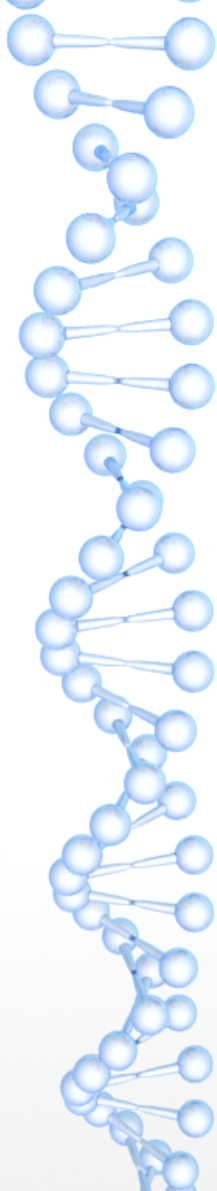


GEO Deposit



Club Bioinfo
07 Novembre 2019
Jean-Baptiste Claude

GEO Deposit



The screenshot shows the top navigation bar of the NCBI GEO website. On the left is the NCBI logo, and on the right is the GEO logo (Gene Expression Omnibus). Below the logos are navigation links for 'GEO Publications', 'FAQ', 'MIAME', and 'Email GEO'. A breadcrumb trail reads 'NCBI » GEO » Info » Submitting data' and a 'Login' link is on the far right. The main heading is 'Submitting data'. The text explains that GEO accepts various high-throughput genomic data and offers assistance with submissions. A list of data types includes microarrays, high-throughput sequencing, and other methods like NanoString and SAGE. A yellow warning box at the bottom states that users are responsible for complying with Human Subject Guidelines when submitting human data.

NCBI » GEO » Info » Submitting data Login

Submitting data

GEO accepts many categories of high-throughput functional genomic data, including all array-based applications and some high-throughput sequencing data.

We aim to make data deposit procedures as straightforward as possible and will provide as much assistance as you require to get your data submitted to GEO. If you have problems or questions about submission, [e-mail us](#) with a brief description of the type of data you are trying to submit, and one of our curators will quickly get back to you.

Data types

- Submit microarray
- Submit high-throughput sequencing
- Submit other (includes NanoString, RT-PCR, traditional SAGE)

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



Requirements

- You must have a MyNCBI account
- GEO requires raw data, processed data and metadata.
- Submit with GEOarchive spreadsheet is strongly recommended

NCBI Account Settings

Email

jean-baptiste.claude@ens-lyon.fr (confirmed)

This email is used for delivery of saved searches and recovery of password for your native NCBI account.

Change



Accessibility

- Your GEO submissions can remain private until a manuscript citing the data is published.
- You can allow reviewers anonymous access to your private records.
- You can update or edit your existing GEO records at any time.



How do I create a GEO account?

- You will need both a My NCBI account and an accompanying My GEO Profile to submit data.
- Submitters are asked to complete a My GEO Profile form that provides the contact information to be used by GEO curators
- The My NCBI account can be used to submit additional data in the future



NGS deposit

Submitting high-throughput sequence data to GEO

- Assembling your submission
 - Metadata spreadsheet
 - Processed data files
 - Raw data files
- Uploading your submission
- General Information
 - Data provisions, standards and administration
 - Categories of sequence submissions accepted by GEO



Submission

Uploading your submission

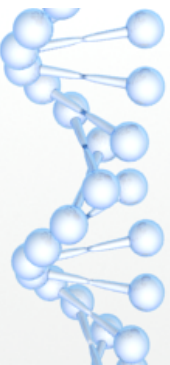
There are two steps for submission:

1. Transfer all your files to the GEO
FTP server

Transfer Files

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

Notify GEO



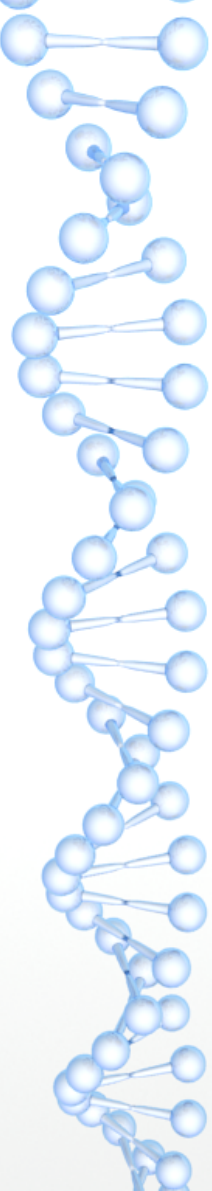
FTP

GEO File Transfer Protocol (FTP)

Step 1. Your personalized upload space is: **uploads/claude_jb_eBy1K8Wo**

Step 2. Transfer files to your personalized upload space according to FTP upload instructions below

▶ Transfer Files



Step 2. Transfer files to your personalized upload space according to FTP upload instructions below

▼ Transfer Files

- a. Create a new folder on your computer that has a meaningful name (e.g. `geo_submission_march13`) and place all of your submission files into the folder. If your submission is comprised of several datasets (e.g. ChIPseq, RNAseq, RRBS, etc) it is OK to organize the files for each data type into its own subfolder (e.g. `geo_submission_mar13/ChIPseq`).
- b. Confirm the size of your folder and **contact us if your submission exceeds 1 terabyte in size**. Please do not proceed with an upload larger than 1 terabyte until you hear back from GEO.
- c. For PC/Mac OS users we recommend transferring files with the free third-party software, [FileZilla Client](#). Please see below for detailed examples and other options.
- d. For LINUX/UNIX users, we recommend transferring files with 'ncftp' or 'lftp', but you can also use 'ftp', 'sftp', or 'ncftpput'. Please see below for detailed examples.
- e. Our FTP server credentials are:

host address	ftp-private.ncbi.nlm.nih.gov
username	geoftp
password	rebUzyi1
- f. After connecting, you **must navigate** to your personalized upload space: `uploads/claude_jb_eBy1K8Wo`
- g. After navigating to your personalized upload space, transfer the meaningfully-named submission folder from your computer to our server.
- h. Notify us (Step 3) and list the meaningfully-named folder in your notification. Do not proceed to Step 3 (below) until your transfer has completed.

Submit to GEO

You are logged in under the **claude_jb** account. Messages from GEO regarding your submission will be sent to the following email address(es): **jean-baptiste.claude@ens-lyon.fr**. If necessary, [visit your account](#) to edit your contact information. See [submitter accounts](#) for more details.

Use this form to either:

- Notify GEO about your [FTP file transfer](#) (suitable for [high-throughput sequencing](#) or [large microarray](#) submissions and updates)

Is your FTP file transfer to GEO complete?

- Yes, all my data have finished transferring

Do not proceed with this form until all components of your submission are fully transferred and ready for us to process.

Name(s) of the directory or files deposited

Submission kind

- new
- update or revision

When this submission should be released to the public ([more information about release dates](#))

- Release immediately following curation
- Release on specified date (up to 3 years from today)

Comment to GEO staff (optional)

Submit

- No, I need help



seq_template_v2.1.xls

High-throughput sequencing metadata template (version 2.1).
All fields in this template must be completed.
Templates containing example data are found in the METADATA EXAMPLES spreadsheet tabs at the foot of this page.
Field names (in blue on this page) should not be edited. Hover over cells containing **field names** to view field content guidelines.
Human data. If there are patient privacy concerns regarding making data fully public through GEO, please submit to [NCBI's db](#)

SERIES

This section describes the overall experiment.

title	downregulation of DDX5 and DDX17
summary	RNAseq : MCF-7 cells were transfected with siRNA targeting both DDX5 and DDX17
overall design	siRNA control (2 replicates) + siRNA DDX5/DDX17 (2 replicates)
contributor	Cyril, Bourgeois
contributor	Didier, Auboeuf
contributor	Clara, Benoit-Pilven
contributor	Sophie, Terrone
supplementary file	
SRA_center_name_code	[optional]



SAMPLES

This section lists and describes each of the biological Samples under investigation, as well as any protocols that are specific to individual Samples.
 # Additional "processed data file" or "raw file" columns may be included.

Sample name	title	source name	organism	characteristics	molecule	description	processed data	raw file	raw file
Sample 1	siCTL_N1	cells	Homo sapiens	MCF7	total RNA			siCTL_N1_GGCTAC_R1.fastq.gz	siCTL_N1_GGCTAC_R2.fastq.gz
Sample 2	siCTL_N2	cells	Homo sapiens	MCF7	total RNA			siCTL_N2_CTTGTA_R1.fastq.gz	siCTL_N2_CTTGTA_R2.fastq.gz
Sample 3	siDDX5_17_N1	cells	Homo sapiens	MCF7	total RNA			siDDX5_17_N1_AGTCAA_R1.fastq.gz	siDDX5_17_N1_AGTCAA_R2.fastq.gz
Sample 4	siDDX5_17_N2	cells	Homo sapiens	MCF7	total RNA			siDDX5_17_N2_AGTCC_R1.fastq.gz	siDDX5_17_N2_AGTCC_R2.fastq.gz

PROTOCOLS

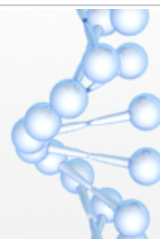
Any of the protocols below which are applicable to only a subset of Samples should be included as additional columns of the SAMPLES section instead.

growth protocol	
treatment protocol	MCF-7 cells were transfected with siRNA targeting both DDX5 and DDX17 RNA helicases
extract protocol	total RNA were extracted as described previously (Dardenne Cell Rep 2014)
library construction protocol	RNA libraries were prepared for sequencing using standard Illumina protocols. rRNA were depleted with the "TruSeq Stranded Total RNA with RiboZero Gold" kit.
library strategy	RNA-Seq 2x125bp

DATA PROCESSING PIPELINE

Data processing steps include base-calling, alignment, filtering, peak-calling, generation of normalized abundance measurements etc...
 # For each step provide a description, as well as software name, version, parameters, if applicable.
 # Include additional steps, as necessary.

data processing step	Data generated with an Illumina HiSeq 2500 platform
data processing step	Sequenced reads were trimmed for adaptor sequence with cutadapt and low quality bases (Q<20) with prinseq
data processing step	Mapping and count with Tophat
data processing step	Splicing analysis with FARLINE pipeline (publication in progress)
data processing step	
genome build	hg18
processed data files format and content	excel file containing splicing events differentially observed in DDX5-17 Vs control (events : acceptor, donor, mutually exclusive, multi exon skipping)



For each file listed in the "processed data file" columns of the SAMPLES section, provide additional information below.

PROCESSED DATA FILES

file name	file type	file checksum
analyse_stat_siDDX5_17-siCTL_recap.xls	xls	c21f2fed706da474e31ca18eb63fbf97

For each file listed in the "raw file" columns of the SAMPLES section, provide additional information below.

RAW FILES

file name	file type	file checksum	instrument model	read length	single or paired-end
siCTL_N1_GGCTAC_R1.fastq.gz	fastq	5936438dc15c3418b889c7a56a738303	Illumina HiSeq	125	paired-end
siCTL_N1_GGCTAC_R2.fastq.gz	fastq	e64aefff34eb6cf8e329226bc422ee00	Illumina HiSeq	125	paired-end
siCTL_N2_CTTGTA_R1.fastq.gz	fastq	8ccc9415642d2a796d6ed1df506d7738	Illumina HiSeq	125	paired-end
siCTL_N2_CTTGTA_R2.fastq.gz	fastq	761e2f8faf96d0f06e1dd782e90155af	Illumina HiSeq	125	paired-end
siDDX5_17_N1_AGTCAA_R1.fastq.gz	fastq	63863d50b0b3e8678a0ba8fd3d1a8b26	Illumina HiSeq	125	paired-end
siDDX5_17_N1_AGTCAA_R2.fastq.gz	fastq	6eafb63a1a9c34c17f5c6cfc5a3a7e11	Illumina HiSeq	125	paired-end
siDDX5_17_N2_AGTTCC_R1.fastq.gz	fastq	f5e06e037e03935e9306efe9f78411ef	Illumina HiSeq	125	paired-end
siDDX5_17_N2_AGTTCC_R2.fastq.gz	fastq	d6ace64f9e5a3a507d79da66d8f752d9	Illumina HiSeq	125	paired-end

For paired-end experiments, list the 2 associated raw files, and provide average insert size and standard deviation, if known. For SOLID experiments, list the 4

PAIRED-END EXPERIMENTS

file name 1	file name 2	average insert size	standard deviation
siCTL_N1_GGCTAC_R1.fastq.gz	siCTL_N1_GGCTAC_R2.fastq.gz		
siCTL_N2_CTTGTA_R1.fastq.gz	siCTL_N2_CTTGTA_R2.fastq.gz		
siDDX5_17_N1_AGTCAA_R1.fastq.gz	siDDX5_17_N1_AGTCAA_R2.fastq.gz		
siDDX5_17_N2_AGTTCC_R1.fastq.gz	siDDX5_17_N2_AGTTCC_R2.fastq.gz		

create checksum

- `cd [yourDirectory]`
- `md5sum * > md5sum.csv`

```
jeanbaptiste@Schrodinger:~/Desktop/Club_bioinfo_GEO_deposit_20191109$ md5sum *
a74be9479873edfd81466714c417ed48 1200px-US-NLM-NCBI-Logo.svg.png
f61bc77f055f091544702037f1041e94 20170125_Depot_siCTL_siDDX5-27_seq_template_v2.1.xls
e638332d686cd330fa31f6e3a7c7c70e Club_bioinfo_GEO_20191109.odp
a848f76948b9b3eba61e15955e735f29 geo_main.gif
77b0090a6644518c4867368d0fb2c414 Screenshot from 2019-11-07 11-30-05.png
3e42553e35a26fbac210065eb03064ec Screenshot from 2019-11-07 12-13-47.png
79b45674f7e0bb6389af3cbd99ffd425 Screenshot from 2019-11-07 12-17-03.png
c171dc1f2c85a8c99edce7daf8bfaebb Screenshot from 2019-11-07 12-20-02.png
25c8ec3e6c2d0d6006342fabab54ad38 Screenshot from 2019-11-07 12-24-09.png
bfe025b4fb5d89f4c942998715f48e28 Screenshot from 2019-11-07 12-25-04.png
0194392ff09b919356c1690ce9741179 Screenshot from 2019-11-07 12-26-29.png
fd5b563c820e41ee91034c8c07287f1a Screenshot from 2019-11-07 12-28-58.png
255aabf6d0d6448c3a94350f7ff4ef86 Screenshot from 2019-11-07 12-32-02.png
29a6d1159e0f624bfea51db44a374be8 Screenshot from 2019-11-07 12-33-47.png
4241944e79c083b00bc55e337b24babd seq_template_v2.1.xls
```



Data File Compression

- gzip and bzip2 (i.e. files ending with a .gz or .bz2 extension).
- Never compress binary files (e.g., BAM, bigWig, bigBed)
- DO NOT upload ZIP archives (files with a .zip extension).



SRA

- (...) 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.