

NRSP-8 BIOINFORMATICS COORDINATION PROGRAM 2022 ACTIVITIES

Supported by Regional Research Funds, Hatch Act
James Reecy, James Koltes, and Fiona McCarthy,
Joint Coordinators

OVERVIEW: Coordination of the NIFA National Animal Genome Research Program's (NAGRP) Bioinformatics is primarily based at, and led from, Iowa State University (ISU), with additional activities at the University of Arizona (UA), and is supported by NRSP-8. The NAGRP is made up of the membership of the Animal Genome Technical Committee, including the Bioinformatic Subcommittee.

FACILITIES AND PERSONNEL: James Reecy (ISU) James Koltes (ISU), and Fiona McCarthy (UA) serve as Co-Coordinators. Iowa State University and University of Arizona provide facilities and support.

OBJECTIVES: The NRSP-8 project was renewed as of 10/01/18, with the following objectives: 1. Advance the quality of reference genomes for all agri-animal species by providing high contiguity assemblies, deep functional annotations of these assemblies, and comparison across species to understand structure and function of animal genomes; 2. Advance genome-to-phenome prediction by implementing strategies and tools to identify and validate genes and allelic variants predictive of biologically and economically important phenotypes and traits; and 3. Advance analysis, curation, storage, application, and reuse of heterogeneous big data to facilitate genome-to-phenome research in animal species of agricultural interest.

PROGRESS TOWARD OBJECTIVE 3: Advance analysis, curation, storage, application, and reuse of heterogeneous big data to facilitate genome-to-phenome research in animal species of agricultural interest.

The following describes the project's activities over this past year.

Multi-species support

The Animal QTLdb, CorrDB, NAGRP Bioinformatics Tools, and the NAGRP data repository have been actively supporting the research activities for multiple species. The QTLdb has been accommodating active curation of QTL/association data for seven species (cattle, catfish, chicken, horse, pig, rainbow trout, and sheep). In 2022, a total of 22,320 new QTL/association data were curated into the database, bringing the total number of curated data to 258,290 QTL/associations. Currently, there are 36,725 curated porcine QTL, 193,641 curated bovine QTL, 18,313 curated chicken QTL, 129 curated goat QTL, 2,649 curated horse QTL, 4,504 curated sheep QTL, and 2,329 curated rainbow trout QTL in the database (<https://www.animalgenome.org/QTLdb/>). An additional 2,735 correlations (increase by species: cattle: 640; chicken: 380; pig: 1,321; sheep: 394) and 459 heritability data (increase by species: cattle: 134; chicken: 58; goat: 3; horse: 1; pig: 160; sheep: 103) were curated into the Animal CorrDB in 2022. Currently there are a total of 26,839 correlation data on 806 traits and 4,778 heritability data on 950 traits in 6 livestock animal species. Note that these data updates are accompanied by a decrease of trait data owing to our recent works on adopting a “trait variant” system for data curation using trait modifiers and quantifiers. The new structure allows us to manage the extended trait/modifier information at the experiment level, as “trait variants”. This system has simplified the curation and management of such trait information. The numbers of

trait reductions are: QTL/association: -81.6%; correlation: -52.9%; and heritability: -79.9% (See our Poster #PO0143 at the PAG meeting for more details).

Ontology development

Our previously implemented ontology hierarchy display tool has facilitated our efforts expanding and exploring the use of the Vertebrate Trait (VT) Ontology, Livestock Product Trait (LPT) Ontology, Clinical Measurement Ontology (CMO), and other ontology hierarchies. This tool has been further integrated with other web portals for Animal QTLdb, VT, LPT, and CMO project websites.

This past year we continued to focus on the integration of the Animal Trait Ontology into the Vertebrate Trait Ontology (<http://bioportal.bioontology.org/ontologies/VT>). Ten (10) dataset updates were released to the public throughout 2022. We have continued working with the Rat Genome Database to integrate ATO terms that are not applicable to the Vertebrate Trait Ontology into the Clinical Measurement Ontology (<http://bioportal.bioontology.org/ontologies/CMO>). Traits specific to livestock products continue to be incorporated into a Livestock Product Trait Ontology (LPT) (<http://bioportal.bioontology.org/ontologies/LPT>). Three (3) LPT updates were released during 2022. Fifteen (15) updates of Livestock Breed Ontology (LBO; <https://www.animalgenome.org/bioinfo/projects/lbo/>) were made. We have also continued mapping the cattle, pig, chicken, sheep, and horse QTL traits to the VT, LPT, and Clinical Measurement Ontology (CMO) to help standardize the trait nomenclature used in both QTLdb and CorrDB. We have worked with the developer of AgroPortal (<http://agroportal.lirmm.fr/>) to streamline the new updates of VT, LPT, and LBO data into AgroPortal. The ontology data for the VT, LPT, and LBO are also available through Github (<https://github.com/AnimalGenome/vertebrate-trait-ontology>, <https://github.com/AnimalGenome/livestock-product-trait-ontology>, and <https://github.com/AnimalGenome/livestock-breed-ontology>, respectively) where users can automate their data updates. Anyone interested in helping to improve the ontologies is encouraged to contact James Reecy (jreecy@iastate.edu), Cari Park (caripark@iastate.edu), or Zhiliang Hu (zhu@iastate.edu). The VT/LPT/CMO cross-mapping has been well employed by the Animal QTLdb, CorrDB, and VCMAP tools. Annotation to the VT is also available for rat QTL data in the Rat Genome Database and for mouse strain measurements in the Mouse Phenome Database. We have also continued to integrate information from multiple resources, e.g. FAO - International Domestic Livestock Resources Information, Oklahoma State University - Breeds of Livestock web site, and Wikipedia, as well as requests from community members.

Expanded Animal QTLdb functionality

While the database supports multiple genome builds for all livestock species, we have adopted a mechanism to designate one “default genome build” per species (in alignment with NCBI/Ensembl). All curated QTL/association data continue to be automatically ported to NCBI, Ensembl, UCSC genome browser, and Reuters Data Citation Index in a timely fashion. Users can fully utilize the browser and data mining tools at NCBI, Ensembl, and UCSC to explore animal QTL/association data. Efforts were continually made, working with our counterparts at these institutions, to eliminate any glitches that arose during the automated or semi-automated data porting process. In addition, we have continued to improve existing and add new QTLdb

curation tools and user portal tools. We have made significant efforts to improve the curator tools in order to facilitate trait-variant related data curations. This involved more than 15 web portal scripts serving 47+ curator views/functions.

Further developments of Animal Trait Correlation Database (CorrDB)

Our efforts to overhaul and redevelop functionality to improve the usability of CorrDB have continued. The web portal has been updated with additions and/or modifications of its correlation network visualization options with the use of Graphviz dot graph and Cytoscape web tools, and newly added correlation and heritability data batch downloads provide more user-friendly access to the curated data.

We continued to strengthen the data quality control procedures to help improve data quality. This is partly reflected in a recall and re-curation of 348 previously entered pig correlation data. As reported in earlier sections, in 2022, correlation and heritability data continued to be curated throughout the year.

Facilitating research

The Data Repository for the aquaculture, cattle, chicken, horse, pig, and sheep communities to share their genome analysis data has been proven to be very useful (<https://www.animalgenome.org/repository>). While new data is continually being curated, we have gradually scaled down the support for hosting supplementary files for publications for more sensible use of the NRSP8 bioinformatics funds. As of April 5, 2022, all valid data (a total of 796 data files, 84.25 GB in size, for 92 manuscripts published in 27 scientific journals) have been transferred to Open Science Framework (OSF; <https://osf.io>) for better long-term data security. Appropriate web visit redirections for each data set have been set up on the current site to forward the incoming traffic to the new URL at OSF.

The data downloads from the repository generated over 290 GB of data traffic in 2022. Throughout the year, there were over 110 communication records on our helpdesk AnimalGenome.ORG to handle users' inquiries/requests for services affecting community research activities and the use of our services. Provided assistance ranged from data transfer and hosting, data deposition, data curation, web presentation, and data analysis, to software applications, code development, advice for tool developments, etc.

In 2022, we actively worked with external working groups and consortia to facilitate bioinformatics support to the community. One such effort was to work with the AgBioData group on recommendations for extending the GFF3 specification for improved interoperability of genomic data (<https://doi.org/10.48550/arXiv.2202.07782>). Another example is to work with scientists from University of Colorado, University of Sydney, and Lawrence Berkeley National Laboratory on development of the Vertebrate Breed Ontology (VBO; <https://github.com/monarch-initiative/vertebrate-breed-ontology>). We believe these efforts will directly or indirectly benefit the livestock genome research community.

Community support and user services at AnimalGenome.ORG

We have been maintaining and actively updating the NRSP-8 species web pages for each of the six NRSP-8 species. We continue to host mailing lists/websites for various research groups in the NAGRP community (<https://www.animalgenome.org/community/>). This includes groups like

AnGenMap (with about 3,000 subscribers from 67 countries/regions of the world), FAANG international consortium working groups (with 8 working group mailing lists and websites), and CRI-MAP users (670+ subscribers), new meetings, and user bulletin boards to facilitate these meetings, among other user forums.

The Functional Annotation of ANimal Genomes (FAANG) website (<https://www.faang.org/>) has been re-developed in 2022 to support a transition of earlier working groups to new “task forces” for more focused FAANG activities. The FAANG site serves not only as a FAANG-related information hub, but also as a platform for this international consortium’s communication, collaboration, organization, and interaction. It serves over 660 members and 8 working groups and sub-groups, with 10 listserv mailing lists, bulletin board, database, and tools for membership and working group management. The actively hosted materials include meeting minutes, tools/protocols for FAANG activities, incorporation and use of data portal hosted at EBI, presentation slides, and video records of scientific meetings and related events, all interactively available to members through the web portal.

Site maintenance

We have worked with Iowa State University IT in a transition to a faster and more secure network at our data center. We have consequently adopted new protocols and procedures to work with collaborators on data transfer and collaborative works involved in getting through the new firewall. Efforts were made to improve data backup, security, and availability. This was accomplished by better use of the resources for shared workloads, better data security and network security, and improved protocols for data backup, management, and inventories.

Reaching out

We have been sending periodic updates to more than 3,000 users worldwide (<https://www.animalgenome.org/community/angenmap/>) to inform the animal genomics research community of the news and updates regarding AnimalGenome.org. “What’s New on AnimalGenome.ORG web site” emails were sent out 3 times in 2022, consistent with the pace/pattern of the past 18 years (<https://www.animalgenome.org/bioinfo/updates/>).

Publications:

Zhi-Liang Hu, Carissa A. Park, and James M. Reecy (2023). An implementation of new approaches to extend livestock trait ontologies for practical curation management of QTL, association, correlation, and heritability data. Plant & Animal Genome Conference 30, January 13-18, 2023. Town & Country Convention Center, San Diego, CA.

Zhi-Liang Hu, Carissa A. Park, and James M. Reecy (2022). Bringing the Animal QTLdb and CorrDB into the future: meeting new challenges and providing updated services. Nucleic Acids Research, Volume 50, Issue D1, pages D956-D961.

Zhi-Liang Hu, Carissa A. Park, and James M. Reecy (2022). A database structural improvement for efficient trait variation curation in Animal QTLdb and CorrDB. The 12th World Congress on Genetics Applied to Livestock Production (WCGALP), Rotterdam, The Netherlands, July 3-8, 2022.

Sabrina Toro, Nicolas Matentzoglou, Kathleen R Mullen, Nicole Vasilevsky, Halie M Rando, Melissa Haendel, Christopher J Mungall, Zhi-Liang Hu, Gregoire Leroy, Imke Tammen,

Frank W Nicholas. (2022). Classifying Animal Breeds with the Vertebrate Breed Ontology (VBO). International Conference on Biomedical Ontology, 2022.

Surya Saha, Scott Cain, Ethalinda K. S. Cannon, Nathan Dunn, Andrew Farmer, Zhi-Liang Hu, Gareth Maslen, Sierra Moxon, Christopher J Mungall, Rex Nelson, Monica F. Poelchau (2022). Recommendations for extending the GFF3 specification for improved interoperability of genomic data. arXiv:2202.07782 [q-bio.OT].