tssports : integrates tsRNA feature extraction, differential analysis, and extensive visualization in one

https://github.com/Xia-Youmei/tssports-master

tssports can perform small RNA classification, differential gene analysis, principal component analysis (PCA), and visualize the related functions and mechanisms of tsRNA through pie chart, volcano chart, heat map, MA chart etc.

Installation

You can install the development version of *tssports* like so:

```
if (!requireNamespace("devtools", quietly = TRUE)) {
  install.packages("devtools")
}

devtools::install_github('Xia-Youmei/tssports-master')
```

Alternatively, the latest version can be installed locally from Git-hub: https://github.com/Xia-Youmei/tssports-master, and use the R package for your own data.

```
if (!requireNamespace("remotes", quietly = TRUE)) {
  install.packages("remotes")
}

remotes::install_local("file_address/tssports-master-main.zip",upgrade =
F,dependencies = T)
```

Note: For most use cases it is not necessary to install the *tssports* package locally, If you have a bad Internet connection, you can choose this method, and because of the sample data is large, please be patient for a few minutes.

To start working with the package *tssports*, you can load it in the R environment with the following command.

```
library('tssports')
```

Example

There is an example downloaded from Gene Expression Omnibus under the accession code GSE144666, it has 6 data sets that are SRR11004011-13 (RNAs were sequenced with standard protocols), SRR11004020-22 (RNAs were treated with T4PNK, then AlkB before cDNA library construction), organism is Mus musculus, tissue is brain.

First, you need to determine the working directory located in the dataset address:

```
setwd("./examples")
```

sportsV1.1 processed and generate the output files into one folder, you can determine the working directory located in the dataset address, it will automatically prossessing those contain "-miR-|-mir-|-let-" keywords annotation commented out, and automatically add reads of same genes at the end of the file, the filename ended in “_output_collapse_miRNA.txt” is obtained.

```
collapse_mature_mirna_reads()
```

This function will result in 3 unified files: 1. sports_combined_sample_fragments_counts_matrix_all.txt; 2. sports_combined_sample_fragments_counts_matrix_0.5.txt; 3. sports_combined_sample_fragments_annotation.txt. It can also read the entire folder, automatically identify and process the files ending in “_miRNA.txt” in the folder, first it will collect and sort out all genes to form the first file contains gene sequence, gene sequence length, whether to match to the genome, annotated genes; After that, all the same gene sequences in the input file were sorted into one file, and the screening condition was set as the probability of each fragment not being zero in the sample was greater than 0.5 to form the second file. The third file is the original file that collates all the same gene sequences in the input file into one file without setting screening conditions and gives users sufficient follow-up custom analysis.

```
combine_read_counts()

1.sports_combined_sample_fragments_counts_matrix_all
Fragment    Length  Match_Genome    Annotation
TCGCTGCGATCTATTGAAAGTCAGCCCTCGACACAAGGGTTTGT    44  Yes 28S-rRNA
TCACAGTGAACCGGTCTCTTTAA 23  NO  piRNA
CGCGACCTCAGATCAGACGT    20  NO  28S-rRNA
TCGGATCCGTCTGAGCTTGGCTTT    24  NO  piRNA
TCACAGTGAACCGGTCTCTTAA  22  NO  piRNA
TCTTTGGTTATCTAGCTGTATGTT    24  NO  piRNA
GGCTGGTCCGAAGGTAGTGAGTTATCTCAATT    32  Yes RNY1-YRNA
TCAGTCGGTCCTGAG 15  Yes 28S-rRNA
TGGGCTGTAGTGCGCTATGC    20  Yes misc_RNA
CTGGGCTGTAGTGCGCTATGC   21  Yes misc_RNA
CGCTGCGATCTATTGAAAGTCAGCCCTCGACACAAGGGTTTGT 43  Yes 28S-rRNA
CGCGACCTCAGATCAGAC  18  NO  28S-rRNA
CATTGATCATCGACACTTCGAACGCACTTGCGGCCCCGGGT   41  NO  5.8S-rRNA

2.sports_combined_sample_fragments_counts_matrix_0.5.txt
Sequence    SRR11004011 SRR11004012 SRR11004013 SRR11004020 SRR11004021
SRR11004022
TCGCTGCGATCTATTGAAAGTCAGCCCTCGACACAAGGGTTTGT    269161  143901  82697   49357
96480   132035
TCACAGTGAACCGGTCTCTTTAA 204627  136474  72396   2679    5694    38834
CGCGACCTCAGATCAGACGT    93017   119623  66666   30488   33792   143835
TCGGATCCGTCTGAGCTTGGCTTT    66940   50898   49442   25380   38967   73429
TCACAGTGAACCGGTCTCTTAA  45249   32876   14415   503 997 8267
TCTTTGGTTATCTAGCTGTATGTT    35142   28972   27240   1199    3243    9601
GGCTGGTCCGAAGGTAGTGAGTTATCTCAATT    32900   20960   14730   52832   34279
44606
mmu-miR-3960    0   0   0   12  8   2
mmu-miR-3473g   0   0   0   2   1   1
mmu-miR-6987-5p 0   0   0   2   3   1
```

```
mmu-miR-1951    0   0   0   1   1   3

3.sports_combined_sample_fragments_annotation.txt
Sequence    SRR11004011 SRR11004012 SRR11004013 SRR11004020 SRR11004021
SRR11004022
TCGCTGCGATCTATTGAAAGTCAGCCCTCGACACAAGGGTTTGT    269161  143901  82697   49357
96480   132035
TCACAGTGAACCGGTCTCTTTAA 204627  136474  72396   2679    5694    38834
CGCGACCTCAGATCAGACGT    93017   119623  66666   30488   33792   143835
TCGGATCCGTCTGAGCTTGGCTTT    66940   50898   49442   25380   38967   73429
TCACAGTGAACCGGTCTCTTAA  45249   32876   14415   503 997 8267
TCTTTGGTTATCTAGCTGTATGTT    35142   28972   27240   1199    3243    9601
GGCTGGTCCGAAGGTAGTGAGTTATCTCAATT    32900   20960   14730   52832   34279
44606
mmu-miR-511-5p  0   0   0   0   0   1
mmu-miR-505-3p  0   0   0   0   0   1
mmu-miR-7074-5p 0   0   0   0   0   1
mmu-miR-7229-5p 0   0   0   0   0   1
mmu-miR-6984-3p 0   0   0   0   0   1
```

This function requires the user to input the set of the last two digits in the SRR filename of the experimental group to distinguish the experimental group from the control group. For example, 20:22 is required for difference analysis in the example data. This function will automatically recognize the sports_combined_sample_fragments_counts_matrix_0.5.txt file output from the previous function. After processing, you will get 4 files, 1. Match all fragment annotations to genes, and get the file sports_counts_all.txt; 2. Match all the annotations to genes, and processed with the DESeq2 R package to obtain the differentially expressed genes, and obtain the file sports_deg_fDR005_2fc_all.txt. 3. All fragments were annotated as genes, and the differentially expressed genes were obtained after processing with DESeq2, and the reads were normalized to cpm (Counts per million) value to obtain the file sports_cpm_fdr005_2fc_all.txt. 4. Save all the differential genes calculated by DESeq2, without distinguishing log2FoldChange and padj, and get the file sports_DEG_all.txt;

```
getdegs(20:22)

1.sports_counts_all.txt
    SRR11004011 SRR11004012 SRR11004013 SRR11004020 SRR11004021 SRR11004022
28S-rRNA    269161  143901  82697   49357   96480   132035
piRNA   204627  136474  72396   2679    5694    38834
RNY1-YRNA   32900   20960   14730   52832   34279   44606
misc_RNA    28044   35480   18645   13025   21257   15968
5.8S-rRNA   22734   13872   2363    21703   16378   11993
45S-rRNA    12263   12825   15348   12838   27474   53615
mature-tRNA-Gly-GCC_5_end;mature-tRNA-Gly-CCC_5_end 10286   8054    6349
15440   24174   38846
5S-rRNA 8157    19663   8814    4571    8266    13170

2.sports_deg_fDR005_2fc_all.txt
ID  baseMean    log2FoldChange  lfcSE   stat    pvalue  padj
mature-tRNA-His-GTG 30035.3716708793    9.18307904145374    0.580202283242111
15.8273748771543    2.01452256749788e-56    2.96940626449187e-53
pre-tRNA-Leu-CAA    21805.5636328603    9.85974407854832    0.658354292658678
14.9763496471346    1.04818154091913e-50    7.72509795657397e-48
mature-tRNA-Arg-CCT;mature-tRNA-Arg-CCG 468.48506156553 7.40567460160257
0.655365046544287   11.3000756458593    1.31105017035912e-29
6.44162650369779e-27
```

```
mature-tRNA-Arg-ACG_CCA_end 43397.7312599756    7.76515581481599
0.692681783635114   11.2102786564782    3.630486126182e-29  1.33783413749807e-26
pre-tRNA-Tyr-GTA    505.057106284187    9.2287311054882 0.828001518415102
11.1457900743384    7.50750007057211e-29    2.21321102080466e-26
mature-tRNA-Pro-TGG_5_end;mature-tRNA-Pro-CGG_5_end;mature-tRNA-Pro-AGG_5_end
1946.81081659445    9.928241334204  1.04537621880354    9.49729021535149
2.15423010298028e-21    5.29222528632156e-19
mature-tRNA-Arg-ACG 1850.07541775039    6.35805962609316    0.67452163520523
9.42602771245233    4.25913920030078e-21    8.96853025891908e-19

3.sports_cpm_fdr005_2fc_all.txt
ID  SRR11004011 SRR11004012 SRR11004013 SRR11004020 SRR11004021 SRR11004022
28S-rRNA    14.3200235328116    13.5965848351574    13.0443714276675
14.6978424726186    14.9935127374893    14.6639604791408
piRNA   13.9245729094712    13.5201406814239    12.8524707211966
10.4952997919691    10.9115022019083    12.89856566266
RNY1-YRNA   11.2882200871046    10.8179018026627    10.5560816430719
14.7959962196167    13.5006881015406    13.0984586794382
misc_RNA    11.057924529273 11.5769419854079    10.8959117870999
12.776022553392 12.8113773297272    11.6166995041221
5.8S-rRNA   10.7552419878891    10.2228471945342    7.92102213175833
13.5125514135644    12.4352583087336    11.2038607265397
45S-rRNA    9.86541494133145    10.109728768677 10.6153362976346
12.7551626430638    13.1814608784966    13.36382994117
mature-tRNA-Gly-GCC_5_end;mature-tRNA-Gly-CCC_5_end 9.61208227726735
9.43932057807877    9.34318987544876    13.0213797611129    12.9968713182784
12.8990113391204
5S-rRNA 9.27799186112964    10.7257992535941    9.81583605097291
11.2657012714733    11.4490159975246    11.3388688232466

4.sports_DEG_all.txt
ID  baseMean    log2FoldChange  lfcSE   stat    pvalue  padj
28S-rRNA    117369.725044431    0.382472823614208   0.644227936830634
0.593691769245269   0.552718289120833   0.673819151856151
piRNA   50767.1086731965    -2.40366212934164   0.765431138235381
-3.14027220643652   0.00168790918441722 0.00846251067289452
RNY1-YRNA   45719.3346323336    2.44151119683615    0.707386312647438
3.45145382824646    0.000557575147822217    0.00336997078222502
misc_RNA    22395.6619710466    0.67498992210017    0.6596268775384
1.02329050723206    0.306170510703025   0.448158225199861
5.8S-rRNA   19601.5728991033    1.97867701752085    0.882338784732432
2.24253659904668    0.0249267126520048  0.0724693776115484
45S-rRNA    24872.0434625502    2.26075136993462    0.569544671009807
3.96940132180727    7.20534381777695e-05    0.000689654336844366
mature-tRNA-Gly-GCC_5_end;mature-tRNA-Gly-CCC_5_end 21344.3984569
2.86332800867512    0.563247054211885   5.08360938111179    3.70329152216592e-07
9.00411658654179e-06
5S-rRNA 9981.49076365101    0.648247200530487   0.639794072975421
1.01321226299543    0.310958784289966   0.450329597790607
```

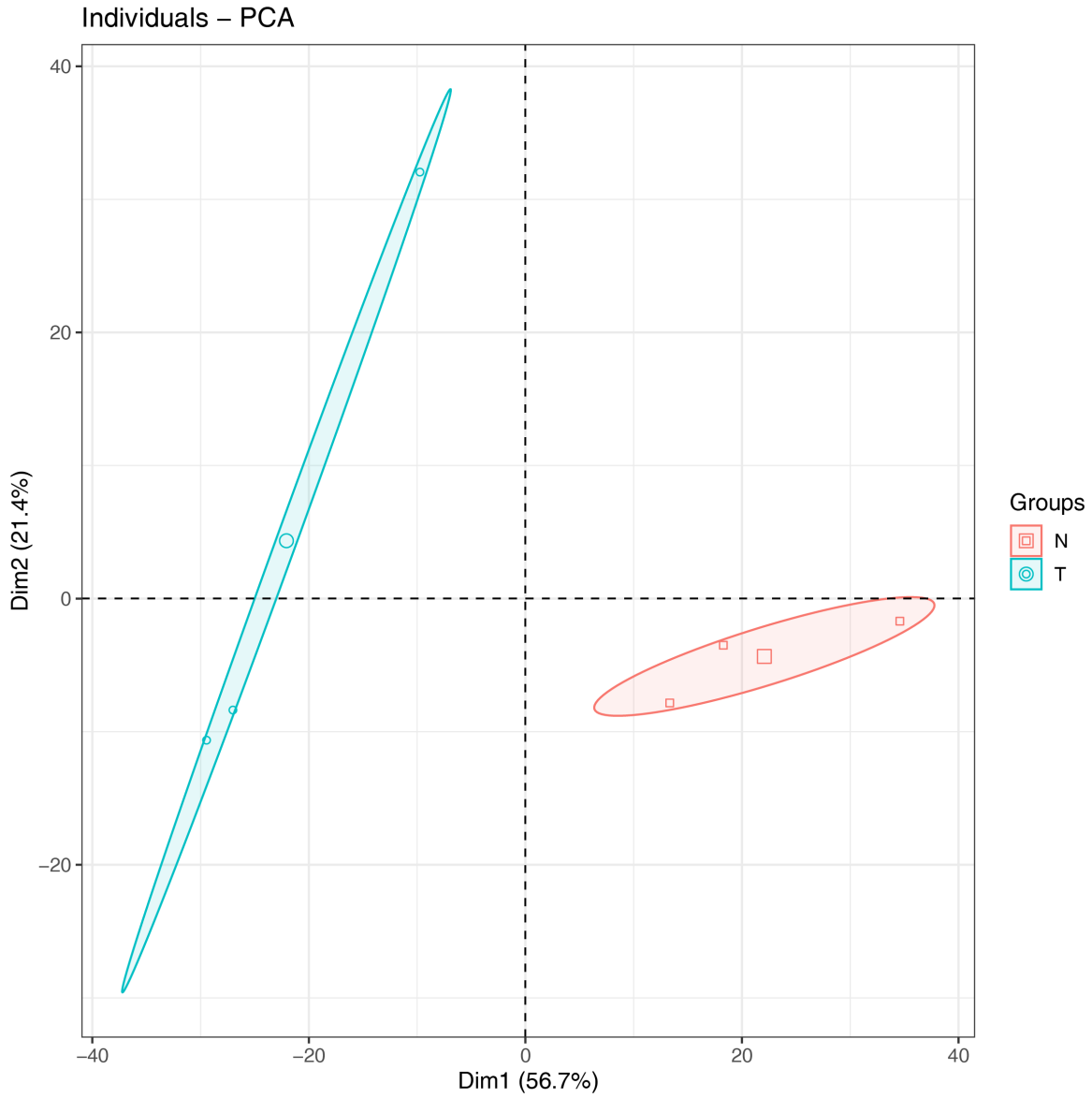
This function will automatically identify sports_DEG_fdr005_2fc_all.txt file in the folder, and categorizing the miRNAs, extract contains keywords "-miR-|-let-" for miRNA_diff. TXT file, The tsRNA_diff.txt file containing the keyword "tRNA", the rsRNA_diff.txt file containing the keyword "rRNA", and the ysRNA_diff.txt file containing the keyword "YRNA".

```
tsRNA_ann_classify()
```

```
1.miRNA_diff.txt
ID  baseMean    log2FoldChange  lfcSE   stat    pvalue  padj
mmu-miR-2137    87.8349839348083    5.9563172036757 0.799484930810467
7.45019321081833    9.3203633809477e-14 7.63234201306495e-12
mmu-miR-153-3p  316.296855176786    -5.03376515234459   0.750489390856289
-6.70731020807795   1.98244544430187e-11    1.12389407111575e-09
mmu-miR-340-5p  1374.52373372071    -4.43397551622456   0.754895118332545
-5.87363119530906   4.26351320188798e-09    1.39653743546286e-07
mmu-miR-101a-3p 13763.2199746233    -4.86630273572994   0.87811726649203
-5.5417458708792    2.9947070804607e-08 8.82839647319815e-07
mmu-miR-690 1186.05912973371    3.84780460164948    0.713714629912904
5.39123683385759    6.99743788822742e-08    1.983504509086e-06
mmu-miR-9-3p    2296.27478279842    -4.20391814855742   0.810478351339246
-5.18695920947278   2.13755526571688e-07    5.62635082440478e-06
mmu-miR-136-3p  124.245832937843    -5.32869206253836   1.04845251393674
-5.08243529554823   3.72626263079409e-07    9.00411658654179e-06

2.tsRNA_diff.txt file
ID  baseMean    log2FoldChange  lfcSE   stat    pvalue  padj
mature-tRNA-His-GTG 30035.3716708793    9.18307904145374    0.580202283242111
15.8273748771543    2.01452256749788e-56    2.96940626449187e-53
pre-tRNA-Leu-CAA    21805.5636328603    9.85974407854832    0.658354292658678
14.9763496471346    1.04818154091913e-50    7.72509795657397e-48
mature-tRNA-Arg-CCT;mature-tRNA-Arg-CCG 468.48506156553 7.40567460160257
0.655365046544287   11.3000756458593    1.31105017035912e-29
6.44162650369779e-27
mature-tRNA-Arg-ACG_CCA_end 43397.7312599756    7.76515581481599
0.692681783635114   11.2102786564782    3.630486126182e-29  1.33783413749807e-26
pre-tRNA-Tyr-GTA    505.057106284187    9.2287311054882 0.828001518415102
11.1457900743384    7.50750007057211e-29    2.21321102080466e-26
mature-tRNA-Pro-TGG_5_end;mature-tRNA-Pro-CGG_5_end;mature-tRNA-Pro-AGG_5_end
1946.81081659445    9.928241334204  1.04537621880354    9.49729021535149
2.15423010298028e-21    5.29222528632156e-19
mature-tRNA-Arg-ACG 1850.07541775039    6.35805962609316    0.67452163520523
9.42602771245233    4.25913920030078e-21    8.96853025891908e-19

3.rsRNA_diff.txt
ID  baseMean    log2FoldChange  lfcSE   stat    pvalue  padj
45S-rRNA    24872.0434625502    2.26075136993462    0.569544671009807
3.96940132180727    7.20534381777695e-05    0.000689654336844366
rRNA    12.4284548829842    2.37423147013067    0.937010532600679
2.53383648051529    0.0112821365345229  0.0402660272442777

4.ysRNA_diff.txt
ID  baseMean    log2FoldChange  lfcSE   stat    pvalue  padj
RNY3-YRNA   1275.60735378423    2.23955706124081    0.603348544285474
3.71187944754726    0.000205725970976646    0.00156309320216277
RNY1-YRNA   45719.3346323336    2.44151119683615    0.707386312647438
3.45145382824646    0.000557575147822217    0.00336997078222502
```

This function will automatically identifies the sports_counts_all.txt file in the folder, selects the top 1000 genes by multiple change, and generates a PDF file of the principal component analysis using the ggplot2 R package.

```
pca(20:22)
```



function of visualization

pie_plot_tsRNA_aa.pdf: This function will automatically identifies the tsRNA_diff.txt file in the folder, and uses the ggplot2 R package to draw the pie chart of different amino acid classes of tsRNA, including Glu, Gly, Val, and Ser, generate the pie_plot_tsRNA_aa.pdf file.

pie_plot_tsRNA_end.pdf: This function will automatically identify the tsRNA_diff.txt file in the folder, and use the ggplot2 R package to draw the pie chart of different tsRNA end categories, including 5'end, 3'end, CCA end, generate the pie_plot_tsRNA_end.pdf file.

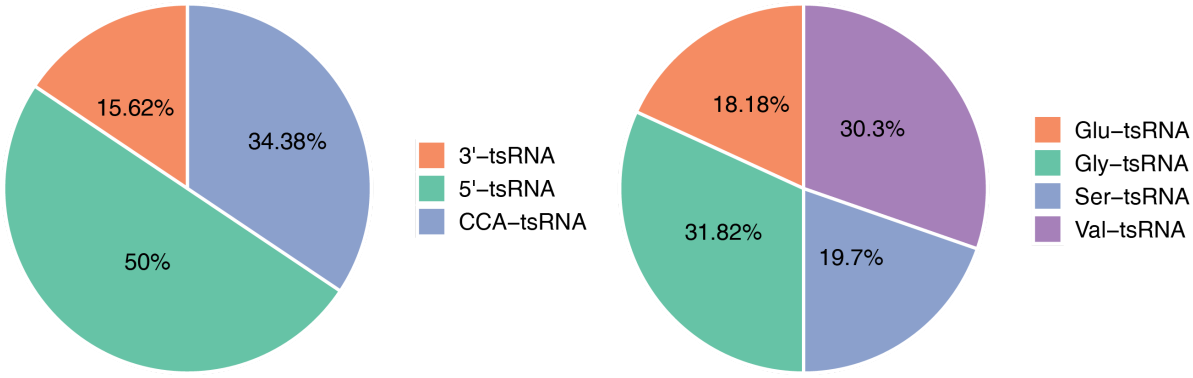
maplot.pdf: This function will automatically identifies the sports_deg_all.txt file in the folder and uses the ggplot2 R package to draw the MA map of sports output differentially expressed genes, generating the "maplot.pdf" file.

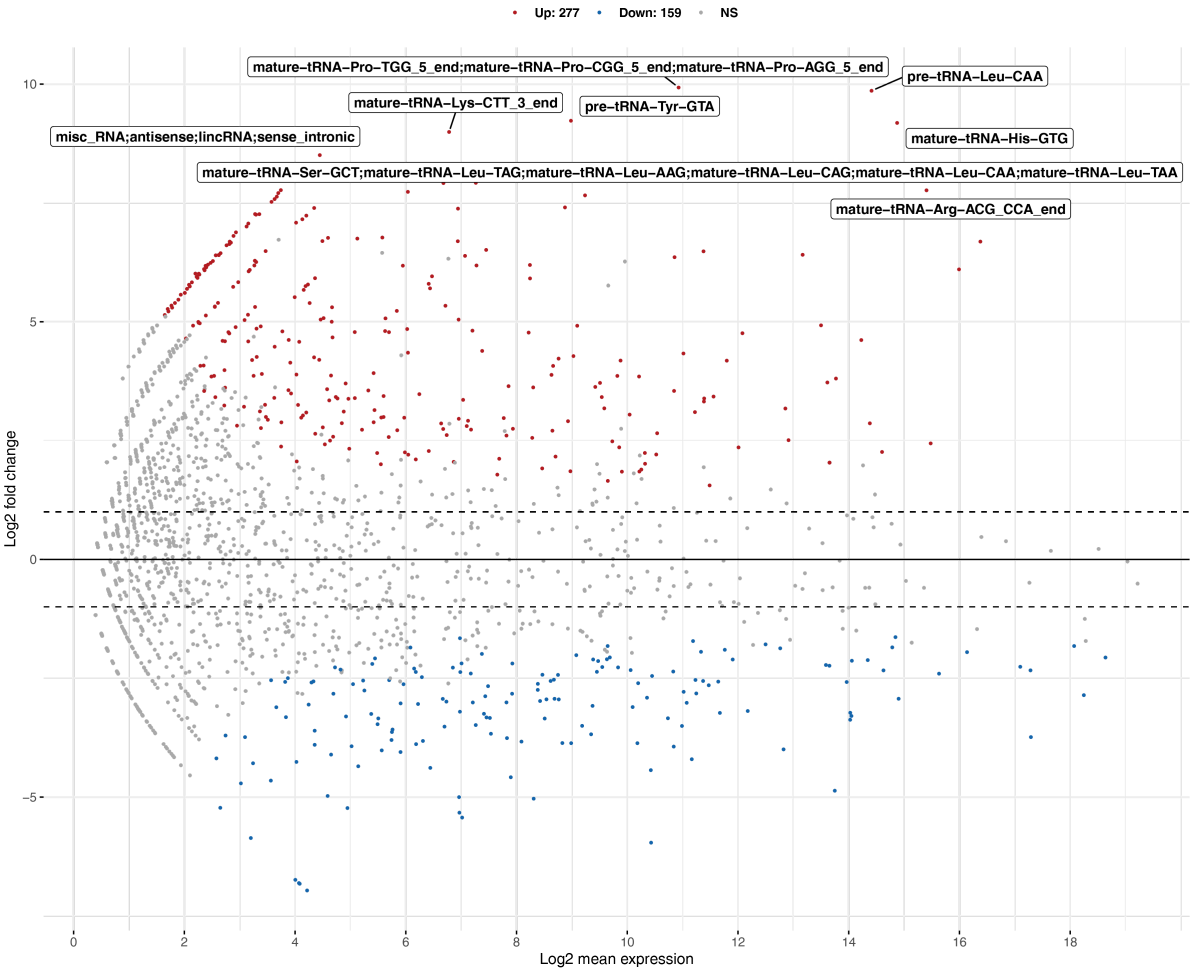
heatmap_plot.pdf: This function will automatically identifies the sports_cpm_fdr005_2fc_all.txt and sports_DEG_fdr005_2fc_all.txt files in the folder, selects and uses the ggplot2 R package to draw a heat map of the differentially expressed genes that sports outputs, Generate the heatmap_plot.pdf file.

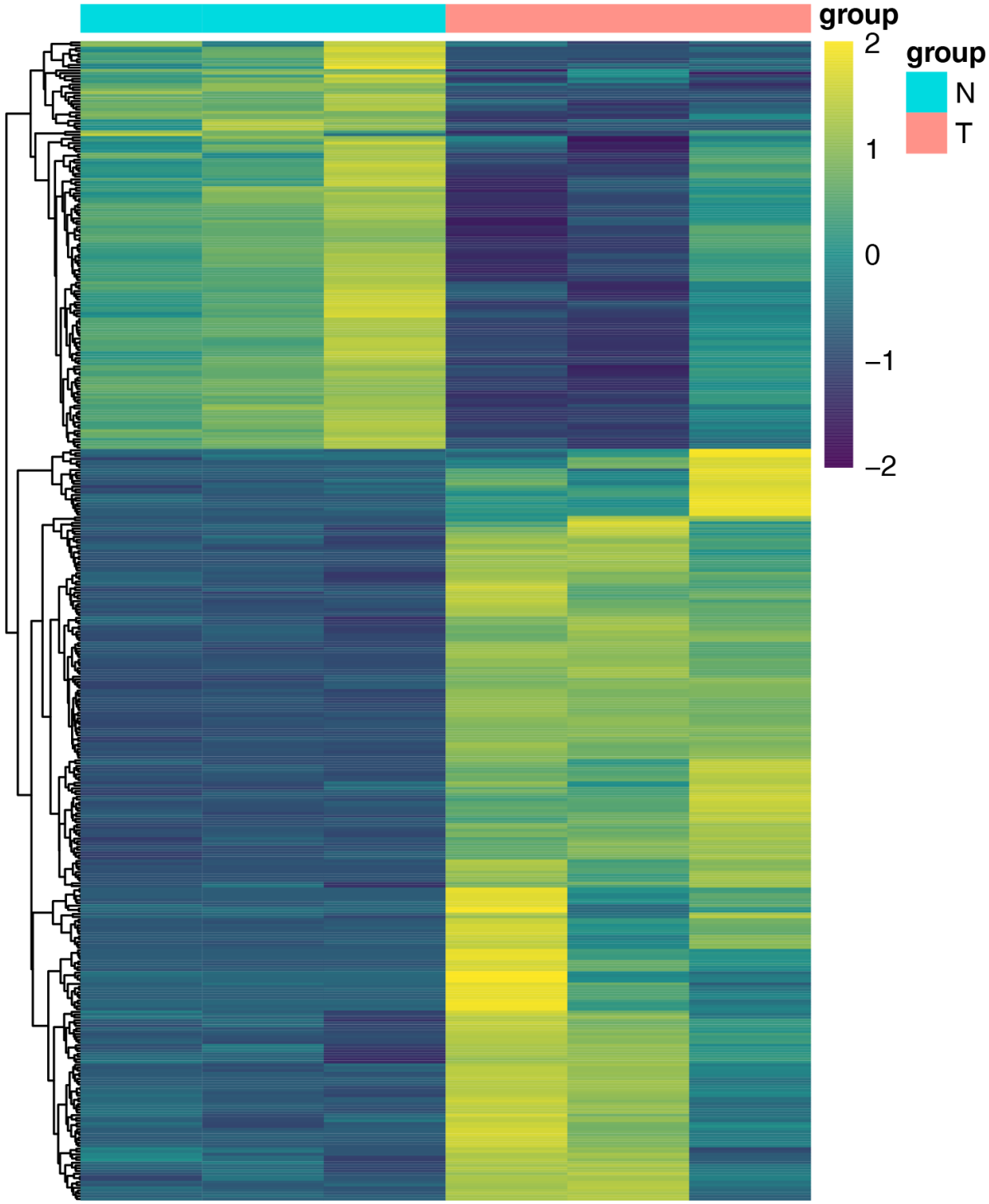
volcano_plot: This function will will automatically identify the sports_DEG_all.txt file in the folder and use the ggplot2 R package to draw the volcano plot of sports output differentially expressed genes, log2FC>1, $p \leq 0.05$, generate the "maplot.pdf" file.

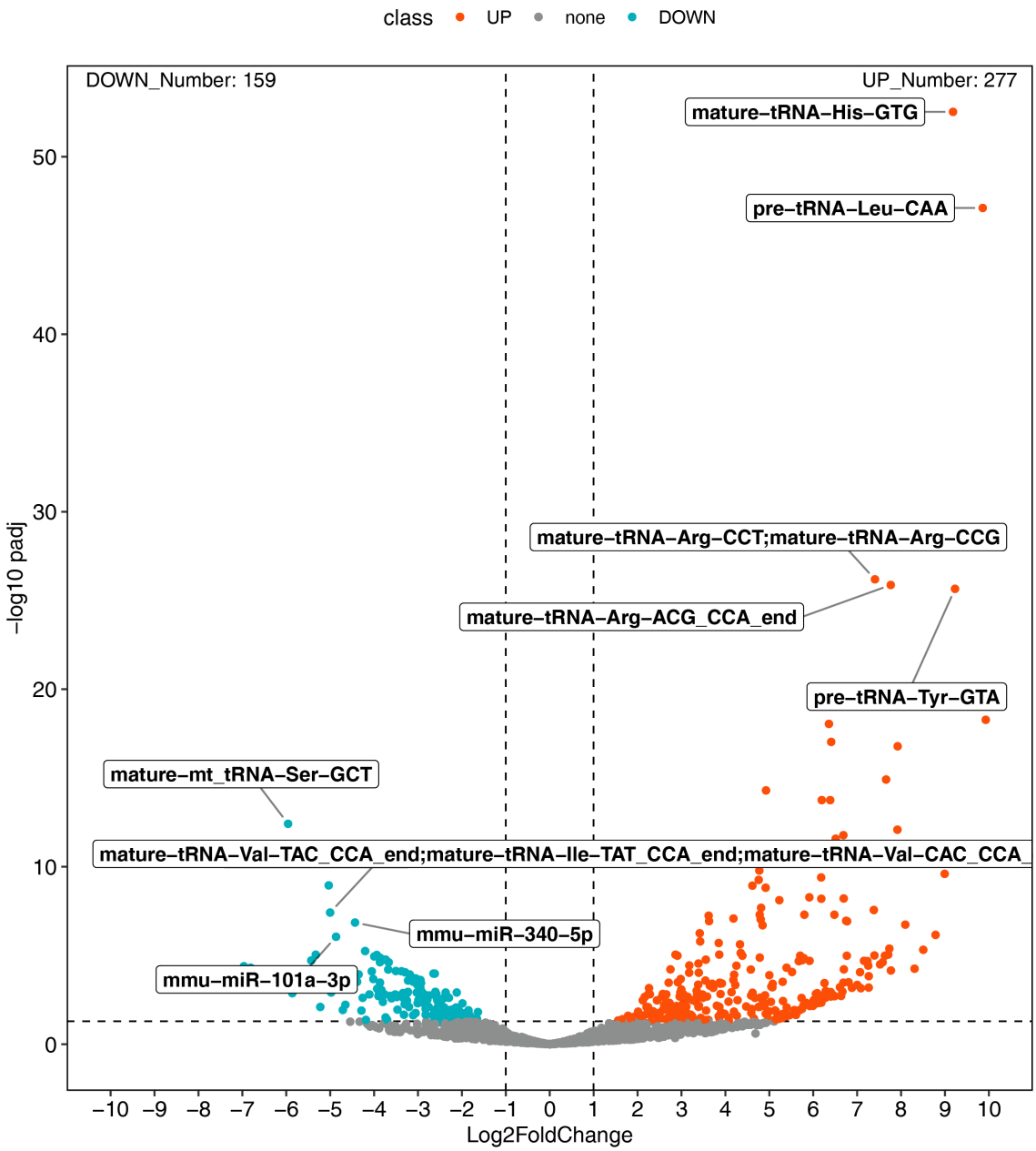
```
visualization()
```

Those pictures shown below are the sample data output, and the format is adjusted by Adobe Illustrator software.






group
2
1
0
−1
−2
group
N
T


class
UP
none
DOWN
DOWN_Number: 159
UP_Number: 277
mature–tRNA–His–GTG
pre–tRNA–Leu–CAA
mature–tRNA–Arg–CCT;mature–tRNA–Arg–CCG
mature–tRNA–Arg–ACG_CCA_end
pre–tRNA–Tyr–GTA
mature–mt_tRNA–Ser–GCT
mature–tRNA–Val–TAC_CCA_end;mature–tRNA–Ile–TAT_CCA_end;mature–tRNA–Val–CAC_CCA
mmu–miR–340–5p
mmu–miR–101a–3p
−log10 padj
Log2FoldChange