**Data Management Plan**

**General Description of Data to be Managed**

Sequence data from *SpeciesName* DNA samples will be produced using an Illumina HiSeq 2500 or PacBio RS II instrument. Other data include RNA-Seq sequences from several tissue types for *SpeciesName.* The total size of this data is expected to be approximately 200 Gigabyte. All file formats used in this proposal will be in flexible, text-based, non-propriety format. Raw data files generated by high throughput sequencing machines are in FASTQ file format. Raw sequence data can be assembled resulting in another FASTQ file or aligned to an assembly resulting in a Sequence *Alignment/*Map (SAM) alignment formatted file. Alignment files can be used to generate Variant Call Files (VCF), which report Single Nucleotide Polymorphisms (SNPs) and Insertions and Deletions (InDels). Alignment files, VCF files and any file that contains a genomic element with a start and end position can be stored in Generic Feature Format (GFF) and displayed using a Genome Browser. No data will contain Personally Identifiable Information that is restricted by law or national security. The PI and Co-PIs will have a copy of this data management plan.

**Points of Contact**

The overall point of contact for the data collection and data management is Andrew Severin at ISUGIF; he will take responsibility for adherence to the data management plan, verifying the quality of the data, data documentation, data storage and data disaster recovery activities. Any data collection questions regarding the sequencing data can be answered by Andrew Severin. Questions regarding data collection from the biological system (*SpeciesName* samples and methodology) can be answered by ResearchersName, whose expertise is on the biological system and/or performed the collection.

**Data Stewardship**

GIF has a webserver which hosts the GIF website, dokuwiki, and Genome Browser (GBrowse). Dokuwiki is used as a secure online notebook to record the commands and programs used in the genome assembly, annotation and comparisons. The wiki page is made available to all members of the research team to provide feedback, error checking and enhance coordination, collaboration, communication between members of our team. A local installation of GBrowse is used to visualize the next generation sequencing data to ascertain the quality of the assembly and annotation and enable the exploration of marker data, QTLs, and alignments.

# **Data Documentation**

ISUGIF utilizes DokuWiki as an online notebook to document data analyses. DokuWiki is a simple to use, highly versatile and Open Source wiki. All information containing to the project is placed in this centralized location. Access to this wiki is granted to collaborators via login and passwords. A typical wiki page entry includes a detailed project description, the download location of the original data and how it was transferred, an initial original data characterization (size, number of reads), the location and version of all accessory data required to perform analyses (genome version, genome download location, database version and download location, etc.), and a logical progression of commands that steps the original data through the analyses to final products (raw reads to assembly, annotation, genetic map, QTL analysis, Genome Browser tracks, RNA-Seq, etc). With this documentation it is straightforward to reproduce or replicate our analyses since all methods, exact commands and versions of programs used are documented.

# **Data Sharing**

Upon project completion, all raw data will be deposited into the National Center for Biotechnology Information’s Short Read Archive (SRA) (<http://www.ncbi.nlm.nih.gov/sra>) which is a public, open access genomic database. After deposit, SRA will be responsible for keeping the data updated to ensure continued access. Andrew Severin commits to maintaining up-to-date contact information with NCBI so they can be notified and consulted on changes.

In compliance with the terms and conditions that govern USDA NIFA funded projects in Aquaculture, under Article 9, Section C, sub-section – Community Resource Projects, results will be made immediately available (pre-publication) for free and unrestricted use by the scientific community upon resource quality verification. This will be achieved by making raw data available along with subsequent assemblies for download from the website on the webserver owned by ISUGIF as the web resource is developed. The initial web resource describing the project and expected data will be in place prior to acquisition of sequencing data.

A website and genomic toolkit for *SpeciesName* species will also be created to consolidate the genomic resources and information generated from this research. The website will include all raw data, assemblies, annotations, genetic markers and QTL generated by the proposed work. There will also be a web-based Genome Browser that consolidates this information and allows users to visualize the data on the genomes at the nucleotide level. Any scripts that are generated to automate the processing of this data will be provided in at least one of the following methods: as supplementary data on the Genome Informatics Facility website, on GitHub and/or in the primary publication of the new resource. All data will be made publically available without restrictions following verification of quality.

# **Initial Data Storage and Protection**

During and after the research period, all data are redundantly backed up using RAID-6 storage to prevent catastrophic loss of data. A key part of our data security will be hosting of the data on RAID 6 storage boxes that are on a private IP address and therefore only accessible to people with accounts and passwords on the ISU campus or through the campus Virtual Private Network (VPN). Additionally, only members of the Genome Informatics Facility group will have access to the folder containing the raw and analyzed data during the exploratory to analysis phases of the project. Data will be provided to other members of the project team through password-protected access to folders containing the data.

Raw data will be managed by the Genome Informatics Facility (GIF) and stored on redundantly backed up 60 Terabyte RAID boxes maintained by the High Performance Computing (HPC) facility on campus. Raw data will be backed up immediately upon receipt. Scripts that generate analyzed data will be backed up daily and analyzed data will be backed up weekly. This model ensures rapid data recovery after the unlikely catastrophic loss of the primary RAID box. An additional copy of the raw data along with the scripts used to generate results will be kept in the GIF, which is located in a separate building from the HPC facility.

GIF has an archive server with 132 TB of storage. This machine serves as tertiary backup of all raw data and scripts that generate data analyses. This secondary site backup ensures the safety and integrity of the raw data and analyses in case of catastrophic failure at the primary site of data analysis in the High Performance Computing Facility at Iowa State University.

# **Long-Term Archiving and Preservation**

As noted in the Data Sharing section, all raw data will be deposited in the NCBI SRA data archive where it will be maintained and served to the public. Redundant backups of biologic data and biological tissue will be maintained by …