Available online at www.sciencedirect.com

SCIENCE DIRECT

Genomics 84 (2004) 785–795

GENOMICS

www.elsevier.com/locate/ygeno

The influence of neighboring-nucleotide composition on single nucleotide polymorphisms (SNPs) in the mouse genome and its comparison with human SNPs

Fengkai Zhang[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 20 May 2004; accepted 28 June 2004
Available online 8 September 2004

Abstract

We analyzed the neighboring-nucleotide composition of 433,192 biallelic substitutions, representing the largest public collection of SNPs across the mouse genome. Large neighboring-nucleotide biases relative to the genome- or chromosome-specific average were observed at the immediate adjacent sites and small biases extended farther from the substitution site. For all substitutions, the biases for A, C, G, and T were 0.21, 2.63, 0.71, and −3.55%, respectively, on the immediate adjacent 5′ site and −3.67, 0.75, 2.69, and 0.23%, respectively, on the immediate adjacent 3′ side. Further examination of the six categories of substitution revealed that the neighboring-nucleotide patterns for transitions were strongly influenced by the hypermutability of dinucleotide CpG and the neighboring effects on transversions were complex. Probability of a transversion increased with increasing A + T content of the two immediate adjacent sites, which was similarly observed in the human and *Arabidopsis* genomes. Overall, the bias patterns for the neighboring nucleotides in the mouse and human genomes were essentially the same; however, the extent of the biases was notably less in mice. Our results provide the first comprehensive view of the neighboring-nucleotide effects in the mouse genome and are important for understanding the mutational mechanisms and sequence evolution in the mammalian genomes.

Keywords: Single nucleotide polymorphisms; Neighboring-nucleotide effects; Mouse; Human; CpG effect

Single nucleotide polymorphisms (SNPs) are the most abundant genetic variants in the genomes. As of April 2004, 11.8 million SNPs in 24 species, including 9.1 million SNPs in the human genome, have been recorded and annotated in the dbSNP database of the National Center for Biotechnology Information (NCBI). The nucleotide variations observed in today's genomes result from the combinatory evolutionary processes such as error-prone DNA replication and repair, recombination, genetic drift, and selection on the naturally occurring mutations [1,2]. A comprehensive understanding of nucleotide variation in the human and other model organisms is essential for unraveling the genetic basis of phenotypes, functions, and diseases [3]. The substitution patterns at polymorphic sites and the sequence context in a local environment of SNPs reflect the mutability of the sequence and, therefore, are important for studying the mechanisms of mutation and the evolution of genome sequences. Initially, the analyses of sequence variations and the influence of the neighboring nucleotides have been limited to the pseudogenes and functional regions, because the former can provide a neutral estimate in the genome and the latter have benefited from the abundant but opportunistic collection of data in the coding regions [4–6]. As a result of the rapid advancement of sequencing technology during the past decade, the number

* Corresponding author. Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, PO 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 © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2004.06.015

of SNPs and the sizes of genome sequences have increased exponentially. The available genome-wide SNPs and their sequence data provide us an unprecedented opportunity to examine the neighboring-nucleotide effects on the single nucleotide polymorphisms in mammals. Our recent study of 2.6 million SNPs across the human genome revealed a large bias relative to the genome average at the two adjacent sites and a small bias that could extend farther [7]. These results not only support the patterns observed based on the substitutions from a large number of coding regions, but also for the first time reveal that a small bias could extend to the other neighboring sites of SNPs [6]. The influence of the local DNA-sequence environment on the SNPs was further examined and an overrepresentation of the small fragments containing dinucleotide CpG was uncovered at the polymorphic sites in the human genome [8].

With the near completion of the mouse and rat genomes, SNP discovery in the rodent genomes is of great interest not only because of their genetic analysis but also because of their comparative nature with humans [9,10]. Currently, there are approximately half a million SNPs deposited in the dbSNP database, representing the largest public collection of the single nucleotide variants in the mouse genome. These SNPs were generated and deposited mainly by the Whitehead Institute. The remaining SNPs were deposited by Celera Genomics, the Mouse Genetics Program and Bioinformatics group, and Roche, Inc. (Mouse Phenome Database, http://www.aretha.jax.org/pub-cgi/phenome/mpdcgi?rtn=projects/details&id=146). Over half of the mouse SNPs were identified from five strains (129S1/SvImJ, BALB/cByJ, C3H/HeJ, C57BL/6J, and CZECHII/EiJ) by the Whitehead Institute. For the remaining SNPs, the mouse strains used for SNP discovery varied among the different sources. The inbred laboratory strains used for the SNP discovery originated from three subspecies groups, *Mus musculus domesticus, M. m. musculus,* and *M. m. castaneus,* which derived from their common ancestor ~1 million years ago [11,12]. The alleles at the polymorphic sites therefore reflect the evolutionary process of the nucleotide changes among these lineages.

Although previous studies based on a very limited number of loci have showed similar substitution patterns in rodents, no genome-wide study has been done [13,14]. In this study, we analyzed the recently released mouse SNP data (dbSNP build 119) to examine the substitution patterns and neighboring-nucleotide effects representative of the whole mouse genome and each chromosome. To evaluate the actual extent of the nucleotide bias, we normalized with respect to the averaged nucleotide proportion in the mouse genome and the relevant chromosome. We next analyzed the neighboring-nucleotide biases and patterns separately for transitional and transversional substitutions and for each of the six categories of substitution because transitions are expected to be influenced by CpG effect, while transversions are free of such effect. The results obtained from the mouse data were subsequently compared with the human data.

Results and discussion

Distribution of substitutions

We analyzed 433,192 mouse SNPs that are biallelic and have chromosome information. There were 147,731 A/G substitutions, 146,975 C/T substitutions, 37,403 A/C substitutions, 37,383 G/T substitutions, 36,114 A/T substitutions, and 27,586 C/G substitutions. Here, each type of substitution combined the nucleotide change from both directions (e.g., A → G and G → A) because the direction of a substitution is generally unknown for the dbSNP data. Instead of four types (e.g., combining A/G and C/T as one type), we analyzed six types of substitution in the dbSNP database to indicate the unique sequence orientation of the SNPs. Correspondingly, the proportions for the six categories of the substitution were 34.10, 33.93, 8.63, 8.63, 8.34, and 6.37% (Table 1). Both substitutions A/G and C/T accounted for approximately one-third of all the data and each of the other types (A/C, G/T, A/T, and C/G) accounted for less than 10%. Moreover, the frequency of the A/C substitution was close to that of G/T, a pattern also observed between A/G and C/T, reflecting the complementary strand symmetry. When we compared the distribution of substitutions in the mouse genome with that in the human genome, the proportion for each substitution was similar in both genomes with one notable exception (Table 1). The least frequently observed substitution was between nucleotides C and G (6.37% of total substitutions) in the mouse SNPs; it was between A and T (7.42%) in the human SNPs. This difference was also observed using Celera's mouse RefSNP data (release 3.4, http://www.celera.com/), which had approximately 2.8 million single nucleotide substitutions (Table 1). Therefore, this is not likely due to the small mouse data set, but rather indicates that the mutation rates between A and T and

Table 1
Distribution of nucleotide substitutions in the mouse and human genomes

Substitution	Mouse		Human	
	dbSNP	Celera RefSNP	dbSNP	Celera RefSNP
A/G (%)	34.10	33.35	33.17	33.28
C/T (%)	33.93	33.33	33.18	33.21
A/C (%)	8.63	9.13	8.69	8.76
G/T (%)	8.63	9.08	8.74	8.77
A/T (%)	8.34	8.83	7.42	7.31
C/G (%)	6.37	6.29	8.80	8.66
Ts/Tv[a]	2.13	2.00	1.97	1.98

The number of SNPs used was 433,192 (dbSNP mouse build 119), 2,780,789 (Celera mouse RefSNP release 3.4), 5,683,336 (dbSNP human build 119), and 3,580,926 (Celera human RefSNP release 3.2).

[a] The ratio of transition (Ts) over transversion (Tv).

between C and G are different between the human and the mouse genomes.

Table 1 shows that the ratio of transition over transversion was approximately 2 in the mouse genome, a value widely observed in the human genome [4,6,15–17].

Bias at the polymorphic sites

Given that two nucleotides at each variant site have the same allele frequency, the proportions of nucleotides at the polymorphic sites were 25.54, 24.47, 24.55, and 25.45% for A, C, G, and T, respectively. To examine the nucleotide bias at the polymorphic sites, the overall nucleotide composition in the mouse genome was obtained by a computer program using the genomic sequences downloaded from NCBI (build 32). The proportions of the four nucleotides were A 29.12%, C 20.87%, G 20.87%, and T 29.14% in a total of 2.70×10^9 nucleotides in the mouse genome. Therefore, the biases were −3.58, 3.60, 3.68, and −3.69%, respectively, relative to the whole genome.

The bias pattern at the substitution sites is consistent with what we observed in the human genome, which was −4.89, 4.90, 4.90, and −4.91, respectively. However, the extent of the deviation from the mouse genome average was approximately 26% less than that in the human genome. Note that the G + C content in the mouse genome was 41.74%, higher than that (40.89%) in the human genome.

Neighboring-nucleotide effects

More than half of the mouse SNPs in the dbSNP database had at least 250 bp flanking sequence at one side. The observed frequency of each nucleotide neighboring the polymorphic site, ranging from the immediate adjacent site to 250 bp away at each side and at the both sides, was obtained (Supplementary Table 1). The proportions of the four nucleotides were 29.33 (A), 23.50 (C), 21.58 (G), and 25.59% (T), at the 5′ immediate adjacent site, and 25.45 (A), 21.62 (C), 23.56 (G), and 29.37% (T), at the 3′ immediate adjacent site. The proportions for each of the four nucleotides are different at the 5′ site and at the 3′ site and are different from the polymorphic site. Moreover, the nucleotide C (23.50%) on the immediate 5′ adjacent site occurred more frequently than the genome average of 20.87%, and the nucleotide G (23.56%) on the immediate 3′ adjacent site occurred more frequently than the genome average of 20.87%. Next, the frequency of each nucleotide at the flanking sites was normalized by subtracting the corresponding genome average value (Supplementary Table 2). Fig. 1 shows the bias patterns after normalization. In general, the nucleotide patterns at the 5′ side of the substitution complemented what we observed at the 3′ side. The proportions at the immediate adjacent site at each side (−1 and + 1) showed the largest bias relative to the average in the mouse genome. For example, at the + 1 site, the frequency of G was 2.69% higher than the genome average and the frequency of A was 3.67% lower than the genome average. One striking feature is that, in terms of the absolute value, nucleotide A had the largest bias at almost all the sites at the 3′ side, and correspondingly, nucleotide T had the largest bias at almost all the sites at the 5′ side.

The bias patterns observed in the mouse genome are similar to those in the human genome; however, the extent of the bias in the mouse genome is notably less at the first two adjacent sites. In humans, for example, the frequency of C is 4.91% higher than the human genome average at the −1 site, and 2.61% higher at the −2 site. This compares to the bias of 2.63% at the −1 site and 0.78% at the −2 site in mice. Two features stood out when we compared the neighboring-nucleotide biases in these two genomes. First, away from the polymorphic site, there was a trend for nucleotides C and G to have a positive bias and nucleotides A and T had a negative bias, which is especially remarkable if the two sides combine. Second, the bias was the largest at

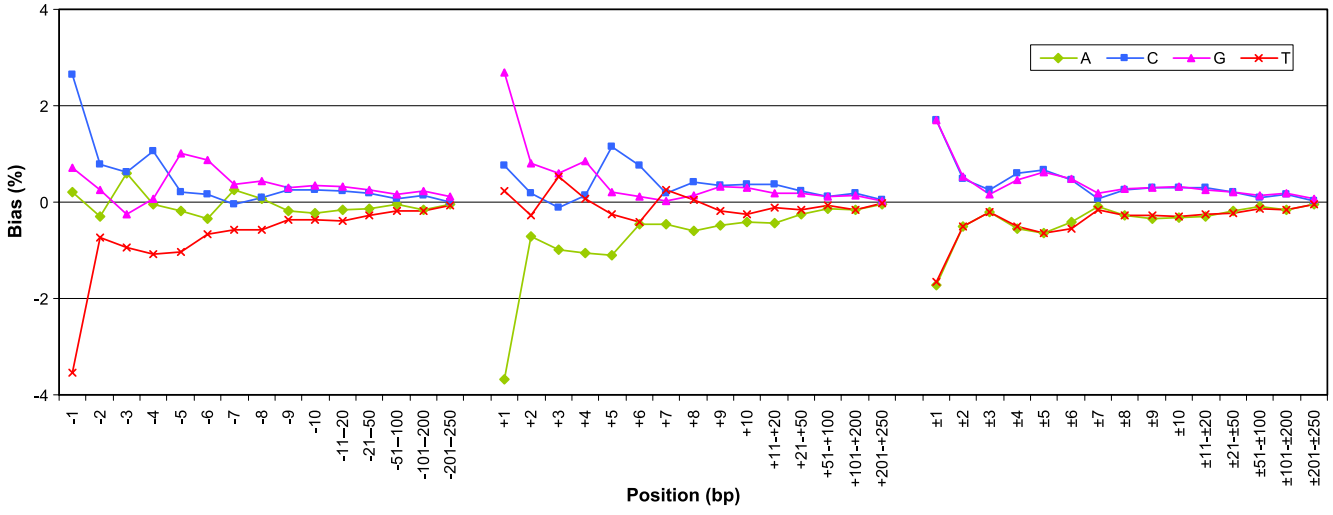

Fig. 1. Proportion bias of the neighboring nucleotides. A minus sign represents the 5′ side, plus sign the 3′ side, and plus/minus sign two-sided.

the immediate adjacent sites; it fluctuated within a moderate range at sites 2–6, decreased toward 0 at site 7, and then fluctuated within a smaller range and gradually decreased to 0 away from the polymorphic site. The reasons to cause such patterns in the human and mouse genome are unknown to us.

The excess of transitions and strong frequency bias of nucleotides G at the + 1 site and C at the −1 site are attributable mainly to the abundant hypermutable methylated 5′CpG3′ across the mouse and human genomes [5,9]. Approximately 80% of the dinucleotide CpGs in the mouse and human genomes are methylated at position 5 of the cytosine [18]. Deamination of methylated cytosine yields thymine and leads to dinucleotide TpG and, in the complement strand, CpA. Therefore, it greatly increases the presence of nucleotide G at the 3′ adjacent site and C at the 5′ adjacent site [6]. Indeed, when we examined 16 possible doublets in the mouse reference genome sequence, the proportion of CGs was only 0.84%, 3.52% below the expected value of 4.36%. On the other hand, the proportion of either TGs or CAs was 1.38% above the expected, reflecting the overrepresentation of substitutions CG → TG and CG → CA in the mouse SNP data. Furthermore, the proportion of CGs in the human genome was 0.99%, which is higher than that (i.e., 0.84%) in the mouse genome. This may partially explain the stronger bias (e.g., G at site + 1) in the human genome. An additional analysis revealed that CpG dinucleotides were overrepresented in the DNA small fragments (e.g., 6 bp) overlapping the polymorphic sites compared with the other part of the flanking sequences (unpublished results). This pattern has been reported in humans [8]. The CpG effect on the neighboring sites will be further addressed below in the category of transition.

Neighboring effects at the chromosome level

The G + C content in the mouse genome, which varied from approximately 39.28% on chromosome X to 43.82% on chromosome 11 (Table 2), is higher than that in the human genome. However, the distribution of the G + C content is tighter in the mouse genome [9]. The nucleotide bias at the adjacent sites was next examined for each chromosome using the chromosome-specific average. The results showed a marked excess of C at position −1 and G at position + 1 and conversely a marked decrease of T at position −1 and A at position + 1. At the −1 site, nucleotide T had a larger bias than C on all chromosomes except for chromosome 17. Similarly, at the + 1 site, nucleotide A had a larger bias than G on all chromosomes. At the substitution site, each chromosome had an excess of C and G but a decrease in A and T.

Table 2 indicates that, in the mouse genome, chromosomes with higher G + C content in general had a larger proportion bias of nucleotide C at the −1 site (C_{-1}) and G at the + 1 site (G_{+1}). Supplementary Fig. 1 shows the relationship between G + C content difference on each chromosome from the genome average (41.74%) and the corresponding proportion bias of C_{-1} and G_{+1}. The relationship can be formulated using a simple linear regression model as

$$\Delta C_{-1} = 2.66 + 0.34(GC - 41.74)R = 0.66,$$

for C_{-1}, and

$$\Delta G_{+1} = 2.74 + 0.49(GC - 41.74)R = 0.78$$

for G_{+1}, where ΔC_{-1} is the bias for C_{-1}, ΔG_{+1} is the bias for G_{+1}, and GC is the G + C content for the chromosome. The linear regressions above are statistically significant ($p = 0.001$ for ΔC_{-1} and $p < 0.0001$ for ΔG_{+1}). Comparing the linear relationship in humans, the equations above had smaller values for linear coefficient, intercept, and the coefficient of determination R^2, suggesting that overall CpG effect in the mouse genome may not be as strong as in the human genome.

For each chromosome, we then computed the bias for C_{-1} using the A/G substitutions only and for G_{+1} using the C/T substitutions only. The extent of bias was different on the autosomes and chromosome X. The bias of C_{-1} was within a range of 4.50–7.27% above the chromosome-specific average on 19 autosomal chromosomes with a mean value 5.91%, but it was only 2.68% on chromosome X. Similarly, the bias of G_{+1} was within a range of 4.67–8.01% on autosomes with a mean value 5.92%, but only 1.23% on chromosome X.

Neighboring effects on the categories of substitution

Hypermutable dinucleotide CpG affects only the transitional substitutions; therefore, we obtained the observed frequencies and examined the neighboring nucleotide effects separately for transitions and transversions and for each of the six categories of substitution: A/G, C/T, A/C, G/T, A/T, and C/G. When we compared the proportions of the four nucleotides at the neighboring sites, nucleotides C and G occurred more frequently in the transitions; conversely nucleotides A and T occurred more frequently in the transversions. For example, the proportion of nucleotide C at the −1 site was 24.90% for transitions, 4.39% higher than that for transversions (20.51%, Table 3). Accordingly, the bias for the nucleotide C at the −1 site was 4.03 and −0.36% relative to the mouse genome average for transitions and transversions, respectively. As in humans, transitions had larger biases to the mouse genome average than transversions.

For the six categories of substitution, the proportions of the nucleotides in the first two positions, which in general had large biases, are shown in Table 3. The nature of the observed frequencies varied greatly among the six substitution categories, especially at the immediate adjacent sites. As for all substitutions, the bias for each category of substitution is not so strong compared in humans. For example, the frequency of C at site −1 in the category of

Table 2
Nucleotide–nucleotide bias at the substitution site and immediate adjacent sites for each chromosome (Chr.)

Chr.	G + C (%)[a]	–1				0				+1			
		A (%)	C (%)	G (%)	T (%)	A (%)	C (%)	G (%)	T (%)	A (%)	C (%)	G (%)	T (%)
1	41.12	0.28	2.47	1.09	–3.84	–4.11	4.13	3.74	–3.76	–3.80	1.13	2.22	0.46
2	42.09	–0.05	3.25	0.79	–3.99	–3.43	3.57	3.42	–3.56	–3.42	0.62	2.90	–0.10
3	40.51	0.17	2.54	0.59	–3.30	–3.96	4.00	4.23	–4.26	–4.06	1.01	1.88	1.18
4	42.32	0.15	2.36	1.06	–3.57	–3.45	3.39	3.28	–3.22	–4.30	0.70	2.92	0.68
5	42.53	–0.08	3.01	1.01	–3.94	–3.45	3.27	3.49	–3.31	–4.13	0.65	3.19	0.29
6	41.44	0.21	2.29	1.19	–3.69	–3.61	3.60	4.01	–4.00	–3.57	0.99	2.53	0.04
7	43.27	0.07	2.37	1.52	–3.97	–3.06	3.09	2.81	–2.84	–3.79	1.32	2.73	–0.26
8	42.34	0.33	2.81	0.52	–3.67	–3.50	3.48	3.16	–3.14	–3.85	0.42	2.95	0.47
9	42.69	0.04	3.00	0.60	–3.64	–3.33	3.40	3.23	–3.30	–3.53	0.91	3.18	–0.55
10	41.48	0.30	3.10	0.14	–3.54	–3.84	3.63	3.98	–3.77	–3.36	0.51	3.02	–0.18
11	43.82	0.23	3.04	0.63	–3.90	–2.62	2.68	3.18	–3.23	–3.91	0.53	3.34	0.04
12	41.69	1.28	2.60	0.06	–3.93	–3.24	3.59	3.78	–4.13	–2.89	0.33	2.17	0.39
13	41.58	0.39	2.62	0.58	–3.58	–3.88	3.89	3.51	–3.51	–4.07	0.87	3.14	0.06
14	41.02	–0.06	2.47	0.81	–3.21	–3.67	3.83	4.36	–4.52	–3.28	0.73	2.30	0.25
15	41.93	0.25	2.40	0.78	–3.42	–3.52	3.36	3.75	–3.58	–4.33	0.58	2.98	0.77
16	40.95	0.56	2.14	0.16	–2.86	–3.74	3.83	3.82	–3.91	–2.99	0.54	2.15	0.30
17	42.80	–0.38	3.48	0.35	–3.45	–3.14	3.12	3.33	–3.31	–4.61	0.85	3.70	0.06
18	41.46	–0.06	2.75	1.00	–3.69	–3.76	3.88	3.83	–3.95	–3.80	0.25	3.46	0.09
19	42.77	–1.07	3.84	1.36	–4.13	–3.52	3.14	3.34	–2.95	–4.98	1.05	3.86	0.07
X	39.28	0.60	1.40	1.20	–3.20	–4.26	3.78	5.59	–5.11	–2.57	1.33	1.27	–0.03
Genome	41.74	0.21	2.63	0.71	–3.55	–3.59	3.60	3.68	–3.69	–3.67	0.75	2.69	0.23

[a] The average G + C content on chromosome.

substitution A/G was 26.71% in mice, significantly less than the 33.46% in humans. However, this frequency is still 5.84% higher than the mouse genome average and 9.65% higher than that observed for substitution C/G (i.e., 17.06%). On the other hand, the frequency of A at site −1 was 34.61% for substitution C/G, 11.66% higher than that observed for substitution G/T (i.e., 22.95%).

Fig. 2 shows the bias patterns of neighboring nucleotide proportions for each substitution category. The normalized bias pattern for the transitions (Figs. 2A and 2B) is remarkably different from that for the transversions (Figs. 2C–2F). For transitions, the pattern was strongly influenced by the hypermutability of dinucleotide CpG. In the category of substitution A/G, a large excess of C was observed at site −1 (5.84%). At the 5′ side, the nucleotide G is consistently higher than the genome average, while nucleotide T is consistently lower than the genome average. At the 3′ side, the proportion of A was 4.95% below the genome average at site +1, but it increased to 3.07% above the genome average at site +2. Interestingly, the proportion of T showed the opposite pattern; it was 1.94% higher than expected at +1, but 3.59% lower than

Table 3
Proportion of neighboring nucleotides for each category of substitution

Type	Position	A (%)	C (%)	G (%)	T (%)	Position	A (%)	C (%)	G (%)	T (%)
Transition	–1	28.77	24.90	21.70	24.63	+1	24.63	21.71	24.91	28.76
	–2	28.22	22.03	22.06	27.70	+2	27.80	21.90	22.08	28.22
A/G	–1	26.41	**26.71**	21.79	25.09	+1	**24.18**	21.57	23.18	31.08
	–2	31.16	20.98	24.57	**23.29**	+2	32.19	19.13	23.13	25.55
C/T	–1	31.13	23.10	21.60	**24.17**	+1	25.08	21.85	**26.67**	26.41
	–2	25.27	23.08	19.53	32.12	+2	**23.37**	24.69	21.02	30.91
Transversion	–1	30.53	20.51	21.32	27.65	+1	27.21	21.43	20.68	30.67
	–2	30.12	20.85	19.09	29.93	+2	29.70	19.23	20.84	30.23
A/C	–1	31.20	23.41	21.99	**23.40**	+1	31.17	23.38	17.42	28.03
	–2	32.06	20.79	19.48	27.67	+2	30.73	19.47	20.45	29.35
G/T	–1	28.27	17.26	23.34	31.13	+1	**22.95**	21.80	23.54	31.71
	–2	28.98	20.49	19.82	30.71	+2	27.31	20.08	20.78	31.83
A/T	–1	29.03	23.50	19.13	28.34	+1	27.61	19.24	23.60	29.55
	–2	32.46	17.11	17.31	33.13	+2	32.75	17.62	17.16	32.46
C/G	–1	**34.61**	17.06	20.54	27.80	+1	27.12	21.17	17.42	**34.29**
	–2	25.99	**26.31**	19.93	27.77	+2	27.55	19.86	**26.28**	26.31

The bias equal to or greater than 5% is in bold.

expected at +2. These observations suggest that A at +1 has a negative influence on the substitution rate of A/G, whereas T at +1 has a positive influence. Similar patterns were observed in the category of substitution C/T, for example, there was a large excess of G at site +1 (5.80%).

As in humans, the nature of neighboring effects on transversions is complex. First, there is a trend that two nucleotides involved in a substitution likely occur more often than expected at the neighboring sites. This trend was observed ranging from site 1 to as far as 250 bases at each side. However, some exceptions existed, espe-

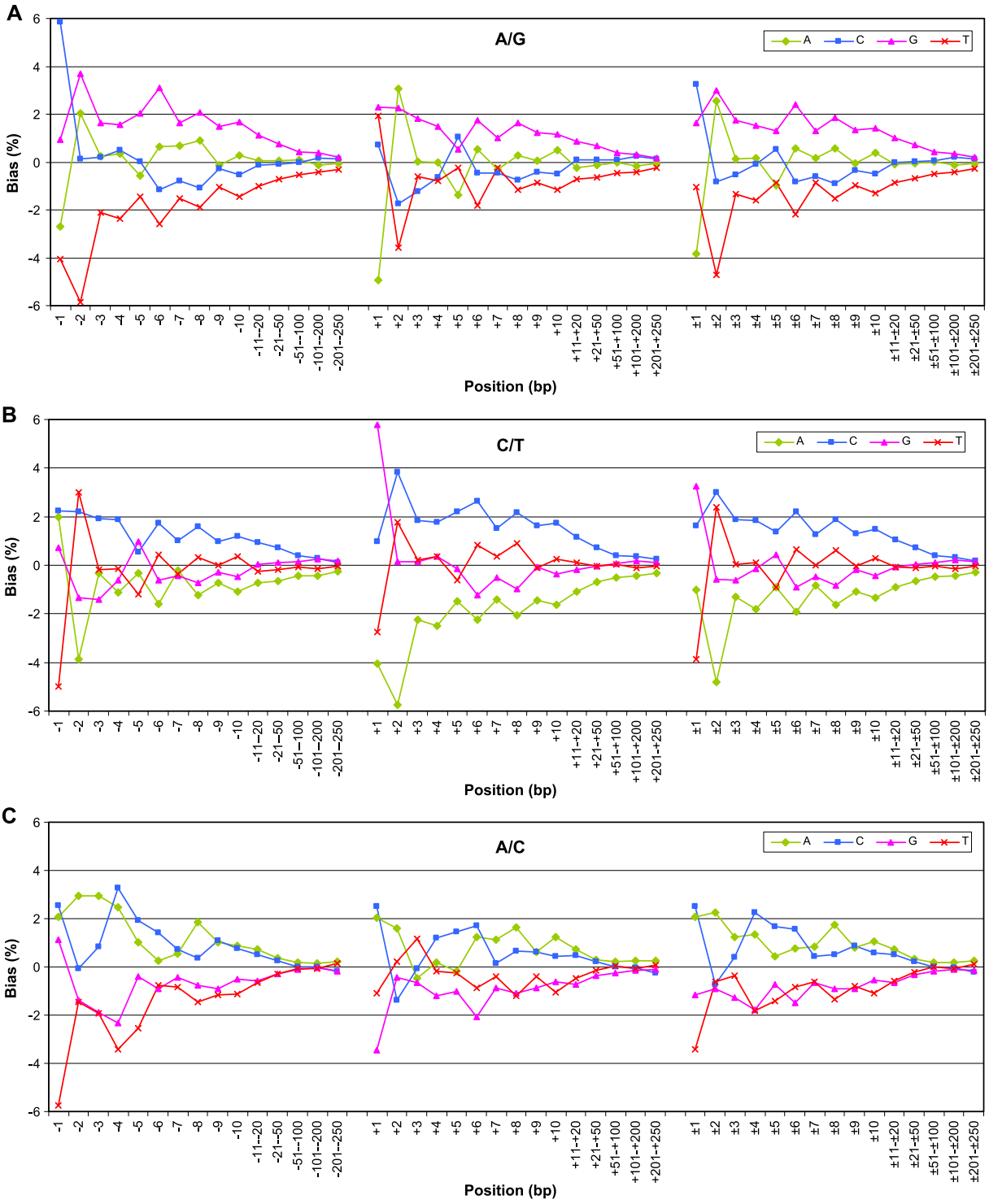

Fig. 2. Neighboring-nucleotide effects on the six categories of substitution. The nucleotide bias is normalized by the averaged values in the genome. A minus sign represents the 5′ side, plus sign the 3′ side, and plus/minus sign two-sided.

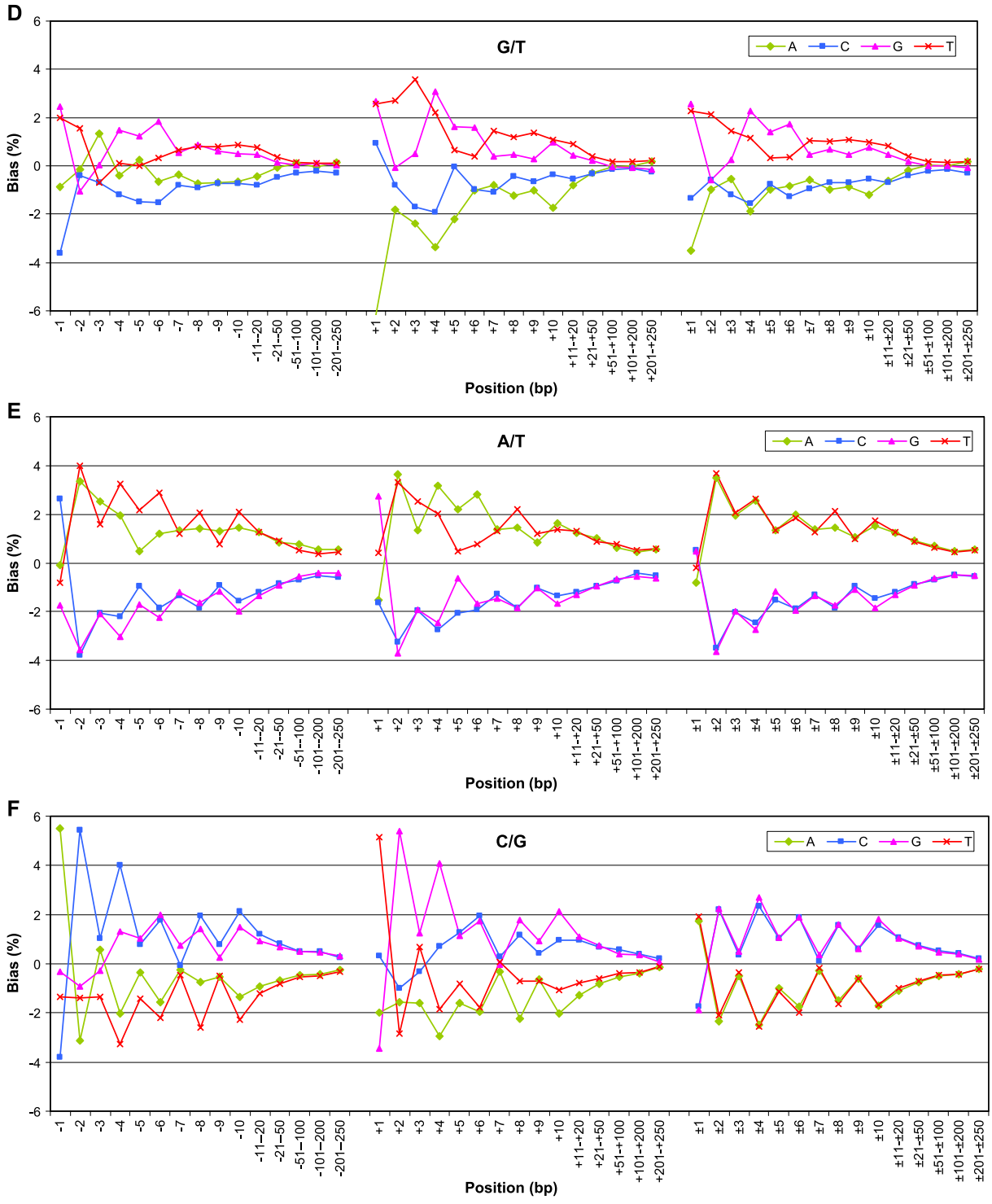

Fig. 2 (*continued*).

cially at the immediate adjacent sites in the categories of C/G and A/T. They are discussed below. Second, the same bias patterns were observed for substitutions A/C and G/T (Figs. 2C and 2D). This is because nucleotide A pairs to T and C pairs to G. For example, nucleotide T at the −1 site occurred 5.74% less often than the average in the category A/C, and correspondingly, nucleotide A at the +1 site occurred 6.17% less often than the average in the category G/T (Figs. 2C and 2D). Third, the patterns in the A/T and C/G categories are complementary to each other. In the category of A/T substitutions, the proportion of G was lower than the genome average at all 3′ sites except for site +1, which had 2.73% positive bias (Fig. 2E). Strikingly, the proportion of G decreased

sharply to 3.71% below the genome average at site +2. For the C/G substitution, opposite symmetric bias patterns were observed between nucleotides A and C at the 5′ flanking sites and between G and T at the 3′ flanking sites (Fig. 2F). Nucleotide A was 5.49% above the genome average at the −1 site, but was 3.14% below the average at the −2 site. On the other hand, nucleotide C was 3.81% below the genome average at the −1 site but 5.44% above the genome average at the −2 site. This sharp difference at the two immediate adjacent sites was unique for the C/G substitutions, which had similarly been observed in the human genome. Finally, although these biases became progressively smaller, small biases could still be observed at the far sites of the flanking sequences.

Influence of A + T content on transversion

Early studies of substitutions in the noncoding and coding regions in the plant chloroplast genomes found that the transversional substitutions were more likely to occur when the two immediate adjacent sites had two nucleotides A or T [19,20]. This pattern was also observed in humans, although the influence was not as strong [7]. In this study, we examined the probability of the occurrence of a transversion in the mouse genome and compared it to the other genomes. Table 4 shows the proportions of transversions in 16 categories grouped by A + T content at the two immediate adjacent sites using SNP data from three genomes: mouse, human, and *Arabidopsis*. In the mouse genome, the proportion of transversions was largest (39.9%) when TNA occurred and smallest (25.6%) when CNG occurred. The proportion of transversions was higher for two sites flanked with an A + T context equal to 2 (36.5%), lower for an A + T context equal to 1 (29.9%), and lowest for an A + T context equal to 0 (29.3%). The same probability pattern was observed in the human and *Arabidopsis* genomes based on the data used. Among these three genomes, the influence of A + T context on transversion is the weakest in the mouse genome and the strongest in the *Arabidopsis* genome (Table 4). Interestingly, the extent of the influence of A + T context on transversion is significantly higher in *Arabidopsis* and in the plant chloroplast than in the mammalians [19,20]. This large difference is probably attributable to the higher A + T content in the *Arabidopsis* genome (63.97%) and in the sequences used in the chloroplast genome (63.3–73.5%) because A + T-rich regions, which are highly unstable, have a higher proportion of transversions [19]. A further investigation of the SNPs in the A + T-rich regions and more SNP data from other species should provide us more genetic information and, therefore, help us to understand the mechanisms of transversional substitutions and the genome sequence evolution.

Table 4
Proportion of transversions (Tv) grouped by A + T content at adjacent sites

A + T context	Type	Mouse Tv (%)	Human[a] Tv (%)	*Arabidopsis*[b] Tv (%)
0	CNC[c]	31.0	32.0	44.6
	CNG	25.6	19.4	46.8
	GNC	30.1	40.8	33.6
	GNG	31.4	32.0	43.4
	Subtotal	29.3	29.4	42.9
1	ANC	34.0	41.2	45.1
	ANG	27.0	23.6	42.7
	CNA	28.1	30.7	47.6
	CNT	27.1	23.7	46.4
	GNA	30.6	39.0	40.9
	GNT	33.7	41.3	45.1
	TNC	31.0	38.7	41.6
	TNG	28.7	30.6	47.0
	Subtotal	29.9	32.0	44.7
2	ANA	35.5	38.5	51.8
	ANT	35.6	33.8	53.4
	TNA	39.9	45.9	54.6
	TNT	36.0	38.6	52.5
	Subtotal	36.5	38.7	53.0

[a] Proportion of transversion was obtained using 5,683,336 human SNPs from dbSNP database (build 119).
[b] Proportion of transversion was obtained using 37,342 SNPs from Monsanto *Arabidopsis* Polymorphism Collection [25].
[c] N denotes any substitution (A/G, C/T, A/C, G/T, A/T, and C/G).

Ranking of nucleotide proportion at adjacent sites

Table 5 presents the ranking of the nucleotide proportion at sites −1 and +1. The same ranking system was used as in our previous study of human SNPs [7]. The symbol >> denotes greater than 5% difference between two nucleotide proportions, > denotes 1–5% difference, and ≈ denotes less than 1% difference. For all substitutions, the rankings of the observed proportions were A > T > C > G at the −1 site and T > A > G > C at the +1 site. The rankings for the transitions and transversions were different and the rankings for each substitution categories varied.

Because the occurrence of nucleotide substitution depends on the actual context of the genome sequence, a more informative approach to ranking the nucleotide bias is to normalize the nucleotide proportions by the genome average values. As shown in Table 5, the rankings were different from those by observed proportions for all substitutions and for each category. The overall rankings of proportion bias were C > G ≈ A > t at the −1 site and G > C ≈ T > a at the +1 site. Here, an uppercase letter denotes an observed proportion greater than the genome average and a lowercase letter denotes an observed proportion lower than the genome average. For transitions, the order was essentially the same as the one observed for all substitutions, reflecting the strong effect of the hypermutability of CpG (Table 5). The rankings varied among the different categories of transversions. Overall, the rankings for transversions were A ≈ G ≈ c > t at the −1 site and T ≈ C ≈ g > a at the +1 site.

Table 5
Ranking of nucleotide proportions at immediate adjacent sites

Type	Observed proportion		Proportion bias	
	–1	+1	–1	+1
Substitution	A > T > C > G[a]	T > A > G > C	C > G ≈ A > t[b]	G > C ≈ T > a
Transition	A > C ≈ T > G	T > G ≈ A > C	C > G > a > t	G > C > t > a
A/G	C ≈ A > T > G	T >> A > G > C	C > G > a > t	G ≈ T > C >> a
C/T	A >> T > C > G	G ≈ T > A > C	C ≈ A > G >> t	G > C > t > a
Transversion	A > T >> G ≈ C	T > A >> C ≈ G	A ≈ G ≈ c > t	T ≈ C ≈ g > a
A/C	A >> C ≈ T > G	A > T > C >> G	C ≈ A > G >> t	C ≈ A > t > g
G/T	T > A > G >> C	T >> G ≈ A > C	G ≈ T > a > c	G ≈ T > C >> a
A/T	A ≈ T >> C > G	T > A > G > C	C > a ≈ t ≈ g	G > T > a ≈ c
C/G	A >> T >> G > C	T >> A >> C > G	A >> g > t > c	T > C > a > g

[a] >>, greater than 5% difference between two nucleotide proportions; >, 1–5% difference; and ≈, less than 1% difference.
[b] A lowercase letter denotes a negative bias of the observed nucleotide proportion compared to the genome average.

The rankings of observed proportion in the mouse genome are essentially the same as observed in the human genome. The rankings of the proportion bias differed somewhat between the mouse and the human data, especially for the transitional substitutions. For transversions, the overall rankings of the proportion bias were the same in the mouse and human genomes, but differed slightly for each category. The results indicated that ranking by the observed proportions and the proportion biases may be influenced by the sequence context (e.g., A + T) and by the hypermutability of CpG, respectively, in the mouse genome.

SNP sample size

Similar bias patterns at the neighboring sites were observed using 433,192 mouse SNPs and 5,683,336 human SNPs in this study and using 2.6 million human SNPs in the previous study [7]. One may ask how many SNPs are sufficient to represent the bias patterns in the mouse or human genomes. A program was developed to select randomly a subset of the mouse or human SNPs, to obtain the biases of the neighboring nucleotides, and to compare the biases obtained from the entire data set. The initial analysis showed that 20,000 mouse or human SNPs were able to present the same bias patterns in the mouse or human genomes, respectively.

Conclusion

In conclusion, we investigated the patterns of neighboring-sequence context of the SNPs across the mouse genome using the largest publicly available data set. Large neighboring-nucleotide bias relative to the genome average was observed at the immediate adjacent sites and small bias was observed at the other adjacent sites. The examination of six categories of the substitutions indicated that transitions were strongly influenced by the hypermutability of dinucleotide CpG, and transversions might help us to understand other molecular mechanisms of the point mutations. Signature of the CpG effects was further detected at the chromosome level. In general, one chromosome having higher G + C content shows the larger proportion bias of nucleotide G at the +1 site and C at the −1 site. In the mouse, human, and *Arabidopsis* genomes the probability of a transversion increased with the A + T context equal to 2 at the two immediate adjacent sites and decreased with the A + T context equal to 0. However, the influence of A + T context on transversion in the mouse and human genome was not as strong as in the plant system. The overall rankings of nucleotide bias were C > G ≈ A > t at the −1 site and G > C ≈ T > a at the +1 site. Since the SNPs used in this study were identified genome-wide, the observed patterns were not limited to particular genes or specific regions of the genome and they should represent a comprehensive view of the effects of the neighboring nucleotides on substitutions during the evolutionary processes of the mouse genomes. Compared with humans, the neighboring bias patterns in mice are generally the same for all substitutions and for each category. However, the extent of the biases, although strong, is not as strong as that in the human genome. This reflects the different genomic sequence context after an approximately 75 million-year divergence of the human and mouse lineages [9]. Further comparative analyses on the specific genomic categories (e.g., exonic regions) or syntenic regions between these two genomes will shed more light on the nucleotide substitutions and genome evolution.

Materials and methods

Data sources

The dbSNP database was selected among various data sources because (1) all the SNPs, their flanking sequences, and other annotation information are publicly available; (2) the SNP data were from all the chromosomes except chromosome Y; and (3) the number of SNPs was sufficient for data analysis at the genome level. Mouse and human SNP data were downloaded from ftp://ftp.ncbi.nih.gov/snp/

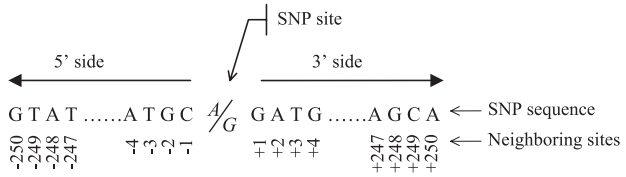

Fig. 3. Annotation of the flanking sites of a SNP.

(build 119, released on February 4, 2004). There were 498,463 and 7,231,721 reference (rs) SNPs registered in the mouse and human dbSNP databases, respectively. Approximately 94% of the mouse SNPs and 47% of the human SNPs have been validated. In addition, many studies have reported the high validation rates for the SNPs deposited in the dbSNP database [9,12,21–24]. Only those SNPs that are biallelic (i.e., SNP class 1) and can be mapped to the chromosome were analyzed; therefore, 433,192 mouse SNPs and 5,683,336 human SNPs were used. Mouse and human genome sequences were downloaded from ftp://ftp.ncbi.nih.gov/refseq/ (mouse genome build 32 released in October 2003, human genome build 34 version 2 released on January 28, 2004).

We downloaded SNP data for *Arabidopsis thaliana* from the Monsanto *Arabidopsis* Polymorphism Collection (http://www.arabidopsis.org/Cereon/index.jsp). There were 37,342 single nucleotide substitutions across the genome. We used the flanking sequences for each SNP provided in the database. For these data, approximately 100% of the validation rate was obtained using a resequencing approach. More details are available on the above mentioned Web site and in Jander et al. [25].

Analysis of the SNP data

For a better comparison, the same approach and annotations in our human study were used [7]. Fig. 3 shows the polymorphic site of a SNP and its 5′ and 3′ flanking sequences. To represent the nucleotides at each neighboring site, we labeled the position at the 5′ side, the 3′ side, and the two sides combined as a negative number, a positive number, and a "±" sign, respectively. For example, +1 stands for the 3′ immediate adjacent nucleotide of the polymorphic site and ±1 for the average of the two immediate adjacent sites. The frequencies of nucleotides A, C, G, and T at each neighboring site were computed using the 5′ and 3′ flanking sequences of SNPs. Next, the proportion of each nucleotide at each site was normalized by its average value in the mouse genome. Among the mouse SNPs, approximately 99% have at least 100 bp at one flanking side, 58% have at least 250 bp at one flanking side, but only about 3% have more than 250 bp. Therefore, we scored the number of neighboring-nucleotides for each site from site 1 to site 250. The proportion of each nucleotide was calculated for the first 10 bp at each side and the averaged proportion of each nucleotide was calculated for the ranges of 11–20, 21–50, 51–100, 101–200, and 201–250 bp. To examine further the bias, the nucleotide frequencies of each site were obtained for each chromosome, for transitions (A/G and C/T) and transversions (A/C, G/T, A/T, and C/G), and for each of the six categories of substitution.

The programs and other information are available at http://www.people.vcu.edu/~zzhao/snp_neighbors/.

Acknowledgments

We are indebted to the people who collected, annotated, and managed those data for public access, especially the NCBI's dbSNP team and Monsanto *Arabidopsis* Polymorphism Collection team. We thank Yixi Zhong for his assistance in the early phase of this study and two anonymous reviewers for the insightful comments on a previous draft. This work was supported by a startup fund of the Virginia Commonwealth University.

Appendix A. Supplementary Materials

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2004.06.015.

References

[1] A.L. Hughes, et al., Widespread purifying selection at polymorphic sites in human protein-coding loci, Proc. Natl. Acad. Sci. USA 100 (2003) 15754–15757.
[2] Z. Zhao, Y.-X. Fu, D. Hewett-Emmett, E. Boerwinkle, Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution, Gene 312 (2003) 207–213.
[3] F.S. Collins, E.D. Green, A.E. Guttmacher, M.S. Guyer, A vision for the future of genomics research, Nature 422 (2003) 835–847.
[4] 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.
[5] D.N. Cooper, M. Krawczak, The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions, Hum. Genet. 85 (1990) 55–74.
[6] 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.
[7] Z. Zhao, E. Boerwinkle, Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome, Genome Res. 12 (2002) 1679–1686.
[8] D.J. Tomso, D.A. Bell, 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 (2003) 303–308.
[9] R.H. Waterston, et al., Initial sequencing and comparative analysis of the mouse genome, Nature 420 (2002) 520–562.
[10] R.A. Gibbs, et al., Genome sequence of the Brown Norway rat yields insights into mammalian evolution, Nature 428 (2004) 493–521.
[11] L. Silver, Mouse Genetics, Oxford Univ. Press, New York, 1995.

[12] C.M. Wade, et al., The mosaic structure of variation in the laboratory mouse genome, Nature 420 (2002) 574–578.
[13] T. Gojobori, W.H. Li, D. Graur, Patterns of nucleotide substitution in pseudogenes and functional genes, J. Mol. Evol. 18 (1982) 360–369.
[14] K.H. Wolfe, P.M. Sharp, Mammalian gene evolution: nucleotide sequence divergence between mouse and rat, J. Mol. Evol. 37 (1993) 441–456.
[15] R.D. Blake, S.T. Hess, J. Nicholson-Tuell, The influence of nearest neighbors on the rate and pattern of spontaneous point mutations, J. Mol. Evol. 34 (1992) 189–200.
[16] 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.
[17] J.C. Venter, et al., The sequence of the human genome, Science 291 (2001) 1304–1351.
[18] F. Antequera, Structure, function and evolution of CpG island promoters, Cell. Mol. Life Sci. 60 (2003) 1647–1658.
[19] B.R. Morton, V.M. Oberholzer, M.T. Clegg, The influence of specific neighboring bases on substitution bias in noncoding regions of the plant chloroplast genome, J. Mol. Evol. 45 (1997) 227–231.
[20] B.R. Morton, The influence of neighboring base composition on substitutions in plant chloroplast coding sequences, Mol. Biol. Evol. 14 (1997) 189–194.
[21] K. Lindblad-Toh, et al., Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse, Nat. Genet. 24 (2000) 381–386.
[22] G. Marth, et al., Single-nucleotide polymorphisms in the public domain: how useful are they? Nat. Genet. 27 (2001) 371–372.
[23] R. Sachidanandam, et al., A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms, Nature 409 (2001) 928–933.
[24] D.E. Reich, S.B. Gabriel, D. Altshuler, Quality and completeness of SNP databases, Nat. Genet. 33 (2003) 457–458.
[25] G. Jander, et al., Arabidopsis map-based cloning in the post-genome era, Plant Physiol. 129 (2002) 440–450.