Genetic Data Submissions to GRIIDC

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 or Biotechnology Information (NCBI; ncbi.nlm.nih.gov) already deals with most of these data types. Instead of GRIIDC reinventing the wheel, data should be deposited to NCBI databases such as GenBank, the Gene Expression Omnibus (GEO) database, Sequence Read Archive (SRA) database. Investigators generating this data should know how to do this and it will be 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

NCBI has requirements for what is necessary for a successful submission and uses a somewhat standardized vocabulary (surprising, there is considerable variation of names of the some genes). However, they do not always require the capture of some information. In particular, depending on the type of submission, Geo-referencing information may not be captured.

Note – Samples and data should ALWAYS be Geo referenced (Latitude, Longitude, Depth, date collected) regardless of NCBI submission requirements. Ideally place this information in the NCBI submissions (e.g., latitude and longitude can be added some times to the "source" field), and then have a table relating accession number to sample (and if needed georeference information) should be provided to GRIIDC. Even if all this is present in the NCBI submission, a table of Accession numbers should be provided to GRIIDC with short sample descriptions.

Below please find 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 through put sequence 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 a given individual can be determined. Ploidy level and populations of individual should be clearly discernable or presented in a separate file.

General info

http://emboss.sourceforge.net/docs/themes/SequenceFormats.html
http://www.ncbi.nlm.nih.gov/sra

example: http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?study=SRP022154
http://trace.ncbi.nlm.nih.gov/Traces/study/?acc=SRP022154