

How To ClusterAutoClass

Introduction

Clustering and particularly unsupervised clustering, is more and more popular for informatics processes dealing with high feature content data (lots of parameters). Classification of such datasets has become more and more challenging, leading to a strong desire for automated classification techniques. A variety of options exist for such analyses. Different classification options are suited to a variety of different data types and classification problems. This plugin currently focuses on a very simple approach to classification which is customizable, easy to understand, reproducible, and extensible to any clustering algorithm which produces a categorical parameter.

ClusterAutoClass is a plugin for FlowJo and for SeqGeq, which uses researcher selected hallmark features to take an educated guess regarding the class of cluster populations, and automatically applies these labels (aka “annotations”) to a new set of identical cluster subpopulations. It does this by recursively testing the expression of each hallmark parameter with a cluster and across clusters to find outlier high expressors of the parameters. In doing so, ClusterAutoClass can automatically classify subpopulations even if the hallmark expression is not uniform, or if the hallmark expression is multi-modal (expressed in more than one cell class). The classification algorithm is also capable of calling a cluster “UNKNOWN” for clusters whose ID is ambiguous either because of a collision of hallmark features, or an insufficient brightness of signal.

Limitations

The current iteration of ClusterAutoClass is limited to classification based on positive/bright signal from a *single* hallmark parameter per class. The plugin also requires that researchers select a parent population of clusters where each cluster subpopulation is gated ahead of time, and based on a categorical parameter. Because ClusterAutoClass is based on previous clustering results, the annotations can never be more accurate or more nuanced than the clustering algorithm’s results allow, and poor clustering will likely lead to poor classification results.

Best results will require some expression of all hallmark genes within the data matrix. A “good” model’s class-hallmark pairing will represent all major cell types within the dataset, without relatively deep or nuanced subsetting. Typically a successful model will contain 5-12 classes of cell types depending on the panel of parameters available and heterogeneity of a given dataset.

Set Up

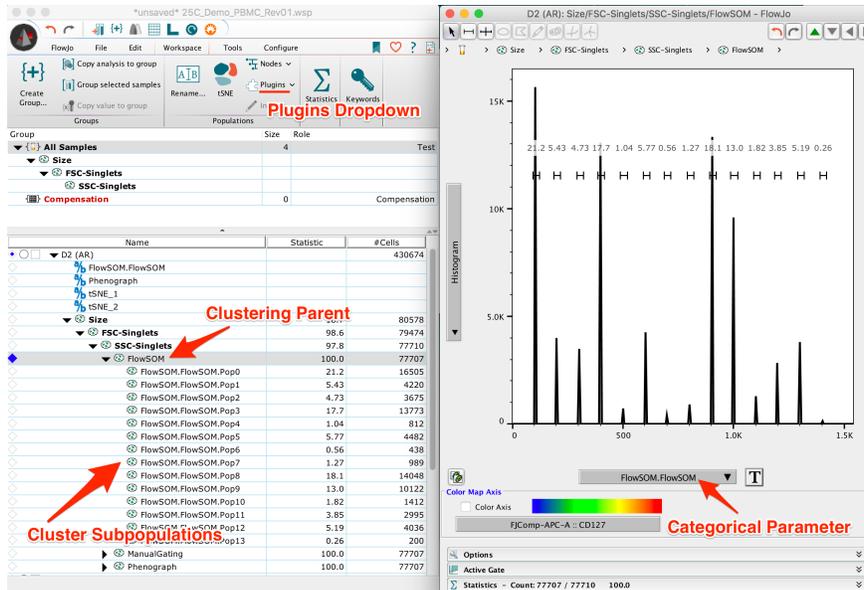
There are general instructions and tips on how to set up plugins in FlowJo’s [technical documentation](#).

This plugin can installed using the following steps:

1. Place the JAR file into a “plugins” folder.
2. Ensure that the plugins folder path is correctly set within the Diagnostics section of FlowJo or SeqGeq’s Preferences (heart icon in the Workspace)
3. Restart the application.

Usage

Select the parent of clustering subpopulations. Go to the workspace tab of your workspace, select the plugins dropdown, and choose ClusterAutoClass from the dropdown.

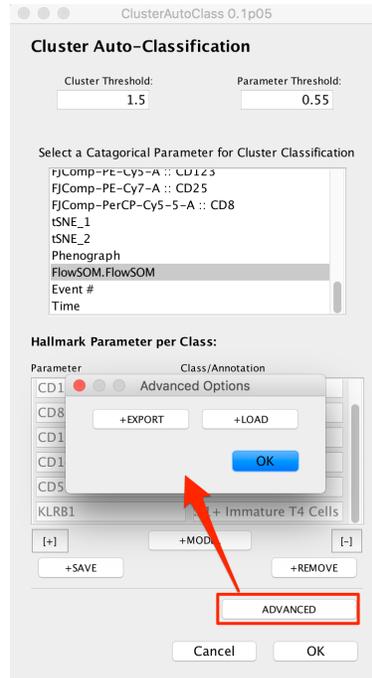


A window will appear prompting researchers to select their categorical parameter, which defines the clusters of interest, and to create a model for their cells of interest. Within this dialog: - the first section sets thresholds for unknown clusters based on outlier hallmark expression - next researchers will need to select their clustering categorical parameter - finally parameter/model pairs can be entered or edited



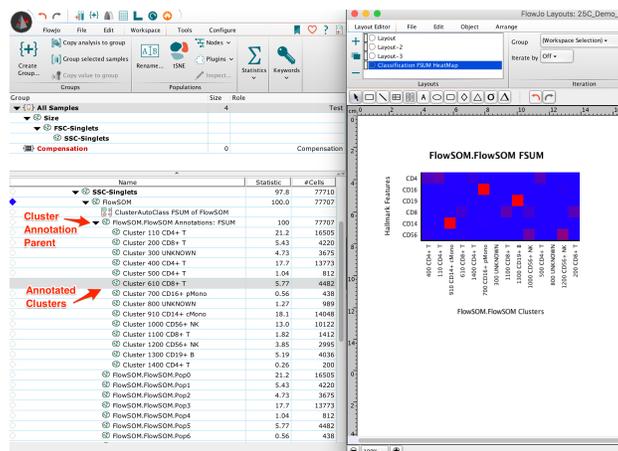
Note: Default models are included with the plugin for some specific parameter panels and sample types. Models generated within the plugin can be saved for later or repeat use, or deleted in case they are found to be irrelevant or erroneous.

At the bottom of the dialog is an advanced section where researchers can save their current list of models for sharing with others, or load model library JSON files from others.



Interpreting Results

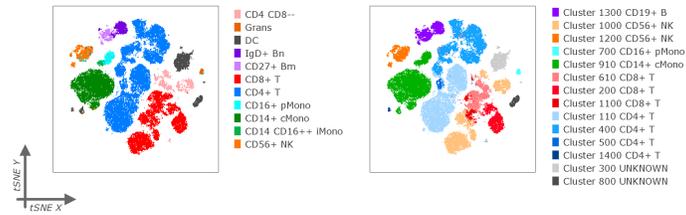
Once classification is completed a new parent population should be created along with annotated cluster subpopulation predictions. A heatmap illustrating normalized expression of hallmark features across clusters will be added to the Layout Editor. Researchers should be careful to confirm subpopulation identification independently.



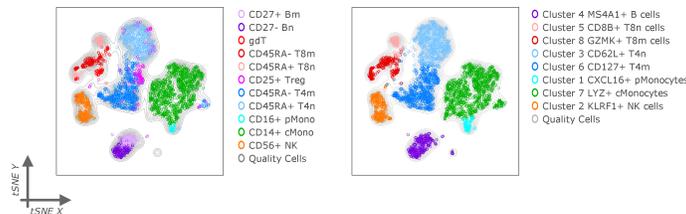
For many datasets the same class may be attributed to many clusters. In order to facilitate deeper sub-classification, researchers may choose to combine such clusters using Boolean “OR” gate logic, and then recursively re-run the plugin using a model specific to their grouped phenotype.

Outputs and Comparisons

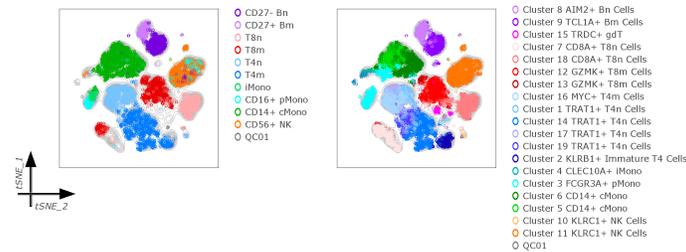
Results from clustering followed by AutoClassification using a default built in model, and manually annotated subpopulations in FlowJo using a 25 color panel of flow cytometry data on human PBMCs (below).



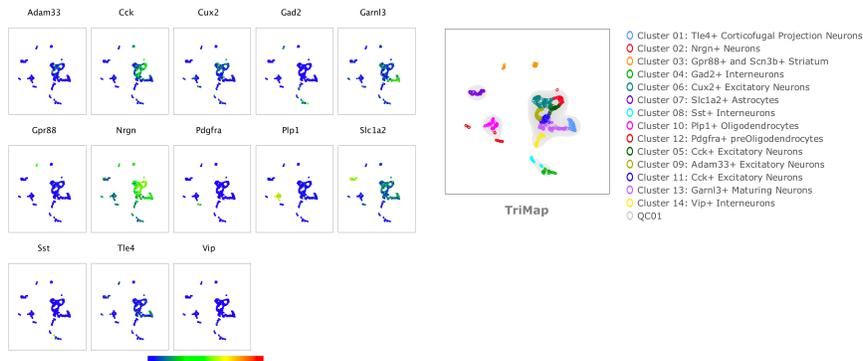
Similar to above, comparison of oligo-tagged antibody plus whole transcriptome (WTA) sequencing data in SeqGeq using a built in “default” model (below).



Comparison of oligo-tagged antibody WTA data concatenated with targeted sequencing plus *BDTM* AbSeq sequencing data (below).



Hallmark geneset relative to cell types developed as a default model from extensive literature search, and expert analysis of sorted mouse neuron nuclei WTA data (below).



Support

Please write to flowjo@bd.com with any questions.