mirTarRnaSeq

Mercedeh Movassagh

1 November 2022

Package

mirTarRnaSeq 1.6.0

Contents

1 Introduction

mirTarRnaSeq is a package for miRNA and mRNA interaction analysis through regression and correlation approaches supporting various modeling approaches (gaussian, poisson, negative binomial, zero inflated poisson or negative binomial models for the data). miRNA and mRNA sequencing data are analysed from the same experiment (condition or time point data) and mRNA targets of the miRNAs from same experiment will be identified using statistical approaches.

The example data set for the first approach is 25 matching miRNA and mRNA EBV positive samples from TCGA identified as high EBV load based on Movassagh et al, Scientific Reports, 2019 paper. We attempt to identify the EBV miRNA targets on EBV genome (part1).

The second example set set is the simulated mouse fibroblast differentiated to muscle cells in three time points. Here, we try to identify mRNA targets of miRNA expressed at various time points (parts 2 and 3).

1.1 Data upload

1.1.1 mirTarRnaSeq accepts data in dataframe or table formats

- For the first approach (Part1) we use a table of differential expressed mRNA genes in EBV from TCGA stomach cancer samples with high levels of EBV miRNA expression.
- Next we use a list of normalized (tpm) EBV miRNA expression data from the same samples. The user has the option to use count data and model accordingly or use tzTrans() function for zscore normalization and then model.
- For the second part (Part2, Part3) of the experiment we are using two tables of differentially expressed mRNA and miRNA sequencing fold change results from mouse time point specific differentiation experiments.
- The example data are also available at https://doi.org/10.5281/zenodo.6371713.

2 Part1 - miRNA mRNA regressions across sample cohorts

The users can utilize TPM/RPKM or count data for this section of analysis as long as if the data is normalized (TPM/RPKM) is used for the mRNA expression files the miRNA expression files are also normalized (TPM). The format of the files for miRNA and mRNA needs to be odd the class dataframe with the sample names as colnames and mRNA names as rownames. Note, for this section of analysis we have only included all differentially expressed mRNAs for the analysis but the user can included all mRNAs. However, the user should realize if they choose to use all mRNAs expressed the analysis will take a longer time.

2.1 Uploading data into the application. The example data can be found in the test folder under package.

```
# Helper function to access test files in extdata
get_test_file <- function(file_name) {
  if (missing(file_name)) {
    return(system.file("extdata", "test", package="mirTarRnaSeq"))
```

```
    }
    return(system.file("extdata", "test", file_name, package="mirTarRnaSeq"))
}

# Read input files
DiffExp<-read.table(get_test_file("EBV_mRNA.txt.gz"), as.is=TRUE, header=TRUE, row.names=1)
miRNAExp<-read.table(get_test_file("EBV_miRNA.txt.gz"), as.is=TRUE, header=TRUE, row.names=1)
```

2.2   Get miRanda file

We currently support miRanda runs (potential miRNA target parsing by score,interaction energy, and interaction or miRNA length) on seven species ("Human", "Mouse", "Drosophila", "C.elegans", "Epstein_Barr" (EBV), "Cytomegalovirus" (CMV) and "Kaposi_Sarcoma" (KSHV)). We also support the viral miRNAs targeting human genes for EBV "Epstein_Barr_Human", CMV "CMV_Human" and KSHV "KSHV_Human". 1) Here we first import the relevant miRanda file . 2) We only keep targets which are also targets of EBV miRNAs based on our EBV miRanda file. Note, for other organisms (not provided by the pacakged) the appropriate format for miRanda input file is V1 column: name of the miRNA, V2 column: name of the predicted gene, V3: column score/rank of the miranda interaction (or any other score for interaction if TargetScan (for TargetScan prediction user can provide context++ score and threshold accordingly) is used, note these three columns are required), V4 column: folding energy of miRNA-mRNA interaction. V5 column: target Identity value and V6 column: miRNA idnetity value. (Note, all these values are available after miRanda analysis but V1-V3 must be provided not matter the prediction method)

```
miRanda <- getInputSpecies("Epstein_Barr", threshold = 140)
DiffExpmRNASub <- miRanComp(DiffExp, miRanda)
```

2.3   Select miRNA

```
miRNA_select<-c("ebv-mir-BART9-5p")
```

2.4   Combine the mRNA and miRNA file and define boundaries ans specify which mRNA and miRNA files in the combined file.

```
Combine <- combiner(DiffExp, miRNAExp, miRNA_select)
geneVariant <- geneVari(Combine, miRNA_select)
```

Running various univariate models(1 to 1 miRNA-mRNA relationships) with various miRNA-mRNA distribution assumptions for either 1 miRNA and 1 mRNA relationship chosen/selected by the user or across all miRNAs and mRNAs in the expression dataframes. Note,the users can choose to run any of the available distribution model assumptions (glm_poisson, glm_gaussian, glm_nb (negative binomial), glm_zeroinfl (zero inflated model or zero inflated negative binomial)).

2.5 Run a one to one miRNA/mRNA gaussian regression model (univariate model for 1 miRNA and 1 mRNA).

Here LMP_1 is EBV mRNA and the ebv-mir-BART9-5p is the miRNA and we are running a glm poisson model.

```
j <- runModel(`LMP-1` ~ `ebv-mir-BART9-5p`,
              Combine, model = glm_poisson(),
              scale = 100)
# Print P value of poisson model of association between between LMP_1 and
# ebv-mir-BART9-5p
print(modelTermPvalues(j))
#> `ebv-mir-BART9-5p`
#>       9.344338e-06
```

2.6 Running Gaussian model over all individual miRNA mRNA models (univatiate model for every mRNA and miRNA relationship across the input dataset with Gaussian distribution assumptions)

```
blaGaus <- runModels(Combine,
                     geneVariant, miRNA_select,
                     family = glm_gaussian(),
                     scale = 100)
# Plot the regression relationship for the mRNA and miRNAs (BHLF1 is the EBV
# mRNA). Note, these are standard quality check plot outputs from plot
# regression.
par(oma=c(2,2,2,2))
par(mfrow=c(2,3),mar=c(4,3,3,2))
plot(blaGaus[["all_models"]][["BHLF1"]])
plot(modelData(blaGaus[["all_models"]][["BHLF1"]]))
# To comprehend the performance of all models we look at Akaike information
# criterion (AIC) across all miRNA-mRNA model performances and then look at
# the density plots. Note, the models with lower comparitive AICs have better
# performace.
G <- do.call(rbind.data.frame, blaGaus[["AICvalues"]])
names(G) <- c("AIC_G")
# Low values seems like a reasonable model
plot(density(G$AIC_G))
```
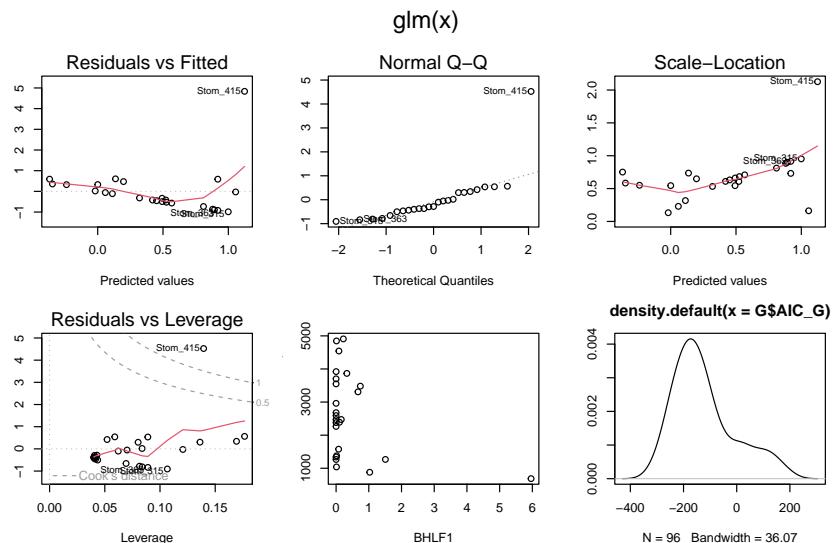
```
# Print out the AIC of all miRNA-mRNA models ( All observed AIC values for the
# miRNA-mRNA models)
GM <- melt(G)
#> No id variables; using all as measure variables
```

2.7 Running poisson model over all individual miRNA mRNA models (univatiate model for every mRNA and miRNA relationship across the input dataset with poisson distribution assumptions)

```
blaPois <- runModels(Combine,
                     geneVariant, miRNA_select,
                     family = glm_poisson(),
                     scale = 100)
par(oma=c(2,2,2,2))
par(mfrow=c(2,3),mar=c(4,3,3,2))
plot(blaPois[["all_models"]][["LMP-2A"]])
plot(modelData(blaPois[["all_models"]][["LMP-2A"]]))
P <- do.call(rbind.data.frame, blaPois[["AICvalues"]])
names(P) <- c("AIC_Po")
PM <- melt(P)
#> No id variables; using all as measure variables
```

glm(x)

Residuals vs Fitted | Normal Q–Q | Scale–Location

Residuals vs Leverage
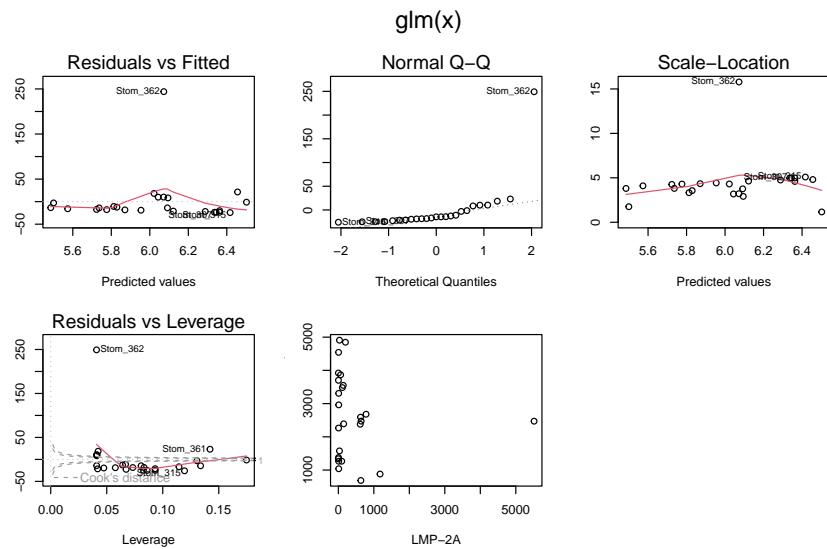
2.8 Running negative binomial model over all individual miRNA mRNA models (univatiate model for every mRNA and miRNA relationship across the input dataset with negative bionomial distribution assumptions).

```
blaNB <- runModels(Combine,
                   geneVariant, miRNA_select,
                   family = glm_nb(), scale = 100)
par(mar=c(4,3,3,2))
plot(modelData(blaNB[["all_models"]][["BALF1"]]))
```
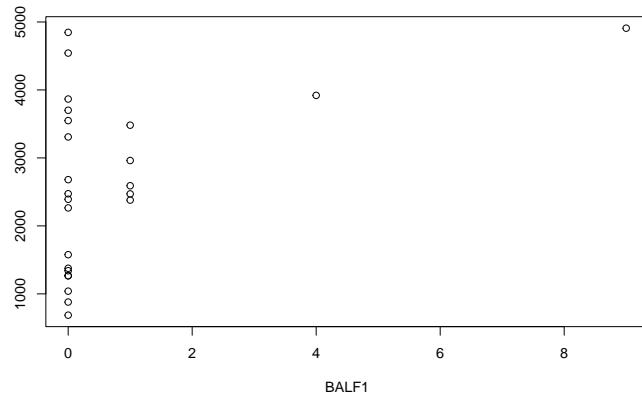
```
B <- do.call(rbind.data.frame, blaNB[["AICvalues"]])
names(B) <- c("AIC_NB")
BM <- melt(B)
#> No id variables; using all as measure variables
```

2.9 Running zero inflated negative binomial model over all individual miRNA mRNA models (univatiate model for every mRNA and miRNA relationship across the input dataset with zero inflated negative bionomial distribution assumptions).

```
blazeroinflNB <- runModels(Combine, geneVariant,
                           miRNA_select,
                           family = glm_zeroinfl(dist = "negbin"),
                           scale = 100)
# To test AIC model performance
ZNB <- do.call(rbind.data.frame, blazeroinflNB[["AICvalues"]])
names(ZNB) <- c("AIC_ZNB")
par(mar=c(4,3,3,2))
plot(density(ZNB$AIC_ZNB))
```

density.default(x = ZNB$AIC_ZNB)

N = 69   Bandwidth = 14.92
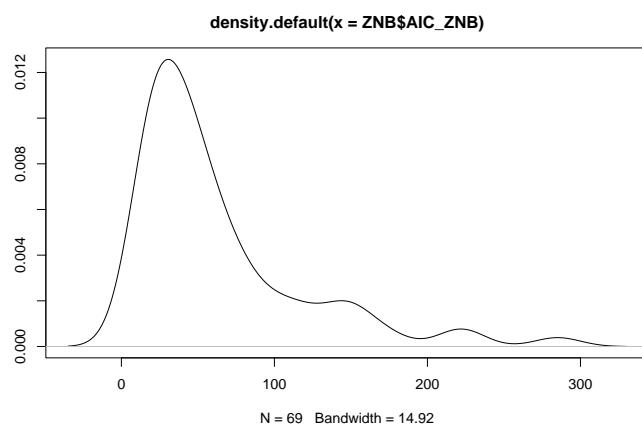
```
ZNBM<-melt(ZNB)
#> No id variables; using all as measure variables
```

2.10 Running zero inflated poisson binomial model over all individual miRNA mRNA models (univatiate model for every mRNA and miRNA relationship across the input dataset with zero inflated poisson distribution assumptions).

```
blazeroinfl <- runModels(Combine, geneVariant,
                         miRNA_select,
                         family = glm_zeroinfl(),
                         scale = 100)
# To test AIC model performance
Zp <- do.call(rbind.data.frame, blazeroinfl[["AICvalues"]])
names(Zp) <- c("AIC_Zp")
par(mar=c(4,3,3,2))
plot(density(Zp$AIC_Zp))
```

density.default(x = Zp$AIC_Zp)

N = 80   Bandwidth = 22.82
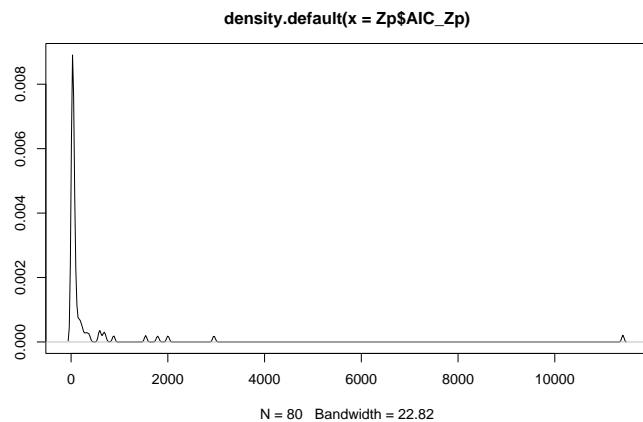
```
ZpM <- melt(Zp)
```

2.11   Including Plots for all models to decide which to use

```
bindM <- rbind(PM, BM, GM, ZpM, ZNBM)
p2 <- ggplot(data = bindM, aes(x = value, group = variable,
                               fill = variable)) +
  geom_density(adjust = 1.5, alpha = .3) +
  xlim(-400, 2000)+
  ggtitle("Plot of of AIC for ebv-mir-BART9-5p regressed all mRNAs ")+
  ylab("Density")+ xlab ("AIC Value")
p2
```

Plot of of AIC for ebv−mir−BART9−5p regressed all m
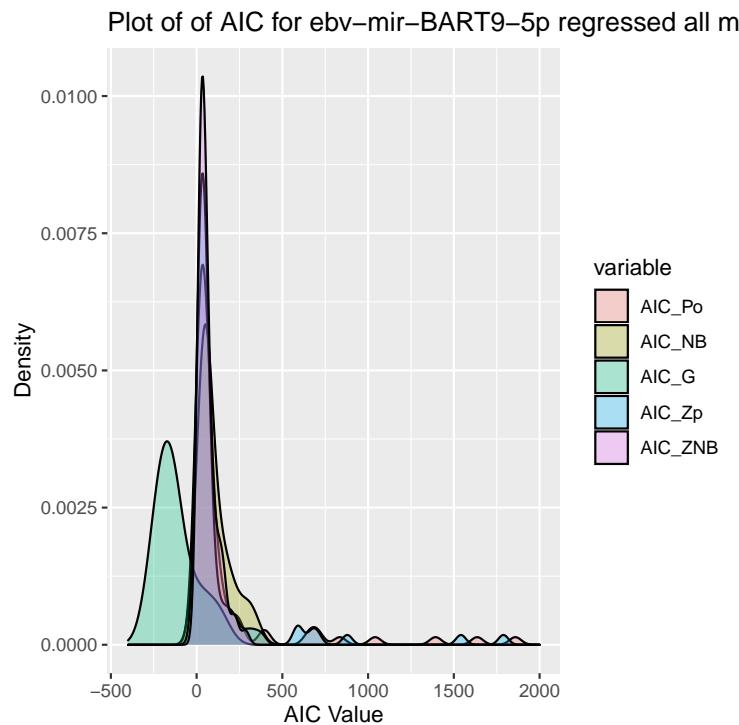
2.12 The user can decide to use runModels() with glm_multi() (with multi and inter mode options)

When using glm_multi(), where all available models will be run, the AICs will be compared and the best model will be chosen based on the miRNA-mRNA model AIC score. In the example bellow we are using the mode= "multi" option for combination of 2 miRNAs (multivariate model) for interaction model the user can choose the mode= "inter" option. Note all_coeff parameters defaults TRUE all interactions (only negative miRNA-mRNA relationships) are reported. More comments on the mode is provided in the next section of the vignette. If all miRNA-mRNA relationships are wanted this parameter can be set to false.

```r
miRNA_select<-c("ebv-mir-BART9-5p","ebv-mir-BART6-3p")
Combine <- combiner(DiffExp, miRNAExp, miRNA_select)
geneVariant <- geneVari(Combine, miRNA_select)
MultiModel <- runModels(Combine, geneVariant,
                        miRNA_select, family = glm_multi(),
                        mode="multi", scale = 10)
# Print the name of the models used for the analysis (note the printed outputs
# are the number of models ran by various models based on the AIC scores using
# the glm_multi())
print(table(unlist(lapply(MultiModel$all_models, modelModelName))))
TRUE
TRUE glm_gaussian        glm_nb
TRUE           95             1
```

2.13 GLM multi and GLM inter

In GLM multi (multinomial), (model="multi"), the user can choose to run as many selected miRNAs they choose against all mRNAs datasets.In this example we select two miRNAs. The user can also select particular mRNAs to run this analysis on. We recommend the user always chooses the number of miRNAs for the multinational models and not run it across all the dataset as the analysis could take a very long time. If the user chooses to use more than two miRNAs for the multinomial model they should assign specific miRNAs and we recommend running it on a high memory machine. In GLM multi (synergy) model (mode="inter"), miRNA interactions. In this example we select two miRNAs. The user can also select particular mRNAs to run this analysis on. We recommend the user always chooses the number of miRNAs for the multinational models and not run it across all the dataset as the analysis could take a very long time. If the user chooses to use more than two miRNAs for the multinomial model they should assign specific miRNAs and we recommend running it on a high memory machine.

```r
miRNA_select<-c("ebv-mir-BART9-5p","ebv-mir-BART6-3p")
Combine <- combiner(DiffExp, miRNAExp, miRNA_select)
geneVariant <- geneVari(Combine, miRNA_select)
InterModel <- runModels(Combine,
                        geneVariant, miRNA_select,
                        family = glm_multi(
                          models=list(glm_gaussian,
                          glm_poisson())),mode="inter", scale = 10)

# Print the name of the models used for the analysis (note, you can see although
```

```
# we have defined for the models to be run using either gaussian or poisson
# options, mirTarRnaSeq chooses the poisson model for all as it performs with a
# better AIC (lower AIC value) for all miRNA-mRNA interactions)
print(table(unlist(lapply(InterModel$all_models, modelModelName))))
TRUE
TRUE glm_gaussian
TRUE           96
```

2.14 Running all miRNA and mRNA combinations at the same time

Note, for "inter" and "multi" mode options we only support combination of 2 if more than two relationships are of interest, we recommend selecting the miRNAs and running the previously described runModels function due to complication of the models and time consumption.

```
vMiRNA<-rownames(miRNAExp)
# Note, the user can run all miRNAs but for speed reasons we have chosen the
# first 5 here for mirnas input for the analysis.
All_miRNAs_run<-runAllMirnaModels(mirnas =vMiRNA[1:5] ,
                                  DiffExpmRNA = DiffExpmRNASub,
                                  DiffExpmiRNA = miRNAExp,
                                  miranda_data = miRanda,prob=0.75,
                                  cutoff=0.05,fdr_cutoff = 0.1, method = "fdr",
                                  family = glm_multi(), scale = 2, mode="multi")
TRUE 1: ebv-mir-BART1-3p, ebv-mir-BART1-5p
TRUE 2: ebv-mir-BART1-3p, ebv-mir-BART10-3p
TRUE 3: ebv-mir-BART1-3p, ebv-mir-BART10-5p
TRUE 4: ebv-mir-BART1-3p, ebv-mir-BART11-3p
TRUE 5: ebv-mir-BART1-5p, ebv-mir-BART10-3p
TRUE 6: ebv-mir-BART1-5p, ebv-mir-BART10-5p
TRUE 7: ebv-mir-BART1-5p, ebv-mir-BART11-3p
TRUE 8: ebv-mir-BART10-3p, ebv-mir-BART10-5p
TRUE 9: ebv-mir-BART10-3p, ebv-mir-BART11-3p
TRUE 10: ebv-mir-BART10-5p, ebv-mir-BART11-3p


#select significant genes
hasgenes <- lapply(All_miRNAs_run, function(x) nrow(x$SigFDRGenes)) > 0
All_miRNAs_run <- All_miRNAs_run[hasgenes]
print(table(unlist(lapply
                (All_miRNAs_run[[1]][["FDRModel"]][["all_models"]],
                  modelModelName))))
TRUE
TRUE          glm_gaussian               glm_nb glm_zeroinfl_poisson
TRUE                   79                    5                    1
# Print specific models for specific miRNAs (in this example the significant
# multivariate model for ebv-mir-BART1-5p and ebv-mir-BART11-3p )
print(
  table(
    unlist(lapply(
      (All_miRNAs_run[["ebv-mir-BART1-5p and ebv-mir-BART11-3p"]]
```

```
                       [["FDRModel"]]
                       [["all_models"]]),
         modelModelName))))
TRUE
TRUE         glm_gaussian                glm_nb  glm_zeroinfl_negbin
TRUE                  79                      3                    2
TRUE glm_zeroinfl_poisson
TRUE                   1
```

2.15 One2manySponge

miRNA-mRNA sparse partial correlation prediction using elastic net and compatibility with SPONGE package.Note, we do not recommend using this method for low number of samples (20 or less), or in miRNAs/mRNAs with low variance as the sample size increases the confidence in this analysis increases same goes for miRNA and mRNA variance. Optimal analysis using this method is in 100 or more samples in high variance miRNAs/mRNAs. For viral analysis or samples with low mRNA miRNA variance we recommend using part1 of mirTarRnaSeq or narrowing down the miRNAs and mRNAs to high variance ones.

```
# Make miRanda file compatible with SPONGE package mir_predicted_targets()
sponge_mir_input<-miranda_sponge_predict(miRNAExp,DiffExp,miRanda)

#Perform sparse partial correlation for miRNA-mRNA relationships
one2many_output<-one2manySponge(miRNAExp,DiffExp,sponge_mir_input)
#> Loading required package: Matrix
#>
#> Attaching package: 'Matrix'
#> The following objects are masked from 'package:tidyr':
#>
#>     expand, pack, unpack
#> Loaded glmnet 4.1-4
#> Loading required package: foreach
#> Loading required package: rngtools
```

3 Part2 - Identify miRNA mRNA correlations across 3 or more time points

Note for this analysis we need fold change data for time points or control versus condition. Hence, a differential expression (DE) analysis needs to be performed before proceeding this analysis (These values should be provided for all miRNA and mRNA in the DE expression and not only the significantly DE miRNAs/mRNAs). Here we are looking at differential expression (DE) files between three time points. The format of each timepoint/control vs condition file needs to be Gene/miRNA names as the first column, log2FC or logfoldchange (FC), (or any other FC metrics as long as for both miRNA, and mRNA the same metrics is used) for column two. The pvalue assigned to the gene(mRNA) expression after the differential expression analysis on the third column.For the miRNA file, the user needs to assign Gene names on the first column and the representative log2FC or logfoldchange (FC) on the second column.

Load files from test directory or you can load individually and feed them in separately in a list: list[(mRNA1,mRNA2,mRNA)]

```
files <- local({
  filenames <- list.files(path=get_test_file(), pattern="^.*\\.txt\\.gz$",
                          full.names=TRUE)
  files <- lapply(filenames, read.table, as.is=TRUE, header=TRUE, sep="\t")
  names(files) <- gsub("^.*/(.*)\\.txt\\.gz$", "\\1", filenames)
  return(files)
})
```

3.1   Get mRNAs

```
mrna_files <- files[grep("^mRNA", names(files))]
```

3.2   Get mRNAs with particular fold change

```
mrna_files <- files[grep("^mRNA", names(files))]
mrna <- one2OneRnaMiRNA(mrna_files, pthreshold = 0.05)$foldchanges
```

3.3   Get all miRNAs

```
mirna_files <- files[grep("^miRNA", names(files))]
mirna <- one2OneRnaMiRNA(mirna_files)$foldchanges
```

3.4   Get mRNA miRNA correlation

```
corr_0 <- corMirnaRna(mrna, mirna,method="pearson")
```

3.5   Make a background distribution correlation

```
outs <- sampCorRnaMirna(mrna, mirna,method="pearson",
                        Shrounds = 100, Srounds = 1000)
```

3.6   Plot density plots

Density plot for background and corrs in our data. Note grey is the background distribution and red is the actual data.

```
#Draw density plot
mirRnaDensityCor(corr_0, outs)
```

density.default(x = corrS)
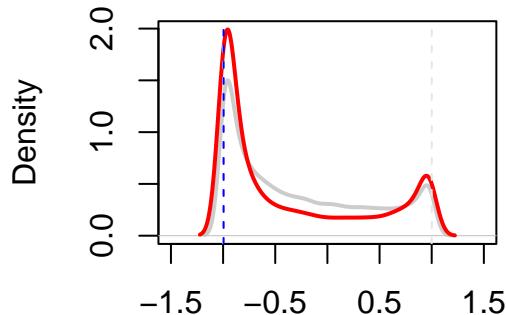
N = 100000   Bandwidth = 0.05945

3.7   Get correlations below threshold

```
#Identify significant correlation
sig_corrs <- threshSig(corr_0, outs,pvalue = 0.05)
```

3.8   Get mouse miRanda data

```
#Import concordant miRanda file
miRanda <- getInputSpecies("Mouse", threshold = 150)
```

3.9   mRNA miRNA correlation heatmap

Correlation heatmap for cor equal or less than -0.7. Note upperbound for heatmap should be always less than the correlation threshold.

```
#Extract your target correlations based on miRanda and correlation threshold.
newcorr <- corMirnaRnaMiranda(mrna, mirna, -0.7, miRanda)
mirRnaHeatmap(newcorr,upper_bound = -0.6)
```

Default mRNA miRNA heatmap
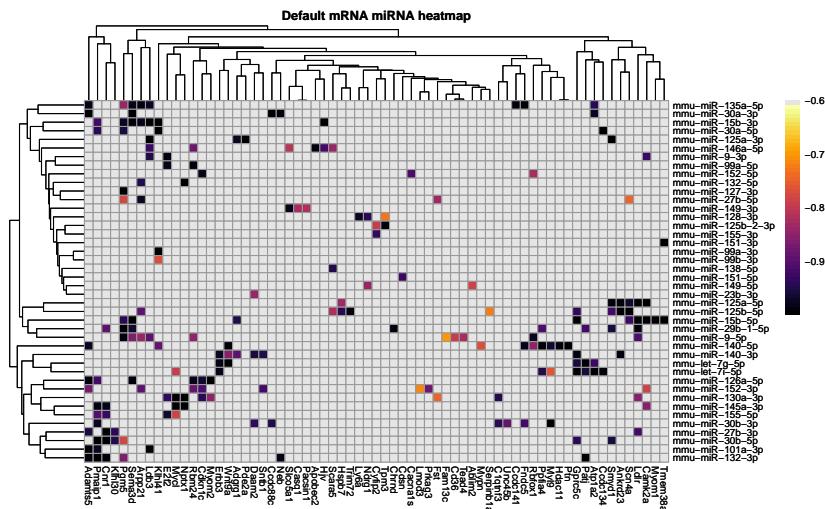
3.10 Get intersection of miRanda

Get miRanda intersection and significant miRNA and mRNA interactions and the plot it.

```
#Make final results file for significant
#correlations intersecting with miRanda file
results <- miRandaIntersect(sig_corrs, outs, mrna, mirna, miRanda)
#Draw correlation heatmap
p<- mirRnaHeatmap(results$corr,upper_bound =-0.99)
p
```


Default mRNA miRNA heatmap
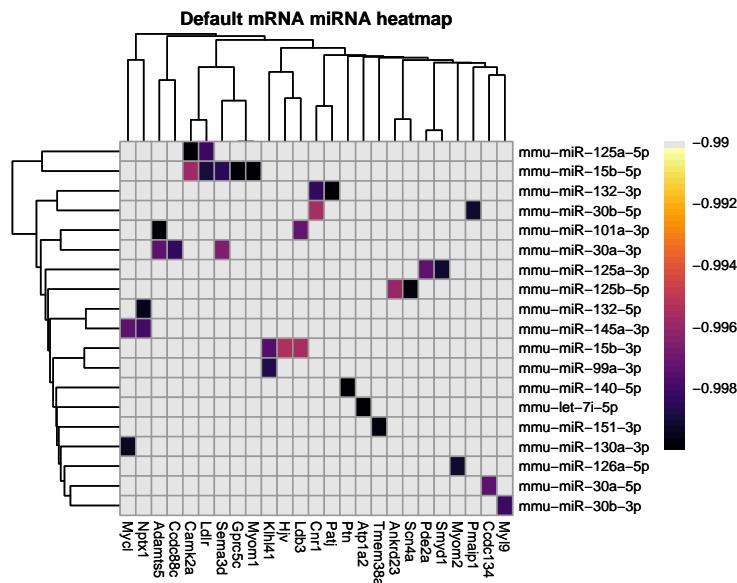
4 Part3 - Identify significant miRNA mRNA relationships for 2 time points

4.1 Import data

```
files <- local({
  filenames <- list.files(path=get_test_file(), pattern="^.*\\.txt\\.gz$",
                          full.names=TRUE)
  files <- lapply(filenames, read.table, as.is=TRUE, header=TRUE, sep="\t")
  names(files) <- gsub("^.*/(.*)\\.txt\\.gz$", "\\1", filenames)
  return(files)
})
```

4.2 Only look for time point difference 0-5

```
mirna_files <- files[grep("^miRNA0_5", names(files))]
mrna_files <- files[grep("^mRNA0_5", names(files))]
```

4.3 Get fold changes above thereshold

```
# Parse Fold Change Files for P value and Fold Change.
mrna <- one2OneRnaMiRNA(mrna_files, pthreshold = 0.05)$foldchanges
mirna <- one2OneRnaMiRNA(mirna_files)$foldchanges
```

4.4 Estimate miRNA mRNA differences based on Fold Change

```
# Estimate the miRNA mRNA FC differences for your dataset
inter0 <- twoTimePoint(mrna, mirna)
```

4.5 Make background distribution

```
#Make a background distribution for your miRNA mRNA FC differences
outs <- twoTimePointSamp(mrna, mirna,Shrounds = 10 )
```

4.6 miRanda data import

```
#Import concordant miRanda file
miRanda <- getInputSpecies("Mouse", threshold = 140)
```

4.7 Identify relationships below threshold

```
#Identify miRNA mRNA relationships bellow a P value threshold, default is 0.05
sig_InterR <- threshSigInter(inter0, outs)
```

4.8 miRanda intersection with results

```
#Intersect the mirRanda file with your output results
results <- mirandaIntersectInter(sig_InterR, outs, mrna, mirna, miRanda)
```

4.9 Make dataframe and plots

```
#Create a results file for heatmap
final_results <- finInterResult(results)
#Draw plots of miRNA mRNA fold changes for your results file
par(mar=c(4,4,2,1))
drawInterPlots(mrna,mirna,final_results)
```
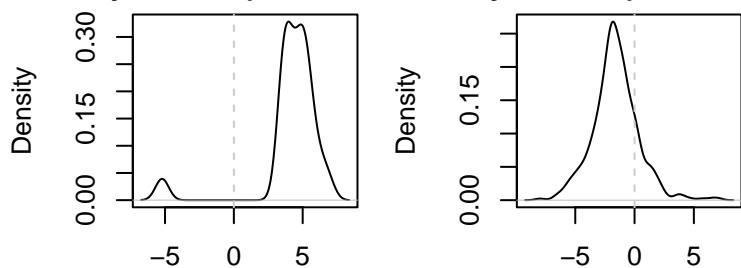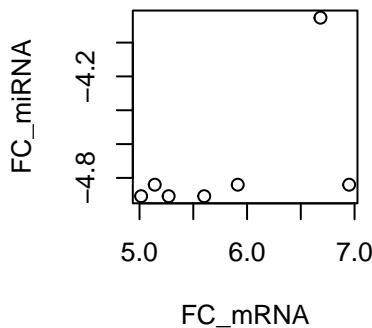

4.10 mRNA miRNA heatmap of miRNA mRNA FC differences

Heatmap for p value significant miRNA mRNA fold change differences when compared to backgound

```
CorRes<-results$corrs
#Draw heatmap for miRNA mRNA significant differences
#Note: you do not have to use the upper_bound function unless you want
#investigate a particular range for miRNA mRNA differences/relationships
mirRnaHeatmapDiff(CorRes,upper_bound = 9.9)
```

Default mRNA miRNA heatmap