NRSP-8 BIOINFORMATICS COORDINATION PROGRAM 2019 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 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.

RNA transcript diversity among nine tissues was assessed using poly(A) selected single-molecule long-read isoform sequencing (Iso-seq) and Illumina RNA sequencing (RNA-seq) from a single White cross-bred pig. Across tissues, a total of 67,746 unique transcripts were observed, including 60.5% predicted protein-coding, 36.2% long non-coding RNA and 3.3% nonsense-mediated decay transcripts. On average, 90% of the splice junctions were supported by RNA-seq within tissue. A large proportion (80%) represented novel transcripts, mostly produced by known protein-coding genes (70%), while 17% corresponded to novel genes. On average, four transcripts per known gene (tpg) were identified; an increase over current EBI (1.9 tpg) and NCBI (2.9 tpg) annotations and closer to the number reported in human genome (4.2 tpg). Our new pig genome annotation extended more than 6000 known gene borders (5′ end extension, 3′ end extension, or both) compared to EBI or NCBI annotations. We validated a large proportion of these extensions by independent pig poly(A) selected 3′-RNA-seq data, or human FANTOM5 Cap Analysis of Gene Expression data. Further, we detected 10,465 novel genes (81% non-coding) not reported in current pig genome annotations. More than 80% of these novel genes had transcripts detected in > 1 tissue. In addition, more than 80% of novel intergenic genes with at least one transcript detected in liver tissue had H3K4me3 or H3K36me3 peaks mapping to their

promoter and gene body, respectively, in independent liver chromatin immunoprecipitation data. These validated results show significant improvement over current pig genome annotations.

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 2019, a total of **14,229** new QTL/association data were curated into the database, bringing the total number of data to **178,491** QTL/associations. Currently, there are 30,170 curated porcine QTL, 130,407 curated bovine QTL, 10,944 curated chicken QTL, 2,413 curated horse QTL, 3,099 curated sheep QTL, and 584 curated rainbow trout QTL in the database. https://www.animalgenome.org/QTLdb/). Continued efforts to also curate data into the Animal CorrDB resulted in an addition of 6,123 correlation data and 1,047 heritability data in 5 animal species. Currently there are a total of **18,638** correlations data on **641** traits, and **3,264** heritability data on **934** traits in **5** livestock animal species. The NAGRP data repository continue 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 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) continues to provide collaborative researchers convenient tools to create, maintain, and manage their sites with complete control.

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). Five (5) updated versions of the data set were released to the public throughout 2019. 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). Three (3) updates of Livestock

Breed Ontology (LBO; https://www.animalgenome.org/bioinfo/projects/lbo/). 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. 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 have been 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. In addition, we have continued to improve existing and add new QTLdb curation tools and user portal tools. The new efforts included accommodating multiple genomes for QTL/association mapping/curation; the use of DOI (Digital Object Identifiers) as an effective solution to link data to their full-text publications, renovation of curator/editor tools to accommodate the database structural changes for inclusion of eQTL data. The curated Epistatic data and pleiotropic data are made public with dynamic links to related data to aid data mining. A new collaborative curation environment has also been developed for multiple curators and editors from different locations to collaboratively work together in a semi-transparent workflow. An improved version of QTL/association data enrichment analysis tool has been made available that allows users to select the traits and genome regions of interest for analysis. Other improvements and developments as an on-going process are continually being carried out.

Further developments of Animal Trait Correlation Database (CorrDB)

To continue our major overhaul re-developments of the CorrDB, we began in 2019 to build some frequently used queries into the user interfaces to improve user experience and allow quicker and more direct access to data. The CorrDB works continue to feature a 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 newly developed 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 2019, correlation data and heritability data were continue to be curated. The public data 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 informed the community of a better data repository resource (Open Science Framework, OSF, https://osf.io/) for better long term data security Despite of this scale down, we have added 43 new data files to the repository, in 2019.

The data downloads from the repository generated over 2.2TB of data traffic in 2019. Throughout the year, over 130 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. Our involvement ranged from data transfer and hosting, data deposition, 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, among other user forums (https://www.animalgenome.org/community).

The Functional Annotation of ANimal Genomes (FAANG) website (https://www.faang.org/) is hosted by AnimalGenome.ORG. The 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 11 working groups and sub-groups, with 14 listserv mailing lists, a bulletin board, and a database for membership and working group management. The actively hosted materials include meeting minutes, presentation slides, and video records of scientific meetings and related events, all interactively available to members through the web portal. The "Funding Opportunities" information service has been improved to accommodate varying situations and to allow scientists to engage in open or private discussions to facilitate collaborations. Increases in the number of web hits and data downloads continued in 2019. AnimalGenome.org received over 3.8 million web hits from 346,600 individual sites (visitors), resulting in about 2.5 TB of data downloads.

Site maintenance

We have retired the old servers acquired in 2004 and 2008, and consolidated services and developmental platforms to the current Dell PowerEdge server and Dual Quad Core Xeon server. Efforts were made in order to better use the resources for shared workloads and better data security and network security, improved data backup schemes, virtual machine management, customer portal hosting, databases, and web services, etc.

Reaching out

We have been sending periodic updates to over 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 2019, in a consistent pace/pattern over the past 5 years.

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:

Beiki H, Liu H, Huang J, Manchanda N, Nonneman D, Smith TPL, Reecy JM, Tuggle CK. Improved annotation of the domestic pig genome through integration of Iso-Seq and RNA-seq data. BMC Genomics. 2019 May 7;20(1):344. doi: 10.1186/s12864-019-5709-y.