FLAMEMetacluster Documentation

Module Name: FLAMEMetacluster
Description: Match populations/clusters across all samples in the dataset, assembles a table of all parameters in all samples
Author: Xinli Hu (Broad Institute), gp-help@broad.mit.edu

Summary:

FLAME (FLow analysis with Automated Multivariate Estimation) uses finite mixture model clustering techniques with novel algorithms and models to define and characterize discrete populations in flow cytometric data [1]. A pipeline of GenePattern modules implements the method: FLAMEPreprocess, FLAMEMixtureModel, FLAMEChooseOptimalClusterNumber, FLAMEMetacluster, FLAMEContourDataGenerator, and FLAMEViewer.

FLAMEMetacluster module takes the results from FLAMEChooseOptimalClusterNumber, where each sample has been clustered into subpopulations using the optimal number of clusters, and matches the subpopulations so that a given population can be identified uniformly across all samples. Cluster results for FLAMEMetacluster are similar in format to cluster results for FLAMEChooseOptimalClusterNumber; however, in the FLAMEMetacluster results, each population/cluster is represented using a consistent color and position across all samples.

When samples in a dataset belong to different classes, the module performs "metaclustering" in two steps: within-class metaclustering and cross-class metaclustering. When all samples belong to a single class, only the first step (within-class metaclustering) is used.

Within-class metaclustering: If samples in the dataset belong to more than one class (i.e. phenotypic class, healthy/control status, treatment groups, etc.), matching within the class is carried out. The module first constructs a "template" sample, by pooling the cluster centers of all samples in the class, and clustered to determine the "typical" cluster centers of this class. Then each sample is compared and aligned to the template individually in a bipartite manner. The module works under the assumption that all samples belonging to the same class have similar biological characteristics, therefore are expected to have similar populations. The class membership of samples should be provided in a two-column text file.

Cross-class metaclustering: The two "typical cluster centers" templates from both classes are matched in a bipartite manner. Cluster membership assignments for all samples are reassigned to reflect final, matched clusters. If all samples in the set belong to one class, cross-class metaclustering is not performed.

This module depends heavily on a well quality-controlled set of data, where outliers are eliminated and technical variation across samples within each phenotypic class is well minimized.

References:

1. Saumyadipta Pyne, Xinli Hu, Kui Wang, Elizabeth Rossin,Tsung-I Lin, Lisa M. Maier, Clare Baecher-Allan, Geoffrey J. McLachlan, Pablo Tamayo, David A. Hafler, Philip L. De Jager, and Jill P. Mesirov. (2009). Automated High-dimensional Flow Cytometric Data Analysis. **citation?**

Parameters:

Name	Description
optimal g mixture model	A .zip file containing the optimal mixture modeling result of each sample. This is the output of the FLAMEChooseOptimalClusterNumber module.
sample class	A two-column text file, where the first column contains sample names (without .fcs or .txt filetype appended), and the second column the corresponding class names. There should be one header row.
estimation increment	Used only for skew distributions. Step is used to calculate the number of iterations of the Expectation-Maximization(EM) algorithm to perform when computing the Maximum Likelihood(ML) estimate. ML estimation is used to obtain the optimal parameters of the parametric mixture model distribution for each cluster. The value must be greater than 0 and less than 1; the default value is 0.5. The smaller the increment, the more accurate the estimation, but the slower the estimation step. Must be the same value used in the FLAMEMixtureModel.
output intermediate results	Choose whether to output the intermediate metaclustering results, such as within-class matching results. (default:no)
output prefix	A prefix for output files.

Output Files:

1. A zip file containing
   a. Final clustering results for each sample at its optimal cluster number, where each population/cluster is represented using a consistent color and

position across all samples. (The FLAMEMixtureModel documentation describes the clustering results files.)
b. a .gct file (table) containing all/features parameters of all samples. Each row is one feature and each column is one sample.
c. a .cls file that provides the matching class information for the .gct file.
d. a .xls file containing all/features parameters of all samples. Each row is one feature and each column is one sample.

The Features are:

prop(n)	proportion of cells in cluster n
mus(m.n)	location parameter of cluster n in dimension m
vars(m1m2.n)	covariance of cluster n along dimensions m1 and m2
alpha(m.n)	skew of cluster n in dimension m
scale(n)	scale in cluster n

2. A zip file containing all intermediate metaclustering results, such as within-class matching results.

Example Data:
The example data is a subset of the data described in Pyne et al. (2009) [1]:

Input parameter	**Value**
optimal g mixture model	Output from FLAMEChooseOptimalClusterNumber: ftp://ftp.broad.mit.edu/pub/genepattern/example_files/FLAME/SMALLphospho.OptimalG.zip
sample class	ftp://ftp.broad.mit.edu/pub/genepattern/example_files/FLAME/SMALL_phospho.AllSamples.txt
estimation increment	0.1
output intermediate results	no

Error Messages:

The FLAME module may report informational “errors” and generate an stderr.txt file. Serious errors end with the phrase “Execution halted.” Messages that do not end with that phrase can be ignored.

Platform dependencies:

Task type:	FLAME
OS:	Any
Language:	R (2.7.0 or later)