RBioFS: Machine learning (random forest)-based gene selection [User Manual]

Hanane Hadj-Moussa and Jing Zhang, 2016.

***Note:** This techbull will refer you various websites and to the RBioFS webpage that is hosted on our lab website (http://kenstoreylab.com/?page_id=2542).

I. Installing R and RStudio

To run RBioFS you must first install R on your computer, then install RStudio the user interface for R. Links to both of these programs are on the lab website (www.kensotreylab.com) → Research → Research Tools → RBioFS (http://kenstoreylab.com/?page_id=2542).

A. Installing "R"

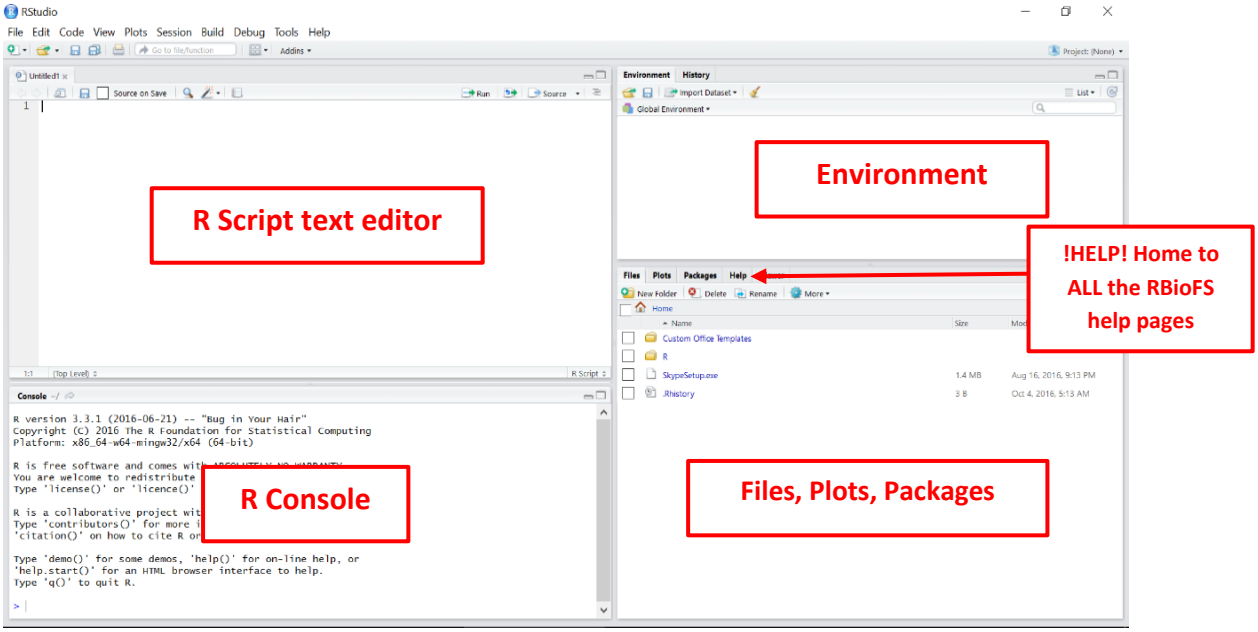
1) To install '**R**' visit (https://www.r-project.org) and then select your CRAN preferred CRAN mirror, we will use the University of Toronto's (http://cran.utstat.utoronto.ca/).
2) Download the version of R that corresponds with your operating system. For example, if you are working on a Windows computer click **Download R for Windows → install R for the first time → Download R 3.3.1 for Windows**
3) Save the installation .exe file → open the file → follow the R for Windows 3.3.1 Setup Wizard's simple installation instructions.
   ***Note:** You should install all the installation components and make sure to accept the default start-up options.

B. Installing RStudio

1) To install **RStudio** visit (https://www.rstudio.com/). Click **RStudio → Desktop → Download RStudio Desktop**
2) Download the RStudio version that corresponds with your operating system. For example, if you are working on a Windows computer click **RStudio 0.99.903 - Windows Vista/7/8/10**
3) Save the installation .exe file → open the file → follow the RStudio Setup Wizard's simple installation instructions.
   ***Note:** Once the program has installed you should locate your RStudio shortcut and place it on your desktop.

II. Setting up RBioFS

Open the RStudio application, this will be the platform you will use to run RBIoFS. The picture below is a breakdown of the different RStudio quadrants and panels. **RStudio is case-sensitive.**



1) To setup the R Script text editor go to **File → New File → R Script**. The text editor is where you should prepare your R script before running your commands in the console. The text editor also allows you to run multiple commands at the same time (as long as you highlight all the commands you want to run).
2) To use the latest version of RBioFS we need to first install **devtools,** this package contains some of the various dependencies that RBioFS needs to run. This is a one-time installation, to install 'devtools' rewrite the command below in the text editor.

```
install.packages("devtools")
```

*Note: You must be connected to the internet. You should rewrite all the commands written in this techbull and not just copy them into the text editor. Microsoft word has altered the formatting and syntax of these codes and if you copy/paste this into RStudio, your command will not run as R Script has unique quotation marks. If you copy/paste then make sure you rewrite your quotation marks.

3) To run your R commands, you have two options:
   a. **R Script text editor**, (preferred): Highlight the command you want to run and press Ctrl+Enter. Or place your text cursor on the command you want to run and press Ctrl+Enter. If you separate your command into different lines then make sure all the different commands are highlighted and that they can
   b. **Console**: Copy your command from your text editor to the Console and press Enter.
4) Once your command has been successfully installed you should receive a message that reads something like this:

```
The downloaded binary packages are in
C:\Users\Hanane\AppData\Local\Temp\Rtmp4gUWDW\downloaded_packages
```

5) Run the following command to install **Bioconductor** (needed to run RBioFS), otherwise skip this step:

```
source("https://bioconductor.org/biocLite.R")
biocLite()
```

For help with installing Bioconductor, visit (https://www.bioconductor.org/install/)

6) Now you are ready to install the RBioFS package. Run the following command:

```
devtools::install_github("jzhangc/git_RBioFS/RBioFS", repos =
  BiocInstaller::biocinstallRepos())
```

7) You may run into a few errors as the current version of **devtools** does not contain all the dependencies that RBioFS needs, which means you need to manually install the missing packages.

Here are a few examples of errors and solutions, in this first one I was missing the **scales** package so I installed it manually:

```
Error in loadNamespace(i, c(lib.loc, .libPaths()), versionCheck = vI[[i]])
: there is no package called 'scales'
  install.packages("scales")
```

After I installed the missing package (**scales)** I re-ran the RBioFS installation command in step 5.

```
Error in loadNamespace(j <- i[[1L]], c(lib.loc, .libPaths()), versionCheck
= vI[[j]]) : there is no package called 'sandwich'
```

Now, I am missing the package (**sandwich)**, so I installed it with the command below and then re-ran the RBioFS installation command from step 5. Repeat this process until you have installed all the missing packages.

```
install.packages("sandwich")
```

8) The next step is to setup your working directory, this is where you will place all your input data files and where all your plots, stats, and graphs will be exported to. In this example my working directory file is on my desktop and it's called "RBioFS stuffs". To setup your working directory run the following command and replace "C:\Users\Hanane\Desktop" with your folder address:

For Windows: `setwd("C:\\Users\\Hanane\\Desktop\\RBioFS stuffs")`

For Mac and Linux: `setwd("C:\Users\Hanane\Desktop\RBioFS stuffs")`

9) In the 'Files, Plots, and Packages' quadrant of the graph you should select 'Packages' and then find RBioFS in the list. Once you click on RBioFS you will be redirected to a page with the RBioFS documentation and all the help files for the different functions. If you get stuck using one of the commands use these help files to trouble shoot.

10) If your RBioFS stops working or there is an error that you have spent the last 3 weeks troubleshooting, then you should email the developer jzhangcad@gmail.com and ask him nicely for help.

File layout

Format: csv

Filter

	SampleID	Condition	dgl-miR-let-7f-5p	dgl-miR-1a-5p	dgl-miR-1b-5p	dgl-miR-10b-5p	dgl-miR-16-3p	dgl-miR-18a-3p	dgl-miR-20a-5p	dgl-miR-21a-3p	dgl-miR-22-5p	dgl-miR-23a-5p	dgl-miR-26a-5p	dgl-miR-27a-5p
Control.1.1	Control.1.1	control	0.11190995	0.002383542	0.002820877	0.05379724	0.009063634	0.08077205	0.13679458	0.6656084	0.03721601	0.03380627	0.46552676	0.01998478
Control.1.2	Control.1.2	control	NA	0.003903241	0.003641357	NA	0.008476861	0.07031616	0.15414938	0.6286915	0.03491059	0.03459015	0.49653253	0.01969414
Control.2.1	Control.2.1	control	0.10776942	0.003421941	0.002178203	0.04467675	0.009056922	0.09235496	0.13899680	0.8224800	0.04485433	0.03801253	0.47693178	0.02621919
Control.2.2	Control.2.2	control	NA	0.003404393	0.003237942	NA	0.011286178	0.09299734	0.15677298	0.7959151	0.04781522	0.03705494	0.51859333	0.02683424
Control.3.1	Control.3.1	control	0.06178795	0.001899035	0.001980194	0.06600755	0.011137533	0.10414508	0.14287847	1.3669317	0.04601862	0.06251752	0.30628346	0.02966780
Control.3.2	Control.3.2	control	NA	0.002950592	0.003332281	NA	0.010973620	0.10856769	0.15407398	1.3620858	0.05105282	0.06552714	0.28648931	0.02691668
Control.4.1	Control.4.1	control	0.10155391	0.005370459	0.002657407	0.05360694	0.009956980	0.12158187	0.25091706	1.0141519	0.04111087	0.04040757	0.47788463	0.03005951
Control.4.2	Control.4.2	control	NA	0.006303695	0.003616004	NA	NA	0.13678671	0.21003632	1.0395482	0.03817431	0.04649464	0.49774289	0.03127912
Torpor.1.1	Torpor.1.1	torpor	0.06219157	0.001090033	0.001017847	0.02511691	0.009941763	0.04249195	0.06427207	0.4138845	0.03260618	0.01915703	0.26681574	0.01096216
Torpor.1.2	Torpor.1.2	torpor	NA	0.001005065	0.001634836	NA	0.009138224	0.04076100	0.05176534	0.4310749	0.03062009	0.01967695	0.08531767	0.01248937
Torpor.2.1	Torpor.2.1	torpor	0.05019095	0.000885359	0.000927175	0.02827481	0.006133757	0.03639792	0.06829312	0.2954151	0.03084727	0.01537112	0.22947118	0.01150312
Torpor.2.2	Torpor.2.2	torpor	NA	0.000951887	0.001027896	NA	0.006946379	0.03589682	0.07208993	0.3040495	NA	0.01594061	0.23718431	0.01067739
Torpor.3.1	Torpor.3.1	torpor	0.03720122	0.000477179	0.001842508	0.03805831	0.007366451	0.03983002	0.04850687	0.3277977	0.04667655	0.02111562	0.16513594	0.01052796
Torpor.3.2	Torpor.3.2	torpor	NA	0.000945255	0.001767801	NA	0.007212473	0.03639792	0.04059032	0.3329776	0.03001228	0.02522184	0.16333062	0.01022526
Torpor.4.1	Torpor.4.1	torpor	0.07417936	0.001973448	0.001550475	0.03300582	0.008382250	0.04152143	0.09470313	0.3784183	0.02936817	0.01976743	0.36551915	0.01410959
Torpor.4.2	Torpor.4.2	torpor	NA	0.001998136	0.001930635	NA	0.008330294	0.04298568	0.09555982	0.3832769	0.03218875	0.02031878	0.36130509	0.01453386

Load the file into R

```
data <- read.csv("data.csv", stringsAsFactors = FALSE)
rownames(data) <- data$SampleID
fsdata <- data[-c(1:2)]
```

(Optional) Data imputation (Random Forest method) and quantile normalization

```
data <-  rbioIMP(data[-c(1:2)],  data$Condition, data$SampleID, method = "rf", iter = 10, ntree = 501) # imputation
data <- rbioNorm(data, correctBG = FALSE) # normalization
fsdata <- t(data)
```

Set up target variable

```
tgt <- as.factor(rep(c("control", "torpor"), each = 8))
```

Initial FS function

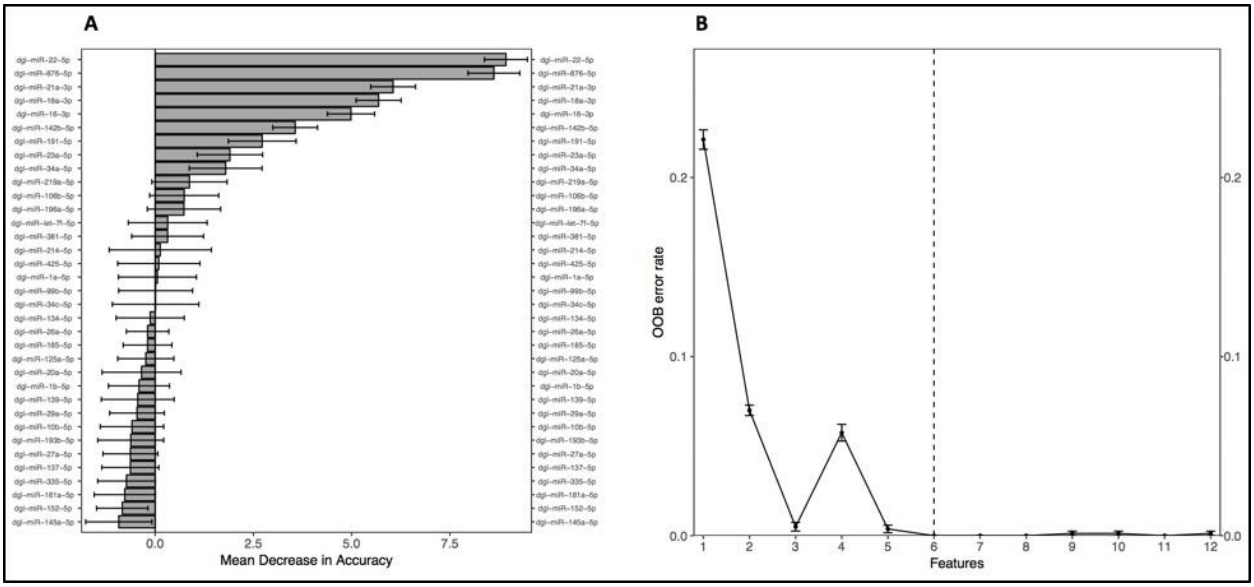
```
rbioRF_initialFS(objTitle = "cvt", fsdata, tgt, nTree = 501, errorbar = "SD", errorbarWidth = 0.4, yTxtSize = 6) # output list name: cvt_initial_FS
```

SFS-like FS function

```
rbioRF_SFS(objTitle = "cvt", cvt_initial_FS$matrix_initial_FS, tgt, nTree = 501, symbolSize = 1, xLabel = "Features", yLabel = "OOB error rate" ) # output list name: cvt_SFS
```

Output

Figures



A – VI ranking after the **rbioRF_initialFS()** step; B – OOB error after the **rbioRF_SFS()** step

Files (.txt)

Initial FS step outputs the results into **cvt.initialFS.txt**. See a truncated portion below:

```
$matrix_initial_FS
           dgl-miR-22-5p dgl-miR-876-5p dgl-miR-21a-3p dgl-miR-18a-3p dgl-miR-16-3p dgl-miR-142b-5p dgl-miR-191-5p
Control.1.1    0.02988938    0.058773318       0.6464010     0.07383765   0.007886430       0.03189701     0.7837999
Control.1.2    0.02988938    0.049030504       0.6464010     0.05877332   0.007886430       0.03189701     0.7837999
Control.2.1    0.03852812    0.052134880       0.7837999     0.07383765   0.007886430       0.02437997     0.5277544
Control.2.2    0.03852812    0.052134880       0.6464010     0.06503953   0.007886430       0.02437997     0.5277544
Control.3.1    0.02988938    0.049030504       0.7837999     0.07383765   0.009068414       0.02772046     0.6464010
Control.3.2    0.03189701    0.046021137       0.7837999     0.07383765   0.007886430       0.02772046     0.5277544
Control.4.1    0.03189701    0.046021137       0.7837999     0.07383765   0.007886430       0.02772046     0.5277544
Control.4.2    0.02988938    0.009068414       0.7837999     0.08295620   0.007886430       0.03189701     0.5277544
Torpor.1.1     0.04903050    0.027720457       0.6464010     0.06503953   0.014003746       0.07383765     0.7837999
Torpor.1.2     0.04602114    0.007886430       0.5277544     0.05877332   0.017259277       0.03346258     0.7837999
Torpor.2.1     0.04602114    0.009068414       0.6464010     0.05877332   0.010759519       0.04903050     0.7837999
Torpor.2.2     0.04903050    0.033462584       0.5277544     0.05213488   0.009068414       0.03189701     0.7837999
Torpor.3.1     0.08295620    0.029889377       0.5277544     0.05877332   0.010759519       0.03189701     0.7837999
Torpor.3.2     0.04108325    0.027720457       0.5277544     0.05213488   0.010759519       0.03583229     0.6464010
Torpor.4.1     0.03852812    0.027720457       0.5277544     0.06190643   0.009068414       0.05213488     0.7837999
Torpor.4.2     0.04602114    0.027720457       0.5277544     0.05213488   0.007886430       0.02988938     0.7837999
           dgl-miR-23a-5p dgl-miR-34a-5p dgl-miR-219a-5p dgl-miR-106b-5p dgl-miR-196a-5p
Control.1.1    0.02772046     0.04602114      0.012874604      0.08295620      0.01400375
Control.1.2    0.02772046     0.04108325      0.010759519      0.10606246      0.01400375
Control.2.1    0.03189701     0.04903050      0.012874604      0.09671394      0.01400375
Control.2.2    0.03346258     0.04602114      0.014003746      0.09671394      0.01075952
Control.3.1    0.04108325     0.03346258      0.007886430      0.10606246      0.01287460
Control.3.2    0.04108325     0.05213488      0.009068414      0.05877332      0.01287460
Control.4.1    0.02988938     0.05213488      0.009068414      0.10606246      0.01725928
Control.4.2    0.03583229     0.04903050      0.010759519      0.07383765      0.01725928
Torpor.1.1     0.02437997     0.05213488      0.007886430      0.12602691      0.01971243
Torpor.1.2     0.02988938     0.05213488      0.009068414      0.12602691      0.01287460
Torpor.2.1     0.02988938     0.06503953      0.007886430      0.10606246      0.02437997
Torpor.2.2     0.02437997     0.07383765      0.007886430      0.10606246      0.01287460
Torpor.3.1     0.03583229     0.04108325      0.007886430      0.04602114      0.01971243
Torpor.3.2     0.03852812     0.04903050      0.007886430      0.12602691      0.01287460
Torpor.4.1     0.02437997     0.06190643      0.010759519      0.04602114      0.01400375
Torpor.4.2     0.02437997     0.05877332      0.010759519      0.12602691      0.01400375

$feature_initial_FS
 [1] "dgl-miR-22-5p"  "dgl-miR-876-5p" "dgl-miR-21a-3p" "dgl-miR-18a-3p" "dgl-miR-16-3p"  "dgl-miR-142b-5p"
 [7] "dgl-miR-191-5p" "dgl-miR-23a-5p" "dgl-miR-34a-5p" "dgl-miR-219a-5p" "dgl-miR-106b-5p" "dgl-miR-196a-5p"

$recur_vi_summary
                      Target        Mean        SD        SEM Rank
dgl-miR-22-5p       dgl-miR-22-5p  8.92779547 0.5460166 0.07721840    1
dgl-miR-876-5p     dgl-miR-876-5p  8.61995805 0.6601663 0.09336161    2
dgl-miR-21a-3p     dgl-miR-21a-3p  6.05579645 0.5698496 0.08058890    3
dgl-miR-18a-3p     dgl-miR-18a-3p  5.68667940 0.5734619 0.08109976    4
dgl-miR-16-3p       dgl-miR-16-3p  4.98389685 0.6000285 0.08485684    5
dgl-miR-142b-5p   dgl-miR-142b-5p  3.56454204 0.5680016 0.08032756    6
dgl-miR-191-5p     dgl-miR-191-5p  2.72445429 0.8626225 0.12199324    7
dgl-miR-23a-5p     dgl-miR-23a-5p  1.90164165 0.8327736 0.11777197    8
dgl-miR-34a-5p     dgl-miR-34a-5p  1.79434591 0.9264354 0.13101776    9
```

SFS step outputs the results into **cvt.SFS.txt**. See a truncated portion below:

```
$selected_features
[1] "dql-miR-22-5p"   "dql-miR-876-5p"  "dql-miR-21a-3p"  "dql-miR-18a-3p"  "dql-miR-16-3p"   "dql-miR-142b-5p"

$feature_subsets_with_min_OOBerror_plus_1SD
[1]  6  7  8 11

$OOB_error_rate_summary
   Features    Mean          SD         SEM
1         1 0.22125 0.038320114 0.005419282
2         2 0.07000 0.020516295 0.002901442
3         3 0.00500 0.017127970 0.002422261
4         4 0.05750 0.033023646 0.004670249
5         5 0.00375 0.014993621 0.002120418
6         6 0.00000 0.000000000 0.000000000
7         7 0.00000 0.000000000 0.000000000
8         8 0.00000 0.000000000 0.000000000
9         9 0.00125 0.008838835 0.001250000
10       10 0.00125 0.008838835 0.001250000
11       11 0.00000 0.000000000 0.000000000
12       12 0.00125 0.008838835 0.001250000

$SFS_matrix
           dql-miR-22-5p dql-miR-876-5p dql-miR-21a-3p dql-miR-18a-3p dql-miR-16-3p dql-miR-142b-5p
Control.1.1    0.02988938    0.058773318       0.6464010     0.07383765   0.007886430       0.03189701
Control.1.2    0.02988938    0.049030504       0.6464010     0.05877332   0.007886430       0.03189701
Control.2.1    0.03852812    0.052134880       0.7837999     0.07383765   0.007886430       0.02437997
Control.2.2    0.03852812    0.052134880       0.6464010     0.06503953   0.007886430       0.02437997
Control.3.1    0.02988938    0.049030504       0.7837999     0.07383765   0.009068414       0.02772046
Control.3.2    0.03189701    0.046021137       0.7837999     0.07383765   0.007886430       0.02772046
Control.4.1    0.03189701    0.046021137       0.7837999     0.07383765   0.007886430       0.02772046
Control.4.2    0.02988938    0.009068414       0.7837999     0.08295620   0.007886430       0.03189701
Torpor.1.1     0.04903050    0.027720457       0.6464010     0.06503953   0.014003746       0.07383765
Torpor.1.2     0.04602114    0.007886430       0.5277544     0.05877332   0.017259277       0.03346258
Torpor.2.1     0.04602114    0.009068414       0.6464010     0.05877332   0.010759519       0.04903050
```

Quick start guide for all-in-one command version

File layout

Format: csv

Filter

	SampleID	Condition	dgl-miR-let-7f-5p	dgl-miR-1a-5p	dgl-miR-1b-5p	dgl-miR-10b-5p	dgl-miR-16-3p	dgl-miR-18a-3p	dgl-miR-20a-5p	dgl-miR-21a-3p	dgl-miR-22-5p	dgl-miR-23a-5p	dgl-miR-26a-5p	dgl-miR-27a-5p
Control.1.1	Control.1.1	control	0.11190995	0.002383542	0.002820877	0.05379724	0.009063634	0.08077205	0.13679458	0.6656084	0.03721601	0.03380627	0.46552676	0.01998478
Control.1.2	Control.1.2	control	NA	0.003903241	0.003641357	NA	0.008476861	0.07031616	0.15414938	0.6286915	0.03491059	0.03459015	0.49653253	0.01969414
Control.2.1	Control.2.1	control	0.10776942	0.003421941	0.002178203	0.04467675	0.009056922	0.09235496	0.13899680	0.8224800	0.04485433	0.03801253	0.47693178	0.02621919
Control.2.2	Control.2.2	control	NA	0.003404393	0.003237942	NA	0.011286178	0.09299734	0.15677298	0.7959151	0.04781522	0.03705494	0.51859333	0.02683424
Control.3.1	Control.3.1	control	0.06178795	0.001899035	0.001980194	0.06600755	0.011137533	0.10414508	0.14287847	1.3669317	0.04601862	0.06251752	0.30628346	0.02966780
Control.3.2	Control.3.2	control	NA	0.002950592	0.003332281	NA	0.010973620	0.10856769	0.15407398	1.3620858	0.05105282	0.06552714	0.28648931	0.02691668
Control.4.1	Control.4.1	control	0.10155391	0.005370459	0.002657407	0.05360694	0.009956980	0.12158187	0.25091706	1.0141519	0.04111087	0.04040757	0.47788463	0.03005951
Control.4.2	Control.4.2	control	NA	0.006303695	0.003616004	NA	NA	0.13678671	0.21003632	1.0395482	0.03817431	0.04649464	0.49774289	0.03127912
Torpor.1.1	Torpor.1.1	torpor	0.06219157	0.001090033	0.001017847	0.02511691	0.009941763	0.04249195	0.06427207	0.4138845	0.03260618	0.01915703	0.26681574	0.01096216
Torpor.1.2	Torpor.1.2	torpor	NA	0.001005065	0.001634836	NA	0.009138224	0.04076100	0.05176534	0.4310749	0.03062009	0.01967695	0.08531767	0.01248937
Torpor.2.1	Torpor.2.1	torpor	0.05019095	0.000885359	0.000927175	0.02827481	0.006133757	0.03639792	0.06829312	0.2954151	0.03084727	0.01537112	0.22947118	0.01150312
Torpor.2.2	Torpor.2.2	torpor	NA	0.000951887	0.001027896	NA	0.006946379	0.03589682	0.07208993	0.3040495	NA	0.01594061	0.23718431	0.01067739
Torpor.3.1	Torpor.3.1	torpor	0.03720122	0.000477179	0.001842508	0.03805831	0.007366451	0.03983002	0.04850687	0.3277977	0.04667655	0.02111562	0.16513594	0.01052796
Torpor.3.2	Torpor.3.2	torpor	NA	0.000945255	0.001767801	NA	0.007212473	0.03639792	0.04059032	0.3329776	0.03001228	0.02522184	0.16333062	0.01022526
Torpor.4.1	Torpor.4.1	torpor	0.07417936	0.001973448	0.001550475	0.03300582	0.008382250	0.04152143	0.09470313	0.3784183	0.02936817	0.01976743	0.36551915	0.01410959
Torpor.4.2	Torpor.4.2	torpor	NA	0.001998136	0.001930635	NA	0.008330294	0.04298568	0.09555982	0.3832769	0.03218875	0.02031878	0.36130509	0.01453386

Install RBioFS

```
install.packages("devtools") # (optional) if no devtools is installed
devtools::install_github("jzhangc/git_RBioFS/RBioFS", repos = BiocInstaller::biocinstallRepos())
```

Set working directory

```
setwd("working directory")
```

One line command

```
rbioFS(file = "test.csv", impute = TRUE, imputeMethod = "mean", quantileNorm = TRUE, nTree = 501,
initialFS_errorbar = "SD", plot = TRUE)
```