



Submitting Eukaryotic Genomes to GenBank

Preparing eukaryotic genomes for submission to GenBank

http://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 includes:

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

The highlighted steps will be described below. Additional details are available in the online documentation:

http://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 three ways to register a BioProject:

- singly by way of <https://submit.ncbi.nlm.nih.gov/subs/bioproject/>
- during a BioSample registration (see section below) <https://submit.ncbi.nlm.nih.gov/subs/biosample/>, or
- during a WGS genome submission <https://submit.ncbi.nlm.nih.gov/subs/wgs/>

The 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 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.

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 the 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
 - ◇ singly or in batch through the BioSample portal <https://submit.ncbi.nlm.nih.gov/subs/biosample/>
 - ◇ singly during a WGS genome submission <https://submit.ncbi.nlm.nih.gov/subs/wgs/>

The BioSample identifier, SAMN#####, is included with each 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
No Assembly	X	X	R	R				
Transcriptome	X	O ²	X	R	O ²			
Multiple related genomes (e.g., multiple strains)	X		X	R				
Genome assembly:								
Traditional contigs	X	R				X		O
Traditional contigs + annotation	X	R				X	X*	X
Gapped submission	X	R					X	
Gapped submission + annotation	X	R					X	O ¹

NOTE:

X = submit

R = 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

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. For details, see:

http://www.ncbi.nlm.nih.gov/books/NBK47537/#File_Format_Guide_B.BAM_Binary_Sequence.

Transcriptome options

- BAM file should be submitted to SRA.
- Assembled sequences should be in Sequin format (.sqn) with reference to SRA runs (SRR#) to TSA. More information is available at: <http://www.ncbi.nlm.nih.gov/genbank/tsaguide>.

Details on BAM format is at: <http://www.ncbi.nlm.nih.gov/books/NBK47537>

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

BAM file plus optional vcf file of SNPs should be submitted to SRA. For more information on .vcf format, see:

http://www.ncbi.nlm.nih.gov/SNP/docs/dbSNP_VCF_Submission.pdf

<http://www.1000genomes.org/node/101>

Genome assembly options

There are two basic types of assemblies, both can be submitted with or without annotation:

- Traditional, submitted as contigs plus an optional AGP file to assemble scaffolds or chromosomes
- Gapped, submitted as scaffolds with Ns that represent gaps converted to assembly_gap features

(cont.)

Genome assembly options (cont.)

Regardless of types, these genome assemblies, along with the genome assembly metadata, are to be submitted via the WGS submission portal (<https://submit.ncbi.nlm.nih.gov/subs/wgs/>). The metadata includes the assembly method, date or version the program was run, the approximate genome coverage, and the relevant sequencing technology used. 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 traditional without annotation:

- Submit contig FASTA files that have:
 - ◊ 10,000 sequences per file maximum
 - ◊ Contigs lengths are >200bp each
 - ◊ no foreign sequence contamination (the results of this screen will be reported back, and 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, include that information in the fasta define, using brackets such as: [chromosome=I] or [plasmid-name=unnamed1].

More details are available at: <http://www.ncbi.nlm.nih.gov/genbank/wgs.submit>

For gapped without annotation:

- Create .sqn files and convert runs of Ns that represent gaps in the FASTA files to assembly_gap features with the correct linkage evidence.
- Tools and files need are:
 - ◊ The command line program tbl2asn, v23.0 or higher, available from ftp://ftp.ncbi.nlm.nih.gov/toolbox/ncbi_tools/converters/by_program/tbl2asn/
 - ◊ FASTA files (specifications are the same as traditional without annotation)
 - ◊ Template file with submitter information is at <http://www.ncbi.nlm.nih.gov/WebSub/template.cgi>
 - ◊ Optional Genome-Assembly-Data structured comment, which can also be provided in the WGS submission form <https://submit.ncbi.nlm.nih.gov/structcomment/genomes/>
- More information are at
 - http://www.ncbi.nlm.nih.gov/genbank/wgs_gapped
 - <http://www.ncbi.nlm.nih.gov/genbank/wgs.submit>

Ns that represent gaps can be easily converted to assembly_gap features if all of 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
 - ◊ whether runs of 100 Ns represent gaps of unknown size
 - ◊ that no other Ns represent gaps of unknown size
 - ◊ all runs of "ambiguous base Ns" must be shorter than any run of Ns that represents a gap
 - ◊ all the gaps must be 'within scaffolds', not 'between scaffolds'

The "Gapped Format for Genome Submission" (http://www.ncbi.nlm.nih.gov/genbank/wgs_gapped) provides the tbl2asn command line example for each of the scenarios described above. 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 run:

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

Additional options for the gap-type and linkage evidence are also listed in this online documentation.

Gapped and annotation (including complex gap cases):

This is like the previous situation. The exception is that the annotation is provided in .tbl files. Instructions and tbl specifications are available online:

http://www.ncbi.nlm.nih.gov/genbank/eukaryotic_genome_submission_annotation
http://www.ncbi.nlm.nih.gov/genbank/eukaryotic_genome_submission

(cont.)

Gapped and annotation (cont.)

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;
- 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. Alternatively spliced genes have a single gene feature that extends from the 5'-most to 3'-most feature, and each mRNA has its own CDS even if multiple CDS have the same translation. See an example at: http://www.ncbi.nlm.nih.gov/genbank/eukaryotic_genome_submission_annotation#Alternativelysplicedgenes

CDS product names must conform to SwissProt guidelines, as per <http://www.uniprot.org/docs/genameprot> and http://www.ncbi.nlm.nih.gov/genbank/eukaryotic_genome_submission_annotation#CDS. See the "Annotation FYI" section of the http://www.ncbi.nlm.nih.gov/genbank/wgs_gapped page for the prohibitions about features crossing gaps. 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. They are not allowed in scaffold sequences. 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. Run tbl2asn:

```
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 this document: <http://www.ncbi.nlm.nih.gov/genbank/wgs.submit#Fix>

Traditional and annotation:

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

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

The annotated scaffolds are created like the Gapped 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 by WGS submission document (<http://www.ncbi.nlm.nih.gov/genbank/wgs.submit#agp>), the AGP file should be made according to the AGP2.0 specifications: http://www.ncbi.nlm.nih.gov/projects/genome/assembly/agp/AGP_Specification.shtml

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. The format of AGP files can be validated on this NCBI web page: http://www.ncbi.nlm.nih.gov/projects/genome/assembly/agp/agp_validate.cgi

For more extensive validation of AGP files locally, a standalone command line program agp_validate is available by anonymous FTP, as explained here: http://www.ncbi.nlm.nih.gov/projects/genome/assembly/agp/AGP_Validation.shtml Run the program with the -help to obtain details of the available arguments and their appropriate input format.

Technical assistance

Technical assistance is available through the following email aliases:

sra@ncbi.nlm.nih.gov
biosamplehelp@ncbi.nlm.nih.gov
genomeprj@ncbi.nlm.nih.gov
genomes@ncbi.nlm.nih.gov
info@ncbi.nlm.nih.gov

for questions on SRA, vcf and BAM submission
 for questions on BioSample registration
 for BioProject registration related question
 for questions about transcriptome and genome assemblies
 for other general questions related to submission