TCGA EM like TNBC Validations

Linhua Wang

2022-09-28

Data loading

Read top 100 E markers and top 100 M markers, and the TCGA triple-negative breast cancer tumor samples.

```
tnbrca.count <- read.table("nature11412-s2/TCGA_BRCA_TN_RPKM_libnorm.txt")
em_markers <- read.csv("top100_em_markers.csv")
em_markers <-em_markers[order(em_markers$marker),]
em_markers <- em_markers[,-1]
row.names(em_markers) <- em_markers$gene
genes <- row.names(em_markers)
length(genes)
```

```
## [1] 200
```

```
genes <- genes[(genes %in% row.names(tnbrca.count))]
length(genes)
```

```
## [1] 185
```

```
tnbrca.count <- tnbrca.count[genes,]
dim(tnbrca.count) ## 185 markers in the TCGA data, 128 TNBC tumor samples exists
```

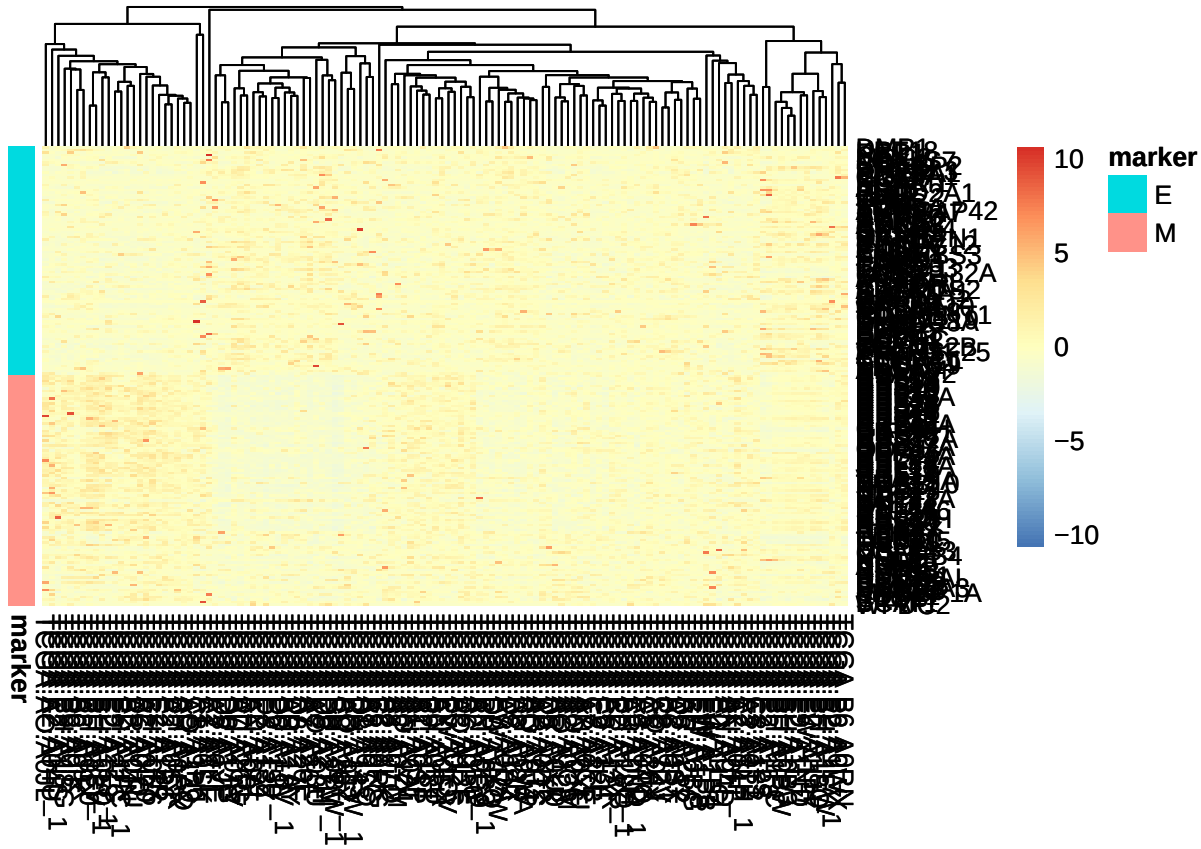
```
## [1] 185 128
```

```
em_markers <- em_markers[row.names(tnbrca.count),]
```

Exploration

Cluster TNBC samples

```
pt = pheatmap(tnbrca.count, scale='row',
              annotation_row  = data.frame(row.names = row.names(em_markers),
                                           marker = em_markers$marker), cluster_rows = F)
print(pt)
```

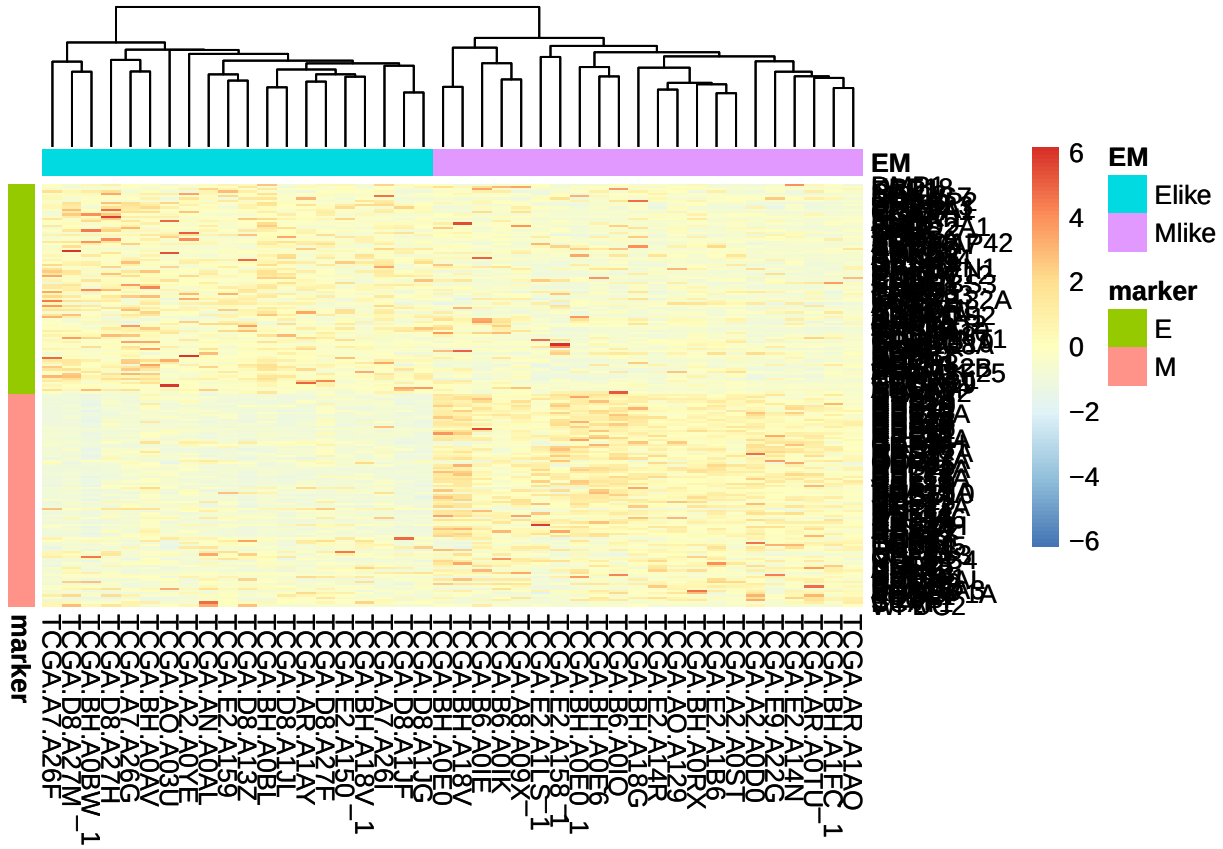



```
# pdf("heatmap_wo_subset_EM-like_samples.pdf", width=7, height=7)
# grid::grid.newpage()
# grid::grid.draw(pt$gtable)
# dev.off()
```

Select E- and M-like samples based on the E/M markers

```
sample_clust <- factor(cutree(pt$tree_col, k=13))
sample_clust_df <- ifelse(sample_clust==3,"Mlike", ifelse(sample_clust==6, "Elike", "Others"))
sample_clust_df <- data.frame(row.names=names(sample_clust), 'EM'=sample_clust_df)

sample_cluster_df.1 <- data.frame(row.names = row.names(sample_clust_df)[sample_clust_df$EM %in% c("Eli
                                  EM = sample_clust_df[sample_clust_df$EM %in% c("Elike", 'Mlike'),])
tnbrca.count.1 <- tnbrca.count[,row.names(sample_cluster_df.1)]
```

```
pt2 <- pheatmap(tnbrca.count.1, scale='row',
         annotation_row  = data.frame(row.names = row.names(em_markers),
                                      marker = em_markers$marker),
         annotation_col = sample_cluster_df.1,
         cluster_rows = F)
print(pt2)
```




```
dim(tnbrca.count.1)
```

```
## [1] 185  42
```

```
table(sample_cluster_df.1$EM) # end up with 20 E-like tumors and 22 M-like tumors
```

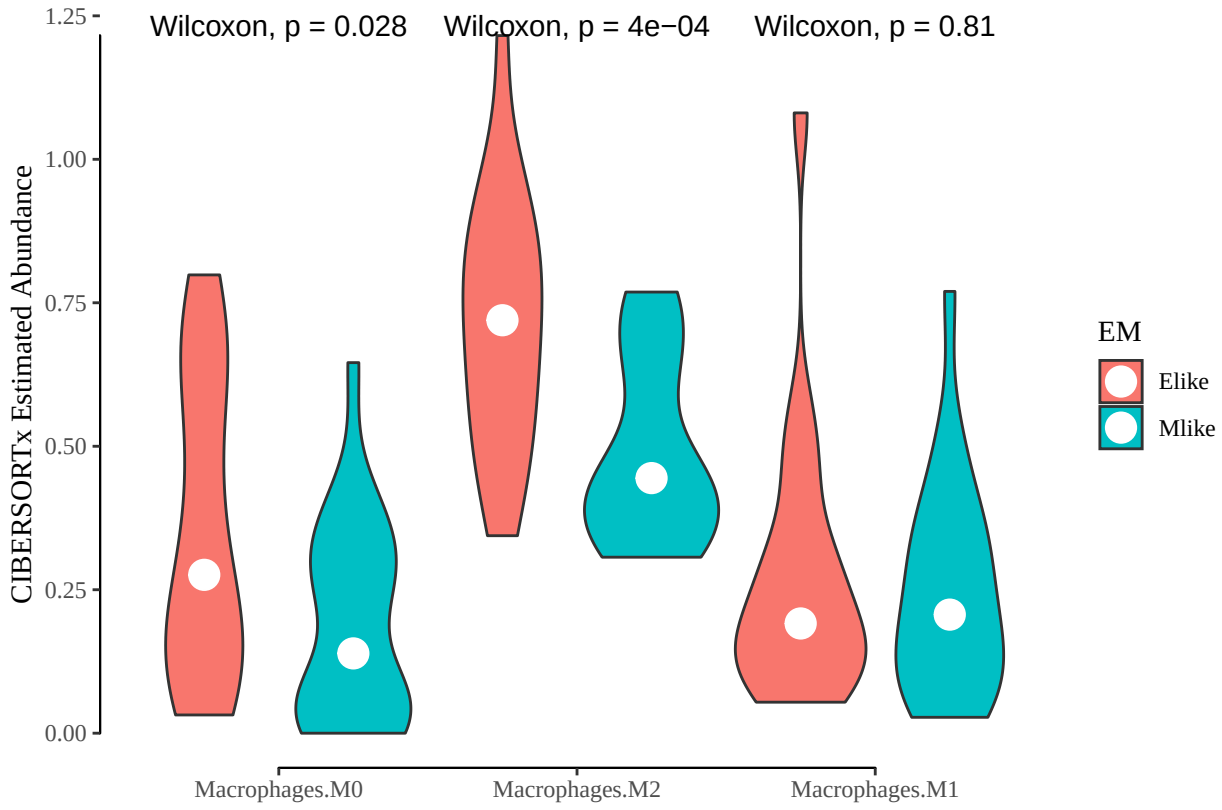
```
##
## Elike Mlike
##    20    22
```

```
cibersort.res <- read.csv("CIBERSORTx_TCGA_BRCA_TN_RPKM_libnormcsv.csv", row.names = 1)
rnames <- c()
for (rn in row.names(cibersort.res)){
  rnames <- c(rnames, gsub( '-', '.', rn))
}
row.names(cibersort.res) <- rnames

cibersort.res <- cibersort.res[, c('Macrophages.M2', 'Macrophages.M1', 'Macrophages.M0')]
cibersort.res <- cibersort.res[row.names(sample_cluster_df.1),]
cibersort.res['EM'] <- sample_cluster_df.1$EM
cibersort.res <- cibersort.res[complete.cases(cibersort.res),] # remove NA due to 0 macrophages
table(cibersort.res$EM)
```

```
##
## Elike Mlike
##    20    22
```

```
library(ggpubr)
```

```
library(reshape2)
library(ggthemes)
cibersort.res.long <- cibersort.res
cibersort.res.long$identifier <- row.names(cibersort.res.long)
cibersort.res.long <- melt(cibersort.res.long, id.vars = c("identifier", "EM"))
cibersort.res.long$variable <- factor(cibersort.res.long$variable, levels=c('Macrophages.M0', 'Macropha

p3 <- ggplot(cibersort.res.long, aes(y=value, x=variable, fill=EM)) +
  geom_violin(position=position_dodge(1)) +
  stat_summary(fun=median, geom="point", size=5, color="white",position=position_dodge(1)) +
  stat_compare_means(method = "wilcox.test") +labs(y= "CIBERSORTx Estimated Abundance", x = "", size=25
  theme_tufte() + geom_rangeframe()
print(p3)
```




```
pdf("TCGA_TNBC_Macrophages_subtype_proportions_EM.pdf")
print(p3)
dev.off()
```

```
## pdf
##   2
```