Isak Sylvin
Swedish University of Agriculture (SLU)
2014-08-24

NCBI Data Submission

Table of Contents

To what database should I submit my data?

The lecture today will briefly mention where most data types should go. In case you have a very special type of data (or just poor memory) check out the following links:

1. *Navigate to https://submit.ncbi.nlm.nih.gov/ and see if you can figure it out*
2. *Else check out http://www.ncbi.nlm.nih.gov/guide/howto/submit-data/ for more info*

What if I really, really can't find an answer after this course?

Make sure you've really checked the NCBI webpage for help. Maybe you even googled "site: http://www.ncbi.nlm.nih.gov WHAT IS COVERAGE". As a final resort you can mail NCBI at info@ncbi.nlm.nih.gov .

Creating an NCBI account

1. Navigate to NCBI homepage: www.ncbi.nlm.nih.gov
2. Use the "*Sign in to NCBI"* in the top right
3. Register using "*Register for an NCBI account"* hyperlink to the left
4. Open up your e-mail and validate your registration; remember to check your spam folder!

GenBank submission

Submission through BankIt

Start at the submission portal: https://submit.ncbi.nlm.nih.gov/ . Click the Genbank link followed by the BankIt link. Then press the "Sign in to use BankIt" link in the top right.

Note: Sequence must be in fasta (not fastq) and include definition lines.

>Seq2 [organism=Mus musculus] ***Mouse strain BMC2/3 cytochrome b (cytb) complete CDS***
ttatatcgatatgacacccgggatatacagatattagggata

Definition line is featured in bold. Failure provide them will mean the submission will either be delayed or outright refused.

Use *"Sign in to use BankIt"*, top-left

If you have an account, log in with it, otherwise see "*How do I get a NCBI account?"*

Click *"New submissions"*. The following provide some outlines for each tab:

Contact

Don't need to fill in all fields, like fax numbers. Notice that BankIt asks you to press continue again even if the fields you skip are optional.

Reference

Add authors and at least the title of the paper to write.

Sequencing tech

Assembly name is optional, NCBI adds its own if you don’t provide one. You should be able to find the coverage of your sequence by either looking at the output of the assembler, or tallying with a bash script.

Nucleotide

Molecule type, topology and genomic completeness are all very important but luckily very intuitive. Don’t worry about providing a correct nucleotide sequence amount as BankIt counts it for you in case you leave it blank.

Submission category

Either you created the data from scratch or you improved someone’s work.

Source modifiers

Organelle/Location is per default genomic and doesn’t need to be selected.

Add as many as source modifiers as possible. Clicking the optional ‘include primers’ checkbox opens up the primers taö. You can use a tab delimited file (if data varies between samples, or you don’t want to use the web form). This is an example tab delimited file:

Sequence_ID	Specimen_voucher	Collected_by	Collection_date	Country	Identified_by	Lat_Lon
Seq1	MKP 334	C. Grant	31-Jan-2001	USA	C. Grant	13.57 N 24.68 W
Seq2	MKP 1230	S. Tracy	28-Feb-2002	Slovakia	C. Grant	13.24 N 24.35 W

Commonly used Source Modifiers

- **Clone** - Name of clone from which sequence was obtained.
- **Collection_date** - Date the specimen was collected.
  In format **DD-Mon-YYYY**, that is 2-digit date, three-character abbreviation of month, and 4-digit year, (*e.g.*, 11-Feb-2002).
  Mon-YYYY and **YYYY** are alternate formats to use when date information is less complete.
- **Country** - The country where the sequence's organism was located. May also be an ocean or major sea. Additional region or locality information must be after the country name and separated by a ':'. For example: USA: Riverview Park, Ripkentown, MD
- **Host** - When the sequence submission is from an organism that exists in a symbiotic, parasitic, or other special relationship with some second organism, the 'host' modifier can be used to identify the name of the host species.

- **Isolate** - Identification or description of the specific individual from which this sequence was obtained.

- **Isolation source** - Describes the local geographical source of the organism from which the sequence was obtained.

- **Specimen_voucher** - An identifier of the individual or collection of the source organism and the place where it is currently stored, usually an institution.

This should be provided using the following format 'institution-code:collection-code:specimen-id'. specimen-id is mandatory, collection-code is optional; institution-code is mandatory when collection-code is provided. Examples:

  - 99-SRNP
  - UAM:Mamm:52179
  - personal collection:Joe Smith:99-SRNP
  - AMCC:101706

- **Strain** - Strain of organism from which sequence was obtained.

The following source modifiers are available to further describe the sequences in a BankIt set:

- **Altitude** - Altitude in metres above or below sea level of where the sample was collected.

- **Authority** - The author or authors of the organism name from which sequence was obtained.

- **Bio_material** - An identifier for the biological material from which the nucleotide sequence was obtained, with optional institution code and collection code for the place where it is currently stored.

This should be provided using the following format **'institution-code:collection-code:material_id'**. material_id is mandatory, institution-code and collection-code are optional; institution-code is mandatory when collection-code is present.

This qualifier should be used to annotate the identifiers of material in biological collections which include zoos and aquaria, stock centers, seed banks, germplasm repositories and DNA banks.

- **Biotype** - Variety of a species (usually a fungus, bacteria, or virus) characterized by some specific biological property (often geographical, ecological, or physiological). Same as biotype.

- **Biovar** - See biotype

- **Breed** - The named breed from which sequence was obtained (usually applied to domesticated mammals).

- **Cell_line** - Cell line from which sequence was obtained.

- **Cell_type** - Type of cell from which sequence was obtained.
- **Chemovar** - Variety of a species (usually a fungus, bacteria, or virus) characterized by its biochemical properties.
- **Clone** - Name of clone from which sequence was obtained.
- **Collected_by** - Name of person who collected the sample.
- **Collection_date** - Date the specimen was collected.
  In format **DD-Mon-YYYY**, that is 2-digit date, three-character abbreviation of month, and 4-digit year, (*e.g.*, 11-Feb-2002).
  Mon-YYYY and **YYYY** are alternate formats to use when date information is less complete.
- **Country** - The country where the sequence's organism was located. May also be an ocean or major sea. Additional region or locality information must be after the country name and separated by a ':'. For example: USA: Riverview Park, Ripkentown, MD
- **Cultivar** - Cultivated variety of plant from which sequence was obtained.
- **Culture_collection** - Institution code and identifier for the culture from which the nucleotide sequence was obtained, with optional collection code.

This should be provided using the following format **'institution-code:collection-code:culture-id'**. culture-id and institution-code are mandatory.

This qualifier should be used to annotate live microbial and viral cultures, and cell lines that have been deposited in curated culture collections.

- **Dev_stage** - Developmental stage of organism.
- **Ecotype** - The named ecotype (population adapted to a local habitat) from which sequence was obtained (customarily applied to populations of Arabidopsis thaliana).
- **Forma** - The forma (lowest taxonomic unit governed by the nomenclatural codes) of organism from which sequence was obtained. This term is usually applied to plants and fungi.
- **Forma_specialis** - The physiologically distinct form from which sequence was obtained (usually restricted to certain parasitic fungi).
- **Fwd_primer_name** - name of forward PCR primer
- **Fwd_primer_seq** - nucleotide sequence of forward PCR primer
- **Genotype** - Genotype of the organism.
- **Haplogroup** - Name for a group of similar haplotypes that share some sequence variation
- **Haplotype** - Haplotype of the organism.

- **Host** - When the sequence submission is from an organism that exists in a symbiotic, parasitic, or other special relationship with some second organism, the 'host' modifier can be used to identify the name of the host species.
- **Identified_by** - name of the person or persons who identified by taxonomic name the organism from which the sequence was obtained
- **Isolate** - Identification or description of the specific individual from which this sequence was obtained.
- **Isolation source** - Describes the local geographical source of the organism from which the sequence was obtained.
- **Lab_host** - Laboratory host used to propagate the organism from which the sequence was obtained.
- **Lat_Lon** - Latitude and longitude, in decimal degrees, of where the sample was collected.
- **Note** - Any additional information that you wish to provide about the sequence.
- **Pathovar** - Variety of a species (usually a fungus, bacteria or virus) characterized by the biological target of the pathogen. Examples include Pseudomonas syringae pathovar tomato and Pseudomonas syringae pathovar tabaci.
- **Pop_variant** - name of the population variant from which the sequence was obtained
- **Rev_primer_name** - name of reverse PCR primer
- **Rev_primer_seq** - nucleotide sequence of reverse PCR primer
- **Specimen_voucher** - An identifier of the individual or collection of the source organism and the place where it is currently stored, usually an institution.

This should be provided using the following format 'institution-code:collection-code:specimen-id'. specimen-id is mandatory, collection-code is optional; institution-code is mandatory when collection-code is provided. Examples:

  - 99-SRNP
  - UAM:Mamm:52179
  - personal collection:Joe Smith:99-SRNP
  - AMCC:101706

- **Serogroup** - Variety of a species (usually a fungus, bacteria, or virus) characterized by its antigenic properties. Same as serogroup and serovar.
- **Serotype** - See Serogroup
- **Serovar** - See Serogroup

- **Sex** - Sex of the organism from which the sequence was obtained.
- **Strain** - Strain of organism from which sequence was obtained.
- **Sub_species** - Subspecies of organism from which sequence was obtained.
- **Subclone** - Name of subclone from which sequence was obtained.
- **Subtype** - Subtype of organism from which sequence was obtained.
- **Substrain** - Sub-strain of organism from which sequence was obtained.
- **Tissue_lib** - Tissue library from which the sequence was obtained.
- **Tissue_type** - Type of tissue from which sequence was obtained.
- **Type** - Type of organism from which sequence was obtained.
- **Variety** - Variety of organism from which sequence was obtained.

Primers

Different reaction sets = Primers separated by multiple reactions

Features

Features should be provided in a document in Plain ASCII. It can be provided in the web form but it's really, really tedious.

Every sequence is divided into sections. Every row is a pair of identifier and then value.

If a feature is reversed, so are the indexes.

< > means incomplete (partial features) meaning they start and stop upsteams and downstreams of the nucleotide positions respectively.

All genes should include a gene index which is positioned so gene = 5'UTR+CDS+3'UTR.

If you get unsure about how to annotate something you can always mail info@ncbi.nlm.nih.gov .

```
>Feature Seq1
<1    >1050    gene
               gene        ATH1
<1    1009    CDS
               product      acid trehalase
               product      Athlp
               codon_start   2
<1    >1050    mRNA
               product      acid trehalase

>Feature Seq2
2626  2590    tRNA
2570  2535
```

```
                product     tRNA-Phe

>Feature Seq3
1080  1210  CDS
1275  1315
                product     actin
                note        alternatively spliced
1055  1210  mRNA
1275  1340
                product     actin
1055  1340  gene
                gene        ACT
1055  1079  5'UTR
1316  1340  3'UTR
```

Review and correct

You may download your complete set as a zip file.

Do not press "Finish submission" as it sends your test to NCBI!

Finally in case you haven't received an automatic reply, your genbank accession number or final records you can always mail gb-admin@ncbi.nlm.nih.gov to see the status of your project.

Sequin

Sequin can be used locally on your machine. Just download

http://www.ncbi.nlm.nih.gov/Sequin/download/seq_download.html

make a directory and move the file there and doubleclick on the sequin executable. A bunch of files should be generated. Start sequin.exe. If you need a fasta file to use as template you can download one from http://hpc.ilri.cgiar.org/~isylvin/seqFasta.fasta

Metadata creation

BioSample

Start at the submission portal: https://submit.ncbi.nlm.nih.gov/ . Click the BioSample link.
In the future in case you want to create multiple BioSamples at once there's a link to download a "batch template". Open the file in Excel and add info to it, at least to all columns marked with an *. If you're unsure about what to put in each field, use https://submit.ncbi.nlm.nih.gov/biosample/template/?package=MIGS.eu.human-associated.4.0&action=definition as a reference.

Sample type

Choose "Genome, metagenome or marker sequences" per default. Just make sure your data is MIxS compliant (Minimum information about (x) sequences).

Attributes

Pick something nice for sample name.

For “isolation and growth condition” you’ll be needing a PMID or similar URL for the protocol/SOP.

The ”reference for biomaterial” requires an uploaded report as well.

For “geographic location” most country names exist. However, a full list is located here: *http://www.insdc.org/documents/country-qualifier-vocabulary*

Overview

Make sure you don’t hit SUBMIT as we’re mostly fooling around with our entries

BioProject

Start at the submission portal: https://submit.ncbi.nlm.nih.gov/ . Click the BioProject link.

Please make sure to create a BioSample before you start on a BioProject.

Use https://submit.ncbi.nlm.nih.gov/ and select “BioProject”

Submitter

Boring and trivial. Just make sure you get the organization and department right.

Project type/ Target

It is highly suggested to use Google to find the definitions of the terms you’re unsure of. Even if the data is submitted properly, inputting incorrect classifications into the necessary data might poison further studies.

Make sure you fill in as many optional fields as possible during a real run.

BioSample

Use the format SUBxxxxx: BIOSAMPLENAME.

Publications

One of the few steps that is actually validated. Use a PubMed id like “25107883” to continue.

A doi is very similar to an ISBN, and is much more general than a PubMed id (PMID).

Overview

Make sure you don’t hit SUBMIT as we’re mostly fooling around with our entries

Submitting GEO data

Start at the submission portal: https://submit.ncbi.nlm.nih.gov/ . Click the microarray GEO link.

If you ever need professional help you can mail geo@ncbi.nlm.nih.gov .

First of all press the “Submit” button on the page. Then upload the data and fill out the form. The preferred format is GEOarchive.
Make sure you don’t hit SUBMIT as we’re mostly fooling around with our entries

If you want to doublecheck your SOFT or MiniML formatted data, use http://www.ncbi.nlm.nih.gov/geo/submission/depslip.cgi?subm=0 and actually submit to it to test your format.

Submitting SRA data

Start at the submission portal: https://submit.ncbi.nlm.nih.gov/ . Click the SRA link. Bear in mind all SRA studies need a BioProject or at least an associated BioSample.

Use either login route, preferably NIH. Sometimes the login is down. Enter an alias (project description) and possibly an internal comment.

Click “Set new experiment”

Alias is the experiment name; title is used to call out individual records from the experiment.

Add library info about how the data was sequenced.

Pipeline refers to **all** the bioinformatical programs used to manipulate the data.

Links and attributes lets you add links like DDBJ.

Save then click new run. Add new files in a format like fastq or bam.