BFRMNormalize

Jeffrey T. Chang, Michael L. Gatza, Joseph E. Lucas, Quanli Wang, and Mike West

18 October 2010

Summary: This module normalizes the expression values of one or more gene expression data sets. If multiple data sets are given, it will also merge them into a single, clean data set. Briefly, it first merges all data sets and identifies (using a principal component analysis) structure in the Affymetrix control probe sets that represent technical variation, or noise, across the merged file. It then uses BFRM to remove the contribution of the noise principal components from the expression profile of each gene.

This algorithm works only on Affymetrix data sets, and the control probe sets (that start with *affx_*) should not be filtered.

For more information on BFRM, see:

> Carvalho C, *et al*. High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. *Journal of the American Statistical Association*. 103:1438-1456, 2008.

For more information on its use in merging files, see:

> Gatza *et al*. "A pathway-based classification of human breast cancer." *Proc Natl Acad Sci USA*. 107(15):6994-9, 2010.

Parameter: num_factors

This determines the number of principal components of noise to remove from the merged data set. The higher the number here, the more aggressively the module removes noise. However, useful signal may also be removed. It may be necessary to try different values to optimize the noise removed without cutting into the signal. As a general default, we recommend 15.

Parameters: file<*num*>

These are gene expression data sets to be merged. They can be given as PCL, GCT, and some other data formats.

For help, please email: jeffrey.chang@duke.edu.