

# Ganatum: a graphical single-cell RNA-seq analysis pipeline

*User Manual*

February 20, 2017

## Contents

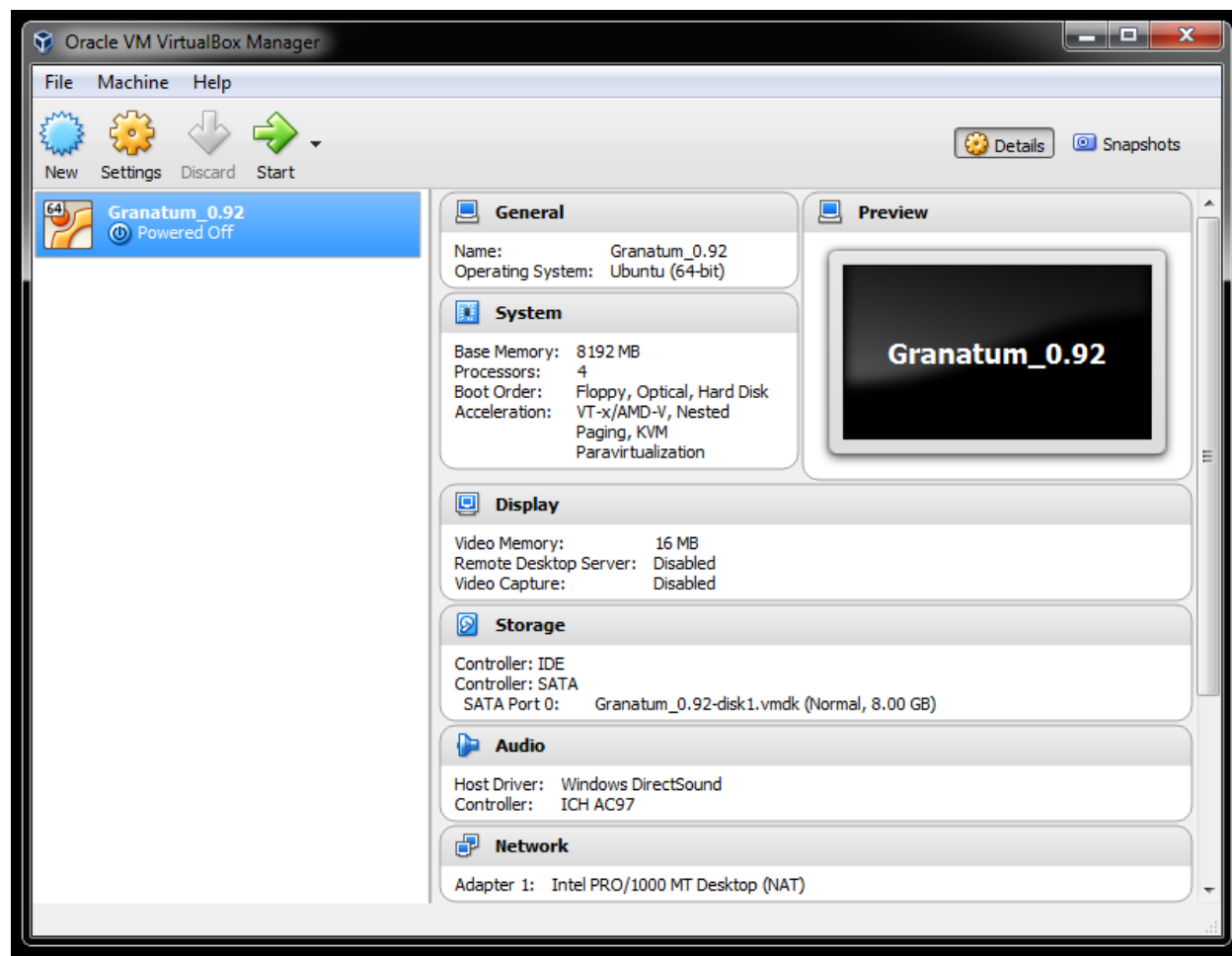
1. Introduction .....	1
2. Setup .....	1
3. Upload.....	3
4. Batch-effect removal .....	5
5. Outlier removal .....	6
6. Normalization.....	8
7. Gene filtering .....	10
8. Clustering .....	11
9. Differential expression.....	13
10. Protein network .....	16
11. Pseudo-time .....	17
12. References .....	18

## 1. Introduction

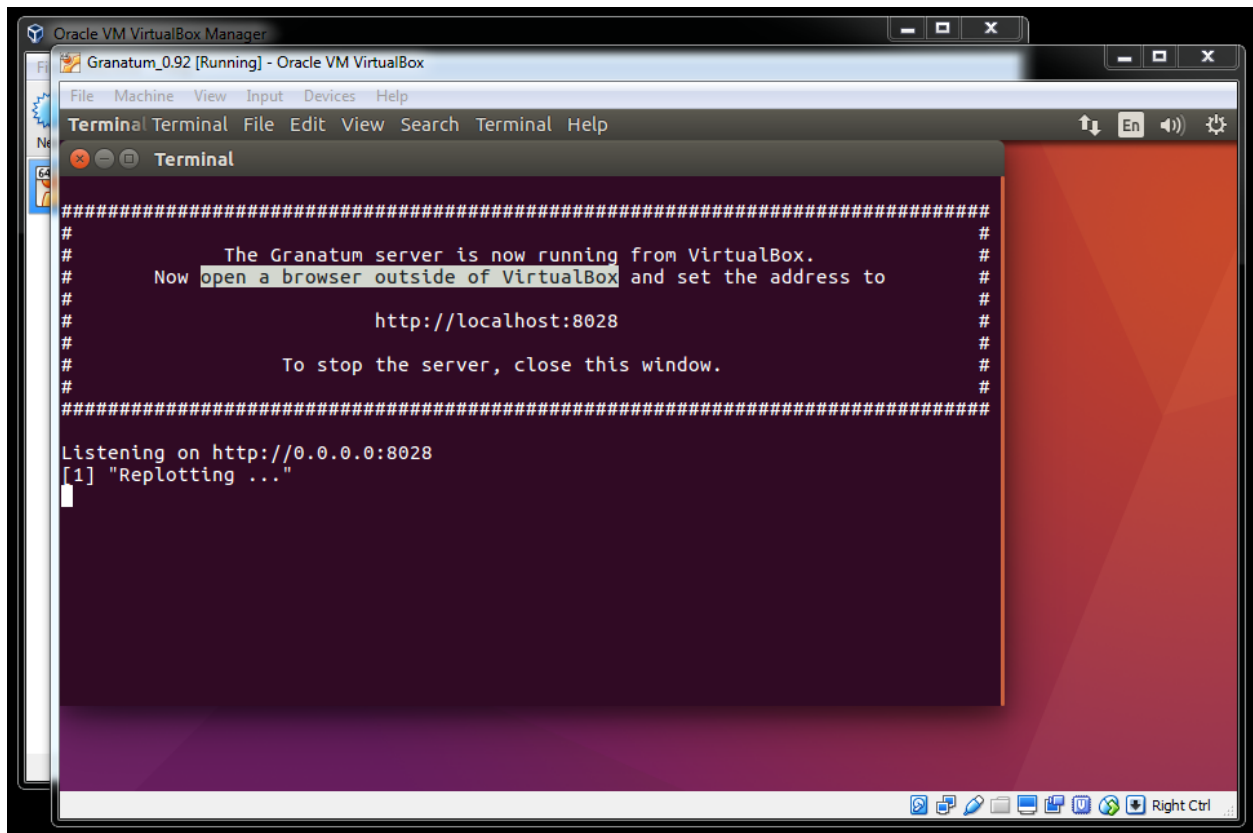
Granatum is graphically driven analysis platform for single-cell high-throughput RNA sequencing (scRNA-seq) data. The technology of scRNA-seq allows for the detection of distinct expression profiles between individual cells from heterogeneous populations. Applications of this technology have included detecting the expression differences between cancer/normal cells in heterogeneous samples [1], identifying differences between primary and metastatic cancer cells [2], and tracing the path of cell fates in development over time [3]. This tutorial will walk through Granatum's graphical interface for scRNA-seq analysis. It includes procedures for uploading data, removing batch effects, removing cell outliers, normalizing expression levels, filtering genes, clustering cells, identifying differentially expressed genes, visualizing protein network interaction, and constructing a pseudo-time path.

## 2. Setup

Begin Granatum setup by installing VirtualBox (<https://www.virtualbox.org/>) on a computer with at least 16 GB RAM and preferably an Intel i7 processor or equivalent. Start VirtualBox and import our pre-configured Granatum server file (VirtualBox Appliance): click "File"→"Import Appliance...", select our Granatum.ova file, and click through the remaining "Import Appliance" steps. A new "Granatum" entry will appear in your Appliance list.



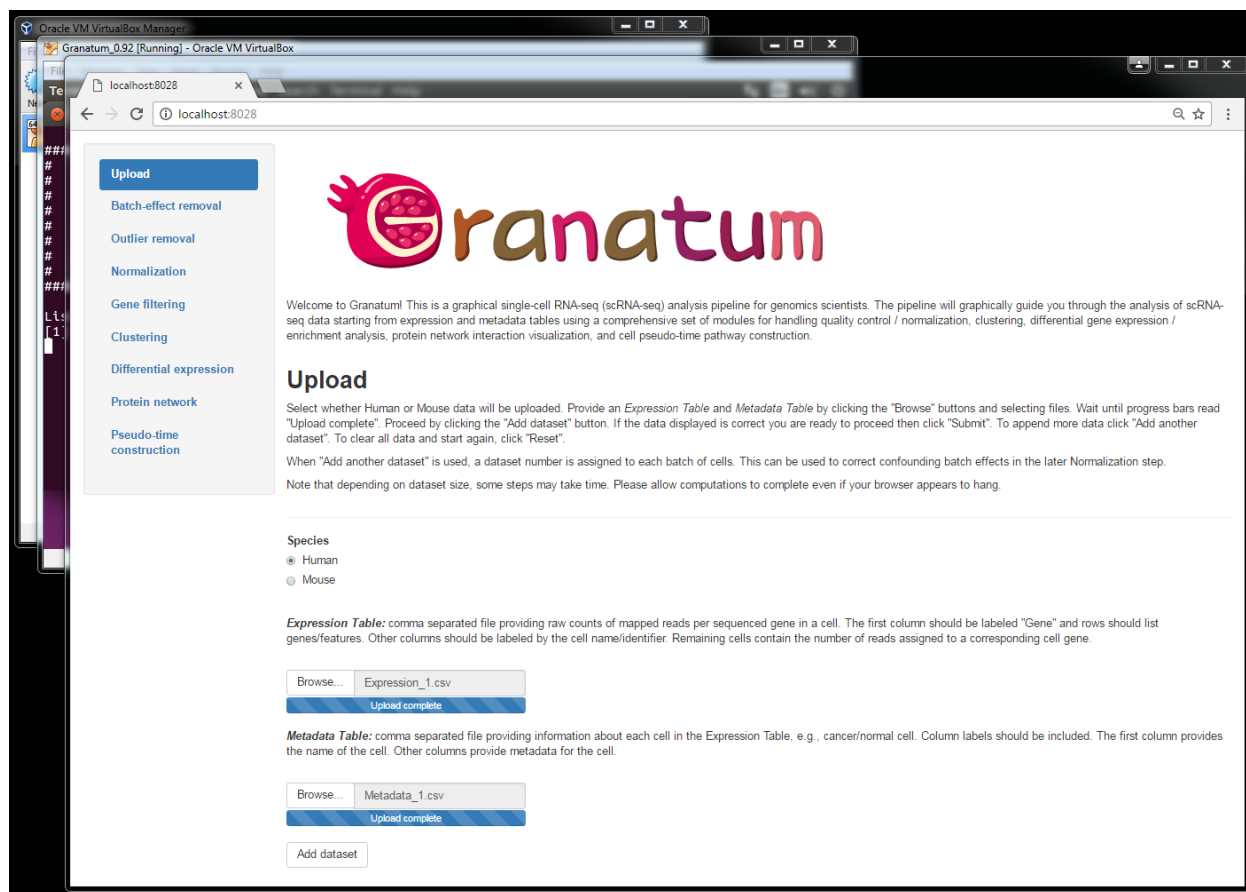
Double-click this entry to start the Appliance. You will be presented with an Ubuntu desktop in the Appliance, with a window showing Granatum startup messages. Wait for a message indicating that the server is running.



Now you are ready to open a web browser, e.g., Chrome, from outside of the VirtualBox Appliance. Set this browser to the following address:

<http://localhost:8028/>

The Granatum welcome and Upload page will be presented.



### 3. Upload

Granatum requires two files per dataset – an **Expression Table** file and a **Metadata Table** file. Both tables are formatted as comma separated files. The **Expression Table** first column first row entry should be left blank or be labeled “Gene”. The remaining columns should have cell identifiers in the first row. The remaining rows of the first column should have gene identifiers. Other entries should provide raw number of reads mapped to each gene for each cell. The **Metadata Table** first row provides column labels. Rows in the first column provide the same cell identifiers as in the expression table columns. The remaining columns may include information about each cell, e.g., “primary” or “metastatic”.

To input the data, first choose what species was sequenced (Human or Mouse), then select table files by clicking the “Browse” buttons. For this example we are using results from Kim, et al. [2], which will show a segregation between primary and metastatic renal cancer cells. Here, we have split the dataset into three sets of files (ending \_1.csv, \_2.csv, and \_3.csv) corresponding to different cell sources (patient vs. PDX and primary vs. metastatic). Once both status bars indicate “Upload complete” the “Add dataset” button can be clicked. This will bring you to a dataset preview page, showing the most recently uploaded data.

Upload

Batch-effect removal

Outlier removal

Normalization

Gene filtering

Clustering

Differential expression

Protein network

Pseudo-time construction

## Upload

Select whether Human or Mouse data will be uploaded. Provide an *Expression Table* and *Metadata Table* by clicking the "Browse" buttons and selecting files. Wait until progress bars read "Upload complete". Proceed by clicking the "Add dataset" button. If the data displayed is correct you are ready to proceed then click "Submit". To append more data click "Add another dataset". To clear all data and start again, click "Reset".

When "Add another dataset" is used, a dataset number is assigned to each batch of cells. This can be used to correct confounding batch effects in the later Normalization step.

Note that depending on dataset size, some steps may take time. Please allow computations to complete even if your browser appears to hang.

Summary of datasets uploaded

Dataset	Number of genes	Number of samples
1	19924	36
Total num of distinct genes: 19924		Total num of samples: 36

Expression Table

Metadata Table

Show 10 entries

Search:

Gene	PDX_mRCC_SC_1	PDX_mRCC_SC_79	PDX_mRCC_SC_4	PDX_mRCC_SC_87	PDX_mRCC_SC_34	PDX_mRCC_SC_89	PDX_mRCC_SC_65
A1BG	0	0	0	0	0	0	0
A1CF	0	0	0	0	0	0	0
A2M	0	0	0	0	0	0	0
A2ML1	0	0	0	0	0	0	24
A2MP1	0	0	0	0	0	0	0
A3GALT2	0	0	0	0	0	0	0
A4GALT	0	0	0	0	0	0	0
A4GNT	0	0	0	0	0	0	0
AAAS	0	0	161	0	0	28	0
AACS	86	0	0	0	12	32	9

Gene

PDX\_mRCC\_SC\_1

PDX\_mRCC\_SC\_79

PDX\_mRCC\_SC\_4

PDX\_mRCC\_SC\_87

PDX\_mRCC\_SC\_34

PDX\_mRCC\_SC\_89

PDX\_mRCC\_SC\_65

Showing 1 to 10 of 19,924 entries

Previous

1

2

3

4

5

...

1993

Next

Add another dataset

Reset

Submit

Click the tabs (Expression Table or Metadata Table) to switch between previews of Expression/Metadata inputs. If the data looks correct, click "Submit". If something looks wrong click "Reset" and all data will be purged. If you wish to append an additional batch of data click "Add another dataset" to select and upload the additional batch. This will be relevant in the Batch-effect removal stage.

**Upload**

Select whether Human or Mouse data will be uploaded. Provide an *Expression Table* and *Metadata Table* by clicking the "Browse" buttons and selecting files. Wait until progress bars read "Upload complete". Proceed by clicking the "Add dataset" button. If the data displayed is correct you are ready to proceed then click "Submit". To append more data click "Add another dataset". To clear all data and start again, click "Reset".

When "Add another dataset" is used, a dataset number is assigned to each batch of cells. This can be used to correct confounding batch effects in the later Normalization step. Note that depending on dataset size, some steps may take time. Please allow computations to complete even if your browser appears to hang.

Species

☒ Human

☐ Mouse

**Expression Table:** comma separated file providing raw counts of mapped reads per sequenced gene in a cell. The first column should be labeled "Gene" and rows should list genes/features. Other columns should be labeled by the cell name/identifier. Remaining cells contain the number of reads assigned to a corresponding cell gene.

Browse... Expression\_2.csv

Upload complete

**Metadata Table:** comma separated file providing information about each cell in the Expression Table, e.g., cancer/normal cell. Column labels should be included. The first column provides the name of the cell. Other columns provide metadata for the cell.

Browse... Metadata\_2.csv

Upload complete

Add dataset

Summary statistics for each dataset that has been uploaded are presented to help you keep track of what has been entered.

**Upload**

Select whether Human or Mouse data will be uploaded. Provide an *Expression Table* and *Metadata Table* by clicking the "Browse" buttons and selecting files. Wait until progress bars read "Upload complete". Proceed by clicking the "Add dataset" button. If the data displayed is correct you are ready to proceed then click "Submit". To append more data click "Add another dataset". To clear all data and start again, click "Reset".

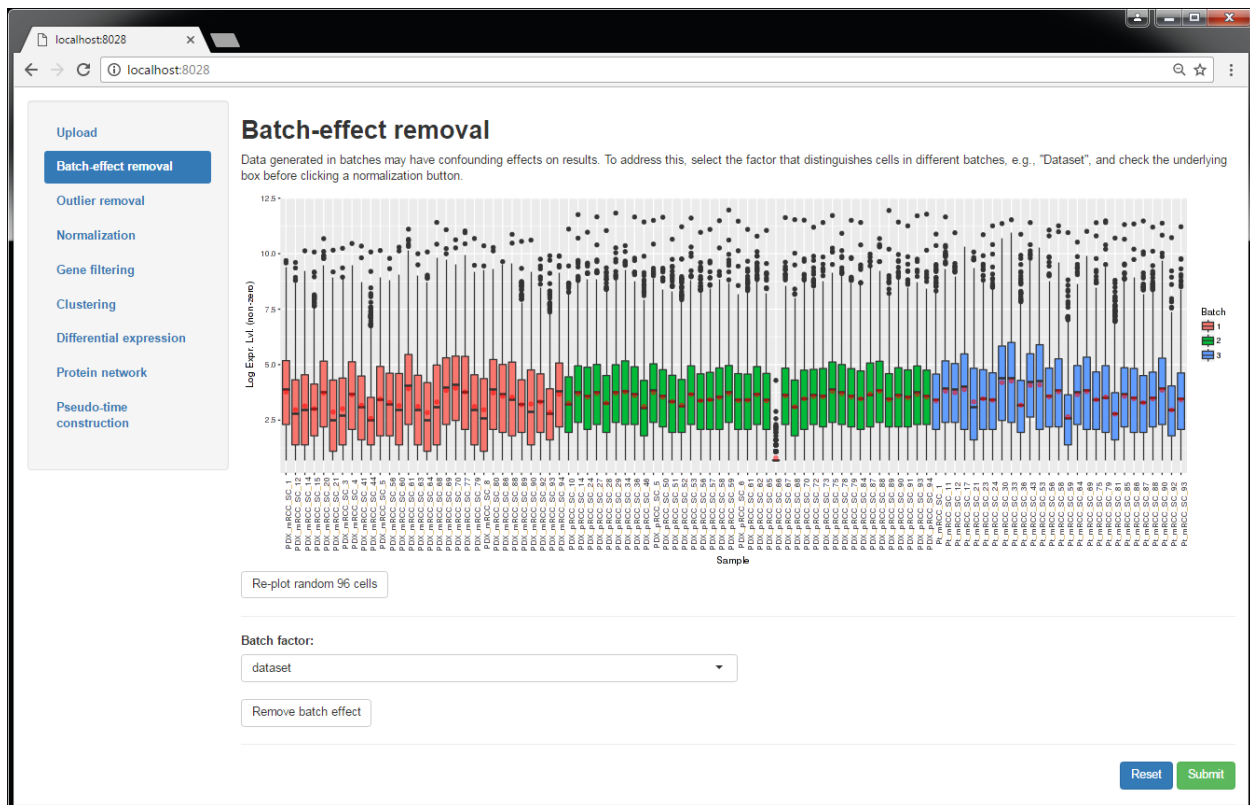
When "Add another dataset" is used, a dataset number is assigned to each batch of cells. This can be used to correct confounding batch effects in the later Normalization step. Note that depending on dataset size, some steps may take time. Please allow computations to complete even if your browser appears to hang.

Summary of datasets uploaded

Dataset	Number of genes	Number of samples
1	19924	36
2	19924	47
3	19924	35
Total num of distinct genes: 19924		Total num of samples: 118

## 4. Batch-effect removal

Data generated in batches may have confounding effects on results. Here we display box plots of expression levels for cells and allow for batch corrections. The orange dots indicate geometric means. For computational reasons, levels for up to 96 randomly selected cells are shown. To re-plot another random selection of cells click "Re-plot random 96 cells". To address batch effects, select the factor that distinguishes cells in different batches, e.g., "dataset", and click the "Remove batch effect button". To go back to the original uploaded data, click "Reset". Once ready to move to the next step, click "Submit".



## 5. Outlier removal

Some cells may have been damaged or had problems in library preparation and/or sequencing. In this step, problematic cells can be identified and removed. Two plots cluster cells according to their expression profiles either by PCA (left plot) or t-SNE (right plot) dimensionality reduction method. To change how cells are colored/labeled make a selection from the “Cell labels” drop down list. Cells (points) lying outside of clusters can be manually selected from one or both plots simultaneously. Selected cells will gain a “halo”. To clear selections click “De-select all”.



localhost:8028

localhost:8028

Upload

Batch-effect removal

Outlier removal

Normalization

Gene filtering

Clustering

Differential expression

Protein network

Pseudo-time construction

## Outlier removal

Remove cells with unusual expression levels, possibly caused by damage at capture, poor cell health, or problematic library preparation or sequencing. These cells may have expression patterns that confound the main results and should therefore be removed before downstream analysis. By default, plots are made using the top 50 expressed genes in each cell; to use all genes de-select the box labeled "Cluster using only top expressed genes".

Select outliers automatically by clicking "Auto-identify" and then setting parameters in pop-up box. To select/de-select outliers manually, click on points (representing cells) in the interactive plots. Remove selected cells by clicking "Remove selected". To de-select all cells click "De-select all". To start again with all cells click "Reset". Proceed by clicking "Submit".

Cell labels (from metadata)

Type

PC2

7.5

5.0

2.5

0.0

-2.5

-10

0

10

20

30

40

PC1

PDX\_meta

PDX\_primary

Pt\_meta

tSNE2

7

4

-4

-1

0

4

tSNE1

PDX\_meta

PDX\_primary

Pt\_meta

☒ Cluster using only top expressed genes

Auto-identify

Remove selected

De-select all

Reset

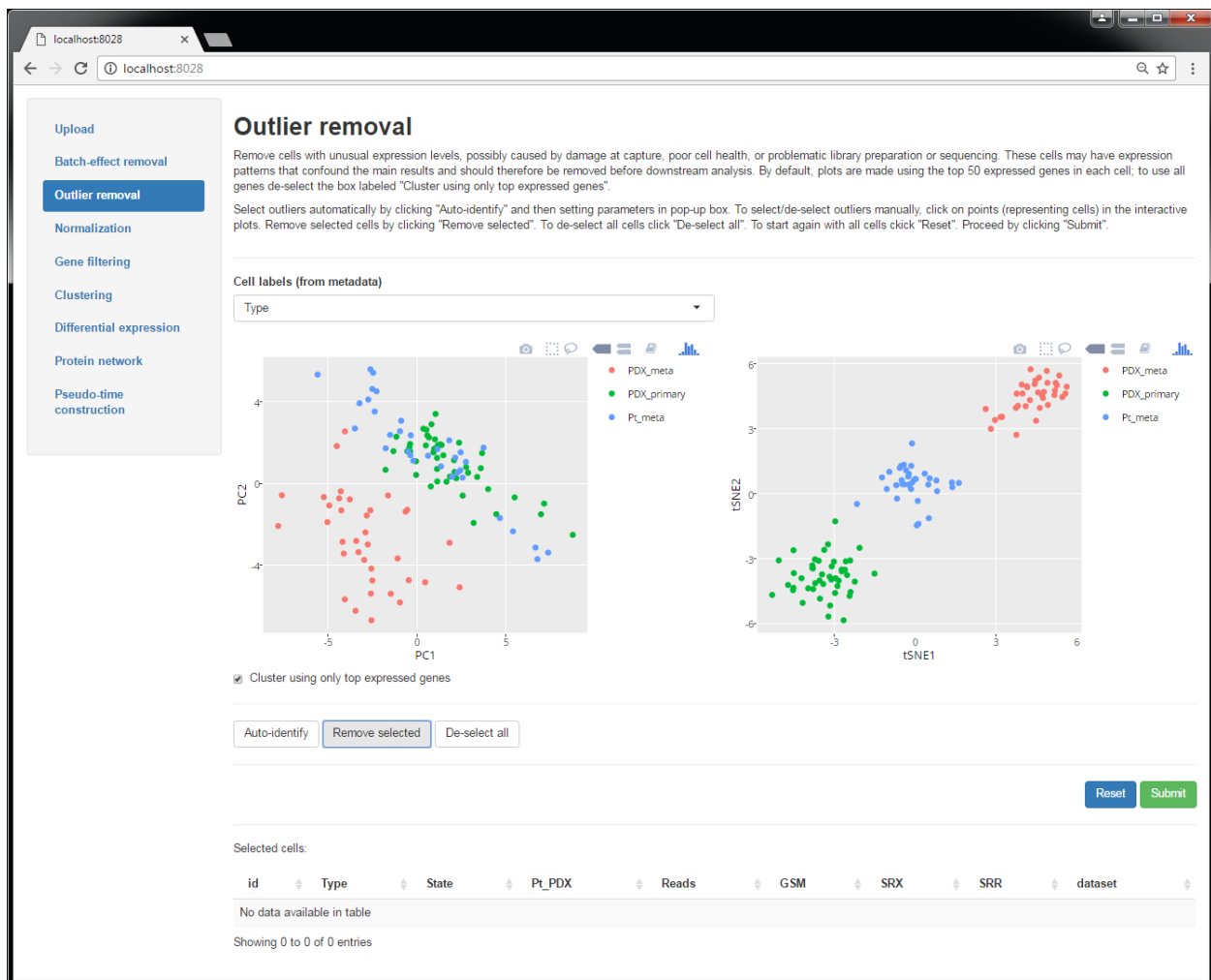
Submit

Selected cells:

id	Type	State	Pt_PDX	Reads	GSM	SRX	SRR	dataset
Pt_mRCC_SC_5	Pt_meta	mRCC	Pt	36234	GSM1887310	SRX1253756	SRR2431431	3
PDX_pRCC_SC_66	PDX_primary	pRCC	PDX	7603	GSM1887283	SRX1253736	SRR2431411	2

Showing 1 to 2 of 2 entries

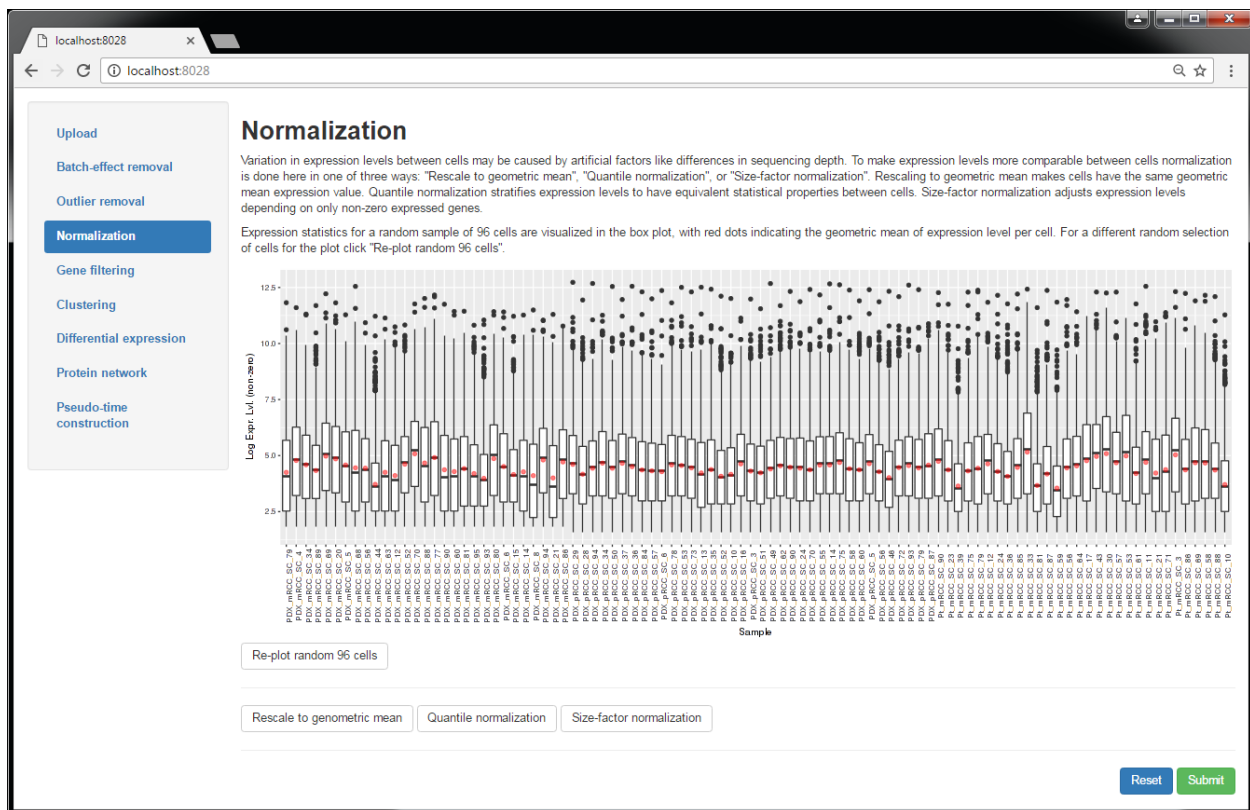
Cells can also be automatically removed using the “Auto-identify” button. Once cells to be removed are selected click “Remove selected” to remove them before proceeding.



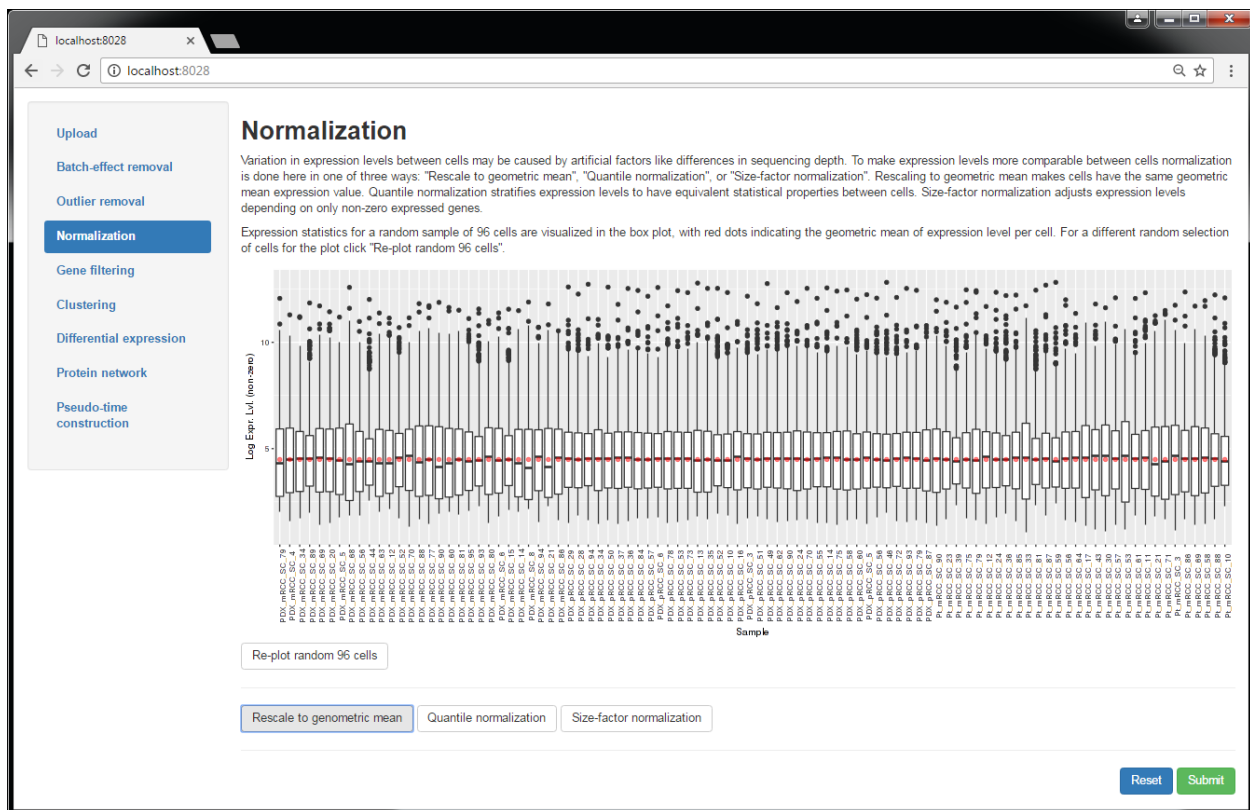
To reset to original graphs click “Reset”. Proceed by clicking “Submit”.

## 6. Normalization

To make better comparisons between the expression profiles of different cells/batches, normalize their expression levels here. The box plot indicates expression levels for individual cells, with an orange dot indicating the geometric mean.



For computational reasons, values for up to 96 randomly selected cells are shown. To plot another random selection, click "Re-plot random 96 cells". Your input metadata may provide information about which cells were processed together in a batch. In the Pate, et al. generated data individual cells were sequenced from five groups. To correct for potential bias arising from differences between the processing of batches we can select the metadata category and click the box labeled "Perform ComBat". Next, to normalize expression levels across all cells click one of three buttons representing a normalization method: "Rescale to geometric mean", "Quantile normalization", or "Size-factor normalization". Click "Rescale to geometric mean" and the orange dots will align.



Click "Reset" if you would like to start with original input expression levels again. Click "Submit" to proceed to the next step with the (normalized) data.

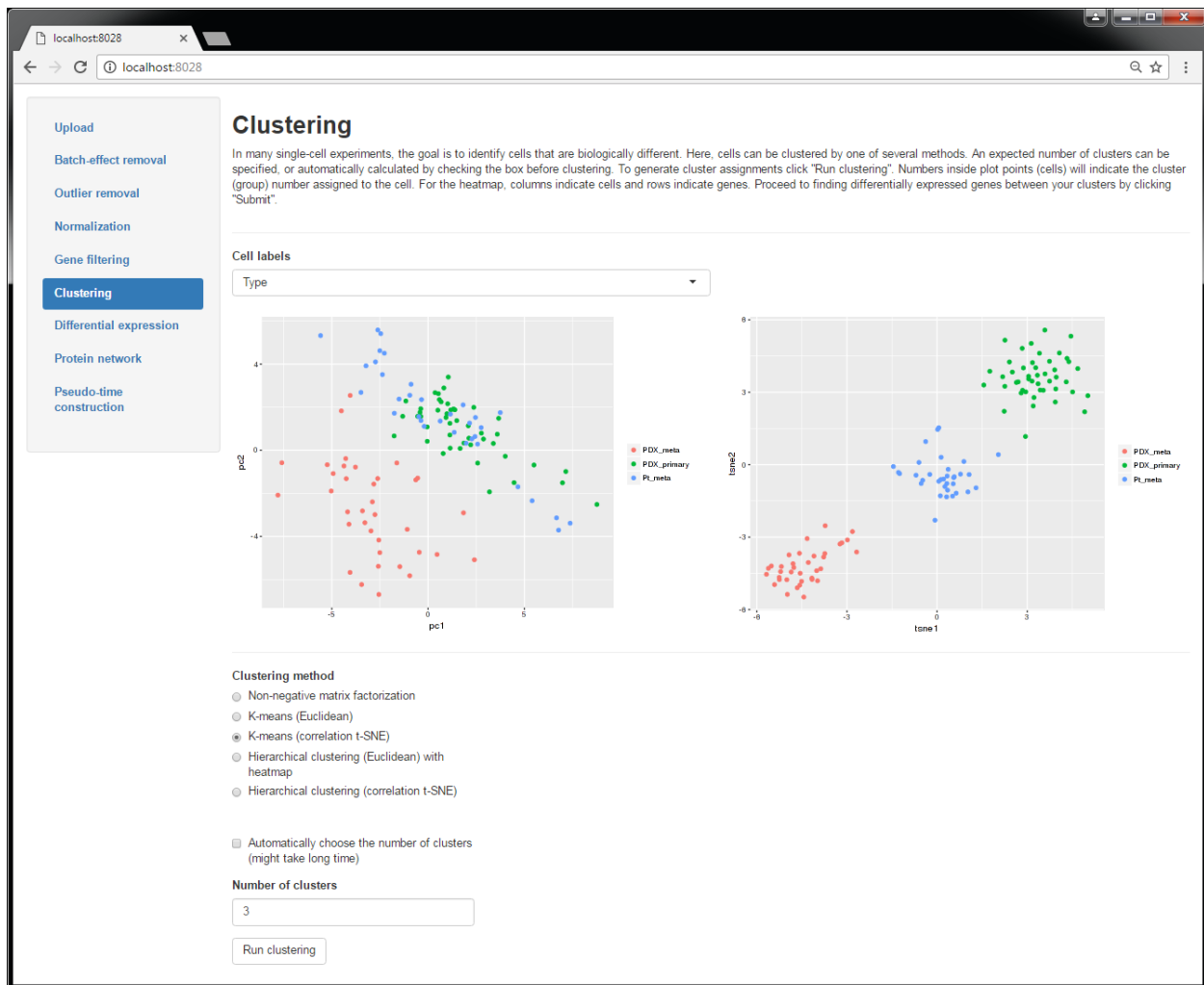
## 7. Gene filtering

In the "Gene filtering" step lowly expressed and low variably expressed genes can be filtered out by moving the "Log Mean Expression Threshold" and "Dispersion Fit Threshold" sliders rightward, respectively. We recommend keeping at least 2,000 genes (number is listed under "Post-filtering number of genes") to keep some methods relevant, like differential gene expression analysis. Once satisfied with filtering parameters click "Submit" to proceed with the filtered gene set.



## 8. Clustering

Initially, just the clustered cells with cell labels from input metadata are shown.



To calculate cluster assignments select a clustering method from the list, then choose to automatically estimate number of clusters by clicking the checkbox or enter a specific number before clicking "Run clustering". Once clustering is completed, the cluster assignments are indicated by numbers within the plot points.

localhost:8028

localhost:8028

Upload

Batch-effect removal

Outlier removal

Normalization

Gene filtering

Clustering

Differential expression

Protein network

Pseudo-time construction

## Clustering

In many single-cell experiments, the goal is to identify cells that are biologically different. Here, cells can be clustered by one of several methods. An expected number of clusters can be specified, or automatically calculated by checking the box before clustering. To generate cluster assignments click "Run clustering". Numbers inside plot points (cells) will indicate the cluster (group) number assigned to the cell. For the heatmap, columns indicate cells and rows indicate genes. Proceed to finding differentially expressed genes between your clusters by clicking "Submit".

Cell labels

Type

Clustering method

☐ Non-negative matrix factorization
 ☐ K-means (Euclidean)
 ☒ K-means (correlation t-SNE)
 ☐ Hierarchical clustering (Euclidean) with heatmap
 ☐ Hierarchical clustering (correlation t-SNE)

☐ Automatically choose the number of clusters (might take long time)

Number of clusters

3

Run clustering

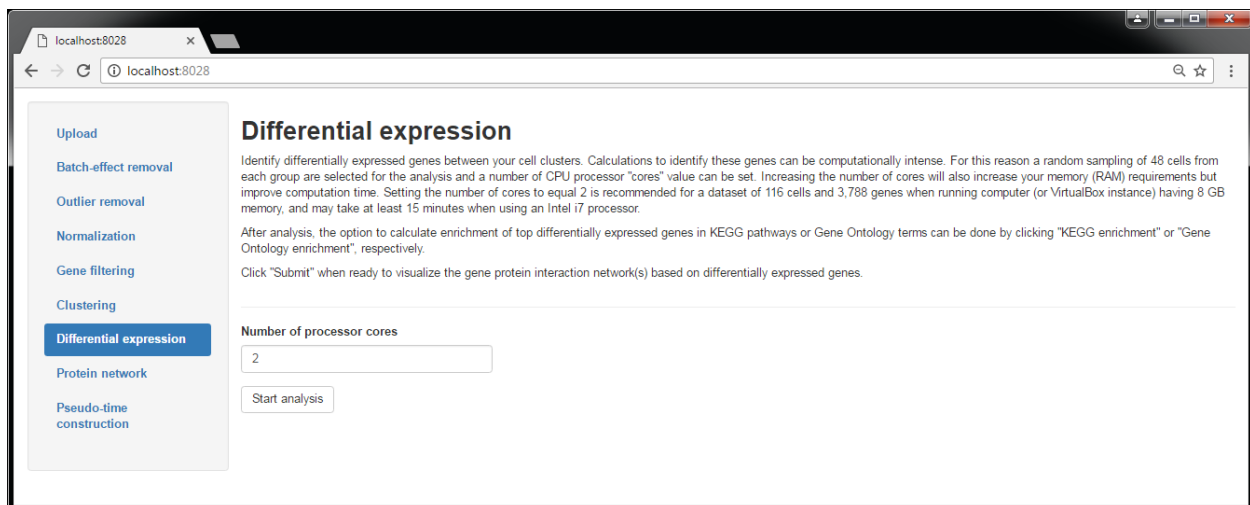
Submit

Click "Submit" to proceed to the next step.

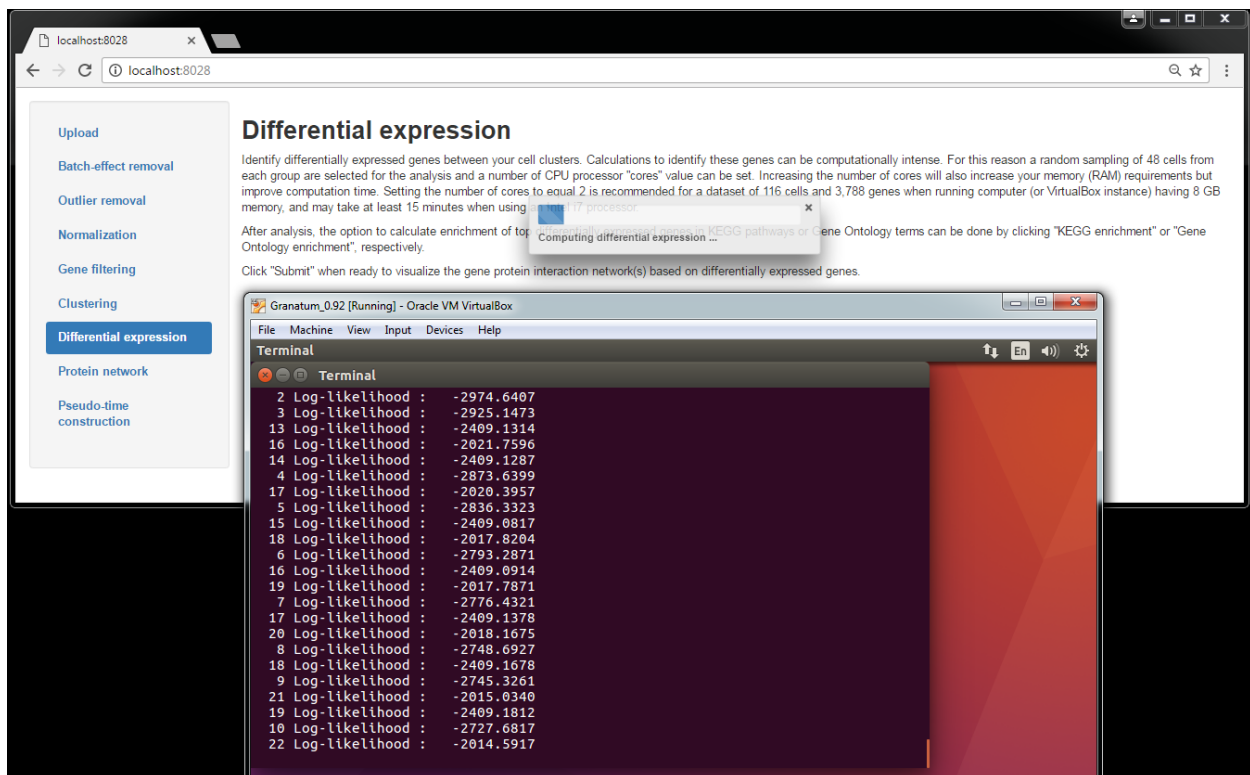
## 9. Differential expression

At the start of the "Differential expression" stage you will be asked to provide a "Number of processor cores", which will depend on your computer hardware. Due to high memory (RAM) requirements at this step, which will increase when using more cores, only 1 or 2 cores are recommended.

13

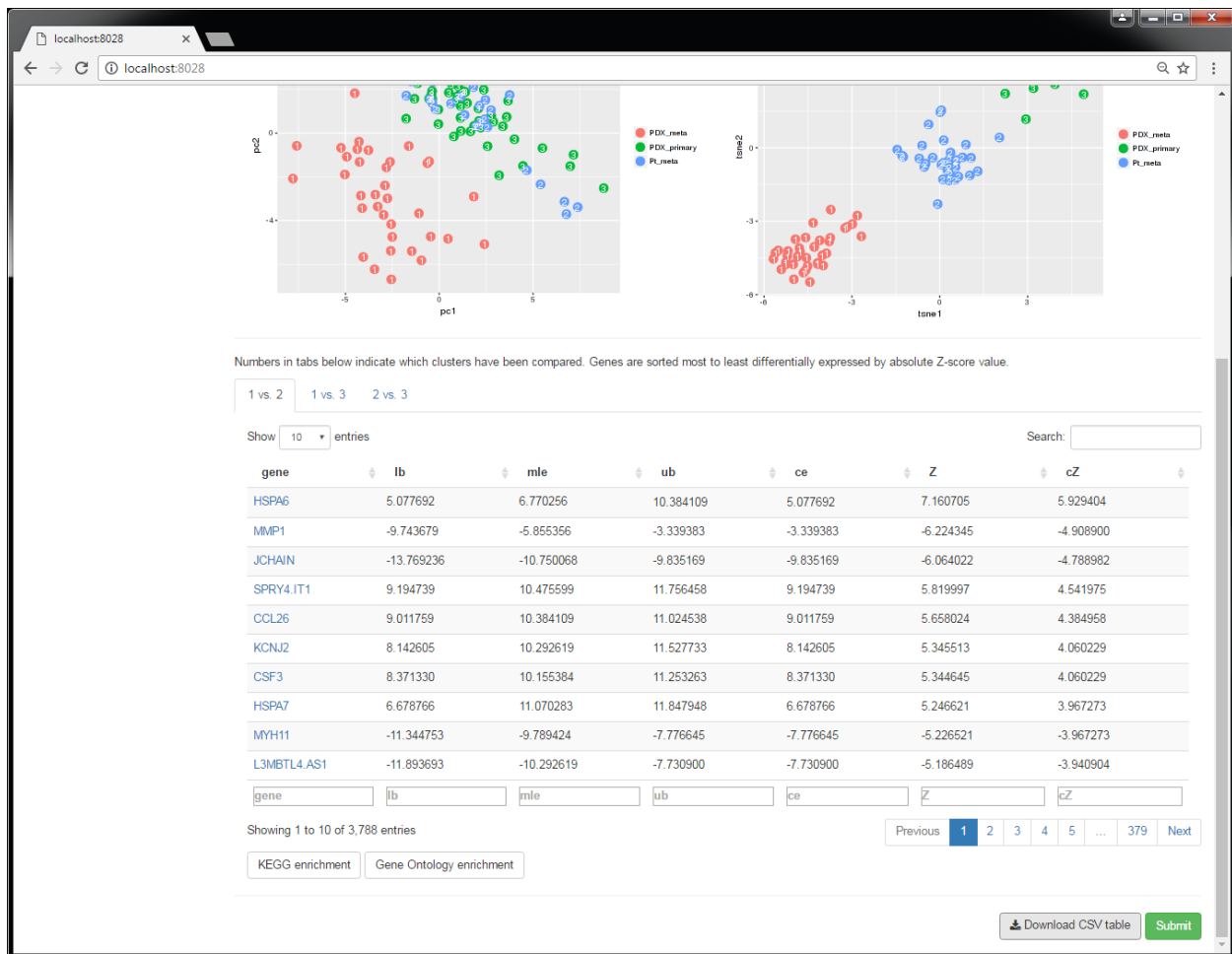


Click “Start analysis” to begin calculations. This step may take the most time, e.g., 30 minutes for three clusters using 116 cells and 3,788 genes. Messages will be displayed in the VirtualBox Appliance while processing takes place.

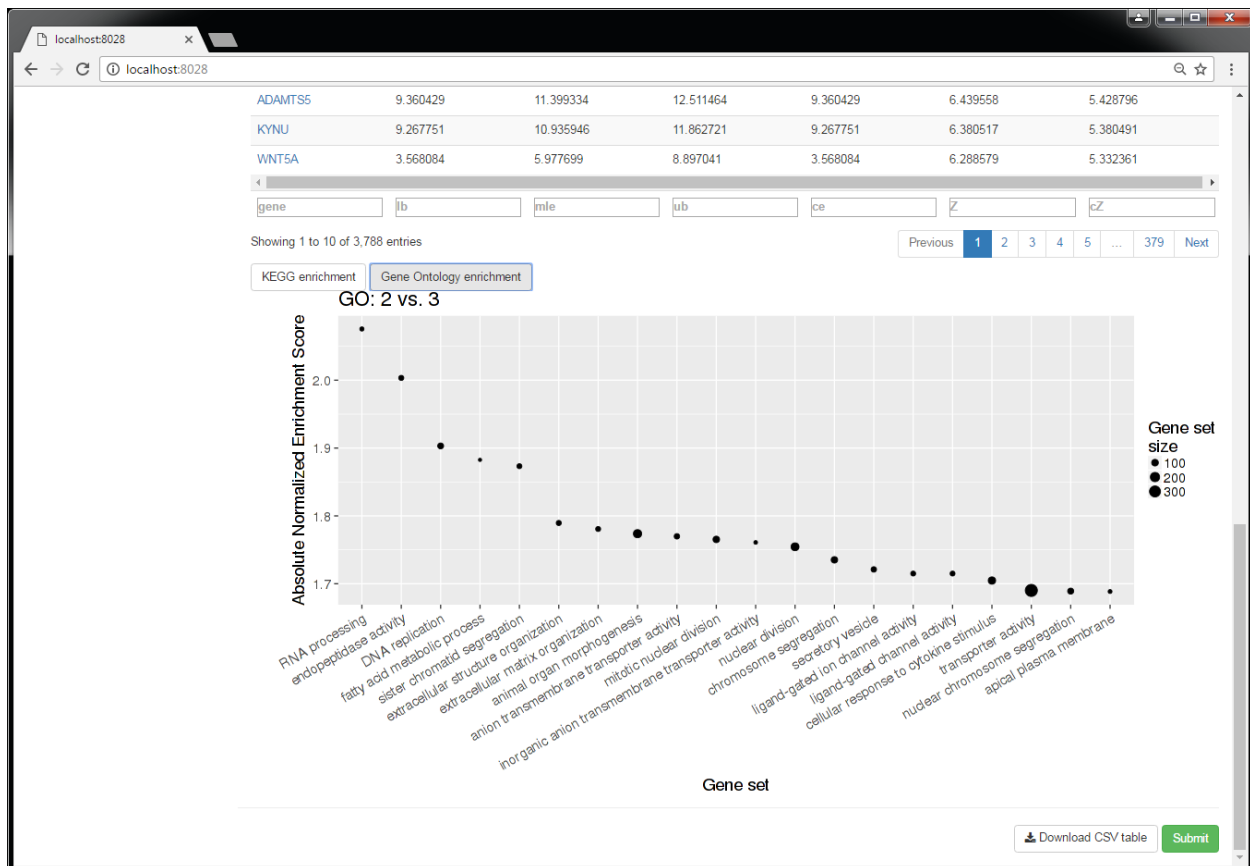


Once differentially expressed genes have been identified they will be displayed in tabs below the plots, sorted by absolute Z score value. Numbers in the tabs identify which clusters (from the plots) are being compared. This table can be downloaded using the “Download CSV table” button.





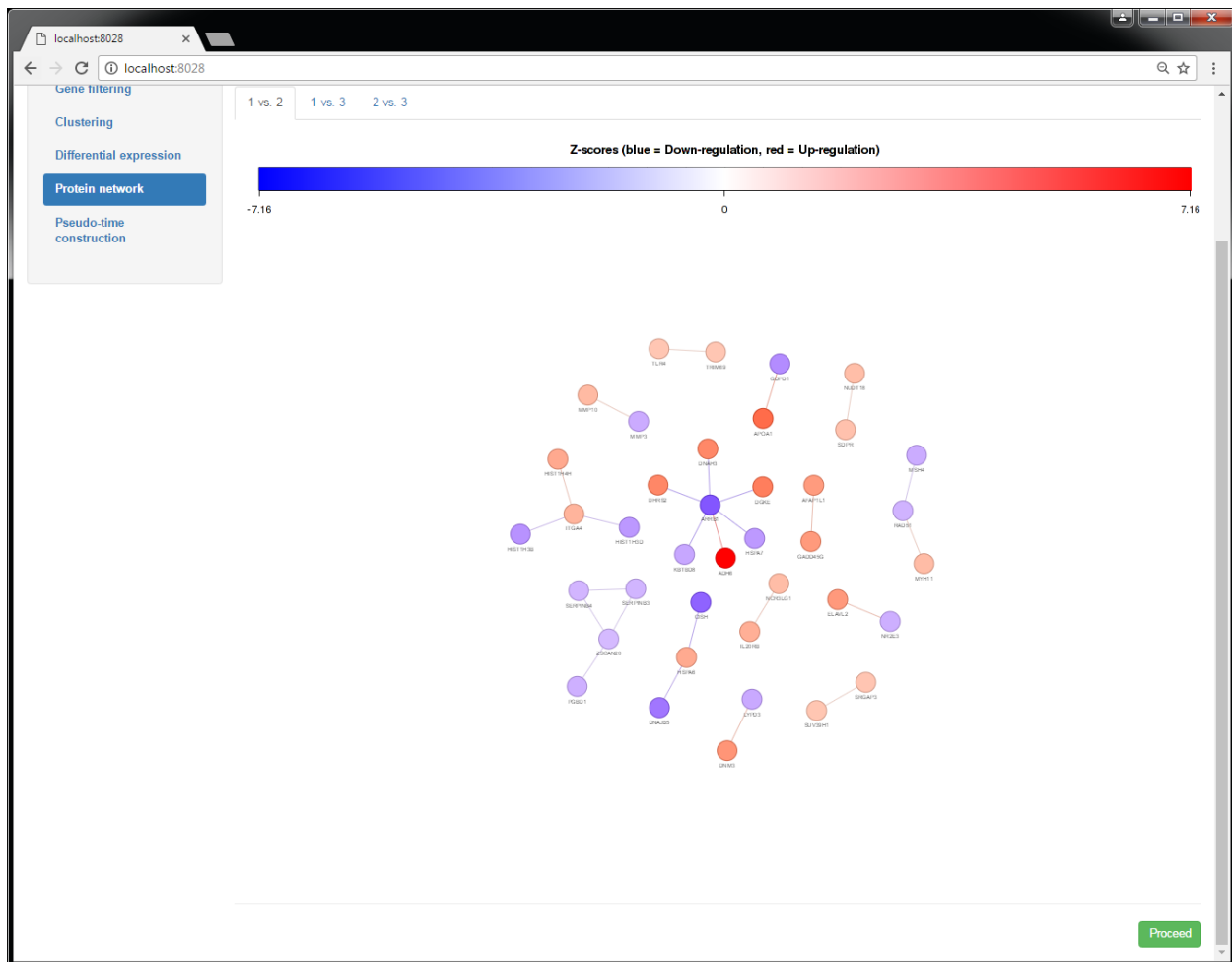
The enrichment of differentially expressed in KEGG pathways or Gene Ontology terms can be calculated for the selected tab by clicking the “KEGG enrichment” or “Gene Ontology enrichment” buttons, respectively.



Click "Submit" to proceed to the next step.

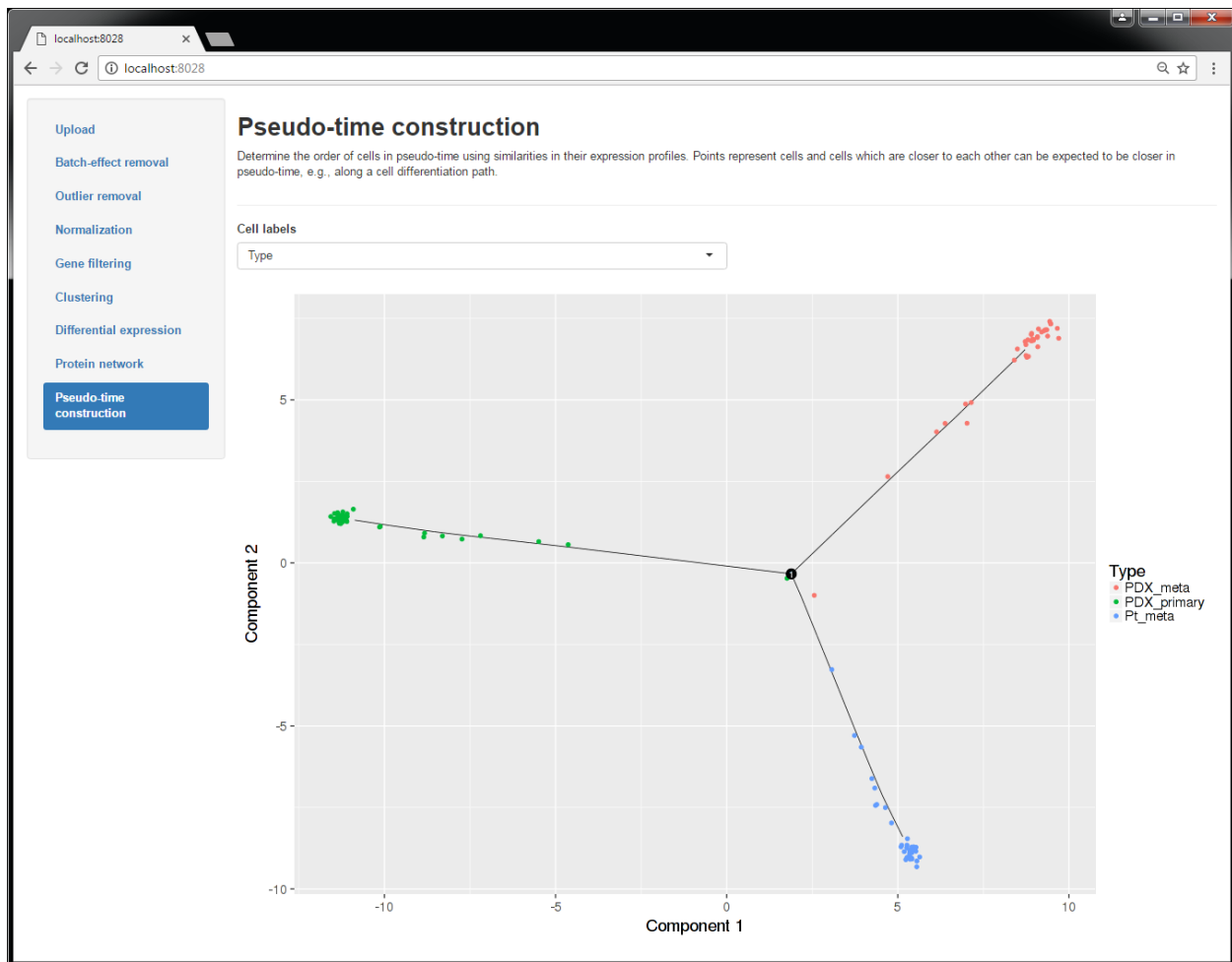
## 10. Protein network

Protein-protein interactions, e.g., publication-supported biochemical reactions, between the proteins encoded by top differentially expressed genes are displayed. Here is where you can examine the co-expression profile of associated genes. Tabs indicate which clusters of cells are being compared. Plot points represent proteins encoded by differentially expressed genes and lines represent a documented interaction. Colors represent the degree of under-/over-expression as indicated by the color bar at the top. Move your mouse wheel to zoom in and out. Points can be selected and moved to see them better in dense networks. Go to the next step by clicking "Proceed" at the bottom right of the page.



## 11. Pseudo-time

In this final step, determine the order of cells in pseudo-time using similarities in their expression profiles. Points represent cells and cells which are closer to each other can be expected to be closer in pseudo-time, e.g., along a cell differentiation path.



## 12. References

1. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed BV, Curry WT, Martuza RL *et al*: **Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma**. *Science* 2014, **344**(6190):1396-1401.
2. Kim K-T, Lee HW, Lee H-O, Song HJ, Jeong DE, Shin S, Kim H, Shin Y, Nam D-H, Jeong BC *et al*: **Application of single-cell RNA sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma**. *Genome Biology* 2016, **17**(1):80.
3. Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, Desai TJ, Krasnow MA, Quake SR: **Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq**. *Nature* 2014, **509**(7500):371-375.

## MISC/TRASH:

Upload

Gene filtering

Outlier removal

Normalization

Clustering

Differential expression

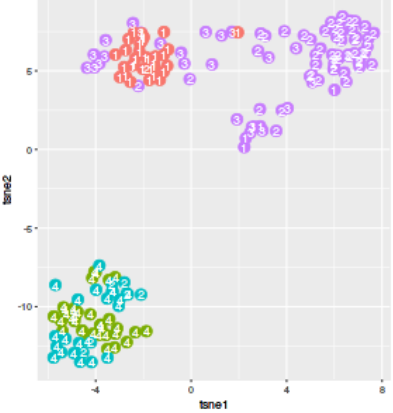

Protein network

Pseudo-time construction

### Clustering

In many single-cell experiments, the goal is to identify cells that are biologically different. Here, cells can be clustered by one of several methods. An expected number of clusters can be specified, or automatically calculated by checking the box before clustering. To generate cluster assignments click "Run clustering". Numbers inside plot points (cells) will indicate the cluster (group) number assigned to the cell. For the heatmap, columns indicate cells and rows indicate genes. Proceed to finding differentially expressed genes between your clusters by clicking "Submit".

Cell labels (from metadata):  
time\_point



Clustering method  
☒ Non-negative matrix factorization  
☐ K-means (Euclidean)  
☐ K-means (correlation t-SNE)  
☐ Hierarchical clustering (Euclidean) with heatmap  
☐ Hierarchical clustering (correlation t-SNE)  
  
☐ Automatically choose the number of clusters (might take long time)  
  
Number of clusters  
4  

Run clustering

Submit