Available online at www.sciencedirect.com

SCIENCE DIRECT®

Gene 366 (2006) 316 – 324

GENE
SECTION
EVOLUTIONARY GENOMICS

www.elsevier.com/locate/gene

Sequence context analysis of 8.2 million single nucleotide polymorphisms in the human genome

Zhongming Zhao [a,b,c,*], Fengkai Zhang [a]

[a] *Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA 23298, USA*
[b] *Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23284, USA*
[c] *Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China*

Received 4 May 2005; received in revised form 29 August 2005; accepted 30 August 2005
Available online 28 November 2005
Received by J.G. Zhang

Abstract

We analyzed *n*-mers (*n*=3–8) in the local environment of 8,249,446 human SNPs and compared their distribution with that in the genome reference sequences. The results revealed that the short sequences, which contained at least one CpG dinucleotide, occurred more frequently in the local SNP sequences than in the genome sequences. To exclude the hypermutability effect of the methylated CpG dinucleotides on the sequence context of SNPs, we examined the distribution patterns for each of the six categories of substitution. We observed the similar pattern (i.e., CpG-containing *n*-mers vs. non-CpG-containing *n*-mers) in SNP categories A/G, C/T and C/G but the opposite pattern in category A/T. We next identified 34,928 putative CpG islands in the human genome and located 133,591 SNPs within these islands. In the CpG islands, CpG SNPs were 3.92-fold less prevalent relative to the presence of CpG dinucleotides. Conversely, in the human genome, the frequency of CpG dinucleotides at the polymorphic sites was 6.09 times that in the genome reference sequences. These results support the previous views of mutational suppression at the CpG sites in the CpG islands and hypermutability of the methylated CpG dinucleotides that are prevalent in the non-CpG island sequences in the human genome. Our study represents a comprehensive investigation of the sequence context of SNPs in the human genome and in human CpG islands.

Keywords: SNPs; Short sequence; CpG dinucleotide; CpG islands; GC content

1. Introduction

Nucleotide substitution does not occur randomly in both the non-coding and coding regions and its rate depends on the sequence context (Siepel and Haussler, 2004). A detailed examination of the non-random sequence context where the genetic polymorphism occurs is important for understanding the mechanism of mutation (e.g., hotspot), protein–DNA interaction and genome sequence evolution. The early analyses of sequence variations in mammalian genomes and their influence of neighboring nucleotides have been limited to pseudogenes and functional regions; this is largely due to the limited data that were available at that time (Li et al., 1984; Blake et al., 1992; Hess et al., 1994; Krawczak et al., 1998). Genome-wide, or chromosome-wide, examination of the sequence compositions of SNPs was made possible with the recent identification of millions of SNPs in the human and mouse genomes (Zhao and Boerwinkle, 2002; Tomso and Bell, 2003; Zhang and Zhao, 2004; Fryxell and Moon, 2005). In our recent analyses of 2.6 million human SNPs and 0.4 million mouse SNPs, we revealed a large bias relative to the genome average at the two adjacent sites, as well as a small bias that could extend further from the polymorphic site (Zhao and Boerwinkle, 2002; Zhang and Zhao, 2004). Our results demonstrate a non-random sequence fashion of SNPs that are surviving in today's genomes. However, these studies are based on the observation of whole genomic regions. A detailed and comparative analysis of the sequence context patterns in specific genomic categories, such

Abbreviations: SNPs, single nucleotide polymorphisms; bp, base pairs; kb, kilobase pairs; nt, nucleotide; GC, guanine–cytosine; CGIs, CpG islands.

* Corresponding author. Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, P.O. Box 980126, Richmond, VA 23298-0126, USA. Tel.: +1 804 828 8129; fax: +1 804 828 1471.
E-mail address: zzhao@vcu.edu (Z. Zhao).

0378-1119/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2005.08.024

as CpG islands and exonic regions, would reveal more important biological information on this non-random sequence fashion of SNPs.

A CpG island is a cluster of CpG dinucleotides in the GC-rich regions, usually ~1 kb long (Lander et al., 2001). Approximately 80% of the CpG dinucleotides are methylated in the human genome; however, they often remain nonmethylated in CpG islands (Antequera and Bird, 1993; Antequera, 2003). An early analysis of the human genome sequence revealed approximately 50,000 CpG islands in the human genome, approximately 29,000 in the repeat-masked sequences (Lander et al., 2001). These estimates are under revision (e.g., Waterston et al., 2002). CpG islands were identified in the promoter regions of approximately 50% of the genes in vertebrate genomes (Antequera and Bird, 1993). Experiments have shown that the methylation of promoter CpG islands plays an important role in regulating gene expression and cell differentiation (Takai and Jones, 2002). Although many studies have been performed to identify CpG islands, to estimate the CpG mutation rates, and to examine the methylation modulation on disease genes (e.g., Antequera and Bird, 1993; Fryxell and Zuckerkandl, 2000; Ponger and Mouchiroud, 2002; Takai and Jones, 2002; Abdolmaleky et al., 2004), little is known about the distribution and pattern of genetic variation in CpG islands, or in the promoter-associated CpG islands (start CpG islands) in the human genome.

Tomso and Bell (2003) performed the first large-scale survey of the distribution of 6-bp fragments (6-mers) overlapping the human SNP sites and in their flanking sequences. They found that the 6-mers containing CpG dinucleotides were ~6.7-fold over-represented at the polymorphic sites versus in the flanking sequences. Their results supported the effect of the hypermutability of the methylated CpG. While their study revealed some interesting genetic information in the sequence context, they investigated only partial genome sequences (chromosomes 1, 2, 3, 19, and X) and limited SNP data (~1.9 million). Today the number of human SNPs available has increased to more than 10 million in the dbSNP database. Tomso and Bell's approach can be improved by comparing the distribution of short sequences surrounding the polymorphic sites with that in the genome sequences, rather than that in the SNP flanking sequences since the flanking sequences are near the polymorphic sites and thus biased (Zhao and Boerwinkle, 2002). Finally, there is some uncertainty of the accuracy on their identification of CpG islands (Tomso and Bell, 2003). They used a typical 401-bp sequence for most SNPs to scan the CpG islands using an algorithm by Takai and Jones (2002). However, Takai and Jones suggested a minimum length of 500 bp for searching a potential CpG island.

The relatively good quality of the human genome sequence and annotations of SNPs made it possible for us to perform a comprehensive analysis. Aiming to draw some general conclusions, we investigated and compared the distributions of several short fragments (3–8 nt) surrounding the 8.2 million polymorphic sites and in the whole human genome sequence. To distinguish the hypermutability effect of the methylated CpG dinucleotides on the sequence context of SNPs, we examined the distribution patterns of the short sequences for each type of SNP. We used genome contig sequences to identify the CpG islands, and then identified SNPs that were located in the putative CpG islands according to the mapping annotations. Next, we obtained and compared the nucleotide compositions in the CpG islands with those at the polymorphic sites or in the whole genome sequence.

2. Materials and methods

2.1. SNP and sequence data

The dbSNP database has the largest SNP data deposits in the public domain (Wheeler et al., 2005). It has related genome mapping information, such as chromosome positions and flanking sequences. We downloaded the human SNP data in XML format from ftp://ftp.ncbi.nih.gov/snp/ on November 13, 2004 (Build 123, released on November 3, 2004). A total of 10,079,771 human reference SNPs were available. The human genome sequence build 35 (version 1) was released on August 26, 2004. We downloaded the contig sequences in the human genome from ftp://ftp.ncbi.nih.gov/refseq/.

We wrote Perl scripts to parse the annotation information in the NCBI dbSNP XML files. We selected SNPs that were non-insertion/deletion, biallelic (i.e., only two nucleotides at a polymorphic site) and had unique hits to the genome reference sequences. There were 8,249,480 SNPs that satisfied the criteria above. For each SNP, we extracted the following annotation information from the XML data file: alleles (e.g., A/G), 5′ and 3′ flanking sequences, chromosome information, contig id and contig location. Next, we tested if these annotations were accurate and consistent with the genome reference sequences that we downloaded. We randomly chose 1000 SNPs covering all chromosomes and then performed blast searches of the SNP flanking sequences over the genome reference sequences. We wrote a Perl script to assess if the search results for each SNP matched the corresponding annotation information (e.g., allele, contig id, location, flanking sequences). The results showed that all SNPs we tested ($n=1000$) matched the annotations, indicating a high accuracy of the annotations. Finally, we reformatted these SNPs using the following two steps. First, if the length of each flanking sequence of a SNP was more than 200 bp, we trimmed it to be 200 bp; on the other hand, if the length was less than 200 bp, we extended the flanking sequence to be 200 bp using the corresponding contig sequence. Second, if the orientation of a SNP was annotated “minus”, we obtained and used its reverse complementary sequences (including SNP site and flanking sequences). This resulted in a total of 8,249,446 SNPs that were used in our analysis.

2.2. Identification of CpG islands

We used the algorithm in Takai and Jones (2002) for the identification of CpG islands. This algorithm was implemented in the CpG island searcher program (CpGi130), which we downloaded from http://cpgislands.usc.edu/ (Takai and Jones, 2003). We first evaluated the search criteria using the human contig sequences. Using these search criteria of GC content $\geq 55\%$, $\mathrm{Obs}_{\mathrm{CpG}}/\mathrm{Exp}_{\mathrm{CpG}} \geq 0.65$ and length ≥ 500 bp, we identified 34,928 CpG island sequences, among which 913 were in the human chromosome 22. These numbers are close to others in previous reports (Ewing and Green, 2000; Lander et al., 2001; Takai and Jones, 2003). Next, we categorized SNPs in the CpG islands by matching the position of a SNP with the location of a CpG island in a contig sequence. We identified 133,591 SNPs in the CpG islands.

2.3. Short sequence analysis in genome and SNP local sequences

A program package written in Java was used to compute the observed and expected frequencies for single nucleotides (i.e., A, C, G and T), dinucleotides (e.g., CG), and short fragments (*n*-mers) in a DNA sequence. Given a sequence (e.g., human genome reference sequences), the program first calculates the observed frequencies for single nucleotides A, C, G and T and the 16 possible dinucleotides. The program then calculates the expected frequency for each dinucleotide by $E_{ij}=f_if_j$ where f_i and f_j are the observed frequencies for nucleotide i and j in that sequence, respectively. We used this program package to obtain the observed and expected frequencies of dinucleotides in the human genome, SNP flanking sequences and CpG island sequences.

The frequency of short sequences (*n*-mers) surrounding the polymorphic sites and in the genome sequences were obtained using a sliding window approach (Tomso and Bell, 2003). For *n*-mer analysis, the local sequence of each SNP was defined by $2\times n-1$ nucleotides overlapping the polymorphic site. For example, in the 6-mer analysis, we extracted 11 nucleotides across the polymorphic site for each SNP, including the 5 nucleotides immediately adjacent to the polymorphic site at the 5′ side and at the 3′ side, respectively. Each SNP was treated equally, regardless of its location in the genome sequences. The two alleles of a SNP were equally separated because the allele frequency for each SNP is generally uncertain. Next, a sliding window of size *n* was applied to count the occurrence of *n*-mers in the SNP local sequences, the other part of the SNP flanking sequences and the genome reference sequences. The window slid one nucleotide each time. The frequency of each *n*-mer was calculated by the number of its counts divided by the total number of counts of *n*-mers. We also obtained the frequencies of *n*-mers for each type of SNP using the approach described above. The programs and processed SNP data are available upon request.

3. Results

3.4. Distribution of the short sequences surrounding SNP sites and in the genome

We analyzed 8,249,446 SNPs which are non-insertion/deletion, biallelic, have a unique hit to the human genome reference sequences, and have the same orientation as the reference sequences. We used the annotation in dbSNP to represent SNP types, for example A/G represents the nucleotide changes between A and G. The proportions of six types of SNP (A/G, C/T, A/C, G/T, A/T and C/G) and the ratio (1.91) of transition over transversion were essentially the same as that observed in our previous study (Zhang and Zhao, 2004), indicating consistency of the data.

We analyzed the short fragments (e.g., *n*-mers, $n=3$–8) in the local sequences of 8,249,446 SNPs and the human genome reference sequences. Fig. 1 shows the distribution of all possible 256 (4^4) tetramers plotted by its frequency observed in the SNP local sequence versus in the genome sequences. The distribution patterns for other *n*-mers are similar (e.g., the distribution of 6-mers is shown in Supplementary Fig. S1). For each *n*-mer analysis, we separated the *n*-mers containing at least one CpG dinucleotide (labeled in red in Fig. 1) from those without any CpG dinucleotide (labeled in blue in Fig. 1). We named these two categories CpG group and non-CpG group, respectively. The results are summarized in the following two points.

First, we noted a trend for *n*-mers that were frequently observed in the genome sequences were also frequently observed in the SNP local sequences. This was more striking when we looked within each of the two groups (Fig. 1). Given that each fragment in the SNP local sequences occurs with the same frequency as in the genome sequences, it should be plotted along the diagonal line — the slope of the line should be 1 and the intercept should be 0. Using a simple linear regression model, the distribution of all tetramers were represented by $f_{\mathrm{SNP}}=0.0019+0.53f_{\mathrm{genome}}$ ($R^2=0.81$) where f_{SNP} and f_{genome} denote the frequency of a tetramer in the SNP local sequences and in the genome sequences, respectively. The slope was much less than 1, while the intercept was about half of the average frequency value of a tetramer ($1/256=0.0039$), indicating that the tetramers with high frequency in the genome sequences had relatively less presence in the SNP local sequences.

Second, the distributions for CpG and non-CpG groups were notably distinct (Fig. 1). In the tetramer analysis, the regression line was represented by $f_{\mathrm{SNP}}=0.0008+2.61f_{\mathrm{genome}}$ for the CpG group and $f_{\mathrm{SNP}}=0.0014+0.60f_{\mathrm{genome}}$ for the non-

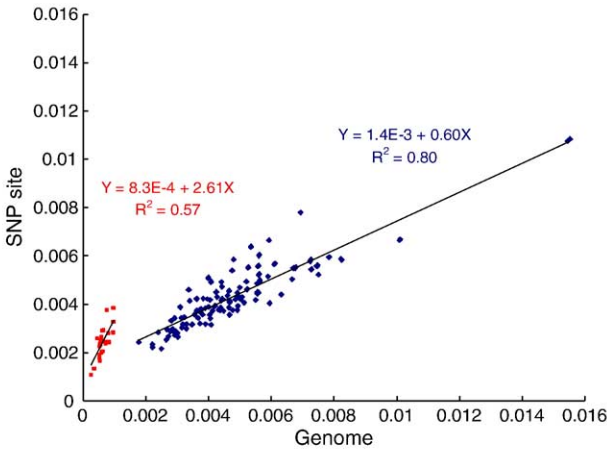

Fig. 1. The distribution of possible 256 tetramers in the local sequences of SNPs versus in the human genome sequences. The tetramers containing at least one CpG dinucleotide are labeled in red and the remaining tetramers are labeled in blue.

Table 1
Linear regression for *n*-mer fragments

Size (*n*)	CpG group			Non-CpG group		
	Intercept (α)	Slope (β±95% CI)[a]	R^2	Intercept (α)	Slope (β±95% CI)[a]	R^2
3	−2.7E−3	6.00 (±7.10)	0.42	7.8E−3	0.48 (±0.10)	0.63**
4	8.3E−4	2.61 (±0.68)	0.57**	1.4E−3	0.60 (±0.04)	0.80**
5	1.6E−4	2.35 (±0.17)	0.75**	2.5E−4	0.70 (±0.02)	0.85**
6	3.6E−5	2.09 (±0.05)	0.82**	3.7E−5	0.79 (±0.01)	0.86**
7	8.9E−6	1.83 (±0.02)	0.87**	3.1E−6	0.86 (±0.01)	0.84**
8	2.3E−6	1.61 (±0.01)	0.89**	−5.5E−7	0.92 (±0.00)	0.81**

The CpG group includes the fragments containing at least one CpG dinucleotide (labeled in red in Fig. 1) and non-CpG group includes the fragments without any CpG dinucleotide (labeled in blue in Fig. 1). The linear regression for these two groups is expressed by $f_{\text{SNP}}=\alpha+\beta f_{\text{genome}}$ where f_{SNP} and f_{genome} denote the frequency of each *n*-mer observed in the SNP local sequences and in the genome sequences.

[a] CI: confidence interval.

** $p<0.0001$.

CpG group. The linear regression for these two groups in other *n*-mer analysis is summarized in Table 1. In each *n*-mer analysis, the regression line was different between the CpG group and non-CpG group. In the CpG group, when n was smaller, the slope was larger. For example, it was 1.61 for 8-mers but 6.00 for 3-mers (Table 1). This is consistent with our previous report that the two immediate adjacent sites of SNPs had strong biases relative to the human genome average (Zhao and Boerwinkle, 2002). In the non-CpG group, the above trend is opposite. Note that the slope values were less than 1 for all the *n*-mers, indicating that the short sequences in the non-CpG group are in general less prevalent in the local sequences of SNPs.

3.2. Distribution of *n*-mers by SNP types

In our *n*-mer analysis, we used the genome reference sequences for background information. Correspondingly, for the SNPs with an orientation of minus in the dbSNP database, we used the reverse complementary sequences. That is, for all SNPs we analyzed, their alleles and flanking sequences had the same orientation as the genome reference sequences. We first obtained the distribution patterns for the six types of SNP (A/G, C/T, A/C, G/T, A/T and C/G). For each *n*-mer analysis, the distribution of the SNP types A/G and C/T, as well as, A/C and G/T were essentially the same because of the complementary strand symmetry in DNA sequences. Therefore, we combined A/G and C/T as one type, and A/C and G/T as one type, and presented the results in a total of four types (i.e., A/G|C/T, A/C|G/T, A/T and C/G). Table 2 shows the linear regression for the 6-mers and 4-mers in the CpG group and the non-CpG group. In addition, the distribution of 6-mers for four SNP types is displayed in Fig. 2. In the CpG group, the slope value was much greater than 1 for SNP types A/G, C/T and C/G, but much less than 1 for SNP type A/T. In contrast, in the non-CpG group, the slope value was much greater than 1 for SNP type A/T, but less than 1 for SNP types A/G, C/T and C/G.

3.3. CpG islands in the human genome

A total of 34,928 CpG islands were identified in the human genome reference sequences by the CpGi130 program with the criteria of length ≥500 bp, G+C content ≥55% and $\text{Obs}_{\text{CpG}}/\text{Exp}_{\text{CpG}} \geq 0.65$. The average length of the CpG island sequences was 1094 bp. The average GC content in the CpG island sequences was 62.02%, this compares to the 40.60% in the non-CpG island sequences and the 40.91% in the total genome sequences. The average $\text{Obs}_{\text{CpG}}/\text{Exp}_{\text{CpG}}$ ratio was 0.72 in the CpG island sequences, three times that (0.24) in the human genome sequences.

Table 3 shows the proportion of the CpG island sequences, measured by single nucleotides, CpG dinucleotides and CpG SNPs, on each chromosome and in the whole genome. The proportion of the CpG island sequences varied among

Table 2
Linear regression for SNP types

SNP type	CpG group			Non-CpG group		
	Intercept (α)	Slope (β±95% CI)[a]	R^2	Intercept (α)	Slope (β±95% CI)[a]	R^2
6-mers						
A/G|C/T	3.8E−5	2.55 (±0.09)	0.73 **	4.6E−5	0.77 (±0.03)	0.40 **
A/C|G/T	3.1E−5	1.06 (±0.07)	0.43 **	−6.3E−5	1.18 (±0.07)	0.30 **
A/T	1.2E−5	0.21 (±0.03)	0.19 **	−5.3E−4	2.66 (±0.08)	0.57 **
C/G	4.9E−5	2.36 (±0.11)	0.62 **	1.5E−4	0.44 (±0.04)	0.16 **
4-mers						
A/G|C/T	9.5E−4	3.17 (±1.03)	0.46 **	7.2E−4	0.83 (±0.18)	0.30 **
A/C|G/T	6.1E−4	1.27 (±0.72)	0.22 *	1.3E−4	1.03 (±0.30)	0.20 **
A/T	2.4E−4	0.08 (±0.24)	0.01	−5.9E−3	2.29 (±0.24)	0.63 **
C/G	7.2E−4	3.62 (±0.70)	0.71 **	3.0E−3	0.33 (±0.20)	0.05 *

The CpG group includes the fragments containing at least one CpG dinucleotide (labeled in red in Fig. 2) and non-CpG group includes the fragments without any CpG dinucleotide (labeled in blue in Fig. 2). The linear regression for these two groups is expressed by $f_{\text{SNP}}=\alpha+\beta f_{\text{genome}}$ where f_{SNP} and f_{genome} denote the frequency of each *n*-mer observed in the SNP local sequences and in the genome sequences.

[a] CI: confidence interval.

** $p<0.0001$.

* $p=0.001$.

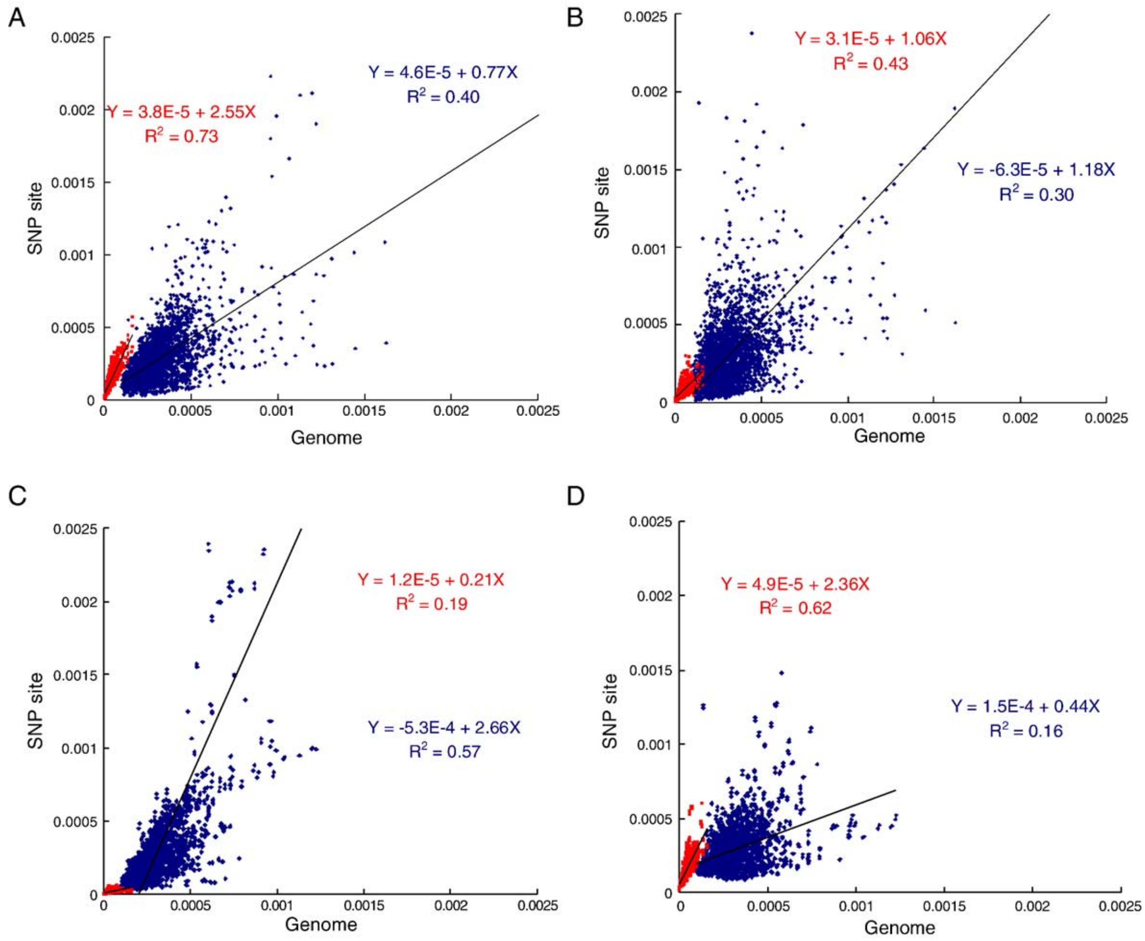

Fig. 2. The distribution of 6-mers by SNP types. The 6-mers containing at least one CpG dinucleotide are labeled in red and the remaining 6-mers are labeled in blue. For easy comparison, we made the scale in each figure the same. This made a few 6-mers invisible because of high frequency. The full distribution is displayed in Supplementary Fig. S2.

chromosomes, ranging from 0.93% on chromosome 4 to 5.48% on chromosome 19. On average, CpG island sequences accounted for only 1.45% of the human genome. However, contrary to the small proportion of the CpG island sequences in the human genome, CpG dinucleotides in the CpG islands accounted for 10.35% of all CpG dinucleotides in the human genome. In a simple comparison, CpG dinucleotides in the CpG islands are 7.16-fold over-abundant relative to the whole human genome. This is shown in particular, on chromosome 19, which has the highest GC content and gene density (Venter et al., 2001; Zhao et al., 2003). Chromosome 19 stood out as being different from the other chromosomes with respect to the proportion of nucleotides and CpG dinucleotides in the CpG islands. About 5.5% of sequences on chromosome 19 were located within CpG islands. The CpG dinucleotides accounted for 20.07%, nearly 2.5 times that on chromosome Y (7.09%) or chromosome 3 (7.98%). Overall, in the human genome, there is a strong correlation of proportion of CpG island sequences or proportion of CpG dinucleotides with the GC content on a chromosome. That is, a chromosome with higher GC content has more nucleotides and CpG dinucleotides in the CpG island sequences. Using a simple linear regression model, this correlation is represented as follows: $Y=1.43+0.33\times(\mathrm{GC}-40.91)$ ($R^2=0.82$, $p<0.0001$) between the proportion of CpG island sequences on a chromosome (Y) and its GC content (GC, %) difference from the genome average, and $Y=9.80+0.91\times(\mathrm{GC}-40.91)$ ($R^2=0.82$, $p<0.0001$) between the proportion of CpG dinucleotides in the CpG islands on a chromosome (Y) and its GC content (GC, %) difference from the genome average.

3.4. CpG SNPs in the CpG islands

There were 2,236,053 SNPs that occurred at the CpG sites. We examined the distribution of CpG SNPs in the islands. As shown in Table 3, approximately 2.64% of all CpG SNPs in the human genome occurred in the CpG islands. However, this proportion is much lower than that of the CpG dinucleotides. Assuming nucleotide changes occur randomly in whole genome sequence, CpG SNPs in the CpG islands were underrepresented approximately 3.92-fold than in the human genome. Similarly, there is a correlation between the proportion of CpG SNPs in the islands on a chromosome (Y) and its GC content (GC, %) in human chromosomes; this is represented by $Y=2.49+0.37\times(\mathrm{GC}-40.91)$ ($R^2=0.69$, $p<0.0001$). This relationship indicates that a chromosome with higher GC content has more

Table 3
Proportions of nucleotides (nt), CpG dinucleotides (CpGs) and CpG SNPs from CpG islands in the human genome

Chromosome	GC content (%)	Nucleotides (%)	CpG dinucleotides (%)	CpG SNPs (%)[a]	CpGs/nt
1	41.75	1.62	11.32	2.71	6.99
2	40.25	1.15	9.02	2.20	7.85
3	39.69	0.95	7.98	1.83	8.39
4	38.24	0.93	8.23	2.05	8.89
5	39.52	1.10	9.00	2.29	8.20
6	39.62	1.20	9.34	2.65	7.81
7	40.75	1.46	10.14	2.86	6.93
8	40.17	1.15	8.97	2.20	7.77
9	41.35	1.54	10.72	2.81	6.98
10	41.58	1.39	9.57	2.52	6.91
11	41.57	1.44	10.52	2.44	7.31
12	40.79	1.38	9.69	2.34	7.00
13	38.53	1.00	8.31	2.33	8.30
14	40.89	1.39	10.27	2.45	7.37
15	42.24	1.59	10.65	2.34	6.68
16	44.79	2.46	12.84	3.69	5.22
17	45.58	3.06	14.73	4.35	4.81
18	39.79	1.15	9.00	2.36	7.85
19	48.34	5.48	20.07	7.18	3.66
20	44.13	1.92	11.58	3.11	6.04
21	40.88	1.41	9.39	2.79	6.68
22	47.92	2.90	13.24	3.82	4.56
X	39.49	0.99	8.11	1.53	8.19
Y	39.85	0.95	7.09	0.77	7.44
Genome	40.91	1.45	10.35	2.64	7.16

The proportion was calculated by the observed numbers in the CpG islands over the numbers in the corresponding chromosome or genome sequences.

[a] CpG SNPs include six possible nucleotide changes at CpG dinucleotide.

CpG SNPs that occur in the CpG islands, which is in agreement with the higher proportion of CpG dinucleotides (Table 3).

The proportions of the six possible types of CpG SNPs, i.e., [A/C]G, [C/G]G, [C/T]G, C[A/G], C[C/G] and C[G/T], were different in the CpG island sequences versus in other parts of the genome sequences. In the CpG islands, the proportion of substitution of [C/T]G and C[A/G] was 60.15% and 60.45%, respectively. This is significantly ($p \approx 0$, χ^2 test) lower than that of 80.76% and 80.65%, respectively, was found in the non-CpG island sequences. Conversely, the proportion of the other four types ([A/C]G, [C/G]G, C[C/G] and C[G/T]) in the CpG islands was significantly ($p \approx 0$, χ^2 test) higher than that in the non-CpG island sequences (Supplementary Table S1).

3.5. Dinucleotides in the CpG islands and human genome

Table 4 shows the distribution of the 16 possible dinucleotides in the CpG island sequences and the whole human genome. The proportion of the 16 dinucleotides at the polymorphic sites (SNP sites) was obtained by including two immediate adjacent nucleotides. The proportion of the 16 dinucleotides varied among three sequence categories: (1) CpG island sequences, (2) human genome sequences and (3) SNP sites. However, the extent of the difference among the 16 dinucleotides at the SNP sites was notably less than that in the CpG island sequences or the genome sequences. We next compared the frequency ratio of each dinucleotide observed at the SNP sites versus the sequences in the CpG islands or the whole genome. In the CpG islands, the ratio of CpG dinucleotide was 0.97. Interestingly, the ratios for the other three dinucleotides containing only G and/or C (i.e., GC, CC and GG) were ~0.5, the lowest among the 16 dinucleotides.

Table 4
Dinucleotides in the CpG islands and the human genome

Dinucleotide	CpG islands (CGIs)				Genome			
	Expected (%)[a]	Observed (%)	SNP sites (%)[b]	SNP/CGIs[c]	Expected (%)[d]	Observed (%)	SNP sites (%)	SNP/genome[e]
AA	3.62	4.58	6.80	1.49	8.72	9.77	6.79	0.69
AC	5.89	4.60	6.78	1.47	6.04	5.03	6.62	1.31
AG	5.90	6.99	6.26	0.89	6.04	6.99	6.25	0.89
AT	3.62	2.86	8.05	2.82	8.73	7.72	7.65	0.99
CA	5.89	6.19	6.96	1.12	6.04	7.25	6.52	0.90
CC	9.58	10.71	5.27	0.49	4.18	5.21	5.48	1.05
CG	9.59	7.07	6.84	0.97	4.18	0.99	6.01	6.09
CT	5.89	6.99	6.27	0.90	6.05	7.00	6.24	0.89
GA	5.90	5.98	5.09	0.85	6.04	5.93	5.70	0.96
GC	9.59	9.69	4.92	0.51	4.18	4.27	5.34	1.25
GG	9.61	10.72	5.28	0.49	4.19	5.21	5.45	1.05
GT	5.89	4.61	6.80	1.47	6.05	5.05	6.58	1.30
TA	3.62	2.29	5.81	2.54	8.73	6.56	6.31	0.96
TC	5.89	5.96	5.09	0.85	6.05	5.94	5.73	0.96
TG	5.89	6.21	6.98	1.12	6.05	7.27	6.53	0.90
TT	3.62	4.56	6.81	1.49	8.74	9.80	6.78	0.69

[a] The expected frequency of a dinucleotide f_{ij} was calculated by $f_i \times f_j$, where f_i and f_j were the observed frequencies of nucleotides i and j in the CpG island sequences.

[b] For each SNP, the two immediate adjacent sites were included to calculate the frequency of dinucleotides.

[c] Ratio of the proportion of a dinucleotide observed at SNP sites versus in the CpG island sequences.

[d] The expected frequency of a dinucleotide f_{ij} was calculated by $f_i \times f_j$, where f_i and f_j were the observed frequencies of nucleotides i and j in the human genome sequences.

[e] Ratio of the proportion of a dinucleotide observed at SNP sites versus in the human genome sequences.

This is probably due to the high proportion of these dinucleotides observed in the CpG island sequences (Table 4). This may also explain the high ratios of dinucleotide AT (2.82) and TA (2.54) because AT and TA were rarely observed in the CpG island sequences (Table 4). In the genome, the ratio of CpG dinucleotide at the polymorphic sites versus in the genome sequences was sharply higher (6.09). This reflects the very low presence of CpG dinucleotides in the human genome, i.e., the ratio of the observed over the expected frequency of CG dinucleotides in the human genome was only 0.24 (0.99/4.18).

4. Discussion

4.1. Summary

In this study, we investigated the short sequences in the local environment of 8,249,446 SNPs and compared them to those in the human genome sequences. Two distinct groups were observed among the *n*-mers ($n = 3–8$) we examined. First, the short sequences containing at least one CpG dinucleotide (CpG group) occurred more frequently in the local sequences surrounding the polymorphic sites than in the genome sequences, or in other parts of the SNP flanking sequences. Second, the short sequence without any CpG dinucleotide (non-CpG group) occurred more frequently in non-polymorphic sequences than in the SNP local sequences. Further analysis of the six types of SNP revealed that the distribution for the SNP types A/G and C/T, as well as, A/C and G/T were essentially the same, therefore they were combined. However, the distribution patterns among A/G|C/T, A/C|G/T, A/T and C/G varied greatly, suggesting the preference of the specific local sequence environment by each type of SNP. We next investigated the nucleotide composition and distribution of SNPs in the CpG islands versus the human genome. The results revealed that CpG dinucleotides were ~6.09-fold over-represented at the polymorphic sites than in the genome sequences. In the CpG islands, CpG dinucleotides had nearly the same representation at the polymorphic sites with that in the CpG island sequences. Furthermore, SNPs that occurred at the CpG dinucleotides were 3.92-fold less prevalent than expected in the CpG islands. While our results, in general, support previous analyses by Tomso and Bell (2003), this study represents a more comprehensive and accurate examination of sequence context in the local environment of SNPs and CpG islands.

4.2. Mutational mechanisms and natural selection

These results have implications on the mutational mechanisms and natural selection. The sequence context analysis in the local environment of SNPs or each type of SNP indicates that substitution events are sequence-dependent. Overall, a short sequence containing CpGs is more likely to be represented in the local sequence of SNPs than in the non-polymorphic sequences. However, the distribution of short sequences for SNP types C/G and A/T showed the opposite pattern. This supports the previous view that an AT-rich sequence may have more likelihood of nucleotide changes between A and T while a GC-rich sequence may have more likelihood of nucleotide changes between C and G (Li, 1997). Since the CpG effect does not involve these two types of SNP, their distribution patterns may be useful for further study of mutational mechanisms. Moreover, the examination of the distribution of dinucleotides and SNPs in the CpG islands and the other part of the genome supports the commonly known CpG effects in the human genome. In the human genome, ~80% of the CpG dinucleotides are methylated; in contrast, they often remain nonmethylated in CpG islands (Antequera and Bird, 1993; Antequera, 2003). The mutation rate at methylated CpG dinucleotide sites was estimated to be ~10- to 50-fold higher than other transitional changes (Sved and Bird, 1990; Fryxell and Moon, 2005). Therefore, the CpG dinucleotides are more likely to be observed at the polymorphic sites in the non-CpG island sequences. This effect finally leads to the great deficiency of CpG dinucleotides in the human genome. The signature of the CpG effects was further detected when we compared the proportion of the six types of CpG SNPs (Supplementary Table S1). The proportion of [C/T]G or C[A/G] in the non-CpG island sequences was significantly higher than that in the CpG island sequences.

In the CpG islands, the suppression of polymorphisms at the CpG dinucleotides might be due to (purifying) selection. This is because (1) the CpG effect does not play a role in the CpG islands and (2) the CpG islands generally overlap the promoter regions of genes, which are generally under selection pressure (Larsen et al., 1992; Antequera and Bird, 1993; Ponger et al., 2001). A mutation occurring in the promoter region of a gene may influence its function and, subsequently, it is likely eliminated from the population. Recently, an examination of the minor allele frequency (MAF) of SNPs in the CpG islands and non-CpG island noncoding regions in 65 genes indicated the possible weak purifying selection (Freudenberg-Hua et al., 2003). Further comparative analysis of the MAF of SNPs in the start CpG islands, other CpG islands and non-CpG island sequences may provide more evidence for this.

4.3. Verified SNPs versus all SNPs

We analyzed 8,249,446 SNPs that were non-insertion/deletion, biallelic and had a unique hit to the genome reference sequences. Approximately 58.4% (4,815,863) of them was validated according to the annotation information in the dbSNP database. We wondered if the same sequence context information could be obtained by only using verified SNPs. Our analysis showed essentially that they yield the same results, albeit the slope values tended to be slightly smaller. For example, in the CpG group, the slope values were 2.03 for the 6-mers and 2.54 for the 4-mers by using verified SNPs only; they were 2.09 and 2.61, respectively, by using all SNPs (Table 1).

4.4. SNP local versus flanking sequences

To investigate the unique distribution patterns of short sequences surrounding the polymorphic sites we compared them with those occurred in the genome reference sequences. In this approach, the background information – the distribution of

Table 5
Linear regression for 6-mers in SNP local sequences versus flanking sequences

Length (bp)[a]	CpG group			Non-CpG group		
	Intercept (α)	Slope (β±95% CI[b])	R^2	Intercept (α)	Slope (β±95% CI b)	R^2
20	4.6E−5	1.68 (±0.05)	0.78 **	7.7E−5	0.67 (±0.01)	0.91 **
100	4.5E−5	1.70 (±0.05)	0.78 **	4.3E−5	0.77 (±0.01)	0.90 **
200	4.3E−5	1.77 (±0.05)	0.79 **	4.0E−5	0.78 (±0.01)	0.89 **
Genome	3.6E−5	2.09 (±0.05)	0.82 **	3.7E−5	0.79 (±0.01)	0.86 **

[a] The length of the flanking sequences at each side for each SNP. The flanking sequence at each side of a SNP was further separated as two parts in the analysis: SNP local sequences and the remaining SNP flanking sequences.
[b] CI: confidence interval.
** $p<0.0001$.

n-mers in the human genome sequences – included the "biased" SNP flanking sequences. Since SNP sites accounted for only ~0.3% of the genome sequences, the bias was minimal. Conversely, the effect will be larger when we use the SNP flanking sequences for the background information because the neighboring-nucleotide biases could extend as far as 200 bp in the SNP flanking sequences (Zhao and Boerwinkle, 2002). To reveal the extent of the effect, we obtained the frequencies of *n*-mers in the SNP flanking sequences, which were formatted to have the length of 20, 100, and 200 bp at each side, respectively. Only the flanking sequences excluding the SNP local sequences were used (see Materials and methods). The details of the distribution of 6-mers in the SNP local sequences versus the flanking sequences are summarized in Table 5. The results show that, although there was no remarkable difference in the non-CpG group, the slope values in the CpG group decreased when the length of the flanking sequence decreased. This indicates that the difference of the distribution in the SNP local sequences versus flanking sequences became smaller when the shorter flanking sequences were used. Therefore, it suggests that the genome sequences are the better option for background information. Interestingly, in Tomso and Bell's study using flanking sequences (Tomso and Bell, 2003), the slope value for the CpG group was 2.08, which is the same as what we observed using genome sequences, but higher than when using the flanking sequences. This difference is likely caused by the small data set in Tomso and Bell's study.

4.5. Identification of CpG islands

The distribution of short sequences in the CpG islands relies on the accuracy of the identification of CpG islands. There are mainly two algorithms for CpG island identification. One is the original criteria proposed by Gardiner-Garden and Frommer (1987), and the other, which was modified from the original criteria, was proposed by Takai and Jones (2002). The second algorithm was implemented in the CpG island searcher program (CpGi130). This program has been applied to the NCBI (Wang and Leung, 2004). In our study, we also used the CpGi130 program to search for CpG islands in the human genome. By using stringent criteria, this algorithm can identify the CpG islands that are more likely associated with the 5′ regions of genes and can effectively remove the most Alu repeats, which have relatively high GC content (53%) and ratio of $\mathrm{Obs}_{\mathrm{CpG}}/\mathrm{Exp}_{\mathrm{CpG}}$ (0.62) (Ponger et al., 2001; Takai and Jones, 2002). The number of CpG islands we identified in the human genome (34,928) and in chromosome 22 (913) were close to those in previous reports (Ewing and Green, 2000; Lander et al., 2001; Takai and Jones, 2003).

Acknowledgements

We are grateful to the people who collected, annotated and managed those data for public access, especially the NCBI's dbSNP team. This project was supported by Thomas F. and Kate Miller Jeffress Memorial Trust Fund.

Appendix A. Supplementary materials

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2005.08.024.

References

Abdolmaleky, H.M., et al., 2004. Methylomics in psychiatry: modulation of gene–environment interactions may be through DNA methylation. Am. J. Med. Genet., Part B Neuropsychiatr. Genet. 127, 51–59.

Antequera, F., Bird, A., 1993. Number of CpG islands and genes in human and mouse. Proc. Natl. Acad. Sci. U. S. A. 90, 11995–11999.

Antequera, F., 2003. Structure, function and evolution of CpG island promoters. Cell. Mol. Life Sci. 60, 1647–1658.

Blake, R.D., Hess, S.T., Nicholson-Tuell, J., 1992. The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. J. Mol. Evol. 34, 189–200.

Ewing, B., Green, P., 2000. Analysis of expressed sequence tags indicates 35,000 human genes. Nat. Genet. 25, 232–234.

Freudenberg-Hua, Y., et al., 2003. Single nucleotide variation analysis in 65 candidate genes for CNS disorders in a representative sample of the European population. Genome Res. 13, 2271–2276.

Fryxell, K.J., Zuckerkandl, E., 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. Mol. Biol. Evol. 17, 1371–1383.

Fryxell, K.J., Moon, W.J., 2005. CpG mutation rates in the human genome are highly dependent on local GC content. Mol. Biol. Evol. 22, 650–658.

Gardiner-Garden, M., Frommer, M., 1987. CpG islands in vertebrate genomes. J. Mol. Biol. 196, 261–282.

Hess, S.T., Blake, J.D., Blake, R.D., 1994. Wide variations in neighbor-dependent substitution rates. J. Mol. Biol. 236, 1022–1033.

Krawczak, M., Ball, E.V., Cooper, D.N., 1998. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am. J. Hum. Genet. 63, 474–488.

Lander, E.S., et al., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Larsen, F., Gundersen, G., Lopez, R., Prydz, H., 1992. CpG islands as gene markers in the human genome. Genomics 13, 1095–1107.

Li, W.H., Wu, C.I., Luo, C.C., 1984. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J. Mol. Evol. 21, 58–71.

Li, W.-H., 1997. Molecular Evolution. Sinauer Associates, Sunderland, MA.

Ponger, L., Mouchiroud, D., 2002. CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. Bioinformatics 18, 631–633.

Ponger, L., Duret, L., Mouchiroud, D., 2001. Determinants of CpG islands: expression in early embryo and isochore structure. Genome Res. 11, 1854–1860.

Siepel, A., Haussler, D., 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol. Biol. Evol. 21, 468–488.

Sved, J., Bird, A., 1990. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. Proc. Natl. Acad. Sci. U. S. A. 87, 4692–4696.

Takai, D., Jones, P.A., 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc. Natl. Acad. Sci. U. S. A. 99, 3740–3745.

Takai, D., Jones, P.A., 2003. The CpG island searcher: a new WWW resource. In Silico Biol. 3, 235–240.

Tomso, D.J., Bell, D.A., 2003. Sequence context at human single nucleotide polymorphisms: overrepresentation of CpG dinucleotide at polymorphic sites and suppression of variation in CpG islands. J. Mol. Biol. 327, 303–308.

Venter, J.C., et al., 2001. The sequence of the human genome. Science 291, 1304–1351.

Wang, Y., Leung, F.C., 2004. An evaluation of new criteria for CpG islands in the human genome as gene markers. Bioinformatics 20, 1170–1177.

Waterston, R.H., et al., 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562.

Wheeler, D.L., et al., 2005. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 33, D39–D45.

Zhang, F., Zhao, Z., 2004. The influence of neighboring-nucleotide composition on single nucleotide polymorphisms (SNPs) in the mouse genome and its comparison with human SNPs. Genomics 84, 785–795.

Zhao, Z., Boerwinkle, E., 2002. Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. Genome Res. 12, 1679–1686.

Zhao, Z., Fu, Y.-X., Hewett-Emmett, D., Boerwinkle, E., 2003. Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. Gene 312, 207–213.