

BioProject Frequently Asked Questions:

- [Submission Questions](#)
 - [What is a BioProject?](#)
 - [Under what circumstances is it necessary to register a BioProject?](#)
 - [How do I submit to BioProject?](#)
 - [What information should I provide about my BioProject?](#)
 - [Do I need to make a separate BioProject for every type of data?](#)
 - [What is Project Data Type?](#)
 - [What is Sample Scope?](#)
 - [When should I choose 'Multiisolate' as the scope for a BioProject?](#)
 - [What types of validation must my BioProject pass?](#)
 - [When will I receive my BioProject accession number?](#)
 - [When will my BioProject record be released?](#)
 - [Will NCBI apply further curation to my BioProject records?](#)
 - [How do I update my BioProject?](#)
 - [Should I cite BioProject accession numbers in my manuscript?](#)
 - [How do I get a locus_tag prefix for annotating a genome assembly?](#)
 - [How do I create an Umbrella BioProject?](#)

Submission Questions

What is a BioProject?

A BioProject is a collection of biological data related to a single initiative originating from a single organization or from a consortium. A BioProject record provides users a single place to find links to the diverse data generated for that project and deposited into the archival databases maintained by members of the INSDC. Typical examples of a BioProject include a multiisolate project for sequencing multiple strains of a bacterial species, or a monoisolate project for the genome and transcriptome of a particular organism. The description you supply about this research effort is important for providing context to your experimental data.

Under what circumstances is it necessary to register a BioProject?

BioProject registration is required as part of data deposit to several NCBI primary data archives including [SRA](#), [TSA](#) and [WGS](#). Typically, a BioProject is registered first or during the submission of a genome assembly being submitted to WGS. The BioProject is assigned a BioProject accession number (PRJNAxxxxxx) which is referenced when submitting the corresponding BioSamples and experimental data to archival databases. Use the same BioProject accession for related data, eg the raw reads that are submitted to SRA and a genome assembly of those reads that is submitted to GenBank/WGS. At this time, BioProject submission is not required for [GEO](#) or [dbGaP](#); deposit to those databases triggers automatic creation of BioProject records.

How do I submit to BioProject?

Several submission routes are supported:

[BioProject Submission Portal](#) to pre-register a project before submitting data

- Online wizard that supports single submission using web forms.

[Genome Submission Portal](#) to register a project while submitting a prokaryotic or eukaryotic genome

- Online wizard that supports single or batch submission using web forms
- **Most submitters should use this method or the next one**

[SRA Submission Portal](#) to register a project while submitting sequence reads to SRA

- Online wizard that supports single or batch submission using web forms
- **Most submitters should use this method or the previous one**

[XML deposit](#) to pre-register

- Programmatic API deposit in XML format. Suitable only when data is stored in an inhouse database or LIMS, and from which valid BioProject XML can be generated. Here are the [instructions and schemas](#).

What information should I provide about my BioProject?

You need to indicate the type of project and the sample scope, which are defined in the [Glossary](#) and below.

A description of the project is also required. Provide comprehensive information that will allow users to fully understand your research study.



Although it is not required, it is highly recommended that you include the grant(s) associated with the research effort. A new feature in September 2015 is that you can look up your NIH grant during the submission process. For non-NIH grants you'll need to provide the grant ID and title as well as the funding agency.

Depending upon the scope of the project, the organism is required. For example, a monoisolate BioProject requires the genus and species, but you should also provide the infraspecific identifier (strain, breed, cultivar or isolate) that will be registered in BioSample for that BioProject.

Do I need to make a separate BioProject for every type of data?

No, you do not. You should organize your BioProjects the most appropriate way for your research effort. For example, if you are creating both transcriptome and genome assemblies of an organism, then you could register a single "Genome sequencing and assembly" BioProject and submit all of the data with that BioProject. Once the data are public, the BioProject will be automatically updated with links to the data and the additional project type will be added. Be sure to include all the goals of the project in the Description.

What is Project Data Type?

"Data Type" or "Project Data Type" is a general label indicating the initial primary study goal(s). You must select one, but can select multiple goals. Note that the value selected now does not limit the sort of data that can be associated with this BioProject later. A BioProject can have any sort of data linked to it, regardless of the initially selected "Project Data Type".

"Genome sequencing" is set automatically as the Project Data Type of BioProjects that are created during submission of prokaryotic or eukaryotic genomes. "Raw sequence reads" is set automatically as the Project Data Type of BioProjects that are created during submission of sequences reads to SRA.

See the [Help documentation](#) for more information about Project Data Type.

What is Sample Scope?

"Sample scope" indicates the scope and purity of the biological sample used for the study. Select the most appropriate value:

- Monoisolate: a single organism (eg, animal, cultured cell-line, inbred population) is being studied in this research effort.
- Multiisolate: multiple individuals of the same species are being studied in this research effort.
- Multi-species: multiple species are being studied in this research effort.
- Environment: the species content of the sample is not known because the nucleic acid was directly isolated from an environmental sample for analysis. This is used for metagenome studies.
- Synthetic: the sample is synthesized in a laboratory.
- Other: the scope was not defined.

"Monoisolate" is set automatically as the Scope of BioProjects that are created during submission of a single prokaryotic or eukaryotic genome. "Multispecies" is set automatically as the Scope of BioProjects that are created during batch submission of genomes or the submission of sequences reads to SRA.

When should I choose 'Multiisolate' as the scope for a BioProject?

Choose Multiisolate as the Scope when the goal of the research is to compare multiple individuals or strains of the same species, eg, in a 'Variation' or 'Genome sequencing and assembly' project. Choose Multispecies when different species are being examined. Choose Monoisolate if the goal is to make a single genome or transcriptome assembly, even if more than one individual was the source of the DNA or RNA.

What types of validation must my BioProject pass?

Beyond providing the required information on the submission web pages, the only validation is that the BioProject cannot be a duplicate. BioProjects from the same submitter are unique if any of these is different:

- Organism name, strain or isolate
- Project type
- Grant
- Organizations, eg a different Consortium
- External links to non-NCBI resources
- Title (this is usually auto-generated from the organism and project type for monoisolate projects)

When will I receive my BioProject accession number?

If your submission passes validation, you can expect to receive a BioProject accession number(s) within a few minutes by email.

When will my BioProject record be released?

During submission, you are presented two options for releasing your BioProject to the public. If you select 'Release immediately upon curation' the records will be released within a few hours of having a valid organism name. If you select 'Release on a specified date', the BioProject will be released on the date you specify or upon the release of any data that reference that

BioProject accession, whichever is first. At this time, we do not have a mechanism in place for you to view your records before release.

Will NCBI apply further curation to my BioProject records?

No, BioProject is a submitter-driven repository. Submitters are responsible for the content and accuracy of their records, and for ensuring that sufficient information has been provided to allow users to fully interpret their study. BioProject submissions must pass basic validation rules and taxonomy review. Otherwise, records are generally not subject to further curation.

How do I update my BioProject?

At this time, it is necessary for submitters to write to bioprojecthelp@ncbi.nlm.nih.gov to request updates and withdrawals as necessary. Please note that when BioProjects are updated, the Submission Overview page in the Submission Portal will not reflect this change. That page is only a record of the initial submission, and does not display changes made in the BioProject database.

Should I cite BioProject accession numbers in my manuscript?

No, typically, you should cite the accession numbers that are assigned to your data submissions, e.g. the GenBank, WGS or SRA accession numbers. If individual BioProjects do need to be referenced, state that "The data have been deposited with links to BioProject accession number PRJNAXxxxxx in the NCBI BioProject database (<https://www.ncbi.nlm.nih.gov/bioproject/>)."

How do I get a locus_tag prefix for annotating a genome assembly?

A locus_tag prefix is automatically assigned to each BioProject/BioSample pair of a "Genome sequencing and assembly" project, but you must register the BioProject first and then register the BioSample(s) associated with that BioProject. The locus_tag prefixes are reported back in the [BioProject submission portal](#). If there are multiple prefixes, they are reported there in a file named "locustagprefix.txt". If there are problems, write to bioprojecthelp@ncbi.nlm.nih.gov.

If you request to have a prokaryotic genome annotated by NCBI's [Prokaryotic Genome Annotation Pipeline](#) (PGAP), then you need to have a BioProject and BioSample registered for that genome. PGAP will ensure that there is a locus_tag prefix for the genome when the pipeline is run, and the registered locus_tag prefix will be reported back in the BioProject submission portal.

How do I create an Umbrella BioProject?

If you want to cluster several of your data-level projects under an Umbrella, write to bioprojecthelp@ncbi.nlm.nih.gov with details about what projects you want to cluster and why.