Multi-transcripts toolbox

I. Description and Motivation

Several approaches exist for studying and measuring gene expression. Microarrays, which is still the most used, and RNA sequencing, which are gaining ground and becoming the technology of choice for new experiments.

The aim is to perform statistical analysis on merged data from these two technologies, despite their differences in structure and nature.

This tool can simulate RNA-seq and microarrays normalized data, normalize raw RNA-seq data (count data), and standardize microarrays and/or RNA-seq data and then merge into a single data set.

It contains 4 modules:

- **microarray_simul.r**: it allows to generate microarrays data. The simulated data have similar characteristics compared to the microarrays data produced by the Affymetrix® platform, after normalization;
- **rnaseq_simul.r**: it allows to simulate the RNA-seq count data and then normalizes them. There are three normalization methods implemented: DESeq2, edgeR, and VOOM;
- **normalisation.rna_seq.r:** it allows to normalize RNA-seq count data. Three normalization methods are available**:** DESeq2, edgeR and VOOM;
- **rnaseq_microarray_fusion.r:** it allows to standardize and then merge standardized RNA-seq and/or microarrays data (real or simulated). Three standardization methods are available: Z-score, Z-score Robust and Quantile Normalization.

II. Deployment and usage in R

II.1 "microarray_simul.r"

This function allows to generate microarrays data with two conditions and known characteristics. These data have similar behavior as those obtained with Affymetrix® platform, after normalization (Log2 intensity).

II.1.1 Arguments

- **gene_number:** an integer specifying the number of genes to be simulated. Default to 10,000;
- **samples_n1:** an integer specifying the number of phenotype 1 (condition 1) samples to be simulated. Default to 75;
- **samples_n2**: an integer specifying the number of phenotype 2 (condition 2) samples to be simulated. Default to 75;
- **diff_genes_ratio**: the proportion of differentially expressed genes. Default to 0.1;
- **up_ratio**: the proportion of up-regulated genes among differentially expressed genes. Default to 0.5;
- **m1**: a decimal number corresponding to average expression difference between condition 2 and condition 1. Default to 1.4;
- **m2**: similar to m1, it allows to have 2 levels of difference, for example high and moderate. Default to 0.8;
- **seed:** an integer used as seed for generating random number, it permits to generate reproducible data. By default, none is set.

II.1.2 Note and details

If the user supplies a decimal number instead of an integer for the first three parameters, the value will be rounded.

The function will not run and will return error messages in the following cases:

- one of the numeric parameters is not numeric;
- one of the integer parameters is negative;
- the number of genes to be simulated is zero;
- the number of samples is zero;
- at least one proportion parameter is not between 0 and 1.

II.1.3 Value

The output is a tab-delimited text file containing a dataset with gene_number rows and (samples_n1+samples_n2+1) columns. The first column contains gene names. First ones are the up–regulated genes, then down-regulated genes, then the genes that are not differentially expressed.

II.1.4 Usage in Rscript

RScript microarray_simul.r --gene_number 1000 --samples_n1 20 --samples_n2 20 --up_ratio 0.5 --diff_genes_ratio 0.1 –m1 1.4 --m2 0.8 --seed 123

II.2 "rnaseq_simul.r"

This function is used to simulate RNA-seq count data and to normalize them. Three normalization methods are available: DESeq2[1][2][3] , edgeR[4][5] and VOOM[6][7].

II.2.1 Arguments

- **gene_number:** an integer specifying the number of genes to be simulated. Default to 10,000;
- **samples_n1:** an integer specifying the number of phenotype 1 (condition 1) samples to be simulated. Default to 75;
- **samples_n2**: an integer specifying the number of phenotype 2 (condition 2) samples to be simulated. Default to 75;
- **diff_genes_ratio**: the proportion of differentially expressed genes. Default to 0.1;
- **up_ratio**: the proportion of up-regulated genes among differentially expressed genes. Default to 0.5;

[1] S Anders and W Huber. Differential expression analysis for sequence count data. Genome Biology 2010;11:R106.

[2] S Anders, W Huber. Differential expression of RNA-Seq data at the gene level – the DESeq package. Last revision 2016-01-12.

[3] DESeq2paper. [http://www-huber.embl.de/DESeq2paper]. [En ligne]

[4] MD Robinson, DJ McCarthy, GK Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010

[5] edgeR: Empirical Analysis of Digital Gene Expression Data in R. https://bioconductor.org/packages/release/bioc/html/edgeR.html. [En ligne]

[6] Charity W Law, Yunshun Chen, Wei Shi and Gordon K Smyth. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology. 15:R29, 2014.

[7] Limma: Linear Models for Microarray Data. http://bioconductor.org/packages/release/bioc/html/limma.html. [En ligne]

- **fc_file**: The path to a text file containing fold-changes in column, with a header. If no file is provided, the default fold-changes will be taken, either as is, or with a random sampling to get the right number of DE and up-regulated genes;
- **rnaseq_norm**: a character indicating the normalization method for RNA-seq. Available methods are "DESeq2", "edgeR" and "VOOM". Default to "DESeq2";
- **seed**: an integer used as seed for generating random number, it permits to generate reproducible data. By default, none is set.

II.2.2 Note and details

Similar checks as for function "microarray_simul.r" will be made. See section II.1.2.

II.2.3 Value

The output is a tab-delimited text file containing a dataset with gene_number rows and (samples_n1+samples_n2+1) columns. The first column contains gene names, first ones are the up–regulated genes, then down-regulated genes, then the genes that are not differentially expressed.

II.2.4 Usage in Rscript

Rscript rnasrq_simul.r --gene_number 1000 --samples_n1 20 --samples_n2 20 --up_ratio 0.5 --diff_genes_ratio 0.1 --seed 123

II.3 "normalisation.rna_seq.r"

It allows normalization of RNA-seq count data. Three methods are available: DESeq2, edgeR and VOOM.

II.3.1 Arguments

- **count_file**: a path to tab-delimited text file containing a matrix of non-negative integers;
- **design**: a path to a text file containing a condition vector in column (qualitative variable) describing the plan of the experiment (condition1 / condition2), samples need to be in same order as in count_file;
- **rnaseq_norm**: a character indicating the normalization method for RNA-seq. Available methods are "DESeq2", "edgeR" and "VOOM". Default to "DESeq2".

II.3.2 Value

The function returns a tab-delimited text file containing the normalized RNA-seq data matrix according to the chosen normalization method. The output matrix has the same dimensions as the input matrix. Also, they have the same names of rows and columns.

II.3.3 Usage in Rscript

Rscript normalisation.rna_seq.r --gene_number --samples_n1 20 --samples_n2 20 --up_ratio 0.5 --diff_genes_ratio 0.1

II.4 "rnaseq_microarray_fusion.r"

II.4.1 Arguments

- **standardisation**: a character indicating the standardization method. Available methods are zscore, robust_zscore and quantile. Default to zscore ;
- **all_genes**: a logical parameter indicating whether the function should return all genes or just the one in common. Default to TRUE;
- **tables**: a character string containing the paths to the datasets to be standardized and merge, separated by commas. The default value is "MicroArray_simulation.txt,RNA-seq_simulation.txt".

II.4.2 Note and details

The function will perform a pretreatment process to make the data usable:

- Check if some patient's names are shared between datasets. if there are any duplicate, we add a suffix ("_ti", where i indicates the ith dataset) at the end of the name;
- Check the nature of data: if the data are not numeric, the function halts and an error is produced.

II.4.3 Value

The function returns a tab-delimited text file containing the matrix of merged data processed with the chosen standardization method.

The number of columns of the output data is equal to the sum of total samples plus one. The first column contains gene names. The number of rows depends on whether user wants to keep all genes or only common genes.

II.4.4 Usage in Rscript

Rscript rnaseq_microarray_fusion.r --standardisation zscore --tables MicroArray_simulation.txt,RNASeq_simulation.txt

III. Deployment and usage in Galaxy workflows

To use Galaxy[8], you must create a user account on the computing cluster of BiRD platform: http://www.pf-bird.univ-nantes.fr/demande-de-compte-birdcluster-1354976.kjsp?RH=1442585061597

IV. Software dependencies

R packages:

- optparse: https://CRAN.R-project.org/package=optparse
- edgeR: https://www.bioconductor.org/packages/release/bioc/html/edgeR.html
- DESeq2: https://www.bioconductor.org/packages/release/bioc/html/DESeq2.html
- preprocessCore: https://www.bioconductor.org/packages/release/bioc/html/preprocessCore.html

[8] https://galaxy-bird2.univ-nantes.fr