Submitting Sequencing Data

Table of content:

Related Material:

- Tutorial Slides containing screenshots
- Video of Webinar from 7/11/2018 (1 hour, 5 min)

Updates

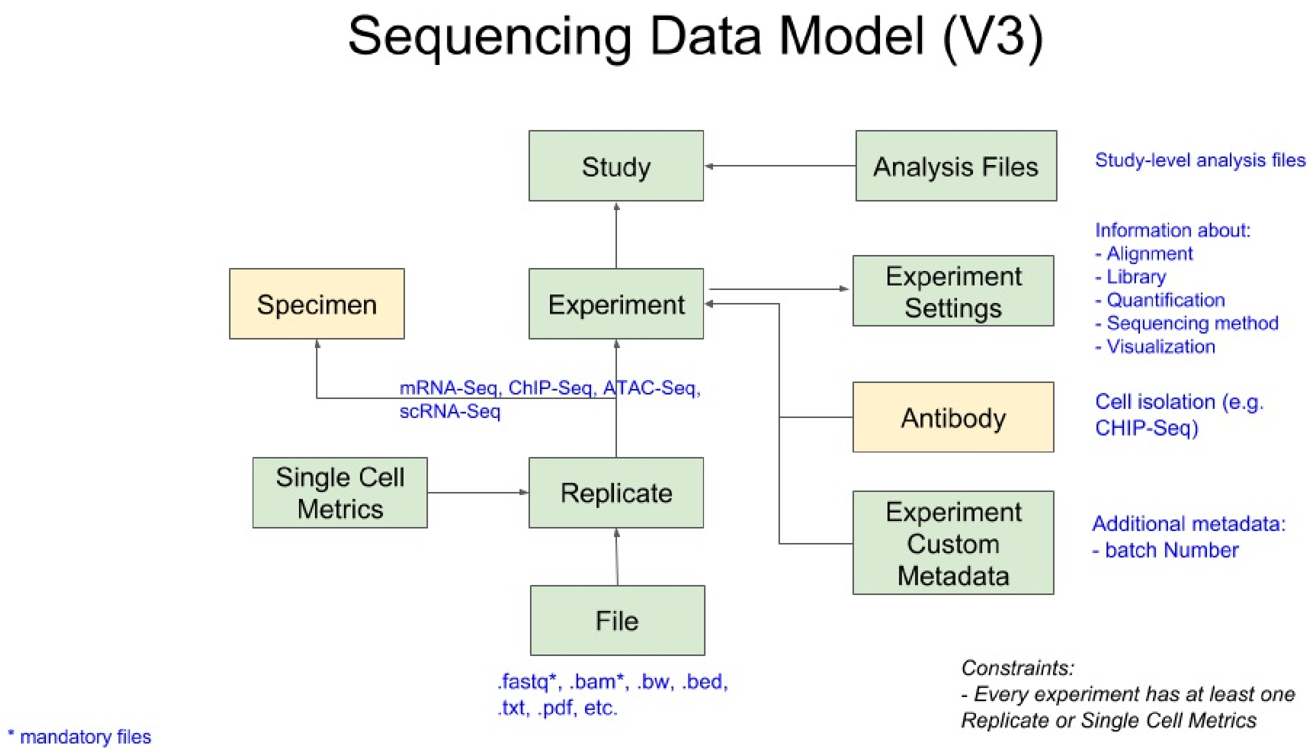
The data model has been updated on 02/01/2019 to unify the replicate and bio-sample metadata among the bulk and single cell RNASeq data. Users can use the same process to submit all sequencing type. The main difference is that single cell data will have "Single Cell Metrics" associated with a replicate. This is not the case for bulk RNASeq data.

Submitting sequencing data (RNA-Seq, ChIP-Seq, ATAC-Seq, scRNA-seq)

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.

2. Sequencing data model


Sequencing Data Model (V3)
Study
Analysis Files
Study-level analysis files
Experiment Settings
Information about:
- Alignment
- Library
- Quantification
- Sequencing method
- Visualization
Specimen
Experiment
mRNA-Seq, ChIP-Seq, ATAC-Seq, scRNA-Seq
Antibody
Cell isolation (e.g. CHIP-Seq)
Single Cell Metrics
Replicate
Experiment Custom Metadata
Additional metadata:
- batch Number
File
.fastq*, .bam*, .bw, .bed, .txt, .pdf, etc.
Constraints:
- Every experiment has at least one Replicate or Single Cell Metrics
* mandatory files

The sequencing data model was updated to V3 on 02/01/19. Deprecated data models can be found here:

- Sequencing Data Model V2

3. Create metadata associated with sequencing assays using data explorer

- Create studies. Fill in the detail in the `Study` form. Please make sure to fill in all the mandatory fields including `Summary` and `Overall Design` of the study. Hover over the field names on that left that have dotted lines underneath for description of the fields. Once all the fields are filled in, click submit to create a study.
- Create experiments. On the detail page of the study, click `Add` on top of the `Experiment` table section to add new experiments. If you don't see any table listed on the page, please click `Show All Related Records` at the upper right corner of the page. Fill in the detail of your experiment, then click submit to create an experiment. Note that multiple experiments can be created simultaneously by clicking the plus button at the upper right corner.
- Create experiment settings. On the detail page of the experiment, click `Add` on top of the `Experiment Settings` table section to add the detail related to the data processing.
- Create Antibodies. On the detail page of the experiment, click `Add` on top of the `Antibody` table to add the detail related to the antibodies used for cell isolation in the experiment.
- Create replicates. On the detail page of the experiment, click `Add` on top of the `Replicate` table to add new replicates.
    - While creating a new replicate, click the place holder for the `Specimen RID`; this will pop-up a window displaying a list of existing specimens. Select from an existing specimen or create a new one to capture the bio-sample used in the experiment. Note that the plus sign at the upper right corner of the `Specimen` pop-up window allows you to create a new specimen. Be sure to provide as much information as you can related to your bio specimen.

  - After a specimen is created, make sure that `Anatomical Sources`, `Specimen Allele` and `Specimen Mouse Strain Contributing to Specimens` are added (if applicable) by clicking `Add` on top of the corresponding tables.

- For single cell RNASeq data, create Single Cell Metrics. On the detail page of an single cell RNASeq replicate, click `Add` on top of the `Single Cell Metrics` table to add new entry.

- Notes:

  - Please make sure that all the relevant records are properly created. `Record Status` associated with your record will inform you whether all mandatory records are "Complete".
  - By default, the `Curation Status` associated with individual record is by default set to `In preparation` which allow you to keep refining or editing your records. It will not be released to the public. Once all the records are ready to submit, please change the `Curation Status` to "Submitted". All records that are "Complete" and "Submitted" will be routed to the bio-curator for final review and finally release to the public. Until the `Curation Status` is marked as "Release", the data will not be visible to the public.
  - For convenience of the data submitter, all records (e.g. Replicates, Specimen, Files) associated with an experiment will have the same `Curation Status` as the 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 experiments that are released co-existing with experiments that are not.

4. Uploading sequencing and analysis files

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

There are 2 different ways to upload files to our data repository:

1. Through the browser GUI.

- Replicate-level files: On the `Replicate` detail page, click `Add` on top of the `File` table section to add sequencing and analysis files associated with a specific replicate. Normally, users will need to upload the actual files to the Hub. For sequencing files that are archived in other permanent repositories (e.g. GEO), a URL to get to the archive can be provided. For human-protected sequencing file stored in dbGaP, please provide `dbGaP Accession ID`.
- Study-level files: On the `Study` detail page, click `Add` on top of the `Study Analysis File` table section to add new analysis files associated with your study.

2. Through DERIVA client tools. This approach is recommended in the case that there are many very large files (e.g. bigger than 5 GB) to upload. You will need to install the client tool on your system and prepare your directory structure.

4.1. Supported file extensions

Please follow the following file extension convention below. Remember the fastq and bam files are mandatory:

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

4.2. Preparing your files on disk

The upload tools will use the names of files and directories (folders) to determine what kind of files you're uploading and which data records to attach them to. We support the following conventions:

- 4.2.1. Sequencing or analysis files associated with a replicate
- 4.2.2. Study analysis files

4.2.1. Preparing replicate-level sequencing files

The layout of folders for sequencing files is:

```
$userid
  \- deriva
    \- Seq
      \- <Study Internal ID>
        \- <Experiment Internal ID>
          \- <Biological Replicate Number>_<Technical Replicate Number>_<Custom
Text>.R1.fastq.gz
          \- <Biological Replicate Number>_<Technical Replicate Number>_<Custom
Text>.R2.fastq.gz
          \- <Biological Replicate Number>_<Technical Replicate Number>_<Custom Text>.bam
          \- <Biological Replicate Number>_<Technical Replicate Number>_<Custom Text>.bed
          \- <Biological Replicate Number>_<Technical Replicate Number>_<Custom Text>.bw
          \- <Biological Replicate Number>_<Technical Replicate Number>_<Custom Text>.txt
            or
          \- <Replicate RID>_<Custom Text>.R1.fastq.gz
          \- <Replicate RID>_<Custom Text>.R2.fastq.gz
          \- <Replicate RID>_<Custom Text>.bam
          \- <Replicate_RID>_<Custom Text>.bed
          \- <Replicate_RID>_<Custom Text>.bw
          \- <Replicate RID>_<Custom Text>.txt
```

where

- `deriva` is the name of our software
- `Seq` is a subfolder of `deriva`. This indicates that everything within is the mass sequencing data (e.g. non single-

cell)

- `<Study Internal ID>` is what is specified in the study metadata e.g. `NPC_stability`
- `<Experiment Internal ID>` is what is specified in the sample metadata e.g. `mNPC_RNA`
- `<Biological Replicate Number>` is the biological replicate number associated with the replicate e.g. `1`
- `<Technical Replicate Number>` is the technical replicate number associated with the replicate e.g. `1`
- `<Replicate RID>` is the replicate RID e.g. `Q-Y500`
- `<Custom Text>` is other metadata associated with the file e.g. `sorted`
- `<File Extension>` is one of the file extensions that we current supported. *File Extension* tells the system what type of file is being uploaded.

Example 1:

If you have a study called "NPC_stability" with experiments "mNPC_RNA" and "mNPC_ATAC", you'd create two folders `deriva/Seq/NPC_stability/mNPC_RNA` and `deriva/Seq/NPC_stability/mNPC_ATAC` (on Windows, the paths would be `deriva\Seq\NPC_stability\mNPC_RNA` and `deriva\Seq\NPC_stability\mNPC_ATAC`). All the sequencing files are then placed into the respective experiment folders. In the example below, we use *Biological Replicate Number* and *Technical Replicate Number* to name the files in experiment "mNPC_RNA" and use the Replicate *RID* to name the files in the experiment "mNPC_ATAC". Both file naming conventions will be accepted by the client tool. See actual examples of metadata and files in the NPC_stability Study.

```
$userid
  \- deriva
    \- Seq
      \- NPC_stability
        \- mNPC_RNA
          \- 1_1.R1.fastq.gz
          \- 1_1.R2.fastq.gz
          \- 1_1_sorted.bed
          \- 1_1_normalized_profile.bw
          \- 1_1.kpm.txt
          \- ...
        \- mNPC_ATAC
          \- Q-Y5CC.R1.fastq.gz
          \- Q-Y5CC.R2.fastq.gz
          \- Q-Y5CC_sorted.bed
          \- Q-Y5CC_normalized_profile.bw
          \- Q-Y5CC.kpm.txt
          \- ...
```

Example 2: If you have a single-cell RNA study called "mouse_SC_RNASeq" with experiments "m1_e11_cortex" and "m2_p0_cortex", you'd create two folders `deriva/scRNASeq/mouse_SC_RNASeq/m1_e11_cortex` and `deriva/scRNASeq/m2_p0_cortex` (on Windows, the paths would be `deriva\scRNASeq\mouse_SC_RNASeq\m1_e11_cortex` and `deriva\scRNASeq\m2_p0_cortex`). All the sequencing files are then placed into the respective experiment folders. In the example below, we use *Biological Replicate Number* and *Technical Replicate Number* to name the files in experiment "m1_e11_cortex" and use the Single Cell Metrics *RID* to name the files in the experiment "m2_p0_cortex". Both file naming conventions will be accepted by the client tool. See actual examples of metadata and files in the mouse_SC_RNASeq Study.

```
$userid
  \- deriva
```

```
    \- Seq
      \- mouse_SC_RNASeq
        \- m1_e11_cortex
          \- 1_1.R1.fastq.gz
          \- 1_1.R2.fastq.gz
          \- 1_1_sorted.bed
          \- 1_1_normalized_profile.bw
          \- 1_1.kpm.txt
          \- ...
        \- m2_p0_cortex
          \- Q-Y4HM.R1.fastq.gz
          \- Q-Y4HM.R2.fastq.gz
          \- Q-Y4HM_sorted.bed
          \- Q-Y4HM_normalized_profile.bw
          \- Q-Y4HM.kpm.txt
          \- ...
```

4.2.2. Preparing study-level analysis files

If there are analysis files that are results of combining data from different experiments together, those files can be added to the `Study Analsysis File` table. Follow the following directory layout to upload those files through the Deriva client tools.

The layout of folders for analysis files to be linked at the Study level is:

```
$userid
  \- deriva
    \- Seq
      \- <Study Internal ID>
        \- <file 1>
        \- <file 2>
```

where

- `deriva` is the name of our software
- `Seq` is a subfolder of `deriva`. This indicates that everything within is sequencing data
- `<Study Internal ID>` is what is specified in the study metadata e.g. `mouse_SC_RNASeq`
- `<file 1>`, `<file 2>` are files that you want to uploaded so they can be linked to the study.

Example:

If there are analysis files that are results of combining data from different experiments together, those files can be added

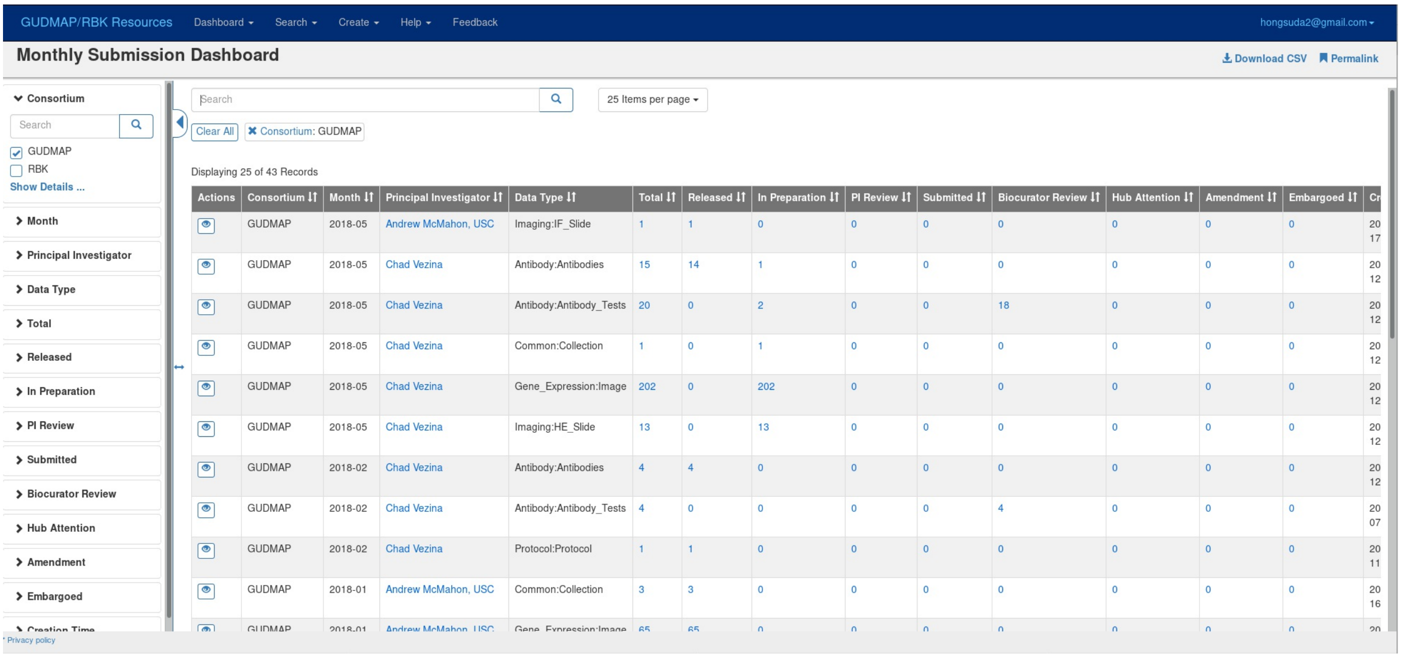
```
$userid
  \- deriva
    \- Seq
      \- mouse_SC_RNASeq
        \- analysis.xls
        \- cluster.html
```

4.3. Install and Run Deriva Client Tools

Follow the directions for Uploading files via Deriva client tools. The tools expect the file and directory naming conventions described above.

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	2017
👁	GUDMAP	2018-05	Chad Vezina	Antibody:Antibodies	15	14	1	0	0	0	0	0	0	2012
👁	GUDMAP	2018-05	Chad Vezina	Antibody:Antibody_Tests	20	0	2	0	0	18	0	0	0	2012
👁	GUDMAP	2018-05	Chad Vezina	Common:Collection	1	0	1	0	0	0	0	0	0	2012
👁	GUDMAP	2018-05	Chad Vezina	Gene_Expression:Image	202	0	202	0	0	0	0	0	0	2012
👁	GUDMAP	2018-05	Chad Vezina	Imaging:HE_Slide	13	0	13	0	0	0	0	0	0	2012
👁	GUDMAP	2018-02	Chad Vezina	Antibody:Antibodies	4	4	0	0	0	0	0	0	0	2012
👁	GUDMAP	2018-02	Chad Vezina	Antibody:Antibody_Tests	4	0	0	0	0	4	0	0	0	2007
👁	GUDMAP	2018-02	Chad Vezina	Protocol:Protocol	1	1	0	0	0	0	0	0	0	2011
👁	GUDMAP	2018-01	Andrew McMahon, USC	Common:Collection	3	3	0	0	0	0	0	0	0	2016
👁	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
```