RNA-Seq Data Analysis Suite Manual

Nikhil Kumar

May 5, 2017

Contents

RNA-Seq Data Analysis Suite (RAS) is an end to end pipeline aimed to guiide you through the entire RNA-Seq Analysis process. It uses many openly available bioinformatic tools.

1 Installation

1.1 MongoDB

Install mongodb from instruction on https://docs.mongodb.com/v3.2/tutorial/install-mongodb-on-ubuntu/

```
sudo apt-key adv --keyserver hkp://keyserver.ubuntu.com:80 --recv EA312927
echo "deb http://repo.mongodb.org/apt/ubuntu xenial/mongodb-org/3.2 multiverse"
```

Change xenial to your code name

```
sudo apt-get update
sudo apt-get install -y mongodb-org
```

Start the MongoDB service

```
sudo service mongod start
```

1.2 Fastqc

Download fastqc here: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

```
chmod 755 fastqc
sudo ln -s /path/to/FastQC/fastqc /usr/local/bin/fastqc
```

1.3 Ubuntu Packages

```
sudo apt-get install libcurl4-openssl-dev libxml2-dev libkrb5-dev
sudo apt-get librtmp-dev libssl-dev graphviz
```

1.4 Other software

1. python2
2. R
3. pip2
4. virtualenv

5. screen
6. zip
7. fastq-dump
8. fastqc
9. hisat2
10. cufflinks
11. cuffquant
12. cuffnorm
13. samtools

1.5 Download and Install

```
git clone https://github.com/nikhil/RAS
cd RAS
chmod +x setup.sh
./setup.sh
```

1.6 Start the server

After installing you need to start the server.

```
./start_server.sh
```

1.7 Stop the server

If you need to stop the server

```
./stop_server.sh
```

2 Tutorial

2.1 Access the server

Go to [your ip]:5000 on your server then click on 'Single-cell RNA-Seq'. Then click 'Assign Job'

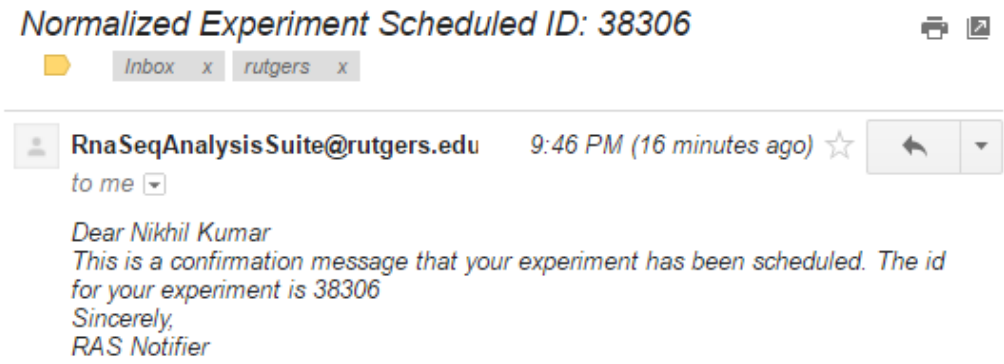
2.2 Assign Job

Enter your name and email. Provide a name for each condition and a link for each replicate. All link and files uploaded should be a sra file. In this tutorial we will use 4 cells from GSE71485. In the GEO experiment page, you will need to click on the correct sample, click on the ftp link,find the sra file, and then copy the link. The samples in this list also include the hyperlinks for your reference.

1. Sample C10
2. Sample C110
3. Sample C111
4. Sample C118

Here is how the assign job input form should look.

The first email will confirm that the process has been schedules.



Normalized Experiment Scheduled ID: 38306

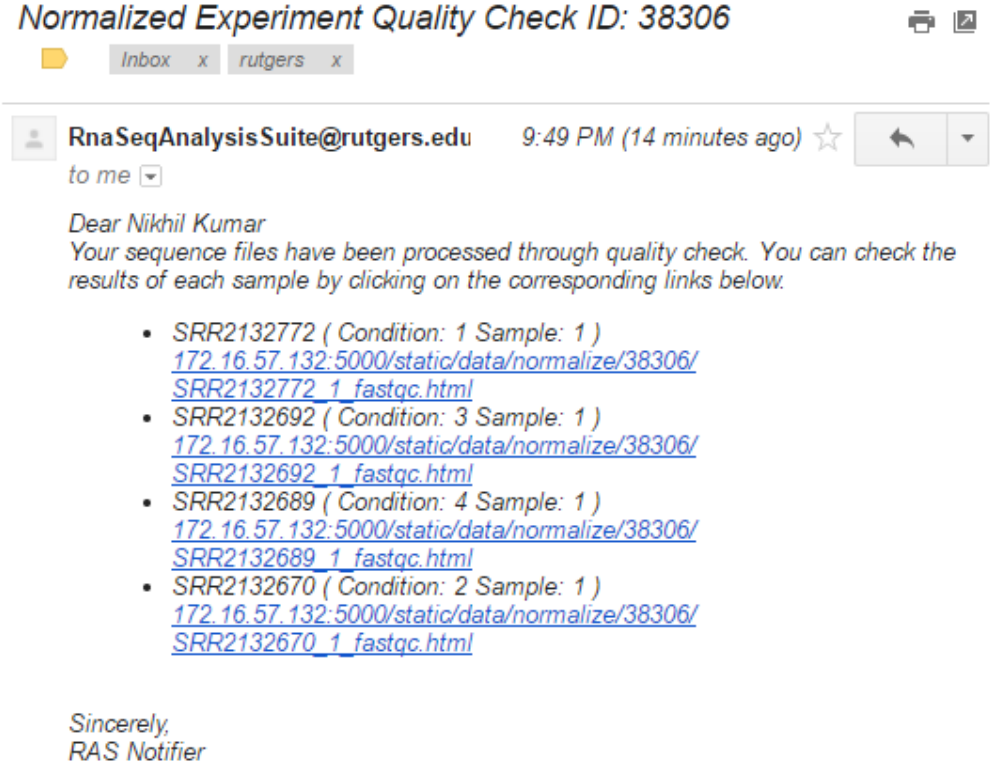
Inbox x rutgers x

RnaSeqAnalysisSuite@rutgers.edu 9:46 PM (16 minutes ago)

to me

Dear Nikhil Kumar
This is a confirmation message that your experiment has been scheduled. The id for your experiment is 38306
Sincerely,
RAS Notifier

The second email will contain links to the quality score of the samples



Normalized Experiment Quality Check ID: 38306

Inbox x rutgers x

RnaSeqAnalysisSuite@rutgers.edu 9:49 PM (14 minutes ago)

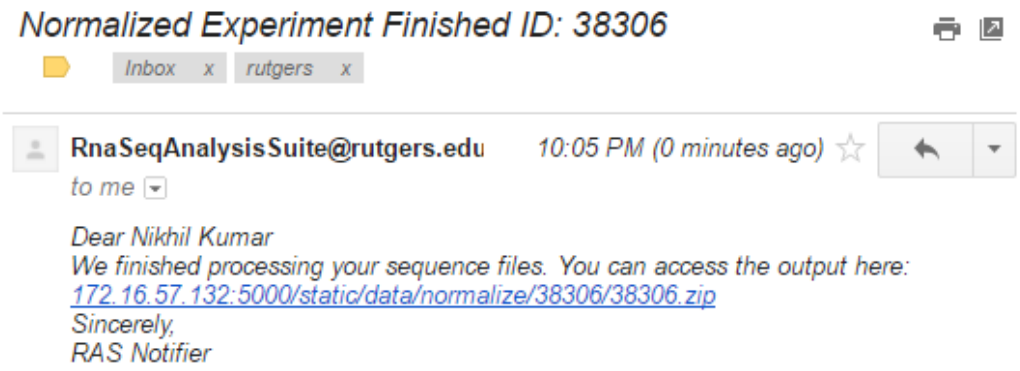
to me

Dear Nikhil Kumar
Your sequence files have been processed through quality check. You can check the results of each sample by clicking on the corresponding links below.

- *SRR2132772 (Condition: 1 Sample: 1)*
  172.16.57.132:5000/static/data/normalize/38306/SRR2132772_1_fastqc.html
- *SRR2132692 (Condition: 3 Sample: 1)*
  172.16.57.132:5000/static/data/normalize/38306/SRR2132692_1_fastqc.html
- *SRR2132689 (Condition: 4 Sample: 1)*
  172.16.57.132:5000/static/data/normalize/38306/SRR2132689_1_fastqc.html
- *SRR2132670 (Condition: 2 Sample: 1)*
  172.16.57.132:5000/static/data/normalize/38306/SRR2132670_1_fastqc.html

Sincerely,
RAS Notifier

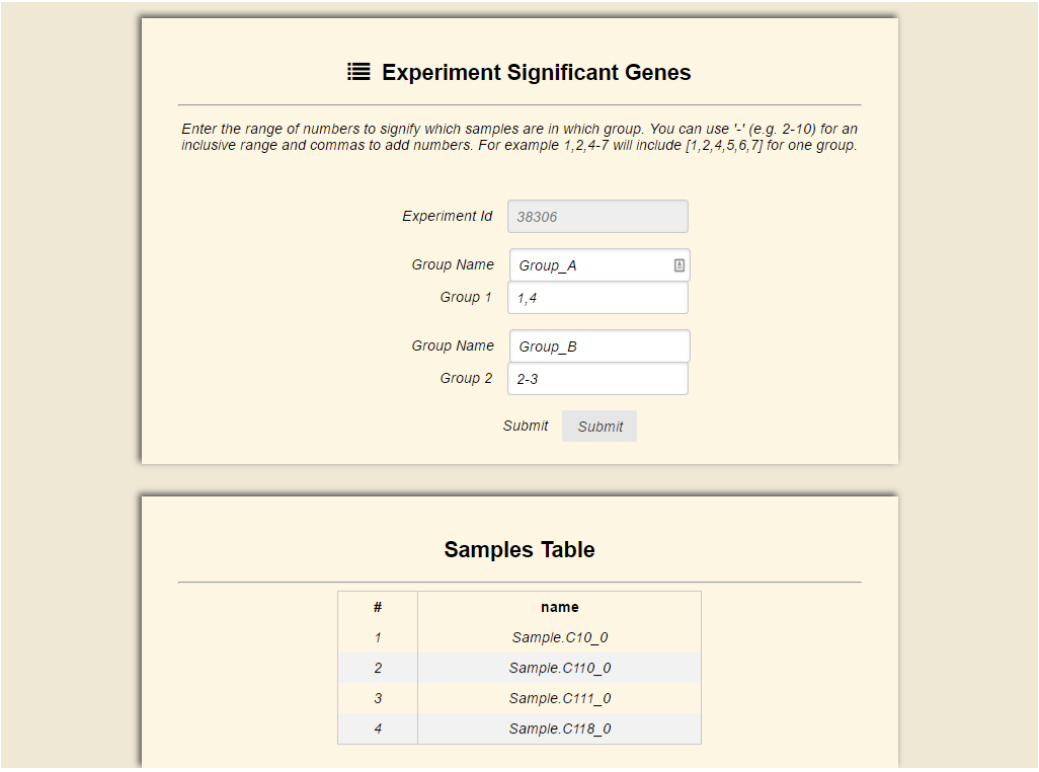
The third email will contain a link to the finished analysis files.



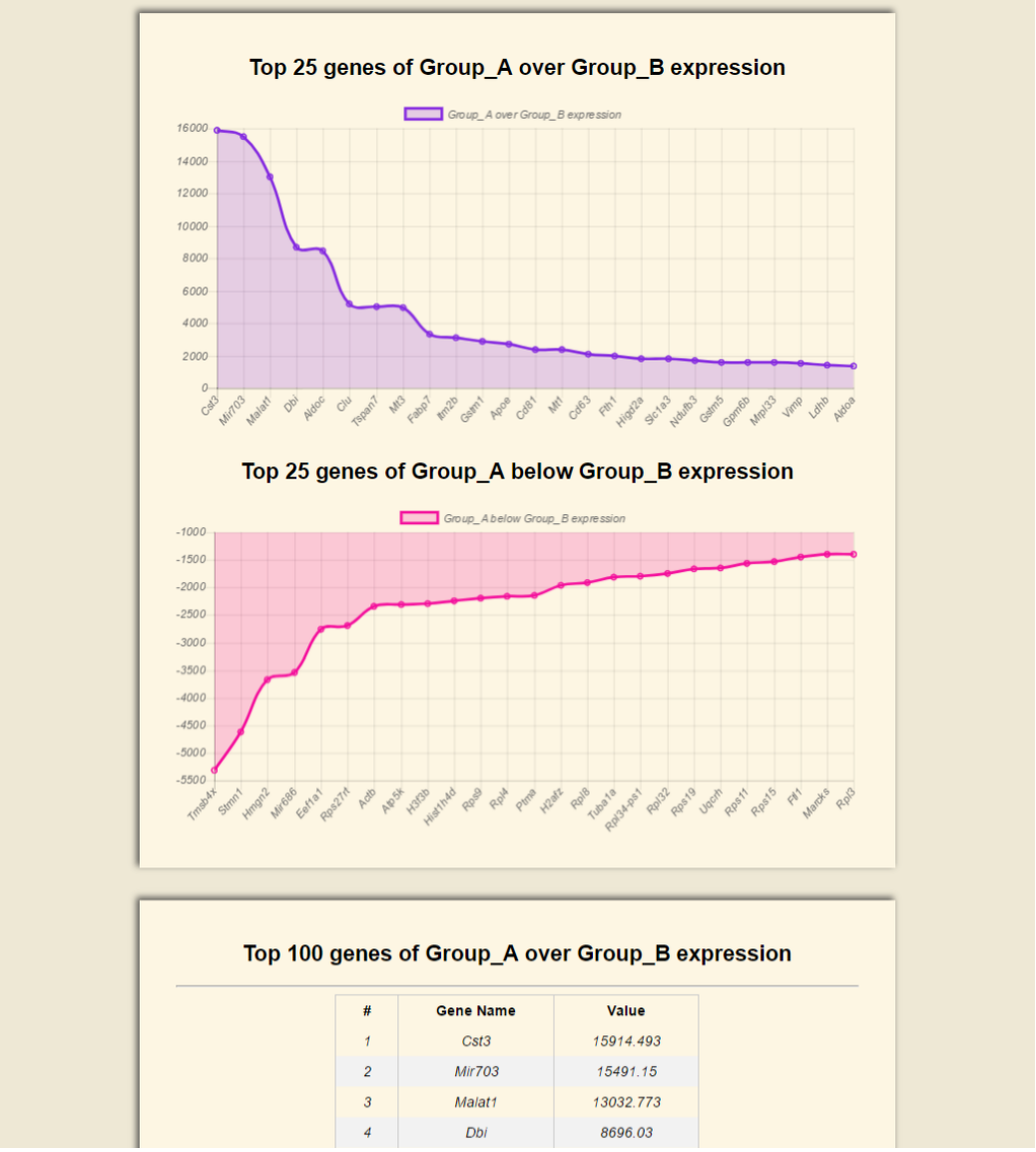
Note: Receiving emails from RAS has worked fine with me from Gmail. The Rutgers email seems to have some issues occasionally in receiving emails from the Sendgrid email server.

2.3 Get Significant Genes [optional]

If you already know which groups the cell are under you can find the genes which have the largest difference in expression across these groups. I use my experiment id that was given to me in the email from RAS.



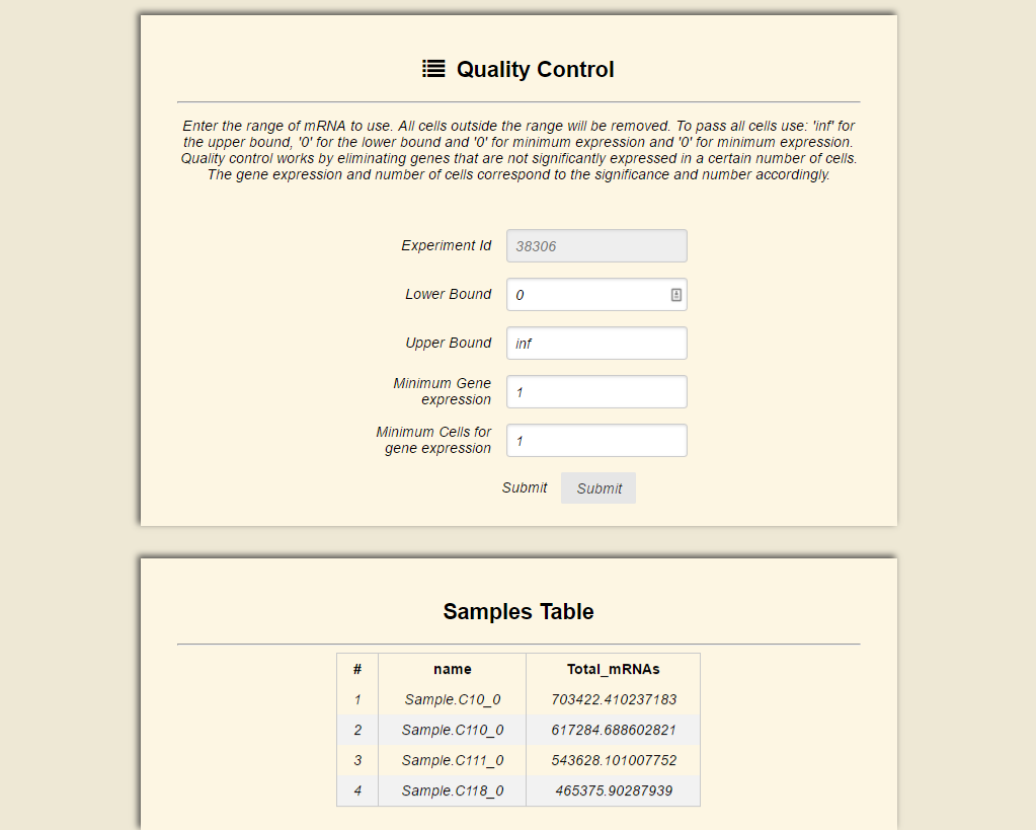
Here is a cropped image of the results. The results also include the top 100 genes expressed above and below the other group.



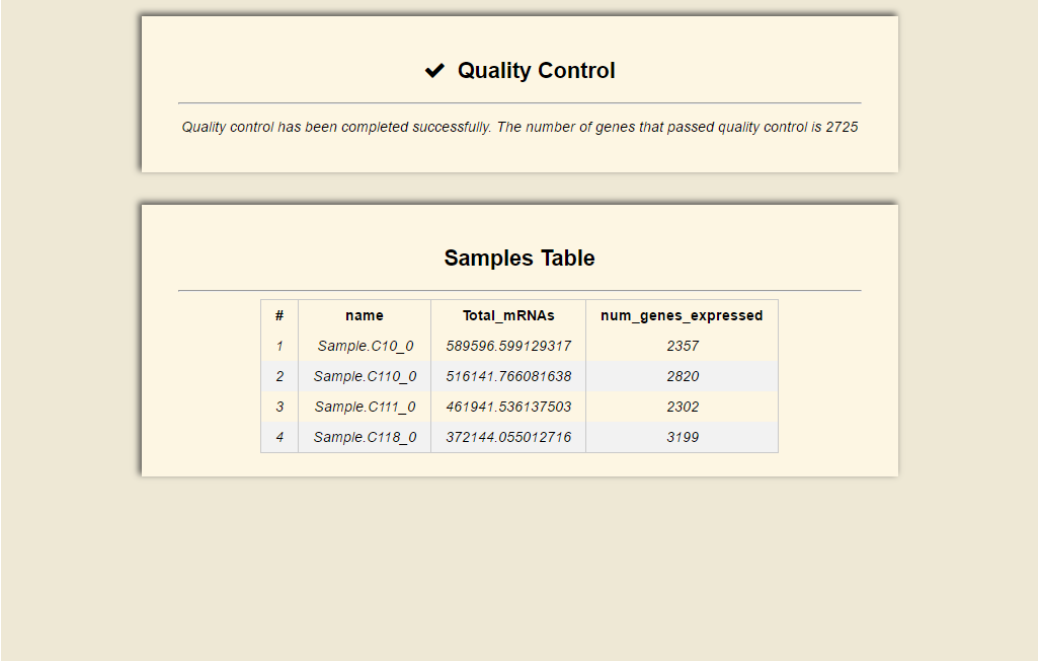
2.4 Run Quality Control [required]

The total mRNA in each sample is provided on the table below. In this page you need to enter the range of mRNA to use. All cells outside the range will be removed. To pass all cells use: 'inf' for the upper bound, '0' for the lower bound and '0' for minimum expression and '0' for minimum expression. Quality control works by eliminating genes that are not significantly

expressed in a certain number of cells. The gene expression and number of cells correspond to the significance and number accordingly. In the tutorial I am passing all cells. For the genes, I filter all the genes which are not expressed above the value of 1 in at least 1 of the cells.



The confirmation page will show show the cells that passed through and the number of genes.

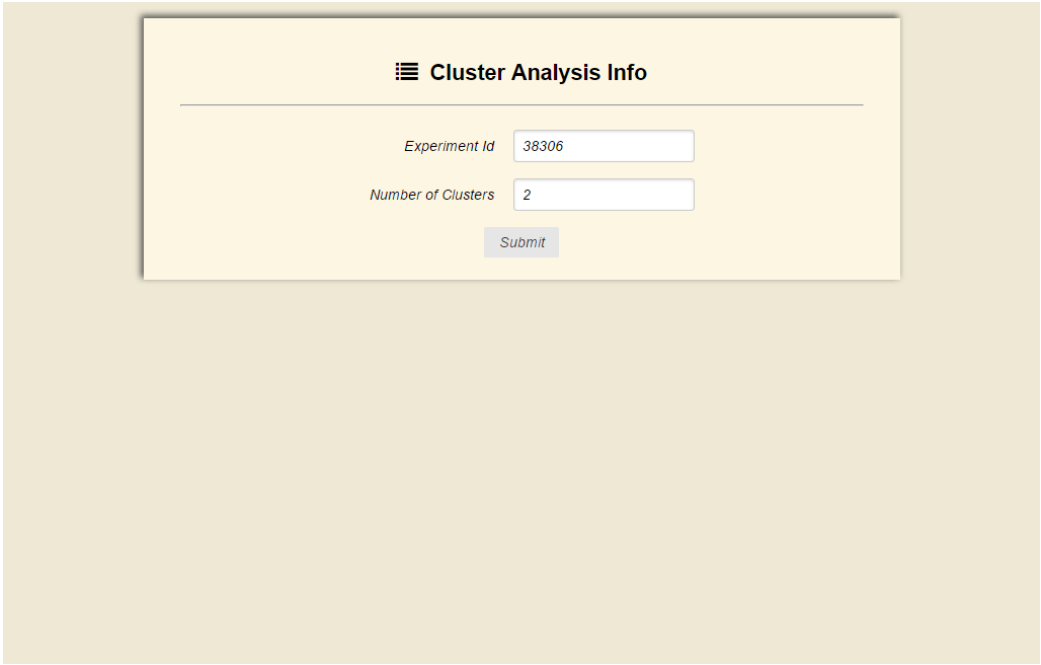


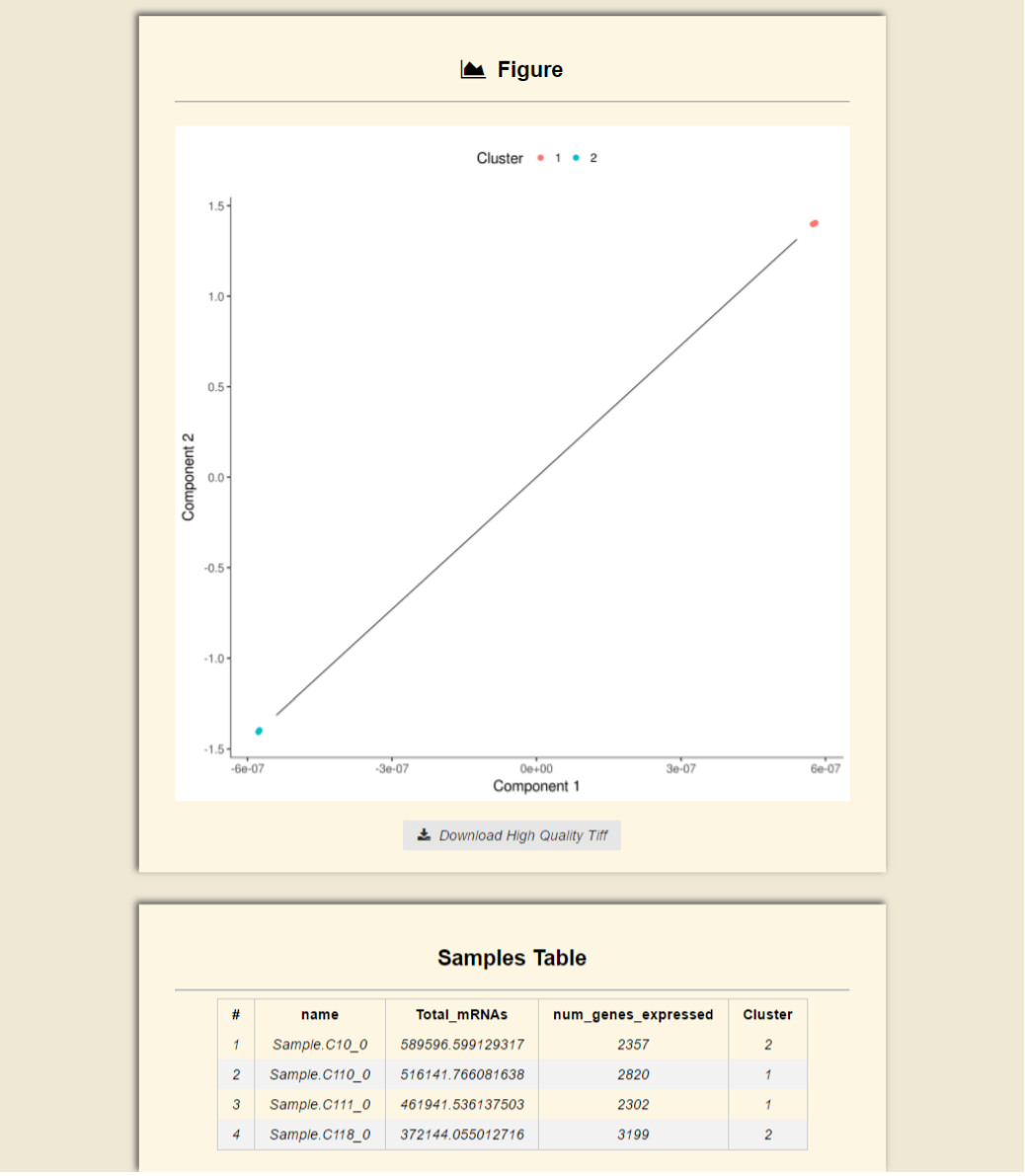
2.5 Trim Data [optional]

I will not trim data in this tutorial. However, you can use trim data to remove any outliers that were not removed from quality control. This process needs to be run after quality control.

2.6 Cluster Analysis [required]

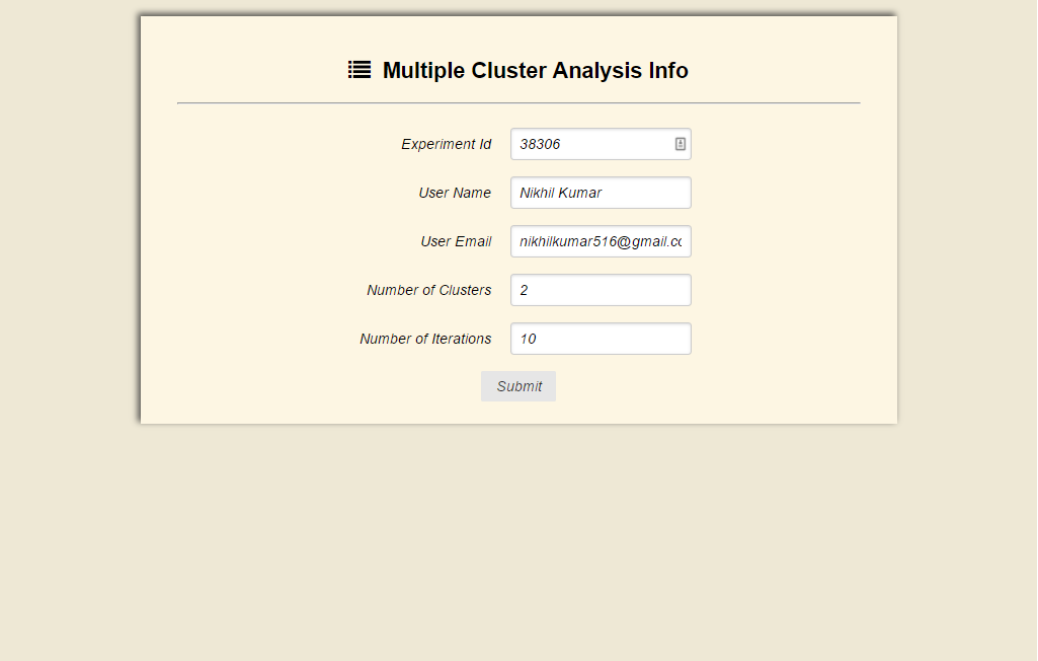
I clustered the data set into 2 groups.


Cluster Analysis Info
Experiment Id
38306
Number of Clusters
2
Submit



2.7 Cluster Validation [optional]

Often clustering analysis gives us different results at different runs. In order to confirm our results we can run the data through multiple iterations of clustering and confirm that our current clustering configuration is the most numerous one.



Name/Label	Type 1
Number	10
Sample.C10_0	1
Sample.C110_0	2
Sample.C111_0	2
Sample.C118_0	1

We can see here that in all 10 iterations the cluster matched my current configuration.

2.8 Decision Tree [optional]

We can make a decision tree to find a significant gene which divides the group well.

Decision Tree

Enter the range of numbers to signify which samples are in which group. You can use '-' (e.g. 2-10) for an inclusive range and commas to add numbers. For example 1,2,4-7 will include [1,2,4,5,6,7] for one group.

Experiment Id: 38306

Group Name: Cluster 1
Group 1 Range: 2-3

Group Name: Cluster 2
Group 2 Range: 1,4

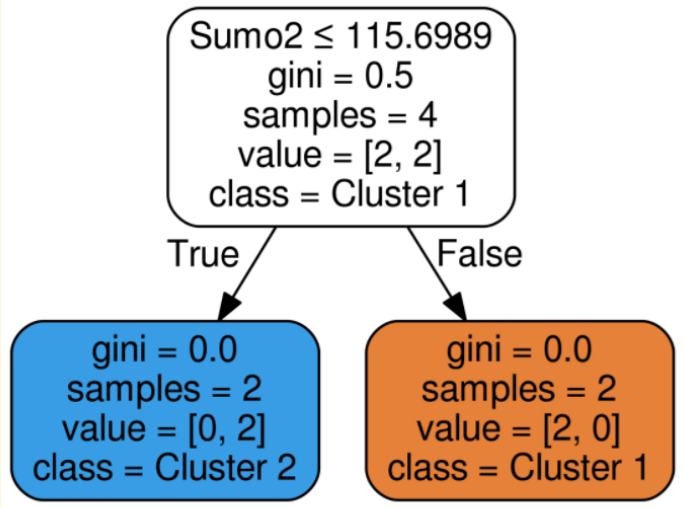
Modify Groups: Add Group | Remove Group

Submit: Submit

Samples Table

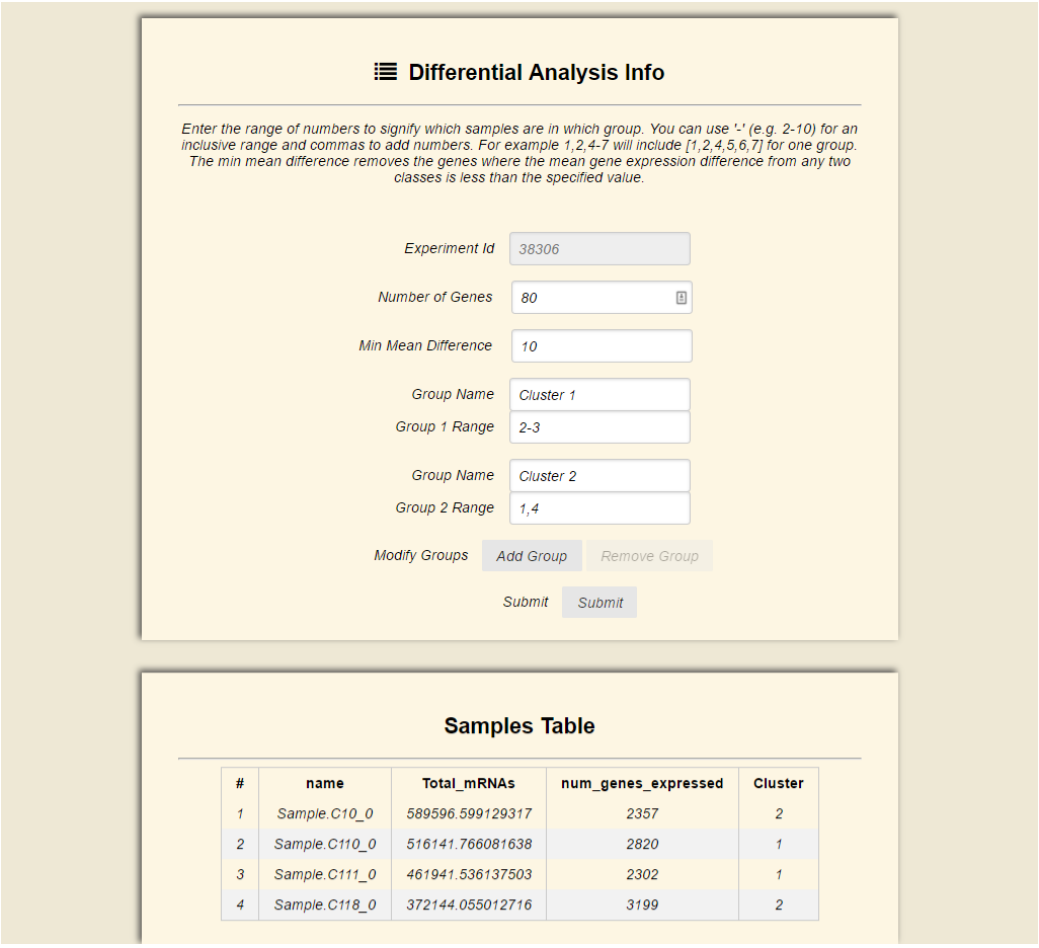
#	name	Total_mRNAs	num_genes_expressed	Cluster
1	Sample.C10_0	589596.599129317	2357	2
2	Sample.C110_0	516141.766081638	2820	1
3	Sample.C111_0	461941.536137503	2302	1
4	Sample.C118_0	372144.055012716	3199	2

Figure

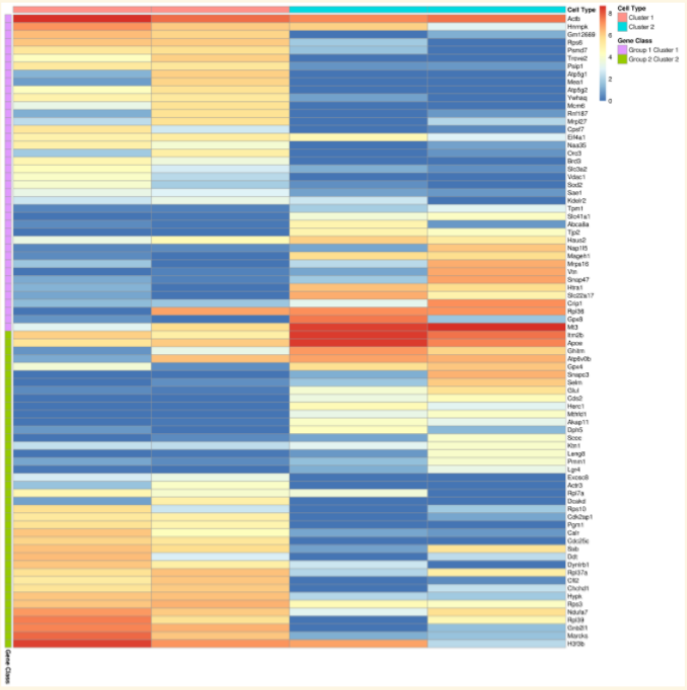


2.9 Manual Heatmap [optional]

We can get a heatmap of differential gene expression across known groups. Here is a cropped image of the results. The results also all the genes and its average expression across the groups.



Figure



Average Gene Expression

#	Gene Name	Cluster 1	Cluster 2
0	Abca8a	0.32638	62.01242
1	Actb	4022.975	1683.665
2	Actr3	32.07774	0.0

2.10 Classify [required]

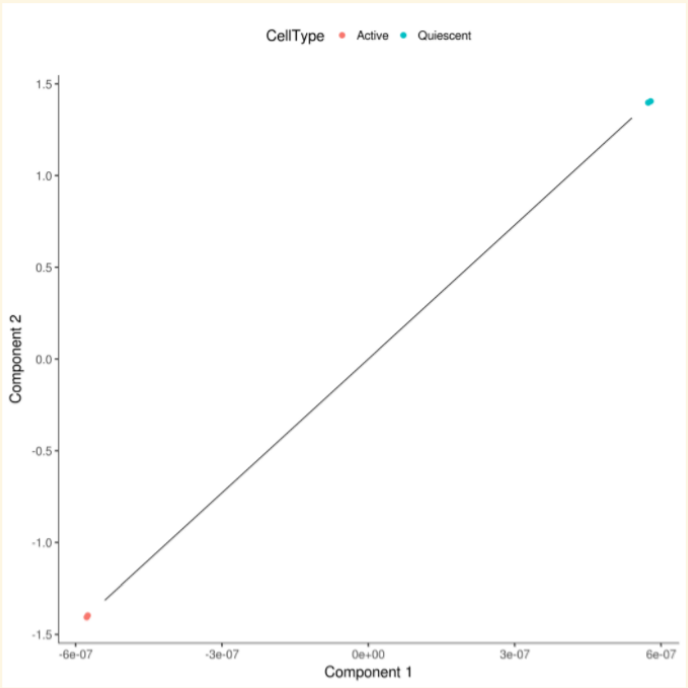
I already know that the cells in cluster 2 are active neural stem cells. I can use the critical gene I found from the decision tree to classify the dataset through gene expression.


User Info
Experiment Id
38306
Label 1
Label Name
Active
Condition 1
Sumo2
<=
115.69
Modify conditions
Add condition
Remove condition
Label 2
Label Name
Quiescent
Condition 1
Sumo2
>
115.69
Modify conditions
Add condition
Remove condition
Modify labels
Add label
Remove label
Submit
Submit

✔ Classification

Classification has completed successfully.

Figure



Download High Quality Tiff

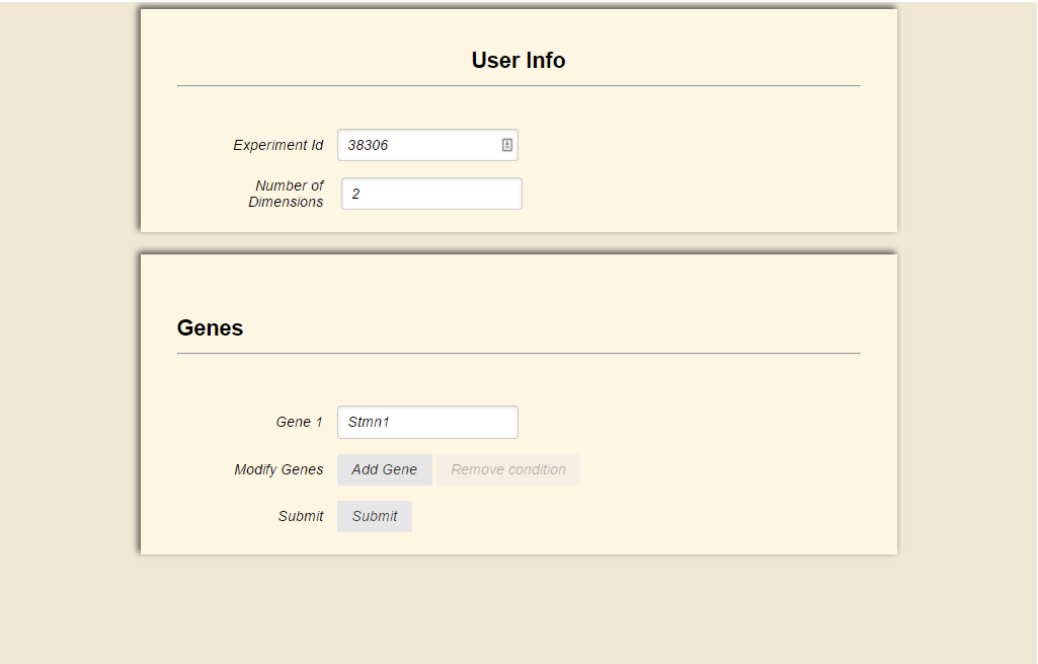
Samples Table

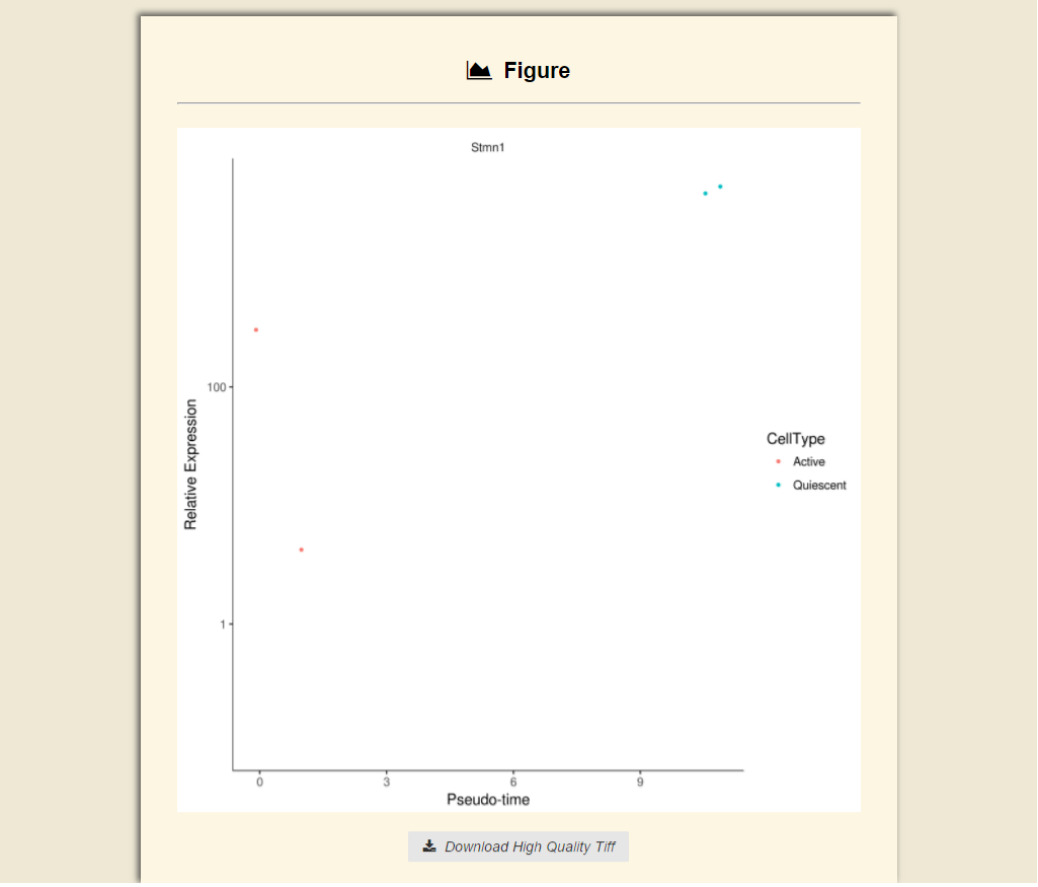
#	name	Total_mRNAs	num_genes_expressed	Cluster	CellType
1	Sample.C10_0	589596.599129317	2357	2	Active
2	Sample.C110_0	516141.766081638	2820	1	Quiescent
3	Sample.C111_0	461941.536137503	2302	1	Quiescent
4	Sample.C118_0	372144.055012716	3199	2	Active

If you choose not to classify through gene expression you will need to classify manually.

2.11 Pseudotime [optional]

You can compare gene expression with the cells ordered in pseudotime. For more information on pseudotime ordering, it is recommended to read the Monocle Manual.



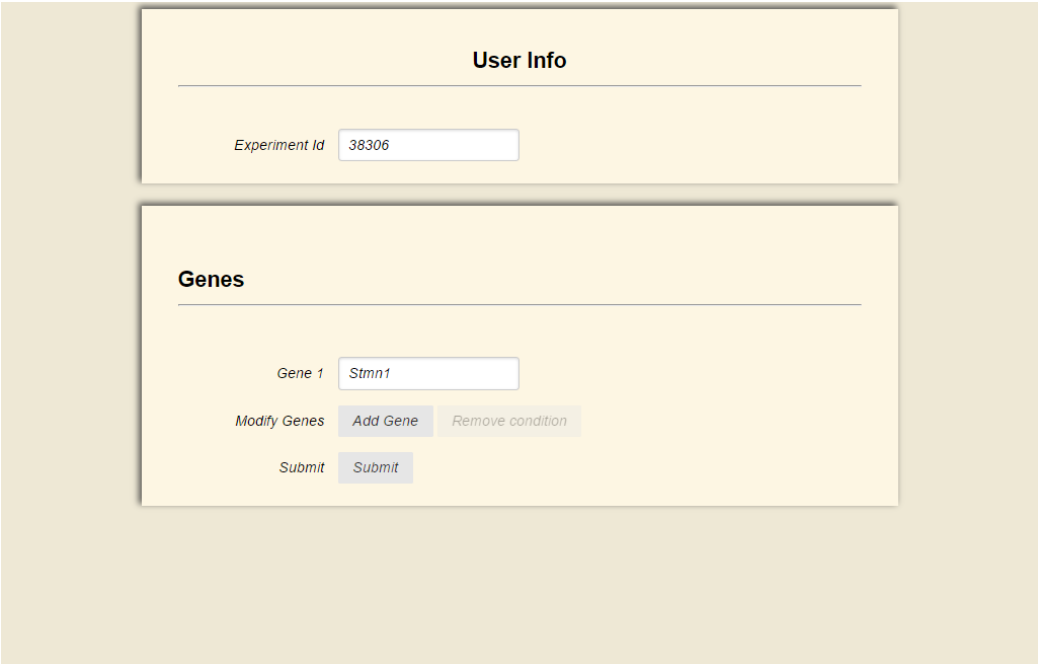


2.12 Classified Heatmap [optional]

This is the same as the manual heatmap. The only difference is that the group and group name will be set based on the classification.

2.13 Differential Gene Analysis [optional]

This shows the gene expression of a specific gene across all the classified groups.


User Info
Experiment Id
38306
Genes
Gene 1
Stmn1
Modify Genes
Add Gene
Remove condition
Submit
Submit

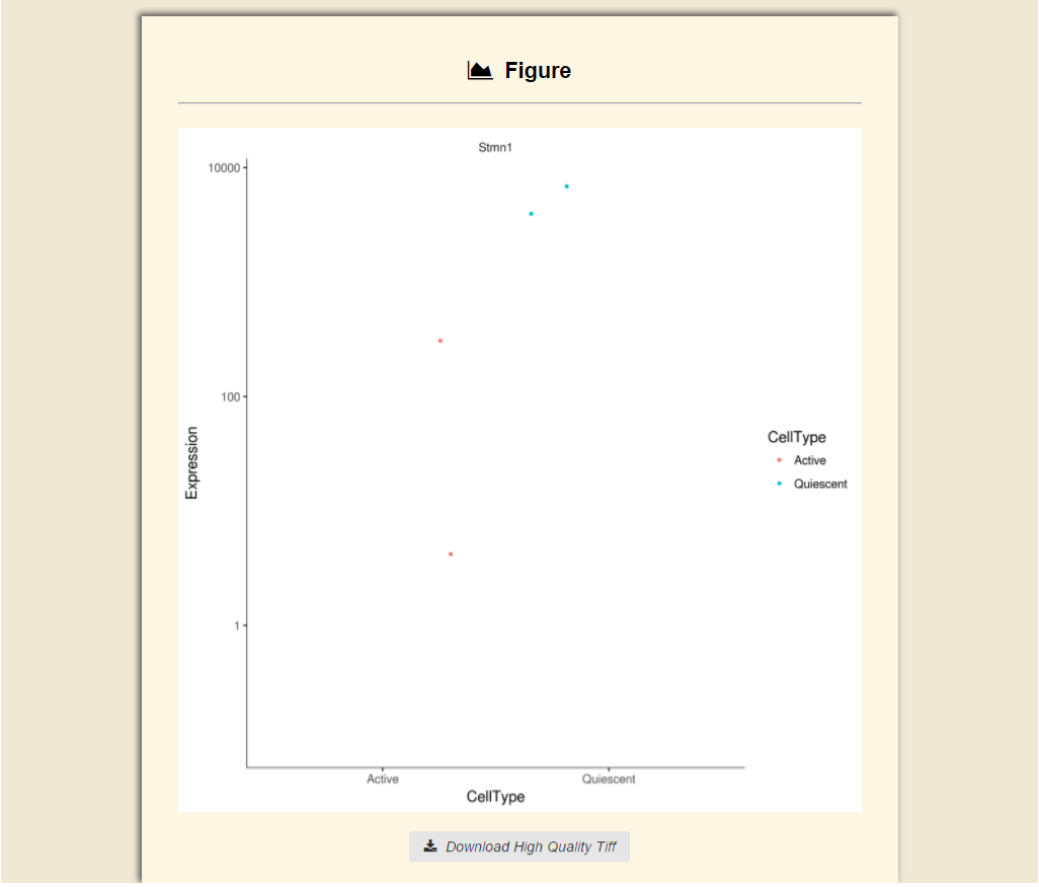
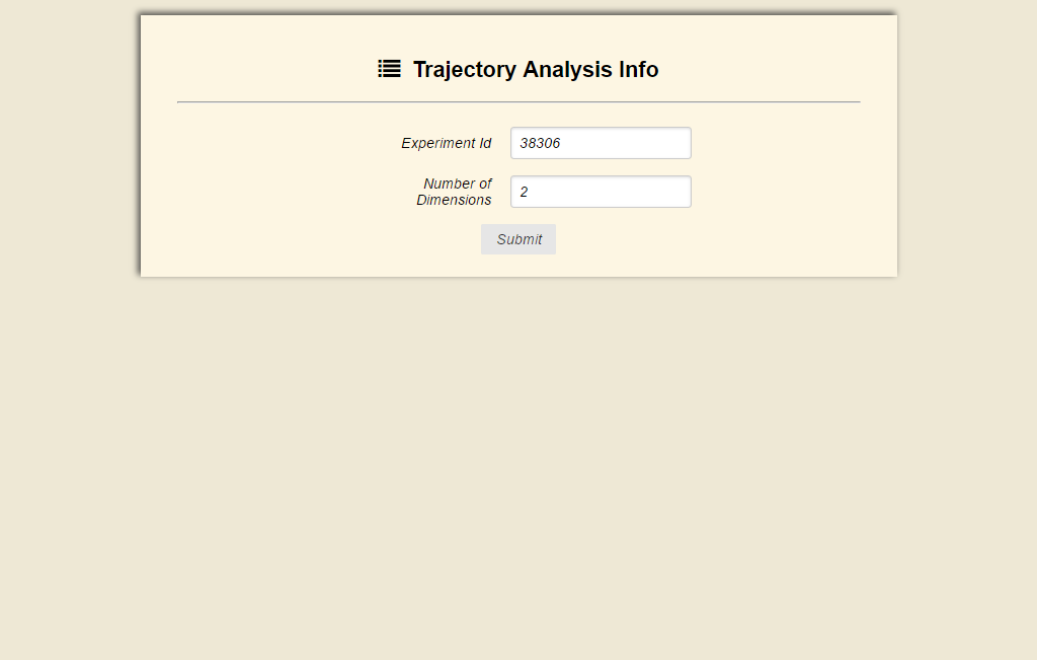


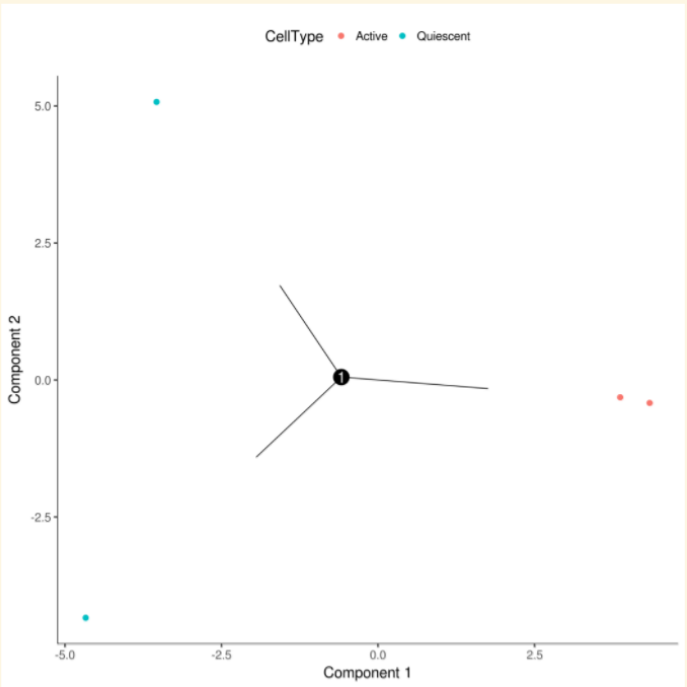
Figure
Stmn1
10000
100
1
Expression
Active
Quiescent
CellType
CellType
Active
Quiescent
Download High Quality Tiff

2.14 Trajectory [optional]

The cells can be ordered in a trajectory. For more information refer to the Monocle manual.



Figure



Download High Quality Tiff