



Submitting Eukaryotic Genomes to GenBank

Preparing eukaryotic genomes for submission to GenBank

https://www.ncbi.nlm.nih.gov/genbank/eukaryotic_submission

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

Introduction

Submission of eukaryotic genomes to GenBank includes registering the metadata, such as the research efforts and source materials, and submitting the appropriate data files. The steps include:

- **Registering the research project**
- **Registering the sample source**
- **Preparing the data files**
- **Submitting the data**

They are described below, and additional details are available in the online documentation:

www.ncbi.nlm.nih.gov/genbank/eukaryotic_submission

Registering Research Effort in BioProject

The BioProject records the research effort of the registering researcher or laboratory and it ties together the discrete datasets generated from that research effort. A BioProject entry can be used for a single experiment or for multiple experiments, e.g., comparison of multiple strains of the same species or analyzing the genome and transcriptome assemblies of the same organism.

There are several ways to register a BioProject:

- During a genome submission submit.ncbi.nlm.nih.gov/subs/genome/, or
- During submission of the reads to SRA submit.ncbi.nlm.nih.gov/subs/sra/, or
- Singly by way of submit.ncbi.nlm.nih.gov/subs/bioproject/, or
- During a BioSample registration (see section below) submit.ncbi.nlm.nih.gov/subs/biosample/

The only time that you need to pre-register a BioProject is when you are submitting with annotation because you need a locus_tag prefix, which is assigned to the BioProject:BioSample pair. The same BioProject identifier, PRJNA#####, is included with each set of data that is submitted, e.g. raw reads, BAM file, VCF file, TSA assembly, and genome assembly.

Registering Details of the Source of Nucleotide Sequences in BioSample

This registration provides the metadata about the source of the RNA or DNA, e.g., the location where the animal lived when the sample was collected, the sex of the animal, and the breed of the animal or the cultivar of the plant, and the isolate number or specimen voucher of the animal or plant.

Basic guidance for BioSample registration are:

- Register a separate BioSample for each unique source, e.g., RNA from the wings is a separate BioSample than RNA from legs if those two sources were sequenced independently
- A genome assembly can have only one BioSample. For a genome assembled from reads of multiple BioSamples, register a new BioSample and indicate which other BioSamples were used to generate the assembly. For example, if the reads from a male and from a female were submitted to SRA separately but the reads were combined to assemble the genome, register a new BioSample for the male plus the female, providing the identifiers of the male and female BioSamples in the new BioSample registration.
- Endosymbionts: Because sequences are annotated by genome, one would need a separate BioProject/BioSample pair for an insect and its endosymbiont. In this case, we recommend indicating that the endosymbiont's BioSample is separate and references the insect BioSample.
- Different aliquots of the sample can be registered either as a single BioSample or as multiple BioSamples, depending upon the research focus.
- BioSample can be registered either
 - ◇ During a genome submission submit.ncbi.nlm.nih.gov/subs/genome/, or
 - ◇ During submission of the reads to SRA submit.ncbi.nlm.nih.gov/subs/sra/, or
 - ◇ Singly or in batch through the BioSample portal submit.ncbi.nlm.nih.gov/subs/biosample/

The only time that you need to pre-register a BioSample is when you are submitting with annotation because you need a locus_tag prefix, which is assigned to the BioProject:BioSample pair. Register the BioProject first, and then include the BioProject's identifier (PRJN#####) during the BioSample registration. A file of the locus_tag prefix(es) for the BioSample(s) within a BioProject is linked to the BioProject submission at submit.ncbi.nlm.nih.gov/subs/bioproject/. Write to genomes@ncbi.nlm.nih.gov if you did not receive a locus_tag prefix in this file after preregistering a BioSample for your BioProject. The same BioSample identifier (SAMN#####) is included with each relevant set of data that is submitted, e.g., reads, BAM file, vcf file, TSA assembly, and genome assembly.

Submission Scenarios for Different Data Types

Data submission table

Submission scenarios and data files (required or optional) to be submitted for each scenario are summarized in the table below followed by additional explanation for each scenario.

Submission Scenarios	BioProject & BioSample Registration	SRA			TSA	WGS/Genome		
		Reads	BAM	.vcf	.sqn	FASTA	.sqn	AGP
Unassembled reads	X	X	C	C				
Transcriptome options	X	X	O ²	C	O ²			
Multiple related unassembled genomes (e.g., multiple strains)	X		X	C				
Genome assembly options:								
Unannotated contigs or scaffolds	X	C				X		O
Annotated contigs or scaffolds without an AGP	X	C					X	
Annotated contigs or scaffolds with an AGP	X	C					X*	O ¹

NOTE:

X = submit **C** = strongly recommended **O** = optional **X*** = .sqn file of annotated scaffolds

1 = required only if scaffolds are assembled into chromosomes

2 = submit either a BAM file, or the reads plus .sqn files of the assembled sequences

Unassembled reads (No assembly)

DNA or RNA-seq reads (NOT the processed FASTA) should be submitted to Sequence Read Archive (SRA). If possible, BAM is preferred over FASTQ, and vcf files are strongly recommended. See www.ncbi.nlm.nih.gov/sra/docs/submit/

Transcriptome options

- BAM file can be submitted to SRA, or
- The raw reads can be submitted to SRA and the assembled sequences can be submitted to TSA/GenBank in ASN (.sqn) or fasta format. The SRA run identifiers (SRR#) will be collected in the TSA submission form. More information is available at: www.ncbi.nlm.nih.gov/genbank/tsaguide.

Common errors to avoid in a TSA submission:

- Sequences are less than 200 bp in length
- Sequences with vector screening hits for Next-Gen sequencing primers, i.e. not properly trimmed
- Sequences with more than 10% N's or 14 consecutive N's not in assembly_gap features format
- Files that are incorrectly formatted or have biologically invalid annotation

Multiple related genomes, reference-guided assemblies of multiple strains that are unassembled

BAM file plus optional vcf file of SNPs should be submitted to SRA. For more information on .vcf format, see: www.ncbi.nlm.nih.gov/SNP/docs/dbSNP_VCF_Submission.pdf and www.1000genomes.org/node/101

Genome assembly options

The fasta sequences of the scaffolds can be submitted along with the information to convert runs of Ns into assembly_gaps in most circumstances. Chromosomes that are assembled from scaffolds need to be submitted as an AGP file, but chromosomes that do not contain scaffold-breaking gaps can be submitted as sequences in either fasta or ASN files, depending on the situation. Annotation should be on the scaffold or chromosome sequences, and must be submitted as ASN (.sqn) files.

Genome assemblies, along with the genome assembly metadata, are to be submitted via the genome submission portal (submit.ncbi.nlm.nih.gov/subs/genome). The required metadata includes the assembly method(s), the date or version the program was run, the approximate genome coverage, and the relevant sequencing technology(ies) used. Optional metadata include the polishing method used, and whether it is a reference-guided assembly (see go.usa.gov/xpvtc). Multiple genomes can be submitted in a single batch submission when specific criteria are met (see go.usa.gov/xpvttr). For genomes with complex annotation, it is useful to submit the FASTA sequences first and request that they be run through the foreign contamination screen, to ensure that any contamination is removed before the submission files are created and annotated.

For genomes with simple gaps and without annotation:

Sequences with Ns that represent gaps can be submitted as fasta files if the following criteria are met:

- Each sequence represents a sequence that occurs biologically in the organism, such as a sub-chromosome scaffold, the contigs were not simply concatenated

(cont.)

Genome assembly options (cont.)

- The unplaced sequences are not collected together into a “chrUn”
- No artificial sequences, such as linkers with multiple stop codons, are present
- The linkage evidence for each gap is the same
- Only a single length is used for gaps of completely unknown size (if such gaps are present)
- All runs of "ambiguous base Ns" are shorter than any run of Ns that represents a gap
- All the gaps are 'within scaffolds', not 'between scaffolds'

Information about whether Ns represent gaps will be collected in the submission form. The default values are:

- 10 or more Ns is a gap (Assembly statistics are always calculated using 10 or more Ns as a gap, regardless of the presence/absence of gaps in the final genome sequence)
- All gaps are of estimated (approximate) size
- "paired-ends" is the linkage evidence connecting the sequences on either side of the gaps (Submitter should change these when the default values are not correct).

Otherwise, provide the contig fasta sequences and create the scaffolds via an AGP file or create scaffold .sqn files as described below.

- Submit contig or scaffold FASTA files that have:
 - ◊ Contigs lengths are ≥ 200 bp each, if they are not part of multi-component scaffolds in an AGP file
 - ◊ no foreign sequence contamination (GenBank reports back screen results, you can resubmit corrected files)
 - ◊ no Ns at the ends of sequences
- Use concise identifiers in the fasta files, e.g. contig00001, contig00002; avoid overtly long identifiers containing information on coverage or length; and avoid any punctuation except underscores.
- For sequences representing a chromosome or belonging to a plasmid, provide that information in the Assignment tab during a single-genome submission. For batch submissions, that information must be in the fasta defines of the sequences, as described in www.ncbi.nlm.nih.gov/genbank/genomesubmit/#batch_assignment.

More details are available at: www.ncbi.nlm.nih.gov/genbank/genomesubmit

Annotated genomes (including complex gap cases):

- Create .sqn files: convert runs of Ns that represent gaps in the FASTA files to assembly_gap features with the correct linkage evidence and include annotation.
- Tools and files needed are:
 - ◊ The command line program tbl2asn (v23.0 or higher) available from go.usa.gov/xpvtF
 - ◊ FASTA files (specifications are the same as above)
 - ◊ Template file with submitter information is at www.ncbi.nlm.nih.gov/WebSub/template.cgi
 - ◊ Optional Genome-Assembly-Data structured comment, which can also be provided in the genome submission form: submit.ncbi.nlm.nih.gov/structcomment/genomes/
 - ◊ Annotation in .tbl files. Instructions and tbl specifications are available online at go.usa.gov/xpvtI and www.ncbi.nlm.nih.gov/genbank/eukaryotic_genome_submission
- More information is at www.ncbi.nlm.nih.gov/genbank/genomesubmit/#sqn
- Annotation might also be provided using .gff files as the input with our beta version converter table2asn_gff. Instructions and .gff specifications are available online at www.ncbi.nlm.nih.gov/genbank/genomes_gff/

For eukaryotic annotation, brief requirements for common features are:

- each rRNA, tRNA, ncRNA needs a gene
- each CDS needs an mRNA and a gene
- each CDS/mRNA pair must share a unique protein_id and transcript_id in the .tbl file
- each gene must have a locus_tag that has the registered locus_tag prefix, and
- each locus_tag must be unique across the genome

It is recommended that the protein_id and transcript_id be based upon the gene's locus_tag. Genes that are alternatively spliced have a single gene feature that extends from the 5'-most to 3'-most end, and each mRNA has its own CDS even if multiple CDS features have the same translation. See an example at: go.usa.gov/xpvtz. CDS product names must conform to the International Protein Nomenclature Guidelines, go.usa.gov/xpvtJ, and go.usa.gov/xpvzq. See the “Annotation FYI” section of the www.ncbi.nlm.nih.gov/genbank/wgs_gapped page for the prohibitions about features crossing gaps.

Ns that represent gaps can be easily converted to assembly_gap features if all the following criteria are met:

- Each sequence represents a sequence that occurs biologically in the organism, such as a chromosome; the contigs were not simply concatenated
- No artificial sequences, such as linkers with multiple stop codons, are present
- The linkage evidence for each gap is the same
- Only a single length is used for gaps of completely unknown size (if such gaps are present)
- All runs of "ambiguous base Ns" are shorter than any run of Ns that represents a gap

(cont.)

Gapped and annotation (cont.)

- All the gaps are 'within scaffolds', not 'between scaffolds'

The "Gapped Format for Genome Submission" (www.ncbi.nlm.nih.gov/genbank/wgs_gapped) page provides the tbl2asn command line examples. Additional options for the gap-type and linkage evidence are also listed in this online documentation. For example, if runs of 10 or more N's are estimated gaps, and shorter runs of N's are just ambiguous bases, and all runs of exactly 100 N's are unknown gaps, and the linkage evidence is paired-ends, then generate the gap features in the .sqn file with the following tbl2asn command line syntax:

```
tbl2asn -p path_to_fsa_files -t template -M n -Z discrep -a r10u -l paired-ends
```

If the sequences are gapped (i.e., are scaffolds), but the simple cases for using tbl2asn to convert the Ns to assembly_gap features do not apply (e.g., there are different kinds of linkage evidence), then the assembly_gap features need to be included in the annotation .tbl file. They are set up like this, with the appropriate gap-type and linkage evidence:

```
100 201 assembly_gap
      gap_type      within scaffold
      linkage_evidence align-genus
```

Note that gap-type "between scaffolds" is allowed only when the sequences are chromosomes. The gapped sequences also must meet these criteria: 1) Each sequence represents a sequence that occurs biologically in the organism, such as a chromosome; the contigs were not simply concatenated; 2) No artificial sequences, such as linkers with multiple stop codons, are present. When the appropriate gap features are incorporated in the .tbl file, the tbl2asn assembly gap command line arguments are omitted, so the tbl2asn command line syntax looks like this:

```
tbl2asn -p path_to_fsa_files -t template -M n -Z discrep
```

Before submission, fix any Errors or FATALs in the .val or discrep files that are produced by referencing instructions given in the "Check the output of the validation and discrepancy report and fix problems" section of this web document:

www.ncbi.nlm.nih.gov/genbank/genomesubmit/#sqn

Annotated genomes with an AGP file:

The annotation must be at the scaffold or chromosome level, so submit:

- FASTA files of the contigs or scaffolds
- an AGP file to assemble the contigs into scaffolds and/or scaffolds into chromosomes, and
- .sqn files of the annotated scaffolds or chromosomes

The annotated scaffolds are created like the "Annotated genomes" examples described early in this document, either the simple cases of using tbl2asn command lines or the complex cases that require including the assembly_gap features in the .tbl files. As described in the WGS submission document (go.usa.gov/xpvz2), the AGP file should be made according to the AGP2.1 specifications as described here: go.usa.gov/xpvzj

The contig identifiers must match the component IDs in column 6 of the AGP files, and the scaffold/chromosome identifiers must match the object IDs in column 1 of the AGP files. AGP files can be validated here: go.usa.gov/xpvzX. For more extensive validation of AGP files locally, use the standalone command line program agp_validate as explained here: go.usa.gov/xpvzP. Use the "-help" flag to see available arguments and their appropriate input format.

Resolved pseudohaplotypes of a diploid/polyploid organism

The pseudohaplotypes of a diploid or polyploid assembly are submitted as individual genome submissions that share the same BioSample but have separate BioProjects which will be connected by an Umbrella BioProject. The submission files are created according to the relevant instructions above, and they can be submitted as separate 'genomes' via the new "Pseudohaplotypes of one or more diploid assemblies" option in the submission portal. During the submission you must define one of the pseudohaplotypes as the primary/principal assembly, even if they are the maternal/paternal haplotypes of the child in a trio. We recommend encoding the principal/alternate or maternal/paternal information in the Assembly Name as a visual clue for users of the data.

Creating or editing a genome submission in GenomeWorkbench

We recently released a new version of GenomeWorkbench that allows you to create or edit a genome submission. See this announcement: go.usa.gov/xpvzS. A video tutorial is available at: www.youtube.com/watch?v=BN9e4ma10kA

Contacts for Technical Assistance

Technical assistance for specific datatypes is available through the following email aliases:

sra@ncbi.nlm.nih.gov	for questions on SRA, vcf and BAM submission
biosamplehelp@ncbi.nlm.nih.gov	for questions on BioSample registration
bioprojecthelp@ncbi.nlm.nih.gov	for BioProject registration related question
genomes@ncbi.nlm.nih.gov	for questions about transcriptome and genome assemblies

For generic submission questions or questions over NCBI resources in general, please contact info@ncbi.nlm.nih.gov.