

---

# STOmics Data Archive

*Release 1.0*

CNGBdb

Nov 21, 2022



# STOMICS DATA STANDARD

<b>1</b>	<b>Project</b>	<b>3</b>
<b>2</b>	<b>STOmics Sample</b>	<b>5</b>
<b>3</b>	<b>Tissue Section</b>	<b>9</b>
<b>4</b>	<b>Experiment &amp; Run</b>	<b>13</b>
<b>5</b>	<b>Analysis</b>	<b>17</b>
<b>6</b>	<b>General guidance</b>	<b>19</b>
<b>7</b>	<b>Project registration</b>	<b>25</b>
<b>8</b>	<b>Offline template submission</b>	<b>27</b>



Welcome to the documents for submission of SpatioTemporal Omics data for [STOmics DB](#). Please use the links to find instructions specific to your needs.

The realization of the importance of the spatial organization and the exact position of molecular features, historically unobtained in bulk and single cell experiments, have driven the technological advancements in spatially resolved transcriptomics.

An approach is to capture transcripts in situ, then perform sequencing ex situ.

[Stereo-seq](#) developed by **BGI Genomics**, and [Visium Spatial Gene Expression](#) developed by **10x Genomics**, are two popular technologies that combined spatial chip and in situ RNA capture technology.



## PROJECT

An overall description of a single research initiative; a project will typically relate to multiple samples and datasets.

## 1.1 General Information

### \*Project title

- Definition: A phrase or short sentence that describes the overall study.

### \*Summary

- Definition: Thorough description of the goals and objectives of this study. The abstract from the associated manuscript may be suitable.

### Relevance

- Definition: The primary general relevance of the project.
- Value syntax: ['Agricultural', 'Medical', 'Industrial', 'Environmental', 'Evolution', 'Model organism', 'Other']

### \*Project data type

- Definition: A general label indicating the primary study goal.
- Value syntax: {'Genome sequencing and assembly', 'Raw sequence reads', 'Genome sequencing', 'Assembly', 'Clone ends', 'Epigenomics', 'Exome', 'Map', 'Metagenome', 'Metagenomic assembly', 'Phenotype or Genotype', 'Proteome', 'Random survey', 'Targeted loci cultured', 'Targeted loci environmental', 'Targeted Locus (Loci)', 'Transcriptome or Gene expression', 'Variation', 'Metabolome', 'STomics', 'Other'}

### \*Sample scope

- Definition:

The scope and purity of the biological sample used for the study.

Choose Multiisolate as the Scope when the goal of the research is to compare multiple individuals or strains of the same species, e.g., in a "Variation" or "Genome sequencing and assembly" project.

Choose Multispecies when different species are being examined.

Choose Monoisolate if the goal is to make a single genome or transcriptome assembly, even if more than one individual was the source of the DNA or RNA.

- Value syntax: ['Monoisolate', 'Multiisolate', 'Multispecies', 'Environment', 'Synthetic', 'Other']

### \*Related projects

- Definition: The projects that are related to this project.

## 1.2 Contributors

- Definition: The main contributors or the leader of the study used as the main contact for the study.

## 1.3 Publications

- Definition: Present the research results of the Project with publications.

Fields for publications
<ul style="list-style-type: none"><li>• Status</li><li>• Title</li><li>• Authors</li></ul>

## 1.4 Experimental protocols

- Definition: Experimental protocols designed for the overall study. It should be documented to contain such as sample prepararion, sample staining and imaging, tissue permeabilization, library construction, sequencing, analysis and visualization, etc. The document should be submitted in Microsoft Word Document (DOCX/DOC) or Portable Document Format (PDF).

## STOMICS SAMPLE

Description of biological source material; each physically unique specimen should be registered as a single sample with a unique set of attributes.

\*sample name :

- Definition: An arbitrary and unique identifier for each sample.
- Note: The sample name is used to associate Sample with other objects.

\*sample title :

- Definition: The sample title for public display is a short, preferably a single sentence, description of the sample.
- Note: The sample title is for public display of Sample.

\*taxonomy ID :

- Definition: The Taxonomy ID indicates the taxonomic classification of the sample (e.g. 9606 for human).

\*organism :

- Definition: The most descriptive organism name for this sample (to the species, if relevant).

\*isolate :

- Definition: Identification or description of the specific individual from which this sample was obtained.

\*tissue

- Definition: Type of tissue the sample was taken from.

\*sex :

- Definition: Physical sex of sampled organism.
- Field Format: text choice
- Expected value: enumeration
- Value syntax: ['male', 'female', 'pooled male and female', 'neuter', 'hermaphrodite', 'intersex', 'not determined', 'not applicable', 'not collected', 'not provided', 'restricted access', 'missing']

\*age :

- Field Format: restricted text
- Expected value: measurement value
- Value syntax: {float} {unit}
- Preferred unit: centuries,days,decades,hours,minutes,months,seconds,weeks,years

\*development stage :

- Definition: Developmental stage at the time of sampling.

\*biomaterial provider :

- Definition: Name and address of the lab or PI, or a culture collection identifier.
- Field Format: free text

\*geographic location :

- Definition: Geographical origin of the sample; use the appropriate name from this list <http://www.insdc.org/documents/country-qualifier-vocabulary>. Use a colon to separate the country or ocean from more detailed information about the location, e.g. "China:Shenzhen" or "China:Hebei:Baoding".
- Field Format: restricted text
- Expected value: country or sea name (INSDC or GAZ):region(GAZ):specific location name
- Value syntax: {term}:{term}:{text}
- Example: Germany:Sylt:Hausstrand

\*collection date :

- Definition: The time of sampling, either as an instance (single point in time) or interval. date/time ranges are supported by providing two dates from among the supported value formats, delimited by a forward-slash character, e.g., 2017/2019; In case no exact time is available, the date/time can be right truncated i.e. all of these are valid times: 2008-01-23T19:23:10+00:00; 2008-01-23T19:23:10; 2008-01-23; 2008-01; 2008; Except: 2008-01; 2008 all are ISO8601 compliant.
- Field Format: restricted text
- Expected value: date and time
- Value syntax: {timestamp}
- Example: 2017/2019, 2008-01-23T19:23:10+00:00, 2008-01-23T19:23:10, 2008-01-23, 2008-01, 2008

collected by :

- Definition: Name of persons or institute who collected the sample.
- Field Format: free text

latitude and longitude :

- Definition: The geographical coordinates of the location where the sample was collected. Specify as degrees latitude and longitude in format "d[d.ddd] N|S d[dd.ddd] W|E", e.g., 38.98 N 77.11 W
- Field Format: restricted text
- Expected value: decimal degrees
- Value syntax: {float} {float}
- Example: 38.98 N 77.11 W

strain :

- Definition: Microbial or eukaryotic strain name.

breed :

- Definition: breed name - chiefly used in domesticated animals or plants.

cultivar :

- Definition: cultivar name - cultivated variety of plant.

ecotype :

- Definition: A population within a given species displaying genetically based, phenotypic traits that reflect adaptation to a local habitat, e.g., Columbia

isolation source :

- Definition: Describes the physical, environmental and/or local geographical source of the biological sample from which the sample was derived.

disease :

- Definition: List of diseases diagnosed; can include multiple diagnoses. the value of the field depends on host; for humans the terms should be chosen from DO (Disease Ontology), free text for non-human. For DO terms, please see <https://www.ebi.ac.uk/ols/ontologies/symp>
- Field Format: free text
- Expected value: disease name or DO
- Value syntax: {term}

disease stage :

- Definition: Stage of disease at the time of sampling.

cell line :

- Definition: Name of the cell line.

cell type :

- Definition: Type of cell of the sample or from which the sample was obtained.

treatment :

description :

- Definition: Description of the sample.



## **TISSUE SECTION**

Description of fresh frozen or formalin fixed & paraffin embedded (FFPE) tissue that has undergone a series of treatments. Cryosectioning, section placement, staining and visualization, then the tissue is permeabilized for reactions to generate a sequencing-ready library.

\*tissue section alias:

- Definition: An arbitrary and unique identifier for each tissue section.
- Note: The tissue section alias is used to associate Tissue Section with other data objects.

\*tissue section ID:

- Definition: A phrase or short sentence for public display.
- Note: The tissue section ID is for public display of Tissue Section.

\*tissue type:

- Definition: It can be fresh frozen or formalin fixed paraffin embedded (FFPE) tissues.
- Field Format: text choice
- Expected value: enumeration
- Value syntax: ['fresh frozen (FF)', 'formalin fixed paraffin embedded (FFPE)']

tissue freezing and embedding:

- Definition: Freezing and embedding may be performed simultaneously, or as separate steps. If fresh tissue is available, simultaneous freezing and embedding may be preferred. Thin tissues that are prone to curling may benefit from simultaneous freezing and embedding.
- Field Format: text choice
- Expected value: enumeration
- Value syntax: ['', 'simultaneously', 'separately']

\*section resource:

- Definition: Describe the section from an anatomical point of view. For human, it may be “sagittal posterior section”, “sagittal anterior section”, etc. For plant, it may be “transverse section”, “tangential longitudinal section”, “radial longitudinal section”, etc.

slice position:

- Definition: Slice position can be described as the relative position between tissue sections. For example, 355/1000 means that 1000 slices have been cut from the sample, and this tissue section is the 355th slice.
- Example: 355/1000

\*cryosectioning temperature:

- Definition: Cryosectioning temperatures impact tissue section integrity. A temperature setting of  $-20^{\circ}\text{C}$  for blade and  $-10^{\circ}\text{C}$  for the specimen head is recommended. The temperature settings depend upon the local conditions, tissue types, and the cryostat used and should be optimized based on the quality of resulting tissue sections.
- Field Format: restricted text
- Expected value: measurement value
- Value syntax: {float} {unit} [deg C]
- Preferred unit: degree Celsius,  $^{\circ}\text{C}$

\*tissue section size:

- Definition: A tissue section of 6.5 mm\*6.5 mm is compatible with Visium Spatial slides. A tissue section of 130 mm\*130 mm is compatible with Stereo-seq Spatial slides.
- Field Format: restricted text
- Expected value: measurement value
- Value syntax: {float} {unit}\*{float} {unit}
- Preferred unit: millimeter\*millimeter, mm\*mm

\*section thickness:

- Definition: Recommended section thickness for most tissue types is 10  $\mu\text{m}$ . Tissues with higher fat content (e.g., breast tissue) may require thicker sections. Visit the 10x Genomics support website for guidance on section thickness for compatible tissue types (<https://support.10xgenomics.com/spatial-gene-expression/sample-prep/doc/specifications-visium-spatial-gene-expression-optimized-tissues>).
- Field Format: restricted text
- Expected value: measurement value
- Value syntax: {integer} {unit}
- Preferred unit: micrometer, m

RIN:

- Definition: RNA Integrity Number (RIN) should be 7 and RNA quality assessment should be done before placing the tissue sections on the Spatial slides. Various factors could lead to low RIN scores, such as specific tissue types, diseased or necrotic tissues, sample preparation and handling.
- Field Format: restricted text
- Expected value: measurement value
- Value syntax: {float}

tissue score:

- Definition: Large tissue samples can be scored during sectioning to generate smaller samples to fit the Capture Areas. Scoring can be done by making a shallow incision (~1 mm deep) on the cutting surface of the tissue with a pre-cooled razor blade.
- Field Format: restricted text
- Expected value: measurement value
- Value syntax: {float}

DV200:

- Definition: DV200 represents the percentage of RNA fragments that are >200 nucleotides in size. Using DV200 to assess FFPE RNA quality and it should be 50%.

- Field Format: restricted text
- Expected value: measurement value
- Value syntax: {integer} {unit}
- Preferred unit: percentage, %

\*staining protocol:

- Definition: It can be immunofluorescent staining, DNA fluorescent staining, or histological staining, etc., which is used to obtain spatial information such as RNA fragments distribution, specific molecules distribution via the Spatial slides.
- Field Format: text choice
- Expected value: enumeration
- Value syntax: ['ssDNA staining', 'H&E Staining', 'IF Staining', 'not determined', 'not applicable', 'not collected', 'not provided', 'restricted access', 'missing']

optimal permeabilization time:

- Definition: For fresh frozen sample, ensure that permeabilization times are optimized for each tissue type. Sub-optimal permeabilization will diminish sensitivity and spatial resolution.
- Field Format: restricted text
- Expected value: measurement value
- Value syntax: {integer} {unit}
- Preferred unit: hours, minutes



## EXPERIMENT & RUN

A description of tissue-sample-specific sequencing library, instrument and sequencing methods. Runs describe the files that belong to the previously created experiments.

### 4.1 Metadata

\*spatial slide:

- Definition: slide serial number.

\*experiment title:

- Definition: Short description that will identify the dataset on public pages. A clear and concise formula for the title would be like: {methodology} of {organism}: {sample info} e.g. RNA-Seq of mus musculus: adult female spleen

\*library name:

- Definition: Short unique identifier for the sequencing library. Each library name MUST be unique!

\*library strategy:

- Definition: Sequencing technique intended for the library.
- Value syntax: ['STOmics\_RNA']

\*library source:

- Definition: The library source specifies the type of source material that is being sequenced.
- Value syntax: ['TRANSCRIPTOMIC SPATIAL']

\*library selection:

- Definition: Method used to enrich the target in the sequence library preparation.
- Value syntax: ['cDNA barcoded with spatial and molecular identifier(FF)', 'ligated probes extended with spatial and molecular identifier(FFPE)']

\*sequencer:

- Value syntax: ['DNBSEQ-G50(MGISEQ-200)', 'DNBSEQ-G400(MGISEQ-2000)', 'DNBSEQ-G400 FAST', 'DNBSEQ-T1', 'DNBSEQ-T5', 'DNBSEQ-T7', 'DNBSEQ-T10', 'DNBSEQ-T10x4', 'DNBSEQ-T20', 'DNBSEQ-T20x2', 'Illumina NovaSeq 6000', 'Illumina HiSeq 4000', 'Illumina HiSeq 3000', 'Illumina HiSeq 2500', 'Illumina NextSeq 500', 'Illumina NextSeq 550', 'Illumina NextSeq 2000', 'Illumina MiSeq', 'Illumina iSeq 100']

\*library layout:

- Definition: The library layout specifies whether to expect single, paired, or other configuration of reads. In the case of paired reads, information about the relative distance and orientation is specified.
- Value syntax: ['paired']

\*nominal size:

- Definition: The average insert size for paired reads.
- Value syntax: {integer}

\*spot layout:

- Definition: a spot descriptor that describes the position of the technical reads (e.g. Spatial barcode/CID, UMI/MID).
- Example: Spatial barcode: read1 1-25; UMI: read1 41-50

## 4.2 Data file

### 4.2.1 FASTQ files

FASTQ format is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. This is the most widely used format in sequence analysis as well as what is generally delivered from a sequencer.

Each sequence requires at least 4 lines:

```
@<identifier and expected information>
<sequence>
+<identifier and other information OR empty string>
<quality>
```

- Identifier and expected information: text string terminated by white space.
- fastq sequence should contain standard base calls (ACTGactg) or unknown bases (Nn) and can vary in length.
- Qualities options:

Decimal-encoding, space-delimited	[0-9]+   <quality>s[0-9]+
Phred-33 ASCII	[!\"#\$%&'()*+,-./0-9;:<=>?@A-I]+
Phred-64 ASCII	[@A-Z[\]^_`a-h]+

Quality string length should be equal to sequence length.

### Paired-end FASTQ

Paired-end data submitted in FASTQ format should be submitted as separate files for forward and reverse reads, in which the reads are in the same order.

## 4.2.2 BAM files

BAM is a compressed binary version of the Sequence Alignment/Map (SAM) format (see [SAMv1](#)) that is used to represent aligned sequences.

The BAM format file generated by STOmics Analysis Workflow (SAW) can be downloaded at <https://hub.docker.com/r/stomics/saw>) is more suitable for reading, writing and storage of spatial transcriptome big data.

SAW **mapping** BAM adds custom tags in the BAM optional field to record reads coordinates, CID and MID information. **count** BAM adds annotation information in the tag field. Custom tags are described in the table below.

Tag	Description
Cx:i	The x coordinate of CID.
Cy:i	The y coordinate of CID.
UR:Z	The hexadecimal representation of uncorrected binary-encoded MID.
XF:Z	Mapping region on the reference genome. Valid value: 0=EXONIC, 1=INTRONIC, 2=INTERGENIC.
GE:Z	Annotated gene name.
GS:Z	'+' or '-', indicating forward/reverse strand respectively.
UB:Z	The hexadecimal representation of count corrected binary-encoded MID.

Example of **mapping** BAM:

```
E100026571L1C009R00301275185 16 1 3000095 255 26M121066N74M * 0 0
GGCTTTTTTTTTTTTTTTTTTTTTTTTTCTAAATATTGGGTTTATTAGCACCAT-
GATAACTGTAT
ATTAATTTGCACTGACTGTCATAACAAAATACG+:GFFGGFGFFGFFGFGGFFGFFFCFGFCFG
GGFGGFGFFFGGFGGFFFGGFFGFFGFGFFGFGFFFGGFFFGFFFGGFFGGFFGFEF
NH:i:1 HI:i:1 AS:i:88 nM:i:0 Cx:i:4826 Cy:i:11598 UR:Z:6FA29
```

Example of **count** BAM:

```
E100026571L1C002R00703943265 1040 1 3082766 255 11M132671N89M * 0 0 CTGCT-
GCAGCTTTTTTTCTTTGAGATTTATTTTATGCTATGTGTATGGGTATTTGCCTGCATAT

ATGTCTATGCACCATGTGTGTGCAGTGCTTGAGFFFFFECGFDCFGDGDFFEE@EEGIBFGGCGFFGA

CGFCGFFDGDGFFFFFFFEGCDFCGFFGG@FFF=EFFDGGGGGFDGFFFGGGFGFFGGGFFGGGDFG
NH:i:1 HI:i:1 AS:i:88 nM:i:0 Cx:i:7767 Cy:i:18052 UR:Z:7AE49 XF:i:0 GE:Z:Xkr4 GS:Z:-
UB:Z:79E49
```

## 4.2.3 Reference files

### Reference fasta

FASTA format is the most basic format for reporting a sequence. A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The definition line (define) is distinguished from the sequence data by a greater-than (>) symbol at the beginning. The word following the ">" symbol is the identifier of the sequence, and the rest of the line is the description (optional).

**Example:**

```
>gi|129295|sp|P01013|OVAX_CHICK GENE X PROTEIN (OVALBUMIN-RELATED)

QIKDLLVSSSTDLDITLVNNAIYFKGMWKTAFAEDTREMPPHVTQESKPVQMMCMNNSFNVAT
```

LPAEKMKILELPPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTPNPTMEKRRVKVYLPQMKIEEKY  
NLTSVLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMESEDGIEMAGSTGVIEDIKHSPESQFR  
ADHPFLFLIKHNPTNTIVYFGRYWSP

## Reference annotation

A 9-column annotation file conforming to the GFF, GFF3 or GTF specifications can be used for reference annotation submission.

General Feature Format (GFF) is a tab-delimited text file that holds information any and every feature. Everything from CDS, microRNAs, binding domains, ORFs, and more can be handled by this format. It consists of one line per feature, each containing 9 columns of data, plus optional track definition lines.

There have been many variations of the original GFF format and many have since become incompatible with each other. The latest accepted format (GFF3) has 9 required fields, though not all are utilized (either blank or a default value of '.').

The Gene transfer format (GTF) is a file format used to hold information about gene structure. It is a tab-delimited text format based on the general feature format (GFF), but contains some additional conventions specific to gene information.

The basic characteristics of the file formats are described at:

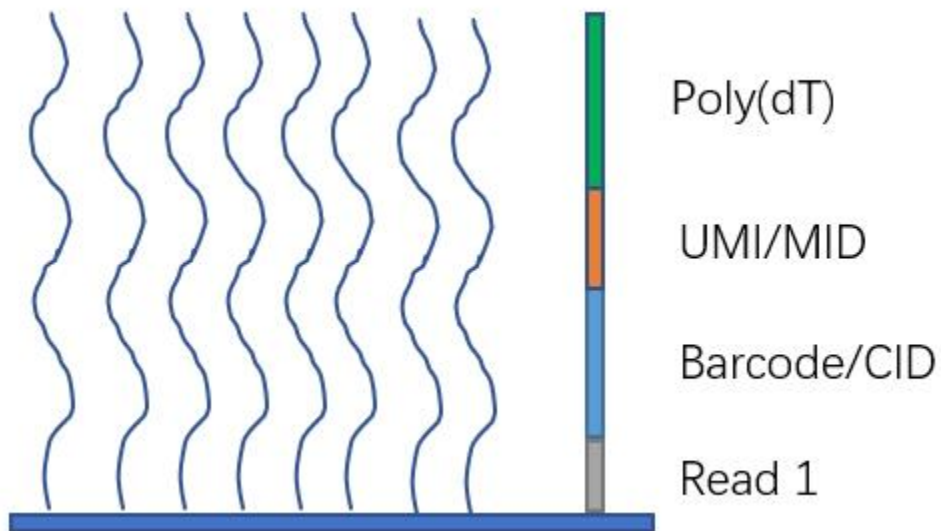
- **GFF:** <https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md>
- **GTF:** <http://mblab.wustl.edu/GTF22.html>

## ANALYSIS

Analysis represents a collection of STOmics data. It should contain spatial positions, gene expression matrices, visualized images, cell annotation, etc.

## 5.1 Spatial Positions

Spatial coordinate index data, contains Spatial Barcode/Coordinate Identity (CID) and it's position coordinate.



## 5.2 Gene Expression Information

Gene expression information is usually given in the form of matrix, which records the number of UMIs/MIDs associated with a feature and a barcode/CID.

## 5.3 Visualization Images

There are a series of images for tissue detection, for example,

- brightfield or fluorescence images acquired by imaging system,
- registered microscopic image,
- downsampled versions of the original, full-resolution image.

## 5.4 Cell Annotation

Cell identification and segmentation performed to define each cell population based on marker genes, cell morphology, etc.

## 5.5 Other Downstream Analysis Data

Data sets generated by downstream analysis such as marker identification, cluster annotation, differential expression, etc. Scripts are also included.

## GENERAL GUIDANCE

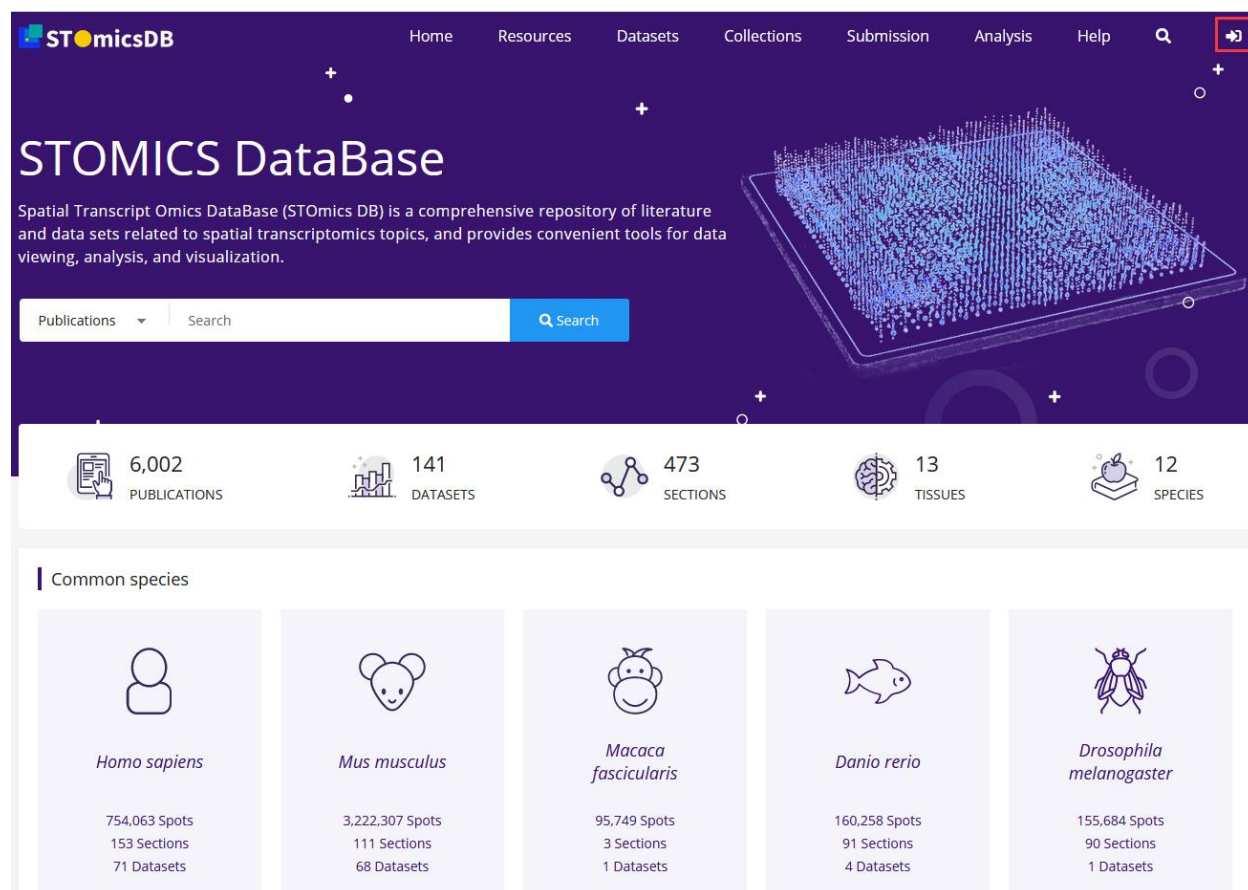
Welcome to the general guidance for the STOmics data submission. Please take a moment to view this introduction before you begin your submission.

### 6.1 Getting Started

#### 6.1.1 Register a Submission Account

Before you can submit data to [CNGBdb](#), you must register a CNGBdb account.

To submit STOmics data, please navigate to [STOmics DB](#), you will be presented with the below interface.



**STOMICS DataBase**

Spatial Transcript Omics DataBase (STOmics DB) is a comprehensive repository of literature and data sets related to spatial transcriptomics topics, and provides convenient tools for data viewing, analysis, and visualization.

Publications ▾ Search

6,002 PUBLICATIONS   141 DATASETS   473 SECTIONS   13 TISSUES   12 SPECIES

**Common species**

Species	Spots	Sections	Datasets
<i>Homo sapiens</i>	754,063	153	71
<i>Mus musculus</i>	3,222,307	111	68
<i>Macaca fascicularis</i>	95,749	3	1
<i>Danio rerio</i>	160,258	91	4
<i>Drosophila melanogaster</i>	155,684	90	1



By password

By phone

E-mail/Username

Password



Remember me

[Forget password](#) | [Signup](#)

Login

Other login methods



Wechat



BGI



Github

Please choose the manner that suitable for you to register, and remember the account and password.

**Note:** Each manner your registration corresponds to an account, regardless of whetherless it is registered by the same person.

## 6.1.2 Register a Submitter

To start submitting STOmics data, submitter who claimed ownership of the data should be registered first.

1 Submitter 2 Submission type 3 Submission application 4 Study 5 Overview

Please ensure that the information you filled in is accurate and valid. After submission, you cannot modify it yourself. If you need to modify it, please contact [datasubs@cngb.org](mailto:datasubs@cngb.org).

**Submitter**

Chinese name English name \* E-mail \* Phone number

\* Department \* Organization Submitting organization URL

\* Street \* City State/Province \* Country

\* Verification code | Get code

Save and next

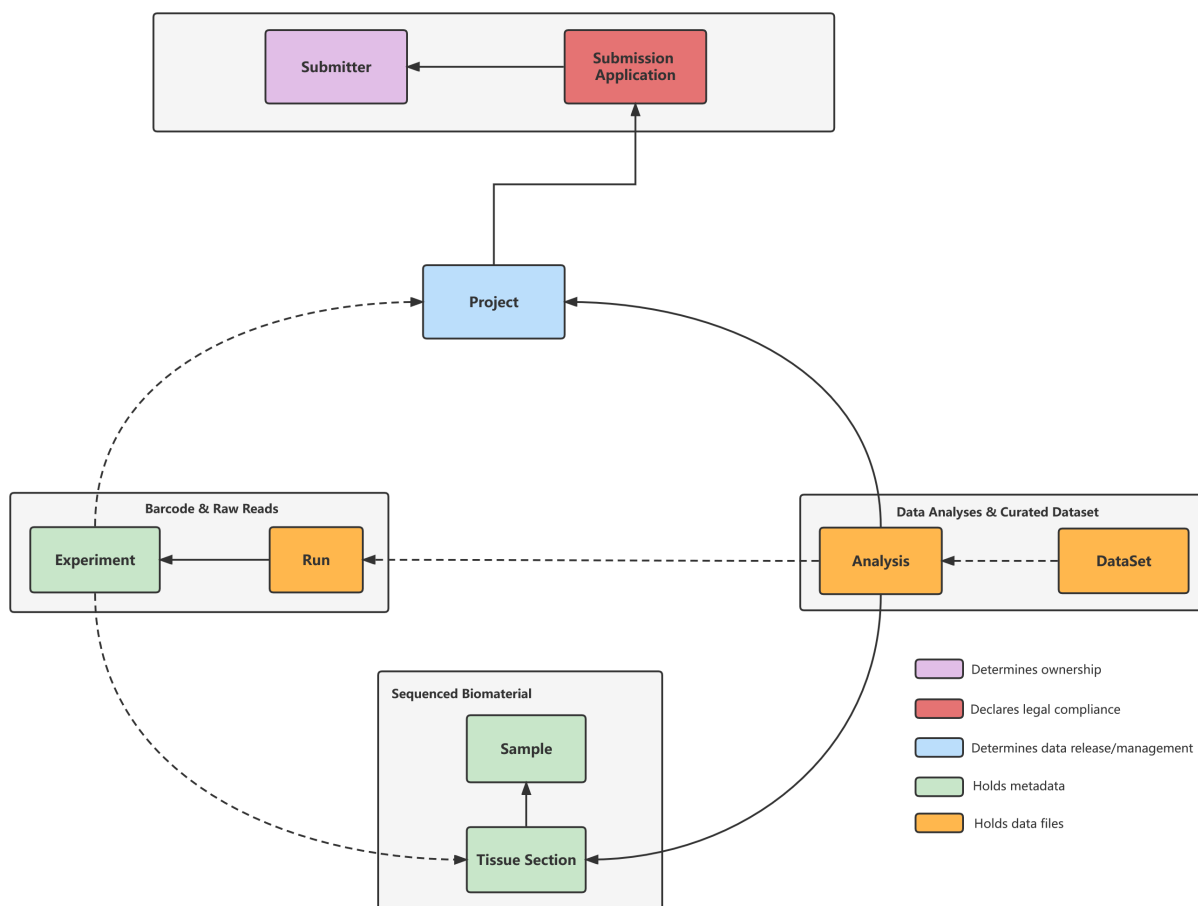
- Chinese name and English name should be chosen at least one to fill in.
- The fields with \* are mandatory.
- Click **Get code** to get the **Verification code** from your mobile phone, because you have filled in the phone number on this page.

### Note:

- Please be more careful to register submitter. Once submitted, it cannot be modified by yourself, and it will be applied to all subsequent submitted data in your account by default.
- If you need to modify it, please contact [datasubs@cngb.org](mailto:datasubs@cngb.org).
- If you found that the submitter has not been reviewed, please contact us immediately. It will affect your creating a new submission.

### 6.1.3 Metadata Model

Submissions are represented using a number of different metadata objects. Before submitting STOmics data, it is important to familiarise yourself with the metadata model. This will determine what you need to submit.



- **Submitter:** A person who owns the data.
- **Submission Application:** Legal compliance statement of your data.
- **Project:** A project groups together submitted data and controls its management. A project accession is typically used when citing submitted data.
- **Sample:** A sample contains information about the sequenced source material. Samples are always associated with a taxonomy.
- **Tissue Section:** A tissue section represents a slice cyosectioned from the sample.
- **Experiment:** An experiment contains information about a sequencing experiment including library and instrument details.
- **Run:** A run is part of an experiment and refers to data files containing sequence reads.
- **Analysis:** An analysis contains secondary analysis results derived from sequence reads. An analysis is typically a collection of STOmics data.

### 6.1.4 Accession Numbers

Completed submissions results in accession numbers. A set of rules describing the format of the accessions are shown below.

Object	Accession format	Examples
Submission	“sts” + 7 numerals	sts0000001
Project	“STT” + 7 numerals	STT0000001
Sample	“STSA” + 7 numerals	STSA0000001
Tissue Section	“STTS” + 7 numerals	STTS0000001
Experiment	“STEP” + 7 numerals	STEP0000001
Run	“STRN” + 7 numerals	STRN0000001
Dataset	Coming soon...	Coming soon...

---

**Note:** Not all accessions become available in the browser and not all can be used in publications.

---

### 6.1.5 How to cite

The top-level Project accession can be cited as follows.

The data that support the findings of this study have been deposited into STOmics DB of China National GeneBank DataBase (CNGBdb) [1] with accession number STTXXXXXXX.

[1] Chen FZ, You LJ, Yang F, et al. CNGBdb: China National GeneBank DataBase. Hereditas. 2020;42(08):799-809. doi:10.16288/j.ycz.20-080.



## PROJECT REGISTRATION

Before to register your project, you should fill in a submission application first, which mainly used to declare the legal compliance of the project data.

### 7.1 Submission Application

#### 7.1.1 Submission Application

Submission application declares the data generated from project is legal compliance, especially the Human Genetic Resources (HGR) information involved.

##### Data Access Manner

There are two manners to manage your data,

- One is **Public**. It means all your information submitted associated with the project will be released at the `release date`.
- The other one is **Controlled**. It means that the metadata of the project will be released at the `release date`, and the data files will never be released. The data files should be controlled access.

---

**Note:** `release date` can be as much as 2 years beyond the present date.

---

##### Resources

There are some important information should be provided like:

- Principal investigator
- Project cooperation entity (multiple)
- Data type (multiple choices)
- Sample type (multiple choices)
- Human microbiome sample and data collection (if refers to human metagenome)
- Human genetic resources information
- Data collection entity or preservation entity

---

**Important:** Please fill in the above information with careful according to the actual situation of your project!

---

Last but not the least, tick off the commitment agreement if you have read and understood it.

## 7.2 Spatial Technology

There are two popular technologies:

- [Stereo-seq](#) developed by **BGI Genomics**, and
- [Visium Spatial Gene Expression](#) developed by **10x Genomics**

for your choice.

---

**Note:** Once the spatial technology you have choose, it cannot be modified, you can only create another new submission.

---

## 7.3 Project information

You can register a new project, or use an already registered project.

### 7.3.1 Project submission

The project mainly involves the following information:

- \*Project title
- \*Summary
- \*Project data type (multiple choices)
  - STOmics selected by default.
  - If no data type listed suitable for you, you can choose “other”, and provide additional information to describe your project data type.
- \*Sample scope (drop-down menu)
- Relevance (drop-down menu)
- \*Contributors (multiple)
- Publications (multiple)
  - It contains the publication status, the article title and authors.
- Related projects (multiple)
  - The projects that are related to this project can be listed here. The related project accessions will be shown on the project details page when this project is public.
- \*Experimental protocols file
  - The document should be submitted in Microsoft Word Document (DOCX/DOC) or Portable Document Format (PDF).

## OFFLINE TEMPLATE SUBMISSION

There are some information to submitted with offline template:

- Sample
- Tissue Section
- Experiment & Run (if sequencing reads choose to be submitted)
- STOmics Analysis
- Other

These templates can be downloaded at <https://ftp.cnbg.org/pub/stomics/>.

### 8.1 Sample registration

In the **STOmics sample** template, the green fields are mandatory, and the yellow fields are optional.

- **sample name**  
User-defined name for the sample. It is unique for each sample. It cannot be modified if submitted, because it only used for objects association in the database.
- **sample title**  
The sample title is for public display as you want.
- **taxonomy ID and organism**  
Species information for your research subjects. The **taxonomy ID** and **organism** should be consistent with each other. For example, the **taxonomy ID** is 10090, and the **organism** is *Mus musculus* for the house mouse, which can be retrived [here](#).
- **sex**  
It need to choose from the following list:  
['male', 'female', 'pooled male and female', 'neuter', 'hermaphrodite', 'intersex', 'not determined', 'not applicable', 'not collected', 'not provided', 'restricted access', 'missing']
- **age**  
It is restricted text. The value syntax is '{float} {unit}'.
- **geographic location**  
It is restricted text. It expected to fill in country or sea name (INSDC or GAZ):region(GAZ):specific location name. The value syntax is '{term}:{term}:{text}', e.g. "China:Shenzhen" or "China:Hebei:Baoding".

- **collection date**

It is restricted text. It expected to fill in date and time, for example, 2008-01-23T19:23:10+00:00; 2008-01-23T19:23:10; 2008-01-23; 2008-01; 2008; 2017/2019.

The value can not be a future time.

- **latitude and longitude**

It is restricted text. Specify as degrees latitude and longitude in format “d[d.dddd] N|S d[dd.dddd] W|E”, e.g., 38.98 N 77.11 W.

For more detailed explanation for the fields, please refer to the standard below:

---

**Important:**

- The number of the Sample filled in the template cannot be greater than 100, and the template file cannot be greater than 10MB.
  - **sample name** cannot be modified. **sample title** can be modified because it is for publication.
  - Each of your samples must have differentiating information (excluding **sample name**, **sample title**, and **description**). This check was implemented to encourage submitters to include distinguishing information in their samples. If it is necessary to represent true biological replicates as separate Samples, you might add an ‘aliquot’ or ‘replicate’ attribute, e.g., ‘replicate = biological replicate 1’, as appropriate.
  - The **taxonomy ID** and **organism** should be consistent with each other.
  - **geographic location** and **latitude and longitude** (if filled in) should be consistent with each other.
  - If need to modify Sample, the assigned sample accession numbers can not be modified.
- 

## 8.2 Tissue Section registration

There are two ways/templates to register Tissue Section. You can choose one to submit your tissue sections.

- One is registered with **sample name**,

You can register Tissue Section with **sample name**, which associated a tissue section to a sample.

- The other one is registered with **sample accession**.

You can also register Tissue Section with **sample accession** if you get the sample accession number, which is start with ‘STSA’.

In the **Tissue section** template, the green fields are mandatory, and the yellow fields are optional.

- **tissue section alias**

User-defined name for the tissue section. It is unique for each tissue section. It cannot be modified if submitted, because it only used for objects association in the database.

- **tissue section ID**

The tissue section ID is for public display as you want.

- **tissue type**

It need to choose from the following list:

['fresh frozen (FF)', 'formalin fixed paraffin embedded (FFPE)']

- **tissue freezing and embedding**

It need to choose from the following list:

['', 'simultaneously', 'separately']

- **section thickness**

It is restricted text. The value syntax is '{integer} {unit}'. The unit is usually 'm'.

- **RIN**

Numbers are allowed. Supporting up to 1 digits after the decimal point.

- **tissue score**

Numbers are allowed. Supporting up to 2 digits after the decimal point.

- **DV200**

It is restricted text. The value syntax is '{integer} {unit}'. The unit is usually '%'.

- **staining protocol**

It need to choose from the following list:

['ssDNA staining', 'H&E Staining', 'IF Staining', 'not determined', 'not applicable', 'not collected', 'not provided', 'restricted access', 'missing']

For more detailed explanation for the fields, please refer to the standard below:

---

**Important:**

- The number of the Tissue Section filled in the template cannot be greater than 100, and the template file cannot be greater than 10MB.
  - `tissue section alias` cannot be modified. `tissue section ID` can be modified because it is for publication.
  - If need to modify Tissue Section, the assigned tissue section accession numbers can not be modified.
- 

## 8.3 Data files preparation

### 8.3.1 File name restrictions

---

**Important:**

- File names should NOT include any sensitive information (these will appear publicly).
  - File names should be unique (DO NOT upload subdirectories containing identically-named files).
  - Avoid whitespace and special characters in file names. Use only alphanumerals [A-Z, a-z, 0-9], underscores [`_`] and dots [`.`].
- 

### 8.3.2 File upload options

There are three ways to upload your data files, the login details can be find [here](#).

#### File Transfer Protocol (FTP)

---

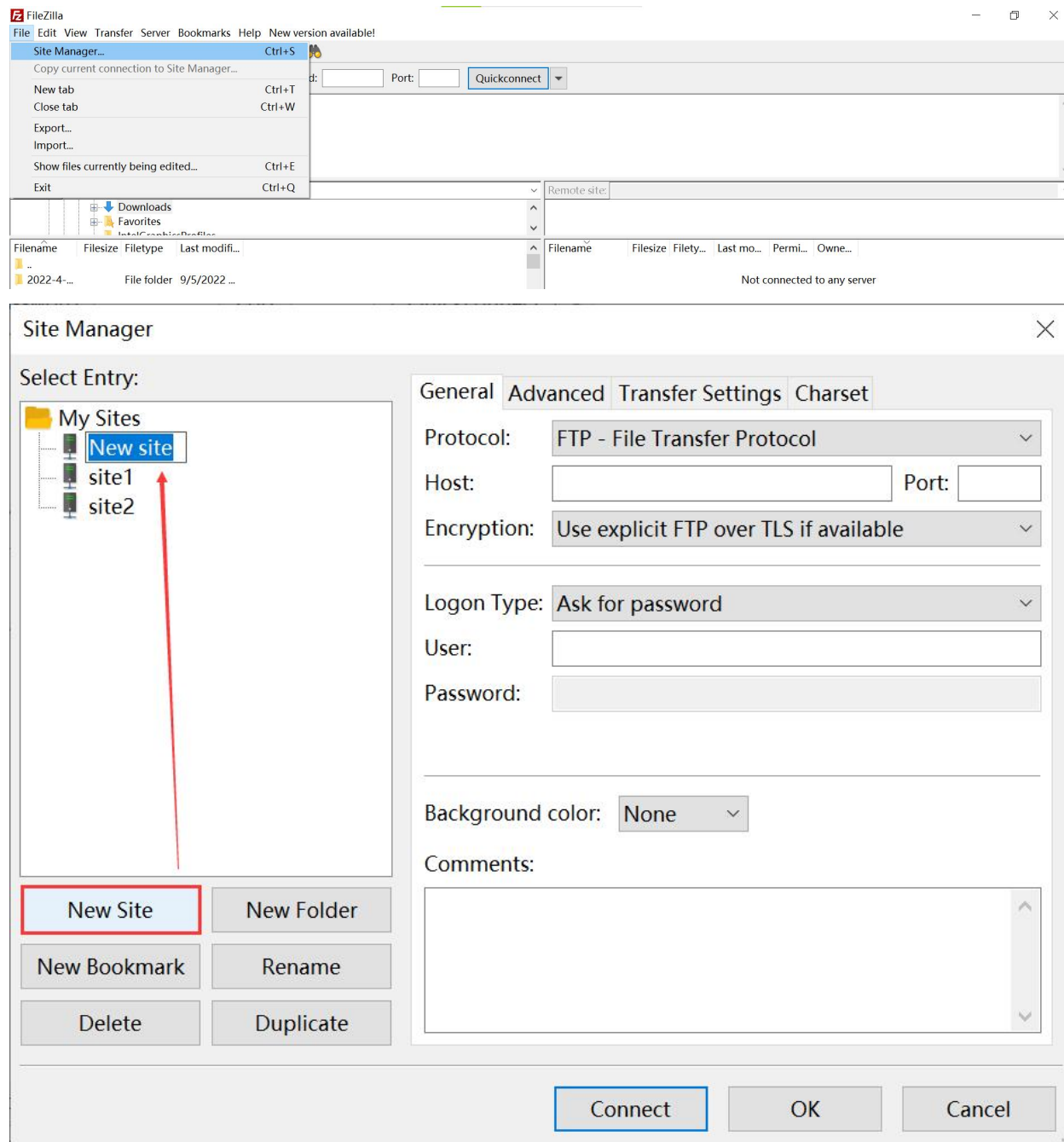
**Tip:**

- Please use passive & binary modes when transferring files.
  - The FTP server is a temporary storage space. Files will be moved to an internal location for archive and assigning of accessions.
  - Files deposited on the FTP site are not displayed under 'My Submissions' on the web interface. The web interface only displays accessioned submissions.
  - You must submit the metadata on the web interface. If no metadata is submitted within two months, the data files will be automatically deleted.
-

## Using third party FTP clients

Many reliable **FTP clients** can be found on Internet. For example, [Filezilla](#). Please refer to its documentation for usage instructions and troubleshooting tips.

1. Open Filezilla after installation. Register a new site, and rename it, such as “CNGB-ftp”.



2. Some configuration is required before use.

Site Manager ✕

Select Entry:

- My Sites
  - CNGB-ftp
    - site1
    - site2

**General** Advanced Transfer Settings Charset

Protocol: 1 FTP - File Transfer Protocol

Host: 2 ftp.cngb.org Port: 21

Encrypti 3 Only use plain FTP (insecure)

Logon Type: Ask for password

User: 4 ngb\_

Password:

Background color: None

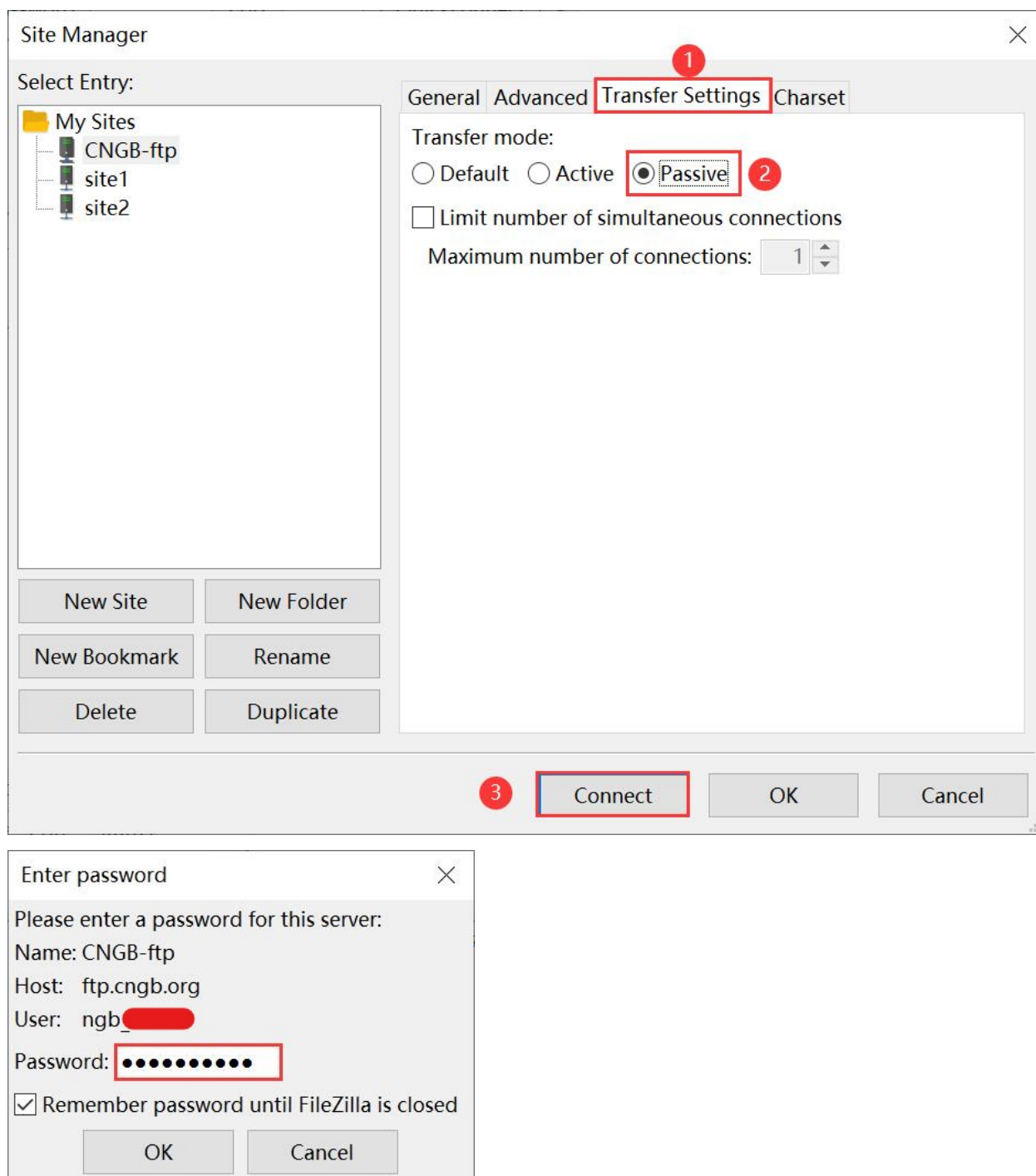
Comments:

New Site New Folder

New Bookmark Rename

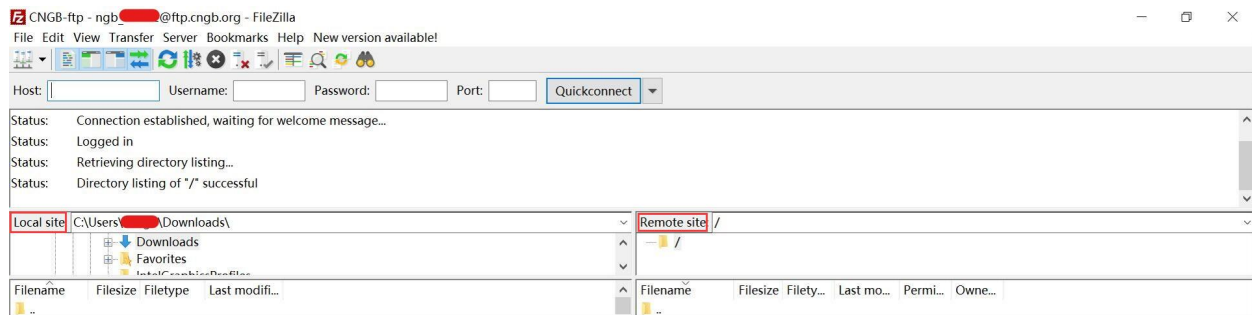
Delete Duplicate

Connect OK Cancel

**Important:**

- The host (ftp.cngb.org), username (ngb\_\*\*\*\*\*), password can be found at <https://db.cngb.org/mycngbdb/services>, which is assigned by default.

3. After successful connection, You can now transfer files by dragging your folder containing all submission files from the 'Local site' window and dropping into your personalized upload space ('Remote site' window).



The use of Filezilla in **Windows** and **Mac OS** is similar, and you can refer to the above steps.

## Using FTP command to transfer files

**FTP command** can be executed in **Linux/Unix, Mac OS Terminal**.

```
# Establish FTP connection
ftp ftp.cngb.org

# Go to the local directory containing your submission files
lcd local_path_to_your_files

# Use the put command to place one file (or mput for multiple files) into the FTP
↪ directory
put file_name
mput *
```

other commands you may use:

```
ls # to list the names of the files in the current remote directory

mkdir # to make a new directory within the current remote directory

cd # to change directory on the remote machine

pwd # to find out the pathname of the current directory on the remote machine

rmdir # to remove (delete) a directory in the current remote directory

delete # to delete (remove) a file in the current remote directory (same as 'rm' in UNIX)

quit # to exit the FTP environment (same as 'bye')
```

## Aspera Command Line

You may use the following command to upload files via Aspera Command-Line:

```
ascp -i <path/to/key_file> -P33001 -QT -l100m -k1 -d <path/to/folder/containing_files>
↪ aspera_*****@183.239.175.39: /
```

where:

- <path/to/key\_file> must be an absolute path, e.g.: /home/keys/aspera.openssh
- <path/to/folder/containing\_files> needs to specify the local folder that contains all of the files to upload.

Get the [key file](#). Please do not share this key file. Do not include this information or key file on a public page.

If you upload data files and do not submit them on the web interface, they will be automatically deleted two months according to the database record.

**Stay tuned for more useful upload functions!**

- Computer Cluster (Exclusively for BGI employees)

## 8.3.3 MD5 Checksum

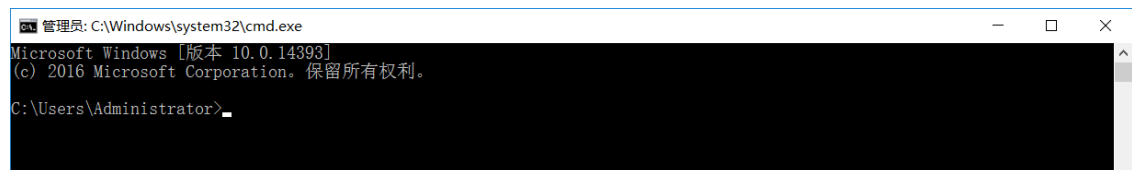
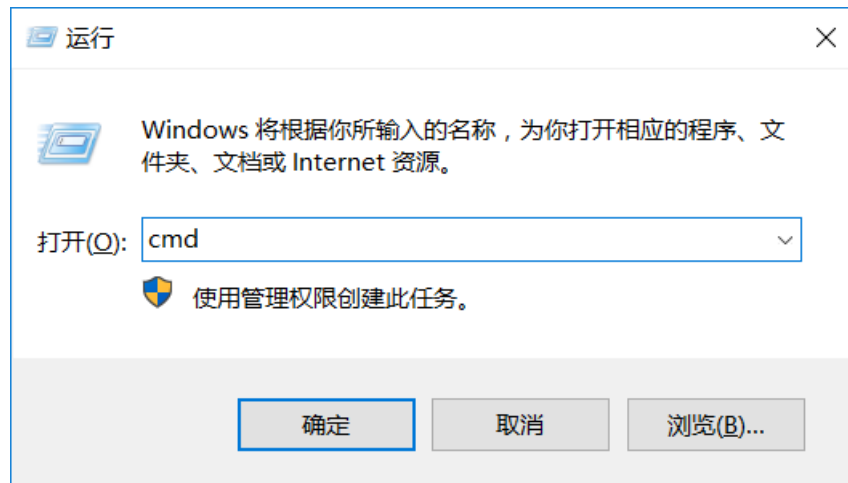
### MD5 Checksum

This is a 32-character alphanumeric string (e.g. 9F6E6800CFAE7749EB6C486619254B9C) that can be computed for each file with native command line tools md5 (**Mac OS X**) or md5sum (**Linux**).

For **Windows** users, there are several ways.

#### 1. Using command line program for Windows

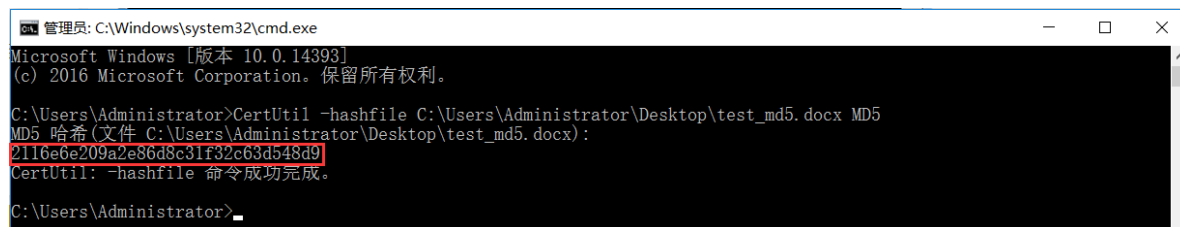
- Press the Windows icon + R, the following interface appears, enter cmd to open the program.



- Enter the following command to calculate the MD5 value:

```
CertUtil -hashfile Path\filename MD5
```

For example,



```
管理员: C:\Windows\system32\cmd.exe
Microsoft Windows [版本 10.0.14393]
(c) 2016 Microsoft Corporation。保留所有权利。

C:\Users\Administrator>CertUtil -hashfile C:\Users\Administrator\Desktop\test_md5.docx MD5
MD5 哈希(文件 C:\Users\Administrator\Desktop\test_md5.docx):
2116E6E209A2E86D8C31F32C63D548D9
CertUtil: -hashfile 命令成功完成。

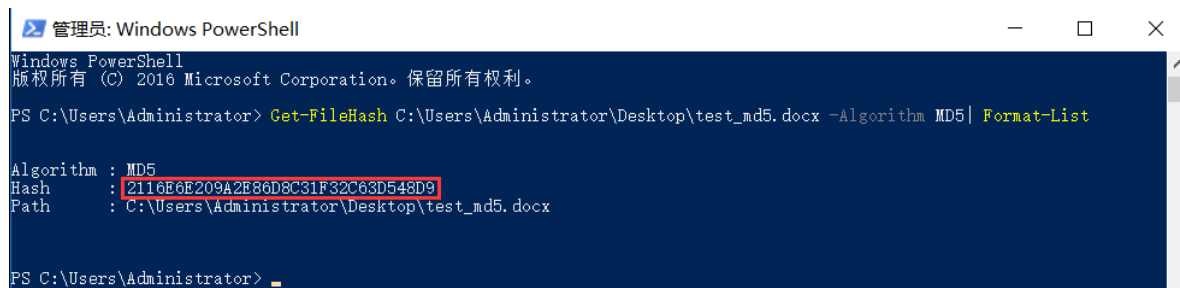
C:\Users\Administrator>
```

## 2. Using Windows PowerShell

Open Windows PowerShell, enter the following command to calculate the MD5 value:

```
Get-FileHash Path\filename -Algorithm MD5 | Format-List
```

For example,



```
管理员: Windows PowerShell
Windows PowerShell
版权所有 (C) 2016 Microsoft Corporation。保留所有权利。

PS C:\Users\Administrator> Get-FileHash C:\Users\Administrator\Desktop\test_md5.docx -Algorithm MD5 | Format-List

Algorithm : MD5
Hash      : 2116E6E209A2E86D8C31F32C63D548D9
Path      : C:\Users\Administrator\Desktop\test_md5.docx

PS C:\Users\Administrator>
```

## 3. Using the third party tools, e.g. Fsum Frontend.

# 8.4 Experiment and Run

The sequencing reads can be submitted in [the STOmics submission portal](#).

The template consists of four parts:

1. **Metadata**, describing associated tissue section, spatial slides, library and sequencing information.

Metadata is required and cannot be left blank.

tissue section alias: multiple values are supported, separated by commas.

spatial slide: multiple values are supported, separated by commas.

library name is unique for each library.

Fixed value or drop-down options have been given for some fields.

## 2. Fastq data files

.fastq.gz, .fastq.bz2, .fq.gz, .fq.bz2 are accept for fastq format data files.

MD5 values for these files should be filled in the template.

The fastq format is the most commonly submitted.

### 3. Aligned data files

The file name needs to be suffixed with `.bam`. The md5 value is also required.

### 4. Reference data files

They are not mandatory. Sequence and annotation are supported if available.

There are two ways to submit them.

- The reference accession in the public repository. For example, GRCh38.p14, NCBI Homo sapiens Annotation Release 108, GENCODE 40.
- Custom data file and its md5 value.

`.fa`, `.fasta`, `.fna`, `.fna.gz`, `.fa.gz`, `.fasta.gz`, `.fna.bz2`, `.fa.bz2`, `.fasta.bz2` are accept for sequence submission. `.gff.gz`, `.gff3.gz`, `.gtf.gz` are accept for annotation data submission.

For more detailed explanation for the fields, please refer to the standard below:

---

#### Important:

- The number of rows in the template cannot be greater than 800, and the template file cannot be greater than 10MB.
  - All file names and MD5 values cannot be repeated in the template, except reference data.
  - The data files that have been submitted cannot be submitted again, judged according to the MD5 value.
  - Both fastq data and aligned data must be submitted at least one.
  - If need to modify, the assigned accession numbers can not be modified.
- 

## 8.5 Analysis

STOmics Analysis data must be submitted in [the STOmics submission portal](#).

Two popular spatial technologies are supported to submit Spatial Transcriptomic data.

In the template, the purple fields are mandatory, the blue fields are conditional, and the orange fields are optional.

The template is mainly to fill in the name of the relevant data file, the restrictions are as follows.

### 8.5.1 Stereo-seq

**Spatial positions:** A binary file that records positions of Coordinate identity (CID) on the Stereo chip. Stereo chip mask, suffixed with `.h5`, `.bin`.

**Matrices:** `.gem`, `.gef`, `.gem.gz`, `.tsv`, `.tsv.gz`, `.txt`, `.txt.gz` are accepted for matrix (raw feature-spot matrix, filtered feature-spot matrix) submission. The filtered feature-spot matrix should be provided and its bin size (`bin size (matrix)`) is also needs to be provided.

**Annotation:** Define each cell population according to the marker gene, cell morphology, etc. `.csv`, `.txt`, `.tsv`, `.csv.gz`, `.txt.gz`, `.tsv.gz` are accepted.

There are two types of annotation files.

- Bin. It is mandatory, and bin size also needs to be provided. (`cell annotation, bin size (annotation)`)
- Cell bin. It is conditional. It can be left blank or fill in “not applicable”. (`cell annotation: cell bin`)

**Images:** images taken by microscope (microscope slide image) and its corrected images (registered image). .jpg, .jpeg, .png, .tiff, .tif, .tiff.gz, .tif.gz are accepted. They are optional. Both are required if provided. And “not applicable” is accepted for no information.

**Report:** It is optional. The file name needs to be suffixed with .html.

**MD5 list:** It should be provided and must be named with your submission ID, for example, sts\*\*\*\*\*.md5.list. MD5 values of all files listed in the template should be provided in this file. The file has two columns, **file name** and **MD5 value** in order, separated by spaces or tabs.

## 8.5.2 Visium Spatial Gene Expression

**Spatial positions:** .csv, .csv.gz are accepted, for example, tissue\_positions\_list.csv.

**Matrices:** tar.gz, .tar.bz2, .h5 are accepted for matrix (raw feature-barcode matrices, filtered feature-barcode matrices) submission. The filtered feature-barcode matrices should be provided.

**Annotation:** Define each cell population according to the marker gene, cell morphology, etc. .csv, .txt, .tsv, .csv.gz, .txt.gz, .tsv.gz are accepted.

**scale factors:** It supported json file, for example, scalefactors\_json.json.

**Images:** There are two types of images: high resolution tissue image and low resolution tissue image. The latter is mandatory.

**Report:** It is optional. The file name needs to be suffixed with .html or .csv.

**MD5 list:** It should be provided and must be named with your submission ID, for example, sts\*\*\*\*\*.md5.list. MD5 values of all files listed in the template should be provided in this file. The file has two columns, **file name** and **MD5 value** in order, separated by spaces or tabs.

## 8.5.3 Relevant instructions

For more detailed explanation for the fields, please refer to the standard below:

---

**Important:**

- The number of rows in the template cannot be greater than 100, and the template file cannot be greater than 10MB.
  - All file names cannot be repeated in the **Stereo-seq** template, expect Stereo chip mask, summary report and MD5.list.
  - All file names cannot be repeated in the **Visium spatial** template, expect summary report and MD5.list.
  - In the MD5 list, MD5 values must be unique, and file name should not be repeated.
  - The data files that have been submitted cannot be submitted again, judged according to the MD5 value.
  - Each row in the **Stereo-seq** template represents a dataset with a unique combination of *tissue section alias + Stereo chip mask + filtered feature-spot matrix + bin size (matrix)*.
  - Each row in the **Visium spatial** template represents a dataset with a unique combination of *tissue section alias + tissue position + filtered feature-barcode matrices*.
-

## 8.6 Other

Other data types not listed above can be submitted in **Other** template. Scripts are also accepted.

The file name, file type, MD5 value, and its description should to be listed in the template.

---

### Important:

- The number of rows in the template cannot be greater than 100, and the template file cannot be greater than 10MB.
  - All file names and MD5 values cannot be repeated in the template.
  - The data files that have been submitted cannot be submitted again, judged according to the MD5 value.
- 

## 8.7 Character limitation

These templates are restricted to be filled in English. The following special characters can be used.

- The Greek alphabet

Letter	Symbol	Letter	Symbol	Letter	Symbol
alpha		iota		rho	
beta		kappa		sigma	
gamma		lambda		tau	
delta		mu		upsilon	
epsilon		nu		phi	
zeta		xi		chi	
eta		omicron		psi	
theta		pi		omega	

- Special characters

Temperature symbol: °

Plus/minus sign: ±

Multiplication sign: ×