International Journal of Bioinformatics Research
ISSN: 0975–3087, E-ISSN: 0975–9115, Vol. 3, Issue 1, 2011, pp-135-141
Available online at http://www.bioinfo.in/contents.php?id=21

INVESTIGATING THE SEX SPECIFIC RATES OF REPLICATION DRIVEN MUTATIONS IN HUMANS USING GENOME-WIDE INDEL MUTATIONS IN HUMAN *ALU* REPEATS

SRIDHAR RAMACHANDRAN*
Department of Informatics, Indiana University Southeast, Indiana, USA
*Corresponding author. E-mail: sriramac@ius.edu

Received: October 13, 2010; Accepted: January 28, 2011

Abstract- Background: To understand the tempo and mode of evolution at the nucleotide level it is important to estimate the spontaneous rate of each mutation type. Many molecular evolutionary studies have concluded that due to the greater number of cell divisions in the male germline than in the female germline, replication-based nucleotide substitutions in primates occur more frequently in males than in females. However, a potential sex bias in mutations other than nucleotide substitutions has not been extensively investigated. The human *Alu* repeats provide an ideal mechanism to further investigate the degree of replication-based indel (insertion and deletion) mutations in the human chromosomes. **Results:** We analyze patterns of small indel mutations (1bp) in the middle poly (A) track of *Alu* repeats across the entire human genome in order to elucidate the processes of mutation and fixation. This analysis adds further support for the accumulation of more mutations in the Y chromosome compared to the X chromosome. We report the male-to-female mutation ratio α in humans as ~1.5. **Conclusion:** Our results suggest that although small indel mutation may be primarily replication driven (as previous studies suggest) the observed value of α does not exceed the threshold necessary to conclude that contributions of replication independent factors are negligible. We also report that, with small indels (1bp) deletions outnumber insertion events. This relative excess of deletions may be an important parameter in the long-term evolution of genomic size.
Keywords– ALU repeats, male-to-female mutation ratio, insertions, deletions, indels

Background

In humans, men have more germ cell divisions than women. The germ-lines are maintained separately from the somatic cells; therefore, the mutations in the gametes can arise only from within the germ cells. If mutations arise primarily from DNA replication errors during germ cell divisions, the mutation rates in males should be higher in males than that in females. Assuming mutations to be the source of genetic variations, a male bias in mutation rates would suggest that evolution is 'male biased'. Even though a number of studies have detected a male-driven evolution in mammals, birds and plants, a precise value of the male-to-female mutation ratio, (α), in humans is incomplete. Knowing the accurate value of human α is critical in understanding whether germline mutations are primarily caused by imperfectly copied DNA during replication or by primarily environmental factors.

With many more rounds of cell division per generation, males accumulate more mutations. In primates, males undergo two-to-six times more germline cell divisions than females [3]. If mutations originate primarily due to errors in replication, then the male-to-female mutation rates (α) should be similar to the male-to-female ratio of germline cell division (c). If the observed value of α is smaller than c then the role of replication independent factors in generating mutations is not negligible. Published molecular evolutionary studies have concluded that the nucleotide substitution rates are higher in males than among females [9,17]. The Y chromosome is transmitted only through the male germ line because it is carried only by males; the X chromosome is transmitted more often through the female germline (because X spends 1/3 of its evolutionary time in males and 2/3 of its time in females) while the autosomes are transmitted equally in the male and female germline. Thus the male-to-female mutation rate ratio, α, can be determined by comparing the mutation rates among the X chromosome, the Y chromosome, and the autosomes [21]. A value of α less than one provides evidence that the mutations under study are selectively neutral (w.r.t. errors due to replication). A value of α between one and the ratio of germline cell division (c) would provide evidence indicating a possible male bias and also the presence of replication-independent factors for the mutations under study. The reported value of

germline cell division in humans is 6 (c = 6) [12]. A value of α greater than c provides evidence confirming the important role of replication errors in the generation of mutations. A value of α much greater than c might imply that errors in DNA replication during germ-cell division are the primary source of mutation and that replication-independent mutagenic factors such as methylation and oxygen radicals play lesser roles [33].

Wide range values is reported for human α in the current literature. Studies that compare the nucleotide substitution rates at homologous regions in primate genes between the sex chromosomes and the autosomes, have reported the value for α as ~5 [11,33]. When large regions (38.6 kb) with no known genes from the X and Y chromosomes were compared in humans, the value of α reported was 1.7 (95% confidence interval 1.15 – 2.87) in primates [2]. A genome wide analysis of Long Interspersed Nuclear Elements (LINES) from the initial sequence of the human genome reported α as ~2 [16]. All possible homologous comparisons between chimpanzee and human chromosomes reported α as ~3 [7]. When noncoding fragment on Y of about 10.4 kilobases (kb) and a homologous region on chromosome 3 in humans, greater apes, and lesser apes were compared, the estimated α was ~5 [18]. Hence, there is compelling evidence that the mutation rate for nucleotide substitution is higher amongst males than among females; however the precise extent of male point mutations remains an issue of debate.

Several reasons can be attributed for the variation in the reported α. Many investigations use homologous genes or strictly sex-linked sequences to calculate α [3,11,33]. Selection could have skewed sequence evolution in the introns and exons thus rendering the investigation to be biased. When sequences across species are compared to calculate α, the pairs under study might lie within chromosomal regions with substantially divergent nucleotide sequences which might skew the result. Also, when closely related sequences are compared, the reported α could be underestimated due to pre-existing polymorphisms. The variation in the reported values of α may be in part attributed to the small size of samples used in the various studies. Interestingly, most of the researches investigating male bias have analyzed point mutations only. While nucleotide substitution models have been studied extensively other mutations like indels have largely been treated as uninformative events. Thus, investigating whether insertions and deletions (indels) occur predominantly in males compared to females provides new insights on the widely accepted male driven evolution hypothesis. For humans knowing the extent of male bias in humans is of interest to evolutionary biologists.

A commonly observed replication error is the replication slippage, which occurs at the repetitive sequences when the new strand mispairs with the template strand. Mononucleotide runs are well-known hot spots for frame shift mutations, with DNA polymerase slippage typically resulting in loss or gain of one or a few nucleotides. Several studies have reported that replication slippage is responsible for many (1bp) small indels [24,34]. Deletions are generated when the replication complex skips across a number of nucleotides and fails to replicate them, whereas insertions are formed when the same region is mistakenly re-replicated. The replication driven origins of small indels in humans is supported by the study of potential indel mutation mechanisms including misalignment of short direct repeats during DNA replication and excision repair-mediated resolution of short inverted repeats [4]. The formation of indels is related to the nucleotide-sequence features in which they occur, such as the occurrence of repetitive motifs. Hence, it is necessary to investigate the male-to-female mutation rate using repeat sequences that harbor repetitive motifs are ancestrally related (that have accumulated indel mutations over time).

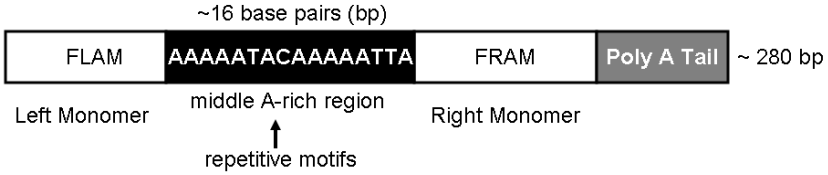

Fig. 1-A Typical Alu element structure

A major category of non-coding repetitive DNA within all mammalian genomes studied to date is the Short Interspersed Nuclear Elements (SINEs) that account for as much as 10% of all genomic sequence. Within the human genome, there are approximately one million copies of the *Alu* family of SINEs alone. *Alus* are 280bp long sequences with no known functionality [25]. *Alus* require forming of an RNA transcript that must then be reverse transcribed and inserted into a new location in the genome [6]. Thus *Alus* are believed to have colonized the genome by a 'copy and paste' mechanism [10] and have actively copied and pasted themselves in the genome at different time periods. Interestingly, there are no known mechanisms that specifically remove *Alu* elements from the genome [29] and

hence *Alus* can be used as effective fossil records. *Alus* have bypassed mutational inactivation, negative selection and/or putative host defense mechanisms that could have limited their expansion [26]. *Alu* elements are therefore a rich source of inter- and intra- species primate genomic variation [1,27,31,32]. As shown in Figure 1, the *Alu* element is a fusion of two free *Alu* monomers, the free left *Alu* monomer (FLAM) and the free right *Alu* monomer (FRAM) [26]. The two monomers are linked by a ~ 16 base pair (bp) poly (A) region. This middle poly (A) track in *Alu*s provides an ideal mechanism to further investigate the degree of replication-based indel (insertion and deletion) mutations in the human chromosomes. In a recent study on indels across the human genome, the majority of single base pair indels were reported as A:T and T:A base pairs, and these two classes together accounted for 84 % of the single base pair indels recorded [20]. Also, the middle poly (A) rich region is free from CpG dinucleotides and its phylogenetic analysis shall avoid chances of spurious variations.

In this study we provide a large scale genetic analysis of *Alu* elements found in the human genome. Analysis of indel patterns in the poly (A) track of the *Alu* elements found in the autosomes and the sex-chromosomes provides an unbiased investigation in calculating α for humans. It allows analysis of large numbers of sequences throughout the genome since it is found on all chromosomes in numbers sufficient for a rigorous statistical analysis. In non-functional sequences the rate of small indel mutations (replication driven mutations) should equal to the rate of mutation, hence the indels accumulated in *Alu* elements found on the Y-chromosomes shall constitute the mutations of paternal origin. Likewise, the number of indels accumulated on the X-chromosomes shall provide us with the mutations of maternal origin. The indels on the *Alu* elements that are found on the remaining 22 autosomes (non-sex-based chromosomes) shall provide us with a statistical baseline. This data is used to calculate the male-to-female mutation rate ratio (α).

Results

Number of *Alu* elements found in the human genome.

Table I shows the result of searching the entire human genome for *Alu* elements. 436562 *Alu* elements in the 22 non sex chromosomes (Autosomes), 6624 *Alu* elements in the X-chromosome and 3628 *Alu* elements in the Y-chromosome were recorded for analysis. Imperfectly copied *Alu*s during recombination were avoided in the search. Only the *Alu* elements with the middle poly (A) track were recorded and analyzed. A total of 7099741 nucleotides in the Autosomes, 107425 nucleotides in the X-chromosome and 59320 nucleotides in the Y-chromosome (all constituting the middle poly (A) regions of the detected *Alu*s) were reported. As shown in Table I deletions outnumber insertions in both Autosomes and the sex chromosomes.

Table I - Number of Alu elements and 1bp Indel events found in the human genome

	Number of				Percentage (%)		
	Alu elements	**Nucleotides**	**Insertions**	**Deletions**	**Insertions**	**Deletions**	**Indels**
Autosomes	436562	7099741	28864	130828	0.4065	1.8427	2.2492
X-Chromosome	6624	107425	475	1727	0.4421	1.6076	2.0498
Y-Chromosome	3628	59320	386	1222	0.6507	2.0600	2.7107

Insertion and Deletion events

After extracting information about the number of insertions, deletions, and length of middle poly (A) of each *Alu* element reported in the data set, the rate ratios are calculated using the three different methods shown below. As shown below the rate ratios Y/X are calculated each using only insertion events, deletion events and both insertion and deletion (Indels) events. The values for percentage indel events were obtained from Table I. Similarly, rate ratios were calculated for Y/A and A/X as shown in Table II.

$$R_{\text{insertion Y/X}} = \frac{\%\ \text{Insertions in Y-Chromosome}}{\%\ \text{Insertions in X-Chromosome}}$$

$$R_{\text{deletions Y/X}} = \frac{\%\ \text{Deletions in Y-Chromosome}}{\%\ \text{Deletions in X-Chromosome}}$$

$$R_{\text{indels Y/X}} = \frac{\%\ \text{Indels in Y-Chromosome}}{\%\ \text{Indels in X-Chromosome}}$$

The male-to-female mutation rate ratio (α)

Having estimated the rate ratios in the Autosomes (A), X chromosome (X) and the Y chromosome (Y), the male-to-female mutation rate ratios are calculated using the simple model of mutation frequencies proposed by Miyata T [21].

$\alpha_{A/X} = \left[\frac{(4\times R)-3}{3-(2\times R)}\right]$; where R is the rate ratio of the mutations in Autosomes and the X-chromosome

$\alpha_{Y/A} = \left(\frac{R}{(2-R)}\right)$; where R is the rate ratio of the mutations in Y-chromosome and the Autosomes

$\alpha_{Y/X} = \left(\frac{(2\times R)}{(3-R)}\right)$; where R is the rate ratio of the mutations in Y-chromosome and the X-chromosome

The calculated values for the male-to-female mutation rate ratio (α) are shown in Table II. We report the $\alpha_{Y/X}$ using combined (both insertion and deletion) indel events (shown in bold in Table II) as our analyzed male-to-female mutation ratio α in humans.

Table II- The male-to-female mutation rate ratio (α) using Indel ratios

	Y/X	$\alpha_{y/x}$	Y/A	$\alpha_{y/a}$	A/X	$\alpha_{a/x}$
Only Insertion events	1.4718	1.9262	1.6007	4.010	0.9194	0.5838
Only Deletion events	1.2814	1.4912	1.1179	1.2673	1.1462	2.2401
Combined Indel events	**1.3224**	**1.5765**	**1.2051**	**1.5163**	**1.0972**	**1.7246**

Discussion

The magnitude of the sex ratio of mutation rate has been a controversial issue, particularly in humans. The observations presented here are a result of investigations on only deletion and insertion mutations as point mutations have a different mechanism of mutagenesis. Because mutations in general and indels in particular are very rare, they are often difficult to measure with precision in a laboratory setting. A common alternative approach is to study substitutions in non-coding DNA. Given their evolutionary history and dearth of functionality, *Alus* offer a nearly ideal substrate for estimation of mutation rates in humans. Additionally, *Alu* repeats based results utilize information gathered over a large number of sites and from the accumulation of mutations over long evolutionary times. Since the α estimated for indel events from the three chromosomal comparisons ($\alpha_{A/X}$, $\alpha_{Y/A}$ and $\alpha_{Y/X}$) are similar (as shown in Table II) it can be inferred that differences between indel rates in the male and female germlines may be the dominant factor influencing the rate of DNA sequence evolution in humans.

Thus, the time DNA sequences spend in the male and female germline determines their overall evolutionary rate. Our estimate of α ~ 1.5 is based on the complete, diverse set of germline indel mutations that accumulated within the large, selectively neutral genomic *Alu* sequences. Our findings propose that indel rates in human males are only mildly higher than in females. Moreover, our findings suggest that sexual differences in indel rates are far less evident than the striking asymmetry observed in the number of cell divisions reported in humans. From the estimated value of α, it can be inferred that the errors in mitotic DNA replication and repair account for only a minority of germline indels in the human genome. As noted by Bohossian HB et al. [2] perhaps DNA replication and repair are unusually accurate in spermatogonial stem cells, which account for most of the excess cell divisions in the male germline. Our findings reflect a difference in numbers of genomic replications coupled to cell divisions per generation in males and females. Our results thus suggest a re-investigation of the model that human mutation rates are directly proportional to the number of cell divisions (c).

The value of α in human can be much smaller than c because the generation time in humans is much longer than the 25 years that was used in estimating the value of *c* for humans [12]. Also, the data for

calculating the number of germ-cell divisions in humans is insufficient to provide a reliable estimate for the value of *c* [17]. If recombination is mutagenic then the value of α can be underestimated from a comparison of *Alu* elements in the autosomes and the sex chromosomes because recombination is absent in the Y chromosome and the recombination rate is lower in the X chromosome than in the autosomes. Another possible reason for the significantly low value of α could be the specially reduced mutation rate in the X chromosome that may have been selected to compensate for its hemizygous state in males [19]. Even substantial variation in mutational rates between chromosomes due to regional differences in GC content, DNA repair, nuclear localization and metabolism may have skewed our results. Finally, it can also be hypothesized that the difference in mutational bias observed is simply from the DNA repair errors in the sperm (because of the higher levels of DNA damage) assuming that the errors in replication are similar for both sex chromosomes. It therefore remains to be demonstrated that other mechanisms do play a role in the observed differences in mutational rates between the sex chromosomes.

Many studies have indicated that indel mutations are related to recombination [5,34]. Also, small indels causing some human genetic diseases were found to originate with the same frequency in males and females [28]. If recombination were to main source of small indel mutations we would expect to see a lower X / Autosomes indel rate ratio. Thus our study supports a view that small and large indels originate by different molecular mechanisms. Sequence comparison between ~ 6kb on the X chromosome and ~ 5kb on the Y chromosome in primates indicated similar indel frequencies, suggesting no sex bias for large (> 1bp) indels in primates [34]. Interestingly, the most parsimonious explanation for our results is that most 1bp indels occur during DNA replication and/or during DNA repair after DNA replication. This is consistent with the hypothesis that DNA replication errors are the major source of small indels.

The reason for substantial variations in primate genome sizes is currently unknown. Indel polymorphisms are of great interest because they can alter human phenotypes. It has been suggested that DNA loss caused by biases in small insertions and deletions (indels) can be a determinant of genomic size [24]. Our findings add further support to the mutational equilibrium model shown in Fig. **2-** (proposed by Petrov DA [24]).

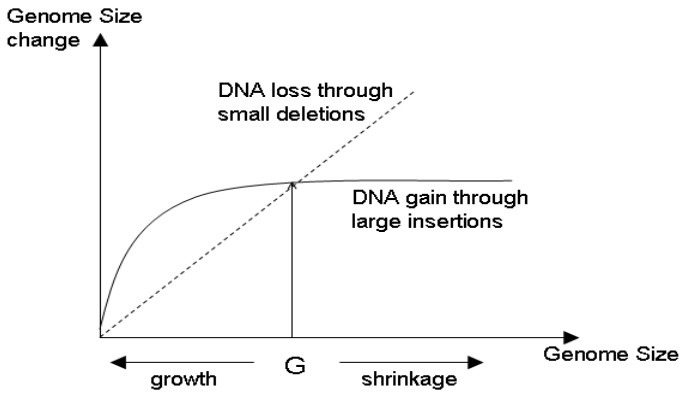

Fig. 2- The Mutational Equilibrium model [24].

The model hypotheses that for small genome sizes the rate of genome size increase is higher than that of DNA loss resulting in genome size growth. However, since the rate of DNA loss through small deletions is shown to grow linearly and thus faster than the rate of DNA gain, for very large genome sizes DNA loss is faster than DNA growth. Therefore, there exists a stable equilibrium at a finite value of genome size (shown as G in Fig. **2-**).

In our analysis, higher prevalence of indels on the Y chromosome compared with X and autosomes are observed for both insertions and deletions. Interestingly, the male-to-female ratio is higher for insertions ($\alpha_{Y/X}$ = 1.9262) than for deletions ($\alpha_{Y/X}$ = 1.4912). Although we cannot rule out coincidence, deletions seem to be a major phenomenon in the generation of sequence diversity. Our results indicate that the mutational pressure at the level of small indels is biased toward DNA loss. If the preferential fixation of small deletions over small insertions is not prevented by selection then all genomes are constantly losing DNA through small indels.

We conclude that although small 1bp indel mutations may be primarily replication driven (as previous studies suggest) the observed value of α does not exceed the threshold necessary to conclude that contributions of replication independent factors are negligible. We also report that, with small indels (1bp) deletions outnumber insertion events. This relative excess of deletions may be an important parameter in the long-term evolution of genomic size.

Material and Methods

Data Acquisition

This study uses the entire human genome data as reported on January 27th 2005 by National Center for Biotechnology Information (NCBI) [22]. The sequences obtained were present in *contigs* of variable length where each *contig* represents a set of contiguous gene cluster present in the chromosome. Each chromosome file was parsed and the *contigs* separated into files. The *contigs* were then cut into smaller parts of 800,000 nucleotides or less for ease in processing. 225 *Alu* sequences were obtained from the Repbase database [14] and from the supplementary material provided at the Genome research website for the article by A.L. Price et al [25].

Data Processing

The study uses the CENSOR, version 1.1, [13], to perform rapid comparison and alignment of reference sequences with the sequence under study. Our study uses 225 *Alu* sequence data file as the reference sequence and the cut up *contigs* of the entire human genome as the sequence under study. CENSOR uses the ratio of mismatches to transitions in combination with alignment and similarity scores to distinguish true homology from accidental similarity between sequences [13]. In our study, CENSOR was used with the default sensitivity settings.

Data Extraction and Analysis

Details about the number of transitions, transversions, matches, mismatches, length, gaps, and type of indels and the rate of substitution was extracted about the FLAM, FRAM and the middle poly (A) track of each *Alu* element found and was recorded using Perl scripts on Censor output files. Statistical analysis on the data was performed using Perl scripts in combination with the JMP statistics software.

Competing Interests Statement

The author declares that he has no competing financial interests.

Acknowledgements

The authors would like to thank Drs. Travis Doom, Dan Krane and Michael Raymer for critically reading the manuscript and providing valuable insights during the process of this investigation. We also thank Indiana University Southeast and Wright State University for providing the funds and resources to carry out the investigation.

References

[1] Bailey J.A., Liu G., Eichler E.E. (2003) *American Journal of Human Genetics*, **73:**823–834.

[2] Bohossian H.B., Skaletsky H., Page D.C. (2000) *Nature*, **406:**622-625.

[3] Chang B.H., Hewett-Emmett D., Li W-H. (1996) *Zoological Studies*, **35:**36 – 48.

[4] Chuzhanova N.A., Annassis E.J., Ball E.V., Krawczak M., Cooper D.N. (2002) *Human Mutations*, **21:**28 – 44.

[5] Crow J.F. (2000) *Nature Review Genetics*, **1:**40 – 47.

[6] Deininger P.L., Batzer M.A. (2002) *Genome Research*, **12:**1455 – 1465.

[7] Ebersberger I., Metzler D., Schwarz C., Paabo S. (2002) *American Journal of Human Genetics*, **70:**1490 – 1497.

[8] Fryxell K.J., Moon W.J. (2004) *Molecular Biology and Evolution*, **22(3):**650 – 658.

[9] Haldane J.B.S. (1935) *Journal of Genetics*, **31:**317 – 326.

[10] Hedges D.J., Callinan P.A., Cordaux R., Xing J., Barnes E., Batzer M.A. (2004) *Genome Research*, **14:**1068 – 1075.

[11] Huang W., Chang B.H., Hewett-Emmett D., Li W-H. (1997) *Journal of Molecular Evolution*, **44:**463 – 465.

[12] Hurst L.D., Ellegren H. (1998) *Trends in Genetics*, **14:**446 – 452.

[13] Jurka J., Klonowski P., Dagman V., Pelton P. (1996) *Computers and Chemistry*, **20(1):**119 – 122.

[14] Jurka J. (2000) *Trends in Genetics,* **9:**418 – 420.

[15] Kimura M. (1980) *Journal of Molecular Evolution*, **16:** 111 – 120.

[16] Lander E.S., Linton L.M., Birren B., et al. (249 co-authors) (2001) *Nature*, **409:** 860 – 921.

[17] Li W-H., Yi S., Makova K. (2002) *Genetics and Development*, **12:**650 – 656, 2002.

[18] Makova K.D., Li W-H. (2002) *Nature*, **416:**624-626.

[19] McVean G.T., Hurst L.D. (1997) *Nature*, **386:**388 – 392.

[20] Mills R.E., Luttig C.T., Larkins C.E., Beauchanp A., Tsui C., Pittard W.S., Devine S.E. (2006) *Genome Research*, **16:**1182 – 1190.

[21] Miyata T., Hayashida H., Kuma K., Mitsuyasa K., Yasunaga T. (1987) Cold Spring Harbor Symposium, *Quantitative Biology*, **52:**863 – 867.

[22] *National Center for Biotechnology Information (NCBI)* [ftp://ftp.ncbi.nih.gov/genbank].

[23] Petrov D.A., Hartl D.L. (1998) *Molecular Biology and Evolution*, **15:** 293 – 302.

[24] Petrov D.A. (2002) *Theoretical Population Biology*, **61:**533 – 546.

[25] Price A.L., Eskin E., Pevzner P.A. (2004) *Genome Research*, **14:**2245 – 2252.

[26] Ramachandran S., Doom T., Raymer M., Krane D. (2006) *In the Proceedings of the*

IEEE Bioinformatics and Bioengineering conference, **BIBE06:**213 - 219

[27] Raya D.A., Xinga J., Hedges D.J., Hall M.A., Laborde M.E., Anders B.A., White B.R., Stoilova N., Fowlkes J.D., Landry K.E., Chemnick L.G., Ryder O.A., Batzer M.A. (2005) *Molecular Phylogenetics and Evolution*, **35:**117–126.

[28] Roberts P.S., Chung J., Jozwiak S., Dabora S., Franz D.N., Thiele E.A., Kwiatkowski D. (2002) *Human Genetics*, **111:**96 – 101.

[29] Roy-Engel A.M., Carroll M.L., El-Sawy M., Salem A-H., Garber R.K., Nguyen S.V., Deininger P.L., Batzer M.A. (2002) *Journal of Molecular Biology*, **316 (5):**1033 – 1040.

[30] Roy-Engel A.M., Salem A-H., Oyeniran O.O., Deininger L., Hedges D.J., Kilroy G.E., Batzer M.A., Deininger P.L. (2002) *Genome Research*, **12:**1333 – 1344.

[31] Salem A-H., Ray D.A., Hedges D.J., Jurka J., Batzer M.A. (2005) *BMC Evolutionary Biology*, **5(18):**1-9.

[32] Shedlock A.M., Takahashi K., Okada N. (2004) *Trends in Ecology and Evolution*, **19 (10):** 545 – 553.

[33] Shimmin L.C., Chang BH-J., Li W-H. (1993) *Nature*, **362:**745 – 747.

[34] Sundstrom H., Webster M.T., Ellegren H. (2003) *Genetics*, **164:**259 – 268.

[35] Taylor J., Tyekucheva J.S., Zody M., Chiaromonte F., Makova K.D. (2006) *Molecular Biology and Evolution*, **23(3):**565-573.