Developmental progress and current status of the Animal QTLdb

Zhi-Liang Hu*, Carissa A. Park and James M. Reecy*

Department of Animal Science, Iowa State University, 2255 Kildee Hall, Ames, IA 50011, USA

ABSTRACT

The Animal QTL Database (QTLdb; http://www. animalgenome.org/QTLdb) has undergone dramatic growth in recent years in terms of new data curated, data downloads and new functions and tools. We have focused our development efforts to cope with challenges arising from rapid growth of newly published data and end users' data demands, and to optimize data retrieval and analysis to facilitate users' research. Evidenced by the 27 releases in the past 11 years, the growth of the QTLdb has been phenomenal. Here we report our recent progress which is highlighted by addition of one new species, four new data types, four new user tools, a new API tool set, numerous new functions and capabilities added to the curator tool set, expansion of our data alliance partners and more than 20 other improvements. In this paper we present a summary of our progress to date and an outlook regarding future directions.

INTRODUCTION

For more than 10 years, the Animal QTL Database (QTLdb; http://www.animalgenome.org/QTLdb) has been a central resource for quantitative trait loci (QTL) and genotype/phenotype association data for an international community of agricultural animal genetic researchers (see Figures 1–3 for curation and usage statistics). According to Google Scholar, the Animal QTLdb has been cited in the literature over 1010 times (as of 5 September 2015). Previously (1–4), we have reported the success of Animal QTLdb tools that allow inter- and intra-species comparisons of QTL/association data and alignments of such data with a number of genome features such as bacterial artificial chromosome (BAC) end sequences, single nucleotide polymorphisms (SNPs), Affymetrix or oligo array elements and the human genome via radiation hybrid (RH) map markers, cytogenetic bands, annotated genes, etc. We also made it possible to narrow down the most likely genomic region for a QTL/association using meta-analysis (4,5).

Our efforts to further develop the database have continued in order to meet the increasing demands of users regarding both data updates and requirements for how the data can be retrieved in a more efficient manner. For example, since our last report on the database in 2013 (4), the total amount of QTL/association data in the QTLdb has increased by greater than 3-fold, reaching 57 414 records (Cattle: 36 693; Chicken: 4676; Horse: 1124; Pig: 13 958; Sheep: 836; Rainbow trout: 127). This is owed both to the diligent efforts of our full-time and volunteer curators and to our newly developed curator tools. A public user interface and private curator/editor environments were created to help improve the user experience. Vigorous efforts have been expended to develop more than 70 new user functions implemented in more than 20 CGI programs. As a matter of fact, a wide spectrum of new developments in Animal QTLdb includes addition of new species, new visualization tools, new data types and new user tools, as well as expanded curator/editor tools. Here we highlight some of these useful improvements.

RECENT DEVELOPMENTS

Since our last report in the January 2013 NAR Database Issue, we have made eight new releases of the database. Each of these releases included both newly curated data and newly added functions (http://www.animalgenome.org/QTLdb/release). Although the dramatic increase in the newly published QTL/association data has posed a challenge for curation, we have kept up with it using improved tools and optimized procedures to accelerate data curation while minimizing the errors introduced during the process. In the meantime, we continue to add new data types to fully embrace the idea of making the Animal QTLdb a central hub for genotype/phenotype annotations of livestock genomes.

New data types

Addition of new data types extends the capability and utility of the Animal QTLdb. We will continue to incorporate data types as they become available.

---
*To whom correspondence should be addressed. Tel: +1 901 212 2820; Fax: +1 515 294 2401; Email: zhu@iastate.edu
Correspondence may also be addressed to James M. Reecy. Tel: +1 515 294 9269; Fax: +1 515 294 2401; Email: jreecy@iastate.edu
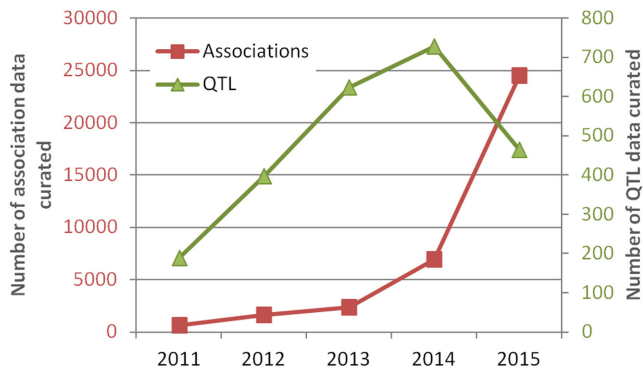
Figure 1. Number of curated QTL and association data per year over the past five years (2015 data are through August).
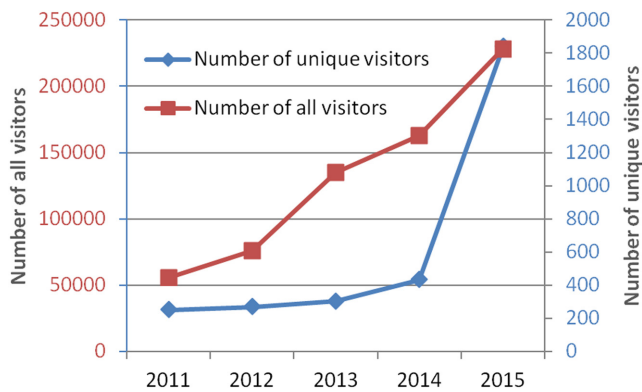

Figure 2. Number of total and unique visitors to the Animal QTLdb over the past five years (2015 data are through August).
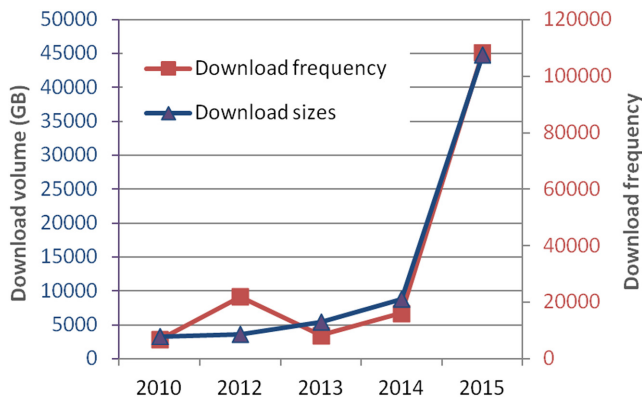

Figure 3. Amount of data downloaded from the QTLdb over the past five years (2015 data are through August).

Signature of selection and copy number variation (CNV) data. Previously, we have added SNP association and eQTL data, and we have now further modified the database to include signatures of selection and CNV data for curation into the QTLdb. In order to accommodate these data types, we needed to add capabilities for the database to record the characteristics of these data. Signatures of selection and CNV are naturally occurring genome features identified only in some animals. They are on somewhat larger scales than SNPs or other types of molecular markers, and

can be structural landmarks for complex traits. Inclusion of such data in QTLdb can be useful to increase the potential for discoveries when combined with genetic networked data analysis. It is our goal to be inclusive of all types of trait mapping features linking phenotypes to genotypes.

Pleiotropic QTL/association data. Pleiotropic QTL data add power and provide more insights for us to understand the genetic architecture of complex traits in situations where multiple traits are involved. To enable pleiotropic QTL curation, links between multiple QTL need to be recorded, along with the evidence for such linkage. Pleiotropic data in the QTLdb will help to illustrate possible interactions between underlying genes and/or genetic and phenotypic correlations between traits.

New species and alignments of linkage maps with their respective genome maps

One advantage of the Animal QTLdb over similar database tools is that the same tools developed for one species can be easily applied to other species and get efficiently reused, therefore maximizing the value of the funding invested in development.

Horse. In 2013, the Horse QTLdb was released. The information required to set up the QTLdb for a new species includes a commonly accepted linkage map, an official genome assembly (optional) and a set of commonly used traits for initial annotations. We have worked with the horse community to combine the two dominant linkage maps, Swinburne (6) and Penedo (7), into a common underlying map as a base map to annotate/curate horse QTL. The horse genome map was set up using build EC_4.0 (8).

Sheep. An official release of the sheep genome assembly became available in late 2012 (version Oar_3.1), which allowed QTL/SNP associations to be visualized against a genome assembly for the first time. With the advent of sheep STS (sequence-tagged site) anchor marker mapping information, we are able to estimate genome coordinates for sheep QTL/association data.

The genome build versions used for each species in the current QTLdb are listed at http://www.animalgenome.org/QTLdb/doc/genome_versions and linked from their mentions throughout the QTLdb web site.

QTL alignments to genome maps

GBrowse *and* JBrowse. Previously, we reported the successful use of GBrowse (9) in the QTLdb for alignment of QTL/association data with other available genome features (4). With the new developments in molecular genomics, it has become more of a challenge to upload large user data files, such as BAM or VCF files, for alignments in GBrowse. To that end, we have implemented usage of the state-of-the-art online alignment tool, JBrowse (10), for users to perform custom quantitative data alignments against QTL/associations, annotated genes and other genome features retrieved from the QTLdb. The format of the alignments includes XYPlot and Density plots. The lightweight

JBrowse greatly helps users avoid internet bottlenecks and perform comparison jobs quickly in the local computing environment.

Development of VT/LPT/CMO ontologies and their mapping to QTL traits

An ontology is a classification methodology that defines a common vocabulary, including the relationships between terms, in a structured way for useful information sharing. In 2013 (4), we reported initial success in applying the Vertebrate Trait Ontology (VT; http://www.animalgenome.org/bioinfo/projects/vt/; (11)), Livestock Product Trait Ontology (LPT; http://www.animalgenome.org/bioinfo/projects/lpt/) and Clinical Measurement Ontology (CMO; http://www.animalgenome.org/bioinfo/projects/cmo/; (12)) to standardize trait classification and comparisons within Animal QTLdb. To date, we have successfully mapped about 98% of animal QTL traits to at least one of the three ontologies. Our goal is to use one or more of the included ontologies to completely represent the livestock traits in the QTLdb. Because animal production traits may be classified in many different ways, the trait terms used to describe the trait may vary due to differences in methods of detection or measurement, scope of description and/or customs. Links to standardized ontology terms provide a common basis for comparison. As the addition of new trait terms to the database is ongoing, we expect the QTLdb ontology development process to continue in the near future.

As a result of these developments, users of the public portal can see new search tools implemented for lookup of QTL/association data by VT, LPT and CMO terms or annotations, along with the reported trait names/descriptions. Users have the option to include or exclude one or more ontologies as search targets. The search results are put in a tabulated form to show how the trait terms are mapped. For interested ontology developers and/or potential collaborators, we also have an option to download the entire mapping table of all three ontologies against the reported QTL/association traits.

New user tools

Chromosome walker. This tool allows the query of neighboring QTL/associations based on an initial user-defined window, which can then be manipulated to observe the change of the QTL/association landscape by a dynamically updated data list. Its utility comes with a companion sliding window, with a user-defined viewing size, slide distance and slide direction (Figure 4). The tool also includes a graphical representation of the current versus the previously selected window positions on a chromosome. Built-in hyperlinks enable users to click to bring the entire set of QTL/association data from a viewing window to visualize alignments in GBrowse. Figure 4 shows an example where a search was targeted for a 50 Mbp window size starting at 45 Mbp on pig chromosome 1, with shift step size at 42 Mbp. This way, users can see neighboring trait data as the window moves. By adjusting the window size, one can narrow down the regions of interest for examination.

Grand search tool. The grand/general search tool (on the top of each species main page and on the top of 'search tools' within each species) was designed to simplify the search process, which previously relied on a number of search windows for input of terms a user may or may not be able to provide, for example, known QTL ID, trait name, key words in a publication, authors, candidate genes, etc. This tool, on the other hand, is made to perform a database-wide search on all aspects of the data, and then return a list of possibilities if there is any match and indicate how many hits there are for each of the above properties. As a result, the tool can help users quickly locate information for exploration or easily find good access points to pursue further data inquiry.

In association with the addition of the grand search tool, several specific search tools were also improved. For example, now the *candidate gene search* and *trait name search* accept partial names/symbols; *publication search* also accepts formatted search strings for a combination of publication properties such as title, author, publication time range, etc. in a predefined format, e.g. 'title:fat authors:Andersson year>2003'. All these tools are located under 'Search Tools' from a QTLdb species page.

Find QTL by genome location and view genome distribution data for a trait. Users often need the ability to quickly access QTL/association data by their genome locations and to obtain an overview of the data based on distribution throughout the genome (Figure 5). The implementation of these tools greatly helps users save time when repeating similar analyses. For best results, these two query tools should be used in combination (follow the 'Search Tools' link on a species main page to locate them in the 'Data Analysis/Query' panel).

User-defined data download tool. In addition to general download where users can download all data from a species for local analysis, now users have the option to download QTL/association data with user-defined scope at several places: (i) from chromosome view window, (ii) from genome location search results window, (iii) from chromosome walk window and (iv) from genome distribution search results window. At each of these pages, the complete QTL/association data set that matches the user query is ready for download. This helps users to filter for the data they need.

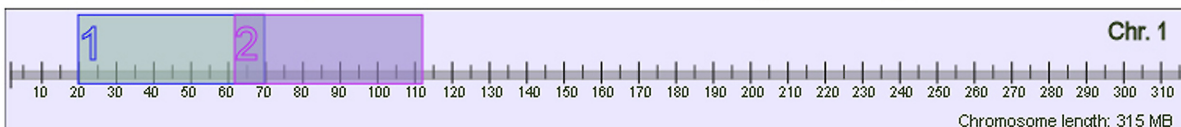### Data alliances under federated database model

Animal QTLdb Data Alliance is a scheme to partner with major community genome databases for data federation. The aim is to take advantage of the strength of each data center and to provide convenience for end users. With newly added members Ensembl (2013) and University of California, Santa Cruz (UCSC) Genome Browser (2014), our current data alliance now has four members: NCBI Entrez GeneDB, Ensembl, Reuters Data Center and UCSC. We have established data stream pipelines to seamlessly update new data to each of these databases in a timely manner upon every QTLdb release. The QTL/association data at each of these sites are integrated into their databases for public access/analysis with their tools. We appreciate the great

Figure 4. Chromosome walker tool, accessible from a main QTLdb species page. Follow the 'Search Tools' link to 'Chromosome Walker' in the 'Data Analysis/Query' panel. Note the data line sections on the graph are trimmed to save space.
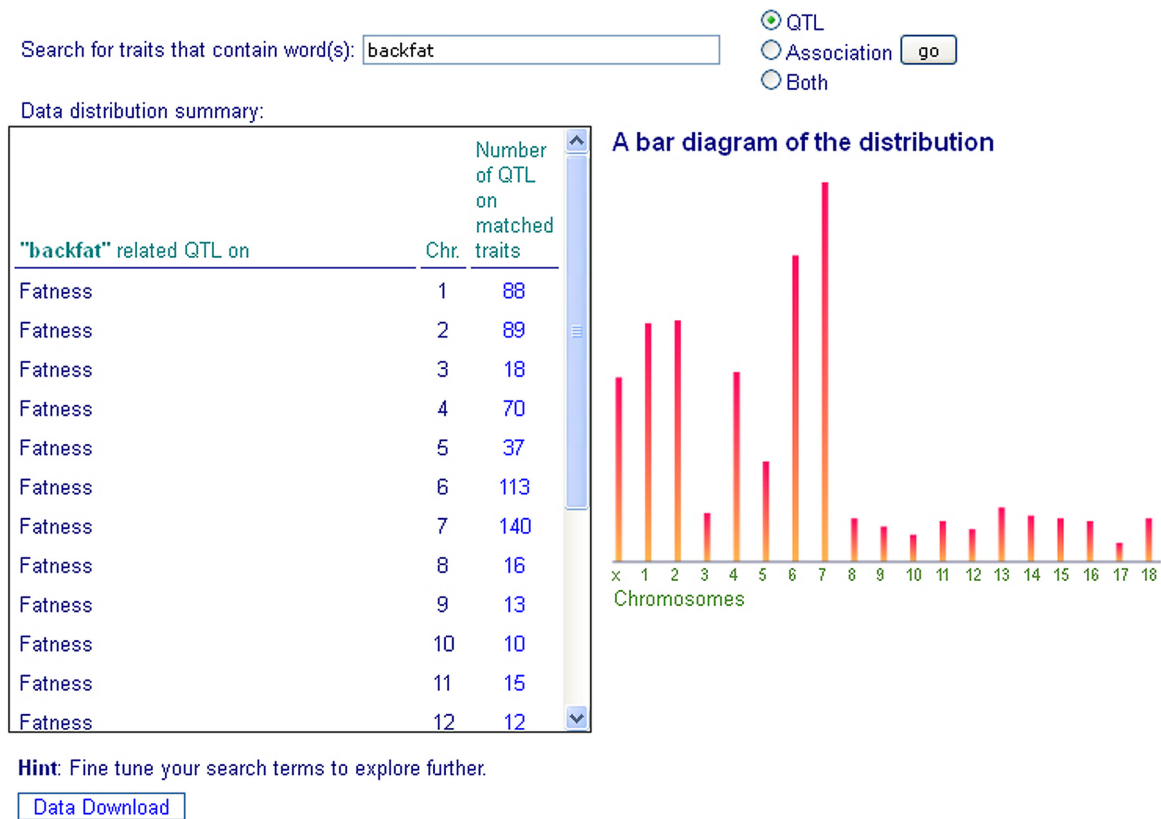
Figure 5. QTL/association data genome distribution tool, accessible from a main QTLdb species page. Follow the 'Search Tools' link to 'View genome distribution data for a trait' in the 'Data Analysis/Query' panel.

synergy in this alliance because users can fully explore the power of their well-developed tools for QTL and association data mining, and we can avoid duplication of effort.

Ensembl. To use, go to http://www.ensembl.org and choose a species to explore, then search for 'phenotype' to list all QTL/association data in both graphical map view and tabulated list view. One can explore further using Ensembl tools.

UCSC. To use, go to https://genome.ucsc.edu/cgi-bin/hgGateway, select 'mammal' (for cattle, horse, pig, sheep) or 'vertebrate' (for chicken), select a genome assembly, and use the search box to search for either locations or trait terms. QTL data display preferences can be found in the 'Mapping and Sequencing Tracks' section.

NCBI. To use, go to http://www.ncbi.nlm.nih.gov/gene and simply put in key words to search. Preferably add species name and use multiple key words to better target the data of interest.

We have also provided options in the 'Search' area for users to directly search any one of the above sites from the Animal QTLdb site.

Application programming interface (API) for programmable access to Animal QTLdb

Active development of the Animal QTLdb has resulted in significant progress in the addition of new database tools and new data. While user feedback has been positive, new demands have also continued to grow. To cope with the constant new challenges we have created an API suite for the Animal QTLdb. This API uses REST architectural framework and XML data mark up language following models used in NCBI's E-utilities (http://www.ncbi.nlm.nih.gov/books/NBK25501/). Access to the API tool is through HTTP on the internet. Our intention is to make the API available for any user with internet access and basic programming ability to make HTTP calls in any programming language.

The API has essentially added scriptable access to the QTLdb, which effectively facilitates more dynamic access to QTLdb data for end users. The main advantage of the API is that it expedites data access and allows implementation of one's own data analysis strategies in real time. The API specifications, sample codes and data representations are publicly available at http://www.animalgenome.org/QTLdb/API/.

Improved curator and editor tools and procedures

Compared to public user tools, the QTLdb curator/editor tools are more actively updated and developed not only to keep up with the need for additional database functions but also to cope with unusual situations from new publications to accommodate smooth and error-free data curation. Here we highlight only a few curator tools that resulted in significant changes to the process of data curation in the QTLdb.

Batch loader. A new challenge in recent years is that many genome-wide association studies (GWAS) have a large number of results appended as tables, and different papers often have different table formats. It makes little sense to manually enter the data row by row during curation, which is prone to introducing errors; on the other hand, it is almost impossible to enforce a unique format only for the sake of data transfer. To this end, we designed a batch loader tool that is both robust in terms of keeping rows and columns in the right locations to ensure proper data identification, and flexible in the way a curator can easily label each column for accurate representation. In addition, we added data identity check protocols to the loader as part of the data quality control procedure within the QTLdb. With this tool, hundreds of rows of data can be easily uploaded for correct curation into the database. This tool has greatly accelerated the process of batch data curation and significantly reduced the backlog of papers in the queue.

External view of pre-release data. More journals now require QTL/association data to be deposited into a database before a paper may be accepted for publication. To accommodate the review of pre-released data (newly curated data that is in 'private' mode) by reviewers/journal editors, we have created a route in the curator account for QTL/association data associated with a reference (paper or report) to be viewable by external people using a unique link produced by the data owner/curator. This helps to expedite the paper review process prior to the release of data when data review is being done within the Animal QTLdb.

Curation of data without official SNP mapping information. Previously, the genome map information was validated only by official SNPs that have dbSNP 'rs' numbers. This was to ensure that the curated genome coordinates remain valid across genome builds (safe transfer of map location to a new assembly). This policy placed many published papers on the waiting list because they used genome map information not represented by official SNPs. We worked around this situation in two ways: (i) when map coordinates are reported on an official genome build but lack official SNP IDs in their study as support, we find official 'rs' SNPs by a location search tool within QTLdb; and (ii) when no official SNP is found to be within a reliable proximity, we allow entry of genome coordinates bundled with genome build version information based on RefSeq accession number of the chromosome (e.g. in the form of `NC_006099.3:10626311-10626312` for a span on chicken chromosome 12). This ensures future reliable transfer of the coordinates to new assemblies. To serve this purpose, we built a local database that ported official dbSNP data to facilitate access to SNP data during the QTL/association data

curation process. The database contains only minimum essential information from dbSNP (this is only for the sake of computing efficiency; the ultimate references to SNPs in the QTLdb still link to dbSNP).

There have also been several minor but important improvements made to the curator/editor tools. These include but are not limited to (i) allowed manual search and import of PubMed references for data curation as a complementary way to introduce new PubMed data besides our automated PubMed search/updates; (ii) added new data review routes and format options for curators/editors to review/update data; (iii) debugged and improved reference-based QTL data status counts and list summary for review/retract data status; and (iv) improved QTL data curation flow route to eliminate ambiguous data status.

Miscellaneous developments

As Animal QTLdb continues to grow and serve as a central repository for livestock QTL/association data, additional improvements are also important to enhance user experience. For example, we have begun to create video tutorials to introduce new tools, and we have been updating the 'Frequently Asked Questions (FAQ)' to address newly emerged problems/concerns/utilities, etc. We use the FAQ as an entry level 'user guide' for end users.

DISCUSSION

While the utility of the Animal QTLdb is evidenced by the growth in number of web page hits, frequency of data downloads and increasing number of citations by scientific publications over the past 10 years, we continue to look ahead and see the potential for Animal QTLdb to be further developed into a resource even more useful to the community, and more robust as part of a federated community database network.

In the earlier days of Animal QTLdb development, we established some rules regarding the essential information needed in order to include a new species in the QTLdb, including a commonly accepted linkage map, an official genome assembly with a set of anchor markers for map alignments with the linkage map and a set of commonly used traits for initial curation annotations, in order to lay out a basic curation template. The basis for this requirement was to make sure that data integrity was maintained throughout data curation, linking and queries in terms of representation of genetic information linking genotypes to phenotypes. We soon realized that such predetermined rules may not be universally applied because they might rule out some data which could prove useful for painting a complete genetic picture. For example, in the case of eQTL studies, trait information may or may not be available; however, expression data links to genes, which indirectly link to phenotypes.

Our successful development efforts are partially driven by interaction with users. Many of the functions we have developed arose from helping users with their specific needs while recognizing that the outcome could have general utility to more users. We highly value our interaction with users, since it serves a dual purpose of confirming that our work

is directly useful for the community while ensuring the new utility is introduced and quickly pushed through the testing and validation process. We also appreciate data curation contributed by users. Although 15 volunteers have curated only 1.14% of the 57 414 data entries as of August 2015, we will continue to improve curation tools for volunteers and encourage their contributions.

Among our efforts, we also successfully set up a mirror site in 2013 at the Huazhong Agricultural University in Wuhan, China. It significantly helped to ease our server/network load from large numbers of visitors from Asian countries. Although it was unsustainable due to remote network administration bottlenecks through local network authorities, it did provide a good model for global collaborations and sharing of resources. We hope to leave the opportunity open for potential new mirror site hosts around the world.

User tools are key to making a database different from a static data repository by providing functions not only to present data in a dynamic and innovative manner but also to facilitate data mining, analysis and discovery. Although we have been successful in our 11 years of development of the Animal QTLdb, we consider our work a starting point for something better. Our focus in the future will be to quickly identify users' needs and determine what is lacking in community databases to come up with possible solutions in a timely manner.

AVAILABILITY

The database contents and online tools are all freely available at http://www.animalgenome.org/QTLdb/.

ACKNOWLEDGEMENTS

REFERENCES

1. Hu,Z.L., Dracheva,S., Jang,W., Maglott,D., Bastiaansen,J., Rothschild,M.F. and Reecy,J.M. (2005) A QTL resource and comparison tool for pigs: PigQTLDB. *Mamm. Genome*, **16**, 792–800.
2. Hu,Z.L., Fritz,E.R. and Reecy,J.M. (2007) AnimalQTLdb: a livestock QTL database tool set for positional QTL information mining and beyond. *Nucleic Acids Res.*, **35**, D604–D609.
3. Hu,Z.L., Park,C.A., Fritz,E.R. and Reecy,J.M. (2010) QTLdb: A comprehensive database tool building bridges between genotypes and phenotypes. In: *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production*, Leipzig.
4. Hu,Z.L., Park,C.A., Wu,X.L. and Reecy,J.M. (2013) Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. *Nucleic Acids Res.*, **41**, D871–D879.
5. Hu,Z., Wu,X. and Reecy,J. (2011) QTL meta-analysis on the fly. *ACM Conference on Bioinformatics, Computional Biology, and Biomedicine (ACM-BCB)*, Chicago.
6. Swinburne,J.E., Boursnell,M., Hill,G., Pettitt,L., Allen,T., Chowdhary,B., Hasegawa,T., Kurosawa,M., Leeb,T., Mashima,S. *et al.* (2006) Single linkage group per chromosome genetic linkage map for the horse, based on two three-generation, full-sibling, crossbred horse reference families. *Genomics*, **87**, 1–29.
7. Penedo,M.C., Millon,L.V., Bernoco,D., Bailey,E., Binns,M., Cholewinski,G., Ellis,N., Flynn,J., Gralak,B., Guthrie,A. *et al.* (2005) International Equine Gene Mapping Workshop Report: a comprehensive linkage map constructed with data from new markers and by merging four mapping resources. *Cytogenet. Genome Res.*, **111**, 5–15.
8. Wade,C.M., Giulotto,E., Sigurdsson,S., Zoli,M., Gnerre,S., Imsland,F., Lear,T.L., Adelson,D.L., Bailey,E., Bellone,R.R. *et al.* (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science*, **326**, 865–867.
9. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
10. Skinner,M.E., Uzilov,A.V., Stein,L.D., Mungall,C.J. and Holmes,I.H. (2009) JBrowse: a next-generation genome browser. *Genome Res.*, **19**, 1630–1638.
11. Park,C.A., Bello,S.M., Smith,C.L., Hu,Z.L., Munzenmaier,D.H., Nigam,R., Smith,J.R., Shimoyama,M., Eppig,J.T. and Reecy,J.M. (2013) The Vertebrate Trait Ontology: a controlled vocabulary for the annotation of trait data across species. *J. Biomed. Semantics*, **4**, 13.
12. Shimoyama,M., Nigam,R., McIntosh,L.S., Nagarajan,R., Rice,T., Rao,D.C. and Dwinell,M.R. (2012) Three ontologies to define phenotype measurement data. *Front. Genet.*, **3**, 87.