Disrupted pathways by miRNAs

This vignette gives some instructions and examples regarding how to apply the miRNA-pathway algorithm to TCGA data.

First, Seq data must be downloaded from the TCGA data portal (https://portal.gdc.cancer.gov/projects). We downloaded Level 3 data sequenced on IlluminaHiSeq platforms for RNASeqV2 (gene expression) and miRNASeq (miRNA expression) for Breast (BRCA), Liver (LIHC), Lung (LUSC), and Prostate (PRAD) cancers. This data is not provided in the repo due to its size. To follow this vignette, download the Seq data for your cancer of choice and place it in whichever directory you wish. You may also download clinical data for exploratory purposes.

This repo is assumed to apply to Breast cancer, in which the raw data files live in rawdata/. Update the file paths and cancer types to wherever the Seq data have been downloaded as necessary.

Directory structure

The directory structure is pretty self-explanatory and consists of three main subdirectories: some provided R data objects in data/ , analysis scripts in scripts/ , and source functions live in source/ .

```
data/
rawdata/
scripts/
source/
```

In addition, source functions are split into two separate files, one called CombineData.R to combine raw data into working R data objects, and another called AnalysisFunctions.R for data analysis.

Data processing

To begin the vignette, run R script 1_CreateDataObjects.R using your working paths to combine raw data files into R data objects and save them. Save them using your preferred naming conventions. Just remember to update variable names in all other scripts to be consistent!

```
R CMD BATCH scripts/1_CreateDataObjects.R
```

Then run R script 2_NormalizeData.R to filter out samples that are not tumor or normal tissue and convert the RNASeq data into Transcripts per Million (TPM).

```
R CMD BATCH scripts/2_NormalizeData.R
```

For the sake of convenience, we have included a Breast Cancer (BRCA) processed ExpressionSet after running 2_NormalizeData.R.

```
data/BRCAEsetTPM.RData
```

The data in this Breast cancer ExpressionSet looks something like this:

```
##        TCGA-E9-A1RD-11A-33R-A157-07 TCGA-E9-A1RC-01A-11R-A157-07
## A1BG                     0.62407478                    2.7469246
## A2LD1                    2.39609655                    0.9146851
## A2ML1                   -4.62138146                   -1.9596998
## A2M                     10.10791034                    6.8788068
## A4GALT                   4.63286154                    4.1560485
## A4GNT                   -0.09720359                   -5.9819807
##        TCGA-AC-A8OP-01A-11R-A36F-07 TCGA-BH-A0B1-01A-12R-A056-07
## A1BG                       3.523406                     3.842809
## A2LD1                      4.081224                     2.508258
## A2ML1                     -5.953398                     1.303028
## A2M                        8.714892                     8.645336
## A4GALT                     5.857923                     4.627991
## A4GNT                     -2.910293                    -3.581874
```

Isomap dimension reduction

In step 3, we run 3_ReducePathways.R on BRCAEsetTPM.RData, the provided ExpressionSet. This script systematically subsets all gene expression data by pathway genes from the KEGG database. For each pathway, it computes the Isomap embedding, the Pathway Activity Summary (PAS). The meat of the script is in:

```r
BRCAPathIso <- mclapply(GEpaths, function(pathway) {
  knn <- findKisomap(pathway, K = seq(4,20), scale = FALSE, plot = FALSE)
  pathway <- scale(pathway)
  distX <- as.matrix(dist(pathway))
  iso <- isomap(distX, ndim = 6, k = knn$k)
  iso <- c(iso, k = knn$k)
  return(iso)
}, mc.cores = 6, mc.preschedule = FALSE)
save(BRCAPathIso, file = "data/BRCAPathIso.RData")
```

Notice this has been parallelized across multiple cores using the mclapply function (utilizing 6 cores in this case). We ramp the free parameter in Isomap, k, to find its "optimal" value as defined in our manuscript using findKisomap(). This computation is fairly intensive for a personal machine and has been run on a computational cluster. You can change the parallelization parameter to suit your computational needs.

We can easily alter the code to run the same k value (k = 8 in this case) for every pathway and thereby remove ramping. This computation will be faster (but not "optimal" in our sense) and can be run on a personal machine.

```r
IsoEXAMPLE <- lapply(GEpaths, function(pathway) {
  pathway <- scale(pathway)
  distX <- as.matrix(dist(pathway))
  iso <- isomap(distX, ndim = 6, k = 8)
  iso <- c(iso, k = 8)
  return(iso)
})
```

miRNA-pathway correlation

In step 4, we correlate the PAS with miRNA expression class-conditionally, for all miRNA-pathway pairs. Afterwards, we compute the difference in tumor vs. normal correlation coefficients as a measure of disruption of pathway activity by miRNAs. Significance is assessed by permutation testing. We can run this code by:

```
R CMD BATCH scripts/4_CorlMiRNAsPathways.R
```

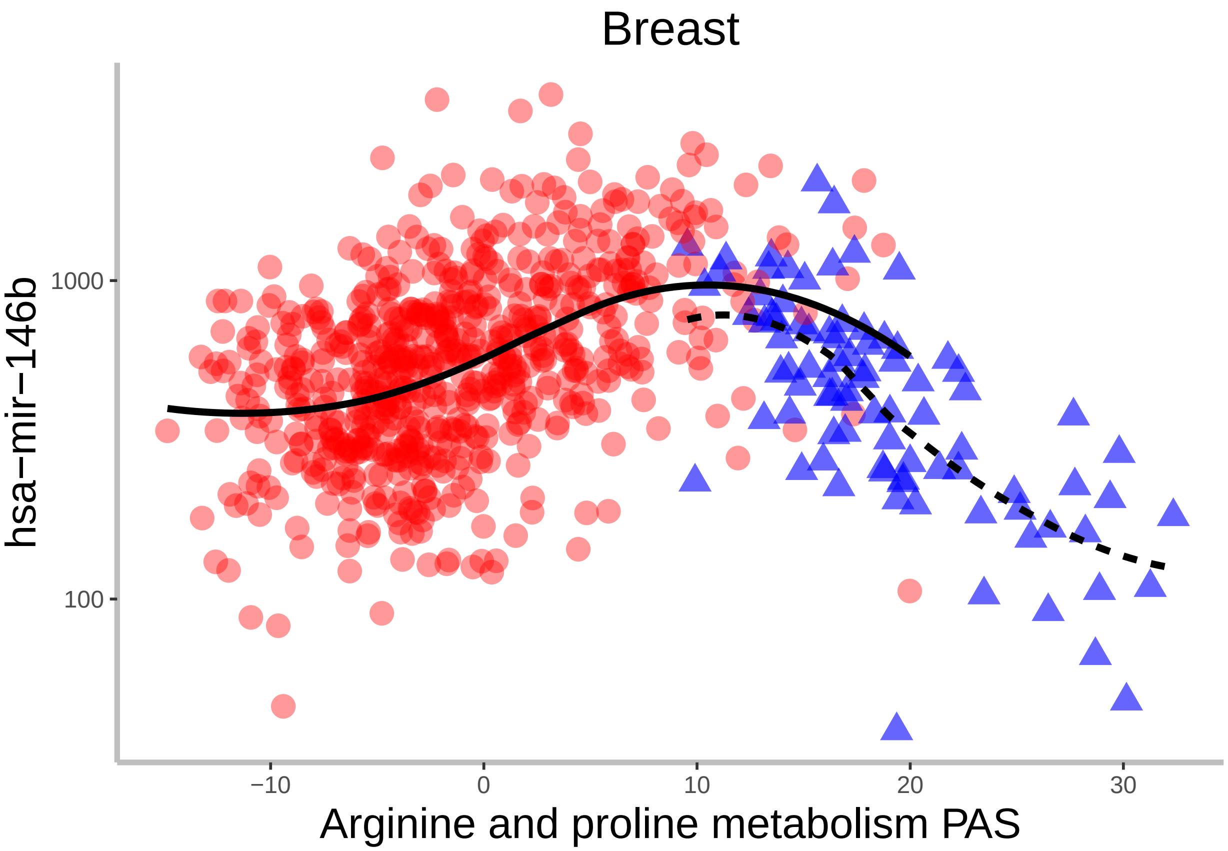
4_CorlMiRNAsPathways.R may take a little while to run due to resampling. The output is a list as shown below,

```
## List of 3
##  $ trueCorlDiffs: num [1:321, 1:223] 0.478 0.48 0.469 0.469 0.872 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:321] "hsa-let-7a-1" "hsa-let-7a-2" "hsa-let-7a-3" "hsa-let-7b" ...
##   .. ..$ : chr [1:223] "04610" "00232" "00983" "01100" ...
##  $ pvals        : num [1:321, 1:223] 0.000999 0.000999 0.000999 0.000999 0.000999 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:321] "hsa-let-7a-1" "hsa-let-7a-2" "hsa-let-7a-3" "hsa-let-7b" ...
##   .. ..$ : chr [1:223] "04610" "00232" "00983" "01100" ...
##  $ commonSamples: chr [1:758] "TCGA-A1-A0SB-01" "TCGA-A1-A0SD-01" "TCGA-A1-A0SF-01" "TCGA-A1-A0SG-01
```

composed of a matrix “trueCorlDiffs” of miRNA-pathway correlation differences, a matrix “pvals” that displays their significance by resampling (p-values), and a vector “commonSamples” of sample names they have in common that were used in their computation.

Making Plots

After having run this analysis on Breast cancer, we now have all the data to plot a figure in the manuscript:


Breast
hsa−mir−146b
Arginine and proline metabolism PAS