
dolphin Documentation

Release 1.0

Alper Kucukural, Nick Merowsky, Alistair Firth

January 26, 2017

1	Dolphin-Docker	3
1.1	Quick Start Guide	3
1.2	Deploying with Docker	5
2	Dolphin-Tools	7
2.1	Quick Start Guide	7
3	Dolphin-UI	11
3.1	Dolphin Quick-start Guide	11
3.2	Excel Import Guide	12
3.3	Fastlane Guide	19
3.4	NGS Browser Guide	25
3.5	Excel Export Guide	32
3.6	NGS Pipeline Guide	33
3.7	NGS Status Guide	39
3.8	NGS Reports Guide	42
3.9	NGS Plots Guide	45
3.10	Table Creator Guide	54
3.11	Dolphin Profile Guide	59
3.12	Dolphin ENCODE Submission Guide	63
3.13	DEBrowser Access Guide	77
3.14	Developer Implementation	80
4	DEBrowser	83
4.1	Quick-start Guide	83
4.2	Local Install Guide	105
4.3	DESeq/DEBrowser	106
5	Indices and tables	111

Contents:

Dolphin-Docker

Contents:

1.1 Quick Start Guide

This guide will walk you through how to start using dolphin workflows.

!!! Warning: Dolphin docker aimed for developers. If you are going to use dolphin, please use ‘**Quick Start Guide for UI** <http://dolphin.readthedocs.io/en/master/dolphin-ui/quickstart.html>’ _ section !!!

1.1.1 General

Clone the docker first: <https://github.com/UMMS-Biocore/dolphin-docker>

1.1.2 Building dolphin-docker

First we need to build dolphin-docker image. Please go to the cloned directory.

```
cd your_path/dolphin-docker
```

```
docker build -t dolphin-docker .
```

1.1.3 Creating an export directory in your host

dolphin-docker uses a directory in your host system to hold information when you exit your docker container. If you are using boot2docker please connect your VM with

```
boot2docker ssh
```

create your export directory and give the full permissions that are going to be used by the container.

```
sudo mkdir /mnt/sda1/export
```

```
sudo chmod 777 /mnt/sda1/export
```

1.1.4 Running dolphin-docker

Dolphin docker has a apache web server that will be used on port 8080 if you run like below.

```
docker run -p 8080:80 -v /mnt/sda1/export:/export -ti dolphin-docker /bin/bash
```

or you can pull latest stable build

```
docker run -p 8080:80 -v /mnt/sda1/export:/export -ti nephantes/dolphin-docker /bin/bash
```

1.1.5 Initialize the system

You need to initialize the system using ‘startup’ command. This will prepare example genome and mysql database in /export directory that are going to be used by dolphin in the first run.

1.1.6 Starting mysql and web server

‘startup’ command will also start mysql and apache web servers. When you run dolphin-docker container. You need to start start using this command.

To reach the applications on apache server please add your docker host ip address into /etc/hosts file

Ex:

```
echo ${DOCKER_HOST}
```

```
tcp://192.168.59.103:2376
```

```
/etc/hosts =>
```

```
192.168.59.103 dolphin.umassmed.edu
```

Now you can use your browser to reach the website using

- <http://dolphin.umassmed.edu:8080/dolphin>

For phpmyadmin

- <http://dolphin.umassmed.edu:8080/phpmyadmin>

1.1.7 Running a test workflow

To run a test workflow please go to directory below;

```
cd /usr/local/share/dolphin_tools/test/
```

```
./run.bash w1.txt
```

1.1.8 Creating needed password variables

After installing dolphin-ui within docker, you are going to want to create a “.salt” file in your config folder with it’s contents similar to this:

```
[Dolphin]  
SALT=  
PEPPER=  
MASTER=  
AMAZON=
```

Make sure you fill in the values for each of these variables with your desired passphrases.

These variables are used for specific passcodes to encrypt and decrypt valuable information that you would want to protect.

1.2 Deploying with Docker

Dolphin-Tools

Contents:

2.1 Quick Start Guide

dolphin-tools

This guide will walk you through creating new pipelines in dolphin step by step.

2.1.1 Getting Started

dolphin-tools need to be located in the processing units. It can be either a HPC system or a standalone machine. All required tools need to be installed in these systems. First we are going to create a simple test pipeline. To be able to create a test pipeline and run it in the cluster we need a workflow file.

2.1.2 Installation

Please check quick start guide for Docker or for HPC please use installation guide for the appropriate cluster section.

After installing dolphin-tools, you are going to want to create a ".salt" file in your default_params folder with it's contents similar to this:

```
[Dolphin]
SALT=
PEPPER=
MASTER=
AMAZON=
```

Make sure you fill in the values for each of these variables with your desired passphrases.

These variables are used for specific passcodes to encrypt and decrypt valuable information that you would want to protect.

2.1.3 Workflow file

Workflow file is a text file and includes the directives about the workflow you want to create. This workflow file can reside either in your client machine or in your cluster. runWorkflow.py script uses this workflow file to run the steps in your target processing units. runWorkflow script is basically a client script to connect a web-api to run the scripts in a distributed environment. In addition to this, this script runs in the client system until all workflow ends successfully.

This workflow file consists of three columns. First column is a name for the step. Second column the command you want to run in your remote machine and the last column is the time interval that the client checks the run status. If the run finished successfully or killed in the cluster or your host machine, the client senses this and continue or exit accordingly. There are examples in tests/pipeline folder.

- Ex:

Here we are going to run three steps in the cluster and runWorkflow script will check the steps in every second, if they finished successfully or not.

2.1.4 Running a workflow

To run a workflow runWorkflow.py file need to be called with the parameters below.

- Usage: runWorkflow.py [options]

Options:

- h, --help** show this help message and exit
- i INPUTPARAM, --inputparam=INPUTPARAM** input parameters for the workflow
- p DEFAULTPARAM, --defaultparam=DEFAULTPARAM** defined parameter file that will be run on cluster
- u USERNAME, --username=USERNAME** defined user in the cluster
- k WKEY, --wkey=WKEY** defined key for the workflow
- w WORKFLOWFILE, --workflowfile=WORKFLOWFILE** workflow filename
- d DBHOST, --dbhost=DBHOST** dbhost name
- o OUTDIR, --outdir=OUTDIR** output directory in the cluster
- f CONFIG, --config=CONFIG** configuration parameter section

please chose your -f option according to your installation. If you are running this on Docker and made your definitions right on your Docker section right. The command should be something like below;

2.1.5 Standart output of a run in Docker

If everything is successfull you need to see an output something like below;

All the services Ended

2.1.6 The directory structure:

- For each step you want to run will be a script under OUTDIR/scripts directory.
- The standard output will be logged under tmp/lsf folder with its PID.std.

There are other log files are about communication with mySQL and LSF logs if you are running them in LSF cluster
* Intermediate submission scripts are in tmp/src folder * If there are other jobs submitted in the steps, they are going to be tracked under track folder to be able to resumed the jobs. But in this test, there is no such jobs.

```
/export/TEST
|-- scripts
|   |-- step1.bash
|   |-- step2.bash
```

```
|  `-- step3.bash
|-- tmp
|   |-- lsf
|   |   |-- 862.jobStatus.log
|   |   |-- 862.std
|   |   |-- 895.jobStatus.log
|   |   |-- 895.std
|   |   |-- 927.jobStatus.log
|   |   `-- 927.std
|   |-- src
|   |   |-- step1.submit.bash
|   |   |-- step1.tmp.bash
|   |   |-- step2.submit.bash
|   |   |-- step2.tmp.bash
|   |   |-- step3.submit.bash
|   |   `-- step3.tmp.bash
|-- track
```


Contents:

3.1 Dolphin Quick-start Guide

This guide is a quick walkthrough for setting up a Dolphin account

3.1.1 Accessing Dolphin

So you want to start using the Dolphin Web tool, but you don't have an account.

Let's go over the steps needed to be taken in order to access the Dolphin Web Service at UMass.

HPCC Access

In order to use the Dolphin Web Service at UMass, you're going to need access to our High Performance Computing Cluster.

Registration for the cluster can be found at [this address](#)

Once the HPCC Admins group receives your registration form, they will send an email to your PI requesting the PI's permission to give you access.

After it's approved you will receive an email from the HPCC Admins group with your HPCC account user name.

Joining the Dolphin Group

Once you have your Cluster account, you're going to want to email to hpcc-support@umassmed.edu to join the galaxy group.

Make sure to CC biocore@umassmed.edu

Log Into Dolphin

In order to make sure you have access to our tools and pipelines, you're going to want to log into 'dolphin.umassmed.edu' to see if you have permissions.

If you are able to log in, then you should be all set.

Dolphin Keys

In order for Dolphin to act on your behalf and use the pipeline tools we have available for you, you're going to need to run a script to give us permissions.

First, log into the cluster terminal using:

`<your_user>@ghpcc06.umassrc.org`

Where `<your_user>` is your username given to you by the cluster.

Once you're logged in, all you have to do is run this script:

```
/project/umw_biocore/bin/addKey.bash
```

This is a one time script that you'll need to run in order to have access to the Dolphin Web Service.

After using this script, it might take a few hours for our system to update before you can log into the Dolphin Web Service. If a day goes by and you're still not able to access the Dolphin Web Service, please contact us at bio-core@umassmed.edu

Project Space requirements

Once you have access to Dolphin, you're going to need some space in order to store your data/results. Make sure you coordinate with HPCC for how much project space you have or require.

You should also consult with your lab on data storage, access rights, and any other additional file information you may need.

Once you have access to Dolphin and have determined whether or not you have enough space, You'll be ready to import your data to our database and begin analyzation!

For more information about importing your data, see the Excel Import Guide.

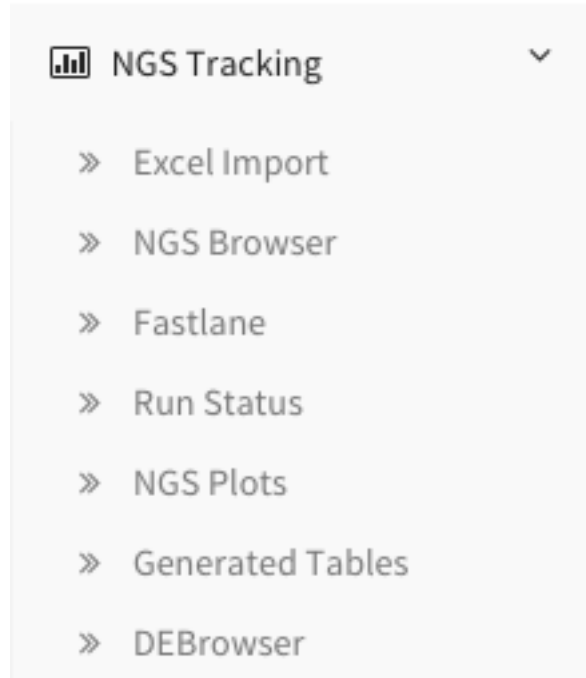
3.2 Excel Import Guide

This guide will walk you through the process of importing via the Excel Import page

3.2.1 Getting Started

First, make sure to have an instance of dolphin available (see Dolphin Docker) as well as an account for the dolphin interface.

Once logged in, click on the 'NGS Tracking' tab on the left, then click on 'Excel Import'.



From here, under the ‘Excel file input’ button, there is a link to download an example excel input. Download the example spreadsheet and you’ll be ready to start adding your data.

You can also download the example spreadsheet here:

```
single input directory
multiple input directories
```

3.2.2 Understanding the Heirarchy

Before we continue, let us discuss the way that your sample information will be stored. Dolphin stores information into 3 specific teirs.

At the top we have the Experiment Series, which essentially stores all of the information pertaining to your series of experiments.

Branching out from the experiment series, you have your set of individual experiments, or in the case, Imports.

Imports will contain all of the samples of that specific experiment as well as any other additional information that would link to your samples.

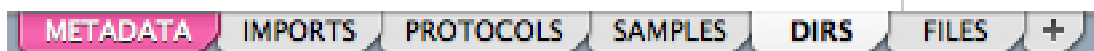
The bottom teir, or samples, are the individual samples obtained with their specific information. This information will help you better understand what each tab of the excel file will contain as we move forward.

3.2.3 Filling Out the Excel File

The example spreadsheet already contains faux information to give users some context to the information they will be replacing.

Certain header cells within the spreadsheet are marked with a red arrow on the top right of the cell. Hovering over these cells with your mouse will give you more details on what each row or column that they demark should include.

The spreadsheet contains multiple tabs that will help you organize your information properly.



If using the single input directory spreadsheet, there will not be a ‘DIRS’ tab within the spreadsheet.

META-DATA:

The first tab of the spreadsheet is the Metadata tab. This tab will reference to specific information about the experiment series at hand.

This tab will ask for information on:

- **title:** The unique title of the Experiment Series (Required)
- **summary:** Summarization of the goals/objectives of the Experiment Series
- **overall design:** Detailed description of the Experiment Series
- **organization:** The organization behind the project
- **lab:** The lab within the organization
- **contributor:** First name, Initial, Last name. You may add additional contributors by creating more ‘contributor’ cells in the A column with the actual contributors in the B column.
- **processed directory:** Full path for the processed output directory. This path is where you want to keep all of the data generated. (Required)
- **amazon bucket:** Optional amazon bucket link

This is the only tab that will ask for information in a single column (column B). The rest of the tabs will ask for information in rows.

Additionally, if using the single input directory spreadsheet, the Metadata tab will include:

- **input directory:** Full path for the input directory. This path contains the fastq files you want to process. (Required)

Note that to add additional Imports/Samples to an already existing Experiment Series, the information about the experiment must be identical to that you are adding to. Processed directory, and amazon bucket does not have to be identical for each submission.

Please note that every field except contributor and amazon bucket are required fields

IMPORTS:

The Imports tab will contain information about lanes/imports of the samples being submitted.

There can be multiple imports within this tab, each one residing on it’s own row.

Information on imports include:

- **Import Name:** The name of the import/experiment being submitted. (Required)
- **Sequencing id:** The id from the sequencing facility.
- **Sequencing facility:** Location of where sequencing took place.
- **Cost:** The cost of the sequencing.
- **Date submitted:** Date of request for sequencing.
- **Date received:** Date of sequencing results.
- **% PhiX requested:** The requested amount of PhiX in lane.
- **% PhiX in lane:** Actual amount of PhiX in lane.
- **# of Samples:** Number of samples within import.

- **Resequenced?:** Was this import resequenced?
- **Notes:** Additional notes about this import.

Please note that import name is required for submission.

PROTOCOLS:

The Protocols tab will contain information about the specific protocols used within the submission.

There can be multiple protocols within this tab, each one residing on it's own row.

Protocol information includes:

- **protocol name:** The name of the protocol. (Required)
- **growth protocol:** Protocols used to grow the organism/cells.
- **extract protocol:** Protocols used to extract/prepare the sequenced material.
- **library construction protocol:** Library construction protocol.
- **crosslinking method:** The crosslinking method if any.
- **fragmentation method:** The fragmentation method if any.
- **strand-specific:** Is this protocol strand specific?
- **library strategy:** Sequencing techniques of this library.

Please note that protocol name is required for submission.

SAMPLES:

The Samples tab will contain information about the samples within the submission.

There can be multiple samples within this tab, each one residing on it's own row.

Sample information includes:

- **Sample name:** The name of the sample. (Required)
- **Import name:** The name of the import in which the sample resides. This import must be present in the Imports tab. (Required)
- **Protocol name:** The name of the protocol in which the sample used. This protocol must be present in the Protocols tab. (Required)
- **barcode:** This samples barcode.
- **title:** Descriptive title for the sample.
- **batch id:** This samples batch id.
- **source symbol:** Symbol used for the Source. Symbol is a 4 character string.
- **source:** Brief description of cell line, biological material, or tissue.
- **organism:** List the organism from which this sample came from.
- **biosample type:** Type of biosample, ie. in vitro.
- **molecule:** Type of molecule extracted from the sample.
- **description:** Added information that pertains to other fields.
- **instrument model:** Sequencing instrument used.
- **average insert size:** Average paired-end insert size.
- **read length:** The length of the reads.

- **Genotype:** The genotype of the sample.
- **Condition Symbol:** Symbols representing the conditions from the condition column. Multiple condition symbols may be present if multiple conditions match the symbols and they are comma separated.
- **Condition:** Specific condition(s) pertaining to the sample. Multiple conditions may be present as long as they are comma separated.
- **concentration:** Concentration of Conditions.
- **treatment manufacturer:** Manufacturer of treatments.
- **Donor:** Name of sample donor, Typically in the D## format.
- **Time:** Time (in minutes) post treatment.
- **Biological Replica:** Biological replica number.
- **Technical Replica:** Technical Replica number.
- **spikeins:** Yes or No based on if spike-ins were introduced into the sample.
- **3' Adapter sequence:** 3' Adapter sequence if present.
- **Notebook reference:** Reference notebook information.
- **notes:** Any other additional notes for the sample.
- **characteristics: newtag:** Biosource characteristic.
- **characteristics: tag:** Biosource characteristic.

Please note that Sample name must be present and the Import name and Protocol name must match one provided in their respected tabs.

DIRS:

Short for directories, this tab indicated all the of directories in which your fastq data are stored. If using the single input directory spreadsheet, this tab will not be present.

There can be multiple entries on this tab.

Directory information includes:

- **Directory ID:** A specified ID to associate to files within the file tab. (Required)
- **Input directory:** Location within the cluster/host machine where the fastq files for this submission are stored. (Required)

FILES:

The Files tab will hold the files associated with either imports or samples.

There can be multiple entries on this tab, as well as multiple entries per import or sample.

File information includes:

- **Sample or Import Name (Enter same name for multiple files):** The sample or import name. These names must be within there respected tabs. (Required)
- **Directory ID:** A specified ID to associate to directories within the dirs tab. (Required)
- **file name(comma separated for paired ends):** The file fastq file name. If paired end, list both files seperated by a comma. (Required)

Please note that these fields are all required, and that if using the single directory input spreadsheet the directory id will not be present.

3.2.4 Preparing for Submission of Your Excel Spreadsheet

Once you've filled out your spreadsheet with all of your desired information, make sure to double check everything is in order. If your file seems to be filled out properly, we're ready to submit.

If you haven't yet already, head back to the excel import page.

Excel file input:

Excel file input

No file chosen

Please choose excel file from your device.

[Download example excel input](#)

Click on the 'Choose File' button to select your excel spreadsheet's path and click open. Once your file has been selected, the file name should appear to the left of the button.

Project Group:

Project Group

umw_biocore

Please select a group you belong for this project

This section will select the group in which you would like to submit this project under. The drop down menu should contain all of the groups that you are a member of.

If you cannot select a group, contact your local administrator or 'biocore@umassmed.edu'.

Who can see?:

Who can see?

- ☐ only me
- ☒ only my group
- ☐ everyone

Please select the security credentials for this import

This section will determine the security credentials of your submission. You will select from one of 3 choices:

- only me
- only my group
- everyone

Your selection will determine who can see the data you will be submitting.

3.2.5 Submission

Once you've followed these above steps, you're ready to hit the submit button.

You will be redirected to the next page where a series of checks overlook your submitted excel spreadsheet to make sure the proper information for submission was submitted.

Each tab within the spreadsheet has its own section for checks, displaying green text if the tests pass.

Yellow text is displayed with helpful information about cells that you may want to fill out, but are completely optional or able to be edited at a later point in time.

If an error occurs from improper excel spreadsheet submission, red text describing the error will appear appear in the specified section.

Submission fails:

Worksheet Information

Checking for proper excel formatting:

1. METADATA

Formatting passed inspection!

2. LANES

Some optional columns missing data, please make sure to add them later if desired

Formatting passed inspection!

3. PROTOCOLS

Some optional columns missing data, please make sure to add them later if desired

Formatting passed inspection!

4. SAMPLES

Sample name does not contain proper characters, please use alpha-numeric characters and underscores (row 4)

Some optional columns missing data, please make sure to add them later if desired

5. FILES

sample/lane name does not match the samples/lanes given (row 4)

sample/lane name does not match the samples/lanes given (row 5)

sample/lane name does not match the samples/lanes given (row 6)

sample/lane name does not match the samples/lanes given (row 7)

sample/lane name does not match the samples/lanes given (row 8)

Excel import aborted due to errors, see above

If your submission fails at least one of the checks in place, the data will not be submitted.

You can then edit your spreadsheet based on the error output text and resubmit the spreadsheet to try again.

Submission passes:

Worksheet Information

Checking for proper excel formatting:

1. METADATA

Formatting passed inspection!

2. LANES

Some optional columns missing data, please make sure to add them later if desired

Formatting passed inspection!

3. PROTOCOLS

Some optional columns missing data, please make sure to add them later if desired

Formatting passed inspection!

4. SAMPLES

Some optional columns missing data, please make sure to add them later if desired

Formatting passed inspection!

5. FILES

Formatting passed inspection!

If your submission passes, each section will contain green text saying that the section passed inspection.

A brief explanation describing what is being inserted/updated within the database is shown, followed by a helpful message that reads:

“We are currently processing your samples to obtain read counts and additional information. You can check the status of these initial runs on your NGS Status page.”

As stated by this message, the samples submitted are now in the initial processing step in which read counts and other additional information is obtained from the samples in order to run further analyses.

You can check the status of this initial run in the ‘Run Status’ page which can be found under ‘NGS Tracking’ on the left menu bar.

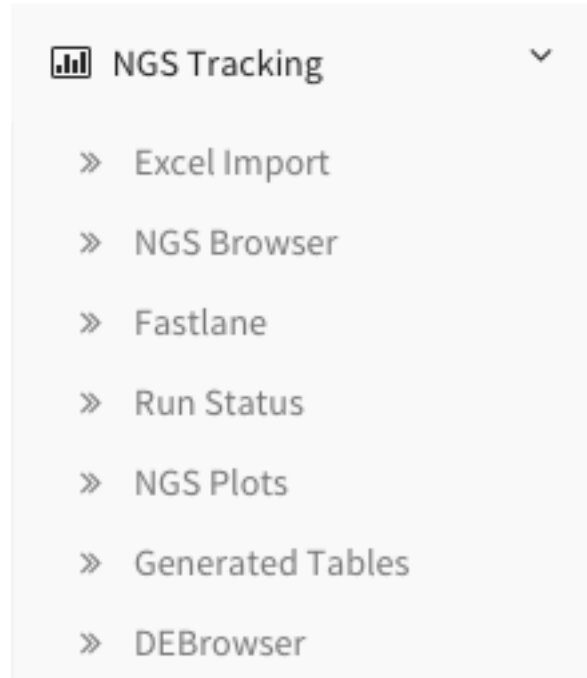
3.3 Fastlane Guide

This guide will walk you through all of your options within the Fastlane page.

3.3.1 Getting Started

First, make sure to have an instance of dolphin available (see Dolphin Docker) as well as an account for the dolphin interface.

Once logged in, click on the ‘NGS Tracking’ tab on the left, then click on ‘Fastlane’.



3.3.2 Fastlane Input

Once you've successfully made it to the Fastlane page, you can start to fill out your submission.

Down the list, from top to bottom and from left to right, the sections are as follows.

Genome Build:

Select one of the genome builds from the corresponding dropdown menu. As more genomes become available, updates will introduce more selection options.

Barcode Separation:

Select whether or not your samples need their barcodes separated or not. Upon selecting yes, 3 new sections to be filled out become visible: Barcode Definitions, Barcode Distance Options, and Barcode Format Options.

Barcode Separation ⓘ <input type="text" value="yes"/>	Mate-paired ⓘ <input type="text" value="yes"/>
Barcode Definitions ⓘ <div> lib_rep1 GATACA lib_rep2 CATATC </div>	
Barcode Distance Options ⓘ <input type="text" value="1"/>	Barcode Format Options ⓘ <input type="text" value="5 end read 1"/>

Mate-paired:

Select whether or not the samples are mate-paired or not.

Barcode Definitions:

This section will appear only if Barcode Separation is selected. This part of the form will ask you for the sample name of each sample, followed by a space and then the barcode.

IE:

Library1 ATCGATCG

AnotherLibrary GGCCTTAA

LastLibrary GCGCGTAC

Samples are to be separated by a new line. An example is shown in light grey within the box of how the sample submission should look.

Barcode Distance Options:

This section will appear only if Barcode Separation is selected. Selection of the barcode distance metric can be selected from the dropdown menu found in this section.

Barcode Format Options:

This section will appear only if Barcode Separation is selected. The dropdown menu within this section will select the whereabouts of the barcode within the reads.

Experiment Series Name:

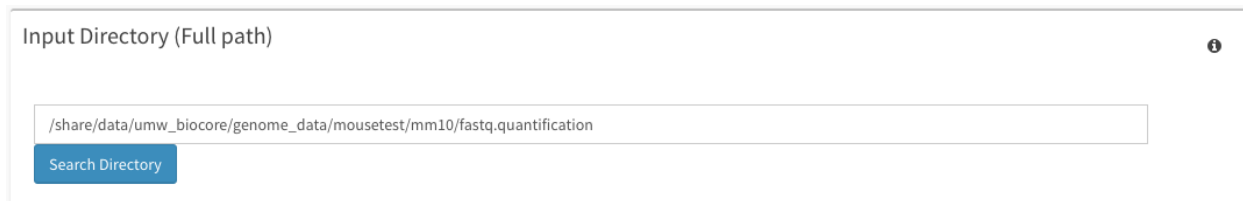
This section will determine the name of the experiment series. If adding to an already created experiment series, please be sure to input the exact name of the experiment series to add to.

Import Name:

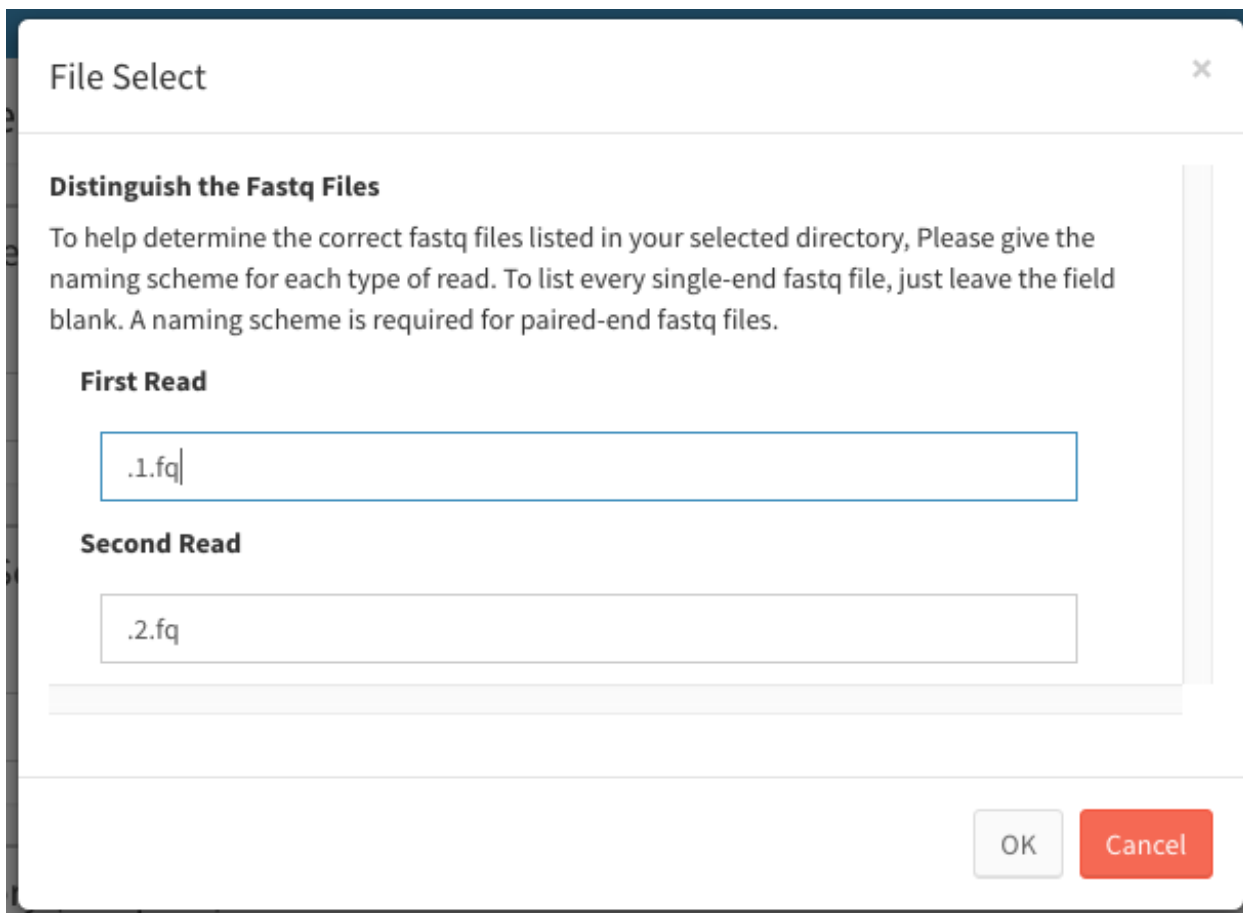
This section will determine the name of the import. If adding to an already created import, please be sure to input the exact name of the import to add to.

Input Directory:

This section will list the exact location of all of the fastq files within the cluster. If files are not all within the same directory, it's advised to either switch to importing via the Excel Import section or to move the files to a unified location. The full path of the directory must be given.

A screenshot of a web form titled 'Input Directory (Full path)'. It features a text input field containing the path '/share/data/umw_biocore/genome_data/mousetest/mm10/fastq.quantification'. Below the input field is a blue button labeled 'Search Directory'. An information icon (i) is located in the top right corner of the form's container.

Additionally, you can search this directory on the cluster for your files. Click on the 'Search Directory' button to bring up a modal which will instruct you to search for your fastq files based on your reads. You will be asked for either all reads or read 1 and read 2 based on your Mate-Paired selection.

A screenshot of a 'File Select' modal window. The title bar says 'File Select' with a close button (X). The main content area is titled 'Distinguish the Fastq Files' and contains the text: 'To help determine the correct fastq files listed in your selected directory, Please give the naming scheme for each type of read. To list every single-end fastq file, just leave the field blank. A naming scheme is required for paired-end fastq files.' Below this text are two input fields. The first is labeled 'First Read' and contains the text '.1.fq'. The second is labeled 'Second Read' and contains the text '.2.fq'. At the bottom right of the modal are two buttons: 'OK' and 'Cancel'.

Input Files:

This section is where you will put sample and file information. This information includes sample names and fastq file names. The input of this section will change based on your previous selections within the form. First, the input for files will be in the manual format.

Input Files

Manual

Directory

Paired End Example:

lane_001_R1.fastq.gz lane_001_R2.fastq

Single End Example:

lane_001.fastq.gz

Upon searching your input directory for files, the directory tab will be selected and filled with your appropriate search terms.

Input Files

Manual

Directory

Read Files

control_rep2.1.fq,control_rep2.2.fq
control_rep3.1.fq,control_rep3.2.fq
exper_rep1.1.fq,exper_rep1.2.fq
exper_rep2.1.fq,exper_rep2.2.fq
exper_rep3.1.fq,exper_rep3.2.fq

Adv Regex Select

Adv Regex Select

Reset

Add All

Add Selection

Show entries

10

Search:

Sample Name	Files Used	Remove
control_rep1	control_rep1.1.fq,control_rep1.2.fq	X

Showing 1 to 1 of 1 entries

Previous

1

Next

From here you can select one or multiple files to create a sample with. Samples are stored in a table under the 'Read Files' list and can be removed from your selection at any time by clicking the red 'X'. The 'Add Selection' button will add whatever files you have highlighted as one sample. The 'Add All' button will attempt to correctly add samples based on your previous searched selection. The 'Reset' button will reset both the Read Files and the selected table back to your original search. Lastly, the 'Adv Regex Select' will attempt to find all of the files with the select search term and pair them to one sample.

You can also change the sample names within the table at any time by selecting the text box and altering the text. It is important to note that if you have selected to carry out barcode separation, that the sample names will be given in the 'Barcode Definitions' section. Thusly, if Barcode Separation is selected, the table will only show the files being used.

Show entries

10

Search:

Files Used	Remove
control_rep1.1.gz,control_rep1.2.gz	

Showing 1 to 1 of 1 entries

Previous

1

Next

If using the manual entry for the samples and files, you can follow these guidelines for input:

No barcode separation, mate-paired:

- If your samples do not require barcode separation and are mate-paired, the input for each sample will be the sample name, first paired-end fastq file, and then the second paired-end fastq file.
- Example: sample_name_1 samplename_1_R1.fastq.gz samplename_1_R2.fastq.gz

No barcode separation, not mate-paired:

- If your samples do not require barcode separation and are not mate-paired, the input for each sample will be the sample name and then the fastq file.
- Example: sample_name_1 samplename_1.fastq.gz

Barcode separation required, mate-paired:

- If your samples require barcode separation and are mate-paired, the input for each sample will be the first paired-end fastq file followed by the second paired-end fastq file.
- Keep in mind that the files must coordinate with with the samples listed in the barcode definitions tab. Thus make sure your files and samples match each line of the two tabs.
- The first sample listed in the Barcode Definitions tab will be the first set of files listed in the Input Files tab.
- Example: samplename_1_R1.fastq.gz samplename_1_R2.fastq.gz

Barcode separation required, not mate-paired:

- If your samples require barcode separation and are not mate-paired, the input for each sample will be the first paired-end fastq file.
- Keep in mind that the files must coordinate with with the samples listed in the barcode definitions tab. Thus make sure your files and samples match each line of the two tabs.
- The first sample listed in the Barcode Definitions tab will be the first file listed in the Input Files tab.
- Example: samplename_1.fastq.gz

It's important to note that samples are to be separated by a new line. An example is shown in light grey within the box of how the sample submission should look. Examples listed within the website will change based on whether or not barcode separation is selected. Bot Paired-end and single land examples are provided within the grey.

Processed Directory:

This section will list the exact location of where the backup/results information will be stored. The full path of the directory must be given.

Amazon Bucket:

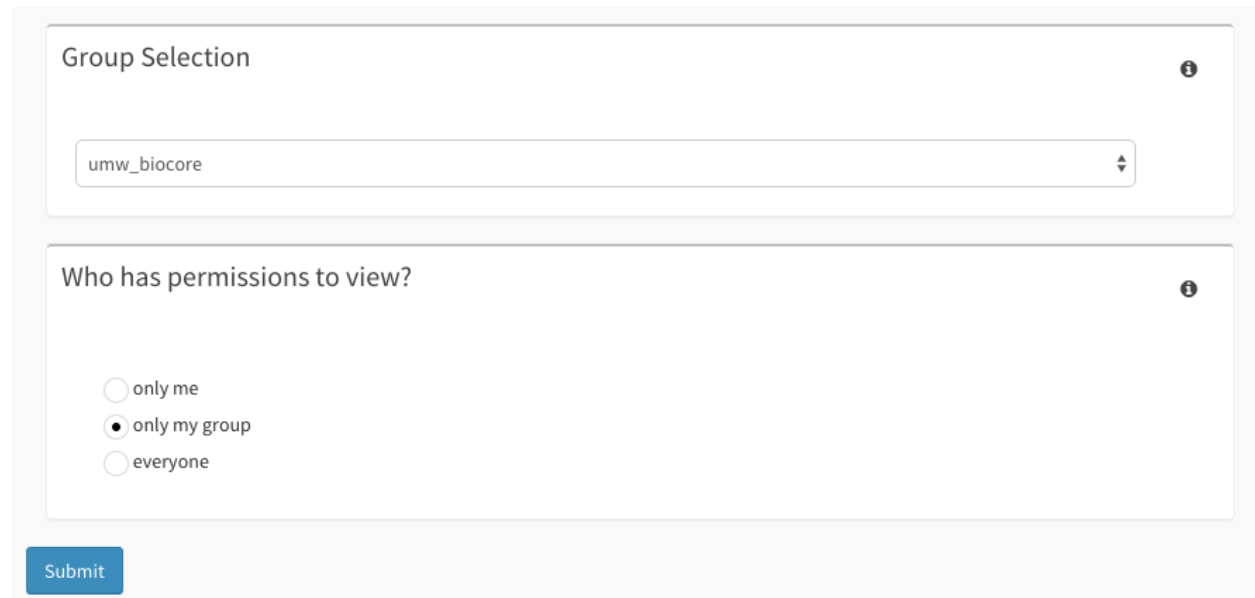
This section the amazon bucket link which you may give to have the data backed up into amazon. Filling out this section is optional.

3.3.3 Submission

Warning

Submitting to an already used processed directory will overwrite your previous run using that directory. This could potentially mess with processed data, resulting in errors and inaccurate data. The only time you should submit to an already used directory is if you are adding reads to previously imported samples. Please be careful when choosing a process directory, unique directories for each new import are suggested.

Before you get ready to submit, make sure to check which group you wish to submit under and who has permissions to view your data.



Group Selection ?

umw_biocore

Who has permissions to view? ?

☐ only me

☒ only my group

☐ everyone

Submit

Once you've filled out all the appropriate information, you're now ready to hit the submit button to start your initial run. After hitting the submit button at the bottom of the page, you will be taken to a submission page that will check your input to make sure everything in the tabs is sufficient. Sample names, whether entered in the Barcode Separation tab or within the Input Files tab, must not exist within the import you are trying to add to. If you're trying to resubmit files using fastlane, it will not allow for submission of samples under the same import with the same name. Please contact your local administrator or biocore@umassmed.edu for help with potential issues.

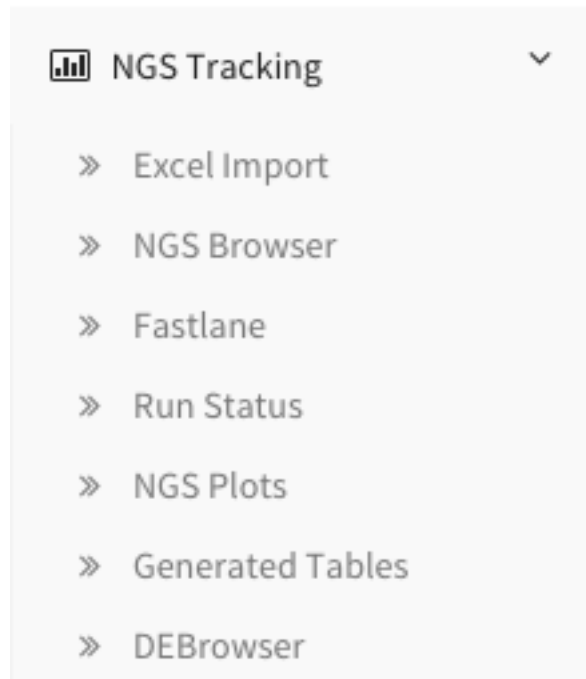
3.4 NGS Browser Guide

This guide will walk you through all of your options within the NGS Browser

3.4.1 Getting Started

First, make sure to have an instance of dolphin available (see Dolphin Docker) as well as an account for the dolphin interface.

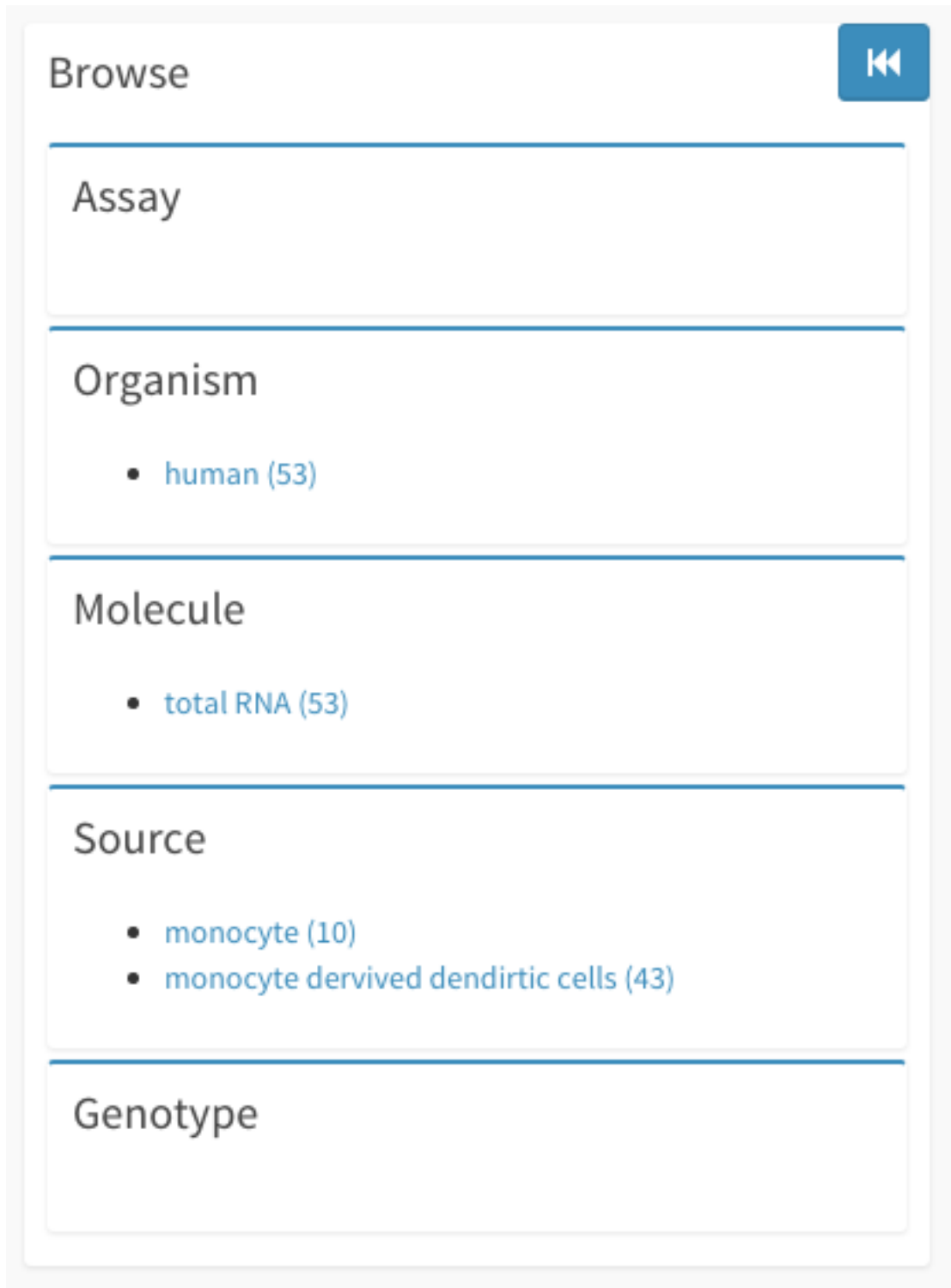
Once logged in, click on the 'NGS Tracking' tab on the left, then click on 'NGS Browser'.



3.4.2 Getting to Know the Browser

The NGS Browser can be broken down into 3 sections:

Browse Catagories:



The browse categories section is located within the top left side of the page. It is marked at the top by the 'Browse' tag.

This section of the browser lets the user sort samples shown based on specific categories that the user can select. Users can select more than one category for additional filtering.

Categories include:

- Assay
- Organism
- Molecule
- Source
- Genotype

The blue rewind button at the top right of the Browse section will bring the user back to a clean search state.

Selected Samples:

Selected Samples

Search:

ID	Sample Name	
35	control_rep1	<input type="button" value="X"/>
36	control_rep2	<input type="button" value="X"/>
37	control_rep3	<input type="button" value="X"/>
38	exper_rep1	<input type="button" value="X"/>
39	exper_rep2	<input type="button" value="X"/>

Showing 1 to 5 of 5 entries

This section of the browser is located right under the Browse category section. As the user selects more Imports/Samples, this box will fill up with the selected samples.

As the box fills with samples, users can click the red 'X' next to a specific sample to unselect that sample. In addition, as samples are selected the 'Clear' button will become active and by clicking this button will remove all of the selected samples.

Data Tables:

id	Series Name	Summary	Design	Selected
1	Barcode Sep Test	please create an experiment summary		<input type="checkbox"/>
2	New Experiment series	example summary	example overall design	<input checked="" type="checkbox"/>

The data tables are the main highlight of the NGS Browser.

Starting at the top right of the page, 3 tables followed by a series of button options will follow. These tables show information about:

- Experiment Series
- Imports
- Samples

There are some helpful tools and information at the top of each table. On the left you have the tables name, and on the right you have an expanding option which will let you see more detailed information about the contents of the specific table followed by an info button that also explains some additional information.

As you move down the table, you can then select how many entries per page you want to view on the left and you can also conduct a real-time search on the right.

Next is the actual contents of the table itself followed by page navigation buttons. If you have the proper permissions, you may edit the contents of the table by clicking on the specific cell within the table. Some fields contain standard text boxes while others will have a searchable dropdown box for you to select previous submissions for that columns category.

For the each of the tables, selection checkboxes are located on the right side of each table row. A helpful ‘!’ button is also placed near these checkboxes to indicate that the Import/Sample is not currently ready for use due to the processing step.

Within the Samples and the Imports table, the column labeled ‘Backup’ lets you know the status of your fastq file backup to AWS. Grey buttons mean there is no backup checksum to compare to your fastq file. Red buttons means that the fastq checksum and the AWS checksum do not match. A blue button lets the user know that the last modification to this upload was over 2 months ago, and the green button lets the user know that file checksum matches the AWS checksum.

Selection Details:

Summary	Groups	Permission
please create an experiment summary	1	15

Within each table, each entry contains a name which is a clickable link. By clicking this link the user will be directed to a new tab with detailed information about that selection.

By clicking on a specific Experiment Series, detailed information about that experiment series will be displayed as well as the Imports/Samples being displayed in the other tables will be from that specific Experiment Series. The same applies if a user selects an Import or a Sample name.

Clicking on a sample for details will produce a table with 4 different tabs. The first tab, named after the samples name, will display all of the information stored about the sample to the user.

Sample

D51_0d

Directory Info

Runs

Tables

Series Name Dendritic Cell Transcriptional Landscape

Lane Name 1stDataSet_and_MoreDepthReseq

Protocol Name DC from monocyte plus LPS treatment

Sample Name D01_MONO_Ctrl_0h

Barcode GTCGTA

Title Donor 51 Day 0

Source monocyte

Organism human

Molecule total RNA

Description Ovation Human FFPE RNA-Seq Multiplex System

Instrument Model Illumina HiSeq 2000

Avg. Insert Size 350

Read Length 100

Condition control

Groups pilot_EC

Permission Only your group

Donor D01

Time 0

Biological Rep 0

Technical Rep 0

The second tab labeled as ‘Directory Info’ will display all of the directory information related to the sample.

Sample

D51_0d

Directory Info

Runs

Tables

Input File(s) Directory:

/project/umw_garberlab/human/DC/1stDataSet_and_MoreDepthReseq

Input File(s):

LPS_libsD51_GTCGTA_L002_R1_001.fastq.gz,LPS_libsD51_GTCGTA_L002_R2_001.fastq.gz

LPS_libsD51_GTCGTA_L002_R1_002.fastq.gz,LPS_libsD51_GTCGTA_L002_R2_002.fastq.gz

LPS_libsD51_GTCGTA_L002_R1_003.fastq.gz,LPS_libsD51_GTCGTA_L002_R2_003.fastq.gz

LPS_libsD51_GTCGTAAT_L002_R1_001.fastq.gz,LPS_libsD51_GTCGTAAT_L002_R2_001.fastq.gz

LPS_libsD51_GTCGTAAT_L002_R1_002.fastq.gz,LPS_libsD51_GTCGTAAT_L002_R2_002.fastq.gz

Processed File(s) Directory:

/farline/umw_garberlab/ngstrack/DC/human/1stDataSet_and_MoreDepthReseq

Processed File(s):

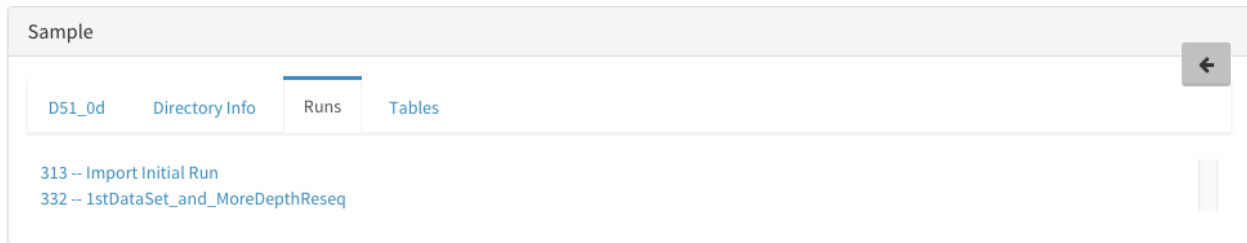
D01_MONO_Ctrl_0h.1.fastq.gz,D01_MONO_Ctrl_0h.2.fastq.gz

Amazon Backup:

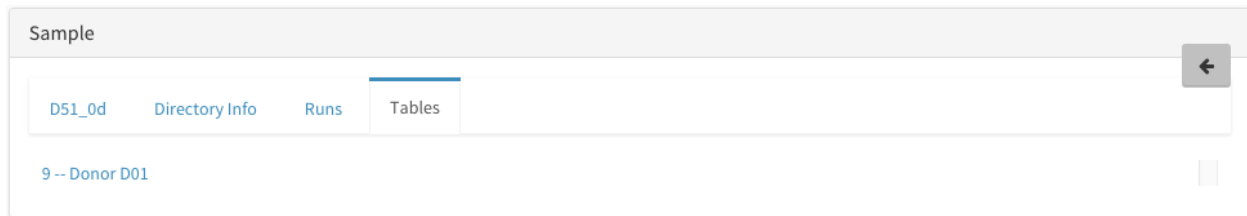
s3://biocorebackup/DC/human/1stDataSet_and_MoreDepthReseq

The third tab labeled as ‘Runs’ will display all of the links to each run report in which the sample has been used that

you have permissions to access (For more information on reports see the NGS Reports Guide).



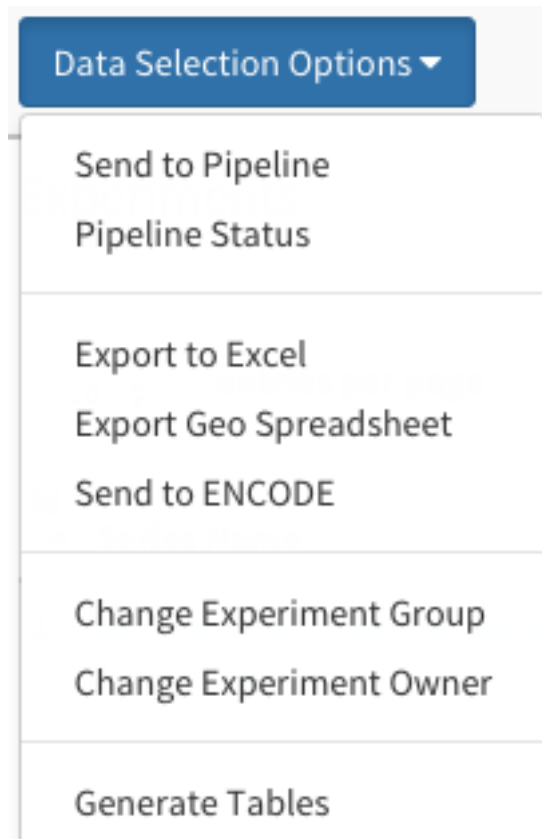
The last tab labeled as ‘Tables’ will display all of the links to custom tables in which the sample has been used that you have permissions to access (For more information on generated tables see the Table Creator Guide).



In the sample detailed information, as long as you have permissions to do so, you can edit the file names and directory paths. Please be warned that editing these file names and directory paths must accurately match the actual file names and paths.

At the top right of each Selection Details tab will be a grey arrow button. This will return the user back to the table portion of the NGS Browser.

Option Buttons:



At the bottom of the page there are a series of buttons that the user can click to perform specific tasks.

- **Send to Pipeline:** Experiment Series/Imports/Samples selected will then be sent to the NGS Pipeline page for further option selection and processing.
- **Pipeline Status:** This button takes the user to the NGS Status page where they can view their current/previous runs.
- **Export to Excel:** This button will take the selected Imports/Samples and save them to an excel spreadsheet for the users convenience (See Excel Export Guide)
- **Export Geo Spreadsheet:** This button takes your selected samples information and fills out an excel spreadsheet for geo submission (some additional information may be required).
- **Send to ENCODE:** *DISABLED* Under Construction
- **Change Experiment Group:** If you're the owner of an experiment series, you can change which group it belongs to as long as you belong to that group.
- **Change Experiment Owner:** If you're the owner of an experiment series, you can transfer ownership to another user within the same group.
- **Delete Selected:** This button will delete the selected Experiment Series/Imports/Samples.

Note that users need the proper permissions to delete a selected Experiment Series/Import/Sample. A message will be displayed upon selecting the Delete Selected button showing the Experiment Series/Imports/Samples that the user has permissions to delete as well as a confirmation text to confirm the deletion.

Before deleting Imports/Samples, please inform your fellow researchers for deleting this information is not recoverable. If you wish to delete Imports/Samples that you do not have permission to delete, contact either the owner of the Import/Sample, your local administrator, or someone at biocore@umassmed.edu.

3.5 Excel Export Guide

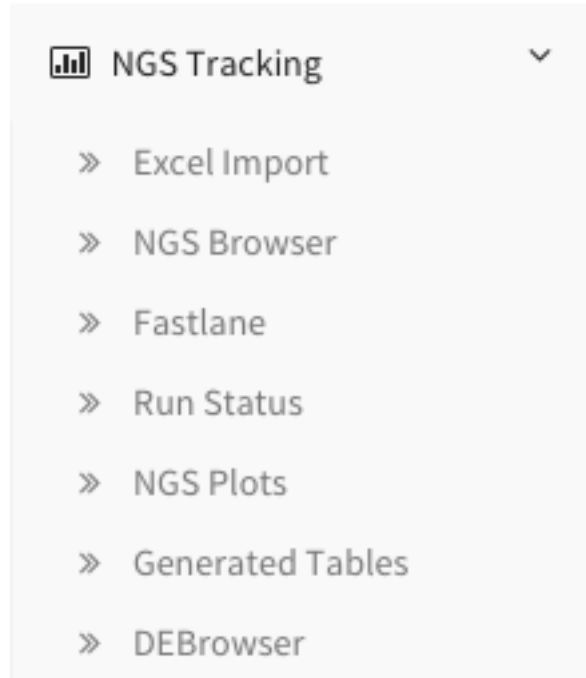
This guide will walk you through the process of exporting selected Imports/Samples through the NGS Browser page.

3.5.1 Getting Started

First, make sure you have an instance of dolphin available (see Dolphin Docker) as well as an account for the dolphin interface.

Next, make sure you have Imports/Samples that are available for you to select to export. This means that the Imports/Samples are viewable by you and that they have underwent the initial processing phase.

Once logged in, click on the 'NGS Tracking' tab on the left, then click on 'NGS Browser'.



3.5.2 From the Browser

Once you've made it to the Browser page, You can then filter the Imports/Samples if you wish in order to help filter your selection options. Once you've selected the Imports/Samples that you wish to export, you may click on the 'Export to Excel' button.

It is important to not that you cannot export Imports/Samples from different Experiment Series. By selecting data that belongs to more than one Experiment Series, an error will occur and the user will receive a warning message that the action is not allowed.

Likewise, if no data is selected and the Export button is pressed, an error will occur and the user will receive a warning message that the action is not allowed.

The file saved will be in the format of: `user _ year _ month _ day _ hour _ minute _ seconds . xls`

Upon downloading the excel export, if you wish to resubmit the files using excel import, make sure you list a processed directory within the spreadsheet.

For more information on the NGS Browser, see the NGS Browser guide.

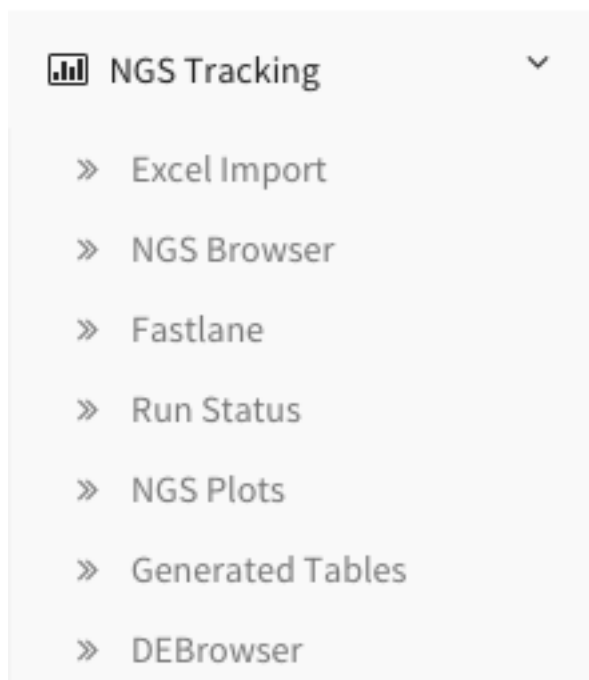
3.6 NGS Pipeline Guide

This guide will walk you through all of your options within the Run Pipeline page.

3.6.1 Getting Started

First, make sure to have an instance of dolphin available (see Dolphin Docker) as well as an account for the dolphin interface.

Once logged in, click on the 'NGS Tracking' tab on the left, then click on 'NGS Browser'.



Once you have selected the samples/imports you wish to analyze, hit the ‘Send to Pipeline’ at the bottom of the screen to begin.

Alternatively, if you’re rerunning a run with different parameters from the status page, you will be taken to the NGS Pipeline page.

3.6.2 NGS Pipeline Parameters

Once you’ve reached the pipeline page, there should be a table at the top of the page with your selection of samples from the browser.

You can explore this table further at your discretion, however the main portion of the page lies under the table within the tabs.

These tabs describe the parameters to the run you will be creating.

The first few tabs are required parameters, and they include:

Name the Run:

This parameter is a brief name you will be giving your run.

This will help you quickly identify which run is which within the status page.

Description:

This parameter is a brief description of the current run you will be creating.

This may help clear up any confusion that might be caused about the run in question later.

Genome Build:

From this dropdown menu, you’ll be able to select the specific genome build you wish to use.

Mate-paired:

Use the dropdown menu to select whether or not your libraries are mate paired or not.

Resume Run:

Select whether this is a fresh run, or a continuation of a previous run.

Output Directory:

This determines where the specific output of the current run should be deposited within the system.

If you're not entirely sure as where to deposit your data, you can contact your local administrator.

User permissions are required within the cluster for the path that you select. An error will show if you do not have proper permissions to add to/create this directory.

FastQC:

Select whether or not you want a FastQC report.

Permissions:

Select who will be able to view this run.

Options include:

- only me
- only my group
- everyone

Group Selection:

Select which group to submit this run under.

A list of all the groups you belong to will be provided.

Submission:

This will determine whether or not to prepare the run for submission to either ENCODE or GEO.

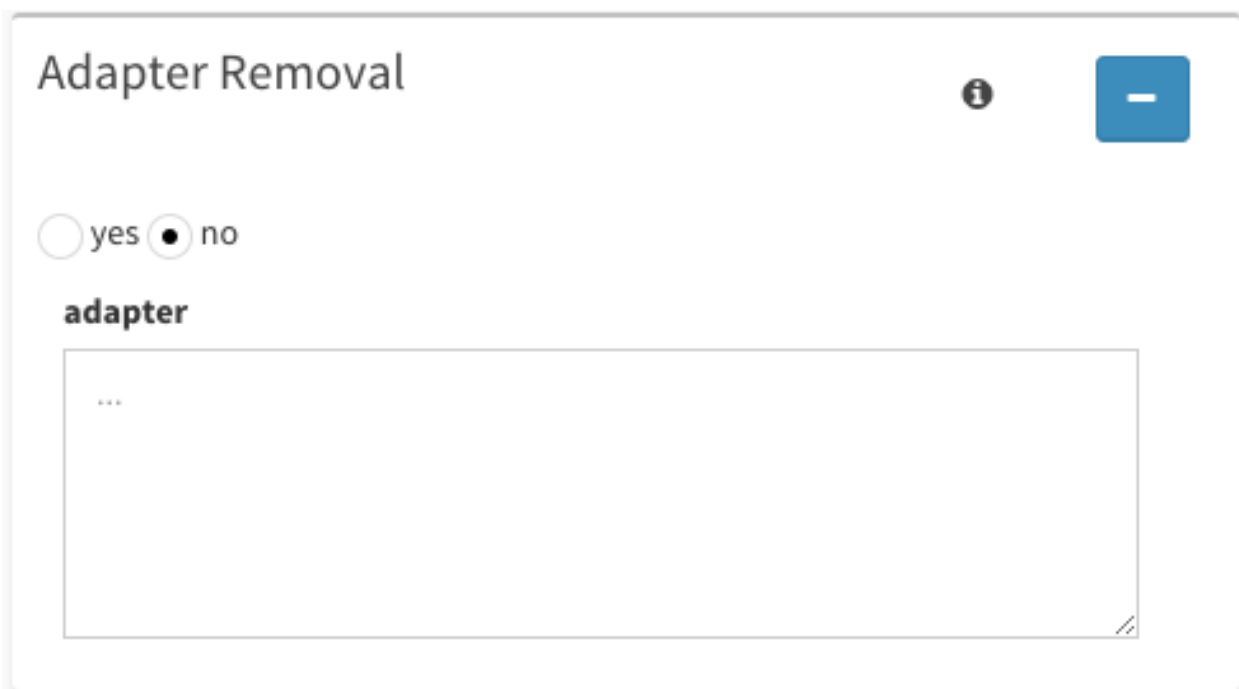
For now, this feature has been disabled.

3.6.3 Optional Parameters

Further down the page, there are tabs with a plus sign on them.



By clicking on the plus sign, you expand the tab and have additional parameter options to choose from.



The expandable tabs include:

Split FastQ:

If you would like the to split the resulting fastq files, you can expand the Split FastQ and select the yes button.

Once yes is selected, you then can select the size of the fastq file based on the number of reads per file.

Quality Filtering:

If your samples require quality filtering, you can expand the Quality Filtering tab and select the yes button.

After you've selected yes, you then can adjust the following quality filtering parameters:

- window size
- required quality
- leading
- trailing
- minlen

Adapter Removal:

If your samples require adapter removal, you can expand the Adapter Removal tab and select the yes button.

Once you've selected yes, you can then add your specific adapters within the adapter text box.

There should be one adapter per line within the text box.

Trimming:

If you would like to trim the reads within your samples selected, you can do so by expanding the Trimming tab and selecting the yes button.

Next, select whether or not the reads are paired-end or single-end.

After you've selected your read type, then enter the 5' length to trim and the 3' length to trim.

If paired-end reads are selected, additionally you will need to supply the 5' and the 3' length of the pair to be trimmed.

Custom Sequence Set:

If you would like to add custom sequence sets, expand the Custom Sequence Set and select the 'Add a Custom Sequence Set' button.

For these custom sequence sets, you will have to supply:

- The custom sequence index directory (full path)
- Prefix of the index file (Example: for index.fasta you would supply 'index')
- Bowtie parameters
- Description of the index
- Whether to filter out the reads mapped to this custom index

Please remove all spacing from the naming of sequences within your .fasta file in order for the pipeline to properly prepare quantification tables.

You can add multiple custom sequence sets if desired.

Common RNAs:

If you would like to map your reads sequentially to common RNAs, Expand the Common RNAs tab and select the yes for the RNAs you would like to map.

Bowtie2 maps your reads sequentially to common RNAs below and then filters the reads mapped out. To change the default parameters, click the *Change Parameters* button and then insert your parameters.

3.6.4 Additional Pipelines

If you would like to add additional features to your current run, you can expand the Additional Pipelines tab and hit the 'Add Pipeline' button.

You can add more than one selection of additional pipelines by clicking the 'Add Pipeline' button again. You can also remove your current pipeline selection by clicking the 'Remove Pipeline' button.

The pipelines added will be carried out from top to bottom.

A new box will appear with a dropdown menu that includes:

RNASeqRSEM:

Selecting the RNASeqRSEM additional pipeline will run RSEM in addition to the current run.

RSEM is an additional pipeline that will estimate gene and isoform expression levels for RNA-Seq data.

Upon selection of RNASeqRSEM, you have the option of selecting additional RSEM command line parameters as well as IGC/TDF or BigWig conversions.

The RSEM pipeline also has an option to generate RSeQC reports. For more information on the RSeQC reports, you can view the program used, 'read-distribution.py' for RSeQC, [here](#).

Tophat:

Selection the Tophat additional pipeline will allow you to run Tophat in addition to the current run.

Tophat is a popular RNA-seq alignment tool used for mapping and splice junction detection.

Upon selecting the Tophat pipeline addition, you have the option of adding additional tophat command line parameters as well as IGC/TDF or BigWig conversions.

The Tophat pipeline also has an option to generate RSeQC reports. For more information on the RSeQC reports, you can view the program used, 'read-distribution.py' for RSeQC, 'here' [_](#).

ChipSeq:

If your samples of interest include ChipSeq data, then the ChipSeq additional pipeline should be selected.

After selection of ChipSeq, some additional parameters are required.

Chip Input Definitions are the names of the files of your chipseq input. Multimappers determines the maximum number of locations reads are allowed to map to.

Tag size, in base pairs, for MACS determines the size of the tags while the Band Width, in base pairs, for MACS determines the size of the sequenced regions.

The Effective genome size, in base pairs, is the size of the mappable part of the genome.

IGV/TDF and BigWig conversion is also selectable.

DESeq:

First, in order to select the DESeq additional pipeline option, you must have already selected the RNASeqRSEM option first.

The DESeq pipeline allows for differential expression analysis to be conducted amongst your samples.

Using the selection boxes labeled 'Condition 1' and 'Condition 2' you can select which samples you wish to check against.

Once you've selected your conditions, you then can determine your Fit Type, and p-adjustment cutoff value as well as whether or not you want the Heatmap and the Fold Change cutoff.

Based on your previous selection from the Common RNAs tab, you can also select which given sequences you want to analyze.

Note that you can select DESeq multiple times, incase you want to run multiple pairwise comparisons on a single run.

BisulphiteMapping:

If you would like to carry out Bisulphite mapping, then the BisulphiteMapping additional pipeline should be selected.

Bisulphite Mapping is a bisulphite sequencing mapping program that indexes only digestion sites.

In addition to running the BSMAP program with it's additional parameters, the user can also run MCall with additional parameters to report statistics such as various bias, confidence intervals, and methylation ratios.

In order to run MethyKit, MCall must first be selected.

DiffMeth:

If you would like to carry out Differential Methylation, then the DiffMeth additional pipeline should be selected.

Differential Methylation lets you compare the Bisulphite Mapping results of samples against other samples within your run.

Using the selection boxes labeled 'Condition 1' and 'Condition 2' you can select which samples you wish to check against.

In order to carry out DiffMeth, the user first has to select the BisulphiteMapping pipeline, select the 'Run MCall' checkbox, and select the 'Run MethyKit' checkbox as they are required to carry out this step.

You may select DiffMeth multiple times, incase you want to run multiple pairwise comparisons on a single run.

HaplotypeCaller:

If you would like to use the Genome Analysis Toolkit's Haplotype Caller to detect SNP variants within your samples, the the HaplotypeCaller additional pipeline should be selected.

In order to run Haplotype Caller, you must have a Tophat pipeline or ChipSeq pipeline already within your list of pipelines in order to generate the bam files needed for this step.

HaplotypeCaller will output .vcf files which will contain possible variants in accordance with the genome selected.

3.6.5 Submission

Once all of your parameters are squared away and you've selected all of the additional options/pipelines that you desire for your run, you may hit the 'Submit' button at the bottom left of the page to submit the run.

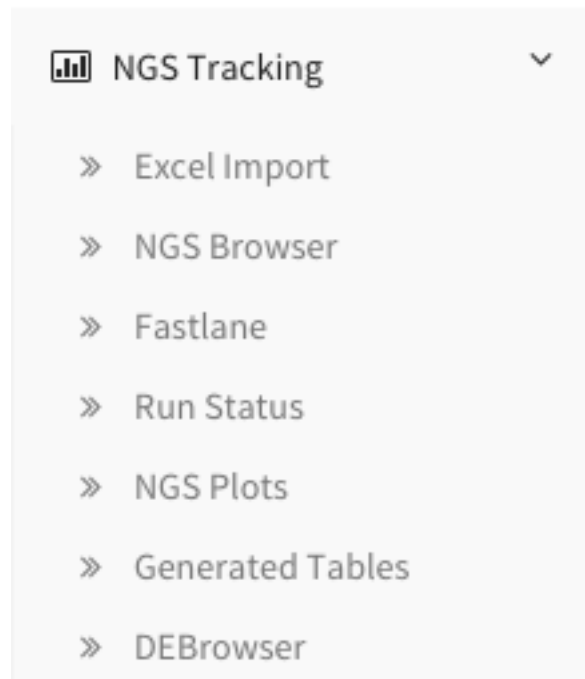
3.7 NGS Status Guide

This guide will walk you through all of your options within the Run Status page.

3.7.1 Getting Started

First, make sure to have an instance of dolphin available (see Dolphin Docker) as well as an account for the dolphin interface.

Once logged in, click on the 'NGS Tracking' tab on the left, then click on 'Run Status'.



3.7.2 Status Page

The Run Status page will display all of your current runs within Dolphin, old and new.

If you have not yet processed any runs, this page will consist of an empty table with no further options of interaction.

However, if you have already started processing some initial runs, the table reads as follows:

ID:

This is a unique run identifier. If contacting your local administrator or someone at biocore@umassmed.edu about a specific run, this number will help identify the run quicker.

Name:

The name given to the specific run.

Output Directory:

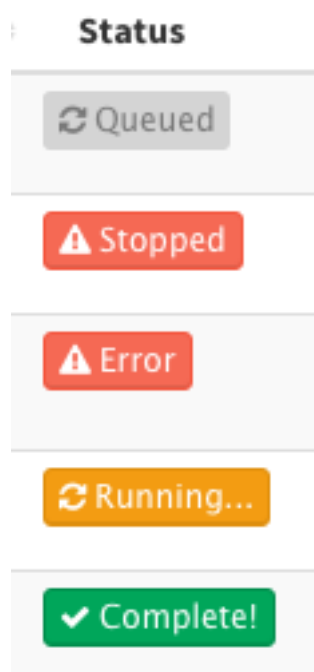
The output directory specified where the run information will be sent within the cluster.

Description:

The description of the run given by the user.

Status:

This is the status of the current run. The run can have 5 different statuses.



Each status corresponds to how the current run is behaving. The statuses and their meanings include:

- *Queued*: A queued run is currently waiting to be executed. Time in the queue depends on the cluster.
- *Stopped*: A stopped run has started to run, but was at one point manually stopped by a user.
- *Error*: An errored run has had something go wrong and could not complete appropriately. By clicking the status button on an error, a window will pop-up with a little more information about the error. If there are logs of the past processes, you can click the 'Adv. Status' button to be taken to the Advanced status page.
- *Running*: A run that is currently running.
- *Completed*: A completed run has finished with its task and the output is ready.

As long as the run isn't queued, you can click on the status button of a run to obtain more detailing information about the run in the Advanced Status page for that specific run.

Options:

The options column contains a clickable options button that will give the user specific options on the specified run.

Stop
Rerun
Resume

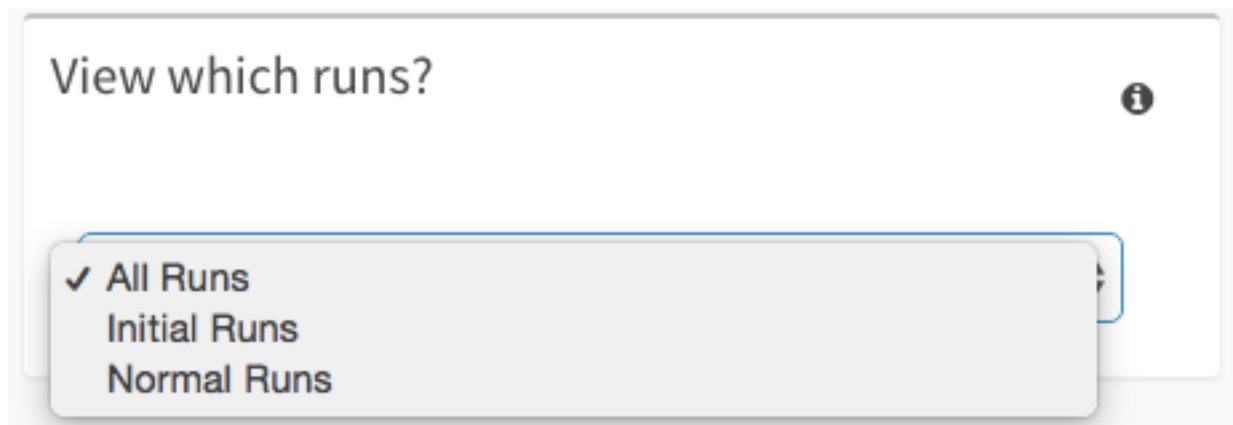
Delete

Each status and run type will have a different set of options.

Rather than describing every set of options for every combination, we will go over what each option does.

- *Delete*: This will delete the current selected run.
- *Cancel*: This will stop a run currently queued. It will not delete the run information.
- *Stop*: This will stop a currently running run. It will not delete the run information.
- *Rerun*: This option will rerun the selected run and allow for parameter changes. If a new directory is specified, it will create a new run.
- *Resume*: This will rerun the selected run without parameter changes.
- *Report Details*: Selecting this option will take you to the selected run's report page.
- *Generate Plots*: Selecting this option will take you to the selected run's plot page.
- *Change Permissions*: If you are the owner of the run, you can change the permissions of the selected run

View Which Runs?:



This dropdown menu will toggle which types of runs you will be viewing within the status table.

The options include:

- *All Runs*: This is the default setting. All of the runs you have permissions to view will be displayed.
- *Initial Runs*: This setting will display all of the initial runs you have permissions to view.
- *Normal Runs*: This setting will display all of the runs that you have permissions to view that are not initial runs.

3.7.3 Advanced Status Page

Upon clicking on a non-queued status, you will be directed to the advanced status page.

This page displays a table with each step processed, or currently being processed within the system of your run.

If the progress bar is green, then the step has fully finished without errors. If the bar is red, it is either currently running or it has received an error within the step.

Upon selecting the ‘Select Service’ button on the right of each step, another table will be shown of all the current subprocesses within this given step.

Along with some additional information, the user can select the ‘Select Job’ button to view the standard output of this current job.

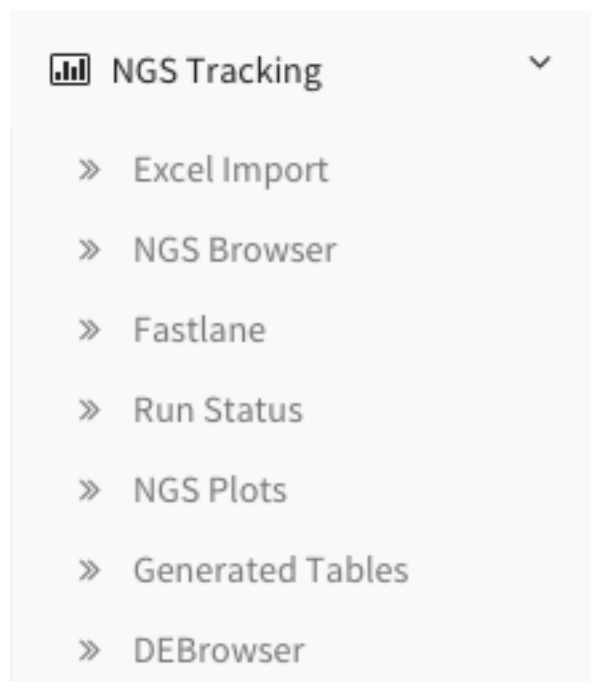
3.8 NGS Reports Guide

This guide will walk you through all of your options within the Reports page.

3.8.1 Getting Started

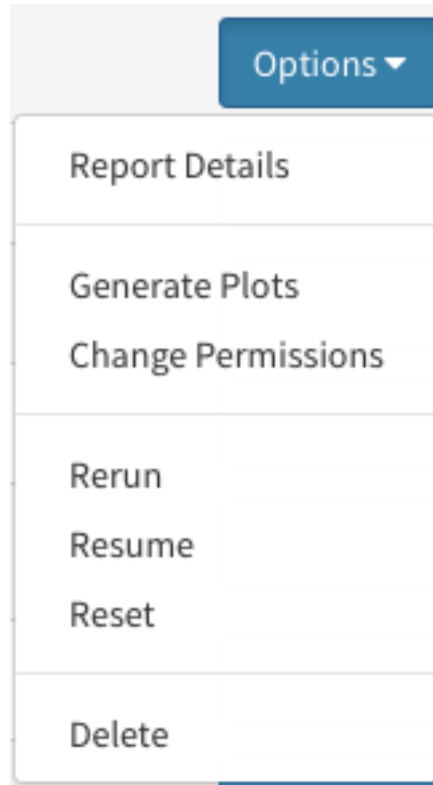
First, make sure to have an instance of dolphin available (see Dolphin Docker) as well as an account for the dolphin interface.

Once logged in, click on the ‘NGS Tracking’ tab on the left, then click on ‘Run Status’.



After navigating to the status page, the next step is to make sure the run whose results you are interested in has completed without errors.

Once you have a completed run, you can select the options button on the far right and select the ‘Report Details’ option.



Selecting this option will bring you to the Report Details page.

3.8.2 Report Details

Upon reaching the Report Details page, the user is greeted with an Initial Mapping Results table.

Initial Mapping Results

Show 10 entries

Search

Libname	Total Reads	rRNA	rmsk	Reads Left	Selected
A1	18,144,346	7,076,495 (39.00 %)	480,032 (2.65 %)	10,587,819 (58.35 %)	<input type="checkbox"/>
A2	15,692,225	4,812,343 (30.67 %)	493,490 (3.14 %)	10,386,392 (66.19 %)	<input type="checkbox"/>
A3	16,128,139	245,833 (1.52 %)	343,452 (2.13 %)	15,538,854 (96.35 %)	<input type="checkbox"/>
A4	10,774,580	243,200 (2.26 %)	347,927 (3.23 %)	10,183,453 (94.51 %)	<input type="checkbox"/>
A5	20,009,284	99,996 (0.50 %)	436,830 (2.18 %)	19,472,458 (97.32 %)	<input type="checkbox"/>
B1	2,271,346	165,145 (7.27 %)	32,199 (1.42 %)	2,074,002 (91.31 %)	<input type="checkbox"/>
B2	2,120,067	230,931 (10.89 %)	37,157 (1.75 %)	1,851,979 (87.35 %)	<input type="checkbox"/>
B3	104,247,650	4,770,145 (4.58 %)	3,527,401 (3.38 %)	95,950,104 (92.04 %)	<input type="checkbox"/>
B4	48,716,066	9,929,651 (20.38 %)	1,065,385 (2.19 %)	37,721,030 (77.43 %)	<input type="checkbox"/>
B5	37,917,514	5,903,201 (15.57 %)	1,348,509 (3.56 %)	30,665,804 (80.88 %)	<input type="checkbox"/>

Showing 1 to 10 of 18 entries

Previous 1 2 Next

--- Select a Result ---

This table displays various information about the samples you've selected for your run and based on the common RNAs that you may have selected to map for, the table will also display the number of reads mapped for each selected.

It's important to note that the reads mapped to RNA are not selected for mapping within the other RNA steps.

If you have indeed selected to map their reads against a specific RNA, the dropdown under the table will also produce additional information in regards to the RNA selected from the dropdown list.

rRNA

Select Data Options

10 entries per page

Search here...

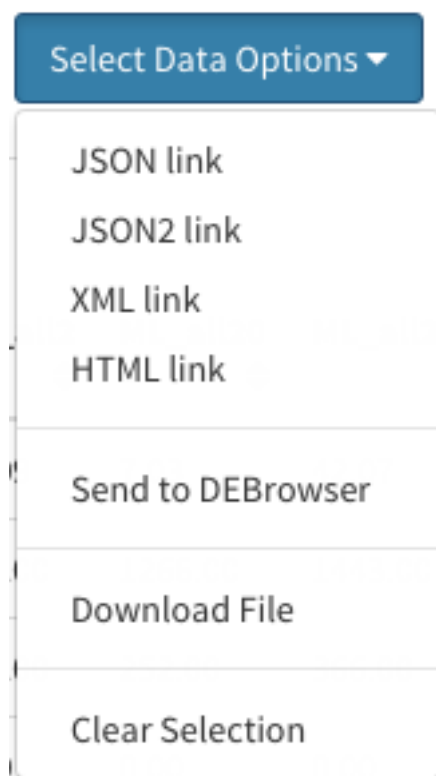
id	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	B6	C1	C2	C3	D1	D2	D3	Tube1
28S	9790590	12328845	726400	789229	289215	425572	572937	7065711	32158096	5475920	1008099	251799	57916	281163	2122646	2665763	3718814	805020
18S	104146	138019	142293	134022	45984	5061	8889	1612717	6079195	991400	226022	39554	9461	44946	338002	409774	595569	76732
5S	551	1471	1651	647	471	59	194	14363	33364	8720	2345	50	7	36	896	632	746	679
5.8S	111	235	4187	4226	1622	5	25	53166	199542	37364	6438	1830	558	1984	16505	19084	30627	2495

← Previous

1

Next →

Once the table is populated, an options button labeled ‘Select Data Options’ will appear and selecting it will give you a variety of options to choose from.



Options include:

- **TYPE Links:** These links will direct the user towards another webpage with the data from the table in the specified TYPE format.
- **Send to DEBrowser:** This link will only appear under rsem/mRNA/tRNA results. It will send the user to the DEBrowser with the selected table as the dataset (For more information on DEBrowser, see the DEBrowser section).
- **Download File:** This link will download the table to a file in a TSV format.
- **Clear Selection:** If the user desires, they can clear the table of it's contents.

Following the Initial Mapping Results there may be additional minimized sections. These sections are available to the user based on the parameters set by the user.



These tabs include:

- FastQC Summary
- Detailed FastQC Results
- RSEM Results
- DESEQ Results
- Picard Metrics
- RSeQC Metrics

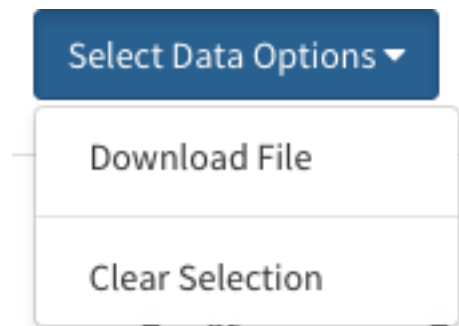
Expanding any of the FastQC tabs will provide a list of links related to either the FastQC summary or the specific sample FastQC results. Clicking one of these links will re-direct you to another page with the specified results.

By expanding the RSEM, DESEQ, Picard Metrics, or RSeQC tabs, you will be greeted with another dropdown menu like that within the Initial Mapping Results tab.

The list provided within the dropdown menu will contain files related to the tab which is expanded.

If you select a TSV file, a table will be populated with the results and another options button will appear to the right of the dropdown list with the same options of that within the Initial Mapping Results table.

Selecting a PDF file will not populate a table, nor will it replace a previously populated table that you have already selected. Instead, selecting a PDF file will change the options available to the user within the 'Select Data Options' button to either 'Download File' or 'Clear Selection'. Selecting the 'Download File' option will redirect you to a new page containing the PDF with the ability to download.



At the very bottom of the page, you have 2 additional buttons.

The 'Return to Status' button will re-direct you to the status page and the 'Go To Plots' button will re-direct you to the plots page.

For more information on the Status page, consult the 'NGS Status Guide'.

For more information on the Plots page, consult the 'NGS Plots Guide'.

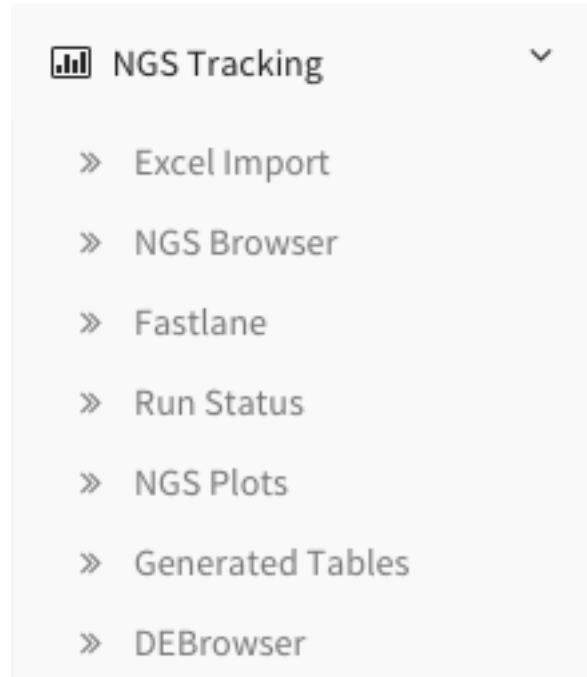
3.9 NGS Plots Guide

This guide will walk you through all of your options within the Plots page.

3.9.1 Getting Started

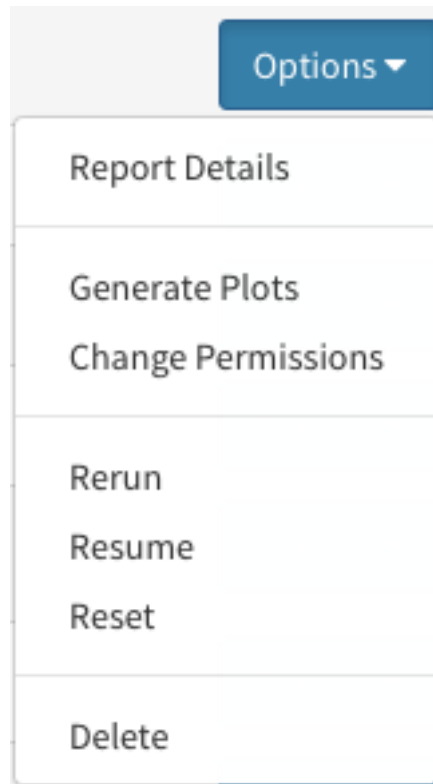
First, make sure to have an instance of dolphin available (see Dolphin Docker) as well as an account for the dolphin interface.

Once logged in, click on the ‘NGS Tracking’ tab on the left, then click on ‘Run Status’.



After navigating to the status page, the next step is to make sure the run whose results you are interested in has completed without errors.

Once you have a completed run, you can select the options button on the far right and select the ‘Generate Plots’ option.



Selection this option will bring you to the Plots page.

Alternatively, You can visit the Reports page for a specified run and click on the 'Go To Plots' button at the bottom of the page to go to the plots page.

Also, you can view your generated tables from the Generated Tables page by selecting to go to plots from the desired table. Please see the 'Table Creator Guide' for additional information.

3.9.2 Plots Navigation

Upon navigating to the plots page, you will notice that there are 3 major sections to the page. These sections include the Control Panel, the Plots, and the Plots Table.

Control Panel:

Control Panel

Source

Select a Source

X axis
(Log) ☒

Y axis
(Log) ☒

Pseudo
Count

1

Color axis

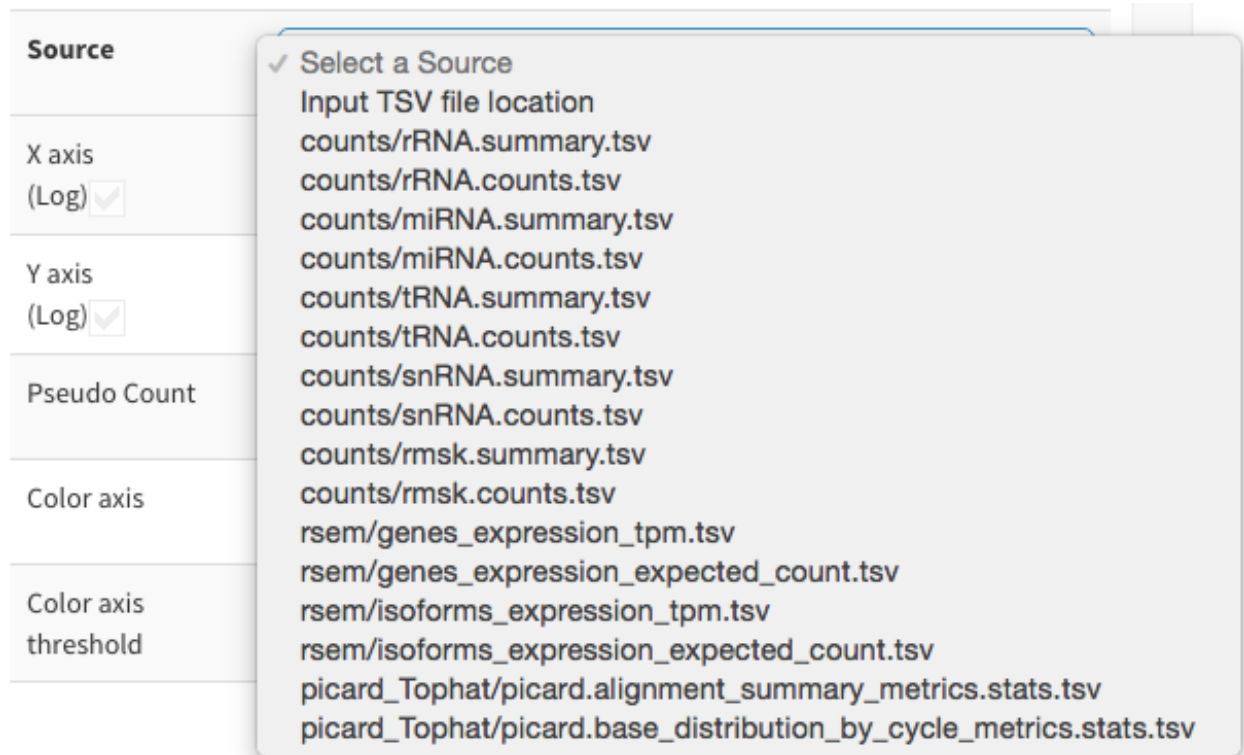
Color axis
threshold

Threshold

Selected
Columns

Query Genes

In order to view the plots of your run, you first need to select the actual data you want to examine. The dropdown menu next to ‘Source’ will give the user a list of potential files to view.



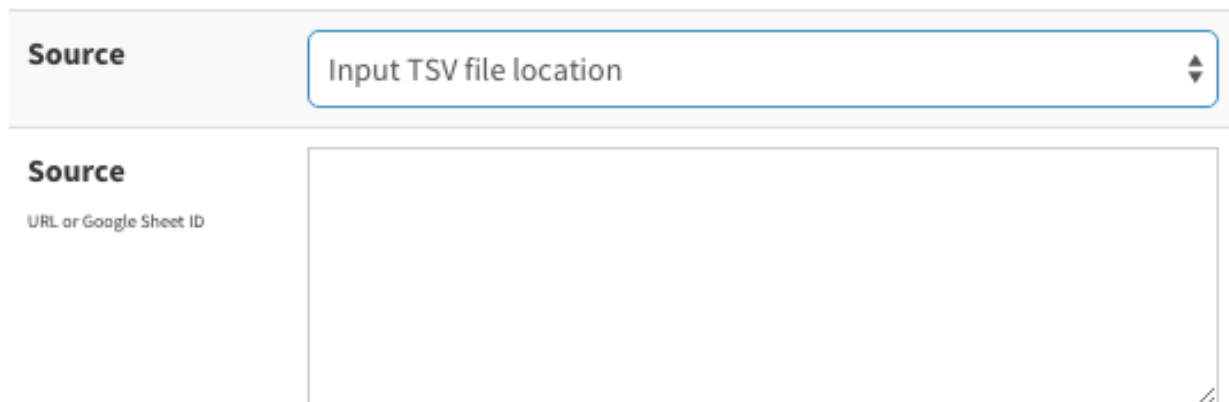
The image shows a web interface with a 'Source' dropdown menu open. The dropdown list includes the following options:

- ✓ Select a Source
- Input TSV file location
- counts/rRNA.summary.tsv
- counts/rRNA.counts.tsv
- counts/miRNA.summary.tsv
- counts/miRNA.counts.tsv
- counts/tRNA.summary.tsv
- counts/tRNA.counts.tsv
- counts/snRNA.summary.tsv
- counts/snRNA.counts.tsv
- counts/rmsk.summary.tsv
- counts/rmsk.counts.tsv
- rsem/genes_expression_tpm.tsv
- rsem/genes_expression_expected_count.tsv
- rsem/isoforms_expression_tpm.tsv
- rsem/isoforms_expression_expected_count.tsv
- picard_Tophat/picard.alignment_summary_metrics.stats.tsv
- picard_Tophat/picard.base_distribution_by_cycle_metrics.stats.tsv

The background interface shows the following controls:

- Source** (dropdown menu)
- X axis** (Log) ☒
- Y axis** (Log) ☒
- Pseudo Count**
- Color axis**
- Color axis threshold**

If the file you want to view isn’t an option in the dropdown menu, you can always manually type in the source of the file by selecting the ‘Input TSV file location’ option.



The image shows the web interface with the 'Source' dropdown menu set to 'Input TSV file location'. Below the dropdown, there is a text input field labeled 'Source' with the placeholder text 'URL or Google Sheet ID'.

Once you’ve selected the file you wish to plot the other various options in the control panel will fill up with the appropriate additional information.

The X axis and Y axis dropdown menus select which variable will be represented for both the X and Y axis within the scatter plot.

The Color axis dropdown menu controls the color separation of data within the scatter plot.

The Color axis threshold has both a slider and an input box where the user can control the color threshold represented within the scatter plot.

In order for the barplot, heatmap, and plot table to map data, a selection from the Selected Columns box must be made.

Multiple selections can be made at the same time.

The Query Genes input allows the user to input specific genes in a comma separated format. After input of a gene list, the user can hit 'Submit' under the input box and the queried genes will be selected on the various plots.

Last, the 'Selected Region' box displays all the genes selected from the scatter plot in a comma-separated format.

Selected Region

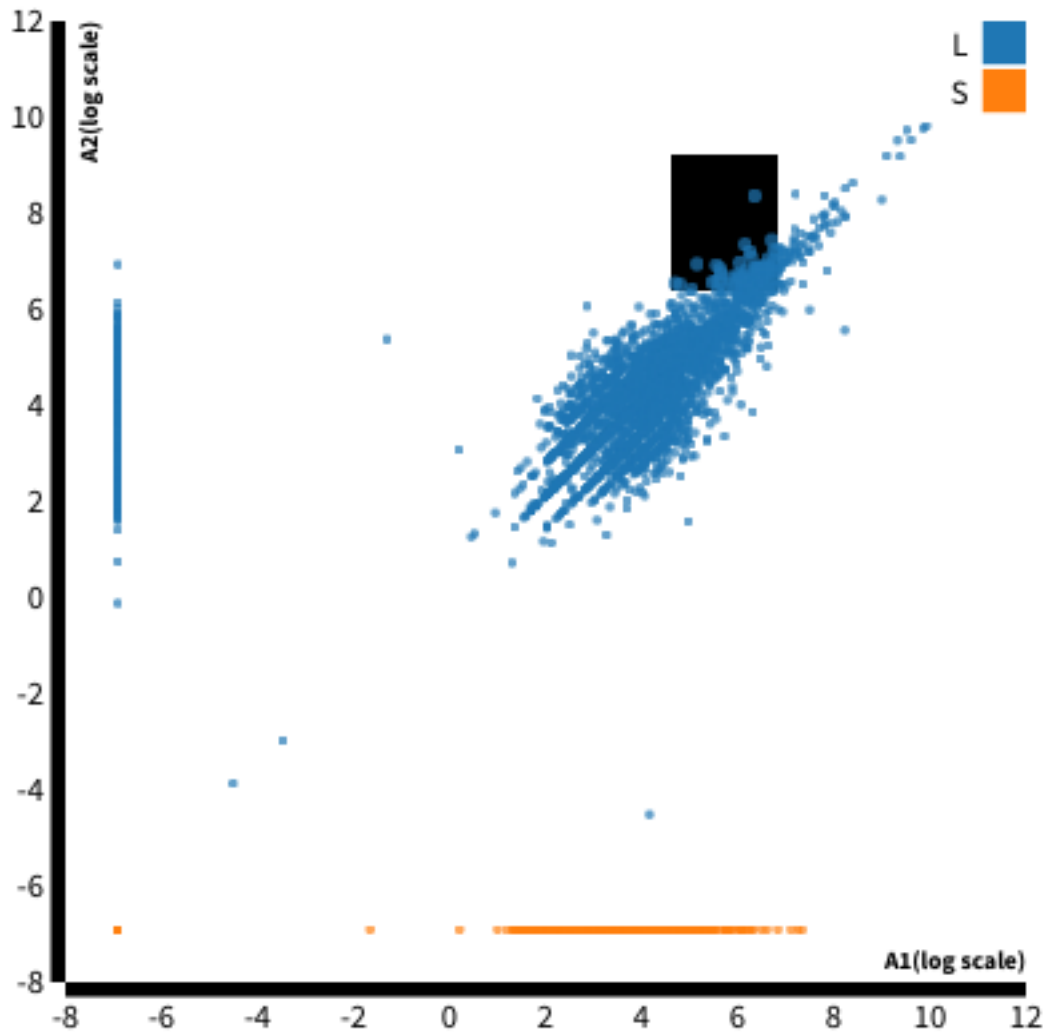
Number: **71**

```
NOL9,MRPS21,INPP5B,PDDC1,RSL1D1,PIGF,SNRNP35,  
MREG,SLC16A4,UBXN2A,MTRNR2L9,MCUR1,RHOU,SRSF  
10,TRMT10B,NBPF3,LOC652276,CXorf56,DENND6A,SNR  
PD3,FAM127A,STAG3L2,TBCE,GINS1,PHAX,LINC00265,R  
PL41,XIAP,FAM27B,RPS2,HOOK3,NMNAT1,LOC730183,A  
RL17A,PNMA6A,TMED4,RPS3A,RNF115,AKIRIN1,LIN52,C  
OQ10B,LINC00657,SLC25A25-  
AS1,ZNF329,ZNF431,METTL7A,DNAL1,SNHG16,TRIM52-  
AS1,NDUFV3,CENPN,RPL26,CDT1,NUDC,GLB1L,ZNF621,  
LOC100134868,HIST2H2AB,GUSBP3,ARL16,GTF2H2B,SL  
C31A1,QPRT,LOC100129917,TTF1,APOBEC3C,ZNF394,L  
MNB1,TMEM223,SNHG7,BCYRN1
```

Plots:

The plots section is divided into 3 specific types of plots: a scatter plot, a barplot, and a heatmap.

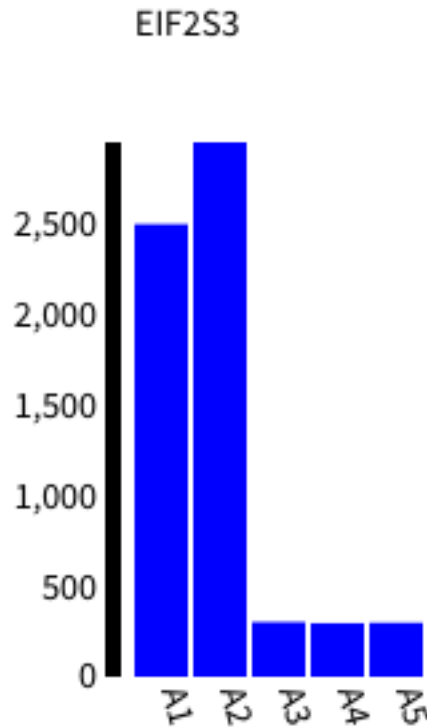
Once you have determined you've input your source file for the plot, a scatter plot will appear within the 'Scatter plot' box based on your specific input variables.



As the scatter plot fills out, you can then click and drag your mouse over data points to select them for both the heatmap selection and the Plot Table selection.



The barplot fills out based on your ‘Selected Column’ section and which point on the scatter plot you mouse over with the cursor.



You can also hover over heatmap rows and the barplot will fill with the corresponding information from the heatmap.

Plots Table:

The plots table fills once a selection has been made within the scatterplot and will fill accordingly.

Table

gene	transcript
NOL9	NM_024654
MRPS21	NM_018997,NM_031901
INPP5B	NM_001297434,NM_005540
PDDC1	NM_182612
RSL1D1	NM_015659
PIGF	NM_002643,NM_173074
SNRNP35	NM_022717,NM_180699,NR_104103
MREG	NM_018000
SLC16A4	NM_001201546,NM_001201547,NM_001201548,NM_001201549,NM_004696

This plots table displays the selection made by the user as well as the values that differ between the samples.

Selections of many points can be expanded by clicking on the 'Show entire table' button at the very bottom of the table list.

Most of the files displayed in the graphs can be obtained from the ‘NGS Reports page’ for download. For more information on how to navigate the Reports page, please check out the ‘NGS Reports Guide’.

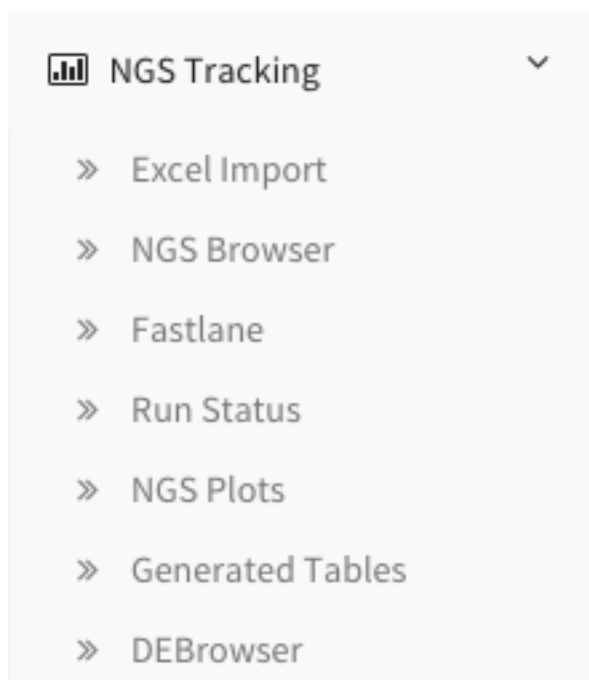
3.10 Table Creator Guide

This guide will walk you through all you need to know about the Table Creator page(s).

3.10.1 Getting Started

First, make sure to have an instance of dolphin available (see Dolphin Docker) as well as an account for the dolphin interface.

Once logged in, click on the ‘NGS Tracking’ tab, followed by clicking on the ‘NGS Browser’ from the dropdown to be sent to the Browser for sample selection.



Additionally, if you want to skip straight to your previous created tables, you can select the ‘Generated Tables’ tab from the ‘NGS Tracking’ tab.

3.10.2 Creating Custom Tables

Sample Selection

Once you’ve successfully made it to the Browser, select samples which have finished runs and also have reports which you would like to have merged.

As soon as you’ve selected all of your samples, you can select the ‘Generate Tables’ button to be sent to the Table Generation page.

Table Generation

Once at the Table Generation page, you will notice that the samples you have selected fill up the ‘Samples Selected’ table. Under the ‘Run ID’ column you can select which run from that samples results you wish to use.

Samples Selected

Show 10 entries

Search:

id	Sample Name	Run ID	Delete
35	control_rep1	Run 13: Run 2 RNA-Seq	
36	control_rep2	Run 13: Run 2 RNA-Seq	
37	control_rep3	Run 13: Run 2 RNA-Seq	
38	exper_rep1	Run 13: Run 2 RNA-Seq	
39	exper_rep2	Run 13: Run 2 RNA-Seq	

Showing 1 to 5 of 5 entries

Previous

1

Next

As you select different runs, you might notice rows in the ‘Report Selection’ box dimming out. The bolded files in this page are the compatible reports in which you can merge while the dimmed out ones are incompatible or missing from one of the runs you have selected.

Once you’ve selected all the appropriate runs, you may then select the report you wish to merge within the ‘Report Selection’ box.

Report Selection

counts/rRNA.summary.tsv

counts/rRNA.counts.tsv

counts/miRNA.summary.tsv

counts/miRNA.counts.tsv

counts/tRNA.summary.tsv

counts/tRNA.counts.tsv

rsem/genes_expression_tpm.tsv

rsem/genes_expression_expected_count.tsv

rsem/isoforms_expression_tpm.tsv

rsem/isoforms_expression_expected_count.tsv

DESeq2RSEM1/alldetected_genes.tsv

DESeq2RSEM1/selected_log2fc_genes.tsv

DESeq2RSEM1/alldetected_isoforms.tsv

picard_Tophat/picard.alignment_summary_metrics.stats.tsv

picard_Tophat/picard.base_distribution_by_cycle_metrics.stats.tsv

picard_Tophat/picard.CollectRnaSeqMetrics.stats.tsv

picard_Tophat/picard.CollectRnaSeqMetrics.hist.tsv

picard_Tophat/picard.insert_size_metrics.stats.tsv

If you forgot any samples that you wanted to add to the table generation, you can maximize the ‘Additional Sample Selection’ and select the samples from the table generated within this box. Checking off a sample will also remove it from the ‘Samples Selected’ table.

Additional Sample Selection

Samples

10 entries per page

Search here...

id	Sample Name	Title	Source	Organism	Molecule	Selected
1	control_rep1	control_rep1				<input type="checkbox"/>
2	control_rep2	control_rep2				<input type="checkbox"/>
3	control_rep3	control_rep3				<input type="checkbox"/>
4	exper_rep1	exper_rep1				<input type="checkbox"/>
5	exper_rep2	exper_rep2				<input type="checkbox"/>
6	exper_rep3	exper_rep3				<input type="checkbox"/>
29	c_rep1	example title 1	example source	human	total RNA	<input type="checkbox"/>
30	c_rep2	example title 1	example source	human	total RNA	<input type="checkbox"/>
31	c_rep3	example title 1	example source	human	total RNA	<input type="checkbox"/>
32	e_rep1	example title 1	example source	human	total RNA	<input type="checkbox"/>

1 to 10 of 18 entries

← Previous

1

2

Next →

[To Created Tables](#)
[Generate Table](#)
[Back to Browser](#)

Additionally, if you want to remove samples, you can also click on the red ‘X’ to the right of each sample within the ‘Samples Selected’ table.

When all the information is selected and you are ready to generate your table, you can then click the ‘Generate Table’ button to be taken to your generated table.

Table Generated

In the ‘Table Generated’ page, you will have the reports of the runs selected merged into one table. This table is much like the tables within the ‘NGS Reports’ page.

Table Generated

10 entries per page

Search here...

File	Total Reads	Unmapped Reads	Reads 1	Reads >1	Total align
control_rep1	24788	24287 (97.98%)	0 (0.00%)	501 (2.02%)	0
control_rep2	16193	15687 (96.88%)	0 (0.00%)	506 (3.12%)	0
control_rep3	17508	17004 (97.12%)	0 (0.00%)	504 (2.88%)	0
exper_rep1	18068	17560 (97.19%)	0 (0.00%)	508 (2.81%)	0
exper_rep2	21520	21000 (97.58%)	0 (0.00%)	520 (2.42%)	0

1 to 5 of 5 entries

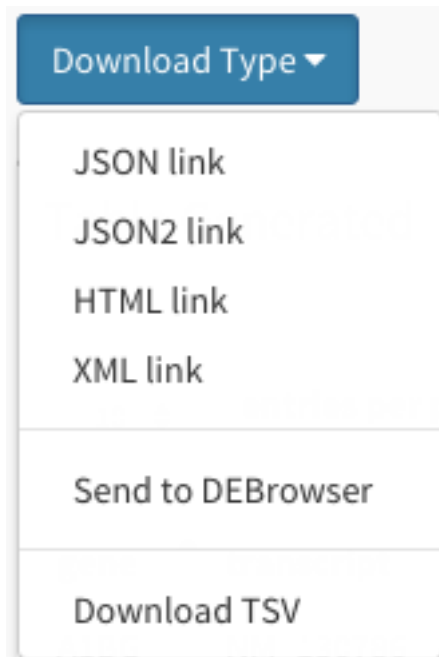
← Previous

1

Next →

For external use, you can select the ‘Download Type’ button to export the table to a variety of formats:

- JSON
- JSON2
- HTML
- XML



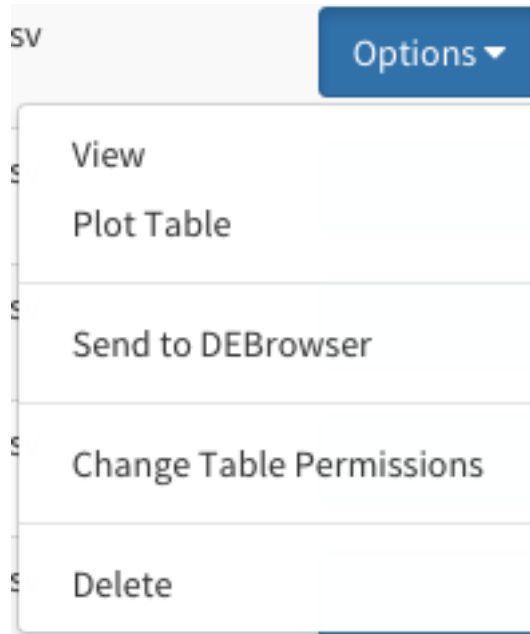
You can also download the raw TSV file of the newly generated table or send the table to the DEBrowser (For more information on DEBrowser, see the DEBrowser section). This will only appear for rsem/mRNA/tRNA tables.

If you wish to save the table, either for easy access later or for plotting capabilities, you can name the table in the ‘Save Table As’ tab and then click the ‘Save Table’ at the bottom of the page.

Saved Tables

Within the ‘NGS Table List’ the user can view all of their merged tables they have created and saved. You can select a few options from the options button to the right of each run:

- View
- Plot Table
- Change Permissions
- Send to DEBrowser
- Delete



Selecting 'View' will take you to the 'Table Generated' page with the report where you can save the table under a new name if you wish.

Selecting the 'Plot Table' will direct you to the 'Plots' page with your generated table selected as the input. You can also visit the 'Plots' page at any time and select any of your generated tables to use for input.

Change Permissions will appear if you are the owner of the generated table. You will be able to change the permissions given to this table by selection this option.

Selecting 'Send to DEBrowser' will allow you to send your table information straight to the DEBrowser (For more information on DEBrowser, see the DEBrowser section). This will only appear for rsem/mRNA/tRNA tables.

Selecting 'Delete' will remove the generated table from your list of tables.

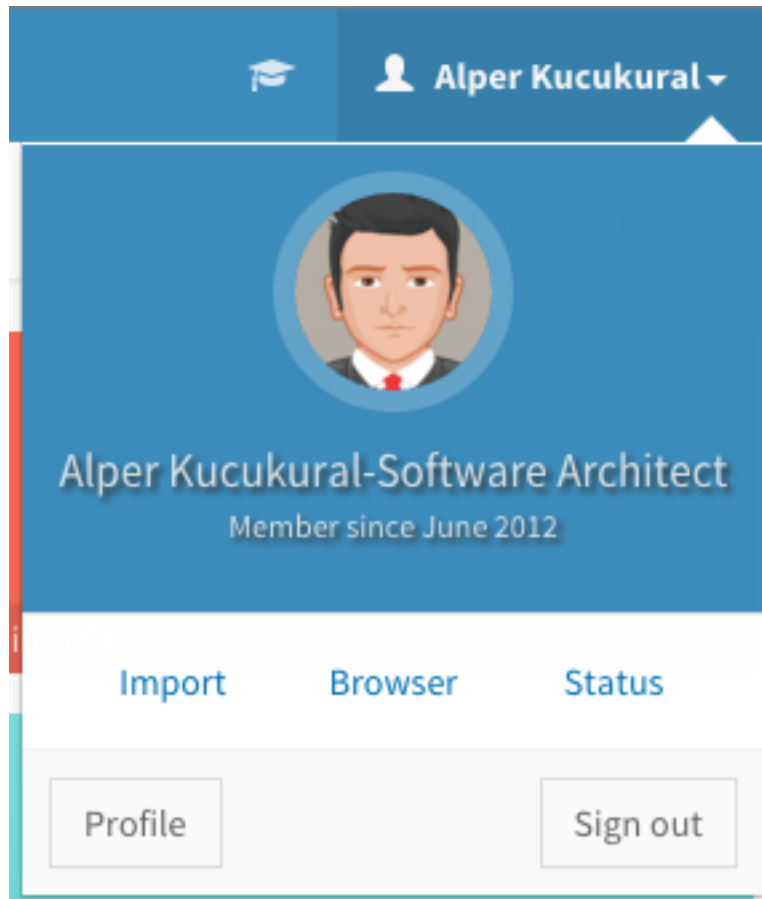
3.11 Dolphin Profile Guide

This guide will walk you through all of your options within the Profile page.

3.11.1 Getting Started

First, make sure to have an instance of dolphin available (see Dolphin Docker) as well as an account for the dolphin interface.

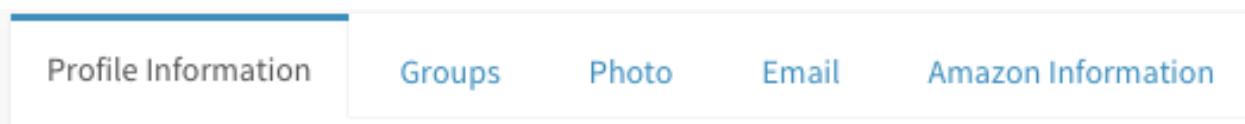
Once logged in, click on the tab in the top left of the screen with your name, then click on 'Profile' once the new menu appears.



3.11.2 Profile Page

Once you've accessed the profile page, you'll notice two main portions of the page.

The first segment is the tab layout.



This is your main form of navigation throughout the profile page.

Following these tabs is your 'Profile Information', or the first selected tab of the navigation tab.

Profile Information	
id	1
username	kucukura
clusteruser	ak97w
role	Software Architect
name	Kucukural,Alper
email	alper.kucukural@umassmed.edu
email_toggle	0
institute	UMassMed
lab	umw_biocore
pass_hash	null
verification	null
memberdate	2014-12-01 11:03:33
date_created	2014-01-01 22:10:52
date_modified	2014-01-01 22:10:52

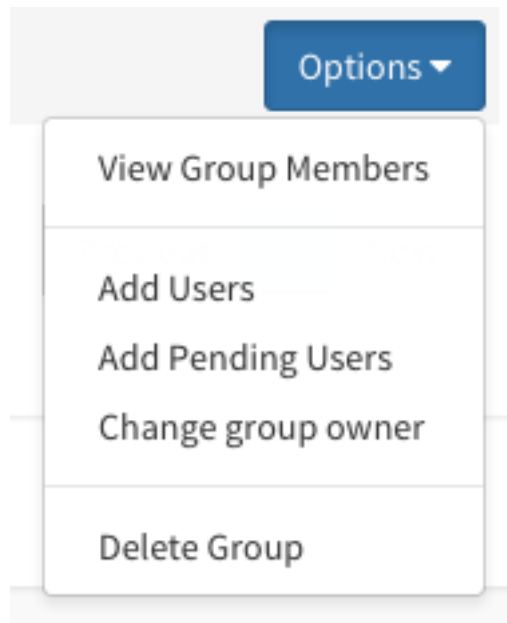
This section will allow you to see various information about your user.

If you select the ‘Groups’ tab, you’ll then be shown a table with all the groups that you are a part of.

Group Information			
<div>20 entries per page</div>		<div>Search here...</div>	
ID	Group Name	Date Created	Options
1	umw_biocore	2014-12-22 20:22:07	Options
82	pilot_EC	2014-12-22 20:22:07	Options
<div>Create a New Group Join a Group</div>		<div>Previous 1 Next</div>	

You have the option of creating groups and requesting to join groups at the bottom left corner of this tab.

Additionally, if you are the owner of a specific group, you have some additional options to the left of that specific group name.



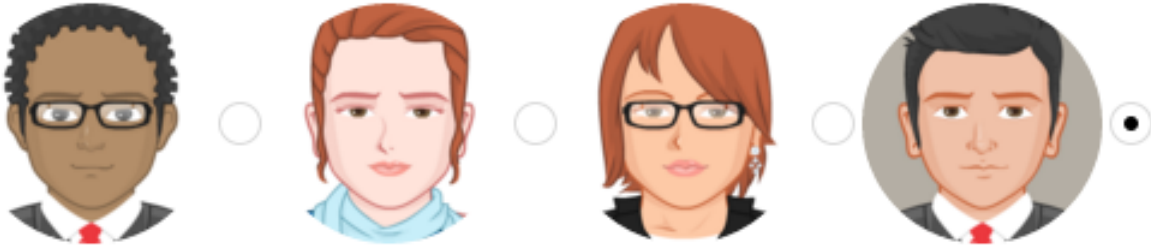
You have the option to view users who have a pending request to join your group, you can also view all of the current users.

If you so desire, you can even transfer the ownership of one group to another member within that group.

And lastly you can also delete the group as well.

The next tab labeled 'Photo' simply allows you to change your stock photo seened at the top left of your screen.

Update your profile picture



The fourth tab allows you to edit your listed email address and decide whether or not we have permission to send you an email upon a run completing.

Update your email settings

<p>Email address</p> <input type="text" value="example@umassmed.edu"/>	<p>Send email for runs?</p> <input type="text" value="No"/>
--	---



The last tab labeled 'Amazon Information' displays your groups current AWS key information. You will only have full access to editing and viewing the key information if you are the owner of the group

Update your Amazon Buckets

20 entries per page

Search here...

Group	Access Key	Secret Key
umw_biocore	*****MFHA	*****W9Bf

Previous 1 Next

Submit Add Amazon

Whenever you have finished making changes to any of these tabs, make sure to select the 'Update' button if one is presented.

3.12 Dolphin ENCODE Submission Guide

3.12.1 Getting Started

Not everyone can submit to ENCODE, you must first have contact with the ENCODE Consortium before considering to design, create, and analyze a project that you are willing to submit to ENCODE. Please visit the [ENCODE website](#) for more details on ENCODE. Not all that apply will be able to submit to ENCODE.

3.12.2 Metadata Collection/Input

Assuming you are to submit to ENCODE, you are going to want to collect a various amount of metadata that surrounds your project. This is a crucial step in submitting to encode and without the proper metadata for submission submission will not be possible.

As a curator for your experiment, you are going to want to collect metadata from all stages of the experiment. This includes:

- Protocol information within the wet lab
- Sequencer and sequencing information
- Analysis programs and versions used
- Outline/Goal of your analysis process

These bullet points encompass a large amount of information, however if you keep proper track of all of this information, this will make submission to ENCODE a smooth process.

More information about which fields are required and which fields you should consider keeping track of can be found at the ENCODE website listed above. Additional information can be found by contacting someone from ENCODE directly, or by keeping in contact with your assigned data wrangler.

3.12.3 ENCODE Submission Process

The Breakdown

In order for ENCODE to retain the metadata you wish to send them, it must be passed to their servers using JSON objects. A JSON object is a form of organizing data that is easy to read and to create. Encode stores information passed by users within linked json objects using aliases, uuids, and accession numbers.

You will be passing JSON objects created from the metadata you have gathered and input into the Dolphin system in order to represent your experiment and the files you will be sending to encode. For more information on what a JSON object is, you can visit [this link](#) for more information.

The Detailed Version

There are 2 major submission phases that include multiple sub-steps along the way. The first phase we shall call the 'overall metadata submission' step. For this phase, you will be submitting metadata that encompasses your overall experiment. For this, we submit 7 specific JSON objects in a specific order:

- Donor JSON
- Experiment JSON
- Treatment JSON
- Biosample JSON
- Library JSON

- Antibody JSON
- Replicate JSON

We submit these JSON objects from top to bottom due to the fact that some of the JSONs require the accession number or alias of a previous step. Without this unique identifier, the JSONs cannot be properly linked and will thus not be correct.

Once the overall metadata step has completed successfully, you can then enter the ‘file data’ submission phase. During this phase you will be submitting file metadata, followed by the actual file, given that the file metadata has been successfully submitted. So for each possible file you want to submit you will send:

- File JSON
- The actual file

Once you have completed all of these phases successfully, your submission to ENCODE will be complete and the next step would be to have your data wrangler look over your submission for validation.

3.12.4 Specific Linkages

Many objects will have more than one direct link to other various objects within your submission. For instance, some samples will come from the same donor and time point, however one could be an ATAC-Seq analysis while the other could be RNA-Seq analysis. These samples would be sharing the same biosample and need to use the same biosample accession number when submitting experiments for them. Biological replicates will also have to share the same experiment accession number when reporting to ENCODE. Dolphin has a way to link specific biosamples and experiments before submission, which will be discussed further in this guide.

It’s important to note that the structure of submission differs between human and non-human samples. For one, the donor for humans refers to the specific human donor, while for other organisms it refers to the strain of the organism. Biosamples from specific time points will also have to reference a parent biosample using the ‘derived_from’ key, thus needing the accession number of that parent biosample for proper submission.

3.12.5 Metadata Objects

Diving deeper into each metadata object being passed, specific metadata will be used to create each JSON. Listing each JSON in order, we include:

- **Donors**
 - *Information gathered with experiment series*
 - * “award”:’grant’
 - * “lab”:’lab’
 - *Information gathered with samples*
 - * “organism”:’organism’
 - * “life_stage”:’life_stage’
 - * “age”:’age’
 - * “sex”:’sex’
- **Experiments**
 - *Information gathered with experiment series*
 - * “award”:’grant’

- * “lab”:’lab’

- *Information gathered with protocols*

- * “assay_term_name”:’assay_term_name’

- * “assay_term_id”:’assay_term_id’

- *Information gathered with samples*

- * “biosample_term_name”:’biosample_term_name’

- * “biosample_term_id”:’biosample_term_id’

- * “biosample_type”:’biosample_type’

- * “description”:’description’

- **Treatments**

- *Information gathered with treatments*

- * “treatment_term_name”:’treatment_term_name’

- * “treatment_term_id”:’treatment_term_id’

- * “treatment_type”:’treatment_type’

- * “amount”:’concentration’

- * “amount_units”:’concentration_units’

- * “duration”:’duration’

- * “duration_units”:’duration_units’

- **Biosamples**

- *Information gathered with experiment series*

- * “award”:’grant’

- * “lab”:’lab’

- *Information gathered with protocols*

- * “starting_amount”:’starting_amount’

- * “starting_amount_units”:’starting_amount_units’

- *Information gathered with samples*

- * “biosample_term_name”:’biosample_term_name’

- * “biosample_term_id”:’biosample_term_id’

- * “biosample_type”:’biosample_type’

- * “organism”:’organism’

- * “derived_from”:’biosample_derived_from’

- * “source”:’source’

- *Information gathered with lanes*

- * “date_obtained”:’date_received’

- **Libraries**

- *Information gathered with experiment series*

- * “award”:’grant’

- * “lab”:’lab’

- *Information gathered with samples*

- * “spike-ins”:’spike_ins’

- * “size_range”:’avg_insert_size’

- *Information gathered with protocols*

- * “nucleic_acid_term_name”:’nucleic_acid_term_name’

- * “nucleic_acid_term_id”:’nucleic_acid_term_id’

- * “extraction_method”:’extraction_method’

- * “crosslinking_method”:’crosslinking_method’

- * “fragmentation_method”:’fragmentation_method’

- **Antibodies**

- *Information gathered with experiment series*

- * “award”:’grant’

- * “lab”:’lab’

- *Information gathered with Antibodies*

- * “source”:’source’

- * “product_id”:’product_id’

- * “lot_id”:’lot_id’

- * “host_organism”:’host_organism’

- * “targets”:’targets’

- * “clonality”:’clonality’

- * “isotype”:’isotype’

- * “purifications”:’purifications’,

- * “url”:’url’

- **Replicates**

- *Information gathered with samples*

- * “biological_replicate_number”:’biological_replica’

- * “technical_replicate_number”:’technical_replica’

To better understand these lists, Let us break them down into how you should be reading them.

- **First Layer**

- **Second Layer**

- * Third Layer

- * Third Layer

- **Second Layer**

- * Third Layer

* Third Layer

The first layer describes which JSON object we are currently creating. The second layer describes which major table from the database that the information will be gathered from. The third layer explains two things, first the JSON field that will act as the key to the object and the second is the field from the database we are gathering to insert as the actual metadata.

3.12.6 File Objects

File metadata submission will have some similar fields and some different. The JSON fields are all dependent on the file type that is being submitted. The JSON object passed for each file type will have the following fields:

- **Fastq files**

- “file_format”:”fastq”
- “run_type”:’run_type’
- “step_run”:’step_run’
- “paired_end”: ‘1’ OR ‘2’
- “output_type”:”reads”
- “read_length”:’read_length’
- “paired_with”:<paired-end alias>
- “derived_from”:<alias of derived file> AND/OR ‘additional_derived_from’

- **BAM files**

- “file_format”:”bam”
- “run_type”:’run_type’
- “step_run”:’step_run’
- “output_type”:”alignments”
- “assembly”:’genome’
- “derived_from”:<aliases of derived files> AND/OR ‘additional_derived_from’

- **TSV files**

- “file_format”:”TSV”
- “run_type”:’run_type’
- “step_run”:’step_run’
- “output_type”:”gene quantifications” OR “transcript quantifications”
- “assembly”:’genome’
- “derived_from”:<aliases of derived files> AND/OR ‘additional_derived_from’

- **bigWig files**

- “file_format”:”bigWig”
- “run_type”:’run_type’
- “step_run”:’step_run’
- “output_type”:”signal of all reads”

- “assembly”:’genome’
- “derived_from”:<aliases of derived files> AND/OR ‘additional_derived_from’
- **bed files**
 - “file_format”:’bed’
 - “run_type”:’run_type’
 - “step_run”:’step_run’
 - “output_type”:’peaks’
 - “assembly”:’genome’
 - “file_format_type”:’narrowPeak’
 - “derived_from”:<aliases of derived files> AND/OR ‘additional_derived_from’

As a reference to the above list, every left field will be the actual key to the JSON and the right field will be the value associated to that key. Values wrapped in double quotes are preset string values, such as file_format’s value being “fastq” for a fastq file, while values wrapped in single quotes are values gathered from the database under that field name. Values defined with the “<>” markers are referencing aliases created using the sample name, lab, step, and file format.

In addition to those fields, each JSON passed will additionally contain these fields:

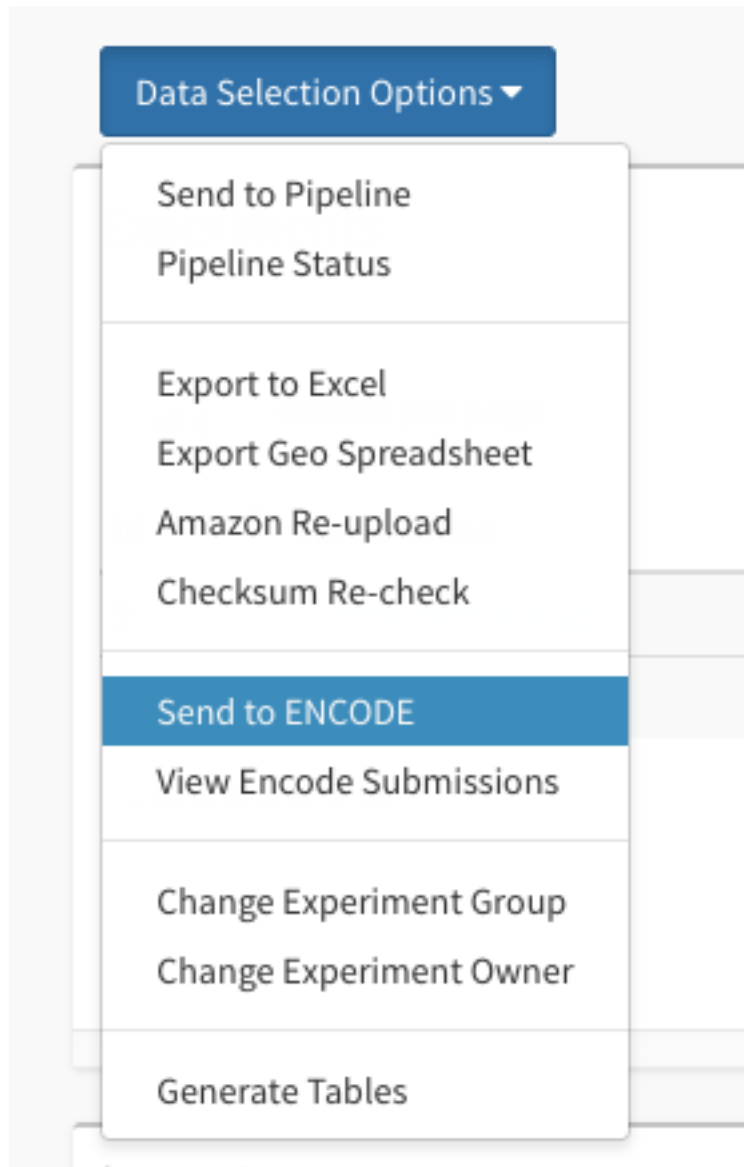
- “dataset”:’experiment_acc’
- “replicate”:’replicate_uuid’
- “file_size”:<byte size of file>
- “md5sum”:<md5sum of file>
- “platform”:’platform’
- “submitted_file_name”:<name of file>
- “lab”:’lab’
- “award”:’grant’
- **“flowcell_details”:**
 - “machine”:’machine_name’
 - “flowcell”:’flowcell’
 - “lane”:’lane’

In this case, any value field marked with the “<>” markers are defined based on the file you are submitting. Dolphin creates files based on sample name, so manual entry is not required.

Upon successfully submitting file metadata, ENCODE will pass back a JSON response with Amazon credentials. These credentials are then used to upload the specific file to their secure Amazon databank.

3.12.7 The Submission Process

As stated above, submitting to ENCODE through Dolphin is easy as long as you have all the proper metadata. You don’t have to worry if you didn’t input all of the information needed during the import since you will be able to edit all of the fields before submission.



In order to start the ENCODE submission process, you are going to want to select the samples you wish to submit within the NGS Browser section of Dolphin. Once you've selected your samples you then can click the 'Data Selection Options' button at the top of the page and select the 'Send to ENCODE' option. This will take you to the ENCODE submission page.

Encode Viewing/Submission Projects and experiments submitted to Encode

Sample Selection	Donors	Experiments	Treatments	Biosamples	Libraries	Antibodies	Replicates	Files
------------------	--------	-------------	------------	------------	-----------	------------	------------	-------

Loading in, you will see a variety of tabs and tables. Each tab represents submission of a specific JSON object and each contains a table with information loaded based on the samples that you have selected. The 'Sample Selection' tab allows you to view the samples you have selected in the top table and add more samples from the bottom table. You can also remove samples by clicking the red 'X' button in the top table for each sample, or manually deselecting the sample from the bottom table. You can also edit specific metadata within each table shown by simply clicking the field you wish to edit and pressing enter when you are finished. It should be noted that not all fields are editable. In addition to editing samples, you can also edit all of your selected samples for that field at once or on a selection basis. Simply

click on the field you wish to edit, edit that field, and click the ‘Change All’ button to change all of your samples for that field at once. You can select specific samples on the right and then click the ‘Change Selected’ in order to change multiple samples at a time.

Selected Samples ?

Show 10 entries

Search:

id	Sample Name	Donor	Organism	Molecule	Lab	Grant	Removal	Selected
9	control_rep3	D25	human	RNA-Seq	manuel-garber		<div><div></div></div>	<input type="checkbox"/>

Showing 1 to 1 of 1 entries

Previous

1

Next

Selected Samples ?

Show 10 entries

Search:

id	Sample Name	Donor	Organism	Molecule	Lab	Grant	Removal	Selected
8	control_rep2	D25	human	RNA-Seq	manuel-garber	U01HG007910		<input type="checkbox"/>
9	control_rep3	D25	human	RNA-Seq	manuel-garber	U01HG007910		<input checked="" type="checkbox"/>
10	exper_rep1	D25	human	RNA-Seq	manuel-garber	U01HG007910		<input checked="" type="checkbox"/>

Showing 1 to 3 of 3 entries

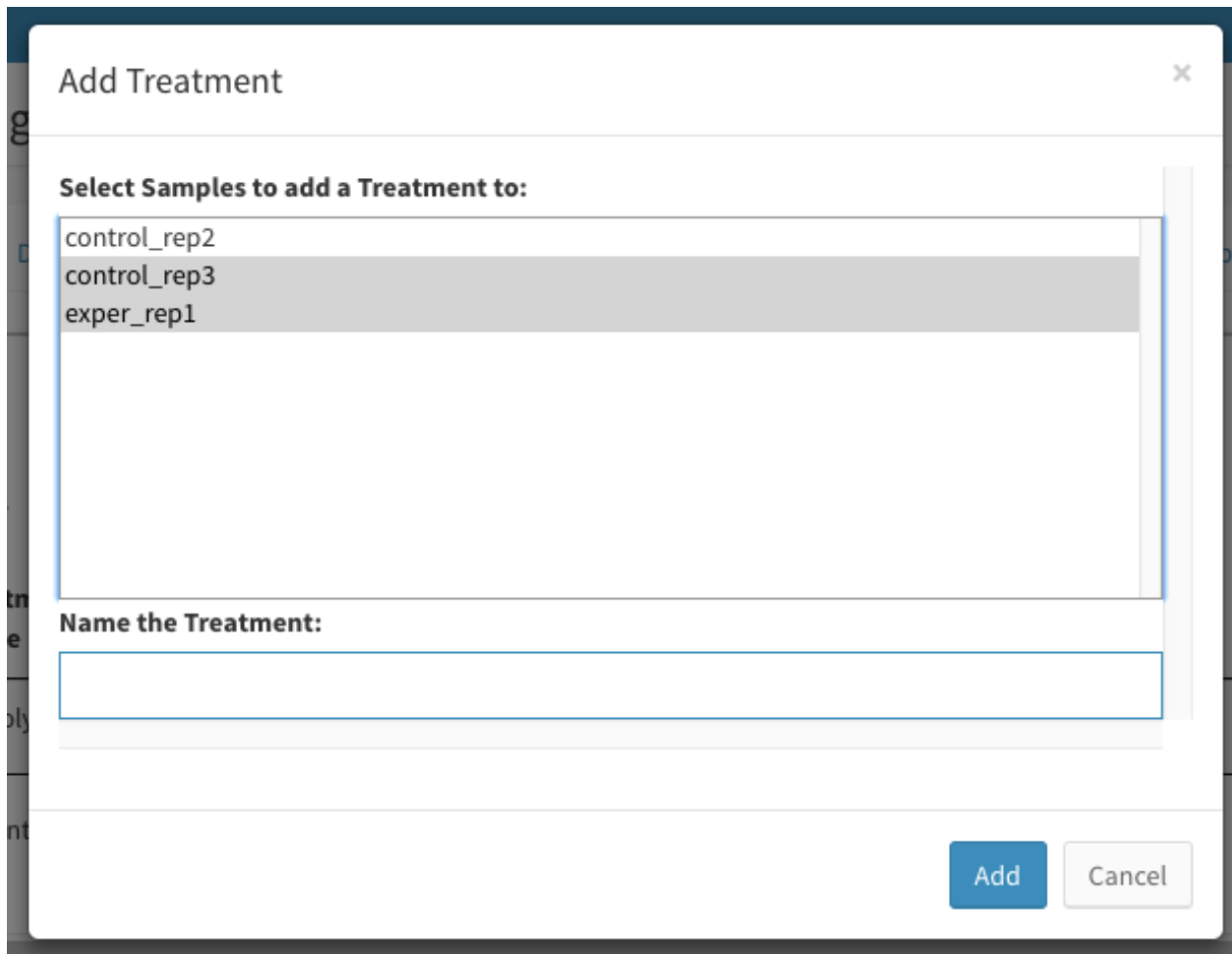
Previous

1

Next

[Change Selected](#)
[Change All](#)

If no treatments are linked to your samples you can create them within the ENCODE submission page on the ‘Treatments’ tab. To create a treatment click on the ‘Add Treatment’ button and a dialog box will pop up. In this dialog box you will select which samples will have this treatment, as well as the name of the treatment. Once you are finished you can click the ‘Add’ button to add the treatment link for your samples and start editing the treatment metadata as you see fit. This same adding and editing strategy applied to the ‘Antibodies’ tab as well.



Add Treatment

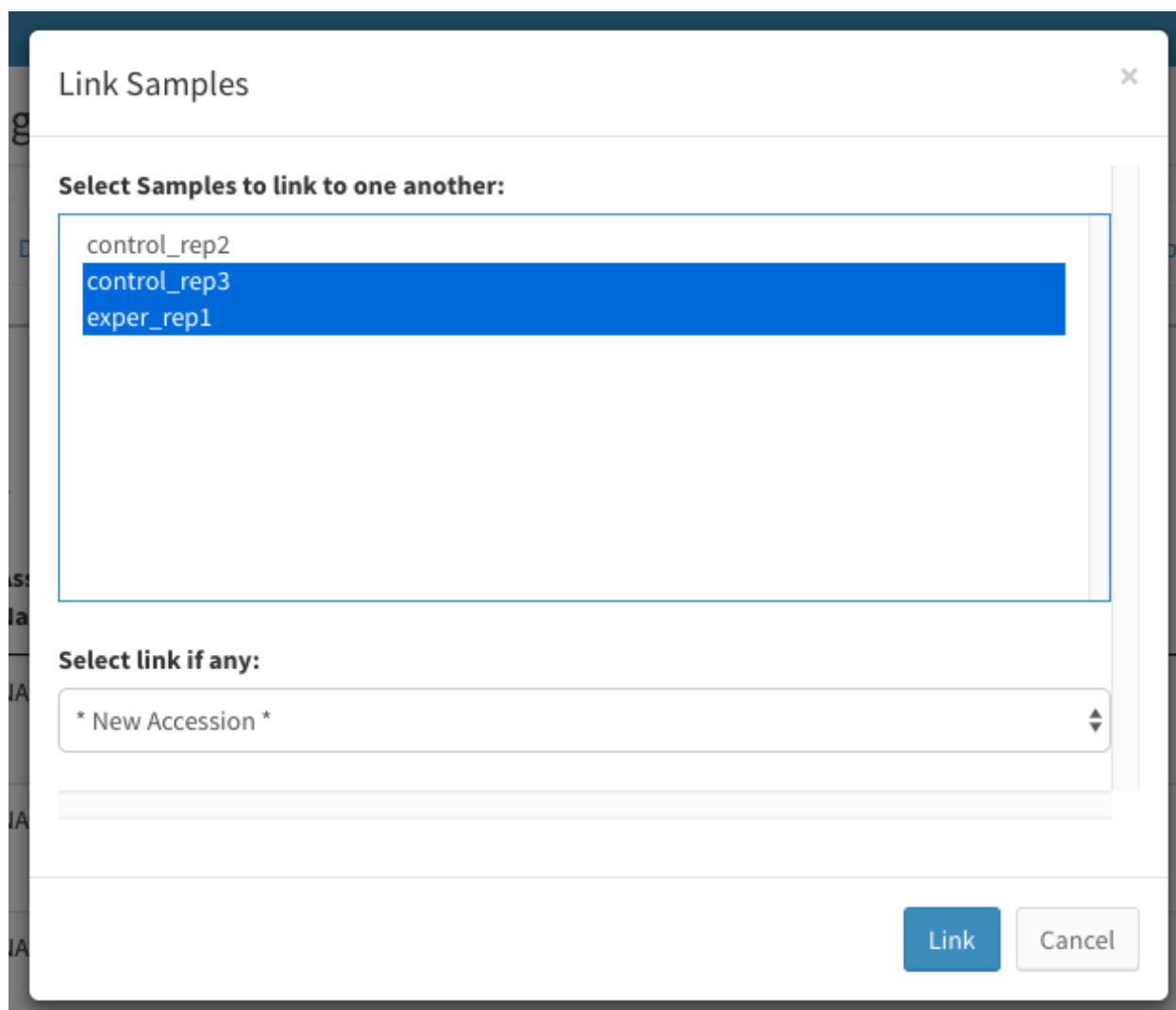
Select Samples to add a Treatment to:

- control_rep2
- control_rep3
- exper_rep1

Name the Treatment:

Add Cancel

As stated above, you can also link biosamples and experiments together via samples in the 'Biosamples' or 'Experiments' tabs. Simply click on the 'Link Biosamples' or 'Link Experiments' tab to bring up a dialog box displaying all of the samples you have selected. From there, select the samples you wish to link together, and then select which accession number you wish to link them under. If no accession number is present, you can select '* New Accession*' to link the samples together without an accession number present.



The 'Files' tab will be the last tab to be filled out before submitting your metadata and files. If you have filled out all of your metadata within the tables, you can go ahead and send the overall metadata without the file metadata and files by selecting the 'Submit Meta-data' button in the bottom left. Within the 'Files' tab there are 4 major components: run selection, file selection, submission order, and previous files submissions.

Run Selection

Show entries

Search:

Sample	Run Selection	Run Directory
control_rep2	encode test tophatsem	/export/out/encode/tophatsem
control_rep3	encode test tophatsem	/export/out/encode/tophatsem
exper_rep1	encode test tophatsem	/export/out/encode/tophatsem

Showing 1 to 3 of 3 entries

Previous **1** Next

File Selection

- /seqmapping/rna/ fastq
- /seqmapping/mirna/ fastq
- /seqmapping/trna/ fastq
- /seqmapping/snRNA/ fastq
- rsem/ bam
- rsem/genes tdf
- rsem/isoforms tdf
- /ucsc_rsem/ bigWig
- /tophat/ bam
- /ucsc_tophat/ bigWig

Submission Order

Show entries

Search:

ID	Parent File ID	File Location	File Type	Step Run	Additional Derived From
No data available in table					

Showing 0 to 0 of 0 entries

Previous Next

Previous Files Submitted

Show entries

Search:

Sample	Run	File	Parent Step	File Acc	File UUID
No data available in table					

Showing 0 to 0 of 0 entries

Previous Next

The run selection section should be looked at first. For each sample you need to determine which dolphin run for each sample has the analysis files you wish to use. Once you've selected all the proper run information, the file selection section will be next. Here you will be able to select multiple file types to submit to ENCODE. Only files produced between all selected runs will be able to be selected. After file type selection, we then can determine the order and the hierarchy of all the files being submitted as well as any other additional information that may need to be provided. Step run and additional derived from parameters should be supplied by your data wrangler once they have created your analysis pipeline objects within the ENCODE system.

The last table, previous file submissions, simply displays previous submissions from the samples that have been selected.

Once all the appropriate information has been filled out, you're ready to submit to ENCODE. Press the submit button which you desire on the bottom left of the screen and confirm the submission to send the data!











3.12.8 Reading the Output

Output for the submission process will take place in a few different locations. For one, you'll be able to view your output if you have access to editing ENCODE information on the main site using the accession numbers given back to the user. However, through dolphin there are two main location as of currently where you will be able to view the output. The submission dialog box will report a condensed version of the JSON submission output to the user and the raw JSON output for the file as well. Additional logs are kept within the dolphin system in the tmp/encode directory via your user name and the time you submitted to ENCODE.









3.12.9 ENCODE Submissions Page

If you click on the bottom right button labeled ‘View Encode Submissions’, you will be taken to a new page with some information about your previous submissions.

ID	Samples	Submission Status	Log File	Resubmission
1	8		docker_2017-01-06-10-23-17.log	Resubmit
2	8		docker_2017-01-06-10-36-17.log	Resubmit
3	9		docker_2017-01-06-10-45-57.log	Resubmit
4	9		docker_2017-01-06-11-25-57.log	Resubmit
5	10		docker_2017-01-06-11-27-45.log	Resubmit
6	11		docker_2017-01-06-11-47-52.log	Resubmit
7	11		docker_2017-01-06-13-23-14.log	Resubmit
8	11		docker_2017-01-06-13-31-12.log	Resubmit
9	11		docker_2017-01-06-13-42-52.log	Resubmit
10	18		docker_2017-01-23-09-42-11.log	Resubmit

Showing 1 to 10 of 24 entries

Previous **1** 2 3 Next

ID	Samples	Submission Status	Log File
1	control_rep2		docker_2017-01-06-10-36-17.log
2	control_rep3		docker_2017-01-06-11-25-57.log
3	exper_rep1		docker_2017-01-06-11-27-45.log
4	exper_rep2		docker_2017-01-06-13-42-52.log
5	berktest6		docker_2017-01-23-12-10-08.log
6	berktest5		docker_2017-01-24-13-57-57.log

Showing 1 to 6 of 6 entries

Previous **1** Next

There are two tables present on this page, the ‘Encode Batch Submissions’ table and the ‘Encode Sample Submissions’ table. The first table, the encode batch submissions table, displays all of the group submissions and the sample numbers

that were submitted together. The submission status column denotes whether or not metadata from all samples is still up to date; green denoting that the metadata is up to date and red showing that it has changed since that submission. In addition the data log's name is displayed for an admin to view until other options have been implemented. You can also choose a sample selection to re-submit which will select all the samples from that submission and send you to the ENCODE submission pipeline. The bottom table shows the exact same information as the top table except it is on a per-sample scale.

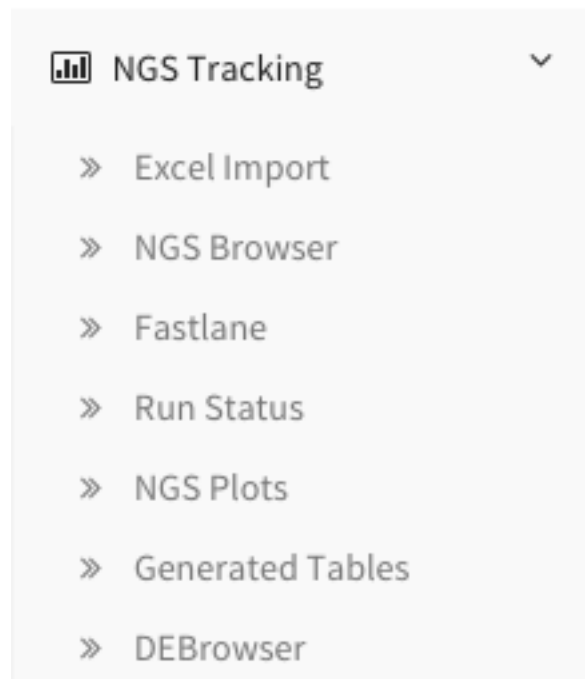
3.13 DEBrowser Access Guide

This guide is a quick guide for the DEBrowser within dolphin.

3.13.1 Accessing DEBrowser

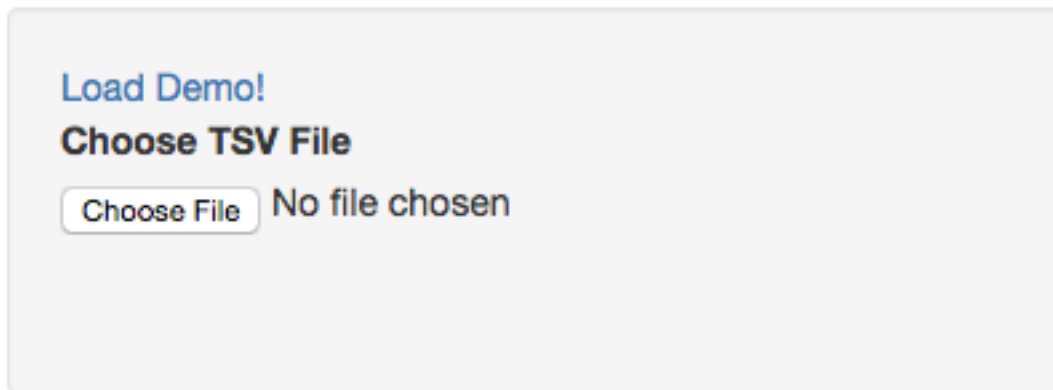
First, make sure to have an instance of dolphin available (see Dolphin Docker) as well as an account for the dolphin interface.

Once logged in, click on the 'NGS Tracking' tab on the left, then click on 'DEBrowser'.



This will bring you to the DEBrowser section within dolphin. This will take you to the base DEBrowser page where you can either try the demo by clicking the 'Load Demo!' to load the demo set we have set up or you can load in locally downloaded tsv files for analysis.

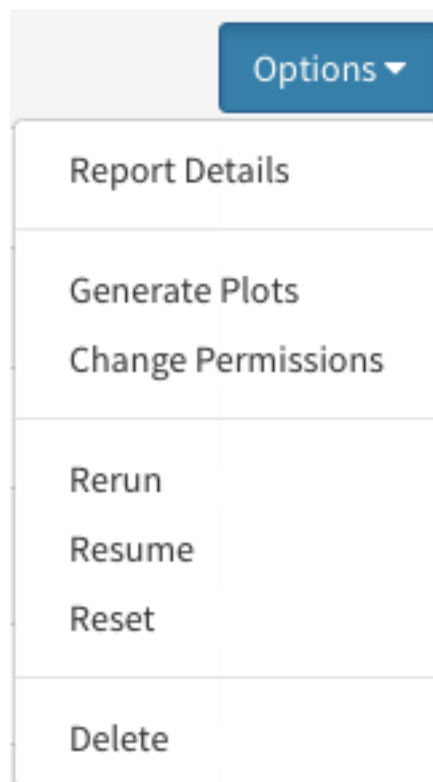
DE Browser



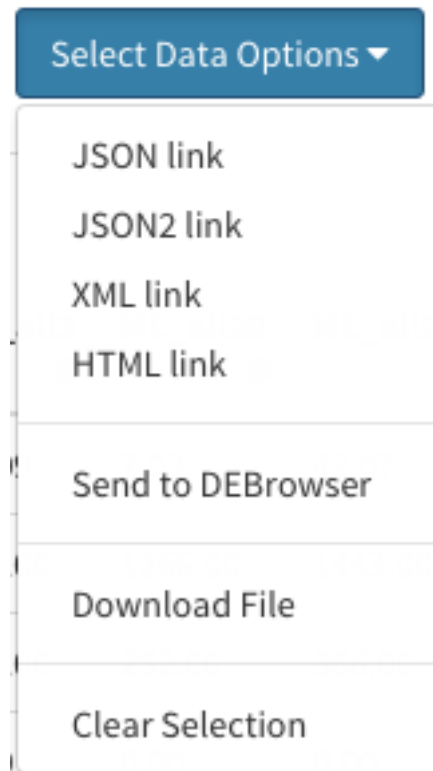
You can also load in RSEM results from runs with the corresponding results in a few easy steps.

From Reports:

You can view your RSEM results straight from your selected run's report page in a few easy steps. First select the run you would like to use and head to that run's report page within the NGS Run Status page.



Next, select the RSEM tab and then select the file you wish to use within the DEBrowser. Once you've selected the file, a dropdown menu will appear and you can then select the 'Send to DEBrowser' option.

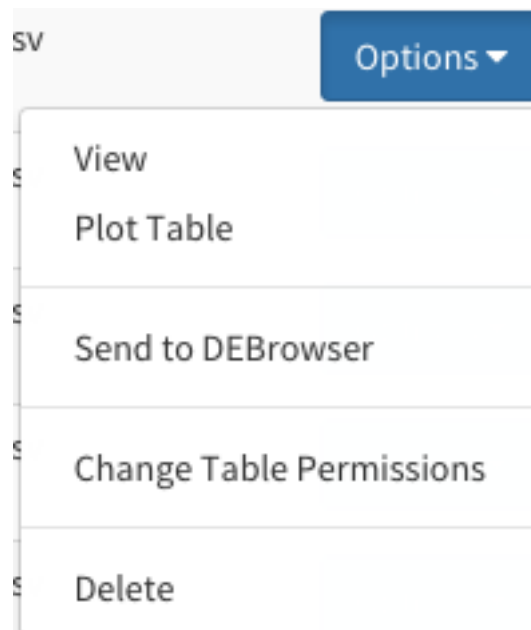


You will then be redirected to the DEBrowser page within dolphin with that data loaded into the browser.

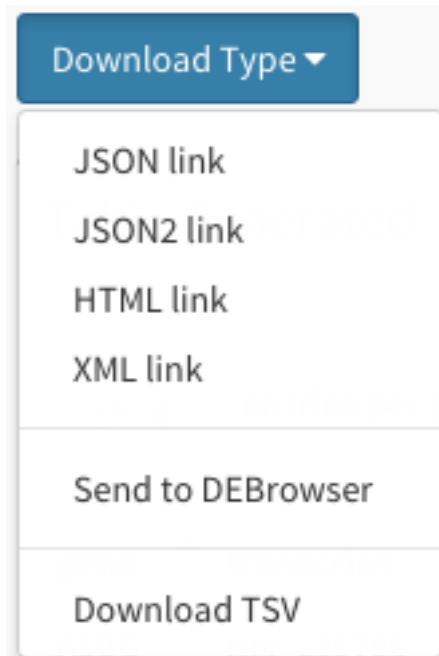
From Generated Tables:

Additionally, you can load custom-made rsem tables from the generated tables section within dolphin. Any generated table which uses a RSEM file can be used within the DEBrowser.

Simply head to the Generated Tables section within dolphin and select the table you wish to use. Then select the 'Send to DEBrowser' option and you will be redirected to the DEBrowser page within dolphin with that data loaded into the browser.



You can also send a table into DEBrowser while viewing the generated table. Just select the Download Type button and then select the ‘Send to DEBrowser’ option.



Additional Information:

Please be patient while the data loads into the Browser for it can take a few seconds before the data is fully loaded and you can run DESeq.

For a more detailed guide on DEBrowser itself, please visit the DEBrowser section.

3.14 Developer Implementation

3.14.1 Dolphin Integration

This section of the documentation is only meant for developers implementing a Dolphin system on their own personal system. This is not to be confused with Dolphin-Docker, which creates an instance of Dolphin using a Virtual Machine. If you are not implementing Dolphin within your own system, please ignore these notes.

3.14.2 Python Dependencies

Dolphin, as you may have noticed, uses python for some scripts in order to generate pipelines or secure data within the database. Here is a list of packages you will need in order to properly implement Dolphin within your system:

- pycrypto
- simple-crypt
- boto
- boto3
- MySQLdb
- ConfigParser

- optparse
- binascii
- subprocess

You can install these modules with the “pip install” feature of python.

Contents:

4.1 Quick-start Guide

This guide is walkthrough for the DESeq Browser from start to finish.

4.1.1 Getting Started

First off, we need to head to the DEBrowser webpage at this url:

<http://debrowser.umassmed.edu/>

Alternatively, if you have the R package installed, you can call these R commands:

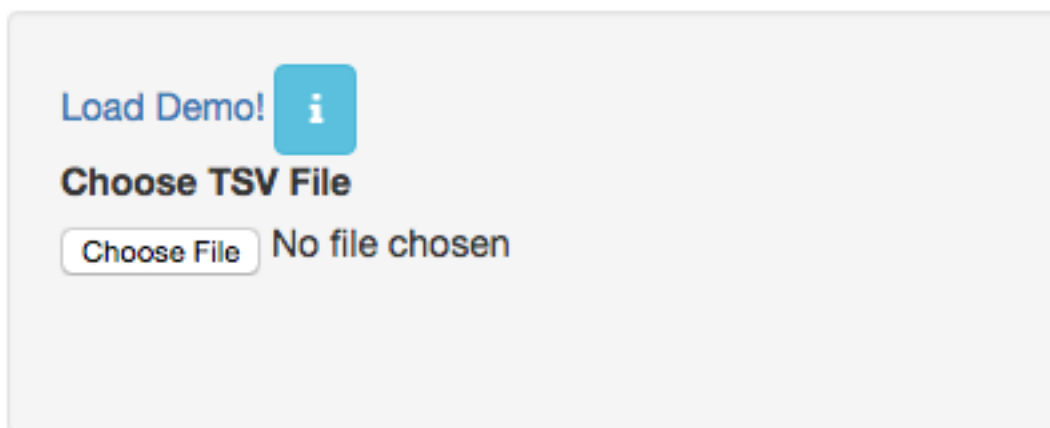
```
library(debrowser)
```

```
startDEBrowser()
```

For more information on installing DEBrowser locally, please consult our Local Install Guide.

Once you've made your way to the website, or you have a local instance of DEBrowser running, you will be greeted with this tab on the left:

DEBrowser



To begin the DESeq process, you will need to select your Data file (TSV format) to be analyzed using DESeq. If you do not have a dataset to use, you can select to use the built in demo by clicking on the ‘Load Demo!’. To view the entire demo data file, you can download this [demo set](#). For an example case study, try our [advanced demo](#).

The TSV files used to describe the quantification counts are similar to this:

IE:

gene	trans	exp1	exp2	cont1	cont2
DQ714	uc007	0.00	0.00	0.00	0.00
DQ554	uc008	0.00	0.00	0.00	0.00
AK028	uc011	2.00	1.29	0.00	0.00

DEBrowser also accepts TSV’s via hyperlink by following a few conversion steps. First, using the API provided by Dolphin, we will convert TSV into an html represented TSV using this website:

<http://dolphin.umassmed.edu/public/api/>

The Two parameters it accepts (and examples) are:

1. source=http://bioinfo.umassmed.edu/pub/debrowser/advanced_demo.tsv
2. format=JSON

Leaving you with a hyperlink for:

http://dolphin.umassmed.edu/public/api/?source=http://bioinfo.umassmed.edu/pub/debrowser/advanced_demo.tsv&format=JSON

Next you will need to encode the url so you can pass it to the DEBrowser website. You can find multiple url encoders online, such as the one located at this web address: <http://www.url-encode-decode.com/>.

Encoding our URL will turn it into this:

http%3A%2F%2Fdolphin.umassmed.edu%2Fpublic%2Fapi%2F%3Fsource%3Dhttp%3A%2F%2Fbioinfo.umassmed.edu%2Fpub%2Fdebrowser%2Fadvanced_demo.tsv%26format%3DJSON

Now this link can be be used in debrowser as:

<http://debrowser.umassmed.edu:443/debrowser/R/>

It accepts two parameters:


1. `jsonobject=http%3A%2F%2Fdolphin.umassmed.edu%2Fpublic%2Fapi%2F%3Fsource%3Dhttp%3A%2F%2Fbioinfo.umassmed.edu%2Fpub%2Fdebrowser%2Fadvanced_demo.tsv%26format%3DJSON`
2. `title=no`

The finished product of the link will look like this:

```
http://debrowser.umassmed.edu:443/debrowser/R/?jsonobject=http://dolphin.umassmed.edu/public/api/?source=http://bioinfo.umassmed.edu/pub/debrowser/advanced_demo.tsv&format=JSON&title=no
```

Inputting this URL into your browser will automatically load in that tsv to be analyzed by DEBrowser!

For more information about the input file, please visit our DESeq/DEBrowser tab within Readthedocs. Once you've selected your file and the upload has completed, you will then be shown the samples listed within your file uploaded as well as a few options.



The screenshot shows a web interface for DEBrowser. At the top, there is a tab labeled "Data Prep". Below it, a section titled "Samples" contains a horizontal list of sample names: "exper_rep1", "exper_rep2", "exper_rep3", "control_rep1", "control_rep2", and "control_rep3". Each name is enclosed in a light gray box. Below the list, there are three buttons: "Go to DE Analysis!", "Go to QC plots!", and "Reset Samples!".

The first option, 'Go to DE Analysis', takes you to the next step within the DESeq workflow. In order to run DESeq on your input data you first need to select which samples will go into your conditions. You can run multiple condition comparisons and view the results separately as well. To remove samples from a condition, simply select the sample you wish to remove and hit the delete/backspace key. To add a sample to a condition you simply have to click on one of the condition text boxes to bring up a list of samples you can add to that comparison. Click on the sample you wish to add from the list and it will be added to the textbox for that comparison.

Data Prep

Samples

exper_rep1 exper_rep2 exper_rep3 control_rep1 control_rep2 control_rep3

Reset Samples!

Please add new comparisons for DE analysis!

Condition 1

exper_rep1 exper_rep2 exper_rep3

Condition 2

control_rep1 control_rep2 control_rep3

Condition 3

exper_rep1 exper_rep2 exper_rep3

Condition 4

control_rep1 control_rep2 control_rep3

Condition 5

exper_rep1 exper_rep2 exper_rep3

Condition 6

control_rep1 control_rep2 control_rep3

Condition 1 Parameters:


DE Method: DESeq2, Fit Type: parametric, Beta Prior: 0, Test Type: Wald, row.sum filter: 10

Condition 3 Parameters:

DE Method: EdgeR, Normalization: TMM, Dispersion: 0, Test Type: exactTest, row.sum filter: 10

Condition 5 Parameters:

DE Method: Limma, Normalization: TMM, Fit Type: ls, Norm. Bet. Arrays: none, row.sum filter: 10

Add New Comparison Remove 

Submit!

The second option, ‘Go to QC plots!’, takes you to a page where you can view quality control metrics on your data input. The page opens with an all-to-all plot displaying the correlation between each sample. Left of this plot is a panel which contains various parameters to alter the look of your plot such as width and height. You can change the type of dataset being viewed within these QC plots by selecting the dataset you want at the top of the left panel. Each dataset can have its own unique parameters to change and alter your QC plots.

In addition to the all-to-all plot, you can also view a heatmap representation of your data as well as a Principal Component Analysis (PCA) plot by selecting the specific plot option on the left panel under ‘QC Plots’. You can also select the type of clustering and distance method for the heatmap produced to further customize your quality control measures.

DEBrowser

QC Plots:

- ☒ All2All
- ☐ Heatmap
- ☐ PCA

Submit!

width

100

700

2,000

100

290

480

670

860

1,050

1,240

1,430

1,620

1,810

2,000

height

100

500

2,000

100

290

480

670

860

1,050

1,240

1,430

1,620

1,810

2,000

corr font size

0.1

2

10

0.1

1.1

2.1

3.1

4.1

5.1


6.1

7.1

8.1

9.1


10

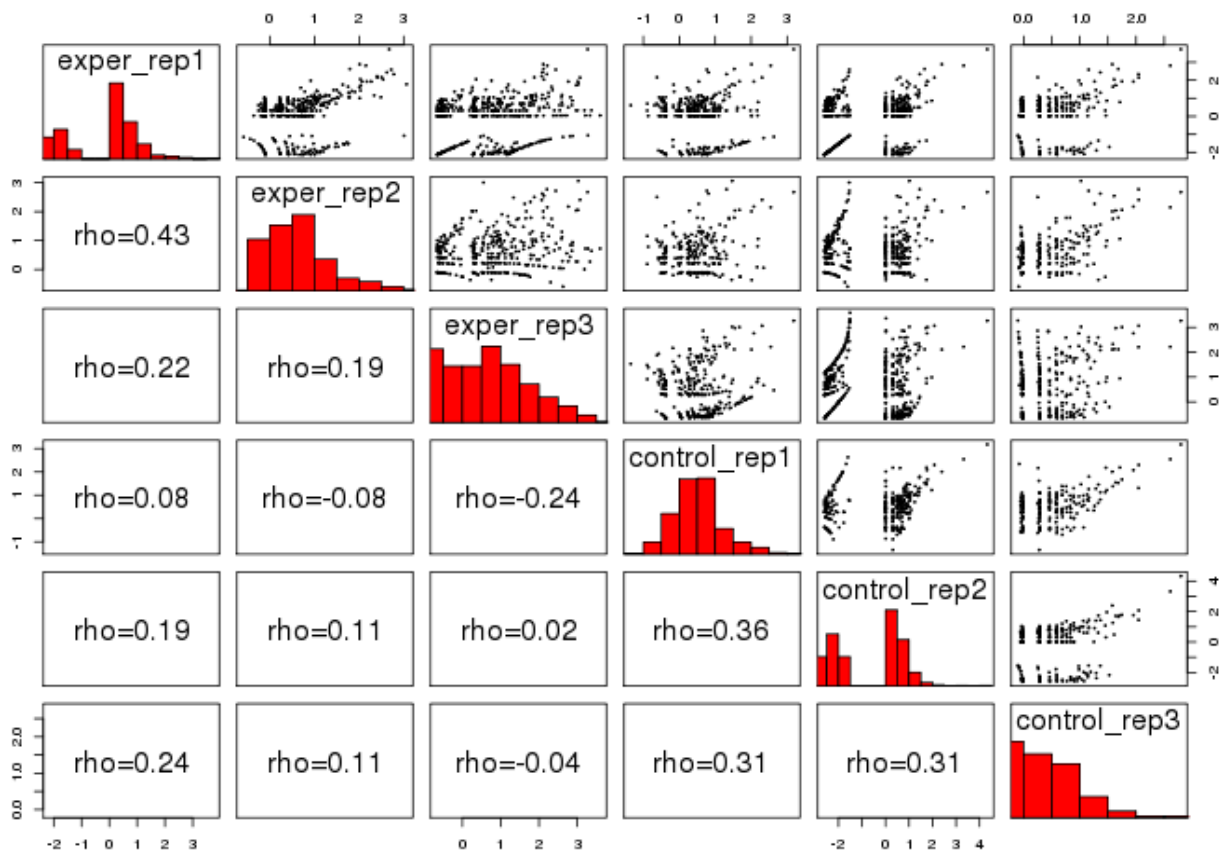
 Download Plot

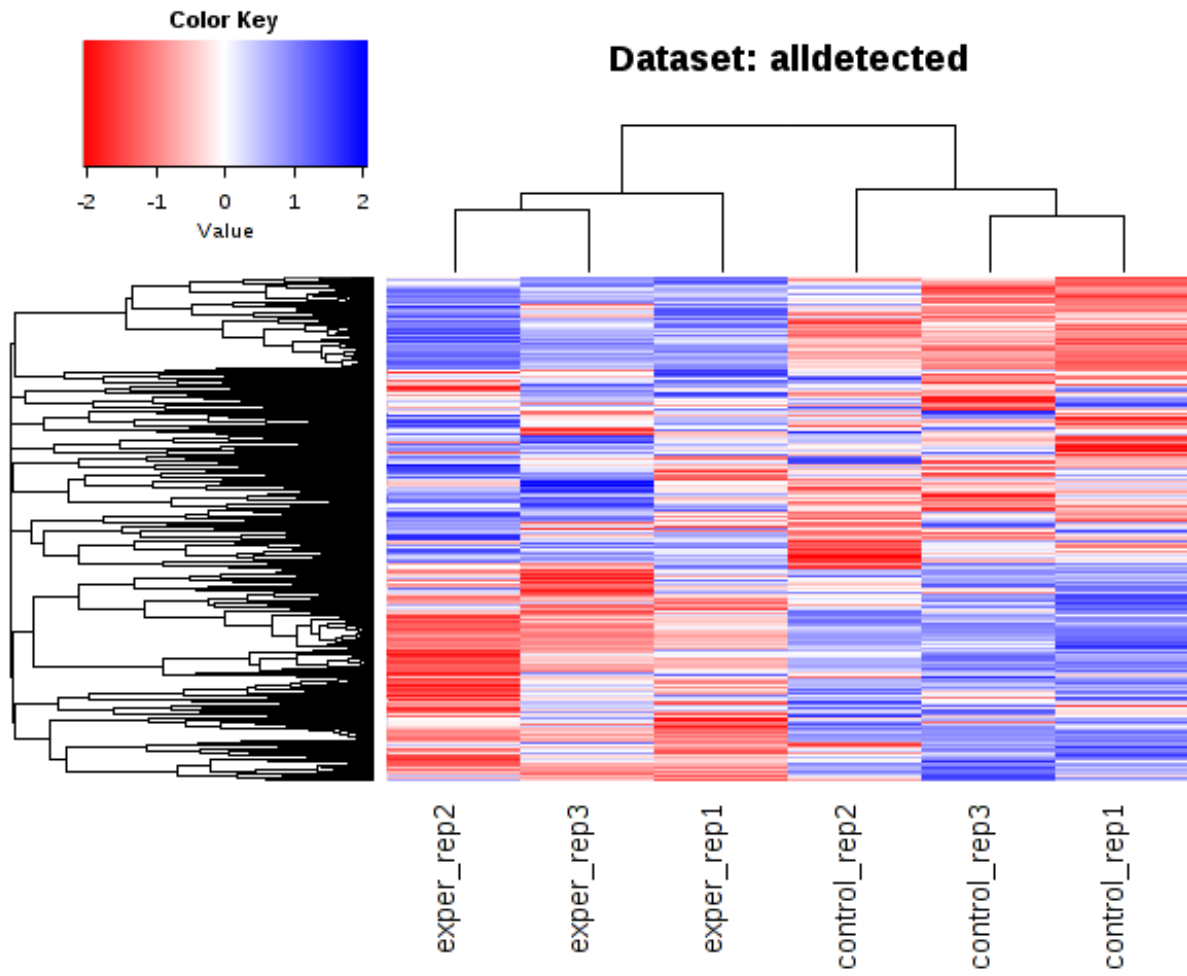
Choose a dataset:

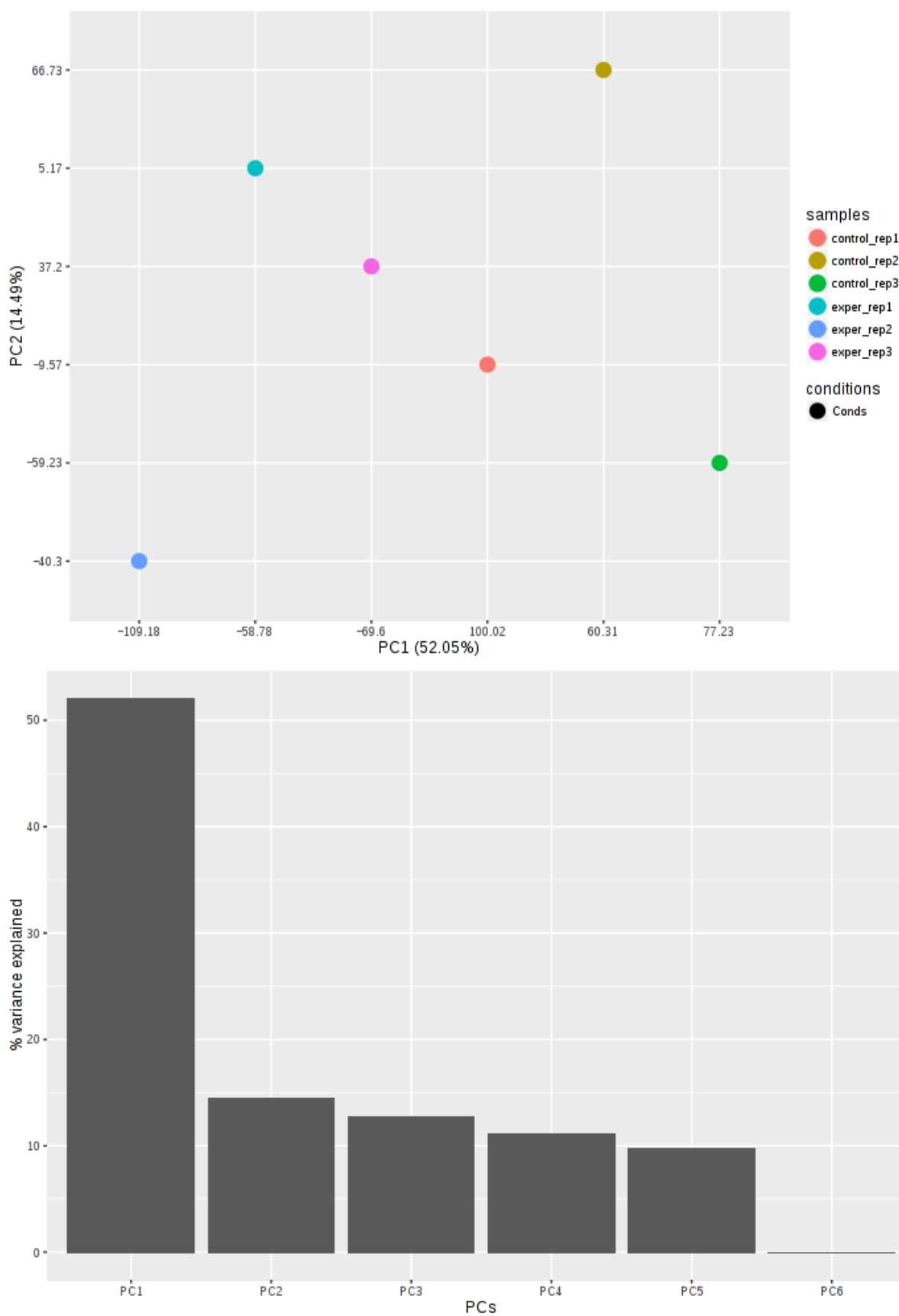
most-varied

▼

 Download Data





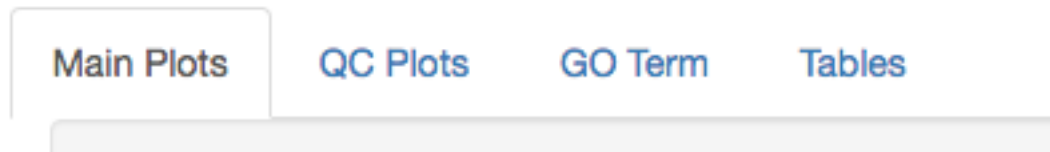


You can also view specific tables of your input data for each type of dataset available and search for a specific geneset by inputting a comma-separated list of genes or regex terms to search for in the search box within the left panel. To view these tables, you must select the tab labeled ‘Tables’ as well as the dataset from the dropdown menu on the left panel.

Once you are happy with your dataset and you have selected your conditions within the ‘DE Analysis’ section, you can then hit ‘Submit!’ to begin.

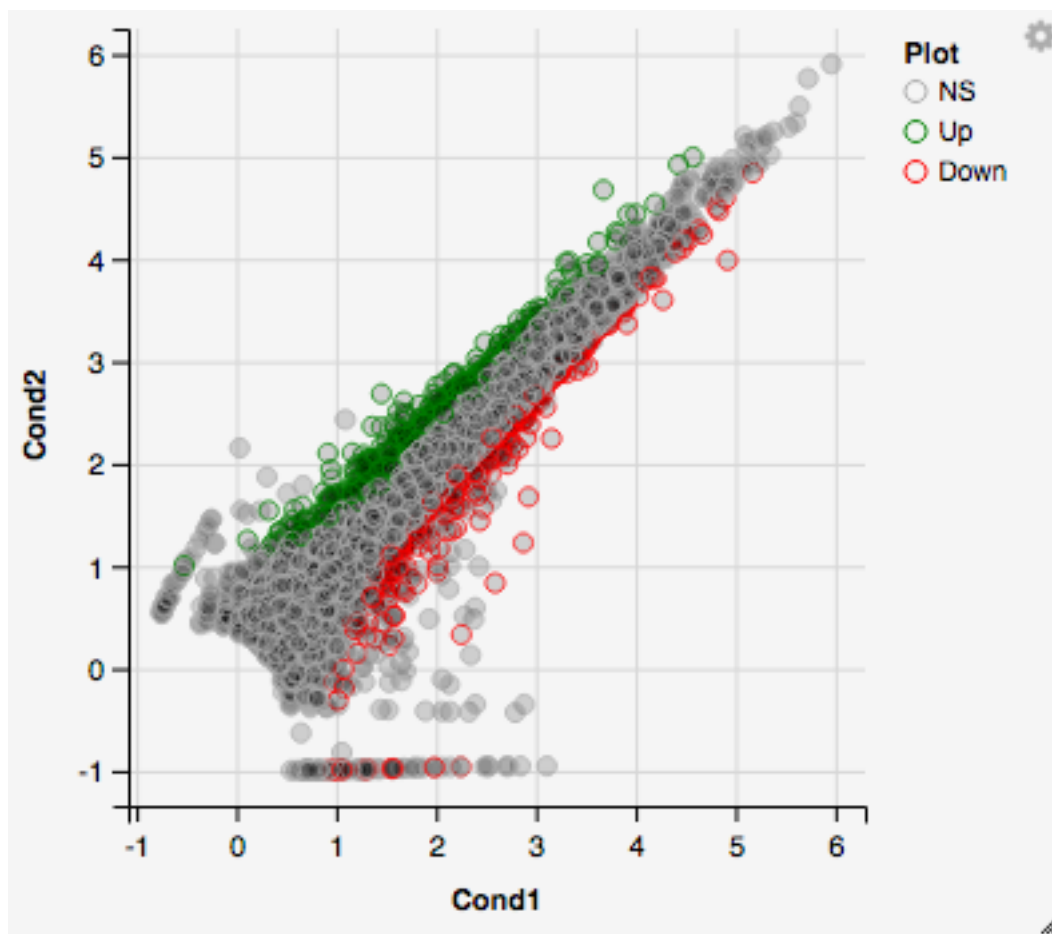
4.1.2 The Main Plots

After clicking on the ‘Submit!’ button, DESeq2 will analyze your comparisons and store the results into separate data tables. Shiny will then allow you to access this data, with multiple interactive features, at the click of a button. It is important to note that the resulting data produced from DESeq is normalized. Upon finishing the DESeq analysis, a tab based menu will appear with multiple options.

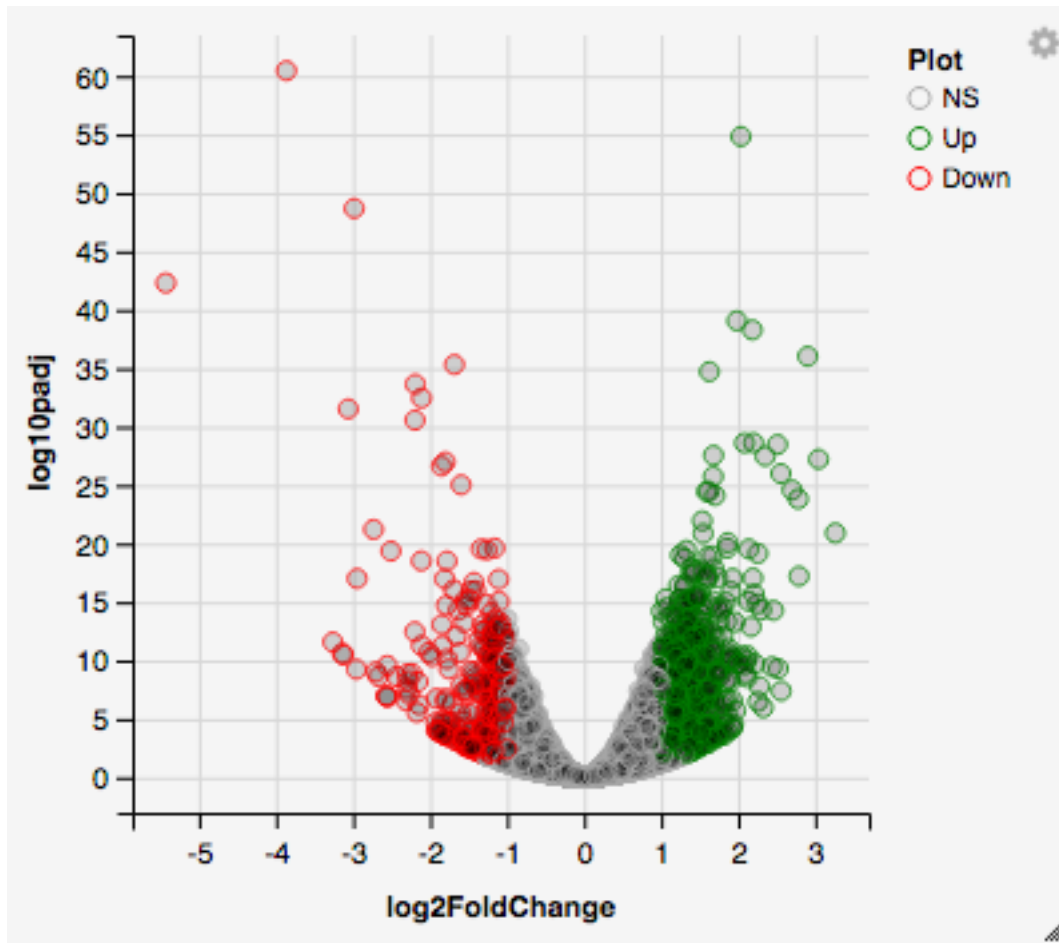


The first tab, the ‘Main Plots’ section, is where you will be able to view the interactive results plots. Plot choices include:

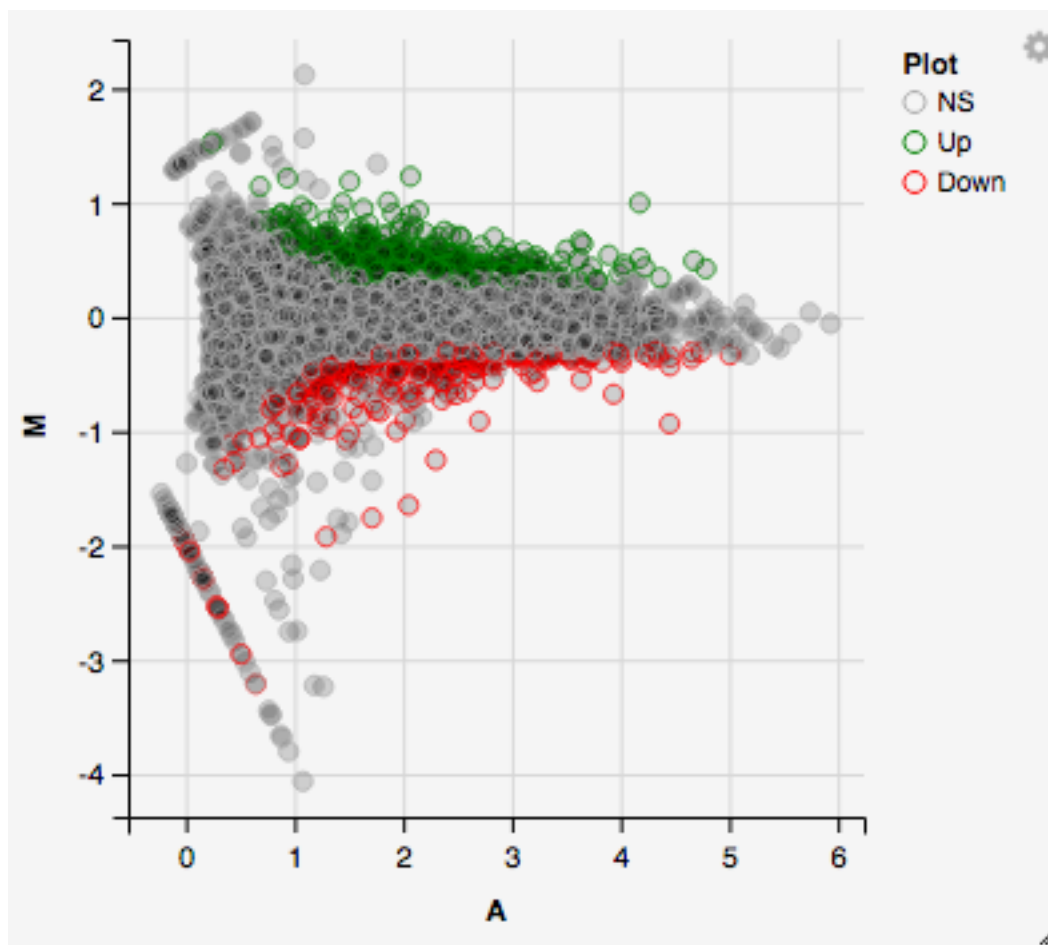
Scatter plot



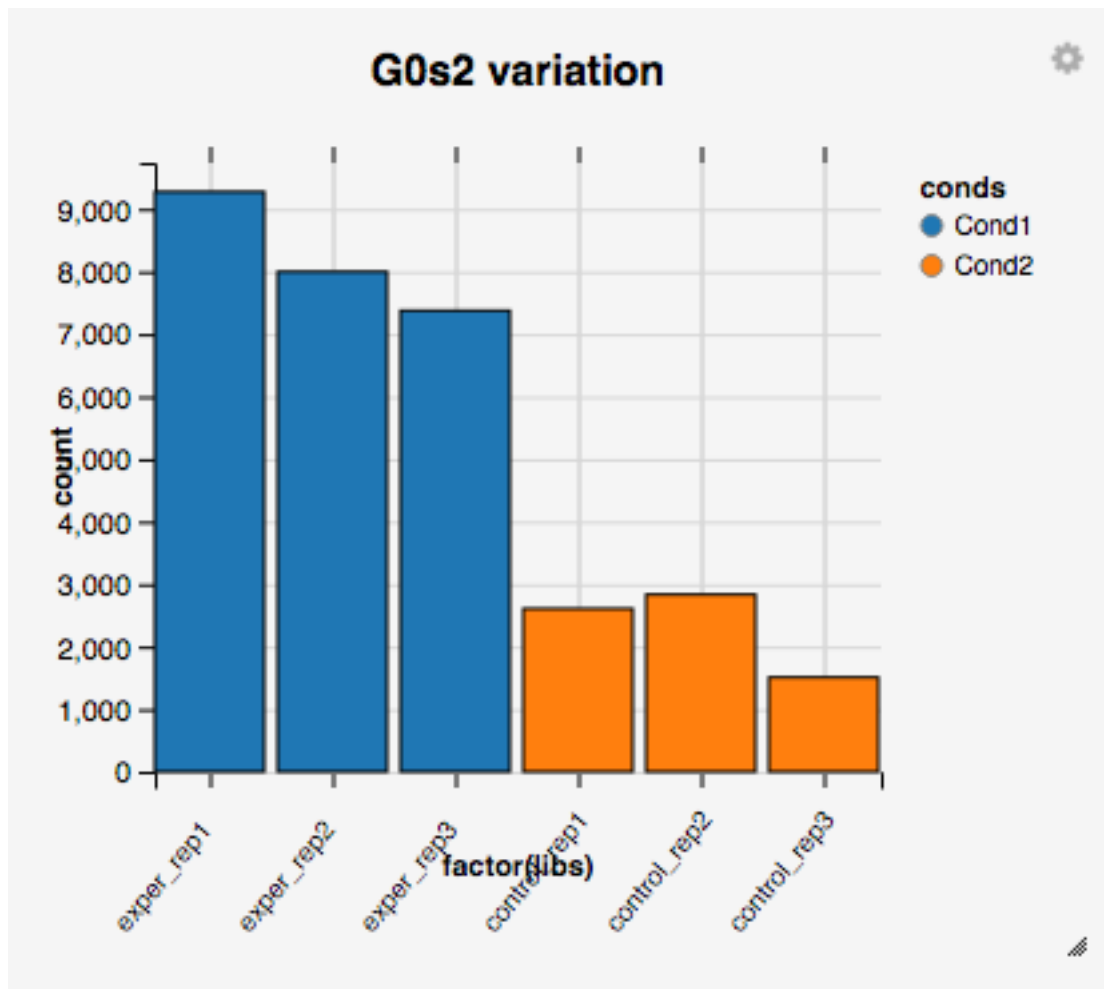
Volcano plot

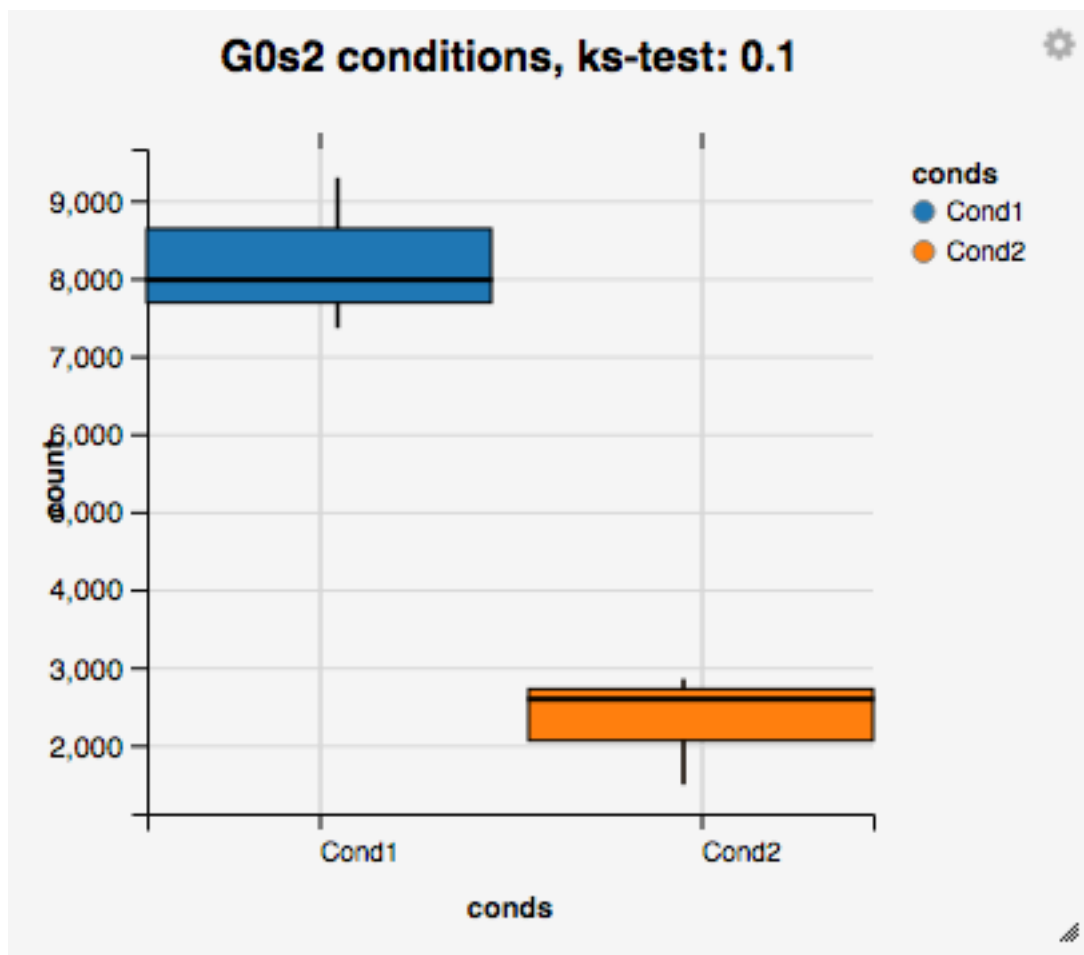


MA plot



You can hover over the scatterplot points to display more information about the point selected. A few bargraphs will be generated for the user to view as soon as a scatterplot point is hovered over.

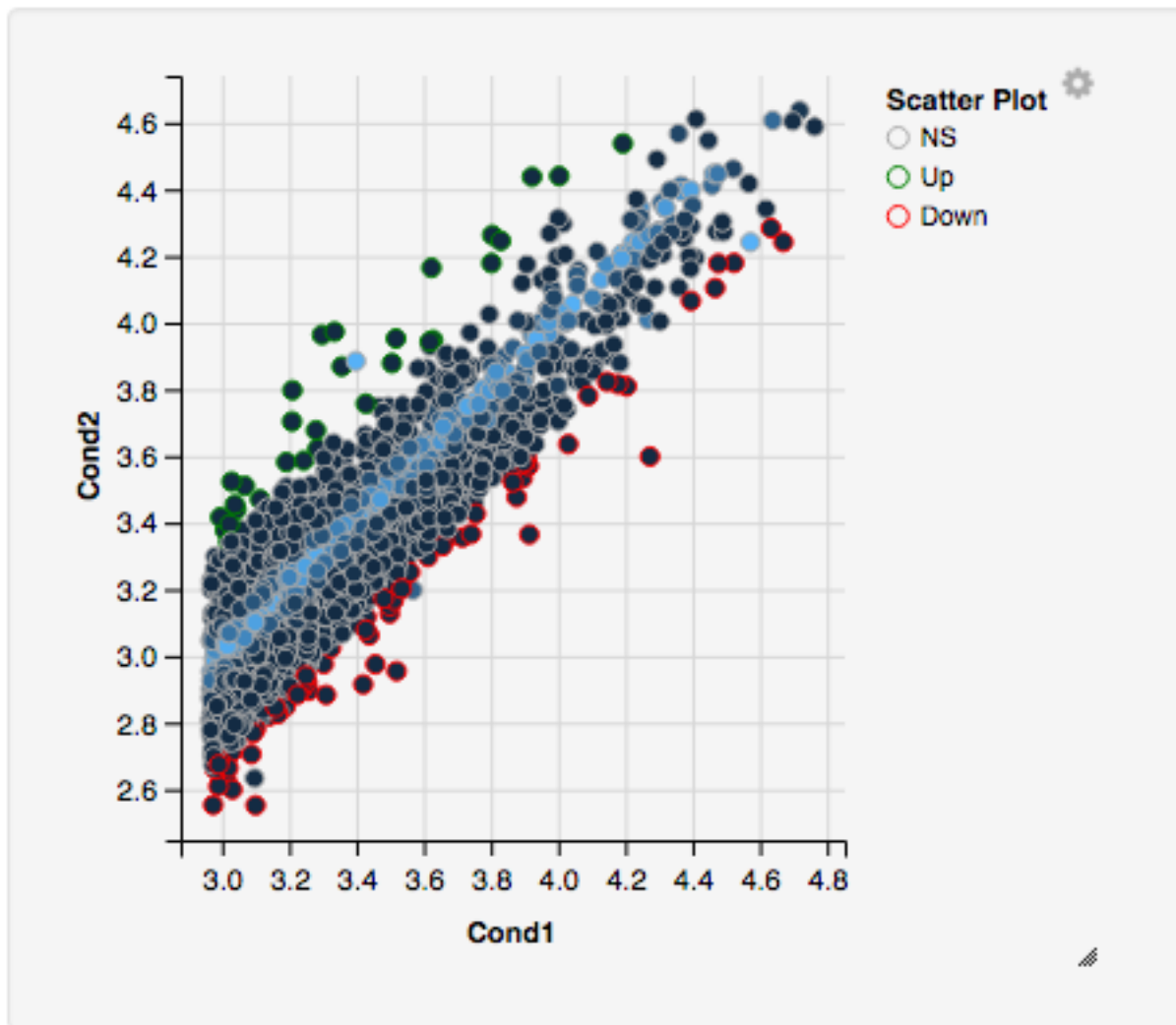




You can also select a specific region within the scatter plot and zoom in on the selected window.



Once you've selected a specific region, a new scatterplot of the selected area will appear on the right




You also have a wide array of options when it comes to fold change cut-off levels, padj cut-off values, which comparison set to use, and dataset of genes to analyze.

Main Plots:

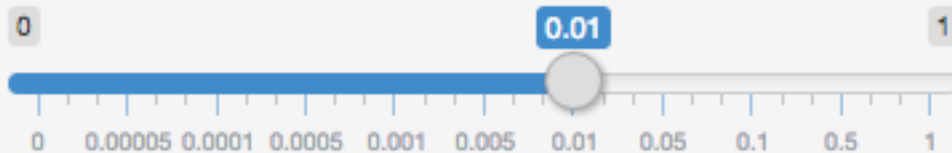
- ☒ Scatter
☐ VolcanoPlot
☐ MAPlot

Submit!**Choose a dataset:**

up+down ▼

 Download Data**Search**

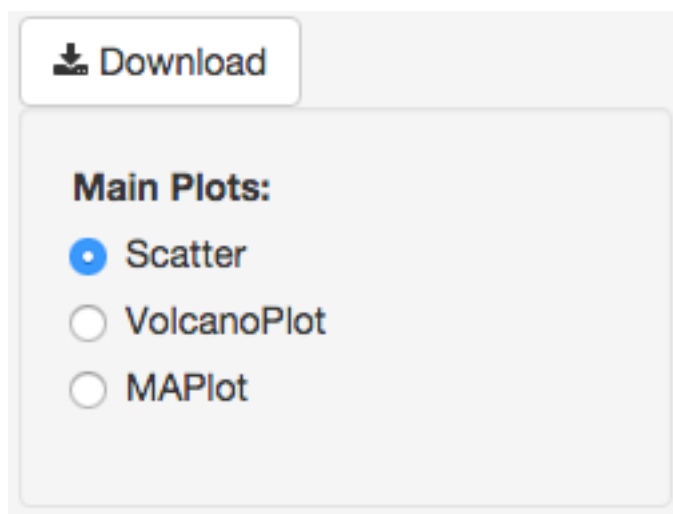
Regular expressions can be used Ex: `^Al => Al..`, `Al$ => ...al`

Filter**padj value cut off**

It is important to note that when conducting multiple comparisons, the comparisons are labeled based on the order that they are input. If you don't remember which samples are in your current comparison you can always view the samples in each condition at the top of the main plots.

Cond1: exper_rep1,exper_rep2,exper_rep3 vs. **Cond2:** control_rep1,control_rep2,control_rep3

If you can select the type of plot at the bottom of the filter tab.



You can download the results in CSV or TSV format by selecting your 'File type' and clicking the 'download' button once you've ran DESeq. You can also download the plot or graphs themselves by clicking on the gear in the upper-left corner of each plot or graph.

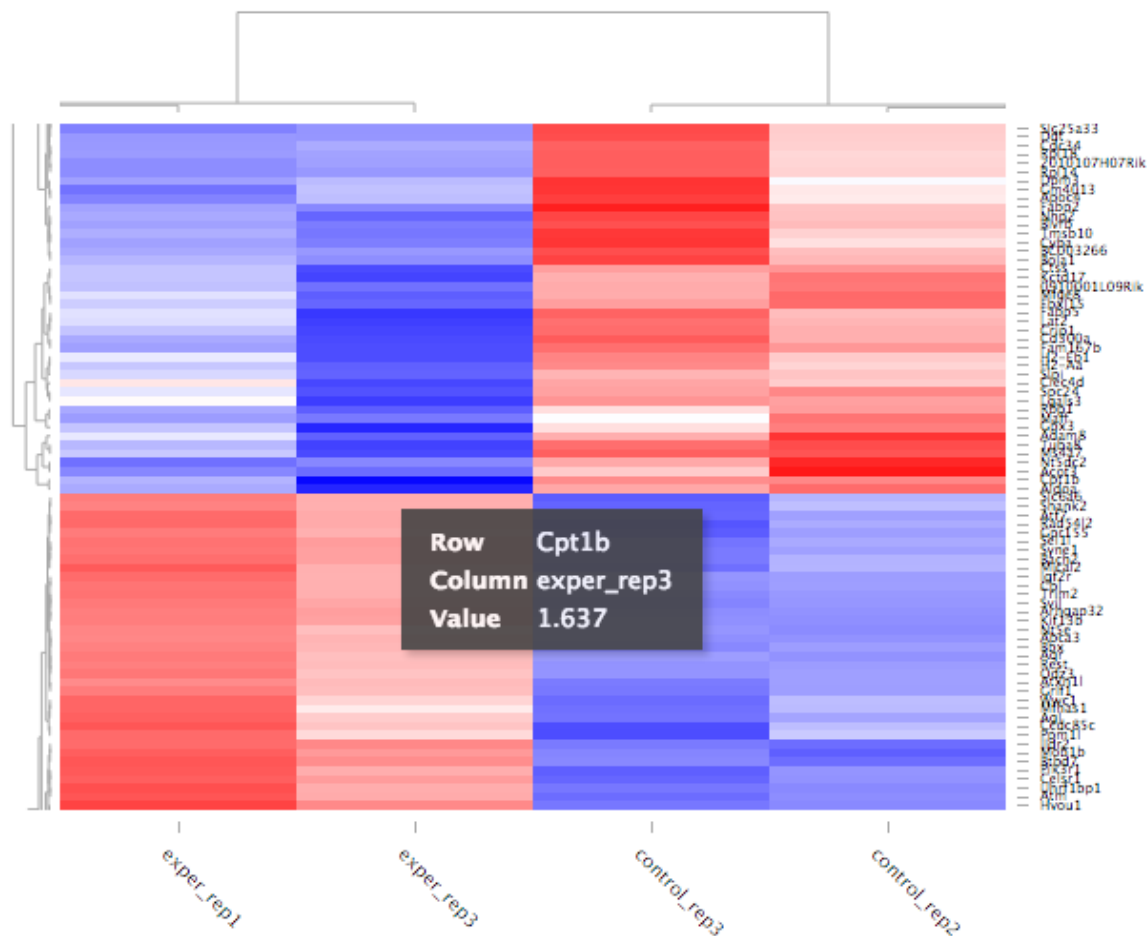
4.1.3 Quality Control Plots

Selecting the 'QC Plots' tab will take you to the quality control plots section. These QC plots are very similar to the QC plots shown before running DESeq and the dataset being used here depends on the one you select in the left panel. In addition to the all-to-all plot shown within the previous QC analysis, users can also view a heatmap and PCA plot of their analyzed data by selecting the proper plot on the left menu. You can also choose the appropriate clustering and distance method you would like to use for this heatmap just about the plot just like in the previous QC section.

For additional information about the clustering methods used, you can consult [this website](#).


For additional information about the distance methods used, you can consult [here](#).

For distances other than 'cor', the distance function defined will be $(1 - (\text{the correlation between samples}))$. Each qc plot also has options to adjust the plot height and width, as well as a download button for a pdf output located above each plot. For the Heatmap, you can also view an interactive session of the heatmap by selecting the 'Interactive' checkbox before submitting your heatmap request. Make sure that before selecting the interactive heatmap option that your dataset being used is 'Up+down'. Just like in the Main Plots, you can click and drag to create a selection. To select a specific portion of the heatmap, make sure to highlight the middle of the heatmap gene box in order to fully select a specific gene. This selection can be used later within the GO Term plots for specific queries on your selection.



4.1.4 GO Term Plots

The next tab, 'GO Term', takes you to the ontology comparison portion of DEBrowser. From here you can select the standard dataset options such as p-adjust value, fold change cut off value, which comparison set to use, and which dataset to use on the left menu. In addition to these parameters, you also can choose from the 4 different ontology plot options: 'enrichGO', 'enrichKEGG', 'Disease', and 'compareCluster'. Selecting one of these plot options queries their specific databases with your current DESeq results.

 Download

Go Plots:

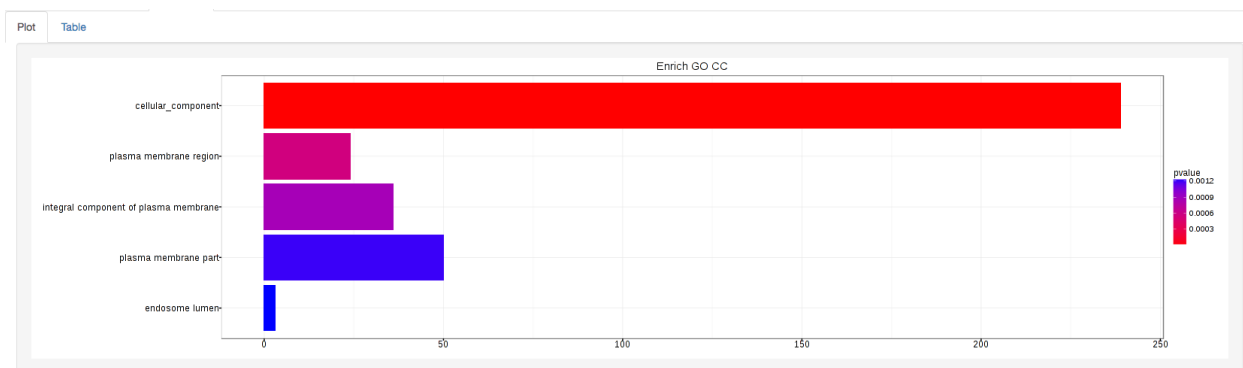
- ☒ enrichGO
- ☐ enrichKEGG
- ☐ Disease
- ☐ compareClusters

Your GO plots include:

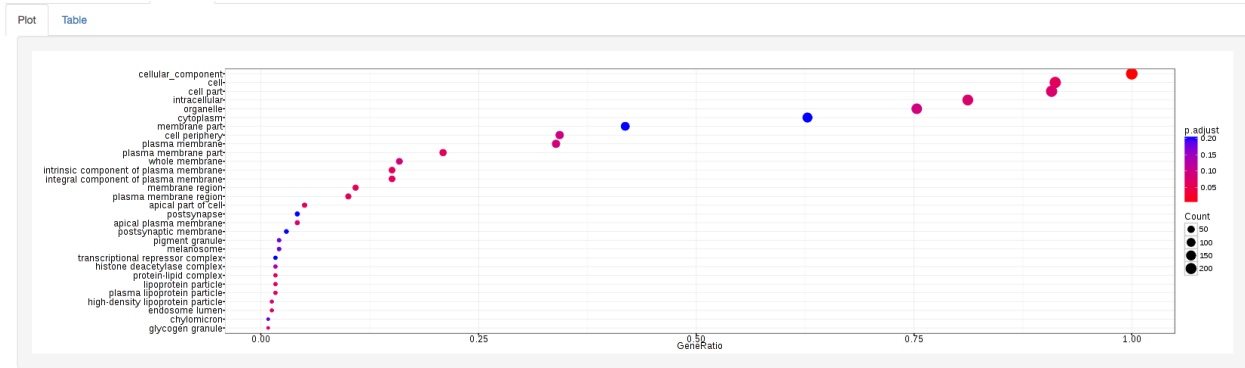
- enrichGO - use enriched GO terms
- enrichKEGG - use enriched KEGG terms
- Disease - enriched for diseases
- compareClusters - comparison of your clustered data

The types of plots you will be able to generate include:

Summary plot:



GOdotplot:



Changing the type of ontology to use will also produce custom parameters for that specific ontology at the bottom of the left option panel.

Once you have adjusted all of your parameters, you may hit the submit button in the top right and then wait for the results to show on screen!

4.1.5 Data Tables

The lasttab at the top of the screen displays various different data tables. These datatables include:

- All Detected
- Up Regulated
- Down Regulated
- Up+down Regulated
- Selected scatterplot points
- Most varied genes
- Comparison differences

Main Plots QC Plots GO Term Tables											
Show 10 entries						Search: <input type="text"/>					
	exper_rep1	exper_rep2	exper_rep3	control_rep1	control_rep2	control_rep3	padj	log2FoldChange	pvalue	foldChange	log10padj
Syng1	7.135273	25.37579	38.917939	7.871420	2.031122e+00	2.798633	2.670881e-03	-1.479199	5.713910e-04	0.3586880	2.573346
Pkp1	23.310344	25.34808	21.862658	5.015264	6.055972e+00	2.796967	6.137991e-04	-1.577093	1.019709e-04	0.3351565	3.211974
AK206687	4.065426	15.31221	9.379557	0.000000	6.271097e-03	0.000000	7.288816e-04	-1.654479	1.248060e-04	0.3176525	3.137343
Rps6ka3	15.280315	12.32422	18.450445	47.740842	7.445891e+01	33.702584	3.250601e-04	1.421371	4.866031e-05	2.6783986	3.488036
D330041H03Rik	80.087331	86.16724	28.416393	6.998361	1.009576e+01	2.745741	2.391427e-09	-2.429347	7.245195e-11	0.1856495	8.621343
Nr2c2	86.363440	24.88276	56.424369	229.305017	1.912994e+02	201.460333	3.523672e-08	1.669908	1.513202e-09	3.1819421	7.453005
Zfhx2	83.320057	54.94153	47.574396	219.240720	2.003584e+02	189.868073	6.859478e-10	1.547058	1.754693e-11	2.9222062	9.163709
Fcgbp	13.176789	13.96867	18.962477	3.278481	1.019879e+00	3.796185	4.634870e-03	-1.412922	1.087597e-03	0.3755504	2.333962
Ly6d	339.117011	334.96468	408.621931	113.570277	1.349614e+02	88.423299	4.973157e-15	-1.628726	3.766736e-17	0.3233737	14.303368
C78339	101.615743	51.38191	130.392025	327.724952	3.262240e+02	294.855319	2.535484e-10	1.612138	5.688760e-12	3.0570451	9.595939

Showing 1 to 10 of 355 entries

Previous 1 2 3 4 5 ... 36 Next

All of the tables tables, except the Comparisons table, contain the following information:

- ID - The specific gene ID
- Sample Names - The names of the samples given and they're corresponding tmm normalized counts
- Conditions - The log averaged values
- padj - padjusted value
- log2FoldChange - The Log2 fold change
- foldChange - The fold change
- log10padj - The log 10 padjusted value

The Comparisons table generates values based on the number of comparisons you have conducted. For each pairwise comparison, these values will be generated:

- Values for each sample used
- foldChange of comparison A vs B
- pvalue of comparison A vs B
- padj value of comparison A vs B

Main Plots QC Plots GO Term **Tables**

Show **10** entries Search:

	exper_rep1	exper_rep2	exper_rep3	control_rep1	control_rep2	control_rep3	foldChange.C1.vs.C2	padjC1.vs.C2	foldChange.C3.vs.C4	padjC3.vs.C4
1110007C09Rik	145.00	211.01	77.00	75.00	52.00	67.00	0.3693497	1.057546e-08	0.3418683	2.589037e-06
1810044D09Rik	73.00	90.00	36.00	32.00	27.00	22.00	0.3562541	3.755707e-06	0.3801446	9.042321e-04
2010003K11Rik	885.00	826.00	403.00	375.00	271.39	334.00	0.3617310	1.140383e-16	0.3338943	1.195500e-11
2410076I21Rik	34.97	64.00	21.97	2.94	13.85	16.00	0.3390859	5.020396e-04	0.3149882	1.752537e-03
2610005L07Rik	248.96	154.06	127.61	1061.71	632.64	468.95	2.6469646	2.756012e-09	3.0490664	2.185518e-08
3930402G23Rik	69.00	49.00	20.00	12.00	8.00	2.00	0.2028349	2.516955e-08	0.2433717	1.730217e-05
4930470H14Rik	36.74	42.01	20.92	242.11	139.65	100.34	3.1135333	1.197587e-08	3.3640596	3.607676e-07
4930565N06Rik	6.01	1.00	0.00	30.01	19.00	12.00	2.6527298	5.414730e-03	2.2782137	4.006262e-02
4932438A13Rik	246.00	234.00	130.00	1434.43	754.31	678.00	3.1780469	1.199287e-19	3.4064309	2.772699e-14
4933411K16Rik	4.00	5.00	5.00	56.96	41.00	15.00	3.1104165	4.043031e-04	3.9322051	1.179784e-04

Showing 1 to 10 of 440 entries Previous **1** 2 3 4 5 ... 44 Next

You can further customize and filter each specific table a multitude of ways. For unique table or dataset options, select the type of table dataset you would like to customize on the left panel under 'Choose a dataset' to view it's additional options. All of the tables have a built in search function at the top right of the table and you can further sort the table by column by clicking on the column header you wish to sort by. The 'Search' box on the left panel allows for multiple searches via a comma-seperated list. You can additionally use regex terms such as "^al" or "*lm" for even more advanced searching. This search will be applied to wherever you are within DEBrowser, including both the plots and the tables.

4.2 Local Install Guide

4.2.1 Quick Local Install

Running these simple command will launch the DEBrowser within your local machine.

Before you start; First, you will have to install R and/or RStudio. (On Fedora/Red Hat/CentOS, these packages have to be installed; openssl-devel, libxml2-devel, libcurl-devel, libpng-devel)

You can download the source code or the tar file for DEBrowser [‘here.<https://github.com/UMMS-Biocore/debrowser/releases>’](https://github.com/UMMS-Biocore/debrowser/releases)_

Installation instructions from source:

1. Install the required dependencies by running the following commands in R or RStudio.

```
source("http://www.bioconductor.org/biocLite.R")
```

```
biocLite("debrowser")
```

2. Start R and load the library

```
library(debrowser)
```

3. Start DE browser

```
startDEBrowser()
```

Once you run ‘startDEBrowser()’ shiny will launch a web browser with your local version of DEBrowser running and ready to use!

For more information about DEBrowser, please visit our Quick-start Guide or our DESeq/DEBrowser section within readthedocs.

4.3 DESeq/DEBrowser

This guide contains a breif discription of DESeq2 used within the DEBrowser

4.3.1 Introduction

Differential gene expression analysis has become an increasingly popular tool in determining and viewing up and/or down experssed genes between two sets of samples. The goal of Differential gene expression analysis is to find genes or transcripts whose difference in expression, when accounting for the variance within condition, is higher than expected by chance. **DESeq2** is an R package available via Bioconductor and is designed to normalize count data from high-throughput sequencing assays such as RNA-Seq and test for differential expression (Love et al. 2014). For more information on the DESeq2 algorithm, you can visit [this website](#) With multiple parameters such as padjust values, log fold changes, and plot styles, altering plots created with your DE data can be a hassle as well as time consuming. The Differential Expression Browser uses DESeq2 coupled with shiny to produce real-time changes within your plot queries and allows for interactive browsing of your DESeq results. In addition to DESeq analysis, DEBrowser also offers a variety of other plots and analysis tools to help visualize your data even further.

4.3.2 Getting Started

In order to conduct differential expression analysis, we first need data to analyze. In order to call DESeq2, we’re going to need gene quantifications and expected counts for those genes. To obtain these quantifications, we typically use **RSEM**, however there are other ways to obtain this data.

The TSV files used to describe the quantification counts are similar to this:

IE:

gene	trans	exp1	exp2	cont1	cont2
DQ714	uc007	0.00	0.00	0.00	0.00
DQ554	uc008	0.00	0.00	0.00	0.00
AK028	uc011	2.00	1.29	0.00	0.00

Where the gene column represent the gene name, the transcript column represents the transcript(s) name (comma separated for multiple), and the rest of the columns are the raw counts for your samples.

4.3.3 DESeq2

For the details please check the user guide. [DESeq2 userguide](#)

DESeq2 performs multiple steps in order to analyze the data you've provided for it. The first step is to indicate the condition that each column (experiment) in the table represent. You can group multiple samples into one condition column. DESeq2 will compute the probability that a gene is differentially expressed (DE) for ALL genes in the table. It outputs both a nominal and a multiple hypothesis corrected p-value (padj) using a negative binomial distribution.

4.3.4 Un-normalized counts

DESeq2 requires count data as input obtained from RNA-Seq or another high-throughput sequencing experiment in the form of matrix values. Here we convert un-integer values to integer to be able to run DESeq2. The matrix values should be un-normalized, since DESeq2 model internally corrects for library size. So, transformed or normalized values such as counts scaled by library size should not be used as input. Please use edgeR or limma for normalized counts.

4.3.5 Used parameters for DESeq2

- **fitType:** either “parametric”, “local”, or “mean” for the type of fitting of dispersions to the mean intensity. See `estimateDispersions` for description.
- **betaPrior:** whether or not to put a zero-mean normal prior on the non-intercept coefficients See `nbinomWaldTest` for description of the calculation of the beta prior. By default, the beta prior is used only for the Wald test, but can also be specified for the likelihood ratio test.
- **testType:** either “Wald” or “LRT”, which will then use either Wald significance tests (defined by `nbinomWaldTest`), or the likelihood ratio test on the difference in deviance between a full and reduced model formula (defined by `nbinomLRT`)
- **rowsum.filter:** regions/genes/isoforms with total count (across all samples) below this value will be filtered out

4.3.6 EdgeR

For the details please check the user guide. [EdgeR userguide](#).

4.3.7 Used parameters for EdgeR

- **Normalization:** Calculate normalization factors to scale the raw library sizes. Values can be “TMM”, “RLE”, “upperquartile”, “none”.
- **Dispersion:** either a numeric vector of dispersions or a character string indicating that dispersions should be taken from the data object. If a numeric vector, then can be either of length one or of length equal to the number of genes. Allowable character values are “common”, “trended”, “tagwise” or “auto”. Default behavior (“auto” is to use most complex dispersions found in data object.
- **testType:** `exactTest` or `glmLRT`. `exactTest`: Computes p-values for differential abundance for each gene between two digital libraries, conditioning on the total count for each gene. The counts in each group as a proportion of the whole are assumed to follow a binomial distribution. `glmLRT`: Fit a negative binomial generalized

log-linear model to the read counts for each gene. Conduct genewise statistical tests for a given coefficient or coefficient contrast.

- **rowsum.filter:** regions/genes/isoforms with total count (across all samples) below this value will be filtered out

4.3.8 Limma

For the details please check the user guide. [Limma userguide](#).

Limma is a package to analyse of microarray or RNA-Seq data. If data is normalized with spike-in or any other scaling, tranforamtion or normalization method, Limma can be ideal. In that case, prefer limma rather than DESeq2 or EdgeR.

4.3.9 Used parameters for Limma

- **Normalization:** Calculate normalization factors to scale the raw library sizes. Values can be “TMM”, “RLE”, “upperquartile”, “none”.
- **Fit Type:** fitting method; “ls” for least squares or “robust” for robust regression
- **Norm. Bet. Arrays:** Normalization Between Arrays; Normalizes expression intensities so that the intensities or log-ratios have similar distributions across a set of arrays.
- **rowsum.filter:** regions/genes/isoforms with total count (across all samples) below this value will be filtered out

4.3.10 DEBrowser

DEBrowser utilizes [Shiny](#), a R based application development tool that creates a wonderful interactive user interface (UI) combined with all of the computing prowess of R. After the user has selected the data to analyze and has used the shiny UI to run DESeq2, the results are then input to DEBrowser. DEBrowser manipulates your results in a way that allows for interactive plotting by which changing padj or fold change limits also changes the displayed graph(s). For more details about these plots and tables, please visit our quickstart guide for some helpful tutorials.

For comparisons against other popular data visualization tools, see the table below.

DEBrowser Comparison

	DEBrowser	MeV	Chipster	Galaxy	CummeRBund
Easy Install	✓	✓	✗	✗	✓
User Friendly	✓	✓	✓	✓	✓
Friendly Learning Curve	✓	✓	✓	✗	✗
Full R Dev Support	✓	✗	✗	✓	✓
DESeq	✓	✓	✓	✓	✓
EDGeR	✓	✓	✓	✓	✓
Limma	✓	✗	✗	✓	✗
CuffDiff	✗	✗	✗	✓	✓
Web Interface	✓	✗	✗	✓	✗
Local Install	✓	✓	✓	✗	✓

4.3.11 References

1. **Love MI, Huber W and Anders S (2014).** Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15, pp. 550. <http://doi.org/10.1186/s13059-014-0550-8>.
2. **Robinson, MD, and Smyth, GK (2008).** Small sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics 9, 321–332.
3. **Ritchie, ME, Phipson, B, Wu, D, Hu, Y, Law, CW, Shi, W, and Smyth, GK (2015).** limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43(7), e47.

Indices and tables

- `genindex`
- `modindex`
- `search`