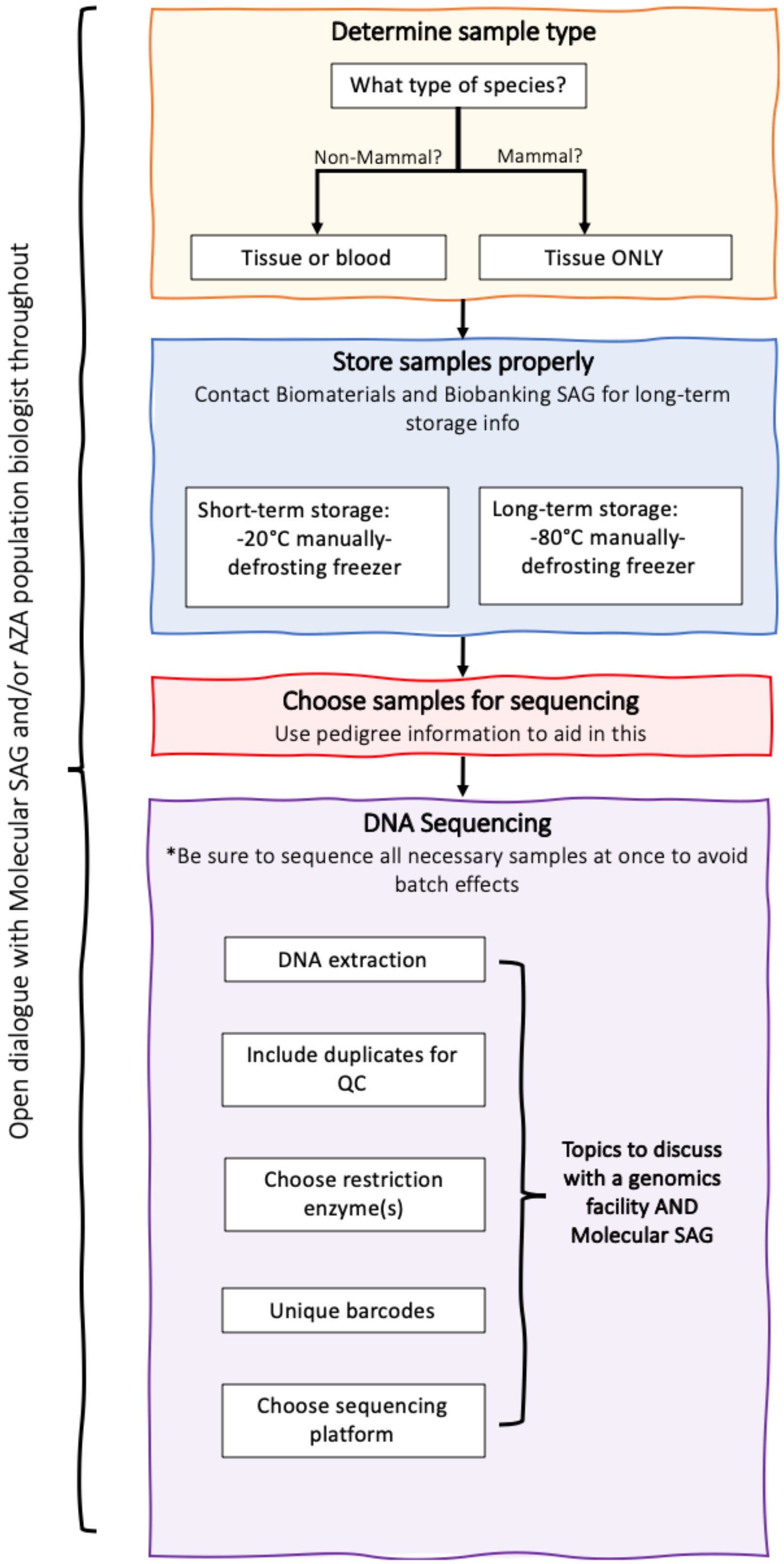
Standardized Data Collection guide

A guide to the collection of genomic data for improved management of AZA's living collections

Samantha Hauser, University of Wisconsin-Milwaukee
Jamie Ivy, San Diego Zoo Global
Andrea Putnam, San Diego Zoo Global
Emily Latch, University of Wisconsin-Milwaukee

Version 1.0. Published June 15, 2020

Standardized Data Collection Workflow



Open dialogue with Molecular SAG and/or AZA population biologist throughout
Determine sample type
What type of species?
Non-Mammal?
Mammal?
Tissue or blood
Tissue ONLY
Store samples properly

Contact Biomaterials and Biobanking SAG for long-term storage info

Short-term storage: -20°C manually-defrosting freezer
Long-term storage: -80°C manually-defrosting freezer
Choose samples for sequencing

Use pedigree information to aid in this

*Be sure to sequence all necessary samples at once to avoid batch effects

DNA extraction
Include duplicates for QC
Choose restriction enzyme(s)
Unique barcodes
Choose sequencing platform
Topics to discuss with a genomics facility AND Molecular SAG

Overview

In this document, we outline a general protocol for standardized collection of genomic data including collecting and storing samples, extracting high quality DNA, mitigating batch effects, and collecting robust sequence data. We recommend partnering with a laboratory with the capacity to generate these data, and discussing the project with an AZA population biologist and/or the Molecular Data for Population Management Scientific Advisory Group (Molecular SAG) first. The topics considered herein provide a basic foundation of information to facilitate discussions with partner labs, to embark on a successful genomics project.

DNA Samples

The ability to accurately estimate relationships to improve a species' pedigree depends in large part on obtaining robust sources of DNA. Captive populations typically have low genetic variation which requires a high number of markers for statistical power. The best sources of DNA are tissue and whole blood samples, in which there is an abundance of DNA and few inhibitors or contaminants. In mammals, mature red blood cells are enucleated and are therefore poor sources of DNA; tissue is typically the best source of mammalian DNA. While there are other sources of DNA that could be used (such as hair, scat, feathers, environmental DNA, etc.) we do not recommend these sources of DNA as they often do not provide sufficient quantities of DNA for high-throughput sequencing.

Blood and tissue samples can be obtained during routine visits with veterinarians in which veterinarians can take a biopsy punch, blood draw or blood prick. We suggest that this become a part of the protocol of routine veterinarian visits, to develop a collection of samples for genetic analyses. We highly recommend biobanking of all samples and coordination to do so with Molecular SAG, Biomaterials and Banking SAG, and any collaborative lab if applicable.

Samples should be collected and stored to preserve the DNA and minimize degradation. For samples collected directly from the animal (i.e., blood or tissue), we recommend that samples be stored in a tube or vial with buffer in a -20°C (short-term) or -80°C (long-term) freezer immediately after collection. Ideally, samples should be frozen in a manually-defrosting freezer; self-defrosting freezers have a freeze-thaw cycle that shears DNA. If freezing is not available, storing tissue in >95% ethanol will protect DNA from degradation and can be stored at room temperature. Storing tissue with silica beads is another viable option to adsorb water to inhibit DNases and growth of DNA-consuming bacteria. Blood may also be smeared onto Nobuto strips or Whatman FTA cards, but these methods are not recommended because 1) the DNA extraction process results in fragmentation of the DNA and poor quality sequences, and 2) they provide only trace amounts of mammalian DNA because mature red blood cells are enucleated in mammals.

[Placeholder/or see other section for choosing samples]

Batch Effects

Batch effects can occur when subsets of samples are processed separately but combined for analysis, notably in reduced representation sequencing such as RAD (Leigh et al., 2018). Assembling datasets from different runs generates different sets of loci that do not completely overlap. It is important to note that there are other methods that are less prone to batch effects such as targeted sequencing in which baits or probes (Ali et al., 2016; McCormack et al., 2016; Suchan et al., 2016) are used to sequence a consistent set of target loci. However, such targeted sequencing requires some knowledge of the genome of the organism and some laboratory optimization. Regardless of the approach used, the most robust sequence data possible will be obtained when all samples are processed together in the same batch. Therefore, we recommend running all samples at once; if that is impossible, we recommend considering targeted sequencing using baits or probes (e.g., myBaits) that allow robust results with many batches.

When it is necessary to run separate batches, ensuring protocols are identical will minimize batch effects. Protocol differences (variations in sample storage, DNA extractions, sequencing preparations, size selection, sequencing platforms, etc.) can result in datasets that are partially or completely incompatible, decreasing the overall analytical power of the dataset and potentially rendering one or more of the datasets unusable for the project.

Sequencing (RAD sequencing)

Accurate estimation of relationships between individual animals in a captive breeding program will require several thousand SNPs (Flanagan & Jones, 2019). There are several approaches that could be used for SNP discovery and genotyping (Ali et al., 2016; Cammen et al., 2016; Narum et al., 2013). We advocate for a reduced representation approach, double-digest restriction-site associated DNA sequencing (ddRADseq; Peterson et al., 2012). Compared to other reduced representation approaches, ddRADseq is better than other RADseq approaches at recovering the same loci from independent samples, and generally is more robust to partially degraded DNA samples. In the ddRADseq approach, DNA is digested with two restriction enzymes, adapters and barcodes are ligated to the resulting fragments, and fragments within a specific size range are selected for sequencing on a high-throughput sequencer (e.g., Illumina HiSeq, NovaSeq; Peterson et al., 2012). The sequencer generates millions of short reads (often paired-end, 150bp reads), which are then aligned to identify variable sites (SNPs). A number of bioinformatic filtering steps (Danecek et al., 2011; Luikart, 2014) are employed to generate a set of thousands of SNPs distributed across the genome that have high statistical power necessary to reconstruct relationships among individuals (e.g., parents, siblings, etc.).

When choosing a partner genomics facility and subsequently coordinating with them, several aspects of the protocol should be discussed:

(a) DNA extractions: Whether the genomics facility is able and willing to extract the DNA. This is an important consideration in choosing a genomics facility, especially if you do not have the capability for DNA extractions in-house.

If you extract the DNA from samples before sending them to the genomics facility, we recommend either phenol-chloroform protocols, which produce high DNA yields but can require some laboratory skill to master, or commercial DNA extraction kits (e.g., Qiagen DNeasy Blood and Tissue Kit), which are typically fast and perform well in most situations. You will also need to standardize your DNA concentrations (e.g., using a Qubit fluorometer) before sending the isolated DNA to the genomics facility; check with the facility to determine the total amount of DNA needed and at what concentration.

Regardless of whom is performing the DNA extraction, you will need to coordinate with the genomics facility to determine how to package the samples (tubes, 96-well plate, with ice packs, etc.) for shipping to the facility.

(b) Duplicates: As quality control, it is recommended to coordinate duplicate samples when sequencing within a run and across runs (if applicable). Duplicates allow evaluation of sequencing consistency and identification of any potential problems in sequencing or sequencing prep that could affect the results. Quality control duplicates should either be set up by you (if you are preparing the DNA samples) or should be discussed with the partner genomics facility.

(c) Restriction Enzymes(s): Restriction enzymes cut double-stranded DNA at a recognized nucleotide sequence, allowing fragmentation of DNA in a predictable fashion. Preparation of a ddRAD sequencing library requires digestion of genomic DNA with a pair of restriction enzymes (typically a 6-cutter and a 4-cutter), to generate an optimal number of fragments in the desired size range (Peterson et al., 2012). The two-enzyme approach utilized by ddRAD sequencing, compared to a one-enzyme plus random shearing method found in other RAD sequencing approaches (Mastretta-Yanes et al., 2015), offers more control over the resulting DNA fragments. The optimal pair of enzymes often differs between taxa. Some genomics facilities may test several pairs of restriction enzymes for you, some will use one enzyme pair for all projects, and some may have you choose up front which enzymes to use. Optimization of restriction enzymes benefits the end result, by generating more fragments in the desired size range and thus more viable fragments to sequence, but can use up precious DNA and resources. Sometimes enzyme pairs that work well in one species can be effective in related taxa, though it can be difficult to predict precisely how the enzymes will fragment the genome. If there is a reference genome for your taxa or a closely related taxa, you could BLAST candidate enzymes' recognition sequences to estimate how frequently the enzymes will cut and give a general sense of how the enzymes might perform . You could also look to broad taxonomic groups (i.e., birds, mammals) to see if a pair of enzymes is frequently successful in that group.

(d) Barcodes: Barcodes are unique DNA sequences added to each sample's DNA fragments during the library preparation process, so that sequences from each individual can be identified and sorted in downstream data analyses. Typically, the genomics facility will have a set of 48-96 unique barcodes that they will add to each of your samples during sequencing preparation. The number of barcodes they have available will dictate the number of samples you can run together, making this an important conversation to have with the genomics facility prior to sample preparation. It is critical to ensure that there are no repeat barcodes within a run or else you will not be able to differentiate your samples.

(e) Size selection: One step in the library preparation is size selection, to choose the size of DNA fragments that will be directly sequenced. Typically, a size range of 300-600bp is optimal. This fragment size range works well with our recommendation to sequence using paired-end 150 runs (PE150; see below for more information on paired-end reads), which sequence 150 base-pairs from each direction of the DNA fragment, generating approximately 300 base-pairs of sequence.

(f) Sequencing platforms: The best sequencing platform for your needs will depend on the number of samples you are sequencing and the amount of sequencing data you want from each sample. Most genomics facilities have several sequencing platforms available and will help you figure out which are appropriate for your needs. Regardless of the sequencing platform, we recommend running paired-end sequencing. By sequencing from both ends of the DNA fragments, paired-end sequencing produces twice the number of reads for the same time and effort in library preparation. Sequences aligned as pairs also produce more accurate read alignments and better variant detection.

Sequencing technologies advance rapidly, with platforms and capabilities changing frequently. Discussions with your partner genomics facility to evaluate available platforms, given your sequencing needs and budget is ideal for choosing the most appropriate platform. Currently, the most commonly used sequencing platforms were developed by Illumina. The NovaSeq platform generates a large quantity of sequencing reads (up to 20 billion reads per run) and thus can accommodate a large number of samples or a small number of samples with a large proportion of the genome sequenced. The HiSeq platform generates a moderate quantity of sequencing reads (2.5-5 billion reads per run). MiSeq or NextSeq sequencers generate the smallest quantity of sequencing reads (130-400 million reads per run) but have very rapid turnarounds. Generally these sequencing platforms are not recommended over HiSeq except for pilot studies or in which very rapid turnaround time for DNA sequences is necessary.