# Submitting Sequencing Data v3

This page provides instructions for adding sequencing data (mRNA-Seq, scRNA-Seq, etc.) to the GUDMAP/RBK Data Explorer.

If you have any questions or feedback, please send them to your Consortium's help email: help@gudmap.org or help@rebuildingakidney.org.

## Overview of the sequencing data submission process

The following are the basic steps you'll take to add metadata records and upload files for sequencing data.

Note that we only require as few fields as possible but have many optional fields to describe data. The more information you provide, the more discoverable is your data; however, it's best to focus on the fields that are most applicable to your data and can help a user reproduce your experiment.

The following are the minimum requirements:

1. Join (if you haven't already) the `kidney-writers` access group.
2. Organize your data by Study, Experiments and Replicates
3. Add Specimen records. You will link them to your replicates.
4. Create metadata records associated with sequencing assays

- 4.1. Create Study records
- 4.2. Create Experiment records
- 4.3. Create Replicate records
- 4.4. Notes about Record Status and Curation Status

5. Upload sequencing and analysis files

- 5.1. Review the supported file extensions
- 5.2. Upload files through the browser OR
- 5.3. Bulk upload through our client tools

6. Review internally and then submit to the Hub when you're finished via the *Curation Status* field.

- 6.1. Data Submission Dashboard

These are some other training materials and other documentation you may find useful:

- Tutorial slides containing screenshots (Sequencing Model V3---latest)
- Advanced Editing Features
- Tips for Submitting Large or Complex Data

Here are the details for each step:

## 1. Join the `kidney-writers` group

- Join the kidney-writer group.

- When you click the kidney-writer group link, if you have never used Globus before, you will be given various choices for logging in: via existing credentials (your institution, Google, or ORCID ID) or by creating a new Globus ID. We recommend using an existing credential if that is available. If you decide to use a login with an email different from the one we invited you with - Globus may ask you to link the two emails/accounts.
- For detailed instruction on how to join different GUDMAP/RBK group, visit Accessing GUDMAP and RBK Resources. If you have *any* problems, please email rbk-ops@rebuildingakidney.org.

Make sure you are logged in before attempting any of the data submissions steps.

## 2. Organize your data by Study, Experiments and Replicates

Our system uses a modular organization of metadata records that link to each other and other assay types to allow for complex searching, filtering and discovery of data.

The following describes how the metadata records for sequencing data are organized:

- A Study record is the top-level "base" record for sequencing data that describes high-level objectives and overall design of the experiments. Study-level analysis files may also be associated with it. A Study contains one or more Experiments.
  - An Experiment record is a group of one or more Replicates with exactly the same experiment methods. Related records include antibodies, custom metadata (optional), and experiment settings.
    - A Replicate contains information about bio-samples as well as their biological and technical replicate numbers. Replicates include:
      - All of the replicate-specific experimental assays, such as sequencing and analysis files.
      - A link to a Specimen record.
      - For single cell RNA-Seq, you may also add a Single Cell Metrics record summarizing the statistics of a replicate.

You can view a complete sequencing metadata model here.

## 3. Add Specimen records

For each slide of tissue used in your sequencing experiments, you will need to link it to a corresponding Specimen record. Create your Specimen records now and note the RID numbers to make it easier to link them to your Replicates later. If the Specimen records are already created, just make a note of their RID numbers. You can find the full documentation for submitting Specimen data here, but here's an overview:

1. Create one or more Specimen records: use this link to go directly to the create form: For GUDMAP | For RBK. If creating multiple Specimen records, click the plus (+) button in the upper right to add more forms. Fill in the fields that make sense for your experiment. After you fill in the form(s), click *Submit*. A new Specimen record is created.
   - You can also go to any Specimen page and click `Create` in the record header (*Search* > *Sequencing Data* > *Specimen*).

2. Add an Anatomical Source: This is a metadata field on the Specimen record that indicates the anatomical region appearing on the slide as a whole. Click "Anatomical Source" in the right Content sidebar to scroll to it, then click `Add` to the right of the field and choose the region.
3. Link additional records: As applicable, scroll down the page or the right Contents sidebar to find the following tables and click `Add`.
   - Specimen Allele: Click the "Allele RID" field to search for and select an existing allele in the system. If it doesn't exist, click the plus sign (+) to add the allele.

- Mouse Strains Contributing to Specimen: Search and select the relevant mouse strains. If it doesn't exist, click the plus sign (+) to add the strain information.

If you need to add the same Anatomical Source to multiple Specimen records:

1. Go to *Search* > *Anatomy Terms* > *Faceted Search* in the menu.
2. Search or filter to find the anatomical region and click its eye icon.
3. Click `Specimens` in the right Content sidebar or scroll down to that table, then click `Add` to the right of the table.
4. In this window, you can filter or search for your Specimens, check their boxes, click `Submit` and now they will be associated with this Anatomical Term for their Anatomical Source.

# 4. Create metadata records associated with sequencing assays

## 4.1. Create Study records

1. Create one or more Study records: use this link to go directly to the create form: For GUDMAP | For RBK. If creating multiple Study records, click the plus (+) button in the upper right to add more forms.
   - You can also go to any Sequencing Study page and click `Create` in the record header (*Search* > *Gene Expression Data* > *Sequencing Data* > *Studies*).

2. Fill in the fields on the form. Mandatory fields are indicated by an asterisk (*), including "Summary" and "Overall Design" of the study. When you are finished, click `Submit`.
3. On the new Study record, scroll to the "Study Analysis File" table and click `Add` to add analysis files associated with your study.

## 4.2. Create Experiment records

1. On the Study record, click Experiments in the Content sidebar or scroll down to the "Experiments" table.
2. Click `Add` on upper right corner of the table and fill in the fields to adequately describe your experiment, then click `Submit`. If you are creating multiple experiments, click the plus (+) button at the upper right corner to add more forms that may be submitted simultaneously.
3. Link additional records. From the Experiment record the minimum required additional records are:
   - Experiment Settings. Click `Add` on top of the Experiment Settings table to enter the details related to the data processing.
   - Antibodies. Click `Add` on top of the `Antibody` table to link to the antibodies used for cell isolation in the experiment.

## 4.3. Create Replicate records

1. While still on the Experiment record, go to the Replicate table and click `Add` to fill out the fields relevant to your replicate. If you are creating multiple replicates, click the plus (+) button at the upper right corner to add more forms that may be submitted simultaneously (see Multi-create).
2. Click the field for Specimen RID to link to one of the Specimens you already created for the bio-sample. This will pop up a window where you can search for the specimen (you can type the RID number into the search field) and select it. If you did not create one yet, click the plus sign (+) at the upper right corner to create a new Specimen record.
3. Enter Biological and Technical Replicate Numbers. For example, the first replicate of an experiment will use the number `1` for both Biological and Technical Replicate numbers.

4. The other fields are optional, but fill out the ones that make sense for the replicate and can help others reproduce the experiment.
5. When you are finished filling out the fields, click `Submit`. Your new Replicate record is created.
6. Link additional records to the Replicate:
   - Files: If you are only adding a few data files, scroll down to the Files table and click `Add`. To upload multiple files at the same time, you can click the plus sign (+) to add more forms. If you are uploading many or very large files, we recommend using the bulk upload method described below.
   - Single Cell Metrics: For single cell RNASeq data, scroll down to the Single Cell Metrics table and click `Add` to include this information.

## 4.4. Notes about Record Status and Curation Status

- If your record is incomplete, the `Record Status` field will list which information is missing.
- By default, new records will have a `Curation Status` set to "In preparation" (draft mode). The public will not be able to see the records until this status is set to "Release" by the Hub.
- Once your records are ready to submit, change the `Curation Status` to "Submitted". All records that are "Complete" and "Submitted" will be routed to the bio-curator for final review and release. Until the `Curation Status` is marked as "Release", the data will not be visible to the public.
- All records (e.g. Replicates, Specimen, Files) associated with an experiment will have the same `Curation Status` as their Experiment. Once the Experiment is released, all those related records will also be released.
- Data submitters can control the curation status of a Study and Experiments within a Study; a Study that is released can have both Experiments that are released and ones that are not.

# 5. Upload sequencing and analysis files

According to the Consortium, the `.fastq` and `.bam` files are mandatory. However, data submitters are encouraged to also upload corresponding analysis files if they are available.

## 5.1. Review supported file extensions

Please follow the file extension convention below:

| Extension | File Type | Description (will appear in file caption) | mandatory |
|---|---|---|---|
| R1.fastq.gz | FastQ | F reads | mandatory |
| R2.fastq.gz | FastQ | R reads | mandatory for Paired-End |
| bam | bam | alignment | mandatory |
| bed | bed | positive regions | optional |
| bw | bigWig | visualization track | optional |
| rpkm.txt | txt | expression value | optional |

| | | | |
|---|---|---|---|
| tpm.txt | txt | expression value | optional |

There are 2 different ways to upload files to our data repository: uploading files through the browser or bulk uploading through our client tools.

## 5.2. Upload files through the browser

This method was already covered in the above sections but here it is again:

- Study-level files: On the Study record, click `Add` to the upper right of the Study Analysis File table to add analysis files associated with your study.
- Replicate-level files: On the Replicate record, click `Add` to the upper right of the File table to add sequencing and analysis files associated with a specific replicate.
  - Normally, users will need to upload the actual data files to the Hub. For sequencing files that are archived in other permanent repositories (e.g. GEO, dbGaP), you also have the option to enter a URL to get to the archive.
  - For human-protected sequencing file stored in dbGaP, please provide a `dbGaP Accession ID`.

This approach is convenient if you only have a few replicates over all. But if you have more than 10, or your files are larger than 5GB, consider the bulk upload method (see the next section).

## 5.3. Bulk upload through our client tools

This approach is recommended when there are many and/or very large replicate-level files (e.g. bigger than 5 GB) to upload.

You will need to:

1. Install the client tools on your system. There are pre-packaged installers for MacOS and Windows available as well as an installer via pip for Linux desktops or remote servers.
2. Organize your directory structure and use a naming convention for your files (You will need to note the Internal IDs you used in your Studies and Experiments).
3. Run the tool for your files to set up a submission job for the files.

This is a very secure and stable service, but if the job is interrupted, the program will be able to retry submission until success. If the job still fails, you can re run the program and it will be able to tell which files were already successfully uploaded and skip them.

If your directory structure and naming convention are correct, the files will be automatically attached to the correct Replicate records.

For full instructions, go to Bulk Upload with DERIVA Client Tools.

(Note that DERIVA is the underlying software of the Data Browser.)

# 6. Review internally and then submit to the Hub

The *Curation Status* field controls the visibility of your record and describes where it is in the Curation Workflow.

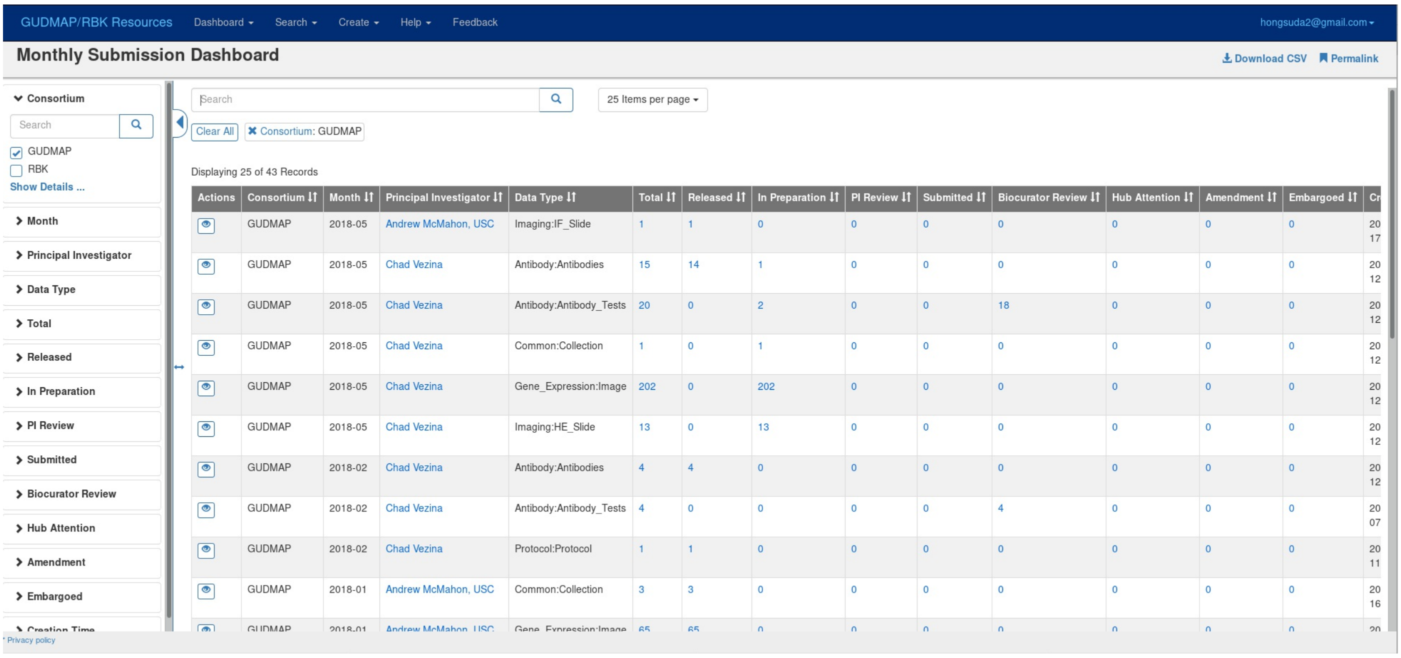
By default, new records have a *Curation Status* of "In Preparation" (ie, draft mode). There is also a status of "PI Review" you may use internally if your lab wants final sign off by your PI or other designated reviewer.

Once you are finished with your submission, make sure the *Curation Status* of the Study record and Experiment records is set to `Submitted`. Generally, this will go into Bio-curator Review. The bio-curator may ask you for further information to ensure the quality of the submission.

When the Hub is satisfied all requirements are met, they will change the status to "Release" and at this point the records are now available to the public.

## 6.1. Data Submission Dashboard

The monthly data submission dashboard is available on the GUDMAP/RBK data browser.



Monthly Submission Dashboard

Displaying 25 of 43 Records

| Actions | Consortium | Month | Principal Investigator | Data Type | Total | Released | In Preparation | PI Review | Submitted | Biocurator Review | Hub Attention | Amendment | Embargoed | Cr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | GUDMAP | 2018-05 | Andrew McMahon, USC | Imaging:IF_Slide | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20<br>17 |
| | GUDMAP | 2018-05 | Chad Vezina | Antibody:Antibodies | 15 | 14 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 20<br>12 |
| | GUDMAP | 2018-05 | Chad Vezina | Antibody:Antibody_Tests | 20 | 0 | 2 | 0 | 0 | 18 | 0 | 0 | 0 | 20<br>12 |
| | GUDMAP | 2018-05 | Chad Vezina | Common:Collection | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 20<br>12 |
| | GUDMAP | 2018-05 | Chad Vezina | Gene_Expression:Image | 202 | 0 | 202 | 0 | 0 | 0 | 0 | 0 | 0 | 20<br>12 |
| | GUDMAP | 2018-05 | Chad Vezina | Imaging:HE_Slide | 13 | 0 | 13 | 0 | 0 | 0 | 0 | 0 | 0 | 20<br>12 |
| | GUDMAP | 2018-02 | Chad Vezina | Antibody:Antibodies | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20<br>12 |
| | GUDMAP | 2018-02 | Chad Vezina | Antibody:Antibody_Tests | 4 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 20<br>07 |
| | GUDMAP | 2018-02 | Chad Vezina | Protocol:Protocol | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20<br>11 |
| | GUDMAP | 2018-01 | Andrew McMahon, USC | Common:Collection | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20<br>16 |
| | GUDMAP | 2018-01 | Andrew McMahon, USC | Gene_Expression:Image | 65 | 65 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 |

# Frequently Asked Questions

## *Question:* 10X generates fastq files in the form of `*.R1_001.fastq.gz` (e.g. my_single_cell.R1_001.fastq.gz). How do I rename file names in bulk from `*.R1_001.fastq.gz` to `*.R1.fastq.gz`?

*Answer*: You can run the following command on the shell prompt to rename files in bulk.

```
# to change *.R1_001.fastq.gz to *.R1.fastq.gz
> for file in *.fastq.gz; do mv -v "$file" $(echo "$file" | sed
's/.R1_001.fastq.gz$/.R1.fastq.gz/'); done
```

## Question: CellRange software expects the sequencing files to be in the form of `*.R1_001.fastq.gz`. How do I change the file names from `*.R1.fastq.gz` to

`*.R1_001.fastq.gz`?

*Answer*: You can run the following command on the shell prompt to rename files in bulk.

```
# to change *.R1.fastq.gz to *.R1_001.fastq.gz
> for file in *.fastq.gz; do mv -v "$file" $(echo "$file" | sed
's/.R1.fastq.gz$/.R1_001.fastq.gz/'); done
```