The Complementary Neighborhood Patterns and Methylation-to-Mutation Likelihood Structures of 15,110 Single-Nucleotide Polymorphisms in the Bovine Genome

Zhihua Jiang^{*,1} Xiao-Lin Wu,² Ming Zhang, Jennifer J. Michal and Raymond W. Wright Jr.

Department of Animal Sciences, Washington State University, Pullman, Washington 99164-6351

Manuscript received April 30, 2008
Accepted for publication June 19, 2008

ABSTRACT

Bayesian analysis was performed to examine the single-nucleotide polymorphism (SNPs) neighborhood patterns in cattle using 15,110 SNPs, each with a flanking sequence of 500 bp. Our analysis confirmed three well-known features reported in plants and/or other animals: (1) the transition is the most abundant type of SNPs, accounting for 69.8% in cattle; (2) the transversion occurs most frequently (38.56%) in cattle when the A + T content equals two at their immediate adjacent sites; and (3) C ↔ T and A ↔ G transitions have reverse complementary neighborhood patterns and so do A ↔ C and G ↔ T transversions. Our study also revealed several novel SNP neighborhood patterns that have not been reported previously. First, cattle and humans share an overall SNP pattern, indicating a common mutation system in mammals. Second, unlike C ↔ T/A ↔ G and A ↔ C/G ↔ T, the true neighborhood patterns for A ↔ T and C ↔ G might remain mysterious because the sense and antisense sequences flanking these mutations are not actually recognizable. Third, among the reclassified four types of SNPs, the neighborhood ratio between A + T and G + C was quite different. The ratio was lowest for C ↔ G, but increased for C ↔ T/A ↔ G, further for A ↔ C/G ↔ T, and the most for A ↔ T. Fourth, when two immediate adjacent sites provide structures for CpG, it significantly increased transitions compared to the structures without the CpG. Finally, unequal occurrence between A ↔ G and C ↔ T in five paired neighboring structures indicates that the methylation-induced deamination reactions were responsible for ~20% of total transitions. In addition, conversion can occur at both CpG sites and non-CpG sites. Our study provides new insights into understanding molecular mechanisms of mutations and genome evolution.

SINGLE-NUCLEOTIDE polymorphisms (SNPs) represent the most abundant form of genetic variation in both plant and animal genomes. For example, SNPs occur every 100–300 bases along the 3-billion-base human genome and represent ~90% of all genetic variation (<http://www.ornl.gov/>). Nucleotide differences in the promoter regions of protein-encoding genes may cause gains/losses of specific regulatory binding sites and result in differential regulation of transcription (JIANG *et al.* 2007). SNPs at intron/exon boundaries may influence the conserved “GU–AG” motifs and modify the resulting polypeptide. Even synonymous SNPs, disregarded in many studies on the basis of the assumption that these are silent, can alter RNA secondary structure (WANG *et al.* 2005) and affect protein conformation and function (SAUNA *et al.* 2007). Therefore, SNPs are important markers that link genes to normal physiological changes, diseases, and responses to pathogens, chemicals, drugs, vaccines, and other agents in humans (RILEY *et al.* 2000; KIM and MISRA 2007). The study of SNPs is also

important in crop and livestock breeding programs. This information can be used to localize genes that affect quantitative traits, identify chromosomal regions under selection, study population history, and characterize/manage genetic resources and diversity (RAFALSKI 2002; DU *et al.* 2007).

Chemically, nucleotides can be grouped into purines (A, G) and pyrimidines (C, T). SNPs within the groups are called transitions and those between the groups are called transversions. Therefore, there are two possible transitions (C ↔ T and A ↔ G) and four possible transversions (A ↔ C, G ↔ T, A ↔ T, and C ↔ G) if we do not consider the directions of mutations. Overall, transition mutations occur most frequently in both plant and animal genomes, for example, ranging from 60% of all mutations in maize (MORTON *et al.* 2006) to 68% in the mouse (ZHANG and ZHAO 2004). It has been widely believed that the hypermutability effects of CpG dinucleotide sites contribute significantly to the increased rate of transitions in both plant and animal genomes as a result of deamination of methylated cytosines (DUNCAN and MILLER 1980). However, the A + T content of the two immediate sites adjacent to the mutation sites is associated with an increased rate of transversion in the

¹Corresponding author: Department of Animal Sciences, Washington State University, Pullman, WA 99164-6351. E-mail: jiangz@wsu.edu

²Present address: Department of Dairy Science, University of Wisconsin, Madison, WI 53706.

nuclear genomes of human, mouse, and Arabidopsis (ZHANG and ZHAO 2004) as well as in the chloroplast genomes of rice and maize (MORTON 1995). Interestingly, the two adjacent nucleotide sites that flank the mutations also show the largest biases compared to the genomewide and chromosome-specific average (ZHAO and BOERWINKLE 2002; ZHANG and ZHAO 2004).

The Bovine Genome Sequencing Project, led by a team at Baylor College of Medicine's Human Genome Sequencing Center (<http://www.hgsc.bcm.tmc.edu/projects/bovine/>), began in 2003. In August 2006, a 7.15× mixed assembly of the draft bovine genome combining whole-genome shotgun (WGS) sequence with BAC sequence was released to the public database at the National Center for Biotechnology Information. Just recently, the project also released 15,110 well-characterized SNPs derived from both dairy and beef cattle. In our study, we first examined genome basics and SNP basics in cattle. We then focused on how SNPs influence their neighborhood patterns of nucleotides and how neighboring-nucleotide compositions affect SNPs in the bovine genome. We confirmed some unique neighborhood features observed in other mammalian species as well as in some plants. We also discovered several novel features about SNPs that have not been reported previously.

MATERIALS AND METHODS

SNPs source, WGS reads, and quality control: SNPs were downloaded from <ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Btaurus/snp/Btau20040927/bovine-snp.txt>. These are high-quality SNPs from the first round of discovery using WGS reads from selected cattle breeds. A total of 15,267 records were released, but there are 157 duplicate calls, thus yielding 15,110 unique SNPs. Sequencing reads were generated from random shotgun libraries from Holstein, Angus, Brahman, Limousin, and Jersey breeds. Reads were compared to the 3x *Bos taurus* genome assembly (Btau20040927), using BLAST with $e=50$ expected cutoff. The criteria for selecting a read for SNP analysis can be viewed in detail at <ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Btaurus/snp/Btau20040927/README>. Basically the procedures for selection of bovine SNPs followed the NQS method developed for initial mapping of human SNPs but all the thresholds were more stringent (K. C. WORLEY, personal communication).

SNPs, flanking nucleotides, and reverse complement: Only these 15,110 unique SNPs were used in this study. For each entry, the type of SNP was well labeled and included 250 bp of 5' flanking sequence and 250 bp of 3' flanking sequence. Basically, we put a SNP in the center of a sequence and labeled the nucleotides at the 5' and the 3' side with consecutive negative and positive numbers toward the end of each flanking sequence. A reverse complementary sequence was made, when necessary, on the basis of the formation of a double-stranded structure by matching base pairs with the target sequence plus a reorientation.

Estimation of nucleotide frequencies using a conjugate beta-binomial hierarchical model: Let x_i be the number of a base (say A, T, C, or G) at a site and N_i be the number of all bases at the site. Here we assume that x_i follows a binomial distribution

$$p(x_i | \theta_i, N_i) \sim \text{dbin}(\theta_i, N_i), \quad (1)$$

where θ_i is the frequency of the base at this site. In the Bayesian specification, a conjugate beta prior is assumed for θ_i ,

$$p(\theta_i | a, b) = \text{beta}(a, b), \quad (2)$$

where hyperparameters a and b are assumed to be known (*i.e.*, their values are arbitrarily given as 1.0 and 3.0, respectively). The posterior distribution of θ_i is also a beta:

$$\begin{aligned} p(\theta_i | x_i, N_i) &\propto p(x_i | \theta_i, N_i) p(\theta_i | a, b) \\ &= \text{beta}(a + x_i, b + N_i - x_i). \end{aligned} \quad (3)$$

In preliminary runs, convergence of iterative simulations was monitored on the basis of multiple sequences with overdispersed starting points, which is described in detail in GELMAN *et al.* (2004). On the basis of the convergence diagnosis results of θ_i , we decided that a single chain consisting of 100,000 iterations be used with a burn-in of 1000 iterations (*i.e.*, samples are discarded during the burn-in period). The autocorrelation function (BOX and JENKINS 1976) was used to measure nonrandomness (or correlation) of sequential draws of a parameter from the conditional distribution (*e.g.*, EBERLY and CARLIN 2000). Empirically, we observed that lag 10 autocorrelations were <0.01 for all chains of θ_i . Thus, it was decided that posterior samples were thinned at every 10 iterations and saved for posterior inference. The 2.5-, 50-, and 97.5-quantiles were calculated for nucleotide frequencies and/or their ratios, using the corresponding posterior samples. The 95% posterior interval encompasses 95% of the posterior distribution, which is a Bayesian formulation analogous to the 95% confidence interval. The above model was implemented using WinBUGS Version 1.4.3 (<http://www.mrc-bsu.cam.ac.uk/bugs>).

Bayesian model selection for transitional patterns: Bayesian model selection is used to compare the two possible configurations of transitional patterns (six SNP types *vs.* four SNP types). Denote $\mathbf{y}$ as a vector containing all observation data. Let M_1 be the model assuming six transitional patterns and M_2 be the model assuming four transitional patterns. Essentially, M_1 is a beta-binomial mixture model with six transitional patterns (C $\leftrightarrow$ T, A $\leftrightarrow$ G, A $\leftrightarrow$ C, G $\leftrightarrow$ T, A $\leftrightarrow$ T, and C $\leftrightarrow$ G) whereas M_2 is a beta-binomial mixture model with four transitional patterns (C $\leftrightarrow$ T/A $\leftrightarrow$ G, A $\leftrightarrow$ C/G $\leftrightarrow$ T, A $\leftrightarrow$ T, and C $\leftrightarrow$ G). Note that data of complementary types are combined in M_2 . To simulate a mixture, portions of each binomial component were known and were equal to the observed portions of each transitional type, and the sum of nucleotide frequencies is fixed to be 1.

The Bayes factor in support of M_2 over M_1 is the ratio between the posterior and the prior odds of the two models (GOODMAN, 1999):

$$B_{21} = \frac{p(\mathbf{y} | M_2)}{p(\mathbf{y} | M_1)} = \frac{p(M_2 | \mathbf{y})/p(M_1 | \mathbf{y})}{M_2/M_1}.$$

Conveniently, the Bayesian information criterion (BIC) can be computed, which is a large sample approximation to twice the logarithm of the Bayes factor (SCHWARZ 1978; RAFTERY 1995). Comparing two models, we observe that

$$\begin{aligned} 2 \log B_{21} &= 2 \log \left(\frac{p(\mathbf{y} | M_2)}{p(\mathbf{y} | M_1)} \right) \\ &= 2 \log \left(\frac{p(\mathbf{y} | M_0)/p(\mathbf{y} | M_1)}{p(\mathbf{y} | M_0)/p(\mathbf{y} | M_2)} \right) \\ &= 2 \log B_{01} - 2 \log B_{02} \approx \text{BIC}_1 - \text{BIC}_2, \end{aligned}$$

FIGURE 1.—The same overall neighborhood patterns flanking all SNPs in cattle (left) and human (right).

where M_0 is a baseline model. Let $\Delta_{\text{BIC}} = \text{BIC}_1 - \text{BIC}_2$. We interpret the BIC difference criteria following RAFTERY (1995):

Weak evidence: $0 \leq \Delta_{\text{BIC}} < 2$

Positive evidence: $2 \leq \Delta_{\text{BIC}} < 6$

Strong evidence: $6 \leq \Delta_{\text{BIC}} < 10$

Very strong evidence: $10 \leq \Delta_{\text{BIC}}$.

Bayesian model selection using BIC was conducted using WinBUGS Version 1.4.3 (<http://www.mrc-bsu.cam.ac.uk/bugs>).

SNP types analysis and chi-square test: In addition to the Bayesian analysis described above, chi-square tests of independence and goodness-of-fit were also used in the data analysis. The former was used to test the hypothesis whether the frequency within types of SNPs is what would be expected, given these marginal N 's. The latter was used to test the hypothesis whether the total number N is distributed evenly between different types of SNPs. The analysis was performed using a web tool called "Calculation for the Chi-Square Test" at <http://www.people.ku.edu/~preacher/chisq/chisq.htm>.

RESULTS

SNP basics and overall neighborhood patterns: As indicated above, a total of 15,110 SNPs were sampled from the bovine genome and used in this study. These SNPs are all biallelic, including 5233 $A \leftrightarrow G$, 5308 $C \leftrightarrow T$, 1232 $A \leftrightarrow C$, 1249 $G \leftrightarrow T$, 928 $A \leftrightarrow T$, and 1160 $C \leftrightarrow G$ substitutions, respectively. Obviously, the number of $A \leftrightarrow G$ substitutions was almost equal to that of $C \leftrightarrow T$ substitutions ($\chi^2 = 0.53$, $P = 0.4651$). As well, the number of $A \leftrightarrow C$ substitutions was also almost the same as that of $G \leftrightarrow T$ substitutions ($\chi^2 = 0.12$, $P = 0.7329$). Both $A \leftrightarrow G$ (34.63%) and $C \leftrightarrow T$ (35.13%) substitutions were most abundant, while each of the other types accounted for <10% ($A \leftrightarrow C$, 8.15%; $G \leftrightarrow T$, 8.27%; $A \leftrightarrow T$, 6.14%; and $C \leftrightarrow G$, 7.67%, respectively) in the bovine genome. We estimated that the ratio of transition over transversion is 2.307, with a 95% posterior interval ranging from 2.227 to 2.388 in cattle. Using Bayesian

analysis, the posterior statistics (mean, standard deviation, and 2.5-, 50-, and 97.5-quantiles) and Markov chain errors of nucleotide frequencies at different flanking sites (*i.e.*, from -250 to $+250$ bp) for all SNPs as well as for all six different types of SNPs are summarized in supplemental Tables 1–7.

There are 16 combinations of nucleotides immediately adjacent to the 5' and 3' sites of mutations, which can be further classified into three groups on the basis of their A + T content: group 0 (A + T content of 0), 5' $C \leftrightarrow C$ 3', 5' $C \leftrightarrow G$ 3', 5' $G \leftrightarrow C$ 3', and 5' $G \leftrightarrow G$ 3'; group 1 (A + T content of 1), 5' $A \leftrightarrow C$ 3', 5' $A \leftrightarrow G$ 3', 5' $C \leftrightarrow A$ 3', 5' $C \leftrightarrow T$ 3', 5' $G \leftrightarrow A$ 3', 5' $G \leftrightarrow T$ 3', 5' $T \leftrightarrow C$ 3', and 5' $T \leftrightarrow G$ 3'; and group 2 (A + T content of 2), 5' $A \leftrightarrow A$ 3', 5' $A \leftrightarrow T$ 3', 5' $T \leftrightarrow A$ 3', and 5' $T \leftrightarrow T$ 3' ($\leftrightarrow$ denotes any substitutions: $C \leftrightarrow T$, $A \leftrightarrow G$, $A \leftrightarrow C$, $G \leftrightarrow T$, $A \leftrightarrow T$, and $C \leftrightarrow G$). We observed that group 2 had the highest transversion frequency of 38.56% compared to both group 0 (25.54%) ($\chi^2 = 149.462$, $P = 0.0000$) and group 1 (26.84%) ($\chi^2 = 179.969$, $P = 0.0000$). On the basis of 500 bp of sequence flanking each of these 15,110 SNPs, we estimated that the average frequencies for nucleotides A, T, C, and G were 0.2859, 0.2883, 0.2142, and 0.2116, respectively, in the bovine genome.

The nucleotide frequencies for each of 10 sites at 5' and 3' sites of SNPs are shown in Figure 1. Figure 1, left, depicts posterior means of nucleotide frequencies obtained using the Bayesian analysis for 15,110 SNPs in cattle, while Figure 1, right, was derived from ZHAO and BOERWINKLE's (2002) study of 2,576,903 SNPs in humans. Interestingly, both species shared the same adjacent neighborhood scenarios flanking the SNPs. Basically, the frequency of nucleotide C gradually increased from a genome average at the -4 site to the highest point at the -1 site, but significantly dropped to a level below the genome average at the $+1$ site. The nucleotide G had a frequency below the genome average at the -1 site, but reached the highest point at the $+1$ site and then gradually decreased to the genome average level at the $+4$ site. Both nucleotides A and T had a reasonable drop in frequency to a level below the genome average, with the former at the $+1$ site and the latter at the -1 site.

(Figure 1). In brief, the overall adjacent scenarios flanking the SNPs were complementary between nucleotides C and G and between nucleotides A and T, respectively.

SNP type-specific neighborhood patterns: The neighborhood patterns of nucleotide distributions for six types of SNPs are illustrated in Figure 2, A–F. In general, posterior means of nucleotide frequencies were in agreement with their counterparts from their frequency averages. Nevertheless, the Bayesian analysis can visually show distributions of nucleotide frequencies in terms of accumulative distributions (*e.g.*, 2.5-, 50-, and 97.5-quantiles) and facilitate a better comparison of their difference in terms of their distributions than its frequency counterpart. The nucleotide frequencies in the neighborhoods of A ↔ G and A ↔ C substitutions were estimated on the basis of their reverse complementary pairs. As a result, we observed that both C ↔ T and A ↔ G transitions had a highly similar neighborhood pattern (Figure 2, A and B) and so did both A ↔ C and G ↔ T transversions (Figure 2, C and D) in terms of nucleotide distributions. This means that the former two transitions and the latter two transversions each shared a reverse complementary neighborhood pattern when the frequencies were calculated on the basis of their original data without the reverse complements. Therefore, we can regroup the current six types of SNPs into four types: C ↔ T/A ↔ G, A ↔ C/G ↔ T, A ↔ T, and C ↔ G. The Bayesian model selection via the BIC further confirmed that the pattern with six types of SNPs is essentially equivalent to that with four types of SNPs (−226.59 *vs.* −228.62). In addition, the model even provided slightly positive evidence ($\Delta_{\text{BIC}} = 2.02$) in favor of the four-transitional pattern over the six-transitional pattern because BIC tends to favor a simpler, more parsimonious model by penalizing the complex model with more parameters.

Among these four types of SNPs, C ↔ T/A ↔ G, A ↔ C/G ↔ T, A ↔ T, and C ↔ G, the neighborhood patterns of nucleotide distributions were, however, quite different from each other (Figure 2). For example, at the −1 site, the trend of nucleotide dynamics was A > T > C > G for the combined C ↔ T/A ↔ G transitions (Figure 2, A and B), T > A > G > C for the combined A ↔ C/G ↔ T transversions (Figure 2, C and D), T > A > C > G for the A ↔ T transversions (Figure 2E), and A > T > G > C for the C ↔ G transversions (Figure 2F), respectively. However, at the +1 site, the trend was G > A > T > C for the combined C ↔ T/A ↔ G transitions (Figure 2, A and B), T > A > G > C for the combined A ↔ C/G ↔ T transversions (Figure 2, C and D), A > T > G > C for the A ↔ T transversions (Figure 2E), and T > A > C > G for the C ↔ G transversions (Figure 2F), respectively.

For the C ↔ T/A ↔ G transitions (Figure 2, A and B), nucleotide A had a high average frequency of 0.3548 (with a 95% posterior interval of 0.3473–0.3624) at the −1 site, but significantly dropped to 0.2358 (0.2292–

0.2426) at the +1 site. In contrast, nucleotide G had a high frequency of 0.3747 ranging from 0.3670 to 0.3825 at the +1 site, but significantly decreased to 0.1855 (0.1794–0.1919) at the −1 site. Nucleotide C also dropped its frequency from 0.2309 with a 95% posterior interval of 0.2243–0.2377 at the −1 site to 0.1657 (0.1599–0.1716) at the +1 site, while nucleotide T kept its frequency relatively consistent, being 0.2289 (0.2223–0.2357) at the −1 site and 0.2239 (0.2206–0.2305) at the +1 site (supplemental Tables 2 and 3).

For the A ↔ C/G ↔ T transversions (Figure 2, C and D), both nucleotides A and T had a frequency drop from the −1 site to the +1 site, but being relatively small for the former nucleotide (0.2763 *vs.* 0.2607 with 95% posterior intervals: 0.2692–0.2834 *vs.* 0.2539–0.2677) while relatively large for the latter nucleotide (0.3460 *vs.* 0.2753 with 95% posterior intervals: 0.3385–0.3536 *vs.* 0.2683–0.2824). However, the frequency increased from 0.1799 (0.1720–0.1840) at the −1 site to 0.2157 (0.2093–0.2223) at the +1 site for nucleotide C and from 0.1999 (0.1937–0.2063) at the −1 site to 0.2484 (0.2416–0.2553) at the +1 site for nucleotide G, respectively (supplemental Tables 4 and 5).

Both nucleotides A and G increased their frequencies from the −1 site to the +1 site, 0.2983 (0.2910–0.3056) *vs.* 0.3605 (0.3576–0.3727) for A and 0.1550 (0.1494–0.1608) *vs.* 0.1953 (0.1890–0.2016) for G when they flank A ↔ T substitutions (Figure 2E and supplemental Table 6), but decreased their frequencies from the −1 site to the +1 site, 0.3410 (0.3335–0.3485) *vs.* 0.2772 (0.2702–0.2844) for A and 0.1981 (0.1918–0.2045) *vs.* 0.1634 (0.1575–0.1639) for G when they flank C ↔ G substitutions (Figure 2F and supplemental Table 7). Adversely, both nucleotides C and T decreased their frequencies from the −1 site to the +1 site, 0.2017 (0.1954–0.2082) *vs.* 0.1380 (0.1326–0.1436) for C and 0.3449 (0.3374–0.3524) *vs.* 0.3015 (0.2943–0.3089) for T when they flank A ↔ T substitutions (Figure 2E and supplemental Table 6), but increased their frequencies from the −1 site to the +1 site, 0.1862 (0.1801–0.1924) *vs.* 0.2126 (0.2062–0.2191) for C and 0.2738 (0.2675–0.2817) *vs.* 0.3469 (0.3394–0.3543) for T when they flank C ↔ G substitutions (Figure 2F and supplemental Table 7).

The overall trends of frequencies for nucleotides A, T, C, and G flanking these four types of SNPs, C ↔ T/A ↔ G, A ↔ C/G ↔ T, A ↔ T, and C ↔ G, were also different (Figure 2). On the basis of the genome averages of nucleotide frequencies described above, we estimated that the genomewide mean ratio of A + T over C + G averaged 1.349 with a 95% posterior interval of 1.344–~1.356. Within the 50 bp of proximal sequences (25 bp from the 5' side and 25 bp from the 3' side) flanking the SNPs, the means (95% posterior intervals) of the ratios were 1.230 (1.224–~1.236) for the C ↔ G substitutions, 1.270 (1.264–~1.276) for the C ↔ T/A ↔ G substitutions, 1.409 (1.403–~1.416) for the A ↔ C/G ↔ T

FIGURE 2.—The neighborhood patterns of nucleotide distributions flanking the C ↔ T (A), A ↔ G (B), G ↔ T (C), A ↔ C (D), A ↔ T (E), and C ↔ G (F) substitutions. Solid lines represent medians of nucleotide frequencies, and top and bottom dotted lines are the corresponding 2.5- and 97.5-quantiles.

FIGURE 3.—Effects of adjacent neighborhood structures on transitional SNP frequencies.

substitutions, and 1.721 (1.712–~1.726) for the A ↔ T substitutions, respectively. As shown in Figure 2E, the highest ratio of 1.72 for A ↔ T substitutions was obviously due to the fact that the frequencies of nucleotides A and T were above, but those of nucleotides C and G were below the genome averages for most sites. However, the lowest ratio of 1.23 for the C ↔ G substitutions was certainly caused by the low frequencies of A and T and the high frequencies of C and G for most sites (Figure 2F). Therefore, A ↔ T substitutions were more abundantly associated with the A + T-rich regions, while the C ↔ G substitutions most frequently occurred in G + C-rich regions.

Neighboring nucleotide structure-associated transitional patterns: As A ↔ G and C ↔ T substitutions were the most abundant forms of SNPs in plant and animal genomes, we further examined how neighboring nucleotide combinations affect their frequencies in the bovine genome. First, we classified these 16 nucleotide combinations for two immediate adjacent sites (labeled as –1 and +1 sites) into two categories: combinations with impossible formation of CpG structures and combinations with possible formation of CpG structures. The former combinations are 5' T⊗A 3', 5' A⊗A 3', 5' T⊗T 3', 5' G⊗C 3', 5' G⊗A 3', 5' T⊗C 3', 5' A⊗T 3', 5' A⊗C 3', and 5' G⊗T 3', while the latter combinations consist of 5' C⊗C 3', 5' G⊗G 3', 5' C⊗A 3', 5' T⊗G 3', 5' A⊗G 3', 5' C⊗T 3', and 5' C⊗G 3'. As illustrated in Figure 3, all combinations that could possibly form CpG structures had higher frequencies of transition SNPs (ranging from 70.13 to 82.86%) than combinations that would not result in formation of CpG structures (ranging from 54.17 to 69.52%) ($P = 0.0000$). In other words, when adjacent nucleotides do not contribute to formation of any potential CpG sites, such combinations increase the occurrence of transversional SNPs.

On the other hand, these 16 nucleotide combinations for two immediate adjacent sites (labeled as –1 and +1

sites) can also be grouped into two categories: no-paired or self-reverse-complementary combinations and paired or reciprocal reverse-complementary combinations. The no-paired group involves 5' A⊗T 3', 5' T⊗A 3', 5' C⊗G 3', and 5' G⊗C 3', while the paired group contains 5' A⊗A 3'–5' T⊗T 3', 5' C⊗C 3'–5' G⊗G 3', 5' A⊗C 3'–5' G⊗G 3', 5' A⊗G 3'–5' C⊗T 3', 5' G⊗A 3'–5' T⊗C 3', and 5' T⊗G 3'–5' C⊗A 3' pairs, respectively. As shown in Table 1, A ↔ G and C ↔ T substitutions occurred equally in all four self-reverse-complementary combinations ($P = 0.1792$ – 0.7925), even when 5' C⊗G 3' had the structure to form CpG dinucleotide sites.

However, among six pairs of reciprocal reverse complementary combinations, only one pair (5' A⊗C 3'–5' G⊗T 3') had an equal occurrence between A ↔ G and C ↔ T transitions ($P = 0.7127$ – 0.8891). Obviously, there is no possibility for this pair to form a CpG site. The remaining paired combinations showed an unequal occurrence between A ↔ G and C ↔ T transitions ($P < 0.05$) (Table 1). When the transitions were constructed in the forms of 5' A⊗A 3', 5' G⊗G 3', 5' A⊗G 3', 5' T⊗C 3', and 5' T⊗G 3', the C ↔ T substitutions were 1.41- to 2.67-fold higher than the A ↔ G substitutions. However, the A ↔ G substitutions were 1.26- to 2.80-fold higher than the C ↔ T substitutions in the forms of 5' T⊗T 3', 5' C⊗C 3', 5' C⊗T 3', 5' G⊗A 3', and 5' C⊗A 3', reflecting a complementary strand symmetry feature to 5' A⊗A 3', 5' G⊗G 3', 5' A⊗G 3', 5' T⊗C 3', and 5' T⊗G 3', respectively (Table 1).

The unequal occurrence between A ↔ G and C ↔ T transitions in these 10 neighboring structures (5' A⊗A 3', 5' T⊗T 3', 5' C⊗C 3', 5' G⊗G 3', 5' A⊗G 3', 5' C⊗T 3', 5' G⊗A 3', 5' T⊗C 3', 5' T⊗G 3', and 5' C⊗A 3') gave us a unique opportunity to partition how many are naturally occurring transitions and how many are caused by methylation conversions. Basically, we assume that the naturally occurring C ↔ T and A ↔ G substitutions remained equal in these combinations, but an excess of C ↔ T (A ↔ G) to A ↔ G (C ↔ T) was the result of the conversion of methylated cytosine to thymine by deamination events. Under such an assumption, we estimated that deamination reactions were responsible for only 20.25% (2135/10,541) of transitions in the bovine genome. The conversion rate was highest (46.41%, 899/1937) in the complementary combination of 5' A⊗G 3'–5' C⊗T 3', followed by 5' C⊗C 3'–5' G⊗G 3' (32.16%, 367/1141), 5' A⊗A 3'–5' T⊗T 3' (29.64%, 466/1572), and 5' C⊗A 3'–5' T⊗G 3' (20.29%, 256/1262), and 5' G⊗A 3'–5' T⊗C 3' had the lowest rate of 14.54% (147/1011) (Table 1). These data also indicated that the conversion of methylated cytosine to thymine by deamination events occurred not only in CpG sites (14.44%, 1522/10,541, genomewide), but also in CpA (4.42%, 466/10,541) and CpC sites (1.39%, 147/10,541), respectively (Table 1).

TABLE 1
Effects of neighboring nucleotide pairs on A ↔ G and C ↔ T substitutions

Nucleotide pairs	Total	A ↔ G		C ↔ T		χ ²	P	CpG
		N	%	N	%			
Self-reverse-complementary combinations								
5' A⊗T 3'	1234	409	33.14	417	33.79	0.08	0.7807	No
5' T⊗A 3'	960	263	27.39	257	26.77	0.07	0.7925	No
5' C⊗G 3'	1237	491	39.69	534	43.16	1.80	0.1792	Yes
5' G⊗C 3'	494	164	33.20	148	29.96	0.82	0.3650	No
Reciprocal-reverse-complementary combination pairs								
5' A⊗A 3'	1242	263	21.18	497	40.02	72.05	0.0000	No
5' T⊗T 3'	1313	522	39.76	290	22.09	66.29	0.0000	No
5' C⊗C 3'	807	385	47.71	181	22.43	73.53	0.0000	Yes
5' G⊗G 3'	790	206	26.08	369	46.71	46.21	0.0000	Yes
5' A⊗C 3'	686	232	33.82	240	34.99	0.14	0.7127	No
5' G⊗T 3'	666	230	34.53	233	34.98	0.02	0.8891	No
5' A⊗G 3'	1243	264	21.24	705	56.72	200.70	0.0000	Yes
5' C⊗T 3'	1214	713	58.73	255	21.00	216.70	0.0000	Yes
5' G⊗A 3'	727	258	35.48	205	28.20	6.07	0.0138	No
5' T⊗C 3'	818	227	27.75	321	39.24	16.12	0.0001	No
5' C⊗A 3'	808	350	43.32	247	30.57	17.77	0.0000	Yes
5' T⊗G 3'	871	256	29.39	409	46.96	35.20	0.0000	Yes

DISCUSSION

Genomes are A + T rich, but SNPs are C ↔ T/A ↔ G rich: The completion of many whole-genome sequencing projects in both plants and animals provides ultimate resources for us to discover unique features related to genome structure, function, and evolution. Genomes are made of four nucleotides A, T, C, and G with an equal frequency between adenine and thymine as well as between cytosine and guanine, which is known as so-called Chargaff's first parity rule. For example, in humans, these four nucleotides were distributed as 29.55% A, 29.54% T, 20.44% C, and 20.46% G among a total of 2.86×10^9 bases (ZHAO and BOERWINKLE 2002), ranging from 26.05% A, 25.98% T, 23.98% C, and 23.97% G on chromosome 22 to 30.98% A, 31.18% T, 18.93% C, and 18.89% G on the Y chromosome (YAMAGISHI and SHIMABUKURO 2008). The mouse genome had 29.12% A, 29.14% T, 20.87% C, and 20.87% G in a total of 2.70×10^9 bases (ZHANG and ZHAO 2004). In our study, we estimated that nucleotides A, T, C, and G were 28.59, 28.83, 21.42, and 21.16% in the bovine genome, which are consistent with those observed in mouse and human genomes. Overall, the mammalian genomes are A + T rich, being ~1.5-fold higher than the G + C content.

Under such an A + T-rich environment, the transition substitutions (C ↔ T/A ↔ G) accounted for roughly two-thirds (65.6%) of the total mutations in the human genome (ZHAO and BOERWINKLE 2002). This number further increased to 68.13% in mouse (ZHANG and ZHAO,

2004) and 69.76% in cattle (this study), but decreased to 58% in Atlantic salmon among the coding SNPs derived from expressed sequences (HAYES *et al.* 2007). Nevertheless, transition is the most abundant form of SNP in both plants and animals. In addition, both C ↔ T and A ↔ G substitutions occur equally, for example, being 32.81 and 32.77% in human (ZHAO and BOERWINKLE 2002), 33.93 and 34.10% in mouse (ZHANG and ZHAO 2004), 35.13 and 34.63% in cattle (this study), and 29 and 29% in Atlantic salmon (HAYES *et al.* 2007).

Among four types of transversions, A ↔ C and G ↔ T also occurred equally, for example, being 8.69 and 8.74% in human, 8.63 and 8.63% in mouse, 8.15 and 8.27% in cattle (our study), and 11 and 13% in Atlantic salmon (HAYES *et al.* 2007). However, the A ↔ T transversions were the least frequent in humans (7.42%) and cattle (6.14%), while the C ↔ G substitutions were the least frequent in mouse (6.37%) and Atlantic salmon (8%) (HAYES *et al.* 2007). Interestingly, when two immediate adjacent sites flanking the SNPs were A + T rich (A + T content of 2, in particular), the frequency of transversional mutations significantly increased in animals, such as cattle (38.56%), human (38.70%), and mouse (36.50%), and in plants, such as Arabidopsis (53%) (ZHANG and ZHAO 2004) and the rice and maize chloroplast noncoding regions (57%) (MORTON 1995).

SNPs can be classified into four types: As indicated above, there are six types of SNPs, including two types of transitions (C ↔ T and A ↔ G) and four types of

transversions ($A \leftrightarrow C$, $G \leftrightarrow T$, $A \leftrightarrow T$, and $C \leftrightarrow G$) if we do not consider the directions of mutations. In this study, we observed that the neighborhood patterns of nucleotides flanking the $C \leftrightarrow T$ and $A \leftrightarrow G$ substitutions including themselves were reverse complementary to each other (Figure 2, A and B), as were those flanking $A \leftrightarrow C$ and $G \leftrightarrow T$ (Figure 2, C and D). These data indicate that a mutation might originally start on a single strand of DNA and then expand to another strand by base pairing during the DNA replication. As both $A \leftrightarrow T$ and $C \leftrightarrow G$ are self-reverse complements, it might be difficult to clearly identify their neighborhood patterns because the sense and antisense sequences flanking these mutations are not actually recognizable.

Although the nucleotide frequencies might be different among different species, the complementary patterns between $C \leftrightarrow T$ and $A \leftrightarrow G$ and between $A \leftrightarrow C$ and $G \leftrightarrow T$ also appeared in human and mouse. In human, the nucleotide ranking orders based on their frequencies were $C \gg A > T \gg G$ at the -1 site and $T \gg A \approx G > C$ at the $+1$ site for $A \leftrightarrow G$ substitutions, which were obviously complementary to $G \gg T > A \gg C$ at the $+1$ site and to $A \gg T \approx C > G$ at the -1 site for $C \leftrightarrow T$ substitutions ($\gg$ denotes greater than a 5% difference between two nucleotide frequencies, $>$ denotes a 1–5% difference, and $\approx$ less than a 1% difference) (ZHAO and BOERWINKLE 2002). The ranks $A > C \gg T > G$ at the -1 site and $A \gg T \approx C \gg G$ at the $+1$ site for $A \leftrightarrow C$ were very likely complementary to $T > G \gg A > C$ at the $+1$ site and $T \gg A > G \gg C$ at the -1 site for $G \leftrightarrow T$ transversions. The complementary ranking orders were also observed in mouse between $C \approx A > T > G$ at the -1 site for $A \leftrightarrow G$ and $G \approx T > A > C$ at the $+1$ site for $C \leftrightarrow T$, between $T \gg A > G > C$ at the $+1$ site for $A \leftrightarrow G$ and $A \gg T > C > G$ at the -1 site for $C \leftrightarrow T$, between $A \gg C \approx T > G$ at the -1 site for $A \leftrightarrow C$ and $T \gg G \approx A > C$ at the $+1$ site for $G \leftrightarrow T$, and between $A > T > C \gg G$ at the $+1$ site for $A \leftrightarrow G$ and $T > A > G \gg C$ at the -1 site for $G \leftrightarrow T$, respectively (ZHANG and ZHAO 2004).

In addition, among four types of SNPs, we observed that the overall frequencies of nucleotide distributions flanking the $C \leftrightarrow G$ substitutions were relatively close between $A + T$ and $C + G$ nucleotides (Figure 2F). The frequency differences between $A + T$ and $C + G$ nucleotides were slightly greater, for the $C \leftrightarrow T$ and $A \leftrightarrow G$ substitutions (Figure 2, A and B), and were larger for the $A \leftrightarrow C$ and $G \leftrightarrow T$ mutations (Figure 2, C and D). Finally, the frequencies for A and T nucleotides were significantly above the genome average, while the frequencies for C and G nucleotides were much below the genome average for the $A \leftrightarrow T$ substitutions (Figure 2E).

The mechanisms causing methylation to mutation seem complicated: The abundant transition polymorphisms, which accounted for approximately two-thirds of the total mutations genomewide, have been largely believed to be caused by the abundant hypermutable methylated dinucleotide 5' CpG 3' (COOPER and KRAWCZAK

1990). Approximately 60–90% of CpG sites might be methylated (BIRD 1986) and deamination of 5'-methylcytosine in such a CpG site leads its conversion to thymine. Indeed, we found that the adjacent neighboring nucleotides with potential to form CpG sites yielded 13.39% more $C \leftrightarrow T$ and $A \leftrightarrow G$ transitions than those structures without any potential to form CpG sites (Figure 3). On the other hand, deamination of 5'-methylcytosine and its subsequent conversion to thymine would not easily occur, because methylation is an important mechanism involved in maintenance of normal genome functions in mammals.

Therefore, we pursued a partition on how many $C \leftrightarrow T$ and $A \leftrightarrow G$ polymorphisms are naturally occurring transitions and how many are caused by methylation conversions. First, we estimated that the deamination events converting 5'-methylcytosine to thymine accounted for ~20% of total transitions and the remaining 80% were naturally occurring transitions in a genome. This claim is supported by similar results observed in the mouse: SNPs occurring at CpG sites were 3.14-fold less prevalent than expected (ZHAO and ZHANG 2006). These data suggest a suppression system for preventing 5'-methylcytosine from deamination in the CpG sites. Second, the 5'-methylcytosine to thymine conversion has different efficiency at the CpG sites, depending on a trinucleotide structure. The conversion rate was highest (46.41%) in the complementary combination of 5' A⊗G 3'-5' C⊗T 3', followed by 5' C⊗C 3'-5' G⊗G 3' (32.16%), and 5' C⊗A 3'-5' T⊗G 3' had the lowest rate of 20.29% (Table 1). Third, the conversion of 5'-methylcytosine into thymine also occurred in non-CpG sites, but was limited in the structures of 5' A⊗A 3'-5' T⊗T 3' (29.64%) and 5' G⊗A 3'-5' T⊗C 3' (14.54%) observed in this study (Table 1). Obviously this is not surprising, because HARDER *et al.* (2004) observed both CpG methylation and non-CpG (CpT, CpC, and CpA) methylation in 21 tumor samples, 16 neurofibromas, and 10 female/10 male leukocytes. All these data indicated that methylations in the genomes are heavily protected for their deamination to thymine, but these observations need to be further validated in other species.

This research was funded by an Emerging Research Issues Internal Competitive Grant from Washington State University, College of Agricultural, Human, and Natural Resource Sciences, Agricultural Research Center, Pullman, Washington.

LITERATURE CITED

- BIRD, A. P., 1986 CpG-rich islands and the function of DNA methylation. *Nature* **321**: 209–213.
- BOX, G. E., and G. JENKINS, 1976 *Time Series Analysis: Forecasting and Control*, Ed. 2. Holden-Day, San Francisco.
- COOPER, D. N., and M. KRAWCZAK, 1990 The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions. *Hum. Genet.* **85**: 55–74.
- DU, F. X., A. C. CLUTTER and M. M. LOHUIS, 2007 Characterizing linkage disequilibrium in pig populations. *Int. J. Biol. Sci.* **3**: 166–178.

- DUNCAN, B. K., and J. H. MILLER, 1980 Mutagenic deamination of cytosine residues in DNA. *Nature* **287**: 560–561.
- EBERLY, L. E., and B. P. CARLIN, 2000 Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. *Stat. Med.* **19**: 2279–2294.
- GELMAN, A., J. B. CARLIN, H. S. STERN and D. B. RUBIN, 2004 *Bayesian Data Analysis*, Ed. 2. Chapman & Hall/CRC, London.
- GOODMAN, S., 1999 Toward evidence-based medical statistics. 2: The Bayes factor. *Ann. Intern. Med.* **130**: 1005–1013.
- HARDER, A., M. ROSCHE, D. E. REUSS, N. HOLTkamp, K. UHLMANN *et al.*, 2004 Methylation analysis of the neurofibromatosis type 1 (NF1) promoter in peripheral nerve sheath tumours. *Eur. J. Cancer* **40**: 2820–2828.
- HAYES, B., J. K. LAERDAHL, S. LIEN, T. MOEN, P. BERG *et al.*, 2007 An extensive resource of single nucleotide polymorphism markers associated with Atlantic salmon (*Salmo salar*) expressed sequences. *Aquaculture* **265**: 82–90.
- JIANG, Z., Z. WANG, T. KUNEJ, G. A. WILLIAMS, J. J. MICHAL *et al.*, 2007 A novel type of sequence variation: multiple-nucleotide length polymorphisms discovered in the bovine genome. *Genetics* **176**: 403–407.
- KIM, S., and A. MISRA, 2007 SNP genotyping: technologies and biomedical applications. *Annu. Rev. Biomed. Eng.* **9**: 289–320.
- MORTON, B. R., 1995 Neighboring base composition and transversion/transition bias in a comparison of rice and maize chloroplast noncoding regions. *Proc. Natl. Acad. Sci. USA* **92**: 9717–9721.
- MORTON, B. R., I. V. BI, M. D. McMULLEN and B. S. GAUT, 2006 Variation in mutation dynamics across the maize genome as a function of regional and flanking base composition. *Genetics* **172**: 569–577.
- RAFALSKI, A., 2002 Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opin. Plant Biol.* **5**: 94–100.
- RAFFERTY, A. E., 1995 Bayesian model selection in social research. *Sociol. Methodol.* **25**: 111–163.
- RILEY, J. H., C. J. ALLAN, E. LAI and A. ROSES, 2000 The use of single nucleotide polymorphisms in the isolation of common disease genes. *Pharmacogenomics* **1**: 39–47.
- SAUNA, Z. E., C. KIMCHI-SARFATY, S. V. AMBUDKAR and M. M. GOTTESMAN, 2007 Silent polymorphisms speak: how they affect pharmacogenomics and the treatment of cancer. *Cancer Res.* **67**: 9609–9612.
- SCHWARZ, G., 1978 Estimating the dimension of a model. *Ann. Stat.* **6**: 461–464.
- WANG, D., A. D. JOHNSON, A. C. PAPP, D. L. KROETZ and W. SADÉE, 2005 Multidrug resistance polypeptide 1 (MDR1, ABCB1) variant 3435C>T affects mRNA stability. *Pharmacogenet. Genomics* **15**: 693–704.
- YAMAGISHI, M. E., and A. I. SHIMABUKURO, 2008 Nucleotide frequencies in human genome and fibonacci numbers. *Bull. Math. Biol.* **70**: 643–653.
- ZHANG, F., and Z. ZHAO, 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., and E. BOERWINKLE, 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., and F. ZHANG, 2006 Sequence context analysis in the mouse genome: single nucleotide polymorphisms and CpG island sequences. *Genomics* **87**: 68–74.

Communicating editor: C. HALEY