RNentropy README (v. 1.1.1)

***INSTALLATION (LINUX)

Just open the archive file “RNentropy_1.1.1.tar.gz” using the command:

```
tar -xvf RNentropy.1.1.1.tar.gz
```

A folder named “RNentropy1.1.1” should appear within your current folder. Now type:

```
cd RNentropy1.1.1
```

*** To use the pre-compiled binary files:

Type:

```
chmod a+x RNentropy
chmod a+x select_results
```

In this way you are flagging the binary files as executable for your OS.
You should now be able to run “RNentropy” and the “select_result”s parser utility on any 64 bit Linux platform.
***

*** To compile from source:

If you prefer to compile your own binary you can find the source code of RNentropy and the parser utility select_results within the “src” folder. Type from within the RNentropy1.1.1 folder:

```
g++ src/RNentropy.cpp -o RNentropy -O3 -lgsl -lgslcblas
g++ src/select_results.cpp -o select_results -O3
```

Please note that in this case you need the Gnu Scientific Libraries
https://www.gnu.org/software/gsl/ already installed in your system.
***

***USING RNentropy

To launch RNentropy just type:

```
./RNentropy -f input_file
```

to perform the global and local sample specificity tests.

RNentropy reads its parameters directly from the headers of the input file.

*****INPUT FILE**

The input file must include an header section followed by tab (or space) separated tabular data.

The following columns are required:

- column with transcript identifiers
- column with genes identifiers (if you have expression values only for genes you can specify the same column for transcripts and genes identifiers)
- a column with an expression measure (e.g. TPM, FPKM or RPKM) for each replicate of each sample.

The order of the columns in the file is not important since you can assign columns to samples from within the header section. Number of replicates should be the same for each sample.

The header section describes the structure of the input file using a set of keywords preceded by the "#" symbol and usually followed by a number (or a comma separated list of numbers) that specify the column(s) they refer to.

This is the list of available keywords and their syntax:

# GENE_COL N
N is the position of the column with genes ids.

# TR_COL   N
N is the position of the column with transcripts ids

# COMMENTS   N1,N2,...,Nn
N1,N2,Nn are the positions of any number of accessory columns that you want reported in the output. They could be for example columns with the genomic position of transcripts. RNentropy will just keep these columns in the output as they are.

# EXP SAMPLE_1   N_R1, N_R2,...,N_RN
This is the keyword to specify the positions of the columns with the expression data from your samples. You need an EXP line for each sample in the input. N_R1,N_R2,...,N_RN are the positions of the columns with expression data for all the replicates from sample SAMPLE_1. You can substitute SAMPLE_1 with any label that is meaningful to your data. Labels must not contain any space or tabular character.

If for example you have two more samples after SAMPLE_1 the header will continue like this:

```
# EXP  SAMPLE_2     N_R1, N_R2,...,N_RN
```

Again, N_R1,N_R2,...,N_RN are the positions of the columns with the expression data for the replicates from sample SAMPLE_2. You can substitute SAMPLE_2 with any label that is meaningful to your data. Labels must not contain any space or tabular character.

And

```
# EXP  SAMPLE_3     N_R1, N_R2,...,N_RN
```

```
# END
```

This keyword marks the end of the header section.

Columns position goes from left to right and the leftmost column has position 1.

***Header section example 1: ***

```
# GENE_COL      3
# TR_COL        2
# COMMENTS      1
# EXP   BRAIN_1 4
# EXP   BRAIN_2 5
# EXP   BRAIN_3 6
# EXP   HEART_1  7
….
# EXP   MUSCLE_1 21
# END
```

In this header the column with gene identifiers is the third one, while the second one contains transcript identifiers. The first column of the file is a comment column to be reported as it is in the output. Then there are 18 samples labeled BRAIN_1, BRAIN_2, BRAIN_3, HEART_1, HEART_2, HEART_3, and so on for KIDNEY, LIVER, LUNG, MUSCLE. In this case each column represents a sample without replicates.

***Header section example 2: ***

```
# GENE_COL  3
# TR_COL        1
# COMMENTS      2
# EXP   BRAIN_1 4,5,6
# EXP   BRAIN_2 7,8,9
# EXP   BRAIN_3 10,11,12
```

END

In this header the column with gene identifiers is the third one while the first one contains transcript identifiers. The second column is a comment column. We have three samples labeled BRAIN_1, BRAIN_2 and BRAIN_3 and each sample has three replicates.

Have a look at the input files within the “example_input” folder for some examples. In sample_input_file_1.txt there are data from 18 samples referring to 6 tissues from 3 individuals without replicates, while in sample_input_file_2 there are three samples from the same tissue (brain) of three different individuals, each sample has 3 replicates.

*****OUTPUT FILES**

RNentropy outputs two tabular files named input_file.main.res and input_file.summary.res, where “input_file” corresponds to the name of your input file. The topmost row of each file are labels that specify the content of the corresponding column. The “.summary.res” file is a more compact version of the output that you find in the “.main.res” file.
Column labels are explained below:

GENE_ID:
column with gene identifiers

TR_ID:
column with transcript identifiers

COMMENT_N:
comment column number N

SAMPLE_1_1:
expression data of SAMPLE_1, replicate 1

SAMPLE_1_N:
expression data of SAMPLE_1, replicate N

SAMPLE_2_1:
expression data of SAMPLE_2, replicate 1

SAMPLE_2_N:
expression data of SAMPLE_2, replicate N

GL_LPV:

negative log of the p-value for the global sample specificity test. Please notice that this is the raw p-value. It should be corrected by using suitable methods, e.g. Bonferroni or Benjamini-Hockberg corrections.

LOC_LPV_SAMPLE_1_1:
log of the p-value for the local sample specificity test referred to sample with label SAMPLE_1, replicate 1. The sign is set to minus when the corresponding expression value is smaller than its expected value, to plus when larger.

LOC_LPV_SAMPLE_1_N:
log of the p-value for the local sample specificity test referred to sample with label SAMPLE_1, replicate N. The sign is set to minus when the corresponding expression value is smaller than expected, to plus when larger.

LOC_LPV_SAMPLE_2_1:
log of the p-value for the local sample specificity test referred to sample with label SAMPLE_2, replicate 1. The sign is set to minus when the corresponding expression value is smaller than expected, to plus when larger.

LOC_LPV_SAMPLE_2_N:
log of the p-value for the local sample specificity test referred to sample with label SAMPLE_2, replicate N. The sign is set to minus whn the corresponding expression value is smaller than expected, to plus when larger.

In the “example_output” folder you will find the RNentropy output files for “sample_input_file_1.txt” and “sample_input_file_2”.txt.

***PARSING RESULTS

You can use the “select_results” parsing utility to pick the “over-expressed” genes from your RNentropy output. The syntax is as follow:

select_results RNentropyfile.summary.res GPV_threshold LPV_threshold sample_num rep_num

Where “RNentropyfile.summary.res*”* is the “summary” results file from a RNentropy run. GPV_threshold and LPV_threshold are the p-value thresholds for the Global (Benjamini-Hockberg corrected) and Local p-values respectively (0.01 is a typical value for both of them). Finally, “sample_num” and “rep_num” are the number of samples and replicates for each sample respectively (the utility works only if all the samples share the same number of replicates).

You will get a “RNentropyfile.summary.res.selected” file with a row for each gene passing the Benjamini-Hockberg corrected GPV threshold and a “SAMPLE_x” column for each sample (x is the sample number). When the expression values for a gene satisfy the over-expression local p-value threshold for all the replicates of “SAMPLE_x” you get a “1” in the corresponding column. On the other hand a “-1” means that the gene seems to be significantly less expressed in all the replicates of the corresponding sample with respect to its overall expression. A 0 or a X means that the local p-value threshold is not satisfied respectively by some or all the replicates of the corresponding sample.

You will also get a “RNentropyfile.summary.res.pmi” file containing the point mutual information and the normalized point mutual information matrices.

***Contacts:

For any question about RNentropy feel free to e-mail us:

giulio.pavesi@unimi.it
federico.zambelli@unimi.it