Available online at www.sciencedirect.com
ScienceDirect
Genomics 88 (2006) 527–534
GENOMICS
www.elsevier.com/locate/ygeno

Mutational spectrum in the recent human genome inferred by single nucleotide polymorphisms

Cizhong Jiang [a], Zhongming Zhao [a,b,*]

[a] *Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, P.O. Box 980126, Richmond, VA 23298-0126, USA*
[b] *Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23284, USA*

Received 3 April 2006; accepted 6 June 2006
Available online 24 July 2006

Abstract

So far, there is no genome-wide estimation of the mutational spectrum in humans. In this study, we systematically examined the directionality of the point mutations and maintenance of GC content in the human genome using ~1.8 million high-quality human single nucleotide polymorphisms and their ancestral sequences in chimpanzees. The frequency of C→T (G→A) changes was the highest among all mutation types and the frequency of each type of transition was approximately fourfold that of each type of transversion. In intergenic regions, when the GC content increased, the frequency of changes from G or C increased. In exons, the frequency of G:C→A:T was the highest among the genomic categories and contributed mainly by the frequent mutations at the CpG sites. In contrast, mutations at the CpG sites, or CpG→TpG/CpA mutations, occurred less frequently in the CpG islands relative to intergenic regions with similar GC content. Our results suggest that the GC content is overall not in equilibrium in the human genome, with a trend toward shifting the human genome to be AT rich and shifting the GC content of a region to approach the genome average. Our results, which differ from previous estimates based on limited loci or on the rodent lineage, provide the first representative and reliable mutational spectrum in the recent human genome and categorized genomic regions.

Keywords: Single nucleotide polymorphisms; Mutational spectrum; Mutation direction; Exons; CpG islands; GC content; Genome evolution

Point mutation does not occur randomly and it depends on many factors such as sequence context, methylation of CpG dinucleotides, and categorized genomic regions [1–5]. One long-standing interest has been to examine the directionality of point mutation and to explore how nucleotide changes shift the nucleotide composition in the different regions of genomes [6–8]. Early studies of 13 mammalian pseudogene sequences revealed that the mutation direction is nonrandom [6,7]. Other surveys from noncoding regions indicated that G/C to A/T (G/C→A/T) changes might occur more frequently than A/T→G/C changes (e.g., Ref. [9]), suggesting that human noncoding and pseudogene sequences have the tendency to become AT-rich. However, these estimates were based on a very limited number of loci and their observations were often inconsistent [10,11]. Moreover, the frequencies of a pair of nucleotide changes based on the DNA strand symmetry, e.g., G→A versus C→T, were not observed to be nearly the same as expected [7]. So far, there is no systematic estimation of the mutational spectrum in the human genome or categorized genomic regions (e.g., intergenic regions, exons, introns, CpG islands).

The GC content in the nonrepetitive fraction of the human genome is ~40.31%; this is lower than that in the mouse (~41.26%) and rat (~41.61%) genomes [12]. Cooper et al. [12] recently inferred the direction of the nucleotide substitutions in the mouse and rat lineages based on large-scale alignments of human, mouse, and rat genomic sequences. Their results showed an excess of A/T→G/C changes (50.3% in rats and 47.2% in mice) over G/C→A/T (35.0% in rats and 37.9% in mice). Although their results determined a relative increase in GC content in the rat genome, the excess of AT→GC in rodents is opposite to what has been previously observed in noncoding or pseudogene sequences; thus, further investigations are warranted. Moreover, rats and mice diverged ~12–24 million years

* Corresponding author. Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, P.O. Box 980126, Richmond, VA 23298-0126, USA. Fax: +1 804 828 1471.
E-mail address: zzhao@vcu.edu (Z. Zhao).

0888-7543/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2006.06.003

(Myr) ago from their common ancestor, while humans and mice diverged ~75 Myr ago [13]. Such a cross-species comparison has limitations in inferring the mutation direction in the human genome because recurrent mutations may have occurred after the divergence of humans and rodents. Indeed, higher substitution rates were found in the rodent lineage compared to the human lineage and only a small portion of these genomes could be aligned, e.g., ~30% of the rat genome was aligned to the mouse genome [13].

At present, more than 10 million single nucleotide polymorphisms (SNPs) have been discovered in the human genome. This provides us with an alternative approach to examine the mutation patterns in the human genome. The mutation direction for each SNP can be inferred by comparing the SNP with its outgroup chimpanzee sequences. This approach is unique since the mutation direction inferred by a SNP and its outgroup sequence is likely to be true because the divergence time between humans and chimpanzees is ~6 Myr ago and most SNPs occurred in a relatively recent time period, i.e., less than 1 Myr ago [9,14]. This should provide more reliable estimates than roughly comparing the aligned genomic sequences. The major limitation is that these polymorphic sites account for only a small portion (~0.3%) of the human genome.

In this study, we first evaluated whether we could infer the mutation direction by comparing human SNPs with chimpanzee sequences only, an approach that has been commonly accepted but never evaluated. This evaluation was conducted by comparing the SNPs in 44 ENCODE regions [15] with two outgroup species, the chimpanzee and baboon. We next systematically inferred the mutation direction in the human genome and categorized genomic regions by comparing two alleles of each human SNP with its ancestral chimpanzee sequence. Finally, we compared the frequencies of G/C→A/T versus A/T→G/C changes and found that the GC content in the human genome was overall not in equilibrium. This study provides a comprehensive genome-wide view of the mutation patterns in recent human history.

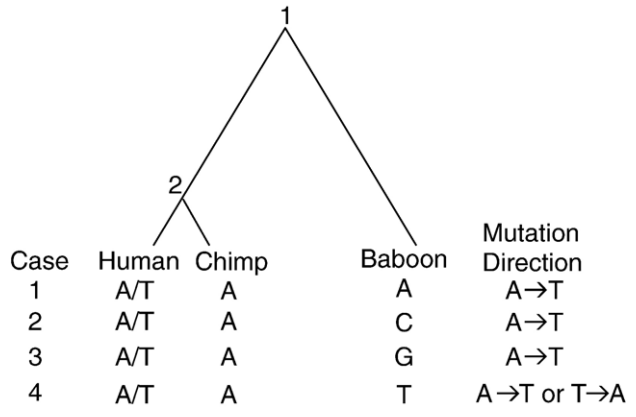

Fig. 1. Inference of mutation direction for human SNPs by maximum parsimony principle. In this example, a human SNP (A/T) has its corresponding chimpanzee (Chimp) allele A. In the first case, the baboon carries the same allele as the chimpanzee (i.e., A). Only one mutation event is needed and this occurred in the branch from node 2 to human. The mutation direction is thus inferred to be A→T. In the second and third cases, the baboon carries an allele (C or G) that is different from both human alleles (A and T). A minimum of two mutational events are needed. However, node 2 should always be allele A no matter how the nucleotide changes in the branches from 1 to 2 or from 1 to baboon. Thus, the mutation direction is inferred to be A→T. In the last case, the baboon carries an allele T. At least two mutational events are needed, but the allele at node 2 cannot be resolved and the mutation direction cannot be inferred.

Table 1
Proportion of inference in each case by ENCODE data

SNP category	Case 1 [a]	Case 2 or 3	Case 4
Transition (33,302 [b])	0.86	0.03	0.11
Transversion (16,645)	0.90	0.07	0.04
Total (49,947)	0.87	0.04	0.09

[a] Four cases are illustrated in Fig. 1.
[b] Number of SNPs.

Results

Inference of mutation direction by ENCODE data

In this study, we inferred the mutation direction of a human SNP by comparing it with its chimpanzee ancestral allele. We first evaluated the reliability of this approach. We retrieved 95,353 human SNPs in 44 ENCODE regions from the UCSC ENCODE browser (http://genome.ucsc.edu/ENCODE/). Among them, 49,947 SNPs were selected because one of their two alleles matched the chimpanzee allele. The mutation direction was inferred by the maximal parsimony principle when these SNPs were compared with their corresponding ancestral alleles in chimpanzees using the baboon as an additional outgroup reference (see Materials and methods). Fig. 1 illustrates four possible cases in the inference of mutation direction for one SNP type (A/T). Table 1 summarizes the proportion of inference in each case for transition, transversion, and all SNPs. Among the 49,947 SNPs analyzed, 87% could be inferred in case 1 and 4% in cases 2 and 3. This accounts for 91% that could be reliably inferred by comparing the ancestral information only in chimpanzees without any other outgroup reference. The proportion was higher for transversional changes (97%) than for transitional changes (89%). This difference was further seen for each SNP category (i.e., A/G, C/T, A/C, G/T, A/T, and C/G; Supplementary Table S1). The higher proportion in case 4 for transitional SNPs might reflect the high transition rate and hypermutability of methylated CpG dinucleotides in vertebrate genomes [16].

According to the maximal parsimony principle, there would be two mutational events in case 4. One mutation would have always occurred in the branch from node 2 to human in Fig. 1. The other mutation might have occurred in one of the other three branches: from node 2 to chimpanzee, from node 1 to node 2, and from node 1 to baboon. Only when the mutation occurred in the branch from node 2 to chimpanzee would the mutation direction inferred be different from that in cases 1–3, i.e., T→A. Since the divergence time between humans and chimpanzees (~6 Myr) is much shorter than that between humans and baboons (~25 Myr) [17,18], assuming the constant point mutation rate in these branches, the probability that the mutation occurred in the branch from node 2 to chimpanzee would be the

Table 2
Frequencies (%) of nucleotide changes in the human genome and categorized genomic regions

Category	GC (%)[a]	No. of SNPs	A→G	T→C	G→A	C→T	A→C	T→G	G→T	C→A	A→T	T→A	G→C	C→G
Genome	40.31	1,785,712	16.7	16.7	17.4	17.4	4.0	4.0	4.2	4.2	3.3	3.3	4.4	4.4
Intergenic regions	40.18	1,142,121	16.8	16.8	16.9	17.0	4.1	4.1	4.3	4.3	3.5	3.5	4.4	4.3
Genes	42.28	339,005	16.6	16.7	18.2	18.1	4.0	3.9	3.8	3.8	2.9	3.0	4.5	4.6
Introns	41.72	418,086	16.8	16.9	17.8	17.8	4.0	4.0	3.8	3.8	3.1	3.1	4.6	4.5
Exons	52.10	12,430	12.2	12.2	26.7	25.5	2.3	2.1	3.2	3.5	1.8	1.7	4.1	4.8
CpG islands	62.07	25,913	9.0	9.0	22.1	22.4	2.6	2.7	6.5	6.1	1.9	2.0	7.8	7.9
Promoter CpG islands	63.38	2731	8.0	9.3	21.7	21.6	2.9	2.8	6.6	6.5	2.0	1.9	8.5	8.2

[a] The average GC content in the sequences we analyzed.

lowest, or much less than 50%. This leads to more than 95.5% (87%+4%+(9%/2)) reliability. In summary, this evaluation indicates that it is very reliable to infer the mutation direction by comparing human SNPs with their chimpanzee sequences only.

Mutational spectrum in the human genome

A total of 1,785,712 human SNPs could be mapped to the chimpanzee genome based on the MegaBLAST search results using the stringent criteria described under Materials and methods. The ancestral alleles of these SNPs were obtained from the alignments. Table 2 summarizes the inferred mutation directions of these SNPs. In the nonrepetitive sequences across the human genome, which is shown in the first row of Table 2, the frequency of each transitional type (e.g., G→A) was ~17%, approximately fourfold that for each transversional type. The most frequently observed nucleotide change was C→T (G→A), which, on average, was 17.4% among all mutations. The frequency of G→A was nearly the same as that of C→T, reflecting the complementary DNA strand symmetry. This feature was also observed in other pairs of nucleotide changes (e.g., A→C and T→G). Moreover, the frequency of C→T (G→A) was slightly higher than that of T→C (A→G). We note that the GC content in the nonrepetitive genomic sequences used in this study was 40.31% (Table 2). Therefore, the differences above were larger after the GC content was normalized to be 50% (data not shown). Finally, among all nucleotide changes, A→T (T→A) changes were the least frequently observed.

Throughout the remaining text, we shall denote G:C→A:T for the pair of mutations G→A and C→T. The other pairs will follow the same rules.

Mutational spectrum in the categorized genomic regions

Table 2 also includes the mutation direction among the categorized genomic regions. The results are summarized as the following four points. First, the mutational spectrum observed in intergenic regions was not much different from that of the whole genome. This is likely due to the large portion of intergenic regions (and introns as well, Table 2) in the human genome. Second, in genic regions, in which GC content (42.28%) was higher than in intergenic regions (40.18%), the frequency of G:C→A:T was moderately higher than that in intergenic regions. This contrasts with a slightly lower frequency of A:T→G:C in genic regions than in intergenic regions. For the transversional changes, except for G:C→C:G, the frequency of each type in genic regions was lower than that in intergenic regions. Third, in exonic regions where the GC content was averaged to be 52.10%, the frequencies of C→T and G→A were 25.5 and 26.7%, respectively. These frequencies stand out as the highest among all categories, even higher than those in the CpG island sequences, which had the highest GC content (62.07%) among the genomic categories. This resulted in lower frequencies of other mutation types, except for G:C→C:G, in exons than in introns.

Finally, in the CpG island sequences mutations from nucleotide G or C to any other nucleotide dominated. For example, G:C→T:A and G:C→C:G had the highest frequencies among the categories. We further analyzed those CpG islands located in the promoter regions. While the mutational spectrum was overall similar to that in the CpG islands, a lower frequency of G:C→A:T was observed, reflecting the absence of the CpG effects in the promoter-associated CpG islands (Table 2).

Table 3
Frequencies (%) of nucleotide changes in intergenic regions

GC (%)[a]	No. of SNPs	A:T→G:C	G:C→A:T	A:T→C:G	G:C→T:A	A:T→T:A	G:C→C:G
<35 (31.36)[b]	205,817	38.1	28.2	9.5	8.2	8.6	7.3
35–40 (37.31)	179,725	35.4	32.1	8.5	8.4	7.1	8.5
40–45 (42.20)	127,842	32.5	35.4	7.8	8.7	6.3	9.3
45–50 (47.18)	75,491	28.8	40.1	6.8	9.0	5.3	10.0
50–55 (52.21)	44,242	25.1	45.1	6.0	9.1	4.5	10.3
55–60 (57.20)	25,170	22.1	48.3	5.0	9.6	3.8	11.3
≥60 (63.92)	15,444	19.0	49.7	4.7	10.9	3.0	12.6

[a] GC content in a nonrepetitive intergenic region was obtained in a window with a size of 500 nucleotides.
[b] The average GC content (%) in the subcategory of intergenic regions.

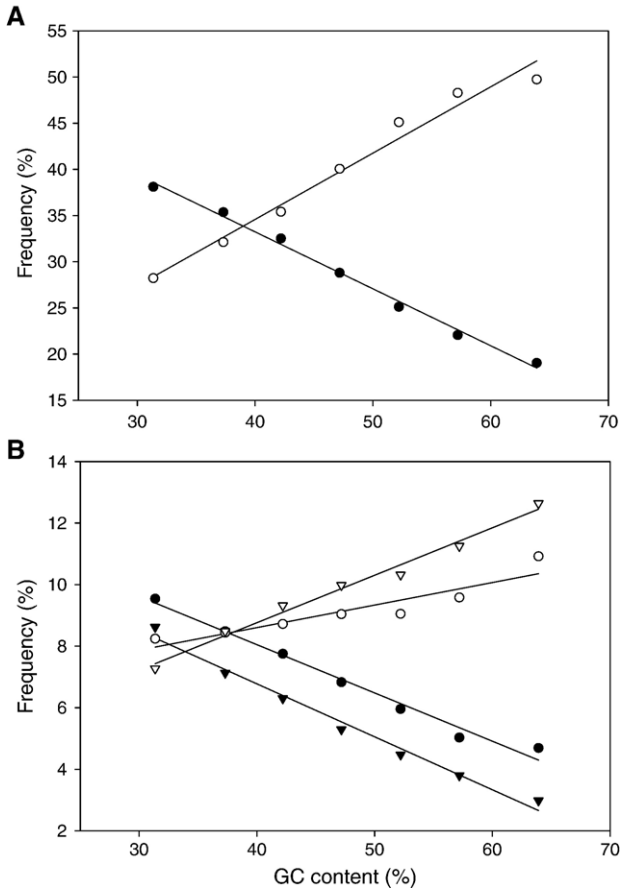

Fig. 2. Linear regression of the frequency of each mutational type versus GC content in intergenic regions. Intergenic regions were subgrouped into seven GC-content bins. (A) Linear regression of the frequency of G:C→A:T (○, R^2=0.98, p<0.0001) and A:T→G:C (●, R^2=0.99, p<0.0001) versus GC content. (B) Linear regression of the frequency of G:C→T:A (○, R^2=0.86, p=0.003), G:C→C:G (▽, R^2=0.99, p<0.0001), A:T→C:G (●, R^2=0.98, p<0.0001), and A:T→T:A (▼, R^2=0.98, p<0.0001) versus GC content. A different scale was used on the y axis in A and B.

Mutational spectrum in intergenic regions

To test whether the mutational spectrum depends on the GC content we further examined the intergenic regions with different GC content since mutations in this category are usually assumed to be selectively neutral. Table 3 shows the mutational spectrum in each subcategory of intergenic regions. Since the frequencies of the two types in each pair were nearly the same, we combined each pair to save space. The results indicate that when GC content increased, the frequencies of changes from C or G to any other nucleotide increased; conversely, the frequencies from A or T to any other nucleotide decreased. A further statistical analysis indicated a significant positive correlation between any mutation type from G/C (i.e., G:C→A:T, G:C→T:A, G:C→C:G) and GC content and, conversely, a significant negative correlation between any mutation type from A/T (i.e., A:T→G:C, A:T→C:G, A:T→T:A) and GC content (Fig. 2). The extent of the changes was stronger in the transitional types than in the transversional types. For example, the frequency of G:C→A:T changes increased from 28.2 (GC content <35%) to 49.7% (GC content ≥60%). Correspondingly, the transition over transversion ratio increased when the GC content increased.

Mutational spectrum after excluding the CpG effects

The mutation rate of methylated CpG to TpG/CpA was estimated to be 10–50 times higher than other transitional changes [16]. Such CpG effects elevate the frequency of G:C→A:T or, more specifically, CpG→TpG/CpA. To examine the CpG effects on the mutational spectrum in the human genome, we compared the mutation direction by (1) removing the SNPs that occurred at the CpG sites and (2) removing CpG→TpG/CpA. These results are shown in Supplementary Tables S2 and S3, respectively. As expected, the frequency of G:C→A:T decreased drastically among all categories after removal, and the extent of the decrease in G:C→A:T after removal of CpG→TpG/CpA was stronger than after removal of the CpG sites. Conversely, the frequencies of all other changes, except for G:C→T:A and G:C→C:G in the (promoter) CpG islands, increased. This resulted in a smaller frequency difference among mutation types. For the total SNPs, the frequency of G:C→A:T decreased from 34.8 to 26.8% after removal of the CpG sites and to 26.4% after removal of CpG→TpG/CpA. This decrease was similarly observed in intergenic and intronic regions (Fig. 3).

Exonic regions showed the strongest decrease among the categories (Fig. 3). The frequency of G:C→A:T decreased by 17.1% when the CpG sites were removed and by 18.6% when CpG→TpG/CpA mutations were removed. The decrease was more than twofold that in intergenic regions and nearly twofold that in introns (Table 2, Supplementary Tables S2 and S3). We next compared the extent of the decrease in exons with that in intergenic regions with similar GC content (50–55%). The results indicate that the extent of the decrease after removal was much stronger in exons than in intergenic regions (Fig. 4A).

Interestingly, in the CpG islands, the frequency of G:C→A:T decreased by only 7.6% after removal of the CpG sites, much less than the 12.4% decrease after removal of CpG→TpG/CpA. This difference became even smaller, i.e., 5.6 and 10.4%, respectively, when we examined the promoter-associated CpG

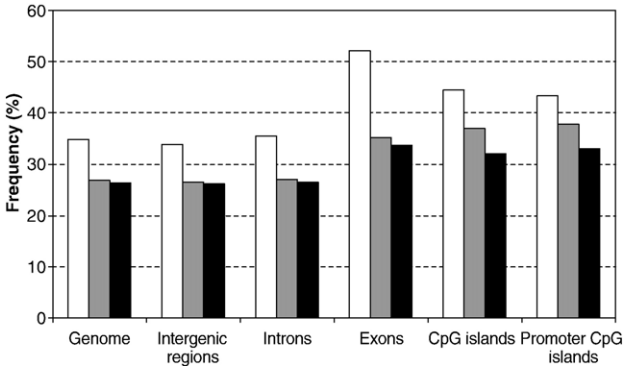

Fig. 3. Frequency of G:C→A:T mutations among the genomic categories. White bar, original frequency; gray bar, after removal of the CpG sites; black bar, after removal of CpG→TpG/CpA.

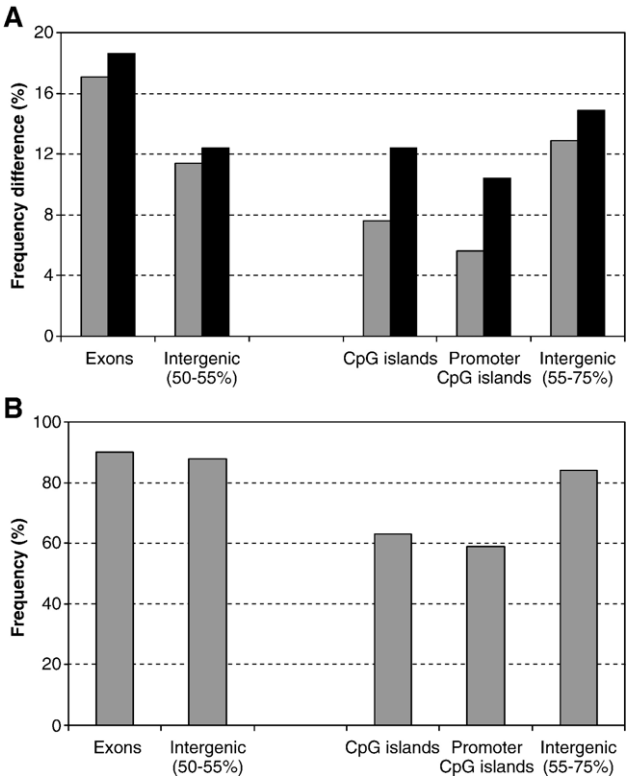

Fig. 4. Effects of mutations at the CpG sites. (A) Frequency changes in exons and CpG islands after removal of the CpG sites or CpG→TpG/CpA. Frequency difference between the original data and after removal of CpG sites is labeled in gray and frequency difference between the original data and after removal of CpG→TpG/CpA is labeled in black. (B) Frequency of CpG→TpG/CpA mutations at the CpG sites.

islands. We next compared the extent of the decrease in the CpG islands with that in intergenic regions with similar GC content (55–75%). In Fig. 4A, the extent of the decrease is much smaller in the (promoter-associated) CpG islands than in the corresponding intergenic regions when the CpG sites were removed (12.9%). However, the extent of the decrease became less remarkable (e.g., 12.4% in the CpG islands vs 14.9% in intergenic regions) when CpG→TpG/CpA mutations were removed. This indicated that mutations other than CpG→TpG/CpA (e.g., CpG→GpG/CpC) at the CpG sites may occur more frequently in the CpG islands than in other genomic regions. Indeed, this was confirmed when we compared the frequency of CpG→TpG/CpA at the CpG sites (Fig. 4B).

GC content disequilibrium in the human genome

Each nucleotide at a site may change to any of the other three nucleotides. Based on the 1,785,712 SNPs in this study, the frequencies of mutations from nucleotide A, T, C, or G to any other nucleotide were 24.0, 24.0, 26.0, and 26.0%, respectively. And these mutations resulted in nucleotides A, T, C, and G by the frequencies 24.8, 24.9, 25.2, and 25.1%, respectively. Further, the frequency of G/C→A/T was 43.1% among all mutations, higher than that (41.4%) of A/T→G/C. This indicated that, in the recent human genome, there is a trend to have more G or C changes to A or T. This trend was found among all genomic categories (Table 4), suggesting that the point mutations may tend to decrease GC content across the human genome.

We next examined whether the GC content in the human genome has been under statistical equilibrium. If selection is not considered, the mean GC content at equilibrium (P_{GC}) is expected to be

$$P_{GC} = v/(u+v), \tag{1}$$

where u and v are the rates of G/C→A/T and A/T→G/C, respectively [19]. This equation can be transformed to

$$u/v = \frac{1-P_{\mathrm{GC}}}{P_{\mathrm{GC}}} \tag{2}$$

For SNP data,

$$u/v = \frac{N_u/f_{\mathrm{GC}}}{N_v/(1-f_{\mathrm{GC}})} = \frac{N_u}{N_v} \times \frac{1-f_{\mathrm{GC}}}{f_{\mathrm{GC}}} \tag{3}$$

where N_u and N_v are the numbers of G/C→A/T and A/T→G/C, respectively, and f_{GC} is the GC content. Eq. (3) becomes Eq. (2) when N_u is close to N_v, which is under the GC content equilibrium. In Table 4, the u/v ratio was 1.54 in the whole genome, which is nearly the same as the expected (1.48) from the GC content at equilibrium. This pattern was similarly found in intergenic and intronic regions. However, the u/v ratio was 1.89 in exons, twice the expected value (0.92). In the CpG islands, the u/v ratio was more than twofold the expected (0.61). In theory, the expected u/v ratio at equilibrium decreases when GC content increases and becomes less than 1 when GC content is higher than 50%. These results suggest, assuming there is no selection, that (1) the GC content in the overall genome, intergenic, or intronic regions is nearly at equilibrium (reexamined in the next paragraph); (2) the GC content in exons and CpG islands is not under equilibrium, with a trend to decrease the GC content; and (3) the point mutations created slightly more G/C→A/T than expected in the overall genome, intergenic, or intronic regions, but much more G/C→A/T in exons and CpG islands.

We next examined the u/v ratios in intergenic regions by different GC content. Table 5 shows that the observed u/v ratio was much smaller than the expected when the GC content was low (e.g., <35%), then became close to the expected when the GC content was close to the genome average, and finally, became greater than the expected when the GC content was

Table 4
Mutations between nucleotides G/C and A/T in the human genome

Category	GC (%)	G/C→A/T (%)	A/T→G/C (%)	u/v	$(1-P_{GC})/P_{GC}$
Genome	40.31	43.1	41.4	1.54	1.48
Intergenic	40.18	42.5	41.8	1.52	1.49
Gene	42.28	43.9	41.1	1.46	1.37
Introns	41.72	43.2	41.7	1.45	1.40
Exons	52.10	59.0	28.8	1.89	0.92
CpG islands	62.07	57.0	23.3	1.49	0.61
Promoter CpG islands	63.38	56.4	22.9	1.42	0.58

higher than 45%. A more detailed analysis of intergenic regions whose GC content was very close to the genome average (e.g., 38–39, 39–40, 40–41, 41–42%) further enhanced this trend (data not shown). The results clearly indicate that the overall observations in Table 4 were not sufficient and the GC content in intergenic regions was not in equilibrium. This implies that the human genome is overall not in equilibrium, with a trend for the GC content of a region to approach the human genome average.

Discussion

We have examined the directionality of the point mutations and maintenance of GC content in the human genome and categorized genomic regions using ~1.8 million high-quality SNPs and their ancestral information. In the human genome, the frequency of C→T (G→A) was 17.4%, which is higher than that (16.7%) of T→C (A→G). The frequency of each type of transition was approximately fourfold that of each type of transversion. In intergenic regions, when the GC content increased, the frequencies of mutations from G or C (G:C→A:T, G:C→T:A, G:C→C:G) increased and, conversely, the frequencies of mutations from A or T (A:T→G:C, A:T→C:G, and A:T→T:A) decreased. In exons, the frequency of G:C→A:T was the highest among the genomic categories. This was contributed mainly by the "excess" mutations at the CpG sites. In contrast, mutations at the CpG sites, or CpG→TpG/CpA mutations, occurred less frequently in the CpG islands compared with similar intergenic regions. Further, the mutations other than CpG→TpG/CpA at the CpG sites may occur more frequently in the CpG islands. The comparison of the rates of G/C→A/T and A/T→G/C suggests that GC content is generally not under equilibrium in the genome. Because most SNPs occurred recently, these results provided a detailed view of the mutational spectrum in the recent human genome. Critically, our estimates provide an updated and reliable mutational spectrum in the human genome.

The frequencies of nucleotide changes in this study were different from the previous estimates by other approaches. First, we compared our results with those based on the 13 mammalian pseudogene sequences [6,7]. Because those results were presented by relative substitution frequencies, for comparison purposes, we recalculated their data by the frequency of each mutation type among all observed mutations (Supplementary Table S4). The frequencies of G→A (22.2%), C→T (20.8%), G→T (7.7%), and C→A (6.5%) were much higher than those of our results, whether in the whole genome, intergenic, or intronic regions (Table 2). Conversely, the frequencies of A→G (8.7%) and T→C (8.0%) were much lower than our results. In those pseudogene sequences, the frequencies in each pair (e.g., A→C and T→G) were not nearly the same as expected, indicating that these estimates were not reliable due to the small number of loci. Second, we compared our results with those from the human SNPs in chromosome 21 [14]. Our reanalysis of their 19,985 SNPs indicated a slightly different mutational spectrum from the one based on our 30,191 SNPs in the same chromosome. However, the mutational spectrum in chromosome 21 was different from the global spectrum in our study, indicating its insufficiency to represent the human genome (Table 2 and Supplementary Table S4). From the data in Watanabe et al. [14], we also found the frequency of G:C→A:T to be lower than that of A:T→G:C, though the opposite has been widely observed previously, as well as in this study [7,9]. This might be due to the small number of SNPs used, the criteria of selection of SNPs in their study, or both. Third, a recent estimate of the mutational spectrum based on the global alignments of human, mouse, and rat genomic sequences indicated that the frequency of G/C→A/T was 35.0% in the rat lineage and 37.9% in the mouse lineage [12]; both were less than the 43.1% estimated based on our human SNPs and the 43.4% estimated based on chimpanzee SNPs (unpublished results). Among them, the frequency of G:C→A:T was 28.3 (rat), 30.7 (mouse), 34.8 (human), and 35.3% (chimpanzee). Interestingly, this trend was just the opposite of the GC content in the nonrepetitive fraction of these four genomes: 41.61 (rat), 41.26 (mouse), 40.31 (human), and 40.02% (chimpanzee) [12]. This may, at least partially, explain the GC content difference in these four genomes.

Our analysis of the mutation patterns based on the SNPs revealed that the GC content is not in equilibrium in the recent human genome. Overall, we observed more G/C→A/T than A/T→G/C mutations, no matter which genomic categories we examined (Table 4). This suggests that the human genome has currently been shifting toward being AT-rich, given that selection is not an factor. Our further analysis of the u/v ratio in intergenic regions revealed a surprising trend: the regions with lower GC content tend to have a lower u/v ratio than the expected and the regions with higher GC content tend to have a higher u/v ratio than the expected. Moreover, we found that the observed u/v ratios in exons and CpG islands were much higher than the expected, suggesting their GC content might become lower by point mutations. A recent study supporting this found a 30–60% higher mutation rate in exons than in the noncoding regions, due mainly to the overabundance of synonymous sites involving the CpG dinucleotides in exons [3]. In the CpG islands, evidence was found to support a loss of ancestral CpG islands in the mouse and human genomes ([20,21] and unpublished results). In this "loss of CpG islands" model, CpG dinucleotides in some islands might become de novo methylated and thus more likely to be mutated to TpG because of the hypermutability of the methylated CpG dinucleotides. Furthermore, the strong difference between the observed u/v ratios and the expected values in exons and CpG islands might also be due to selection pressure and gene conversion [1,8].

Table 5
Mutations between nucleotides G/C and A/T in intergenic regions

Category	GC (%)	G/C→A/T (%)	A/T→G/C (%)	u/v	$(1 - P_{GC})/P_{GC}$
<35%	31.36	36.5	47.6	1.67	2.19
35–40%	37.31	40.6	43.8	1.55	1.68
40–45%	42.20	44.1	40.3	1.50	1.37
45–50%	47.18	49.1	35.6	1.54	1.12
50–55%	52.21	54.1	31.1	1.59	0.92
55–60%	57.20	57.9	27.1	1.60	0.75
≥60%	63.92	60.7	23.7	1.44	0.56

One important issue is how reliable our analysis is. In this study, we applied stringent criteria for the selection of SNPs and mapping of the human SNPs to the chimpanzee genome. Among 8,353,499 available SNPs, we selected only the SNPs that satisfied the following criteria: (1) non-insertion/deletion, (2) biallelic, (3) validated, (4) longer than 100 nucleotides at each flanking side, (5) uniquely mapped in the human non-repetitive sequences, and (6) uniquely mapped in the chimpanzee genome. To determine whether a human SNP uniquely maps in the chimpanzee genome, certain parameters needed to be satisfied in the process of the MegaBLAST search results (see Materials and methods). This resulted in the final set of 1,785,712 SNPs used in this study. The frequencies of two mutation types in a pair (e.g., G→A versus C→T) were nearly the same in our analysis, indicating that the number of SNPs was sufficient. In addition, we had applied more stringent criteria to determine the ancestral information of SNPs and found the results were essentially the same. For example, in the SNP mapping procedure, we extended the criteria described under Materials and methods to require that (1) the alignment length was exactly 201, (2) the identity of the alignment was at least 98%, (3) no gap was in the alignment, and (4) the position of the SNP was at the center (i.e., at position 101). This procedure resulted in 1,072,068 SNPs that generated the same results (data not shown).

In conclusion, we systematically examined the mutational spectrum in the whole human genome and categorized regions using human SNPs. The results revealed how the point mutations had shifted the nucleotide compositions in the recent human genome, including the noncoding, exonic, and CpG island regions. Our results indicate that the previous estimates based on the limited loci were not accurate and not representative in the human genome. The comparative analysis of the *u/v* ratios suggests that the GC content in the categorized human genomic regions is not in equilibrium, with a current trend toward approaching the genome average. Our analysis provides the first detailed estimate of the mutational spectrum in the recent human genome and its categorized genomic regions.

Materials and methods

Inference of mutation direction by ENCODE data

The human, chimpanzee, and baboon sequences in 44 ENCODE regions were downloaded from http://genome.ucsc.edu/ENCODE/ (UCSC human genome assembly hg16). A total of 95,353 human SNPs were also retrieved from the UCSC ENCODE browser.

Multiple sequence alignments for the human, chimpanzee, and baboon were carried out using the Multi-LAGAN program [22]. This program, which is based on a progressive alignment approach and a known phylogenetic tree for the sequences, can generate highly accurate multiple alignments of long genomic sequences within a short computational time [22].

A Perl script was written to extract the corresponding chimpanzee and baboon nucleotides for each human SNP from the alignments. To infer the mutation direction, we compared each human SNP with the chimpanzee allele using the baboon as an additional outgroup. We selected the SNPs of which one of their two alleles matched the chimpanzee allele. Next, the mutation direction was inferred by the maximum parsimony principle, as illustrated in Fig. 1.

Sequence and SNP data

The assembled human and chimpanzee genomic sequences were downloaded from ftp://ftp.ncbi.nih.gov/genomes/ (human build 35 version 1, released on August 26, 2004; chimpanzee build 1 version 1, released on November 23, 2004). To organize the data effectively, for all the genomic elements (e.g., repeats, SNPs, genes, CpG islands) analyzed in this study, we referred their locations to the assembled reference sequences.

To retrieve the repeat information in the genomic sequences, we downloaded two files: "masking_coordinates.gz," which listed the locations of the repetitive sequences on each contig, and "seq_contig.md," which listed the locations and orientations of the contigs across each chromosome. The coordinates of the repetitive sequences in each chromosome were obtained by combining the information in these two files.

The human SNPs in XML format were downloaded from ftp://ftp.ncbi.nih.gov/snp/ (NCBI dbSNP build 124, released on January 6, 2005). A total of 8,353,499 reference SNPs were available; among them, 98.6% (8,234,647 SNPs) were biallelic and uniquely mapped in the genome, and 59.1% (4,934,380 SNPs) were validated. The location of each SNP was determined by comparing the SNP position in the contig sequence and the contig location in the reference sequences. To make sure these annotations were accurate and matched correctly, we tested them using a method from our previous study [23]. To briefly summarize, (1) a set of SNPs was randomly chosen, (2) then, BLAST searches of the SNP flanking sequences were performed over the reference sequences, and (3) the search results were compared with the sequences extracted directly from the reference sequences based on the locations of SNPs.

Finally, we excluded the SNPs in the repetitive sequences by comparing SNP locations and repeat coordinates in the reference sequences. This gave 2,646,337 SNPs in the nonrepetitive sequences. We selected 2,632,415 SNPs whose flanking sequences were at least 100 nucleotides at each side and formatted them to have 100 nucleotides at each side.

Identification of SNPs in exonic, intronic, and intergenic regions

The gene annotations, including the start and end positions of genes and exons, were retrieved from the ENSEMBL database (ftp://ftp.ensembl.org/pub/, version 32.35e, released in July 2005). To identify SNPs in genic, exonic, intronic, and intergenic regions, stringent criteria were applied. First, we utilized only known genes and exons in the known genes. Second, since some genes have alternative transcripts, we selected only those introns that did not contain any exonic sequences. Third, intergenic regions were selected if there was no known or predicted gene overlapping with them. Fourth, we categorized the processed SNPs by comparing SNP locations with the coordinates of each categorized region. We identified 1,706,674, 496,160, 18,368, and 608,748 SNPs in intergenic, genic, exonic, and intronic regions, respectively.

Identification of SNPs in CpG islands

CpG islands were identified using the CpG island searcher program (CpGi130) available at http://cpgislands.usc.edu/ [24]. We used stringent search criteria for GC content ≥55%, Obs_{CpG}/Exp_{CpG} ≥0.65, and length ≥500 bp to screen CpG islands in the human genome sequences. The criteria above can effectively exclude the universal *Alu* repeats, which typically have a sequence length of 300 bp, GC content of 53%, and Obs_{CpG}/Exp_{CpG} ratio of 0.62 [25,26]. A total of 37,503 CpG islands were identified; among them, 4722 CpG islands were located or overlapped in the promoter regions (i.e., 2-kb upstream sequences from the start codon). Next, we identified the processed SNPs in these CpG islands by matching the coordinates of SNPs with those of the CpG islands in the reference sequences. We identified 46,060 SNPs in the CpG islands and 4737 SNPs in the promoter associated CpG islands.

Mapping human SNPs to the chimpanzee genome

Because the sequence similarity between the human and the chimpanzee is very high (e.g., ~99%) [27], we chose MegaBLAST [28] to map human SNPs and their flanking sequences to the chimpanzee genome. This is the standard SNP mapping method in the dbSNP database [29]. The MegaBLAST searches

were performed by the following command: *megablast –X 180 –r 10 –q −20 –R T –U T –F m –e 1e−80 –d database –i input_file –o output_file*.

We filtered the repetitive sequences in the input sequence files, used an E value of −80, and increased the X-drop-off value (X) to 180. This generated the best mapping results (data not shown).

A SNP was considered to "match" when the high-scoring segment pairs (HSPs) satisfied the following criteria: (1) identity was ≥95%; (2) the alignment length was between 196 and 206; (3) the SNP site was located approximately in the center of the alignment, i.e., within 96–106; (4) the corresponding chimpanzee allele matched one of the SNP alleles; (5) the immediately adjacent five nucleotides at the 5′ and 3′ sides of the SNP were identical between the human and the chimpanzee sequences; and (6) only one HSP satisfied these criteria.

Perl scripts were written to retrieve the human SNPs and their corresponding chimpanzee alleles from the satisfactory HSPs and to infer the mutation direction. Perl scripts and the processed data are available upon request.

Acknowledgments

We thank Leng Han, Miaojun Han, and Daekwan Seo for helpful discussions; Kenneth Kendler for valuable advice; Jill Opalesky for critically reading and improving the manuscript; and one anonymous reviewer for helpful comments. This project was supported by the Thomas F. and Kate Miller Jeffress Memorial Trust Fund.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2006.06.003.

References

[1] W.-H. Li, Molecular Evolution, Sinauer, Sunderland, MA, 1997.
[2] M. Krawczak, E.V. Ball, D.N. Cooper, Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes, Am. J. Hum. Genet. 63 (1998) 474–488.
[3] S. Subramanian, S. Kumar, Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes, Genome Res. 13 (2003) 838–844.
[4] A. Siepel, D. Haussler, Phylogenetic estimation of context-dependent substitution rates by maximum likelihood, Mol. Biol. Evol. 21 (2004) 468–488.
[5] F. Zhang, Z. Zhao, The influence of neighboring-nucleotide composition on single nucleotide polymorphisms (SNPs) in the mouse genome and its comparison with human SNPs, Genomics 84 (2004) 785–795.
[6] T. Gojobori, W.H. Li, D. Graur, Patterns of nucleotide substitution in pseudogenes and functional genes, J. Mol. Evol. 18 (1982) 360–369.
[7] W.H. Li, C.I. Wu, C.C. Luo, Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications, J. Mol. Evol. 21 (1984) 58–71.
[8] M.T. Webster, N.G.C. Smith, H. Ellegren, Compositional evolution of noncoding DNA in the human and chimpanzee genomes, Mol. Biol. Evol. 20 (2003) 278–286.
[9] Z. Zhao, et al., Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22, Proc. Natl. Acad. Sci. USA 97 (2000) 11354–11358.
[10] D. Casane, S. Boissinot, B.H. Chang, L.C. Shimmin, W. Li, Mutation pattern variation among regions of the primate genome, J. Mol. Evol. 45 (1997) 216–226.
[11] N. Yu, et al., Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1, Mol. Biol. Evol. 18 (2001) 214–222.
[12] G.M. Cooper, et al., Characterization of evolutionary rates and constraints in three mammalian genomes, Genome Res. 14 (2004) 539–548.
[13] R.A. Gibbs, et al., Genome sequence of the Brown Norway rat yields insights into mammalian evolution, Nature 428 (2004) 493–521.
[14] H. Watanabe, et al., DNA sequence and comparative analysis of chimpanzee chromosome 22, Nature 429 (2004) 382–388.
[15] The ENCODE Project Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project, Science 306 (2004) 636–640.
[16] J. Sved, A. Bird, The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model, Proc. Natl. Acad. Sci. USA 87 (1990) 4692–4696.
[17] G.V. Glazko, M. Nei, Estimation of divergence times for major lineages of primate species, Mol. Biol. Evol. 20 (2003) 424–434.
[18] M. Goodman, The genomic record of humankind's evolutionary roots, Am. J. Hum. Genet. 64 (1999) 31–39.
[19] N. Sueoka, On the genetic basis of variation and heterogeneity of DNA base composition, Proc. Natl. Acad. Sci. USA 48 (1962) 582–592.
[20] F. Antequera, A. Bird, Number of CpG islands and genes in human and mouse, Proc. Natl. Acad. Sci. USA 90 (1993) 11995–11999.
[21] Z. Zhao, F. Zhang, Sequence context analysis in the mouse genome: single nucleotide polymorphisms and CpG island sequences, Genomics 87 (2006) 68–74.
[22] M. Brudno, et al., LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA, Genome Res. 13 (2003) 721–731.
[23] Z. Zhao, F. Zhang, Sequence context analysis of 8.2 million single nucleotide polymorphisms in the human genome, Gene 366 (2006) 316–324.
[24] D. Takai, P.A. Jones, The CpG island searcher: a new WWW resource, In Silico Biol. 3 (2003) 235–240.
[25] L. Ponger, L. Duret, D. Mouchiroud, Determinants of CpG islands: expression in early embryo and isochore structure, Genome Res. 11 (2001) 1854–1860.
[26] D. Takai, P.A. Jones, Comprehensive analysis of CpG islands in human chromosomes 21 and 22, Proc. Natl. Acad. Sci. USA 99 (2002) 3740–3745.
[27] A.G. Clark, et al., Inferring nonneutral evolution from human–chimp–mouse orthologous gene trios, Science 302 (2003) 1960–1963.
[28] Z. Zhang, S. Schwartz, L. Wagner, W. Miller, A greedy algorithm for aligning DNA sequences, J. Comput. Biol. 7 (2000) 203–214.
[29] D.L. Wheeler, et al., Database resources of the National Center for Biotechnology Information, Nucleic Acids Res. 33 (2005) D39–D45.