Development of Animal QTLdb and CorrDB: Resynthesizing Big Data to Improve Meta-analysis of Genetic and Genomic Information

Z.-L. Hu, C.A. Park, & J.M. Reecy[*]

*Department of Animal Science, Iowa State University, 2255 Kildee Hall, Ames, Iowa 50011, USA. *Corresponding email: jreecy@iastate.edu*

Summary

In the age of big data, biological databases must undergo rapid development of their data infrastructure in order to effectively accommodate abundant data in a structured manner to improve metadata analysis. The livestock genetic and phenotypic correlation data from studies carried out in the past 70+ years, and the quantitative trait loci (QTL) mapping results from studies over the past 25+ years, provide a huge amount of information to add new types of annotations to animal genomes. The growth of Animal QTLdb and CorrDB over the past decade provides valuable tools for researchers to utilize a wealth of historical and future phenotype/genotype data to elucidate the genetic mechanisms behind livestock production improvements. Our recent efforts in extensive data curation, data quality maintenance, new web tool developments, and collaborative database expansions provide convenient platforms for data queries and analysis to serve the phenotype/genotype data collection needs of the livestock genetics/genomics community. Through the course of over 13 (QTLdb) and 5 (CorrDB) years of development, applications developed for Animal QTLdb and CorrDB have embraced the big data era when metadata analysis started to demonstrate its power and utility in terms of resynthesis of metadata for improved genetic analysis. To date, there have been 136,137 QTL/association data curated from 1,881 journal articles that represent 1,890 different traits in 6 livestock animal species. We use a strategy to map all QTL/correlation trait data to ontology terms so that they can be linked by underlying information networks. By developing trait-centric and gene-centric views of the QTL/association data, vast amounts of phenotype/genotype data can now be summarized in helpful new ways. In addition, we continue to expand the types of data collected for inclusion. The most recent addition is to include "supplementary data," e.g., original genotypes, phenotypes, near-significant or other association/QTL data from the same experiment that may not be part of official publications. The inclusion of such data may add value to the "big data" pool when meta-analyses are conducted. The most critical developmental work, although not obvious to the public, is the improvement of curation tools and workflow to improve data quality control and maintenance. For example, we have added several new data status types, as well as corresponding procedures to better manage data flow within the curator/editor pipelines. The goals of our ongoing database development are not only to facilitate data collection, curation, and annotation, but also to provide mechanisms to support new types of data reanalysis, combined analysis, and data mining that may lead to new discoveries.

Keywords: livestock, qtldb, corrdb, phenotype, genotype, trait, ontology, database, big data, curation.

Introduction

The rapid progress of animal genomic studies has introduced a wealth of genome information thanks to advancements in new technologies. Whole-genome assemblies for cattle (Elsik et al, 2009; Zimin et al, 2009), chicken (ICGSC, 2004), pig (Groenen et al., 2012), sheep (Jiang et al., 2014), horse (Wade et al, 2009), catfish (Liu et al., 2016), rainbow trout (Berthelot et al., 2014), and other agricultural species have been made available within the last decade, and efforts are ongoing to improve the quality of these assemblies and to functionally annotate gene information to these genomes (FAANG Consortium, 2015). The newly available genome information, upon completion of functional annotation, provides a powerful tool for investigation of the genetic mechanisms behind prolific production in livestock animals. The genetic and phenotypic correlation data from studies carried out in the past 70+ years, and the quantitative trait loci (QTL) mapping results from studies over the past 25+ years, provide a huge amount of information to add new types of annotations to the genomes (Sharma et al., 2015) and to help researchers, livestock producers, and other stakeholders dissect the genetics underlying phenotypic variation. The ongoing development of Animal QTL Database (QTLdb) and Animal Trait Correlation Database (CorrDB) provides excellent tools to facilitate this process. In fact, an average of over half a million annual web visits are made to Animal QTLdb (2010-2017), and a total of 1,450 citations of Animal QTLdb can be found in Google Scholar (Google Scholar Citations, 2017).

QTL/associations represent chromosomal regions linked to complex traits by genetic and association analysis of polymorphic genetic markers against variations in phenotypic trait measurements. Phenotypic and genetic correlations represent co-variations between two traits with regard to trait performance records and their genetic values. We developed Animal QTLdb and CorrDB to house all relevant published data, initially with two primary functions: as a centralized database for easy data retrieval, and as a platform for data comparisons on data collected across different experiments, methods, geographic locations, and times (Hu et al, 2005; Hu and Reecy, 2007, Hu et al., 2013). Through the course of over 13 (QTLdb) and 5 (CorrDB) years of development, applications developed for Animal QTLdb and CorrDB have embraced the big data era when metadata analysis started to demonstrate its power and utility in terms of resynthesis of metadata for updated genetic analysis (Wu et al., 2011; Hu et al., 2013). This brought many opportunities, as well as challenges, to our developmental work to meet user demand for QTL/association/correlation data in an abstractive, detail-ready format, subject to queries for linked genomic data (genome locations, genes, related genome features and variations, related studies, etc.), to provide global networked views of genotypic and phenotypic data.

In this report, we present our most recent progress in development of Animal QTLdb and CorrDB, with a focus on synergistically reusing developed components, combining functionalities, co-developing modules that integrate resources, and most importantly, providing possibilities for users to examine QTL/association-based data in a networked fashion for genetic analysis.

New Developments

Continued curation of QTL/association/correlation data, data entry standards, and database growth

The number of curated data points in the Animal QTLdb has undergone exponential growth over the past 13 years (Figure 1). To date, there have been 136,137 QTL/associations curated from 1,881 journal articles that represent 1,890 different traits in 6 species (Figure 1). The

SNP association data has undergone the largest increase by data type (Figure 2).

Database curation is a demanding job not only in terms of the sheer volume of data processed with minimal errors, but also for the responsibility to maintain the entered data in good standing for its lifetime in the database. In addition to a number of tools to facilitate quality curation workflow (Hu et al, 2016), we have also developed a minimum required information list for Animal QTLdb data entry (https://www.animalgenome.org/QTLdb/doc/minfo). This helps to minimize the gaps between collaborative curation efforts. Besides the internal curation protocols and data flow framework, we also provide steps for data authors to submit their new data in batches through web tools where they can take ownership to manage their data, and get their data integrated into the data flow (https://www.animalgenome.org/QTLdb/doc/batchdata).

Integration of VT/PT/CMO ontologies and their mapping to QTL/correlation traits

In animal science, variations of trait names exist in many forms. It is therefore necessary in database development to manage them using controlled vocabularies, especially when traits are the focus of scrutiny. The use of biological ontologies serves this purpose. The data curation process involves associating standardized trait symbols and names with the best matches to respective trait ontology terms while ideally maintaining the ability to search for the name reported in the literature. We have reported earlier (Hu et al., 2016) the integrated use of Vertebrate Trait Ontology (VT; https://www.animalgenome.org/bioinfo/projects/vt/; Park et al., 2013), Livestock Product Trait Ontology (LPT; https://www.animalgenome.org/bioinfo/projects/lpt/), and Clinical Measurement Ontology (CMO; https://www.animalgenome.org/bioinfo/projects/cmo/; Shimoyama et al., 2012) to manage traits within Animal QTLdb. Now we have expanded the use of these three ontologies to manage traits for the Animal CorrDB as well. In order to do so, we developed a trait mapping tool (Figure 3) to allow traits within the CorrDB to be mapped to one or more of the three ontologies. The links between trait terms in Animal QTLdb and CorrDB can be established through their mutual mapping to the corresponding VT/LPT/CMO ontology terms. Figure 4 and Figure 5 show two screenshots, where correlation data, when available, is provided on a QTL/association data sheet listed by traits, and QTL/association data, when available, is provided on a correlation data table view.

Incorporation of new types of traits – those “modified” with additional attributes

A new challenge we have undertaken is how to accommodate the trait naming method seen in most correlation reports, where a trait name can be conceptually “modified” into a slightly different one depending upon trait measurement conditions. For example, pig litter size can be measured at the sow’s first parity (“first parity litter size”) or second parity (“second parity litter size”), and backfat thickness may be measured by ultrasound or ruler, as well as at different (rib) locations along the back, etc. In this way the same trait may have different forms thus may be treated as a “new” trait for the sake of comparison. Yet, in ontology terms, these are actually the same trait. To cope with this situation, we are currently developing a system to allow “modifiers” to alter traits based on measurement methods, time, anatomical locations, etc., while still retaining the original definition for ontology data management. We call this type of modified traits “trait variants.” To this end, we have been developing a list of modifiers with controlled vocabulary (Figure 6). We then append these modifiers to a trait, making them distinguishable from other variants, yet retaining the root of the term as belonging to a general trait in the context of ontology (Figure 7).

Development of gene-centric and trait-centric views of QTL/association data

With the exponential growth of the data in QTLdb, it is necessary to facilitate users' ability to quickly extract relevant genotype-phenotype data for human-consumable analysis. To achieve this goal, we have developed tools to digest the data into gene-centric and trait-centric views. Figure 8 and 9 are two screenshots, showing how gene-centric and trait-centric views of the QTL/association data are displayed. For example, upon user query, a gene list is given with summaries of gene name, symbol, and any other known names. Dynamic links are embedded for each gene to be expandable to NCBI GeneDB for further detailed information on the gene, while a summary is given for the number of QTL/association data that are associated with each gene. This gives the user options, when needed, to open up the QTL/association data list for exploration or for download. Likewise, the traits on a trait-centric view have direct links with VT/LPT/CMO trait ontology when such mapping is available.

Additional supplementary data associated with published data

"Supplementary data" provides supportive evidence to curated QTL/associations that includes supplementary information useful for combined or other types of meta-analysis in the future but may not be essential data for a current publication. We have modified the database structure to allow supplementary data from a publication to be uploaded "as is," and data file information is annotated to the released data. These data include original genotypes, phenotypes, near-significant or other association/QTL data from the same experiment, supplementary to the curated data. Currently this type of data is not directly available to the public but will be available upon request. The inclusion of these data is part of our efforts to host more complete data collections to fuel future metadata analysis.

Improved curation tools/procedures and data release procedures

Over the years, we have developed a 17-point checklist for data quality control and review, 6-point data release operations, and 5-point post-release operations for each data release cycle (https://www.animalgenome.org/QTLdb/doc/protocols/release). While we have automated these steps as much as possible, human attention is required for many of the operations, because new situations arise and automated scripts would need to be altered as part of the routine. This also provides a mechanism of feedback for curation flow improvements.

Since our last report (Hu et al., 2016), several new curation procedures/protocols have been developed. Data may now be “re-tracked” for valid reasons, placed “on hold” when verification is needed for contradicting or confusing information, “suspended” temporarily when problems occur, or “obsoleted” when data is invalidated. Under certain circumstances we may conditionally release data, for example, to pre-release a set of data to meet publication requirements for authors.

Future Developments

The future developments of both databases will be focused on expansion of data/structure integrations under a federated database concept, with emphasis on collaborative efforts and shared resources and outcomes.

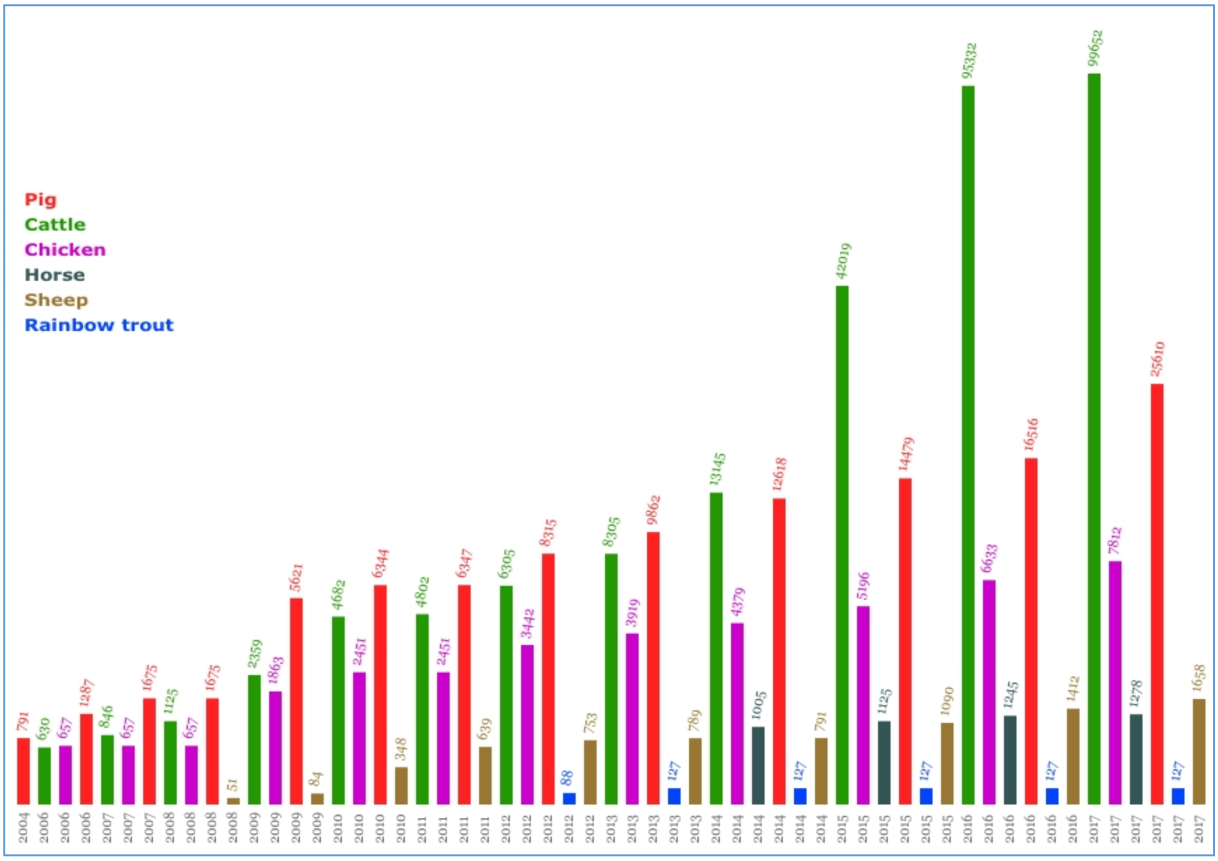


Figure 1. Growth of curated data in the Animal QTLdb by year and by species, represented by data point counts released each year on the Animal QTLdb website. Note that all data are LOG transformed so that the bar graphs are readable in a reasonable window size.

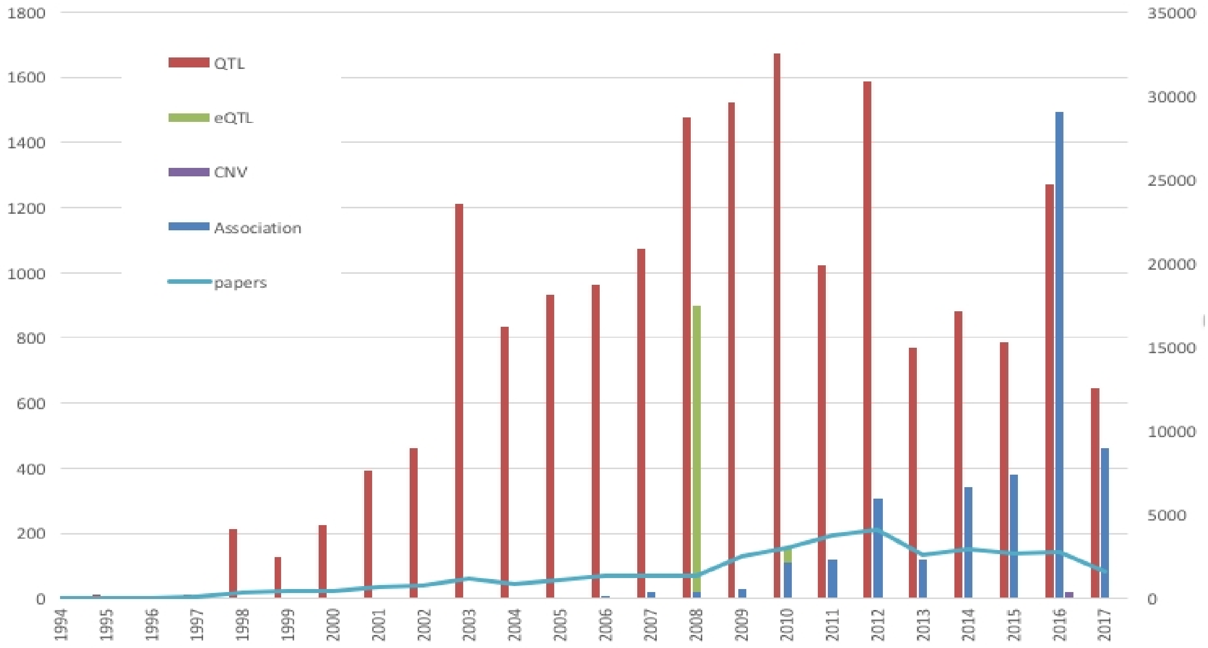
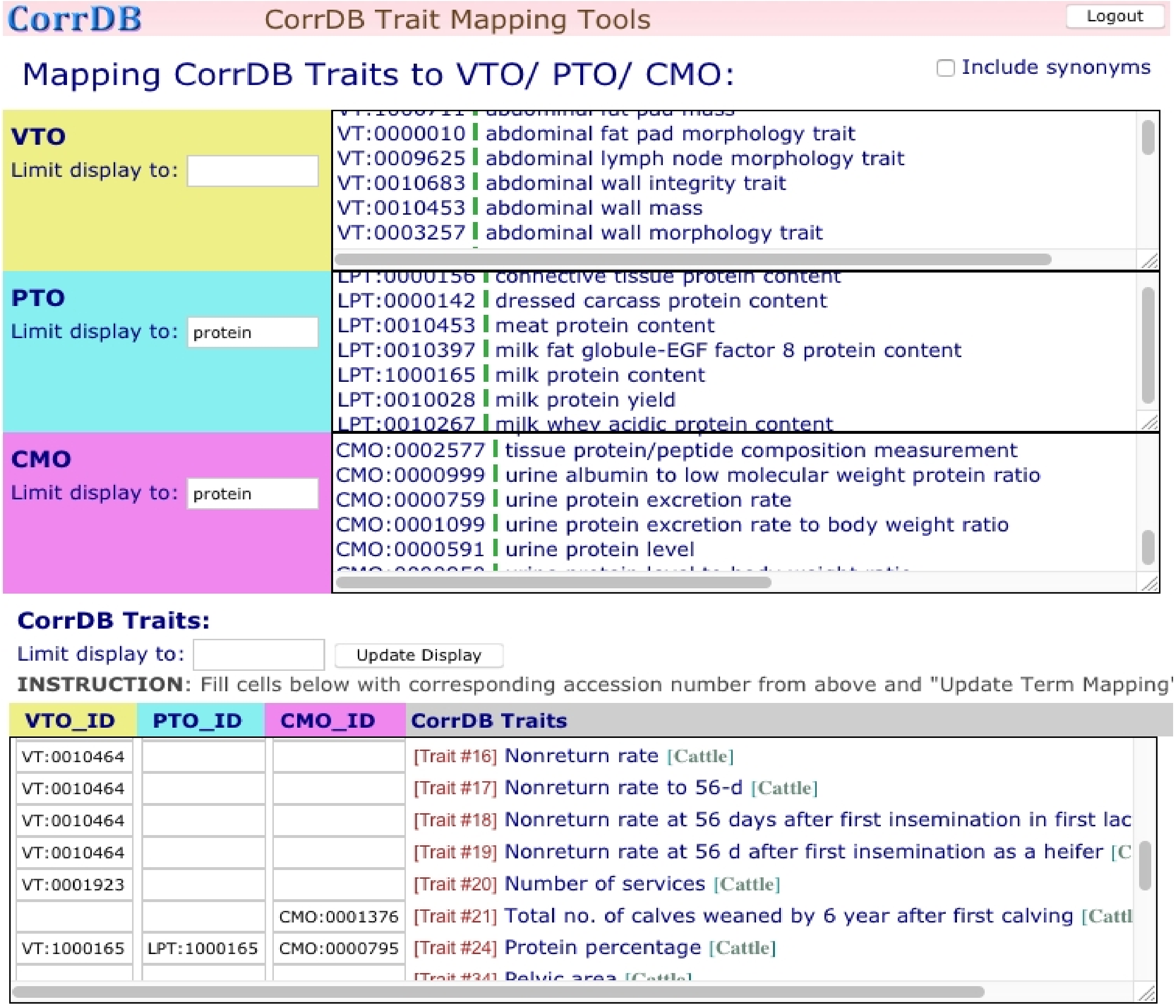


Figure 2. Animal QTLdb data growth by data type. The association data counts are on the right axis, while the rest of the data type counts are on the left axis.



Mapping CorrDB Traits to VTO/ PTO/ CMO:

☐ Include synonyms

VTO
Limit display to:

VT:0000010 abdominal fat pad morphology trait
VT:0009625 abdominal lymph node morphology trait
VT:0010683 abdominal wall integrity trait
VT:0010453 abdominal wall mass
VT:0003257 abdominal wall morphology trait

PTO
Limit display to: protein

LPT:0000156 connective tissue protein content
LPT:0000142 dressed carcass protein content
LPT:0010453 meat protein content
LPT:0010397 milk fat globule-EGF factor 8 protein content
LPT:1000165 milk protein content
LPT:0010028 milk protein yield
LPT:0010267 milk whey acidic protein content

CMO
Limit display to: protein

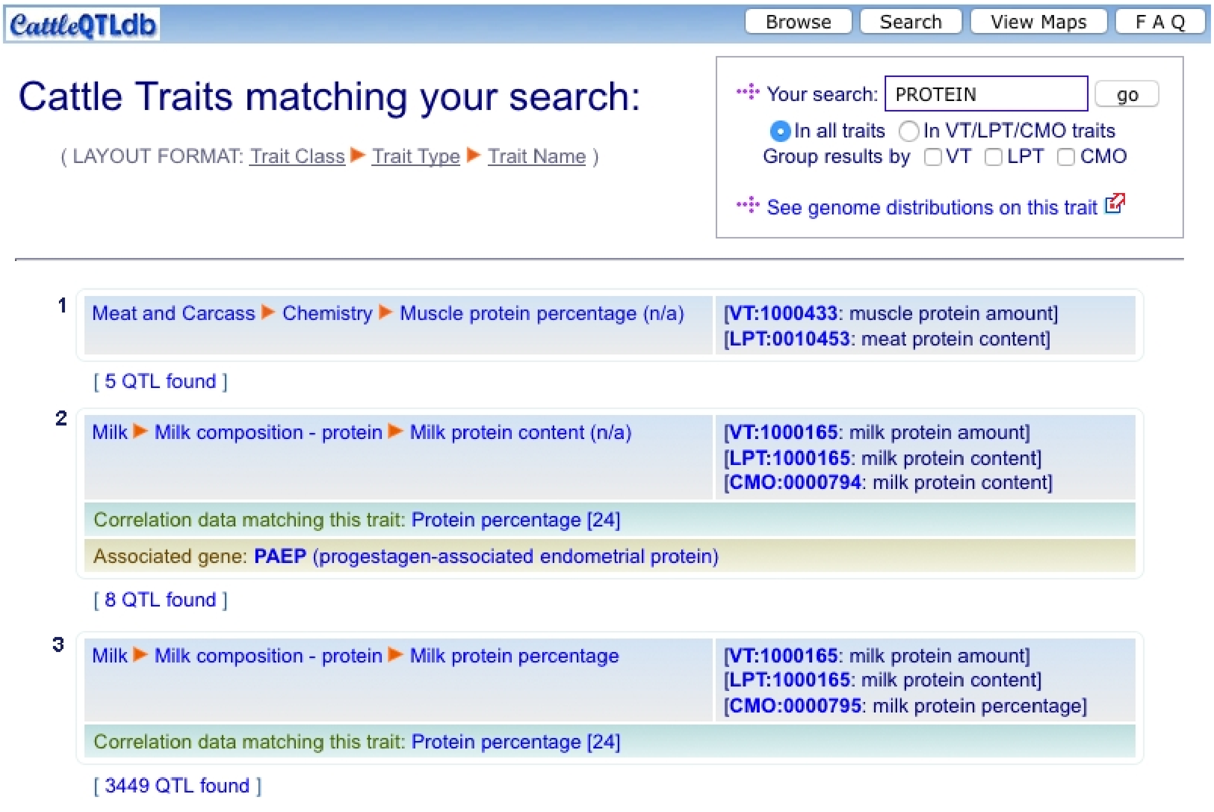
CMO:0002577 tissue protein/peptide composition measurement
CMO:0000999 urine albumin to low molecular weight protein ratio
CMO:0000759 urine protein excretion rate
CMO:0001099 urine protein excretion rate to body weight ratio
CMO:0000591 urine protein level

CorrDB Traits:
Limit display to:

INSTRUCTION: Fill cells below with corresponding accession number from above and "Update Term Mapping'

VTO_ID	PTO_ID	CMO_ID	CorrDB Traits
VT:0010464			[Trait #16] Nonreturn rate [Cattle]
VT:0010464			[Trait #17] Nonreturn rate to 56-d [Cattle]
VT:0010464			[Trait #18] Nonreturn rate at 56 days after first insemination in first lac
VT:0010464			[Trait #19] Nonreturn rate at 56 d after first insemination as a heifer [C
VT:0001923			[Trait #20] Number of services [Cattle]
		CMO:0001376	[Trait #21] Total no. of calves weaned by 6 year after first calving [Cattl
VT:1000165	LPT:1000165	CMO:0000795	[Trait #24] Protein percentage [Cattle]

Figure 3. A screenshot of the trait mapping tool used in the CorrDB. The tool allows comparative viewing of three ontologies against one (CorrDB) trait set, so that the best term can be chosen and ontology development needs evaluated.



Cattle Traits matching your search:

(LAYOUT FORMAT: Trait Class ▶ Trait Type ▶ Trait Name)

Your search: PROTEIN

● In all traits ○ In VT/LPT/CMO traits

Group results by ☐ VT ☐ LPT ☐ CMO

1 Meat and Carcass ▶ Chemistry ▶ Muscle protein percentage (n/a) | [**VT:1000433**: muscle protein amount] [**LPT:0010453**: meat protein content]

[5 QTL found]

2 Milk ▶ Milk composition - protein ▶ Milk protein content (n/a) | [**VT:1000165**: milk protein amount] [**LPT:1000165**: milk protein content] [**CMO:0000794**: milk protein content]

Correlation data matching this trait: Protein percentage [24]

Associated gene: **PAEP** (progestagen-associated endometrial protein)

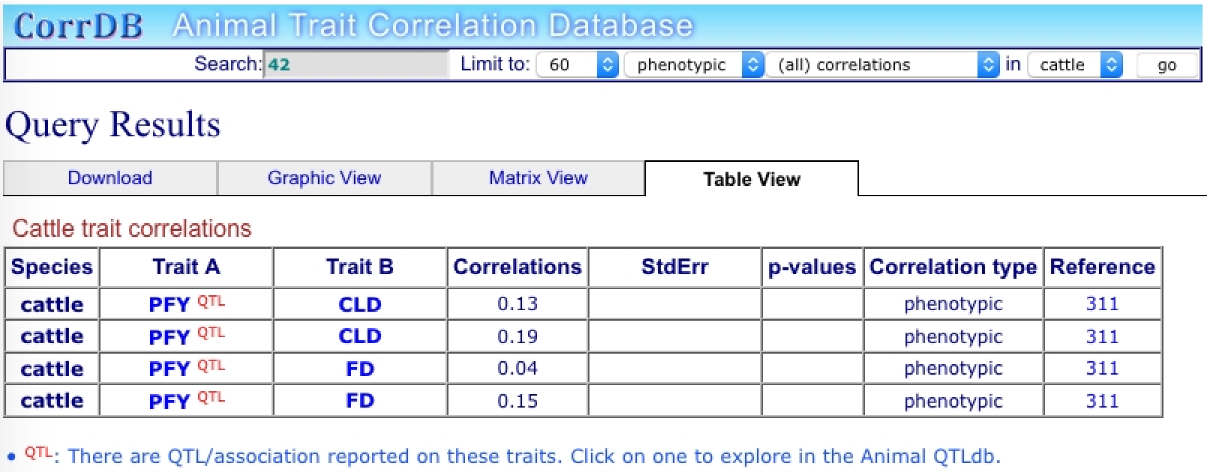
[8 QTL found]

3 Milk ▶ Milk composition - protein ▶ Milk protein percentage | [**VT:1000165**: milk protein amount] [**LPT:1000165**: milk protein content] [**CMO:0000795**: milk protein percentage]

Correlation data matching this trait: Protein percentage [24]

[3449 QTL found]

Figure 4. A QTL/association data view showing links to CorrDB where they exist (highlighted in light green).



Search: 42 Limit to: 60 phenotypic (all) correlations in cattle

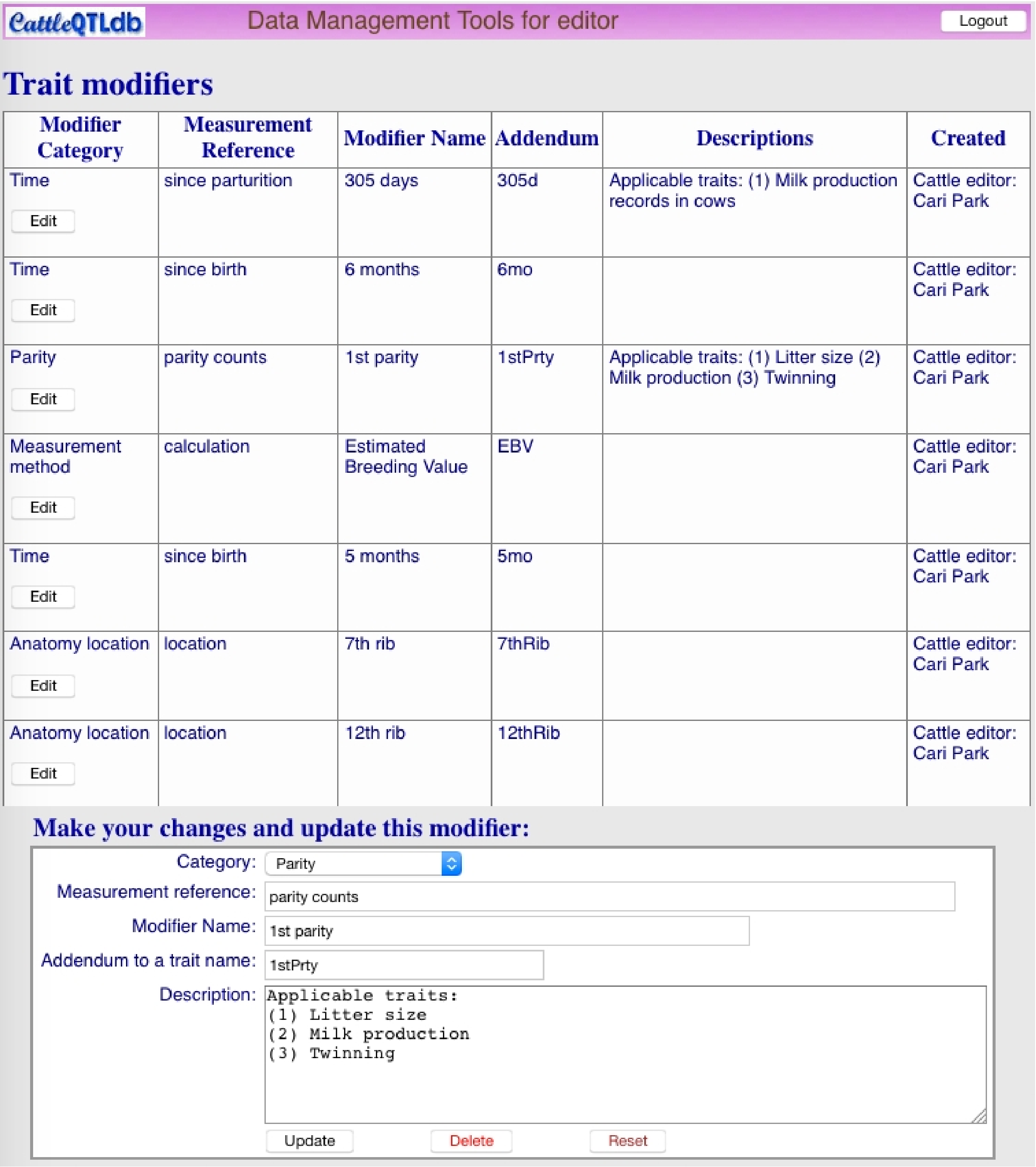
Query Results

Cattle trait correlations

Species	Trait A	Trait B	Correlations	StdErr	p-values	Correlation type	Reference
cattle	**PFY** QTL	**CLD**	0.13			phenotypic	311
cattle	**PFY** QTL	**CLD**	0.19			phenotypic	311
cattle	**PFY** QTL	**FD**	0.04			phenotypic	311
cattle	**PFY** QTL	**FD**	0.15			phenotypic	311

- QTL: There are QTL/association reported on these traits. Click on one to explore in the Animal QTLdb.

Figure 5. A CorrDB view of correlations showing traits with links to existing QTL/association data.



Trait modifiers

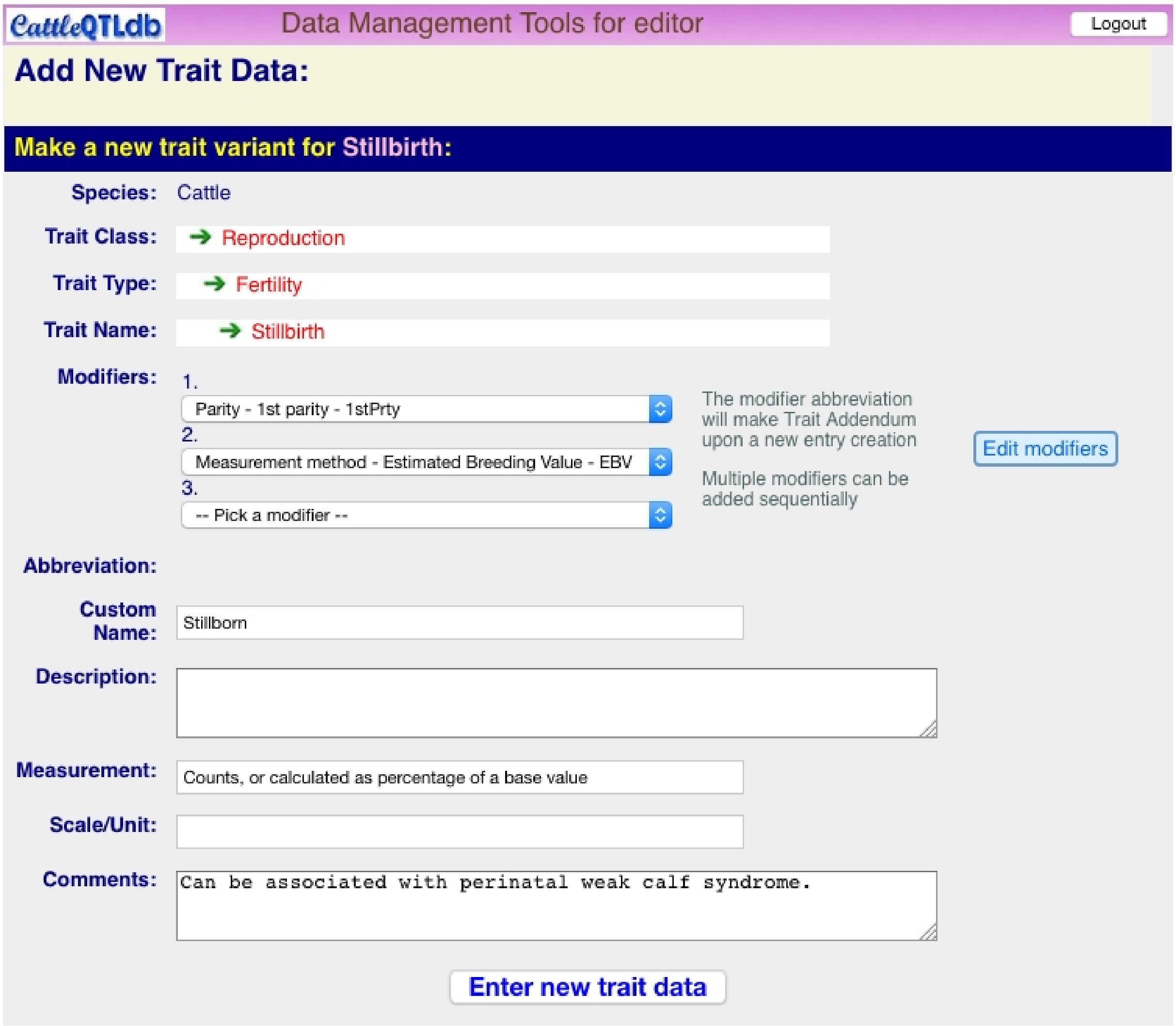
Modifier Category	Measurement Reference	Modifier Name	Addendum	Descriptions	Created
Time Edit	since parturition	305 days	305d	Applicable traits: (1) Milk production records in cows	Cattle editor: Cari Park
Time Edit	since birth	6 months	6mo		Cattle editor: Cari Park
Parity Edit	parity counts	1st parity	1stPrty	Applicable traits: (1) Litter size (2) Milk production (3) Twinning	Cattle editor: Cari Park
Measurement method Edit	calculation	Estimated Breeding Value	EBV		Cattle editor: Cari Park
Time Edit	since birth	5 months	5mo		Cattle editor: Cari Park
Anatomy location Edit	location	7th rib	7thRib		Cattle editor: Cari Park
Anatomy location Edit	location	12th rib	12thRib		Cattle editor: Cari Park

Make your changes and update this modifier:

Category: Parity
Measurement reference: parity counts
Modifier Name: 1st parity
Addendum to a trait name: 1stPrty
Description:

```
Applicable traits:
(1) Litter size
(2) Milk production
(3) Twinning
```

Figure 6. A QTLdb editor window showing how "modifier" attributes are managed with controlled vocabulary and context.



Add New Trait Data:

Make a new trait variant for Stillbirth:

Species: Cattle

Trait Class: → Reproduction

Trait Type: → Fertility

Trait Name: → Stillbirth

Modifiers:
1. Parity - 1st parity - 1stPrty
2. Measurement method - Estimated Breeding Value - EBV
3. -- Pick a modifier --

The modifier abbreviation will make Trait Addendum upon a new entry creation

Multiple modifiers can be added sequentially

Abbreviation:

Custom Name: Stillborn

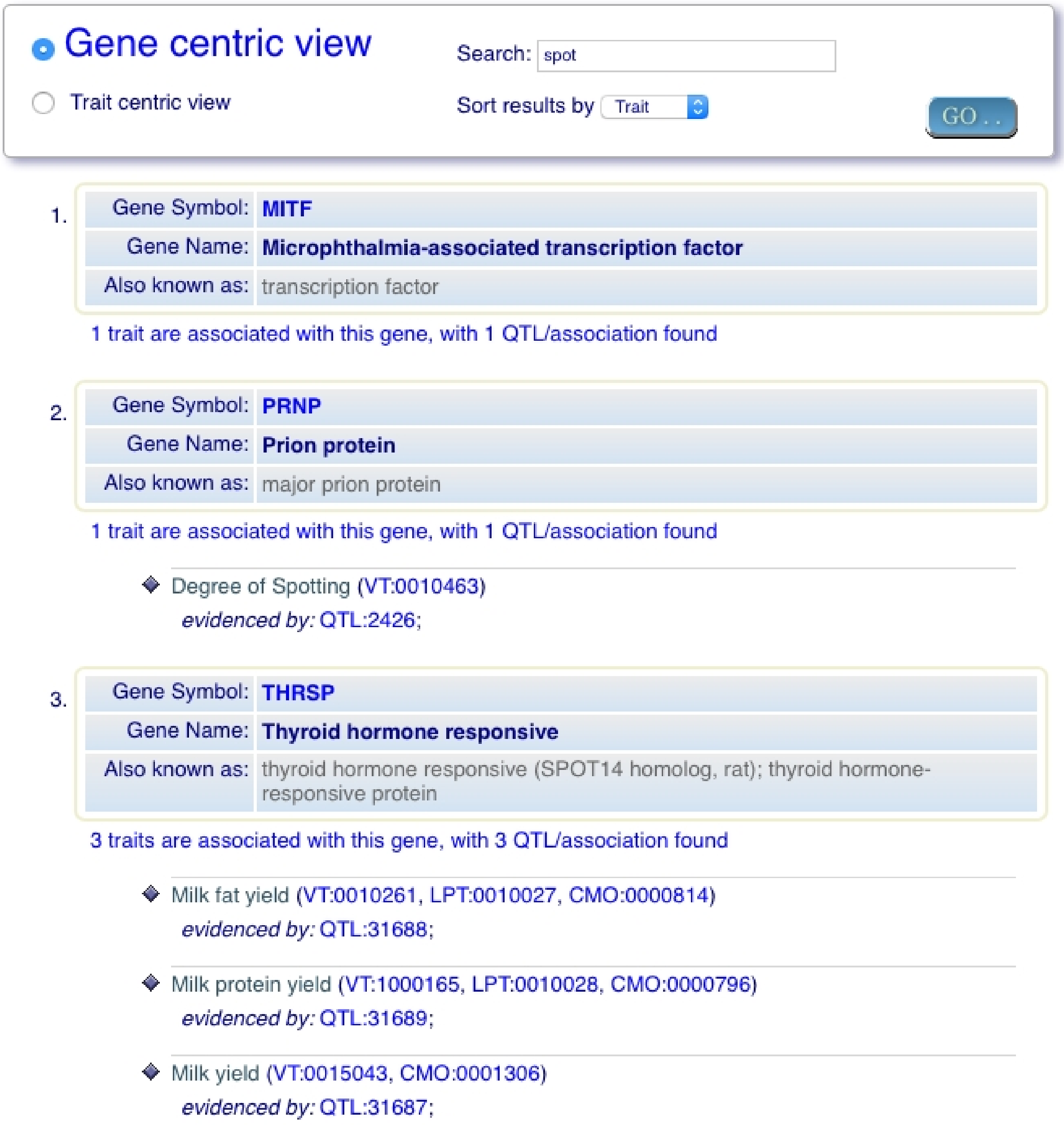
Description:

Measurement: Counts, or calculated as percentage of a base value

Scale/Unit:

Comments: Can be associated with perinatal weak calf syndrome.

Figure 7. A QTLdb editor window showing how a trait variant with modifiers can be created to ensure accurate representation of a trait in a given context.



Search: spot

Sort results by Trait

1\.

Gene Symbol:	**MITF**
Gene Name:	**Microphthalmia-associated transcription factor**
Also known as:	transcription factor

1 trait are associated with this gene, with 1 QTL/association found

2\.

Gene Symbol:	**PRNP**
Gene Name:	**Prion protein**
Also known as:	major prion protein

1 trait are associated with this gene, with 1 QTL/association found

◆ Degree of Spotting (VT:0010463)
evidenced by: QTL:2426;

3\.

Gene Symbol:	**THRSP**
Gene Name:	**Thyroid hormone responsive**
Also known as:	thyroid hormone responsive (SPOT14 homolog, rat); thyroid hormone-responsive protein

3 traits are associated with this gene, with 3 QTL/association found

◆ Milk fat yield (VT:0010261, LPT:0010027, CMO:0000814)
evidenced by: QTL:31688;

◆ Milk protein yield (VT:1000165, LPT:0010028, CMO:0000796)
evidenced by: QTL:31689;

◆ Milk yield (VT:0015043, CMO:0001306)
evidenced by: QTL:31687;

Figure 8. Gene-centric view of animal QTL/association data. Note that the long QTL/association list is hidden upon loading and can be click-expanded.

○ Gene centric view
● Trait centric view

Search: spot

Sort results by Trait

1.

Trait name:	**Degree of Spotting**
Ontology terms:	VT:0010463

2 genes are associated with this trait, with QTL/association found

- **"MITF"** Microphthalmia-associated transcription factor
  evidenced by: QTL:20538;
- **"PRNP"** Prion protein
  evidenced by: QTL:2426;

2.

Trait name:	**Milk fat yield**
Ontology terms:	VT:0010261, LPT:0010027, CMO:0000814

1 gene are associated with this trait, with QTL/association found

3.

Trait name:	**Milk protein yield**
Ontology terms:	VT:1000165, LPT:0010028, CMO:0000796

1 gene are associated with this trait, with QTL/association found

4.

Trait name:	**Milk yield**
Ontology terms:	VT:0015043, CMO:0001306

1 gene are associated with this trait, with QTL/association found

Figure 9. Trait-centric view of animal QTL/association data. Note that the long QTL/association list is hidden upon loading and can be click-expanded.

List of References

Animal QTLdb, 2017. URL: https://www.animalgenome.org/QTLdb. Created: September 2, 2004. Last updated: August 28, 2017.

Archibald A.L., N.E. Cockett, B.P. Dalrymple, T. Faraut, J.W. Kijas, J.F. Maddox, J. C. McEwan, V.H. Oddy, H.W. Raadsma, C. Wade, et al., 2010. The sheep genome reference sequence: a work in progress. Anim Genet. 41(5): 449-53.

Berthelot C., F. Brunet, D. Chalopin, A. Juanchich, M. Bernard, B. Noël, P. Bento, C. Da Silva, K. Labadi, A. Alberti, et al., 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nature Commun. 5: 3657.

Elsik C.G., R.L. Tellam, K.C. Worley, R.A. Gibbs, D.M. Muzny, G.M. Weinstock, D.L. Adelson, E.E. Eichler, L. Elnitski, R. Guigó, et al., 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science. 324(5926): 522-8.

Google Scholar Citations (2017). https://scholar.google.com/scholar.

Groenen M.A., A.L. Archibald, H. Uenishi, C.K. Tuggle, Y. Takeuchi, M.F. Rothschild, C. Rogel-Gaillard, C. Park, D. Milan, H.J. Megens, et al., 2012. Analyses of pig genomes provide insight into porcine demography and evolution. Nature. 491(7424): 393-8.

Hu, Z.-L., and J.M. Reecy, 2007. Animal QTLdb: beyond a repository. A public platform for QTL comparisons and integration with diverse types of structural genomic information. Mamm Genome. 18(1): 1-4.

Hu, Z.-L., C.A. Park, X.L. Wu, & J.M. Reecy, 2013. Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. Nucleic Acids Res. 41(D1): D871-D879.

Hu, Z.-L., D. Kumar, & J. Reecy, 2013. CorrDB: a livestock animal genetic/phenotypic trait correlation database. Plant & Animal Genomes XX Conference, January 11-16, 2013. Town & Country Convention Center, San Diego, CA.

Hu, Z.-L., S. Dracheva, W. Jang, D. Maglott, J. Bastiaansen, M.F. Rothschild, & J.M. Reecy, 2005. A QTL resource and comparison tool for pigs: PigQTLDB. Mamm Genome. 16(10): 792-800.

International Chicken Genome Sequencing Consortium, 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 432(7018): 695-716.

Jiang, Y., M. Xie, W. Chen, R. Talbot, J.F. Maddox, T. Faraut, C. Wu, D.M. Muzny, Y. Li, W. Zhang, et al., 2014. The sheep genome illuminates biology of the rumen and lipid metabolism. Science. 344(6188): 1168-73.

Liu Z., S. Liu, J. Yao, L. Bao, J. Zhang, Y. Li, C. Jiang, L. Sun, R. Wang, Y. Zhang, et al., 2016. The channel catfish genome sequence provides insights into the evolution of scale formation in teleosts. Nature Commun. 7:11757.

MIQAS—Minimum Information for QTLs and Association Studies. URL: http://miqas.sourceforge.net/; Last updated: May 12, 2009; Last accessed: September 21, 2017.

Park, C.A, S.M. Bello, C.L. Smith, Z.L. Hu, D.H. Munzenmaier, R. Nigam, J.R. Smith, M. Shimoyama, J.T. Eppig, & J.M. Reecy, 2013. The Vertebrate Trait Ontology: a controlled vocabulary for the annotation of trait data across species. Journal Biomed Semantics. 4(1): 13.

Sharma, A., J.S. Lee, C.G. Dang, P. Sudrajad, H.C. Kim, S.H. Yeon, H.S. Kang, & S.H. Lee, 2015. Stories and challenges of genome wide association studies in livestock — A review. Asian-Australas J Anim Sci. 28(10): 1371–1379.

Shimoyama M., R. Nigam, L.S. McIntosh, R. Nagarajan, T. Rice, D.C. Rao, & M.R. Dwinell, 2012. Three ontologies to define phenotype measurement data. Front Genet. 3:87.

The FAANG Consortium, 2015. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of ANimal Genomes Project. Genome Biol. 16:57.

Wade C.M., E. Giulotto, S. Sigurdsson, M. Zoli, S. Gnerre, F. Imsland, T.L. Lear, D.L. Adelson, E. Bailey, R.R. Bellone, et al., 2009. Genome sequence, comparative analysis, and population genetics of the domestic horse. Science. 326(5954): 865-7.

Wu, X.L., D. Gianola, Z.L. Hu, & J.M. Reecy, 2011. Meta-analysis of quantitative trait association and mapping studies using parametric and non-parametric models. J Biomet Biostat. S1:001 doi: 10.4172/2155-6180.S1-001.

Zimin A.V, A.L. Delcher, L. Florea, D.R. Kelley, M.C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C.P. Van Tassell, T.S. Sonstegard, et al., 2009. A whole-genome assembly of the domestic cow, Bos taurus. Genome Biol. 10(4): R42.