Recombination-Associated Sequence Homogenization of Neighboring Alu Elements: Signature of Nonallelic Gene Conversion

Alexey Aleshin and Degui Zhi

1Department of Medicine, Division of Hematology, Oncology, David Geffen School of Medicine, University of California, Los Angeles
2Department of Biostatistics, University of Alabama at Birmingham

*Corresponding author: E-mail: dzhi@ms.soph.uab.edu.
Associate editor: Koichiro Tamura

Abstract

Recently, researchers have begun to recognize that, in order to establish neutral models for disease association and evolutionary genomics studies, it is crucial to have a clear understanding of the genomic impact of nonallelic gene conversion. Drawing on previous successes in characterizing this phenomenon over protein-coding gene families, we undertook a computational analysis of neighboring Alu sequences in the genome scale. For this purpose, we developed adjusted comutation rate (acMR), a novel statistical method measuring the excess number of identical point mutations shared by adjacent Alu sequences, vis-à-vis random pairs. Using acMR, we uncovered a remarkable genome-wide sequence homogenization of neighboring Alus, with the strongest signal observed in the pseudoautosomal regions of the X and Y chromosomes. The magnitude of sequence homogenization between Alu pairs is greater with shorter interlocus distance, higher sequence identity, and parallel orientation. Moreover, shared substitutions show a strong directionality toward GC nucleotides, with multiple substitutions tending to cluster within the Alu sequence. Taken together, these observed recombination-associated sequence homogenization patterns are best explained by frequent ubiquitous gene conversion events between neighboring Alus. We believe that these observations help to illuminate the nature and impact of the enigmatic phenomenon of gene conversion.

Key words: Alu, gene conversion, recombination, crossing over, genetic/*genetics, models, genetic, mutation.

Introduction

Recombination-associated genome alterations and their association with numerous genetic disorders have, as of late, drawn increasing attention (e.g., Mefford et al. 2008). These alterations have been shown to drive the evolution of genomes, creating deletions, duplications, and rearrangements. In particular, one such alteration, biased gene conversion, has been recognized to have a profound impact on the mutational landscape of genomes, as well as on the maintenance of the genome’s high GC content (Eyre-Walker 1993 and references since then; see Duret and Galtier 2009 for a recent review).

Gene conversion commonly refers to the recombination-associated DNA heteroduplex mismatch repair, where the template strand (donor) “converts” the “acceptor” strand, thereby resulting in a unidirectional sequence transfer. Traditionally, gene conversion has been considered as being a poorly characterized side product of recombination. However, recent progress has added substantially to our understanding of the molecular mechanisms underpinning gene conversion, as well as its impacts on genome evolution and the etiology of inherited diseases (see Chen et al. 2007 for a recent review).

Gene conversion, much like recombination, can occur between either allelic or nonallelic loci. In this work, we focus on nonallelic gene conversion (NAGC). NAGC, although being a largely neutral process, can nevertheless have a profound impact on the evolution of the genome. Recurring NAGC events among arrays of homologous genes act to create a high degree of allelic polymorphism in gene families. One of the most studied examples of gene conversion is the human major histocompatibility complex (MHC) region, in which the frequency of gene conversion is estimated to be as high as $10^{-5}$ to $10^{-6}$ events per domain per generation (Ohta 1982; Högstrand and Böhme 1994). Gene conversion is suggested to be responsible for the higher than expected number of single-nucleotide polymorphisms (SNPs) in segmental duplication regions as opposed to nonduplicated regions of the human genome (Hurles 2002). Paradoxically, NAGC also acts to homogenize the nucleotide sequences of the donor and acceptor loci. For example, the nucleotide identity among orthologous mouse and human loci in the MHC region is only slightly higher than the value observed for the paralogous MHC loci in human (Ploegh et al. 1981). NAGC is also found to contribute to reticulate evolution within human segmental duplications (Jackson et al. 2005). A more striking example is the palindromic regions located within the human and chimpanzee Y chromosomes, which have been reported (Rozen et al. 2003; Hughes et al. 2010) to contain ultrahigh intrapalindromic sequence identities typically above 99.9%. This observation contrasts sharply with the genomic average of 98.66% sequence identity between orthologous palindromic arms.
Based on a few DNA polymorphism data sets from well-studied genomic regions undergoing concerted evolution, the following coarse scale characteristics of NAGC emerge (Chen et al. 2007). First, most gene conversion events occur between homologous sequences with >95% sequence identity. Second, the rate of gene conversion is positively correlated with the length of the longest uninterrupted sequence identity tract between the donor and acceptor loci, with the minimal processing segment for efficient meiotic homologous recombination in human estimated to be 337–456 bp (Reiter et al. 1998). Thirdly, the rate of gene conversion is inversely correlated with the distance between the donor and acceptor loci (multiple references, see Schildkraut et al. 2005, for an example). These findings are largely confirmed in the human genome by a recent survey (Benovoy and Drouin 2009). Although these observations are instrumental to our understanding of NAGC, it is desirable to be able to examine and characterize NAGC events on a genome-wide basis.

In this work, we study the properties of NAGC via a genome-wide computational sequence analysis of the most abundant repeat element in the human genome—the Alu retrotransposon (more than 1 million copies). Alus have been a popular choice for estimating a variety of genomic evolutionary parameters including nucleotide composition and substitution bias (Arndt et al. 2003; Webster et al. 2005) because of the following technical conveniences: 1) the ubiquitous distribution of Alus in the genome and 2) the relatively weak evolutionary forces exerted upon Alu sequences.

Do NAGC events indeed occur between Alus? The answer is yes. Although individual Alu gene conversion events are extremely difficult to ascertain, they are highly likely to occur due to the following arguments. First, 283-bp long Alus are valid substrates for NAGC, with studies in cultured cells suggesting that NAGC can be initiated by as few as 134–232 bp of perfect sequence overlap (Waldman and Liskay 1988). Additionally, anecdotal cases of Alu gene conversion have been reported (Kass et al. 1995) with (Roy et al. 2000) estimating that gene conversion events could affect 10–20% of the Alus in the Ya5 subfamily. Recent comparisons of the chimpanzee and human genomes (Chimpanzee Sequencing and Analysis Consortium 2005; Sen et al. 2006) revealed that homologous reorganizations between neighboring Alu elements (unequal crossovers) are responsible for about 492–612 deletions in the human genome since the human–chimp split. Considering the fact that the rate of allelic gene conversion is an order of magnitude higher than the rate of crossover (Jeffreys and May 2004), we postulate that there could be thousands of NAGC events between Alus since the human–chimp split. Even so, such simplistic reasoning is likely to underestimate the true rate of gene conversion because Alu-Alu–mediated duplications/deletions are subject to potentially high evolutionary pressure, whereas Alu-Alu–mediated NAGC events are practically free of any selective pressures because such events only result in single-nucleotide substitutions in the repetitive Alu sequences themselves.

In an earlier study (Zhi 2007), one of the authors identified that neighboring Alu pairs in human chromosome 22 shared sequence similarity above what would be expected by chance alone. In particular, a statistical measure (termed comutation) was introduced in order to quantify the surplus rate of nucleotide substitutions shared by neighboring Alus over the background rate. The observed sequence homogenization was hypothesized to be the footprint of gene conversion.

In this study, we extend the previous study (Zhi 2007) to encompass the entire genome. Furthermore, we carry out an extensive characterization of the Alu sequence homogenization in the context of currently accepted gene conversion models and their evolutionary impacts. In particular, we develop a novel statistic to measure the elevation of comutation rate (CMR) within an Alu pair that accounts for the ages of the Alus. As a result, this new statistic effectively characterizes intrachromosomal variation of the Alu sequence homogenization pattern. By documenting these patterns and correlating them with the current understanding of gene conversion, we verify the result of Zhi (2007) that the observed sequence homogenization among neighboring Alu sequences is primarily due to NAGCs. Finally, based on the genome-wide data, we are able to characterize the general phenomenon of NAGC in the human genome with more quantitative detail.

Materials and Methods

Sequence Identification

Sequence data were obtained from the completed human genome sequence (UCSC release May 2004). Co-ordinates of putative Alus were downloaded from the UCSC database and extended by 40 bp on each side to insure that full-length sequences were identified. The excised sequences were aligned against the consensus Alu sequence (RepBase) using basic local alignment search tool (Blast) (Altschul et al. 1990) and only the sequences of length ≥280 bp were retained (full length). In order to minimize inconsistencies in the alignment boundaries, Alus with alignments not covering base 1 and base 283 were excluded from the study, as were Alus with substantial internal deletions or insertions (≥5 bp). Final analysis was performed on 714,317 Alu (714,293 adjacent pairs) sequences, of the 929,838 putative Alu sequences identified by RepeatMasker. All data manipulation was performed using custom scripts written in Perl and R (R-project release 2.4.0).

To identify potentially duplicate Alu pairs, Blast alignment (“bl2seq –p BlastN –G 2 –E 1 –F F”) was performed for each 300 bp segment immediately 5’ to each Alu within an Alu pair. If the flanking regions contained an alignment (about 200,000 instances), the corresponding Alu pair was classified as arising from segmental duplication, otherwise the Alu pair was classified as nonsegmentally duplicated (about 500,000 instances).

Definition of Comutation

Analysis of sequence correlation between neighboring Alus was approximated by the pairwise CMR. The CMR was...
defined as the number of identical mutations away from the consensus sequence shared by neighboring Alus. Only nucleotide substitutions were counted as comutations, with insertions and deletions being excluded from analysis. Furthermore, CMR was calculated only at non-CpG and nondiagnostic positions. We excluded CpG nucleotides because of their high rate of mutation. Diagnostic positions (positions that define Alu subfamilies) were excluded because a majority of the comutations at these positions are byproduct of a common origin. Additionally, nucleotide positions 58, 207, and 278 were excluded because these positions displayed a CMR approximately 100-fold higher than other non-CpG nondiagnostic positions. We speculate that these positions may act to define as yet unidentified Alu subfamilies.

Parameters Associated with Neighboring Alus
In order to observe which factors best correlate with CMR, a number of parameters were calculated for each Alu pair, including

1. Percent sequence similarity—calculated by noting the number of identical nucleotides shared by a pair of Alus divided by the length of the Alu sequence (283 bp).
2. Longest stretch of perfect identity—length of the longest uninterrupted tract of nucleotide identity shared by a pair of Alus.
3. Distance between Alu instances—the distance separating Alus within an Alu pair was defined as the length of Alu (283 bp) plus the length of the region in between the two Alus that is free of full-length Alus. For parallel Alu pairs, this is the distance between the potential recombining sites during a gene conversion. Between antiparallel Alu pairs, the distance between the potential recombining sites varies and this value is the average distance between their recombining sites.

Adjusting CMR for Sequence Similarity
We design three levels of adjustments to account for homoplasy in measuring the CMR: the chromosomal level, the intrachromosomal regional level, and the individual Alu pair level.

At the chromosomal level, following a similar procedure to Zhi (2007), the average elevation of CMR was calculated for each of the 22 autosomes and 2 sex chromosomes. To quantify the degree of adjusted comutation rate (aCMR) elevation above the rate due to homoplasy alone, a bootstrapped estimate of the aCMR for random Alu pairs was generated. For each of the 24 chromosomes, 500 independent bootstrapped trials were conducted using a random sample of 20,000 Alu sequences. The bootstrapped distribution of the aCMR was approximated by a normal distribution. Significance testing was performed using a standard Wald test (Engle 1983).

We define an Alu pair–level statistic, aCMR, to measure the excess in the observed CMR above what would be expected due to chance alone. We normalized the CMR between a pair of Alus by dividing the observed rate by the expected CMR, defined as $k \mu_1 \mu_2$, where $\mu_1$ and $\mu_2$ are the rates of substitutions among nondiagnostic non-CpG positions of Alu1 and Alu2. Assuming the mutations in Alu1 and Alu2 occur independently, $k \mu_1 \mu_2$ estimates the joint probability of having a mutation at the same sites between two Alus. $k$ is a constant reflecting the intrachromosomal mutation rate and the probability of two independent substitutions creating the same base pair at equivalent positions between the two Alus. Although $k$ can be computed from a nucleotide substitution model, in this work, $k$ was chosen to normalize the genomic bootstrapped aCMR to 1, as determined by a sample of 714,293 random Alu pairs. Following adjustment, aCMR is expected to be 1 for random Alu pairs, and an aCMR of $>1$ signifies enriched sequence homogenization.

The aCMR is a more refined adjustment than that in Zhi (2007), where the CMRs were only grossly adjusted by the average CMRs throughout an entire chromosome. The added resolution in aCMR allows our analysis of the sequence homogenization at the level of individual Alu pairs.

Although this adjustment facilitates the characterization of factors affecting individual Alu pairs, this measure is numerically unstable because Alus with very low mutation rates contain comutations are given disproportionate weights. Therefore, we introduce the aCMR-$n$ measure. At the intrachromosomal regional level, we derive an aggregated of CMR among $n$ “contiguous” Alu pairs and compare it against the aggregated estimate of CMR derived from $n$ Alu pairs with the same local nucleotide substitution frequencies. In practice, we found that aCMR-15 maintains the ability to finely resolve intrachromosomal variation of aCMR while maintaining numerical stability.

Biased GC Comutation
The tendency of comutations to be directed toward GC nucleotides was calculated by dividing the number of AT to GC comutations by the total number of comutations. To distinguish the observed trend from an artifactual bias, a random sample of 714,293 Alus was drawn and the same pairwise analysis was performed. Similar analyses were performed for GC to AT comutations.

Average Distance between Comutation Sites
To measure the clusteredness of comutation sites, the average distance between comutations was calculated by dividing the sum of the distances between all pairwise combinations of comutations by the number of such events. Alu pairs having one or less comutations were excluded from this analysis. Again, a random sample of 714,293 Alu pairs was generated in order to approximate the relationship between these two variables that would be expected through chance alone.

Results and Discussions
To minimize potential artifacts, other than point mutations, we focused our analysis on full-length Alu sequences devoid of gross deletions and insertions. After careful filtering, our data set consisted of 714,317 Alus (see Materials
and Methods). For each pair of adjacent full-length Alus, we counted the number of comutations, that is, identical single base pair deviations from the Alu consensus sequence, obtained from RepBase (Jurka 2000). CMR was defined as the number of comutations per base pair per Alu pair. In order to avoid possible biases from well-known patterns of Alu mutations, we limited our analysis to positions that were neither CpG in the Alu consensus sequence nor Alu subfamily–specific nucleotides (non-CpG nondiagnostic).

CMR Is Elevated Across the Genome
We found that the CMR of adjacent Alu pairs is elevated across the genome. The average rate of comutation at non-CpG nondiagnostic positions is 0.00267 per base pair per Alu pair, vis-à-vis, the background rate of 0.00254 as measured by the mean derived from 500 bootstrapping runs using random intrachromosomal Alu pairs (random Alu pairs for short). This elevation of CMR occurs in all chromosomes (supplementary fig. S1, Supplementary Material online) (supplementary table S2, Supplementary Material online). However, we noticed that the mean CMR of random Alu pairs varies widely from one chromosome to the next, reflecting possible chromosome-specific factors influencing the intensity of CMR. The accurate measurement of CMR is confounded by the prevailing mutation rates acting upon individual Alus, which are known to vary within and among chromosomes (Duret 2009). To control for the variation in local mutation rate, we defined the aCMR (aCMR), calculated as the ratio of the observed CMR to the expected CMR between a pair of Alus (see Materials and Methods section for details). The distribution of Alu pairs in relation to various ranges of aCMR is shown in supplementary table S1, Supplementary Material online.

Compared with CMR, aCMR shows much less variability among individual chromosomes (fig. 1; table 1). aCMR of adjacent Alu pairs is elevated across all chromosomes vis-à-vis the background. This elevation is statistically significant (Z value > 2) for all but chromosomes 10 and 21.

The elevation of CMR between adjacent Alu pairs is significantly higher in the sex chromosomes than in any of the autosomes. The elevation of aCMR above the bootstrapped baseline is 0.266 (Z value = 8.47) for chromosome X and 0.683 (Z value = 26.66) for chromosome Y, whereas for the autosomes, the elevation of aCMRs ranges from 0.036 (Z value = 1.19) for chromosome 21 to 0.138 (Z value = 5.35) for chromosome 15, with an average elevation of 0.100 and an average Z value of 3.57.

Intriguingly, the unusually high aCMRs of X and Y chromosomes are largely attributed to the 2.7 M bp-long pseudoautosomal region (PAR)-1. The Z values are 12.03 for X PAR1 and 12.54 for Y PAR1, whereas the Z values are 3.87 and 3.83 for the rest of the chromosomes X and Y, respectively, inline with the genome average of Z values (3.57). These results extend the preliminary observations in Zhi (2007) and confirm that the homogenization of adjacent Alu sequences is a global phenomenon acting throughout the human genome. Thus, we set out to characterize the patterns of this sequence homogenization.

aCMR Decays with Increasing Interlocus Distance
We investigated the correlations of the observed sequence homogenization, as measured by aCMR, with several features of Alu pairs, including their interlocus distance, relative orientation, and sequence similarity. We classified Alu pairs from all 24 chromosomes into different categories according to these features and analyzed their aCMR measures.

We observed that the sequence homogenization is a short-range effect (fig. 2), confirming the similar result in a previous study of chromosome 22 (Zhi 2007). The aCMR is high for tandemly arranged Alu’s (<300 bp) and rapidly falls off with increasing interlocus distance. At a distance of 10 kb, the aCMR approaches that of
random Alu pairs. The observed trend of aCMR decay with distance is consistent with the hypothesis that the observed sequence homogenization is caused by NAGC. However, the short distances between Alus with high aCMR may imply that these Alus reside in Alu-dense regions of the genome. Thus, it is possible that regional Alu density, GC content, and gene density (all of which largely correlate with each other) affect the rate of comutation more than inter-Alu distance does. We tested the Spearman correlation (rho) of aCMR with these genome parameters over 1 Mb windows. We found that aCMR is correlated with Alu density (rho = 0.020, P value = 0.27), GC content (rho = 0.025, P value = 0.17), or gene density (rho = 0.016, P value = 0.38).

Moreover, we observed an intriguing nuance in this overall pattern: Although the trend of aCMR decay is apparent for Alu pairs with distance ranges <700 bp and >3,000 bp, aCMR remains at a relatively stable level for Alu pairs in the 700—3,000 bp distance range. Our data show that Alu-Alu-mediated sequence deletion may have contributed to this trend (see the subsection “Impact of Sequence Homogenization on Percent Identity between Neighboring Alus” for more details).

In addition, the relative orientation of the two Alus comprising an Alu–Alu pair plays a significant role in the magnitude of the observed aCMR. Parallel Alus display a much more significant aCMR vis-à-vis, their antiparallel counterparts, both by visual inspection (fig. 2) and by quantitative measure (P ≤ 10^-4, Wilcoxon rank sum test with continuity correction). This difference is more pronounced in the short distance ranges. This phenomenon may be explained by one of the two explanations. The first explanation is that the relative insertion orientation of sequences determines short-range nonallelic recombination. Presumably, short-range nonallelic recombination between parallel Alus results from the relative slippage in recognition of allelic loci, whereas antiparallel Alus can only be paired following the formation of a DNA loop. However, to the best of our knowledge, there are no previous reports describing the influence of relative orientation of homologous sequences on the frequency of gene conversion.

Alternatively, it may be more plausible to consider the explanation that antiparallel Alu pairs at a close distance may be under strong purifying selection at both DNA and RNA levels. Previous studies have demonstrated that antiparallel Alus at a close distance may create genomic instability (Stenger et al. 2001). Moreover, the homogenization of inverted Alus at a close distance would increase the possibility that the two Alus form a stem when transcribed into an RNA. Recently, it has been reported that stem–loop structures mediated by inverted Alus in the 3’-untranslated region of mRNAs are subjects for RNA editing and cause repression of gene expression (Chen et al. 2008). Given that most regions of human genome are transcribed and that the formation of Alu stem loop may often be harmful, it is likely that antiparallel Alu pairs are also under selection at the RNA level.

### Table 1. Elevation of CMRs Across All Human Chromosomes.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>aCMR</th>
<th>Observed Bootstrapped</th>
<th>Standard Deviation</th>
<th>Elevation</th>
<th>Z</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.085</td>
<td>1.025</td>
<td>0.016</td>
<td>0.06</td>
<td>3.66</td>
<td>1.26E-04</td>
</tr>
<tr>
<td>2</td>
<td>1.057</td>
<td>1.011</td>
<td>0.014</td>
<td>0.047</td>
<td>3.3</td>
<td>4.83E-04</td>
</tr>
<tr>
<td>3</td>
<td>1.045</td>
<td>1.007</td>
<td>0.015</td>
<td>0.038</td>
<td>2.53</td>
<td>5.70E-03</td>
</tr>
<tr>
<td>4</td>
<td>1.046</td>
<td>0.989</td>
<td>0.013</td>
<td>0.057</td>
<td>4.58</td>
<td>2.32E-06</td>
</tr>
<tr>
<td>5</td>
<td>1.044</td>
<td>0.997</td>
<td>0.014</td>
<td>0.047</td>
<td>3.27</td>
<td>5.38E-04</td>
</tr>
<tr>
<td>6</td>
<td>1.056</td>
<td>1.009</td>
<td>0.014</td>
<td>0.047</td>
<td>3.26</td>
<td>5.57E-04</td>
</tr>
<tr>
<td>7</td>
<td>1.057</td>
<td>0.998</td>
<td>0.014</td>
<td>0.059</td>
<td>4.26</td>
<td>1.02E-05</td>
</tr>
<tr>
<td>8</td>
<td>1.035</td>
<td>1</td>
<td>0.013</td>
<td>0.035</td>
<td>2.63</td>
<td>4.27E-03</td>
</tr>
<tr>
<td>9</td>
<td>1.047</td>
<td>1.002</td>
<td>0.015</td>
<td>0.045</td>
<td>3.11</td>
<td>9.35E-04</td>
</tr>
<tr>
<td>10</td>
<td>1.033</td>
<td>1.008</td>
<td>0.015</td>
<td>0.026</td>
<td>1.72</td>
<td>4.27E-02</td>
</tr>
<tr>
<td>11</td>
<td>1.064</td>
<td>1.007</td>
<td>0.014</td>
<td>0.057</td>
<td>4.13</td>
<td>1.81E-05</td>
</tr>
<tr>
<td>12</td>
<td>1.066</td>
<td>1.018</td>
<td>0.016</td>
<td>0.048</td>
<td>2.99</td>
<td>1.39E-03</td>
</tr>
<tr>
<td>13</td>
<td>1.033</td>
<td>0.986</td>
<td>0.015</td>
<td>0.047</td>
<td>3.1</td>
<td>9.68E-04</td>
</tr>
<tr>
<td>14</td>
<td>1.067</td>
<td>1.025</td>
<td>0.015</td>
<td>0.042</td>
<td>2.84</td>
<td>2.26E-03</td>
</tr>
<tr>
<td>15</td>
<td>1.089</td>
<td>1.016</td>
<td>0.014</td>
<td>0.073</td>
<td>5.35</td>
<td>4.40E-08</td>
</tr>
<tr>
<td>16</td>
<td>1.051</td>
<td>0.989</td>
<td>0.016</td>
<td>0.062</td>
<td>3.96</td>
<td>3.75E-05</td>
</tr>
<tr>
<td>17</td>
<td>1.074</td>
<td>1.015</td>
<td>0.016</td>
<td>0.059</td>
<td>3.74</td>
<td>9.20E-05</td>
</tr>
<tr>
<td>18</td>
<td>1.07</td>
<td>0.998</td>
<td>0.015</td>
<td>0.071</td>
<td>4.84</td>
<td>6.49E-07</td>
</tr>
<tr>
<td>19</td>
<td>1.06</td>
<td>0.987</td>
<td>0.014</td>
<td>0.073</td>
<td>5.12</td>
<td>1.53E-07</td>
</tr>
<tr>
<td>20</td>
<td>1.1</td>
<td>1.028</td>
<td>0.016</td>
<td>0.072</td>
<td>4.5</td>
<td>3.40E-06</td>
</tr>
<tr>
<td>21</td>
<td>1.001</td>
<td>0.981</td>
<td>0.016</td>
<td>0.019</td>
<td>1.19</td>
<td>1.17E-01</td>
</tr>
<tr>
<td>22</td>
<td>1.097</td>
<td>1.022</td>
<td>0.017</td>
<td>0.075</td>
<td>4.54</td>
<td>2.81E-06</td>
</tr>
<tr>
<td>X</td>
<td>1.14</td>
<td>1.004</td>
<td>0.017</td>
<td>0.14</td>
<td>8.47</td>
<td>&lt;1.00E-12</td>
</tr>
<tr>
<td>Y</td>
<td>1.23</td>
<td>0.877</td>
<td>0.013</td>
<td>0.358</td>
<td>26.66</td>
<td>&lt;1.00E-12</td>
</tr>
<tr>
<td>X PAR1</td>
<td>1.894</td>
<td>1.171</td>
<td>0.06</td>
<td>0.724</td>
<td>12.03</td>
<td>&lt;1.00E-12</td>
</tr>
<tr>
<td>Y PAR1</td>
<td>1.903</td>
<td>1.155</td>
<td>0.06</td>
<td>0.748</td>
<td>12.54</td>
<td>&lt;1.00E-12</td>
</tr>
<tr>
<td>X non-PAR1</td>
<td>0.983</td>
<td>0.815</td>
<td>0.043</td>
<td>0.168</td>
<td>3.87</td>
<td>5.44E-05</td>
</tr>
<tr>
<td>Y non-PAR1</td>
<td>0.788</td>
<td>0.623</td>
<td>0.043</td>
<td>0.165</td>
<td>3.84</td>
<td>6.15E-05</td>
</tr>
</tbody>
</table>

**NOTE.**—See caption of figure 1 for details.
We found that Alu pairs with highly similar sequences display a higher aCMR than Alu pairs with divergent sequences. We used both the gross sequence percent identity within the entire 283 bp of Alu and the length of longest stretch of uninterrupted sequence identity as surrogate indexes for Alu–Alu similarity. Our results (fig. 3) show that aCMR is significantly enriched in Alu pairs with long stretches of identity. The effect becomes apparent with as few as 35–40 bp of perfect identity and increases considerably with longer stretches. Again, the orientation of the Alu pairs appears to be a significant contributing factor, with parallel Alus displaying a greater elevation of CMR with increasing stretches of sequence identity. Using sequence identity as surrogate index gives essentially similar results (data not shown).

Moreover, previous studies suggest that the minimal length of gene conversion tracts in human genome can be as short as 50 bp (Zangenberg et al. 1995; Papadakis

---

**Fig. 2.** Influence of interlocus distance between Alu pairs on aCMR for parallel Alu pairs and antiparallel Alu pairs. All 384,397 parallel Alu pairs are sorted by their interlocus distances and evenly distributed into 40 bins, and all Alu pairs in a bin are represented by one point, positioned by their average interlocus distance and aCMR. Similarly, all 329,896 antiparallel Alu pairs are also represented by 40 points. The most pronounced decrease in aCMR appears to occur in the 250–700 bp and >3,000 bp range, whereas aCMR is relatively stable in the 700–3,000 bp range.

**Fig. 3.** Analysis of aCMR versus longest stretch of sequence identity shared between adjacent Alus. All 714,293 Alu pairs were binned according to each pair’s maximal length of uninterrupted bp identity, and the averaged aCMR of all Alu pairs in each bin is plotted, with vertical bars representing one standard deviation. Bootstrapped values were calculated for randomly selected Alu pairs grouped by the same bins. Elevation of aCMR, above baseline, is apparent for Alu pairs sharing just 35–40 bp of uninterrupted identity and increases steadily after that. The elevation is more pronounced in parallel Alus, as opposed to their antiparallel counterparts.
Within random pairs when aCMR within neighboring Alu pairs are more clustered than that along the Alu sequence, it is apparent that comutation sites exhibit some complex pattern, possibly due to the non-clusteredness for both neighboring and random Alu pairs as a function of their aCMR. Although the comutation site displays the comutation site clusteredness of Alu pairs as sites (for Alu pairs with two or more comutations). Figure 4 by the average pairwise distance among its comutation sites (for Alu pairs with two or more comutations). A stretch of sequence identity (71–90 bp) being more than twice as likely to present with GC directed comutations as would be expected by chance alone.

Biased gene conversion has been suggested to have a profound impact on the mutational landscape of genomes, especially in terms of maintaining high GC content (Eyre-Walker (1993) and references thereafter). Recent studies (Dreszer et al. 2007; Duret and Arndt 2008) highlighted the recombination-associated gene conversion mismatch repair bias toward GC nucleotides and the implications of this effect on the evolution of genomes. Our observed computational GC-bias provides additional evidence in support of the sequence homogenization by NAGC theory.

There exists an ongoing debate as to whether biased gene conversion causes high GC contents in regions of the genome or high GC content catalyzes gene conversions. In our analysis, we found that regions with high CMRs tend to produce comutations biased toward GC nucleotides. However, neither the GC content of the surrounding region nor the GC content of the Alu’s themselves had any significant correlation with the observed CMR (data not shown). Therefore, our data imply that biased gene conversion is driving isochore evolution and not the other way around. This is again consistent with the present view regarding isochore evolution and maintenance (Duret and Arndt 2008).

Commutations Are Biased Toward Weak-to-Strong Mutations
We further characterize the observed aCMR by examining the nucleotide substitution patterns at the comutation sites. Figure 5 demonstrates a clear weak-to-strong (A/T to G/C) bias for comutations among neighboring Alu pairs. This bias becomes more significant as the length of the longest stretch of uninterrupted sequence identity between neighboring Alu’s is increased, with Alu pairs with the longest stretch of sequence identity (71–90 bp) being more than twice as likely to present with GC directed comutations as would be expected by chance alone.

CMRs Are Higher at Recombination Hot Spots on Sex Chromosomes
We next set out to study the intrachromosomal variation of the aCMR. To control for the high variance of aCMR of individual Alu pairs, we measured the local sequence homogenization of Alus by the aCMR-15 measure, defined as the ratio of the total number of comutations within a stretch of 15 Alu pairs to the total number of estimated comutations among the 15 pairs assuming independence (see Materials and Methods).

When we scanned the entire genome for patterns of local sequence homogenization, the sex chromosomes displayed the most dramatic intrachromosomal variation of the aCMR-15. We found that the aCMR-15 measure accurately recapitulated the structural and functional organization of the sex chromosomes [fig. 6 in Ross et al. (2005)]. Figure 6 highlights the extremely high level of aCMR-15 in the PAR1 of both the X and Y chromosomes. Moreover, aCMR-15 of the Y chromosome drops sharply at the PAR1 boundary and displays an aCMR-15 similar to the chromosomal bootstrapped average.

Figure 4. Analysis of average distance between comutation sites along the Alu sequence. The genome-wide average distance between comutation sites in neighboring Alus is plotted versus aCMR alongside the average bootstrapped values. Only 77,450 Alu pairs with two or more comutations are plotted. Comutations become, on average, closer than expected, by chance alone, as the aCMR increases.
Again, the homogenization by gene conversion theory offers a coherent interpretation of these observations. The PAR regions (including PAR1 and PAR2) exhibit the highest recombination rate in the human genome, according to "the theory of the single obligatory crossover in PAR1" (Burgoyne 1982). PAR2 region is too short for our analysis to show significant signals and therefore has been omitted. The estimated male recombination rate of PAR1 is 4.33–20.48 cM/Mb, about 10 times of that of the autosomes (Flaquer et al. 2008). This explains the extremely high level of aCMR-15 in the PAR1 regions of both X and Y chromosomes. Moreover, neighboring Alus of the XTR1 region of the Y chromosome (just outside of the PAR1 region) still display considerable sequence similarity (data not shown). This is consistent with the fact that this region, until very recently (Ross et al. 2005), has resided on the X chromosome and has therefore undergone NAGC for most of its evolutionary history (thus the considerable sequence similarity). Finally, the aCMR-15 for the remainder of the Y chromosome is below the genomic average. This is again inline with the underlying fact that the non-PAR1 regions of the Y chromosome undergo no recombination.

We also point out that although a high recombination rate is one of the most prominent features of PAR1, thus offering a plausible explanation for the observed high aCMR, it is possible that other features specific to PAR1 may affect Alu homogenization. For example, certain chromatin structures may form in the PAR1 during meiotic sex chromosome inactivation, thereby elevating aCMR.

At the genome-wide scale, we find a correlation (Spearman’s rho = 0.04, P = 0.02) between average aCMR and male recombination rates over 1 Mbp windows. However, we find only marginal patterns of correlation between aCMR-15 and male recombination rate over the entire genome (data not shown). This is likely secondary to the transient nature of recombination hot spots (Ptak et al. 2005; Winckler et al. 2005) (see Discussion for more details).

**Comutations Are Not an Artifact of Tandem Segmental Duplications**

Alus are known to be associated with segmental duplications, among other genome evolutionary events (Jurka et al. 2004). Alus are enriched at or near junctions of segmental duplications, and Alu-Alu-mediated nonallelic

---

**FIG. 5.** Weak-to-strong bias of comutations, stratified by sequence similarity of Alu pairs. Actual and bootstrapped proportions of weak to strong (AT to GC) and strong to weak (GC to AT) are shown. Alu pairs are binned by their longest stretch of sequence identity.
Recombinations have been suggested to be partially responsible for the heightened rate of segmental duplications in the primate lineage (Bailey et al. 2003). Conversely, tandem duplications of short genomic segments containing only a single full-length Alu, including those induced by Alu–Alu unequal crossovers, would create identical neighboring Alus, which theoretically could contribute to the observed elevation in aCMR. Although the results above support the supposition that NAGC is the main driver of the observed aCMR, it is still possible that other genomic events such as segmental duplications could contribute to aCMR. In particular, tandem segmental duplications induced by unequal crossovers between adjacent Alus could explain the elevation in aCMR among Alu pairs, as well as aCMR’s decay with distance and correlation with relative Alu orientation.

In order to gauge the contribution of segmental duplication to the observed aCMR, we analyzed the sequence similarity of the flanking regions surrounding individual Alu copies and classified them into either “duplicated pairs” (Alu pairs in duplicated regions) or “nonduplicated pairs” (Alu pairs in nonduplicated regions).

We found that the aCMR for nonduplicated pairs is 1.063, well above the bootstrapping baseline (1.001). Therefore, we conclude that the observed cmutations are not simply an artifact of tandem segmental duplications. Moreover, the aCMR for duplicated pairs is 1.103, indeed higher than that of nonduplicated pairs ($P < 10^{-6}$). This result confirms the correlation between Alu sequence homogenization and tandem segmental duplications.

**Impact of Sequence Homogenization on Percent Identity between Neighboring Alus**

In order to check if sequence homogenization indeed makes neighboring Alus more similar, we analyzed the sequence identities of Alu pairs. We reported the percent sequence identity of neighboring Alu pairs over the entire 283 nucleotide positions and over only non-CpG nondiagnostic nucleotide positions, respectively. On the genome-wide bases, the mean sequence identity in neighboring Alu pairs is 0.77542 (0.77436 in random pairs) at all 283 nucleotide positions and 0.86590 (0.86558 in random pairs) at non-CpG nondiagnostic positions. These are statistically significant differences ($P < 10^{-15}$ and $P < 10^{-3}$, respectively). In other words, there are 0.00032 less mutations per nucleotide position between neighboring Alu pairs than random Alu pairs.

Given that the global average unadjusted CMR...
elevation at non-CpG nondiagnostic positions is 0.00013 (=0.00267 – 0.00254) per nucleotide position (number extracted from supplementary table S2, Supplementary Material online), we estimate about 34% (=0.00013/0.00032) of the elevation in the pairwise sequence identity among neighboring Alus can be attributed to homogenizing effects of comutations.

Intriguingly, the percent sequence identity elevation and the aCMR elevation correlate differently with the interlocus distance of Alu pairs. As shown in figure 2, as distance between Alus increases, the aCMR decays monotonously. However, this is not the case for the correlation between percent identity and the distance between Alu sequences. As figure 7 shows, the overall percent identity of Alu pairs increases to a maximum at a separation distance of 1,000 bp and then decay in the expected exponential fashion.

The aforementioned trend can be rationalized in the context of previously published literature: Stenger et al. (2001) suggested that highly homologous antiparallel Alus in close proximity tend to create genomic instability and are thus lost from the genome. The only exception to this trend is that parallel Alu pairs at very short distances (<300 bp) have very similar sequences. As this phenomenon only exists for parallel Alu pair, we speculate that this is probably a result of tandem duplication.

**Other Discussions**

By assuming that Alu sequence homogenization is a direct result of NAGC, we estimate the impact of Alu gene conversion on the evolution of overall human genome. As shown in supplementary table S2, Supplementary Material online, CMR between neighboring Alu pairs is elevated by 0.00013 per base pair per Alu pair above baseline. This observation implies that NAGCs are responsible for 5% of the comutations present within Alus in general—the remaining 95% are presumed to arise from independent random mutations. Therefore, we estimate that the impact of NAGC to current studies of mutation rates should be relatively small. Still, it accounts for about 37,000 point mutations in the human genome. Moreover, the extraordinarily high level of aCMR in the highly recombination-rich PAR1 region (fig. 6) suggests that Alu-mediated NAGC plays a major role in shaping this region. Because NAGC creates a nonmonotonic decay of linkage disequilibrium, and thus complicates linkage analyses, we recommend caution when analyzing genetic markers in the Alu sequences from recombination hot spots, especially the PAR1 region.

Given the universal presence of nonallelic Alu gene conversion, whose frequency is correlated with recombination rate, it may be tempting to infer a genome-wide map of recombination rate based on aCMR patterns, especially given the initial promising results in and around the PAR1 region of the sex chromosomes. However, outside of this region, the correlation between the observed homogenizing force and the recombination/gene conversion activity measured by other authors (e.g., Kong et al. 2002; Jeffreys and May 2004) is unclear. We reason that this lack of correlation may be explained by the following. First, it is due to the transient nature of gene conversion hot spots, whose location is known to vary with time and even between individuals (Yu et al. 1996). Furthermore, since we measure the total effect of Alu-mediated gene conversion over all of evolutionary history, our observations reflect the
cumulative sum of Alu-mediated gene conversion events. Second, cCMR caused by NAGC is determined by multiple factors in addition to the frequency of NAGC (recombination rate), including the length of gene conversion tract, the time before the gene conversion event (allowing for accumulation of mutations to be copied around), and the time after the gene conversion event (allowing for independent accumulation of point mutations). These factors may not be uniformly distributed through the genome. Third, as discussed above, contribution to overall mutation from NAGC in Alus is only minor and thus the signal of cCMR is easily buried in the noise from fluctuations of mutation rate in the genome. In summary, cCMR (and cCMR-15) is probably not a sufficient source for deriving a detailed genome-wide recombination map.

The methodology developed in this work is certainly not restricted to Alus. In general, any abundant repetitive sequence in human genome is subject to nonallelic recombination and gene conversion. For example, long interspersed element L1 constitutes ~17% of the human genome sequence, an even larger portion than the Alus (Lander et al. 2001). However, although L1 sequences litter the human genome with >500,000 copies each of length up to 67 kb, most L1 copies exist in drastically truncated forms and their sequence similarities are much lower than that between Alus (Lander et al. 2001). Therefore, NAGCs of L1s are probably much less frequent than that of Alus. Moreover, the human genome harbors a very large collection of repetitive elements. Although other repetitive elements are less abundant and generally more divergent than Alu and L1, they are certainly not immune to NAGC (e.g., Bosch et al. 2004). We believe that more comprehensive and in-depth studies of gene conversion events among a broader range of repetitive elements would be a fertile ground for genomics research. With the accelerated pace of whole genome sequencing efforts, studies of NAGCs from repeats are likely to be applicable to other genomes and comparative genomic studies.

Conclusions

This paper presents a genome-wide characterization of Alu sequence homogenization. We introduced a new statistic, cCMR, which measures the excess of sequence correlation between pairs of Alus above a baseline that would be expected from chance alone. Our analysis uncovered a significant sequence homogenization tendency acting on neighboring Alu pairs throughout the human genome. Moreover, we identified the following characteristics underlying the comutation tendencies of neighboring Alu pairs. First, sequence homogenization decays with increasing distance between Alu pairs. Second, sites of comutations are clustered. Third, sequence homogenization is positively correlated with the length of the maximum stretch of identical sequences between the Alu pairs. Fourth, sequence homogenization peaks at the PAR1 regions of sex chromosomes, which have the highest recombination rate in the human genome. And fifth, comutations are biased toward GC mutations. Taken together, all these findings lead to the hypothesis that Alu-mediated NAGC is a major force responsible for the observed sequence homogenization.

Supplementary Material

Supplementary figure S1 and tables S1 and S2 are available at Molecular Biology and Evolution online (http://mbe.oxfordjournals.org/).

Acknowledgments

We are grateful for Steven E. Brenner for support. We also thank Dmitri Petrov and Peter Arndt for inspiring discussions and Kyle Grimes for critical reading the manuscript. D.Z. is partly supported by the National Institute of Health Pathway to Independence award (RR024163).

References


