SUPPLEMENTARY MATERIALS FOR:

NASQAR: A web-based platform for High-throughput sequencing data analysis and visualization

Ayman Yousif[1], Nizar Drou[1], Jillian Rowe[1], Mohammed Khalfan[2] and Kristin C. Gunsalus[1,2,–]
[1]NYU Abu Dhabi Center for Genomics & Systems Biology, Division of Biological Sciences, Abu Dhabi, United Arab Emirates and
[2]Center for Genomics & Systems Biology, Department of Biology, New York University, New York, 10003, United States.
–To whom correspondence should be addressed.

1) NASQAR platform and implementation details:

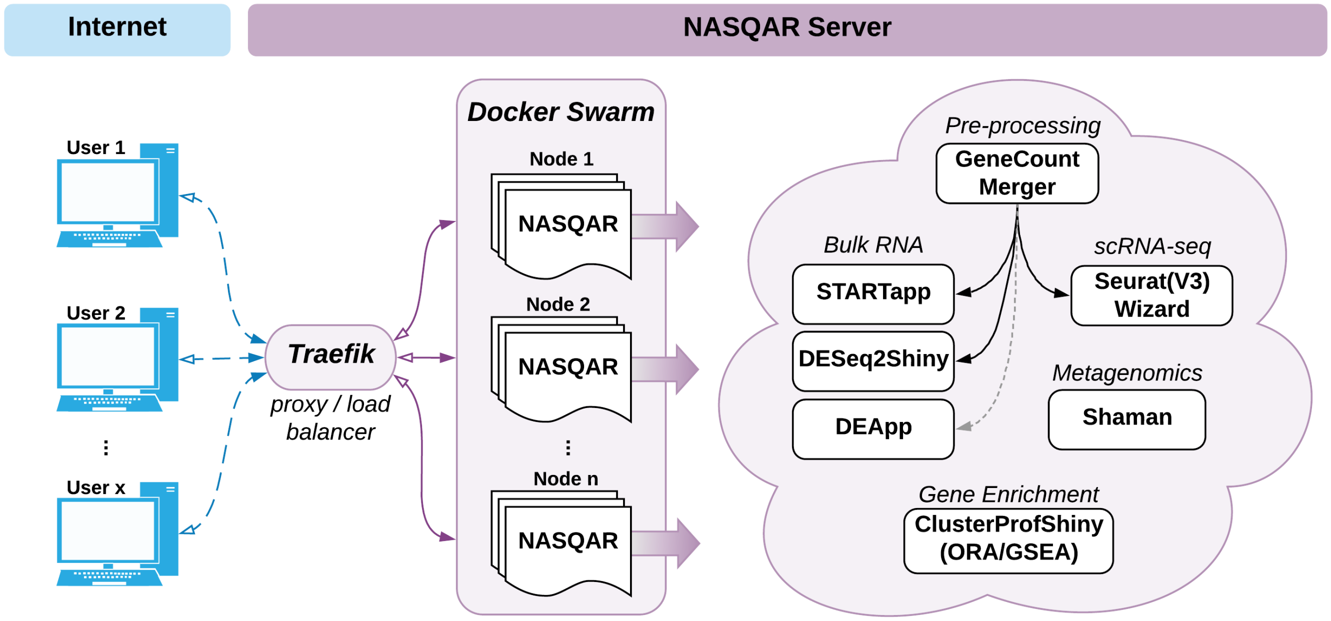


Fig. 1. NASQAR Platform Architecture. A cluster of virtual machines at NYU Abu Dhabi serves NASQAR applications to multiple concurrent users. Applications are containerized and managed on the cluster using Docker and Swarm, while Traefik load-balances requests among available server nodes. Functionality includes merging gene counts, conversion of gene IDs to gene names, analysis of differential mRNA expression, metagenomics analysis, and gene set and functional enrichment analysis. Applications for bulk expression analysis include DESeq2, limma, and EdgeR. Single-cell RNAseq analysis with Seurat Wizards is built on top of the Seurat R package and includes options for filtering, normalization, dimensionality reduction (PCA), clustering, and t-SNE. Enrichment analysis includes applications for Gene Set Enrichment Analysis (GSEA) and Over-representation Analysis (ORA) built using the clusterProfiler R package.

The architectural framework of the NASQAR web platform is illustrated in Figure 1. NASQAR has been deployed on a cluster of virtual machines and is publicly accessible at http://nasqar.abudhabi.nyu.edu/. Docker (Merkel 2014) and Swarm (Soppelsa and Kaewkasi 2017) provide containerization and cluster management, and the Traefik reverse proxy / load balancer (https://traefik.io/) manages requests and maintains sticky user sessions, which is essential for hosting Shiny applications. This framework allows access to multiple users concurrently while providing sufficient resources (RAM/CPU) for the applications. In anticipation of growing computational demand and the addition of more applications, the scalable design makes it relatively easy to increase dedicated resources simply by adding more nodes to the cluster.

A Docker image of NASQAR is publicly available through DockerHub and can be used to deploy the application seamlessly on any system with Docker installed, whether a local computer or a public server. In addition, each application can be installed and launched on its own, saving users from the hassle of satisfying the different software and hardware requirements. The source code is available publicly on GitHub and is actively maintained. All applications have clear user guides with example data sets to help users get started and acclimate quickly.

NASQAR comprises a collection of applications primarily implemented in R, a widely used and freely available statistical programming language (R Core Team 2017). Most of the analysis workflows are built using R libraries for genomics and computation. The front-end design utilizes the R Shiny (Chang *et al.* 2018) library and is supported by JavaScript/CSS to enhance usability and improve overall user experience.

In addition to previously published software, we introduce here several new applications we have developed that wrap around popular analysis packages, such as DESeq2 (Love *et al.* 2014) and Seurat (Butler *et al.* 2018) for bulk and single-cell RNA-seq analysis and visualization, respectively. Since most of the analysis applications in NASQAR require a matrix of gene counts as input, we have also built a convenient tool to assist with preprocessing, GeneCountMerger. Some of the applications have been integrated to provide a seamless transition from data preprocessing to downstream analysis. This implementation gives users the option of using multiple analysis applications without having to modify/reformat the input data set, thus allowing them to easily benchmark and compare the performance of different analysis software packages.

The following is a description of each application currently hosted by NASQAR:

1.1 GeneCountMerger

This preprocessing tool is used to merge individual raw gene count files produced from software such as htseq-count (Pyl *et al.* 2014) or featurecounts (Smyth *et al.* 2013). Options include:

- Merge individual sample count files into one matrix _
- Merge multiple raw count matrices _
- Convert Ensembl gene IDs to gene names _
- Select from available genomes / versions _
- Add pseudocounts _
- Rename sample column headers _
- Download merged counts file in .csv format _
- Seamless transcriptome analysis following count merger (Seurat _Wizard for single-cell RNA analysis; DESeq2Shiny or START (Sklenar *et al.* 2016) for bulk RNA analysis) _

1.2 Seurat Wizards

Seurat Wizards are wizard-style web-based interactive applications to perform guided single-cell RNA-seq data analysis and visualization. They are based on Seurat, a popular R package designed for QC, analysis, and exploration of single-cell RNAseq data. The wizard style makes it intuitive to go back and forth between steps and adjust parameters based on the results/feedback of different outputs/plots/steps, giving the user the ability to interactively tune the analysis. SeuratWizard and SeuratV3Wizard implementations provide support for Seurat versions 2 and 3 (Stuart *et al.* 2019), respectively. _

Another web-based tool for scRNA-seq analysis, IS-CellR(Patel 2018), has recently been described that also utilizes Seurat v2. The SeuratWizard and SeuratV3Wizard take a different approach to design and implementation and follow closely the Seurat Guided Clustering Tutorials devised by the authors (https://satijalab.org/seurat/v3. 0/pbmc3k_tutorial.html). Users can follow the tutorials while using the Wizards and can edit parameters at almost every step, which is instrumental in producing accurate results. A unique feature of the Seurat Wizards is that they can accept as input processed 10X Genomics data files in place of a matrix of gene counts, which eliminates the need for this additional pre-processing step. SeuratV3Wizard integrates several additional features like the UCSC Cell Browser (https://github. com/maximilianh/cellBrowser), enabling users to interactively visualize clusters and gene markers, and the newly published sctransform method (Hafemeister and Satija 2019), which gives users the ability to run the analysis using two slightly different workflows and compare the results. These differences in features and design give the Seurat Wizards more versatility and improve usability in comparison with other publicly available implementations of Seurat.

1.3 Deseq2Shiny

The Deseq2Shiny app is a Shiny wrapper around DESeq2, a popular R package for performing differential mRNA expression analysis of RNA-seq data. This web-based application implements the standard default workflow outlined in_the DESeq2 Bioconductor tutorial (https://bioconductor.org/ packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2. html). This includes normalization, data transformation (e.g., rlog and _vst for clustering), and estimation for dispersion and log fold change. This app follows the same implementation as other apps on NASQAR, whereby users can fine tune the analysis parameters interactively.

1.4 ClusterProfilerShiny

The ClusterProfilerShiny apps wrap the popular clusterProfiler (Yu *et al.* 2012) package, which implements methods to analyze and visualize functional profiles of genomic coordinates, genes, and gene clusters. Users can upload their own data from DESeq2 or import data from the upstream Deseq2Shiny app. These apps allow for quick and easy over-representation analysis (ORA) and gene set enrichment analysis (GSEA) of GO terms and KEGG pathways. Visuals produced include dot plots, word clouds, category net plots, enrichment map plots, GO induced graphs, GSEA plots, and enriched KEGG pathway plots using the Pathview package (Luo and Brouwer 2013).

1.5 Other open-source apps

- START: a web-based RNA-seq analysis and visualization resource. We have modified this application slightly from the open-source version to add options to some plots. We have also integrated it with GeneCountMerger so that once merging gene counts is complete, users can launch the START app and have their merged matrix data loaded automatically. _
- DEApp (Li and Andrade 2017): an interactive web application for differential expression analysis. _

- Shaman (Quereda *et al.* 2016): a Shiny application that enables the identification of differentially abundant genera within metagenomic datasets. It wraps around the Generalized Linear Model implemented in DESeq2. It includes multiple visualizations, and is compatible with common metagenomic file formats. _

2) Launch NASQAR using Docker (Recommended):

The recommended way to get NASQAR running is using Docker. The reason is that applications hosted within NASQAR have many package dependencies (R and OS) that might be tedious and very time consuming for the average user especially when trying to support different OS's (Windows/Linux/OSX). **Prerequisite**: Make sure Docker (version >= 17.03.0-ce, https://docs.docker.com/install/) is installed.

Run NASQAR docker image as follows:

a) docker run -p 80:80 aymanm/nasqarall:nasqar (runs on port 80)
   If you are running this in your personal laptop/PC, access NASQAR using a modern web browser at the following URL http://localhost/
b) docker run -p 8083:80 aymanm/nasqarall:nasqar (runs on port 8083)
   If you are running this in your personal laptop/PC, access NASQAR using a modern web browser at the following URL http://localhost:8083/
c) If you are hosting this as a service at your organization, make sure the specified port is not blocked with a firewall so users can access the service. Execute the same commands as a) and b). Users can access NASQAR using a modern web browser at the following URL http://<server_ip>:port/

Note: All apps in NASQAR can be launched individually. Visit each app's GitHub page for relevant instructions. For 3rd party apps hosted on NASQAR, please refer to their Github repositories.

App GitHub pages:

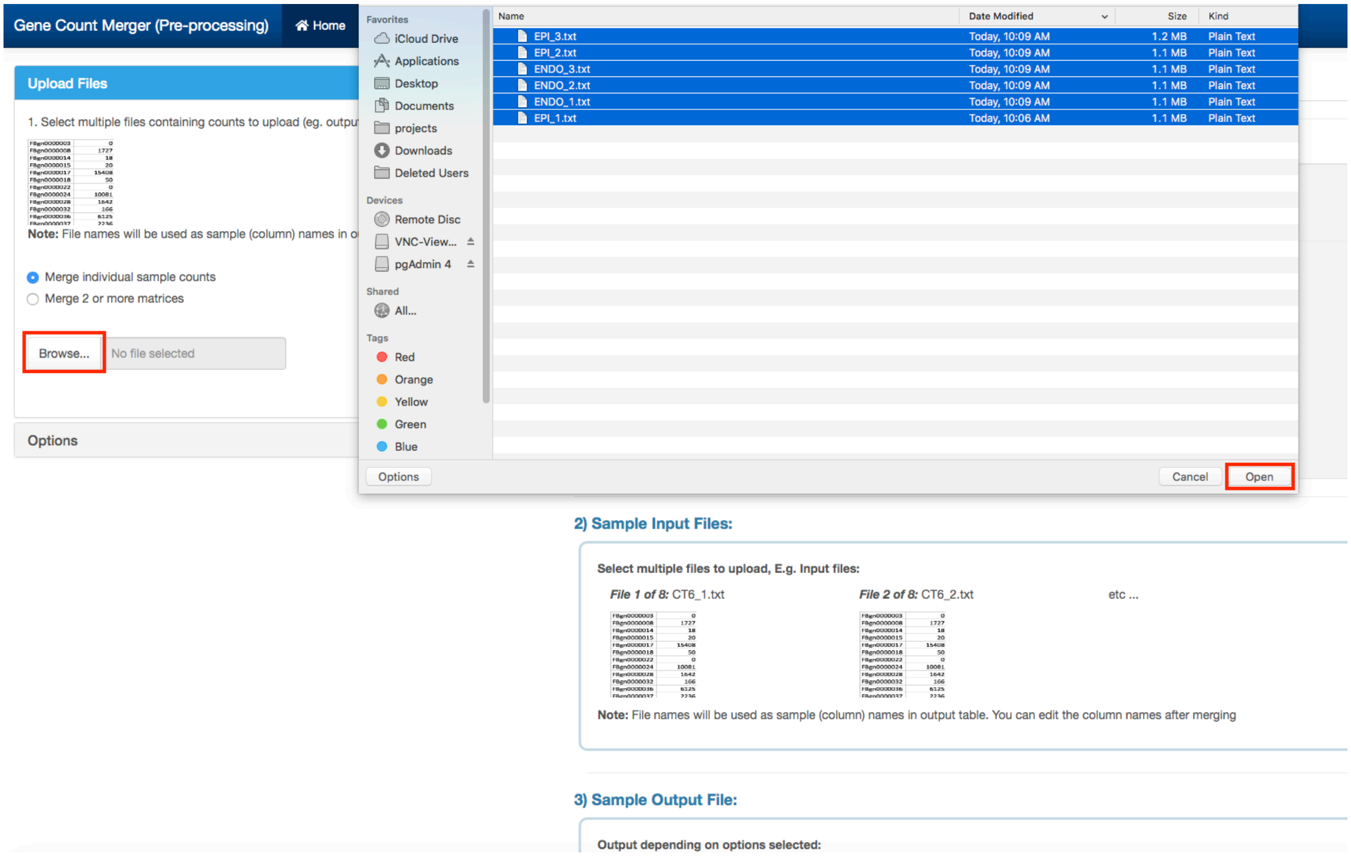
_ SeuratV3Wizard (scRNA): https://github.com/nasqar/seuratv3wizard
_ SeuratWizard (scRNA): https://github.com/nasqar/SeuratWizard
_ deseq2shiny (Bulk RNA): https://github.com/nasqar/deseq2shiny
_ GeneCountMerger (Pre-processing): https://github.com/nasqar/GeneCountMerger
_ ClusterProfShinyGSEA (Enrichment): https://github.com/nasqar/ClusterProfShinyGSEA
_ ClusterProfShinyORA (Enrichment): https://github.com/nasqar/ClusterProfShinyORA
_ NASQAR (main page): https://github.com/nasqar/NASQAR

3) Example Use Case 1 (DGE and GSEA):

In this example we will show, using only your web browser, how you can start with individual sample gene count files (eg output of htseq counts) and carry out Differential Gene Expression (DGE) and Gene Set Enrichment analysis using DESeq2 and clusterProfiler R packages respectively.
The datasets provided in this example use case have been download from the ENSEMBL expression ATLAS (https://www.ebi.ac.uk/gxa/experiments/E-MTAB-970/Results). They belong to the ENSEMBL expression ATLAS experiment title “***Transcription profiling by high throughput sequencing of Sox17.Epi and Endo cells from mouse embryos***” ([dataset] 2016). The data comprises of 6 mouse RNA-seq samples, across two conditions. They were selected as an example dataset only, no other criteria were used in their selection.

Step 1: Merge counts (GeneCountMerger):

a) Download example count files zip from here (https://drive.google.com/file/d/1OB2vojHqscLAHZWGETHdGK0yaZ9r06PT/iew?usp=sharing)
b) Extract zip file
c) Launch GeneCountMerger http://nasqar.abudhabi.nyu.edu/GeneCountMerger/
d) Click browse and select all gene count files:



1. Select multiple files containing counts to upload (eg. outpu

FBgn0000003	0
FBgn0000008	1727
FBgn0000014	18
FBgn0000015	20
FBgn0000017	15408
FBgn0000018	50
FBgn0000022	0
FBgn0000024	10081
FBgn0000028	1642
FBgn0000032	166
FBgn0000036	6125
FBgn0000037	2236

Note: File names will be used as sample (column) names in o

● Merge individual sample counts
○ Merge 2 or more matrices

Name	Date Modified	Size	Kind
EPI_3.txt	Today, 10:09 AM	1.2 MB	Plain Text
EPI_2.txt	Today, 10:09 AM	1.1 MB	Plain Text
ENDO_3.txt	Today, 10:09 AM	1.1 MB	Plain Text
ENDO_2.txt	Today, 10:09 AM	1.1 MB	Plain Text
ENDO_1.txt	Today, 10:09 AM	1.1 MB	Plain Text
EPI_1.txt	Today, 10:06 AM	1.1 MB	Plain Text

2) Sample Input Files:

Select multiple files to upload, E.g. Input files:

File 1 of 8: CT6_1.txt

FBgn0000003	0
FBgn0000008	1727
FBgn0000014	18
FBgn0000015	20
FBgn0000017	15408
FBgn0000018	50
FBgn0000022	0
FBgn0000024	10081
FBgn0000028	1642
FBgn0000032	166
FBgn0000036	6125
FBgn0000037	2236

File 2 of 8: CT6_2.txt

FBgn0000003	0
FBgn0000008	1727
FBgn0000014	18
FBgn0000015	20
FBgn0000017	15408
FBgn0000018	50
FBgn0000022	0
FBgn0000024	10081
FBgn0000028	1642
FBgn0000032	166
FBgn0000036	6125
FBgn0000037	2236

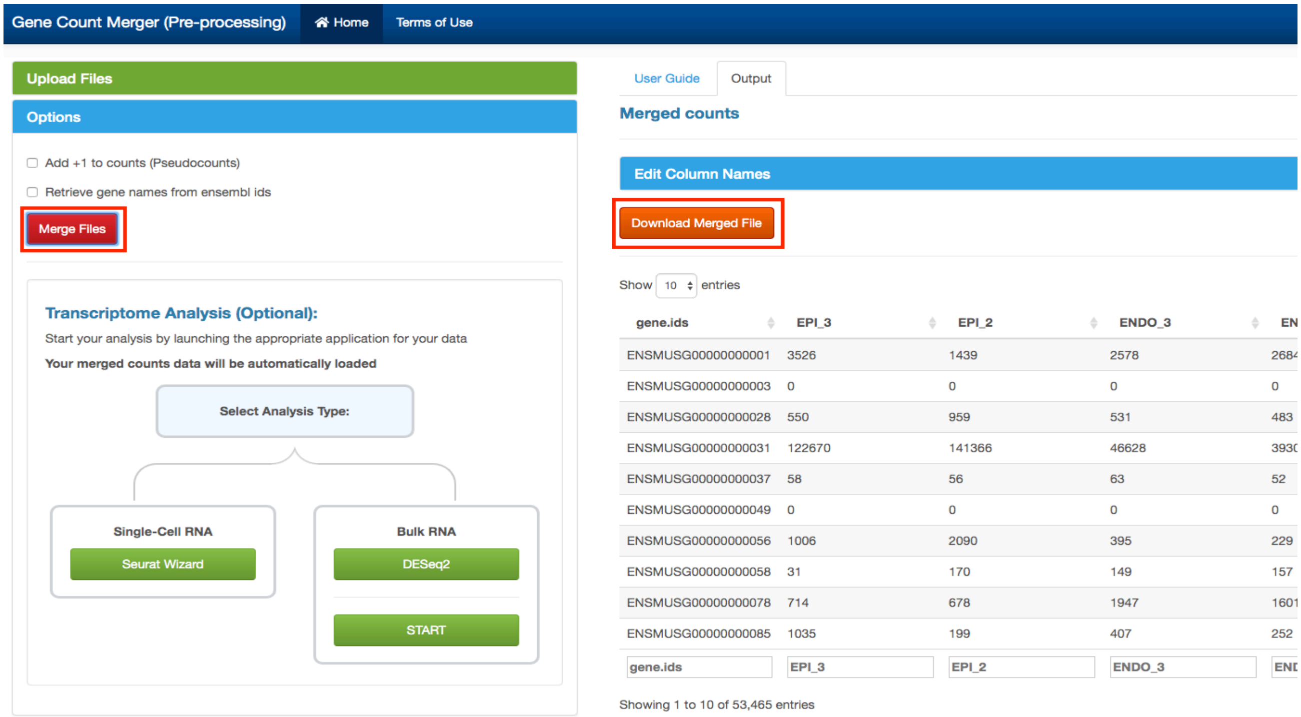
etc ...

Note: File names will be used as sample (column) names in output table. You can edit the column names after merging

3) Sample Output File:

Output depending on options selected:

e) Once loaded, click the red Merge button. Now all counts files have been consolidated into one table. You can now download and save it as a .csv file



□ Add +1 to counts (Pseudocounts)

□ Retrieve gene names from ensembl ids

Merge Files

Transcriptome Analysis (Optional):

Start your analysis by launching the appropriate application for your data

Your merged counts data will be automatically loaded

Merged counts

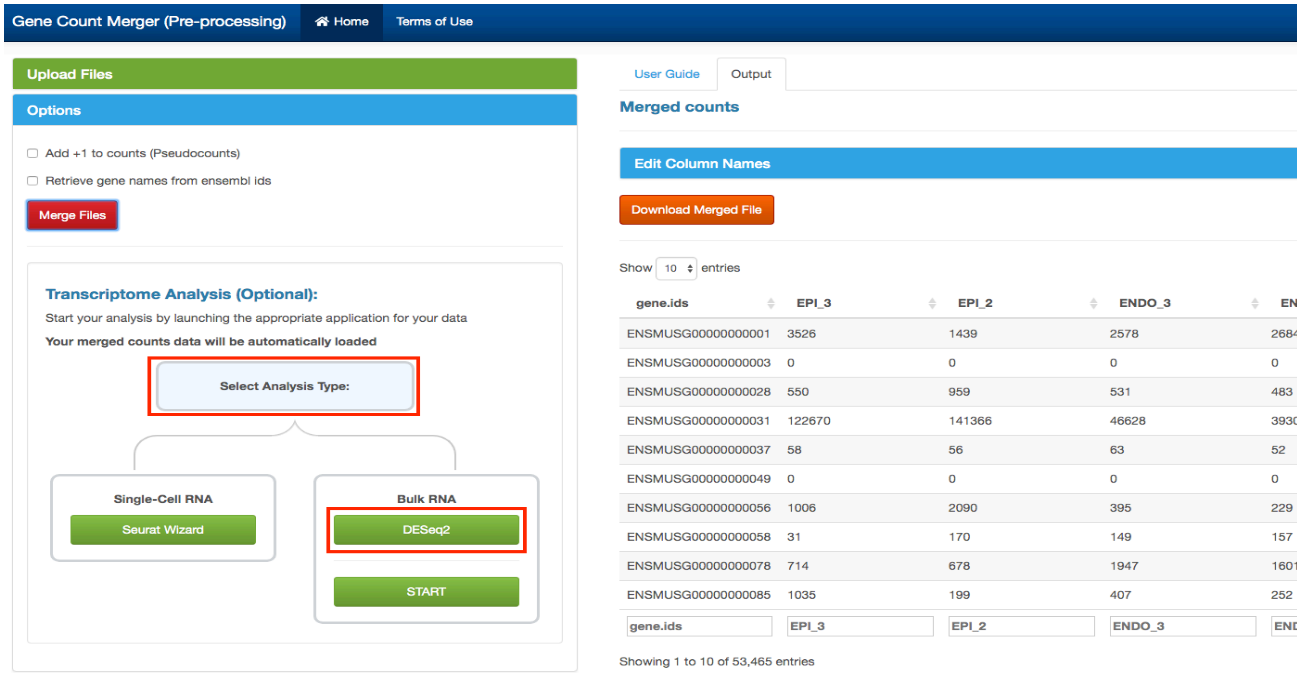
Show 10 entries

gene.ids	EPI_3	EPI_2	ENDO_3	EN
ENSMUSG00000000001	3526	1439	2578	2684
ENSMUSG00000000003	0	0	0	0
ENSMUSG00000000028	550	959	531	483
ENSMUSG00000000031	122670	141366	46628	3930
ENSMUSG00000000037	58	56	63	52
ENSMUSG00000000049	0	0	0	0
ENSMUSG00000000056	1006	2090	395	229
ENSMUSG00000000058	31	170	149	157
ENSMUSG00000000078	714	678	1947	1601
ENSMUSG00000000085	1035	199	407	252
gene.ids	EPI_3	EPI_2	ENDO_3	END

Showing 1 to 10 of 53,465 entries

Step 2: Differential Gene Expression analysis (Deseq2Shiny):

a) Carrying on from the previous step, under “**Select Analysis Type**” select “**DESeq2**”



Gene Count Merger (Pre-processing) | Home | Terms of Use

Upload Files

Options

□ Add +1 to counts (Pseudocounts)
□ Retrieve gene names from ensembl ids

Merge Files

Transcriptome Analysis (Optional):

Start your analysis by launching the appropriate application for your data

Your merged counts data will be automatically loaded

Select Analysis Type:

Single-Cell RNA: Seurat Wizard

Bulk RNA: DESeq2

START

User Guide | Output

Merged counts

Edit Column Names

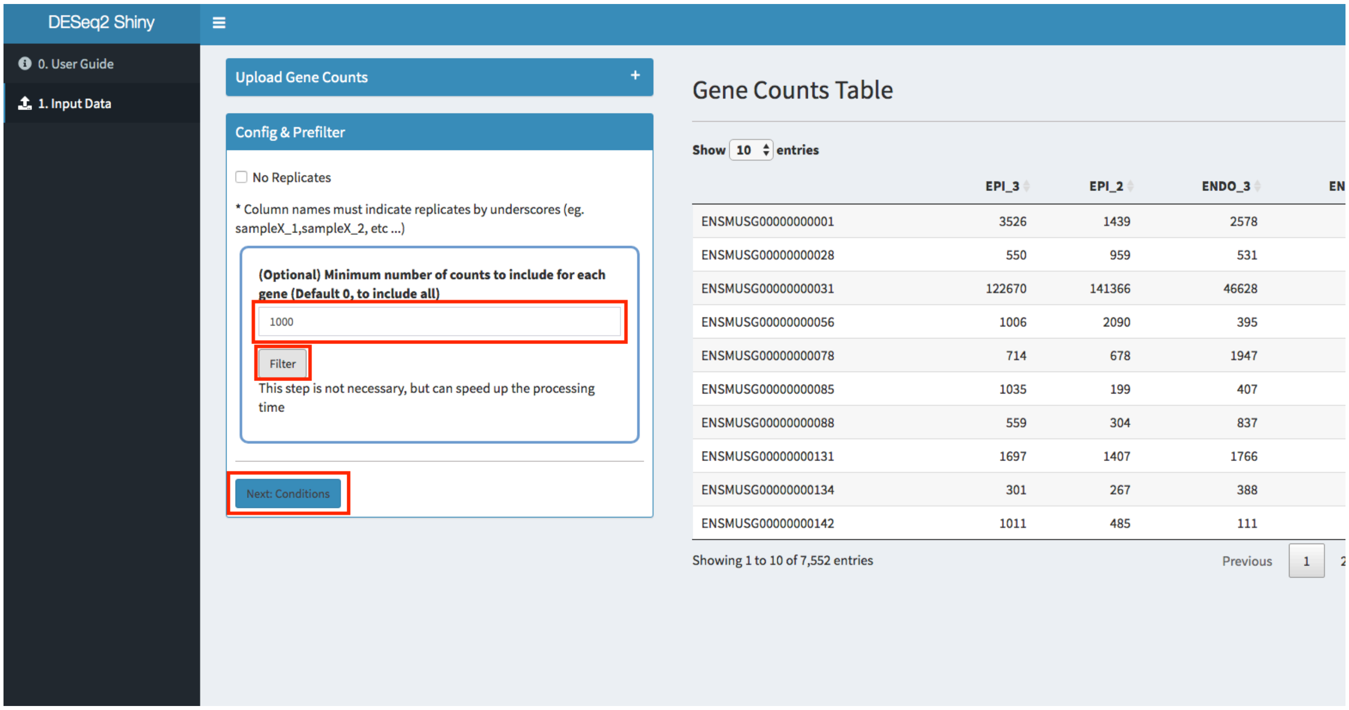
Download Merged File

Show 10 entries

gene.ids	EPI_3	EPI_2	ENDO_3	EN
ENSMUSG00000000001	3526	1439	2578	2684
ENSMUSG00000000003	0	0	0	0
ENSMUSG00000000028	550	959	531	483
ENSMUSG00000000031	122670	141366	46628	3930
ENSMUSG00000000037	58	56	63	52
ENSMUSG00000000049	0	0	0	0
ENSMUSG00000000056	1006	2090	395	229
ENSMUSG00000000058	31	170	149	157
ENSMUSG00000000078	714	678	1947	1601
ENSMUSG00000000085	1035	199	407	252
gene.ids	EPI_3	EPI_2	ENDO_3	END

Showing 1 to 10 of 53,465 entries

b) The Deseq2Shiny app will be launched in a new tab, with all counts data already loaded onto it. You can set the minimum number of counts to 1000 for example (this is mainly for speed up, you can continue with out it) and click “**Filter**”. Then click “**Next: Conditions**”



DESeq2 Shiny

0. User Guide
1. Input Data

Upload Gene Counts

Config & Prefilter

□ No Replicates

* Column names must indicate replicates by underscores (eg. sampleX_1,sampleX_2, etc ...)

(Optional) Minimum number of counts to include for each gene (Default 0, to include all)

1000

Filter

This step is not necessary, but can speed up the processing time

Next: Conditions

Gene Counts Table

Show 10 entries

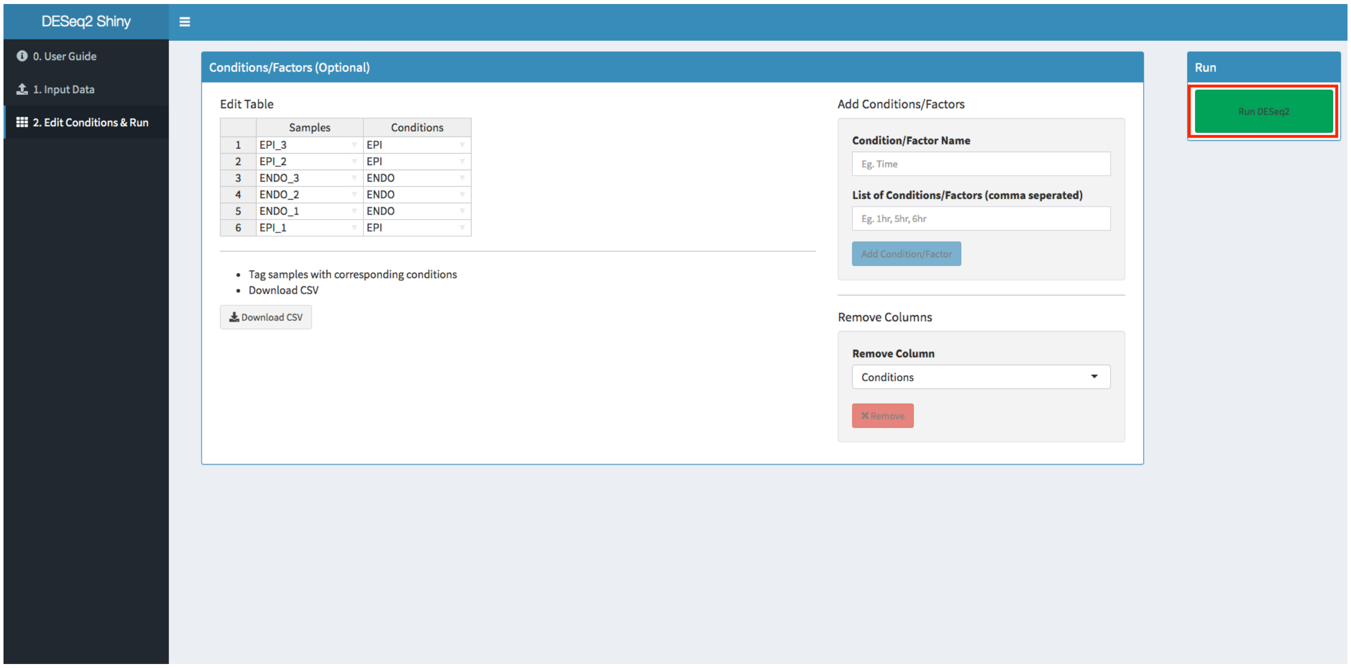
	EPI_3	EPI_2	ENDO_3	EN
ENSMUSG00000000001	3526	1439	2578	
ENSMUSG00000000028	550	959	531	
ENSMUSG00000000031	122670	141366	46628	
ENSMUSG00000000056	1006	2090	395	
ENSMUSG00000000078	714	678	1947	
ENSMUSG00000000085	1035	199	407	
ENSMUSG00000000088	559	304	837	
ENSMUSG00000000131	1697	1407	1766	
ENSMUSG00000000134	301	267	388	
ENSMUSG00000000142	1011	485	111	

Showing 1 to 10 of 7,552 entries

Previous 1 2

c) You can verify the sample/conditions table on the left, then click “**Run DESeq2**”. The replicate/condition IDs are inferred from the naming of the raw counts files. For example, ENDO_1 will be automatically determined to mean that the sample belongs to the condition

"ENDO", and that it is the first replicate. In case a user supplies filenames that are named differently, this page will allow users to tag their replicates to the appropriate conditions.



Conditions/Factors (Optional)

Edit Table

	Samples	Conditions
1	EPI_3	EPI
2	EPI_2	EPI
3	ENDO_3	ENDO
4	ENDO_2	ENDO
5	ENDO_1	ENDO
6	EPI_1	EPI

- Tag samples with corresponding conditions
- Download CSV

Download CSV

Add Conditions/Factors

Condition/Factor Name: Eg. Time

List of Conditions/Factors (comma seperated): Eg. 1hr, 5hr, 6hr

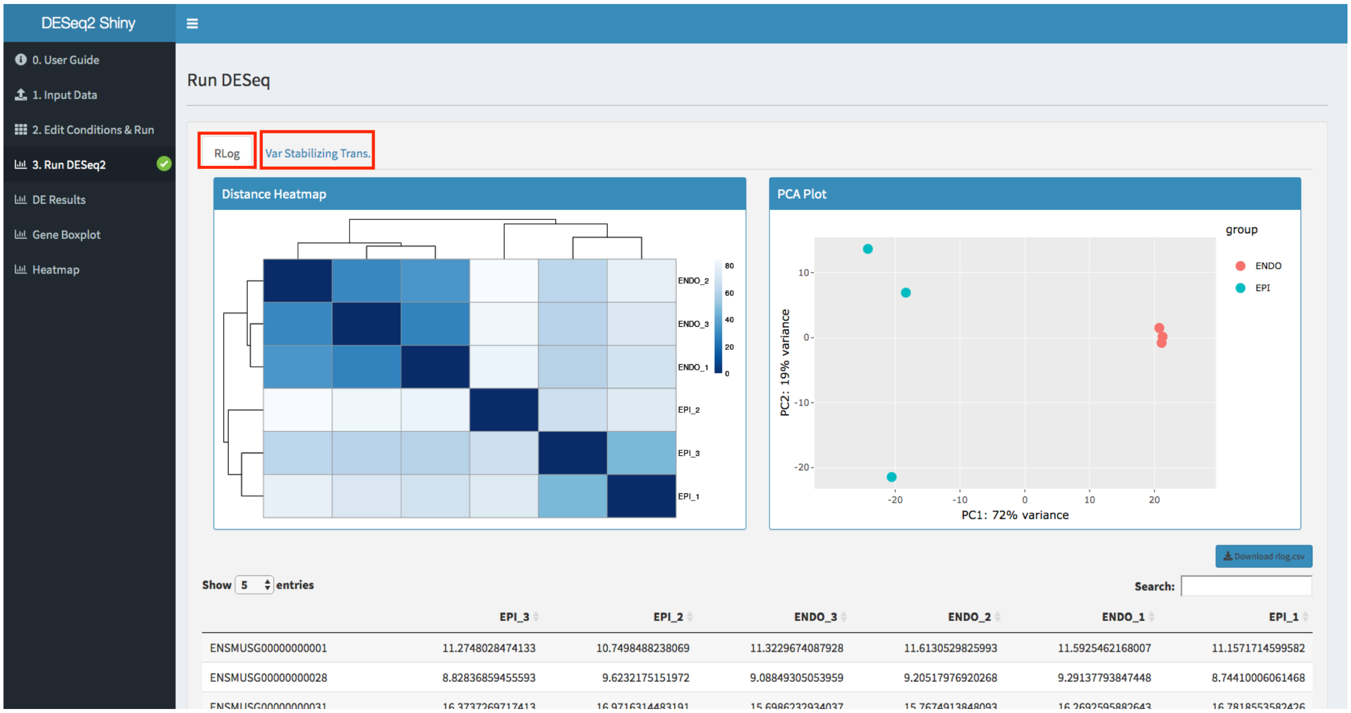
Add Condition/Factor

Remove Columns

Remove Column: Conditions

✖ Remove

d) Once DESeq2 has completed, you will be able to see rlog/vst transformation matrices and PCA and distance heatmap plots.

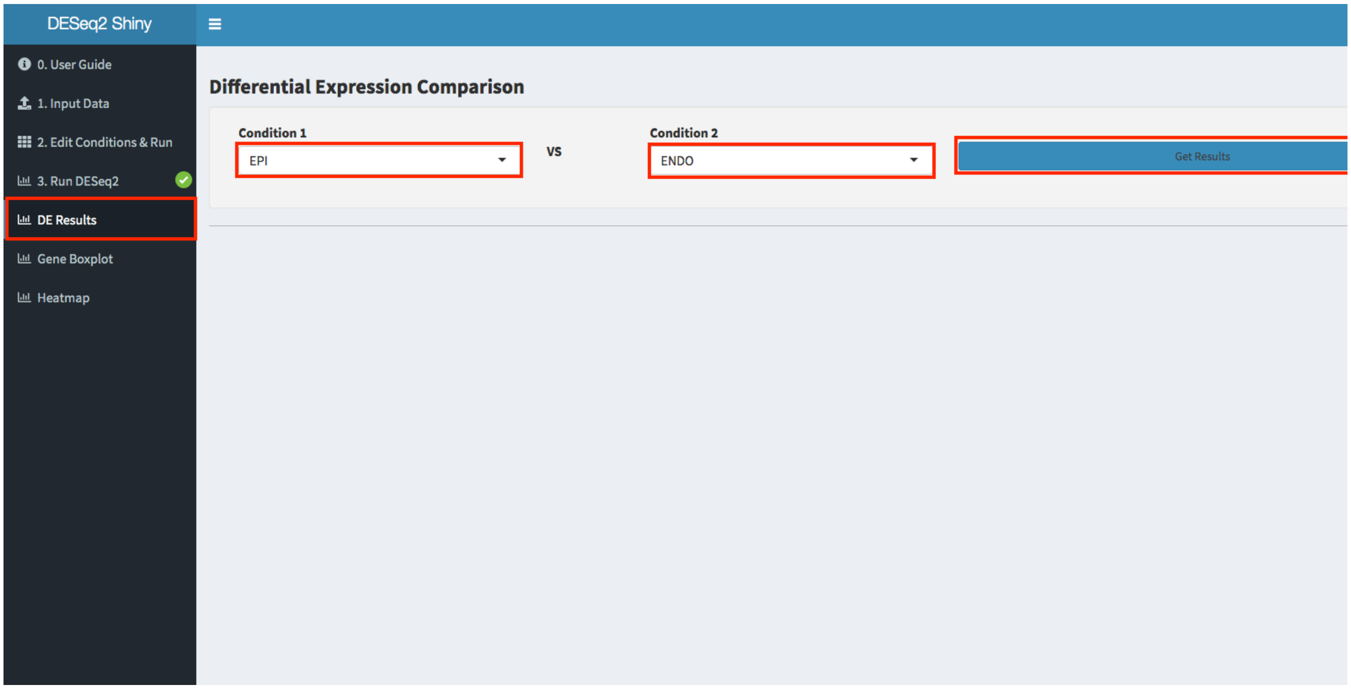


Show 5 entries

Search:

	EPI_3	EPI_2	ENDO_3	ENDO_2	ENDO_1	EPI_1
ENSMUSG00000000001	11.2748028474133	10.7498488238069	11.3229674087928	11.6130529825993	11.5925462168007	11.1571714599582
ENSMUSG00000000028	8.82836859455593	9.6232175151972	9.08849305053959	9.20517976920268	9.29137793847448	8.74410006061468
ENSMUSG00000000031	16.3737269717413	16.9716314483191	15.6986232934037	15.7674913848093	16.2692595882643	16.7818553582426

e) Next, go to “**DE Results**” to be able to run comparisons between different sample conditions. Select two different conditions, for example “EPI” and “ENDO” and then click “**Get Results**”

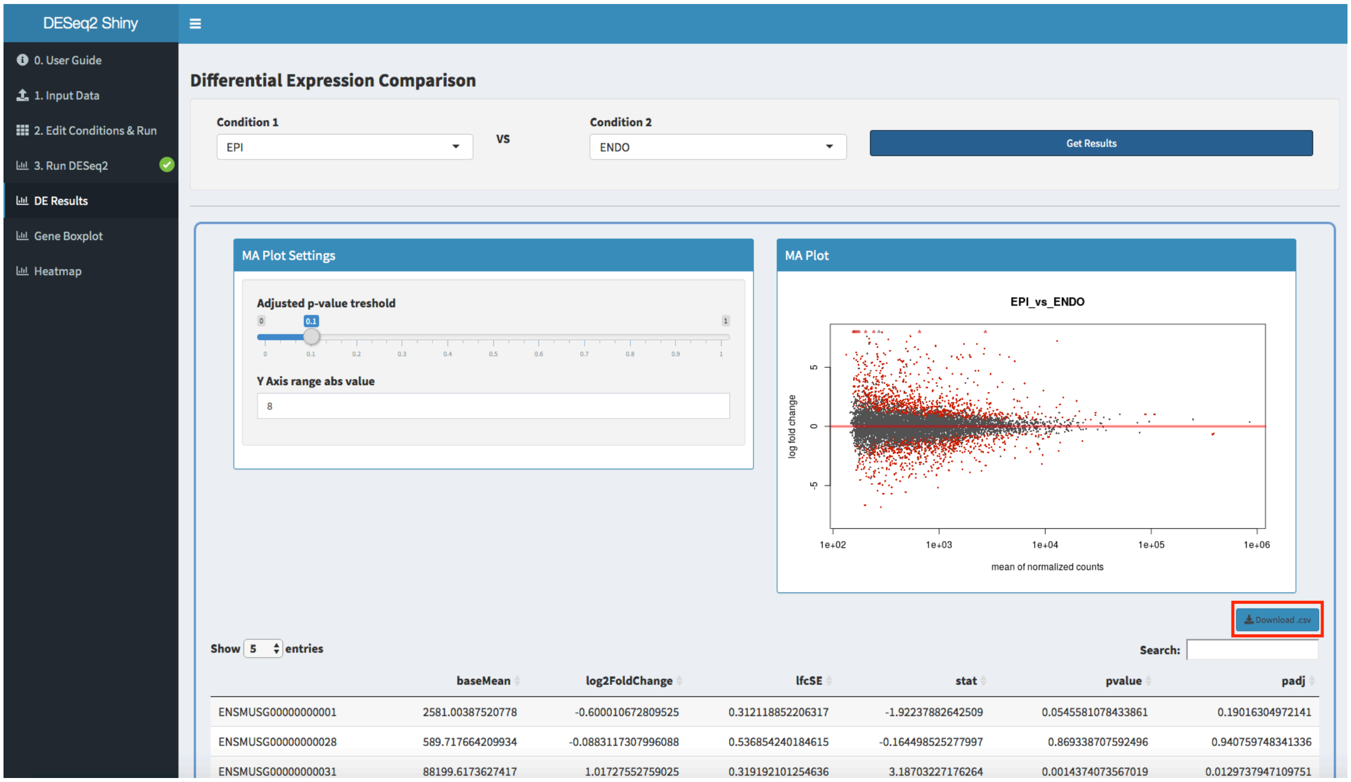


Differential Expression Comparison

Condition 1: EPI

Condition 2: ENDO

f) Scroll down and you can see the DE comparison results. Click “**Download .csv**”. This will save a CSV file locally called “EPI_vs_ENDO.csv”.



Differential Expression Comparison

Condition 1: EPI

Condition 2: ENDO

Adjusted p-value treshold: 0.1

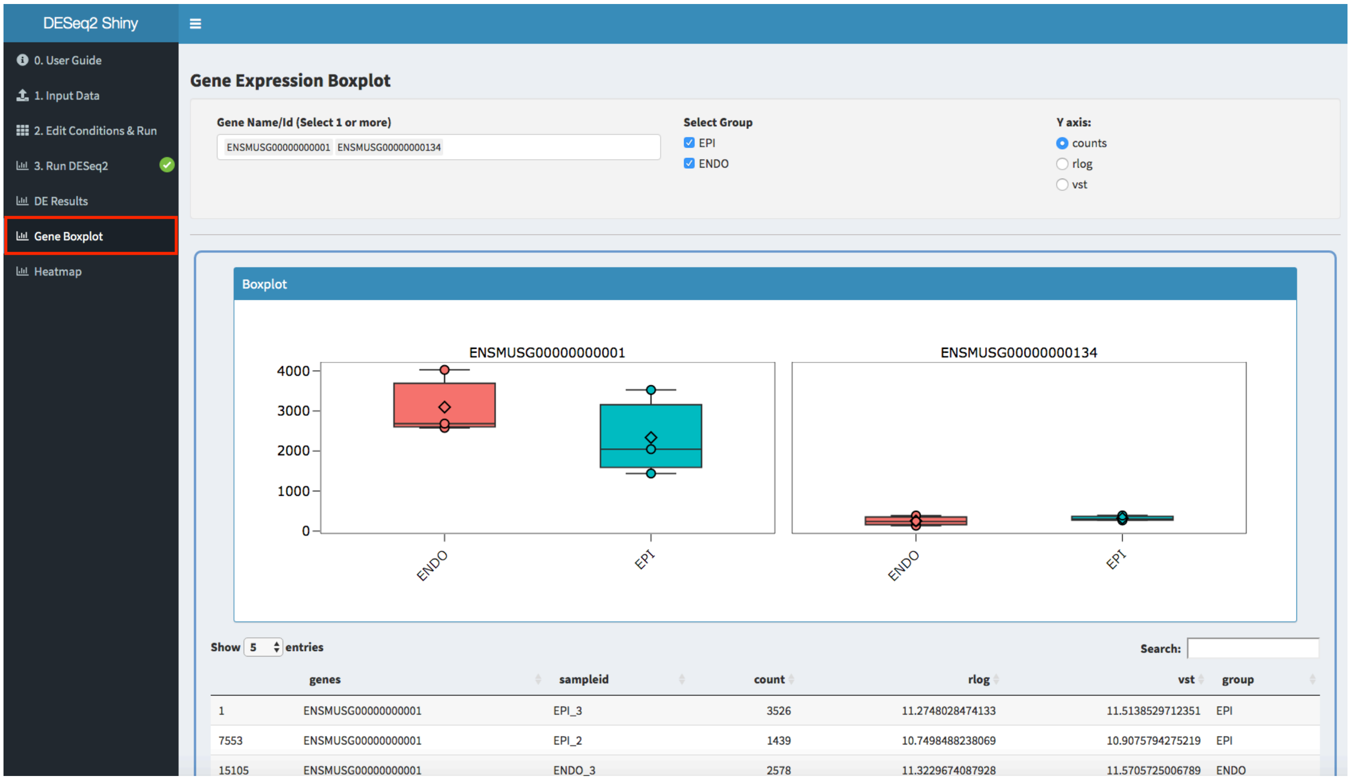
Y Axis range abs value: 8

Show 5 entries

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
ENSMUSG00000000001	2581.00387520778	-0.600010672809525	0.312118852206317	-1.92237882642509	0.0545581078433861	0.19016304972141
ENSMUSG00000000028	589.717664209934	-0.0883117307996088	0.536854240184615	-0.164498525277997	0.869338707592496	0.940759748341336
ENSMUSG00000000031	88199.6173627417	1.01727552759025	0.319192101254636	3.18703227176264	0.0014374073567019	0.0129737947109751

g) There are other visualization plots like gene box plots and clustered heatmap

Box Plots:



Gene Expression Boxplot

Gene Name/Id (Select 1 or more): ENSMUSG00000000001 ENSMUSG00000000134

Select Group
☑ EPI
☑ ENDO

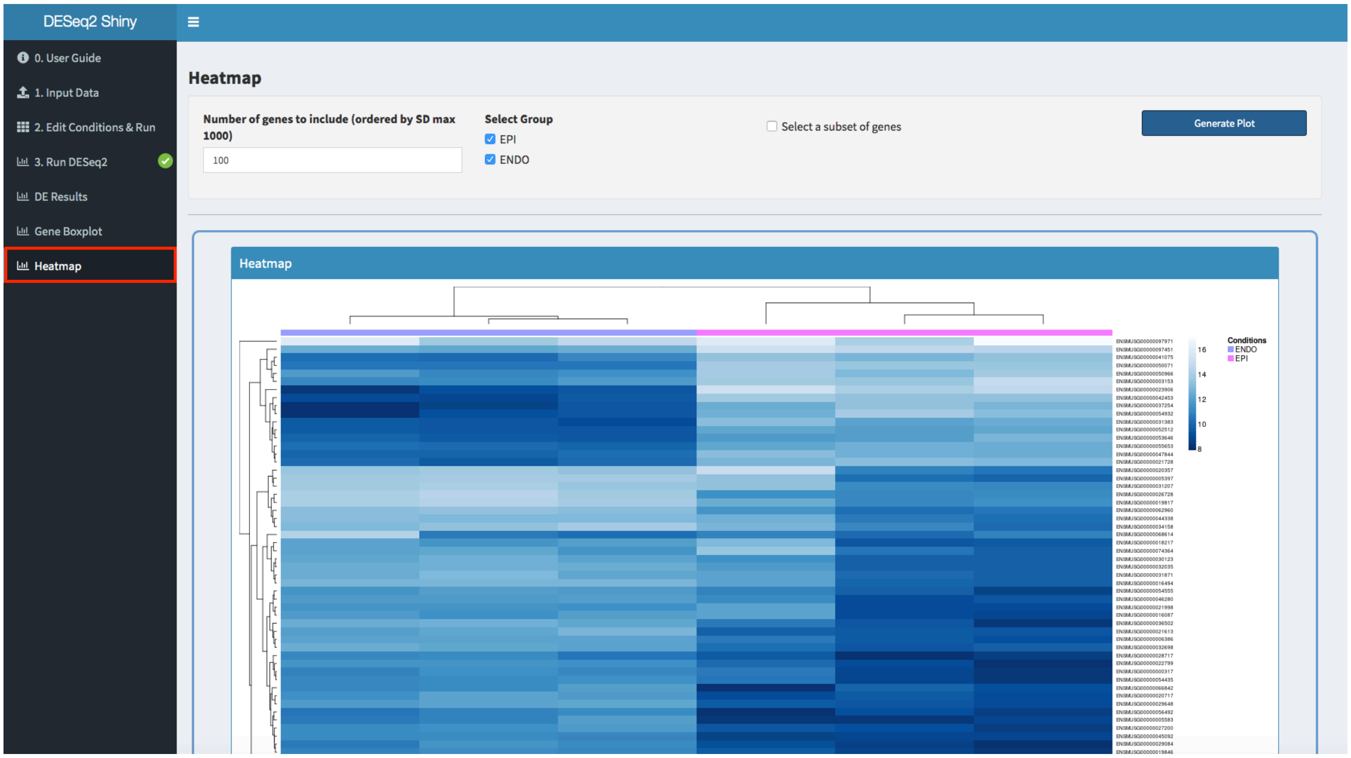
Y axis:
● counts
○ rlog
○ vst

Show 5 entries

Search:

	genes	sampleid	count	rlog	vst	group
1	ENSMUSG00000000001	EPI_3	3526	11.2748028474133	11.5138529712351	EPI
7553	ENSMUSG00000000001	EPI_2	1439	10.7498488238069	10.9075794275219	EPI
15105	ENSMUSG00000000001	ENDO_3	2578	11.3229674087928	11.5705725006789	ENDO

Clustered Heatmap:



Heatmap

Number of genes to include (ordered by SD max 1000): 100

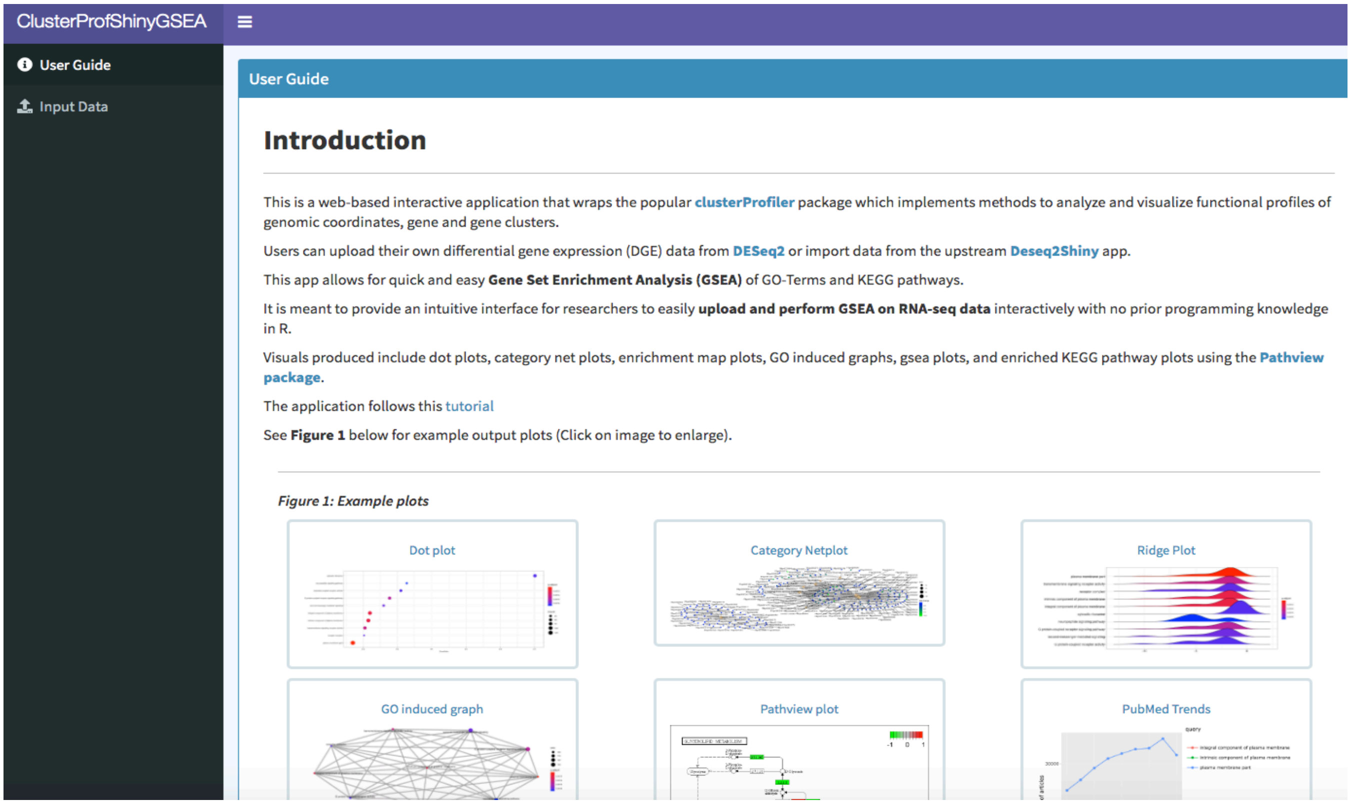
Select Group
☑ EPI
☑ ENDO

□ Select a subset of genes

Generate Plot

Step 3: Gene Set Enrichment Analysis (GSEA):

a) Launch ClusterProfShinyGSEA http://nasqar.abudhabi.nyu.edu/ClusterProfShinyGSEA/



Introduction

This is a web-based interactive application that wraps the popular **clusterProfiler** package which implements methods to analyze and visualize functional profiles of genomic coordinates, gene and gene clusters.

Users can upload their own differential gene expression (DGE) data from **DESeq2** or import data from the upstream **Deseq2Shiny** app.

This app allows for quick and easy **Gene Set Enrichment Analysis (GSEA)** of GO-Terms and KEGG pathways.

It is meant to provide an intuitive interface for researchers to easily **upload and perform GSEA on RNA-seq data** interactively with no prior programming knowledge in R.

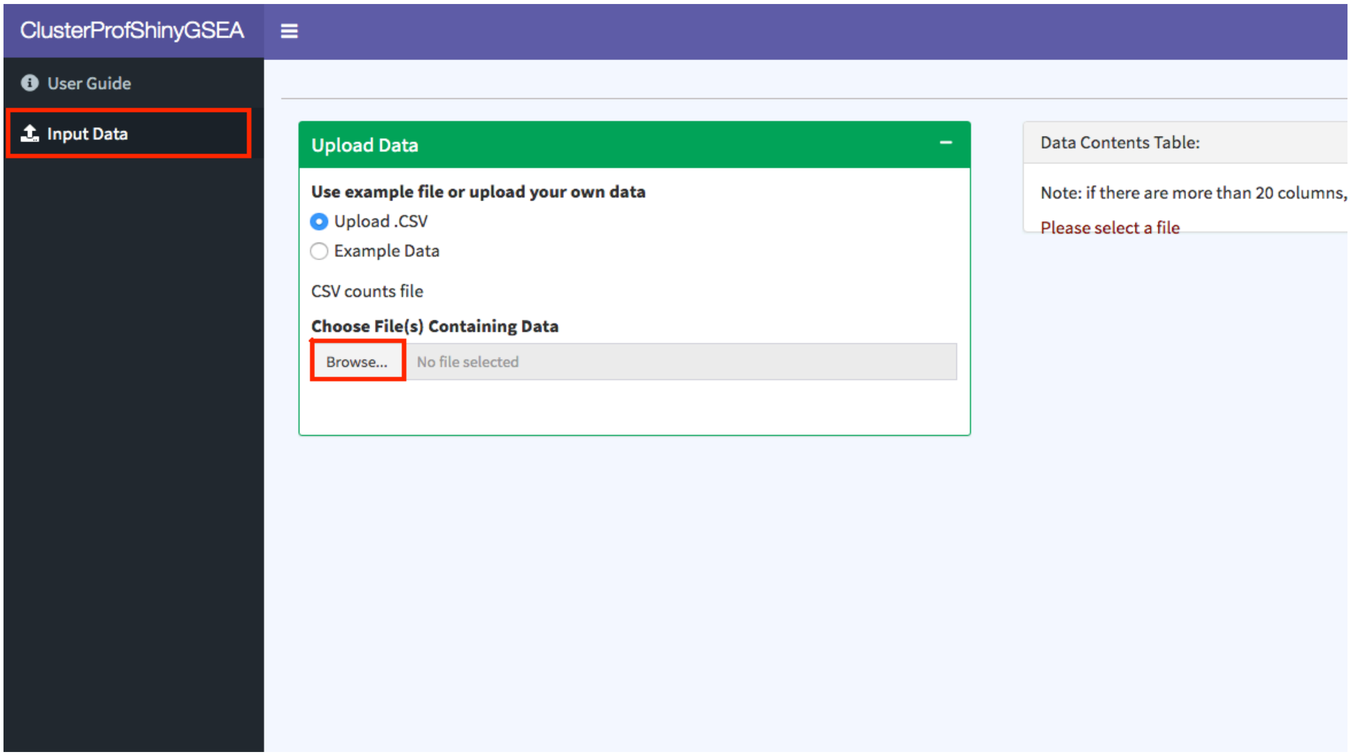
Visuals produced include dot plots, category net plots, enrichment map plots, GO induced graphs, gsea plots, and enriched KEGG pathway plots using the **Pathview package**.

The application follows this tutorial

See **Figure 1** below for example output plots (Click on image to enlarge).

Figure 1: Example plots

b) Go to "**Input Data**" tab and click "**Browse**"



Use example file or upload your own data

● Upload .CSV

○ Example Data

CSV counts file

Choose File(s) Containing Data

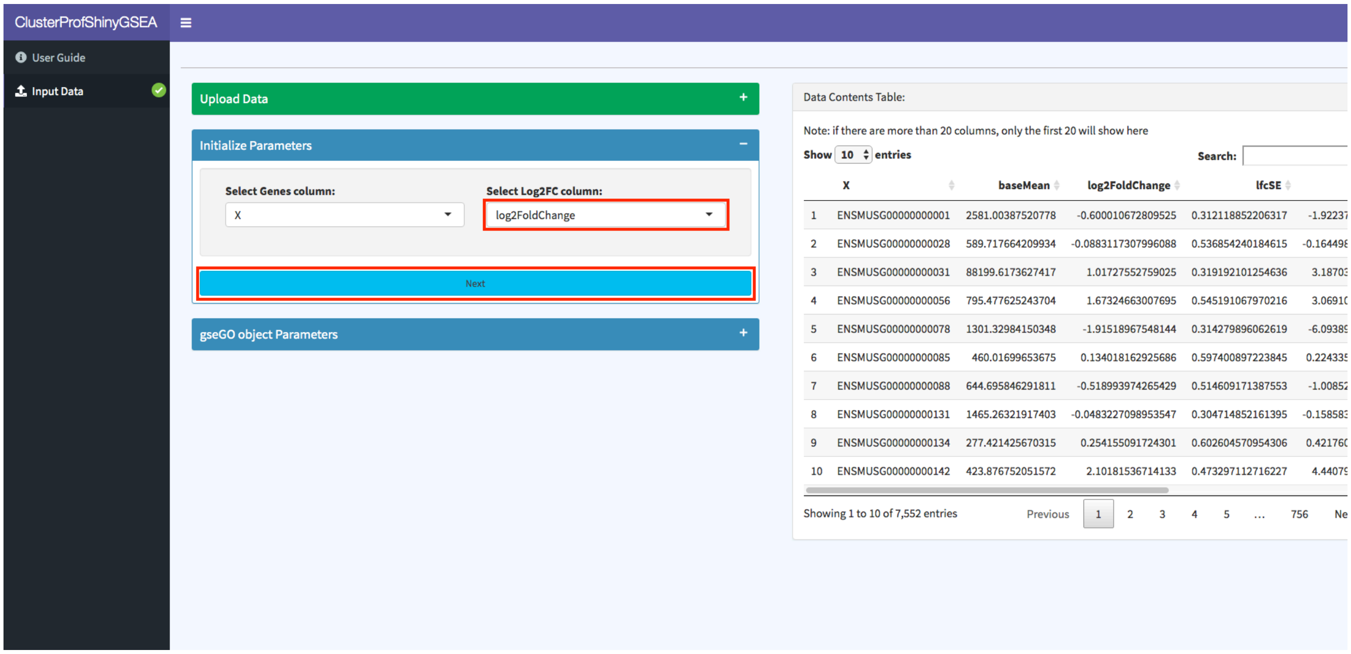
Browse... No file selected

Note: if there are more than 20 columns,

Please select a file

c) Select the DE results file downloaded in the previous section (Section 2, Step f) from Deseq2Shiny (**EPI_vs_ENDO.csv**) to upload.

d) In the tab "Initialize Parameters", make sure to select the correct column name that corresponds to the LogFC column and click "**Next**"



Select Genes column: X

Select Log2FC column: log2FoldChange

Data Contents Table:

Note: if there are more than 20 columns, only the first 20 will show here

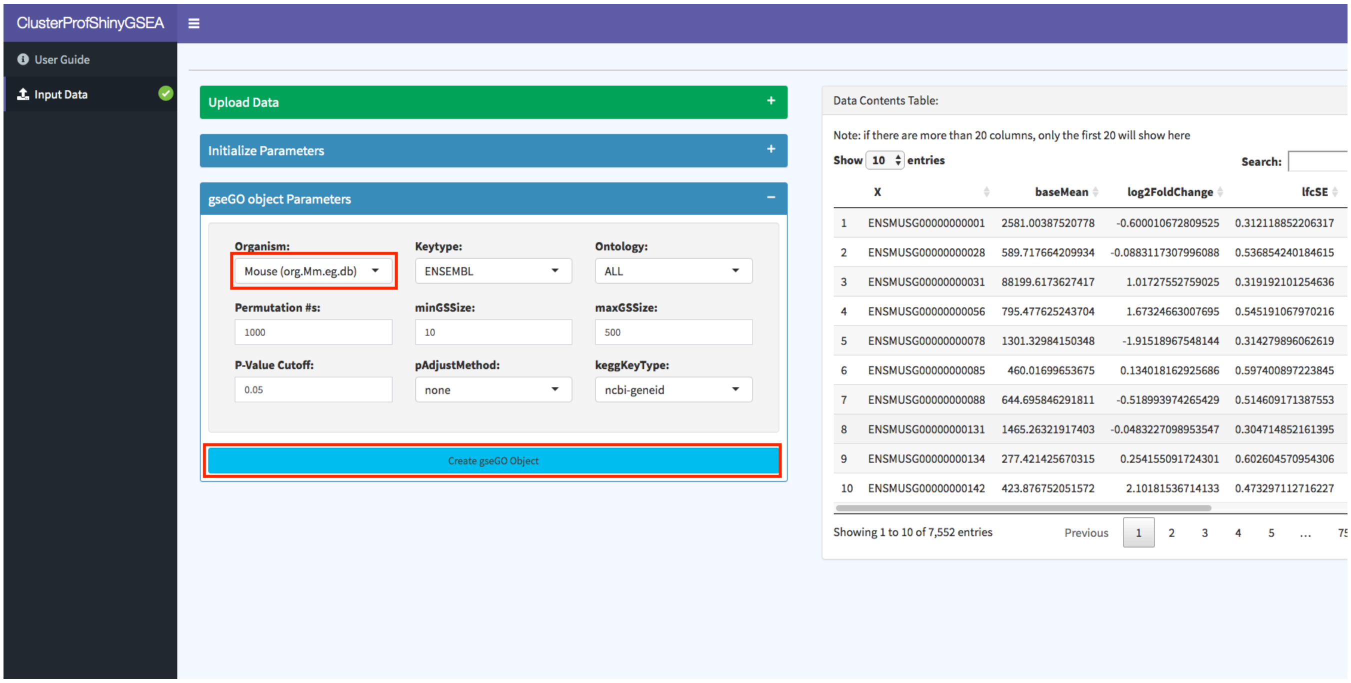
Show 10 entries

Search:

	X	baseMean	log2FoldChange	lfcSE	
1	ENSMUSG00000000001	2581.00387520778	-0.600010672809525	0.312118852206317	-1.92237
2	ENSMUSG00000000028	589.717664209934	-0.0883117307996088	0.536854240184615	-0.164498
3	ENSMUSG00000000031	88199.6173627417	1.01727552759025	0.319192101254636	3.18703
4	ENSMUSG00000000056	795.477625243704	1.67324663007695	0.545191067970216	3.06910
5	ENSMUSG00000000078	1301.32984150348	-1.91518967548144	0.314279896062619	-6.09389
6	ENSMUSG00000000085	460.01699653675	0.134018162925686	0.597400897223845	0.224335
7	ENSMUSG00000000088	644.695846291811	-0.518993974265429	0.514609171387553	-1.00852
8	ENSMUSG00000000131	1465.26321917403	-0.0483227098953547	0.304714852161395	-0.158583
9	ENSMUSG00000000134	277.421425670315	0.254155091724301	0.602604570954306	0.421760
10	ENSMUSG00000000142	423.876752051572	2.10181536714133	0.473297112716227	4.44079

Showing 1 to 10 of 7,552 entries

e) Select "**Mouse (org.Mm.eg.db)**" as the organism and click "**Create gseGO Object**" to start the analysis



Organism: Mouse (org.Mm.eg.db)

Keytype: ENSEMBL

Ontology: ALL

Permutation #s: 1000

minGSSize: 10

maxGSSize: 500

P-Value Cutoff: 0.05

pAdjustMethod: none

keggKeyType: ncbi-geneid

Data Contents Table:

Note: if there are more than 20 columns, only the first 20 will show here

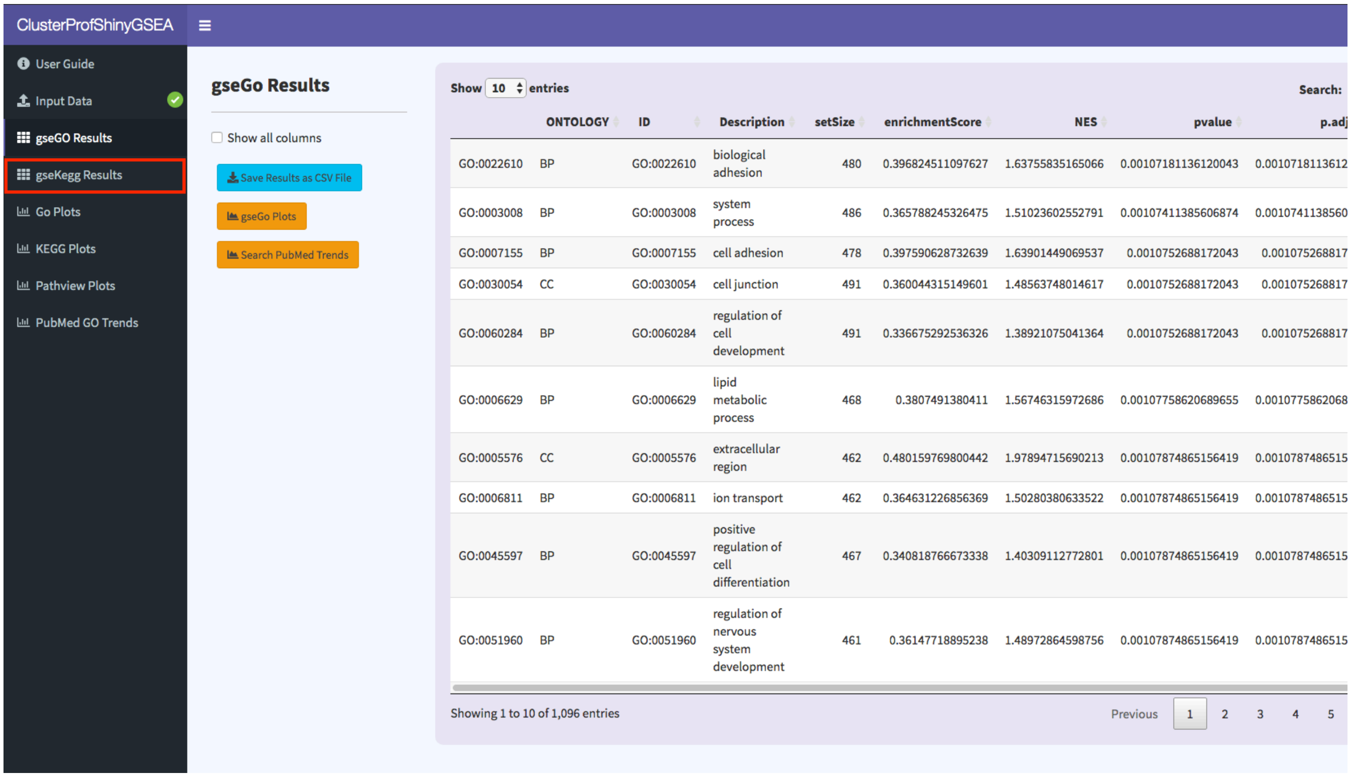
Show 10 entries

Search:

	X	baseMean	log2FoldChange	lfcSE
1	ENSMUSG00000000001	2581.00387520778	-0.600010672809525	0.312118852206317
2	ENSMUSG00000000028	589.717664209934	-0.0883117307996088	0.536854240184615
3	ENSMUSG00000000031	88199.6173627417	1.01727552759025	0.319192101254636
4	ENSMUSG00000000056	795.477625243704	1.67324663007695	0.545191067970216
5	ENSMUSG00000000078	1301.32984150348	-1.91518967548144	0.314279896062619
6	ENSMUSG00000000085	460.01699653675	0.134018162925686	0.597400897223845
7	ENSMUSG00000000088	644.695846291811	-0.518993974265429	0.514609171387553
8	ENSMUSG00000000131	1465.26321917403	-0.0483227098953547	0.304714852161395
9	ENSMUSG00000000134	277.421425670315	0.254155091724301	0.602604570954306
10	ENSMUSG00000000142	423.876752051572	2.10181536714133	0.473297112716227

Showing 1 to 10 of 7,552 entries

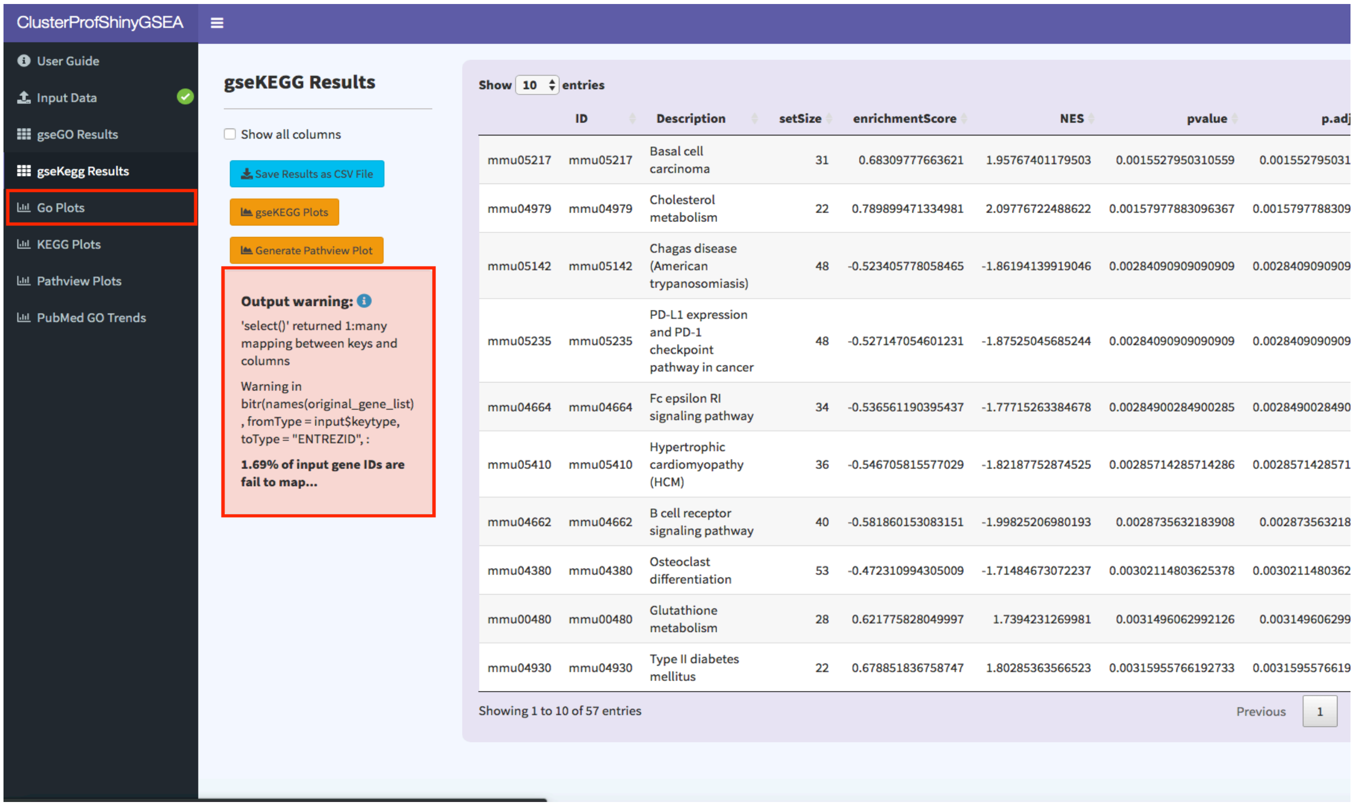
f) Now you can see the gseGO results table. Next go to “**gseKegg Results**” tab to view gseKEGG results table.



	ONTOLOGY	ID	Description	setSize	enrichmentScore	NES	pvalue	p.adj
GO:0022610	BP	GO:0022610	biological adhesion	480	0.396824511097627	1.63755835165066	0.00107181136120043	0.0010718113612
GO:0003008	BP	GO:0003008	system process	486	0.365788245326475	1.51023602552791	0.00107411385606874	0.0010741138560
GO:0007155	BP	GO:0007155	cell adhesion	478	0.397590628732639	1.63901449069537	0.0010752688172043	0.001075268817
GO:0030054	CC	GO:0030054	cell junction	491	0.360044315149601	1.48563748014617	0.0010752688172043	0.001075268817
GO:0060284	BP	GO:0060284	regulation of cell development	491	0.336675292536326	1.38921075041364	0.0010752688172043	0.001075268817
GO:0006629	BP	GO:0006629	lipid metabolic process	468	0.3807491380411	1.56746315972686	0.00107758620689655	0.0010775862068
GO:0005576	CC	GO:0005576	extracellular region	462	0.480159769800442	1.97894715690213	0.00107874865156419	0.0010787486515
GO:0006811	BP	GO:0006811	ion transport	462	0.364631226856369	1.50280380633522	0.00107874865156419	0.0010787486515
GO:0045597	BP	GO:0045597	positive regulation of cell differentiation	467	0.340818766673338	1.40309112772801	0.00107874865156419	0.0010787486515
GO:0051960	BP	GO:0051960	regulation of nervous system development	461	0.36147718895238	1.48972864598756	0.00107874865156419	0.0010787486515

Showing 1 to 10 of 1,096 entries

g) Here are the Kegg results with an output that indicates the percentage of genes that were not mapped. Next go to "**GO Plots**" tab



gseKEGG Results

☐ Show all columns

Output warning:

'select()' returned 1:many mapping between keys and columns

Warning in bitr(names(original_gene_list), fromType = input$keytype, toType = "ENTREZID", :

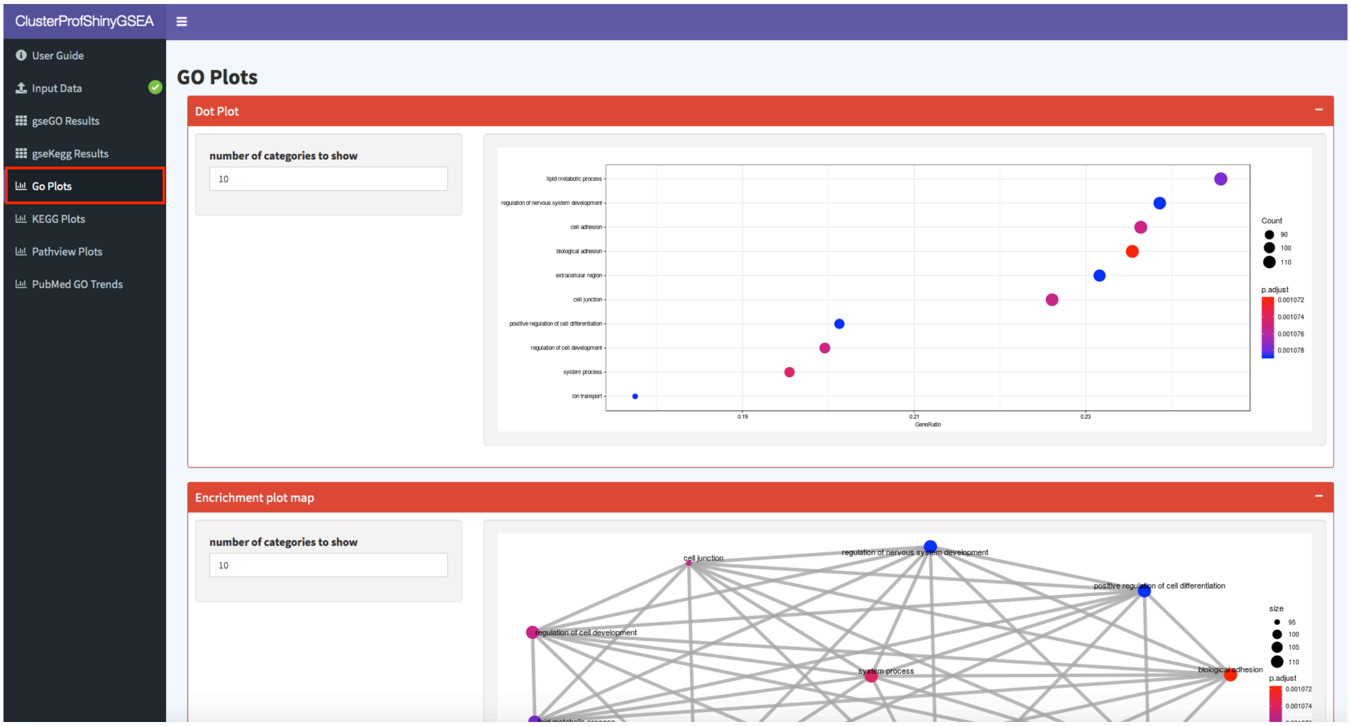
1.69% of input gene IDs are fail to map...

Show 10 entries

	ID	Description	setSize	enrichmentScore	NES	pvalue	p.adj
mmu05217	mmu05217	Basal cell carcinoma	31	0.68309777663621	1.95767401179503	0.0015527950310559	0.001552795031
mmu04979	mmu04979	Cholesterol metabolism	22	0.789899471334981	2.09776722488622	0.00157977883096367	0.0015797788309
mmu05142	mmu05142	Chagas disease (American trypanosomiasis)	48	-0.523405778058465	-1.86194139919046	0.00284090909090909	0.0028409090909
mmu05235	mmu05235	PD-L1 expression and PD-1 checkpoint pathway in cancer	48	-0.527147054601231	-1.87525045685244	0.00284090909090909	0.0028409090909
mmu04664	mmu04664	Fc epsilon RI signaling pathway	34	-0.536561190395437	-1.77715263384678	0.00284900284900285	0.0028490028490
mmu05410	mmu05410	Hypertrophic cardiomyopathy (HCM)	36	-0.546705815577029	-1.82187752874525	0.00285714285714286	0.0028571428571
mmu04662	mmu04662	B cell receptor signaling pathway	40	-0.581860153083151	-1.99825206980193	0.0028735632183908	0.002873563218
mmu04380	mmu04380	Osteoclast differentiation	53	-0.472310994305009	-1.71484673072237	0.00302114803625378	0.0030211480362
mmu00480	mmu00480	Glutathione metabolism	28	0.621775828049997	1.7394231269981	0.0031496062992126	0.003149606299
mmu04930	mmu04930	Type II diabetes mellitus	22	0.678851836758747	1.80285363566523	0.00315955766192733	0.0031595576619

Showing 1 to 10 of 57 entries

h) There are several plots to explore (Dot plot, Enrichment map)



GO Plots

Dot Plot

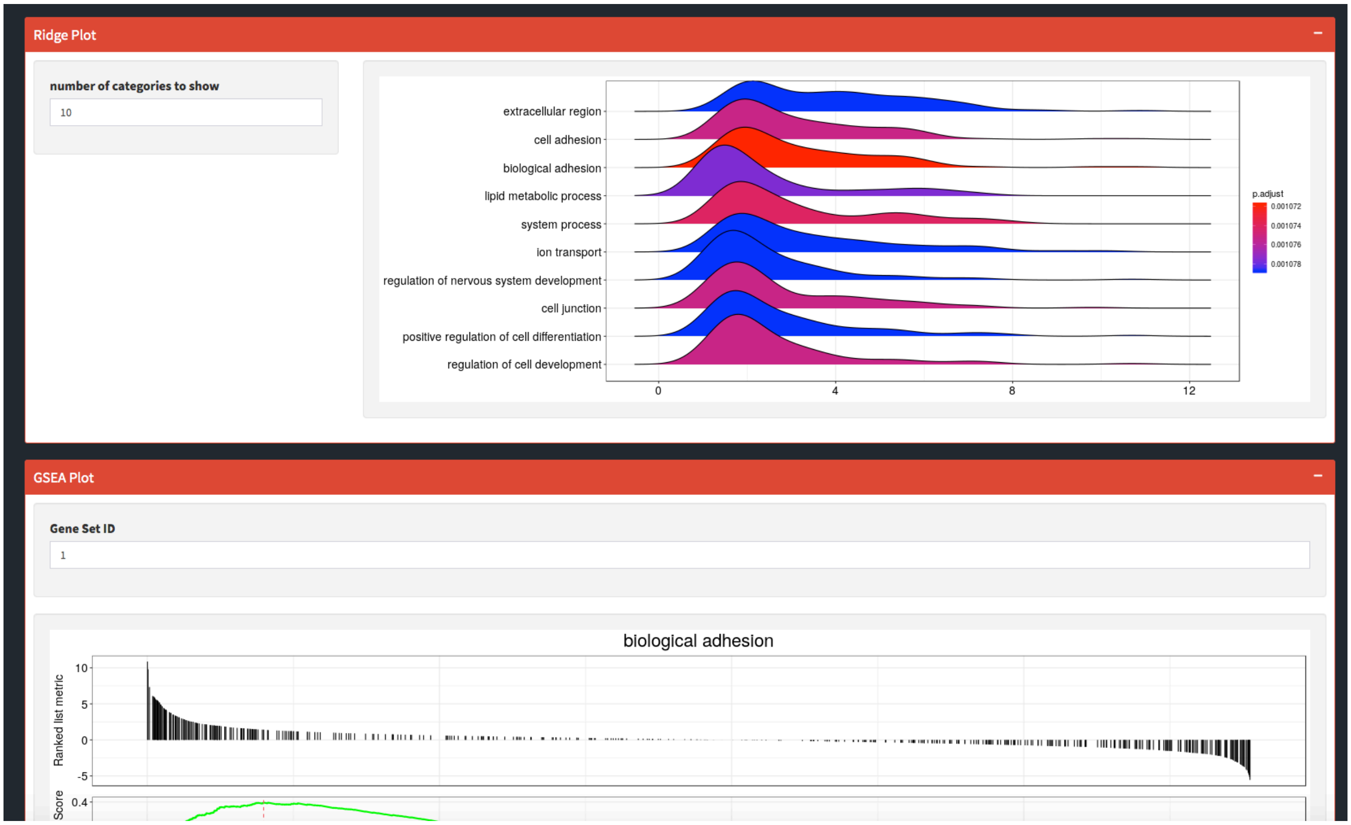
number of categories to show: 10

Enrichment plot map

number of categories to show: 10

cell junction
regulation of nervous system development
positive regulation of cell differentiation
regulation of cell development
system process
biological adhesion
lipid metabolic process
size
95
100
105
110
p.adjust
0.001072
0.001074

i) Ridge plot and GSEA plot



Ridge Plot

number of categories to show

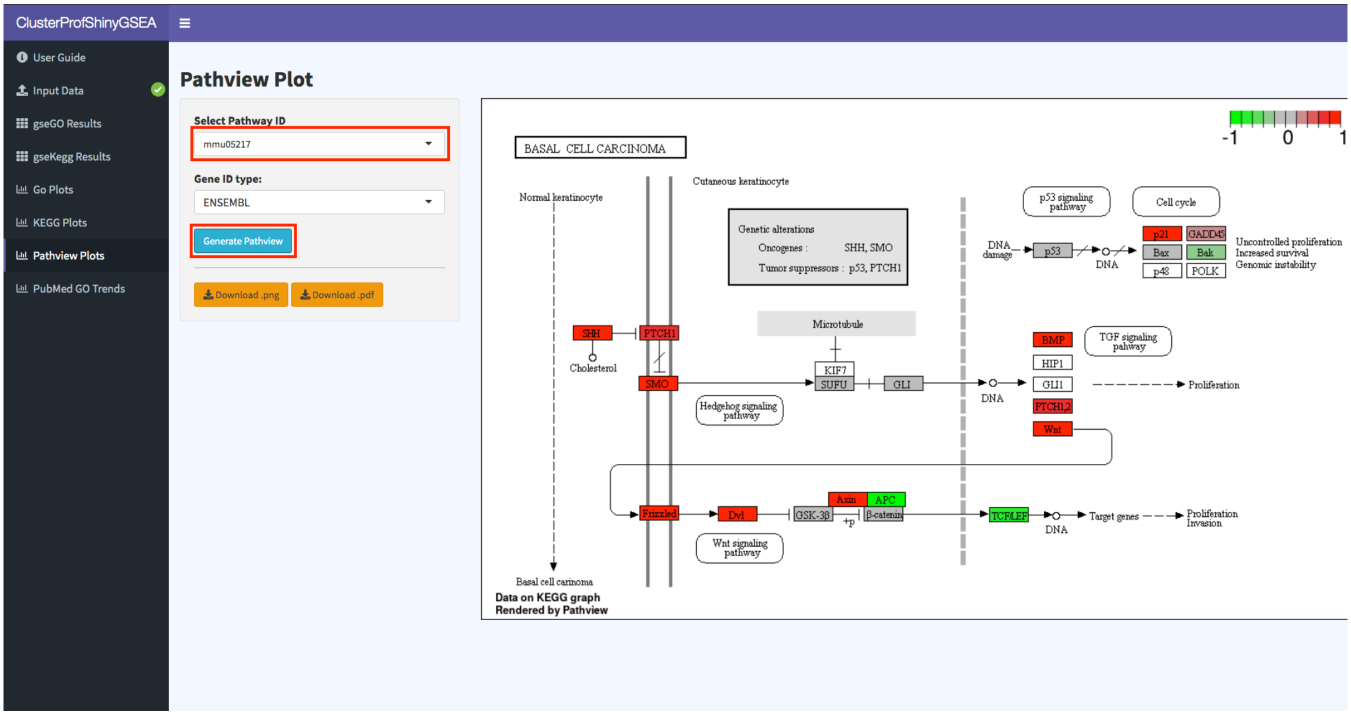
10

GSEA Plot

Gene Set ID

1

j) Next go to "**Pathview Plots**" tab and select gene "**mmu05217**" and click "**Generate Pathview**"



Pathview Plot

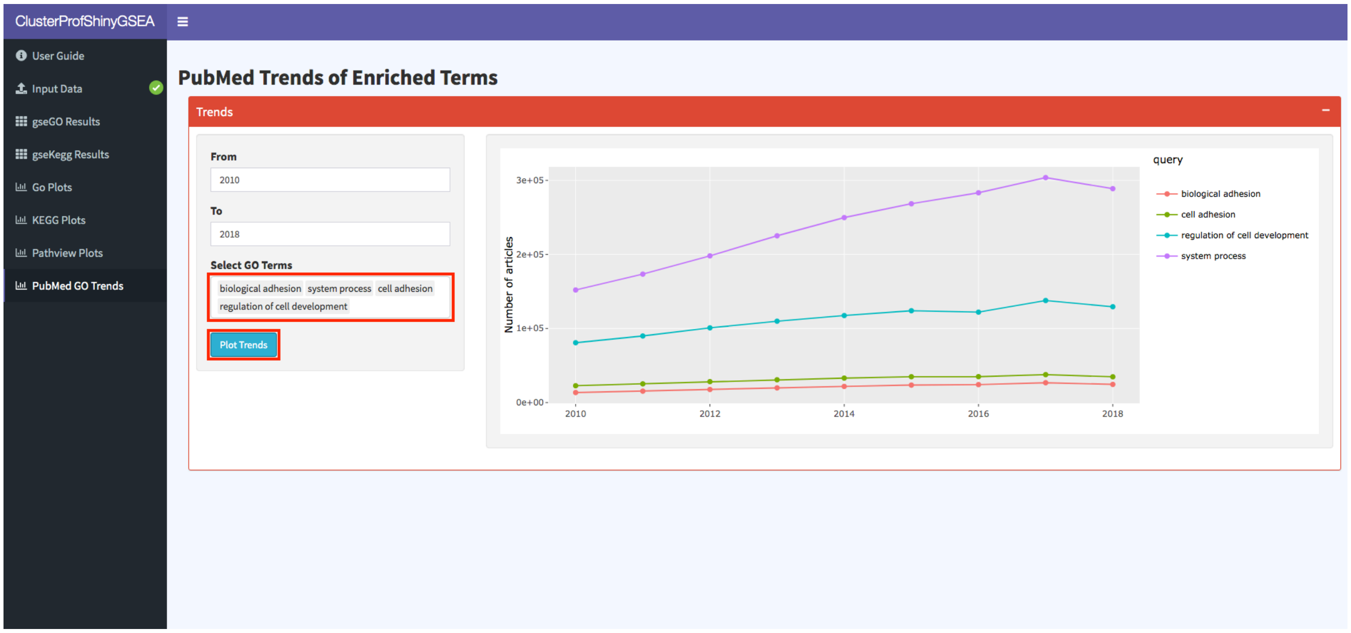
Select Pathway ID

mmu05217

Gene ID type:

ENSEMBL

k) Last step is to check pubmed trends for enriched categories. Select a few GO terms and click “**Plot Trends**”



From

2010

To

2018

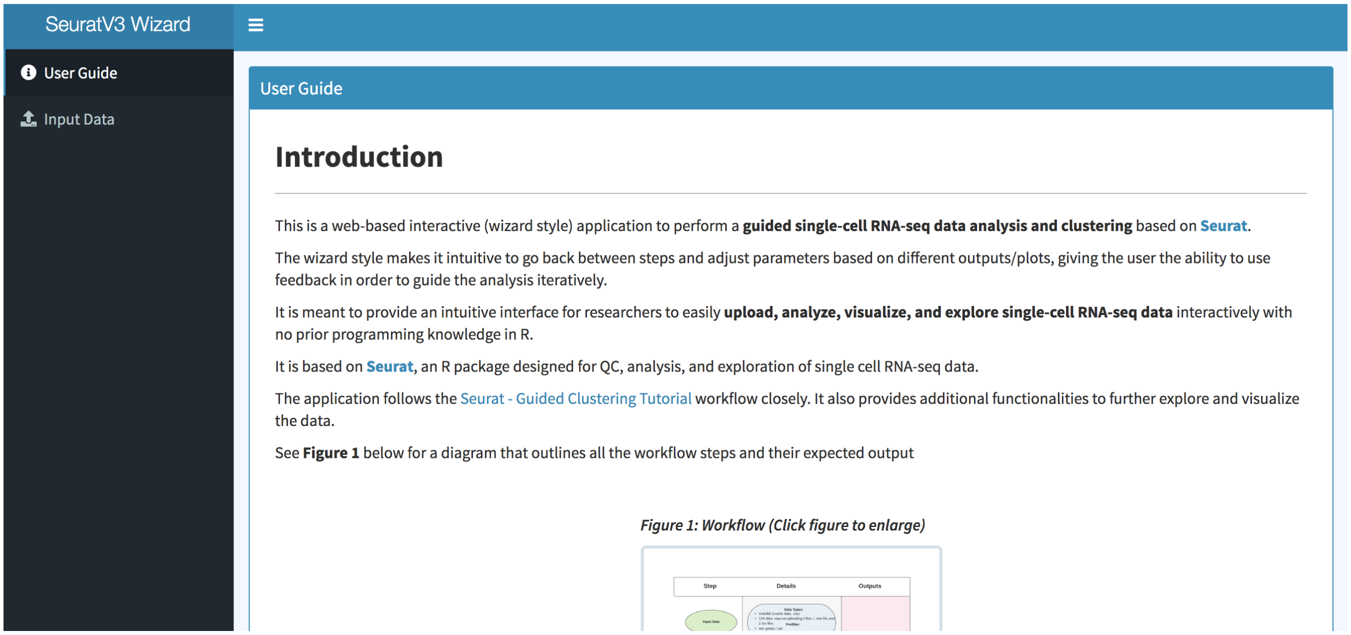
Select GO Terms

biological adhesion system process cell adhesion regulation of cell development

4) Example Use Case 2 (scRNA-seq analysis using Seurat):

In this example we will show how you can easily upload your single-cell RNA seq sample data (10X format, using the 10x genomics cellranger pipeline) and perform a guided single-cell data analysis and visualization using the popular Seurat library. You can follow along using the Seurat tutorial (https://satijalab.org/seurat/v3.0/pbmc3k_tutorial.html), which the SeuratV3Wizard mirrors closely.

a) Download and extract data file from https://s3-us-west-2.amazonaws.com/10x.files/samples/cell/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz
b) Launch SeuratV3Wizard (http://nasqar.abudhabi.nyu.edu/SeuratV3Wizard/)



Introduction

This is a web-based interactive (wizard style) application to perform a **guided single-cell RNA-seq data analysis and clustering** based on **Seurat**.

The wizard style makes it intuitive to go back between steps and adjust parameters based on different outputs/plots, giving the user the ability to use feedback in order to guide the analysis iteratively.

It is meant to provide an intuitive interface for researchers to easily **upload, analyze, visualize, and explore single-cell RNA-seq data** interactively with no prior programming knowledge in R.

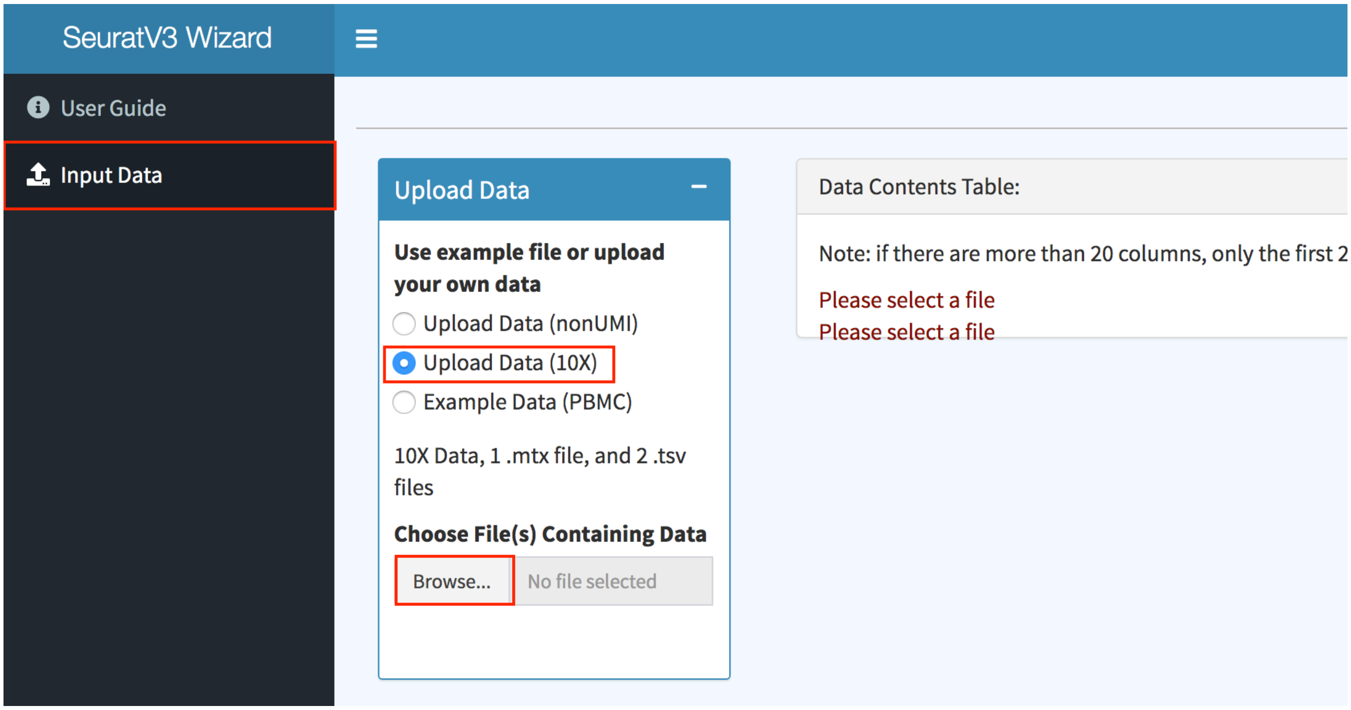
It is based on **Seurat**, an R package designed for QC, analysis, and exploration of single cell RNA-seq data.

The application follows the Seurat - Guided Clustering Tutorial workflow closely. It also provides additional functionalities to further explore and visualize the data.

See **Figure 1** below for a diagram that outlines all the workflow steps and their expected output

Figure 1: Workflow (Click figure to enlarge)

c) Go to “**Input Data**” tab and select “**Upload Data (10X)**”



Upload Data

Use example file or upload your own data

○ Upload Data (nonUMI)
● Upload Data (10X)
○ Example Data (PBMC)

10X Data, 1 .mtx file, and 2 .tsv files

Choose File(s) Containing Data

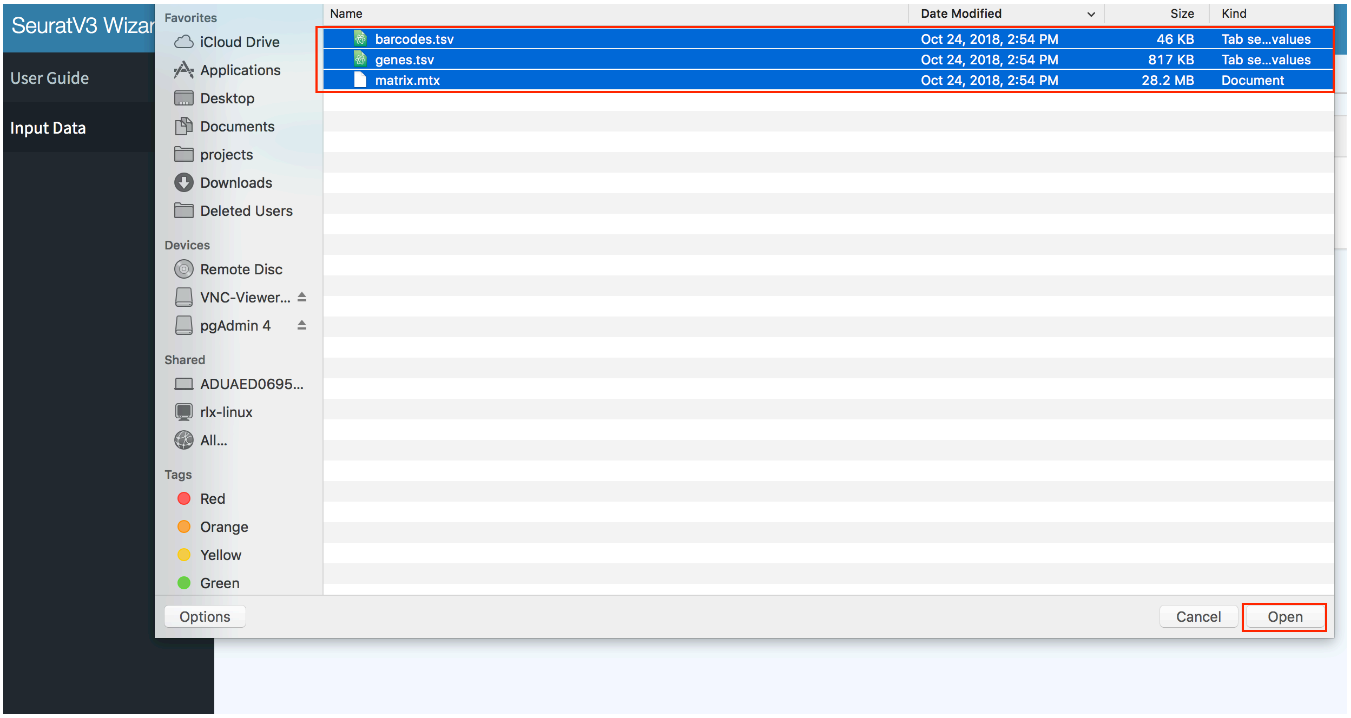
Browse... No file selected

Data Contents Table:

Note: if there are more than 20 columns, only the first 2

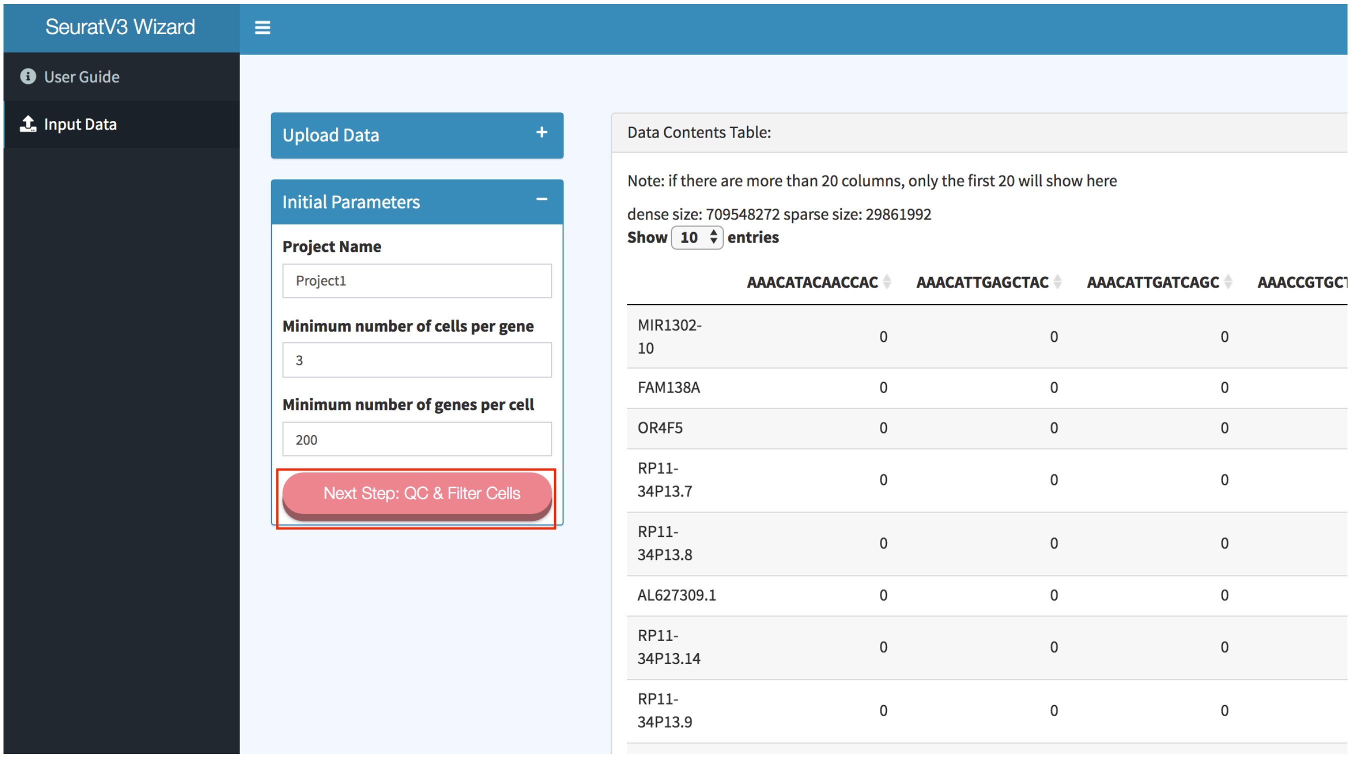
Please select a file
Please select a file

d) Browse for the extracted data files downloaded earlier. Select all 3 files to upload



Name	Date Modified	Size	Kind
barcodes.tsv	Oct 24, 2018, 2:54 PM	46 KB	Tab se...values
genes.tsv	Oct 24, 2018, 2:54 PM	817 KB	Tab se...values
matrix.mtx	Oct 24, 2018, 2:54 PM	28.2 MB	Document

e) Click “**Next Step: QC & Filter Cells**”



Project Name: Project1

Minimum number of cells per gene: 3

Minimum number of genes per cell: 200

Data Contents Table:

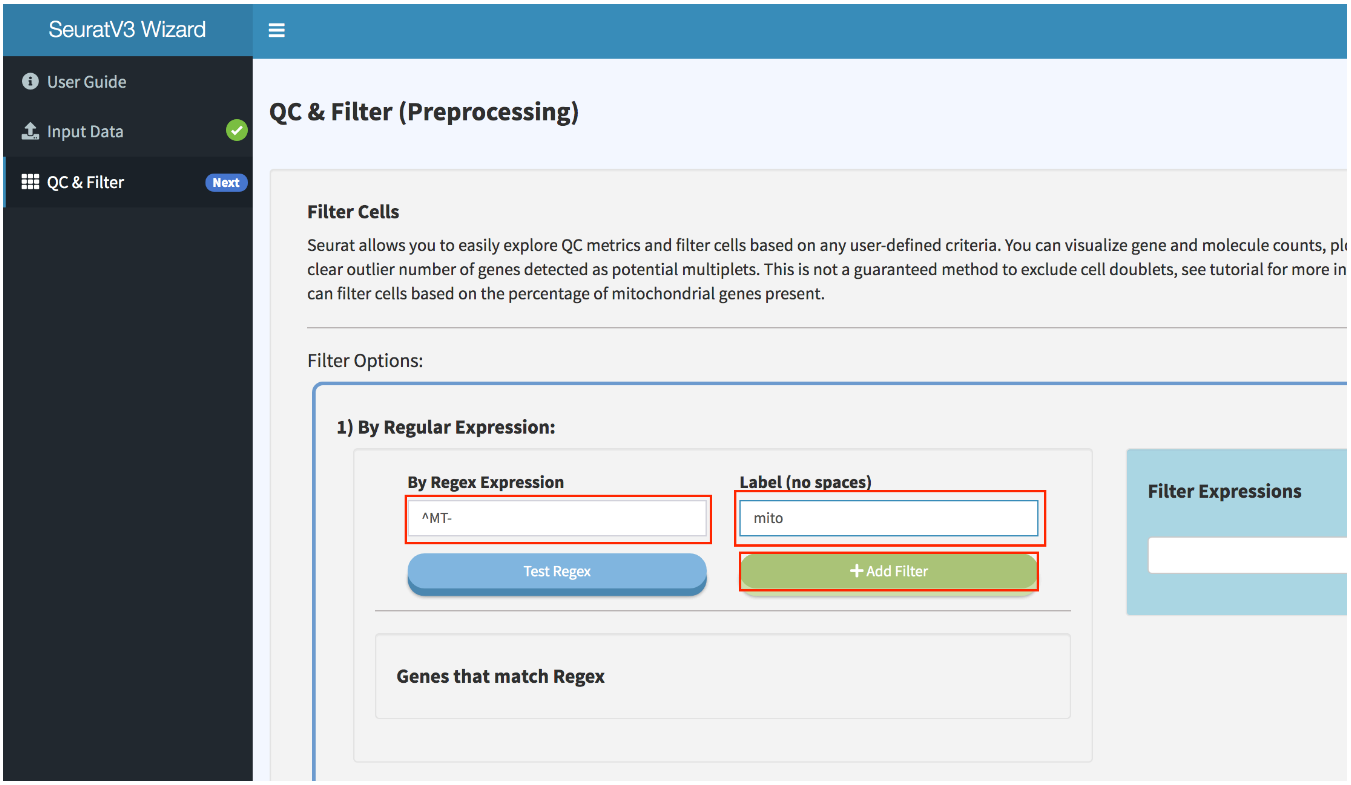
Note: if there are more than 20 columns, only the first 20 will show here

dense size: 709548272 sparse size: 29861992

Show 10 **entries**

	AAACATACAACCAC	AAACATTGAGCTAC	AAACATTGATCAGC	AAACCGTGCT
MIR1302-10	0	0	0	
FAM138A	0	0	0	
OR4F5	0	0	0	
RP11-34P13.7	0	0	0	
RP11-34P13.8	0	0	0	
AL627309.1	0	0	0	
RP11-34P13.14	0	0	0	
RP11-34P13.9	0	0	0	

f) Add a regular expression for mitochondrial genes by adding the **regex** and **label** as seen below, and then click “**Add Filter**”. You can use the “**Test Regex”** tab in order to verify that the regular expression the user supplies works as expected.



QC & Filter (Preprocessing)

Filter Cells

Seurat allows you to easily explore QC metrics and filter cells based on any user-defined criteria. You can visualize gene and molecule counts, pl
clear outlier number of genes detected as potential multiplets. This is not a guaranteed method to exclude cell doublets, see tutorial for more in
can filter cells based on the percentage of mitochondrial genes present.

Filter Options:

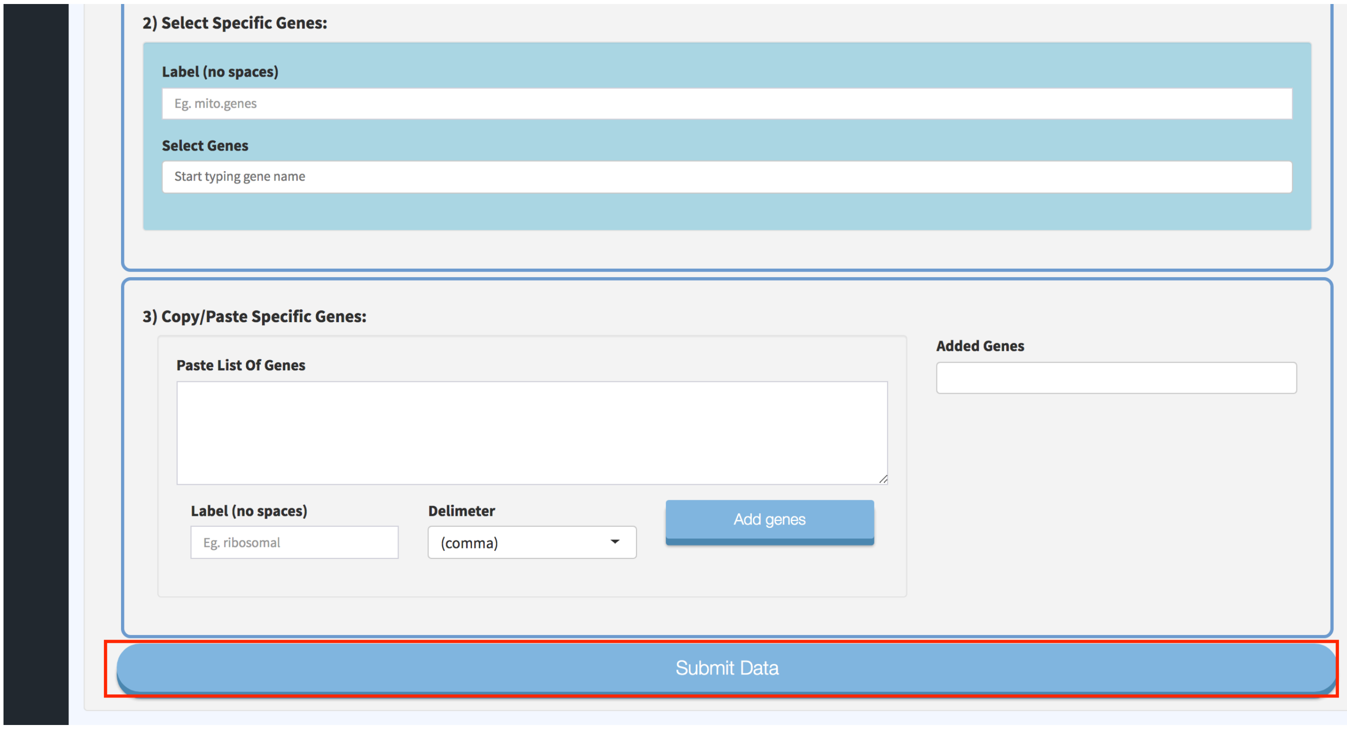
1) By Regular Expression:

By Regex Expression: ^MT-

Label (no spaces): mito

Genes that match Regex

g) Scroll down and click “**Submit Data**”



2) Select Specific Genes:

Label (no spaces)

Eg. mito.genes

Select Genes

Start typing gene name

3) Copy/Paste Specific Genes:

Paste List Of Genes

Added Genes

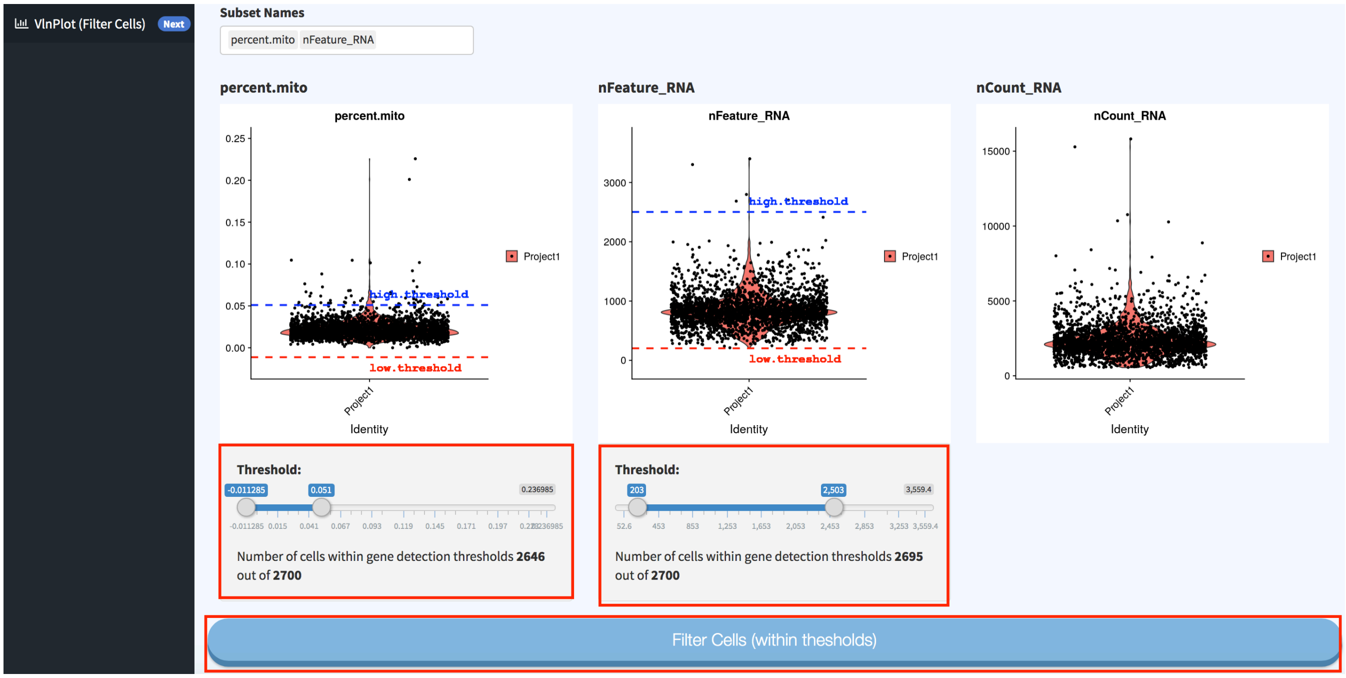
Label (no spaces)

Eg. ribosomal

Delimeter

(comma)

h) Select the low and high thresholds to filter out cells. Click “**Filter Cells (within thresholds)**”. This is an interactive filter that will display the effects of applying different filtering thresholds on the data.



Subset Names

percent.mito nFeature_RNA

Threshold:

-0.011285 0.051 0.236985

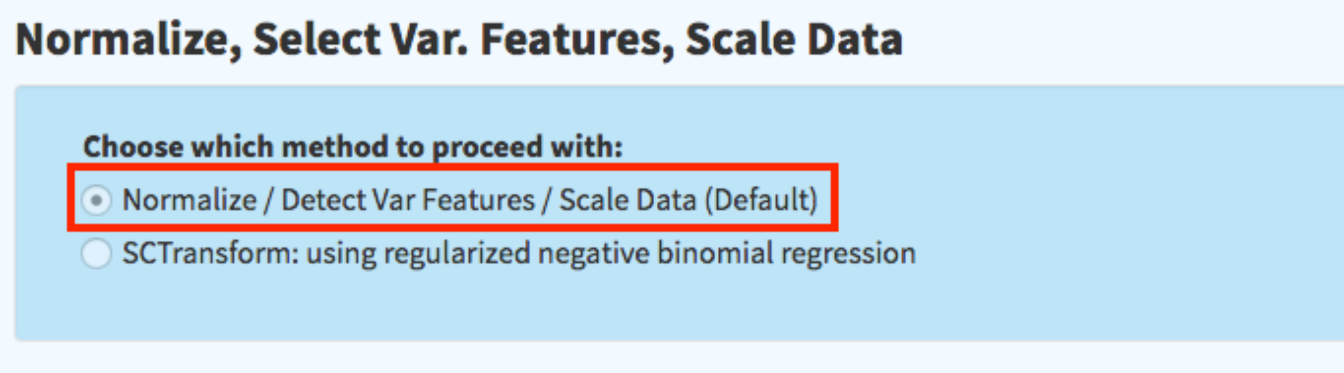
Number of cells within gene detection thresholds **2646** out of **2700**

Threshold:

203 2,503 3,559.4

Number of cells within gene detection thresholds **2695** out of **2700**

i) There are now two options to proceed within the analysis. We will choose the default one so we can mirror the tutorial. The second option is the SCTransform method (https://rawgit.com/ChristophH/sctransform/master/inst/doc/seurat.html)

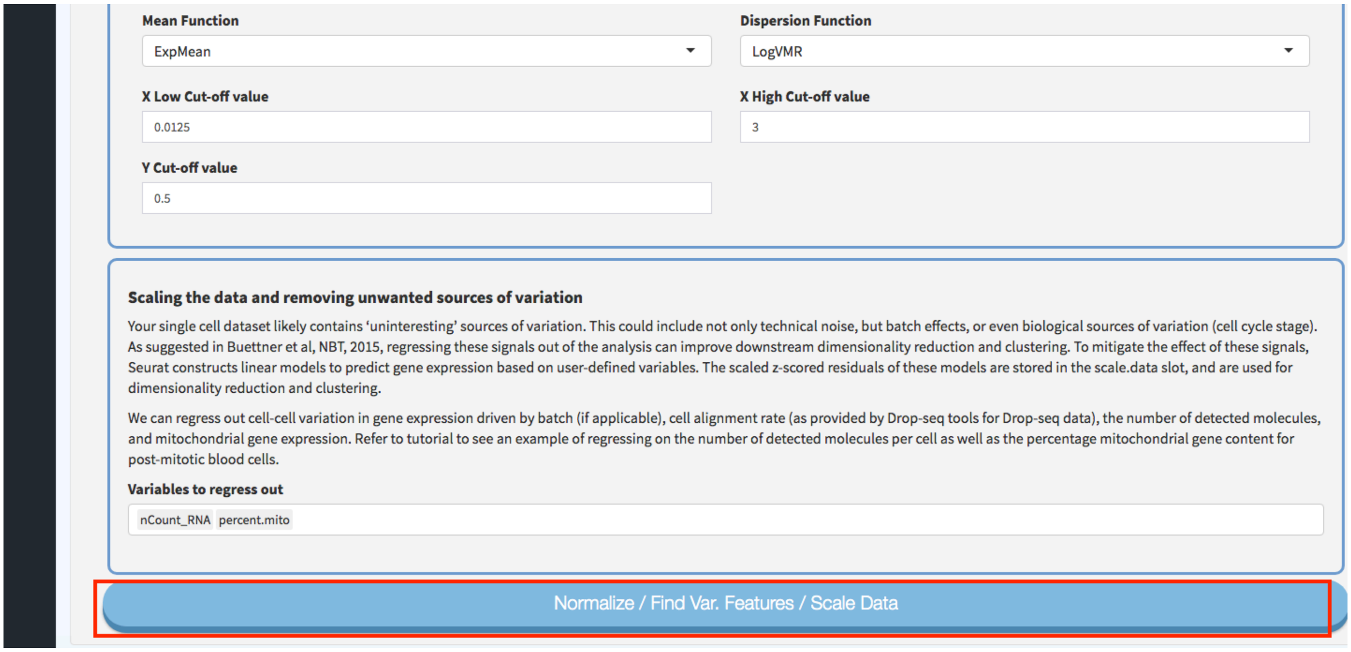


Normalize, Select Var. Features, Scale Data

Choose which method to proceed with:

- ● Normalize / Detect Var Features / Scale Data (Default)
- ○ SCTransform: using regularized negative binomial regression

j) Proceed by clicking “**Normalize / Find Var. Feaatures / Scala Data**”



Mean Function: ExpMean

Dispersion Function: LogVMR

X Low Cut-off value: 0.0125

X High Cut-off value: 3

Y Cut-off value: 0.5

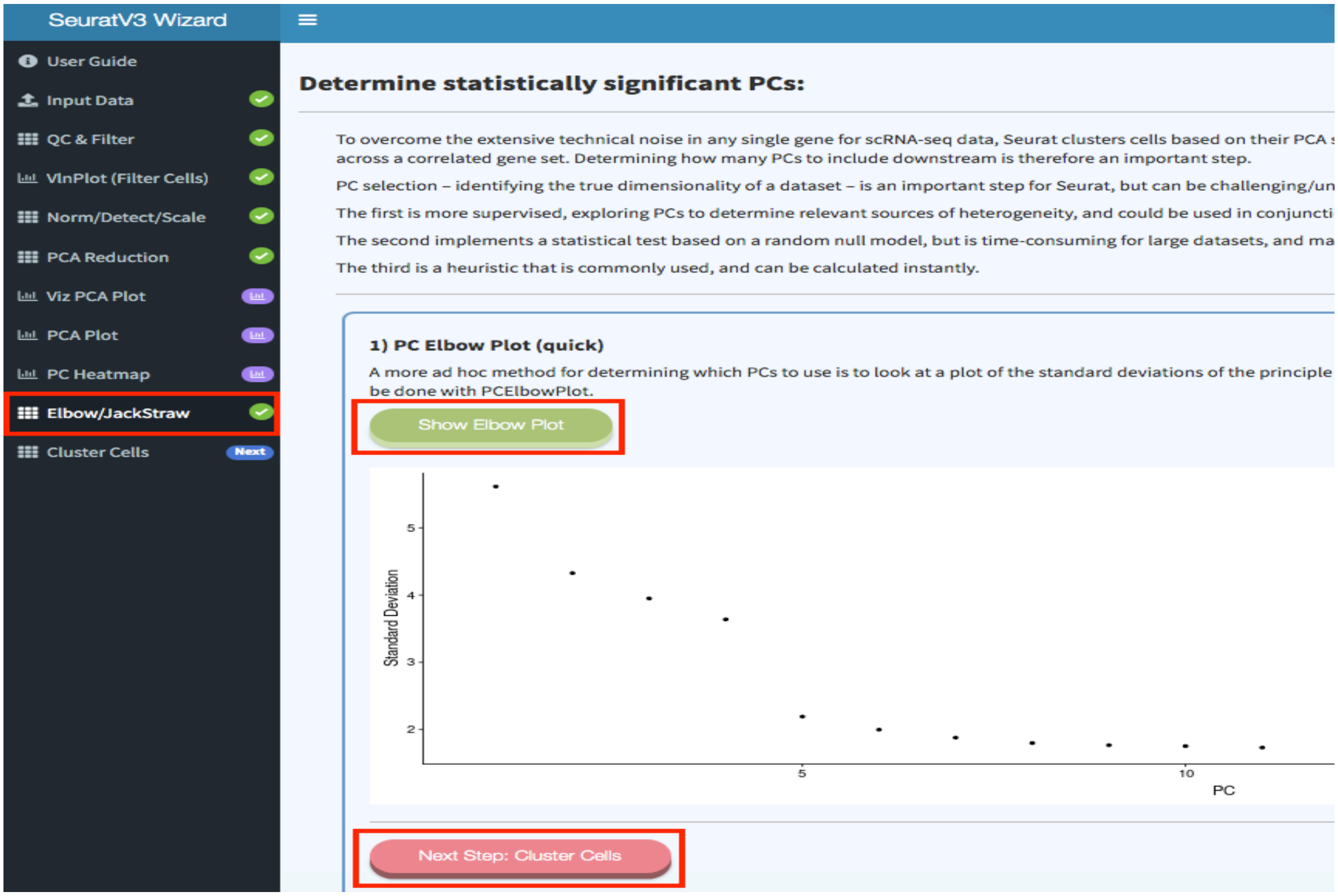
Scaling the data and removing unwanted sources of variation

Your single cell dataset likely contains ‘uninteresting’ sources of variation. This could include not only technical noise, but batch effects, or even biological sources of variation (cell cycle stage). As suggested in Buettner et al, NBT, 2015, regressing these signals out of the analysis can improve downstream dimensionality reduction and clustering. To mitigate the effect of these signals, Seurat constructs linear models to predict gene expression based on user-defined variables. The scaled z-scored residuals of these models are stored in the scale.data slot, and are used for dimensionality reduction and clustering.

We can regress out cell-cell variation in gene expression driven by batch (if applicable), cell alignment rate (as provided by Drop-seq tools for Drop-seq data), the number of detected molecules, and mitochondrial gene expression. Refer to tutorial to see an example of regressing on the number of detected molecules per cell as well as the percentage mitochondrial gene content for post-mitotic blood cells.

Variables to regress out: nCount_RNA percent.mito

k) Now you can follow the default settings in the wizard to go through the steps until "**Elbow/Jackstraw**". Click "**Show Elbow Plot**" and click "**Next Step: Cluster Cells**"



Determine statistically significant PCs:

To overcome the extensive technical noise in any single gene for scRNA-seq data, Seurat clusters cells based on their PCA s across a correlated gene set. Determining how many PCs to include downstream is therefore an important step.

PC selection – identifying the true dimensionality of a dataset – is an important step for Seurat, but can be challenging/un

The first is more supervised, exploring PCs to determine relevant sources of heterogeneity, and could be used in conjuncti

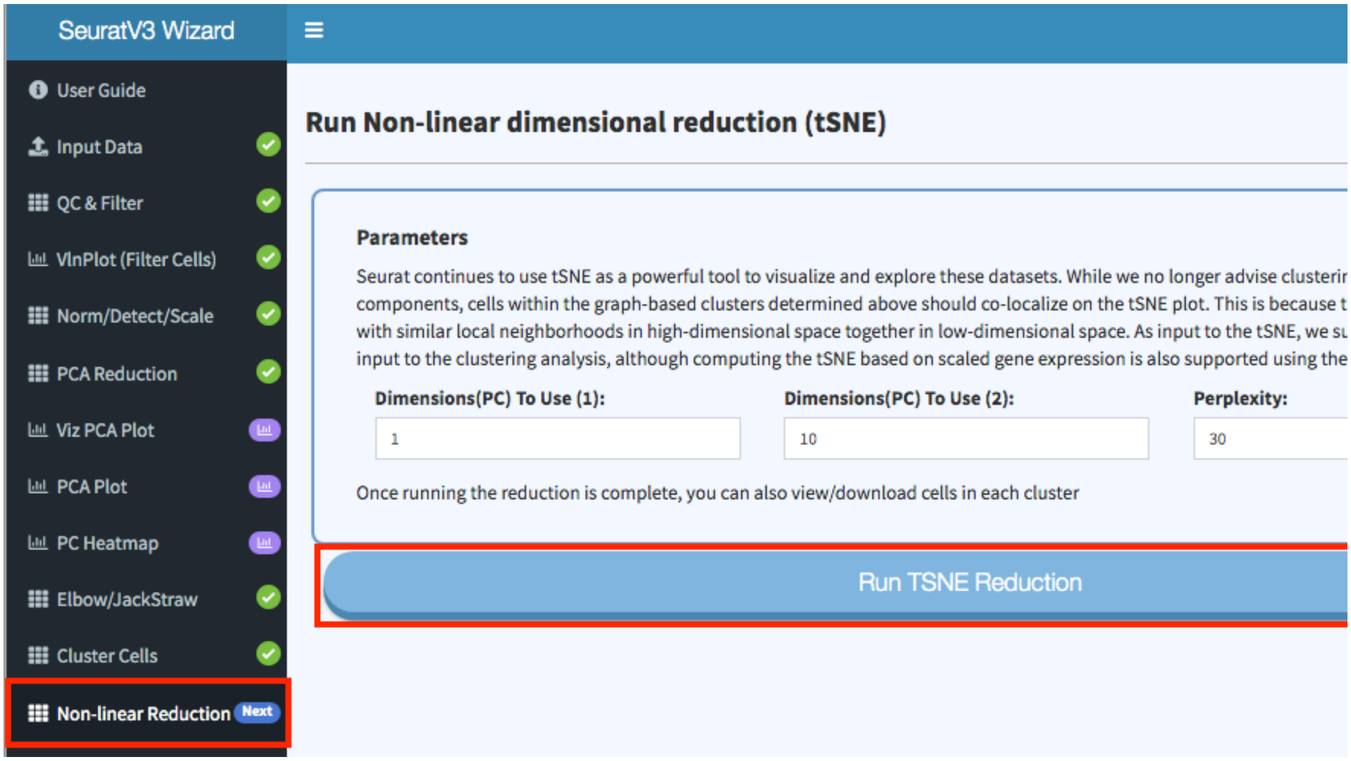
The second implements a statistical test based on a random null model, but is time-consuming for large datasets, and ma

The third is a heuristic that is commonly used, and can be calculated instantly.

1) PC Elbow Plot (quick)

A more ad hoc method for determining which PCs to use is to look at a plot of the standard deviations of the principle be done with PCElbowPlot.

l) Next run “**Cluster Cells**” and go to “**Non-linear Reduction**” tab and “**Run TSNE Reduction**”



Run Non-linear dimensional reduction (tSNE)

Parameters

Seurat continues to use tSNE as a powerful tool to visualize and explore these datasets. While we no longer advise clusterir
components, cells within the graph-based clusters determined above should co-localize on the tSNE plot. This is because t
with similar local neighborhoods in high-dimensional space together in low-dimensional space. As input to the tSNE, we su
input to the clustering analysis, although computing the tSNE based on scaled gene expression is also supported using the

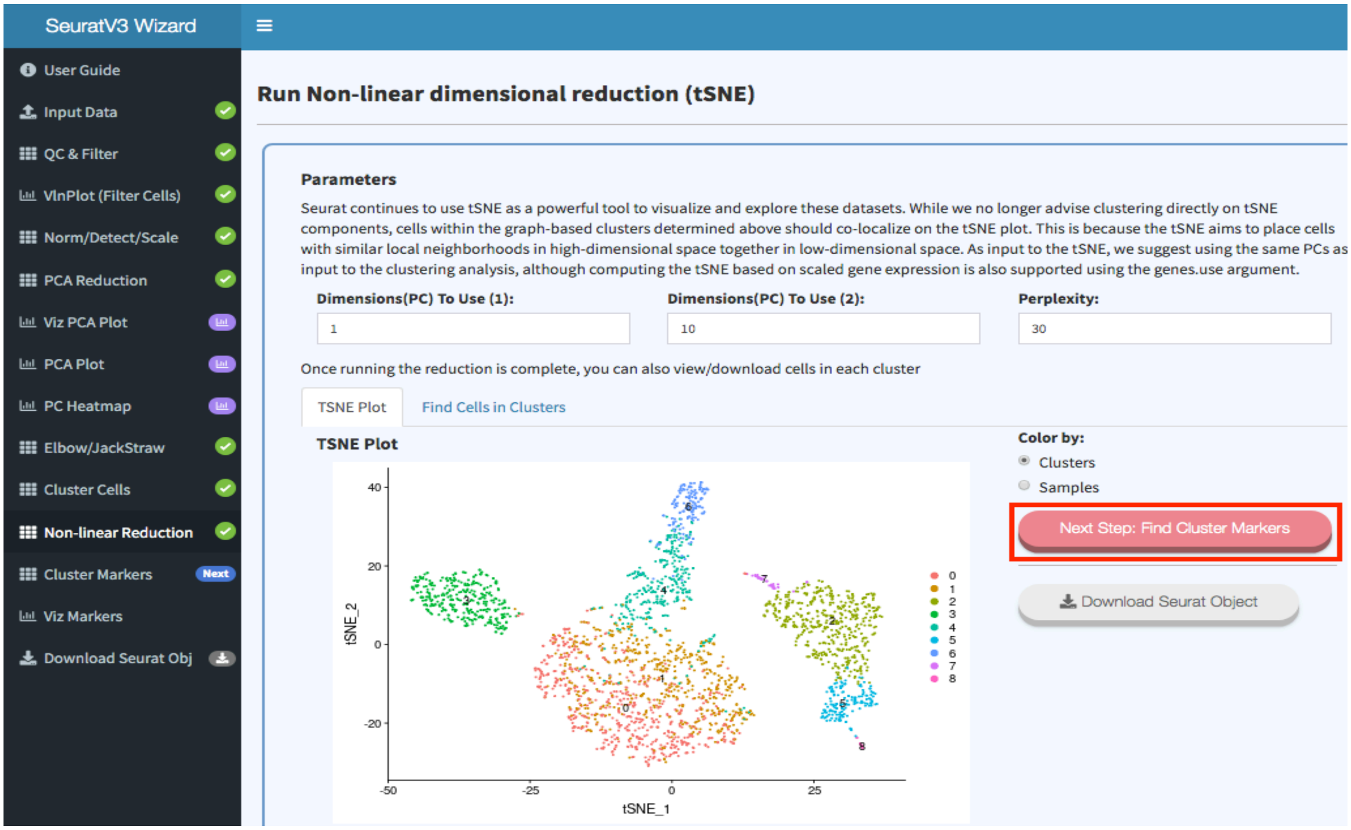
Dimensions(PC) To Use (1): 1

Dimensions(PC) To Use (2): 10

Perplexity: 30

Once running the reduction is complete, you can also view/download cells in each cluster

m) Once tSNE step is done, we can now go to the next step “**Next Step: Find Cluster Markers**” and click “**Find Cluster Markers**”



Run Non-linear dimensional reduction (tSNE)

Parameters

Seurat continues to use tSNE as a powerful tool to visualize and explore these datasets. While we no longer advise clustering directly on tSNE components, cells within the graph-based clusters determined above should co-localize on the tSNE plot. This is because the tSNE aims to place cells with similar local neighborhoods in high-dimensional space together in low-dimensional space. As input to the tSNE, we suggest using the same PCs as input to the clustering analysis, although computing the tSNE based on scaled gene expression is also supported using the genes.use argument.

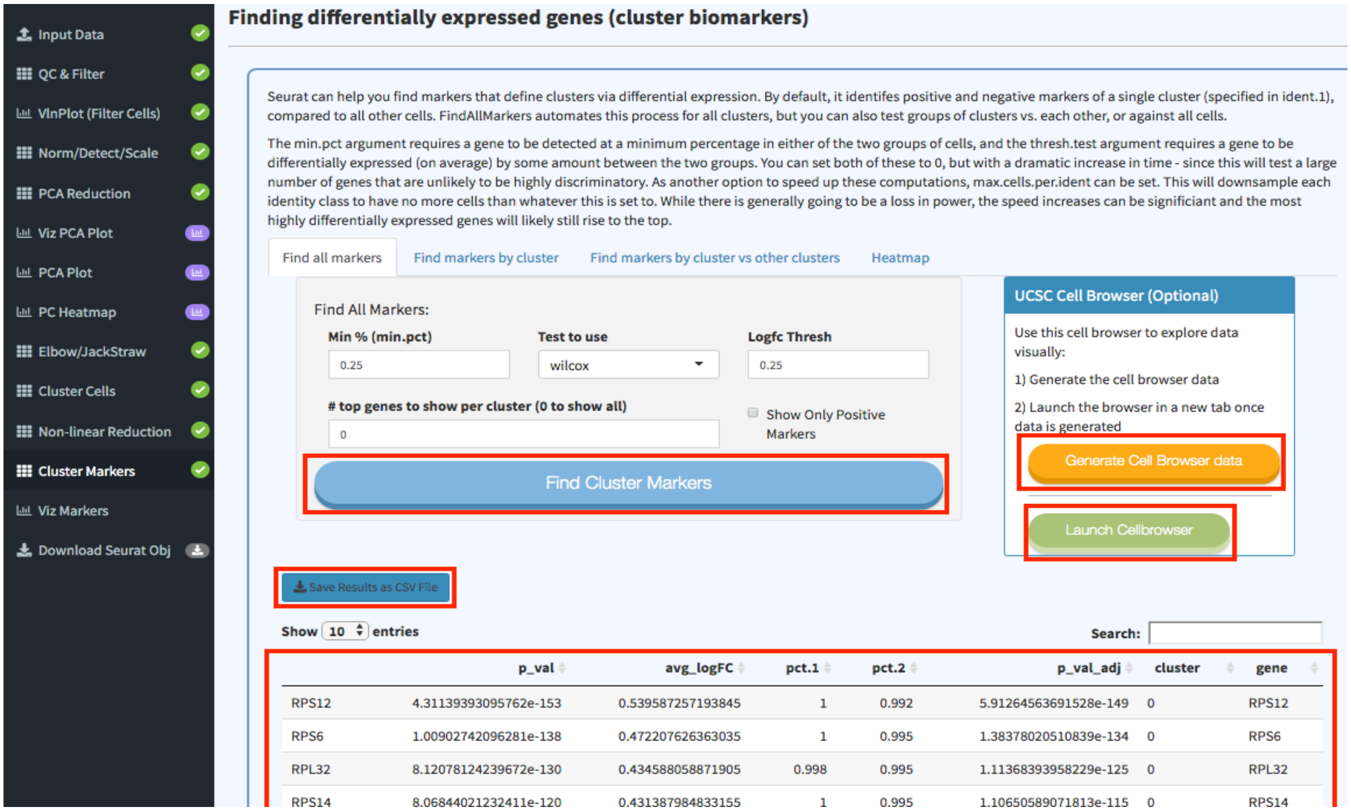
Dimensions(PC) To Use (1): 1

Dimensions(PC) To Use (2): 10

Perplexity: 30

Once running the reduction is complete, you can also view/download cells in each cluster

n) Now you can see the markers table (which is downloadable). There is also an option to explore the clusters and markers visually using UCSC Cellbrowser (https://github.com/maximilianh/cellBrowser). Click “**Generate Cell browser data**”. Once done you can launch the cellbrowser by clicking on “**Launch Cellbrowser**”.



Finding differentially expressed genes (cluster biomarkers)

Seurat can help you find markers that define clusters via differential expression. By default, it identifes positive and negative markers of a single cluster (specified in ident.1), compared to all other cells. FindAllMarkers automates this process for all clusters, but you can also test groups of clusters vs. each other, or against all cells.

The min.pct argument requires a gene to be detected at a minimum percentage in either of the two groups of cells, and the thresh.test argument requires a gene to be differentially expressed (on average) by some amount between the two groups. You can set both of these to 0, but with a dramatic increase in time - since this will test a large number of genes that are unlikely to be highly discriminatory. As another option to speed up these computations, max.cells.per.ident can be set. This will downsample each identity class to have no more cells than whatever this is set to. While there is generally going to be a loss in power, the speed increases can be significiant and the most highly differentially expressed genes will likely still rise to the top.

Find All Markers:

Min % (min.pct): 0.25

Test to use: wilcox

Logfc Thresh: 0.25

top genes to show per cluster (0 to show all): 0

□ Show Only Positive Markers

UCSC Cell Browser (Optional)

Use this cell browser to explore data visually:

1) Generate the cell browser data

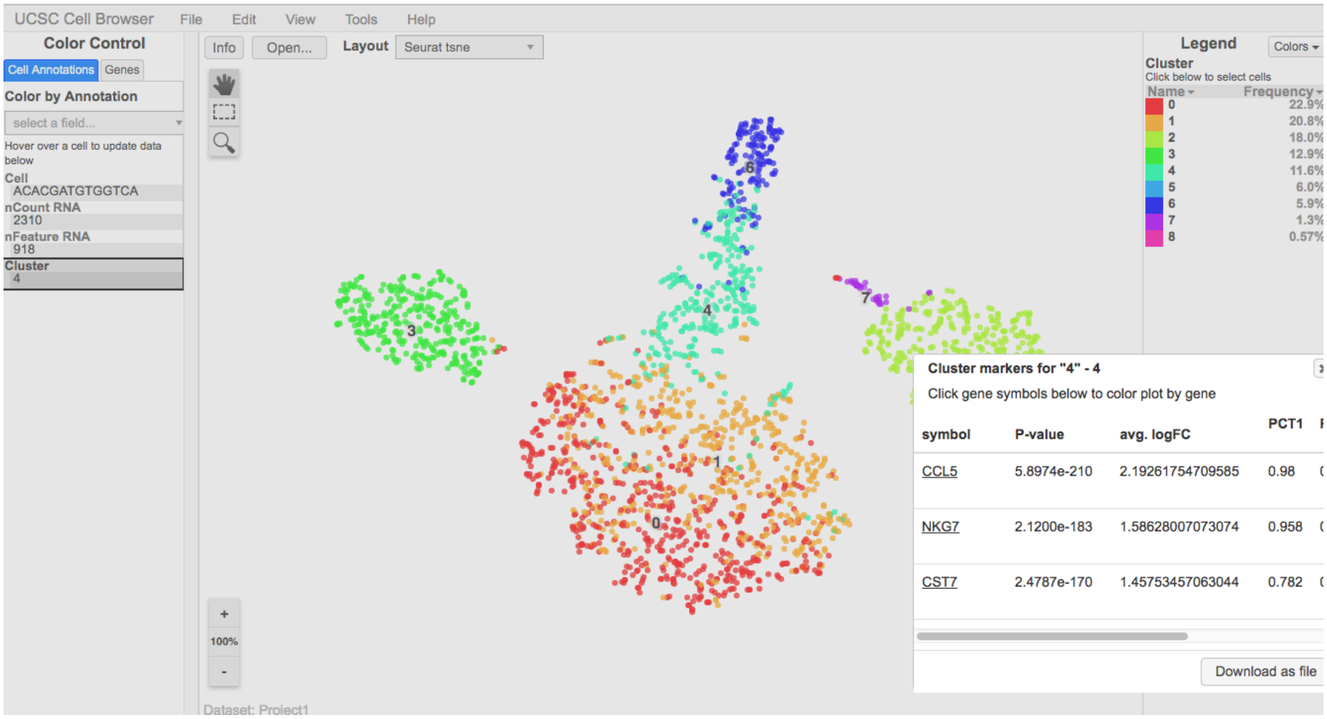
2) Launch the browser in a new tab once data is generated

Show 10 entries

Search:

	p_val	avg_logFC	pct.1	pct.2	p_val_adj	cluster	gene
RPS12	4.31139393095762e-153	0.539587257193845	1	0.992	5.91264563691528e-149	0	RPS12
RPS6	1.00902742096281e-138	0.472207626363035	1	0.995	1.38378020510839e-134	0	RPS6
RPL32	8.12078124239672e-130	0.434588058871905	0.998	0.995	1.11368393958229e-125	0	RPL32
RPS14	8.06844021232411e-120	0.431387984833155	1	0.995	1.10650589071813e-115	0	RPS14

o) A tab is opened with UCSC cellbrowser



Hover over a cell to update data below

Cell
ACACGATGTGGTCA

nCount RNA
2310

nFeature RNA
918

Cluster
4

Cluster

Click below to select cells

Name	Frequency
🟥 0	22.9%
🟧 1	20.8%
🟩 2	18.0%
🟩 3	12.9%
🟩 4	11.6%
🟦 5	6.0%
🟦 6	5.9%
🟪 7	1.3%
🟪 8	0.57%

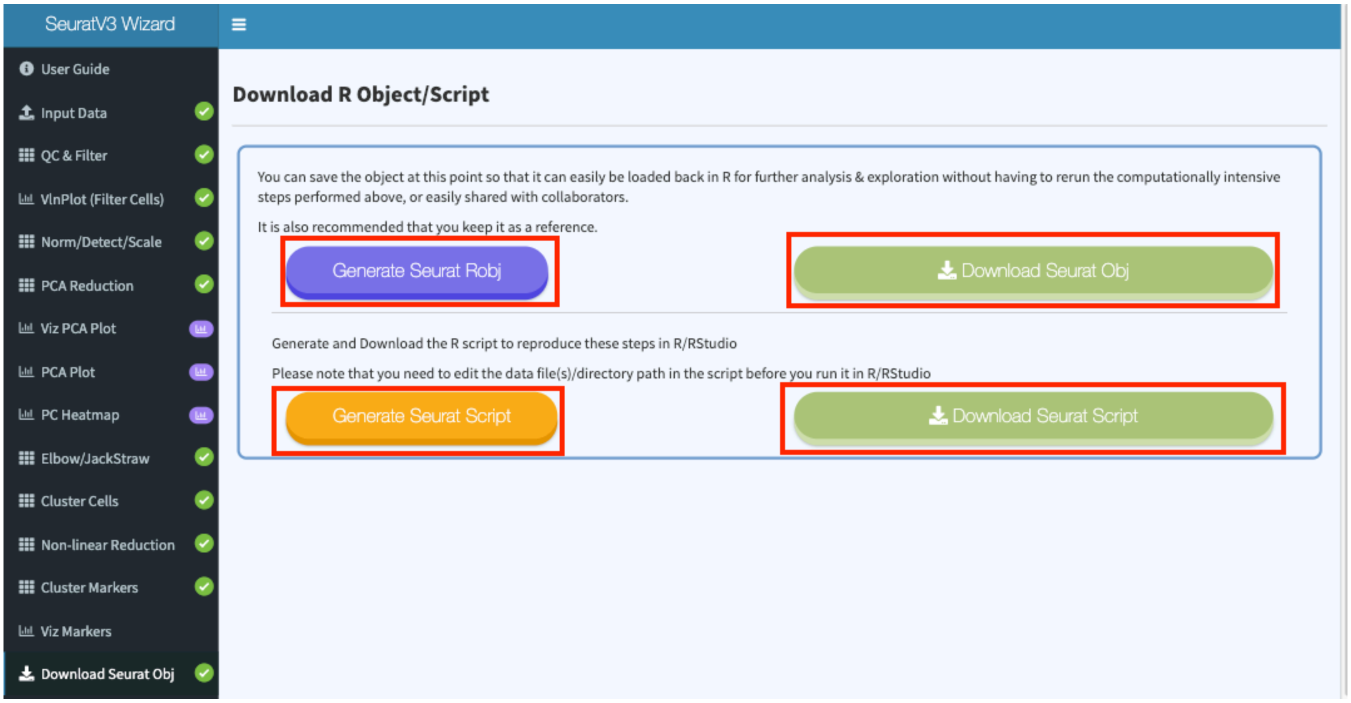
Cluster markers for "4" - 4

Click gene symbols below to color plot by gene

symbol	P-value	avg. logFC	PCT1
CCL5	5.8974e-210	2.19261754709585	0.98
NKG7	2.1200e-183	1.58628007073074	0.958
CST7	2.4787e-170	1.45753457063044	0.782

p) Last step would be to download the Seurat object for reproducibility and further analysis in R. Go to the last tab "**Download Seurat Obj**". Click "**Generate Seurat Obj**", once done click "**Download Seurat Obj**" when it appears.

You can also download the R script used for generating this analysis but clicking **"Generate Seurat Script"**.



Download R Object/Script

You can save the object at this point so that it can easily be loaded back in R for further analysis & exploration without having to rerun the computationally intensive steps performed above, or easily shared with collaborators.

It is also recommended that you keep it as a reference.

Generate and Download the R script to reproduce these steps in R/RStudio

Please note that you need to edit the data file(s)/directory path in the script before you run it in R/RStudio

References

Merkel, D. (2014). Docker: Lightweight linux containers for consistent development and deployment. *Linux J.*, 2014(239).

Soppelsa, F. and Kaewkasi, C. (2017). *Native Docker Clustering with Swarm*. Packt Publishing.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, https://www.R-project.org/.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2018). *shiny: Web Application Framework for R*. R package version 1.1.0.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15, 550.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36, 411.

Pyl, P. T., Anders, S., and Huber, W. (2014). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2), 166–169.

Smyth, G. K., Shi, W., and Liao, Y. (2013). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7), 923–930.

Sklenar, J., Nelson, J. W., Minnier, J., and Barnes, A. P. (2016). The START App: a web-based RNAseq analysis and visualization resource. *Bioinformatics*, 33(3), 447–449.

Stuart, T. and Satija, R. (2019). Integrative single-cell analysis. *Nature Reviews Genetics*, 20(5), 257–272.

Patel, M. V. (2018). iS-CellR: a user-friendly tool for analyzing and visualizing single-cell RNA sequencing data. *Bioinformatics*, 34(24), 4305–4306.

Hafemeister, C. and Satija, R. (2019). Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *bioRxiv*.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterprofiler: an r package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*, 16(5), 284–287.

Luo, W. and Brouwer, C. (2013). Pathview: an R/Bioconductor package for pathway- Li, Y. and

Andrade, J. (2017). Deapp: an interactive web interface for differential expression analysis of next generation sequence data. *Source Code for Biology and Medicine*, 12(1), 2._based data integration and visualization. *Bioinformatics*, 29(14), 1830–1831.

Quereda, J. J., Dussurget, O., Nahori, M.-A., Ghozlane, A., Volant, S., Dillies, M.-A., Regnault, B., Kennedy, S., Mondot, S., Villoing, B., Cossart, P., and Pizarro-Cerda, J. (2016). Bacteriocin from epidemic listeria strains alters the host intestinal microbiota to favor infection. *Proceedings of the National Academy of Sciences*, 113(20), 5706–5711.

Choi, E. , Kraus, M. R., Lemaire, L. A., Yoshimoto, M. , Vemula, S. , Potter, L. A., Manduchi, E. , Stoeckert, C. J., Grapin_Botton, A. and Magnuson, M. A. (2012), Dual Lineage_Specific Expression of Sox17 During Mouse Embryogenesis. STEM CELLS, 30: 2297-2308.

[dataset] (2016). Transcription profiling by high throughput sequencing of Sox17.Epi and Endo cells from mouse embryos, atlas-experiments, V1. http://www.ebi.ac.uk/gxa/experiments/E-MTAB-970.