NormalizeColumns Documentation

Description: Apply a (quantile, scale or Znorm) normalization on the columns of a GCT file.

Author: Dr Mark Cowley (Garvan Institute of Medical Research, Sydney Australia), m.cowley@garvan.org.au

Kevin Ying (Garvan Institute of Medical Research, Sydney Australia), k.ying@garvan.org.au

Date: September, 2011

Release: 4.3

Summary

Apply a (quantile, scale, or Znorm) normalization on a GCT file.

Some normalization advice:

First, your data should already be log-transformed. See the LogTransform module. Second, always look at plots of the distributions of your data before choosing a normalisation method. This GenePattern module produces a picture file containing these plots BEFORE and AFTER normalising.

The boxplots can reveal gross differences in the distributions of the arrays, and the density plots can reveal a little more detail. If each array looks very different, then you will need to use a more extreme form of normalisation. We recommend starting with scale normalisation; if the scale-normalised data look quite different still, then try just quantile normalising. NB, it's not a great idea to normalise, normalised data; so don't scale-norm, and then qnorm this scale-norm'ed data.

Scale normalisation involves scaling the expression levels on each array to have the same median-abolute-deviation (MAD) across arrays. After scale normalisation, the boxes in a boxplot will all line up, but the tails of the distributions may differ. This is appropriate to use if the data have roughly the same shaped distributions, but perhaps the midpoint of the boxes and or the width of the boxes differ between arrays. If your data distributions look very different before normalisation, then Quantile norm may be more appropriate. The scale normalization method was proposed by Yang et al (2001,2002) and is further explained by Smyth and Speed (2003).

Quantile normalisation ensures that the expression levels on each array have the same empirical distribution across arrays. After quantile normalisation, all arrays will have IDENTICAL distribution. It does this by averaging all the distributions together & then forcing each array to fit that distribution. It's appropriate to use when data are quite varied across each array. It's a fairly extreme form of normalisation and has been known to reduce any subtle but true differences that may exist between arrays. Quantile normalization was proposed by Bolstad et al (2003) for Affymetrix-style single-channel arrays and by Yang and Thorne (2003) for two-color cDNA arrays.

Znorm (aka standardization) standardizes the expression levels on each array so that they have a mean=0 and standard deviation=1. This is quite popular in Boston. After Znorm, the mean of each array should line up around 0, and they should have similar box widths. This is a pretty good normalisation to use if you want to combine data from multiple experiments together.

Usage

Supply a file to be normalized as an “input file” and select a normalize method.

Open “plots.png/pdf” to see box and density plots of the original and normalized data.

All missing data points are ignored.

References & Links

Parameters (* = required)

Name	Description
input file*	The GCT file to be normalized
output file*	The name of the normalized GCT file to write out. Default: <input.file_basename>_norm.gct
normalize method*	Normalization method. Default: scale

Input Files

1. **input file**
   A GCT file to be normalized

Output Files

output_norm gct
The normalized GCT

plots.png
Box and density plot of original and normalized GCT

plots.pdf
(This will be exported if the server could not export a png)

stdout.txt
A useful file containing information about the run.

Warning/Error Messages

The first time the module is run, limma libraries are likely to be compiled. This may give error or warning messages, which can be ignored.

"Warning: could not generate png. Contact your system administrator"

The version of R running on the server might not have png support. You might need to recompile with cairo or Xlib support.

In the case that a png could not be generated, a pdf is produced instead.

Example Data

ftp://ftp.broadinstitute.org/pub/genepattern/datasets/all_aml/all_aml_test.gct

Citing this module

Cowley, M.J., Ying, K., *NormalizeColumns* – a GenePattern module for applying a normalization on the columns of a GCT file (not published).

Platform Dependencies

Module type:	Preprocess & Utilities
CPU type:	any
OS:	any Tested on Ubuntu 10.10,
software	*limma (Bioconductor), Xlib (or xvfb) or cairo support*
Language:	R 2.12