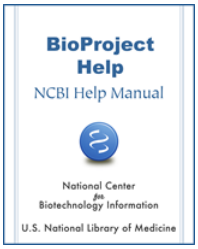
U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: BioProject Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2011-. Glossary. 2011 May 6 [Updated 2011 Nov 9].
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



Glossary

Created: May 6, 2011; Updated: November 9, 2011.

The following terms may be found on the Submission wizard pages, the web Summary report pages, the web full report pages, or as indexed fields in the Advanced Search page.

Project identifier

Project Accession — The format of the BioProject Accession is five alpha-letters followed by one to six numbers. For example PRJNA43021.

Project type

Umbrella project — Umbrella projects are administrative in nature. They are created upon the request of the submitter, a funding agency, or by NCBI staff to group multiple projects that are part of a large initiative or collaboration or funding source. Umbrella projects are indirectly connected to data through the linked primary submission projects. For example, Umbrella projects reflect the general organizational structure of the Human Microbiome Project and the ENCODE project.

Primary submission — Primary submission projects represent, and are linked to, current or future data submissions. Primary submissions include a series of attributes describing the initative that utilize a controlled vocabulary. Each element also includes a single free text option in order to flexibly support a wide range of projects. The free text will be periodically reviewed with the goal of updating the controlled vocabulary list. These fields and controlled vocabularies are described below.

Project data type

Project data type

A general label indicating the primary study goal. These are only relevant for Primary submission projects (not Umbrella projects). Includes:

- Assembly: genome assembly project utilizing already existing sequence data including data that was submitted by a different group
- Clone ends: clone-end sequencing project
- Epigenomics: DNA methylation, histone modification, chromatin accessibility datasets
- Exome: exome resequencing project
- Genome sequencing: whole, or partial, genome sequencing project (with or without a genome assembly)
- Map: - project that results in non-sequence map data such as genetic map, radiation hybrid map, cytogenetic map, optical map, and etc.

- Metagenome: sequence analysis of environmental samples
- Metagenome assembly: a genome assembly generated from sequenced environmental samples
- Other: a free text description is provided to indicate Other data type
- Phenotype or Genotype: project correlating phenotype and genotype
- Proteome: large scale proteomics experiment including mass spec. analysis
- Random Survey: sequence generated from a random sampling of the collected sample; not intended to be comprehensive sampling of the material.
- Targeted locus (loci): project to sequence specific loci, such as a 16S rRNA sequencing
- Transcriptome or Gene expression: large scale RNA sequencing or expression analysis. Includes cDNA, EST, RNA_seq, and microarray.
- Variation: project with a primary goal of identifying large or small sequence variation across populations.

Attributes

Sample Scope

Indicates the scope and purity of the biological sample used for the study.

- Monoisolate: a single animal, cultured cell-line, inbred population,.or possibly a heterogeneous population when a single genome assembly is generated from a pooled sample because multiple individuals are needed to collect enough material and an inbred line is not available; however, this situation is not preferred.
- Multiisolate: multiple individuals that represent distinct sample collections, a population (representative of a species). This is often used for variation or phenotype and genotype studies. This should not be used when multiple genomes will be annotated. Eventually, multiple locus_tag prefixes will be able to be assigned to a single multiisolate genome sequencing project, but currently only a single prefix can be registered per project. Therefore, individual monoisolate projects need to be registered when more than one genome will be annotated.
- Multi-species: sample represents multiple species.
- Environment: the species content of the sample is not known. Generally, nucleic acid is directly isolated from an environmental sample for analysis. This is used for metagenome studies.
- Synthetic: the sample is synthesized in a laboratory.
- Other: specify the sample scope that was used.

Material

Indicates the type of material that is isolated from the sample for use in the study.

- Genome: a whole genome initiative (a specific sub-cellular molecule is not experimentally isolated). May be only the nuclear genome. Use for DNA of a metagenome sample.
- Purified chromosome: one or more chromosomes or replicons were experimentally purified.
- Transcriptome: transcript and/or expression data.
- Phenotype: phenotypic descriptive data.
- Reagent: material studied was obtained by chemical reaction, precipitation.
- Proteome: protein or peptide data.
- Other: specify the material that was used.

Capture

Indicates the scale, or type, of information that the study is designed to generate from the sample material.

- Whole: the project makes use of the whole sample material (most common case). Use this for whole genome sequencing studies, transcriptome studies that are not targeting specific loci, epigenetic studies of a genome, and metagenomes or unbiased transcriptome studies of metagenomes.
- CloneEnds: capturing clone end data.
- Exome: capturing exon-specific data.
- TargetedLocusLoci: capturing specific loci (gene, genomic region, bar code standard).
- RandomSurvey: not using whole sample, an incomplete survey of the sample.
- Other: specify the scale or type of the captured material when none of the above options are correct for your study.

Method

Indicate the general approach used to obtain data.

- Sequence: select Sequence if any sequence data is generated
- Array: select Array if that is the primary method and no sequence data is submitted
- Mass Spectrometry: select Mass Spectrometry if that is the primary method
- Other: specify the method.

Objective

Indicates the project goals with respect to the type of data that will be generated and submitted to an INSDC database. Select all relevant menu options.

- Raw Sequence Reads: submission of raw reads to SRA or Trace repositories
- Sequence: submission of sequence data to standard archival sequence databases (yielding accession.version identifiers; e.g., whole genome shotgun, cDNA sequences, transcript shotgun assemblies)
- Analysis: other analysis not otherwise indicated, includes submission of BAM files
- Assembly: submission of genome assembly (AGP data)
- Annotation: sequence annotation data
- Variation: identification of sequence variation data for submission to dbSNP or dbVAR
- Epigenetic Markers: DNA methylation, histone modification, chromatin accessibility datasets
- Expression: assays of transcript or protein existence or abundance
- Maps: non-sequence based map data; e.g., genetic, radiation hybrid, cytogenetic, etc.
- Phenotype: phenotypic measurements for submission to dbGaP
- Other: specify the other objective