1.50	1.80
L1PA6	4247	943	22.2	415	118	28.4	76	28	36.8	1.66	1.28	1.30

In several species, a higher density of retrotransposons in regions of reduced recombination has been reported (Charlesworth et al., 1992b,a; Hoogland and Biemont, 1996; Boissinot et al., 2001; Bartolome et al., 2002; Dasilva et al., 2002). If we assume that the frequency of ectopic recombination correlates with the recombination rate (Langley et al., 1988), this observation seems to support the ectopic exchange model. However, because the efficacy of selection in low- and non-recombining regions of the genome is limited by Hill-Robertson interactions (Hill and Robertson, 1966) and by Muller's ratchet (Felsenstein, 1974), deleterious retrotransposons are expected to accumulate in regions of reduced recombination whatever the nature of their deleterious effect. In addition, regions of low recombination are known to be relatively gene-poor (Fullerton et al., 2001). In humans, the proportion of full-length L1 elements on the non-recombining region of the Y chromosome was previously reported to be higher than on the autosomes, suggesting that full-length inserts had been subjected to purifying selection (Boissinot et al., 2001). However this observation was based on the analysis of a very small fraction of the genome and the number of L1 elements analyzed did not permit the determination of the basis for selection against full-length elements. Thus, although the accumulation of retrotransposons in low-recombining regions suggests they are indeed deleterious, this observation does not unambiguously support or contradict any of the three selection models.

Here, we tested some of the predictions associated with the three models of selection. If the gene inactivation model is correct, truncated (TR) and full-length (FL) elements should have similar genomic distribution because they both can affect gene function. If the retrotransposition process itself is deleterious, then selection should act only against FL elements (i.e. potentially active elements) and FL elements but not TR ones should accumulate in regions of low recombination. Finally, if the ectopic exchange model is correct, long elements (FL and long TR) should accumulate in regions of low or non-recombination to a greater extent than short TR elements, because they are more likely to mediate ectopic recombination and therefore to be deleterious. We performed a genome-wide analysis of the distribution of L1 retrotransposons relative to the

Table 2
Proportion of full-length (FL) and truncated (TR) elements on human chromosomes

Chromosomes	L1PA2			L1PA3			L1PA4			L1PA5			L1PA6		
	FL	TR	% FL	FL	TR	% FL	FL	TR	% FL	FL	TR	% FL	FL	TR	% FL
1	63	176	26.4	108	470	18.7	93	596	13.5	64	552	10.4	80	279	22.3
2	96	224	30.0	107	511	17.3	99	655	13.1	76	644	10.6	79	295	21.1
3	83	176	32.0	103	512	16.7	111	583	16.0	87	576	13.1	89	257	25.7
4	84	231	26.7	125	533	19.0	111	629	15.0	95	636	13.0	92	271	25.3
5	82	194	29.7	116	490	19.1	106	618	14.6	83	575	12.6	84	255	24.8
6	62	158	28.2	99	408	19.5	89	528	14.4	69	469	12.8	79	250	24.0
7	55	147	27.2	72	317	18.5	71	385	15.6	52	400	11.5	51	203	20.1
8	74	149	33.2	96	360	21.1	80	440	15.4	61	420	12.7	57	161	26.1
9	42	97	30.2	56	277	16.8	44	324	12.0	46	328	12.3	32	139	18.7
10	36	100	26.5	63	281	18.3	58	316	15.5	47	338	12.2	48	139	25.7
11	55	146	27.4	83	320	20.6	86	457	15.8	53	397	11.8	56	183	23.4
12	51	110	31.7	84	309	21.4	67	377	15.1	48	394	10.9	45	154	22.6
13	31	94	24.8	38	214	15.1	19	271	6.6	30	286	9.5	27	117	18.7
14	30	85	26.1	42	199	17.4	36	233	13.4	46	251	15.5	34	91	27.2
15	31	56	35.6	31	162	16.1	32	167	16.1	20	189	9.6	19	79	19.4
16	16	58	21.6	21	106	16.5	24	143	14.4	20	149	11.8	11	44	20.0
17	12	31	27.9	10	98	9.3	6	106	5.4	10	130	7.1	13	79	14.1
18	31	63	33.0	22	154	12.5	22	179	10.9	28	234	10.7	16	95	14.4
19	6	21	22.2	16	64	20.0	17	91	15.7	7	62	10.1	12	49	19.7
20	13	44	22.8	19	94	16.8	8	125	6.0	13	108	10.7	14	46	23.3
21	8	28	22.2	4	63	6.0	6	83	6.7	4	85	4.5	2	44	4.3
22	2	16	11.1	3	26	10.3	8	49	14.0	8	42	16.0	3	29	9.4
Total Autosomes	963	2404	28.6	1318	5968	18.1	1193	7355	14.0	967	7265	11.7	943	3259	22.4
X	96	252	27.6	189	663	22.2	211	790	21.1	135	637	17.5	118	297	28.4
Y	27	49	35.5	41	120	25.5	48	132	26.7	34	74	31.5	28	48	36.8

Please cite this article as: Song, M., Boissinot, S. Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination. Gene (2006), doi:10.1016/j.gene.2006.09.033

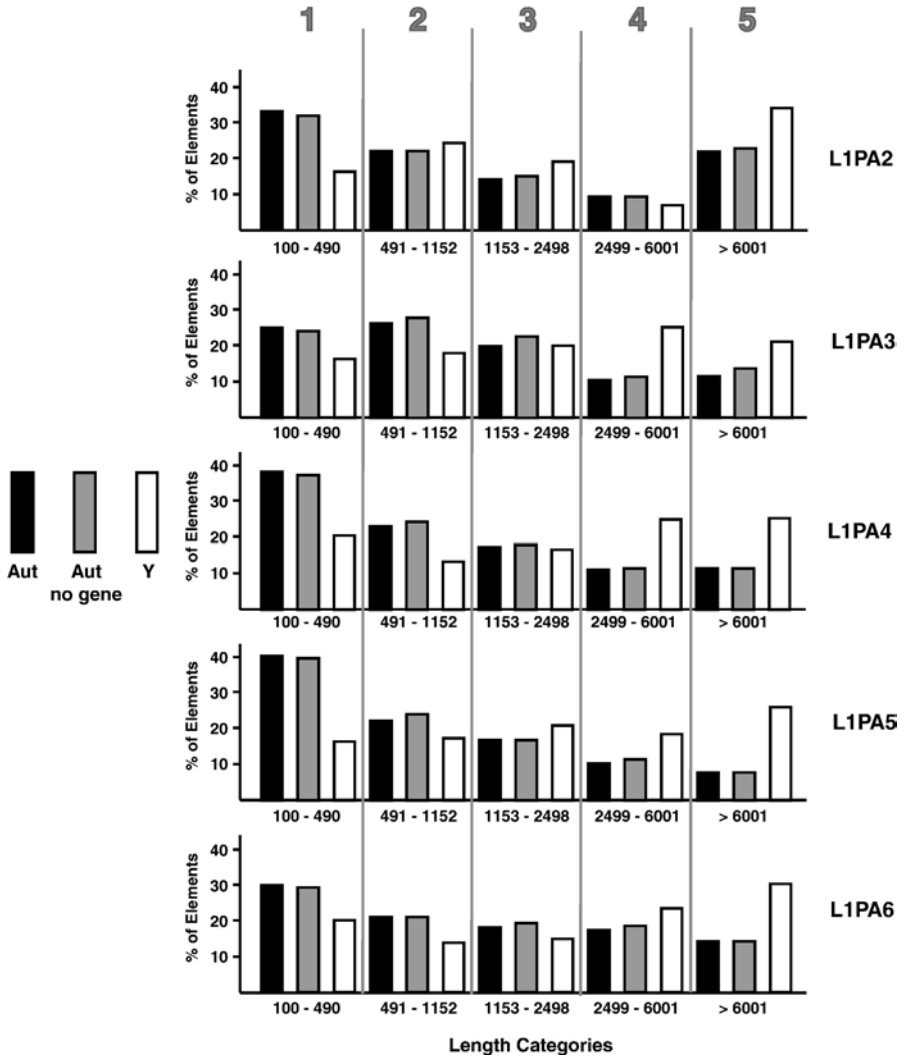

Fig. 1. Length distribution of autosomal and Y-linked L1 elements. The numbers at the top of the figure correspond to the length categories. Elements located outside of genes (Aut no gene) were obtained by eliminating from the data set all L1 elements (between 24.1 and 26.1% of the total depending on the family) which are within a RefSeq gene (obtained from the RefSeq table at http://genome.ucsc.edu).

recombination rate and the length and family classification of the elements. We found that both FL and long TR elements (>1.2 Kb) are more abundant in non- and low-recombining regions of the genome but we did not detect any apparent differences between FL and long TR elements. We also found that the intensity of selection against long elements is proportional to the replicative success of families, suggesting that the more active an L1 family is, the more deleterious its activity is for the host. These observations provide strong support for the ectopic exchange model and suggest that the deleterious effect of L1 elements result principally from their ability to mediate ectopic recombination.

2. Materials and methods

The coordinates of L1 elements belonging to families L1PA6 to L1PA2 were obtained from table RepeatMasker (assembly of April 2003) at http://genome.ucsc.edu. These families have evolved as a single lineage over the last 27 Myr and have produced virtually all the L1 elements that was inserted in the human genome since the split between the Cercopithecidae (Old World monkeys) and Hominidae (Human, apes and gibbons). The age of families L1PA6, L1PA5, L1PA4, L1PA3 and L1PA2 are respectively 26.8, 20.4, 18.0, 12.5 and 7.6 Myr (Khan et al., 2006). The most recent and currently active human L1 family (L1PA1 or L1Hs) was excluded from this analysis because it contains a mixture of fixed and polymorphic elements which may still be subject to selection. In addition, the evolution of the L1PA1 family and the effect of selection have been examined in detail in several recent studies (Myers et al., 2002; Boissinot et al., 2000, 2004, 2006). The length of each element was directly determined from the RepeatMasker table. When an element had an inverted bipartite structure, the two parts of the inverted element were fused into a single continuous element. Elements shorter than 100 bp (i.e. 7.8% of the total number of elements) were excluded from the analysis because family identification of such short elements by RepeatMasker can be inaccurate. Because the length of elements can change after

Please cite this article as: Song, M., Boissinot, S. Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination. Gene (2006), doi:10.1016/j.gene.2006.09.033

Table 3
Proportion of full-length (FL) elements with no internal deletions

Chromosomes	Families	FL elements	FL with deletions	% of deletion-free FL elements
Autosomes	L1PA2	965	867	89.8
	L1PA3	1318	1141	86.6
	L1PA4	1193	980	82.1
	L1PA5	967	737	76.2
	L1PA6	943	609	64.6
X Chromosome	L1PA2	96	87	90.6
	L1PA3	189	164	86.8
	L1PA4	211	161	76.3
	L1PA5	135	109	80.7
	L1PA6	118	78	66.1
Y Chromosome (NRR)	L1PA2	27	26	96.3
	L1PA3	41	29	70.7
	L1PA4	48	36	75.0
	L1PA5	34	22	64.7
	L1PA6	28	18	64.3

insertion through internal insertions and deletions, we considered that an L1 element was FL if its sequence started at position 1 of a FL L1 consensus and ended at the 3′ end of the 3′UTR.

Two estimators of the recombination rate were used. The sex averaged recombination from the deCode Iceland data set (Kong et al., 2002) was used to estimate the recombination rate distribution function for entire chromosomes by optimal quantization (Song et al., 2003). This estimator (RR) is computed in variable window sizes factored by the location and frequency of recombination data. This estimator allows more adaptive inspection of the recombination rate compared to the estimator provided on the UCSC genome browser (UCSCRR) which uses fixed 1 Mb windows to compute average recombination rates. Statistical analyses were performed using both the RR and UCSCRR estimators.

Statistical analyses were performed using programs written in the R language for statistical computing (R Development Core Team, 2004) and the C++ programming language. The mean recombination rate between TR and FL elements was compared using *t*-tests. The classification of L1 elements into length class was done using optimal quantization algorithm based on dynamic programming (Song and Haralick, 2002). The mean recombination rate among length subgroups were compared using analysis of variance and Tukey's Honest Significant Differences (HSD) test for multiple comparisons.

3. Results

3.1. Chromosomal distribution of full-length and truncated L1 elements

We first compared the relative abundance of full-length (FL) and truncated (TR) elements on the autosomes, the X chromosome and the non-recombining region (NRR) of the Y chromosome. Table 1 shows that a larger fraction (between 26 and 37% depending on the family) of L1 elements is FL on the NRR of the Y chromosome than on autosomes and on the X chromosome. The enrichment in FL elements of the Y chromosome relative to autosomes is observed across the five L1 families analyzed here and is statistically significant for all families except L1PA2. The fraction of FL elements on the X chromosome is intermediate (except for family L1PA2), although the ratios of FL elements on the X and autosomes do not significantly differ for families L1PA3, L1PA5 and L1PA6. Depending on the family, there are 1.2 to 2.7 times as many FL elements on the Y than on the autosomes and 1.4 to 2.2 times as many on the Y than on the X. The fraction of FL on the Y is similar among all families ($p>0.12$ for all comparisons using Fisher's exact test) and corresponds to the rate at which those elements are generated by active families (Boissinot et al., 2004). Interestingly, the pseudo-autosomal region (PAR) of the Y, which is subject to one of the highest recombination rates, contains only one FL elements for 25 TR elements (i.e. 4.0% of FL). Therefore, the larger proportion of FL element on the NRR of the Y is most likely due to the lack of recombination of this region rather than some other feature of the Y chromosome such as its mode of transmission. We also analyzed separately the 22 autosomes (Table 2) and found that the fraction of FL elements was remarkably similar among all autosomes with the exception of autosome 21 which has 2 to 4 times (depending on the family) fewer FL elements than other autosomes. This explains why the fraction of FL on autosomes reported here (12 to 29%) is significantly higher than reported earlier using a dataset limited to chromosome 21 and 22 (8% in Boissinot et al., 2001).

On most chromosomes (including most autosomes, the X, and the Y), families L1PA2 and L1PA6 have a larger proportion of FL elements than families L1PA3, L1PA4 and L1PA5 (Tables 1 and 2). For instance, the autosomal fractions of FL L1PA2 and L1PA6 elements are more than twice the fraction of FL L1PA5 elements. This suggests that the proportion of FL elements is not related to the age of a family as families L1PA2 and L1PA6 are respectively the youngest and the oldest families studied here. Instead, it seems the more abundant families show the lowest fractions of FL elements and a significant correlation between the copy number of L1 families and the fraction of FL elements is observed ($r=-0.95$, $p<0.01$). In addition, the difference in the proportion of FL elements between the Y chromosome and autosomes is larger for the more abundant families, suggesting that selection against FL inserts was stronger when L1 activity was high.

We then examined if the autosomal deficit of FL elements was limited to this class of elements or if long truncated

Table 4
Mean recombination rates at genomic locations where autosomal full-length (FL) and truncated (TR) elements are found

Family	Number of elements		Mean RR[a]			Mean UCSCRR[b]		
	FL	TR	FL	TR	*p*-value	FL	TR	*p*-value
L1PA2	965	2410	0.956	1.085	0.010	1.078	1.087	0.380
L1PA3	1318	5983	0.935	1.074	<0.001	1.048	1.069	0.182
L1PA4	1193	7383	1.033	1.100	0.118	1.015	1.117	<0.001
L1PA5	967	7314	0.972	1.121	<0.001	1.077	1.143	0.008
L1PA6	943	3302	0.964	1.042	0.053	1.011	1.166	<0.001
All families	5386	26392	0.977	1.082	<0.0001	1.041	1.095	<0.0001

[a] Recombination rate estimated using the method of Song et al. (2003).
[b] Recombination rate from the UCSC browser.

Please cite this article as: Song, M., Boissinot, S. Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination. Gene (2006), doi:10.1016/j.gene.2006.09.033

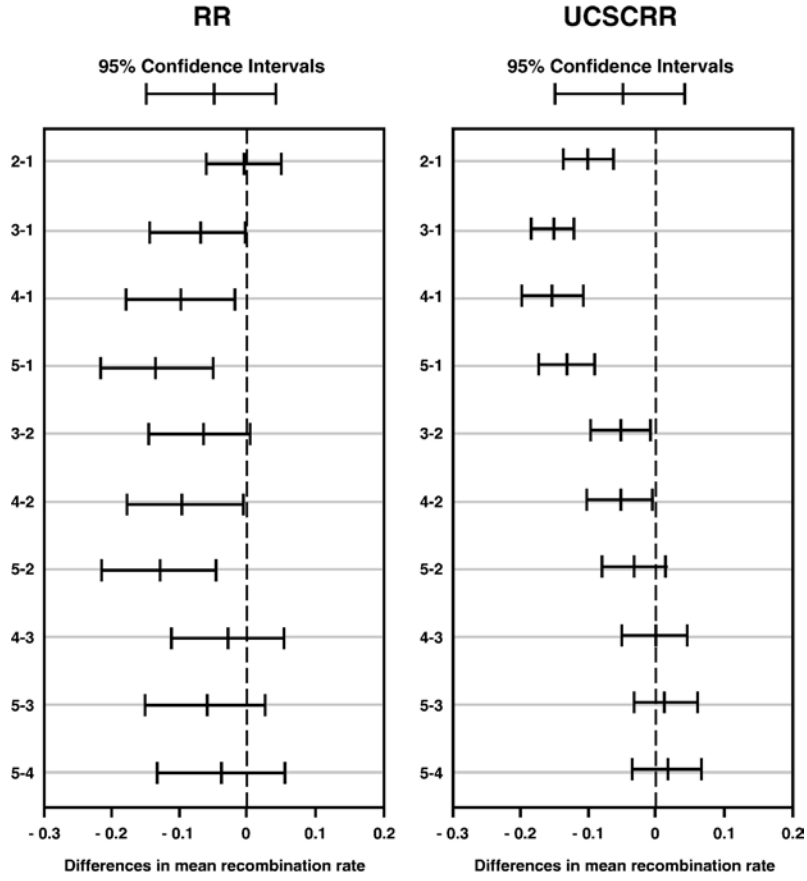

Fig. 2. Tukey's Honest Significant Differences test on the mean recombination rate among length subgroups. The range of each line segment corresponds to the 95% confidence interval of the mean recombination rate difference between the two length subgroups labeled on the left of the segment. The vertical dashed line marks the zero difference location. If an interval contains zero, it implies that there is no significant difference between the two subgroups. The numbers on the vertical axes correspond to length subgroups (1=100–490 bp; 2=491–1152 bp; 3=1153–2498; 4=2499–6001; 5=6002–6183). For example, 5−3 stands for the mean recombination rate of length category 5 minus that of length category 3.

elements were also less abundant on the autosomes than on the Y or X chromosomes. To this end, we divided L1 elements into subgroups corresponding to different length classes using the method described in (Song and Haralick, 2002). This method separates L1 elements into subgroups by length when there is a sudden change in the number of L1s over unit length. We selected the number of subgroups to be five, roughly capturing the overall distribution of length while at the same time assuring that the intervals are not too small for meaningful comparisons. Fig. 1 shows that for each family, the size distribution of the elements differs between the autosomes and the NRR of the Y chromosome ($p<0.025$ for all families using the Chi-square test). On average, autosomes contain a higher proportion of short elements than the Y chromosome. Depending on the family, between 25 and 41% of autosomal L1 elements are severely truncated (<490 bp) whereas only 16 to 20% of Y-linked elements fall in this category. Except for family L1PA2, long TR elements (i.e., elements between 2.5 and 6 Kb) are 1.3 to 2.3 times more abundant on the NRR of the Y than on autosomes. The relative abundance of long TR (>2.5 Kb) and FL elements on autosomes relative to the NRR of the Y are strikingly similar. For instance, long TR and FL L1PA4 elements are similarly abundant on the NRR of the Y (both with a frequency of 25%) and on autosomes (with a frequency of 11%; see Fig. 1). Therefore, it seems that FL or long TR inserts have been subjected to negative selection to the same extent. We repeated the length distribution analysis after excluding all L1 elements located within genes (Fig. 1). We found that the length distribution of autosomal L1 elements outside of genes differs from the Y chromosome distribution to the same extant that the overall autosomal distribution does. Therefore, the low gene content of the Y chromosome can not account for the difference in length distribution between the Y and the autosomes.

Although the above observations suggest that long L1 elements (FL and TR) accumulate on the NRR of the Y, it is also plausible that autosomal L1 elements are becoming shorter than Y-linked elements if internal DNA deletions are occurring more frequently on the autosomes than on the Y chromosome. To test this hypothesis, we determined the fraction of FL elements that are free of deletions (i.e. >6 Kb) on the Y, X, and autosomes (Table 3). As expected, FL elements belonging to older families are less likely to have retained their original length than those belonging to younger families, but for a given family there is no significant difference between elements on the autosomes, the Y, or the X. This indicates that the rate of decay of elements is

Please cite this article as: Song, M., Boissinot, S. Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination. Gene (2006), doi:10.1016/j.gene.2006.09.033

similar across the entire genome and we can exclude a role of DNA deletions as an explanation for the difference in size distribution between the autosomes, the Y, and the X.

3.2. Distribution of autosomal L1 elements relative to the recombination rate

Because of its complete lack of recombination, the NRR of the Y chromosome constitutes an extreme situation. Thus, we determined if the biased distribution of FL, short TR, and long TR elements relative to recombination rate could also be observed on the autosomes. For each autosomal L1 element, we collected the local recombination rate of the genomic region where it is located. Two estimators of the recombination rate were used: RR and UCSCRR (see Materials and Methods). We first compared the local recombination rate for TR and FL elements (Table 4). We found that FL elements are on average in genomic regions with a lower recombination rate than TR elements and that this difference is statistically significant (Table 4). This difference is observed for all five families and using both estimators of the recombination rate, although some comparisons are not statistically significant (Table 4).

We found that the length of the elements and the local recombination rate were negatively correlated and that the correlation is statistically significant (length over RR: $r=-2.41\times10^{-5}$, $p<2\times10^{-16}$; length over UCSCRR: $r=-1.95\times10^{-5}$, $p<2\times10^{-16}$). This suggests that long elements tend to reside in regions of lower recombination than short elements. However, this correlation only indicates a general trend and does not adequately capture subtleties of the relationship between element length and recombination rate. Therefore we compared the mean recombination rate between each of the length subgroups by analysis of variance. We found that the different length subgroups differ significantly in their mean recombination rate (RR: $F=7.54$, $p<0.0001$; UCSCRR: $F=41.09$, $p<0.0001$). To determine which differences among subgroups are responsible for the ANOVA results we used Tukey's Honest Significant Differences (HSD) test for multiple comparisons. Fig. 2 shows the result of Tukey's HSD test on the difference between the mean recombination rates of the subgroups. For the analysis using the RR estimate, the only difference between two consecutive subgroups that approached statistical significance occurs between length subgroups 2 (491–1152 bp) and 3 (1153–2498 bp) and this explains all other significant difference between non-consecutive length subgroups. For the UCSCRR data set, two significant comparisons between consecutive subgroups are observed between length categories 1 (100–490 bp) and 2 (491–1152 bp) and categories 2 (491–1152 bp) and 3 (1153–2498 bp). Therefore, this multiple comparison analysis shows that the most significant difference in local recombination rate takes place between elements smaller and larger than 1.2 Kb.

4. Discussion

We performed a genome-wide analysis of the distribution of L1 elements relative to the local recombination rate and the length and family of the elements. Our analysis shows a strong tendency for long elements to accumulate in non- and low-recombining regions of the genome. This bias is not limited to FL elements but is also observed for long TR elements (>1.2 Kb). A similar distribution bias has been reported in recombination hotspots where long L1 elements are severely under-represented (Myers et al., 2005). Because short and long elements are generated by the same mechanism and because truncation of elements occurs at the time of insertion (Martin et al., 2005), the contrasted distribution of short and long TR elements most likely result from some post-insertional mechanism. As the fraction of deletion-free FL elements is similar across the entire genome for a given family, we can exclude a role of DNA deletions in the pattern of distribution of short versus long elements. Therefore, the difference in the distribution of short and long elements results from different rates of fixation of these elements. As short and long elements are equally affected by genetic drift, we conclude that long elements have been selected out of recombining regions of the genome because of their deleterious effect. We also found that the selection against long TR and FL elements seems more pronounced for larger families of elements (L1PA3, L1PA4, L1PA5) than for the relatively small families (L1PA2 and L1PA6). Therefore, selection against L1 acts in a length-dependent and family (i.e., copy number)-dependent manner.

Among the three possible models for selection against L1 retrotransposons (i.e., gene inactivation, ectopic recombination, and retrotransposition process), ectopic recombination between homologous sequences is the one that best explains the above observations. The chance that an ectopic recombination event will occur depends on the number of homologous sequences in the genome (i.e., the size of the family) (Charlesworth and Langley, 1989; Charlesworth et al., 1994; Pasyukova et al., 2004) and the length of the elements (Hasty et al., 1991; Cooper et al., 1998). Therefore, the intensity of selection against the deleterious effect of ectopic recombination should be positively correlated with both the copy number of L1 families and the length of the elements (Petrov et al., 2003). As our results are consistent with both predictions, we suggest that negative selection against L1 elements is principally due to their ability to mediate ectopic recombination. In addition, our results are consistent with experimental results on ectopic homologous recombination in mammals. We found that L1 elements >1.2 Kb are more likely subject to negative selection than those <1.2 Kb. Cooper et al. (1998) showed that ectopic recombination occurs more frequently between sequences containing 2.5 Kb of homology versus 1.2 Kb, and was not detectable between sequences of 1 Kb or less. L1-mediated ectopic recombination is known to occur in humans and some of these events are responsible for disease-causing genetic rearrangements (Burwinkel and Kilimann, 1998; Segal et al., 1999). However, the elements involved in these recombination events were old (i.e., fixed) and, although such events can be deleterious, they have no bearing on the overall distribution of L1 elements relative to recombination rate because selection can act only against those elements that are polymorphic in the population. In addition, ectopic recombination events seem

Please cite this article as: Song, M., Boissinot, S. Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination. Gene (2006), doi:10.1016/j.gene.2006.09.033

relatively rare as a comparison of the human and chimpanzee genomes identified only 26 and 48 deletions involving adjacent L1 copies, respectively (C.S.A.C., 2005) and, to our knowledge, no case of ectopic recombination involving polymorphic L1 elements has been reported in humans. The apparent rarity of ectopic recombination events in modern humans could result from the low activity of L1 in recent human history and from the small size of currently active human L1 sub-families (Boissinot et al., 2000; Khan et al., 2006). In addition, ectopic recombination events between non-adjacent L1 elements could produce chromosomal rearrangements and large DNA deletions that are so deleterious that they would rarely (or never) be observed in natural populations.

Although the ectopic recombination model is consistent with our data, two other models of selection may play a role in the distribution of L1 elements. First, all size classes of L1 elements could potentially alter the function of genes and a number of *de novo* disease-causing L1 inserts have been described (for a review, see Ostertag and Kazazian, 2001). It is plausible that longer elements might be more disruptive to gene function, for instance, by introducing more transcription termination signals (Perepelitsa-Belancio and Deininger, 2003) or by reducing the amount of transcript produced (Han et al., 2004). However, several observations suggest that insertion into genic regions is unlikely to contribute significantly to the accumulation of long TR and FL elements in low- and non-recombining regions. First, the length distribution of autosomal L1 elements located outside of genes shows the same deficit in FL and long TR elements when compared with the length distribution of Y-linked elements than the distribution including all autosomal elements (Fig. 1). Second, the proportion of long TR and FL elements on autosomes is not related to the density of genes as indicated by the fact that the fraction of FL elements is remarkably similar on all autosomes (Table 2) and shows no significant correlation with the gene density ($p > 0.1$ for all families). For instance, the fraction of FL L1PA4 elements on chromosome 4 (15%), which is gene-poor (~8.2 genes/Mb), is similar to the fraction of FL elements on chromosome 19 (15.7%) which is gene-rich (~44.3 genes/Mb) (Table 2). Additionally, the deleterious effect that an element could have on gene function should be independent of the L1 family and the gene activation model does not predict that elements belonging to higher copy number families should be more deleterious than elements from smaller families.

The validity of the second model, selection against the direct cost of the retrotransposition process, is more difficult to assess because FL elements are capable of both producing the RNAs and proteins necessary for retrotransposition and efficiently mediating ectopic recombination. However, the effect of selection against FL elements does not differ significantly from selection against long TR elements and a cost of retrotransposition need not be invoked to explain selection against FL elements. In addition, a deleterious effect of the retrotransposition process would not account for purifying selection against long TR elements which are incapable of producing the biochemical machinery (RNA and proteins) necessary for retrotransposition. This does not mean that the L1 retrotransposition process is not deleterious for its host. In fact, it has recently been demonstrated that L1 activity causes a large number of DNA double-strand breaks that could be severely deleterious (Gasior et al., 2006).

If long L1 elements are subject to purifying selection, we expect polymorphic long L1 inserts to be found at lower frequency in modern populations than short L1 elements. The frequency distribution of inserts belonging to the currently active Ta1 subfamily was recently examined in human (Boissinot et al., 2006). FL Ta1 elements were found at lower frequency in human populations than TR elements suggesting that FL Ta1-containing alleles are subject to purifying selection. However, the number of polymorphic long (>1.2 Kb) TR Ta1 elements analyzed was very small and it was not possible to determine if these elements are also subject to negative selection. The fact that the Ta1 subfamily imposes a detectable fitness cost on its host, despite its low copy number (<400 copies), suggests that a few hundred copies might constitute the threshold at which an L1 family becomes deleterious. It is likely that long L1 elements were subjected to a far stronger purifying selection when L1 activity was higher (i.e. during the amplification of the L1PA5 to L1PA3 families) and that the genetic load imposed by L1 families was much stronger than it has been in recent human history.

Acknowledgments

The authors are acknowledging the support of the CUNY Institute for Software Design and Development. We thank two anonymous reviewers for their helpful suggestions.

References

Bartolome, C., Maside, X., Charlesworth, B., 2002. On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*. Mol. Biol. Evol. 19, 926–937.

Biemont, C., Tsitrone, A., Vieira, C., Hoogland, C., 1997. Transposable element distribution in *Drosophila*. Genetics 147, 1997–1999.

Boissinot, S., Chevret, P., Furano, A.V., 2000. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Mol. Biol. Evol. 17, 915–928.

Boissinot, S., Entezam, A., Furano, A.V., 2001. Selection against deleterious LINE-1-containing loci in the human lineage. Mol. Biol. Evol. 18, 926–935.

Boissinot, S., Entezam, A., Young, L., Munson, P.J., Furano, A.V., 2004. The insertional history of an active family of L1 retrotransposons in humans. Genome Res. 14, 1221–1231.

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

Burwinkel, B., Kilimann, M.W., 1998. Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease. J. Mol. Biol. 277, 513–517.

Charlesworth, B., Charlesworth, D., 1983. The population dynamics of transposable elements. Genet. Res. 42, 1–27.

Charlesworth, B., Langley, C.H., 1989. The population genetics of *Drosophila* transposable elements. Annu. Rev. Genet. 23, 251–287.

Charlesworth, B., Lapid, A., Canada, D., 1992a. The distribution of transposable elements within and between chromosomes in a population of *Drosophila melanogaster*. I. Element frequencies and distribution. Genet. Res. 60, 103–114.

Charlesworth, B., Lapid, A., Canada, D., 1992b. The distribution of transposable elements within and between chromosomes in a population

ARTICLE IN PRESS

of *Drosophila melanogaster*. II. Inferences on the nature of selection against elements. Genet. Res. 60, 115–130.
Charlesworth, B., Sniegowski, P.D., Stephan, W., 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371, 215–220.
Charlesworth, B., Langley, C.H., Sniegowski, P.D., 1997. Transposable element distributions in *Drosophila*. Genetics 147, 1993–1995.
C.S.A.C., 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437, 69–87.
Cooper, D.M., Schimenti, K.J., Schimenti, J.C., 1998. Factors affecting ectopic gene conversion in mice. Mamm. Genome 9, 355–360.
Dasilva, C., et al., 2002. Remarkable compartmentalization of transposable elements and pseudogenes in the heterochromatin of the *Tetraodon nigroviridis* genome. Proc. Natl. Acad. Sci. U. S. A. 99, 13636–13641.
Felsenstein, J., 1974. The evolutionary advantage of recombination. Genetics 78, 737–756.
Finnegan, D.J., 1992. Transposable elements. Curr. Opin. Genet. Dev. 2, 861–867.
Fullerton, S.M., Bernardo Carvalho, A., Clark, A.G., 2001. Local rate of recombination are positively correlated with GC content in the human genome. Mol. Biol. Evol. 18, 1139–1142.
Furano, A.V., Duvernell, D., Boissinot, S., 2004. L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish. Trends Genet. 20, 9–14.
Gasior, S.L., Wakeman, T.P., Xu, B., Deininger, P.L., 2006. The Human LINE-1 retrotransposon creates DNA double-strand breaks. J. Mol. Biol. 357, 1383–1393.
Han, J.S., Boeke, J.D., 2005. LINE-1 retrotransposons: Modulators of quantity and quality of mammalian gene expression? BioEssays 27, 775–784.
Han, J.S., Szak, S.T., Boeke, J.D., 2004. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature 429, 268–274.
Hasty, P., Rivera-Perez, J., Bradley, A., 1991. The length of homology required for gene targeting in embryonic stem cells. Mol. Cell. Biol. 11, 5586–5591.
Hill, W.G., Robertson, A., 1966. The effect of linkage on the limit to artificial selection. Genet. Res. 8, 269–294.
Hoogland, C., Biemont, C., 1996. Chromosomal distribution of transposable elements in *Drosophila melanogaster*: test of the ectopic recombination model for maintenance of insertion site number. Genetics 144, 197–204.
Kazazian, H.H., 2004. Mobile elements: drivers of genome evolution. Science 303, 1626–1632.
Khan, H., Smit, A., Boissinot, S., 2006. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res. 16, 78–87.
Kong, A., et al., 2002. A high-resolution recombination map of the human genome. Nat. Genet. 31, 241–247.
Lander, E.S., et al., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.
Langley, C.H., Montgomery, E., Hudson, R., Kaplan, N., Charlesworth, B., 1988. On the role of unequal exchange in the containment of transposable element copy number. Genet. Res. 52, 223–235.
Martin, S.L., Li, W.-H.P., Furano, A.V., Boissinot, S., 2005. The structures of mouse and human L1 elements reflect their insertion mechanism. Cytogenet. Genome Res. 110, 223–228.
Myers, J.S., et al., 2002. A comprehensive analysis of recently integrated human Ta L1 elements. Am. J. Hum. Genet. 71, 312–326.
Myers, S., Bottolo, L., Freeman, C., McVean, G., Donnelly, P., 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science 310, 321–324.
Neafsey, D.E., Blumenstiel, J.P., Hartl, D.L., 2004. Different regulatory mechanisms underlie similar transposable element profiles in pufferfish and fruitflies. Mol. Biol. Evol. 21, 2310–2318.
Nuzhdin, S.V., 1999. Sure facts, speculations, and open questions about the evolution of transposable elements copy number. Genetica 107, 129–137.
Ostertag, E.M., Kazazian Jr., H.H., 2001. Biology of mammalian L1 retrotransposons. Annu. Rev. Genet. 35, 501–538.
Pasyukova, E.G., Nuzhdin, S.V., Morozova, T.V., Mackay, T.F., 2004. Accumulation of transposable elements in the genome of *Drosophila melanogaster* is associated with a decrease in fitness. J. Heredity 95, 284–290.
Perepelitsa-Belancio, V., Deininger, P.L., 2003. RNA truncation by premature polyadenylation attenuates human mobile element activity. Nat. Genet. 35, 363–366.
Petrov, D., Aminetzach, Y.T., Davis, J.C., Bensasson, D., Hirsh, A.E., 2003. Size matters: non-LTR retrotransposable elements and ectopic recombination in Drosophila. Mol. Biol. Evol. 20, 880–892.
Segal, Y., et al., 1999. LINE-1 elements at the sites of molecular rearrangements in Alport syndrome-diffuse leiomyomatosis. Am. J. Hum. Genet. 64, 62–69.
Song, M., Haralick, R.M., 2002. Optimally quantized and smoothed histograms. Proceedings of the Joint Conference of Information Sciences, Durham, NC, 2002, pp. 894–897.
Song, M., Boissinot, S., Haralick, R.M., Phillips, I.T., 2003. Estimating recombination rate distribution by optimal quantization. Proceesings of IEEE Computational Systems Bioinformatics Conference, 2003, pp. 403–406.
R Development Core Team, 2004. R: A language and environment for statistical computing. In: computing, R.F.f.s. (Ed.). Austria, Vienna, 2004.
Waterston, R.H., et al., 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562.

Please cite this article as: Song, M., Boissinot, S. Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination. Gene (2006), doi:10.1016/j.gene.2006.09.033