



Submitting Data to NCBI

Depositing different types of data for public access through NCBI

<http://www.ncbi.nlm.nih.gov/guide/howto/submit-data/>

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

Data submission in general

NCBI databases accept many types of data from biomedical research projects. These range from biological sequences, microarray data, chemical substances and their biological activities, to complete scientific manuscripts. Since each of these diverse data types has unique requirements, NCBI has specialized submission processes for each one. This handout provides a quick guide to NCBI data submission.



Data types accepted

Data accepted for deposition into various NCBI databases can be divided into two main categories: sequence data and non-sequence data.

- Sequence data refer to **a)** traditional sequences, **b)** assembled genomes, **c)** sequences with third party annotation, **c)** sequence variation data including those with asserted clinical significance, and **d)** other types of sequences.
- Non-sequence data come from **a)** microarrays, **b)** human clinical studies, **c)** data for chemical substances and their biological assays, **d)** detailed descriptions of genetic tests, and **e)** descriptive data (known as metadata) for biologic research projects (such as genome sequencing projects).

Submission of traditional sequences

Traditional sequences are individual sequences or small batches of sequences experimentally determined by the submitters that represent specific loci from one or more organisms. The records should be annotated to provide information on features present in the sequences, e.g. CDS, repeats, functional domains, etc. For batch sequence data obtained by a single read from a collection of sequences from a library, annotation is not needed.

Starting Sequence Type	Detailed description and links
Single nucleotide sequence OR Nucleotide sequences for <i>different</i> genes or loci	The sequences should be contiguous bases of cDNA or genomic DNA, presented in the IUPAC code and with feature annotation. Sequences may be prepared and submitted by the BankIt online tool or using the standalone tool Sequin and submitted by email or file upload through SequinMacroSend . BankIt www.ncbi.nlm.nih.gov/WebSub/?tool=genbank Sequin www.ncbi.nlm.nih.gov/Sequin/index.html SequinMacroSend www.ncbi.nlm.nih.gov/LargeDirSubs/dir_submit.cgi
A set of nucleotide sequences for the <i>same</i> gene or locus	These include: population studies (sequences for a single species from multiple samples or individuals), phylogenetic studies (sequences from multiple species), and environmental samples (such as cultured or uncultured bacteria or metagenomic samples). Sequence sets are most easily prepared for submission using Sequin and submitted through email to gb-sub@ncbi.nlm.nih.gov or file upload through SequinMacroSend . Sequin www.ncbi.nlm.nih.gov/Sequin/index.html SequinMacroSend www.ncbi.nlm.nih.gov/LargeDirSubs/dir_submit.cgi
Barcode of Life sequences	Mitochondrial cytochrome oxidase I sequences that are part of the Barcode of Life initiative can be submitted using the online customized BarCode Submission Tool. BarCode Submission Tool www.ncbi.nlm.nih.gov/WebSub/?tool=barcode

Submission of batch sequences

These are sequences obtained from single-read high throughput sequencing projects. They include cDNA from Expressed Sequence Tags (EST) or clone ends from genomic libraries (Genome Survey Sequences or GSSs).

Starting Sequence Type	Detailed description and links
Batches of Sequences	No specialized tool exists to prepare this group of sequences. These sequences should be converted to a specific format and submitted through email or ftp upload as described in the relevant submission pages: EST www.ncbi.nlm.nih.gov/dbEST/how_to_submit.html GSS www.ncbi.nlm.nih.gov/dbGSS/how_to_submit.html

Submission of assembled genomes

Submission of genome sequences, in various stages of assembly/annotation, requires special submission procedures because of their complexity and large size, especially for large genomes from higher eukaryotes.

Starting Sequence Type	Detailed description and links
Small complete genomes	These include chloroplasts, mitochondria, plasmids, phage and viral sequences that do not require Locus_tag or Genome Project registration. They should be submitted as traditional sequences using Sequin standalone submission tool. Sequin www.ncbi.nlm.nih.gov/Sequin/index.html
Large complete genomes	This subcategory contains chromosome and plasmids from bacteria, archaea, or eukaryotic organisms. Prokaryotic and eukaryotic genomes are processed differently. Prokaryotic www.ncbi.nlm.nih.gov/genbank/genomesubmit.html Eukaryotic www.ncbi.nlm.nih.gov/genbank/eukaryotic_genome_submission.html Questions on specific submissions not addressed by help documentation should be sent to: genomes@ncbi.nlm.nih.gov
Incomplete genomes	These sequences are often contigs and/or scaffolds assembled from whole genome shotgun (WGS) sequences, and should be prepared using the standalone tool tbl2asn, which can also be used to batch process annotated features into Sequin files. Introduction www.ncbi.nlm.nih.gov/projects/genome/assembly/submission/ WGS www.ncbi.nlm.nih.gov/genbank/wgs.html tbl2asn www.ncbi.nlm.nih.gov/genbank/tbl2asn2.html Additional help is available by writing to: genomes@ncbi.nlm.nih.gov
High Throughput Genome Sequences (HTGSs)	These sequences are from large-scale clone-based (bacterial artificial chromosomes, BACs) genome sequencing projects, processed and released quickly into GenBank upon submission. These sequences should be submitted through the HTGS system and require prior communication with NCBI staff. HTGS www.ncbi.nlm.nih.gov/HTGS/subinfo.html

Submission of next generation sequence data

Next generation sequence data should be submitted to the Sequence Read Archive (SRA) database.

Starting Sequence Type	Detailed description and links
Next generation sequence data	Detailed instruction for SRA submission is online at: www.ncbi.nlm.nih.gov/books/NBK47529/ Questions on SRA submission should be addressed to: sra@ncbi.nlm.nih.gov

Submission of third party annotation sequences

Many of the genomic sequences in the public sequence database lack feature annotation. This limits the usefulness of these sequences. To help make these sequences become more useful by increasing their information content, NCBI accepts sequences re-annotated by third parties. These submissions must be based on experimental evidence and/or other comprehensive analysis that has passed formal peer-review in the form of scientific publication.

Starting Sequence Type	Detailed description and links
New sequence annotation for a <i>non-RefSeq</i> record submitted to GenBank by someone else	The Third Party Annotation (TPA) database accepts annotation of existing GenBank records when the submitter has experimental or inferential evidence that will be published in a peer-reviewed biological journal. More information is available in the following online documents. TPA submission policy www.ncbi.nlm.nih.gov/Genbank/TPA.html TPA FAQ www.ncbi.nlm.nih.gov/genbank/tpafaq.html

Submission of sequence variation data

Sequence variation data can be classified into two large groups according to the length of the variations. Small-scale variations such as single nucleotide polymorphisms, insertions, deletions and repeats less than 50 bases in length are accepted by dbSNP. Large genomic sequence variations not covered by dbSNP are processed through dbVar.

Starting Sequence Type	Detailed description and links
Single Nucleotide Variations or Short Nucleotide Polymorphisms	<p>Single nucleotide polymorphisms as well as short insertions and deletions (50bp or shorter) may be submitted to dbSNP.</p> <p>The NCBI dbSNP submission processing system now accepts Variant Call Format (VCF). The flexible VCF format can be customized to accommodate requirements of the data being submitted and can be used to submit common variations, their associated genotypes and annotation as well as variation data collected over multiple populations in various scales.</p> <p>A major benefit of submission in VCF format is that it allows the submission of variations with “asserted positions” [see NOTE] on a defined genomic assembly rather than flanking sequences as a means to locate variations. Using asserted positions allows for much greater accuracy in variation mapping onto the genome assembly, its annotated RNA and protein products, as well as subsequent remapping to future updates of the assembly than do flanking sequences.</p> <p>We strongly encourage our submitters to adopt the VCF submission format. General information about the VCF format and detailed instructions for creating a VCF formatted submission are available at: www.ncbi.nlm.nih.gov/projects/SNP/docs/dbSNP_VCF_Submission.pdf</p> <p>Questions on submission to dbSNP can be addressed to: snp-admin@ncbi.nlm.nih.gov</p> <p>NOTE: An “asserted position” is a statement, or assertion, based on experimental evidence that a variant is located at a particular position. Preferably, all variant asserted positions should be submitted on a sequence accession that is part of an assembly housed in the NCBI Assembly Resource, even though a RefSeq or GenBank accession for an asserted position not associated with an assembly are still accepted, but they will not appear on maps or graphical presentations of the assembly.</p>
Genome Structural Variations	<p>Data for genome structural variations, copy number variation (CNV) and insertions and deletions should be submitted to dbVar. More details on submission to dbVar are at: www.ncbi.nlm.nih.gov/dbvar/content/submission/</p>
Clinical Variation	<p>Human nucleotide and protein variations related to phenotypes with supporting evidence should be submitted to ClinVar database. Submission page can be found here: www.ncbi.nlm.nih.gov/clinvar/docs/submit/</p>

Submission of sequence-based reagents

Sequence-based reagents used in molecular biology research including primer pairs used in PCR amplifications, small RNA molecules used in RNA interference studies, and primer sets for use in expression studies, etc., should be submitted to the Probe database.

Starting Sequence Type	Detailed description and links
Primers, siRNAs, or probes	<p>Sequence based reagents, such as primer, nucleotide-based probe sequences, and RNAi, etc., should be submitted to the Probe Database.</p> <p>Probe Submission Portal www.ncbi.nlm.nih.gov/genome/probe/doc/Submitting.shtml</p>

Submission of non-sequence data

NCBI also accepts deposit of certain types of non-sequence data. These include data from MicroArray studies, clinical studies, chemical substances and results from their biological analyses, certain categories of manuscripts, and general biological project data.

Starting Data Type	Detailed Descriptions and Links
MicroArray Data	<p>MicroArray data, with the exception of those generated by clinical studies, may be submitted to Gene Expression Omnibus (GEO) using the GEO submission page. GEO submission portal: www.ncbi.nlm.nih.gov/geo/info/submission.html</p> <p>Functional genomics studies that examine gene expression, regulation or epigenomics using methods such as RNA-Seq, miRNA-Seq, ChIP-Seq or methyl-Seq may also be submitted to GEO as described. Other data accepted by GEO: www.ncbi.nlm.nih.gov/geo/info/seq.html</p>
Genotype and Phenotype Association Study Data	<p>Data from genotype and phenotype studies, such as genome wide association studies (GWAS), can generate a wide array of data types such as:</p> <ul style="list-style-type: none"> • Phenotype measures: demographic, clinical, exposure to factors • Genotype measure: SNPs, CNVs, imputed genotypes, MAF, raw array data • Next generation sequence data (brokered by SRA) • Medical Images: MRI, CT scans • Study documents • Association analysis results <p>These data should be submitted to dbGaP. dbGaP submission guideline is at: www.ncbi.nlm.nih.gov/books/NBK5297/</p>
Genetic Tests	<p>Voluntary submission of genetic test information by test providers should be deposited to Genetic Testing Registry (GTR) through the common submission portal, which is linked from the GTR submission help document: www.ncbi.nlm.nih.gov/gtr/docs/submit/</p>
Chemical Substances & Their Biological Activity Assay Data	<p>BioAssay data and chemical substance information should be submitted to PubChem via the PubChem Deposition Gateway. PubChem Deposition Gateway: pubchem.ncbi.nlm.nih.gov/upload/</p> <p>PubChem Deposition Help: pubchem.ncbi.nlm.nih.gov/upload/docs/upload_help.html</p>
Scientific Manuscripts	<p>Manuscripts or publications, from NIH-funded research may be deposited into the PubMed Central Database through NIHMS submissions system to comply with NIH Public Access Policy.</p> <p>Details on how to comply with NIH Public Access Policy: publicaccess.nih.gov/ NIH Manuscript Submission: www.nihms.nih.gov/</p>

Technical assistance

Many submission pages contain email help to the specific groups responsible for processing the submissions. Questions on submission of traditional sequences and for NCBI resources in general can also be sent to NCBI Service Desk at info@ncbi.nlm.nih.gov.