

How-To-Seurat

What is Seurat?

Seurat is an extremely popular pipeline for analyzing single cell RNA Sequencing (scRNA-Seq) data. The R package is developed and maintained by the Satija lab.

The tool is capable of performing some simple quality control, dimensionality reduction, Louvain clustering and differential expression analysis for marker genes per cluster. We've also implemented some tools for automatically labeling populations based on simple phenotyping parameters, for certain model systems.

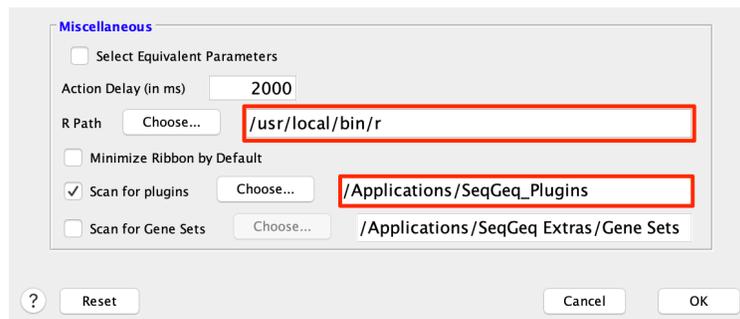
We've implemented this tool as a plugin in SeqGeq (v1.6+) in order to make these features available for SeqGeq users and simplify the process of producing results from the Seurat pipeline as much as possible.

Version 4.0.1+ of this SeqGeq plugin now includes 'weighted-nearest-neighbors' (WNN) analysis, an unsupervised framework to allow for an integrative analysis of multiple modalities, known as multimodal analysis. These different modalities can come from the proteome, transcriptome, epigenome, or genome. With the combination of this information, or modalities, from the same single cells we can now define cellular states based on multiple data types.

Please review our [video](#) and [documentation](#) for installing SeqGeq Plugins. 

Installation

Having downloaded the plugin, place it (the "Seurat.jar" file) into your SeqGeq Plugins directory. Install R version 4.0 (or higher), and set the R path in the Diagnostics section of SeqGeq's preferences:



Note: In the latest version of this plugin, all dependencies should be installed automatically within R. For troubleshooting, or in case this automated install does not function, the commands to install packages required for Seurat are:

```
install.packages(c('gplots', 'RColorBrewer', 'dplyr', 'data.table', 'Matrix', 'Seurat',  
                  'cowplot', 'reshape2', 'ggplot2', 'tidyr', 'sctransform', 'tools',  
                  'BiocManager', 'viridisLite'))
```

```
BiocManager::install('multtest')
```

The Seurat plugin version 4.0.1 (and higher) requires the Seurat package 4.0.4 (or higher).

Usage

Standard Seurat Pipeline

To run the plugin, select the population of interest, visit the Workspace tab of the SeqGeq workspace, and select **Seurat** from the plugins drop-down menu. You will then be presented with the main window where you can select from a Standard Seurat analysis pipeline or a Multimodal Pipeline. The standard Seurat pipeline will create one assay from the selected cells and features to perform QC(optional), normalization and principal component analysis, clustering and dimensionality reduction with UMAP or tSNE, differential expression analysis for the clusters found, and classification of clusters from selected model(optional).

We recommend to run the Quality Control steps directly in SeqGeq first to filter out outlier cells and dimly expressed features. Then use highly dispersed genes as the input features for Seurat. If you choose to perform QC in Seurat we recommend that all of the biological features measured are used as the input.

QC options in Seurat Selecting the QC check-box will open up additional options to perform some basic quality control steps. Entering a value in **%Genes** will exclude cells that are not expressing this percentage of features. Similarly, the **%Cells** will exclude features that are not being expressed in this percentage of cells.

Selecting the check-box next to **Library Filter** will apply a cutoff threshold to exclude outlier cells based on the 1st decile and 99th quantile of features expressed in each cell. Again, this step is best performed in SeqGeq where you can visually gate on events.

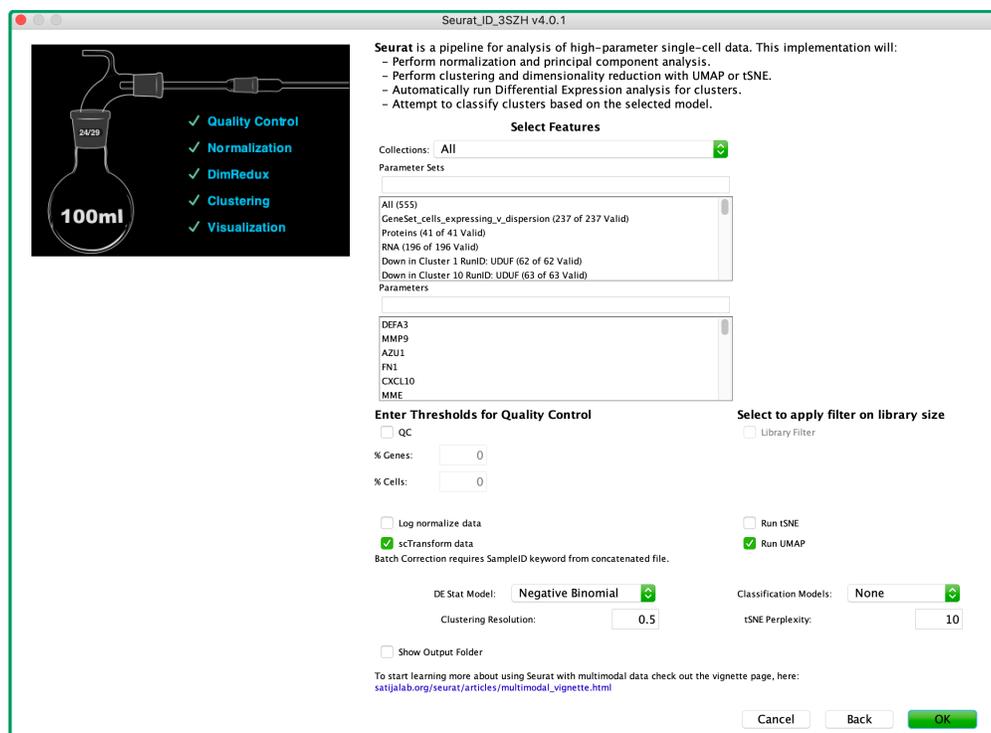


Figure 1: Seurat standard pipeline UI

There are two methods for normalization within the Seurat plugin. You can select the check-box next to **Log normalize data** to perform log normalization in Seurat where feature counts for each cell are divided by the total counts for that cell and multiplied by a scale factor of 10000. This result is then natural-log transformed. The second method, **scTransform data**, is the recommend method for normalization and variance stabilization of molecular count data. This method uses “regularized negative binomial regression” and has been shown to better represent data that are no longer influenced by technical variables, and still

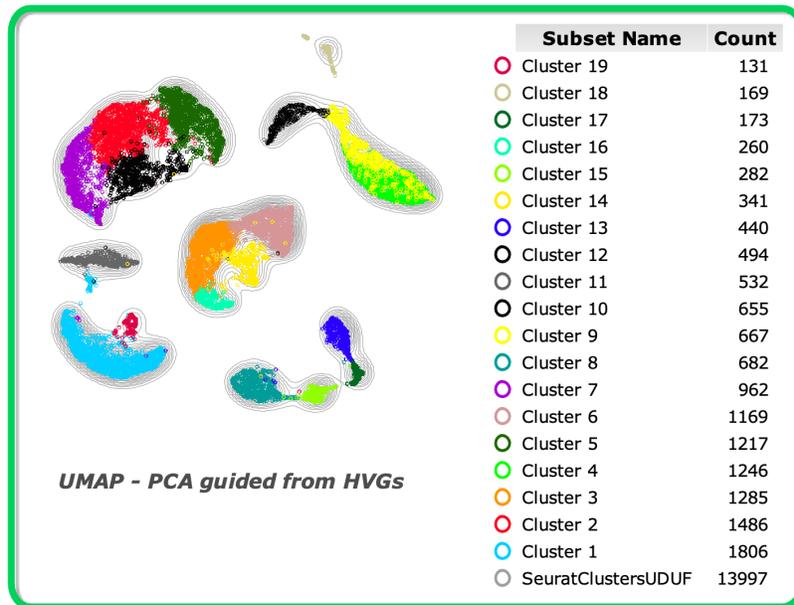
retain heterogeneity within distinct biological states after running `sctransform`. You can read more about this method here: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1874-1>

Select a dimensionality reduction option by selecting the check box next to either **Run tSNE** or **Run UMAP**. When complete, you will see an SeqGeq graph window showing the dimensionally reduced space.

Choose a test method for determining differentially expressed features between each cluster by choosing one from the drop-down menu next to **DE Stat Model**. If your data represents one of the models available you can select the model from the **Classification Models** drop-down menu in order to automatically label populations. You also have control to change the **Clustering resolution** by entering a value from 0.1 - 10. Larger numbers here will result in more clusters being returned from the Louvain clustering. Optimal resolution often increases with larger data sets.

Selecting the box to **Show Output Folder** will open the folder containing the output from running the R script. This is helpful when troubleshooting as a text file is generated that contains the output from running the R script which may contain the failure message.

The plugin will need to run for a little while, after calculations are complete it will generate a pair of tSNE and/or UMAP derived parameters within your data matrix, as well as a set of graph based clusters (and corresponding categorical parameter). Finally the plugin will produce a collection of gene sets representing the top most differentially expressed genes from each cluster.



Multimodal Analysis

The multimodal pipeline is an exciting and new feature to the plugin and available if you have R-4.0.1+ and Seurat 4.0.4 and or later. The multimodal Seurat pipeline will create two separate assays from the selected cells and features to perform QC(optional), normalization and principal component analysis, clustering and dimensionality reduction with UMAP or tSNE, differential expression analysis for the clusters found, and classification of clusters from selected model(optional).

The two modalities are used together to give more power to the analysis and help identify clusters that may be missed when just using one modality, such as RNA features alone.

Before starting the plugin, create two separate gene sets in SeqGeq to identify the RNA and antibody measurements. You can create a new static gene set in SeqGeq and use the search feature to make two unique gene sets where one contains all of the protein measurements and another which does not contain any

of them. When starting the plugin you will notice the left parameter selector is used to choose the RNA features and the right parameter selector is used to select the protein measurements.

The remaining options in the plugin are the same as the standard pipeline. You can choose to perform QC in the Seurat pipeline or use SeqGeq. We recommend to run the Quality Control steps directly in SeqGeq first to filter out outlier cells and dimly expressed features. Then use highly dispersed RNA genes and all of the proteins as the input features for Seurat multimodal analysis. If you choose to perform QC in Seurat we recommend that all of the biological features measured are used as the input.

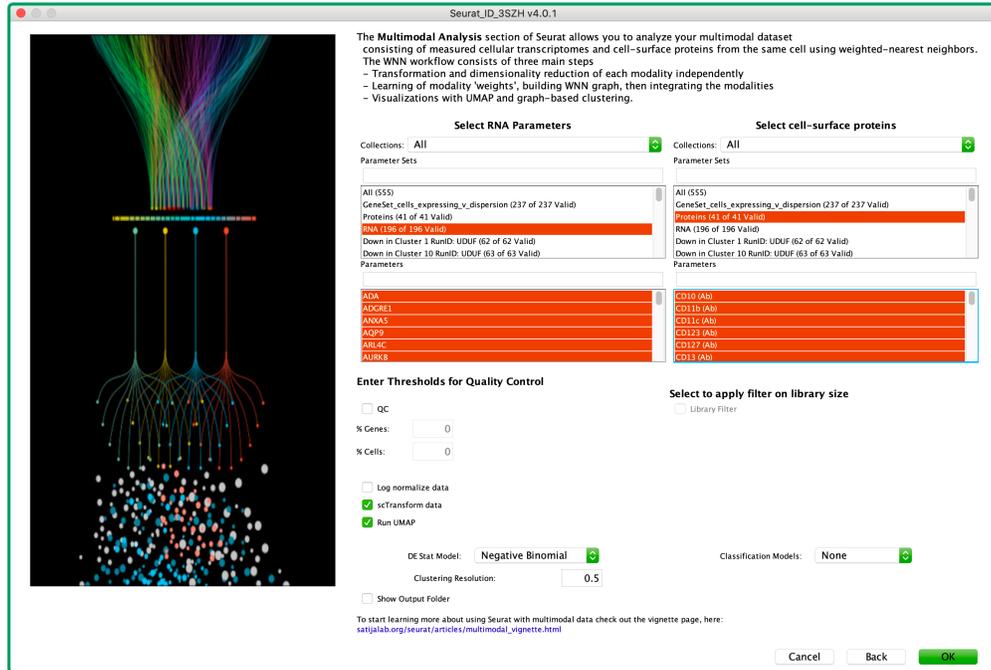
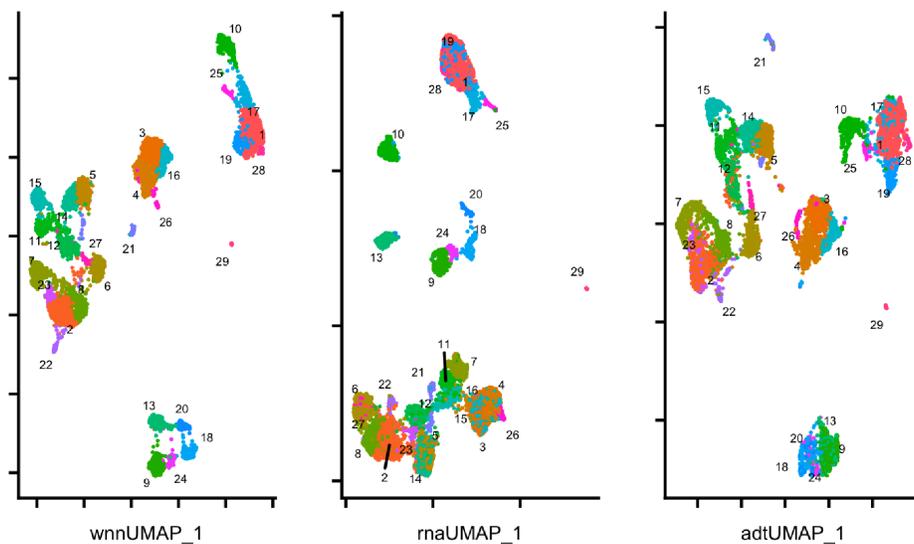


Figure 2: Multimodal analysis pipeline UI.

You may notice that the results return more clusters than if you run the Seurat pipeline with RNA features alone and vice versa. This is thanks to weighted-nearest neighbor analysis where the k-nearest neighbors of cells for each modality are identified, and expression levels of neighbors are compared for within, and across modalities. This can help to identify which measurements may benefit from additional weighting to help improve cell clustering and classification in downstream steps. For each cell, weights are calculated for each modality. The weights indicate how important each modality is for defining cellular identity. With these learned weights, the modalities can be combined to help identify and cluster cells.

This weighted information is then presented in a weighted-nearest neighbor UMAP graph with improved identification of cell states. The three UMAP plots below show the differences between a WNN UMAP, RNA UMAP, Protein UMAP. The weighted-nearest neighbor analysis shows improved identification for some subsets that are not resolved when looking at the transcriptome or proteome alone. The WNN analysis can help represent the abundance of information collected from both modalities.



If you have any questions or concerns regarding this or any other plugin for SeqGeq, we would welcome your feedback: <mailto:seqgeq@bd.com> 

References:

1. Source (Seurat Github): <https://github.com/satijalab/seurat> Seurat is developed and maintained by the Satija lab, and is released under the GNU Public License (GPL 3.0)
2. Hao Y, Hao S, Anderson-Nissen E, Gottardo R, Smibert P, Satija R, et. al 2021, "Integrated Analysis of multimodal single-cell data", *Cell* vol 184, no. 13, pp. 3573-3587.E29 <https://doi.org/10.1016/j.cell.2021.04.048>