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

User Manual

February 28, 2017

Contents

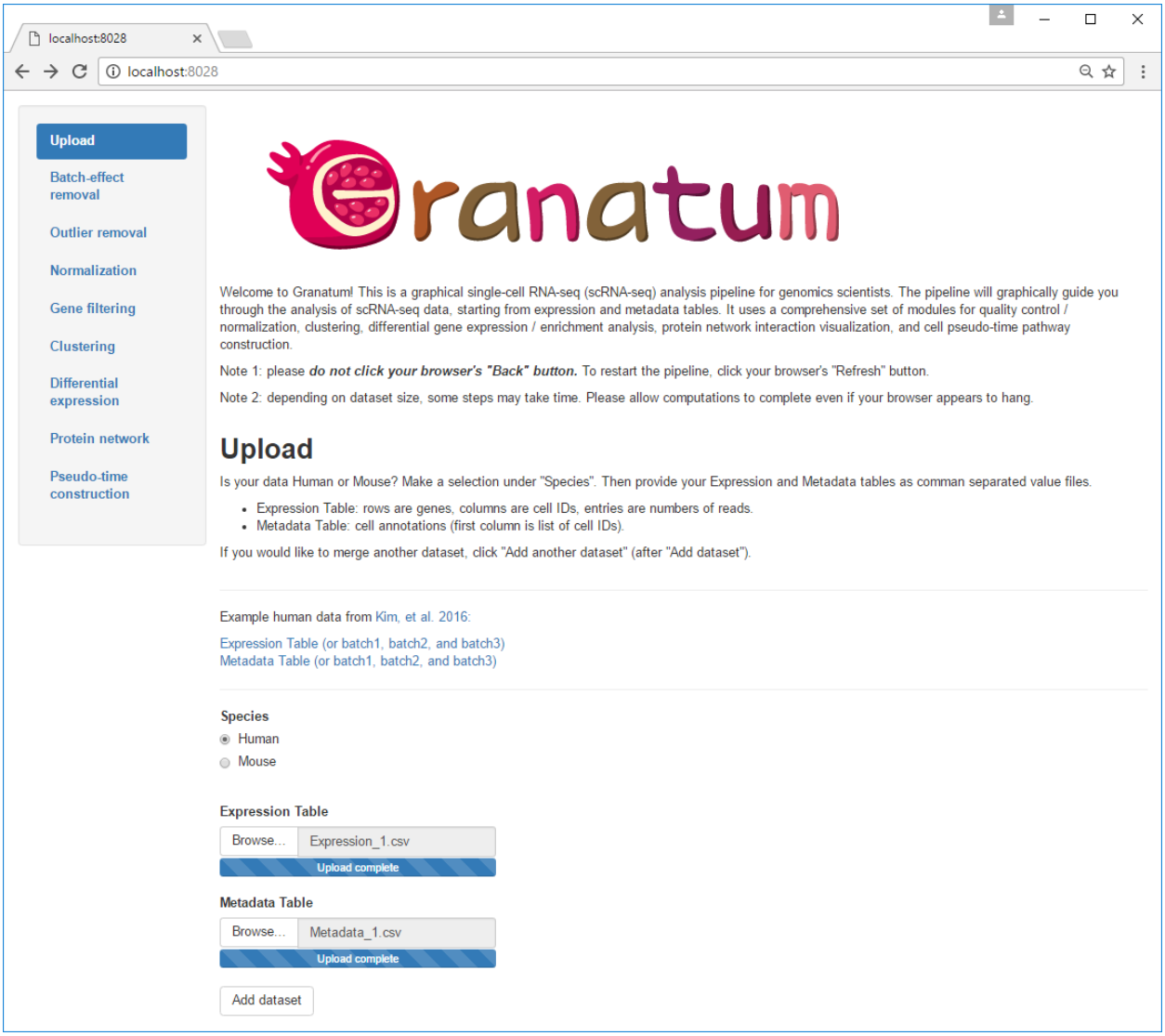
1. Introduction

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.

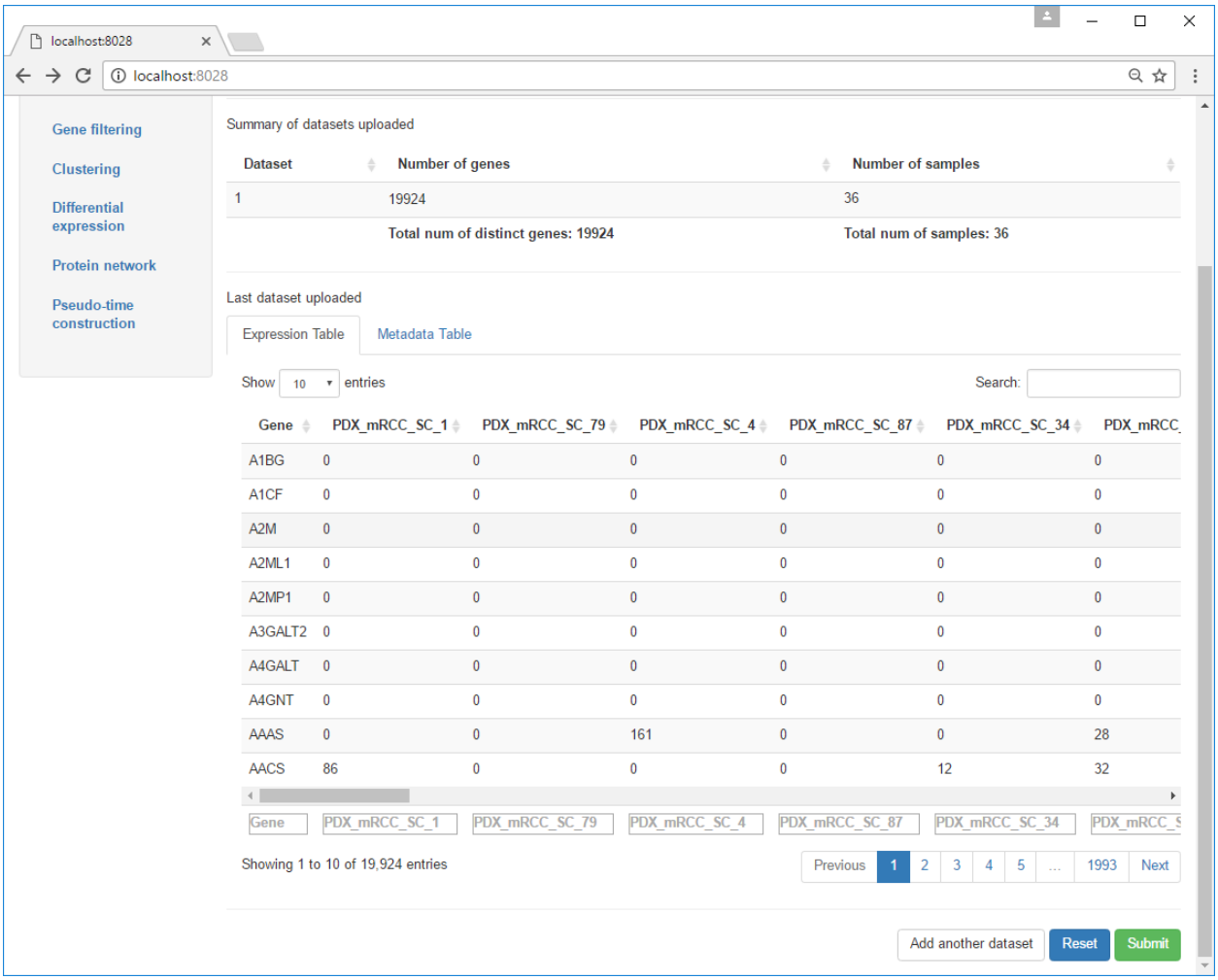
2. Upload

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".

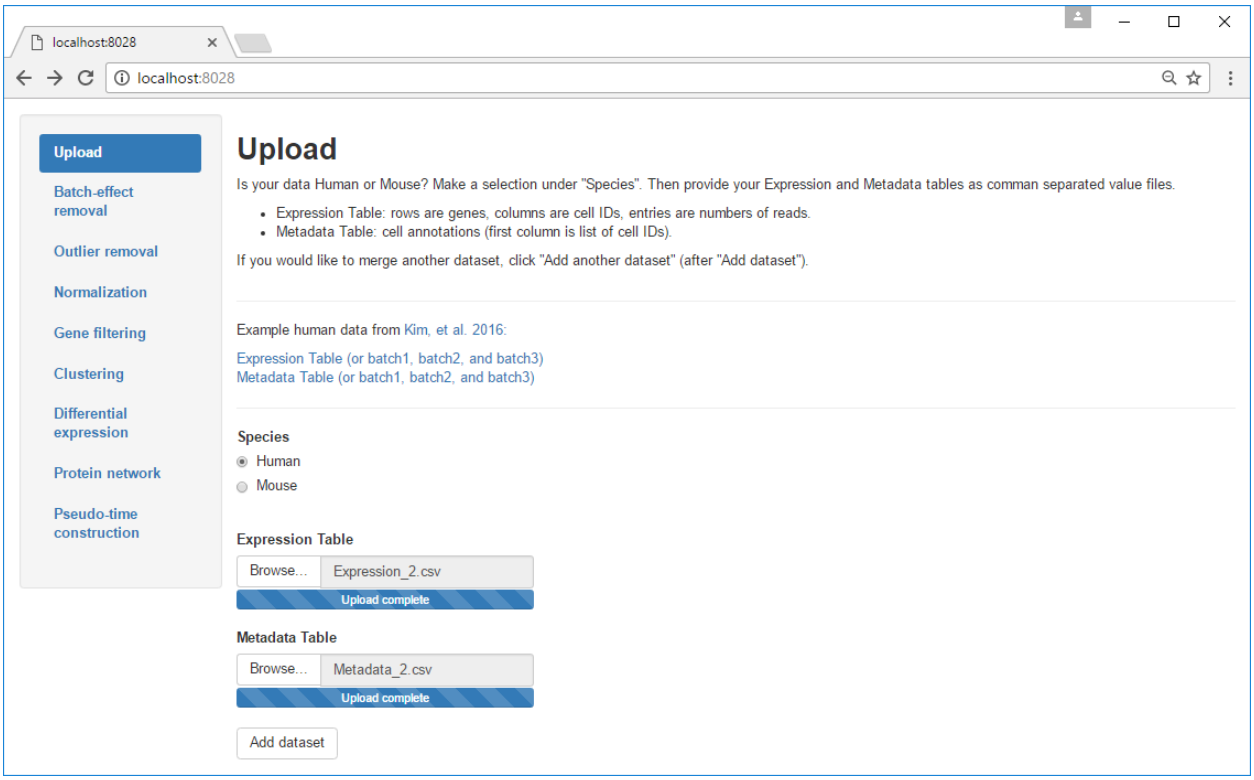
To input the data, first choose what species was sequenced (Human or Mouse), then select table files by clicking the "Browse" buttons. For this example we are using results from Kim, et al. [2], which will show a segregation between primary and metastatic renal cancer cells.



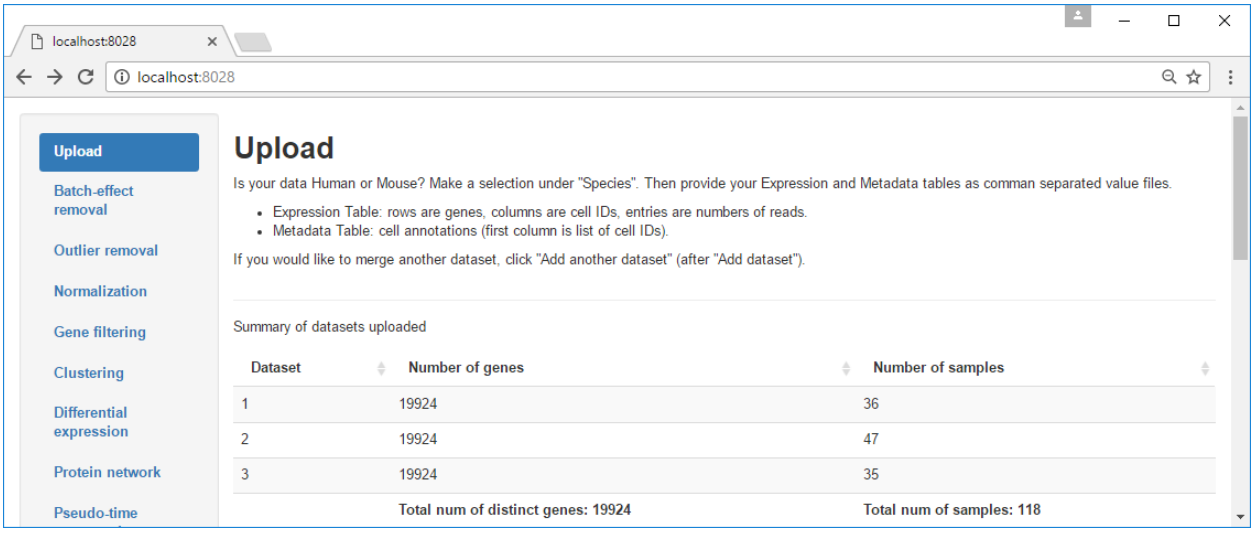
Here, we have split the dataset into three sets of files (ending _1.csv, _2.csv, and _3.csv) corresponding to different cell sources (patient vs. PDX and primary vs. metastatic). Once both status bars indicate “Upload complete” the “Add dataset” button can be clicked. This will bring you to a dataset preview page, showing the most recently uploaded data.



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.



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



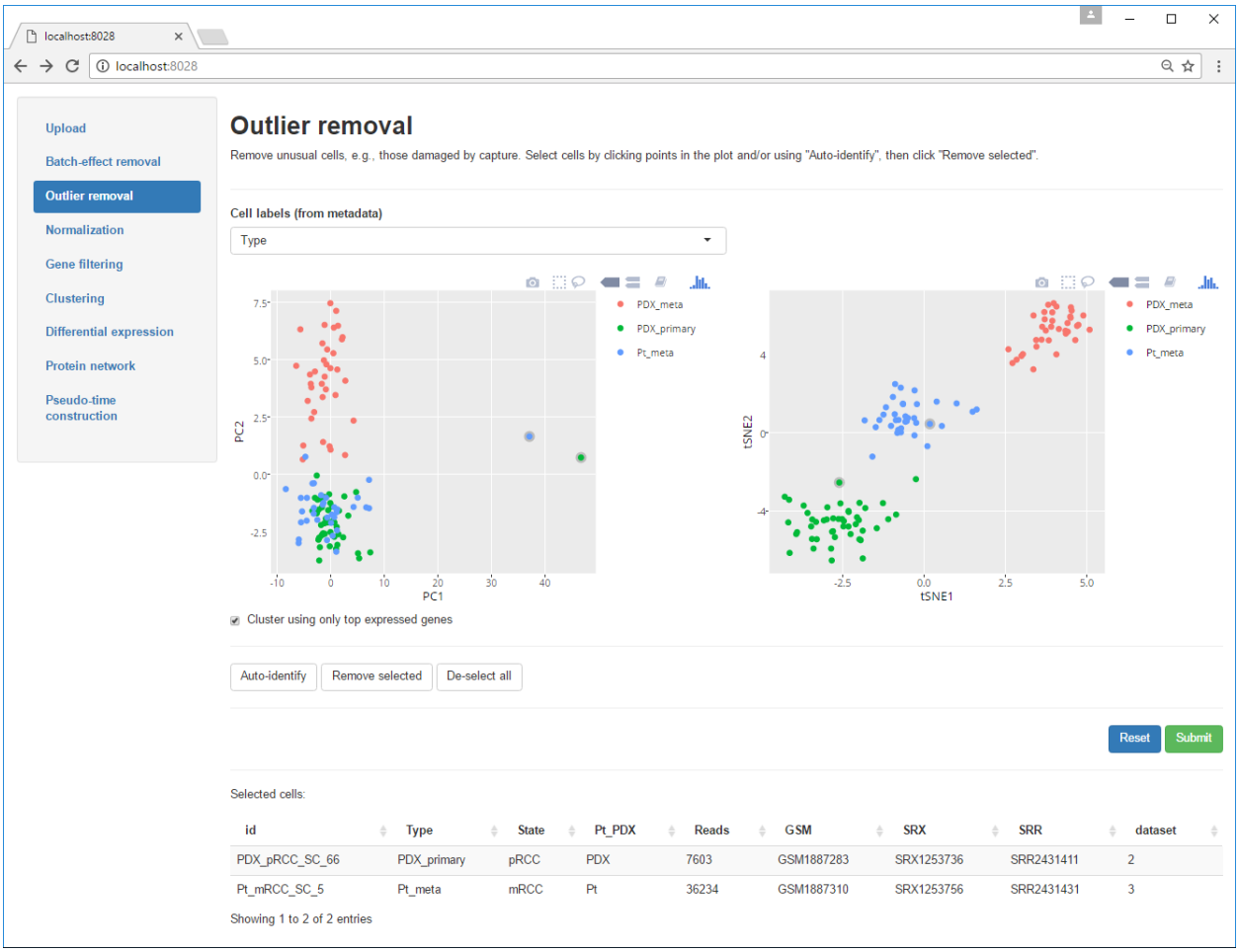
3. Batch-effect removal

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”.

4. 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’.

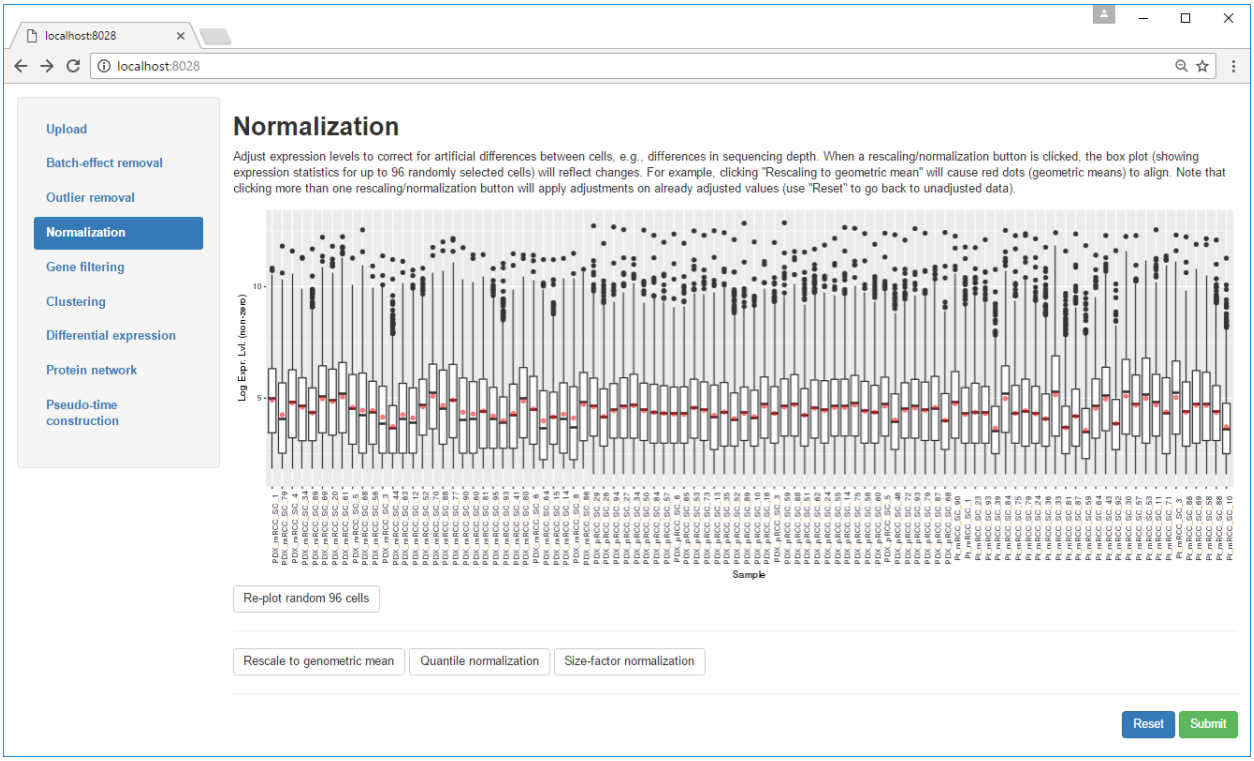


Cells can also be automatically removed using the “Auto-identify” button. Once cells to be removed are selected click “Remove selected” to remove them before proceeding.

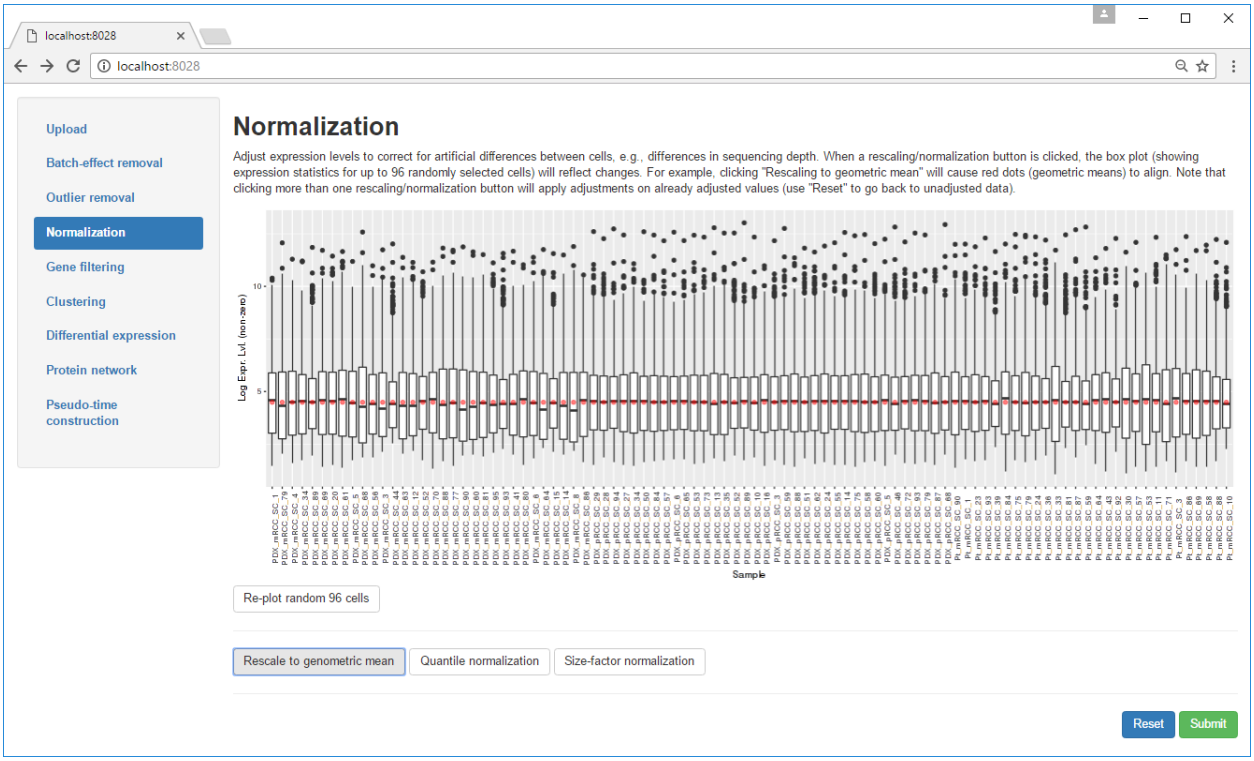
To reset to original graphs click “Reset”. Proceed by clicking “Submit”.

5. 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 genometric mean”, “Quantile normalization”, or “Size-factor normalization”. Click “Rescale to geometric mean” and the orange dots will align.



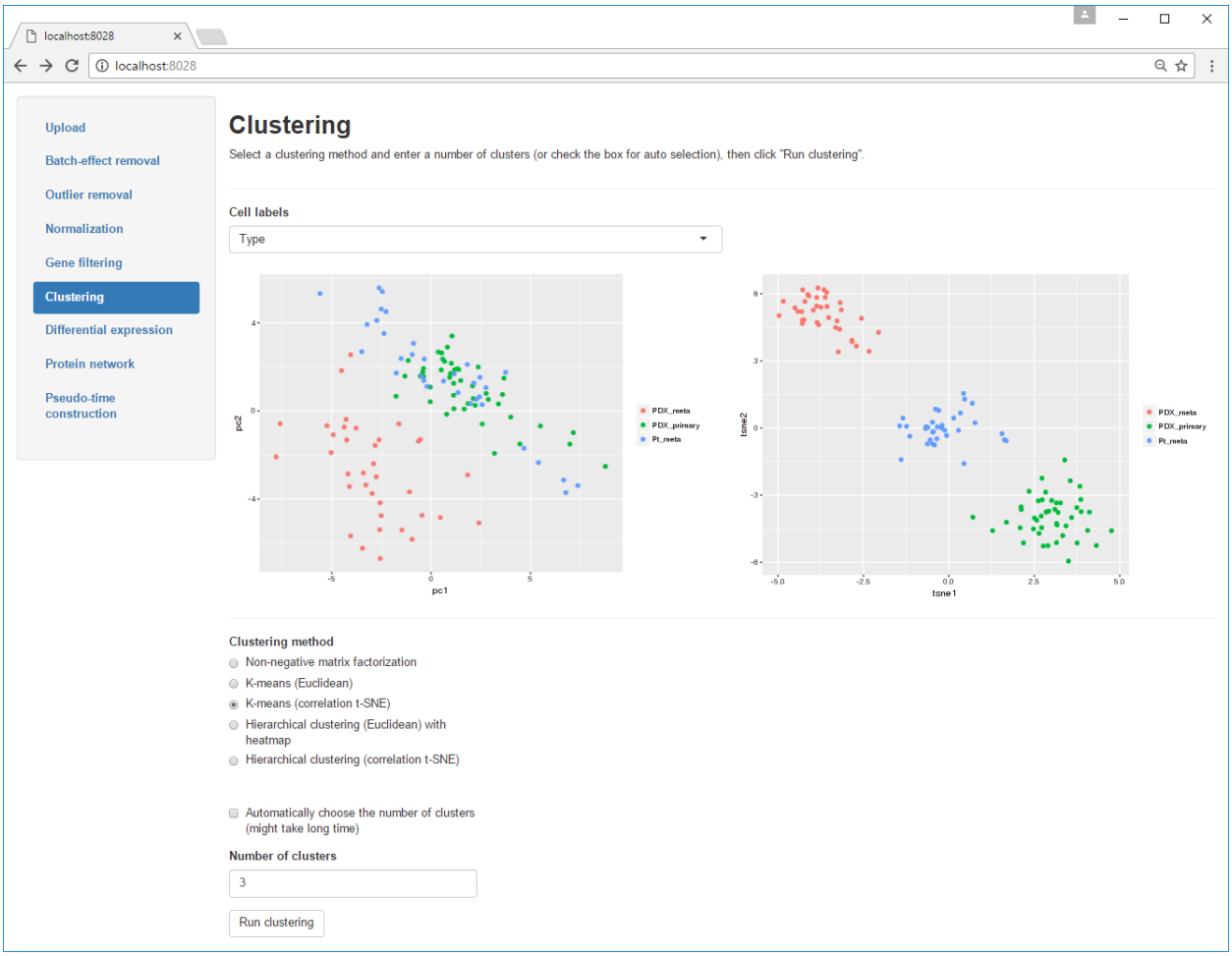
Click “Reset” if you would like to start with original input expression levels again. Click “Submit” to proceed to the next step with the (normalized) data.

6. 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.

7. Clustering

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

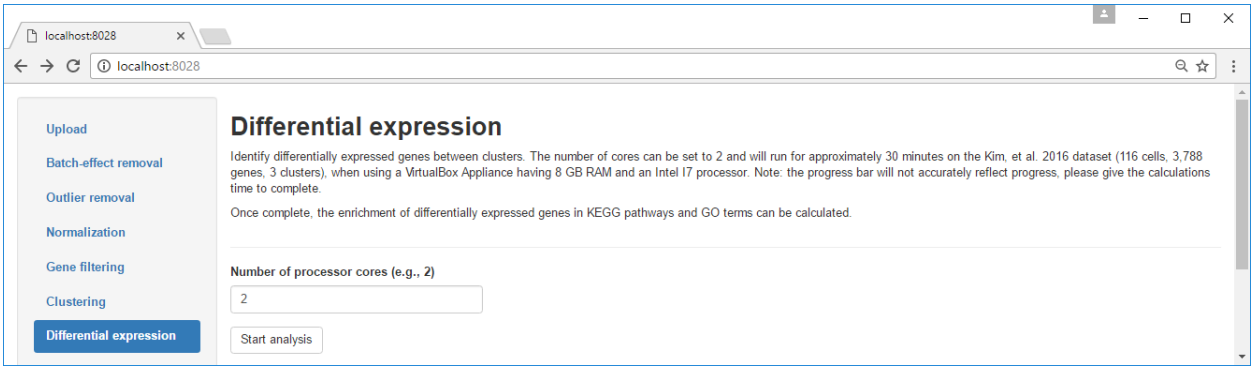


To calculate cluster assignments select a clustering method from the list, then choose to automatically estimate number of clusters by clicking the checkbox or enter a specific number before clicking “Run clustering”. Once clustering is completed, the cluster assignments are indicated by numbers within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

Click “Submit” to proceed to the next step.

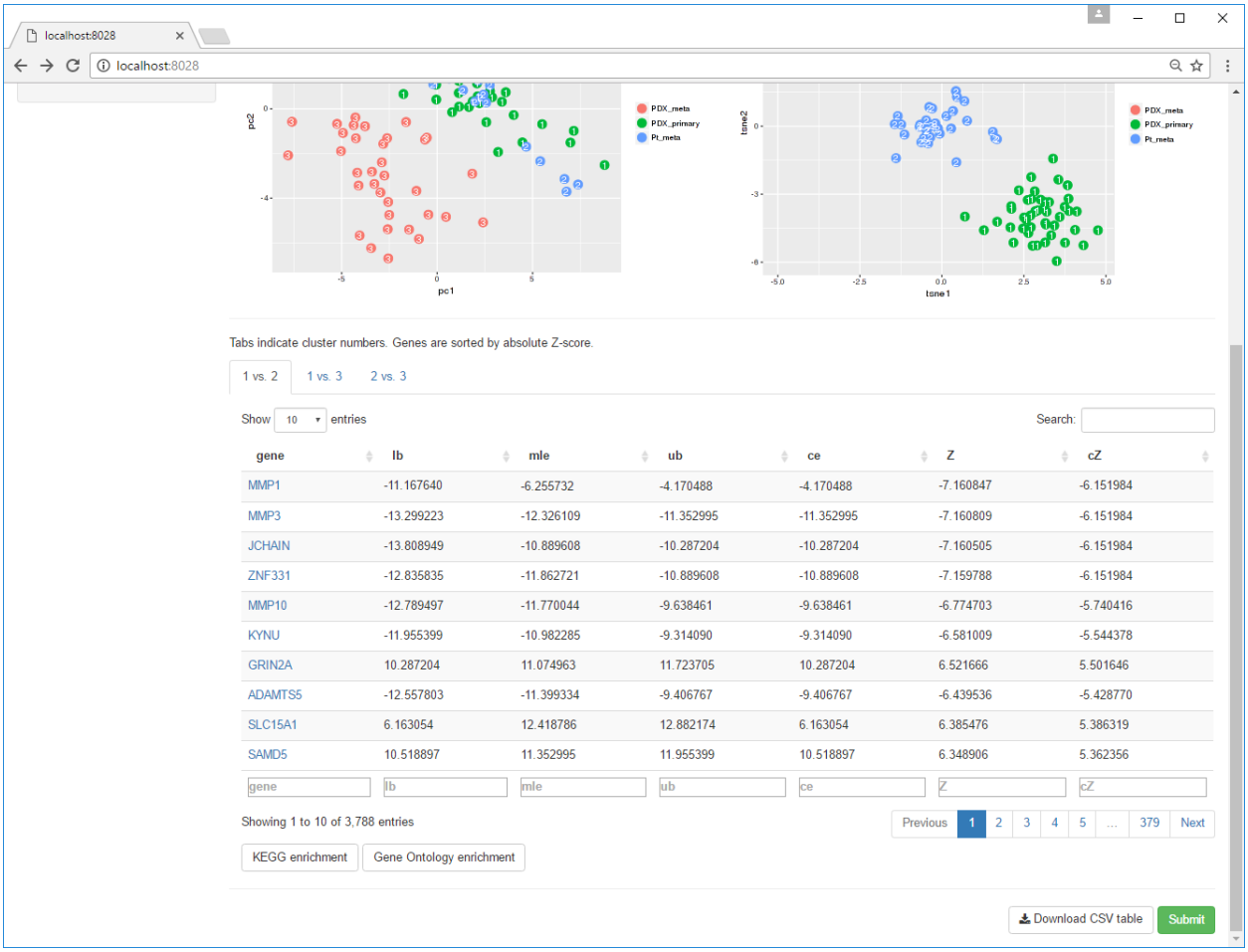
8. Differential expression

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.

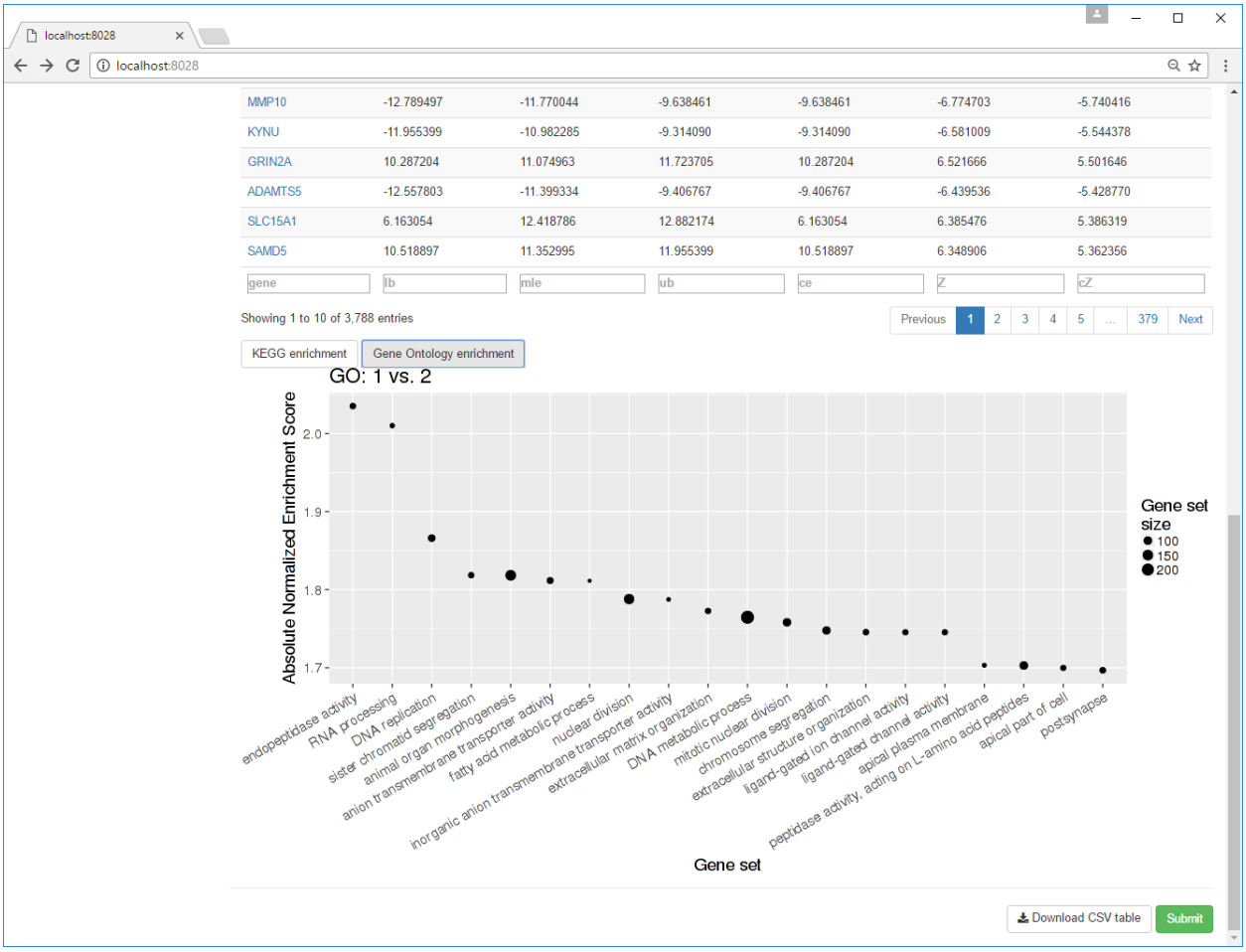


Click “Start analysis” to begin calculations. This step may take the most time, e.g., 30 minutes for three clusters using 116 cells and 3,788 genes. Once differentially expressed genes have been identified they

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.



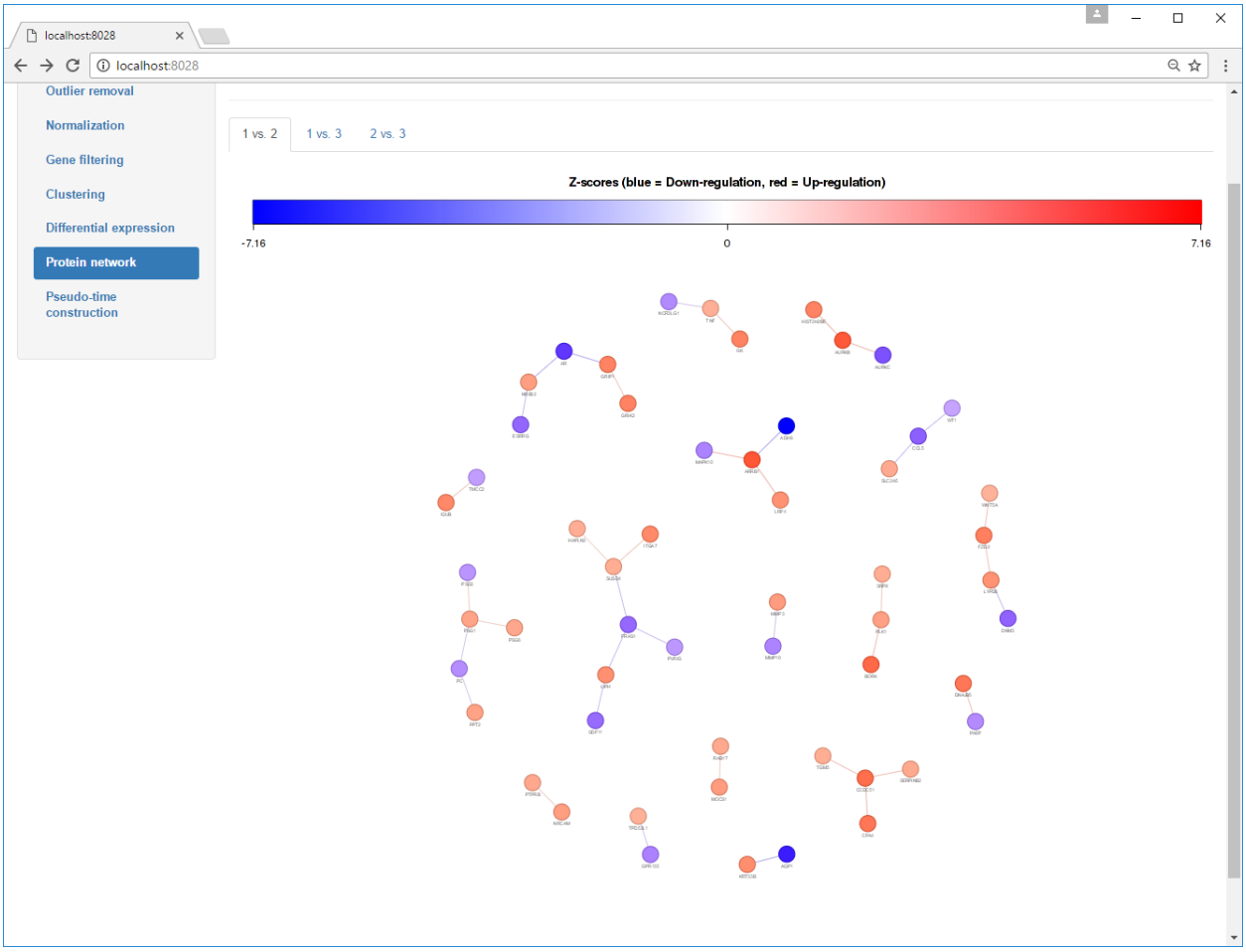
The enrichment of differentially expressed in KEGG pathways or Gene Ontology terms can be calculated for the selected tab by clicking the “KEGG enrichment” or “Gene Ontology enrichment” buttons, respectively.



Click “Submit” to proceed to the next step.

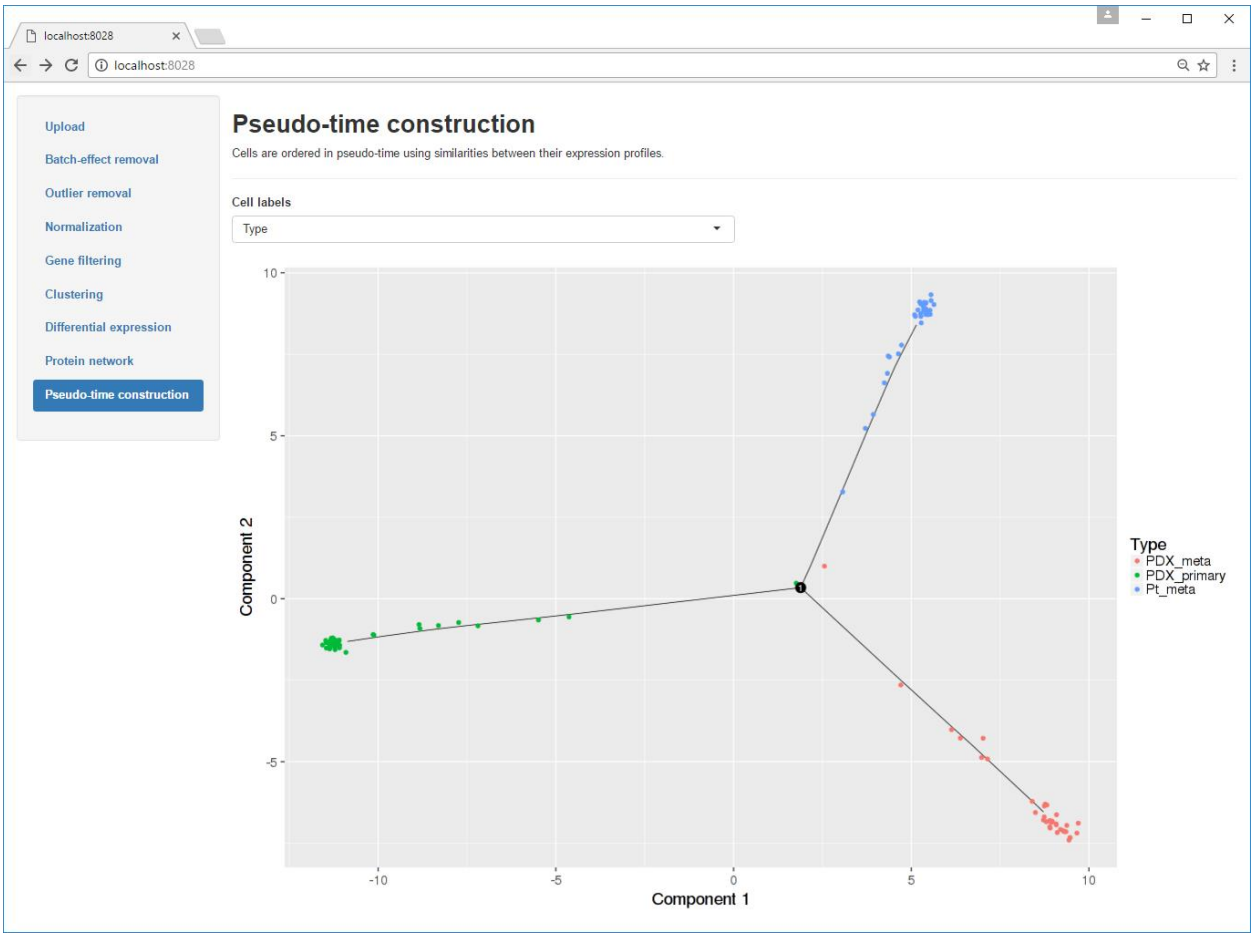
9. Protein network

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.



10. 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.



11. 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.