11-26-2004

Alu element mutation spectra: Molecular clocks and the effect of DNA methylation

Jinchuan Xing
Ctr. Bio-Modular Microsystems

Dale J. Hedges
Ctr. Bio-Modular Microsystems

Kyudong Han
Ctr. Bio-Modular Microsystems

Hui Wang
Ctr. Bio-Modular Microsystems

Richard Cordaux
Ctr. Bio-Modular Microsystems

See next page for additional authors

Recommended Citation

Xing, J., Hedges, D., Han, K., Wang, H., Cordaux, R., & Batzer, M. (2004). Alu element mutation spectra: Molecular clocks and the effect of DNA methylation. *Journal of Molecular Biology, 344* (3), 675-682. https://doi.org/10.1016/j.jmb.2004.09.058

doi:10.1016/j.jmb.2004.09.058 J. Mol. Biol. (2004) 344, 675–682

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Alu Element Mutation Spectra: Molecular Clocks and the Effect of DNA Methylation

Jinchuan Xing, Dale J. Hedges, Kyudong Han, Hui Wang Richard Cordaux and Mark A. Batzer*

Department of Biological Sciences, Biological Computation and Visualization Center, Center for Bio-Modular Microsystems, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803, USA

**Corresponding author*

In primate genomes more than 40% of CpG islands are found within repetitive elements. With more than one million copies in the human genome, the *Alu* family of retrotransposons represents the most successful short interspersed element (SINE) in primates and CpG dinucleotides make up about 20% of *Alu* sequences. It is generally thought that CpG dinucleotides mutate approximately ten times faster than other dinucleotides due to cytosine methylation and the subsequent deamination and conversion of C→T. However, the disparity of *Alu* subfamily age estimations based upon CpG or non-CpG substitution density indicates a more complex relationship between CpG and non-CpG substitutions within the *Alu* elements. Here we report an analysis of the mutation patterns for 5296 *Alu* elements comprising 20 subfamilies. Our results indicate a relatively constant CpG *versus* non-CpG substitution ratio of ~6 for the young (*Alu*Y) and intermediate (*Alu*S) *Alu* subfamilies. However, a more complex non-linear relationship between CpG and non-CpG substitutions was observed when old (*Alu*J) subfamilies were included in the analysis. These patterns may be the result of the slowdown of the neutral mutation rate during primate evolution and/or an increase in the CpG mutation rate as the consequence of increased DNA methylation in response to a burst of retrotransposition activity ~35 million years ago.

Keywords: CpG methylation; Alu; primate evolution; molecular clock; transposable elements

Introduction

Deoxycytosine methylation is the most common epigenetic modification within vertebrate genomes. It has been implicated in many important functions, such as regulation of gene expression, control of development and repressing transposable elements.[1,2] On the other hand, aberrant DNA methylation patterns and methylation-induced mutations have also been associated with multiple diseases, particularly with the development of cancers.[2,3] Deoxycytosine methylation primarily occurs at the cytosine of a CpG dinucleotide, generating 5-methyl cytosine (5mC). In different vertebrate genomes, 60–90% of the CpG dinucleotides contain 5mC.[4] The 5mC at CpG sites mutates unidirectionally to thymine by spontaneous deamination at a much higher transition rate compared to non-CpG dinucleotides.[5] This leads to a rapid decay of CpG sites, which is believed to be the cause of the observed deficiency in CpG dinucleotides and the corresponding increase in TpG and CpA dinucleotide frequency in vertebrate genomes. In the human genome, CpG dinucleotides are present only at about 20% of their expected frequency.[6]

In primate genomes more than 40% of CpG islands are found within repetitive elements,[6] and they are generally heavily methylated. *Alu* elements represent the most successful short interspersed element (SINE) within primates. With more than 1.1 million copies, *Alu* elements comprise ~10% of the human genome by mass.[6,7] Several properties of *Alu* elements provide for a unique opportunity to examine the CpG mutation history in the genome. First, the CpG content in *Alu* elements accounts for up to 30% of the total 5mC sites in the human genome.[8] Second, based on shared diagnostic nucleotide sites, *Alu* elements can be classified

Abbreviations used: myr(s), million year(s); SINE, short interspersed element.

E-mail address of the corresponding author: mbatzer@lsu.edu

into specific subfamilies, which have expanded at different times during primate evolution.[9] The CpG mutation pattern of individual *Alu* subfamilies can thus be used as "snapshots" of the CpG mutational history at different time periods throughout primate evolution. Third, *Alu* elements are generally considered neutral loci that are not subject to selective pressure once fixed in the genome.[7] Finally, unlike pseudogenes and other genetic markers, the over 1.1 million copies of *Alu* elements are distributed widely throughout the genomic landscape. These elements not only provide a representative coverage of a variety of genomic locations, but also allow us to minimize sampling errors through the use of a large dataset.

It is generally accepted that CpG mutations in *Alu* elements occur about ten times faster than non-CpG mutations, and traditionally a ten times faster molecular clock has been applied for CpG mutations.[10,11] However, the age estimations obtained by CpG mutations or non-CpG mutations based on this ratio provide results that are appreciably and systematically different.[12–14] This discrepancy suggests that the tenfold higher substitution rate may not accurately reflect the relationship between the CpG and non-CpG substitution density in *Alu* elements.

Here we report the analysis of 5296 *Alu* elements belonging to 20 *Alu* subfamilies in the human genome. This dataset allows us to trace the CpG mutation history back to the beginning of primate evolution, about 65 million years (myrs) ago.[15] This study is based on the largest dataset gathered to date for the analysis of 5mCpG mutation patterns and aims at contributing to a better understanding of the CpG decay process during primate evolution as well as providing an updated, more accurate estimate of the neutral mutation rate disparity between CpG and non-CpG dinucleotides.

Results

Substitution densities in *Alu* elements

The CpG and non-CpG substitution densities were analyzed for 5296 *Alu* elements from 20 subfamilies based on sequence alignments. The sample size of each subfamily, observed CpG and non-CpG substitution densities, and the observed ratios of CpG/non-CpG substitution density, *R*, of different subfamilies are shown in Table 1. For most of the young *Alu*Y subfamilies, the ratio *R* varies from 4.80 to 9.27. The single exception is the *Alu*Ya5a2 subfamily ($R = 43.77$). The *Alu*Ya5a2 subfamily appears to be extremely young (only ten mutations in the whole subfamily) and has the smallest sample size (33 elements) among all the subfamilies examined (Table 1). This suggests that the high *R* of *Alu*Ya5a2 may be due to sampling error arising from the limited number of mutations present in this very young subfamily. This is further

Table 1. Summary of *Alu* subfamilies examined

Alu subfamily	Sample size	CpG substitution density (SD) (%)	Non-CpG substitution density (SD) (%)	CpG/non-CpG substitution ratio (*R*)
Ya5a2[a]	33	0.58 (1.21)	0.01 (0.07)	43.77
Ya8[b]	33	1.45 (1.91)	0.16 (0.24)	9.27
Yb9[c]	53	1.44 (1.90)	0.22 (0.37)	6.54
Ya5[d]	488	3.16 (4.82)	0.50 (0.77)	6.73
Yc1[e]	232	1.79 (2.89)	0.29 (0.48)	6.13
Yb8[d]	290	2.87 (3.66)	0.48 (0.67)	7.37
Yb7[f]	153	1.43 (1.77)	0.30 (0.43)	4.82
Yd6[g]	97	1.76 (2.07)	0.37 (0.44)	4.80
Yg6[h]	156	2.64 (3.01)	0.40 (0.46)	6.55
Yi6[h]	106	4.93 (4.10)	0.64 (0.96)	7.77
Yd3[g]	193	12.74 (6.18)	1.89 (0.98)	6.73
Ye5[i]	139	10.65 (5.00)	1.73 (0.86)	6.17
Yd[j]	915	17.32 (7.03)	2.24 (1.24)	7.73
Sp[j]	209	29.75 (5.98)	4.25 (1.40)	7.01
Sc[j]	169	30.42 (6.73)	4.57 (1.52)	6.67
Sg[j]	510	31.21 (6.56)	5.13 (1.75)	6.08
Sq[j]	340	33.74 (6.30)	5.92 (1.87)	5.71
Sx[j]	423	35.82 (6.09)	6.61(2.15)	5.42
Jb[j]	399	41.40 (5.95)	10.64 (2.43)	3.90
Jo[j]	358	44.92 (5.72)	13.11 (2.82)	3.43

SD, standard deviation.
[a] Ref. 32.
[b] Ref. 33.
[c] Ref. 34.
[d] Ref. 14.
[e] Ref. 35.
[f] Ref. 13.
[g] Ref. 12.
[h] Ref. 36.
[i] Ref. 37.
[j] Ref.; this study.

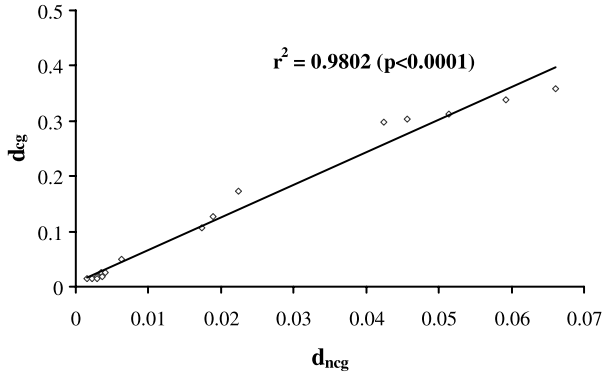

Figure 1. CpG substitution densities (d_{cg}) and non-CpG substitution densities (d_{ncg}) of *Alu*Y and *Alu*S subfamilies exhibit a linear correlation. Correlation coefficient (r-square) value and p value are shown.

corroborated by descriptive statistics, which indicate that the *Alu*Ya5a2 R value lies well over three interquartile ranges beyond the third quartile of the dataset. The *Alu*Ya5a2 subfamily was therefore excluded from the subsequent analyses, yielding an average R for *Alu*Y subfamilies of 6.72. For the intermediate *Alu*S subfamilies, R ranged from 5.42 to 7.01 with an average of 6.18, similar to *Alu*Y subfamilies. By contrast, the oldest *Alu*J subfamilies exhibited substantially lower R values, with an average of 3.67 (Table 1).

The CpG mutation rate in *Alu* elements remains constant for young and intermediate *Alu* subfamilies

By excluding *Alu*J subfamilies, the expected CpG and non-CpG substitution density of the young *Alu*Y and intermediate *Alu*S subfamilies showed a good linear correlation ($r^2 = 0.98$, $p < 0.0001$) (Figure 1). This result indicates that during the time period that gave rise to *Alu*Y and *Alu*S subfamilies nucleotide substitution can be treated as an approximately linear process, and the slope of the correlation (5.89) suggests that the CpG and non-CpG substitution densities exhibited a relatively constant approximately sixfold difference. One exception is *Alu*Ya8, which exhibited a CpG to non-CpG substitution density ratio of 9.27. This deviation of the *Alu* Ya8 value is not surprising, however, given that, similar to *Alu*Ya5a2, *Alu* Ya8 has a small sample number ($N = 33$) and contains relatively few mutations for accurate estimation of relative decay rates.

We next designated the ratio of CpG to non-CpG mutation rates as alpha (μ_{cg}/μ_{ncg}). There is a subtle but important distinction to be made between R and alpha. R is the observed ratio of the substitution rates while alpha is the ratio of the absolute mutation rates. In a simple linear model, the mutation rate is equal to the substitution rate, thus these two values would be equivalent. However, once back-substitution of nucleotides and unidirectional CpG mutation are incorporated, the R value (d_{cg}/d_{ncg}) changes as a function of time while alpha, the ratio of the actual mutation rates μ_{cg}/μ_{ncg}, remains constant.

Since R equals alpha under a linear model, we re-estimated the age of *Alu*Y and *Alu*S subfamilies based on CpG sites using alpha = 6 and compared the results with both non-CpG-based age estimates and CpG-based age estimates using alpha = 10 (Table 2). Figure 2 illustrates that for all *Alu* subfamilies examined age estimates based on alpha = 6 are more consistent with non-CpG-based age estimates than those based on the previously used alpha = 10, with the exception of *Alu* Ya8 as noted above.[10–12,14] For example, for *Alu*Ya5, one of the largest *Alu*Y subfamilies, the age estimates based on non-CpG substitution density and CpG substitution density with alpha = 6 or alpha = 10 were 2.33, 2.63 and 1.58 myrs, respectively. For the *Alu* Sg subfamily the estimates were 34.23, 34.68 and 20.81 myrs, respectively. This analysis therefore suggests that a relatively constant sixfold higher

Table 2. Age estimates of *Alu*Y and *Alu*S subfamilies

Alu subfamily	Age non-CpG (SD) (myrs)	Age CpG (alpha = 6) (SD) (myrs)	Age CpG (alpha = 10) (SD) (myrs)
Ya8	1.07 (1.62)	1.61 (2.12)	0.97 (1.27)
Yc1	1.95 (3.19)	1.99 (3.21)	1.20 (1.93)
Yb9	1.47 (2.47)	1.60 (2.11)	0.96 (1.27)
Yd6	2.45 (2.92)	1.96 (2.30)	1.18 (1.38)
Yg6	2.68 (3.06)	2.94 (3.35)	1.76 (2.01)
Yb7	2.00 (2.87)	1.59 (1.97)	0.95 (1.18)
Yb8	2.33 (2.73)	2.89 (2.93)	1.73 (1.76)
Ya5	2.33 (2.67)	2.63 (2.87)	1.58 (1.72)
Yi6	4.24 (6.41)	5.48 (4.55)	3.29 (2.73)
Yd3	12.61 (6.50)	14.15 (6.86)	8.49 (4.12)
Ye5	11.53 (5.75)	11.83 (5.56)	7.10 (3.33)
Yd	14.96 (8.27)	19.24 (7.82)	11.54 (4.69)
Sp	28.31 (9.35)	33.06 (6.64)	19.83 (3.99)
Sc	30.44 (10.14)	33.80 (7.48)	20.28 (4.49)
Sg	34.23 (11.64)	34.68 (7.29)	20.81 (4.37)
Sq	39.44 (12.44)	37.49 (7.01)	22.49 (4.20)
Sx	44.10 (14.31)	39.80 (6.77)	23.88 (4.06)

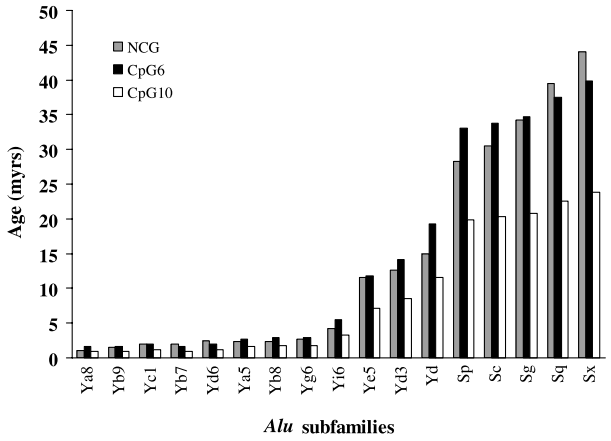

Figure 2. Age estimates for *Alu*Y and *Alu*S subfamilies based on different parameters. The *Alu* age estimates based on non-CpG substitution density, and CpG substitution density with alpha = 6 and alpha = 10.

CpG *versus* non-CpG substitution rate can generally be used in *Alu* subfamily age estimation and other applications for the CpG decay in *Alu* elements less than 50 myrs old.

Increased complexity for CpG *versus* non-CpG decay in older subfamilies

To further examine the relationship between CpG and non-CpG mutations, mean CpG substitution densities were plotted against mean non-CpG substitution densities for each *Alu* subfamily (Figure 3(a) and (b)). To obtain an estimate of the expected ratio of CpG *versus* non-CpG decay over time, the substitution model of Jukes–Cantor[16] was used to determine the substitution density at non-CpG sites. More modern analytical models (e.g. K2P) proved difficult to implement due to the complexity of dissociating CpG and non-CpG transitions and accounting for the recreation of CpG sites, these additional factors were therefore incorporated through computer simulation (see below). Due to the unidirectional decay of CpG dinucleotides, the deficit of CpG dinucleotides over time will result in a non-linear decay of CpG sites. Thus, for CpG sites the decay kinetics can be approximated by $d_{cg} = 1 - e^{(-\mu_{cg}t)10}$ in which μ_{cg} is the mutation rate at CpG sites (see Materials and Methods).

The neutral non-CpG mutation rate of $\mu_{ncg} = 0.0015$ substitutions/site per myr[17,18] and the CpG

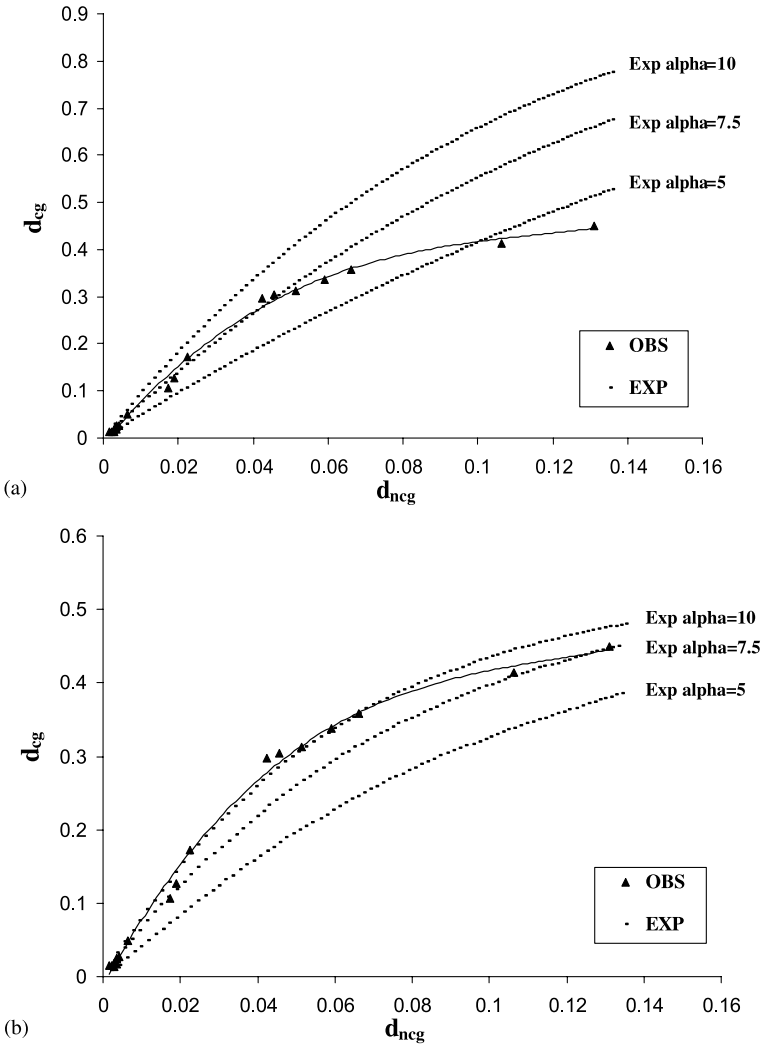

Figure 3. (a) CpG substitution density *versus* non-CpG substitution density in examined *Alu* subfamilies as compared to analytical Jukes–Cantor model. CpG substitution densities (d_{cg}) are plotted against non-CpG substitution densities (d_{ncg}). Triangles represent observed *Alu* subfamily values and a polynomial trendline is shown for the observed data. The broken line represents expected densities based on non-CpG and CpG mutation analytical models over 100 myrs with alpha values of 5, 7.5 and 10 used for comparison (see Materials and Methods). Subfamilies *Alu*Jo and Jb are indicated on the plot. (b) CpG substitution density *versus* non-CpG substitution density in examined *Alu* subfamilies as compared to simulation of *Alu* evolution. Triangles represent observed *Alu* subfamily values and a polynomial trendline is shown for the observed data. The broken line represents expected densities based on non-CpG and CpG mutation computer simulation over 80 myrs with alpha values of 5, 7.5 and 10 used for comparison.

mutation rate calculated as $\mu_{cg} = (\mu_{ncg} \times \text{alpha})$ were used to estimate the expected CpG and non-CpG substitution densities over time using multiple values of alpha (5, 7.5, 10). Since CpG-based age estimates of *Alu* subfamilies in the literature (using alpha = 10) appear to systematically be lower than non-CpG-based age estimates,[12,13] we used an alpha value of ten for our upper boundary. When compared to the expected values, the observed curve for the majority of *Alu* subfamilies fits best when a parameter of alpha = 7.5 is used. However, the two oldest *Alu*J subfamilies showed an apparent decrease in CpG decay that was more pronounced than what was expected under a constant alpha model (Figure 3(a)).

In order to incorporate the regeneration of CpG dinucleotides through back mutation and the disparity between transition and transversion rates, a computer simulation was developed wherein a group of *Alu* elements was allowed to evolve neutrally under a K2P model. The simulation allowed for the adjustment of alpha, the ratio of CpG to non-CpG mutation rates, as well as the rate of non-CpG mutation. Expected CpG substitution densities *versus* non-CpG substitution densities over 80 myrs were simulated with different alpha values (5, 7.5 and 10) and plotted on the graph (Figure 3(b)) (see Materials and Methods for details). The expected curves represent the average of five independent simulation runs. Due to the population size of the elements being simulated, standard deviations were small (on the order of 10^{-3}) and were not indicated on the graph. The resulting simulated expectation curves were similar to that observed in the analytical model. Although they more closely approximated the shape of the observed data, discrepancies were still apparent, indicating further factors are likely involved in the long term CpG decay process than were incorporated in either model.

Discussion

Due to the unidirectional decay of CpG dinucleotides, the deficit of CpG dinucleotides over time will result in a non-linear decay of CpG dinucleotides. In our survey of over 5000 elements that represent most known *Alu* subfamilies, we observed the decay of *R* as expected for the majority of subfamilies. However, we found that the two oldest *Alu* subfamilies, *Alu*Jo and *Alu*Jb, showed a substantially sharper decline in *R* when compared to expectations. This result indicates that our simple model, which held the actual CpG/non-CpG mutation ratio (alpha) and neutral mutation rate to be constant, was not adequate for longer evolutionary time-scales (> 50 myrs).

Multiple factors could plausibly account for the deviation in the decay observed in older *Alu* subfamilies. One factor that would explain a change of alpha is that the neutral non-CpG mutation rate did not remain constant during primate evolution. A molecular clock slowdown during primate evolution has previously been proposed by Goodman *et al.*,[19] and additional studies have provided further support for this hypothesis.[20–22] Such a change in the non-CpG mutation rate may have been the result of an increase in the average generation time of primates that resulted in fewer cell divisions and consequently fewer mutations per calendar year (generation time effect).[19] Other factors, like slower metabolic rates or improved DNA repair systems could have also contributed to lowered levels of mutation in the Hominidae lineage.[20] On the other hand, since the rapid CpG decay is primarily due to DNA methylation, factors that influence the neutral mutation rate (generation time, metabolic rate etc.) may not have as much influence on the CpG mutation rate as on the non-CpG mutation rate.

To investigate this hypothesis, the simulation was adjusted to use different non-CpG mutation rates to account for the possible neutral mutation rate slowdown during primate evolution. Using different neutral mutation rates[20] at non-CpG sites and a constant alpha value (see Materials and Methods), Figure 4(a) illustrates that the simulation model can more closely approximate the observed data points.

On the other hand, there may have also been an increase in the CpG mutation rate during primate evolution. It is well established that the majority of *Alu* elements were integrated into primate genomes 35–50 myrs ago.[7,23] For example, the major *Alu*S subfamilies account for about 62% of the whole *Alu* family and are thought to have expanded throughout primate genomes during this period.[15] A burst of processed pseudogenes[24] and retroviruses[25] may have also occurred during the same time period, indicating an overall retrotransposon explosion at that time. Although the reason for the amplification burst remains unknown, the dramatic increase of retroelement proliferation must have had a substantial impact on primate genomes. As a result, primates likely underwent selective pressure for developing mechanisms to suppress retrotransposon expansion. Increased levels of DNA methylation may have been one such mechanism. Methylation can control retroelements in multiple ways,[1] such as repressing the transcriptional activity of LINEs[26] and increasing the mutation rate in CpG-rich retroelements such as *Alu* elements resulting in retrotranspositional quiescence.[7] One possible explanation for our data is that the CpG decay in the *Alu* family occurred in two major stages: first, there was a relatively slow CpG decay rate during early primate evolution. Next, due to the retroelement explosion 35–50 myrs ago, the rate of CpG methylation increased to counteract retroelement proliferation, which resulted in a higher CpG decay rate. Subsequent to this period, the CpG decay rate has remained effectively constant and high. By holding the neutral non-CpG mutation rate constant at 0.0015 substitutions/site per myr and altering the alpha parameter from an initial

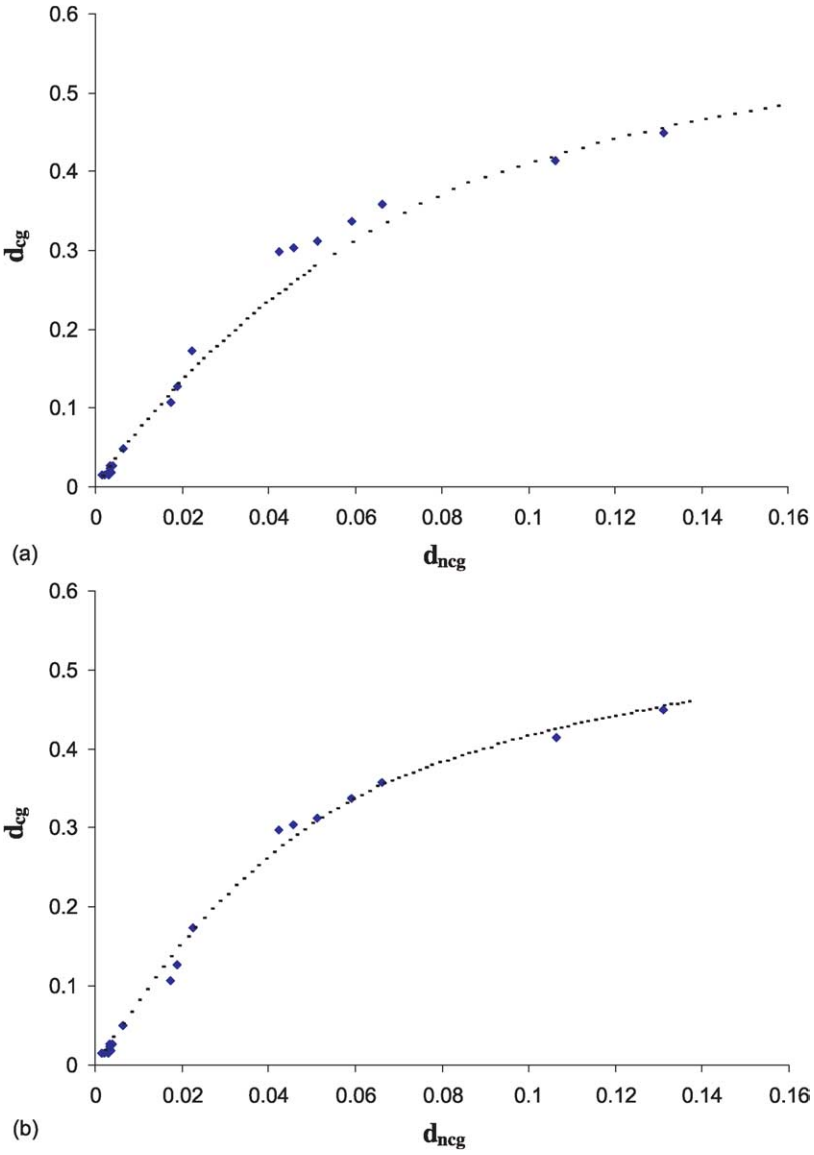

Figure 4. (a) CpG substitution density *versus* non-CpG substitution density in examined *Alu* subfamilies with a neutral mutation rate slowdown model. Diamonds represent observed subfamily values and the broken line represents expected CpG and non-CpG substitution densities based on non-CpG mutation slowdown model. Neutral mutation rate changes from 0.0035 substitutions/site per myr to 0.00117 substitutions/site per myrs at approximately 35 myr. (b) CpG substitution density *versus* non-CpG substitution density in examined *Alu* subfamilies with CpG mutation rate increase model. Diamonds represent observed subfamily values and the broken line represents expected CpG and non-CpG substitution densities based on CpG mutation rate increase model. CpG/non-CpG mutation ratio, alpha, changes from 10.0 to 7.0 at approximately 35 myrs.

value of 10.0 to 7.0 at the 35 myrs retrotransposition burst period in our simulations, we can obtain an approximate fit for the observed data (Figure 4(b)).

Our results indicate that several factors are likely involved in the observed relationship between CpG and non-CpG decay in the human lineage. While a non-CpG mutation rate slowdown is well-documented, it is unlikely that this change is solely responsible for the observed decay pattern, as indicated by Figure 4(a). An additional possibility is that recruitment of the deoxycytosine methylation machinery is more efficient in elements that possess a larger proportion of CpG dinucleotides, perhaps due to a proximity effect. This would lead to decreased methylation and thereby decreased mutation of older, more decayed, *Alu* inserts compared to younger, CpG-rich elements, further exacerbating the non-linearity of the decay process. The molecular mechanism underlying methylation is an active area of research, and it is difficult at present to assess the likelihood and/or extent of such a proximity-based methylation effect. It is also important to consider that we are viewing primate mutational history through the lens of the human lineage, it will be interesting to see if a similar mutation pattern can be observed in the genomes of other non-human primate genomes.

Conclusion

Since *Alu* elements represent 10% of the human genome and are heavily methylated, the pattern we report here may reflect the general CpG decay pattern in humans. Our results indicate that multiple processes will need to be accounted for in order to adequately measure CpG decay over extended evolutionary periods (>50 myrs). However, our results lend strong support for an approximately sixfold difference in CpG *versus* non-CpG mutation acting over recent human evolution.

Materials and Methods

Data collection and multiple alignments

Alu elements from 12 *Alu*Y subfamilies (Ya5, Ya5a2, Ya8, Yb7, Yb8, Yb9, Yc1, Yd3, Yd6, Ye5, Yg6 and Yi6) were obtained from previously published data (see Table 1 for details). *Alu*Yd subfamily members and a random subset of seven *Alu*S and *Alu*J subfamilies were retrieved from human genomic sequences from the July 2003 release of the UC Santa Cruz draft sequence† using Perl scripts and output from "Repeatmasker"‡ software. Alignments among members of each subfamily were obtained using Clustal X.[27] The resulting multiple alignments were subjected to further manual adjustments by removing insertions and poly(A) tails from all *Alu* elements. Elements that contained deletions larger than 50 bp and/or that could not be faithfully aligned were also excluded from the analysis. The alignment of each subfamily is available at our website§ under publications. Substitutions in individual *Alu* elements were recorded based on the consensus sequence of each subfamily[9,28] and divided into "CpG" and "non-CpG" substitutions using a PERL script. For CpG sites, only C to T or G to A mutations were counted. The CpG substitution density (d_{cg}) and non-CpG substitution density (d_{ncg}) were obtained by dividing the total number of observed substitutions by the number of CpG and non-CpG sites, respectively, based on the consensus sequence of each subfamily. The ratio of d_{cg}/d_{ncg} is denoted as R.

Expected substitution density in *Alu* elements using an analytical model of *Alu* sequence evolution

The substitution model of Jukes–Cantor[16] was used to determine the substitution density at non-CpG sites. The non-CpG substitution density within *Alu* elements can be described as: $d_{ncg} = 3/4(1 - e^{-4/3*\mu_{ncg}*t})$, where μ_{ncg} is the neutral mutation rate at non-CpG sites and t is the time in years of the *Alu* elements integration (or age of the elements). For CpG sites the decay kinetics can be described as $d_{cg} = 1 - e^{(-\mu_{cg}t)}$ in which μ_{cg} is the mutation rate at CpG sites.[29] Thus, the expected ratio of CpG to non-CpG dinucleotide substitution density, designated here as R_{exp} is the quotient of the equations for d_{ncg} and d_{cg} above:

$$R_{exp}(t) = [1 - e^{(-\mu_{cg}t)}]/[3/4(1 - e^{-4/3*\mu_{ncg}*t})]$$

Expected substitution density in *Alu* elements using computational simulations of *Alu* sequence evolution

One hundred perfect copies of *Alu*Y consensus were initially generated. For each evolutionary time increment (50,000 years), *Alu* elements accumulate nucleotide substitutions at designated mutation rates. For the non-CpG sites, the mutation process was simulated using a two-parameter reversible mutation model (K2P)[30] with mutation rate of 0.0015 substitutions/site per myr[17,18] and a 4× ratio of transitions to transversions.[31] For the CpG dinucleotide, the CpG mutation rate was calculated as $\mu_{cg} = (\mu_{ncg} \times \text{alpha})$ to simulate the expected CpG substitution densities over time using multiple values of alpha (5, 7.5, 10). Once mutated, CpG locations revert to the non-CpG rate unless they are reconstituted through back mutations.

For the neutral mutation rate of the slowdown model, different neutral mutation rates were used during different windows of evolutionary time.[20] In detail, $\mu_{ncg} = 0.0010$ substitutions/site per myr from 1 myrs to 35 myrs and $\mu_{ncg} = 0.0035$ substitutions/site per myr from 36 myrs to 80 myrs. The non-CpG mutation rate was held constant as 0.00117 substitutions/site per myr, which is the estimated value for recent human evolution.[17] The expected CpG mutation densities and non-CpG substitution densities were plotted over 80 myrs with one myr intervals.

† http://genome.ucsc.edu/
‡ http://www.repeatmasker.org/
§ http://batzerlab.lsu.edu

Acknowledgements

We thank Drs Randall Garber and David Ray for critical reading and suggestions during the preparation of this manuscript. This research was supported by the Louisiana Board of Regents Millennium Trust Health Excellence Fund HEF (2000-05)-01 (M.A.B.), National Science Foundation BCS-0218338 (M.A.B.) and EPS-0346411 (M.A.B.) and the State of Louisiana Board of Regents Support Fund (M.A.B.).

References

1. Schmid, C. W. (1998). Does SINE evolution preclude Alu function? *Nucl. Acids Res.* **26**, 4541–4550.
2. Paulsen, M. & Ferguson-Smith, A. C. (2001). DNA methylation in genomic imprinting, development, and disease. *J. Pathol.* **195**, 97–110.
3. Greenblatt, M. S., Bennett, W. P., Hollstein, M. & Harris, C. C. (1994). Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. *Cancer Res.* **54**, 4855–4878.
4. Tweedie, S., Charlton, J., Clark, V. & Bird, A. (1997). Methylation of genomes and genes at the invertebrate-vertebrate boundary. *Mol. Cell Biol.* **17**, 1469–1475.
5. Coulondre, C., Miller, J. H., Farabaugh, P. J. & Gilbert, W. (1978). Molecular-basis of base substitution hotspots in *Escherichia coli*. *Nature*, **274**, 775–780.
6. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J. *et al*. (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
7. Batzer, M. A. & Deininger, P. L. (2002). Alu repeats and human genomic diversity. *Nature Rev. Genet.* **3**, 370–379.
8. Schmid, C. W. (1991). Human Alu subfamilies and their methylation revealed by blot hybridization. *Nucl. Acids Res.* **19**, 5613–5617.
9. Batzer, M. A., Deininger, P. L., Hellmann-Blumberg, U., Jurka, J., Labuda, D., Rubin, C. M. *et al*. (1996). Standardized nomenclature for Alu repeats. *J. Mol. Evol.* **42**, 3–6.
10. Labuda, D. & Striker, G. (1989). Sequence conservation in Alu evolution. *Nucl. Acids Res.* **17**, 2477–2491.
11. Batzer, M. A., Kilroy, G. E., Richard, P. E., Shaikh,

T. H., Desselle, T. D., Hoppens, C. L. *et al*. (1990). Structure and variability of recently inserted Alu family members. *Nucl. Acids Res.* **18**, 6793–6798.
12. Xing, J. C., Salem, A. H., Hedges, D. J., Kilroy, G. E., Watkins, W. S., Schienman, J. E. *et al*. (2003). Comprehensive analysis of two Alu Yd subfamilies. *J. Mol. Evol.* **57**, S76–S89.
13. Carter, A. B., Salem, A.-H., Hedges, D. J., Nguyen Keegan, C., Kimball, B., Walker, J. A. *et al*. (2004). Genome wide analysis of the human Alu Yb lineage. *Hum. Genomics*, **1**, 167–178.
14. Carroll, M. L., Roy-Engel, A. M., Nguyen, S. V., Salem, A. H., Vogel, E., Vincent, B. *et al*. (2001). Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. *J. Mol. Biol.* **311**, 17–40.
15. Zietkiewicz, E., Richer, C., Sinnett, D. & Labuda, D. (1998). Monophyletic origin of Alu elements in primates. *J. Mol. Evol.* **47**, 172–182.
16. Jukes, T. H. & Cantor, C. R. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism* (Munro, H. N., ed), pp. 21–132, Academic Press, New York.
17. Nachman, M. W. & Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*, **156**, 297–304.
18. Miyamoto, M. M., Slightom, J. L. & Goodman, M. (1987). Phylogenetic relations of humans and African apes from DNA sequences in the psi eta-globin region. *Science*, **238**, 369–373.
19. Goodman, M. (1985). Rates of molecular evolution: the hominoid slowdown. *Bioessays*, **3**, 9–14.
20. Bailey, W. J., Fitch, D. H., Tagle, D. A., Czelusniak, J., Slightom, J. L. & Goodman, M. (1991). Molecular evolution of the psi eta-globin gene locus: gibbon phylogeny and the hominoid slowdown. *Mol. Biol. Evol.* **8**, 155–184.
21. Yi, S., Ellsworth, D. L. & Li, W. H. (2002). Slow molecular clocks in Old World monkeys, apes, and humans. *Mol. Biol. Evol.* **19**, 2191–2198.
22. Page, S. L. & Goodman, M. (2001). Catarrhine phylogeny: noncoding DNA evidence for a diphyletic origin of the mangabeys and for a human-chimpanzee clade. *Mol. Phylogenet. Evol.* **18**, 14–25.
23. Britten, R. J. (1994). Evidence that most human Alu sequences were inserted in a process that ceased about 30 million years ago. *Proc. Natl Acad. Sci. USA*, **91**, 6148–6150.
24. 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**, R74.
25. Sverdlov, E. D. (2000). Retroviruses and primate evolution. *Bioessays*, **22**, 161–171.
26. Yu, F. Z. N., Schumann, G. & Stratling, W. H. (2001). Methyl-CpG-binding protein 2 represses LINE-1 expression and retrotransposition but not Alu transcription. *Nucl. Acids Res.* **29**, 4493–4501.
27. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl. Acids Res.* **25**, 4876–4882.
28. Jurka, J., Krnjajic, M., Kapitonov, V. V., Stenger, J. E. & Kokhanyy, O. (2002). Active Alu elements are passed primarily through paternal germlines. *Theor. Popul. Biol.* **61**, 519–530.
29. Sved, J. & Bird, A. (1990). The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc. Natl Acad. Sci. USA*, **87**, 4692–4696.
30. Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120.
31. Li, W.-H. (1997). *Molecular Evolution*, pp. 62–63, Sinauer Associates, Sunderland, MA.
32. Roy, A. M., Carroll, M. L., Nguyen, S. V., Salem, A. H., Oldridge, M., Wilkie, A. O. *et al*. (2000). Potential gene conversion and source genes for recently integrated Alu elements. *Genome Res.* **10**, 1485–1495.
33. Roy, A. M., Carroll, M. L., Kass, D. H., Nguyen, S. V., Salem, A. H., Batzer, M. A. *et al*. (1999). Recently integrated human Alu repeats: finding needles in the haystack. *Genetica*, **107**, 149–161.
34. Roy-Engel, A. M., Carroll, M. L., Vogel, E., Garber, R. K., Nguyen, S. V., Salem, A. H. *et al*. (2001). Alu insertion polymorphisms for the study of human genomic diversity. *Genetics*, **159**, 279–290.
35. Garber, R. K., Hedges, D. J., Herke, S. W., Hazard, N. W. &Batzer, M. A. (2005). The Alu Yc1 subfamily: sorting the wheat from the chaff. *Cytogenet. Genome Res.*, in press.
36. Salem, A. H., Kilroy, G. E., Watkins, W. S., Jorde, L. B. & Batzer, M. A. (2003). Recently integrated Alu elements and human genomic diversity. *Mol. Biol. Evol.* **20**, 1349–1361.
37. Salem, A. H., Ray, D. A., Xing, J., Callinan, P. A., Myers, J. S., Hedges, D. J. *et al*. (2003). Alu elements and hominid phylogenetics. *Proc. Natl Acad. Sci. USA*, **100**, 12787–12791.

Edited by J. Karn

(Received 3 September 2004; received in revised form 21 September 2004; accepted 22 September 2004)