

BioProject Help

Last Updated: November 9, 2011



National Center for Biotechnology Information (US)
Bethesda (MD)

National Center for Biotechnology Information (US), Bethesda (MD)

NLM Citation: BioProject Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2011-.

This book contains information on the NCBI BioProject Database, a searchable collection of complete and incomplete (in-progress) large-scale sequencing, assembly, annotation, and mapping projects for cellular organisms.

Table of Contents

BioProject Help 1

 Introduction..... 1

 BioProject Quick Start Guide..... 2

 Display options 4

 Submitting to BioProject..... 9

Glossary 11

 Project identifier..... 11

 Project type 11

 Project data type..... 11

 Attributes 12

BioProject Help

Kim Pruitt, Ph.D.,¹ Karen Clark, Ph.D.,² Tatiana Tatusova, Ph.D.,³ and Ilene Mizrahi, Ph.D.⁴

Created: May 4, 2011; Updated: November 9, 2011.

Introduction

The BioProject resource is a redesigned, expanded, replacement of the NCBI Genome Project resource. The redesign adds tracking of several data elements including more precise information about a project's scope, material, and objectives. Genome Project identifiers are retained in the BioProject as the ID value for a record, and an Accession number has been added. Other changes include a more flexible approach to grouping projects and addition of data elements including fields for funding source and general relevance categories. The web site presentation has been redesigned and some of the data elements that can be tracked in the new database design will be added to the public display once data has accumulated. In addition, database content is exchanged with other members of the International Nucleotide Sequence Database Collaboration (INSDC).

The BioProject database provides an organizational framework to access metadata about research projects and the data from those projects which is deposited, or planned for deposition, into archival databases maintained by members of the INSDC. The resource supports a variety of projects in terms of type and complexity, ranging from a focused genome sequencing project to a large international collaboration with multiple sub-projects such as sequencing, collecting genotype/phenotype data, calling sequence variants, or assaying epigenetic information. Data submitted to INSDC-associated databases cross-reference the BioProject identifier to support navigation between the Project and the project's datasets. Therefore, the BioProject resource provides a reliable mechanism, for a variety of complex cases, to access specific datasets that can be difficult to find due to volume, sequential submissions over the course of a project, or submissions of distinct data types to multiple archival databases.

The definition of a set of related data, a 'project' is very flexible and supports the need to define a complex project and various distinct sub-projects using different parameters. For example, BioProject records can be established for:

- Genome sequencing and assembly
- Metagenomes
- Transcriptome sequencing and expression
- Targeted locus sequencing
- Genetic or RH Maps
- Epigenetics
- Phenotype or Genotype
- Variation detection

The database is not limited by taxonomy and as such includes information for studies of eukaryotes, prokaryotes, and environmental samples. Registration for a BioProject accession is encouraged for projects that result in a very large volume of data submissions, submissions from multiple members of a collaboration, or submissions to multiple archival databases. Registration for a BioProject accession is discouraged for small datasets for which the results are found in one (or a small number) of accession numbers such as a single viral or organelle genome

Author Affiliations: 1 NCBI; Email: pruittk@ncbi.nlm.nih.gov. 2 NCBI; Email: kclark@ncbi.nlm.nih.gov. 3 NCBI; Email: tatiana@ncbi.nlm.nih.gov. 4 NCBI; Email: mizrahi@ncbi.nlm.nih.gov.

 Corresponding author.

sequencing study. A BioProject ID is required for some database submissions including dbVar, SRA, and GenBank microbial and eukaryotic genomes.

The database defines two types of projects: 1) Primary submission projects are directly associated with submitted data and may be registered by submitters of that data using the NCBI submission portal; 2) Umbrella projects reflect a higher-level organizational structure for larger initiatives or provide an additional level of data tracking that has been requested by a NIH institute. These projects are created by request, typically by a funding source. An Umbrella project may group projects that are part of a single collaborative effort but represent distinct studies that differ in methodology, sample material, or resulting data type. Complex studies may be represented with more than one layer of Umbrella project such that a highest-level Umbrella project is linked to one or more sub-project Umbrella projects which in turn are linked to one or more Primary submission projects that describe the data in more detail. As described below, the resource supports navigating up and down this hierarchy from any starting point within the hierarchy, as well as navigating to peer projects, e.g., those with a common Umbrella.

BioProject Quick Start Guide

BioProject records can be accessed by query, by browsing, or by following a link from another NCBI database.

Links to BioProject records may be found in several databases including dbVar, Gene, Genomes, GEO, and Nucleotide which includes GenBank or RefSeq nucleotide sequences for which there is a registered BioProject identifier.

To access BioProjects by browsing, follow the link on the home page (Figure 1) to browse “[By project attributes](#)”. This page (Figure 2) supports browsing the database content by major organism groups, data attributes, or project data type. The table includes links to the NCBI Taxonomy database where additional information about the organism may be available and to the BioProject record where more information about the research project is reported.

The database can also be accessed by direct query. Queries can be entered from the BioProject home page, or by selecting the BioProject database name from the search menu found on the NCBI home page or other NCBI databases. The BioProject database can be queried using any text term or by restricting the query to a specific category using the Limits page or the Advanced Search page. Use the “Search Builder” menu available on the Advanced Search page to explore what fields are indexed and to build a complex Boolean query. Please refer to the [Entrez Help](#) chapter for a general description of Limits and Advanced Search pages.

Search Tips

You may search BioProject like any other NCBI database, namely by:

- Searching for an organism name
- Searching by the database accession or ID
- Searching for any word
- Restricting a search to a specific field using Limits
- Using the Advanced Search page to build a query restricted by multiple fields

The Limits page can be used to search by Project Type, Project Attributes, or Organism or Metagenome Groups.

See the Advanced Search page to explore the indexed fields and properties, view your search history, save search results, and view search details. Field restricted searching can be performed using the Search Builder. Restricted searches may also be entered manually by following the term with the name of the field in square brackets “[]”, as shown in the following examples. To see the indexed terms, choose a field from the pull down menu, then click on “Show index”.



Figure 1. The BioProject home page. The top portion of the page includes the standard NCBI search interface with links to the Limits and Advanced search pages. The main body of the page includes links to help documentation, the submission portal which is used to register new projects, the FTP site, an interface to find projects by browsing, example projects, and related resources.

Here are some representative searches:

NCBI Resources ☒ How To ☒ My NCBI Sign In

BioProject

Search:

First Previous **Shown: 1 - 100 out of 622 items** Next Last

Project Accession	Organism/Name <input type="button" value="All"/>	TaxID	Project Type <input type="button" value="Primary submission"/>	Project Data Type <input type="button" value="Transcriptome or Gene expression"/>	Date
PRJNA113	Pyropia yezoensis	2788	Primary submission	Transcriptome or Gene expression	2003/02/25
PRJNA114	Emiliania huxleyi	2903	Primary submission	Transcriptome or Gene expression	2003/03/05
PRJNA115	Alexandrium tamarense	2926	Primary submission	Transcriptome or Gene expression	2003/03/05
PRJNA137	Fusarium sporotrichioides	5514	Primary submission	Transcriptome or Gene expression	2003/02/25
PRJNA143	Eimeria tenella str. Houghton	413949	Primary submission	Transcriptome or Gene expression	2003/02/25
PRJNA152	Babesia bovis	5865	Primary submission	Transcriptome or Gene expression	2003/02/25
PRJNA160	Amblyomma americanum	6943	Primary submission	Transcriptome or Gene expression	2003/02/25
PRJNA165	Apis mellifera ligustica	7469	Primary submission	Transcriptome or Gene expression	2003/02/25
PRJNA194	Glossina morsitans morsitans	37546	Primary submission	Transcriptome or Gene expression	2003/02/25
PRJNA362	Emericella nidulans	162425	Primary submission	Transcriptome or Gene expression	2003/02/25
PRJNA1450	Citrus sinensis	2711	Primary submission	Transcriptome or Gene expression	2003/11/10
PRJNA10631	Aedes aegypti	7159	Primary submission	Transcriptome or Gene expression	2004/03/10
PRJNA10720	Arabidopsis thaliana	3702	Primary submission	Transcriptome or Gene expression	2004/05/03

Figure 2. The interface to find projects by browsing includes options to restrict by organism kingdom, primary project type attributes (umbrella projects vs. primary submissions), or specific project data type. The display shown is restricted to primary submissions and transcriptome projects and is sorted by the Project Accession.

Find BioProjects by...	Search text example(s)
A species name	Saccharomyces cerevisiae[organism]
Project data type	"metagenome"[Project Data Type]
Project data type and Taxonomic Class	"transcriptome"[Project Data Type] AND Insecta[organism]
Publication	"10473380"[PMID]
Submitter organization, consortium, or center	JGI[Submitter Organization]
Sample scope and material used	"scope environment"[Properties] AND "material transcriptome"[Properties]
A BioProject database identifier	PRJNA33823 or PRJNA33823[bioproject] or 33823[uid] or 33823[bioproject]

Display options

NCBI's Entrez system supports alternate display options for each of its databases. The options available can be browsed by clicking on *Display Settings*. The options presented in the *Display Settings* window depend on whether you are viewing a set of results, or just one record. In the former case, the *Display settings* window also provides choices for controlling the number of query results to display on the page.

BioProject offers four display settings, three of which are listed in the Display Settings menu, and the fourth is available by clicking on a record title. These display options are:

The screenshot shows the NCBI BioProject search interface. At the top, there's a navigation bar with 'NCBI', 'Resources', and 'How To'. The main search bar contains 'disease' and a 'Search' button. Below the search bar, there are links for 'Save search', 'Limits', and 'Advanced'. The 'Display Settings' section shows 'Summary, 20 per page'. A 'Limits Activated' warning indicates 'Project type: Primary submission'. The 'Results' section shows '1 to 20 of 887' results. The first three results are listed, each with a checkbox, a numbered title, taxonomy, project data type, attributes, and accession/ID. On the right, there are sections for 'Find related data' (with a database selector and 'Find items' button) and 'Search details' (with a search box containing 'disease[All Fields] AND Primary submission[Project_Type]' and a 'Search' button). At the bottom right, there's a 'Recent activity' section showing 'Your browsing activity is empty.'

NCBI Resources How To My NCBI Sign In

BioProject BioProject disease Search

Save search Limits Advanced Help

Display Settings: Summary, 20 per page Send to: Filters: Manage Filters

Limits Activated: Project type: Primary submission Change Remove

Results: 1 to 20 of 887 << First < Prev Page 1 of 45 Next > Last >>

☐ Enterobacter mori LMG 25706

1. Enterobacter mori LMG 25706 Genome sequencing project
Taxonomy: [Enterobacter mori LMG 25706](#)
Project data type: RefSeq Genome
Attributes : Scope: Monoisolate; Material: Genome; Capture: Whole; Method Type: Other
NCBI
Accession: PRJNA75365 ID: 75365

☐ Genome-wide Analysis of Chronic Lymphocytic Leukemia

2. Genome-wide Analysis of Chronic Lymphocytic Leukemia
Organism: Homo sapiens
Taxonomy: [Homo sapiens \(human\)](#)
Project data type: Other
Attributes : Scope: Multiisolate; Material: Other; Capture: Other; Method Type: Other
Trustees of Columbia University in the City of New York
Accession: PRJNA75357 ID: 75357

☐ Pseudomonas syringae pv. actinidiae TP6-1

3. Pseudomonas syringae pv. actinidiae TP6-1 Genome sequencing
Taxonomy: [Pseudomonas syringae pv. actinidiae TP6-1](#)
Project data type: Genome sequencing
Attributes : Scope: Monoisolate; Material: Genome; Capture: Whole; Method Type: Sequencing
NZGL
Accession: PRJNA74977 ID: 74977

Find related data Database: Select Find items

Search details disease[All Fields] AND Primary submission[Project_Type] Search See more...

Recent activity Turn Off Clear

Your browsing activity is empty.

Figure 3. Summary display. A query for the term “disease”, with Limits activated to restrict to Primary submissions, returns 887 results. The first 3 results are shown in this view. Note the upper left link to ‘Display settings’ can be used to change the display to report a list of project accessions, or to show a larger number of results per page. Each result is numbered sequentially and includes the organism name or label, which is linked to view the full report, the project title, the taxonomy name, the project type, attributes, a member of the submitting group or consortium name, and the project accession and identifier.

Summary

When you submit a query, the results are shown in the Summary format (or ‘docsum’) as shown in Figure 3. In the Summary format, each result is numbered, and a check box is provided at the left of the record. The check box enables you to select which of the records in the retrieval set that you want to review in another format, according to your selection in the Display Settings box. If none are checked, then all results are displayed in the selected format; this is the same as having all the boxes checked.

The text of the Summary includes the Project name or label (which is often the organism name), title, Taxonomy, Project data type, Attributes, the project source (if multiple, then only one is listed here), and the BioProject accession and ID. The Project name or label is linked to the full report page and the Taxonomy term is linked to NCBI’s taxonomy database.

NCBI Resources ▾ How To ▾ My NCBI Sign In

BioProject BioProject 131[uid] Search

Save search Limits Advanced Help

Display Settings: ▾ Send to: ▾

Name: *Aspergillus fumigatus* Af293 Accession: PRJNA131 ID: 131
Title: WGS sequencing of strain Af293

[J. Craig Venter Institute](#) and the [Sanger Institute](#) sequenced the *Aspergillus fumigatus* strain Af293 genome at 10.5X coverage using whole genome shotgun (WGS) sequencing. The draft genome was finished, resulting in a final assembly that contains 19 contigs in eight supercontigs.

Project Data Type: Genome sequencing; **Locus Tag Prefix:** AFUA

Attributes: Scope: Monoisolate; Material: Genome; Capture: Whole; Method type: Sequencing;

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide	28
SRA Experiments	1
Protein Sequences	9630
PUBLICATIONS	
Pubmed	1

Genome assemblies, organelles and plasmids:

Name	GenBank
Chromosome 1	CM000169.1
Chromosome 2	CM000170.1
Chromosome 3	CM000171.1
Chromosome 4	CM000172.1
Chromosome 5	CM000173.1
Chromosome 6	CM000174.1
Chromosome 7	CM000175.1
Chromosome 8	CM000176.1
Whole Genome Shotgun Assembly	AAHF00000000

Related RefSeq Project
 PRJNA14003 : Reference genome sequence

Publications:

- Nierman WC *et al.*, "Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*.", *Nature*, 2005 Dec 22;438(7071):1151-6

Lineage: Eukaryota; Fungi; Dikarya; Ascomycota; Pezizomycotina; Eurotiomycetes; Eurotiomycetidae; Eurotiales; Trichocomaceae; mitosporic Trichocomaceae; *Aspergillus*; *Aspergillus fumigatus*; *Aspergillus fumigatus* Af293

Submission:
 Registration date: 1-Jun-2005
 J. Craig Venter Institute
 - Sanger Institute

See Genome Information for *Aspergillus fumigatus*

NAVIGATE ACROSS
 5 additional projects are related by organism.

Related information

- Nucleotide
- Project
- Protein
- PubMed
- Related Genes
- SRA
- Taxonomy

Related Resources

- GOLD

LinkOut to external resources

- GOLDCARD: Gc00277 [Genomes On Line Database]
- SILVA LSU Database [SILVA]
- SILVA SSU Database [SILVA]

Recent activity

Turn Off Clear

- Aspergillus fumigatus* Af293 BioProject (Genome Project)
- 131[uid] (1) BioProject (Genome Project)

See more...

Search details


131[uid]

Search

See more...

Figure 4. Full Report page. The BioProject accession and ID, organism name or label, project title, and project description (when provided) are shown at the top of the page. Below that, details about the project type and attributes, tabular report of available project data, citations (when provided), taxonomic lineage, and submitter information. The upper portion of the page includes navigation tools to facilitate navigating to the related Genomes resource which focuses on taxonomically organized genome sequencing projects, or to a linked Umbrella project (not shown here), or to 'peer' projects that share a link to the same Umbrella project or by shared taxonomy. If a genome assembly is represented by both a INSDC genome sequencing project, and a RefSeq genome project, then the correspondence between these projects is also indicated, as shown in this view ('Related RefSeq Project').

A.

NIH Human Microbiome Project (HMP) Roadmap Project encompasses the following 4 sub-projects:		
Project Type		Number of Projects
Umbrella project 		4
BioProject accession	Organism	Title
PRJNA48489	Human Microbiome Project (HMP) 16S rRNA Gene Diversity	Examining the diversity of 16S ribosomal RNA genes in the human microbiome for the Human Microbiome Project (HMP) (NIH Human Microbiome Consortium)
PRJNA46305	Human Microbiome Project (HMP) Demonstration Projects	The human microbiome and human health and disease (NIH Human Microbiome Consortium)
PRJNA43017	Human Microbiome Project (HMP) Metagenome Projects	Deeper shotgun sequencing of human microbiome samples for the Human Microbiome Project (HMP) (NIH Human Microbiome Consortium)
PRJNA28331	Human Microbiome Project (HMP) Reference Genomes	Genomes of microorganisms that have been isolated in and on the human body, to be used as Reference Genomes for the Human Microbiome Project (HMP) (NIH Human Microbiome Consortium)

B.


Human Microbiome Project (HMP) Metagenome Projects encompasses the following 3 sub-projects:		
Project Type		Number of Projects
Metagenome 		3
BioProject accession	Organism	Title
PRJNA48475	human metagenome	Human Microbiome Project (HMP) Metagenomic WGS Projects, deeper sequencing of the human microbiome samples: Mock Pilot (NIH Human Microbiome Consortium)
PRJNA48477	human metagenome	Human Microbiome Project (HMP) Metagenomic WGS Projects, deeper sequencing of the human microbiome samples: Clinical Pilot (NIH Human Microbiome Consortium)
PRJNA48479	human metagenome	Human Microbiome Project (HMP) Metagenomic WGS Projects, deeper sequencing of the human microbiome samples: Production Phase (NIH Human Microbiome Consortium)

Figure 5. A. The Umbrella Project table displayed for BioProject accession PRJNA43021. The table indicates that PRJNA43021 is grouping four sub-projects that are also Umbrella project types (arrow). Clicking the BioProject accession navigates to that projects report page. B. The Umbrella table displayed for BioProject accession PRJNA43017. The table reports Primary submission projects, of type Megagenome (arrow) that are grouped under PRJNA43017.

Accessions List or BioProject ID List

These display only the BioProject accession number or the BioProject ID respectively, for the query result set.

Full Reports

The full report page can be accessed by following the hyperlinked project name presented in the top row of each result returned in the Summary display. The Full Report display for Primary submission projects, as shown in Figure 4, includes the project name and/or title, a text description of the project (when provided), the project data type and specific project attributes, a project data section with data links, citations relevant to the project, taxonomic lineage, and information about the submitting group. Navigation tools are provided near the top of the report to facilitate navigation to NCBI's taxonomically organized Genomes resource, 'up' to higher-level Umbrella projects, or 'across' to other BioProject records that are related by organism, or via a common

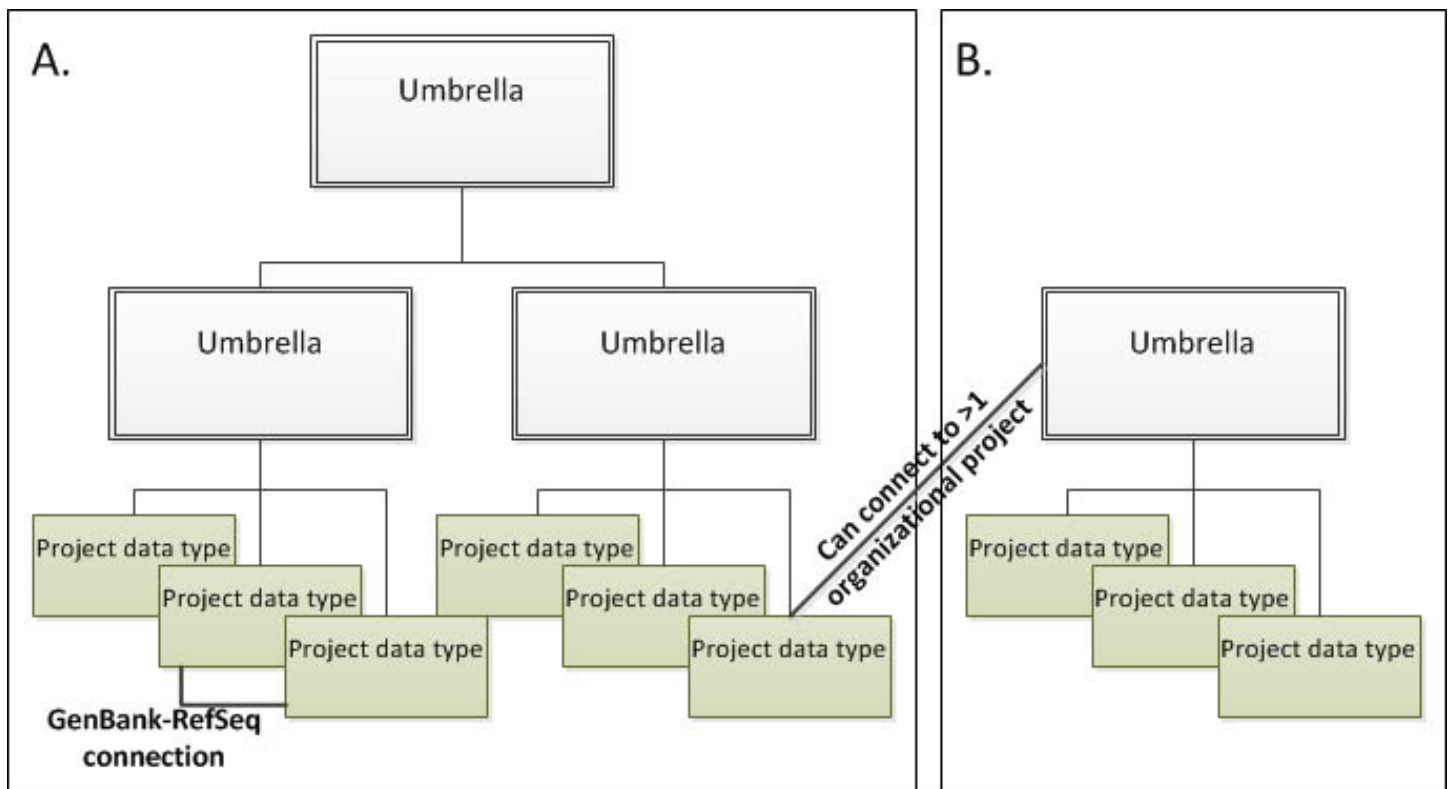


Figure 6. Schematic diagram of BioProject hierarchies. A. Large initiatives which have distinct sub-projects may have more than one level of Umbrella project. For example, a top-level Umbrella project groups all components of the initiative; mid-level Umbrella projects reflect distinct branches of the project (such as sequencing vs. epigenetics); and several Submission projects denote distinct project data types (e.g., genome sequencing, transcriptome, epigenetics, etc.). B. Other initiatives may be organized under a single Umbrella project with one or many submitted projects that are connected to data. Note that a given submission project may have no connection to any hierarchical Umbrella projects, or may be connected to more than one organizational layer, and there may be connections directly between submitted projects such as the indicated RefSeq to GenBank link.

Umbrella project. Umbrella project reports pages include a tabular report (when relevant) listing Umbrella sub-projects (see the HMP report PRJNA43021), or listing Primary submission projects, organized by Project data type, that it is grouping (Figure 5) (see the HMP Reference Genome project PRJNA28331). For all report pages, the right column provides navigation links and standard Entrez functions such as browsing history.

Some large initiatives are represented by more than one layer of umbrella projects (see Figure 6); for instance, a top-most level may identify the largest definition of the collaboration; a second level of umbrella projects identify the primary categories of data production; and finally a third layer represents the projects that actually generate the data that is submitted. The Human Microbiome project is an example of this type of complex hierarchy where the top-most project, PRJNA43021, represents the most inclusive definition of the initiative, and a secondary level (such as PRJNA28331) identifies a major sub-project to sequence multiple reference genomes each of which has a distinct project accession.

Genome sequencing projects may include a table that reports the accession numbers for assembled chromosomes, linkage groups, or other replicating molecules (such as organelles or plasmids), as well as the master accession for whole genome sequencing (WGS) projects.

When the experimental data for a BioProject is submitted to archival databases, it contains the BioProject accession which links the data to the BioProject. The Project Data table presents data counts from databases at NCBI that have links to the displayed BioProject record; if the displayed record is an Umbrella project, then the Data table presents a sum of data links for the grouped sub-projects. The counts are hyperlinked to the NCBI

database indicated when a component (non-Umbrella) project is displayed. Hyperlinks are currently provided on Umbrella projects only in those cases when the database indicated holds links to a single component BioProject grouped under the Umbrella project; consequently, hyperlinks are not always present on Umbrella projects.

Submitting to BioProject

Projects can be registered with the BioProject database using the [Submission portal](#) (access the link from the home page). The Submission portal requires authentication and provides several login options, including National Institutes of Health eRA Login and other NIH logins. A login for the NCBI PDA (Primary Data Archives) system can be created here, if the user does not have any of the other types of accounts. Once logged in, the submission wizard provides a list of previously created submissions with some simple status information and a button to initiate a new submission. When making a new submission, the wizard presents a series of pages where information about the project can be entered. Required fields are marked with an asterisk (*) and simple validation identifies missing required content with red highlighting and warns of data attribute combinations that are observed less frequently. In-line help can be presented by hovering over the blue '?' icons. The submission wizard pages must be completed in the order presented, but after that it is possible to navigate back to previous pages using tabs available along the top of the page. The content of each page is saved by clicking on the 'Continue' button located at the bottom of each page. To edit content on a previous page, the 'Continue' button must be clicked to save your changes. A submission may be started, set aside, and completed at a later time by signing back into the Submission wizard and selecting the incomplete project.

The page tabs presented by the Submission wizard are:

- **Submitter** – the name and email information is auto-filled if logging on using a NIH-based login approach and should identify the person who is entering the data in the form.
- **General info** – this page collects general descriptive information about the project, its relevance, whether it is part of a large initiative that has already registered with the BioProject resource, related web resources that are specific to the project, funding information, and information about the consortium or center name and/or data provider.
- **Project data type** – this page collects more specific information about the Sample Scope, Material, and other attributes (Capture and Method), as well as the Objective or goals of the project being registered. See the Glossary for descriptions of the attributes.
- **Target** – this page collects organism information (for projects focusing on an identified organism) or labeling information for projects that encompass multiple species whether identified or not (an environmental sample). An optional section, Biological Properties, collects general properties for samples that represent a single organism. This information is readily known for model organisms, but is not as readily available for lesser known organisms. Information about the number of chromosomes, genome size, mode of reproduction, and general habitat provides some useful context for other scientists who may be interested in the project and submitted data.
- **Publications** – this page collects publication information specific to the registered project. A publication identifier is required. A PubMed ID is preferred, but lacking that then a DOI may be supplied.
- **Overview** – this page presents a summary of the provided information. Click the 'Submit' button at the bottom of the page to complete the submission.

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 initiative 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