Accounting for Regions of High and Low G+C Content Found in Human Genomic DNA

Eric C. Rouchka[1]
TR-ULBL-2014-01

April 30, 2014

[1]University of Louisville
Speed School of Engineering
Department of Computer Engineering and Computer Science
Duthie Center for Engineering Room 208
Louisville, Kentucky, USA 40292

eric.rouchka@louisville.edu

UNIVERSITY OF LOUISVILLE Bioinformatics Laboratory Technical Report Series

Bioinformatics Research

Accounting for Regions of High and Low G+C Content Found in Human Genomic DNA

Eric C. Rouchka[1,*]

[1]Department of Computer Engineering and Computer Science, University of Louisville, Duthie Center for Engineering, Room 208, Louisville, KY, USA

UNIVERSITY OF LOUISVILLE BIOINFORMATICS LABORATORY TECHNICAL REPORT SERIES REPORT NUMBER **TR-ULBL-2014-01**

ABSTRACT

Motivation: Increased availability of finished human genomic sequence data has made it possible to analyze human genomic organization at the sequence level. Examination of sequence data indicates regions of high and low G+C content exist within the human genome. Different hypotheses are given examining why these regions are present in the human genome, including widely studied hypotheses stating these regions are maintained in the human genome via various mechanisms.

Results: Preliminary tests of one of these hypotheses strongly suggest high and low G+C regions have not been maintained by the presence of repetitive elements with a high or low G+C content within them. Examination of a mutational hypothesis supports the conclusion that compositional mutation biases the evolution of the human genome. However, observed mutation biases do not seem to maintain regions of high G+C content. Rather, preliminary results indicate different substitution rates are in effect in different regions of the genome, presenting the current mosaic view of the genome.

Conclusion: The preliminary study of composition specific substitution rates in repetitive elements and pseudogenes suggest features inserted under less selective pressure appear to be mutating towards a higher A+T composition with a rate dependent upon the local G+C context at the insertion site.

1 INTRODUCTION

It has been proposed vertebrate genomes are made up of compositionally homogeneous DNA segments based on G+C content (Macaya et al., 1976; Cuny et al., 1981). These regions, known as isochores, have been studied for nearly 30 years using experimental density gradient centrifugation techniques (Bernardi, 1993; Macaya et al., 1976).

Recent sequence level studies argue the human genome is not nearly as homogeneous as Bernardi's 5-class system of isochore classification might lead one to believe (Rouchka and States, 2002; IHGSC, 2001; Nekrtenko and Li, 2000). While these studies have brought to light that a strict five-class system based on G+C content may not be the best approach for sequence segmentation in human DNA, all of the authors seem to agree large regions of long-range variation in high and low G+C content are present in the human genome.

At least two categories of theories have emerged to account for these regions. The first category, the maintenance hypotheses, states regions of high and low G+C content are present in the human genome due to various poorly specified mechanisms promoting compositional maintenance. The second category hypothesizes regions of high and low G+C content are observed within the human genome due to regional variations in mutational rates.

1.1 Overview of Maintenance Hypotheses

Several theories have been proposed arguing in support of maintenance mechanisms (see Eyre-Walker and Hurst, 2001, and Bernardi, 2000, for reviews). The two main arguments stem from a selectionist hypothesis that a selective process is at work to promote G+C compositional regions and a neutralist hypothesis stating no selection is occurring. The neutralist theories can be broken down into two camps, those subscribing to biased gene conversion theories and those who believe some sort of mutational mechanism was at work.

1.1.1 Selectionist Hypothesis

The selectionist argument suggests high and low G+C regions arise due to selective advantages. G:C base pairs contain three hydrogen bonds while A:T base pairs contain two, and thus G:C base pairs should provide greater stability at higher temperature levels (Wada and Suyama, 1986). The argument for the presence of high and low G+C regions in warm-blooded vertebrates due to selection stems from the apparent observation of an "isochore" structure in mammals and birds, while genomes of cold-blooded vertebrates in-

*To whom correspondence should be addressed.

cluding fish and amphibians are devoid of such structure (Bernardi, 1993). An increase in G+C content could provide thermodynamic stability against degradation by heat (Bernardi, 2000; Ohama et al., 1987). A conflicting study (Galtier and Lobry, 1997) shows this may not be the case. Bernardi (2000) suggests this lack of correlation could be due to other selective factors such as DNA-binding proteins (Robinson *et al*., 1998) and thermostable chaperonins (Taguchi *et al*., 1991) that act to stabilize genomic DNA. This hypothesis of high/low G+C structure as a selective advantage to homeothermy has additionally been questioned due to the apparent presence of an "isochore" structure in the genomes of some cold-blooded vertebrates (Hughes *et al*., 1999), indicating the strong possibility that "isochore" evolution predated homeotherm evolution.

The most unfortunate property of the selectionist hypothesis as presented is that it cannot be easily tested using scientific rigor. While the argument may have some merit, it appears to be grounded more at a philosophical rather than factual level. Therefore, the selectionist hypothesis as proposed by Bernardi is not considered and tested.

1.1.2 Biased Gene Conversion

The biased gene conversion (BGC) hypothesis states regions of the human genome have been maintained at a higher (lower) G+C composition due to a bias in A|T→ G|C (G|C→A|T) gene conversion events (Galtier *et al*., 2001). Biased gene conversion plays a potential role in the maintenance of high G+C regions due to the high G+C content of recombination hotspots such as regions encoding ribosomal operons, tRNAs and histones (Galtier *et al*., 2001). Galtier *et al*. (2001) suggest the BGC hypothesis could account for the bias in G|C→A|T vs. A|T→G|C mutations found within single nucleotide polymorphisms (Eyre-Walker, 1999).

1.1.3 Mutational Bias

The mutational bias hypothesis states these regions are maintained by biases in mutational mechanisms favoring A|T→G|C mutations in G+C rich regions and G|C→A|T mutations in G+C poor regions. Filipski (1987) studies the correlation between coding regions and their surrounding G+C content and codon usage, a phenomenon now well studied (Knight, Freeland and Landweber, 2001; D'Oniofro and Bernardi, 1992). Filipski argues regions of differing G+C content have been maintained due to mutational biases caused by the actions of the more error prone β polymerase acting in G+C rich chromatin regions.

Wolfe, Sharp and Li (1989; also see Wolfe, 1991) suggest compositional biases could be due to differences in replication conditions. High G+C regions replicate early in the S-phase of the cell cycle when dGTP and dCTP is high in the dNTP pools. As the S-phase progresses, the dGTP and dCTP concentrations decrease, and low G+C regions replicate. As a result, A|T→G|C mutations are more likely to occur early in S-phase replication (or in high G+C regions) and C|G→A|T mutations are more likely to occur later in low G+C regions. Casane *et al*. (1997) performed similar experiments on processed pseudogenes and the surrounding non-coding regions in primates. Their results show the ratio of the G|C →A|T mutation rate to the A|T→G|C mutation rate varied according to the G+C content of the genomic position, indicating a mutational bias was at work.

Francino and Ochman (1999) suggest high and low G+C regions result from mutation events. Their results indicate the ratio of G|C→A|T to A|T→G|C mutations produces strikingly different results when the composition of the genes and pseudogenes is considered.

Ohama *et al*. (1987) examine the G+C composition of the streptomycin operon in two separate bacterial organisms with different overall G+C content. The high G+C content of the *M. luteus* genome affects the G+C composition of the *str* operon which has a mean G+C content of 67%, much higher than found in *E. coli* (51%), which has a lower genomic G+C content. In addition, 95% of all wobble bases in the *M. luteus str* operon are either G or C compared to only 52% in *E. coli*.

Fryxell and Zuckerkandl (2000) suggest context dependent mutational bias is possibly due to cytosine deamination, which decreases in rate two-fold for each 10% increase in G+C content. This implies the higher the G+C content, the lower the rates of C→T and G→A mutations will be, and similarly, lower G+C content will produce a higher rate of C→T and G→A mutations. This bias could be due to a higher concentration of methylation/deamination enzymes in regions of lower A+T composition. Cytosine deamination would then function as a positive feedback loop, promoting maintenance of both high and low G+C regions.

1.2 Overview of Regional Variation in Mutation Hypotheses

In 1972, Cox argued the spontaneous mutation rate within mammalian DNA varies over the entire genome. This conflicted the previous assumption that mutation rates were uniform throughout genomes (Sueoka, 1962). More recent studies illustrate variation in mutation rates across a genome (Castresana, 2002; Casane et al., 1997; Wolfe, Sharp and Li, 1989).

Regions of high and low G+C could arise due to regional variations in mutation rates. A hypothesis studied herein is the human genome evolved from a G+C rich ancestral genome. As discussed in the results, substitution rates within the human genome appear to have moved the genome to-

wards A+T richness. This rate would appear to have been slower, but nonetheless present, in regions of high G+C. The variability in the mutation rate hypothesis suggests regions of high G+C are seen in the present view of the human genome due to their location in regions of low mutation while regions of low G+C tend to be located in mutational hot spots.

1.3 Understanding Large-scale G+C Variation

The interest in understanding large-scale G+C variation within the human genome led to the exploration of two hypotheses designed to test the maintenance hypothesis. The first hypothesis states high and low G+C regions are maintained by the presence of repetitive elements with a high or low G+C content within them. The second hypothesis tested was that a compositional bias for mutation rates exists which promotes the maintenance of such regions.

2 METHODS

2.1 Exploration of Two Maintenance Hypotheses

One of the shortcomings of previous studies into the mechanisms suggesting maintenance of regions of high and low G+C content is they are largely based on looking at genic regions, which constitutes only a small fraction of the human genome (Gardiner, 1996). In order to understand regions of high and low G+C composition more completely, potential maintenance of these regions was studied by looking at repetitive elements and processed pseudogenes, which are two features in the human genome less likely to be under selective pressure. Such an approach may rule out other evolutionarily advantageous mechanisms at work.

2.1.1 Hypothesis 1: Regions of High/Low G+C Result from Repetitive Element Composition

Previous studies show the densities of certain types of repetitive elements are not uniform throughout the human genome (Belle and Eyre-Walker, 2002; IHGSC, 2001; Pavlíček *et al*., 2001; Matasi, Labuela and Bernardi, 1998; Jabbari and Bernardi, 1998). The pattern of distribution of G+C rich short interspersed (SINE) elements (the mean G+C content of the representative ALUs is 52%) and G+C poor long interspersed (LINE) elements (L1 elements are 37% G+C) is particularly intriguing (Belle and Eyre-Walker, 2002; IHGSC, 2001; Eyre-Walker and Hurst, 2001). SINEs and LINEs both incorporate the LINE transcription mechanism (Jurka, 1997; Feng et al., 1996). It would be thought that the insertion at a TTTT/A cleavage site would promote SINEs and LINEs both within A+T rich regions due to an increased likelihood of finding such a site. However, LINEs tend to be found in A+T rich regions, while SINEs are found in more G+C rich regions (IHGSC, 2001; Eyre-Walker and Hurst, 2001), although more recent ALUs are more evenly distributed in the genome (Eyre-Walker and Hurst, 2001).

One potential explanation leading to the appearance of high and low G+C regions in the human genome is regions of G+C variation are caused by the presence of repetitive elements within them. Under this hypothesis, regions of high G+C will exist in the human genome due to a high density of G+C rich SINEs within them. Similarly, regions of low G+C should be observed due to the high density of G+C poor LINEs in these areas. If repeats alone were responsible for regional variation, there should be no correlation between regional G+C content and the G+C content of the unique sequence contained within.

Calculating Repetitive and Non-repetitive G+C Composition

The G+C composition of the region as a whole was compared to the G+C composition of the repetitive and potentially unique (non-repetitive) regions. If repetitive elements were the driving force behind the overall G+C composition, there should be a higher correlation between the G+C content of the repetitive elements and the G+C content of the overall region. At the same time, the G+C content of the unique regions should remain neutral and randomly vary based on the G+C content of the repetitive elements.

This hypothesis was explored by examining the Goldenpath December 2001, assembly of the human genome (Kent and Haussler, 2001; `genome.ucsc.edu`). Only contigs mapped to a particular chromosome were considered. Known repeats from the `Repbase` database version 6.10 (Jurka, 2000) were masked out using `RepeatMasker` (A. Smit and P. Green, unpublished). Each contig was run through `RepeatMasker` twice. One run was performed in the slower, native settings for the detection of low complexity and simple repeats (using the `-int` option). The second run took advantage of the `-w` option, which incorporates `wublastn` (W. Gish, 1996-2001) as the underlying alignment algorithm (Bedell, Korf and Gish, 2000).

2.1.2 Hypothesis 2: Mutational Biases Revisited

As previously discussed, one of the hypotheses for high and low G+C region maintenance is it is due to biological mechanisms favoring compositional bias in mutation rates. Previous studies have focused on a limited set of genes and pseudogenes within human and primate populations (Filipski, 1987; Wolfe, Sharpe and Li, 1989; Casane *et al*., 1997). In order to work around selection mechanisms that may play a role, two elements less likely to be under selective pressure were studied: processed pseudogenes and repetitive

elements. In an ideal case, the rate of A|T→G|C and G|C→A|T mutations would be compared when elements derived from the same ancestor were placed in differing neighborhoods of G+C concentration. However, it is not always possible to determine whether a mutation has occurred within the ancestor or the descendant sequence. Therefore, the rate of A|T→G|C and G|C→A|T substitutions were studied as to how they related to the surrounding G+C composition.

2.2 Studying Compositional Bias in Processed Pseudogenes

For the purpose of the study, it was assumed that the gene locus existed first, and then at some point in the evolutionary history of humans, the pseudogene arose. Once the gene and pseudogene were in place, they could evolve and mutate independently of one another. However, genes are under selective pressure, so there are expected to be fewer mutations within them than in neutrally mutating pseudogenes. When a nucleotide difference is observed between a gene and pseudogene, it is more likely to have occurred within the pseudogene. An exception would be when a mutation occurred in the third codon position (wobble base). Synonymous wobble base mutations are not expected to alter the fitness of the genic region in a significant way.

Details on how to incorporate directional information is given in the Discussion. However, the directionality of the mutation is not nearly as important as whether or not it changes the overall G+C composition of the gene or pseudogene. Therefore, substitutions were reported as A|T→G|C and G|C→A|T where the nucleotide of the gene was listed first, and the nucleotide of the pseudogene second. If the original nucleotide is an A or T in the gene and the nucleotide in the pseudogene is a C or G, the effect will be the same as if the original gene nucleotide was a C or G that mutated to an A or T over time. Thus, the rates of A|T→G|C and G|C→A|T substitutions are compared when the gene is in one G+C composition and the pseudogene is in another. This allows a determination of whehter a compositional bias in substitution rates within genes and pseudogenes potentially exists.

2.2.1 Obtaining Pseudogene Data

Potential processed pseudogenes were obtained by searching individual mRNA entries of RefSeq (Pruitt and Maglott, 2001) against the University of California-Santa Cruz's Goldenpath assembly of the human genome using `wublastn`. For the data sets, `RefSeq` was downloaded on April 18, 2002, when 15,199 human mRNAs were available. The December 2001 `Goldenpath` assembly was used.

`RefSeq` entries mapping to more than one location were likely to contain both a native locus location as well as one or more other locations that were potential paralogs or pseudogenes. For entries with multiple loci, each individual `BLAST` *HSP* was assigned a score S_{HSP} (Equation 1) equal to the fractional percentage identity multiplied by the fraction of the mRNA that the HSP covered. L_{HSP} is the length of the HSP and L_{REFSEQ} is the length of the RefSeq entry. Scores for all of the HSPs occurring within a single locus (a total of n HSPs) were summed into a single score, S_{LOCUS} (Equation 2).

$$S_{HSP} = \%id * \frac{L_{HSP}}{L_{REFSEQ}} \tag{1}$$

$$S_{LOCUS} = \sum_{i=1}^{n} S_{HSP_i} \tag{2}$$

The locus with the highest (optimal) S_{LOCUS} score was considered the native locus. All other loci were treated as potential candidates for paralogs and pseudogenes, both processed and unprocessed. Each HSP within an alignment should roughly correspond to an alignment of exonic regions. RefSeq hits were filtered to only contain entries where the native locus contained at least three HSPs to increase the likelihood that at least one intron (two exons) was in the native gene. This reduced the problem of differentiating between paralogs, unprocessed pseudogenes and processed pseudogenes corresponding to single exon genes. An additional restriction that the non-native loci contains only a single HSP was applied since processed pseudogenes have intronic regions spliced out and they should map continuously with the RefSeq mRNA. A final restriction required non-native loci to align within 20 basepairs (bp) of the 3' end of the RefSeq sequence, since processed pseudogenes are often truncated at the 5' end. While these restrictions would not detect all of the processed pseudogenes within the human genome, the reported gene-pseudogene pairs had a greater likelihood of being true positives.

Gene and pseudogene pairs were separated into four categories based on their G+C content (Table I). The categories are: (LOW, LOW), (LOW, HIGH), (HIGH, LOW), and (HIGH, HIGH). The first element in the ordered pair represents the regional G+C composition flanking the gene while the second represents the regional G+C composition flanking the pseudogene. These neighboring compositions were calculated taking 25 kb on both sides of the gene or pseudogene. Regions containing less than 41% G+C were categorized as LOW, while regions containing greater than 44% G+C were categorized as HIGH. The total neighborhood size of 50-kb (25-kb on two ends) was used to maintain consistency with Bernardi's earlier density gradient centrifugation experiments. Additionally, boundaries of 41% and

Table I: Number of Genes and Pseudogenes Found.

GENE G+C	Pseudogene G+C	Number of Genes	Number of Pseudogenes
HIGH	LOW	242	564
HIGH	HIGH	233	464
LOW	LOW	173	250
LOW	HIGH	52	79
TOTALS		**700**	**1,357**

44% G+C were chosen to correspond with major breakpoint divisions within Bernardi's isochore definitions.

2.2.2 Calculation of Gene-Pseudogene Substitution Rates

Once the genes and pseudogenes were separated into the appropriate category, they were aligned using Sim4 (Florea *et al*., 1998). Whenever a mismatch appeared between the gene and pseudogene, it was treated as a substitution event. The annotated coding sequence (CDS) was parsed out of each RefSeq entry. Substitutions in the CDS were recorded and separated into wobble and non-wobble base positions. Anything outside the CDS was labeled a non-coding substitution, which were separated into 5' UTR mutations and 3' UTR mutations depending on their relationship to the start and end of the CDS.

Once all of the alignments were made, the number of each of the 16 substitution events (gathered from the Cartesian product A x B where A,B = {A, C, G, T} and A represents the nucleotide in the gene and B represents the corresponding nucleotide in the processed pseudogene) were calculated for the following categories: coding regions, wobble bases, non-wobble coding bases, non-coding regions, 5' UTRs and 3' UTRs.

2.3 Approaches to Looking at Mutation and Substitution Events

Effective base conversion rates, u and v, are described by Sueoka (1962) as the rates of conversion at any given point in the genome from A|T→G|C and G|C→A|T nucleotides, respectively. They are explained in terms of the observed inherited rates of nucleotide substitution within a single organism from generation to generation. These values are used in more recent studies to measure the mutation rates within different genomic regions (Piganeau *et al*., 2002; Smith and Eyre-Walker, 2001; Casane *et al*., 1997; Gu and Li, 1994).

Using these models as guidelines, the rate of A|T→G|C substitutions (u) was calculated as the probability that a G or C nucleotide was found at a given location in the pseudogene, conditioned on the nucleotide in the gene being an A or T. In addition, the rate of G|C→A|T substitutions (v) was calculated as the probability that an A or T nucleotide was found at a given location in the pseudogene, conditioned on the nucleotide in the gene being a C or G.

The G+C bias (f) was calculated as $f=u/(u+v)$ (Piganeau *et al*., 2002). A measure of the A+T bias can be obtained as $1-f$. The G+C bias ranges from 0 to 1, where 0 indicates there are no A|T→G|C substitutions in a region for a given feature, 1 indicates there are no G|C→A|T substitutions, and 0.5 indicates equal A|T→G|C and G|C→A|T substitution rates. A G+C bias less than 0.5 indicates a region will drift to A+T richness over time, while a value greater than 0.5 indicates a drift towards G+C richness.

In order to test for compositional bias in substitution rates, a ratio of the G+C bias in high G+C regions (f_{HIGH}) to the G+C bias in low G+C regions (f_{LOW}) was computed. A ratio, r, consistently greater than 1 indicates a compositional bias in substitution rates was likely to exist, where high G+C regions acquired more G's and C's over time and low G+C regions were adding more A's and T's over time. A ratio less than 1 on a consistent basis indicates there was likely to be a negative correlation where G+C rich regions would be mutating towards A+T and A+T rich regions would be mutating towards G+C. If the ratios randomly fluctuate above and below 1, a compositional bias for substitution rates cannot be demonstrated for the feature being studied.

2.4 Studying Compositional Bias in Repetitive Elements

A large portion (over 45%; IHGSC, 2001) of human genomic DNA has been derived from the dispersion of transposable elements throughout the genome (Prak and Kazazinan, 2000; Smit, 1999). There are approximately 868,000 copies of LINEs in the human genome, making up over 20% of the total genomic sequence. In addition, there are over 1.5 million copies of SINEs, accounting for over 13% of the genome (IHGSC, 2001). Due to the large abundance of repetitive elements in the human genome, substitution rates within them were studied to determine if a compositional bias for substitution potentially existed in these segments.

2.5 Detecting Repetitive Elements

Instances of SINE and LINE repeats were located within the human genome using RepeatMasker release 6/19/01 with Repbase update 6.6 repeat definitions. Once the contigs were masked, the generated `.out` files containing tables of repeat information were parsed. Files were generated to group together the Goldenpath contig name, contig location and orientation of the repeat instances for each type of repeat. The repeat regions were extracted from the contigs,

Table II: Comparison of G+C bias in instances of repeat families. The second and third columns contain the G+C bias ***f=u/(u+v)*** (where ***u*** was the rate of A|T→G|C substitutions and ***v*** was the rate of G|C→A|T substitutions) calculated for instances of repeats occurring in HIGH and LOW G+C regions, respectively.

Repeat Family	HIGH G+C	LOW G+C	RATIO HI:LOW
AluYa5	0.3821	0.4077	0.937
AluYb8	0.5121	0.5440	0.941
AluYc	0.2486	0.2467	1.008
AluY	0.2479	0.2397	1.034
AluSg1	0.2017	0.2091	1.036
L1PA6	0.3018	0.2883	1.047
L1PA3	0.3217	0.3057	1.053
L1PA4	0.3418	0.3222	1.061
L1	0.2888	0.2708	1.066
L1PA2	0.4242	0.3955	1.073
AluSq	0.2332	0.2173	1.073
AluSc	0.2333	0.2160	1.080
AluSp	0.2109	0.1952	1.080
L1PA8A	0.3291	0.2978	1.105
L1PA7	0.2989	0.2687	1.112
L1PA5	0.3500	0.3134	1.117
L1PB1	0.3428	0.3003	1.141
L1PA10	0.3493	0.3020	1.157
L1PA8	0.3367	0.2870	1.173
L1PA11	0.3601	0.3049	1.181
L1MA3	0.3373	0.2829	1.192

Repeat Family	HIGH G+C	LOW G+C	RATIO HI:LOW
L1MA2	0.3439	0.2870	1.198
L1PB2	0.3549	0.2930	1.211
L1PA15	0.3227	0.2659	1.214
L1PB3	0.3213	0.2621	1.226
L1PA14	0.3469	0.2830	1.226
LAMA4A	0.3369	0.2743	1.228
L1PA13	0.3646	0.2968	1.229
L1MA4	0.3370	0.2741	1.229
L1PA16	0.3376	0.2701	1.250
L1MB4	0.3527	0.2810	1.255
L1ME1	0.3489	0.2758	1.265
L1PA17	0.3275	0.2579	1.270
L1PB4	0.3511	0.2739	1.282
L1MB8	0.3558	0.2770	1.284
L1MA9	0.3616	0.2811	1.286
L1MB7	0.3659	0.2756	1.327
L1MA8	0.3691	0.2780	1.328
L1MB2	0.3635	0.2732	1.331
L1MC1	0.3836	0.2811	1.365
L1MB5	0.3838	0.2726	1.408
L1MB3	0.4035	0.2808	1.437

and the G+C content of the surrounding 50-kb (25 kb on each side) window was noted. Each instance of a repeat was placed into one of two files for each repeat type based on whether the G+C content of the surrounding window was less than 41% or greater than 44%, labeled low and high G+C, respectively. Those repeat elements falling in the intermediate range of 41% to 44% G+C were discarded from the study.

Repetitive element families and subfamilies with the greatest number of instances currently detectible in the human genome were studied. The resulting data set analyzed included eight ALU families/subfamilies and 34 LINE families/subfamilies (Table II).

2.6 Calculating Repetitive Element Substitution Rates

With repetitive elements, it is difficult to assign directionality for each mutation since it cannot easily be determined which copy of a repeat was present first in a genome, and whether or not a second repeat was derived as a direct ancestor. In addition, once a copy is in place, it mutates and evolves independently of its parent copy. One possible scenario is that a C or G nucleotide is observed at one position in a copy of an element situated in a region of high G+C composition. At the same time, an A or T could be observed at the same position when a copy of the element was found in a low G+C region.

The Repbase-defined consensus was taken as the ancestral repeat element. Such an approach is justified in the sense that the consensus sequence has been derived to be the best approximation of the original transposable element that generated a given repeat subfamily (Jurka, 1998). Such an approach assumes a master/slave model of repetitive element propagation (Shen, Batzer and Deringer, 1991; see Discussion). Substitution rates were measured as the difference from the Repbase sequence.

Each instance of a given repetitive element was compared against the Repbase consensus sequence using `wublastn` with the parameters `-S2=200 -S=250`. Using the default `wublastn` scoring parameters of `+5,-4` for matches and mismatches, this corresponds to an ungapped alignment of at least 50 bp at 100% identity, or 78 bp at 80% identity.

The total number of substitution events FROM the Repbase consensus TO the instance of the repeat was noted. The total substitution events for repeat instances in low G+C (<41%) and high G+C (>44%) were calculated. The rate of A|T→G|C (*u*) and G|C→A|T (*v*) substitutions were computed as well as the G+C bias (*f*) for two categories: HIGH and LOW for each of the repetitive element families studied. HIGH represents those repetitive regions occurring in >44%

G+C regions and LOW represents those repeats occurring in <41% G+C regions. A ratio of the HIGH:LOW G+C biases was calculated for each repeat family studied. A ratio greater than 1 indicates that the rate of A|T→G|C vs. G|C→A|T mutations is likely to be higher in high G+C regions.

2.7 Repeats on Chromosome Y

As previously discussed, one potential problem is the mutational bias and biased gene conversion theories are not necessarily mutually exclusive. In order to address this concern, another study examining only instances of repetitive elements occurring on chromosome Y was performed. Chromosome Y contains a non-recombining region making up over 95% of the chromosome (Tilford *et al*., 2001). The non-recombining region of chromosome Y does not recombine with chromosome X or any other chromosome (Lahn, Pearson and Jegalian, 2001). Non-recombining regions will not allow for gene conversion, and biased gene conversion could not be the cause of any biases in G+C composition that are observed in such regions. Analysis on chromosome Y was limited to Alu and LINE elements having at least five different instances in LOW G+C regions and five different instances in HIGH G+C regions. The G+C bias was calculated for instances occurring in HIGH and LOW G+C for repetitive elements fitting this criterion. In addition, the ratio of the HIGH:LOW G+C biases was computed.

3 RESULTS

3.1 Repetitive Element Composition

A total of 51.6% of the bases were masked, indicating they contained repetitive sequence structure. For each of the 2,992 GoldenPath contigs, the G+C composition of the overall, masked, and unmasked regions was recorded. Figure 1 shows the resulting plot for contigs greater than 250 KB in length. The G+C composition of each overall contig was compared to the G+C composition of the masked regions and unmasked regions. Correlation coefficients and *t*-scores were calculated for each of these comparisons. In the case of the masked/overall comparison, the correlation coefficient of 0.9620 yielded a *t*-score of 192.55. For the unmasked/overall comparison, the correlation coefficient is 0.9532, corresponding to a *t*-score of 172.45. In each of these cases, the *t*-score was much greater than the critical value of 2.58 (using a *p*-value of 0.995; $\alpha = 0.005$).

A positive correlation between the G+C content of the masked regions and the overall contigs was expected. This is due to the previously reported positive correlation between increasing genomic G+C content and G+C rich SINE elements and the negative correlation between increasing genomic G+C content and the density of A+T rich LINE elements (IHGSC, 2001; Eyre-Walker and Hurst, 2001).

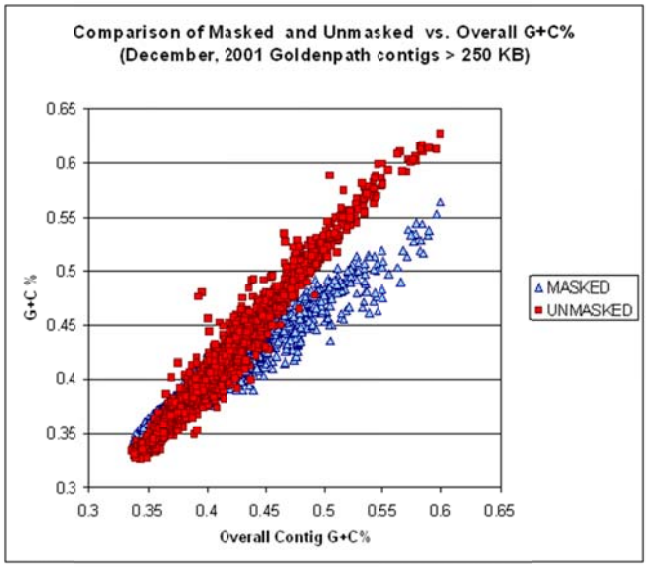

Figure 1: Comparison of G+C Content. Shown in this figure is the comparison of the G+C content of masked (repetitive) and unmasked (potentially non-repetitive) regions compared to the overall G+C content for each of the 1,927 Goldenpath contigs greater than 250 KB in length from the December, 2001 build. The x-axis represents the overall G+C content of each contig. Regions were masked using RepeatMasker (A. Smit and P. Green, unpublished)

However, such a strong positive correlation between the overall G+C content and the G+C content of the unmasked regions was not expected. Since the unique regions were highly correlated with the overall G+C content, it cannot be concluded that the G+C content of repetitive regions was responsible for the variable G+C content within the human genome.

It could be postulated there was some sort of mechanism for preferential insertion of low G+C repetitive elements into genomic regions of low G+C, while high G+C repetitive elements were inserted into genomic regions high in G+C content. However, as previously discussed (Feng *et al*., 1996), SINEs and LINEs use the same mechanism of insertion. This indicates both SINEs and LINEs would be preferentially located in regions of low G+C. Eyre-Walker and Hurst (2001) show this is the case when only recently inserted SINE elements are considered. Pavlíček *et al*. (2001) propose older SINE insertions may tend to be found in higher G+C regions if the excision of ALUs was fast enough to remove new copies before they had a chance to fixate in the population. They discuss the possibility of positive selection of the CpG rich ALUs in G+C rich regions due to hypomethylation in germline cells. In addition, it is suggested there are different recombination rates that could be affected by the short length of SINE elements (on the order of 300 bases) when compared to LINE elements (several KB long).

The parameters in RepeatMasker have been designed so potentially interesting, unique regions are not falsely masked as repetitive. As a result, sufficiently divergent repetitive elements will not be detected. Repetitive elements closest in identity to the Repbase consensus are detected. These result from more recently active transposable elements within the human genome. Since recent transposable events do not lead to the creation and maintenance of regions of high and low G+C content within the human genome, it is unlikely ancient copies of the same repetitive elements would have any different effect. In fact, these ancient copies should behave in the same manner due to the same mechanisms of insertion. In addition, Repbase consensus sequences have been carefully constructed to address the problem of detecting diverse repeats by representing the best available approximation of the elements that generated the repeats (Jurka, 1998).

The variance of the G+C content in unmasked regions was small. Ancient copies of repeats currently undetected are expected to have properties similar to detected repeats. If methods to detect these repeats were available, a migration of the data points in Figure 1 from the unmasked fraction to the masked fraction would result. This migration should have little effect on the correlation between the unique region and overall contig G+C% due to the low variance.

Low copy number and uncharacterized repeats in the human genome will not be detected using RepeatMasker since they are found only in a small portion of the genome. Since Repbase has been carefully examining and collating information on repetitive elements within the human genome (which has been available since February, 2001 (IHGSC, 2001)), it is unlikely there are any high copy number repeats that remain uncharacterized. Since any remaining uncharacterized repeat families or subfamilies will likely have a relatively low copy number and constitute a low percentage of the human genome, they should contribute little information into the origin and maintenance of high and low G+C regions within the human genome.

Based on the information gathered, the first hypothesis should be rejected. Regions of G+C content within the human genome do not appear to result from the presence of repetitive elements; rather it appears as though the presence of regions of high and low G+C concentration determines the density of certain repetitive elements within the human genome.

3.2 Gene-Pseudogene Mutational Bias

In order to test for a possible compositional bias for substitution rates in gene-pseudogene pairs, two different comparisons were made: one where the gene originated in a low G+C region, and one where the surrounding content of the gene was high G+C. In each comparison, two different cases were examined. The first case involved the pseudogene occurring in a low G+C region, and the second case was when the pseudogene was in a high G+C region.

If a compositional bias for substitution rates exists, the G+C bias, *f*, would be expected to increase as the G+C context of the pseudogene increases. This would indicate the ratio of A|T→G|C to G|C→A|T mutations increase as the surrounding G+C context increases. In order to test this hypothesis, the G+C bias, *f*, was calculated for the four cases defined by the Cartesian product A x B where A,B={HIGH, LOW} and A=G+C context of the gene; B=G+C context of the pseudogene. The resulting G+C biases were labeled as follows: f_1 = {LOW,LOW}; f_2 = {LOW,HIGH}; f_3 = {HIGH,LOW}; f_4 = {HIGH,HIGH}. In order to test for potential compositional biases for substitution rates, the ratios $r_1 = f_2:f_1$ and $r_2 = f_4:f_3$ were calculated. If a compositional bias exists, the values of r_1 and r_2 would be expected to be greater than 1. The results are listed in Table III.

Table III: Comparison of G+C bias in gene and pseudogene pairs for genes in low (A) and high G+C content (B).

A)

Gene in Low G+C	Pseudogene HIGH G+C Bias	Pseudogene LOW G+C Bias	Ratio of HIGH:LOW
5' UTR	0.4632	0.3797	1.220
CDS	0.4288	0.3309	1.296
WOBBLE	0.4721	0.3467	1.362
3' UTR	0.3674	0.3132	1.173

B)

Gene in High G+C	Pseudogene HIGH G+C Bias	Pseudogene LOW G+C Bias	Ratio HIGH:LOW
5'UTR	0.4721	0.4032	1.171
CDS	0.4159	0.3600	1.155
WOBBLE	0.5710	0.5036	1.134
3' UTR	0.4376	0.3765	1.162

For each of the features studied, the values of r_1 and r_2 were greater than 1, with r_1 ranging from 1.173 to 1.362 and r_2 ranging from 1.134 to 1.175, indicating A|T→G|C and G|C→A|T substitutions were 17-36% higher in the first case, and 13-17% higher in the second case. These increases indicate that, when pairs of genes and pseudogenes were examined, there appeared to be a compositional bias for substitutions.

When the G+C bias, *f*, was compared in 5' UTRs, CDS, non-wobble CDS, and 3' UTRs (Table III), the value was always less than 0.5, indicating these portions of the

pseudogenes had higher rates of G|C→A|T substitutions than A|T→G|C substitutions no matter what the original gene and pseudogene G+C contexts were. As a result, as pseudogenes aged, these regions tended towards A+T regardless of the surrounding G+C content. However, the rate of this substitution trend was slowed when the surrounding region was G+C rich.

Substitutions found within non-wobble coding positions are likely to have occurred within the pseudogene since most mutations within the first two codon positions of a gene will cause a change to the amino acid encoded by that codon. Such a change can affect the fitness of the gene. Therefore, the results listed in Table 2 suggesting a compositional bias for substitution has occurred within non-wobble coding regions is likely to have an associated directionality.

The study of gene and pseudogene pairs indicates there was a strong possibility of a compositional bias for substitution rates. However, the rate of A|T→G|C substitutions was always less than the rate of G|C→A|T substitutions. This indicates pseudogenes within the human genome were likely to accumulate more A+T sequence over time regardless of the surrounding G+C context. However, as the G+C context of the pseudogene increased, the rate of this change slowed. As a result, a compositional bias in substitution rates was observed, but this rate cannot be the determining factor for maintaining regions of low and high G+C composition.

3.3 Repeat Instance Substitution Bias

Table II lists the resulting G+C biases calculated for each of the repeat families for the instances in low and high G+C. For the Alu repeat families studied, the ratio ranged from 0.937 to 1.080. Six of the eight Alu families had ratios greater than 1 (with the exception of the AluYa5 and AluYb8 families). This suggests for six of these families, a slight compositional bias for mutation rates exists. All 34 of the LINE families studied had ratios greater than 1. In fact, these ratios tended to be larger than the ratios for Alu families, ranging from 1.047 for the L1PA6 family, to 1.437 for the L1MB3 family. These values show the LINE families have a potentially stronger compositional bias for mutation rates.

The G+C biases for nearly all of the repetitive families were much less than 0.5, yielding results similar to the gene-pseudogene substitution rates. This indicates no matter what the surrounding G+C content is for an instance of a repetitive element, the repeat copy will likely drift towards A+T richness over time. Since the ratios were greater than 1 (indicating there was a compositional bias for substitution rates), the rate of drift should be slower when the surrounding G+C content is higher. These results indicate there seems to be a compositional bias for substitution rates; however, this bias is unlikely be the cause for the maintenance of high G+C regions containing the features studied.

3.4 Repeats on Chromosome Y

A total of five Alu and twelve LINE families were studied on chromosome Y (Table IV). The only repeat subfamily with a ratio less than 1 was the AluY subfamily, the youngest repeat studied with an age less than 1 million years old (IHGSC, 2001). The other 16 repetitive element families on chromosome Y likely have a compositional bias affecting substitution rates. G+C biases were significantly less than 0.5, indicating repetitive elements on chromosome Y likely tend toward A+T richness over time. Since 95% of chromosome Y is not subject to recombination, it is unlikely the compositional bias for substitution rates within repetitive elements on chromosome Y was due to biased gene conversion. Although it cannot be certain that biased gene conversion does not largely contribute on other chromosomes, the results observed for chromosome Y were consistent with the previous repeat study. As a result, biased gene conversion is thought to contribute little to the observed compositional bias.

Table IV: Comparison of G+C bias for repeats round on chromosome Y.

Repeat Family	HIGH G+C	LOW G+C	RATIO HIGH:LOW
AluY	0.2194	0.2210	0.993
L1PA2	0.3916	0.3898	1.005
L1PB1	0.2876	0.2836	1.014
AluSq	0.2165	0.2099	1.031
L1MA9	0.3008	0.2778	1.083
L1PA4	0.3321	0.3046	1.090
L1PA14	0.2740	0.2503	1.095
L1PA3	0.3350	0.2957	1.133
AluSp	0.2185	0.1910	1.144
AluSc	0.2478	0.2150	1.153
AluSx	0.2614	0.2122	1.231
L1	0.3238	0.2472	1.310
L1MB7	0.4076	0.2866	1.422
L1PA7	0.3784	0.2547	1.486
L1MA8	0.4465	0.2984	1.497
L1PB4	0.4125	0.2563	1.609
L1PA15	0.4394	0.2570	1.709

3.5 Testing for Drift to an A+T Rich Genome Using Long Terminal Repeats (LTRs)

The results of looking at gene/pseudogene pairs and instances of repetitive elements suggest elements inserted into the human genome are likely to mutate towards a higher A+T composition over time. This phenomenon was observed when comparing the rate of A|T→G|C and G|C→A|T substitutions and was independent of the G+C content of the surrounding region. In order to test this hypothesis, ele-

ments inserted at different points in time were studied to determine whether or not older elements tend to be more A+T rich.

One class of repetitive elements of particular interest is those caused by LTR retroviral integration events. These elements are useful to study since the mechanism of LTR retroviral integration produces two identical long terminal repeats (LTRs) which flank the 5' and 3' end of the virus (Lodish *et al*., 1995). The divergence between the 5' and 3' LTRs can be used to calculate an approximate integration date for any particular instance (Tristem, 2000).

Approximately 1.3% of the human genome is composed of human endogenous retroviral (HERV) elements, representing roughly half of the LTRs found in humans (Smit, 1996). One recent study looked at classification and integration age of the various HERV families (Tristem, 2000). This study estimates the HERV-H, HERV-K, and HERV-L families have the largest copy number in the human genome.

3.6 Detecting Copies of HERVs

Representative sequences for HERV-H, HERV-K, and HERV-L families as described by Tristem (2000) were obtained from `GenBank` (Benson, et al., 2002). The accessions obtained were as follows: D11078 (HERV-H) (Hirose *et al*., 1993); M14123 (HERV-K) (Ono *et al*., 1986); and X89211 (Corodonnier, Casella and Heidmann, 1995). Each of these sequences was searched against the December 2001 release of the Goldenpath assembly of the human genome using `wublastn`. Score cutoff parameters of `-S=2000` and `-S2=2000` were used to filter spurious hits. A score of 2000 using the default `wublastn` scoring scheme of `+5,-4` requires a 400 bp ungapped alignment at 100% identity, or a 625 bp ungapped alignment at 80% identity. In addition, the parameter `-gapw=2000` was used to close longer alignment gaps.

The search matched 1001 HERV-H locations, 409 HERV-L locations, and 723 HERV-K locations. However, many of these instances were truncated, missing one or both of the LTR sequences due to recombination events leading to a solitary LTR (Prak and Kazazian, 2000). These matches were manually filtered to include only full-length copies. The resulting datasets included 14 HERV-H, 21 HERV-K, and 72 HERV-L copies.

3.6.1 Determining Insertion Age and G+C Composition

`GenBank` records for the HERV-H, HERV-K and HERV-L representative sequences contain various annotations including the 5' and 3' LTR sequences. The representative 5' and 3' LTR sequences were extracted and placed into separate files. Each of the full-length copies were searched against the appropriate 5' LTR using `wublastn` with the parameters `-S=300 -S2=300 -gapw=200`. Since the 5' and 3' LTRs should be identical at the time of insertion, searching full-length repeats for the presence of the 5' or 3' LTR should produce the same results. The 5' LTR was arbitrarily chosen, which in every instance located the 3' LTR as well. The resulting `wublastn` output was parsed to extract the 5' LTR sequence and 3' LTR sequence. These were aligned to each other using `wublastn` with the parameters `-S=200 -S2=200 -gapw=128`. The approximate edit distance for each instance was determined based on the number of mismatched bases in the alignment of the 5' and 3' LTRs. Gaps were ignored.

After the 5' and 3' LTRs were located in each full-length copy, the G+C content of the repeat copy was calculated. The edit distance for each instance was compared to the G+C content to see if more distant elements tend to be more A+T rich. Figure 2 shows a graph plotting the G+C composition against the percent divergence for the 72 full-length HERV-L copies. For this figure, the percent divergence was calculated as the percentage of mismatching bases when the 5' and 3' LTRs were aligned. The assumption is the higher the percent divergence, the older the insertion date will be.

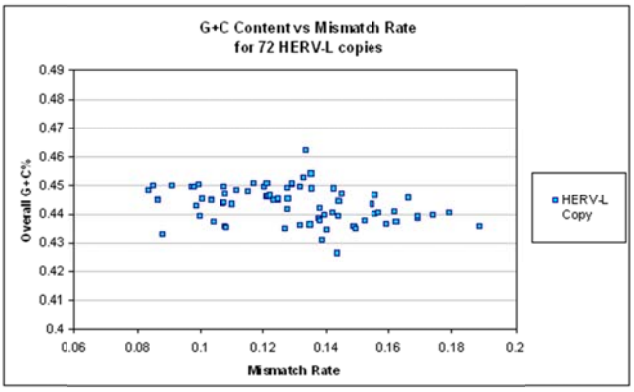

Figure 2: Plot of divergence rate vs. G+C composition in HERV-L repeats. Shown in this figure is a plot of the divergence rate versus the overall G+C percentage for each of the 72 full-length HERV-L copies found within the human genome. The divergence rate (x-axis) is calculated as the percentage of bases mismatched in an alignment between the 3' and 5' LTRs of the HERV-L copy. The overall G+C percentage (y-axis) is based on the G+C content of the complete HERV-L copy.

A correlation coefficient was calculated to determine whether or not a correlation exists between the edit distance and the G+C content. An *r*-value of -0.3279 was calculated for the 72 HERV-L instances, indicating a slight negative correlation between the LTR divergence and the repeat G+C content. This suggests the older the date of insertion, the greater the accumulation of A's and T's will be. A *t*-score

was calculated for the *r*-value of -0.3279 with 72 instances to determine the level of significance for this correlation coefficient. The resulting *t*-score was -2.946. Using 70 degrees of freedom and a two-tailed test, this *t*-score yields a *p*-value of 0.0087, indicating the observed correlation is likely to exist between the insertion date and G+C content.

While a correlation between the insertion date and G+C content has been demonstrated with the HERV-L repeat family, it would be useful to locate more instances of high copy number elements in which the relative date of insertion can be determined. There are two main difficulties in obtaining such data for human LTR retrotransposons. The first problem is homologous recombination events often remove one or both of the LTRs (Prak and Kazazian, 2000). In addition, the human genome contains relatively few LTR elements (Smit, 1996), many of which are solitary LTRs. Next to the HERV families of LTR retrotransposons, the mammalian apparent LTR-retrotransposon (MaLR) superfamily is the most interesting to study. However, most of the LTR copies from the MaLR superfamily are found as solitary LTRs in the genome (Smit, 1993), making it difficult to determine an insertion date.

It has been shown through examination of full-length copies of the HERV-L family of LTR retrotransposons that a correlation between the relative insertion date of an element and its G+C content likely exists. This upholds the previously described observations of mutation rates in gene/pseudogene pairs and instances of repetitive elements. Such a result was not expected, yet it leads to an interesting conclusion.

4 DISCUSSION

4.1 Shortcomings in Determining Fixed Mutation Directionality

Gene – Pseudogene Pairs. One of the shortcomings of the approach of looking at mutation rates in the gene-pseudogene case is the direction in which a substitution has occurred cannot be inferred with a high degree of certainty. A fairly good idea of the direction of mutation is obtained in the gene-pseudogene case, since genes are under high selective pressure, and therefore are likely to have fewer mutations than pseudogenes. However, there are regions such as synonymous wobble bases, where mutations can occur in genic regions with little consequence to fitness. One method of getting around this would involve constructing an evolutionary phylogeny of the genes in the data set using sequences from three or more related species. This would allow us to determine with greater confidence what the original nucleotide was in the human gene, and therefore directionality could be assigned more reliably, although still not with absolute certainty.

Such an approach is taken by Wolfe, Sharp and Li (1989) and Casane *et al*. (1987). While such a study may not currently be possible on a large set of genes due to the lack of large scale genomic sequence information for comparative species, it will shortly be possible in this era of genomics. Assemblies of the human, mouse and rat genomes (genome.ucsc.edu) are already available and other complete genomes are likely to become available in the not too distant future.

Repetitive elements. Repeats within the human genome are thought to have evolved in one of two ways (Shedlock and Okada, 2000). The master gene model (Shen, Batzer and Deringer, 1991) suggests only a few Alu loci are capable of amplification, and all subsequent copies found within the genome are direct descendants from these loci. The multiple source gene model (Matera and Hellman, 1990) states offspring copies of repetitive elements may also be amplified.

Depending on which model actually holds for the human genome, the study of substitution rates in repetitive element instances has some potential pitfalls as well. Substitution rates were calculated from the Repbase defined sequence to the copies found in the human genome. If the master gene model was the actual mechanism, the assumptions made should be correct to the degree that the Repbase sequences were the actual master genes. However, if the multiple source gene model was the mechanism, some of the substitutions reported could actually be due to a single substitution occurring at some point in time in an intermediary copy, which subsequently proliferated throughout the genome.

Since the issue of which mechanism was involved is hard to resolve, we cannot be completely confident in assuming the master gene model was the only mechanism at work. At the same time, comparing substitutions to the Repbase defined consensus sequences is promising, since the Repbase repeats have been arduously studied. Therefore, while intermediary subfamilies may still exist, it seems likely a majority of substitutions observed between the Repbase sequence and a particular copy in the genome are due to accumulated substitution events in the genomic loci rather than a long line of mutational intermediaries.

4.2 Repeat Composition

The resulting studies of repetitive elements give insight into how regions of high and low G+C content are maintained within the human genome. Examination of repetitive elements indicates their G+C content was not the driving factor into the appearance of high and low G+C regions. Rather it appears the unique sequence DNA mirrors the G+C pattern of the surrounding sequence. Thus, the repeat composition and distribution hypothesis cannot be accepted as the cause

for the maintenance of high and low G+C regions within the human genome.

4.3 Compositional Bias

The studies of G+C biases found in gene and pseudogene pairs as well as instances of repetitive elements indicate a high likelihood for compositional biases in substitution rates existing within the human genome. However, this compositional bias cannot be the cause for maintaining high and low G+C regions. This is due to the observed G+C biases suggesting the human genome is mutating towards A+T richness independently of the surrounding G+C content.

The ratio of G|C→A|T to A|T→G|C observed substitution rates is much higher in regions of high A+T. Such a finding suggests the human genome evolved from a G+C rich ancestral genome, and regions of high and low G+C arose as a result of the variance in mutation rates where some regions (high A+T regions) mutated faster than others (high G+C regions).

One of the difficulties with the selectionist, biased gene conversion, and mutational bias hypotheses is they are not mutually exclusive. For instance, it is possible a substitutional bias could be observed due to biased gene conversion. It is also possible substitutional biases are observed since they provide evolutionary advantages, and therefore fall under a selectionist hypothesis. Biased gene conversion could also provide advantageous changes, which would fall under selectionist theories.

Pseudogenes and repetitive elements are features likely to be under less selective pressure. In these regions, the bias observed is unlikely to have been caused by selection. The study of repetitive elements on the non-recombining chromosome Y yields similar results. This indicates biased gene conversion is not likely to be the cause of the compositional biases in substitution we observe in these regions.

As described earlier, this context dependent substitution rate could be caused by mechanisms involved in DNA synthesis such as the fidelity of α and β polymerases (Filipski, 1987), modification in the components of DNA synthesis (Muto and Osawa, 1987), or cytosine deamination (Fryxell and Zuckerkandl, 2000). Of course these mechanisms must be tied to germline cells in order for the mutations to become fixed in the population.

4.4 Shift Towards an A+T Rich Genome

Perhaps the most intriguing result of the substitutional bias study was that the G+C biases for nearly all of the cases looked at were less than 0.5, indicating no matter what the surrounding G+C context was, the rate of A|T→G|C substitutions seemed to be higher than the rate G|C→A|T substitutions. Such a result suggests over time, regions under less selective pressure within the human genome evolve into more A+T rich regions. The rate of this evolution appeared to be slower in high G+C regions, although it was still observed. The study of LTR retrotransposons within the human genome supports these results, since older copies tended to contain a higher A+T concentration.

4.4 Maintenance of High G+C Regions

The results suggest the human genome began from an ancestral genome higher in G+C composition that has evolved into a progressively lower G+C genome. However, the regions studied involved those features (pseudogenes and repetitive elements) less likely to be involved in selection. Since there are regions of high G+C content observed within the human genome, there is likely to be some other mechanism at work to preserve these regions. One explanation is that the presence of functionally and structurally important features in these regions makes the genome less tolerant of changes in their G+C composition. This would explain the high association between increasing G+C content and a higher gene density (Zoubak, Clay and Bernardi, 1996). If this is the case, the selectionist (and possibly biased gene conversion) hypotheses would hold true for these regions.

Comparing the G+C content of conserved and non-conserved regions in mouse and human could test this hypothesis. It is postulated conserved regions would have a higher G+C composition than non-conserved regions, if some sort of selection maintained high G+C regions. Otherwise, these regions would be subject to the compositional bias in substitution rates that are observed, and therefore the overall genome should mutate towards a higher A+T genome.

ACKNOWLEDGEMENTS

We would like to thank David States and Zhengyan Kan for helpful insight and review of this article. In addition, ER would like to thank members of the University of Louisville's Bioinformatics Research Group (BRG) for their support.

REFERENCES

Bedell, J.A., Korf, I., and Gish, W. 2000. MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics*, **16:** 1040-1041.

Belle, E.M., and Eyre-Walker, A. 2002. A test of whether selection maintains isochores using sites polymorphic for Alu and L1 element insertions. Genetics, **160:** 815-817.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., and Wheeler, D.L. 2002. GenBank. *Nucleic Acids Research*, **30:** 17-20.

Bernardi, G. 1993. The isochore organization of the human genome and its evolutionary history -- a review. *Gene*, **135:** 57-66.

Bernardi, G. 2000. Isochores and the evolutionary genomics of vertebrates. *Gene*, **241:** 3-17.

Casane, D., Boissinot, S., Chang, B.H., Shimmin, L.C., and Li, W. 1997. Mutation Pattern Variation Among Regions of the Primate Genome. *Journal of Molecular Evolution*, **45:** 216-226.

Castresana, J. 2002. Estimation of genetic distances from human and mouse introns. *Genome Biology*, **3:** 0028.1-0028.7.

Corodonnier, A., Casella, J.F., and Heidmann, T. 1995. Isolation of novel human endogenous retrovirus-like elements with foamy virus-related pol sequence. *Journal of Virology*, **69:** 5890-5897.

Cox, E.C. 1972. On the Organization of Higher Chromosomes. *Nature New Biology*, **239:** 133-134.

Cuny, G., Soriano, P., Macaya, G., and Bernardi, G. 1981. The Major Components of the Mouse and Human Genomes. *European Journal of Biochemestry,* **115:** 227-233.

D'Onofrio, G., and Bernardi, G. 1992. A universal compositional correleation among coding positions. *Gene,* **110:** 81-88.

Eyre-Walker, A. 1999. Evidence of Selection on Silent Site Base Composition in Mammals: Potential Implications for the Evolution of Isochores and Junk DNA. *Genetics*, **152:** 675-683.

Eyre-Walker, A., and Hurst, L.D. 2001. The evolution of isochores. *Nature Reviews Genetics,* **2:** 549-555.

Feng, Q., Moran, J.V., Kazazian, H.H.Jr., and Boeke, J.D. 1996. Human L1 Retrotransposons encodes a conserved endonuclease required for retrotransposition. *Cell*, **87:** 905-916.

Filipski, J. 1987. Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. *FEBS Letters,* **217:** 184-186.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller, W. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Research,* **8:** 967-974.

Francino, M.P., and Ochman, H. 1999. Isochores results from mutation not selection. *Nature,* **400:** 30-31.

Fryxell, K.J., and Zuckerkandl, E. 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Molecular Biology and Evolution,* **17:** 1371-1383.

Galtier, N., and Lobry, J.R. 1997. Relationships Between Genomic G+C Content, RNA Secondary Structures, and Optimal Growth Temperature in Prokaryotes. *Journal of Molecular Evolution*, **44:** 632-636.

Galtier, N., Piganeau, G., Mouchiroud, D., and Duret, L. 2001. GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. *Genetics*, **159:** 907-911.

Gardiner, K. 1996. Base composition and gene distribution: critical patterns in mammalian genome organization. *Trends in Genetics,* **12:** 519-524.

Griffiths, D.J. 2001. Endogenous retroviruses in the human genome sequence. *Genome Biology*, **2:** 1017.1-1017.5.

Gu, X., and Li W.-H. 1994. A Model for the Correlation of Mutation Rate with GC Content and the Origin of *GC*-Rich Isochores. *Journal of Molecular Evolution*, **38:** 468-475.

Hirose, Y., Takamatsu, M., and Harada, F. 1993. Presence of env genes in members of the RTVL-H family of human endogenous retrovirus-like elements. *Virology*, **192:** 52-61.

Hughes, S., Zelus, Z., and Mouchiroud, D. 1999. Warm-Blooded Isochore Structure in Nile Crocodile and Turtle. *Molecular Biology and Evolution*, **16:** 1521-1527.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature,* **409:** 860-921.

Jabbari, K., and Bernardi, G. 1998. CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families. *Gene,* **224:** 123-128.

Jurka, J. 1998. Repeats in genomic DNA: mining and meaning. *Current Opinion in Structural Biology,* **8:** 333-337.

Jurka, J. 2000. Repbase update: a database and an electronic journal of repetitive elements. *Trends in Genetics,* **16:** 418-420.

Kent, J.W., and Haussler, D. 2001. Assembly of the Working Draft of the Human Genome with GigAssembler. *Genome Research,* **11:** 1541-1548.

Knight, R.D., Freeland, S.J., and Landweber, L.F. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biology,* **2:** research0010.1-0010.13.

Lahn, B.T., Pearson, N.M., and Jegalian, K. 2001. The human Y chromosome, in the light of evolution. *Nature Reviews Genetics*, **2:** 207-216.

Lodish, H., Baltimore, D., Berk, A., Zipursky, S.L., Matsudaria, P., and Darnell, J. 1995. *Molecular Cell Biology*. New York: Scientific American Books.

Macaya, G., Thiery, J.P., and Bernardi, G. 1976. An approach to the organization of eukaryotic genomes at a macromolecular level. *Journal of Molecular Biology,* **108:** 237-254.

Matassi, G., Labuda, D., and Bernardi, G. 1998. Distribution of the mammalian-wide interspersed repeats (MIRs) in the isochores of the human genome. *FEBS Letters,* **439:** 63-65.

Matera, A.G., Hellman, U., and Schmid, C.W. 1990. A transpositionally and transcriptionally competent Alu subfamily. *Molecular Cell Biology*, **10:** 5424-5432.

Muto, A., and Osawa, S. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. *Proceedings of the National Academy of Sciences USA,* **84:** 166-169.

Nekrutenko, A., and Li, W.H. 2000. Assessment of compositional heterogeneity within and between eukaryotic genomes. *Genome Research,* **10:** 1986-1995.

Ohama, T., Yamao, F., Muto, A., and Osawa, S. 1987. Organization and Codon Usage of the Streptomycin Operon in *Micrococcus luteus*, a Baterium with a High Genomic G+C Content. *Journal of Bacteriology,* **169:** 4770-4777.

Pavlíček, A., Jabbari, K., Pačes, J., Pačes, V., Hejnar, J., and Bernardi, G. 2001. Similar integration but different stability of Alus and LINEs in the human genome. *Gene*, **276:** 39-45.

Piganeau, G., Mouchiroud, D., Duret, L., and Gautier, C. 2002. Expected Relationship Between the Silent Substitution Rate and the GC Content: Implications for the Evolution of Isochores. *Journal of Molecular Evolution*, **54:** 129-133.

Prak, E.T., and Kazazian, Jr., H.H. 2000. Mobile elements and the human genome. *Nature Reviews Genetics*, **1:** 134-144.

Pruitt, K.D., and Maglott, D.R. 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research,* **29:** 137-140.

Robinson, H., Gao, Y., Mccray, B.S., Edmondson, S.P., Shriver, J.W., and Wang, A.H.J. 1998. The hyperthermophile chromosomal protein Sac7d sharply kinks DNA. *Nature*, **392:** 202-205.

Rouchka, E.C., and States, D.J. 2002. Compositional Analysis of Homogeneous Regions in Human Genomic DNA. Technical Report, Washington University Department of Computer Science, WUCS-2002-2.

Shedlock, A.M., and Okada, N. 2000. SINE insertions: powerful tools for molecular systematics. *Bioessays*, **22:** 148-160.

Shen, M.R., Batzer, M.A, and Deninger, P.L. 1991. Evolution of the master Alu gene(s). *Journal of Molecular Evolution*, **33:** 311-320.

Smith, N.G.C., and Eyre-Walker, A. 2001. Synonymous Codon Bias Is Not Caused by Mutation Bias in G+C Rich Genes in Human. *Molecular Biology and Evolution*, **18**: 982-986.

Smit, A.F.A. 1993. Identification of a new, abundant superfamily of mammalian LTR-transposons. *Nucleic Acids Research*, **21:** 1863-1872.

Smit, A.F.A. 1996. The origin of interspersed repeats in the human genome. *Current Opinion in Genetics & Development*, **6:** 743-748.

Smit, A.F. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current Opinion in Genetics And Development* **9:** 657-663.

Sueoka, N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. *Proceedings of the National Academy of Sciences USA*., **48:** 582-592.

Taguchi, H., Konishi, J., Ishii, N., and Yoshida, M. 1991. A chaperonin from a thermophilic bacterium *Thermus thermophylus*, that controls refolding of several thermophilic enzymes. *The Journal of Biological Chemistry*, **266:** 22411-22418.

Tilford, C.A., Kuroda-Kawaguchi, T., Skaletsky, H., Rozen, S., Brown, L.G., Rosenberg, M., McPherson, J.D., Wylie, K., Sekhon, M., Kucaba, T.A., et al. 2001. A physical map of the human Y chromosome. *Nature*, **409:** 943-945.

Tristem, M. 2000. Identification and Characterization of Novel Human Endogenous Retrovirus Families by Phylogenetic Screening of the Human Genome Mapping Project Database. *Journal of Virology*, **74:** 3715-3730.

Wada, A., and Suyama, A. 1986. Local stability of DNA and RNA secondary structure and its relation to biological functions. *Progress in Biophysics and Molecular Biology*, **47:** 113-157.

Wolfe, K.H. 1991. Mammalian DNA Replication: Mutation Biases and the Mutation Rate. *Journal of Theoretical Biology*, **149:** 441-451.

Wolfe, K.H., Sharp, P.M., and Li, W-H. 1989. Mutation rates differ among regions of the mammalian genome. *Nature,* **337:** 283-285.

Zoubak, S., Clay, O., and Bernardi, G. 1996. The gene distribution of the human genome. *Gene,* **174:** 95-102.