UMAP on TCGA Breast Cancer

Parissa Amin

2023-12-27

Loading Data

Loading local bulk RNA seq data

```
library(umap)
```

```
## Warning: package 'umap' was built under R version 4.2.3
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.2.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
exp <- read.delim("D:/BulkRNA/Data/data_mrna_seq_v2_rsem.txt" )
exp<- na.omit(exp)
exp<- data.frame(t(exp))

colnames(exp) <- exp[ 1,]
exp<- exp[-1:-2, ]
```

Loading clinical data

```
PatientInfo <- read.delim("D:/BulkRNA/Data/tcga_clinical_data.tsv")
PatientInfo$Sample.ID <- gsub("-" ,"." , PatientInfo$Sample.ID)
```

Assigning ER status from clinical data to expression data

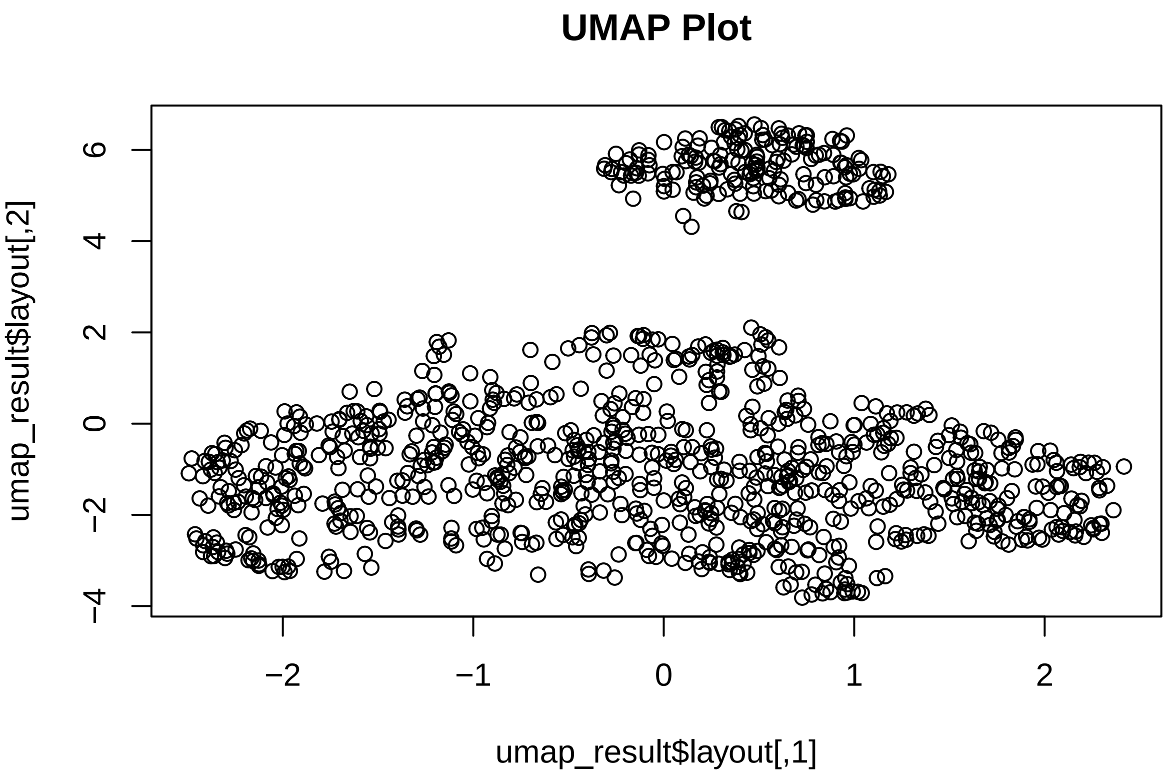
```
exp$ER.status <- PatientInfo$ER.Status.By.IHC[match(rownames(exp), PatientInfo$Sample.ID)]
exp<- na.omit(exp)
exp<- exp[exp$ER.status!="Indeterminate",]
```

```
#log-transform the expression data
# Convert to numeric and then apply log2 transformation

log_exp <- as.data.frame(lapply(exp[,-20531], as.numeric))
log_exp<-as.data.frame(log2(log_exp+1)) # Adding 1 to avoid log(0)
```

UMAP

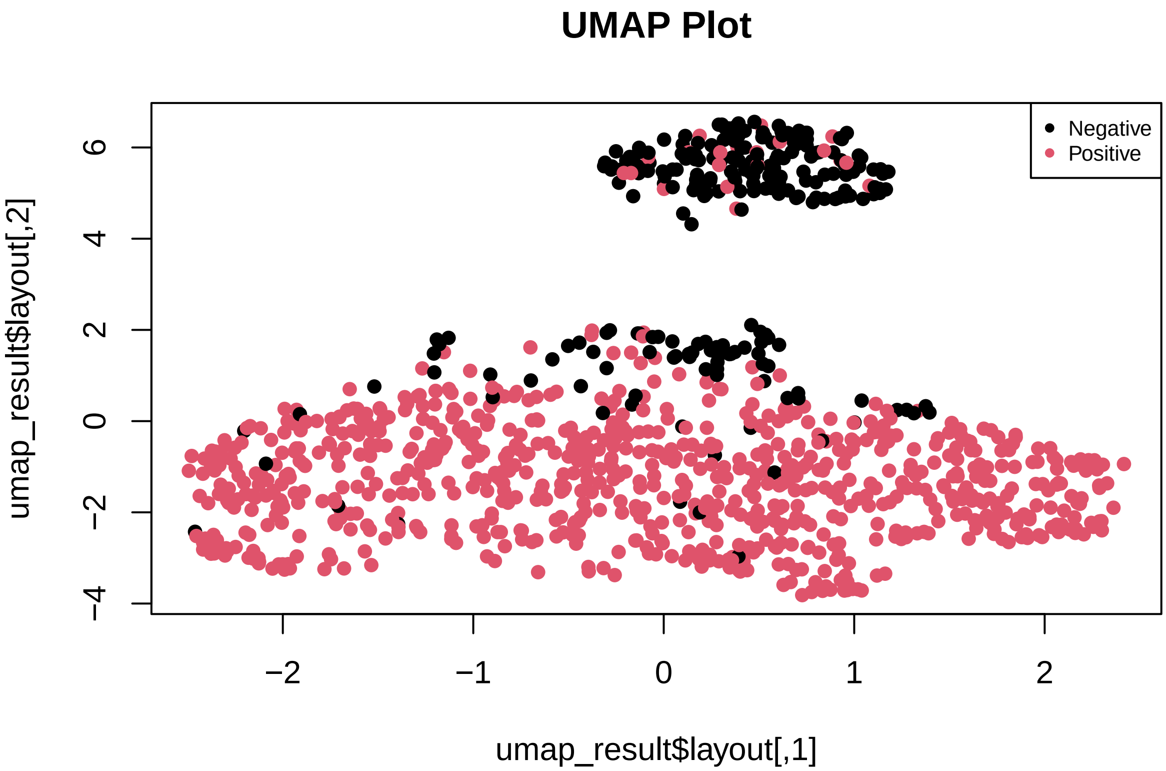
Running UMAP with 30 NN

```
umap_result <- umap(log_exp, n_neighbors =30)
plot(umap_result$layout, main="UMAP Plot")
```



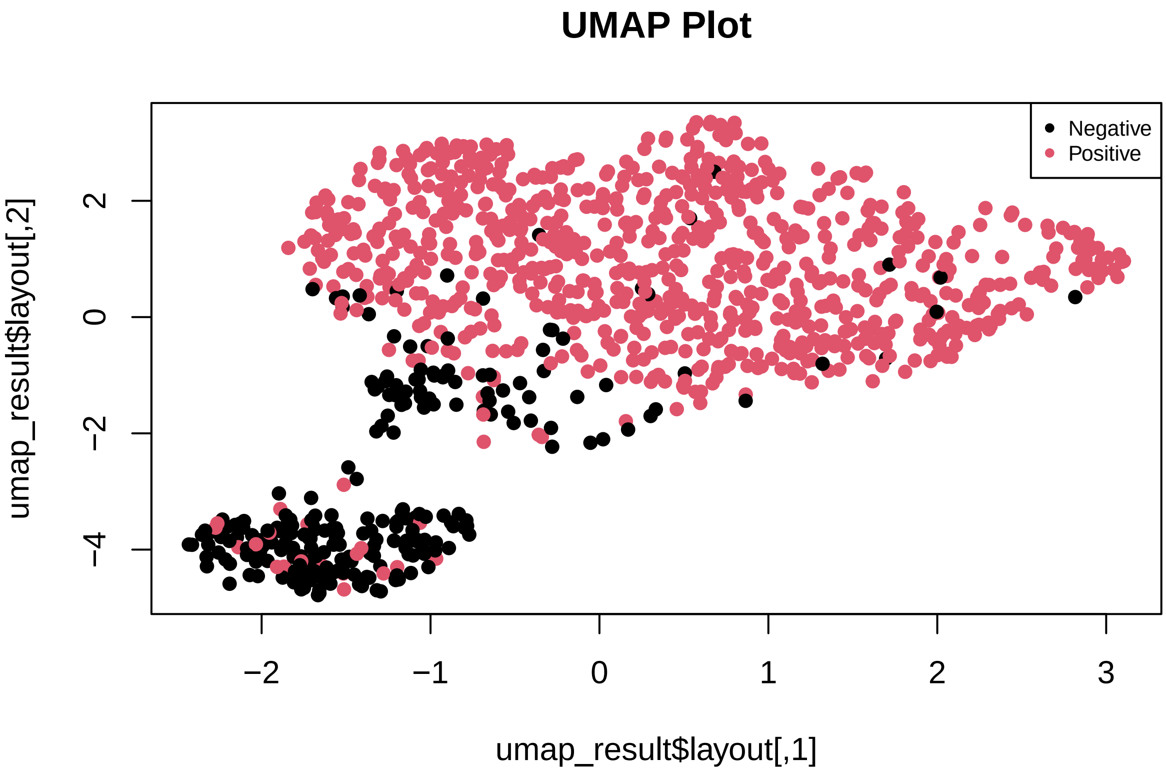
Adding ER status in UMAP

```
labels <- exp$ER.status
labels <- as.factor(labels)

plot(umap_result$layout, col = as.numeric(labels), pch = 16, main = "UMAP Plot")
# Add a legend if labels are factors
if (is.factor(labels)) {
  legend("topright", legend = levels(labels), col=1:length(levels(labels)), pch=16, cex=0.65)
}
```



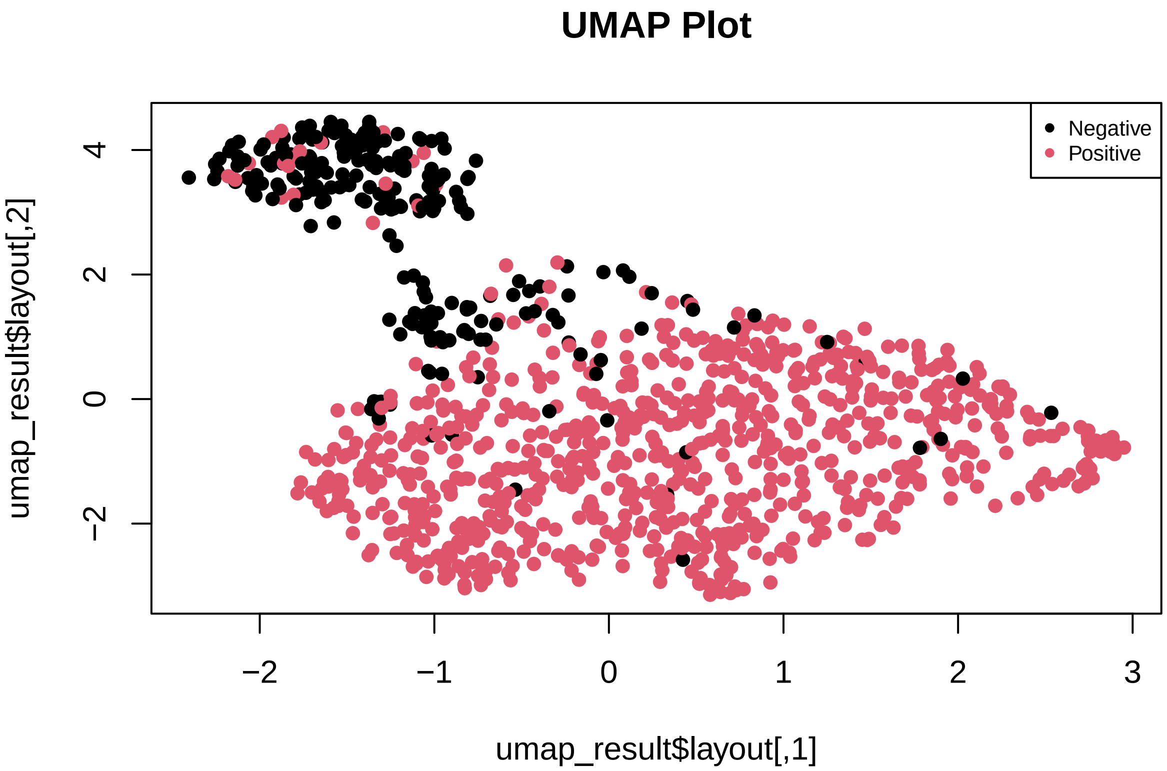
Running UMAP with new, 70 NN

```
umap_result <- umap(log_exp, n_neighbors =70)

plot(umap_result$layout, col = as.numeric(labels), pch = 16, main = "UMAP Plot")
if (is.factor(labels)) {
  legend("topright", legend = levels(labels), col=1:length(levels(labels)), pch=16, cex=0.65)
}
```



Running UMAP with new, 120 NN

```
umap_result <- umap(log_exp, n_neighbors =120)

plot(umap_result$layout, col = as.numeric(labels), pch = 16, main = "UMAP Plot")
if (is.factor(labels)) {
  legend("topright", legend = levels(labels), col=1:length(levels(labels)), pch=16, cex=0.65)
}
```



Conclusion

We ran UMAP using different values for NN. In all cases, UMAP seperates samples based on their ER status, showing that breast tumors are in fact include two major clusters: ER+ and ER- subtypes.