Sequence context analysis in the mouse genome: Single nucleotide polymorphisms and CpG island sequences

Zhongming Zhao\textsuperscript{a,b,c,*}, Fengkai Zhang\textsuperscript{a}

\textsuperscript{a} Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA 23298, USA
\textsuperscript{b} Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23284, USA
\textsuperscript{c} Key Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China

Received 24 May 2005; accepted 20 September 2005
Available online 28 November 2005

Abstract

A genome-wide view of sequence mutability in mice is still limited, although biologists usually assume the same scenario for mice as for humans. In this study, we examined the sequence context in the local environment of 482,528 mouse single nucleotide polymorphisms (SNPs). We found that CpG-containing short sequences, in general, had more representation in the local sequences of SNPs compared to the genome sequences. The extent of this overrepresentation was stronger in mice than in humans, which is inconsistent with previous observations of the weaker neighboring-nucleotide biases on mouse SNPs. To exclude the CpG effect, we compared the distribution patterns of short sequences among the six categories of SNPs. The results revealed an even stronger pattern in the CpG-containing group for C/G substitution compared to for A/G or C/T substitutions. We next performed the first genome-wide sequence context analysis of SNPs in the mouse CpG islands. SNPs occurring at CpG sites were 3.14-fold less prevalent than expected, suggesting the suppression of methylation-dependent deamination in the CpG islands. The extent of this suppression was less in mice than in humans. Finally, compared with humans, the observations of a greater deficit of CpG dinucleotides, a stronger overrepresentation of CpG-containing \textit{n}-mers surrounding the polymorphic sites, and a higher SNP/genome ratio of CpG dinucleotides in the mouse genome support the “loss of CpG islands” model in the mouse lineage.

Keywords: Single nucleotide polymorphisms; SNPs; Short sequence; Mouse; Human; CpG islands; CpG dinucleotide; CpG effect

A detailed view of the nonrandom distribution of sequence context surrounding the nucleotide polymorphisms in the mammalian genome is important for understanding the mechanism of point mutations, protein–DNA interactions, and genome sequence evolution. The nonrandom fashion of mutations was initially observed with the strong excess of transition over transversion from the limited sequence data from pseudogenes and functional regions [1,2]. The approximate 2:1 ratio of transitional over transversional substitution, which has been widely observed in mammalian genomes, can be attributed mainly to the \textasciitilde 10-to 50-fold higher mutation rate of mC to T than other transitional changes [3,4]. This mechanism, which is commonly known as the CpG effect, should result in an uneven distribution of nucleotides at the neighboring sites of polymorphisms during evolution. Indeed, strong neighboring-nucleotides biases, in particular strong positive biases of nucleotide G at the 3' immediate adjacent site and C at the 5' immediate adjacent site, were observed among 7271 human coding single nucleotide polymorphisms (SNPs), 2.6 million human SNPs, and 0.4 million mouse SNPs [5–7]. However, the extent of the biases was notably less in mice than in humans, calling for further investigation [7].

The genome-wide analyses above [5–7] were based mainly on the examination of nucleotide compositions at each flanking site of SNPs. Although this simple approach could easily detect the biases at the neighboring sites, in particular at the immediate adjacent sites, the information discovered at each neighboring site was isolated and, thus, limited. Analysis of short sequences surrounding the polymorphic sites would be a
islands, which are clusters of CpG dinucleotides in the GC-rich regions, are particularly interested in the SNP analysis of mouse CpG islands. In this study, we were generally under selection pressure while noncoding regions in the CpG dinucleotides among genomic categories, (2) coding regions are the nucleotide changes in the islands. In addition, CpG islands have been identified in mice. For example, an analysis of short sequences surrounding polymorphic sites may reveal not only the specific sequence environment that favors point mutation occurrence and survival but also the characteristics with respect to the polymorphism in the genomic categories (e.g., promoter regions, exonic regions, splicing sites). Recently, Tomso and Bell [12] investigated the distribution of the 6-mer at 1.9 million human polymorphic sites versus their flanking sequences and revealed abundant biological information. For example, the fragments containing CpG dinucleotides were overrepresented at the polymorphic sites in the overall human genome. To our knowledge, no such studies have been performed in the mouse or other mammalian genomes.

Another limitation in the previous neighboring-nucleotide studies [6,7] is the lack of analysis of the SNPs in the categorized genomic regions such as promoters and exons. Such information is more important because (1) mutational effects rely on sequence characteristics (e.g., GC content, methylation of CpG dinucleotides) among genomic categories, (2) coding regions are generally under selection pressure while noncoding regions in general are not, and (3) the information in the functional regions may be further applied to disease studies. In this study, we were particularly interested in the SNP analysis of mouse CpG islands, which are clusters of CpG dinucleotides in the GC-rich regions [13]. The reasons are given below.

First, CpG dinucleotides in the mouse CpG islands are often unmethylated. The CpG effect would generally not influence the nucleotide changes in the islands. In addition, CpG islands reside in approximately 60% of the promoters in the mouse genome and may participate in the regulation of gene expression and cell differentiation by the methylation mechanism [13,14]. This suggests possible functional constraints in the CpG island sequences [15]. How these factors influence the genetic polymorphisms in the overall mouse genome is unclear. Second, although the number of genes seems to be almost the same in the mouse and human genomes, many fewer CpG islands have been identified in mice [16,17]. For example, an initial analysis found ~15,500 and ~27,000 in the nonrepetitive portions of the mouse and human genome sequences [16]. The evolutionary forces causing such a large difference between these two lineages remain largely unknown. One explanation was the smaller fraction of extremely high GC content in the mouse genome [16]. Obviously, this observation itself needs to be further investigated. The second explanation was based on a “loss of CpG islands” model in the mouse genome [17]. Under this model, CpG dinucleotides in some islands became de novo methylated. Because of the hypermutability of methylated CpG dinucleotides, their GC content decreased and this finally led to non-CpG island sequences. This model was supported by an examination of the mutation patterns in three genes [17] and later at least in another gene, Dmalp [18]. However, according to this model, the GC content in the mouse genome would become lower. This is not true because the GC content in the mouse genome is slightly higher than that in the human genome [16]. To explore how the CpG islands have evolved since the split of the human and mouse lineages, a genome-wide comparative analysis of sequence context patterns for nucleotide variations in the mouse and human CpG islands is warranted.

In this study, we report an investigation into the sequence context of 482,528 mouse SNPs, representing the largest publicly available SNP data (dbSNP build 123). First, an analysis of short sequences surrounding SNPs revealed that the CpG-containing n-mers were abundantly represented in the local environment of polymorphic sites. Second, aiming to distinguish the CpG effect, we examined the distribution of short sequences for each of the six categories of substitution. We next screened the CpG islands in the mouse genome and mapped SNPs to these putative islands. Finally, we examined the mutational patterns and explored the CpG island maintenance mechanisms in the mouse genome by comparing the distribution of sequence context in the mouse CpG islands, their SNPs, and the whole genome. This study provides detailed information on the influences of local sequences on polymorphisms and evidence supporting the loss of CpG islands model in the mouse genome.

Results

Distribution of short sequences

A total of 482,528 mouse SNPs that were non-insertion/deletion, were biallelic, and had a unique hit in the mouse genome reference sequences were extracted from the downloaded data (dbSNP build 123). To investigate the sequence preference of SNPs in the mouse genome, we compared the distribution of the short fragments (n-mers, n = 3–8) in the local sequences of these SNPs with that in the mouse genome reference sequences. For an n-mer analysis, the local sequence of each SNP was defined by 2 × n − 1 nucleotides symmetrically overlapping the polymorphic site (see Materials and methods). Here, we present the results of the 6-mer analysis first. Fig. 1 displays the distribution of all possible 4096 (4^6) 6-mers plotted by the frequency observed in the SNP local sequences versus the frequency in the mouse genome sequences. Under a simple random mutation model, each fragment in the SNP local sequences is expected to occur with the same frequency as that in the genome sequences. Therefore, the n-mer fragments should scatter along the diagonal line. Fig. 1 shows an overall downward skewed distribution. This was largely attributable to those n-mers that were highly frequently observed in the mouse genome. A further examination found that the GC content among those 6-mer fragments with high frequency was generally lower than among those with low frequency. For example, the GC content among the top 10%
most frequently observed 6-mers was on average 33.4%, this compared with 67.9% among the top 10% least frequently observed 6-mers. The distribution above indicates that the AT-rich short sequences in general have less representation of nucleotide polymorphisms relative to their average in the mouse genome.

Next, we divided all the 6-mers into two exclusive groups: (1) the CpG group, which included 6-mers containing at least one CpG dinucleotide (labeled red in Fig. 1), and (2) the non-CpG group, which included the remaining 6-mers (labeled blue in Fig. 1). The distributions of these two groups were clearly distinct. Using a simple linear regression model, the distribution of 6-mers was represented by \( f_{\text{SNP}} = 1.4 \times 10^{-5} + 2.71f_{\text{genome}} (R^2 = 0.87) \) for the CpG group and \( f_{\text{SNP}} = 9.6 \times 10^{-5} + 0.62f_{\text{genome}} (R^2 = 0.88) \) for the non-CpG group. Here, \( f_{\text{SNP}} \) and \( f_{\text{genome}} \) denote the frequency of a 6-mer in the SNP local sequences and in the genome sequences, respectively.

The distribution patterns in the 6-mer analysis were similarly observed in other n-mer analyses; however, the extent of difference between CpG and non-CpG groups varied. Table 1 summarizes the linear regression for each n-mer analysis. As in humans [19], we observed a trend for the slope values in the CpG group to become larger when \( n \) was smaller. However, for each n-mer analysis, the slope value in the CpG group was consistently larger in mice than in humans; this difference was conversely observed in the non-CpG group. For example, in the CpG group, the slope value in 4-mer analysis was 3.88 in mice and 2.61 in humans. Note that the slope values in the non-CpG group were within a more limited range (0.45 to 0.62) in mice than in humans (0.48 to 0.92).

**Short sequence analysis by SNP categories**

The above distribution patterns in the CpG group reflected the CpG effect in the mouse genome. Since the hypermutability of CpG in the mouse genome does not affect the transversional substitutions (i.e., A/C, G/T, A/T, and C/G), we examined the distribution of n-mers for each category of SNPs. The distribution of the 6-mers, plotted by the frequency in the local sequences of each SNP category versus the mouse genome sequences, is displayed in Supplementary Fig. S1. As expected, the distribution patterns for the categories A/G and C/T, as well as A/C and G/T, were essentially the same because of the symmetric base-pair characteristics of DNA. However, the distribution patterns among the others were remarkably different. Table 2 summarizes the linear regression for the 6-mers in the CpG group and the non-CpG group for each SNP category. In the CpG group, the slope value was the highest for category C/G (3.66) and the lowest (0.54) for category A/T. This sharp difference was conversely observed in the non-CpG group (i.e., the slope value was 0.52 and 1.65, respectively). Surprisingly, in the CpG group, the slope value for category C/G was even larger than that for category A/G or C/T (Table 2). This difference was not observed in the analysis of human SNPs [19]. In humans, the slope value in the CpG group was 2.36 for category C/G, less than that (2.55) for category A/G or C/T. Moreover, the slope values for categories A/C and G/T were close to 2 in the CpG group; this is nearly twice that (1.06) observed in humans. The difference among the six SNP categories in the 6-mer analysis was similarly observed in other n-mer analysis; with a trend toward a larger difference among the six categories when \( n \) became smaller. For example, in the CpG group of the tetramers, the slope value was 5.65 for category C/G, nearly eight times that (0.71) compared to category A/T (Supplementary Table S1).

### Table 1

<table>
<thead>
<tr>
<th>Size (n-mer)</th>
<th>CpG group</th>
<th>Non-CpG group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intercept (a)</td>
<td>Slope (b)</td>
</tr>
<tr>
<td>3</td>
<td>-4.5 × 10⁻³</td>
<td>7.33</td>
</tr>
<tr>
<td>4</td>
<td>1.4 × 10⁻⁴</td>
<td>3.88</td>
</tr>
<tr>
<td>5</td>
<td>4.7 × 10⁻⁵</td>
<td>3.18</td>
</tr>
<tr>
<td>6</td>
<td>1.4 × 10⁻⁵</td>
<td>2.71</td>
</tr>
<tr>
<td>7</td>
<td>4.5 × 10⁻⁶</td>
<td>2.31</td>
</tr>
<tr>
<td>8</td>
<td>1.7 × 10⁻⁶</td>
<td>1.84</td>
</tr>
</tbody>
</table>

The CpG group includes the fragments containing at least one CpG dinucleotide (labeled red in Fig. 1) and non-CpG group includes the fragments without any CpG dinucleotide (labeled blue in Fig. 1). The linear regression for these two groups is expressed by \( f_{\text{SNP}} = a + bf_{\text{genome}} \), where \( f_{\text{SNP}} \) and \( f_{\text{genome}} \) denote the frequency of each n-mer observed in the SNP local sequences and in the genome sequences.
The proportion was calculated by the observed numbers in the CpG islands over the numbers in the corresponding chromosome or genome sequences. CpG SNPs includes six possible nucleotide changes.

\[ \text{Obs}_{\text{CpG}}/\text{Exp}_{\text{CpG}} \text{ ratio} = 0.74 \text{ in the CpG island sequences, } 3.9 \text{ times that (0.19) in the mouse genome sequences. } \]

### Cpg islands in the mouse genome

A total of 20,713 Cpg islands were identified, with an average length of 1034 bp. The average GC content for the Cpg island sequences was 60.46%; this compared with 41.55% in the non-Cpg island sequences and 41.71% in the whole genome sequences we used. The average \( \text{Obs}_{\text{CpG}}/\text{Exp}_{\text{CpG}} \) ratio was 0.74 in the CpG island sequences, 3.9 times that (0.19) in the mouse genome sequences.

Table 3 shows the proportion of the predicted Cpg island sequences, measured by single nucleotides, Cpg dinucleotides, and Cpg SNPs, on each chromosome and in the mouse genome. The Cpg SNPs included six possible nucleotides changes at the Cpg site, i.e., [A/C]G, [C/G]G, [C/T]G, C[A/G], C[C/G], and C[G/T]. We found 111,777 Cpg SNPs in the mouse genome; among them, 2498 were in the Cpg islands. On average, Cpg island sequences accounted for only 0.85% of the mouse genome, less than the 1.45% in the human genome [19]. However, in contrast to the small proportion of the Cpg island sequences in the mouse genome, the Cpg dinucleotides within the Cpg islands accounted for 7.00% of the total Cpg dinucleotides in the mouse genome. Finally, approximately 2.23% of the mouse Cpg SNPs occurred in the Cpg islands. This proportion is remarkably lower than the proportion of Cpg dinucleotides, but higher than the proportion of Cpg island sequences in the mouse genome. Assuming that nucleotide changes occur randomly in a sequence, SNPs were underrepresented by approximately 3.14-fold at Cpg sites in the Cpg islands. The ratio was less than that (i.e., 3.92) in the human genome.

We next compared the proportion of each type of Cpg SNP. In the Cpg islands, the proportion of substitution [C/T]G and C[A/G] was 56.10 and 52.78%, respectively (Supplementary Table S2). This is significantly \( p \approx 0, \chi^2 \text{ test} \) lower than the corresponding values (78.24 and 78.37%) in the non-Cpg island sequences. Since the proportion of substitution [C/T]G (C[A/G] on other DNA strand) in the non-Cpg island sequences reflects the overall CpG effect in the mouse genome, the significantly lower proportion in the Cpg islands indicated a strong suppression of the \( \text{mCpG} \rightarrow \text{TpG} \) mutations in mouse Cpg islands.

#### Distribution of dinucleotides

We examined the distribution of the 16 possible dinucleotides in the mouse: (1) in the genome sequences, (2) in Cpg island sequences, and (3) at the SNP sites. The proportion of each dinucleotide varied among these three categories (Table 4). The proportion of CG dinucleotides in the mouse genome sequences was 0.83%, the lowest among all 16 dinucleotides. In contrast, the proportion of CG dinucleotides was 6.89% in the Cpg islands and 6.00% at the SNP sites. In the Cpg island sequences, the proportion of the dinucleotide CG was lower than that of CC, GG, or GC, although, as expected, all four dinucleotides had high frequency in the Cpg islands (Table 4). The range of the proportion values among the 16 dinucleotides at the SNP sites was much smaller than that in the genome sequences or in the Cpg island sequences. This pattern was further shown by the ratio of the observed and expected proportion of each dinucleotide. The ratio was only 0.19 for the Cpg dinucleotide...
in the genome, indicating its great deficiency in the mouse genome. The ratio of CpG dinucleotide at the SNP sites over genome sequences was 7.23, which indicated a strong bias for CpG dinucleotides at the polymorphic sites relative to the genome sequences.

**Discussion**

At present, the knowledge of the sequence mutability in the mouse genome is still very limited; despite this, biologists usually assume the same scenario for mice as humans. Our previous study of 0.4 million mouse SNPs first revealed that the extent of the biases at the two adjacent sites was notably less in mice than in humans [7]. In this study, we have further examined the sequence context of the mouse in the local environment of SNPs, CpG islands, and whole genome sequences. We analyzed 482,528 mouse SNPs, which represented the largest publicly available data set so far. Two distinct groups were observed for all n-mer (n = 3–8) analyses: (1) the short sequences in the CpG group were observed more frequently in the local sequences surrounding the polymorphic sites than in the genome sequences, while (2) the short sequences in the non-CpG group were observed less frequently in the sequences surrounding the polymorphic sites. This pattern, which was similarly observed in the human genome, reflected the strong CpG effect in the mammalian genome. To exclude the CpG effect, we further examined the six categories of SNPs. The results revealed that short DNA fragments in the CpG group were even more abundant in the local sequences of C/G compared to the A/G or C/T substitutions, while the opposite pattern was observed for the A/T substitution. This new finding suggests a strong preference of the sequence environment by the different categories of substitution. Further comparison of those top-ranking over-or underrepresented short sequences in the local environment of SNPs with other characteristic sequences (e.g., motifs, protein binding sites, splicing sites) may provide more insights on mutational mechanisms, natural selection, and function.

We performed the first genome-wide sequence context analysis of the genetic polymorphisms in the mouse CpG islands. The results revealed that CpG dinucleotides, which presented only \( \sim 19\% \) of the expected in the mouse genome, were approximately 7.23-fold more abundant at the polymorphic sites than in the whole genome. On the other hand, SNPs occurring at CpG sites were 3.14-fold less prevalent than what was expected in the CpG islands. These detailed results in mice extended the previous views of the hypermutability of methylated CpG dinucleotides in the non-CpG island sequences in the human genome and the suppression of the polymorphisms at CpG dinucleotides in the CpG islands [6,7,12,14].

While it is not surprising to observe similar sequence context patterns in the mouse and human genomes, some major differences existed. Considering the large size of the data analyzed, these differences may have important implications for mammalian genome sequence evolution. First, the slope values in the CpG group using the mouse data were notably larger than those in humans. For example, in the tetramer analysis, the slope value was 3.88 in mice and 2.61 in humans [19]. Second, for all six categories of SNPs, the slope values for the CpG group in mice were greater than those in humans; conversely, the slope values for the non-CpG group in mice were generally smaller than those in humans. Third, it appears that the distribution of the CpG group was separated more from that of the non-CpG group in the mouse analysis. This was observed for all SNPs and each category of SNPs. The three points above suggests that CpG-containing sequences in the mouse genome have more representation of polymorphisms. This seems to suggest further a stronger CpG effect in mice than in humans. However, this inference is inconsistent with our previous neighboring-nucleotide bias analysis [7], in which we found that the extent of the biases for nucleotide G at the 3' immediate adjacent site in mice was much less than that in humans. The reasons for these
incongruent observations are unknown to us. Fourth, SNPs occurring at the CpG sites were 7.23-fold more abundant than expected in the mouse genome; this compared with a 6.09-fold abundance in the human genome. Finally, CpG SNPs were 3.14-fold less prevalent relative to the presence of CpG dinucleotides in the CpG islands in the mouse genome; this compared with a 3.92-fold less prevalence in the human genome (or 6.84-fold in an early analysis [12]). The difference, though moderate, suggests that the extent of the suppression of methylation-dependent deamination in the mouse CpG islands may be less than that in the human genome.

The comparative results above consistently support the previous inference of a loss of ancestral CpG islands in the mouse lineage [17]. Under this model, CpG dinucleotides in some islands became de novo methylated in the mouse germ line. Because the mutation rate of $^{32}$PcG → TpG was estimated at $\sim$10-to 50-fold higher than other transitions [3], it resulted in a greater deficit of CpG dinucleotides in the mouse genome (0.19) than in the human genome (0.24), more representation of $n$-mers in the CpG group in the local sequences of SNPs, and a higher SNP/genome ratio of CpG dinucleotide (7.23 in mice versus 6.09 in humans, Table 4). So, why is the GC content in today’s mouse genome higher than that in the human genome? One possible explanation is that the CpG effect in the human non-CpG island sequences was stronger than in the mouse non-CpG island sequences, which is supported by our previous analysis [7]. This finally led to a lower GC content in the human genome. Readers may need to be cautious of these interpretations. More detailed analysis of mutation patterns (e.g., CG → TG/CA) in the CpG islands located in the orthologous regions would provide more evidence.

There are a few limitations in this study. First, because the direction of a substitution is generally unknown for the dbSNP data, we could not distinguish the patterns for the $C \rightarrow T$ (G → A) substitution from those of other substitutions. Second, we assumed an equal allele frequency for each SNP, that is, we selected: (1) non-insertion/deletion, (2) biallelic, (3) having a unique hit to the genome reference sequences, and (4) having a minimum of 7 nucleotides in each $n$-mer in the SNP local sequences. We expect this had a minor effect in our study because the average chromosomes used for the mouse SNP discovery is $\sim$2.6 (Mouse Phenome Database, http://www.jax.org/phenome/). The effect may be examined indirectly by using the human SNP data generated in the HapMap project [20]. Third, the distribution of the sequence context in the CpG islands relies on an accurate identification of CpG islands in the mouse genome. Although there is still a lack of a standard definition of a CpG island, Takai and Jones’ algorithm can effectively exclude the universal Alu repeats in the genome and identify the CpG islands that are more likely to be associated with the 5’ regions of genes [14]. This algorithm has been applied to the NCBI database [21]; therefore, we also used the algorithm to screen the CpG islands in this study. Finally, the number of SNPs we used is much smaller than that of human SNPs. It is possible that the distribution of the $n$-mers (e.g., slope values) may be slightly different, although the overall patterns are expected to be the same.

In conclusion, these analyses provide a detailed view of the sequence context patterns observed in the local environment of SNPs, CpG islands, and whole genome sequences in mice. First, in the local environment of SNPs, we found that the extent of the overrepresentation of CpG-containing short sequences in the local environment of polymorphisms in the mouse genome was stronger than that in the human genome. Second, a further examination of sequence context among the six categories of substitution uncovered greatly varied distribution patterns, indicating the strong preference of the local sequence environment by each substitution category. Third, in the CpG islands, we found that SNPs occurring at CpG sites were 3.14-fold less prevalent than expected, suggesting the suppression of the polymorphisms at CpG dinucleotides in the CpG islands. However, the extent of the suppression in the mouse CpG islands was less than that in humans. Finally, some evidence, such as stronger overrepresentation of CpG-containing short sequences in the local environment of SNPs in mice than in humans, appeared to support the loss of CpG islands model in the mouse lineage.

Materials and methods

Mouse SNP and sequence data

We downloaded the mouse 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 581,577 reference SNPs were available across the mouse genome; among them, 92.0% were validated. The mouse genome sequence build 33 was released in June 2004. We downloaded the contig sequences in the mouse genome from http://genome.ucsc.edu/ on October 15, 2004.

A Perl script was developed to parse the annotation information in the dbSNP XML files. Only those SNPs that satisfied the following criteria were selected: (1) non-insertion/deletion, (2) biallelic, (3) having a unique hit to the genome reference sequences, and (4) having a minimum of 7 nucleotides in each flanking sequence of a SNP. This resulted in 482,528 SNPs that we used for further analysis. For each SNP, we extracted the following annotation information from XML data files: (1) alleles (e.g., A/G), (2) 5’ and 3’ flanking sequences, (3) chromosome information, (4) contig ID, and (5) contig location.

Identification of CpG islands and their SNPs

Identification of the CpG islands was carried out using the CpG island searcher program (CpgIs130, http://cpgislands.usc.edu/) [22]. The CpgI130 software implemented Takai and Jones’ algorithms [14] for identifying CpG islands from genomic sequences without filtering the repeats. We used stringent search criteria of GC content $\geq$55%, Obs$_{CpG}$/Exp$_{CpG}$ $\geq$ 0.65, and length $\geq$500 bp to screen CpG islands in the mouse genome sequences. The algorithm based on the criteria above can effectively exclude the Alu repeats, which typically have a sequence length of 300 bp, GC content of 53%, and Obs$_{CpG}$/Exp$_{CpG}$ ratio of 0.62 [23].

We next identified the SNPs that locate in the CpG islands by comparing the SNP locations in the contig sequences with the position index of the CpG islands. To do that, we first examined if annotations of contig position in the dbSNP database matched the version of the genome reference sequences. We randomly chose 1000 SNPs across the genome. The flanking sequences of these SNPs were used to perform a BLAST search over the genome reference sequences. According to the search results, we found that all SNPs matched the annotation information. A total of 6457 SNPs were then identified in the CpG islands.

Short sequence analysis in the genome and SNP local sequences

A program package (available upon request) was used to compute the observed and expected frequency values for single nucleotides (i.e., A, C, G, and T), dinucleotides (e.g., CG), and other short sequences ($n$-mers) in the mouse genome, SNP flanking sequences, and CpG island sequences,
respectively. This program package was written in Java and was run in a Dell workstation in the Linux environment.

For the \( n \)-mer analysis, a modified method of Tomso and Bell [12] was used to obtain the frequency of each \( n \)-mer surrounding the polymorphic sites and in the mouse genome reference sequences. First, for each SNP, we extracted the local sequence, which was defined by \( 2 \times n - 1 \) nucleotides symmetrically overlapping the polymorphic site. For example, in the 6-mer analysis, we obtained 5 nucleotides from the 5' flanking sequences, 5 nucleotides from the 3' flanking sequences, and the nucleotide at the polymorphic site. Next, a sliding window of size \( n \) was applied to count the occurrence of \( n \)-mers in the SNP local sequences and genome sequences, respectively. The frequency of each \( n \)-mer was calculated by the number of its counts divided by the total number of counts of \( n \)-mers. The frequencies of \( n \)-mers were also obtained for each category of SNPs using the above approach.

Acknowledgments

We are indebted to the people who collected, annotated, and managed the data for public access, especially the NCBI dbSNP team. This project was supported by the A.D. Williams Trust Funds and 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.2005.09.012.

References

[19] Z. Zhao, F. Zhang, Sequence context analysis of 8.2 million single nucleotide polymorphisms in the human genome, Gene in press.