Fall 9-19-2024

# Animal Breeding-RCN: Farm Animal Genomics Collective

Noelle E. Cockett
*Utah State University*, noelle.cockett@usu.edu

**Data Management Plan**

Senior personnel listed on the *Animal Breeding-RCN: Farm Animal Genomics Collective* proposal are committed to responsible data management best practices for data management, including data analysis, sharing, and re-use. Because the overarching goal of the RCN Collective is to broadly share data, information and products produced within this project, participants of the RCN Collective will be expected to support responsible data management and routinely meet or exceed federal and community data management policy and best practices. This includes agreement to abide by the Toronto principles for data release, the Fort Lauderdale policy for rapid pre-publication release of data sets, and federal data sharing policies and requirements, including the *USDA Guidelines for Data Management Planning* for data release.

*Expected Data Types.* While we can predict many of the data types that will be collected in this project, the innate nature of a project with diverse collaborative activities in a rapidly changing field such as genomics means that not all data types are known in advance. Therefore, while only known data types are described below, the farm animal genomics community has a demonstrated history of adapting to new data types and ensuring these data can be shared and analyzed. For example, a recent NRSP-8 policy for data sets with no established repositories outlines the use of the Open Science Framework (OSF). Similar policies or existing policy adaptations will be developed as needed within the RCN.

Types of data expected in this project include:

(a) *Genomic and genetic sequence data.* This includes genomics data for pangenomes, expression datasets, related regulatory sequence data (e.g., histone marks, chromatin accessibility) and variant/genotyping data types.

(b) *Functional annotation.* Information about the role of genes and regulatory elements in physiological function, phenotypes and traits.

(c) *Haplotype, LD and epistasis.* These data may include population and individual sequence variants (genotypes), genome-wide association results, and DNA structure or interactions (3D genome structure).

(d) *Phenotypic data.* Curated and labeled sensor data that include the sensor output raw data (e.g.; images, accelerometer output, etc.) and the corresponding annotations (e.g.: animal activity/posture associated with an image, animal ID, etc.).

(e) *Analyses outputs*. Analyses performed under each activity will result in output data. For instance, sequence analyses typically produce alignment and variant call files.

In addition to these data types, the project will also produce workflows, genomics methods, reports, white papers, and scientific publications. These products will be publicly available and freely disseminated to the research community via the RCN website.

*Data Format.* We will use community standards for data formatting to ensure that the information provided by this proposal is easily shared and integrated with existing data sets. Genomics data will be formatted at fasta and gtf files. While there is no sanctioned file format for functional annotation and phenotypic data, there is the expectation that BigWig and BigBed will be used for ChIP-Seq and chromatin accessibility data, and GTF will be used for RNA-Seq expression quantification. Variant data and analysis files for information like haplotypes, LD and epistasis will use gVCF, .bim, .bed, and similar file types. Phenotypic data files will be specific for the sensor generating the data; for instance, .jpeg for images and some type of ASCII file (.json, .csv) for their annotations.

*Data Storage and Preservation.* Temporary data storage and sharing for this project is available through institutional data resources, commonly used back up, archiving and cloud storage facilities, and data sharing initiatives such as OSF and Dryad. All code and workflows will be shared through gitHub. The RCN will establish a policy of releasing data sets to public sequence repositories as soon as they have passed project quality checks (QC); this ensures that all data generated within this project will be preserved.

*Data Sharing and Public Access.* All data and metadata generated directly from this project (i.e. paid using project funds) will be freely and publicly available without restriction, and data sets will be submitted to appropriate public data archives as soon as they pass testing, quality checks, and documentation. To ensure continuity beyond the funding cycle, the RCN will encourage all participants to maintain their own OSF and gitHub repositories; the RCN Executive Committee will review each repository and provide feedback on compliance with minimum standards. In addition, a general repository that links to each contributor's repository will be available through the RCN website. In this way a large number of small but cross-linked repositories will be maintained using the "free" subscription services to OSF and gitHub. Thus, continuity is ensured beyond the lifetime of this grant.

*Roles and Responsibilities.* The RCN Executive Committee is ultimately responsible for ensuring DMP implementation. However, all participants in the Collective are expected to follow best practices for data management. Several of the lead personnel are trained (or have trained skilled personnel) in data formatting, analysis, sharing, archiving and re-use. As part of this project, RCN Research Area co-leaders with the necessary expertise will provide mentorship of the Collective members to ensure they are trained in and apply best practices for data management. Moreover, the RCN Executive Committee will develop a detailed public data management plan as data is generated with project funds. The plan will be updated with the generation of new data types. All members of the Collective will be responsible for periodically reviewing and observing the detailed data management plan.