BcSNPdb: Bovine Coding Region Single Nucleotide Polymorphisms Located Proximal to Quantitative Trait Loci

Sunjin Moon¹, Hyoung Doo Shin², Hyun Sub Cheong², Hye Young Cho², Sohg Namgoong², Eun Mi Kim², Chang Su Han², Samsun Sung¹ and Heebal Kim¹,*

¹Laboratory of Bioinformatics and Population Genetics, School of Agricultural Biotechnology, Seoul National University, Seoul 151-742, Korea
²Department of Genetic Epidemiology, SNP Genetics Inc., Rm 1407, 14th floor, B-dong, Woolim Lion’s Valley, 371-28, Gasan-dong, Geumcheon-Gu, Seoul, Korea 153-803

Received 25 August 2006, Accepted 12 September 2006

Bovine coding region single nucleotide polymorphisms located proximal to quantitative trait loci were identified to facilitate bovine QTL fine mapping research. A total of 692,763 bovine SNPs was extracted from 39,432 UniGene clusters, and 53,446 candidate SNPs were found to be a depth >3. In order to validate the in silico SNPs experimentally, 186 animals representing 14 breeds and 100 mixed breeds were analyzed. Genotyping of 40 randomly selected candidate SNPs revealed that 43% of these SNPs ranged in frequency from 0.009 to 0.498. To identify non-synonymous SNPs and to correct for possible frameshift errors in the ESTs at the predicted SNP positions, we designed a program that determines coding regions by protein-sequence referencing, and identified 17,735 nsSNPs. The SNPs and bovine quantitative trait loci informations were integrated into a bovine SNP data: BcSNPdb (http://snugenome.snu.ac.kr/BtcSNP/). Currently there are 43 different kinds of quantitative traits available. Thus, these SNPs would serve as valuable resources for exploiting genomic variation that influence economically and agriculturally important traits in cows.

Keywords: Bovine single nucleotide polymorphism, Non-synonymous SNP, QTL

Introduction

As the number of genes encoding a trait increases, it becomes more difficult to model the inheritance of the trait via Mendelian genetics (She et al., 2004). Mutations associated with rare diseases are usually recessive and low in frequency in the population as the result of high selection pressure against the deleterious allele (Ramensky et al., 2002). In contrast to rare diseases, the genetic basis for complex traits has been difficult to determine because they are caused by common polymorphisms that are dispersed over multiple genes (Hirschhorn, 2005). Thus, single nucleotide polymorphisms (SNPs) are the most frequent and important forms of DNA variation for quantitative trait loci (QTL) mapping.

Reliable computational methods, including PolyBayes (Sachidanandam et al., 2001) for expressed sequence tag (EST) data and PolyPhred (Nickerson et al., 1997) for genome data, have been used to identify SNPs. The PolyPhred program was developed to detect the presence of heterozygous SNPs using fluorescence-based sequencing of PCR products from a genomic sequence. However, it is compatible with the Phred/Phrap/Consed pipeline, and using the program in this pipeline has successfully extracted SNPs from porcine (Fahrenkrug et al., 2002), bovine (Stone et al., 2002), and chicken (Nagaki et al., 2004) ESTs, by tagging homozygous SNPs. One of the main limitations of extracting SNPs from a broad range of EST data is the availability of sequence trace files, which are needed to determine the quality of each sequence position. To overcome this, the PolyPhred developer suggested the use of SudoPhred (http://www.phrap.org), which gives each sequence a quality score. In addition, Barker et al. (Day et al., 2004) developed the AutoSNP program to address this problem. One of the most important benefits of extracting candidate SNPs from EST data is the acquisition of non-synonymous SNPs (nsSNP), or SNPs that result in a change in the amino acid sequence of the encoded protein (Nagaki et al., 2004). Recently, an interactive bovine in silico SNP (IBISS) database was developed from bovine ESTs (Hawken et al., 2004). Although the IBISS is a valuable resource for bovine QTL...
mapping, it does not yet include a large amount of bovine QTL mapping results. Thus, to facilitate bovine QTL fine mapping research, we identified bovine in silico SNPs from ESTs of nsSNPs, and used linkage mapping marker and bovine genome scaffold data to obtain various QTL mapping results.

Materials and Methods

Data source. A nonredundant set of bovine (Bos taurus) gene-oriented clusters was obtained from the UniGene database (National Center for Biotechnology) (Wheeler et al., 2005). A UniGene cluster representing a unique gene consists of sequences expressed in various tissues. Bovine UniGene Build #74 contains 39,432 clusters. The Mar. 2005 Bos taurus draft genome assembly (Btau_2.0) was obtained from the UCSC genome browser. The assembly represents about 17.7 Gb of sequence and 6.2 Gb coverage of the bovine genome. Bos taurus EST sequences were obtained from the CattleGene Index (The Institute for Genomic Research).

SNP identification by sequence data mining. Sequences of each UniGene cluster were converted into reference trace using the program SudoPhred in the PolyPhred software package (Nickerson et al., 1997). Each base of the reference sequence was assigned a Phred quality score of 30. Each cluster was assembled into multiple contigs with the default parameter of Phrap. Then potential heterozygous sites were detected by comparing sequence traces using PolyPhred 5.02 and a quality threshold of 30, a rank threshold of 6, and a genotype tag. We selected only the homozygous tag of the PolyPhred genotype. The 5'- and 3'-flanking sequences were aligned to the bovine genome using BLAT (Kent, 2002), and identities higher than 97% were regarded as potential SNPs.

Identification of amino acid-changing SNPs. To identify non-synonymous SNPs (nsSNPs) and to correct for possible frameshift errors in the ESTs at the predicted SNP positions, we designed a program that determines coding regions by protein-sequence referencing, based on a method that identifies nsSNPs from chicken ESTs (Nagaki et al., 2004). The computational process, which is written in Python script and can be viewed at http://snugenome.snu.ac.kr/, uses a standalone BLASTX program to search the nonredundant protein database with an E-value of 10^{-5}. The mixture was incubated at 95°C for 5 min in a reaction mixture and then analyzed by electrophoresis in an ABI Prism 3100 Genetic Analyzer. The results were statistically analyzed using the software GeneScan and Genotyper (Applied Biosystems).

Results

Identification of candidate SNPs from bovine sequences data. A total of 692,763 bovine SNPs was extracted from 39,432 UniGene clusters, and 53,446 candidate SNPs were found to be a depth >3. The depth of alignment, or the number of ESTs per contig, ranged from 3 to 302, and the average number of ESTs per contig was 17.0 (Fig. 1). The percentages of transition substitution types (A/G and C/T; 62%) were higher than transversions (T/A, T/G, C/A, and C/G; 38%), whereas the percentages of A/G, C/T, T/A, T/G, C/A, and C/G substitution types were 31, 31, 7, 12, 9, and 10%, respectively. The average frequency of polymorphisms in the bovine expressed genes was one per 724 bp of contig sequence. Most contigs had multiple SNPs averaging 13 SNPs per contig.

Fig. 1. Number of bovine SNP plotted against the depth of alignment.
Determination of amino acid-changing SNPs by double-screening methods. A more detailed analysis of candidate bovine SNPs was performed by modifying a previous study (Nagaki et al., 2004). Self- and cross-referencing protein sequences were obtained. If nothing was found in the mammalian protein sequence database at NCBI, then the ESTScan program was used to estimate the coding regions for all contig sequences that contained candidate bovine SNPs (Iseli et al., 1999). This method is four times more precise than using BLASTX in BLASTALL on the nonredundant (nr) database alone (Fig. 2A). In addition, we found that the specificity and sensitivity of the ESTScan were equivalent to those of BLASTX (Fig. 2B). About 86% of the 53,446 candidate SNPs were located in the corresponding best-hit protein sequence or predicted coding sequence. In total, 17,735 (38%) were predicted to be amino acid-altering mutations, and 6,581 (14%) were synonymous, which do not change protein sequences. A total of 18,290 (39%) fell into the transcribed but untranslated region (UTR), whereas 2,602 and 15,688 were located at 5' and 3' UTR respectively. The other 14 candidate SNPs were classified as erroneous or unknown.

Qualification of predicted SNP. In total, 186 animals representing 14 breeds and 100 mixed breeds were analyzed. Genotyping of 40 randomly selected candidate SNPs revealed that 43% of these SNPs ranged in frequency from 0.009 to 0.498. The other 57% of these SNPs were monomorphic or expected to be rare SNPs.

Construction of a bovine candidate SNP database with QTL. We integrated all of the information related to the SNP data (http://snugenom.e.snu.ac.kr/) including SNP type, genomic location, UniGene cluster ID, mRNA sequence, and LocusLink ID, when available; best-hit protein sequence found in human, mouse, and yeast genomes; the QTL region that the SNP located; and tissues in which the gene was expressed and flanking sequences 50 nt upstream and downstream. This information is searchable by index keyword, region, and batch query, and a user-friendly graphical interface makes browsing the SNP database easy. SNPs located in the region related to dairy and beef traits can also be searched. Ten QTL traits are available: birth weight, slaughter weight, hot carcass weight, udder balance, yearling weight, adjusted yearling weight, adjusted weaning weight, protein yield, dressing percentage, and adjusted fat. These QTL marker data are derived from the Bovine QTL Viewer site, which contains public domain bovine QTL data. The statistics of the database are summarized in Table 1.

Discussion

We built a bovine SNP database based on a reliable SNP identification method. The database contains more than 50,000 SNPs from a nonredundant set of gene-oriented sequence clusters. In silico SNP prediction is limited because most sequence variations are due to sequencing error during high-throughput sequencing projects. Furthermore, contig sequences are typically obtained from a redundant sample source. For these reasons, the depth of an SNP locus does not directly indicate allele frequency. Therefore, the majority of false-positives with low depth are not screened (Nagaki et al., 2004). We validated the candidate SNP data from the

<table>
<thead>
<tr>
<th>Type</th>
<th>Synonymous</th>
<th>Non-synonymous</th>
<th>5' UTR</th>
<th>3' UTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>QTL</td>
<td>4340</td>
<td>5849</td>
<td>895</td>
<td>11239</td>
</tr>
<tr>
<td>Non-QTL</td>
<td>2011</td>
<td>2735</td>
<td>419</td>
<td>5341</td>
</tr>
</tbody>
</table>

Table 1. Classification of bovine SNP
randomized sample of SNPs, and more than half had a monomorphic locus. This result is consistent with a recent study on silico SNP identification, which reported a validation rate of 50% (Hawken et al., 2004). Keeping these validation statistics in mind, the higher number of nsSNPs may have been because most of them were false-positives. This large amount of non-synonymous mutations may occur when it is assumed that the nucleotide substitution is random at any codon (Mendrzyk et al., 2006). An alternative explanation might be that the sampling biases may have been considered target genes related to agriculturally important traits in cows. These genes would be subject to strong artificial selection via domestication (Innan & Kim, 2004; Wright et al., 2005). Thus, different pressures of strong purifying selection may have acted on the same genes of different breeds. In the human genome, nsSNPs are overrepresented in the extended LD region (Hinds et al., 2005). The average SNP frequency of polymorphism also has to be adjusted to one SNP per 1.5 kilobases. This frequency is moderate compared to that of humans (one SNP per 1,000-2,000 bp), the mouse (one SNP per 250-20,000 bp), and the dog (one SNP per 900-1,500 bp) (Sachidanandam et al., 2001; Wade et al., 2002; Lindblad-Toh et al., 2005).

We classified each SNP according to whether it alters amino acid sequences, which would likely modify protein function, using a double-screening method (Nagaki et al., 2004). The optimal method of identifying SNP types in ESTs is self-species referencing to protein sequences, but self-species referencing in the bovine, as well as other economically important animals, is difficult because there are insufficient protein sequence data in public databases. To predict SNP type accurately with high throughput, we designed a double-screening strategy using self- and cross-species protein referencing, in addition to the ESTScan program. We demonstrated that this method is efficient and precise in extracting candidate nsSNPs from bovine EST data. The procedure used to construct this comprehensive bovine SNP database was designed to allow frequent updates.

Traditional breeding programs select phenotypic differences among breeds. The increasing public demand for higher-quality meat and milk has led to the development of more cost-effective methods of production. The discrepancy in abilities among breeds may stem from nsSNPs in genes that are responsible for phenotypic traits (Cohen-Zinder et al., 2005; Stone et al., 2005), and the QTLs appear to be more complex. As our bovine SNP database gains more qualitative traits of the dairy and beef that producers consider of high importance, it should serve as a valuable resource for exploiting genomic variation that influence economically and agriculturally important traits in cows. These SNPs would serve as potential markers for selection in breeding, control of animal disease, and to enhance food quality.

Acknowledgments This work was supported by a grant from the BioGreen 21 Program of the Korean Rural Development Administration and by a graduate fellowship provided by the Ministry of Education through the Brain Korea 21 project.
References


