miRNA Regulation in Breast Cancer

Grace Newman, Sabrina Chow, Bhuvna Murthy

August 7, 2020

miRcore summer camp 2020

Disease Information

What is the significance ((background and problems) of the disease you are analyzing?

Breast cancer is one of the most well-known cancers out there. It's the second most common cancer diagnosed in women living in the US. There are multiple stages, I through IV. If the cancer is allowed to develop all the way to stage IV, the prognosis is much grimmer. The 5-year survival rate drops dramatically, so it's really important to diagnose breast cancer early and even perform preventative measures.

Genetics is an important part of that because 5-10% of breast cancer cases are hereditary (i.e. passed down in genes). The BRCA gene is known as the breast cancer gene because a mutation in BRCA1 or BRCA2 indicates a person is a strong risk for breast cancer.

Reading in Data

Read in the cancer miRNA expression text file that you downloaded from the Google Drive

```
dir()
```

```
##  [1] "4-2 randomForest - sig.Rmd" "BRCA.txt"
##  [3] "cacamp-2-empty.Rmd"         "cacamp-3-
empty.Rmd"
##  [5] "cacamp-ifelse-empty.Rmd"    "correlation and
ttest.Rmd"
##  [7] "final_presentation.html"
"final_presentation.Rmd"
##  [9] "full_BRCA_miR_data.txt"     "ggplot_basic.Rmd"
## [11] "glm-full.Rmd"               "miR_targets.txt"
## [13] "miRcore day 3 quiz.Rmd"     "miRcore day 4
quiz.Rmd"
## [15] "miRcore-day-3-quiz.html"    "Monday_R.Rmd"
## [17] "R day 3.Rmd"                "R_reference.pdf"
## [19] "Rbase.R"                    "rIntro.Rmd"
## [21] "Sample1_data.txt"           "small_data.txt"
## [23] "smalldata.Rmd"              "vector.Rmd"
```

```
brca = read.table("BRCA.txt", row.names = 1, header =
T, stringsAsFactors = F)
brcaM = as.matrix(brca)
```

A. What are the dimensions of your text file?

941 rows by 100 columns

B. How many Tumor and Control samples are in your dataset?

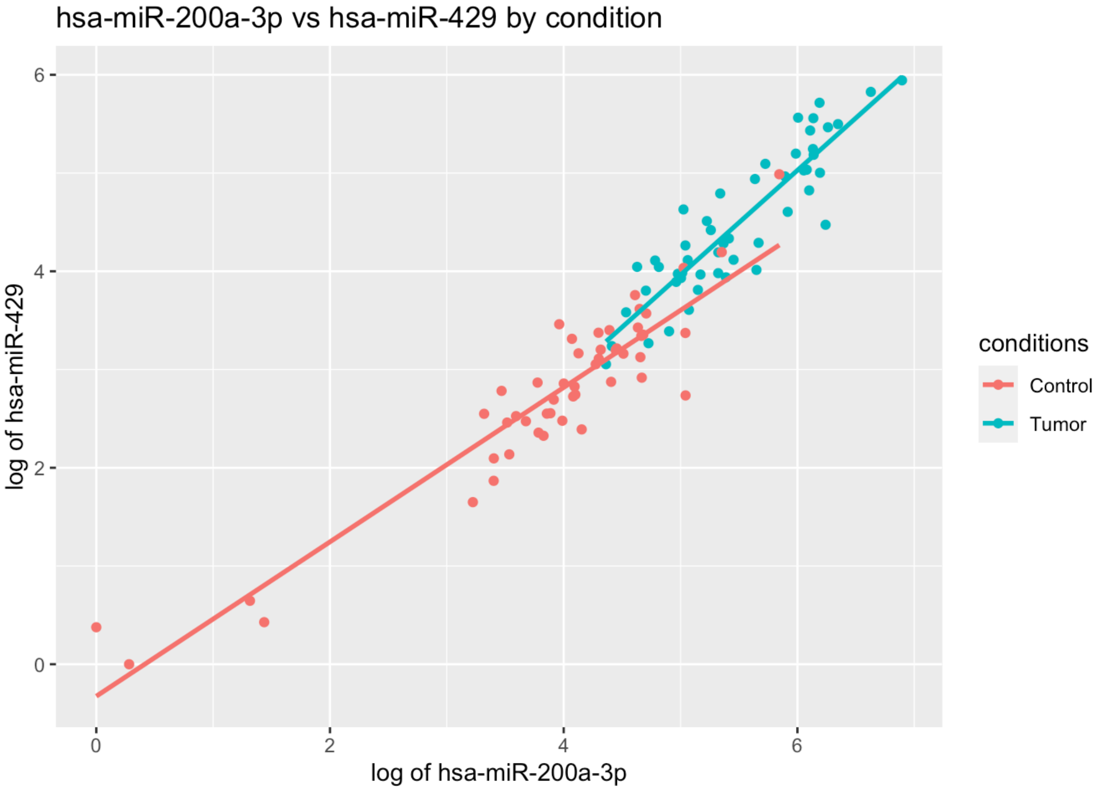
50 tumor, 50 control

Plotting

B. Make a scatterplot for hsa-miR-200a-3p and has-miR-429

using all samples in your data file. Then compute the Pearson correlation.

```
library(ggplot2)

brca_t = data.frame(t(brca))
conditions = c(rep("Tumor", 50), rep("Control", 50))
conditions = as.factor(conditions)
brca_t = cbind(brca_t, conditions)

hsa200 = brca_t[,"hsa.miR.200a.3p"]
hsa429 = brca_t[,"hsa.miR.429"]

ggplot(brca_t, aes(x=log(hsa200+1), y=log(hsa429+1),
color=conditions)) +
  geom_point() +
  labs(title="hsa-miR-200a-3p vs hsa-miR-429 by
condition",
       x="log of hsa-miR-200a-3p",
       y="log of hsa-miR-429") +
  geom_smooth(method=lm, se=F)
```



```
# Pearson correlation:

cor(hsa200, hsa429)
```

```
## [1] 0.935492
```

T Tests

A. For each miRNA: calculate mean, standard deviation, and the 2-tail student t-test p-value of Tumor and Control groups. Save all results in a table.

```
num_miRs <- length(rownames(brcaM))

c_means = vector()
t_means = vector()
c_sds = vector()
t_sds = vector()
pvals = vector()
```

```
for(i in 1:num_miRs){
  c_means <- c(c_means, mean(brcaM[i,51:100]))
  t_means <- c(t_means, mean(brcaM[i,1:50]))

  c_sds <- c(c_sds, sd(brcaM[i,51:100]))
  t_sds <- c(t_sds, sd(brcaM[i,1:50]))

  p_value <- t.test(brcaM[i,1:50],brcaM[i,51:100])
$p.value
  pvals <- c(pvals, p_value)
}

full_stats = data.frame()
full_stats <- cbind(c_means, t_means, c_sds, t_sds,
pvals)
#View(full_stats)
rownames(full_stats) <- rownames(brcaM)
```

B. Find 50 most significant miRs.

```
full_stats_ordered = full_stats[order(pvals),]
sig_brca = data.frame()
sig_brca = full_stats_ordered[1:50,]
```

C. Find miRNAs of p < 0.00001.

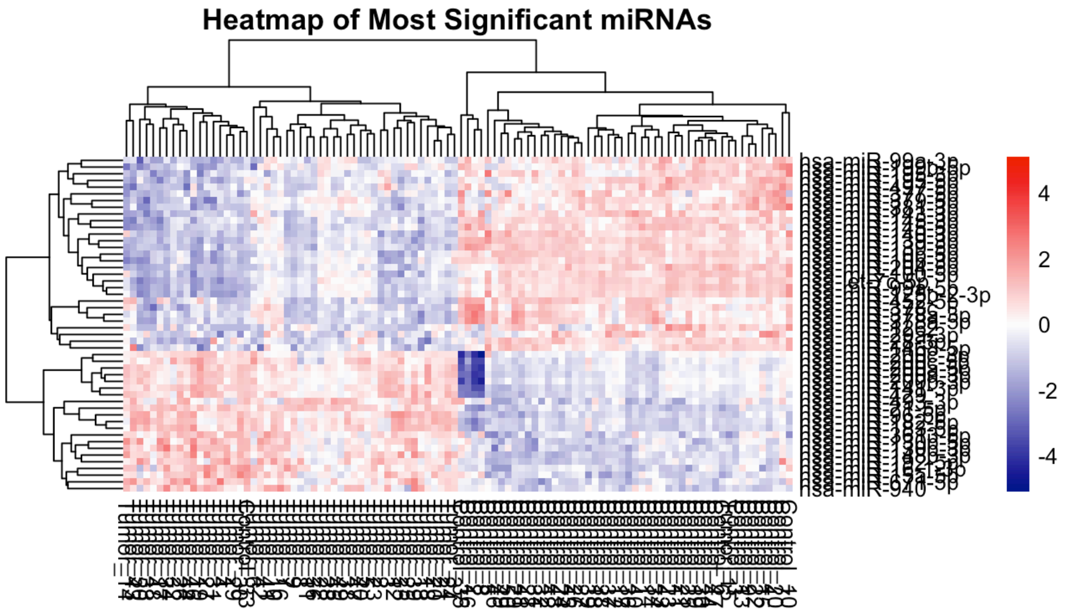
```
goodpval = vector('numeric')
indices = vector('numeric')

for (i in 1:941){
  p_value <- t.test(brca[i,1:50],brca[i,51:100])
$p.value
  if (!is.na(p_value) && p_value < 0.00001){
    goodpval <- c(goodpval, p_value)
    indices <- c(indices, i)
  }
}
```

```
indices
```

```
##  [1]   4  11  16  22  26  27  68  72  78  85 111 119
120 135 136 137 139 140 143
## [20] 147 148 149 157 159 160 176 179 189 193 211 212
224 230 231 239 244 245 246
## [39] 247 248 249 253 264 317 320 326 327 352 447 448
449 452 462 463 477 482 486
## [58] 530 560 565 566 567 569 571 574 576 627 640 651
657 658 672 673 675 678 792
## [77] 799 800 801 831 836 874 887 919 922 930 936 938
939
```

D. Make a heatmap of miRNAs with p < 0.00001.

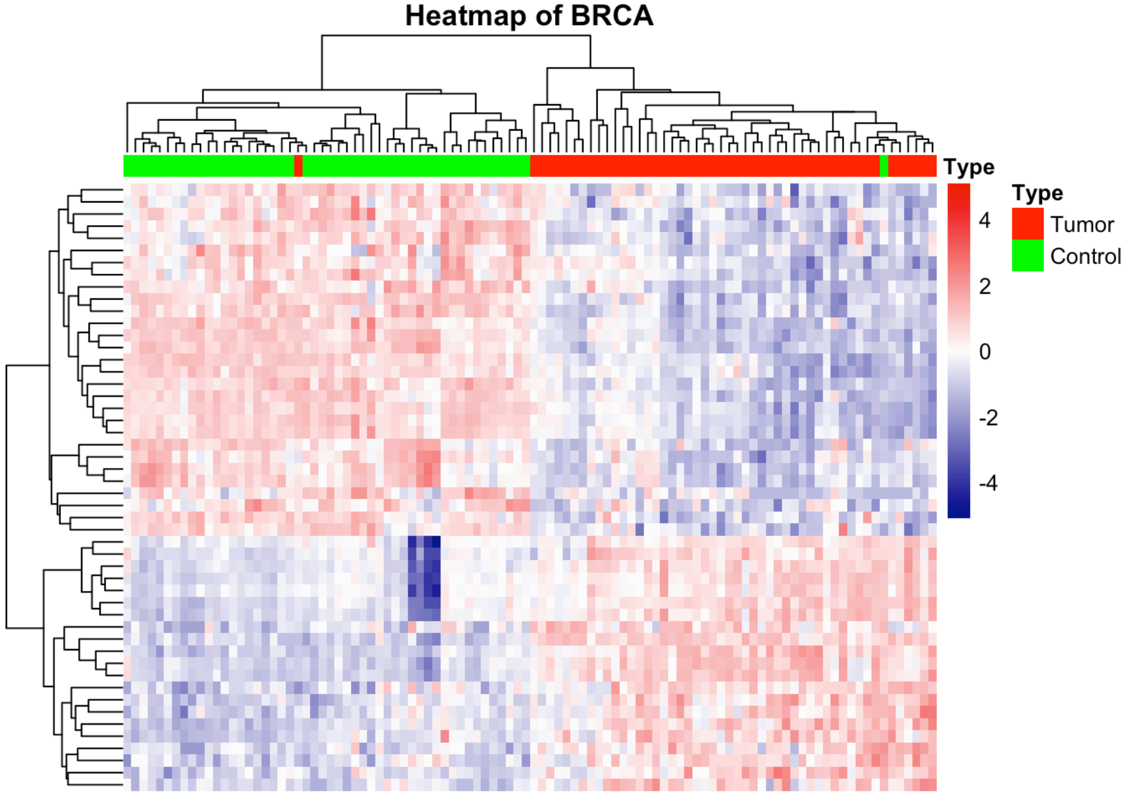
```
library(pheatmap)

sig_brca_data <- data.frame()
sig_brca_data <- brca[rownames(sig_brca),]
sig_brcaM_data <- as.matrix(sig_brca_data)
s_heatmap <- log((sig_brcaM_data+1),2)
#View(sig_brca_data)
#head(sig_brca_data)

pheatmap(s_heatmap,
        cluster_cols = T,
        scale = "row",
        cellwidth = 3,
        cellheight = 3,
        color = colorRampPalette(c("darkblue",
"white","red2"))(256),
        border_color = NA,
        main = "Heatmap of Most Significant miRNAs")
```



E. Does the heatmap cluster tumor and control samples correctly?

Below is an error-checked version of the heatmap.

```
sig_brca_data <- data.frame()
sig_brca_data <- brca[rownames(sig_brca),]
sig_brcaM_data <- as.matrix(sig_brca_data)
sig_brcaM_data = sig_brcaM_data[rowSums(sig_brcaM_data) > 0,]
col.features = data.frame(Type = factor(rep(c("Tumor","Control"), each=50))) #Making each equal to length
rownames(col.features) = colnames(sig_brcaM_data)
pheatmap(log2(sig_brcaM_data + 1),
         scale = "row",
         clustering_distance_cols = "correlation",
         color = colorRampPalette(c("darkblue", "white", "red2")) (256), border_color = NA,
show_rownames = F, show_colnames = F,main="Heatmap of
BRCA",  gaps_col=length(all_tumors),
annotation_col=col.features,
annotation_colors=list(Type
=c(Tumor="red",Control="green")))
```



Correlations

A. Find correlations between the most significant miRNA with 49 other miRNAs for 1) all samples.

Our most significant miRNA is hsa-miR-10b-5p.

```
corval = vector()
sig_correlations = data.frame()
for(i in 2:50) {
  corval = c(corval, cor(brcaM[rownames(sig_brca)[1],],
                         brcaM[rownames(sig_brca)[i],]))
}
```

```
sig_correlations = cbind(rownames(sig_brca), corval)
#View(sig_correlations)
```

B. Among A. can you find some patterns between miRNAs? What are your findings? (example: to find the most correlated miRNA with the most significant miRNA)

The most correlated miRNA with hsa-miR-10b-5p is hsa-miR-145-3p with a value of ~0.875.

Random Forest

A. Create a random forest using the steps from the random forest .html file to classify your data. What is your accuracy?

Setting up data table to run random forest:

```
library(randomForest)

sig_miRs = vector()
num_miRNAs = length(brca)
tumor_samples = 1:50
control_samples = 51:100
for(i in 1:num_miRNAs) {
  tumor_vector = brca[i, tumor_samples]
  control_vector = brca[i, control_samples]
  p_value = t.test(tumor_vector, control_vector)
$p.value
  if(!is.na(p_value) && p_value < 0.0001) {
    sig_miRs = c(sig_miRs, i)
  }
}
#length(sig_miRs)

sbrca = brca[sig_miRs,]
#dim(brca)
sbrca = t(sbrca)
```

```
conditions = rownames(sbrca)
rownames(sbrca) = vector()

for(i in 1:length(conditions)) {
  if(grepl("Tumor", conditions[i], fixed = TRUE) ==
TRUE) {
    conditions[i] = "Tumor"
  } else {
    conditions[i] = "Control"
  }
}

sbrca = as.data.frame(sbrca)
sbrca = cbind(conditions, sbrca)
sbrca[1:10, 1:5]
```

```
##    conditions hsa-let-7c-5p hsa-let-7g-3p hsa-
miR-100-5p hsa-miR-103a-2-5p
## 1       Tumor      1199.961     76.981396
2121.914          2.189044
## 2       Tumor      1524.170     55.120415
1039.381          2.936742
## 3       Tumor      1166.842     28.869432
1385.517          1.938992
## 4       Tumor       594.546     80.737490
1207.962          1.078297
## 5       Tumor      4366.677      8.328753
9635.928          0.876711
## 6       Tumor      2929.552      9.105156
4718.907          0.128242
## 7       Tumor      3268.027     47.606227
2707.554          1.600209
## 8       Tumor      3984.783     52.456013
3117.653          0.535265
## 9       Tumor      2604.580     27.690366
1803.486          1.203929
```

```
## 10      Tumor      5391.681     93.268706
4235.399          0.832756
```

```
colnames(sbrca)[1] = "Condition"
```

Running the actual tests:

```
set.seed(123)
num_rows = length(conditions)
train_samples = sample(1:num_rows, round(0.75 *
num_rows), replace = FALSE)

colnames(sbrca) = gsub("-",".",colnames(sbrca))
train_data = sbrca[train_samples,]
test_data = sbrca[-train_samples,]
#dim(train_data)
#dim(test_data)

train_data$Condition = as.factor(train_data$Condition)
test_data$Condition = as.factor(test_data$Condition)
brca_forest = randomForest(Condition ~ .,
                            ntree = 100,
                            data = train_data)
train_predictions = predict(brca_forest, train_data)
table(train_predictions, train_data$Condition)
```

```
##
## train_predictions Control Tumor
##           Control      38     0
##           Tumor         0    37
```

```
test_predictions = predict(brca_forest, test_data)
table(test_predictions, test_data$Condition)
```

```
##
## test_predictions Control Tumor
##          Control      11     0
##          Tumor         1    13
```

B. Change variables in the random forest run and check if the results are different. What are the parameters that you can

achieve a better accuracy?

We changed the test so that we were using a 75/25 split instead of a 70/30.

Analysis

miRNAs

What are your most significant miRNAs? Are they up-regulated or down-regulated? Are they highly correlated with each other?

```
ratios = log2(sig_brca[,"t_means"]/
sig_brca[,"c_means"])
sig_brca = cbind(sig_brca, ratios)
# View(sig_brca)
```

Bonus: split the dataset up between upregulated and downregulated genes.

```
upregulated = data.frame()
downregulated = data.frame()
up = vector()
down = vector()

for(i in 1:50){
  if(sig_brca[i,"ratios"] > 0){
    upregulated = rbind(upregulated, sig_brca[i,])
    up = c(up,i)
  } else if(sig_brca[i,"ratios"] < 0){
    downregulated = rbind(downregulated, sig_brca[i,])
    down = c(down,i)
  }
}
colnames(upregulated) = colnames(sig_brca)
colnames(downregulated) = colnames(sig_brca)
upregulated <- cbind(rownames(sig_brca)[up],
upregulated)
downregulated <- cbind(rownames(sig_brca)[down],
downregulated)

# length(rownames(downregulated)) #29
# length(rownames(upregulated)) #21

# View(upregulated)
# View(downregulated)
```

Most significant have a NEGATIVE ratio, meaning they are downregulated, but there are still some upregulated genes. There were a total of 29 downregulated genes and 21 upregulated genes.

Genes

A. What are the common gene targets of your significant miRNAs?

```
miR_targets = read.table("miR_targets.txt",
stringsAsFactors = F)
# View(miR_targets)

down_genes = vector()
for(i in 1:29) {
  down_genes = c(down_genes,
miR_targets[grep(rownames(downregulated)[i],
miR_targets[,1]),2])
}
# down_genes

up_genes = vector()
for(i in 1:21) {
  up_genes = c(up_genes,
miR_targets[grep(rownames(upregulated)[i],
miR_targets[,1]),2])
}
```

```
# up_genes

common_up_targets =
up_genes[which(duplicated(up_genes)==T)]
common_down_targets =
down_genes[which(duplicated(down_genes)==T)]

#View(common_up_targets)
#View(common_down_targets)
```

B. Are any of your genes targeted by these top miRNAs known mutations of the disease you are researching?

Yes: mTOR, RAP1, STMN1, MYB, E2F1, NF1, RAC1, APC, ACKR3, CHEK2, VEGFA, ATM, BRCA1, BRCA2.

Conclusion

What do your findings suggest to you about potential diagnosis/ treatment targets for the disease you are researching?

Significant downregulation of hsa-miR-10b-5p and related miRNAs appears to correlate with increased likelihood of breast cancer presence in patients. This is perhaps related to the resulting increased protein synthesis in mutated BRCA1 and BRCA2 genes, along with misregulation of the tumor-suppressing mTOR pathway. Further study is necessary to verify this correlation.