Single cell analysis tips & tricks

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 the challenges. To minimize the impact of the large sizes some tips & tricks are highlighted below.

The purpose of the tips & tricks is to enable work with very large data sets also on normal computer and laptops. A faster computer with more memory will be faster. Specific hardware requirements are presented in the Qlucore system requirements.

Example: The workflows and examples below were tested with a data set of 20 000 samples and 10 000 variables, on a 3 years old laptop with 8 GB of RAM and an Intel core I7 processor.

We recommend that you always use the latest version of the program. In December 2021 it is: Qlucore Omics Explorer version 3.8 (xxx).

LOAD THE DATA SET AND FIRST PLOT

If you have 10X Genomics Data there is a specific Template that makes import and normalization easy. Check the Documentation and Help Manager for information or read the use "How to Load 10x data" document.

All import options in QOE support very large data sets.

A very large data set will initially consume time in two ways. First the actual data loading time (read from disc) 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 30 seconds. *There is no other option but to wait for the first plot to be ready.*

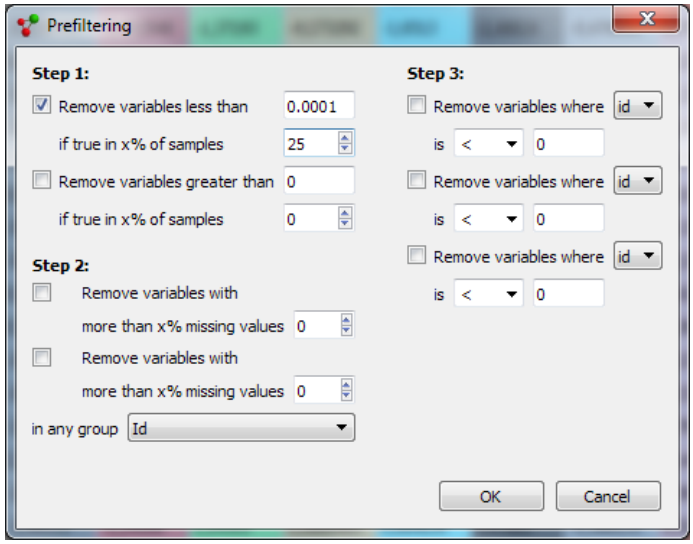
Overall it is a good strategy to do data reductions and apply filters from the table view and then switch to plots for inspection and further investigation.

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.

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 datamining, 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.

VARIANCE FILTERING AND STATISTICS

Variance filtering is useful both for large and small data sets to remove variables with little variance.

Statistical tools such as t-test or ANOVA works fine on very large data sets.

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 then 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 compute 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.

T-SNE AND UMAP
UMAP and t-SNE are plots that shows the result transformations. It includes calculations which are significant and can take time.

KMEANS++ CLUSTERING
Works fine, but it takes some time.

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.