Unit 4. Submitting sequences to databases

Every database has unique sequence submission portals.

- NCBI: 3 ways of submission: submission portal, Sequin, BankIT, tbl2an (ftp system- outdated)
- EBI-EMBL European Nucleotide Archives: WebIN
- DDBJ: Nucleotide Sequence Submission System (NSSS)

Type of DNA sequence accepted:

- Raw data reads/ Simple sequences
- Assembly reads
- Large scale metadata
- High Throughput Whole genome sequence (HTWGS)
- Genome Annotations
- EST

  All data is exchanged daily between all three as part of International Nucleotide Sequence Database Collaboration

  We will focus on NCBI data submission system:

1. Submission Portal: different types of sequence can be submitted
   a. **Genbank Submission Protal**: Specific Prokaryotic and Eukaryotic Gene Submission portal: This submission tool is for the specific data types indicated below:
      i. Prokaryotic rRNA/IGS
      ii. Eukaryotic nuclear rRNA/ITS
      iii. Eukaryotic organelle rRNA
      iv. Metazoan (multicellular animal) COX1
      v. Influenza, Norovirus or Dengue virus

    b. Sequence Retrieval Archives (SRA): The archive accepts data from all branches of life as well as metagenomic and environmental surveys.
        i. GEO: Gene Expression Omnibus : Submit RNA-seq, ChIP-seq, and other types of gene expression and epigenomics datasets
        ii. BioProject & BioSample: sequencing information from a single experiment/ consortia
    c. Transcriptome Shotgun Assembly: computationally assembled transcribed RNA sequences from next generation sequencing technologies Submit unassembled, high throughput sequencing reads
2. **BankIT**: This submission option is for genomic DNA (e.g. protein-coding gene, regulatory element), transcripts (e.g. mRNA, ncRNA), or small genomes (organelle, plasmid, and phage and other viral) from any organism. Artificial sequences (cloning/expression vector) as well as annotated or assembled third party sequences can also be submitted here. Customized BankIT allows submission of Primer sequences, DNA barcode etc
3. ***Tbl2asn** is a command-line program that automates the creation of sequence records for submission to GenBank. It uses many of the same functions as Sequin but is driven generally by data files. Tbl2asn generates .sqn files for submission to GenBank. Additional manual editing is not required before submission. Tbl2asn is available by anonymous File Transfer Protocol (FTP). Useful for WGS sequences.*
4. ***Sequin:** stand-alone software tool developed by the NCBI for submitting and updating entries to the GenBank sequence database which is **phased out**. It is capable of handling simple submissions that contain a single short mRNA sequence, and complex submissions containing long sequences, multiple annotations, gapped sequences, or phylogenetic and population studies. A single Sequin file should contain less than 10,000 sequences for maximum performance. Larger submissions should be made with tbl2asn*

GenBank Sequence Submission Policy

1. the GenBank database is intended for new sequence data that is determined by and annotated by the submitter
2. sequences built or derived from other GenBank primary data intended for the Third Party Annotation (TPA) database may be submitted through BankIt
3. the following types of submissions are **NOT** acceptable:

a. sequences less than 200 nucleotides long, unless they represent complete exons, non-coding RNAs (ncRNAs), microsatellites or ancient DNA
b. non-contiguous sequences that have been artificially joined; for example, multiple exons without their intervening introns or without a 'gap' of internal NNNs representing any missing sequence
c. protein-only sequences
d. single sequences that are a mix of molecule types, such as mix of genomic and mRNA sequence data
e. sequences without a physical counterpart (consensus sequences)

How to: Submit sequence data to NCBI

Starting with...	NOTES	SUBMISSION TOOLS & HELP DOCUMENTS
	Simple Sequence Submissions	
Single nucleotide sequence *or* Several nucleotide sequences for *different* genes or loci	Contiguous bases of cDNA or genomic DNA, but should not be complete genomes. Complete genomes should be submitted via the appropriate protocol indicated below. Records with simple annotation may be submitted by BankIt or Sequin, while records with complicated annotation may be more easily submitted via Sequin.	BankIt or Sequin
Group of nucleotide sequences for the *same* gene or locus	Includes: • population studies (sequences for a single organism) • phylogenetic studies (sequences for multiple organisms) • environmental samples (such as cultured or uncultured bacteria or metagenomic samples)	BankIt or Sequin
Batches of Sequences	Includes: • Expressed Sequence Tags (ESTs) • Genome Survey Sequences (GSSs)	Batch submit guidance page
	Genomic Assembly Submissions	

Starting with...	NOTES	SUBMISSION TOOLS & HELP DOCUMENTS
Small complete genomes	Includes chloroplasts, mitochondria, plasmids, phages, and viruses *(Locus_tag or BioProject registration is NOT required.)*	Sequin
Large complete genomes	Includes paired chromosome and plasmids, as well as bacterial or eukaryotic chromosomes Questions regarding a specific submission that are not answered in the documented instructions can be sent to genomes@ncbi.nlm.nih.gov .	Prokaryotic Genomes submission Eukaryotic Genomes submission
Incomplete genomes	These can be whole genome shotgun (WGS) sequences. WGS submissions should be prepared using the tbl2asn or Sequin tools. For assistance contact genomes@ncbi.nlm.nih.gov .	Assembly submission information & Examples WGS submissions
High Throughput Genome Sequences (HTGSs)	The clones (e.g. BACs) of large-scale clone-based genome sequencing projects that are to be released quickly into GenBank can be submitted via the HTGS system. Sequences that are to be kept confidential or are few in number should be submitted as described above for *Single nucleotide sequences*. HTGS submissions require prior communication with NCBI staff, so please read about the HTGS submission process for details.	HTGS submissions
	Other Submission Types	

Starting with...	NOTES	SUBMISSION TOOLS & HELP DOCUMENTS
Barcode of Life sequences	Mitochondrial cytochrome oxidase I sequences that are part of the Barcode of Life initiative can be submitted using a customized Bankit.	Barcode submit page
New sequence annotation for a *non-RefSeq* record submitted to GenBank by someone else	Third Party Annotation (TPA) submissions can be created for annotation of existing GenBank records when the submitter has *experimental or inferential evidence* that will be published in a peer-reviewed biological journal. Please read about the TPA database and its submissions policies before submission.	TPA information TPA FAQs
Computationally assembled transcript sequences	These records, based on those that have already been submitted to dbEST, SRA or the Trace Archive, may be candidates for submission to the Transcriptome Shotgun Assembly (TSA) repository.	TSA information
Variations or Polymorphisms[1]	Single nucleotide polymorphisms as well as short insertions and deletions (<50bp) should be submitted to dbSNP, while large structural variations and copy number variation (CNV) data should be submitted to dbVar. Please note that human variations/polymorphisms with *clinical relevance* should be submitted to a specialized Human Variation Batch submission process using HGVS nomenclature.	Variation Submission Portal
Primers, siRNAs, or probes	Primer or nucleotide-based probe sequences should be submitted to the Probe Database.	Probe submit page

Starting with...	NOTES	SUBMISSION TOOLS & HELP DOCUMENTS
High throughput sequences	The Sequence Read Archive (SRA) accepts reads from high throughput sequencing instruments. Some submissions include sets of SRA reads as part of a comprehensive package. *For the specific datasets described below,* please initiate submissions with the appropriate archive: • *Human sequence or metagenome sequence data* derived from clinical isolates or from sources with privacy concerns should be submitted to dbGaP. • *Functional genomics studies* that examine gene expression, regulation or epigenomics (using methods such as RNA-Seq, miRNA-Seq, ChIP-Seq or methyl-Seq) should be submitted to GEO. • *Transcript survey sequence assemblies* should go to the Transcriptome Shotgun Assembly (TSA) archive. • *Non-human and environmental metagenomics data* should go to the Metagenome archive. • *Whole genome sequence assemblies* should be submitted to WGS. • *Capillary traces* should be deposited in the Trace Archive. • *Sequences from the Barcode of Life project* should be submitted to Barcode. Curators of these resources will assist submitters in sending the data to SRA during the submission process.	For data types not mentioned to the left, submit directly to SRA: SRA submit page SRA submission guidance

What is needed to submit a sequence through BankIt?

(See Submission Requirements page for more detail)

- registration through the MyNCBI Login System (register on Sign in page through link above)
- sequence data can be either cut-and-pasted as text or uploaded as file (multiple sequences must be in a FASTA format)
- date for public release (immediate or at a specified future date)
- basic information (authors and a working title) for a corresponding reference paper
- name(s) of the organism(s) from which the sequence data were isolated and any other related descriptive data
- sequence features (for example: CDS, gene, rRNA, tRNA, with nucleotide intervals and product names)

GenBank Sequence Submission Policy

1. the GenBank database is intended for new sequence data that is determined by and annotated by the submitter
2. sequences built or derived from other GenBank primary data intended for the Third Party Annotation (TPA) database may be submitted through BankIt
3. the following types of submissions are **NOT** acceptable:
    - sequences less than 200 nucleotides long, unless they represent complete exons, non-coding RNAs (ncRNAs), microsatellites or ancient DNA
    - non-contiguous sequences that have been artificially joined; for example, multiple exons without their intervening introns or without a 'gap' of internal NNNs representing any missing sequence
    - primer-only sequences
    - protein-only sequences
    - single sequences that are a mix of molecule types, such as mix of genomic and mRNA sequence data
    - sequences without a physical counterpart (consensus sequences)

Updates to Sequences that are already submitted

BankIt does not have an update option. Please see options for updating a submitted or existing GenBank record.

Sequin--A DNA Sequence Submission Tool

Sequin home

FAQs

Help Documentation

Wizard Documentation

Download Sequin

Download Instructions

Update Instructions

GenBank Submissions Handbook

NCBI is phasing out support of the Sequin submission tool. Please submit your data usingBankIt, Submission Portal or tbl2asn. See Submission Tools for details on the appropriate tool

What Is Sequin

Sequin is a stand-alone software tool developed by the NCBI for submitting and updating entries to the GenBank sequence database. It is capable of handling simple submissions that contain a single short mRNA sequence, and complex submissions containing long sequences, multiple annotations, gapped sequences, or phylogenetic and population studies. A single Sequin file should contain less than 10,000 sequences for maximum performance. Larger submissions should be made with tbl2asn .

How to Get Sequin

Sequin 15.50 is currently available from the NCBI. Sequin runs on Macintosh, PC/Windows, and UNIX computers. Instructions for downloading and installing the program are provided. The program itself, along with its on-line help documentation, is available by anonymous FTP.

Sequin Help Documentation

A window containing the Sequin Help Documentation is opened when the Sequin program is launched. The contents of this scrolling window change as you move within the Sequin program, presenting you with help documentation appropriate for the section of Sequin you are presently visiting. This documentation is also available in a World Wide Web format. Detailed instructions for the various Sequin Wizards are also available.

Annotation Using A Table

A five-column, tab-delimited table of feature locations and qualifiers can be used to import annotation into an existing Sequin submission. This is the same table format that must be used to annotate features when creating a submission using tbl2asn.

Network-Aware Sequin

Sequin can be used in one of two modes, stand-alone or network-aware. In the network-aware mode, the program can exchange data between any computer connected to the Internet and the NCBI.

SequinMacroSend

The SequinMacroSend tool allows the submission of very large Sequin files directly. Those files that may be truncated during mailing with conventional maile rs, including large population sets or complete plasmids or small genomes, can be sent using this method.

Tbl2asn

The tbl2asn command line program is available via ftp and is designed as an alternative to the Sequin program for generating large single submissions (complete genomes) containing a great deal of annotation. It can also be used to generate a batch submission containing thousands of individual sequences. More detailed instructions about using this function are provided. When submitting a complete bacterial genome, please review the genome guidelines.

Sequin Revision History

The Sequin revision history contains a reverse chronological list of changes made to the Sequin program.

Prokaryotic and Eukaryotic Genomes Submission Guide

Both WGS and non-WGS genomes, including gapless complete bacterial chromosomes, can be submitted via the Submission Portal. You will be asked to choose whether the genome being submitted is considered WGS or not. The differences for GenBank purposes are:

non-WGS

- Each chromosome is in a single sequence and there are no extra sequences
- Each sequence in the genome must be assigned to a chromosome or plasmid or organelle
- Plasmids and organelles can still be in multiple pieces.

WGS

- One or more chromosomes are in multiple pieces and/or some sequences are not assembled into chromosomes

In both cases

- There can still be gaps within the sequences; you will supply that information in the submission
- Plasmids and organelles can still be in multiple pieces.
- Internal sequences must be arranged in the correct order and orientation.
- Sequences concatenated in unknown order are not allowed.

Table of Contents

Submit a single genome

This is the simplest submission route because you just fill in a web form in the Submission Portal and upload fasta (or sqn) files of the genome sequences. You will need to:

- Provide the BioProject created for this research effort, e.g., during submission of the reads to SRA OR register a new BioProject during the genome submission.
- Provide the BioSample created during submission of the reads to SRA OR register a BioSample during the genome submission
- Assert whether this is a WGS or non-wgs genome assembly
- Upload fasta sequences of the genome (or .sqn, file if the genome is annotated)
- Upload optional AGP file(s) to assemble scaffolds (unplaced or unlocalized) and/or chromosomes from the submitted sequences. This is for WGS only. Remember that you can submit the gapped scaffolds themselves instead of submitting contigs plus an AGP file
- Provide information in response to prompts during the genome submission (see the common metada section):
    - Genome Assembly Data and other information about this genome assembly
    - Gap Information (What the Ns represent)
    - Chromosome and plasmid assignments. Every sequence in a non-wgs genome must have a chromosome or plasmid assignment and every chromosome must be submitted as a single sequence.
    - Authors and a title (for fasta submissions)
    - Release date (immediately after processing OR a specific date. Release will be on that date or upon publication, whichever is first)
    - Optional request for annotation of prokaryotic genomes by PGAP

Submit a batch of genomes

This submission route allows you to submit as many as 400 WGS or non-wgs genomes in a single batch submission. In this route you choose Batch/multiple in the Genome Submission Portal , fill in the web form, upload a Genome Info file with genome metadata, and upload or preload fasta files (or sqn files if there is annotation) of the genome sequences. All the genomes within a batch must:

- Be part of the same BioProject
- Be either WGS or non-wgs, not a mix of both types
- Have the same (initial) release date
- Have the same gap/Ns information
- Contain either fasta files or ASN (.sqn) files, not a mix of file types. We recommend submitting fasta files unless the submission needs to include annotation or the Genome-Assembly-Data structured comment
- Have a single file for each genome, including any plasmid or organelle sequences
- Have a separate file for each genome, not all the genomes together
- Request PGAP annotation or not (only relevant for prokaryotic genomes)
- Be just a single layer (= no AGP file(s))

You will need to:

- Provide the BioProject created for this research effort, e.g., during submission of the reads to SRA OR register a new BioProject during the genome submission.
- Provide the BioSamples that were preregistered, eg during submission of the reads to SRA OR register BioSamples during the genome submission
- Include assignment (ie, chromosome, plasmid or organelle) information about the sequence in the fasta files (see the Additional requirements for batch submissions section)
- Upload or preload fasta sequences (or sqn files for annotated genomes) of the genomes. Each genome is in a separate single file, uniquely named, but the files can be archived together
  - An option for batch submission is to preload the files of genome assemblies before beginning the submission, rather than uploading them in the browser during the submission. You can preload using Aspera, the FTP protocol or Filezilla. Detailed instructions for using the preload option for genome submissions are at How to preload files.
- Upload a Genome Info table with information specific to each genome

- Provide this information in response to prompts on the web pages during the genome submission (see the common metadatasection):
  - Gap Information (What the Ns represent)
  - Authors and a title (for fasta submissions)
  - Release date (immediately after processing OR a specific date. Release will be on that date or upon publication, whichever is first)
  - Optional request for annotation of prokaryotic genomes by PGAP

Events

1. Only if you will be submitting a genome with annotation and you have not yet registered a BioProject and BioSample for this genome, then you will register the genome sequencing project with the BioProject and BioSample databases so that a locus_tag prefix will be assigned to the BioProject:BioSample pair. If you have already registered a BioProject and BioSample for this genome, eg when submitting the reads to SRA, then a locus_tag prefix should have already been assigned. A file of the locus_tag prefix(es) for the BioSamples within a BioProject is linked to the BioProject submission. Write to genomes@ncbi.nlm.nih.gov if you did not receive a locus_tag prefix. Do not register a duplicate BioProject or BioSample for the same genome. Provide these preregistered BioProject and BioSample accessions in the genome submission. Remember that annotation is optional for genome submissions. If you are submitting a genome without annotation, even if you will be requesting PGAP annotation, then you'll create the BioSample (and BioProject, if necessary) during the genome submission. Genomes sequenced as part of the same research effort can belong to a single BioProject, so it's common to create a BioProject during the submission of one genome and then include that BioProject during the submission of additional genomes.

2. Make the genome assembly data files.

   - Unannotated genomes just need fasta files
   - Annotated genomes need to make .sqn file submissions by running the command line program tbl2asn, and then fixing Errors and Fatals that are indicated in the .val and discrep files. Failure to do this will cause serious delays in processing.

3. If you have higher-level assembly information, scaffolds and/or chromosomes, then generate an AGP file to build those objects from the wgs-contigs.

4. If you are submitting a batch of genomes (maximum of 400 per batch), then create a Genome Info file.
5. Submit via the new Genomes (was WGS) in the Submission Portal
6. What happens after submission

Submission Files

Fasta files

Put the sequences file into fasta format

IMPORTANT: Additional requirements for batch submissions

.sqn files

These are generally required only when the submitter wants to include annotation. Annotation is optional for GenBank genome submissions.

see details

AGP file (optional)

AGP files provide the ordering and orientation information to construct scaffolds from contigs, or to construct chromosomes from scaffolds and/or contigs. However, remember that we do accept the gapped scaffolds themselves as the basic sequences of the genome. If you choose to submit a multi-layer submission with and AGP file, then know that the AGP file defines these genome assemblies, so be sure to include all wgs-contigs that are considered to be part of the genome in the AGP file. However, if the sequences in the fasta (or .sqn) files are already the scaffolds or chromosomes, then do not make an AGP file.

see details

Genome Info table

The Genome Info table is required for batch submissions and is used to provide the Genome Assembly Data of each. Download the Genome Info file template. The instructions are on the first tab of this file and the template is on the second tab. Complete the second tab (Genome_Data), then save the worksheet as a Text (Tab-delimited) file -- (use 'File, Save as, Save as type: Text (Tab-delimited)').

see details

Metadata required for all genome submissions

BioProject

The BioProject contains the description of the research effort, relevant grant(s), and has links to the public data for the proejct. Each genome must belong to a BioProject, and genomes sequenced as part of the same research effort can belong to a single BioProject. Use the same BioProject for the sequence reads and genome assembly made from those reads; do not create duplicate BioProjects. If a new BioProject is necessary for unannotated (or PGAP-annotated) genomes, then registering during the genome submission process is simplest. However, genomes submitted with annotation will need to be pre-registered so that a locus_tag prefix can be assigned to the BioProject/BioSample pair and used to identify each gene within that genome uniquely. A file of the locus_tag prefix(es) for the BioSamples within a BioProject is linked to the BioProject submission. Write to genomes@ncbi.nlm.nih.gov if you did not receive a locus_tag prefix after preregistering a BioSample for your BioProject.

BioSample

The BioSample contains the source information of the sample that was sequenced. Use the same BioSample for the sequence reads and genome assembly made from those reads; do not create duplicate BioSamples. Registering a new BioSample can be done during the genome submission process for unannotated (or PGAP-annotated) genomes; however, genomes submitted with annotation will need to be pre-registered to get a locus_tag prefix. Include the registered BioProject when you register the BioSample so that a locus_tag prefix is assigned to the pair. You'll find the locus_tag assignment(s) in a file linked to the BioProject submission.

Genome Assembly Data and other information about a genome assembly

- **Assembly method** : Name of the assembly algorithm(s)
- **Assembly method version or date** : version of the algorithm or date it was run
- **Genome coverage** : The estimated base coverage across the genome, eg 12x.
- **Sequencing technology** : sequencing platform(s) used
- **Assembly date** : Optional. Year, month or day the assembly was made. Date formats: YYYY-MM-DD; YYYY-MM; YYYY
- **Assembly name** : Optional and not usually relevant for prokaryotes. This is a short name suitable for display that does not include the organism name eg, LoxAfr_3.0 for a Loxodonta africana assembly, version 3.0
- **Full or Partial Genome in the sample** : the answer is nearly always "yes, Full". Choose "no, partial" only if a subset of the sample was deliberately selected, eg just exomes or a single chromosome of a eukaryote or only the non-repetitive regions of the genome
- **Reference genome** : If this is NOT a de novo assembly, you will need to provide the accession.version and/or the assembly name of the genome assembly that was used as the reference guide for this assembly
- **Update** : accession of the genome being updated, when appropriate
- **bacteria_available_from** : Optional. For prokaryotes provide a name and physical address (not email) of the lab or PI, or a culture collection identifier where scientists could obtain this bacterial culture

Gap Information: What the Ns represent

- The minimum number of consecutive Ns that represents a gap (must be 10 or less). Be aware that the 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.
- The number of Ns that represents a gap of completely unknown length (usually 0; sometimes 100 or another value)
- The evidence used to assert that the sequence on either side of the gap is linked (usually paired-ends)
- This information is collected in the submission form for individual and batch submissions. Default answers are those that have been most commonly submitted. Be sure to select the correct answer when the defaults are incorrect for the genome(s) being submitted.

Chromosome and plasmid assignments

Plasmid and chromosome names rules

Submit the genomes to the Genome Submission Portal

All files must be submitted via the Submission Portal. Choose "Single" or "Batch/multiple" genomes. Answer the questions and upload the necessary files Review the summary page and click the "Submit" button. The submission will be given a 'SUB' temporary identifier which you can use in correspondence before an accession number is assigned to the genome submission.

What happens next

Once we receive your genome submission, a few automated validations are run and a member of our staff conducts an initial review. If no significant issues are found, the genome will be assigned an accession number.

If there are problems

Submission statuses in the submission portal

If you elected to hold your genome until a particular date (or publication, whichever is first), we ask that you provide us with the expected publication date and also notify us in a timely manner of the upcoming publication and the relevant citation details. This will allow us to coordinate the release of your genome with the appearance of the paper. Please provide at least two weeks' notice of any upcoming publication.

NOTE: As of January 2017, genomes will be released on their release date without additional communication, as is the normal GenBank policy. Be sure to request an extension of the release date if the genome is not yet published and you wish to continue to keep it confidential.

Requesting PGAP annotation of prokaryotic genomes

Requests for annotation by the Prokaryotic Genomes Annotation Pipeline is a step during submission of the genome to GenBank. Prepare a regular GenBank genome submission and request PGAP annotation during the submission process by clicking on the box

"Annotate this prokaryotic genome in the NCBI Prokaryotic Annotation Pipeline before being released". The annotated genome will be posted back to the Submission Portal for your review. You may edit the file and resubmit that to GenBank; however, this is not required and is generally not recommended, as it will slow processing and may introduce problems that you would need to fix.

What is tbl2asn?

Tbl2asn is a command-line program that automates the creation of sequence records for submission to GenBank. It uses many of the same functions as Sequin but is driven generally by data files. Tbl2asn generates .sqn files for submission to GenBank. Additional manual editing is not required before submission.

Tbl2asn is available by anonymous FTP. Copy the right version for your platform, then uncompress the file, rename it to "tbl2asn", and set the permissions, as necessary for the platform.

Additional details are provided in the GenBank Submission Handbook

6 types of input data files

REQUIRED

1. Template file containing a text ASN.1 Submit-block object (suffix .sbt).
2. Nucleotide sequence data in FASTA format (suffix .fsa).
3. Feature Table (suffix .tbl). [Required only if including annotation]

OPTIONAL

4. Quality Scores (suffix .qvl.)
5. Protein sequence (suffix .pep). (These are rarely needed.)
6. Source Table (suffix .src.)

Generating the .sqn file for submission

- The minimum requirements to generate a Sequin file using tbl2asn are one .sbt file and one or more .fsa files.
- The files are placed in a source directory and a series of command-line arguments are used to generate the .sqn files.

- Tbl2asn will generate a .sqn for every .fsa file in the directory, plus any of the corresponding optional files that may be present. The other files must have the same file name prefix as their corresponding .fsa. (for example helicase.fsa and helicase.tbl).

Submitting HTG Sequences

If you are new to HTG submissions please start here.

The HTG division contains unfinished DNA sequences generated by the high-throughput sequencing centers using **traditional clone-based Sanger sequencing**.

Draft genomes sequenced using non-clone based whole genome shotgun sequencing are not appropriate for HTG, these should be submitted as a WGS submission as described at www.ncbi.nlm.nih.gov/Genbank/wgs.html. NextGen sequences and should not be submitted to HTG instead these should be submitted to the Sequence Read Archive. For general submission information, please start here.

The submission process for HTGs is quite different from that for other direct submissions. The goal of the process is to make new and updated sequences available to the public in a timely fashion. Thus, NCBI will perform only very basic validation checks of HTGs, and submitters must check their records carefully before submission. Furthermore, because sequences will be released to the public as soon as processing is finished, it is presently not the standard procedure to indicate a "hold until published" (HUP) date on which they should be released. If a HUP date is necessary, the submitter should please contact the database staff about submitting through an alternate route.

Sequencing centers that will be submitting HTGs to NCBI should email htgs-admin@ncbi.nlm.nih.gov to establish an FTP account. Prepared records should be transferred to this site, where they will be retrieved daily by the NCBI staff. These records should not be emailed to the NCBI. Submitted HTG sequences must be written in ASN.1 format. **PLEASE NOTE** most nucleotide sequence submissions to NCBI do NOT use an FTP account. The instructions here are specific to high-throughput genomic sequences (HTG).

Submission Tools

There are currently two ways to create HTG records:

1. The Sequin program **Sequin will soon be retired and if you are using Sequin for HTG submissions, you should switch to tbl2asn.** Sequin contains a setting that allows genome centers to prepare HTG submissions. Sequin reads in a FASTA sequence file (or an Ace Contig file with Phrap sequence quality values) and a Sequin submission template file (to get contact and citation information). Users then enter additional information into a Sequin form, the same information that they would enter at the command line in fa2htgs (see below). Sequin generates the ASN.1 file for submission.

2. The tbl2asn tool tbl2asn is a command-line program that has replaced the deprecated program fa2htgs. tbl2asn reads in a FASTA sequence file (or an Ace Contig file with Phrap sequence quality values), a Sequin submission template file (to get contact and citation information), and a series of command-line arguments (to get additional information). tbl2asn then generates the ASN.1 file for submission. tbl2asn can be incorporated into scripts to facilitate expedient processing of records.