# QTLdb: A Comprehensive Database Tool Building Bridges between Genotypes and Phenotypes

*Zhi-Liang Hu*[1], Carissa A. Park, Eric R. Fritz and James M. Reecy

## Introduction

Quantitative Trait Loci (QTL) provides a way to associate segments of genome locations with quantitative traits which represent the majority of economically important phenotypes measured in livestock animals. Over the past 15 years or so, thousands of QTL have been detected in pigs, cattle, chickens, and sheep, among other species. These data allow researchers to narrow down genomic regions and identify the genetic factors that contribute to trait variations (Mehar, 2004; Hocking, 2005; Rothschild *et al*., 2007). However, there is a bottleneck between mapped QTL and gene discovery (Womack, 2005) which must be overcome in order for animal geneticists to use this information in the genetic improvement of livestock.

It is a challenge for researchers to quickly comprehend the large number of QTL data and make most efficient use of these data in a relatively short time. For example, QTL information within the public domain is scattered throughout many publications; each publication is from an independent study with potentially different statistical analysis methods and with different experimental animal populations; the QTL traits are defined and measured in many different ways in different laboratories and/or countries; and so on. Our Pig QTLdb (Hu *et al*. 2005) represented a significant step toward establishing a QTL data repository for housing and comparing QTL within one species. This was a successful attempt in terms of consolidating the wealth of all publicly available QTL results from different laboratories (Rothschild *et al*., 2007). Furthermore, to facilitate gene discoveries using the mapped QTL information, we have expanded the database into the Animal QTLdb to encompass additional livestock species (Hu *et al*., 2007) and added more tools for map alignments of structural genomics information such as radiation hybrid (RH) markers, microarray elements (including oligonucleotide probes and Affymetrix elements), single nucleotide polymorphisms (SNPs), and other types of markers. We have also added a set of data curation web tools. Additionally, the introduction of ontology for improving the consistency of trait nomenclature and organizing traits for easy database management has led to the development of Animal Trait Ontology (ATO; Hughes *et al*., 2008).

Rapid progress of genomic research in the past 20 years has enriched genomic databases with vast amounts of information surrounding the central dogma of biology. However, the genetic information flow from genotypes to phenotypes has not been well represented in currently available public databases. The Animal QTLdb is filling this gap to some extent. In this paper,

[1] Department of Animal Science, Center for Integrated Animal Genomics, Iowa State University, 2255 Kildee Hall, Ames, Iowa, 50011, USA

we report the current progress in the development of the comprehensive database tools built into QTLdb as part of our continued efforts to enhance the genetic information links between genotypes and phenotypes. New functions are continuously added to serve this purpose. The most recent additions include utilization of the generic genome browser (GBrowse; Stein *et al*., 2002) to align gene transcripts, mRNA and annotated gene information from public databases and other types of users' map information to the mapped QTL, and the virtual comparative map (VCMap; Kwitek *et al*., 2010) to facilitate the comparison of QTL across species.

## Materials and methods

**Data and data curation.** The QTLdb is designed to accept either curated public data from journal papers or private laboratory reports that are in the process of publication. More than 50 parameters/data types are subject to collection in reporting a QTL. These data includes QTL location, anchoring marker information, various test statistics, QTL effects, and traits and their measurement information, as reported earlier (Hu *et al*., 2007). We have recently added a number of new data types to enhance our ability to be more inclusive in QTL data collection. These new data types include "association" data for candidate gene or single marker association data; "eQTL" from microarray based QTL scan analysis; "test scale" to differentiate genome-wise, chromosome-wise, comparison-wise and experiment-wise QTL reports; "test model" to indicate epistatic, and maternal or paternal imprinted QTL; new test statistics such as Bayes value and likelihood ratio, etc. We have also added animal breed information for future breed-associated QTL analysis. The backbone maps to record QTL are mainly from USDA-Meat Animal Research Center (MARC; for pigs and cattle) and Wageningen University (for chicken). Reported QTL locations are interpolated to the backbone maps via anchoring markers.

The QTLdb has a three-tiered data curation structure so that curators, editors and database administrators can work together and share responsibilities in a workflow to ensure data quality and smooth process control. A set of new data debugging tools, process control mechanisms, and functions for the ease of use of the tools are also developed through data curation lessons learned in the past few years.

**QTL data transformation**. QTL are mapping features recorded as linkage distances. In order for GBrowse to display QTL and for users to easily port QTL data for customized analysis, we established a process to convert the QTL linkage locations (centimorgan, cM) to their physical locations (mega base-pair, Mbp). The data conversion is a mathematical process built in a Perl script where interpolation or extrapolation is performed with reference to the nearest common anchoring marker locations on both maps.

**Platform and software**. The QTLdb is built on a RedHat Linux platform with MySQL (version 14.12) as the backend relational database and Apache (2.2.13) as the web server. Perl (5.8.8) was used to program the web interface for user-controlled data presentations and

interactive curator tools for data entries. Some lightweight PHP codes were also developed to serve some minor web functions.

## Results and discussion

We have made 10 QTLdb public releases since the database was first established in 2004. As of December 31, 2009, 9,927 QTL from 486 publications have been entered and these QTL represent 996 different traits in four livestock species (Table 1). On average, the database receives 500 hits daily and about 8 GB of data are downloaded annually. The QTLdb has been cited by over 100 peer-reviewed journal articles over the past five years. Currently, the database has 30 registered curators (Cattle: 11; Chicken: 5; Pigs: 14), and data curation is mainly performed by professional curators under the NRSP-8 Bioinformatics Coordinator, and through contributions from volunteer curators. Our goal is to release the new data every 3 months when enough new data has been entered.

**Table 1. Summary of the Animal QTLdb content in terms of the number of QTL in the collection, number of published papers the data were curated from, and number of traits they represent.** [a]

| Species | Number of QTL | Number of publications | Number of Traits |
|---|---|---|---|
| Pig | 5,621 | 237 | 546 |
| Cattle | 2,359 | 142 | 212 |
| Chicken | 1,863 | 92 | 208 |
| Sheep [b] | 84 | 15 | 30 |
| Total | 9,927 | 486 | 996 |

[a.] The data count was as of December 31, 2009. [b.] The sheep QTL data is being curated by Jill Maddox's group at the University of Melbourne, Australia.

The QTLdb design, developments and implementations have been reported previously (Hu *et al*., 2005, 2007a, 2007b). In this paper, we only highlight the main progress made recently.

**GBrowse and map alignments**. We have added more alignment capabilities to the QTLdb since our last report (Table 2). The different types of structural genome features include SNP, microarray elements, RH markers, microsatellites, BAC/FPC etc., and most recently the high density Illumina SNP chip data. The alignment data addition was mainly for cattle and pigs as there is more information available.

While our ultimate goal is to align all possible genome features to QTL, it has been a challenge to do so within the current QTLdb, due to the sheer amount of genome data and the limited capacity of QTLdb to handle the whole genome data. To meet this challenge, we have introduced the Generic Genome Browser (GBrowse; Stein et al., 2002) to house and align complex genome features. Shown in Figure 1 is an example of a GBrowse window displaying bovine chromosome 21 (56-62Mbp) where annotated genes, transcripts, coding regions, QTL,

and bovine 50k SNPs are aligned. GBrowse has been implemented for displaying QTL data for cattle, chicken and pigs. The gene information is directly downloaded from NCBI Entrez database, and the QTL data is converted from QTLdb by translating the linkage distances to physical distances using the nearest common anchor markers information.

**Table 2. Summary of the current status of the map alignments in Animal QTLdb in terms of diverse types of structural genomics information.**

| Species | Gene/ transcript | RH marker | BAC/ FPC | SNP | Illumina [a] SNP Chip | Microarray Elements | | Human map |
|---|---|---|---|---|---|---|---|---|
| | | | | | | Affy | Oligo | |
| Pig | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Cattle | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Chicken | Yes | n/a | Yes | Yes | n/a | n/a | n/a | n/a |

[a.] These features are aligned using GBrowse.

Currently, we are supporting both QTLdb web viewer and GBrowse viewer at the same time. This provides users with options on which to use to best serve their needs. To this aim, we have built into each viewer dynamic hyper-links so that any QTL data point in either viewer can be linked to that in the other view by a simple mouse click. Another advantage with GBrowse is that users can directly upload their own GFF map data for alignment experiments.

**Figure 1**. An example of a GBrowse window displaying bovine chromosome 21 (56-62Mbp) where annotated genes, transcripts, coding regions, QTL, and bovine 50k SNPs are aligned.

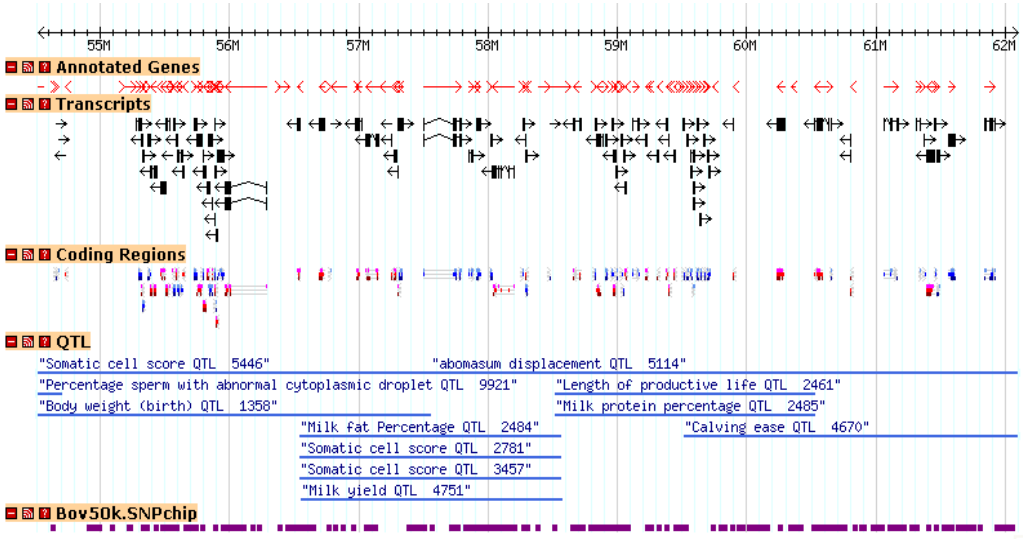
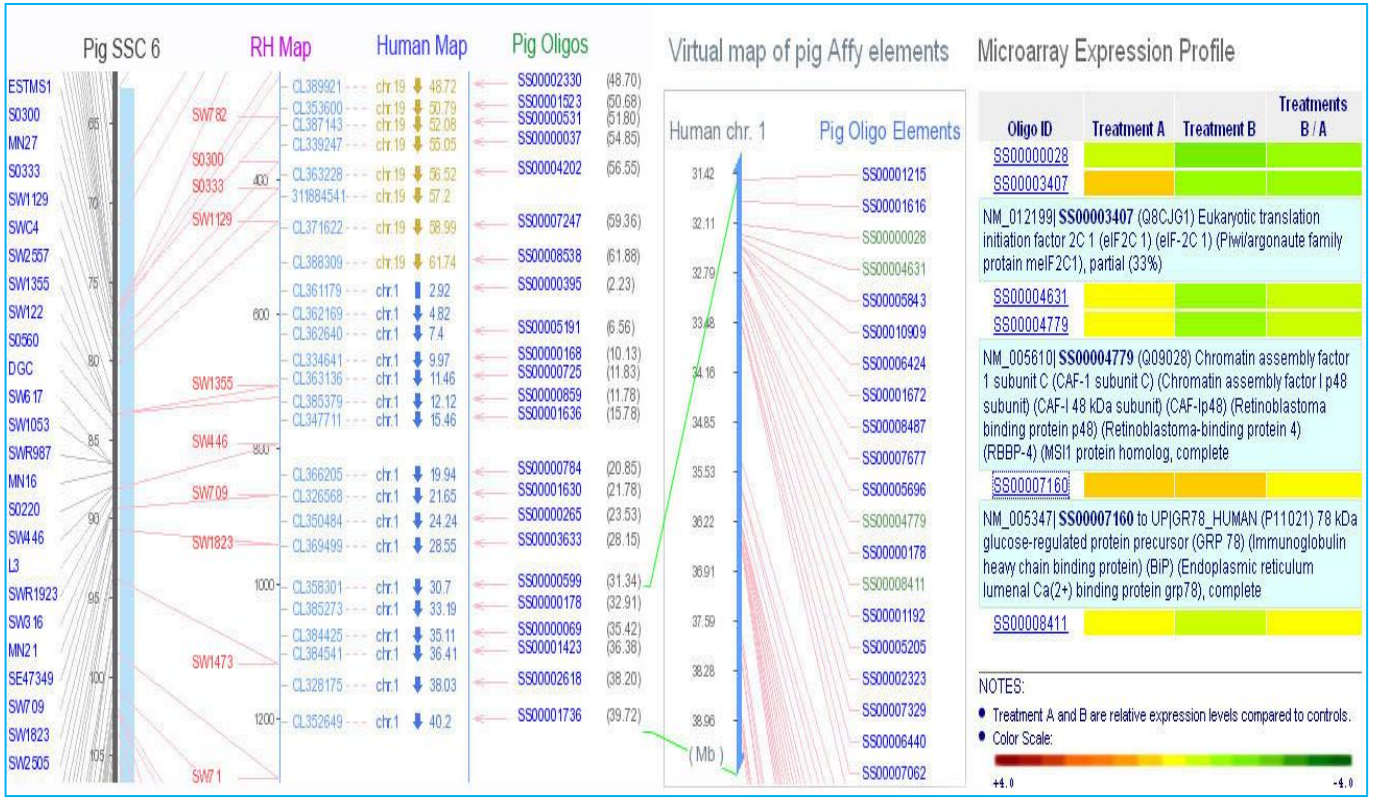


**Figure 2**. A snap shot of pig chromosome 6 showing an interesting QTL region (light blue highlighted portion to the left). The pig chromosome is aligned with human chromosomes 1, 18 and 19 via RH comparative mapping. To the right of the graph are Affymetrix expression profiles [a] linked to their oligo locations on the human genome by BLAST.




[a] The Affymetrix microarray expression data was kindly provided by Shu-hong Zhao of the Huazhong Agriculture University.

One objective for the application of the QTL map alignment was to locate the Affymetrix microarray elements on the QTL map so that expression profiles can be directly viewed against the QTL locations on the chromosome. In Figure 2 is shown a snap shot of pig chromosome 6 aligned with human chromosomes 1, 18 and 19 via RH comparative map. The pig genome sequence assembly was not available when this work was done; therefore the Affymetrix oligo mapping to pig linkage map was done by BLAST alignment of the oligonucleotide probe sequences against the human genome to determine their comparative pig genome locations. It should be pointed out that as the pig genome assembly is becoming available, direct alignment to the genome shall be included to improve the alignment results.

**Data downloads**. QTL data can be directly downloaded from QTLdb by links on its web interface. We have made it convenient for users to customize the data downloads by providing a button on the QTL display page, so that one can download not only the whole data set but also QTL data shown within the users' current view. We also made QTL data available for download in both GFF and tab-delimited formats.

**Data synchronization with NCBI**. When the Pig QTLdb was first developed, we worked with NCBI to transfer our QTL data via automatic submission into the LocusLink database. Since NCBI has phased out LocusLink and developed a new GeneDB within the Entrez database (Maglott et al., 2005), recently we were able to work out a new scheme with NCBI, so that the cattle, chicken and pig QTL data can be synchronized into the GeneDB as Gene Reference Into Function (GeneRIF) records. Having the QTL data in the NCBI GeneDB allows their being automatically matched to updated public sequence data by e-PCR. Synchronization of unique identifiers for each QTL within both QTLdb and GeneDB makes it possible for a QTL in either database to be easily linked to its entry in the other database. Currently, the QTL data synchronization is triggered upon each QTLdb release. A standard operating procedure has been established, and a set of scripts was developed for data debug before the release, to ensure smooth data release and data synchronization after the release.

**Candidate gene and association data**. In the Pig QTLdb, we collected "candidate gene" information as part of QTL data entry. In other words, in a QTL study, a known gene might be closely associated with one of the markers in the QTL scan, or a marker itself may represent a mutation in the neighborhood of a gene. We add the gene information in the "comments" field of the QTL entry as additional information. Now we have expanded the "candidate gene" data collection to include results from association analysis (in the context of QTLdb, we treat it as a special "single point QTL"). As such, significant association results by (anonymous) markers are also subject to collection in this category. To date, there are 864 such cattle data that have been collected. We plan to continue to expand our efforts in this direction, to include similar data to associate genotypes and phenotypes, such as high density SNP chip whole genome association results.

**Enhanced curator tools**. Scientific data curation can be tedious, which can hinder consistent output of good quality work. We have added a few more tools to reduce the work load of

curators, to help curators concentrate better on the scientific part of the job, therefore reducing the error rate. We have added a program to periodically query PubMed for new publications on the targeted subject, and check against the QTLdb to filter out those that have already been curated. This not only helps curators standardize the literature query, automatically bringing potentially useful data into a format for simplified database entry, but also maintains a working list for keeping track of curation progress by allowing curators to manage the list, to share workloads, and to avoid overlap by different curators.

**Use of ontology in the QTLdb**. Ontologies use controlled vocabularies to describe objects and the relationships between them in a formal manner. We have successfully introduced ontology to organize and manage animal traits within QTLdb, which led to the development of the Animal Trait Ontology (ATO; Hughes *et al*., 2008) and more recently Vertebrate Trait Ontology (VT; unpublished data). As we incorporate breed information into the QTLdb, we also plan to use an ontology to manage the breed data, with a vision that this may provide a prototype of a possible future "breed ontology".

**VCMap**. As a collaboration between Iowa State University, the Medical College of Wisconsin, and University of Iowa, Virtual Comparative Map (VCMap, http://bioneos.com/VCMap/; Kwitek *et al*., 2010) is being developed to assist with comparative views of QTL between species. This will be a useful addition to the QTLdb in terms of assisting comparative genome information mining for promising QTL regions, using information from well-studied species such as mouse, rat, or humans.

**Going further**. The QTLdb has been actively developed for over 7 years. We realize this will be a continuous process as new methods to discover genotype/phenotype associations, new types of data for analysis, and new directions to follow in terms of functional genomics studies continue to emerge in the post-genome era. To meet this challenge, we are in the process of adding a number of data types (e.g., eQTL and association results) and parameters (e.g., test scale, test model, and test statistics; see "Materials and Methods" for details) for data collection, and new functions to accommodate the new data. We will continue to improve the QTLdb, with a vision to assist future combined QTL data mining such as meta analysis.

The focus of our QTLdb development has been on expanding its utility and its peripheral tools for better comparison, confirmation, and location of QTL on a chromosome, in order to aid the search for the most probable sites for genes responsible for economically important traits in livestock. More than 90% of the web pages in the QTLdb are dynamically served through CGI programs to present data, draw maps "on the fly" and link each data point to related data resources (internally or externally). Our work also emphasizes that the database tools are built with the capability to allow users to interactively use it, to mine QTL for feasible candidate genes using references from enhanced structural information alignments and refined comparative mapping results.

Current and future initiatives will include extending the utility of Animal QTLdb by integrating QTL information from more species by working with livestock, rat and mouse communities. Continued efforts are also being made to include new types of data, e.g. copy number variations, segmentation duplications, eQTL, genome wide SNP associations, etc., as they become available, in order to strengthen the links between phenotypes and genotypes.

## Conclusion

From a computer science perspective, QTLdb makes use of relational database structure and ontology management methods, with tools that have been developed to create an interactive data repository to help curators and users via worldwide web. From a biology perspective, it is not only a powerful QTL data comparison, data mining and collaboration aid, but it also provides key links among existing genome databases in terms of genetic information flow between genotypes and phenotypes, which is useful for meta-analysis and network analysis in terms of system biology. Our work on the development of Animal QTLdb has provided a useful database tool not only for the research community, but also for the animal production industry and the general public.

## Acknowledgements

We thank Drs. Svetlana Dracheva and Wonhee Jang from NCBI for their assistance with incorporation of QTL data from the Animal QTLdb into the Gene database of the Entrez database family. We also wish to thank Drs. Max Rothschild and John Bastiaansen for their useful discussions in the early stages of Pig QTLdb development.

## References

Hocking, P.M. (2005). *World's Poultry Science Journal*, **61**:215-226.

Hu, Z-L. and Reecy J.M. (2007b). *Mammalian Genome*, Volume **18**, 1-4.

Hu, Z-L., Dracheva, S., Jang, W-H. *et al*. (2005). *Mamm. Genome*, **16**, 792–800.

Hu, Z-L., Fritz E.R, and Reecy J.M. (2007a). *Nucleic Acids Res.*, **35**, D604–D609.

Hughes, L.M., Bao J., Hu Z-L., *et al*. (2008). *Journal of Animal Science*, **86**:1485-1491.

Kwitek, A., Davis S., Shimoyama M., *et al*. (2010). *Plant & Animal Genomes XVIII Conference, January 9-13, 2010, Town & Country Convention Center, San Diego, CA.*

Maglott, D., Ostell J., Pruitt K.D., *et al*. (2005). *Nucleic Acids Res*., **33**: D54–D58.

Mehar, S. Khatkar, Peter C. *et al*. (2004). *Genet Sel Evol*. **36**(2): 163–190.

Rothschild, M.F., Hu Z-L., and Jiang Z-H. (2007). *Int. J. Biol. Sci.,* **3**:192-197.

Stein, L.D., Mungall C., Shu S., *et al*. (2002). *Genome Res.,* **12**: 1599-610.

Womack, J.E. (2005). *Genome Res*., **15**, 1699–1705.