PAM50 Data Download

Kevin Cissie

2024-08-28

```
#Setting the working Directory
setwd("D:/IDC_MAGE/PAM50 data Download_Classification")
getwd()
```

```
## [1] "D:/IDC_MAGE/PAM50 data Download_Classification"
```

```
# OBJECTIVE 1
#PART 2 : DATA CATEGORIZATION USING PAM50
#First we're going to download PAM50 classified data from the TCGA and use
this data to categorize our IDC data
#Since we used cases (used these as barcodes to obtain data matching the
manifest file corresponding to the IDC data), We're going to use a package
TCGAutils to convert these cases to barcode ids that are identifiable for the
PAM50 data download package of TCGAbiolinks

#Loading libraries
#Loading libraries
library(TCGAutils) # this package contains the TCGAcode which converts cases
ids to barcodes
library(TCGAbiolinks)
library(readxl)
```

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

```
#Loading the cases barcodes from the IDC gdc_manifest file
cases_barcodes <-
read.csv("D:/IDC_MAGE/IDC_TCGA_Datadownload/IDCCases_barcodes.csv")
head(cases_barcodes)
```

```
##   X                           x
## 1 1 TCGA-A2-A04U-01A-11R-A115-07
## 2 2 TCGA-AN-A04A-01A-21R-A034-07
## 3 3 TCGA-A7-A13D-01A-13R-A12P-07
## 4 4 TCGA-BH-A201-01A-11R-A14M-07
## 5 5 TCGA-A2-A04R-01A-41R-A109-07
## 6 6 TCGA-AN-A03X-01A-21R-A00Z-07
```

```
#Converting these barcodes into a list
cases_barcodes <- cases_barcodes$x #Extracting these from the x column
cases_barcodes <- c(cases_barcodes) # making these into a list
head(cases_barcodes)
```

```
## [1] "TCGA-A2-A04U-01A-11R-A115-07" "TCGA-AN-A04A-01A-21R-A034-07"
## [3] "TCGA-A7-A13D-01A-13R-A12P-07" "TCGA-BH-A201-01A-11R-A14M-07"
## [5] "TCGA-A2-A04R-01A-41R-A109-07" "TCGA-AN-A03X-01A-21R-A00Z-07"

#Converting these cases ids into sample barcodes
IDC_samplebarcodes <- TCGAbarcode(cases_barcodes)
head(IDC_samplebarcodes) # the barcodes displayed are shorter than the cases
barcodes

## [1] "TCGA-A2-A04U" "TCGA-AN-A04A" "TCGA-A7-A13D" "TCGA-BH-A201" "TCGA-A2-
A04R"
## [6] "TCGA-AN-A03X"

#Using TCGAquery_subtype to download the PAM50 data corresponding to the BRCA
project
#PAM50 classified data corresponding to the IDC barcodes
#Obtaining TCGA samples with their categorized PAM50 subtypes
PAM50_BRCAdata <- TCGAquery_subtype("BRCA") # Downloads pam50 subtypes of
BRCA  Tumor

## brca subtype information from:doi.org/10.1016/j.ccell.2018.03.014

head(PAM50_BRCAdata) # Column 11 contains the BRCA-subtypes classified using
PAM50

## # A tibble: 6 × 24
##   patient      Tumor.Type Included_in_previous_mark…¹ vital_status
days_to_birth
##   <chr>        <chr>      <chr>                       <chr>        <chr>
## 1 TCGA-3C-AAAU BRCA       NO                          Alive        -20211
## 2 TCGA-3C-AALI BRCA       NO                          Alive        -18538
## 3 TCGA-3C-AALJ BRCA       NO                          Alive        -22848
## 4 TCGA-3C-AALK BRCA       NO                          Alive        -19074
## 5 TCGA-4H-AAAK BRCA       NO                          Alive        -18371
## 6 TCGA-5L-AAT0 BRCA       NO                          Alive        -15393
## # ℹ abbreviated name: ¹Included_in_previous_marker_papers
## # ℹ 19 more variables: days_to_death <chr>, days_to_last_followup <chr>,
## #   age_at_initial_pathologic_diagnosis <dbl>, pathologic_stage <chr>,
## #   Tumor_Grade <chr>, BRCA_Pathology <chr>, BRCA_Subtype_PAM50 <chr>,
## #   MSI_status <chr>, HPV_Status <chr>, tobacco_smoking_history <chr>,
## #   `CNV Clusters` <chr>, `Mutation Clusters` <chr>,
## #   `DNA.Methylation Clusters` <chr>, `mRNA Clusters` <chr>, …

# Saving out the PAM50 subtypes file
write.csv(PAM50_BRCAdata, "PAM50_BRCAdata.csv")

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

# Filter the data frame
PAM50_IDCdata <- PAM50_BRCAdata %>%
  filter(patient %in% IDC_samplebarcodes)
head(PAM50_IDCdata)

## # A tibble: 6 × 24
##   patient      Tumor.Type Included_in_previous_mark…¹ vital_status
days_to_birth
##   <chr>        <chr>      <chr>                       <chr>        <chr>
## 1 TCGA-3C-AALI BRCA       NO                          Alive        -18538
## 2 TCGA-3C-AALJ BRCA       NO                          Alive        -22848
## 3 TCGA-3C-AALK BRCA       NO                          Alive        -19074
## 4 TCGA-A1-A0SD BRCA       YES                         Alive        -21793
## 5 TCGA-A1-A0SF BRCA       YES                         Alive        -19731
## 6 TCGA-A1-A0SH BRCA       YES                         Alive        -14595
## # ℹ abbreviated name: ¹Included_in_previous_marker_papers
## # ℹ 19 more variables: days_to_death <chr>, days_to_last_followup <chr>,
## #   age_at_initial_pathologic_diagnosis <dbl>, pathologic_stage <chr>,
## #   Tumor_Grade <chr>, BRCA_Pathology <chr>, BRCA_Subtype_PAM50 <chr>,
## #   MSI_status <chr>, HPV_Status <chr>, tobacco_smoking_history <chr>,
## #   `CNV Clusters` <chr>, `Mutation Clusters` <chr>,
## #   `DNA.Methylation Clusters` <chr>, `mRNA Clusters` <chr>, …

total_rows <- nrow(PAM50_IDCdata)
total_rows   #This returns 595 entries(samples) which corresponds to the
number of unique samples

## [1] 595

#Counting occurences of each BRCA-subtype

BRCAsubtype_counts <- as.data.frame(table(PAM50_IDCdata$BRCA_Subtype_PAM50))

# Spreading the counts into a wide format
library(tidyr)

## Warning: package 'tidyr' was built under R version 4.2.3

BRCAsubtype_counts <- spread(BRCAsubtype_counts, Var1, Freq)

# View the result
print(BRCAsubtype_counts)
```

```
##   Basal Her2 LumA LumB Normal
## 1   133   58  257  129     18
```

```
library(ggplot2)
```

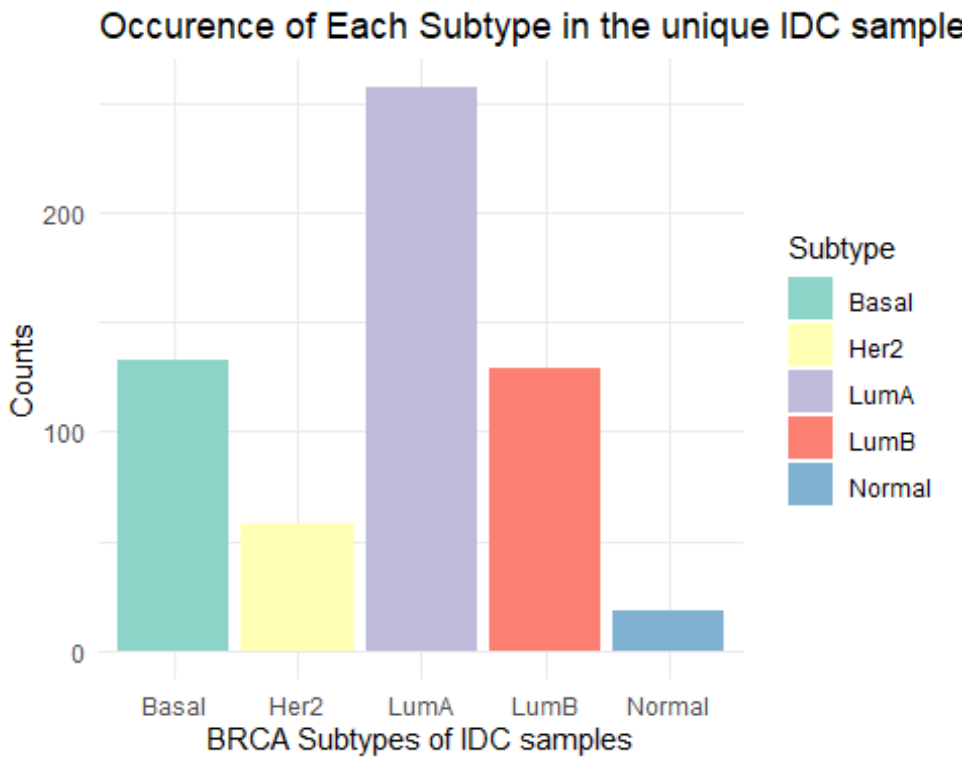
```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
#converting the subtypes categorization into a long format
BRCAsubtype_counts_long <- gather(BRCAsubtype_counts, key = "Subtype", value
= "Count")
print(BRCAsubtype_counts_long)
```

```
##   Subtype Count
## 1   Basal   133
## 2    Her2    58
## 3    LumA   257
## 4    LumB   129
## 5  Normal    18
```

```
write.csv(BRCAsubtype_counts_long, "BRCAsubtypes_countsocurrences.csv")

# Plot the counts of each subtype
plot <- ggplot(BRCAsubtype_counts_long, aes(x = Subtype, y = Count, fill =
Subtype)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Occurence of Each Subtype in the unique IDC samples", x =
"BRCA Subtypes of IDC samples", y = "Counts") +
  scale_fill_brewer(palette = "Set3")
# Save the displayed plot to a PDF file
ggsave("BRCA_Subtype_Plot.pdf", plot = plot, width = 8, height = 6)
# Display the plot
print(plot)
```

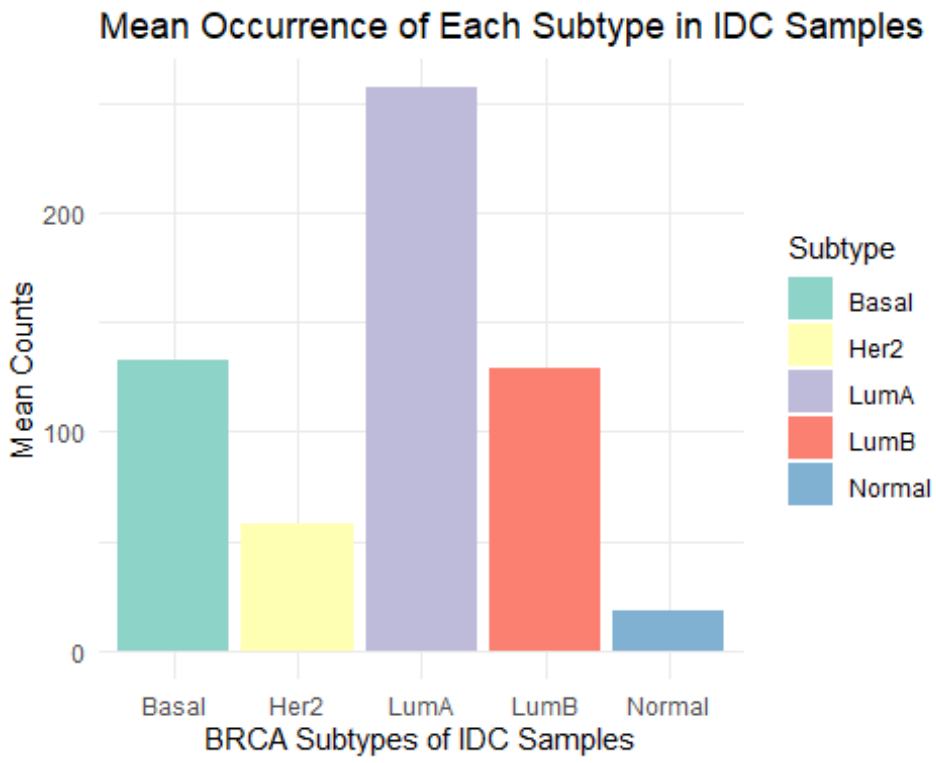



```r
#Calculating the means of counts and doing the plots
# Calculate the mean of counts for each subtype
mean_counts <- BRCAsubtype_counts_long %>%
  group_by(Subtype) %>%
  summarise(Mean_Count = mean(Count))

# Plot the mean counts of each subtype
mean_plot <- ggplot(mean_counts, aes(x = Subtype, y = Mean_Count, fill =
Subtype)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Mean Occurrence of Each Subtype in IDC Samples", x = "BRCA
Subtypes of IDC Samples", y = "Mean Counts") +
  scale_fill_brewer(palette = "Set3")

# Save the mean plot to a PDF file
ggsave("BRCA_Subtype_Mean_Plot.pdf", plot = mean_plot, width = 8, height = 6)

# Display the mean plot
print(mean_plot)
```


Mean Occurrence of Each Subtype in IDC Samples
Mean Counts
200
100
0
Basal
Her2
LumA
LumB
Normal
BRCA Subtypes of IDC Samples
Subtype
Basal
Her2
LumA
LumB
Normal