Assignment 2: Gene Expression Analysis & Interpretation

Conor Heffron - 23211267

Introduction

- In this report, I will analyse a publicly available dataset based on clinical breast cancer data. Breast cancer is the most diagnosed cancer in women. There are several subtypes of diseases characterized by different genetic drivers for cancer risk and tumour growth. The human epidermal growth factor receptor 2 amplified (HER2: ERBB2 / ERBB2IP) breast cancer is one of the most aggressive subtypes. In addition, I will investigate HER3 (ERBB3), HER4 (ERBB4), PIK3C2B, MDM4, LRRN2, NFASC, KLHDC8A, and CDK18 gene mutations. Although there are targeted therapies that have been developed to treat these cancer cases, the response rate ranges from 40% - 50%. I will download, decompress, clean and process the TCGA RNASeq data for breast cancer from cbioportal and identify the differentially expressed genes between ERBB2 / ERBB2IP, ERBB3, ERBB4, PIK3C2B, MDM4, LRRN2, NFASC, KLHDC8A, and CDK18 cancer tumours.

Note

- The dataset can be downloaded from this link:
  - https://www.cbioportal.org/study/summary?id=brca_tcga_pan_can_atlas_2018.

Methods Overview

- The methods to import data are from the `rio` package. To manipulate, analyse and query the data the `tidyverse` package includes several libraries. In particular, I have heavily used the `dplyr` package and methods such as **filter** to generate summary tables after data analysis and enrichment processes which are described and commented in the code chunks in an incremental fashion. I have implemented and imported a utility script written in R to assist in the loading, analysis, and aggregation of the TCGA data. The analysis was completed in a step by step fashion to help with my biological interpretation of the results of this analysis. This helped with the selection of features and values for deeper analysis and investigation of smaller subsets of samples.

Biological Interpretation

- The BRCA1 gene mutation is heavily associated with breast cancer. People who carry this gene mutation, have a hightened risk of developing cancer over time. Carriers of the BRCA1 gene often develop triple-negative, basal-like, aggressive breast tumours. Hormone signalling is pertinent in the inception of BRCA1 mutant breast cancers. Progesterone (PR) levels are clearly higher in BRCA1 mutation carriers and they have a higher risk of developing breast cancer with a low survival rate.
- HER2 is a member of the human Epidermal Growth Factor Receptor (EGFR) family, which actuates the signalling pathways that promote cell proliferation & survival by dimerization with other EGFR family members. HER2 breast cancers are likely to benefit from chemotherapy and treatment targeted to HER2.
- EGFR is a protein located on cells that help them to grow. A mutation in the EFGR gene can compel excessive growth which can cause cancer.
- There are different breast cancer groups taken into account during the TCGA data analysis segments of this report. The main groups include Luminal tumours (A & B). Luminal A are tumours that are Oestrogen+ (ER+) & PR+ & HER2-. Luminal A breast cancers benefit from hormone therapy & may also benefit from chemotherapy. Luminal B breast cancerts can be HER- or HER+ & ER+. HER2 breast cancers are PR+.
- HER3 is becoming a prominent biomarker for breast cancers (HER3 mRNA is expressed as Luminal tumours or ER+) as it is essential for cell survival in Luminal A and Luminal B but not basal normal mammary epithelium (basal like or triple negative breast cancers). Triple negative is the most aggresive form of breast cancer as they can groq and spread more quickly. The most difficult to treat compared to other invasive types of breast cancer because the cancer cells do not have the Oestrogen or Progesterone receptors or enough of the HER2 protein to make hormone therapy or targeted HER2 drugs work.
- HER4 expression in Oestrogen receptor-positive breast cancer is associated with decreased sensitivity to tamoxifen treatment and reduced overall survival of postmenopausal women.

Incremental Analysis, Code & Results

- The following graphics and summaries have the corresponding code chunks that shows how my analysis of the TCGA data evolved as I noticed patterns related to ER+, HER2, and upgraded/downgraded gene mutations.

Load packages, functions / methods and scripts

```r
library(knitr)
library(readr)
library(rio)
library(tools)
library(conflicted)
library(dplyr)
library(tibble)
suppressMessages(suppressWarnings(library(DESeq2)))
library(ggplot2)

# resolve conflicts
suppressMessages(suppressWarnings(conflict_prefer("filter", "dplyr")))
suppressMessages(suppressWarnings(conflict_prefer("lag", "dplyr")))
suppressMessages(suppressWarnings(conflict_prefer("count", "dplyr")))
suppressMessages(suppressWarnings(conflict_prefer("select", "dplyr")))
suppressMessages(suppressWarnings(conflicts_prefer(GenomicRanges::setdiff)))

suppressMessages(suppressWarnings(source("assignment-2-utils.R")))
```

Note

- Download the dataset and save to working directory (WD), see link to zip / tarball at https://www.cbioportal.org/study/summary?id=brca_tcga_pan_can_atlas_2018.

```r
path_wd <- "/Users/conorheffron/Desktop/assignment-2/"
setwd(path_wd)
```

Untar the folder and extract the files

```
dir_name <- "brca_tcga_pan_can_atlas_2018"
extension <- ".tar.gz"
untar(paste(dir_name, extension, sep=""), files = NULL, list = FALSE, exdir = ".",
      extras = NULL, verbose = FALSE,
      restore_times =  TRUE,
      support_old_tars = Sys.getenv("R_SUPPORT_OLD_TARS", FALSE),
      tar = Sys.getenv("TAR"))
```

Important

- Read the RNA Sequence data file: `data_mrna_seq_v2_rsem.txt`

```
data_mrna <- import_data(dir_name, "^data_mrna_seq_v2_rsem.txt", 0)
```

```
[1] "data_mrna_seq_v2_rsem.txt - importing data"
```

Important

- Read the Patient Data file: `data_clinical_patient.txt`

```
data_clinical <- import_data(dir_name, "^data_clinical_patient", 4)
```

```
[1] "data_clinical_patient.txt - importing data"
```

Important

- Read the Copy Number Aberrations (CNA) Data: `data_cna.txt`

```
data_cna <- import_data(dir_name, "^data_cna", 0)
```

```
[1] "data_cna_hg19.seg is not needed for import..."
[1] "data_cna.txt - importing data"
```

! Important

- Read the Samples Data: `data_clinical_sample.txt`

```
data_clinical_sample <- import_data(dir_name, "^data_clinical_sample", 4)
```

```
[1] "data_clinical_sample.txt - importing data"
```

! Important

- Create metadata using the Seq IDs of ERBB2+.

```
keep <- !duplicated(data_mrna$data_mrna_seq_v2_rsem[, 1])
temp_df_mrna <- data_mrna$data_mrna_seq_v2_rsem[keep,]
temp_df_mrna <- rownames_to_column(as.data.frame(t(data_mrna$data_mrna_seq_v2_rsem |> fi

colnames(temp_df_mrna) <- temp_df_mrna[1,]
df_mrna_seq <- temp_df_mrna[-c(1, 2),]
df_mrna_seq <- df_mrna_seq |> dplyr::rename(PATIENT_ID_REF = Hugo_Symbol)
df_mrna_seq <- df_mrna_seq |> relocate(PATIENT_ID_REF)
df_mrna_seq[, 2:5] <- sapply(df_mrna_seq[, 2:5], as.numeric)
rownames(df_mrna_seq) <- NULL
df_mrna_seq <- df_mrna_seq %>% rename_with(~ paste(., "SEQ", sep = "_"))
df_mrna_seq$PATIENT_ID <- substr(df_mrna_seq$PATIENT_ID_REF_SEQ, 1, nchar(df_mrna_seq$PA
df_mrna_seq <- df_mrna_seq |> relocate(PATIENT_ID)
```

! Important

- Create metadata using the CNA level IDs of ERBB2+ features etc.

```
temp_cna_df <- data_cna$data_cna
df_cna_ids <- rownames_to_column(temp_cna_df, "row_names")
df_cna_ids <- setNames(data.frame(t(temp_cna_df[,-1])), temp_cna_df[,1])

erbb2_cols <- df_cna_ids[, grepl("ERBB", names(df_cna_ids)) | grepl("FAM72C", names(df_c

erbb2_cols$PATIENT_ID_REF <- rownames(erbb2_cols)
erbb2_cols <- erbb2_cols |> relocate(PATIENT_ID_REF)
rownames(erbb2_cols) <- NULL
erbb2_cols = erbb2_cols[-1,]
erbb2_cols$PATIENT_ID <- substr(erbb2_cols$PATIENT_ID_REF, 1, nchar(erbb2_cols$PATIENT_I
```

Important

- Match the RNA Seq data with the CNA ids & the Patient Data
  - Pathway Enrichment (Combination of enriched patient, sample, CNA and RNA Sequence data)

```
# Merge RNA Seq data with CNA data  (ERBB2+ and other gene IDs meta data)
df_clin <- merge(x = df_mrna_seq, y = erbb2_cols, by = "PATIENT_ID", all = TRUE)

# Merge result with clinical patient data (data enrichment)
df_clin <- merge(x = df_clin, y = data_clinical$data_clinical_patient, by = "PATIENT_ID"

# Merge in sample data by patient ID
df_clin <- merge(x = df_clin, y = data_clinical_sample$data_clinical_sample, by = "PATIE
```

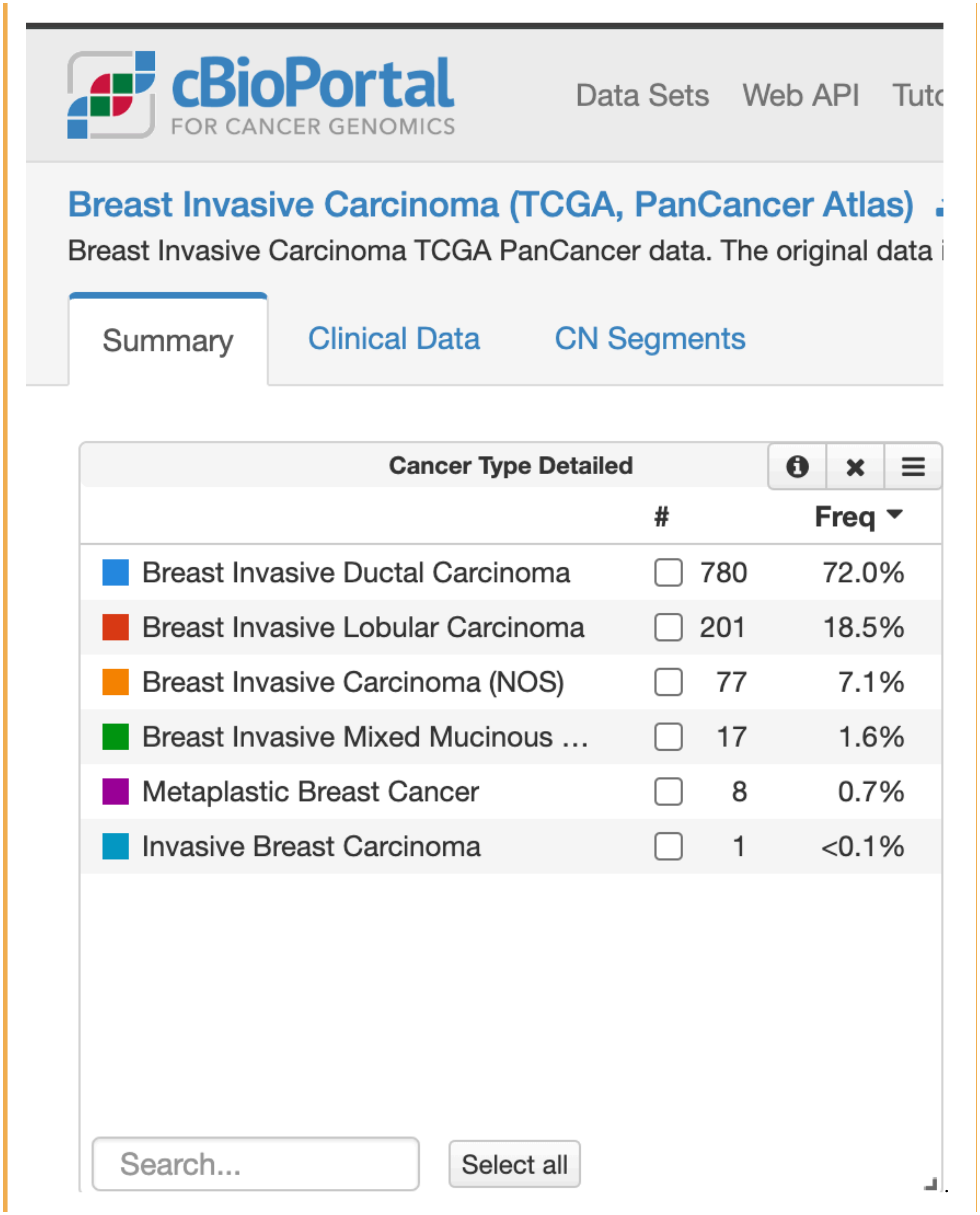
Note

- Check for top 10 mutations and have ER+ counts ready for amplified comparison (sums)

```r
temp_cna_df <- data_cna$data_cna
temp_cna_df[temp_cna_df < 0] <- 0
r_sums_cna <- temp_cna_df %>%
  mutate(rowsums = select(., -c(1:2)) %>% rowSums(na.rm = TRUE))
r_sums_cna_ss <- select(r_sums_cna, c(Hugo_Symbol, rowsums))
all_r_sums_cna <- r_sums_cna_ss[order(r_sums_cna_ss$rowsums, decreasing = T),]
ebbr_r_sums_cna <- all_r_sums_cna |> filter(grepl("ERBB", Hugo_Symbol))
```

⚠ Warning

- **Equivalent Summary Table Snippet**
    - (First High Level breakdown, followed by further breakdown with SEQ data and then ER+ data)



Breast Invasive Carcinoma (TCGA, PanCancer Atlas)

Breast Invasive Carcinoma TCGA PanCancer data. The original data i

Cancer Type Detailed

	#	Freq ▾
🟦 Breast Invasive Ductal Carcinoma	□ 780	72.0%
🟥 Breast Invasive Lobular Carcinoma	□ 201	18.5%
🟧 Breast Invasive Carcinoma (NOS)	□ 77	7.1%
🟩 Breast Invasive Mixed Mucinous …	□ 17	1.6%
🟪 Metaplastic Breast Cancer	□ 8	0.7%
🟦 Invasive Breast Carcinoma	□ 1	<0.1%

```r
count_agg(data_clinical_sample$data_clinical_sample, "CANCER_TYPE_DETAILED", n_results=2
```

CANCER_TYPE_DETAILED	n	Freq
Breast Invasive Ductal Carcinoma	780	72
Breast Invasive Lobular Carcinoma	201	19
Breast Invasive Carcinoma (NOS)	77	7
Breast Invasive Mixed Mucinous Carcinoma	17	2
Metaplastic Breast Cancer	8	1
Invasive Breast Carcinoma	1	0

```r
count_agg(df_clin, "CANCER_TYPE_DETAILED", n_results=20, digits=2)
```

CANCER_TYPE_DETAILED	n	Freq
Breast Invasive Ductal Carcinoma	780	71.96
Breast Invasive Lobular Carcinoma	201	18.54
Breast Invasive Carcinoma (NOS)	77	7.10
Breast Invasive Mixed Mucinous Carcinoma	17	1.57
Metaplastic Breast Cancer	8	0.74
Invasive Breast Carcinoma	1	0.09

```r
count_agg(df_clin |> filter(ERBB2_SEQ > 0 & ERBB2 > 0), "CANCER_TYPE_DETAILED", n_result
```

CANCER_TYPE_DETAILED	n	Freq
Breast Invasive Ductal Carcinoma	268	81.71
Breast Invasive Lobular Carcinoma	37	11.28
Breast Invasive Carcinoma (NOS)	16	4.88
Breast Invasive Mixed Mucinous Carcinoma	4	1.22
Metaplastic Breast Cancer	3	0.91

⚠ Warning

- **Pie Charts** from https://www.cbioportal.org/study/summary?id=brca_tcga_pan_can_atlas_2018 replicated as Summary Tables:

```
count_agg(df_clin, "OS_STATUS", n_results=20, digits=2)
```

OS_STATUS	n	Freq
0:LIVING	933	86.07
1:DECEASED	151	13.93

```
count_agg(df_clin, "SEX", n_results=20, digits=2)
```

SEX	n	Freq
Female	1072	98.89
Male	12	1.11

```
count_agg(df_clin, "ETHNICITY", n_results=20, digits=2)
```

ETHNICITY	n	Freq
Not Hispanic Or Latino	877	80.90
	169	15.59
Hispanic Or Latino	38	3.51

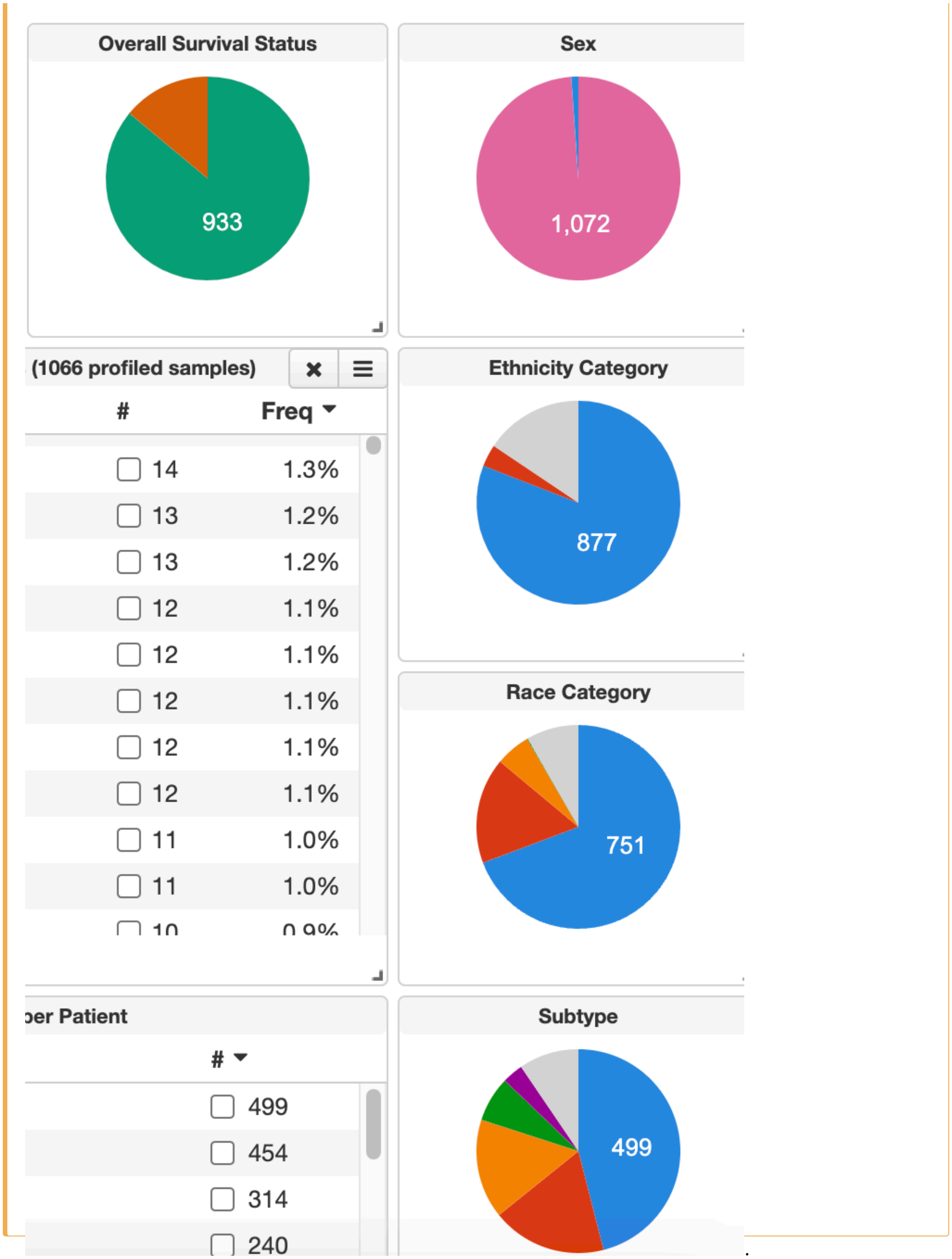
```
count_agg(df_clin, "RACE", n_results=20, digits=2)
```

RACE	n	Freq
White	751	69.28
Black or African American	182	16.79
	90	8.30
Asian	60	5.54
American Indian or Alaska Native	1	0.09

```
count_agg(df_clin, "SUBTYPE", n_results=20, digits=2)
```

SUBTYPE	n	Freq
BRCA_LumA	499	46.03
BRCA_LumB	197	18.17
BRCA_Basal	171	15.77
	103	9.50
BRCA_Her2	78	7.20
BRCA_Normal	36	3.32

- **Equivalent Charts Snippet**



Overall Survival Status

Sex

(1066 profiled samples) ✖ ☰

#	Freq ▾
□ 14	1.3%
□ 13	1.2%
□ 13	1.2%
□ 12	1.1%
□ 12	1.1%
□ 12	1.1%
□ 12	1.1%
□ 12	1.1%
□ 11	1.0%
□ 11	1.0%
□ 10	0.9%

Ethnicity Category

Race Category

per Patient

▾
□ 499
□ 454
□ 314
□ 240

Subtype

❗ Important

- **Not Amplified Summary Tables by other enrichment features**
    - Cancer type, cancer sub type, patient cancer status.

```
count_agg(df_clin, "CANCER_TYPE_ACRONYM", n_results=20, digits=2)
```

CANCER_TYPE_ACRONYM	n	Freq
BRCA	1084	100

```
count_agg(df_clin, "SUBTYPE", n_results=20, digits=2)
```

SUBTYPE	n	Freq
BRCA_LumA	499	46.03
BRCA_LumB	197	18.17
BRCA_Basal	171	15.77
	103	9.50
BRCA_Her2	78	7.20
BRCA_Normal	36	3.32

```
count_agg(df_clin, "PERSON_NEOPLASM_CANCER_STATUS", n_results=20, digits=2)
```

PERSON_NEOPLASM_CANCER_STATUS	n	Freq
Tumor Free	870	80.26
	123	11.35
With Tumor	91	8.39

❗ Important

- **ER+ Summary Tables**

```
count_agg(df_clin, "ERBB2", n_results=20, digits=2)
```

ERBB2	n	Freq
0	481	44.37
-1	260	23.99
1	206	19.00
2	123	11.35
NA	14	1.29

```
count_agg(df_clin, "ERBB2IP", n_results=20, digits=2)
```

ERBB2IP	n	Freq
0	592	54.61
-1	281	25.92
1	187	17.25
NA	14	1.29
-2	10	0.92

```
count_agg(df_clin, "ERBB3", n_results=20, digits=2)
```

ERBB3	n	Freq
0	701	64.67
1	218	20.11
-1	149	13.75
NA	14	1.29
2	2	0.18

```
count_agg(df_clin, "ERBB4", n_results=20, digits=2)
```

ERBB4	n	Freq
0	710	65.50
-1	253	23.34
1	93	8.58
NA	14	1.29
-2	7	0.65
2	7	0.65

! Important

- **ERBB2 Amplified data grouped by other columns**

```
count_agg(df_clin |> filter(ERBB2 > 0 & ERBB2_SEQ > 0), "CANCER_TYPE_ACRONYM", n_results
```

CANCER_TYPE_ACRONYM	n	Freq
BRCA	328	100

```
count_agg(df_clin |> filter(ERBB2 > 0 & ERBB2_SEQ > 0), "SUBTYPE", n_results=20, digits=
```

SUBTYPE	n	Freq
BRCA_LumA	113	34.45
BRCA_LumB	93	28.35
BRCA_Her2	62	18.90
BRCA_Basal	29	8.84
	28	8.54
BRCA_Normal	3	0.91

```
count_agg(df_clin |> filter(ERBB2 > 0 & ERBB2_SEQ > 0), "PERSON_NEOPLASM_CANCER_STATUS",
```

PERSON_NEOPLASM_CANCER_STATUS	n	Freq
Tumor Free	261	79.57
	36	10.98
With Tumor	31	9.45

! Important

- **Amplified by ERBB2 & MRNA Seq**

```
count_agg(df_clin |> filter(ERBB2 > 0 & ERBB2_SEQ > 0), "ERBB2", n_results=20, digits=2)
```

ERBB2	n	Freq
1	206	62.8

2	122	37.2

- **Amplified by ERBB2IP & MRNA Seq**

```
count_agg(df_clin |> filter(ERBB2IP > 0 & ERBB2IP_SEQ > 0), "ERBB2IP", n_results=20, dig
```

ERBB2IP	n	Freq
1	187	100

Important

- **Amplified by ERBB3 & MRNA Seq**

```
count_agg(df_clin |> filter(ERBB3 > 0 & ERBB3_SEQ > 0), "ERBB3", n_results=20, digits=2)
```

ERBB3	n	Freq
1	218	99.09
2	2	0.91

- **Amplified by ERBB4 & MRNA Seq**

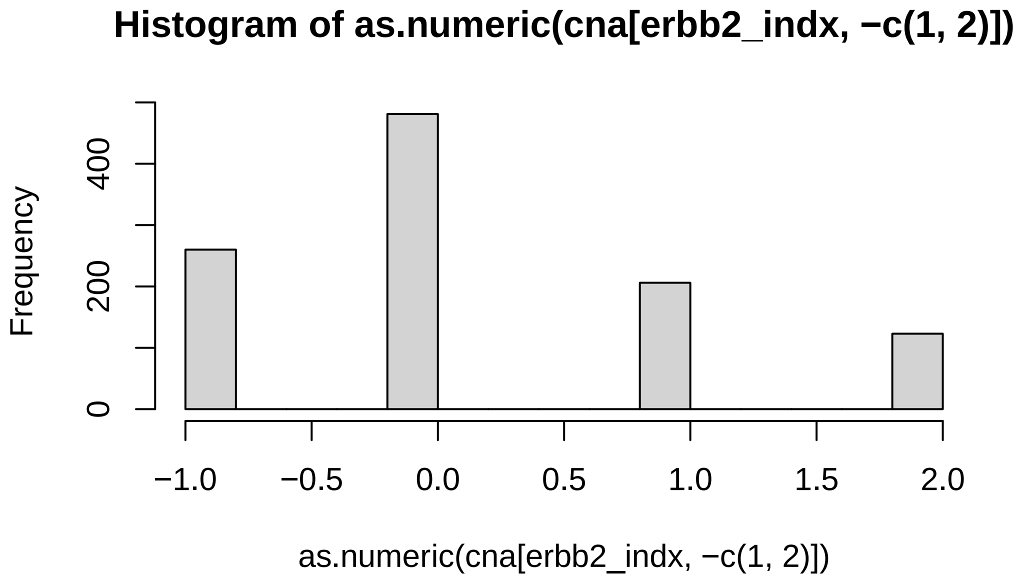
```
count_agg(df_clin |> filter(ERBB4 > 0 & ERBB4_SEQ > 0), "ERBB4", n_results=20, digits=2)
```

ERBB4	n	Freq
1	10	100

Warning

- Load guide script and compare with count variable `test_meta_erbb2_length`.

```
suppressWarnings(source("Assignment_Guide.R"))
```



- **Verify** guide script count samples amplified by ERBB2 matches my code.
- The counts now match after adding SEQ data filter for ERBB2 column (`ERBB2_SEQ > 0`)

```
test_meta_erbb2_length <- length(meta_erbb2[meta_erbb2[,"ERBB2Amp"] == 1])
test_meta_erbb2_length
```

```
[1] 328
```

```
length(meta_erbb2[meta_erbb2[,"ERBB2Amp"] == 0])
```

```
[1] 740
```

```
length(meta_erbb2[meta_erbb2[,"ERBB2Amp"] == 0]) + length(meta_erbb2[meta_erbb2[,"ERBB2A
```

```
[1] 1068
```

```
dim(rna_cna_sub)
```

```
[1] 20512  1068
```

```
test_meta_erbb2_length == dim(df_clin |> filter(ERBB2_SEQ > 0 & ERBB2 > 0))[1]
```

```
[1] TRUE
```

Differential Expression Analysis

- **BRCA HER2+: Amplified by ERBB2 & Cancer Type Detailed Summary Table**

```
count_agg(df_clin |> filter(ERBB2_SEQ > 0 & ERBB2 > 0 & SUBTYPE == "BRCA_Her2"), "CANCER
```

CANCER_TYPE_DETAILED	n	Freq
Breast Invasive Ductal Carcinoma	57	91.94
Breast Invasive Carcinoma (NOS)	2	3.23
Breast Invasive Lobular Carcinoma	2	3.23
Metaplastic Breast Cancer	1	1.61

- **BRCA HER2+: Amplified by ERBB2IP & Cancer Type Detailed Summary Table**

```
count_agg(df_clin |> filter(ERBB2IP_SEQ > 0 & ERBB2IP > 0 & SUBTYPE == "BRCA_Her2"), "CA
```

CANCER_TYPE_DETAILED	n	Freq
Breast Invasive Ductal Carcinoma	7	87.5
Breast Invasive Lobular Carcinoma	1	12.5

- **BRCA HER2+: Amplified by ERBB3 & Cancer Type Detailed Summary Table**

```
count_agg(df_clin |> filter(ERBB3_SEQ > 0 & ERBB3 > 0 & SUBTYPE == "BRCA_Her2"), "CANCER
```

CANCER_TYPE_DETAILED	n	Freq
Breast Invasive Ductal Carcinoma	17	80.95
Breast Invasive Lobular Carcinoma	3	14.29
Breast Invasive Carcinoma (NOS)	1	4.76

ℹ Note

- ERBB4 not included as it is not relevant and no amplified results to summarise.

- **BRCA HER2: ERBB2 Summary Tables**
- Removing sequence data filter because `*_SEQ` filter for HER2- does not return any results

```r
count_agg(df_clin |> filter(SUBTYPE == "BRCA_Her2"), "ERBB2", n_results=20, digits=2)
```

ERBB2	n	Freq
2	55	70.51
-1	8	10.26
0	8	10.26
1	7	8.97

```
count_agg(df_clin |> filter(SUBTYPE == "BRCA_Her2"), "ERBB2IP", n_results=20, digits=2)
```

ERBB2IP	n	Freq
-1	35	44.87
0	35	44.87
1	8	10.26

- **BRCA HER2: ERBB3 Summary Table**

```
count_agg(df_clin |> filter(SUBTYPE == "BRCA_Her2"), "ERBB3", n_results=20, digits=2)
```

ERBB3	n	Freq
0	47	60.26
1	20	25.64
-1	10	12.82
2	1	1.28

- **BRCA HER2: ERBB4 Summary Table**

```
count_agg(df_clin |> filter(SUBTYPE == "BRCA_Her2"), "ERBB4", n_results=20, digits=2)
```

ERBB4	n	Freq
0	39	50.00
-1	22	28.21
1	17	21.79

- **BRCA HER2: Cancer Type Detailed Summary Table**

```r
count_agg(df_clin |> filter(SUBTYPE == "BRCA_Her2"), "CANCER_TYPE_DETAILED", n_results=2
```

CANCER_TYPE_DETAILED	n	Freq
Breast Invasive Ductal Carcinoma	72	92.31
Breast Invasive Lobular Carcinoma	3	3.85
Breast Invasive Carcinoma (NOS)	2	2.56
Metaplastic Breast Cancer	1	1.28

- **BRCA HER2: Patient Status Summary Table**

```
count_agg(df_clin |> filter(SUBTYPE == "BRCA_Her2"), "OS_STATUS", n_results=20, digits=2
```

OS_STATUS	n	Freq
0:LIVING	63	80.77
1:DECEASED	15	19.23

- **BRCA HER2: MDM4 Summary Table**

```
count_agg(df_clin |> filter(SUBTYPE == "BRCA_Her2"), "MDM4", n_results=20, digits=2)
```

MDM4	n	Freq
1	52	66.67
0	15	19.23
2	10	12.82
-1	1	1.28

- **BRCA HER2: LRRN2 Summary Table**

```
count_agg(df_clin |> filter(SUBTYPE == "BRCA_Her2"), "LRRN2", n_results=20, digits=2)
```

LRRN2	n	Freq
1	52	66.67
0	15	19.23
2	10	12.82
-1	1	1.28

- **BRCA HER2: PIK3C2B Summary Table**

```
count_agg(df_clin |> filter(SUBTYPE == "BRCA_Her2"), "PIK3C2B", n_results=20, digits=2)
```

PIK3C2B	n	Freq
1	52	66.67
0	15	19.23
2	10	12.82
-1	1	1.28

❗ Important

- **Normalize data using DESeq2 and Run DE gene analysis, generate PCA plots**

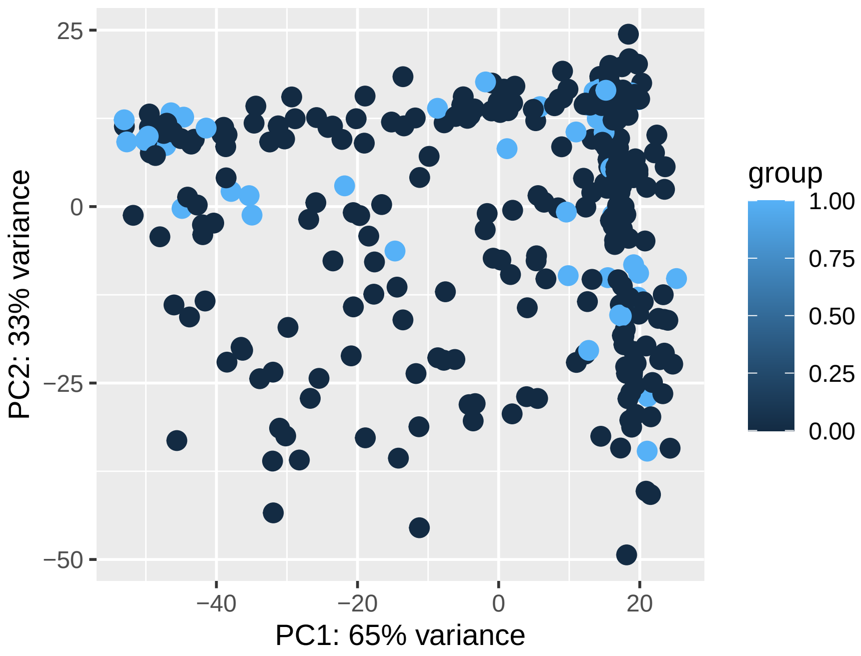
- **DE Seq Run 1 (ERBB2)**
- The 2 principal components are `ERBB2_SEQ` & `MDM4_SEQ` for `ERBB2` DE Seq Run grouped by patient status (`0` for living & `1` for deceased)

```
# Status is 1 or 0 which maps -> 0:LIVING & 1:DECEASED
de_ls1 <-
  pre_process_df(df_clin |> mutate(Status = as.numeric(substr(OS_STATUS, 1, 1))) |> filt

                   select(
                     c(
                       Status,
                       ERBB2_SEQ,
                       ERBB2IP_SEQ,
                       ERBB3_SEQ,
                       ERBB4_SEQ,
                       MDM4_SEQ,
                       LRRN2_SEQ,
                       PIK3C2B_SEQ
                     )
                   ))
dds_run1 <-
  suppressMessages(suppressWarnings(DESeqDataSetFromMatrix(
    countData = de_ls1$countdata,
    colData = de_ls1$coldata,
    design = ~ ERBB2_SEQ
  )))
 suppressMessages(suppressWarnings(de_seq_run("Status", dds_run1)))
```

```
log2 fold change (MLE): ERBB2 SEQ
Wald test p-value: ERBB2 SEQ
DataFrame with 8 rows and 6 columns
               baseMean log2FoldChange       lfcSE      stat      pvalue
              <numeric>      <numeric>   <numeric> <numeric>   <numeric>
ERBB2_SEQ   4.43262e+04    2.64257e-05 6.82781e-07 38.703108 0.00000e+00
MDM4_SEQ    1.07397e+03   -3.19709e-06 4.14565e-07 -7.711912 1.23946e-14
```

```
ERBB4_SEQ   8.70415e+02   -1.00166e-05 1.56319e-06 -6.407794 1.47640e-10
LRRN2_SEQ   6.71901e+02   -5.03708e-06 1.14855e-06 -4.385605 1.15664e-05
ERBB2IP_SEQ 2.47022e+03   -1.78001e-06 4.26535e-07 -4.173187 3.00368e-05
ERBB3_SEQ   7.39463e+03   -1.70765e-06 5.27955e-07 -3.234462 1.21872e-03
PIK3C2B_SEQ 9.46785e+02    1.10020e-06 4.76158e-07  2.310584 2.08558e-02
Status      1.70048e-01   -7.42672e-07 3.84788e-06 -0.193008 8.46952e-01
                   padj
              <numeric>
ERBB2_SEQ   0.00000e+00
MDM4_SEQ    4.95786e-14
ERBB4_SEQ   3.93708e-10
LRRN2_SEQ   2.31327e-05
ERBB2IP_SEQ 4.80588e-05
ERBB3_SEQ   1.62496e-03
PIK3C2B_SEQ 2.38352e-02
Status      8.46952e-01
```

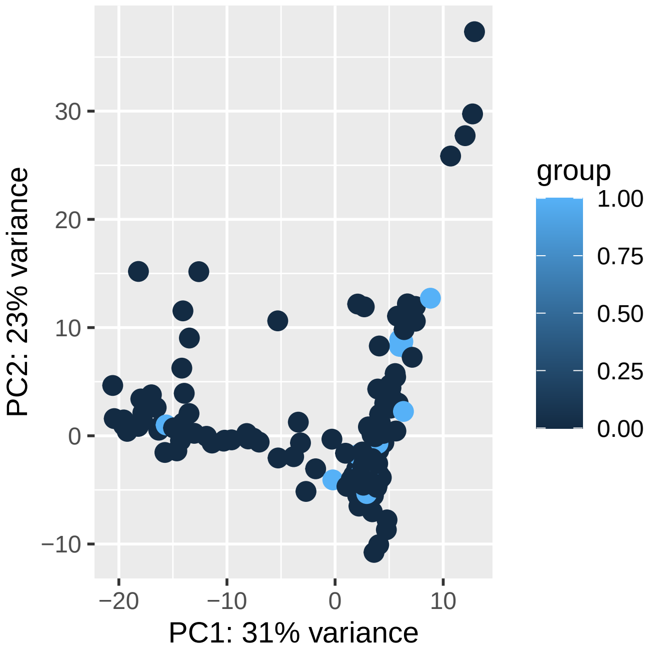


- **DE Seq Run 2 (ERBB2IP)**
- The 2 principal components are `ERBB2IP_SEQ` & `PIK3C2B_SEQ` for `ERBB2IP` DE Seq Run grouped by patient status (`0` for living & `1` for deceased)

```
de_ls2 <-
  pre_process_df(df_clin |> mutate(Status = as.numeric(substr(OS_STATUS, 1, 1))) |> filt
                   select(
                     c(
                       Status,
                       ERBB2_SEQ,
                       ERBB2IP_SEQ,
                       ERBB3_SEQ,
                       ERBB4_SEQ,
                       MDM4_SEQ,
                       LRRN2_SEQ,
                       PIK3C2B_SEQ
                     )
                   ))
dds_run2 <-
  suppressMessages(suppressWarnings(DESeqDataSetFromMatrix(
    countData = de_ls2$countdata,
    colData = de_ls2$coldata,
    design = ~ ERBB2IP_SEQ
  )))
suppressMessages(suppressWarnings(de_seq_run("Status", dds_run2)))
```

```
log2 fold change (MLE): ERBB2IP SEQ
Wald test p-value: ERBB2IP SEQ
DataFrame with 8 rows and 6 columns
               baseMean log2FoldChange       lfcSE      stat      pvalue
              <numeric>      <numeric>   <numeric> <numeric>   <numeric>
ERBB2IP_SEQ 3.02377e+03    1.73541e-04 3.19770e-05  5.427064 5.72885e-08
PIK3C2B_SEQ 8.93973e+02   -1.58682e-04 3.44888e-05 -4.600976 4.20516e-06
LRRN2_SEQ   7.82808e+02   -3.25024e-04 7.71064e-05 -4.215267 2.49482e-05
ERBB2_SEQ   1.83024e+04   -3.77534e-04 1.06985e-04 -3.528854 4.17363e-04
ERBB4_SEQ   1.00909e+03    2.74506e-04 8.87036e-05  3.094640 1.97052e-03
ERBB3_SEQ   7.91247e+03    8.90916e-05 4.60256e-05  1.935697 5.29048e-02
MDM4_SEQ    1.14282e+03   -3.17019e-05 3.90457e-05 -0.811919 4.16838e-01
Status      1.41211e-01   -2.82167e-04 1.28899e-03 -0.218906 8.26723e-01
                   padj
              <numeric>
ERBB2IP_SEQ 4.58308e-07
PIK3C2B_SEQ 1.68206e-05
LRRN2_SEQ   6.65286e-05
ERBB2_SEQ   8.34727e-04
```

```
ERBB4_SEQ   3.15283e-03
ERBB3_SEQ   7.05398e-02
MDM4_SEQ    4.76386e-01
Status      8.26723e-01
```

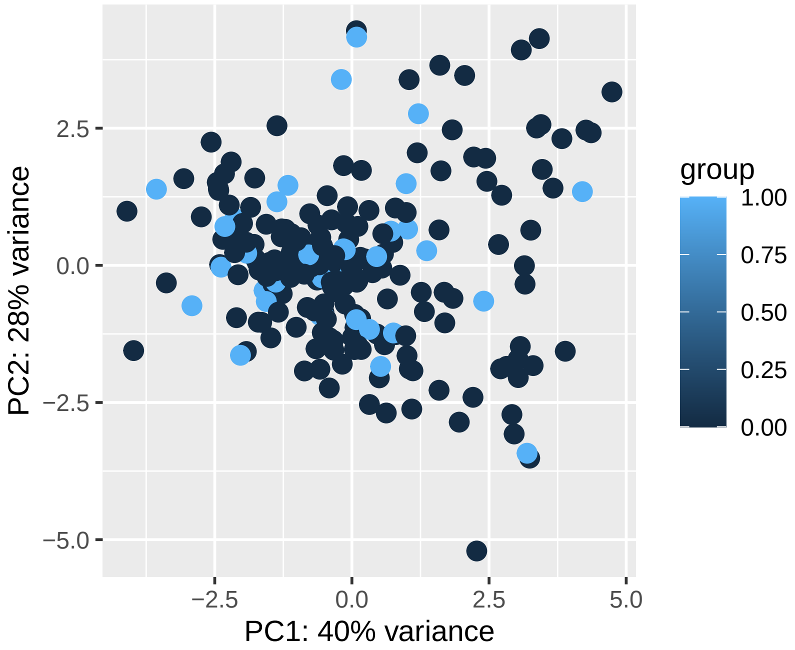


- **DE Seq Run 3 (ERBB3)**
- The 2 principal components are `ERBB3_SEQ` & `MDM4_SEQ` for `ERBB3` DE Seq Run grouped by patient status (`0` for living & `1` for deceased)

```r
de_ls3 <-
  pre_process_df(df_clin |> mutate(Status = as.numeric(substr(OS_STATUS, 1, 1))) |> filt
                   select(
                     c(
                       Status,
                       ERBB2_SEQ,
                       ERBB2IP_SEQ,
                       ERBB3_SEQ,
                       ERBB4_SEQ,
                       MDM4_SEQ,
                       LRRN2_SEQ,
                       PIK3C2B_SEQ
                     )
                   ))
dds_run3 <-
  suppressMessages(suppressWarnings(DESeqDataSetFromMatrix(
    countData = de_ls3$countdata,
    colData = de_ls3$coldata,
    design = ~ ERBB3_SEQ
  )))
suppressMessages(suppressWarnings(de_seq_run("Status", dds_run3)))
```

```
log2 fold change (MLE): ERBB3 SEQ
Wald test p-value: ERBB3 SEQ
DataFrame with 8 rows and 6 columns
               baseMean log2FoldChange       lfcSE      stat      pvalue
              <numeric>      <numeric>   <numeric> <numeric>   <numeric>
ERBB3_SEQ   9.78153e+03    8.00922e-05 6.35230e-06 12.608375 1.89868e-36
MDM4_SEQ    1.09083e+03   -2.95370e-05 7.76117e-06 -3.805738 1.41382e-04
LRRN2_SEQ   6.45159e+02   -7.78044e-05 2.00852e-05 -3.873720 1.07186e-04
PIK3C2B_SEQ 8.81717e+02   -2.88337e-05 7.79687e-06 -3.698111 2.17210e-04
ERBB4_SEQ   9.76102e+02    5.60030e-05 2.43415e-05  2.300721 2.14074e-02
Status      1.60005e-01   -6.04383e-05 7.56041e-05 -0.799405 4.24056e-01
ERBB2IP_SEQ 2.49392e+03    4.53947e-06 8.03103e-06  0.565241 5.71910e-01
ERBB2_SEQ   1.99983e+04    1.03948e-05 2.44181e-05  0.425701 6.70326e-01
                   padj
              <numeric>
ERBB3_SEQ   1.51894e-35
MDM4_SEQ    3.77018e-04
LRRN2_SEQ   3.77018e-04
PIK3C2B_SEQ 4.34420e-04
```

```
ERBB4_SEQ   3.42518e-02
Status      5.65408e-01
ERBB2IP_SEQ 6.53611e-01
ERBB2_SEQ   6.70326e-01
```



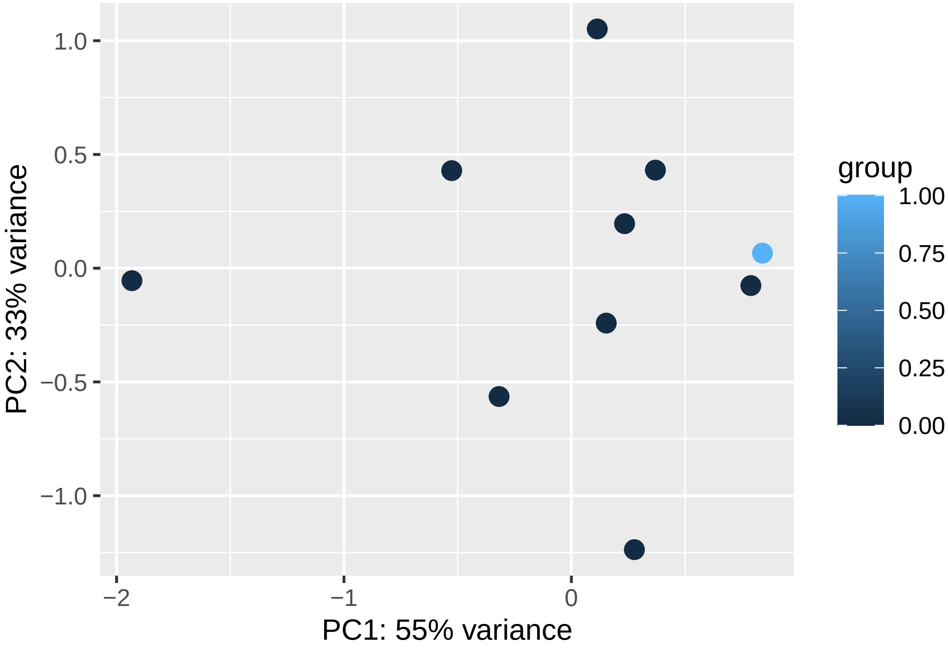
- **DE Seq Run 4 (ERBB4)**
- The 2 principal components are `ERBB4_SEQ` & `MDM4_SEQ` for `ERBB4` DE Seq Run grouped by patient status (`0` for living & `1` for deceased)

```
de_ls4 <-
  pre_process_df(df_clin |> mutate(Status = as.numeric(substr(OS_STATUS, 1, 1))) |> filt
                   select(
                     c(
                       Status,
                       ERBB2_SEQ,
                       ERBB2IP_SEQ,
                       ERBB3_SEQ,
                       ERBB4_SEQ,
                       MDM4_SEQ,
                       LRRN2_SEQ,
                       PIK3C2B_SEQ
                     )
                   ))
print(de_ls4$coldata)
```

```
      Status ERBB2_SEQ ERBB2IP_SEQ ERBB3_SEQ ERBB4_SEQ MDM4_SEQ LRRN2_SEQ
 [1,]      0      3577        3600      4916      1908      745       158
 [2,]      0      7586        1774      6981      2436     1292       393
 [3,]      0      4512        2000      3210      1916      946      2320
 [4,]      0      2638        2217      4095      2249     1022       854
 [5,]      0      7792        1811      6973      1174     1067       928
 [6,]      0      4312        1838      7305      1252      612        64
 [7,]      0      4163        3550      7711      1877      739      1302
 [8,]      0      5016        2462      7892      1228      678       454
 [9,]      0      2062        4450      3205      6078     1424       127
[10,]      1      8411        1846      8236      1301      904       981
      PIK3C2B_SEQ
 [1,]         926
 [2,]         876
 [3,]         525
 [4,]         644
 [5,]         753
 [6,]        1140
 [7,]        1482
 [8,]        1295
 [9,]         755
[10,]        1118
```

```
dds_run4 <-
  suppressMessages(suppressWarnings(DESeqDataSetFromMatrix(
    countData = de_ls4$countdata,
    colData = de_ls4$coldata,
    design = ~ ERBB4_SEQ
  )))
suppressMessages(suppressWarnings(de_seq_run("Status", dds_run4)))
```

```
log2 fold change (MLE): ERBB4 SEQ
Wald test p-value: ERBB4 SEQ
DataFrame with 8 rows and 6 columns
               baseMean log2FoldChange       lfcSE       stat      pvalue
              <numeric>      <numeric>   <numeric>  <numeric>   <numeric>
ERBB4_SEQ   2220.831633    5.27406e-04 7.66146e-05  6.8838885 5.82405e-12
MDM4_SEQ     936.774611    2.43890e-04 7.57410e-05  3.2200518 1.28167e-03
ERBB2_SEQ   4743.502364   -2.45933e-04 9.18585e-05 -2.6773035 7.42174e-03
ERBB2IP_SEQ 2593.073566    2.72591e-04 1.11572e-04  2.4431823 1.45584e-02
ERBB3_SEQ   5868.304396   -1.86969e-04 8.83662e-05 -2.1158412 3.43583e-02
LRRN2_SEQ    701.828546   -4.42582e-04 2.78488e-04 -1.5892305 1.12008e-01
PIK3C2B_SEQ  935.070295   -5.23827e-05 1.18121e-04 -0.4434672 6.57428e-01
Status         0.081226   -6.52253e-05 1.14539e-03 -0.0569459 9.54588e-01
                   padj
              <numeric>
ERBB4_SEQ   4.65924e-11
MDM4_SEQ    5.12670e-03
ERBB2_SEQ   1.97913e-02
ERBB2IP_SEQ 2.91168e-02
ERBB3_SEQ   5.49733e-02
LRRN2_SEQ   1.49344e-01
PIK3C2B_SEQ 7.51346e-01
Status      9.54588e-01
```

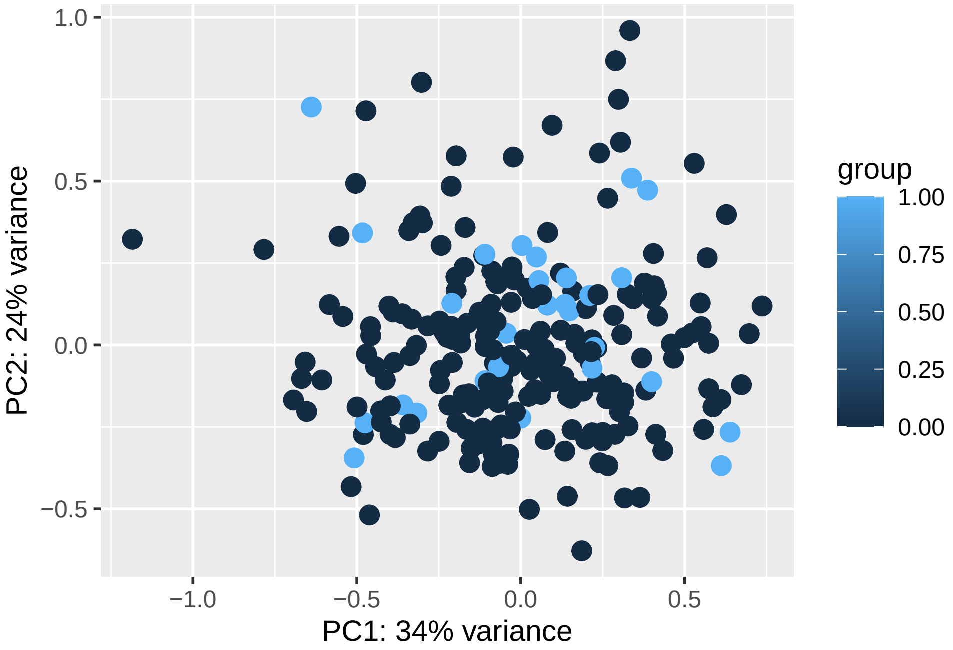


- **DE Seq Run 5 (MDM4)**
- The 2 principal components are `MDM4_SEQ` & `ERBB2IP_SEQ` for `MDM4` DE Seq Run grouped by patient status (`0` for living & `1` for deceased)

```
de_ls5 <-
  pre_process_df(df_clin |> mutate(Status = as.numeric(substr(OS_STATUS, 1, 1))) |> filt
                 select(
                   c(
                     Status,
                     ERBB2_SEQ,
                     ERBB2IP_SEQ,
                     ERBB3_SEQ,
                     ERBB4_SEQ,
                     MDM4_SEQ,
                     LRRN2_SEQ,
                     PIK3C2B_SEQ
                   )
                 ))
dds_run5 <-
  suppressMessages(suppressWarnings(DESeqDataSetFromMatrix(
    countData = de_ls5$countdata,
    colData = de_ls5$coldata,
    design = ~ MDM4_SEQ
  )))
suppressMessages(suppressWarnings(de_seq_run("Status", dds_run5)))
```

```
log2 fold change (MLE): MDM4 SEQ
Wald test p-value: MDM4 SEQ
DataFrame with 8 rows and 6 columns
               baseMean log2FoldChange       lfcSE      stat      pvalue
              <numeric>      <numeric>   <numeric> <numeric>   <numeric>
MDM4_SEQ    1413.862881    5.86591e-04 5.18331e-05 11.316922 1.08205e-29
ERBB2IP_SEQ 2428.981197   -1.47597e-04 6.88055e-05 -2.145130 3.19425e-02
LRRN2_SEQ    758.637500   -2.98945e-04 1.82434e-04 -1.638643 1.01288e-01
PIK3C2B_SEQ  911.947137   -1.35110e-04 8.24171e-05 -1.639349 1.01141e-01
ERBB2_SEQ   5385.630705   -1.07329e-04 8.53769e-05 -1.257124 2.08709e-01
Status         0.122042   -2.34863e-04 9.36742e-04 -0.250724 8.02028e-01
ERBB3_SEQ   6003.815103   -2.68901e-05 7.02650e-05 -0.382695 7.01946e-01
ERBB4_SEQ    945.032164    8.18780e-05 2.59663e-04  0.315324 7.52516e-01
                   padj
              <numeric>
MDM4_SEQ    8.65638e-29
ERBB2IP_SEQ 1.27770e-01
LRRN2_SEQ   2.02575e-01
PIK3C2B_SEQ 2.02575e-01
```

```
ERBB2_SEQ   3.33934e-01
Status      8.02028e-01
ERBB3_SEQ   8.02028e-01
ERBB4_SEQ   8.02028e-01
```

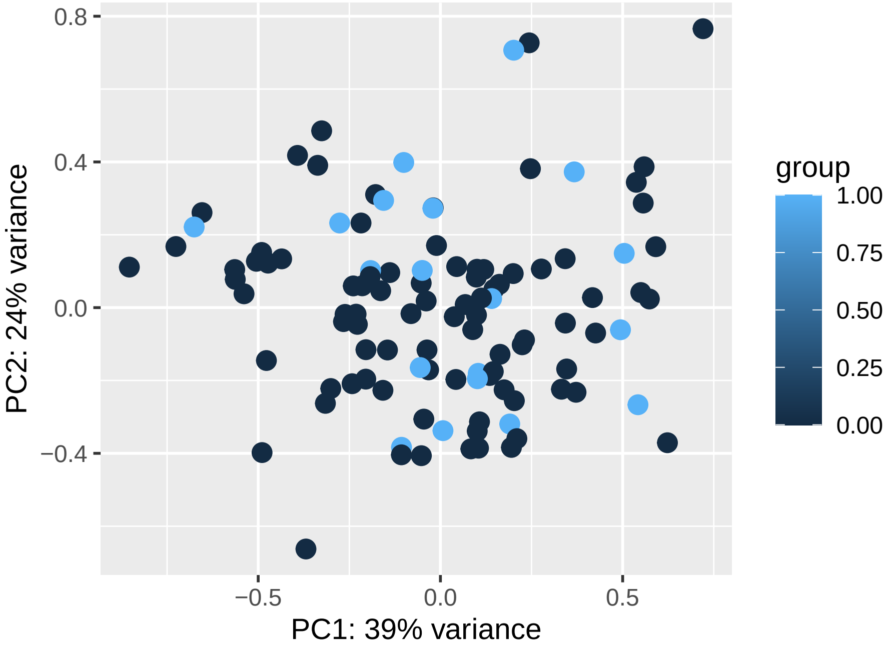


- **DE Seq Run 6 (LRNN2)**
- The 2 principal components are `LRRN2_SEQ` & `ERBB2IP_SEQ` for `LRNN2` DE Seq Run grouped by patient status (`0` for living & `1` for deceased)

```
de_ls6 <-
  pre_process_df(df_clin |> mutate(Status = as.numeric(substr(OS_STATUS, 1, 1))) |> filt
                   select(
                     c(
                       Status,
                       ERBB2_SEQ,
                       ERBB2IP_SEQ,
                       ERBB3_SEQ,
                       ERBB4_SEQ,
                       MDM4_SEQ,
                       LRRN2_SEQ,
                       PIK3C2B_SEQ
                     )
                   ))
dds_run6 <-
  suppressMessages(suppressWarnings(DESeqDataSetFromMatrix(
    countData = de_ls6$countdata,
    colData = de_ls6$coldata,
    design = ~ LRRN2_SEQ
  )))
suppressMessages(suppressWarnings(de_seq_run("Status", dds_run6)))
```

```
log2 fold change (MLE): LRRN2 SEQ
Wald test p-value: LRRN2 SEQ
DataFrame with 8 rows and 6 columns
              baseMean log2FoldChange       lfcSE      stat      pvalue
             <numeric>      <numeric>   <numeric> <numeric>   <numeric>
LRRN2_SEQ   1690.86375    5.94369e-04 5.19608e-05 11.438809 2.67533e-30
ERBB2IP_SEQ 2174.58617   -1.28748e-04 6.96626e-05 -1.848162 6.45789e-02
ERBB3_SEQ   5619.76897   -1.33413e-04 7.27702e-05 -1.833345 6.67513e-02
ERBB2_SEQ   5784.72708   -6.99742e-05 6.03491e-05 -1.159491 2.46256e-01
PIK3C2B_SEQ  841.08082   -7.59215e-05 6.91094e-05 -1.098570 2.71956e-01
ERBB4_SEQ    814.68223    2.25254e-04 2.49301e-04  0.903544 3.66237e-01
Status         0.18505   -3.91644e-04 6.73050e-04 -0.581895 5.60638e-01
MDM4_SEQ    1100.85652   -2.82411e-05 7.30647e-05 -0.386521 6.99111e-01
                   padj
              <numeric>
LRRN2_SEQ   2.14027e-29
ERBB2IP_SEQ 1.78003e-01
ERBB3_SEQ   1.78003e-01
ERBB2_SEQ   4.35129e-01
```

```
PIK3C2B_SEQ 4.35129e-01
ERBB4_SEQ   4.88316e-01
Status      6.40729e-01
MDM4_SEQ    6.99111e-01
```

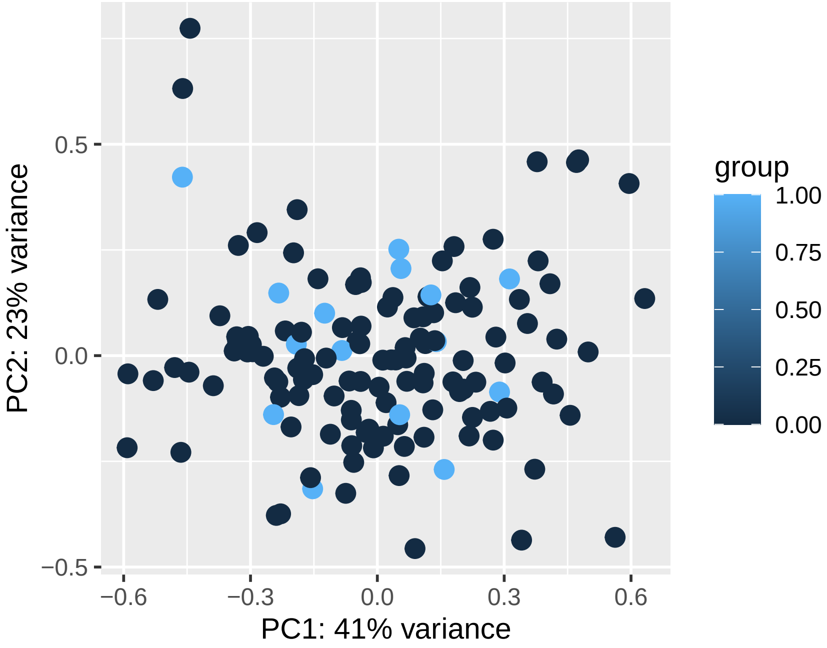


- **DE Seq Run 7 (PIK3C2B)**
- The 2 principal components are `PIK3C2B_SEQ` & `ERBB2_SEQ` for `PIK3C2B` DE Seq Run grouped by patient status (`0` for living & `1` for deceased)

```
de_ls7 <-
  pre_process_df(df_clin |> mutate(Status = as.numeric(substr(OS_STATUS, 1, 1))) |> filt
                 select(
                   c(
                     Status,
                     ERBB2_SEQ,
                     ERBB2IP_SEQ,
                     ERBB3_SEQ,
                     ERBB4_SEQ,
                     MDM4_SEQ,
                     LRRN2_SEQ,
                     PIK3C2B_SEQ
                   )
                 ))
dds_run7 <-
  suppressMessages(suppressWarnings(DESeqDataSetFromMatrix(
    countData = de_ls7$countdata,
    colData = de_ls7$coldata,
    design = ~ PIK3C2B_SEQ
  )))
suppressMessages(suppressWarnings(de_seq_run("Status", dds_run7)))
```

```
log2 fold change (MLE): PIK3C2B SEQ
Wald test p-value: PIK3C2B SEQ
DataFrame with 8 rows and 6 columns
               baseMean log2FoldChange       lfcSE      stat      pvalue
              <numeric>      <numeric>   <numeric> <numeric>   <numeric>
PIK3C2B_SEQ 1305.258863    0.000822108 0.000093869  8.758029 1.98694e-18
ERBB2_SEQ   5831.200415   -0.000413143 0.000144945 -2.850340 4.36725e-03
ERBB3_SEQ   5958.388530   -0.000302321 0.000138666 -2.180213 2.92417e-02
ERBB2IP_SEQ 2370.047650   -0.000158254 0.000124985 -1.266186 2.05447e-01
ERBB4_SEQ    851.489384   -0.000775636 0.000542085 -1.430838 1.52477e-01
MDM4_SEQ    1175.744825    0.000214832 0.000140258  1.531688 1.25599e-01
LRRN2_SEQ    700.423822   -0.000439717 0.000327689 -1.341871 1.79638e-01
Status         0.111083   -0.000508982 0.002370282 -0.214735 8.29974e-01
                   padj
              <numeric>
PIK3C2B_SEQ 1.58956e-17
ERBB2_SEQ   1.74690e-02
ERBB3_SEQ   7.79779e-02
ERBB2IP_SEQ 2.34796e-01
```

```
ERBB4_SEQ    2.34796e-01
MDM4_SEQ     2.34796e-01
LRRN2_SEQ    2.34796e-01
Status       8.29974e-01
```



❗ Important

- **Obtain Deferentially Expressed Genes**

- **Top 10 Deferentially Expressed Genes Ranked (Upgraded)**

```
knitr::kable(all_r_sums_cna[c(1:10),])
```

	Hugo_Symbol	rowsums
1313	FAM72C	974
1386	SRGAP2D	969

2094	MDM4	912
2093	PIK3C2B	910
2095	LRRN2	908
2096	NFASC	908
2103	KLHDC8A	907
2104	LEMD1-AS1	907
2108	CDK18	907
2090	PLEKHA6	906

```
# Hugo_Symbol   row_sums
# MDM4  912
# PIK3C2B   910
# LRRN2 908
# NFASC 908
# KLHDC8A   907
# CDK18 907
# ** denotes have SEQ data AND CNA data
```

- **ER+ Deferentially Expressed Genes Ranked (Upgraded)**

```
knitr::kable(ebbr_r_sums_cna)
```

Hugo_Symbol	rowsums
ERBB2	452
ERBB3	222
ERBB2IP	187
ERBB4	107

- **18 Downgraded Deferentially Expressed Genes Ranked**
    - `TNFSF` gene mutations (The Tumour Necrosis Factor Superfam) occur three times (1 combination) in the 18 downgraded ranked gene mutations. This is significant as these gene mutations could also be targeted for breast cancer treatment.

```r
knitr::kable(all_r_sums_cna[c((dim(all_r_sums_cna)[1])[1]:(dim(all_r_sums_cna)[1]-18)),]
```

	Hugo_Symbol	rowsums
18970	SOX15	52
18969	MPDU1	52
18967	SNORA67	52
18966	CD68	52
18965	SNORD10	52
18964	SNORA48	52
18963	EIF4A1	52
18961	SENP3	52
18960	SENP3-EIF4A1	52
19033	MYH2	53
19032	MYH1	53
19031	MYH4	53
18976	EFNB3	53
18975	WRAP53	53
18971	SHBG	53
18968	FXR2	53
18962	TNFSF13	53
18959	TNFSF12	53
18958	TNFSF12-TNFSF13	53

- **Summary Table per Selected Gene Mutation from Top 10 list (6x)**

```
count_agg(df_clin, "MDM4", n_results=20, digits=2)
```

MDM4	n	Freq
1	722	66.61
0	239	22.05
2	95	8.76
-1	14	1.29
NA	14	1.29

```
count_agg(df_clin, "PIK3C2B", n_results=20, digits=2)
```

PIK3C2B	n	Freq
1	724	66.79
0	240	22.14
2	93	8.58
NA	14	1.29
-1	13	1.20

```
count_agg(df_clin, "LRRN2", n_results=20, digits=2)
```

LRRN2	n	Freq
1	720	66.42
0	239	22.05
2	94	8.67
-1	16	1.48
NA	14	1.29
-2	1	0.09

```
count_agg(df_clin, "NFASC", n_results=20, digits=2)
```

NFASC	n	Freq
1	718	66.24
0	239	22.05
2	95	8.76
-1	17	1.57
NA	14	1.29
-2	1	0.09

```
count_agg(df_clin, "KLHDC8A", n_results=20, digits=2)
```

KLHDC8A	n	Freq
1	715	65.96
0	244	22.51
2	96	8.86
-1	14	1.29
NA	14	1.29
-2	1	0.09

```
count_agg(df_clin, "CDK18", n_results=20, digits=2)
```

CDK18	n	Freq
1	713	65.77
0	244	22.51
2	97	8.95
-1	15	1.38
NA	14	1.29
-2	1	0.09

Important

- **Pathway Enrichment Analysis**
  - Create base data frame for amplified data (to filter down results) and then data frame for each ERBB2+ and top gene mutation columns amplified

```
df_clin_amp_erbb_plus <- df_clin |> filter(ERBB2 > 0 | ERBB2IP > 0 | ERBB3 > 0 | ERBB2IP

df_clin_amp_erbb2 <- df_clin |> filter(ERBB2 > 0 & ERBB2_SEQ > 0)
df_clin_amp_erbb2ip <- df_clin |> filter(ERBB2IP & ERBB2IP_SEQ > 0)
df_clin_amp_erbb3 <- df_clin |> filter(ERBB3 > 0 & ERBB3_SEQ > 0)
df_clin_amp_erbb4 <- df_clin |> filter(ERBB4 > 0 & ERBB4_SEQ > 0)

df_clin_amp_top_features <- df_clin |> filter(MDM4 > 0 | PIK3C2B > 0 | LRRN2 > 0 | NFASC

df_clin_amp_mdm4 <- df_clin |> filter(MDM4 > 0 & MDM4_SEQ > 0)
df_clin_amp_pik3c2b <- df_clin |> filter(PIK3C2B & PIK3C2B_SEQ > 0)
df_clin_amp_lrrn2 <- df_clin |> filter(LRRN2 > 0 & LRRN2_SEQ > 0)
df_clin_amp_nfasc <- df_clin |> filter(NFASC > 0 & NFASC_SEQ > 0)
df_clin_amp_klhdc8a <- df_clin |> filter(KLHDC8A > 0 & KLHDC8A_SEQ > 0)
df_clin_amp_cdk18 <- df_clin |> filter(CDK18 > 0 & CDK18_SEQ > 0)
```

Important

- Get the variance stabilized transformed expression values.

```
erbbp_ls <- c(var(df_clin_amp_erbb2$ERBB2), var(df_clin_amp_erbb2ip$ERBB2IP), var(df_cli
matrix_erbbp <- matrix(erbbp_ls)
rownames(matrix_erbbp) <- c("ERBB2", "ERBB2IP", "ERBB3", "ERBB4")
colnames(matrix_erbbp) <- c("Variance")
matrix_erbbp
```

```
           Variance
ERBB2   0.234317894
ERBB2IP 1.008887832
ERBB3   0.009049398
ERBB4   0.000000000
```

```
# Show sorted matrix variance values in descending order
matrix_erbbp[order(matrix_erbbp[,1],decreasing=T),]
```

```
    ERBB2IP       ERBB2       ERBB3       ERBB4 
1.008887832 0.234317894 0.009049398 0.000000000 
```

```
erbb_seq_ls <- c(var(df_clin_amp_erbb2$ERBB2_SEQ), var(df_clin_amp_erbb2ip$ERBB2IP_SEQ),
matrix_erbb_seq <- matrix(erbb_seq_ls)
rownames(matrix_erbb_seq) <- c("ERBB2_SEQ", "ERBB2IP_SEQ", "ERBB3_SEQ", "ERBB4_SEQ")
colnames(matrix_erbb_seq) <- c("Variance")
matrix_erbb_seq
```

```
              Variance
ERBB2_SEQ   4036630410
ERBB2IP_SEQ    1186963
ERBB3_SEQ     20891406
ERBB4_SEQ      2114973
```

```
# Show sorted matrix variance values in descending order
matrix_erbb_seq[order(matrix_erbb_seq[,1], decreasing=T),]
```

```
  ERBB2_SEQ   ERBB3_SEQ   ERBB4_SEQ ERBB2IP_SEQ 
 4036630410    20891406     2114973     1186963 
```

```
# Other Top Mutations (6 from Top 10)
top_6_ls <- c(var(df_clin_amp_mdm4$MDM4), var(df_clin_amp_pik3c2b$PIK3C2B), var(df_clin_
matrix_top_6 <- matrix(top_6_ls)
rownames(matrix_top_6) <- c("MDM4", "PIK3C2B", "LRRN2", "NFASC", "KLHDC8A", "CDK18")
colnames(matrix_top_6) <- c("Variance")
matrix_top_6
```

```
          Variance
MDM4    0.11255187
PIK3C2B 0.14802490
LRRN2   0.10687089
NFASC   0.09014085
KLHDC8A 0.00000000
CDK18   0.10565544
```

```
# Show sorted matrix variance values in descending order
matrix_top_6[order(matrix_top_6[,1],decreasing=T),]
```

```
   PIK3C2B       MDM4      LRRN2      CDK18      NFASC    KLHDC8A 
0.14802490 0.11255187 0.10687089 0.10565544 0.09014085 0.00000000 
```

Conclusion

- Gene Mutations `PIK3C2B`, `MDM4`, and `LRRN2` are a good choice of gene IDs to target based on my analysis for treatment pathways. The amplified value frequencies and eventual variance values sorted in descending order from the available clinical & sequence data emphasizes this.
- Phosphatidylinositol 4-Phosphate 3-Kinase, Catalytic Sub-Unit Type 2 Beta Gene (`PIK3C2B`). The PIK3C2B gene plays a part in hormone positive breast cancer cases. A mutation in the PIK3C2B gene can cause cells to split and replicate uncontrollably. It contributes to the growth of many cancers such as Metastatic Breast Cancer (MBC). If the tumour has a PIK3C2B mutation, then new treatments that specifically target this mutation could be used for treatment.
- Mouse Double Minute 4 Homolog (`MDM4`) as a regulator of P53 is a protein coding gene. MDM4 promotes breast cancer and can impede the transcriptional activity of p53. The evidence is that MDM4 plays a notable part in breast cancer formation, progression and prognosis. It is reasonable to suggest this should be a targeted pathway.
- MDM4 is a critical regulator of the tumour supressor p53. it restricts p53 transriptional activity & enables MDM2's E3 ligase activity toward p53. These functions of MDM4 are vital for normal cell function and a true response to stress. The MDM2 gene is a gene whose product binds to p53 and regulates its functions. A differential expression of MDM2 gene in relation to Oestregen receptor status was found in human breast cancer cell lines. MDM4 is a rational target for treating breast cancers with mutated p53. It is a key driver of triple negative cancers.
- Leucine Rich Repeat Neuronal 2 (`LRRN2`) was found to be amplified and overexpressed in breast cancer along with MDM4.

Note

```
top_6_seq_ls <- c(var(df_clin_amp_mdm4$MDM4_SEQ), var(df_clin_amp_pik3c2b$PIK3C2B_SEQ)
matrix_top_6_seq <- matrix(top_6_seq_ls)
rownames(matrix_top_6_seq) <- c("MDM4", "PIK3C2B", "LRRN2", "NFASC", "KLHDC8A", "CDK18
colnames(matrix_top_6_seq) <- c("Variance")
matrix_top_6_seq
```

```
          Variance
MDM4     182025.63
PIK3C2B   83973.54
LRRN2    435329.73
NFASC   1153196.62
KLHDC8A 1275971.18
CDK18    192181.73
```

```
# Show sorted matrix variance values in descending order
matrix_top_6_seq[order(matrix_top_6_seq[,1],decreasing=T),]
```

```
   KLHDC8A      NFASC      LRRN2      CDK18       MDM4    PIK3C2B
1275971.18 1153196.62  435329.73  192181.73  182025.63   83973.54
```

Github

- https://github.com/conorheffron/gene-expr

References

- https://pubmed.ncbi.nlm.nih.gov/29617662/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5916809/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6590701/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4614407/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6047885/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3832208/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5439375/
- https://pubmed.ncbi.nlm.nih.gov/10963602/