NRSP-8 BIOINFORMATICS COORDINATION PROGRAM 2020 ACTIVITIES

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

February 20, 2021

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 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.

PROGRESS TOWARD OBJECTIVE 2: Facilitate the development and sharing of animal populations and the collection and analysis of new, unique, and interesting phenotypes.

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 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 2020, a total of 33,301 new QTL/association data were curated into the database,

bringing the total number of curated data to 211,792 QTL/associations. Currently, there are 31,455 curated porcine QTL, 160,659 curated bovine QTL, 12,783 curated chicken QTL, 2,472 curated horse QTL, 3,562 curated sheep QTL, and 861 curated rainbow trout QTL in the database. https://www.animalgenome.org/QTLdb/). In 2020, we also laid the groundwork to have the goat linkage maps, genome maps, SNP mapping data, trait management space, and other required information ready for goat QTL/associations to be curated into the database. Continued efforts to also curate data into the Animal CorrDB resulted in an addition of 2,398 correlation data and 1,055 heritability data in 5 animal species. Currently there are a total of 21,036 correlations data on 754 traits, and 4,320 heritability data on 1,127 traits in 5 livestock animal species. The NAGRP data repository continues to play an active role hosting genomics study data for the community.

The collaborative site at CyVerse continues to play an integral role as a backup/recovery site, sharing some the web traffic load (e.g. http://i.animalgenome.org/jbrowse), and a platform for developmental experiments. New data sources and species continue to be updated. The virtual machine site to host the Online Mendelian Inheritance in Animals (OMIA) database (Dr. Frank Nicholas at the University of Sydney; http://omia.animalgenome.org/) and the Hybrid Striped Bass website (Benjamin Reading of North Carolina State University; http://stripedbass.animalgenome.org/annotator/index) continue to be used by researchers.

Ontology development

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). Seven (7) dataset updates were released to the public throughout 2020. 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), which is available on NCBO's BioPortal (http://bioportal.bioontology.org/ontologies/LPT). Six (6) 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 Vertebrate Trait Ontology (VT), LPT, and Clinical Measurement Ontology (CMO) to help standardize the trait nomenclature used in the QTLdb. A semi-automated data release pipeline was developed to minimize the manual steps involved in new data upload and version release to BioPortal.ORG and GitHub with AnimalGenome.ORG as a new data sync hub. The VT data download is available through the Github portal (https://github.com/AnimalGenome/vertebrate-trait-ontology) where users can automate their data updates. Anyone interested in helping to improve the ATO/VT 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

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. The efforts continued to accommodate multiple genomes for QTL/association mapping/curation. Other improvements included the standardization of data links across species for external databases (db_xref) for both QTLdb and CorrDB; improved editor/curator tools to aid SNP name/ID look up and batch annotation for QTL/association data curation; and more improvements on eQTL data display and batch annotations. More improvements and developments as an on-going process are continually being carried out.

Further developments of Animal Trait Correlation Database (CorrDB)

Our efforts to overhaul or re-develop the CorrDB continued. The new outcome is a re-designed web interface for users to more easily access data by species (the front page). Internally, standardization of program configurations for parameters and functions will help to streamline future tool development and debugging efforts. The CorrDB works continue to feature co-development with the QTLdb for shared use of resources and tools, such as trait ontology development and management, literature management, breed ontology management, and bug reporting tools for improved data quality control. The improved CorrDB curator tools are available to the public for any user to register for an account to curate correlation data. As reported in earlier sections, in 2020, correlation data and heritability data continued to be curated. The public web portals continue to undergo improvement.

Facilitating research

The Data Repository for the aquaculture, cattle, chicken, horse, pig, and sheep communities to share their genome analysis data has proven to be very useful and has been actively used (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. We have redirected the community to a better data repository resource (Open Science Framework, OSF, https://osf.io/) for better long-term data security. In 2020, the data sharing platform was still actively used for researchers to share data individually or in a group.

The data downloads from the repository generated over 2.27 TB of data traffic in 2020. Throughout the year, over 87 cases were handled through our helpdesk at AnimalGenome.ORG which include 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.

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 have been hosting a couple dozen mailing lists/websites for various research groups in the NAGRP community (https://www.animalgenome.org/community/). This includes groups like AnGenMap, FAANG international consortium, and CRI-MAP users, new meetings, and user bulletin boards to facilitate these meetings, among other user forums (https://www.animalgenome.org/community).

The Functional Annotation of ANimal Genomes (FAANG) website (https://www.faang.org/) website has been continually developed and maintained to actively support the 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 500+ members and 12 working groups and sub-groups, with 14 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. Increases in the number of web hits and data downloads continued in 2020. AnimalGenome.org received over 2.9 million web hits from 388k individual sites (visitors), resulting in about 3.6 TB of traffic or data downloads.

Site maintenance

We have further consolidated services and developmental platforms to the current Dual Quad Core Xeon Linux server. 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 2020, consistent with the pace/pattern of the past 16 years (https://www.animalgenome.org/bioinfo/updates/).

PLANS FOR THE FUTURE

OBJECTIVE 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.

We will continue to analyze "omics" data to help better annotate livestock genomes.

OBJECTIVE 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.

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.

We will continue to work with bovine, mouse, rat, and human QTL database curators to develop minimal information for publication standards. We will also work with these same database groups to improve phenotype and measurement ontologies, which will facilitate transfer of QTL information across species. We will continue working with U.S. and European colleagues to develop a Bioinformatics Blueprint, similar to the Animal Genomics Blueprint recently published by USDA-NIFA, to help direct future livestock-oriented bioinformatic/database efforts.

Publications:

Hamid Beiki, James E. Koltes, Zhi-Liang Hu and James M. Reecy (2020). Analysis of Divergent Transcription Factors across Different Cattle Tissues Reveals Their Interesting Biological Functions. Plant & Animal Genomes XXVII Conference, January 11-15, 2020. Town & Country Convention Center, San Diego, CA.

Zhi-Liang Hu, Carissa Park, James M. Reecy (2020). An Update on Database Growth and Improvements of Animal QTLdb and CorrDB. Plant & Animal Genomes XXVII Conference, January 11-15, 2020. Town & Country Convention Center, San Diego, CA.