International Journal of Bioinformatics Research
ISSN: 0975–3087, E-ISSN: 0975–9115, Vol. 3, Issue 1, 2011, pp-142-147
Available online at http://www.bioinfo.in/contents.php?id=21

INVESTIGATING THE MALE-DRIVEN EVOLUTION HYPOTHESIS USING GENOME WIDE POINT 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: There is substantial evidence that replication-based nucleotide substitutions in primates occur more frequently in males than in females. There has been disagreement about the extent of this disparity. The human Alu repeats provide an ideal mechanism to further investigate the degree of replication-based error by providing tighter statistical boundaries on the male-to-female mutation ratio, α, in humans. **Results:** We analyze patterns of point mutations in Alu repeats across the entire human genome in order to elucidate the processes of mutation and fixation. This analysis provides substantial statistically bounded support for the accumulation of more point mutations in the Y chromosome compared to the X chromosome. We report a 99.99% confidence interval for human α between 1.280 and 1.289. Our results suggest that compared with eggs, sperm tend to carry a greater number of point mutations accumulated primarily during the production of gametes. **Conclusion:** Our results suggest that although 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.

Keywords–ALU repeats, male-to-female mutation ratio

Background

In humans, the germ-lines are maintained separately from somatic cells and the mutations in the gametes can arise only from the germ cells. There are many more cell divisions in spermatogenesis than in oogenesis and assuming that new mutations arise due to DNA replications, mutations should originate more frequently in males than in females. Therefore, replication-dependant difference between the male and female germ-lines in humans could lead to gender specific mutation rates. Even though a number of studies have detected a male-driven evolution among 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 and subsequent failure of DNA repair mechanisms.

Many molecular evolutionary studies have concluded that the nucleotide substitution rates are higher in males than among females [7,15]. With many more rounds of cell division per generation, males accumulate more mutations. In primates, males undergo two-to-six times more germ-line 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.

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 (a X chromosome 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 [18]. A value of α less than one provides evidence that the mutations under study are selectively neutral (with respect to 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) [10].

A value of α greater than *c* provides evidence confirming the important role of replication errors in the generation of substitution (point) mutations. A value of α much greater than *c* might imply that errors in DNA replication during germ-cell division is the primary source of mutation and that replication-

Copyright © 2011, Bioinfo Publications

independent mutagenic factors such as methylation and oxygen radicals are not important [27].

There are a wide range of values for human α currently reported in the literature. In studies that compare the nucleotide substitution rates at homologous regions in primate genes between the sex chromosomes and the autosomes, the value for α has been reported as ~5 [9,27]. When large regions (38.6 kb) with no known genes from the X and Y chromosomes were compared in humans, the value of α obtained 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 [14]. All possible homologous comparisons between chimpanzee and human chromosomes reported α as ~3 [5]. 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 [16]. 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 [3,9,27]. Selection could have skewed sequence evolution in the introns and exons leading to a biased estimate. When sequences across species are compared to calculate α, the pairs under study might lie within chromosomal regions with substantially divergent nucleotide sequences which also 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. Thus, it is necessary to investigate the male-to-female mutation rate using selectively neutral sequences that are ancestrally related (that have accumulated mutations without having undergone gene conversion).

A major category of non-coding 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 [20]. Propagation of *Alus* requires forming of an RNA transcript that must then be reverse transcribed and inserted into a new location in the genome [4]. Thus *Alus* are believed to have colonized the genome by a 'copy and paste' mechanism [8] 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 [23] 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 [21]. *Alu* elements are therefore a rich source of inter- and intra- species primate genomic variation [1,22,25,26].

In this study we provide a large scale genetic analysis of *Alu* elements found in the human genome. Analysis of substitution patterns in *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 nonfunctional sequences the rate of nucleotide substitution can be expected to be approximately equal to the rate of mutation; hence the mutations accumulated in *Alu* elements found on the Y-chromosomes constitute the mutations of paternal origin. Likewise, the number of mutations accumulated on the X-chromosomes provides us with the mutations of maternal origin. The mutations 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 shall then be used to calculate the male-to-female mutation rate ratio (α).

Results

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

Table 1 shows the result of searching the entire human genome with four different sets of requirements. When the search was conducted with no restrictions on size or type (Data Set 1), a total of 666259 *Alu* elements were found in the human genome. However, when only *Alu* elements greater than or equal to 200bp long were considered (Data Set 3),to avoid imperfectly copied *Alus* during recombination, if any, 409988 *Alu* elements were reported. Data from the *Alu* elements obtained with and without size restraints were then also masked for hypermutable CpG dinucleotides (Data Sets 2 & 4) to remove potential confounding factors. Fryxell KJ and Moon WJ [6] point out that CpG dinucleotides are mutational hotspots that mutate at a high rate because cytosine is vulnerable to deanimation (removal of the amino group). As *Alus* are unusually CpG rich they can be potential targets for genomic methylation. Thus we mask CpG dinucleotides to avoid a chance of spurious variations via this mechanism. The four data sets shown in Table 1 were each investigated separately to check the consistency of the reported results. The results are consistent for each run and the most restricted data set (Data Set 4) is used in reporting the results of this study henceforth.

Table 1 - Number of Alu elements found in the human genome

Data Set	**Size Restriction**	**CpG Mask**	***Alu* elements in the Human Genome**		
			Autosomes	**X-Chromosome**	**Y-Chromosome**
1	No	No	650935	10034	5290
2	No	Yes	650935	10034	5290
3	Yes	No	400642	6046	3300
4	**Yes**	**Yes**	**400642**	**6046**	**3300**

The Kimura rate of substitution

After extracting information about the number of transitions, transversions, and length of each element reported in the data set, the Kimura rate of substitution [13] is calculated using the two different methods shown below. Taking into consideration the huge sample size (409988 *Alu* elements analyzed), even a small difference in results is of statistical significance. It is necessary to correct for multiple substitutions using the Kimura model because assuming that *Alu* elements of the same subfamily were inserted into the genome at the same time could misstate their degree of difference.

Kimura rate of substitution = $0 \cdot 5 \times LN\left(\frac{1}{1-(2\times(P-Q))}\right) + 0.25 \times LN\left(\frac{1}{1-(2\times Q)}\right)$ where LN = natural*log*

where $P = \left(\frac{total\ number\ of\ transitions}{Alu\ element\ size}\right)$ and $Q = \left(\frac{total\ number\ of\ transversions}{Alu\ element\ size}\right)$, Method 1

where $P = \left(\frac{total\ number\ of\ transitions}{total\ number\ of\ sites}\right)$ and $Q = \left(\frac{total\ number\ of\ transversions}{total\ number\ of\ sites}\right)$, Method 2

Table 2- Kimura rate of substitution distribution values for Data Set 4 using confidence intervals.

C.I.= Mean +/- Z*S.E.								
LCI: Lower Confidence Interval. UCI: Upper Confidence Interval.			95 % Confidence		99 % Confidence.		99.99% Confidence.	
S.E.	***CpG Masking Size Restriction***	Std. Dev.	**LCI**	**UCI**	**LCI**	**UCI**	**LCI**	**UCI**
Method (A)								
Std. Deviation / √ (# *Alu* elements found)	**Autosomes**	0.0415303	0.07545	0.07570	0.07541	0.07574	0.07532	0.07583
	X-Chromosome	0.040606	0.07121	0.07326	0.07089	0.07358	0.0702	0.07427
	Y-Chromosome	0.0428136	0.08331	0.08624	0.08290	0.08670	0.08187	0.08768
Method (B)								
Std. Deviation / √ (# nucleotides)	**Autosomes**	0.0415303	0.07557	0.07558	0.07556	0.07559	0.07556	0.07559
	X-Chromosome	0.040606	0.07217	0.07230	0.07214	0.07232	0.07212	0.07234
	Y-Chromosome	0.0428136	0.08468	0.08487	0.08465	0.08489	0.08461	0.08493

Copyright © 2011, Bioinfo Publications

Using Method 1, a unique value of P, and Q and was recorded per *Alu* element in the data set, thus determining each elements Kimura rate. The collection of Kimura rates for the *Alu* elements in the autosomes, the X chromosome, and the Y chromosome are each analyzed to obtain mean Kimura rates of substitution characterized by standard deviation and the confidence interval of the mean with different confidence intervals. Two different methods are used while calculating the confidence interval of the mean:

Confidence Interval (C.I.) = Mean +/- Z**S.E*;
where, *S.E.* = Std. Deviation / √ (Number of *Alu* elements found), Method A shown in Table 2
where, *S.E.* = Std. Deviation / √ (Total number of nucleotides), Method B shown in Table 2
and Z depends on the size of the interval being determined.

Consistency of results was checked using Method 2, where each nucleotide was considered as a data point while calculating the Kimura rates each for the autosomes, X chromosome and the Y chromosome.

Table 2 shows the calculated Kimura rates for the Data Set 4 from Table 1 using Method 1 discussed above.

Table 3- The male-to-female mutation rate ratio (α) using extreme kimura rates

LL: Lower Limit on the value of α. **HL**: Higher Limit on the value of α.		$\alpha_{y/x}$		$\alpha_{y/a}$		$\alpha_{a/x}$	
		LL.	**HL.**	**LL.**	**HL.**	**LL.**	**HL.**
Using Kimura rates from Table 2 (A)	95% Confidence Interval	1.2210	1.3540	1.2235	1.337	1.1908	1.4329
	99 % Confidence Interval	1.2028	1.3765	1.2072	1.3522	1.1570	1.4756
	99.99 % Confidence Interval	1.1618	1.4266	1.1731	1.3926	1.0873	1.5731
Using Kimura rates from Table 2(B)	95% Confidence Interval	1.2809	1.2894	1.2438	1.2807	1.2984	1.3131
	99 % Confidence Interval	1.2796	1.2908	1.2724	1.2817	1.2953	1.3173
	99.99 % Confidence Interval	1.2780	1.2924	1.2710	1.2831	1.2932	1.3194

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

Having estimated the kimura rates of substitution 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 [18]:

$$\alpha_{A/X} = \left[\frac{(4\times R)-3}{3-(2\times R)}\right] \; ; \quad \text{where } R = \left(\frac{kimura\, rate\, of\ substitution\, in\ Autosomes}{kimura\, rate\, of\ substitution\, in\, Y\ chromosome}\right)$$

$$\alpha_{Y/A} = \left(\frac{R}{(2-R)}\right) \; ; \quad \text{where } R = \left(\frac{kimura\, rate\, of\ substitution\, in\, Y\ chromosome}{kimura\, rate\, of\ substitution\, in\ Autosomes}\right)$$

$$\alpha_{Y/X} = \left(\frac{(2\times R)}{(3-R)}\right) \; ; \text{ where } R = \left(\frac{kimura\, rate\, of\ substitution\, in\, Y\ chromosome}{kimura\, rate\, of\ substitution\, in\ X\ chromosome}\right)$$

To obtain tight statistical bounds while calculating α, the extreme kimura rates from Table 2 were used (using normal approximation). The male-to-female mutation rate ratio:

α $_{(\text{Lower bound})}$ are calculated using $R = \left(\frac{lower\, bound\ value\, of\ kimura\, rate}{higher\, bound\ value\, of\ kimura\, rate}\right)$, and the male-to-female mutation rate ratio: α $_{(\text{Upper bound})}$ are calculated using $R = \left(\frac{higher\, bound\ value\, of\ kimura\, rate}{lower\, bound\ value\, of\ kimura\, rate}\right)$.

The calculated values for the male-to-female mutation rate ratio (α) are shown in Table 3.

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 point substitutions as deletion and insertion mutations have a different mechanism of mutagenesis. Given their evolutionary history and dearth of functionality, *Alu*s offer a nearly ideal substrate for estimation of substitution rates in humans. Additionally, *Alu* 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 from the three chromosomal comparisons ($α_{A/X}$, $α_{Y/A}$ and $α_{Y/X}$) are similar (as shown in Table 3) it can be inferred that differences between mutation rates in the male and female germlines is 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 ~ 1.285 (99.99% confidence interval 1.280 – 1.289) is based on the complete, diverse set of germline point mutations that accumulated within the large, selectively neutral genomic *Alu* sequences. Our findings propose that, contrary to previous reports, substitution rates in human males are only slightly higher than in females. Moreover, our findings suggest that sexual differences in substitution 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 substitutions 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 on 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 [10]. 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* [15]. 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 low value of α could be the reduced mutation rates in the X chromosome that may compensate for its hemizygous state in males [17]. 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.

The exact magnitude of germ cell division in humans needs further investigation. A recent study by Taylor J et al [28] on CpG and non-CpG sites in the human genome reports the possibility of nonuniform male mutational biases across the genome. Factors, other than sex specific differences in substitution rates that influence the accumulation of substitutions in the human genome also need investigation. Extrapolating the mutation rate data from sex chromosomes to overall sex-specific rates requires more investigation on replication-independent factors.

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) [19]. 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 [12] and from the supplementary material provided at the Genome research website for the article by Price AL et al [20].

Data Processing

The study uses the CENSOR, version 1.1, [11], 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 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 [11]. In our study, CENSOR was used with the default sensitivity settings.

Copyright © 2011, Bioinfo Publications

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 each *Alu* element found and 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 authors declare that they have 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] Deininger P.L., Batzer M.A. (2002) *Genome Research*, **12:**1455 – 1465.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[21] Ramachandran S., Doom T., Raymer M., Krane D. (2006) *In the Proceedings of the IEEE Bioinformatics and Bioengineering conference*, **BIBE06:**213 - 219

[22] 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.

[23] 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.

[24] 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.

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

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

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

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