Journal of Biochemistry and Molecular Biology, Vol. 40, No. 1, January 2007, pp. 95-99

JOURNAL OF Biochemistry and Molecular Biology

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

Sunjin Moon[1], Hyoung Doo Shin[2], Hyun Sub Cheong[2], Hye Young Cho[2], Sohg Namgoong[2], Eun Mi Kim[2], Chang Su Han[2], Samsun Sung[1] and Heebal Kim[1,*]

[1]Laboratory of Bioinformatics and Population Genetics, School of Agricultural Biotechnology, Seoul National University, Seoul 151-742, Korea

[2]Department 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 traits 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

*To whom correspondence should be addressed.
Tel: 82-2-880-4803; Fax: 82-2-883-8812
E-mail: heebal@snu.ac.kr

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 Cattle Gene 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} (Altschul *et al.*, 1997). In the case of null results, coding regions are estimated using the ESTScan program, and a parameter file for human and mammalian genomes are applied to the bovine EST contigs. The cSNPs are then identified by comparing EST contig sequences to the BLASTX results, and the aligned regions of the protein sequence obtained from BLASTX are used as seeds. Each seed alignment is extended to the 5'- and 3'-ends of the contig of the codon unit. Meanwhile, structural information about the query sequence is generated, i.e., the coding untranslated region (UTR), query strands with respect to the RefSeq, the existence or nonexistence of initiation and termination codons, and erroneous sequences that cannot constitute amino acids. Finally, the query SNP variant is mapped and translated into a protein to determine whether the amino acid sequence has been altered.

Qualification of Bovine SNP. To validate the predicted SNP data, we randomly selected 40 SNP and genomic flanking sequences from the data set to calculate statistics of prediction accuracy. Hanwoo (82), Holstein (17), Angus (2), Red Angus (4), Charloais (4), Hereford (6), Limousin (5), Simmental (8), Gelbvieh (3), Jersey (15), Ayshire (15), Brown Swiss (8), Guernsey (14), and Canadienne (3) semen samples were collected from Canadian herds. Random samples of mixed breeds (100) were collected from meat shops in Seoul, Korea (one sample per shop). Genomic DNA was extracted using the manual phenol/chloroform method. SNPs were genotyped using Sing-Base Extension (SBE) and electrophoresis.

Primer extension reactions were performed with SNaPshot ddNTP Primer Extension Kit (Applied Biosystems). To clean up the product of the primer extension reaction, one unit of SAP was added to the reaction mixture and incubated at 37°C for 1 h, followed by 15 min at 72°C for enzyme inactivation. The DNA samples, containing extension products, and Genescan 120 Liz size standard solutions were added to Hi-Di formamide (Applied Biosystems). The mixture was incubated at 95°C for 5 min, followed by 5 min on ice, 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.

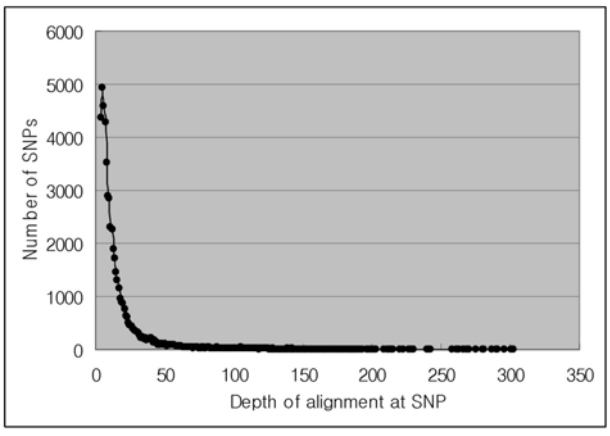

Fig. 1. Number of bovine SNP plotted against the depth of alignment.

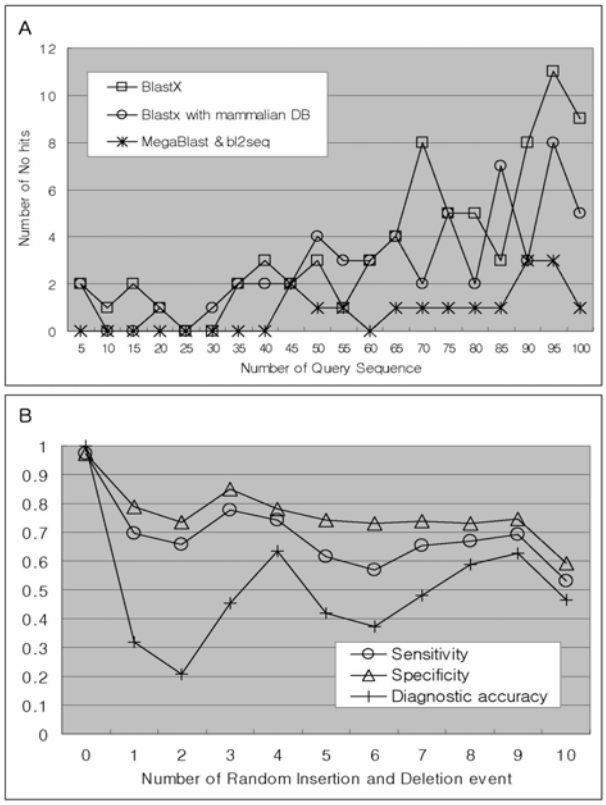

Fig. 2. Evaluation of cSNPer performance. Using a combination of MegaBlast and bl2seq was the best way to determine the corresponding protein sequence accurately (A) and ESTScan or the RefSeq protein sequence yielded similar sensitivity on the random mutation sequences (B).

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://snugenome.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 1. Classification of bovine SNP

Type	Synonymous		Non-synonymous		5' UTR		3' UTR	
	SNP	%	SNP	%	SNP	%	SNP	%
QTL	4340	9.4	5849	12.6	895	1.9	11239	24.2
Non-QTL	2011	4.3	2735	5.9	419	0.9	5341	11.5

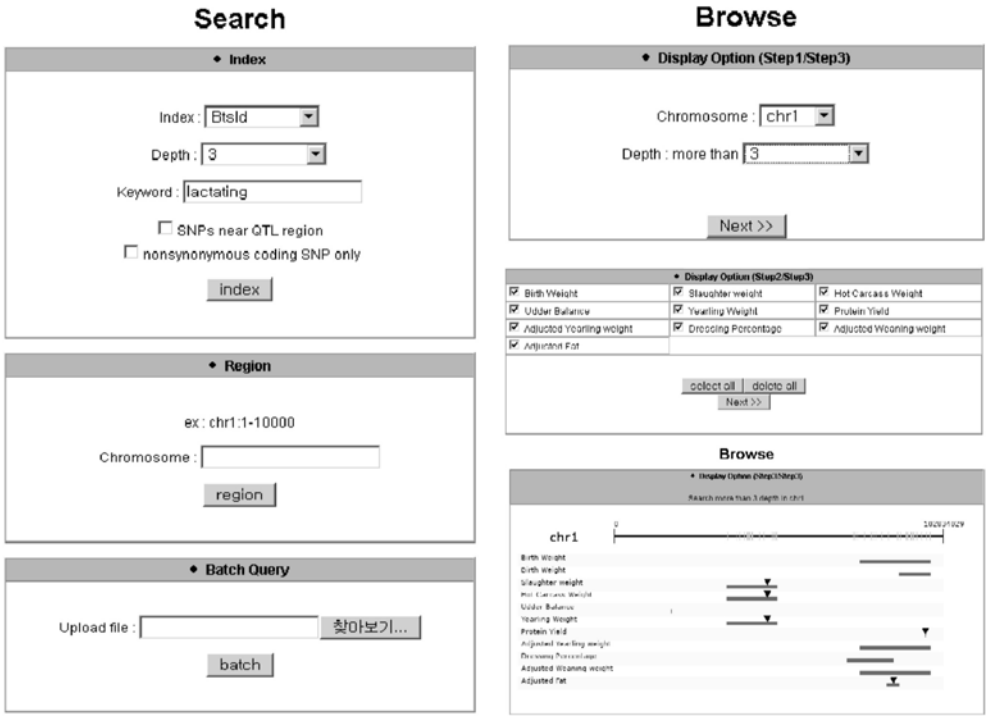

Fig. 3. Web-interface of searching for the bovine SNP database.

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

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.

Cohen-Zinder, M., Seroussi, E., Larkin, D. M., Loor, J. J., Everts-van der Wind, A., Lee, J. H., Drackley, J. K., Band, M. R., Hernandez, A. G., Shani, M., Lewin, H. A., Weller, J. I. and Ron, M. (2005) Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Res.* **15**, 936-944.

Day, I. N., Chen, X. H., Gaunt, T. R., King, T. H., Voropanov, A., Ye, S., Rodriguez, S., Syddall, H. E., Sayer, A. A., Dennison, E. M., Tabassum, F., Barker, D. J., Cooper, C. and Phillips, D. I. (2004) Late life metabolic syndrome, early growth, and common polymorphism in the growth hormone and placental lactogen gene cluster. *J. Clin. Endocrinol. Metab.* **89**, 5569-5576.

Fahrenkrug, S. C., Freking, B. A., Smith, T. P., Rohrer, G. A. and Keele, J. W. (2002) Single nucleotide polymorphism (SNP) discovery in porcine expressed genes. *Anim. Genet.* **33**, 186-195.

Hawken, R. J., Barris, W. C., McWilliam, S. M. and Dalrymple, B. P. (2004) An interactive bovine in silico SNP database (IBISS). *Mamm. Genome* **15**, 819-827.

Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., Frazer, K. A. and Cox, D. R. (2005) Whole-Genome Patterns of Common DNA Variation in Three Human Populations. *Science* **307**, 1072-1079.

Hirschhorn, J. N. (2005) Genetic approaches to studying common diseases and complex traits. *Pediatr. Res.* **57**, 74-77.

Innan, H. and Kim, Y. (2004) Pattern of polymorphism after strong artificial selection in a domestication event. *PNAS* **101**, 10667-10672.

Iseli, C., Jongeneel, C. V. and Bucher, P. (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 138-48.

Kent, W. J. (2002) BLAT--the BLAST-like alignment tool. *Genome Res.* **12**, 656-664.

Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., Clamp, M., Chang, J. L., Kulbokas, E. J., 3rd, Zody, M. C., *et al.* (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803-819.

Mendrzyk, F., Korshunov, A., Toedt, G., Schwarz, F., Korn, B., Joos, S., Hochhaus, A., Schoch, C., Lichter, P. and Radlwimmer, B. (2006) Isochromosome breakpoints on 17p in medulloblastoma are flanked by different classes of DNA sequence repeats. *Genes Chromosomes Cancer* **45**, 401-410.

Nagaki, K., Cheng, Z., Ouyang, S., Talbert, P. B., Kim, M., Jones, K. M., Henikoff, S., Buell, C. R. and Jiang, J. (2004) Sequencing of a rice centromere uncovers active genes. *Nat. Genet.* **36**, 138-145.

Nickerson, D. A., Tobe, V. O. and Taylor, S. L. (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.* **25**, 2745-2751.

Ramensky, V., Bork, P. and Sunyaev, S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* **30**, 3894-3900.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., *et al.* (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928-933.

She, X., Horvath, J. E., Jiang, Z., Liu, G., Furey, T. S., Christ, L., Clark, R., Graves, T., Gulden, C. L., Alkan, C., Bailey, J. A., Sahinalp, C., Rocchi, M., Haussler, D., Wilson, R. K., Miller, W., Schwartz, S. and Eichler, E. E. (2004) The structure and evolution of centromeric transition regions within the human genome. *Nature* **430**, 857-864.

Stone, R. T., Casas, E., Smith, T. P., Keele, J. W., Harhay, G., Bennett, G. L., Koohmaraie, M., Wheeler, T. L., Shackelford, S. D. and Snelling, W. M. (2005) Identification of genetic markers for fat deposition and meat tenderness on bovine chromosome 5: development of a low-density single nucleotide polymorphism map. *J. Anim. Sci.* **83**, 2280-2288.

Stone, R. T., Grosse, W. M., Casas, E., Smith, T. P., Keele, J. W. and Bennett, G. L. (2002) Use of bovine EST data and human genomic sequences to map 100 gene-specific bovine markers. *Mamm. Genome* **13**, 211-215.

Wade, C. M., Kulbokas, E. J., 3rd, Kirby, A. W., Zody, M. C., Mullikin, J. C., Lander, E. S., Lindblad-Toh, K. and Daly, M. J. (2002) The mosaic structure of variation in the laboratory mouse genome. *Nature* **420**, 574-578.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Helmberg, W., *et al.* (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **33**, 39-45.

Wright, S. I., Bi, I. V., Schroeder, S. G., Yamasaki, M., Doebley, J. F., McMullen, M. D. and Gaut, B. S. (2005) The effects of artificial selection on the maize genome. *Science* **308**, 1310-1314.