NL transformation package tutorial

Transforms initial low sequence depth RNA-seq data using reference high depth RNA-seq data from gene-based pooling of all samples together

Usage

NormTable <- Main_function(EMatr, bin_width_log, threshold_log_ref, show_ref_distr=FALSE, plotfigs = TRUE)

Arguments

'EMatr'. Numeric, matrix. Initial low sequence RNA-seq reads. Each column corresponds to individual sample.

'bin_width'. Numeric. Bin width in logarithmic scale is used to create a histogram of RNA-seq calling and furder NL approximation. Value 0.3 is recommended.

'threshold_log_ref'. Numeric. Empirically chosen cutoff value for the reference distribution in a logarithm scale. Values 3.0 – 4.5 are recommended.

'show_ref_distr'. Logical. If set to TRUE, then the reference distribution will be shown from summing up calls across individual genes for all samples, no normalization will be apllied. This regime is usefull to choose 'threshold_log_ref' value for reference curve. If set to FALSE, normalization will be done.

'plotfigs'. Logical. If set to TRUE, differential and cumulative distributions for samples, barplots of library depths will be plotted when 'show_ref_distr' is TRUE

Libraries

'NormalLaplace ' (David Scott, Jason Shicong Fu and Simon Potter)

'data.table';

'Bolstad2'

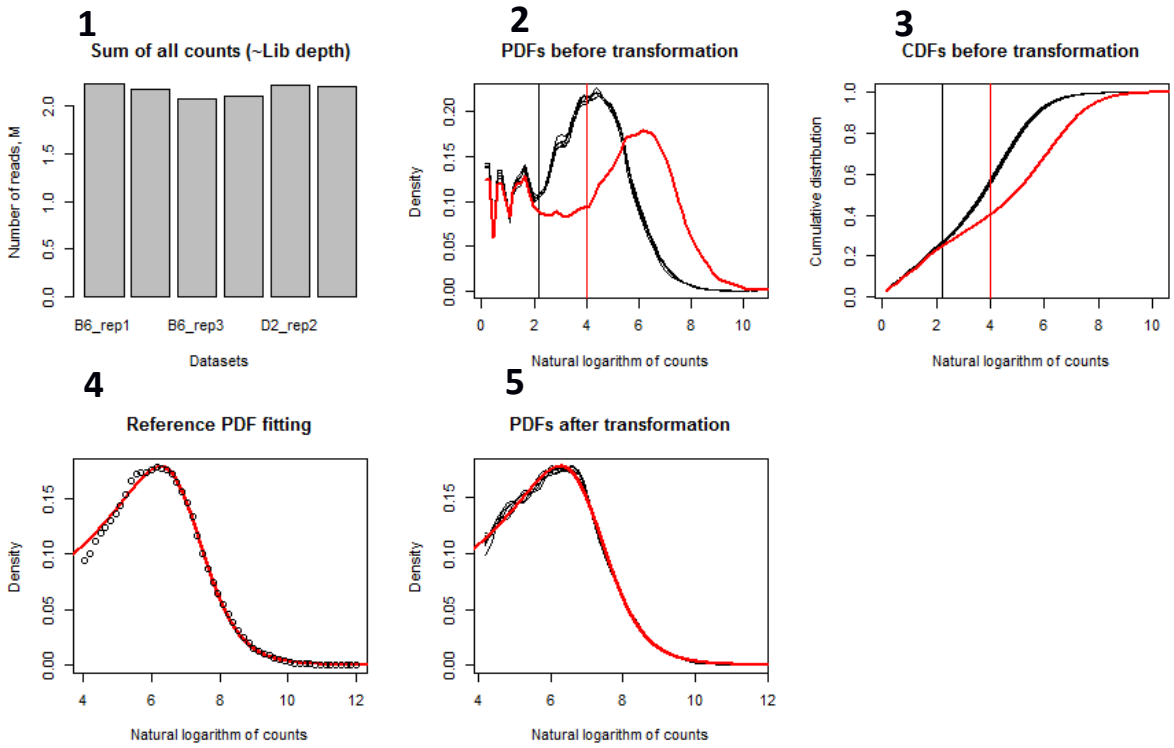
'MESS'

Output

'NormTable'. Numeric. Normalized RNA-seq data table of the same size as initial 'EMatr'.

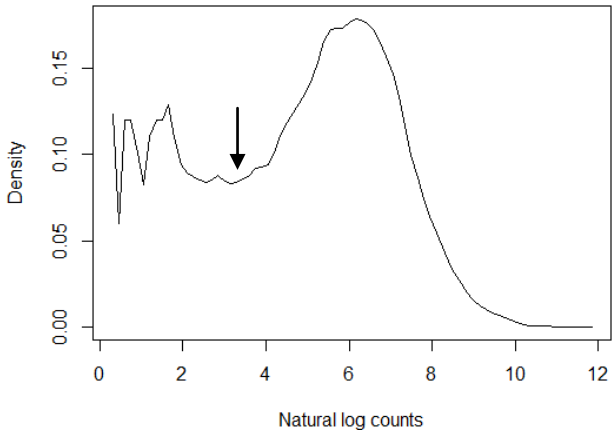
Graphs (see picture below). All counts are natural logarithm transformed.
1) Barplots showing library depths. Each bar corresponds to an individual sample read column from initial input dataset 'EMatr'
2) Probability distributions of reads from initial datasets (black) and pooled reference distribution (red). Vertical red line shows the threshold value ('threshold_log_ref'). Curves to the right from thresholds will be used for NL approximation. Vertical black lines shows threshold values calculated for not normalized distributions, obtained from the condition that the areas under the cut parts of the both not normalized and reference curves are preserved .
3) Cumulative distributions of reads from initial datasets(black) and pooled reference distribution after thresholding (red). Vertical lines show the threshold values the same as at 3). Horizontal line shows the parts of the curves cut below the line after threshholding.
4) Reference distribution (circles) fitted with Normal-Laplace function(red curve).

5) Probability distributions of reads after normalization. Red curve – NL fit of reference distribution, black curves – sample reads distributions after normalization.



Usage

1) Create matrix EMatr, so that colomns correspond to samples and rows to genes.
2) Main_function(EMatr, bin_width_log=0.3, threshold_log_ref=4.0, show_ref_distr= TRUE, plotfigs = TRUE). The output will be a reference distribution below:



Choose the threshold value and proceed to the next step.

3) NormTable <- Main_function(EMatr, bin_width_log=0.3, threshold_log_ref=4.0, show_ref_distr=FALSE, plotfigs = TRUE). The output will be a normalized matrix and if 'plotfigs' is set to TRUE, five figures as shown above.