Asian-Aust. J. Anim. Sci.
Vol. 23, No. 11 : 1399 - 1404
November 2010

www.ajas.info

An Integrated Genomic Resource Based on Korean Cattle (Hanwoo) Transcripts

Dajeong Lim[1, 2, a], Yong-Min Cho[2, a], Seung-Hwan Lee[2], Samsun Sung[1], Jungrye Nam[1], Duhak Yoon[2], Younhee Shin[3], Hye-Sun Park[2] and Heebal Kim[1, *]

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

ABSTRACT : We have created a Bovine Genome Database, an integrated genomic resource for *Bos taurus*, by merging bovine data from various databases and our own data. We produced 55,213 Korean cattle (Hanwoo) ESTs from cDNA libraries from three tissues. We concentrated on genomic information based on Hanwoo transcripts and provided user-friendly search interfaces within the Bovine Genome Database. The genome browser supported alignment results for the various types of data: Hanwoo EST, consensus sequence, human gene, and predicted bovine genes. The database also provides transcript data information, gene annotation, genomic location, sequence and tissue distribution. Users can also explore bovine disease genes based on comparative mapping of homologous genes and can conduct searches centered on genes within user-selected quantitative trait loci (QTL) regions. The Bovine Genome Database can be accessed at http://bgd.nabc.go.kr. (**Key Words :** Korean Cattle (Hanwoo), Bovine, EST, Transcript, Genome Database)

INTRODUCTION

The bovine genome project was started at the Baylor College of Medicine Human Genome Sequencing Center and the British Columbia Genome Sequencing and Mapping Platform at the British Columbia Cancer Agency (BCCA) in December 2003 (http://www.hgsc.bcm.tmc.edu/projects/bovine). The first assembly was based on 3.3-fold coverage of the bovine genome, and by 2005, the first whole-genome radiation hybrid (WG-RH) map for cattle was constructed using a 5000-rad RH panel (Womack et al., 1997), 319 ordered microsatellites, and 768 ESTs at low resolution. Currently, the Btau_4.0 genome build (7.1-fold coverage) is available; it was released in August 2008.

The cow is the first ruminant mammal to have its genome sequenced. It provides a reference point for identifying domesticated animals that may be better suited to a particular market or environment, using genomic studies.

Before the bovine genome sequencing project, a genetic linkage map of the bovine genome had been constructed that provided a basis for further mapping (Barendse et al., 1994; Bishop et al., 1994). Bovine genome sequencing has been underway since 2003, and various types of databases related to bovine genome resources have been developed. A bovine genomic database was constructed and included in the ArkDB of the Roslin Institute. This database provides comprehensive and generic public repositories for genome mapping data from farmed species and other animals (Hu et al., 2001). Information on the bovine genome and related data has been primarily obtained from the European Bioinformatics Institute (EBI) (http://www.ensembl.org/Bos_taurus/index.html) and the US National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/projects/genome/guide/cow). The Bovine Genome Database (http://genomes.arc.georgetown.edu/bovine) supported a QTL viewer, chromosome gBrowse (Stein et al., 2002), gene annotation, and a BAC map, which included overall genome information on Hereford breeds as model bovines from a bovine genome database consortium.

With high coverage of genomic sequences, polymorphisms such as SNPs can be found that can serve as genetic markers for future linkage and association studies.

* Corresponding Author: Heebal Kim. Tel: +82-2-880-4803, Fax: +82-2-873-2271, E-mail: heebal@snu.ac.kr

[2] Division of Bioinformatics and Population Genetics, National Institute of Animal Science, RDA, Suwon 441-706, Korea.

[3] Insilicogen, Inc., Suwon Chomdan Venture Valley, 958, Suwon 441-813, Korea.

[a] The first two authors should be regarded as joint First Authors.

Received April 22, 2009; Accepted August 30, 2009

The NCBI dbSNP collection recently contained more than 2,223,000 *Bos taurus* RefSNP clusters (rs numbers). Panzitta et al. also described the Bovine SNP Retriever, a web facility for ready retrieval of bovine single-nucleotide polymorphism (SNP)-related information that allows searches centered on user-defined sequences (Panzitta et al., 2002).

For transcriptome data, a cattle EST project (http://titan.biotec.uiuc.edu/cattle/cattle_project.htm) was undertaken at the University of Illinois and provides functional data for about 23,000 EST clusters, as well as gene ontology (GO) annotations based on BLASTN and the human UniGene database regarding attributes of a sequence or EST cluster and orthologous gene relationships. There were also EST-related databases: the TIGR Gene Indices (http://www.tigr.org/tigrscripts/tgi/T_index.cgi?species=cattle), CSIRO's interactive bovine *in silico* SNP database (http://www.livestockgenomics.csiro.au/ibiss/), and the BtcSNP database (http://snugenome.snu.ac.kr/BtcSNP/) regarding SNP information for bovine coding region SNPs located proximal to QTL. These resources were designed to provide overall genomic information based on sequences from Hereford breeds.

In the present work, we focused on genome information based on Hanwoo transcripts and developed user-friendly search interfaces within the Bovine Genome Database. We collected bovine data from various public databases and generated an integrated map from genome informatics. The Bovine Genome Database displays mapped results and links them to other sources of mapping data. The database also provides transcript data on 55,213 Hanwoo ESTs, which we produced from cDNA libraries of three tissues. Users can also explore bovine disease genes based on comparative mapping of homologous genes and make searches centered on genes within user-selected quantitative trait loci (QTL) regions.

MATERIALS AND METHODS

Database integration

The UniGene datasets were downloaded from the *Bos taurus* (Build #92) UniGene cluster of NCBI. We also downloaded dbESTs from NCBI and stored tissue library information on ESTs for digital gene expression profiling. Bovine genomic sequences (bosTau 4) were obtained from the UCSC genome browser (http://hgdownload.cse.ucsc.edu/goldenPath/bosTau4/bigZips/). GO categorization of the genes was conducted with the GO mapping file of the Gene Ontology Consortium (http://www.geneontology.org) and TC sequences from the TIGR Gene Indices (Release 12.0, http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=cattle). For the identification of disease genes, we used homologene data (ftp://ftp.ncbi.nih.gov/pub/HomoloGene/build60/) and disease terms data from Mouse Genome Informatics (MGI; http://www.informatics.jax.org/). Bovine QTL data were modified from Animal QTLdb (http://www.animalgenome.org/QTLdb/) by the addition of homologous human genes within QTL regions. Human gene homologs in the bovine genome were identified using the Annotation database (http://hgdownload.cse.ucsc.edu/goldenPath/bosTau4/database/) of the UCSC genome browser. Table 1 provides summary statistics for the Bovine

Table 1. Summary statistics, data contents and features of the database

Menu	Data type	Number	Data contents	Features
Sequence Data	EST (Hanwoo)	55,213	NIAS	Provide gene annotation, genomic location, sequence, BLAST result, Gene Ontology, and library information.
	Contigs (Hanwoo)	4,759	NIAS	
	TC sequences	7,773	TIGR (BtGI release 12.0)	
Search By	Marker	381	USDA-MARC	Simple search the name of marker, QTL, and gene.
	QTL	1,123	Animal QTLdb	
	Gene	26,469	NCBI	
Disease Browser	Disease term	3,136	MGI (Mouse Genome Informatics)	Provide bovine disease genes as query, disease name, OMIM accession ID, or specific text.
	Gene	225,289	homologene of NCBI	
Gene Expression	EST	1,515,204	dbEST of NCBI	Identify tissue-specific genes based on digital-gene expression profiling using audic's test.
	UniGene cluster	4,042	UniGene of NCBI	
QTL comparative gene map	QTL	1,125	cattleQTL.gff' of Animal QTLdb	Find candidate human/bovine genes for a QTL of interest through querying chromosome, QTL traits
	Human gene	35,933	UCSC Genome Browser	
Genome Annotation	BES (Korean cattle)	37,759	NCBI	Display alignment data corresponding to the genomic region as simple as clicking on any genomic location.
	Bovine ESTs	1,575,285	UCSC Genome Browser	
	Bovine mRNA	18,473	UCSC Genome Browser	
	Bovine RefSeq	10,122	UCSC Genome Browser	
	GenScan	19,598	UCSC Genome Browser	

Genome Database.

Database content

Sequence data : We generated 55,213 expressed sequence tags (ESTs) from three cDNA libraries of Hanwoo fat, loin, and liver. Liver, intermuscular fat, and *longissimus dorsi* tissues were obtained from a 24-month-old Hanwoo steer immediately after slaughter. cDNA libraries were constructed according to the oligocapped method. Sequencing of cDNA clones and construction of EST datasets were described in Lim et al. (Lim et al., 2009). We retrieved 4,759 contigs and 7,857 singletons from 55,213 ESTs through an assembly procedure. We assigned functions to our transcript data by sequence homology searches against the NCBI non-redundant protein database (ftp://ftp.ncbi.nih.gov/ blast/db/). Our ESTs have been submitted to NCBI dbEST, and the accession numbers are GH296597-GH332022. To annotate bovine ESTs with Gene Ontology, a sequence similarity search was executed against the tentative consensus (TC) sequences of the *Bos taurus* Gene Index (BtGI, release 12.0) using BLASTN (Altschul et al., 1997) with cutoff values of 95% identity, 60% coverage, and an e-value <0.00001 for GO identification. Then, GO categorization of the ESTs was conducted with the GO profile of the Gene Ontology Consortium.

Disease browser : We retrieved homologene data from NCBI and extracted the protein accession IDs from human and mouse data. To identify orthologs from the human, mouse, and bovine homologenes, we used the reciprocal best blast hits algorithm (Wall et al., 2003) to make comparisons with the human, mouse, and bovine sequences using BLASTX and TBLASTN (cutoff: e-value 0.00001). Candidate disease genes from humans and mice were obtained from the Mouse Genome Informatics (MGI).

Gene expression : Significance tests of gene expression profiles between a pair of the cDNA libraries were performed using Audic's teset (Audic and Claverie, 1997). Because the results of multiple tests can generate an unexpected number of type I errors, a false discovery rate (FDR) correction was applied. We used a 0.01 significance level to reject the null hypothesis. We defined the genes as tissue-specific genes if they were explained by the alternative hypothesis, namely that the ratio of ESTs in that tissue for a given gene *was* significantly different from the ratio of ESTs in all other tissues. Bovine contigs in all libraries were subjected to the following criteria: contigs should consist of a minimum of five ESTs, and the enrichment values should be greater than 2.

QTL comparative gene map : We downloaded bovine QTL information from the Animal QTLdb 'QTL locations by bp' file that supported chromosome coordinates, traits, QTL evidence, and publication year. Mapping results of human gene homologs in the bovine genome were retrieved from the UCSC genome browser. We stored human gene information and detected human and bovine genes within the bovine QTL region.

Genome annotation : We estimated the genomic regions of the data sets, our ESTs, contigs, and BESs against the bovine genome using the BLAT (Kent, 2002) program (cutoff: 90% identity, coverage: 90%). We downloaded 37,759 Hanwoo BAC-end sequences (BESs) from NCBI. Bovine sequences in public databases were assigned to genomic regions using the Annotation database (http://hgdownload.cse.ucsc.edu/goldenPath/bosTau4/database/) of the UCSC genome browser, ESTs, mRNA, genScan information, RefSeq, and comparable human genes. We constructed databases of several types of sequence information including mapping results against the bovine genome, gene descriptions, CDS information, and tissue origin.

Database construction and implementation : The database and Web interface were developed using MySQL, PHP, HTML, and Javascript. The standard server requires an Apache 1.3.19, MySQL 4.0.21, and PHP 4.3.9 with GD library 2.0.33 and uses the FreeBSD 5.2 operating system. Data processing and analysis routines were written in Python.

RESULTS AND DISCUSSION

Database interface

The Bovine Genome Database can be accessed at http://bgd.nabc.go.kr. It supports various types of bovine genome data and implements searches for user-selected queries. The interface contains five main parts: Transcript Data, BLAST search, Search By, Application, and Genome Annotation. The Application part consists of three items: Disease Browser, Gene Expression, and a QTL comparative gene map. More detailed information on each section is provided below.

Sequence data : Bovine transcripts were aligned using the BLASTX program against the non-redundant protein database. We applied a strict threshold for the definition of a homologous sequence because it was not a cross-species sequence comparison that removed blast hits with paralogous sequences. To classify transcripts by putative function, GO categories were determined from the number of transcripts in this study and in the TIGR BtGI. Users can obtain basic EST and transcript information from the Sequences Data menu. EST data can be searched by local ID, accession number, or description of homology against the non-redundant NCBI protein database using the BLASTX score, BLAST score, and e-value. If a user clicks the local ID, then its details are reported on the page (i.e., genomic region, GO annotation, and assembled ESTs information). Functional characterization of unidentified

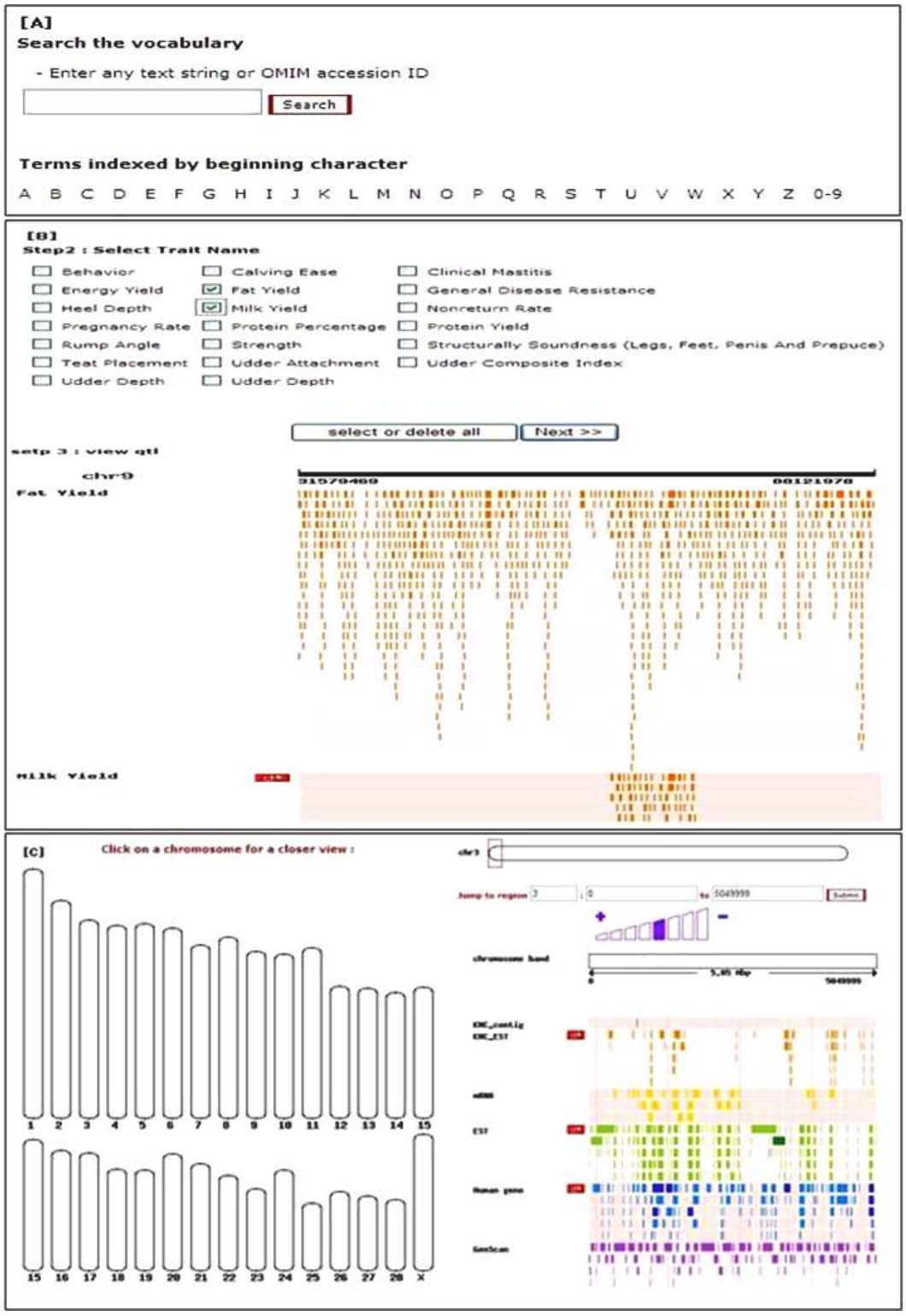

Figure 1. Screenshot showing results from the Bovine Genome Database using the query tool. Users can search for individual genes and sequence information by search options. Of the search interfaces, the Disease Browser enables the user to obtain putative bovine disease transcripts using OMIM accession ID or disease terms [A]. A QTL Comparative Gene Map can show candidate genes between human and bovine within a bovine QTL region. [B]. Genome annotations are shown in the Output Map view according to data type (Korean cattle ESTs, Korean cattle contigs, Korean cattle BESs, all ESTs, all mRNA, human genes, and GenScan results). Genome Browser provides genomic alignments of all data types within a specific genomic region [C].

genes may lead to important findings (Hansen et al., 2004). This menu focused on the functional characterization of Hanwoo ESTs and consensus sequences from the assembly. Hanwoo is a major beef cattle breed in Korea and has superior reproductive abilities. The Korean government has attempted to enhance Hanwoo improvement and to increase meat quality as assessed by marbling score, live body weight, carcass weight, and eye-muscle area. Thus, functional characterization of Hanwoo transcripts may be useful for further genetic studies.

Search by : The Database supports various types of queries such as by EST, transcript, markers, QTL, and gene. The user can readily obtain information from each database. For example, the QTL data retrieved from Animal QTLdb include the locus symbol, the chromosomal position, QTL description, and flanking markers defining the borders of the QTL based on LOD score thresholds. These data were stored in a MySQL table, labeled 'QTL'.

Disease browser : The query can be a disease name, OMIM accession ID, or specific text, and the orthologous gene pairs of the human, mouse, and bovine databases are summarized for the disease term. This feature includes extensive links to relevant resources such as the Entrez Gene and OMIM database. Links to the NCBI database provide detailed information about the disease. This web tool can find bovine gene candidates contributing to diseases such as genetic disorders. We predicted 2,609 human-bovine and 2,575 mouse-bovine ortholog candidate disease genes using 3,136 disease terms of the human disease browser in the MGI database.

Gene expression : We also performed statistical analyses to identify tissue-specific genes based on EST information. Bovine expression data based on EST sequences were analyzed for tissue-specific expression using Audic's test (Audic and Claverie, 1997). The contigs selected were those containing five or more ESTs; 1,649 contigs were retained. There were 325, 210, and 238 genes expressed at a significantly higher level based on the number of transcripts in the fat, loin, and liver data sets, respectively. This likely indicates that, as expected, the tissue-specific genes are related to the specific functions of those tissues. Tissue-specific cDNA library sequences (i.e., expressed sequence tags) yield a detailed snapshot of gene expression and are useful in developing second-generation molecular resources such as microarrays for gene expression profiling (Baumann et al., 2005).

QTL Comparative gene map : The querying process in this application is divided into three steps: select a chromosome, find QTL traits of interest, and display the bovine/human genes within a QTL region. The user can find candidate genes for a QTL of interest through this step. The resulting genes within a QTL region are presented together with a brief description in table form. Identification of bovine/human genes related to QTL traits is important in finding significant regions. This information can also be used as part of the input for statistical analyses such as a meta-analysis of QTLs that will identify the most likely QTL peak/locus by considering QTL data for a target trait across various environments (Goffinet and Gerber, 2000; Grisart et al., 2002).

Genome annotation : Clicking on any genomic location opens a genome browser page that provides genomic alignment data corresponding to the region. Users can click on a chromosome and then specify a genomic position. By default, the genomic alignment shows a region of 1 Mb. The genome browser also features zoom (in/out) buttons. Users can adjust the range by clicking on the chromosome or specifying start and end positions. The Search Results page shows all relevant mapping results with available data types, Hanwoo ESTs, contigs, Hanwoo BESs from NCBI, UniGene from NCBI, all ESTs, mRNAs, genScan results, and human genes. Clicking on any bar opens a pop-up window that provides more detailed information. Various types of sequence data are shown by different colored bars (red: Korean cattle ESTs; orange: Korean cattle contigs; yellow: all mRNA; light green: all EST; light blue: Korean cattle BESs; blue: human genes; light purple: GenScan results).

In summary, we describe a new database of bovine genomic information including Hanwoo data. We intend it to be a knowledgebase that integrates basic information, bioinformatics analysis of transcript sequences (mRNA and EST), QTL information based on published data, and OMIM data related to disease. It should be a valuable resource for increasing the coverage of the bovine genome. The database provides sequence mapping and QTL characterization, information that is useful for identifying genes associated with economically important traits, studying functional genomics, and researching in the field of genetic bovine breeding. In the near future, the database will be updated with Korean cattle SNP, marker information, and gene expression data using microarrays, followed by a specialized bovine genome database for Hanwoo.

ACKNOWLEDGMENTS

This work was supported by a grant (200901FHT020710485) from Agenda, Rural Development Administration, Republic of Korea.

REFERENCES

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17):3389-3402.

Audic, S. and J. M. Claverie. 1997. The significance of digital

gene expression profiles. Genome. Res. 7(10):986-995.

Barendse, W., S. M. Armitage, L. M. Kossarek, A. Shalom, B. W. Kirkpatrick, A. M. Ryan, D. Clayton, L. Li, H. L. Neibergs and N. Zhang. 1994. A genetic linkage map of the bovine genome. Nat. Genet. 6(3):227-235.

Baumann, R. G., R. L. Baldwin, C. Pt. Van Tassell, T. S. Sonstegard and L. K. Matukumalli. 2005. Characterization of a normalized cDNA library from bovine intestinal muscle and epithelial tissues. Anim. Biotechnol. 16(1):17-29.

Bishop, M. D., S. M. Kappes, J. W. Keele, R. T. Stone, S. L. Sunden, G. A. Hawkins, S. S. Toldo, R. Fries, M. D. Grosz and J. Yoo. 1994. A genetic linkage map for cattle. Genetics 136(2):619-639.

Goffinet, B., S. Gerber. 2000. Quantitative trait loci: a meta-analysis. Genetics 155(1):463-473.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid and P. Simon. 2002. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res. 12(2):222-231.

Hansen, C., A. Fu, Y. Meng, E. Okine, R. Hawken, W. Barris, C. Li and S. S. Moore. 2004. Gene expression profiling of the bovine gastrointestinal tract. Genome 47(4):639-649.

Hu, J., C. Mungall, A. Law, R. Papworth, J. P. Nelson, A. Brown, I. Simpson, S. Leckie, D. W. Burt and A. L. Hillyard. 2001. The ARKdb: genome databases for farmed and other animals. Nucleic Acids Res. 29(1):106-110.

Kent, W. J. 2002. BLAT--the BLAST-like alignment tool. Genome. Res. 12(4):656-664.

Lim, Dajeong BM-j, Cho Yong-Min, Yoon Doo-hak, Lee Seung-Hwan, Shin Younhee and Im Seok-ki. 2009. Functional analysis of expressed sequence tags from Hanwoo (Korean cattle) cDNA libraries. Journal of Animal Science and Technology 51:1-8.

Panzitta, F., A. Caprera, I. Merelli, L. Milanesi, J. L. Williams, B. Lazzari and A. Stella. 2008. Mining the bovine genome with the “Bovine SNP Retriever”. J. Hered. 99(6):696-698.

Stein, L. D., C. Mungall, S. Shu, M. Caudy, M. Mangone, A. Day, E. Nickerson, J. E. Stajich, T. W. Harris, A. Arva. 2002. The generic genome browser: a building block for a model organism system database. Genome. Res. 12(10):1599-1610.

Wall, D. P., H. B. Fraser, A. E. Hirsh. 2003. Detecting putative orthologs. Bioinformatics 19(13):1710-1711.

Womack, J. E., J. S. Johnson, E. K. Owens, C. E. Rexroad, 3rd, J. Schlapfer, Y. P. Yang.1997. A whole-genome radiation hybrid panel for bovine gene mapping. Mamm. Genome. 8(11):854-856.