

The Distribution of L1 and Alu Retroelements in Relation to GC Content on Human Sex Chromosomes Is Consistent with the Ectopic Recombination Model

György Abrusán, Hans-Jürgen Krambeck

Max Planck Institute of Limnology, Department of Ecophysiology, August Thienemann Str. 2, 24306 Plön, Germany

Received: 15 November 2005 / Accepted: 30 May 2006 [Reviewing Editor: Dr. Deborah Charlesworth]

Abstract. The distribution of Alu and L1 retroelements in the human genome changes with their age. Active retroelements target AT-rich regions, but their frequency increases in GC- and gene-rich regions of the genome with increasing age of the insertions. Currently there is no consensus on the mechanism generating this pattern. In this paper we test the hypothesis that selection against deleterious deletions caused by ectopic recombination between repeats is the main cause of the inhomogeneous distribution of L1s and Alus, by means of a detailed analysis of the GC distribution of the repeats on the sex chromosomes. We show that (1) unlike on the autosomes and X chromosome, L1s do not accumulate on the Y chromosome in GC-rich regions, whereas Alus accumulate there to a minor extent; (2) on the Y chromosome Alu and L1 densities are positively correlated, unlike the negative correlation on other chromosomes; and (3) in gene-poor regions of chromosome 4 and X, the distribution of Alus and L1s does not shift toward GC-rich regions. In addition, we show that although local GC content of long L1 insertions is lower than average, their selective loss from recombining chromosomes is not the main cause of the enrichment of ancient L1s in GC-rich regions. The results support the hypothesis that ectopic recombination causes the shift of Alu and L1 distributions toward the gene-rich regions of the genome.

Key words: Unequal exchange — Y chromosome degeneration — Non-LTR retroelements

Introduction

Alu and L1 retrotransposons are the most abundant transposable elements (TEs) in our genome, with approximately 1.1 and 0.5 million copies respectively (IHGSC 2001) and, together, contribute ~30% of our genome sequence. The majority of both Alu and L1 insertions are nonfunctional “fossils” that either were already incapable of replication at the time of insertion or became inactive later due to mutations, insertions, or deletions. Only 80–100 L1 copies are active at present (Brouha et al. 2003), and the number of active Alu master sequences is approximately 3%–15% of the new insertions (Cordaux et al. 2004).

The two groups of elements rely on different strategies for survival. L1 elements are autonomous and encode proteins with chaperone (ORF1p), endonuclease, and reverse transcriptase activity (ORF2p), which enables them to replicate and insert independently of host functions. A full-length L1 is ~6500 bp long, but most copies are truncated at their 5' ends, and the mean insertion length of L1s is only 900 bp (IHGSC 2001). The currently active primate specific L1s target gene-poor, AT-rich parts of the genome, which can be explained by the cleavage site of the L1 endonuclease: TTTT/A (Feng et al. 1996), but insertions of old, inactive L1 families are present at similar frequencies across the genome regardless of local GC frequency (IHGSC 2001; Yang et al. 2004).

Correspondence to: György Abrusán;
email: abrusan@mpil-ploen.mpg.de

In contrast to L1s, Alus are short (~300-bp) sequences which, unlike L1s, do not encode their own proteins but parasitise L1s' (Dewannieux et al. 2003; Jurka 1997; Smit et al. 1995). Their replication and insertion depend on the ORF2 protein of L1, and it has been shown that young Alus also target AT-rich parts of the genome and accumulate (more rapidly than L1s) in regions with a high GC content (Gu et al. 2000; IHGSC 2001; Pavlicek et al. 2001; Yang et al. 2004). The consequence of this relatively rapid accumulation is that the chromosomal distribution of Alus and L1s is different, Alus being present mainly in GC-rich regions of the genome (IHGSC 2001; Soriano et al. 1983).

Currently, there is no consensus on the mechanism responsible for the temporal enrichment of Alus and L1s in the GC-rich regions. Gu and colleagues (2000) have argued that the different insertion patterns of Alus and LINEs might be caused directly by Alu/LINE interactions. These authors proposed that Alus could switch their insertion preference, to avoid competition with LINEs for the ORF2 protein. This hypothesis does not require selection to act on the insertions, but recent findings (Hackenberg et al. 2005), and also our results (see below), do not support it. Since GC-rich regions are also gene-rich, it has been proposed that the accumulation of Alus in these regions may be adaptive (IHGSC 2001). However, Brookfield (2001) pointed out that the accumulation of Alus in GC-rich regions is still slower than the time necessary for the fixation of neutral alleles; positive selection is therefore unlikely to be the cause. Interestingly, in rodents even the youngest SINE insertions show strong GC preference, Yang et al. 2004. Bailey et al. (2003) proposed that duplications might contribute to the enrichment of Alus, because duplications occur more frequently in Alu- and GC-rich regions of the genome than elsewhere. However, the frequency of duplications in the human genome is not enough to explain the strong pattern observed (Jurka et al. 2004). Recently Belle et al. (2005) demonstrated that degradation of Alus by short indels in GC-poor regions is also not the cause of selective Alu loss.

Several authors have proposed that the enrichment of repeats in GC-rich regions is the consequence of illegitimate recombination between the repeats (Batzler and Deininger 2002; Brookfield 2001; Hackenberg et al. 2005; IHGSC 2001; Lobachev et al. 2000; Medstrand et al. 2002; Stenger et al. 2001), which can cause deletions and duplications (Fig. 1). Deletions are likely to be more deleterious in gene (and GC-) rich regions than in gene-poor (and AT-rich) regions, because they may delete entire genes or exons. This mechanism can thus eliminate repeats from AT-rich (gene-poor) regions without as many harmful effects, thus leaving a higher abundance in GC-rich regions.

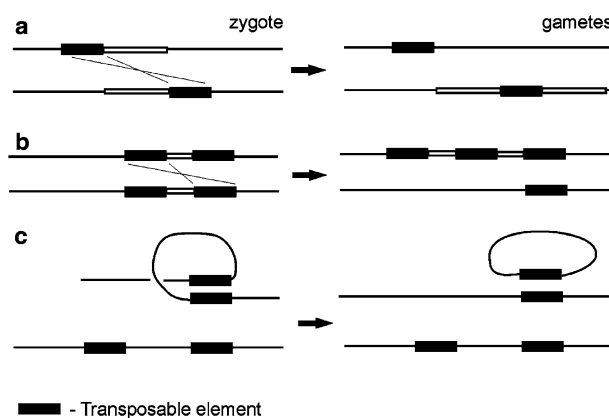


Fig. 1. Chromosomal rearrangements caused by repeats. Open bars mark nonrepetitive regions that undergo duplication. **a** Ectopic recombination between heterozygous repeats leads to a deletion on one chromosome and duplication on the other; either is likely to be deleterious in gene-rich regions. This process is likely to be prevalent in organisms with a low TE abundance but high TE polymorphism, like *Drosophila* (Langley et al. 1988). **b** In mammals most TEs are fixed, and recombination between homozygous TEs probably has minor consequences on fitness, but recombination between nonhomologous repeats leads to duplications and deletions as well. Bailey et al. (2003) have proposed that Alu-mediated duplications could cause the enrichment of Alus in gene-rich regions, because duplications are more frequent in GC-rich regions. However, recently Jurka et al. (2004) concluded that the amount of duplications in the genome is not high enough to explain the GC shift of Alus. **c** Intrachromosomal exchange between repeats. This process removes one repeat with a fragment of surrounding DNA. Since deletions are more likely to be deleterious in gene-rich regions where they can remove exons, the process predicts faster loss of repeats from gene-poor, AT-rich regions of the genome than from GC-rich regions.

It has been shown (MGSC 2002) that in the mouse and human genomes, LINEs and SINEs are found in different chromosomal locations, but human and mouse specific LINE (and also SINE) insertions accumulate in orthologous segments of the genomes; repeat densities in one species are actually correlated more strongly with the density in the other species in the same region than with the local GC content of the insertions. This finding led the authors to the conclusion that GC content is not the direct cause of the different distribution of LINEs and SINEs but is merely correlated with the true cause. A relationship between recombination rate and GC content was observed more than a decade ago (Eyre-Walker 1993). More recently Meunier and Duret (2004) argued that the regional variation of GC content is driven by recombination.

Chromosomal rearrangements caused by repeats occur primarily during meiosis, at least in *Drosophila* (Montgomery et al. 1991). Therefore, ectopic recombination between repeats predicts that, in regions with low recombination rates or with a very low density of selectively important sequences, repeats will not accumulate in GC-rich regions over time. We test this hypothesis using gene-poor regions from human chromosomes 4, X, and Y. At present the Y chro-

mosome experiences no meiotic recombination (with the exception of pseudoautosomal regions), and Medstrand et al. (2002) have already noticed a delay in the accumulation of Alus on the Y. The Y chromosome evolved from an X-like ancestor and originally paired with the ancestor of the X chromosome (reviewed by Charlesworth et al. 2005). The cessation of meiotic recombination between the two chromosomes was gradual and involved several steps leading to “evolutionary strata” on the sex chromosomes, the oldest formed 240–320 million years ago (mya), while the most recent ones formed 30–50 mya (Lahn and Page 1999; Ross et al. 2005; Skaletsky et al. 2003). However, since some Alu and, particularly, L1 families are older than the youngest “evolutionary strata,” at the time of their insertion they could also have inserted into parts of the chromosome which were still recombining. In addition, 10.2 Mb of the Y chromosome sequence was acquired from several autosomes by transposition events in the last 300 million years (Skaletsky et al. 2003), which certainly moved repeats to the Y chromosome. For these reasons we expected reduced enrichment of TEs in GC-rich regions of the Y chromosome, rather than their complete absence.

In this paper we address the following questions.

1. We test whether the change of GC distribution of non-LTR retroelements follows the same temporal pattern on the X chromosome and the male-specific part of the Y chromosome as on the autosomes.
2. Since Alus and L1s use the same protein (the ORF2p of L1s) for replication and insertion, we predict that, in the absence of selection against deleterious deletions, their abundances on the Y chromosome should be positively correlated. We test whether Alus and LINEs accumulate at the same chromosomal locations on the Y and compare the results with those from other chromosomes.
3. If ectopic recombination is the cause of the GC shift of repeats, then the important factors are its frequency and the gene density of the chromosome, which determines the magnitude of deleterious effects of deletions, and both factors are weaker on the Y. They can be separated, comparing regions of the genome that, unlike the Y, do experience meiotic recombination but have a very low gene density, with regions of high gene density but low recombination rate. We test whether the distribution of Alus and L1s shows a shift toward GC-rich regions, using genomic regions which are extremely gene-poor but recombine (Myers et al. 2005), one on chromosome 4 (26.7–38.5 Mb) and one on the X chromosome (86.8–95.7 Mb), and a fragment of the Y chromosome (19–26.7 Mb) which has a higher than average density of genes.
4. It has been shown that mammalian sex chromosomes contain a higher abundance of TEs than autosomes (Baker and Wichman 1990; MGSC 2002) and the proportion of full-length and long L1 insertions is also higher (Boissinot and Furano 2001; Erlandsson et al. 2000; MGSC 2002). In addition, insertions of different lengths are present in regions of different GC content, with AT-rich regions of the genome containing more long L1 insertions (Medstrand et al. 2002; MGSC 2002). Since ectopic recombination is more likely to occur between long insertions than short ones (Hasty et al. 1991; Petrov et al. 2003), the loss of long insertions due to recombination results in an enrichment of LINEs in GC-rich regions, relative to AT-rich regions. We test whether the selective loss of long L1 insertions is a significant factor in the accumulation of old repeats in GC-rich regions.

Methods

The sequence and repeat annotation files (RepeatMasker) of the human genome (hg17; May 2004 assembly) were downloaded from the UCSC genome browser site at <http://genome.ucsc.edu> (Karolchik et al. 2003). The pseudoautosomal regions (PARs) and the X transposed regions (XTRs) were excluded from the sequence of the Y chromosome. Both Alus and LINEs were grouped into cohorts according to their consensus sequences (Jurka and Milosavljevic 1991; Jurka and Smith 1988; Smit et al. 1995). The cohorts were active during different periods of mammalian evolution (IHGSC 2001): AluY (currently active, 0–30 mya); AluS (30–60 mya); AluJ (60–100 mya); L1PA (currently active, 0–65 mya); L1PB (50–80 mya); L1MA (50–100 mya); L1MB (100–150 mya); and L1MC, L1MD, and L1ME (80–150 mya). With the exception of the youngest L1PA and AluY cohorts, in humans all Alu and L1 insertions are “fossils,” and there is no evidence of their activity for millions of years; L1P* cohorts are present only in primates, while L1M* cohorts can be found in most mammals. Analyses of the age, phylogeny, insertion preference, and copy numbers of L1 cohorts can be found in IHGSC (2001), Ohshima et al. (2003), and Smit et al. (1995).

The insertion preference of repeats was calculated using the method of Yang et al. (2004). This takes into account the differences in absolute repeat densities and GC distribution between the chromosomes. First, the frequency distribution of the GC content of the chromosomes (autosomes, X, Y) was calculated by dividing their sequences into nonoverlapping 30-kb windows. The frequency of the G and C nucleotides (GC_{chr}) was calculated in each window. Repetitive sequences were excluded from the windows; consequently the average nucleotide count in the windows was 15 kb. Next, the local GC content of the sequences adjoining individual repeat insertions (GC_{rep}) was calculated. For each insertion, the frequency of G and C nucleotides was counted in adjoining 15-kb windows up- and downstream of the insertions (before the first and after the last position of the repeat), excluding repetitive sequences, to ensure the independence of GC_{rep} from GC_{chr} and local repeat density. Fragmented repeats were counted as one insertion. When calculating the frequency distributions of GC_{chr} and GC_{rep} , bins of 2% were used for the X chromosome, and bins of 3% for the Y chromosome, due to its lower abundance of repeats. The insertion preference of the repeat cohorts is represented by relative

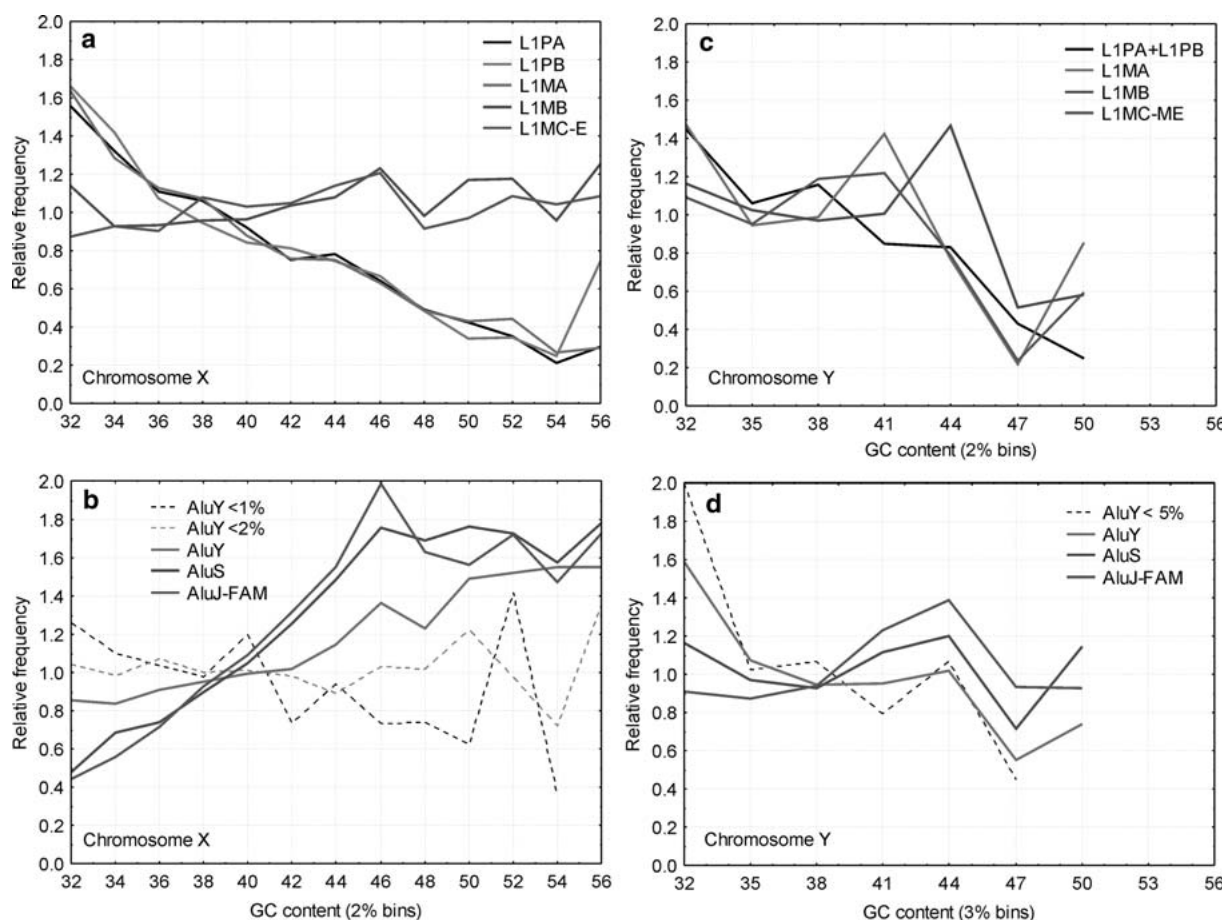


Fig. 2. Frequency of occurrences of different LINE and SINE families in regions of different GC content. The pattern on the X chromosome is similar to the pattern observed on autosomes (Supplementary Fig. 1; IHGSC 2001); at the time of insertion Alus show a similar GC preference to LINES, but accumulate rapidly in GC-rich regions of the genome, while the accumulation of LINES in GC-rich regions characterizes only the older clades. On the

frequencies: the frequency of GC_{rep} in a bin divided by the frequency of GC_{chr} in the same bin. For statistical analyses of Alu distributions, a modified version of the method outlined above was used: the GC_{rep} of each repeat was divided by the median of the GC_{chr} distribution of the sequence analyzed. The resulting distribution was dichotomized; 1 was assigned for each repeat with GC_{rep} equal to or larger than the median, and 0 for repeats with smaller GC_{rep} . The proportions of repeats with a local GC content larger or equal to the median GC_{chr} on each chromosome or genomic region were compared with Tukey tests (Zar 2004). This method allows statistical comparison of the shifts in repeat distributions when the underlying GC distributions of the sequences compared are very different.

Repeat densities on the chromosomes were calculated by dividing the chromosomes into nonoverlapping fragments of 200 kb. In each window, the numbers of Alu, L1PA, and L1PB insertions were counted (the primate specific L1 cohorts); fragmented repeats were counted only once. Repeats lying on the border, or having fragments in two windows, were counted in the window that contains their first nucleotide. Besides the sex chromosomes, the densities of the repeats were also calculated on two autosomes, chromosome 7 and chromosome 21, which have a size and GC content similar to those of the sex chromosomes (chr7-X, chr 21-Y [ICGSC 2004]). The length of L1 insertions was calculated as in

Y chromosome LINES show no accumulation in GC-rich regions, and Alus show only very weak enrichment in GC-rich regions. Dotted lines show the distribution of the youngest Alu insertions of the currently active AluY family, which diverged from their consensus by 1%, 2%, and 5%. Divergence was calculated as in MGSC (2002). On the Y chromosome the higher, 5% cutoff is necessary due to the low abundance of Alus.

ICGSC (2004), as the difference between the first and the last position on the matching consensus sequence.

Results

On the X chromosome and autosomes, both Alus and LIs increase in frequency in GC-rich regions over time (Figs. 2a and b, Supplementary Fig. 1). In contrast to the X chromosome, old LINES on the Y chromosome (L1MB–L1ME cohorts) are present at similar frequencies in AT-rich regions as the currently active L1PA cohort (Fig. 2c). Compared to LINES, the frequency of Alus increases in GC-rich regions over time (Figs. 2d and 3), but the GC shift is delayed (Fig. 2d) and weaker than on the autosomes and X chromosome (Fig. 3). No significant change in repeat distributions was observed in regions with low gene densities on chromosomes 4 and X (Figs. 3 and 7).

Alu and LINE densities are negatively correlated on autosomes and the X chromosome (Spearman

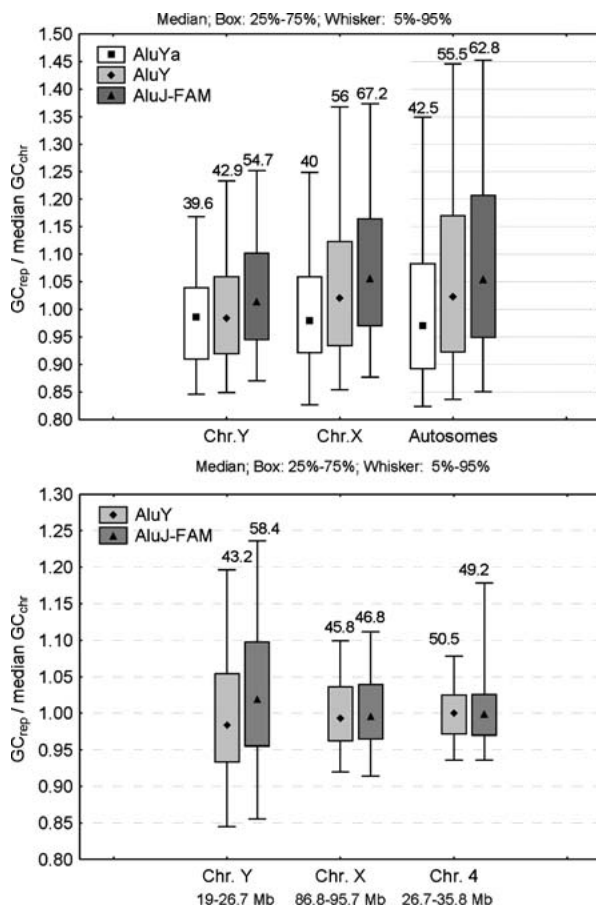


Fig. 3. The shift of Alu cohorts toward GC-rich regions on the analyzed chromosomes and chromosomal regions. The numbers above the whiskers give the frequency (%) of Alus having a local GC content higher than the median of GC_{chr} . The shift toward GC region is highly significant on every chromosome (Tukey's test for proportions, $p < 0.001$), with the exception of the gene-poor regions of chromosomes 4 and X ($p > 0.5$ for both; $n_{chr4} = 724$, $n_{chrX} = 640$). The Y chromosome is significantly different from both the X chromosome and the autosomes ($p < 0.001$).

rank correlations; see legend to Fig. 4), while on the Y chromosome they are positively correlated (Fig. 4a). The strongest negative relationship between Alu density and LINE density is on the X chromosome, where LINES are particularly abundant (Bailey et al. 2000; Ross et al. 2005) and are thought to be involved in its inactivation (Bailey et al. 2000; Chow et al. 2005; Lyon 1998; but see Ke and Collins 2003). On the two autosomes the trend is similar to the trend observed on the X chromosome, but less pronounced (Fig. 4b).

The analysis of the length distribution of L1s in the finished portions of the sex chromosomes (PARs and the X-transposed region excluded from the Y chromosome) shows a qualitatively similar enrichment of full-length L1PA (the currently active LINE cohort) elements on the sex chromosomes (Fig. 5) to that reported by Boissinot et al. (2001). The frequency of full-length (5600- to 6400-bp-long) L1PA insertions is 1.7–2 times higher on the

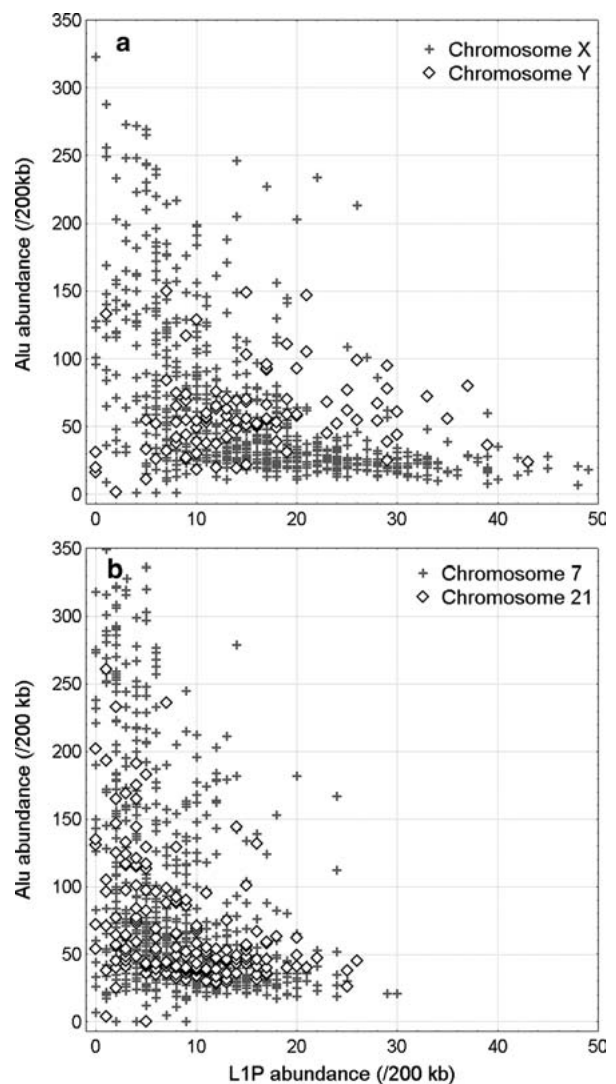


Fig. 4. The abundances of primate-specific L1s (L1PA–L1PB cohorts) and Alu elements in 200-kb nonoverlapping windows. Alu and L1P abundances are strongly negatively correlated on the autosomes and the X chromosome, where the trend is the strongest, but are positively correlated on the Y chromosome. Spearman's rank correlations: Y chromosome, $R = +0.25$, $n = 100$, $p = 0.010$; X chromosome, $R = -0.61$, $n = 761$, $p < 0.001$; chromosome 7, $R = -0.45$, $n = 711$, $p < 0.001$; chromosome 21, $R = -0.45$, $n = 150$, $p < 0.001$.

Y chromosome, and 1.46–1.08 times on the X chromosome, compared with autosomes (Fig. 5). The frequency of full-length elements among the youngest L1PA1 (L1HS)–L1PA5 families is 29.3% on the Y chromosome and 17.2% on the X chromosome (not shown).

Although long insertions are much less frequent in old L1 cohorts than in younger ones, their selective disappearance is not the main cause of the shift of old L1 cohorts toward GC-rich regions (Fig. 6): the variance explained by the negative correlation between the length of the insertions and their local GC content is far less important than the length-independent shift toward a high local GC content (see

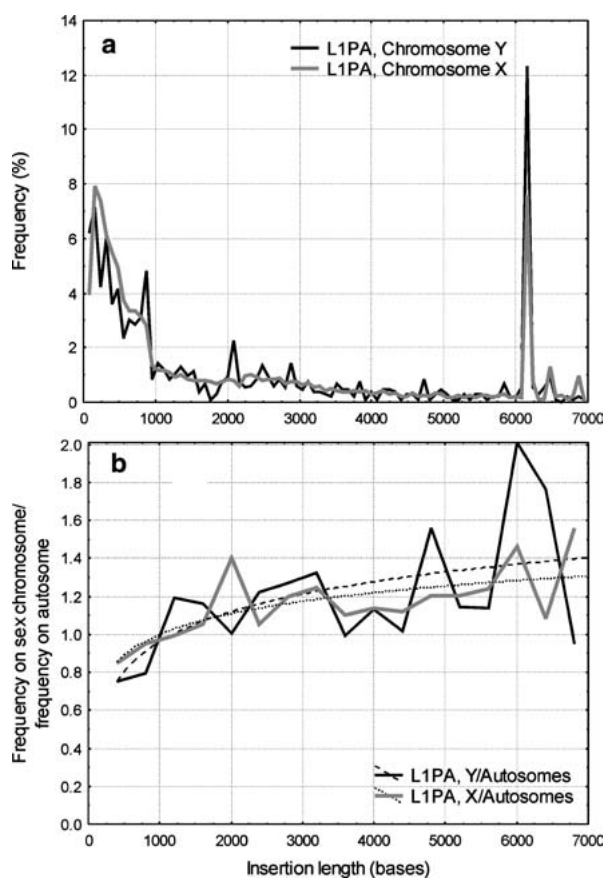


Fig. 5. Higher frequencies of LINES on sex chromosomes. **a** Frequency distribution of insertion sizes of the currently active L1PA cohort on the X and Y chromosomes. The insertions were grouped into bins of increasing insertion size, differing by 80 bases. **b** The frequency of insertions longer than 1000 bp is higher on the sex chromosomes than on autosomes: Mann-Whitney *U*-test ($n = 15$), $p < 0.001$ for both the X and the Y chromosomes. Dotted lines are best-fit (OLS) logarithmic functions. Repeats were grouped into bins differing by 400 bases.

statistics in Fig. 6). However, in the relatively gene-rich (compared to the chromosome average) region of the Y chromosome, long L1 insertions (> 1000 bp) show a weak but significant shift toward regions with a high GC content (Fig. 7).

Discussion

Our results support the hypothesis that ectopic recombination between repeats drives the accumulation of L1s and Alus in GC-rich regions of the genome. This can explain most observed patterns, both on autosomes (and X) and on the Y chromosome. Its implications are: (1) the Y chromosome has the original, approximately unchanged GC distributions of LINES and Alus, which reflects the insertion patterns of the repeats; (2) the AT preference of the recent TE insertions is primarily the result of the target specificity of the ORF2 protein; and (3) the insertion preferences of old LINE (L1MB–L1ME) and Alu

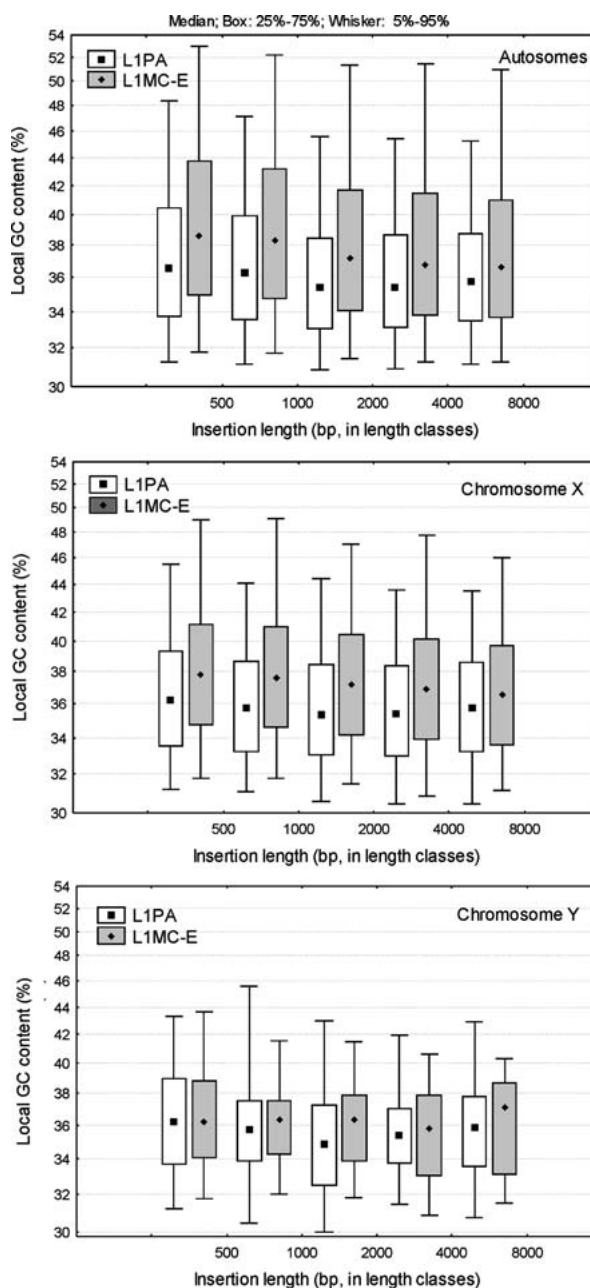


Fig. 6. Relationships between the local GC content and the length of L1 insertions. We used ANCOVA to separate the effect of cohort (categorical predictor) from the effect of insertion length (continuous predictor), but for clarity ($\sim 300,000$ insertions) the data are shown as box-plots. Lengths and local GC contents were log transformed prior to analysis. On autosomes the effect of cohort is almost two times stronger than the effect of length (insertion length, $\beta = -0.09$, $p \ll 0.001$; cohort, $\beta = +0.164$, $p \ll 0.001$; $n = 293,783$); the pattern is similar on the X chromosome (insertion length, $\beta = -0.076$, $p \ll 0.001$; cohort, $\beta = +0.173$, $p \ll 0.001$; $n = 21,224$). On the Y chromosome we found a significant relationship between insertion length and GC preference ($\beta = -0.091$, $p < 0.001$; $n = 1992$) but no significant effect of cohort ($\beta = 0.038$, $p = 0.101$).

(AluS, AluJ) cohorts were similar to those of the currently active cohorts. In agreement with the predictions of ectopic recombination, the gene-poor regions of chromosome 4 and X show no change in Alu

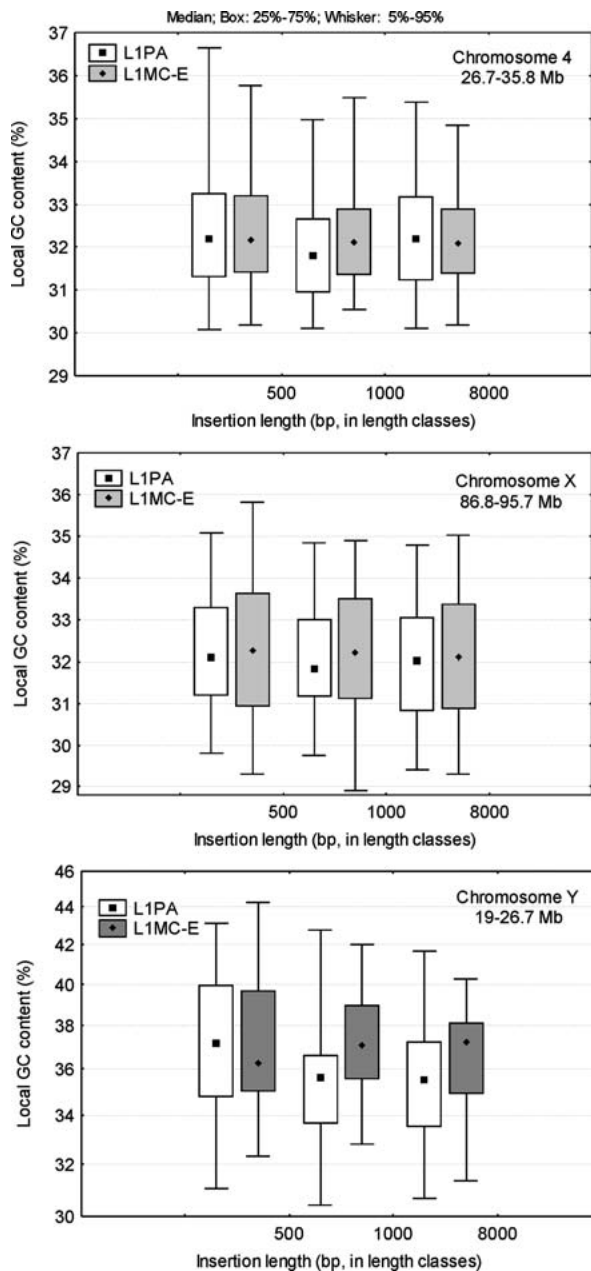


Fig. 7. Relationships between the insertion length of LINES and their local GC content, using chromosomal regions with a low gene density (chromosomes 4 and X) and a relatively gene-dense (compared to the chromosomal average) region from the Y chromosome. We used GLM for statistical analyses. No significant ($p > 0.1$; $n_{\text{chr4}} = 966$, $n_{\text{chrX}} = 1,289$) effects of any factors or their interactions are detected on chromosomes 4 and X, but a weak effect ($p = 0.0131$; $n = 785$) of the interaction between insertion length and cohort is present on the Y chromosome, indicating that the slopes of the regressions of old and young L1 cohorts are different.

and L1 distributions over time, suggesting that the presence of genes is necessary for the GC shift; in their absence recombination alone does not change the distribution of repeats (Figs. 3 and 7). The observation that on the Y chromosome Alus and primate specific L1s accumulate in the same chromosomal regions (Fig. 4) is the expected pattern for

repeats which use the same protein for insertion, and the negative correlation between Alu and L1 abundances on other chromosomes is likely to be the result of subsequent processes (ectopic exchange). The weak enrichment of Alus in GC-rich locations on the Y could be due to the gradual cessation of recombination or even ectopic recombination, because roughly 25% of the euchromatic region of the Y chromosome consists of palindromes, which are older than the human–chimpanzee split and undergo frequent gene conversion (Rozen et al. 2003).

Ectopic recombination occurs more often between long sequences than short ones (Hasty et al. 1991; Petrov et al. 2003; but see Cooper et al. 1998) and this predicts the overrepresentation of long insertions on sex chromosomes. This is clearly the case: the X and, especially, the Y chromosome have higher frequencies of insertions longer than 1000 bp compared with autosomes (Fig. 4). The loss of long L1 insertions is not, however, the main factor that shifts the distribution of old repeats toward GC-rich regions, since on autosomes and the X, all length classes show a comparable shift (Fig. 5). This apparent paradox can be explained by at least two processes: incomplete deletions of repeats may change the length but not necessarily the local GC content of the insertion, and such repeats may be classified among the short insertions; also, the frequency of ectopic recombination is likely to increase with increasing density of repeats, and short L1 repeats (<1000 bp) are much more abundant than long ones on every chromosome. The weak but significant shift of long LINES toward GC-rich regions on the relatively gene-rich (ampliconic) fragment of the Y chromosome (Fig. 7) indicates that recombination between repeats occurs on the Y chromosome as well, but it is either not frequent or not deleterious enough to shift the distribution of the shortest insertions toward GC-rich regions.

An alternative possibility is that the pattern on the autosomes and X chromosome represents the original distribution of repeats and was generated by an unknown process, and it is the Y chromosome that underwent subsequent changes. The Y chromosome has experienced degeneration: it has lost the majority of its genes, and its overall length has decreased to one-third the size of the X chromosome. Its euchromatic region is even smaller, comprising approximately one-half of the chromosome (Skaletsky et al. 2003). The underlying causes of shrinking are not fully understood, but most explanations are related to lack of recombination (reviewed by Charlesworth and Charlesworth 2000). Recombination enables outcrossing parents to produce offspring with fewer deleterious mutations than they have themselves. Therefore in nonrecombining organisms or regions of the genome the accumulation of deleterious muta-

tions is inevitable and can lead to the loss of deleterious and nonfunctional genomic material. The degeneration of the Y chromosome did not affect its entire sequence similarly; GC-rich regions disappeared at a higher rate than AT-rich regions, due to either sequence turnover or deletions, and this has skewed the overall GC distribution of the chromosome compared to X and autosomes (Supplementary Fig. 2). The distribution of L1s shows that most of the oldest cohorts (L1MC, L1MD, L1ME, which were active in the ancestors of primates 80–150 mya) have already disappeared from the Y; the chromosome is dominated by the youngest, primate specific repeats (Supplementary Fig. 3). If repetitive sequences with a high local GC content were deleted on the Y chromosome at a higher rate than euchromatic sequences with the same GC content, it would have led to the depletion of repeats in these regions. Complete deletions leave no signs behind, but partial deletions should result in a more pronounced negative correlation between GC content and insertion length on the Y chromosome than on recombining chromosomes. In the case of LINEs we found no such effect; the standardized slope (β) of the partial regression between insertion length and GC content on the Y chromosome is similar to that of the autosomes (Fig. 6).

Taken together, although the shrinking of the Y chromosome is likely to have influenced the GC distribution of its retroelements, our data do not indicate that this is a major force, while selection against deleterious deletions caused by ectopic recombination explains the observed pattern both on autosomes and sex chromosomes. The two repeat classes behave in a qualitatively similar way; only the speed of the GC shift is different. For heterozygous repeats (Fig. 1a), theory predicts that the likelihood of TEs participating in ectopic recombination events is proportional to the square of the copy numbers of TEs (Langley et al. 1988), and experimental work on *Drosophila* (Montgomery et al. 1991) suggests that ectopic exchange between repeats is most frequent when repeats are heterozygous. It has been proposed (Charlesworth and Charlesworth 1995; Morgan 2001) that in selfing species, where homozygosity is high, TE abundances should be higher due to a reduced frequency of ectopic exchange between repeats. In mammals, similarly to selfing species, the vast majority of L1 and Alu insertions are homozygous (Bennett et al. 2004), in this case because of fixation, and unlike in *Drosophila*, intrachromosomal recombination may be the main force removing TEs from the genome (Fig. 1c). It is unclear how the frequency of intrachromosomal recombination scales with the density of repeats, but again the relationship is likely to be positive and nonlinear. The majority of L1 insertions in the genome predate the appearance of

Alus; only approximately 130,000 L1s date from the time of Alu proliferation. At the peak of their activity (~40 mya) Alu insertions were 25 times more frequent than the coexisting L1s (~180,000 vs. ~7,200 [Abrusan and Krambeck, submitted]). Although a typical LINE insertion is four to five times longer than an Alu, the large difference in their abundances alone may be sufficient to explain the faster shift of Alus toward GC-rich regions.

Acknowledgments. We thank Deborah Charlesworth and both anonymous referees for their constructive reviews, which greatly improved the manuscript, and Thorsten Reusch and Edward Brede for their critical reading of an early draft of this paper. G.A. was supported by a postdoctoral fellowship from the Alexander von Humboldt Foundation and the Max Planck Society.

References

- Bailey JA, Carrel L, Chakravarti A, Eichler EE (2000) Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: The Lyon repeat hypothesis. *Proc Natl Acad Sci USA* 97:6634–6639
- Bailey JA, Liu G, Eichler EE (2003) An Alu transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet* 73:823–834
- Baker RJ, Wichman HA (1990) Retrotransposon Mys is concentrated on the sex-chromosomes—implications for copy number containment. *Evolution* 44:2083–2088
- Batzer MA, Deininger PL (2002) Alu repeats and human genomic diversity. *Nature Rev Genet* 3:370–379
- Belle EMS, Webster MT, Eyre-Walker A (2005) Why are young and old repetitive elements distributed differently in the human genome? *J Mol Evol* 60:290–296
- Bennett EA, Coleman LE, Tsui C, Pittard WS, Devine SE (2004) Natural genetic variation caused by TEs in humans. *Genetics* 168:933–951
- Boissinot SP, Furano AV (2001) Adaptive evolution in LINE-1 retrotransposons. *Mol Biol Evol* 18:2186–2194
- Boissinot S, Entezam A, Furano AV (2001) Selection against deleterious LINE-1-containing loci in the human lineage. *Mol Biol Evol* 18:926–935
- Brookfield JFY (2001) Selection on Alu sequences? *Current Biol* 11:R900–R901
- Brouha B, Schustak J, Badge RM, Lutz-Prigget S, Farley AH, Moran JV, Kazazian HH (2003) Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci USA* 100:5280–5285
- Charlesworth D, Charlesworth B (1995) Transposable elements in inbreeding and outbreeding populations. *Genetics* 140:415–417
- Charlesworth B, Charlesworth D (2000) The degeneration of Y chromosomes. *Philos Trans Roy Soc London Ser B Biol Sci* 355:1563–1572
- Charlesworth D, Charlesworth B, Marais G (2005) Steps in the evolution of heteromorphic sex chromosomes. *Heredity* 95:118–128
- Chow JC, Yen Z, Ziesche SM, Brown CJ (2005) Silencing of the mammalian X chromosome. *Annu Rev Genom Hum Genet* 6:69–92
- Cooper DM, Schimenti KJ, Schimenti JC (1998) Factors affecting ectopic gene conversion in mice. *Mammal Genome* 9:355–360
- Cordaux R, Hedges DJ, Batzer MA (2004) Retrotransposition of Alu elements: how many sources? *Trends Genet* 20:464–467

- Dewannieux M, Esnault C, Heidmann T (2003) LINE-mediated retrotransposition of marked Alu sequences. *Nature Genet* 35:41–48
- Erlandsson R, Wilson JF, Paabo S (2000) Sex chromosomal TE accumulation and male-driven substitutional evolution in humans. *Mol Biol Evol* 17:804–812
- Eyrewalker A (1993) Recombination and mammalian genome evolution. *Proc Roy Soc London Ser B Biol Sci* 252:237–243
- Feng QH, Moran JV, Kazazian HH, Boeke JD (1996) Hum L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* 87:905–916
- Gu ZL, Wang HD, Nekrutenko A, Li WH (2000) Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequence. *Gene* 259:81–88
- Hackenberg M, Bernaola-Galvan P, Carpena P, Oliver JL (2005) The biased distribution of alus in human isochores might be driven by recombination. *J Mol Evol* 60:365–377
- Hasty P, Riveraperez J, Bradley A (1991) The length of homology required for gene targeting in embryonic stem-cells. *Mol Cell Biol* 11:5586–5591
- ICGSC (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432:695–716
- IHGSC (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921
- Jurka J (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retrotransposons. *Proc Natl Acad Sci USA* 94:1872–1877
- Jurka J, Milosavljevic A (1991) Reconstruction and analysis of human Alu genes. *J Mol Evol* 32:105–121
- Jurka J, Smith T (1988) A fundamental division in the Alu family of repeated sequences. *Proc Natl Acad Sci USA* 85:4775–4778
- Jurka J, Kohany O, Pavlicek A, Kapitonov VV, Jurka MV (2004) Duplication, coclustering, and selection of human Alu retrotransposons. *Proc Natl Acad Sci USA* 101:1268–1272
- Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ (2003) The UCSC Genome Browser Database. *Nucleic Acids Res* 31:51–54
- Ke X, Collins A (2003) CpG islands in human X-inactivation. *Ann Hum Genet* 67:242–249
- Lahn BT, Page DC (1999) Four evolutionary strata on the human X chromosome. *Science* 286:964–967
- Langley CH, Montgomery E, Hudson R, Kaplan N, Charlesworth B (1988) On the role of unequal exchange in the containment of TE copy number. *Genet Res* 52:223–235
- Lobachev KS, Stenger JE, Kozyreva OG, Jurka J, Gordenin DA, Resnick MA (2000) Inverted Alu repeats unstable in yeast are excluded from the human genome. *EMBO J* 19:3822–3830
- Lyon MF (1998) X-chromosome inactivation: a repeat hypothesis. *Cytogenet Cell Genet* 80:133–137
- Medstrand P, van de Lagemaat LN, Mager DL (2002) Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res* 12:1483–1495
- Meunier J, Duret L (2004) Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol* 21:984–990
- MGSC (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–562
- Montgomery EA, Huang SM, Langley CH, Judd BH (1991) Chromosome rearrangement by ectopic recombination in *Drosophila melanogaster*—genome structure and evolution. *Genetics* 129:1085–1098
- Morgan MT (2001) Transposable element number in mixed mating populations. *Genet Res* 77:261–275
- 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
- Ohshima K, Hattori M, Yada T, Gojobori T, Sakaki Y, Okada N (2003) Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. *Genome Biol* 4
- Pavlicek A, Jabbari K, Paces J, Paces V, Hejnar J, Bernardi G (2001) Similar integration but different stability of Alus and LINES in the human genome. *Gene* 276:39–45
- Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE (2003) Size matters: non-LTR retroTEs and ectopic recombination in *Drosophila*. *Mol Biol Evol* 20:880–892
- Ross MT, Grafham DV, Coffey AJ et al. (2005) The DNA sequence of the human X chromosome. *Nature* 434:325–337
- Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC (2003) Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* 423:873–876
- Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, Chinwalla A, Delehaunty A, Delehaunty K, Du H, Fewell G, Fulton L, Fulton R, Graves T, Hou SF, Latrielle P, Leonard S, Mardis E, Maupin R, McPherson J, Miner T, Nash W, Nguyen C, Ozersky P, Pepin K, Rock S, Rohlfing T, Scott K, Schultz B, Strong C, Tin-Wollam A, Yang SP, Waterston RH, Wilson RK, Rozen S, Page DC (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423:825
- Smit AFA, Toth G, Riggs AD, Jurka J (1995) Ancestral, mammalian-wide subfamilies of line-1 repetitive sequences. *J Mol Biol* 246:401–417
- Soriano P, Meunier J, Bernardi G (1983) The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes. *Proc Natl Acad Sci USA Biol Sci* 80:1816–1820
- Stenger JE, Lobachev KS, Gordenin D, Darden TA, Jurka J, Resnick MA (2001) Biased distribution of inverted and direct Alus in the human genome: Implications for insertion, exclusion, and genome stability. *Genome Res* 11:12–27
- Yang S, Smit AF, Schwartz S, Chiaromonte F, Roskin KM, Haussler D, Miller W, Hardison RC (2004) Patterns of insertions and their covariation with substitutions in the rat, mouse, and human genomes. *Genome Res* 14:517–527
- Zar JH (2004) *Biostatistical analysis*, 4th ed. Prentice Hall, Upper Saddle River, NJ