NCBI

GEO
Gene Expression Omnibus

GEO Publications | FAQ | MIAME | Email GEO

NCBI » GEO » Info » Submitting high-throughput sequence data to GEO

Login

Submitting high-throughput sequence data to GEO

WARNING: If you are submitting human data, it is your responsibility to comply with Human Subject Guidelines.

Assembling your submission

GEO accepts next generation sequence data that examine quantitative gene expression, gene regulation, epigenomics or other aspects of functional genomics using methods such as RNA-seq, miRNA-seq, ChIP-seq, RIP-seq, HiC-seq, methyl-seq, etc. We process all components of your study, including the samples, project description, processed data files, and we submit the raw data files to the Sequence Read Archive (SRA) on your behalf.

Once you have determined that GEO is an appropriate resource for your data type (see categories of data we do and do not accept), data should be submitted using the spreadsheet-based submission method described below. Alternatively, if your metadata are already in a database, and you can generate and export data in SOFT text format, you may prefer to use SOFT format.

There are three required components for the spreadsheet-based submission method:

1. a metadata spreadsheet
2. processed data files
3. raw data files

Details about each component are described below.

- **Metadata spreadsheet (Updated May 24, 2023)**

  Download metadata spreadsheet

  Metadata refers to descriptive information about the overall study, individual samples, all protocols, and references to processed and raw data file names. Information is supplied by completing all fields of a metadata template spreadsheet. Guidelines on the content of each field are provided within the spreadsheet.

- **Processed data files**

  Processed data are a required part of GEO submissions. The final processed data are defined as the data on which the conclusions in the related manuscript are based. We do not expect standard alignment files (e.g., BAM, SAM, BED) as processed data since conclusions are expected to be based on further-processed data. When standard alignments are the only processed data available, please write to us to inquire about whether your data are suitable for submission to GEO. Requirements for processed data files are not fully standardized and will depend on the nature of the experiment:

  - Expression profiling analysis usually generates quantitative data for features of interest. Features of interest may be genes, transcripts, exons, miRNA, or some other genetic entity. Two levels of data are often generated:

1. raw counts of sequencing reads for the features of interest, and/or
2. normalized abundance measurements, e.g., output from Cufflinks, Cuffdiff, DESeq, edgeR, etc.

- Either or both of these data types may be supplied as processed data. They may be formatted either as a matrix table or individual files for each sample. Provide complete data with values for all features (e.g., genes) and all samples, not only lists of differentially-expressed genes.
- ChIP-Seq and ATAC-seq data might include peak files with quantitative data, tag density files, etc. Common formats include WIG, bigWig, bedGraph. Please leave files in native format and do not paste peak data into Excel.

Features (e.g., genes, transcripts) in processed data files should be traceable using public accession numbers or chromosome coordinates. The reference assembly used (e.g., hg19, mm9, GCF_000001405.13) should be provided in the metadata spreadsheet.

A description of the format and content of processed data files should be provided in the metadata spreadsheet data processing fields.

If you provide WIG, bedGraph, GFF, or GTF files, please refer to the UCSC file format FAQ for format requirements.

- **Raw data files**

Raw data are a required part of GEO submissions. We will submit raw data files to SRA for you. The raw data files should be the original files containing reads and quality scores, as generated by the sequencing instrument (unless the raw files are barcoded/multiplexed, see below for further instructions).

Raw Data File Formats: Acceptable file formats include FASTQ, as well as other formats described in the SRA File Format Guide. Files that do not conform to supported format requirements will be deleted from our systems.

Barcode/Multiplexed Data: Whenever possible, we do require that files be demultiplexed so that each barcoded sample ends up with a dedicated run file. However, for single-cell sequencing studies (e.g. 10x Genomics, Drop-Seq, InDrops), we can support the submission of multiplexed files in cases where these files are required for reanalysis in your pipeline, or when demultiplexing would create an unmanageable number of files.

Paired-end Experiments: We usually expect 2 files per run (4 files per run when sequences and qualities are included in separate files). If submitting FASTQ files, please submit the original unedited files from the Illumina pipeline. Edited files may not be processed correctly by SRA.

MD5 Checksums: We recommend that submitters provide MD5 checksums for their raw data files. The checksums are used to verify file integrity. Checksums can be calculated using the following methods:

- **Unix**: md5sum <file>
- **OS X**: md5 <file>
- **Windows**: Application required. Many are available for free download.

Data File Compression: Individual files can be compressed to speed transfer, but this is not required. Acceptable compression formats are gzip and bzip2 (i.e. files ending with a .gz or .bz2 extension). Never compress binary files (e.g., BAM, bigWig, bigBed), and DO NOT upload ZIP archives (files with a .zip extension).

Uploading your submission

There are two steps for submission:

1. Transfer all your files to the GEO FTP server.
If your files are >5 terabytes, please contact us and do not transfer your files until you hear back from us.

Transfer Files

2. After the FTP transfer is complete, notify GEO using the Submit to GEO web form

Notify GEO

General Information

Data provision and standards

GEO sequence submission procedures are designed to encourage provision of MINSEQE elements:

- Thorough descriptions of the biological samples under investigation, and procedures to which they were subjected
- Thorough descriptions of the protocols used to generate and process the data
- Final processed (or summary) data from which the conclusions in associated manuscripts are based
- Original raw data files containing sequence reads and quality scores, which will be uploaded to NCBI's Sequence Read Archive (SRA) database.

Administration

All standard GEO administration and processing procedures apply to sequence submissions. These include:

- Unique and stable GEO accession numbers are issued to studies; these accessions can be cited in manuscripts
- GEO accession numbers are typically issued within 5 business days after completion of submission
- Data can be held private until publication
- Reviewers can have anonymous access to private records
- Submitters can update their records at any time

More information on these aspects is provided in our FAQ.

Categories of sequence submissions processed by GEO

GEO accepts

Studies concerning quantitative gene expression, gene regulation, epigenetics, or other functional genomic studies.

Examples include:

- mRNA profiling, RNA-seq (example)
- small RNA profiling, miRNA-seq (example)
- ChIP-Seq (example)
- HiC-seq (example)
- methyl-seq, bisulfite-seq (example)

If you have questions about whether GEO can accept your data type, please e-mail GEO.

GEO does not accept

- human data that require controlled access (submit to dbGaP and controlled access SRA)
- transcript assemblies (submit directly to SRA and the Transcriptome Shotgun Assembly Database)
- whole genome sequencing (submit directly to SRA and WGS)
- metagenomic sequencing (submit directly to SRA)
- resequencing, variation or copy number projects (submit directly to SRA and the appropriate NCBI variation resource)
- survey sequencing, whole exome (submit directly to SRA)

For more information about submitting data to NCBI, please refer to the Submission Wizard.