Single cell analysis

INTRODUCTION

Single cell RNA-seq (scRNA-seq) data sets can be significantly large and hence a challenge to analyze. Qlucore Omics Explorer (QOE) includes a suite of tools to assist in managing these challenges. This document aims at providing suggestions tips on how data can be analyzed and also minimize the impact of the large sizes. Even with an optimal approach, a faster computer with more memory is better. Specific hardware requirements are presented in the Qlucore system requirements.

Example one: The workflows and examples below were tested with a data set of 20 000 samples and 10 000 variables, on a laptop with 8 GB of RAM and an Intel core I7 processor.

Example two: For a dataset with 50 000 samples (i.e. cells) and 30 000 variables requires at least 24 GB of RAM.

LOAD THE DATA SET AND START THE ANALYSIS

If you have 10X Genomics Data, there are specific Templates that make import and normalization easy. Check the Documentation and Help Manager for information or read the use “How to Load 10x data” document. The templates offer various pre-filtering options for removal of noise.

For raw reads (counts) the Wizard can be used. There is extensive documentation including films on qlucore.com/resources covering the wizard. Note that we normally recommend setting the gene length to 1.

A very large data set will initially consume time in two ways. First the actual data loading time (read from disc) takes time, and it will be dependent on the speed of the hard drive and the size. *There is no other option but to wait.*

The second time consuming step is to generate the first data representation. For smaller data sets it will be a PCA plot, but for very large data sets a table will be displayed. This can take up to 30 seconds. *There is no other option but to wait for the first presentation to be ready.*

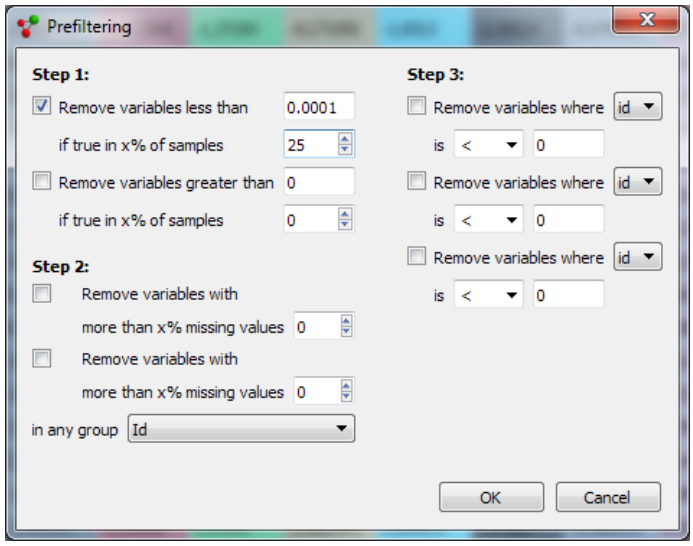
Overall, it is a good strategy to do data reductions and apply filters from the table view and then switch to plots for inspection and further investigation.

VARIABLE PRE-FILTERING

Variable pre-filtering is used to remove variables with certain characteristics. For single cell there will typically be many zeroes in the data. An example set-up to remove variables with zeroes (values less than 0.00001) in 25% of the samples is shown below. The variable pre-filtering is found in the Data tab.

Qlucore Sweden Ideon Science Park, Scheelevägen 17, 223 70 Lund.
Qlucore USA 900 Third Avenue, 29th Floor, New York, NY 10022.

The reduction of variables serves two purposes, it will remove variables with very few measurements and hence keep the variables that can tell us something about the samples, and it will reduce the data size. Think about it as removing noise, i.e. unwanted data at the same time as preserving the signal, i.e. the information of interest. To provide general guidelines on how much to remove and which cut-offs to use is difficult since it will be specific to each experiment set-up.



SUBSAMPLING

Subsampling is used to reduce the number of samples in a data set to reduce size. This can be used as a first or second step or not at all. The Subsampling functionality is found in the **File** menu.

If the purpose of the study is data mining, then it can be time saving to start with a sub sampled version of the data set and then findings or ideas are generated go back to the complete data set.

DIMENSIONALITY REDUCTION & CLUSTERING

Dimensionality reduction and clustering is done to simplify complex, high-dimensional single cell data for effective visualization and analysis while grouping cells into biologically meaningful clusters based on similarities in gene expression profiles. The typical workflow includes Linear Dimensional Reduction followed by Non-Linear Dimensional Reduction

Linear Dimensional Reduction (PCA Plot). Variance filtering can be useful to remove low-expressing or invariant variables that contribute little to the variance of the dataset.

By default, the first three principal components are plotted in a PCA plot to visualize the overall structure of the data revealing clusters, patterns and relationships between cells. The first three principal components usually capture the majority of the variance. Scree plots can be used to examine the percentage of variance explained by each principal component.

Non-Linear Dimensional Reduction. t-SNE and UMAP plots help visualize complex single-cell data by mapping it to the space where cells with similar expression profiles cluster together.

UMAP often preserves both local and global structures more effectively than t-SNE, making it particularly useful for capturing cell relationships and identifying distinct populations.

Note: UMAP and t-SNE are visualization methods that display the results of complex data transformations. These involve significant calculations, which can take time,

IDENTIFYING CELL TYPE

Identifying cell types is an important step to understanding the heterogeneity of tissues, organs or disease states at high resolution. Once clusters are formed using computational methods (e.g. t-SNE, UMAP), the GSEA workbench can be utilized to enrich known gene sets and map them back to clusters for cell type identification.

The first step in identifying cell type is to upload a suitable Gene Set List(s) for your single cell experiment to the Gene Set Lists File Path found in the GSEA Workbench.

Suitable Gene Set Lists for identifying cell types are composed of predefined gene sets that are assembled from known marker genes. Please note that each gene set should correspond to a particular cell type and a Gene Set List can be comprised of multiple gene sets. Predefined gene sets and Gene Set Lists can be downloaded from curated databases such as MSigDB, CellMarker or published literature that provides marker genes for specific cell types.

After suitable Gene Set List(s) have been uploaded to GSEA Workbench File Path, the following workflow can be utilized:

After Linear Dimensional Reduction, select 2D UMAP plot in Method tab, click Calc, create an Annotations called "*UMAP Cluster*" and annotate it with identified clusters (*Cluster 1*, *Cluster 2*, ...)

Launch GESA Workbench, select suitable Gene Set List(s) from the File Path, run Two Group (t-test) for *Cluster 1* against [All] other groups in annotation *UMAP Cluster*, Click Run

1. Identify the gene sets with the highest enrichment scores
2. Check the associated cell type for the gene sets with the highest enrichment scores
3. Color the UMAP based on the most relevant gene set with the highest enrichment score (Change name for *Cluster 1* to the cell type of this gene set in the UMAP Cluster" annotation).
4. Return to step 1 above and complete the for Cluster 2 and each remaining cluster in the *UMAP Cluster* annotation.

Validate cell type assignments with known biology, marker expression levels in the clusters (visualized using violin plots, heatmaps, etc.) or comparing with reference datasets.

KMEANS++ CLUSTERING

Works fine, but it takes some time.

USER INTERFACE RESPONSE TIME

When plots with a lot of information are displayed, they will require a lot of memory from the operating system. When you as a user would like to change content or move to a different user

control it can take some time for the system to free up memory. *You will sometimes have to wait for several seconds.*

GENERAL RULE FOR WHEN A PCA IS INTERACTIVE

PCA is a computationally intensive visualization. For a data set where the lowest value of the number of active samples or the number of active variables, is below 2000, then a PCA plot works with reasonable interaction on most modern computers.

> **Example 1:** an unfiltered data set with 20 000 samples and 1000 variables, works fine.
>
> **Example 2:** an unfiltered data set with 20 000 samples and 3000 variables, will be slow.
>
> **Example 3:** an unfiltered data set with 40 000 samples and 10000 variables that is filtered to 1500 variables using prefiltering, works fine.

GENERAL PLOTS

Heatmaps works fine on large data sets if hierarchical clustering is not used. Remember to uncheck the Auto buttons for the size of samples and variables. If Auto is selected the program will always try to draw the whole data set and it can take time.

It takes some time to generate a heatmap and if the size resolution (sample and or variable) is selected to something smaller than 1 it will take longer time.

Scatter plots are normally fast.