Genomics & Informatics Vol. 5(2) 68-76, June 2007

A Statistical Analysis of SNPs, In-Dels, and Their Flanking Sequences in Human Genomic Regions

Seung Wook Shin[1], Young Joo Kim[2] and Byung-Dong Kim[1,3]*

[1]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-921, Korea, [2]Functional Genomics Research Center, Korea Research Institute of Bioscience and Biotechnology, Daejon 305-333, Korea, [3]Center for Plant Molecular Genetics and Breeding Research, Seoul National University, Seoul 151-921, Korea

Abstract

Due to the increasing interest in SNPs and mutational hot spots for disease traits, it is becoming more important to define and understand the relationship between SNPs and their flanking sequences. To study the effects of flanking sequences on SNPs, statistical approaches are necessary to assess bias in SNP data. In this study we mainly applied Markov chains for SNP sequences, particularly those located in intronic regions, and for analysis of in-del data. All of the pertaining sequences showed a significant tendency to generate particular SNP types. Most sequences flanking SNPs had lower complexities than average sequences, and some of them were associated with microsatellites. Moreover, many Alu repeats were found in the flanking sequences. We observed an elevated frequency of single-base-pair repeat-like sequences, mirror repeats, and palindromes in the SNP flanking sequence data. Alu repeats are hypothesized to be associated with C-to-T transition mutations or A-to-I RNA editing. In particular, the in-del data revealed an association between particular changes such as palindromes or mirror repeats. Results indicate that the mechanism of induction of in-del transitions is probably very different from that which is responsible for other SNPs. From a statistical perspective, frequent DNA lesions in some regions probably have effects on the occurrence of SNPs.

Keywords: single nucleotide polymorphisms, SNPs, Intron, Markov chain

Introduction

The genomes of different individuals typically differ by millions of nucleotides, due in part to inheritance and *de novo* formation of single nucleotide polymorphisms (SNPs). There are more than 3,000,000 known SNPs in the human genome. SNPs have been studied extensively due to the high density genetic markers linked to disease risk, patient outcome, response to specific treatments and treatment side effects (Taylor *et al*., 2001). A set of SNPs suitable for research purposes can be retrieved from public databases, the largest of which is dbSNP. This database contained more than 10 million reference SNP records in 2001 (Sherry *et al*., 2001). With increasing interest in SNPs and mutational hot spots for disease traits, it is becoming more important to define and understand the relationship between SNPs and their flanking sequences.

Mutations mainly occur in one of two ways: as a result of replication errors such as mismatches or frame shifts (Roos *et al*., 1996), or as a result of nucleotide alterations due to mutagens or radiation exposure (Ikehata *et al*., 2004). Mutations can form differently under various constraints, and there are many factors associated with the failure of DNA repair. We believe that SNPs could be the effect of such failures and their original sequences (especially their flanking sequences) could be the cause. Despite the possibility of the *post hoc, ergo propter hoc* fallacy, SNP location can be an important constraint. Firstly, mutation rates vary among regions of the mammalian genomes (Wolfe *et al*., 1989). For example, rates of frameshift mutation for runs of T's and A's are higher than those for runs of C's and G's, and the complementary strands of DNA differ with respect to replication and transcription (Burns *et al*., 1994; Francino *et al*., 1997). Secondly, location can affect SNPs, and the density of SNPs varies with the region involved (The International SNP Map Working Group, 2001). In this study, we aimed to detect statistically significant relationships between SNPs and their flanking sequences, since some flanking sequences may influence particular SNPs.

Studies regarding the linguistic properties of nucleotide sequences were initiated soon after the first long DNA sequences became available (Brendel *et al*., 1986). Since the Markov chain model is useful in identifying biased sequences, we applied this model to detect SNP biases with respect to the flanking sequences. We developed

*Corresponding author: E-mail kimbd@snu.ac.kr
Tel +82-2-880-4933, Fax +82-2-873-5410
Accepted 3 June 2007

methods designed to identify contrast words; those that are significantly over or under-represented by comparison with a model (Burge *et al*., 1992). We present the model employed for these purposes, together with the statistical procedure for identifying biased sequences.

Methods

Statistical methods

The Markov chain model was used to analyze the biases of SNPs and their flanking sequences (Schbath *et al*., 1995). The Markov chain model enabled us to deduce the number of occurrences of SNP biases that are the net effects of the flanking sequences by reading whole given nucleotide sequences. In this research, we assigned different Markov orders, such as 1, 3, 5, and 7, to remove the effects of bias of flanking nucleotides which include mono-, di-, tri-, up to tetra-nucleotides.

Let $W = (w_1\, w_2 \ldots w_m)$ be the word made by concatenation of m nucleotides and *N*(*W*) be its observed count in a sequence, which has a length of m. Under the Markov maximal order model (Leung *et al*., 1996), the expected count of *W*, *E*(*W*), is calculated by

$$\mathrm{E(W)} = \frac{\mathrm{N(w_1 w_2 \ldots\ w_{m-1}) N(w_2 w_3 \ldots w_m)}}{\mathrm{N(w_2 w_3 \ldots\ w_{m-1})}}$$

Having obtained the theoretical expectation for a count of the word, we needed a statistical way to compare it with the actual observed count in a statistically meaningful way. In this work, we used the z value statistics proposed by Schbath *et al.* (1997), where var(W) represents the calculated variance of N(W) - E(W). The main advantage of using the z value is that it follows a reduced normal distribution for a sufficiently large value of N(W) (Schbath *et al*., 1995).

$$z_w = \frac{N(W) - E(W)}{\sqrt{\mathrm{var}(W)}}$$

The z value is the measure of the bias of a word in a text. When the z value drops to zero, it indicates that there is 'no bias' of the word in a genomic text, while a large negative z value indicates that there is an 'under-representation' and a large positive z value indicates that there is an 'over-representation'. For large sequences and counts, the variance for the maximal Markov model can be approximated by the following expression (Schbath *et al*., 1995).

$$\mathrm{var(W)} = \mathrm{E(W)} \times \frac{\mathrm{[N(w_2 w_3 \ldots w_{m\text{-}1}) - N(w_1 w_2 \ldots w_{m\text{-}1})][N(w_2 w_3 \ldots w_{m\text{-}1}) - N(w_2 w_3 \ldots w_m)]}}{\mathrm{N(w_2 w_3 \ldots w_{m\text{-}1})^2}}$$

For each word W, we counted the number of occurrences of the word in the given sequences and computed its expected frequency and variance using the count of smaller words and the previous formula. We then tested whether the z value was significant or not, i.e. if it was compatible with the assumptions of the Markov model. Since we knew that z values are distributed in a normal distribution, we needed to elucidate whether the absolute z value was higher than a given threshold, which is 3.29 for 1% or 1.96 for 5% statistical significance, for a single test.

In this study, the Markov chain model was used to evaluate the significance of a word by taking into account the distribution of words of size m - 2, which measures the significance of an individual word only when all bias caused by the words of smaller sizes are removed. If a motif is degenerate, i.e. not strictly conserved in comparison with its consensus sequence, we can detect less significant results because this approach is used to count only exact words. If we include insertions and deletions in our count, the approach may be more accurate. Another important point lies in the definition of a 'significantly' biased word. Whichever statistic and model are used, the significance of the deviation between *N*(*W*) and *E*(*W*) varies with the counts of *N*(*W*) because the statistical test can distinguish a small deviation with higher accuracy when the counts are larger. For this reason, we also defined and labelled difference value to sequences that had a large z value.

$$Difference = \frac{\sqrt{[N(W) - E(W)]^2}}{[E(W)]}$$

Data sources and programs

Human SNP sequences were obtained from the NCBI database, namely dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) (Benson *et al*., 2000). Because dbSNP is one of the largest SNP databases, providing user-friendly options, those SNPs that were validated by cluster, frequency, submitter, and double hit were selected. For example, 1,600,949 out of 3,245,651 intron sequences were identified to be validated. We excluded those sequences that had flanking sequences that were too short or had a significant number of missing characters denoted by N. There were no identical sequences in the NCBI database. Among eight functional classes in the NCBI database, we primarily analyzed intronic regions. To have an appropriate option, we used SNPs that had a true single nucleotide polymorphism with two alleles. We also tested other SNP classes such as in-dels. The flanking-sequence-checking software was written in C++, which is freely accessible at http://plaza.snu.ac.kr/~best/essay/FCP.zip.

Results

We gleaned some insights from the analysis of SNPs and their flanking sequences in genomic intronic regions. Even though functionally, introns are somewhat disregarded in eukaryotes, some SNPs in intronic regions may play an important role by affecting the expression level of relevant genes or proteins (Liu *et al*., 2004). Some RNA editing in intronic regions are considered to affect splicing sites and modulate splice site selection (Flomen *et al*., 2004).

Biases of the mononucleotides flanking SNPs

We studied genomic intron data because of its abundance in the database, and because there are fewer biological constraints in intronic regions compared to functional or exonic regions (although there are functional regions in introns). Mononucleotides flanking SNPs were analyzed by the Markov chain model to find higher absolute z values (Table 1). A high absolute z value indicated that the SNPs and their flanking nucleotides appeared to be significantly biased in intronic regions. The histogram distribution of the mononucleotides flanking SNPs is very different from a normal distribution (Fig. 1A). A few sequences have values close to zero, meaning that there is no bias. Among the highest 12 SNPs and their flanking mononucleotides shown in the Table 1, all the sequences had the approximate

Table 1. SNPs and their flanking mononucleotides.

SNP*	N(W)	E(W)	Z value	Difference
T[A/G]T	27,373	22,330	47.83	0.226
A[T/C]A	27,045	22,167	46.35	0.220
T[A/T]A	5,570	3,778	40.10	0.474
T[T/C]G	18,019	22,115	-39.43	0.185
A[T/C]C	11,777	15,329	-38.89	0.232
G[A/G]T	11,735	15,239	-38.48	0.230
C[A/G]A	17,923	21,888	-38.30	0.181
T[A/C]T	4,701	3,301	31.62	0.424
T[A/G]G	12,018	14,908	-30.82	0.194
C[T/C]A	12,042	14,894	-30.43	0.191
C[A/T]A	1,654	2,874	-29.96	0.424
A[T/G]A	4,675	3,354	29.65	0.394

*SNPs and their flanking mononucleotides that show the highest absolute z values among 804,527 intron sequences.

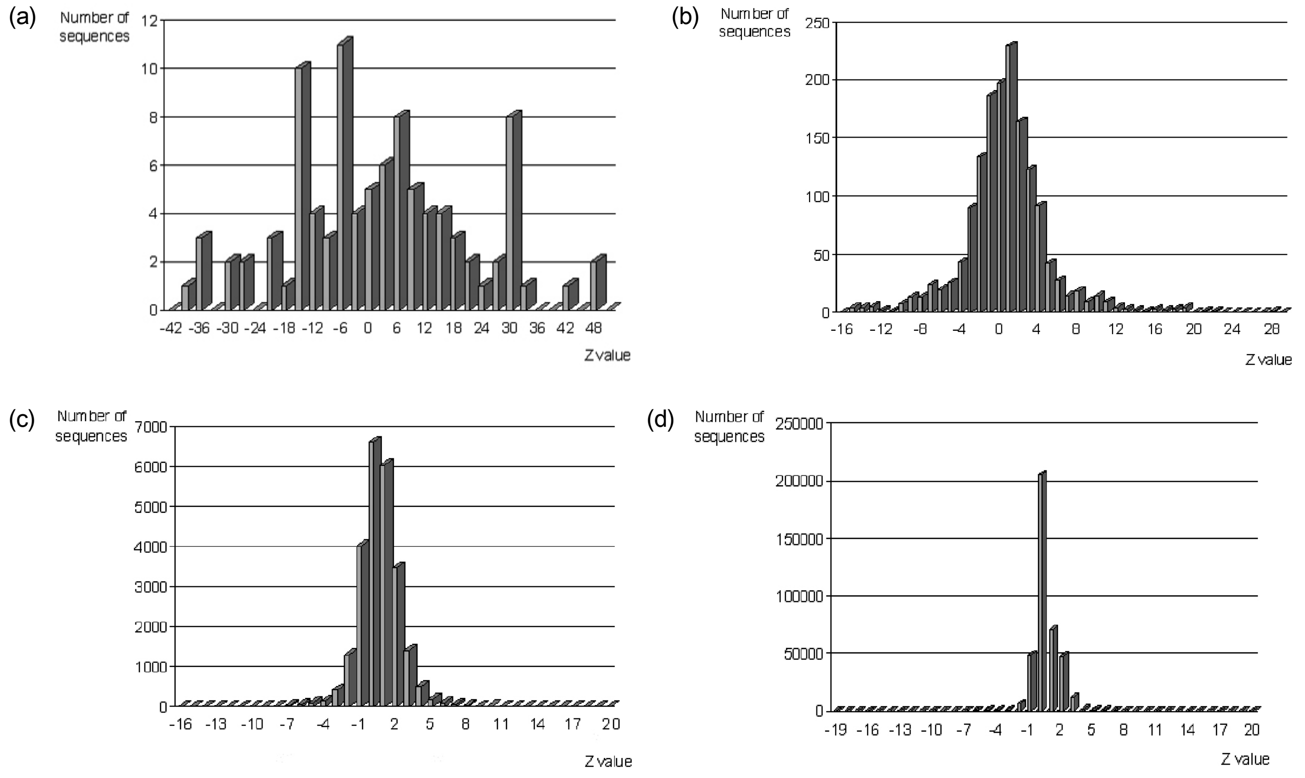

Fig. 1. Diagrams illustrating the shapes of distributions.
Histogram distributions of the z values (X axis) versus the number of sequences. (A) Z values of the mononucleotides flanking SNPs. (B) Z values of the dinucleotides flanking SNPs. (C) Z values of the trinucleotides flanking SNPs. (D) Z values of the tetranucleotides flanking SNPs.

same z value with respect to their reverse or complementary sequences except in the cases of T[A/T]A and A[A/T]T. The z value of T[A/T]A is 40.10, over-represented, and that of A[A/T]T is -12.75, under-represented. Interestingly, the same mononucleotides flanking SNPs such as T[A/G]T, A[T/C]A were over-represented.

Biases of the dinucleotides flanking SNPs

We also used the Markov chain model to analyze SNPs and their flanking dinucleotides. Histogram distribution of the dinucleotides flanking SNPs can be approximated by a normal distribution (Fig. 1B). However, the distribution of the dinucleotides flanking SNPs is wide, and the small peaks around a z value of 18 are due to over-represented sequences in Figure 1B. Interestingly, in this analysis, there was a significant portion of over-represented sequences among sequences that have high absolute z values (Table 2). In most cases, the flanking dinucleotides consisted of double bases like AA, TT, CC, and GG. Also the first base-pair of the left flanking dinucleotides and the last base-pair of the right flanking dinucleotides had a high tendency to be the same as shown in GA[T/C]GG, CC[A/G]TC, CA[T/C]GC, and GC[A/G]TG. This symmetrical pattern of SNPs was consistently observed and hypothesized to be responsible for the over-represented sequences.

The twelve greatest z values in the results of the dinucleotides flanking SNPs were positive, and the greatest z values were from the results of the mononucleotide to the tetranucleotide transition. We presume that biases normally lie on the over-represented sequences in such a way that some flanking sequences stimulate DNA to generate SNPs, and these sequences are under biological constraints that are weakest near the SNP sites, or that it is hard to repair DNA around the over-represented sequences.

Table 2. SNPs and their flanking dinucleotides.

SNP*	N(W)	E(W)	Z value	Difference
TT[A/T]AA	2,927	2,476	27.21	0.182
GA[T/C]GG	2,817	2,076	21.33	0.357
CC[A/G]TC	2,851	2,147	20.07	0.328
CC[T/C]GG	3,461	2,773	18.99	0.248
CA[T/C]GC	3,504	2,795	18.63	0.254
GC[A/G]TG	3,509	2,809	18.37	0.249
CC[A/G]AG	2,219	1,705	18.00	0.301
TC[A/G]CT	1,966	1,461	17.78	0.346
CC[A/G]GG	3,468	2,835	17.43	0.223
AA[A/C]AA	2,187	1,708	16.77	0.280
TT[T/G]TT	1,865	1,496	16.63	0.247

*SNPs and their flanking dinucleotides that show the highest absolute z values among 804,527 intron sequences.

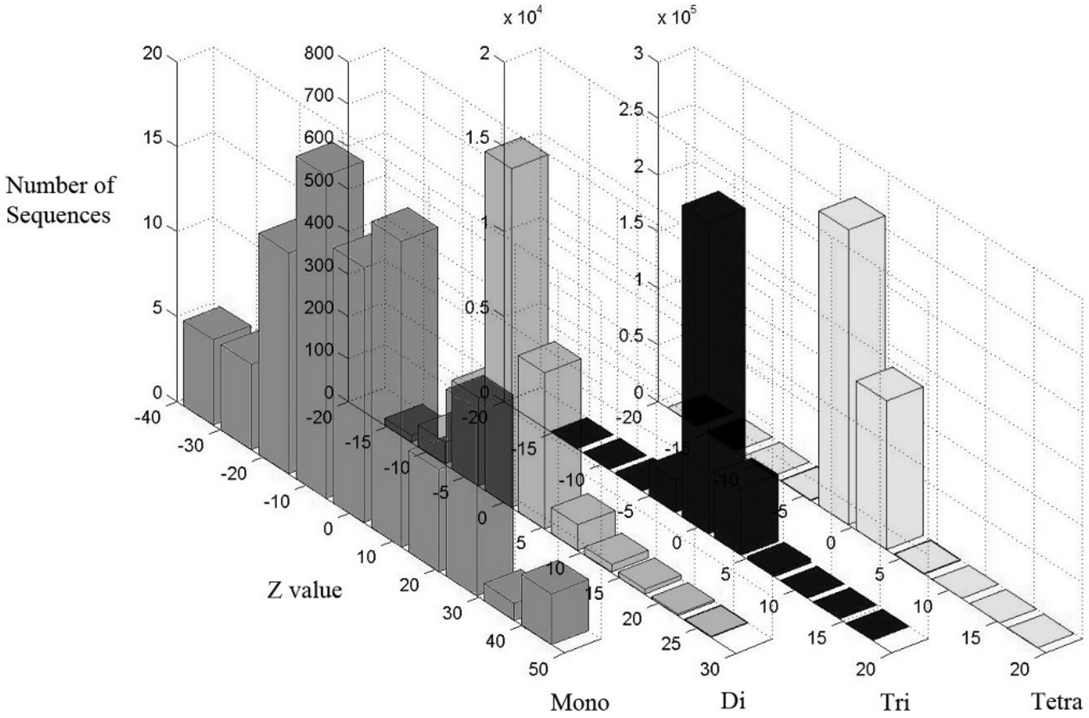

Fig. 2. Comparison between histograms.
Histogram distributions of the z values (X axis) versus the number of sequences from mono to tetra.

Biases of the trinucleotides flanking SNPs

The trinucleotides flanking SNP analysis found that the over-represented sequences were symmetrical around the SNPs and behaved as microsatellites (Table 3). In our test, z values were distributed in a normal distribution. However, the human genome is filled with repeated sequences such as microsatellites. So most symmetrical sequences are slightly over-represented due to their conserved repeat motif. We should endeavour to interpret results carefully when using the Markov chain model. The results of the trinucleotides flanking SNPs were generally similar to those of the dinucleotides flanking SNPs. The histogram distribution of the trinucleotides flanking SNPs can be approximated by a normal distribution (Fig. 1C). This histogram has a narrow bell shape but the distribution of the trinucleotides flanking SNPs is wide because of a few highly over-represented sequences.

Numerous over-represented sequences had large absolute z values. In the cases of CCC[A/G]AGT and ACT[T/C]GGG, both had high z and difference values. Both sequences were parts of specific motifs such as CCTCAGCCTCCC[A/G]AGTAGCT GGGAC and GTCC CAGCTACT[T/C]GGG AGGCTGAGG. They are probably related to Alu repeats (ATCCCAGCACTTTGGGAGGCC GAGG [GenBank:551 542]). There are a few single-base-pair repeat-like sequences such as TTT[T/C]TTT and AAA[A/G]AAA which were also over-represented. However, these showed relatively low difference values. The over-represented sequences showed a much lower complexity. The left and right flanking bases had a tendency to be the same. We also found that, in 16 out of the 30 cases of the highest absolute z value sequences, the first base-pair of the left group of trinucleotides and the last base-pair of the right group of trinucleotides were the same, for example CTC[A/G]CTC, ACA[T/C]GCA, GAC[A/G]GGG, and GTT[T/C]GAG. Usually, they were located in various single base-pair repeat regions or microsatellite regions. Since these sequences had high positive z values, we presume that they were over-represented partly because of their single-base-pair rich sequences and partly because of their microsatellite repeats. For example, GAC[A/G]GGG sequences were not symmetrical but the sequences were G-rich. Although these were single base-pair rich sequences, other sequences turned out to be microsatellites such as CTC[A/G]CTC and ACA[T/C]ACA.

Table 3. SNPs and their flanking trinucleotides.

SNP*	N(W)	E(W)	Z value	Difference
CCC[A/G]AGT	333	163	18.34	1.043
GCC[A/G]GGC	630	391	18.26	0.611
AAG[T/C]GAT	321	169	16.72	0.899
GCC[T/C]GGC	625	410	16.21	0.524
ACT[T/C]GGG	293	150	15.87	0.953
TTT[T/C]TTT	1,050	850	15.68	0.235
ATC[A/G]CTT	357	206	15.61	0.733
GCC[A/G]AGT	66	221	-15.58	0.701
GTG[T/C]GCG	120	44	14.29	1.727
CGC[A/G]CAC	133	52	14.28	1.558
GGG[T/C]GAC	287	168	13.91	0.708
ACT[T/C]GGC	82	218	-13.85	0.624
AAA[A/G]AAA	1,018	854	13.39	0.192
GTC[A/G]CCC	278	167	12.99	0.665
ATA[A/G]ACA	653	514	12.98	0.270
GGC[A/G]TGA	712	533	12.96	0.336
AAA[A/C]AAA	966	841	12.90	0.149
GTT[T/C]GAG	215	117	12.86	0.838
CTC[A/G]AAC	246	143	12.80	0.720
GGC[A/G]TGT	191	341	-12.59	0.440
GAC[A/G]GGG	365	242	12.29	0.508
TCA[T/C]GCC	682	515	12.28	0.324
ACA[T/C]GCC	189	333	-12.22	0.432
TAT[A/G]TGT	406	284	11.89	0.430
CTC[T/C]GTC	392	272	11.63	0.441
TGC[A/G]TGT	311	189	11.58	0.646
GCC[A/G]AGG	435	308	11.56	0.412
ACA[T/C]GCA	284	168	11.53	0.690
ACA[T/C]ACA	436	324	11.47	0.346
CTC[A/G]CTC	279	175	11.33	0.594

*SNPs and their flanking trinucleotides that show the highest absolute z values among 804,527 intron sequences.

Microsatellites are frequently observed in human genomes, and they have conserved repeat motifs. Most symmetrical sequences are slightly over-represented because of their conserved repeat motifs. However, microsatellites are highly polymorphic markers and their length varies due to a high rate of mutation. To some extent, biological constraints do not affect microsatellite regions. Some highly over-represented SNP sequences may be caused by DNA slippage, which results in the loss of proper hydrogen bonding in DNA (Levinson *et al.*, 1987). SNPs can easily be formed by Simple Sequence Repeat (SSR) mutations as well as by hydrogen bond losses. We also found combinations of two microsatellites that were symmetrical around SNPs such as ATA[T/C]ACA and TAT[A/G]TGT. We found many AT and AC repeats in ATA[T/C]ACA sequences. In this case, it was considered that SNPs were formed by two SSR mutations in the upstream and downstream regions of SNPs. This is plausible because SSR mutation rate is much higher than SNP mutation rate, even though the loci of SSR mutations are restricted (Brinkmann *et al.*, 1998). In the case of GTG[T/C]GCG, the assumption of SSR mutation was barely applied because GC repeats are greatly under-represented in the human genome (International Human Genome Sequencing Consortium, 2001). However, GTG[T/C]GCG had a very high difference value. C-to-T transition mutations, the most abundant SNP type, could probably be used to explain this phenomenon best. In humans, DNA methylation occurs predominantly at CpG dinucleotides and plays an essential role in many transcription processes, such as development, and X-chromosome inactivation (Robertson *et al*., 2000). These methylated sites are considered to exhibit high rates of C-to-T transition mutations (Zingg *et al*., 1997).

Biases of the tetranucleotides flanking SNPs

The results from the tetranucleotides flanking SNPs showed a strong relationship with Alu repeats (Table 4). Since Alu repeats are very frequent in the human genome and have conserved motifs, these results showed high difference values despite their low observed and expected counts. No identical sequences were discovered in the NCBI database. Histogram distribution of the tetranucleotides flanking SNPs can be approximated by a normal distribution (Fig. 1D). However, two sequences are highly over-represented. Also, two sequences are highly under-represented as a result of the over-represented sequences. We ceased further inspection after the analysis of the tetranucleotides flanking SNPs due to a lower number of each SNP sequence.

We discovered that the sequences AGGC[A/G]TGAG, CTCA[T/C]GCCT, CCAC[T/C]GCAC, GTGC[A/G]GTGG, TCAC[T/C]GCAA, ATCA[T/C]GCCA, ATCC[A/G]CCTG, CAGG[T/C]GGAT, and GAGG[T/C]GGAG had Alu motifs such as GGGATAACAGGC[A/G]TGAGCCACTGCG and its reverse complement CACGGTGGCTCA[T/C]GCC TGTAATCCC [GenBank: 15145591], GGTGTGAGCCA C[C/T]GCACCCGGCCTG and its reverse complement AGGCTGGAGTGC[A/G]GTGGCTCATTCC [GenBank: 21322214], AGCTCGGCTCAC[T/C]GCAACCTCCGCC [GenBank: 15809152], GAGCCGAGATCA[T/C]GCCAC TGCACTC [GenBank: 33667252], ACCTGGTGATCC [A/G]CCTGCCTCGGCC and its reverse complement GGCCGAGGCAGG[T/C]GGATCACCTGAG [GenBank: 2895321], and GGACTAAGGAGG[T/C]GGAGCTTGC AGT [GenBank: 18450199], respectively. Alleles of almost all sequences were [A/G] or [T/C]. Moreover, these sequences had CpG-like motifs such as [T/C]G, and its reverse complement C[A/G]. Alu repeats are known to be associated with C-to-T transition or RNA editing (Kim *et al*., 2004). Individual Alu elements have 24 or more CpG dinucleotides that are easy to mutate as a result of the deamination of 5-methylcytosine residues (Batzer *et al*., 1993). Alu CpGs account for about one-third of the potential methylation sites in human DNA (Hellmann-Blumberg *et al*., 2005). Alu repeats may also be related to adenosine deaminases acting on RNA (ADAR) editing (Eisenberg *et al*., 2005). This study shows that some of the SNP databases actually display somatic modification, namely A-to-I RNA editing (Knight *et al*., 1996). Alu repeats can be used as the evolutionary mechanism and can actually influence the accumulation of SNPs in the genome (Batzer *et al*., 1995).

Table 4. SNPs and their flanking tetranucleotides.

SNP*	N(W)	E(W)	Z value	Difference
AGGC[A/G]TGAG	341	212	19.51	0.608
TGGC[A/G]TGAG	15	139	-19.47	0.892
CTCA[T/C]GCCT	354	236	18.73	0.500
CTCA[T/C]GCCA	18	126	-17.83	0.857
CCAC[T/C]GCAC	177	104	13.62	0.702
AAAA[A/T]ATAT	385	356	13.05	0.081
TCAC[T/C]GCAA	94	38	12.96	1.474
GTGC[A/G]GTGG	147	84	12.81	0.750
ATCA[T/C]GCCT	20	86	-12.72	0.767
ATCA[T/C]GCCA	109	46	12.71	1.370
ATCC[A/G]CCTG	61	18	12.70	2.389
CAGG[T/C]GGAT	63	19	12.58	2.316
GAGG[T/C]GGAG	230	170	12.35	0.353

*SNPs and their flanking tetranucleotides that show the highest absolute z values among 804,527 intron sequences.

The sequence AAAA[A/T]ATAT had an Alu motif represented by AAAAAAAAA AAA[A/T]ATATATATATAT [Gen-Bank:2275185]. The 3’ terminus of the Alu element almost always consists of a run of A’s that is only occasionally interspersed with other bases (Levanon *et al*., 2004). In addition, individual Alu repeats are also flanked by short (A+T)-rich direct repeated sequences that form when the elements integrate into staggered chromosomal breaks, and are thought to have appeared as a result of the endonucleolytic activity of LINE-derived reverse transcriptase (Jurka, 1997). However, in the case of AAAA[A/T]ATAT data, many simple TA repeats were found after the 3’ oligo(dA)-rich tails of Alu elements.

Biases in in-del data

We analyzed the mononucleotides flanking in-del to test the Markov chain model with different data, and to find the highest absolute z values in in-del data (Table 5), although this model has been used in other studies (Schbath *et al*., 1995; Leung *et al*., 1996; Schbath, 1997; Rocha *et al*., 1998). The number of validated data in in-dels was 6,963. While in-del sequences that have the same left and right mononucleotides were significantly under-represented, sequences that have the different mononucleotides flanking in-dels were over-represented; the opposite result from the mononucleotides flanking SNPs (Table 1). This suggests that the formation of in-del variation is significantly different from that of SNPs and that the Markov chain model is a good

Table 5. In-dels and their flanking mononucleotides that show the highest absolute z values among 804,527 intron sequences.

In-dels	N(W)	E(W)	Z value	Difference
T[M]T	493	629	-7.78	0.216
A[M]A	407	536	-7.74	0.241
C[M]C	182	270	-6.65	0.326
A[M]T	656	559	5.71	0.174
G[M]G	249	328	-5.57	0.241
C[M]A	465	403	4.11	0.154
T[M]A	674	604	4.06	0.116
G[M]C	286	236	3.99	0.212
T[M]G	621	561	3.56	0.107
G[M]T	397	368	2.01	0.079

[M] represents any inserted or deleted nucleotide(s).

Table 6. Distribution of inverted repeats in 6,963 in-dels.

Size of inverted repeats (bps)	No. of stem loop structures	No. of hairpin
11	1	0
10	0	0
9	0	0
8	1	0
7	0	0
6	3	2
5	2	1
4	1	1

model to find different biases in various data sets. In the in-del data, few mirror repeats were observed in the flanking sequences. However, we did find some palindromes and mirror repeats in the alleles of the in-del data. Particularly, many long tandem repeats were found in relatively large alleles of the in-del data. There were 40 in-del sequences, the alleles of which were 24 bps or larger in nucleotide length. We found approximately 8 putative stem loop structures that consisted of 4 or more base-pair hairpins among these 40 alleles (Table 6). For example, the allele of rs16434 ([-/TCCCACGCGAGTGTGGTGGGACCTTG]) has a stem loop structure (underlined) [dbSNP:rs16434]. We found approximately 9 mirror repeats that consisted of 4 or more bp units among the 40 alleles (Table 7). For example, the allele of rs1610903 ([-/GGTGAGGGTGAGGGTATATGGGGAGT]) has mirror repeats (underlined) [dbSNP:rs1610903]. H-DNA can also be formed in mirror repeats (Mirkin *et al.*, 1987).

Discussion

As for the comparison of histogram distributions, biases of the mono-, di-, tri-, and tetranucleotides flanking SNPs could be observed clearly (Fig. 2). The mononucleotides flanking SNPs had an effect on a large percentage of the total number of SNP sequences (Fig. 2). Moreover, the distribution of the dinucleotides flanking SNPs was skewed and long-tailed because of a few over-represented sequences. The distribution of the trinucleotides flanking SNPs was also wide and long-tailed despite the narrow bell shape. In the tetranucleotides flanking SNPs, a few sequences were highly biased. Thus, we concluded that the long-tailed distributions suggest the effects of the flanking sequences on SNPs.

Table 7. Distribution of mirror repeats in 6,963 in-dels.

Size of mirror (bps)	No. of mirror
11	1
10	0
9	1
8	1
7	0
6	1
5	3
4	2

We suggest that three main factors are the cause of SNPs. Firstly, SNPs may occur in some regions where biological constraints are not in effect, to some extent. High polymorphism at microsatellite loci probably represents the fact that SNPs at some microsatellite loci may not be very harmful. Alu repeats are mobile genetic elements and to some degree, biological constraints probably do not affect Alu repeats. In this study, a few flanking sequences that have stem loop structures, and palindrome sequences such as TTGATTTTTT[A/T]AAAAAATCAA [dbSNP:rs4607190] were found even though it was not possible for statistical significance to be inferred. A SNP at the center of the palindrome may not significantly affect its structure.

Secondly, some sequences may cause a great deal of damage to certain nucleotides. It is plausible to think that biological constraints are the most important factor for SNP occurrence. However, at least in a statistical perspective, frequent DNA lesions in some regions can have effects on SNP occurrence in coding regions, and intron regions. We also applied Markov chains for the analysis of sequences in coding regions (data not shown). In both intron regions and coding regions, the greatest z values were positive. If biological constraints were the most important factor of SNP occurrence, the greatest z value would have been negative. The loss of proper hydrogen bonds by mismatches or unusual structures can be caused by microsatellites, palindromes, mirror repeats and A-tracts (McCarthy *et al.*, 1991). DNA slippage is probably the most frequent cause of DNA lesions. A surprising portion of the human genome consists of repetitive DNA sequences such as short tandem repeat sequences. These mutate at a higher rate than the majority of DNA sequences (Asicioglu *et al.*, 2004).

Nucleotides at the loops of palindrome sequences probably have compromised hydrogen bonds.

Thirdly, a variety of repair strategies have evolved to correct DNA lesions and noncanonical DNA structures may hinder some of these. However, some proteins can recognize damaged DNA and repair it effectively. For example, human XPA and RPA probably recognize helical distortions in DNA (Vasquez *et al*., 2002). DNA damage may affect the structure of the double helix, and noncanonical structures are vulnerable to the effects of DNA damage. In in-del data, large in-dels are probably caused predominately by stem structure and H-DNA, as intramolecular hydrogen bonds induce stem structures (Kim, 1985) and hydrogen bonds in H-DNA induce triplex formation. Some large in-dels are also probably caused by long tandem repeats. In H-DNA, one half of the purine strand enters the triplex whereas the other half is unstructured and can form a duplex with the complementary oligonucleotide (Belotserkovskii *et al*., 1992). It is most likely that triplexesin mirror repeats, stem structures and large DNA slippage are the cause of large insertions or deletions. Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells and can induce large-scale deletion (Wang *et al*., 2004). However, the factors that cause SNPs are complicated, and require further research and experimental proofs. The future work for SNP data analysis will include challenges such as the development of statistical methods, and the characterization of the interactions of SNP data.

Acknowledgements

This work was supported by a grant from the Center for Plant Molecular Genetics and Breeding Research (CPMGBR) through the Korea Science and Engineering Foundation (KOSEF) and the Korea Ministry of Science and Technology (MOST). YJ Kim was partially supported by a grant from the Stroke Oriental Medicine Project (M1052701000005N270100000) of MOST.

References

Asicioglu, F., Oguz-Savran, F., and Ozbek, U. (2004). Mutation rate at commonly used forensic STR loci: paternity testing experience. *Dis. Markers*. 20, 313-315.

Batzer, M. A., Rubin, C. M., Hellmann-Blumberg, U., Alegria-Hartman, M., Leeflang, E. P., Stern, J. D., Bazan, H. A., Shaikh, T. H., Deininger, P. L., and Schmid, C. W. (1995). Dispersion and insertion polymorphism in two small subfamilies of recently amplified human Alu repeats. *J. Mol. Biol*. 247, 418-427.

Batzer, M.A., Deininger, P.L., Hellmann-Blumberg, U., Jurka, J., Labuda, D., Rubin, C.M., Schmid, C.W., Zietkiewicz, E., and Zuckerkandl, E. (1996). Standardized nomenclature for Alu repeats. *J. Mol. Evol*. 42, 3-6.

Belotserkovskii, B. P., Krasilnikova, M. M., Veselkov, A. G., and Frank-Kamenetskii, M. D. (1992). Kinetic trapping of H-DNA by oligonucleotide binding. *Nucleic Acids Res.* 20, 1903-1908.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A., and Wheeler, D. L. (2000). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 28, 10-14.

Brendel, V., Beckman, J. S., and Trifonov, E. N. (1986). E. N. Linguistics of nucleotide sequences. *J. Biomol. Struct. Dyn*. 4, 11-21.

Brinkmann, B., Klintschar, M., Neuhuber, F., Hu Hne, J., and Rolf, B. (1998). Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am. J. Hum. Genet*. 62, 1408-1415.

Burge, C., Campbell, A. M., and Karlin, S. (1992). Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl. Acad. Sci. USA* 89, 1358-1362.

Burns, D. P. and Temin, H. M. (1994). High rates of frameshift mutations within homo-oligomeric runs during a single cycle of retroviral replication. *J. Virol.* 68, 4196-4203.

Eisenberg, E., Adamsky, K., Cohen, L., Amariglio, N., Hirshberg, A., Rechavi, G., and Levanon, E. Y. (2005). Identification of RNA editing sites in the SNP Database Eisenberg. *Nucleic Acids Res*. 33, 4612-4617.

Flomen, R., Knight, J., Sham, P., Kerwin, R., and Makoff, A. (2004). Evidence that RNA editing modulates splice site selection in the 5-HT2C receptor gene. *Nucleic Acids Res.* 32, 2113-2122.

Francino, M. P. and Ochman, H. (1997). Strand asymmetries in DNA evolution. *Trends Genet.* 13, 240-245.

Hellmann-Blumberg, U., McCarthy Hintz, M. F., Gatewood, J. M., and Schmid, C. W. (1993). Developmental differences in methylation of human Alu repeats. *Mol. Cell. Biol*. 13, 4523-4530.

Ikehata, H., Nakamura, S., Asamura, T., and Ono, T. (2004). Mutation spectrum in sunlight-exposed mouse skin epidermis: Small but appreciable contribution of oxidative stress-mediated mutagenesis. *Mutat. Res.* 556, 11-24.

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860-921.

Jurka, J. (1997). Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc. Natl. Acad. Sci. USA* 94, 1872-1877.

Kim, B. D. (1985). Four-stranded DNA: An intermediate of homologous recombination and transposition. *Kor. J. Breed*. 17, 453-466.

Kim, D. D., Kim, T. T., Walsh, T., Kobayashi, Y., Matise, T. C., Buyske, S., and Gabriel, A. (2004). Widespread RNA editing of embedded Alu elements in the human transcriptome. *Genome Res.* 14, 1719-1725.

Knight, A., Batzer, M. A., Stoneking, M., Tiwari, H. K., Scheer, W. D., Herrera, R. J., and Deininger, P. L. (1996). DNA sequences of Alu elements indicate a recent replacement of the human autosomal genetic complement. *Proc. Natl.Acad. Sci. USA* 93, 4360-4364.

Leung, M.-Y., Marsh, G. M., and Speed, T. P. (1996). Over- and underrepresentation of short DNA words in herpesvirus genomes. *J. Comput. Biol.* 3, 345-360.

Levanon, E. Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., Fligelman, Z. Y., Shoshan, A., Pollock, S.R., Sztybel, D., Olshansky, M., Rechavi, G., and Jantsch, M. F. (2004). Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat. Biotechnol*. 22, 1001-1005.

Levinson, G. and Gutman, G. A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* 4, 203-221.

Liu, Z., Sun, H. X., Zhang, Y. W., Li, Y. F., Zuo, J., Meng, Y., and Fang, F. D. (2004). Effect of SNPs in protein kinase Cz gene on gene expression in the reporter gene detection system. *World J.Gastroenterol.* 10, 2357-2360.

McCarthy, J. G. and Rich, A. (1991). Detection of an unusual distortion in A-tract DNA using KMnO4: effect of temperature and distamycin on the altered conformation. *Nucleic Acids Res.* 19, 3421-3429.

Mirkin, S. M., Lyamichev, V. I., Drushlyak, K. N., Dobrynin, V. N., Filippov, S. A., and Frank-Kamenetskii, M. D. (1987). DNA H form requires a homopurine-homopyrimidine mirror repeat. *Nature* 330, 495-497.

Robertson, K. D. and Jones, P. A. (2000). DNA methylation: past, present and future. *Carcinogenesis* 21, 461-467.

Rocha, E. P. C., Viari, A., and Danchin, A. (1998). Oligonucleotide bias in Bacillus subtilis: general trends and taxonomic comparisons. *Nucleic Acids Res.* 26, 2971-2980.

Roos, D., de Boer, M., Kuribayashi, F., Meischl, C., Weening, R. S., Segal, A. W., Ahlin, A., Nemet, K., Hossle, J. P., Bernatowska-Matuszkiewicz, E., and Middleton-Price, H. (1996). Mutations in the X-linked and autosomal recessive forms of chronic granulomatous disease. *Blood* 87, 1663-1681.

Schbath, S. (1997). An efficient statistic to detect over- and under-represented words in DNA sequences. *J. Comput. Biol*. 4, 189-192.

Schbath, S., Prum, B., and Turckheim, É. (1995). Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *J. Comput. Biol*. 2, 417-437.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308-311.

Taylor, J. G., Choi, E. H., Foster, C. B., and Chanock, S. J. (2001). Using genetic variation to study human disease. *Trends Mol. Med.* 7, 507-512.

The International SNP Map Working Group. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409, 928-933.

Vasquez, K. M., Christensen, J., Li, L., Finch, R. A., and Glazer, P. M. (2002). Human XPA and RPA DNA repair proteins participate in specific recognition of triplex-induced helical distortions. *Proc. Natl. Acad. Sci. USA* 99, 5848-5853.

Wang, G. and Vasquez, K. M. (2004). Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells. *Proc. Natl. Acad. Sci. USA* 101, 13448-13453.

Wolfe, K. H., Sharp, P. M., and Li, W. H. (1989). Mutation rates vary among regions of the mammalian genome. *Nature* 337, 283-285.

Zingg, J. M. and Jones, P. A. (1997). Genetic and epigenetic aspects of DNA methylation on genome expression, evolution, mutation and carcinogenesis. *Carcinogenesis* 18, 869-882.