Genetic Data Guidance

Genetic and genomic data come in a variety of forms. Many, but not all, of these rely on nucleotide sequence data. Additionally, many questions using genetic data require generation of secondary datasets derived from sequence data. Both the original sequences and the derivative datasets should be captured and submitted to GRIIDC.

Fortunately, the National Center for Biotechnology Information (NCBI; ncbi.nlm.nih.gov) already deals with most of these data types. Data should be deposited to NCBI databases such as GenBank, the Gene Expression Omnibus (GEO) database, or Sequence Read Archive (SRA) database. Investigators generating this data are likely familiar with this process, as it is required by most journals prior to publication. Once NCBI accession numbers are obtained, they should be deposited to GRIIDC with additional relevant information that may not have been captured by NCBI.

Please see http://www.ncbi.nlm.nih.gov/home/submit.shtml for the NCBI submission portal.

NCBI has requirements for what is necessary for a successful submission and uses a standardized vocabulary. However, they do not always require the capture of some information, for example, geo-referencing information.

> Note – Samples and data should ALWAYS be geo-referenced (latitude, longitude, depth, and date collected) regardless of NCBI submission requirements. Ideally, place this information in the NCBI submissions (e.g., latitude and longitude can be added sometimes to the "source" field), and provide a table relating accession number to sample (and if needed, georeferenced information) to GRIIDC.
>
> The information below provides additional information for specific types of genetic data.

Sequence data

The most common type of genetic data generated will be nucleotide sequence data. Data submitted to GRIIDC should include either GenBank accession numbers (for Sanger sequences) or SRA accession number (for high-throughput sequencing data). Barcode data should meet the standards for the "BARCODE" designation for a GenBank accession.

Aligned sequence data

If sequences aligned across different taxa have been used in something like phylogenetic analyses, analyses of nucleotide substitution patterns, or other comparative analyses, they should be submitted to NCBI as, for example, a PopSet in NCBI, Treebase (treebase.org), ARB database for microbes, or the Open Tree of Life.

Genome and transcriptome assemblies and annotation

As appropriate relative to the analyses performed, data files of assembled contigs, assembled scaffolds, and annotated assemblies of FASTA files should be submitted to GRIIDC, in addition to the raw sequence data.

Expression and methylation data

These data should be deposited to the NCBI GEO database and accession numbers provided to GRIIDC with sample information.

SNP and microsatellite data

The SNP or microsatellite data table used in the analyses should be submitted as csv files. The SNP or microsatellite should be clearly labeled in terms of the alleles and individuals such that the alleles present in each individual can be determined. Ploidy level and populations of individual should be clearly discernable or presented in a separate file.