FinalProject

Meha Goyal

2025-10-01

```
data <- read.csv('/Users/mehagoyal/Desktop/Dartmouth Projects/Final Project/counts.csv', check.names = F
```

```
meta_data <- read.csv('/Users/mehagoyal/Desktop/Dartmouth Projects/Final Project/meta_data.csv', check.n
```

```
colnames(meta_data)
```

```
##  [1] "barcode"
##  [2] "patient"
##  [3] "sample"
##  [4] "shortLetterCode"
##  [5] "definition"
##  [6] "sample_submitter_id"
##  [7] "sample_type_id"
##  [8] "tumor_descriptor"
##  [9] "sample_id"
## [10] "sample_type"
## [11] "composition"
## [12] "days_to_collection"
## [13] "state"
## [14] "initial_weight"
## [15] "preservation_method"
## [16] "pathology_report_uuid"
## [17] "submitter_id"
## [18] "oct_embedded"
## [19] "specimen_type"
## [20] "is_ffpe"
## [21] "tissue_type"
## [22] "synchronous_malignancy"
## [23] "ajcc_pathologic_stage"
## [24] "days_to_diagnosis"
## [25] "laterality"
## [26] "treatments"
## [27] "tissue_or_organ_of_origin"
## [28] "age_at_diagnosis"
## [29] "primary_diagnosis"
## [30] "metastasis_at_diagnosis"
## [31] "prior_malignancy"
## [32] "year_of_diagnosis"
## [33] "prior_treatment"
## [34] "diagnosis_is_primary_disease"
## [35] "method_of_diagnosis"
```

```
## [36] "ajcc_staging_system_edition"
## [37] "ajcc_pathologic_t"
## [38] "morphology"
## [39] "ajcc_pathologic_n"
## [40] "ajcc_pathologic_m"
## [41] "classification_of_tumor"
## [42] "diagnosis_id"
## [43] "icd_10_code"
## [44] "site_of_resection_or_biopsy"
## [45] "sites_of_involvement"
## [46] "days_to_last_follow_up"
## [47] "follow_ups_disease_response"
## [48] "race"
## [49] "gender"
## [50] "ethnicity"
## [51] "vital_status"
## [52] "age_at_index"
## [53] "days_to_birth"
## [54] "demographic_id"
## [55] "age_is_obfuscated"
## [56] "country_of_residence_at_enrollment"
## [57] "bcr_patient_barcode"
## [58] "primary_site"
## [59] "project_id"
## [60] "disease_type"
## [61] "name"
## [62] "releasable"
## [63] "released"
## [64] "figo_stage"
## [65] "figo_staging_edition_year"
## [66] "days_to_death"
## [67] "tumor_of_origin"
## [68] "days_to_sample_procurement"
## [69] "last_known_disease_status"
## [70] "tumor_grade"
## [71] "progression_or_recurrence"
## [72] "alcohol_history"
## [73] "exposure_id"
## [74] "year_of_birth"
## [75] "paper_patient"
## [76] "paper_Tumor.Type"
## [77] "paper_Included_in_previous_marker_papers"
## [78] "paper_vital_status"
## [79] "paper_days_to_birth"
## [80] "paper_days_to_death"
## [81] "paper_days_to_last_followup"
## [82] "paper_age_at_initial_pathologic_diagnosis"
## [83] "paper_pathologic_stage"
## [84] "paper_Tumor_Grade"
## [85] "paper_BRCA_Pathology"
## [86] "paper_BRCA_Subtype_PAM50"
## [87] "paper_MSI_status"
## [88] "paper_HPV_Status"
## [89] "paper_tobacco_smoking_history"
```

```
## [90] "paper_CNV.Clusters"
## [91] "paper_Mutation.Clusters"
## [92] "paper_DNA.Methylation.Clusters"
## [93] "paper_mRNA.Clusters"
## [94] "paper_miRNA.Clusters"
## [95] "paper_lncRNA.Clusters"
## [96] "paper_Protein.Clusters"
## [97] "paper_PARADIGM.Clusters"
## [98] "paper_Pan.Gyn.Clusters"
```

```
# head(data)
head(meta_data)
```

```
##                       barcode      patient           sample shortLetterCode
## 1 TCGA-GM-A2DL-01A-11R-A18M-07 TCGA-GM-A2DL TCGA-GM-A2DL-01A              TP
## 2 TCGA-AC-A2QI-01A-12R-A19W-07 TCGA-AC-A2QI TCGA-AC-A2QI-01A              TP
## 3 TCGA-A8-A06R-01A-11R-A00Z-07 TCGA-A8-A06R TCGA-A8-A06R-01A              TP
## 4 TCGA-EW-A1PD-01A-11R-A144-07 TCGA-EW-A1PD TCGA-EW-A1PD-01A              TP
## 5 TCGA-AO-A12D-01A-11R-A115-07 TCGA-AO-A12D TCGA-AO-A12D-01A              TP
## 6 TCGA-AR-A24N-01A-11R-A169-07 TCGA-AR-A24N TCGA-AR-A24N-01A              TP
##            definition sample_submitter_id sample_type_id tumor_descriptor
## 1 Primary solid Tumor    TCGA-GM-A2DL-01A              1          Primary
## 2 Primary solid Tumor    TCGA-AC-A2QI-01A              1          Primary
## 3 Primary solid Tumor    TCGA-A8-A06R-01A              1          Primary
## 4 Primary solid Tumor    TCGA-EW-A1PD-01A              1          Primary
## 5 Primary solid Tumor    TCGA-AO-A12D-01A              1          Primary
## 6 Primary solid Tumor    TCGA-AR-A24N-01A              1          Primary
##                               sample_id   sample_type  composition
## 1 d6e981ee-1afc-4f84-a3be-e87e0d1460cd Primary Tumor Not Reported
## 2 08f10bb1-6dd4-4f36-86a3-1801f9762187 Primary Tumor Not Reported
## 3 7fa91f3b-34db-41ba-8447-d745eebac2db Primary Tumor Not Reported
## 4 9f52ae8b-7fe7-441c-9fb6-c55a9bd1a0d0 Primary Tumor Not Reported
## 5 b6359ede-df10-44b9-a2e2-bbf900eeb241 Primary Tumor Not Reported
## 6 dcc31bf1-9890-4fb1-9677-b5fb77e1fc37 Primary Tumor Not Reported
##   days_to_collection    state initial_weight preservation_method
## 1               2723 released            130             Unknown
## 2                 91 released            140             Unknown
## 3               1184 released            220             Unknown
## 4                112 released            120                 OCT
## 5               1948 released            750                 OCT
## 6               2219 released            310                 OCT
##                   pathology_report_uuid submitter_id oct_embedded specimen_type
## 1 8CF4E53E-924F-4DC6-8748-BC8A1F7D5662 TCGA-GM-A2DL        false  Solid Tissue
## 2 5CF85AF1-D227-430C-99CD-C2A524CFFA29 TCGA-AC-A2QI        false  Solid Tissue
## 3 B6B7939C-63AF-4B52-B494-AA09A9C03871 TCGA-A8-A06R        false       Unknown
## 4 B6D80CA9-F4DA-4DFB-80BF-08BEE17142BD TCGA-EW-A1PD         true  Solid Tissue
## 5 9363D53D-BC3B-4AC0-8185-536C6DB3A6AE TCGA-AO-A12D         true       Unknown
## 6 AAF32BB7-716A-4C0E-AE7E-04F6A5BF5628 TCGA-AR-A24N         true  Solid Tissue
##   is_ffpe tissue_type synchronous_malignancy ajcc_pathologic_stage
## 1   FALSE       Tumor                    No               Stage I
## 2   FALSE       Tumor                    No            Stage IIIA
## 3   FALSE       Tumor                    No             Stage IIB
## 4   FALSE       Tumor                    No             Stage IIA
## 5   FALSE       Tumor                    No             Stage IIA
```

```
## 6   FALSE       Tumor                    No               Stage I
##   days_to_diagnosis laterality
## 1                 0       Left
## 2                 0      Right
## 3                 0      Right
## 4                 0      Right
## 5                 0       Left
## 6                 0       Left
##
## 1
## 2
## 3
## 4
## 5 list("Breast", NULL, NULL, NULL, NULL, "Regional Site");c("First-Line Therapy", "Adjuvant", "Adjuva
## 6
##   tissue_or_organ_of_origin age_at_diagnosis               primary_diagnosis
## 1              Breast, NOS            18475 Infiltrating duct carcinoma, NOS
## 2              Breast, NOS            27865           Lobular carcinoma, NOS
## 3              Breast, NOS            25477 Infiltrating duct carcinoma, NOS
## 4              Breast, NOS            22536 Infiltrating duct carcinoma, NOS
## 5              Breast, NOS            15774 Infiltrating duct carcinoma, NOS
## 6              Breast, NOS            19989 Infiltrating duct carcinoma, NOS
##   metastasis_at_diagnosis prior_malignancy year_of_diagnosis prior_treatment
## 1          No Metastasis               no              2003              No
## 2          No Metastasis               no              2011              No
## 3                   <NA>               no              2007              No
## 4          No Metastasis               no              2010              No
## 5                   <NA>               no              2005              No
## 6                   <NA>               no              2005              No
##   diagnosis_is_primary_disease method_of_diagnosis ajcc_staging_system_edition
## 1                         TRUE         Core Biopsy                          6th
## 2                         TRUE         Core Biopsy                          7th
## 3                         TRUE                <NA>                          5th
## 4                         TRUE         Core Biopsy                          7th
## 5                         TRUE         Core Biopsy                          6th
## 6                         TRUE         Core Biopsy                          6th
##   ajcc_pathologic_t morphology ajcc_pathologic_n ajcc_pathologic_m
## 1               T1c     8500/3           N0 (i-)                M0
## 2                T3     8520/3               N1a                MX
## 3                T2     8500/3               N1a                M0
## 4               T1c     8500/3               N1a                M0
## 5               T1c     8500/3               N1a                M0
## 6                T1     8500/3                N0                M0
##   classification_of_tumor                          diagnosis_id icd_10_code
## 1                 primary 3185a0c5-83e5-5523-81b8-6973708a75b9       C50.9
## 2                 primary 0b952469-20d9-5831-b719-8d55fd60d25b       C50.9
## 3                 primary 2804f895-5d16-5096-9ade-e6852295b87f       C50.9
## 4                 primary 4b25cd81-e384-5a36-9863-4150b90a85dc       C50.9
## 5                 primary 9e050418-bc62-5830-86f9-8b5703a7fd44       C50.9
## 6                 primary f9630094-86ba-5a9e-9e86-556b19c2b338       C50.9
##   site_of_resection_or_biopsy                sites_of_involvement
## 1                Breast, NOS             Breast, Left Upper Outer
## 2                Breast, NOS                         Breast, NOS
## 3                Breast, NOS            Breast, Right Upper Outer
```

```
## 4                 Breast, NOS                          Breast, NOS
## 5                 Breast, NOS Breast, Left Upper Outer;Breast, NOS
## 6                 Breast, NOS             Breast, Left Upper Inner
##   days_to_last_follow_up follow_ups_disease_response         race gender
## 1                   3519               TF-Tumor Free        white female
## 2                    588               TF-Tumor Free        white female
## 3                    547               TF-Tumor Free not reported female
## 4                    424               TF-Tumor Free        white   male
## 5                   2515               TF-Tumor Free        white female
## 6                   3035               TF-Tumor Free        white female
##                ethnicity vital_status age_at_index days_to_birth
## 1 not hispanic or latino        Alive           50        -18475
## 2 not hispanic or latino        Alive           76        -27865
## 3           not reported        Alive           69        -25477
## 4     hispanic or latino        Alive           61        -22536
## 5 not hispanic or latino        Alive           43        -15774
## 6 not hispanic or latino        Alive           54        -19989
##                          demographic_id age_is_obfuscated
## 1 b94ef8b8-13a9-57c5-90ec-d7b640476cd4             FALSE
## 2 de88ff20-a93a-572e-9060-34f64f281360             FALSE
## 3 365d369f-8b9a-5db8-b341-41024f115964             FALSE
## 4 42627028-1318-501d-864a-1a8645ccab14             FALSE
## 5 9585e3b9-74b5-503d-ac8e-3340e56f6a8d             FALSE
## 6 2ed6a99f-7873-5c07-b14c-f5db4727de41             FALSE
##   country_of_residence_at_enrollment bcr_patient_barcode primary_site
## 1                      United States    TCGA-GM-A2DL-01A       Breast
## 2                      United States    TCGA-AC-A2QI-01A       Breast
## 3                            Germany    TCGA-A8-A06R-01A       Breast
## 4                      United States    TCGA-EW-A1PD-01A       Breast
## 5                              <NA>    TCGA-AO-A12D-01A       Breast
## 6                      United States    TCGA-AR-A24N-01A       Breast
##   project_id
## 1  TCGA-BRCA
## 2  TCGA-BRCA
## 3  TCGA-BRCA
## 4  TCGA-BRCA
## 5  TCGA-BRCA
## 6  TCGA-BRCA
##
## 1 Fibroepithelial Neoplasms;Adnexal and Skin Appendage Neoplasms;Adenomas and Adenocarcinomas;Cystic
## 2 Fibroepithelial Neoplasms;Adnexal and Skin Appendage Neoplasms;Adenomas and Adenocarcinomas;Cystic
## 3 Fibroepithelial Neoplasms;Adnexal and Skin Appendage Neoplasms;Adenomas and Adenocarcinomas;Cystic
## 4 Fibroepithelial Neoplasms;Adnexal and Skin Appendage Neoplasms;Adenomas and Adenocarcinomas;Cystic
## 5 Fibroepithelial Neoplasms;Adnexal and Skin Appendage Neoplasms;Adenomas and Adenocarcinomas;Cystic
## 6 Fibroepithelial Neoplasms;Adnexal and Skin Appendage Neoplasms;Adenomas and Adenocarcinomas;Cystic
##                          name releasable released figo_stage
## 1 Breast Invasive Carcinoma       TRUE     TRUE       <NA>
## 2 Breast Invasive Carcinoma       TRUE     TRUE       <NA>
## 3 Breast Invasive Carcinoma       TRUE     TRUE       <NA>
## 4 Breast Invasive Carcinoma       TRUE     TRUE       <NA>
## 5 Breast Invasive Carcinoma       TRUE     TRUE       <NA>
## 6 Breast Invasive Carcinoma       TRUE     TRUE       <NA>
##   figo_staging_edition_year days_to_death tumor_of_origin
## 1                        NA            NA            <NA>
```

```
## 2                       NA            NA            <NA>
## 3                       NA            NA            <NA>
## 4                       NA            NA            <NA>
## 5                       NA            NA            <NA>
## 6                       NA            NA            <NA>
##   days_to_sample_procurement last_known_disease_status tumor_grade
## 1                         NA                      <NA>        <NA>
## 2                         NA                      <NA>        <NA>
## 3                         NA                      <NA>        <NA>
## 4                         NA                      <NA>        <NA>
## 5                         NA                      <NA>        <NA>
## 6                         NA                      <NA>        <NA>
##   progression_or_recurrence alcohol_history exposure_id year_of_birth
## 1                      <NA>            <NA>        <NA>            NA
## 2                      <NA>            <NA>        <NA>            NA
## 3                      <NA>            <NA>        <NA>            NA
## 4                      <NA>            <NA>        <NA>            NA
## 5                      <NA>            <NA>        <NA>            NA
## 6                      <NA>            <NA>        <NA>            NA
##   paper_patient paper_Tumor.Type paper_Included_in_previous_marker_papers
## 1  TCGA-GM-A2DL             BRCA                                      YES
## 2  TCGA-AC-A2QI             BRCA                                       NO
## 3  TCGA-A8-A06R             BRCA                                      YES
## 4          <NA>             <NA>                                     <NA>
## 5  TCGA-AO-A12D             BRCA                                      YES
## 6  TCGA-AR-A24N             BRCA                                      YES
##   paper_vital_status paper_days_to_birth paper_days_to_death
## 1              Alive              -18475                  NA
## 2              Alive              -27865                  NA
## 3              Alive              -25477                  NA
## 4               <NA>                  NA                  NA
## 5              Alive              -15774                  NA
## 6              Alive              -19989                  NA
##   paper_days_to_last_followup paper_age_at_initial_pathologic_diagnosis
## 1                        3519                                        50
## 2                         588                                        76
## 3                         547                                        69
## 4                          NA                                        NA
## 5                        2515                                        43
## 6                        3035                                        54
##   paper_pathologic_stage paper_Tumor_Grade paper_BRCA_Pathology
## 1                Stage_I                NA                  IDC
## 2               Stage_III                NA                  ILC
## 3                Stage_II                NA                  IDC
## 4                    <NA>                NA                 <NA>
## 5                Stage_II                NA                 <NA>
## 6                 Stage_I                NA                  IDC
##   paper_BRCA_Subtype_PAM50 paper_MSI_status paper_HPV_Status
## 1                     LumA               NA              NA
## 2                     LumA               NA              NA
## 3                     LumB               NA              NA
## 4                     <NA>               NA              NA
## 5                     Her2               NA              NA
## 6                     LumB               NA              NA
```

```
##   paper_tobacco_smoking_history paper_CNV.Clusters paper_Mutation.Clusters
## 1                            NA                C3                      C1
## 2                            NA                C1                      C4
## 3                            NA                C6                      C4
## 4                            NA              <NA>                    <NA>
## 5                            NA                C6                      C9
## 6                            NA                C1                      C1
##   paper_DNA.Methylation.Clusters paper_mRNA.Clusters paper_miRNA.Clusters
## 1                             C2                 C1                    C2
## 2                             C1                 C2                    C2
## 3                             C2                 C1                    C3
## 4                           <NA>               <NA>                  <NA>
## 5                             C4                 C2                    C3
## 6                             C2                 C1                    C3
##   paper_lncRNA.Clusters paper_Protein.Clusters paper_PARADIGM.Clusters
## 1                    C1                    C2                     C5
## 2                  <NA>                    C1                     C6
## 3                    C3                    C1                     C4
## 4                  <NA>                  <NA>                   <NA>
## 5                    C1                    C1                     C4
## 6                    C2                    C1                     C5
##   paper_Pan.Gyn.Clusters
## 1                     C3
## 2                     C1
## 3                     C1
## 4                   <NA>
## 5                     C3
## 6                     C1
```

```
num_samples <- ncol(data)
num_samples
```

```
## [1] 1232
```

```
num_genes <- nrow(data)
num_genes
```

```
## [1] 60660
```

```
gene1 <- data[1, -1]
gene2 <- data[2, -1]

gene1_name <- data[1, 1]
gene2_name <- data[2,1]

gene1_df <- data.frame(
  sample = colnames(data)[-1],
  data = as.numeric(gene1)
)

gene2_df <- data.frame(
  sample = colnames(data)[-1],
```

```
  data = as.numeric(gene2)
)

gene1_name
```

```
## [1] "ENSG00000000003.15"
```

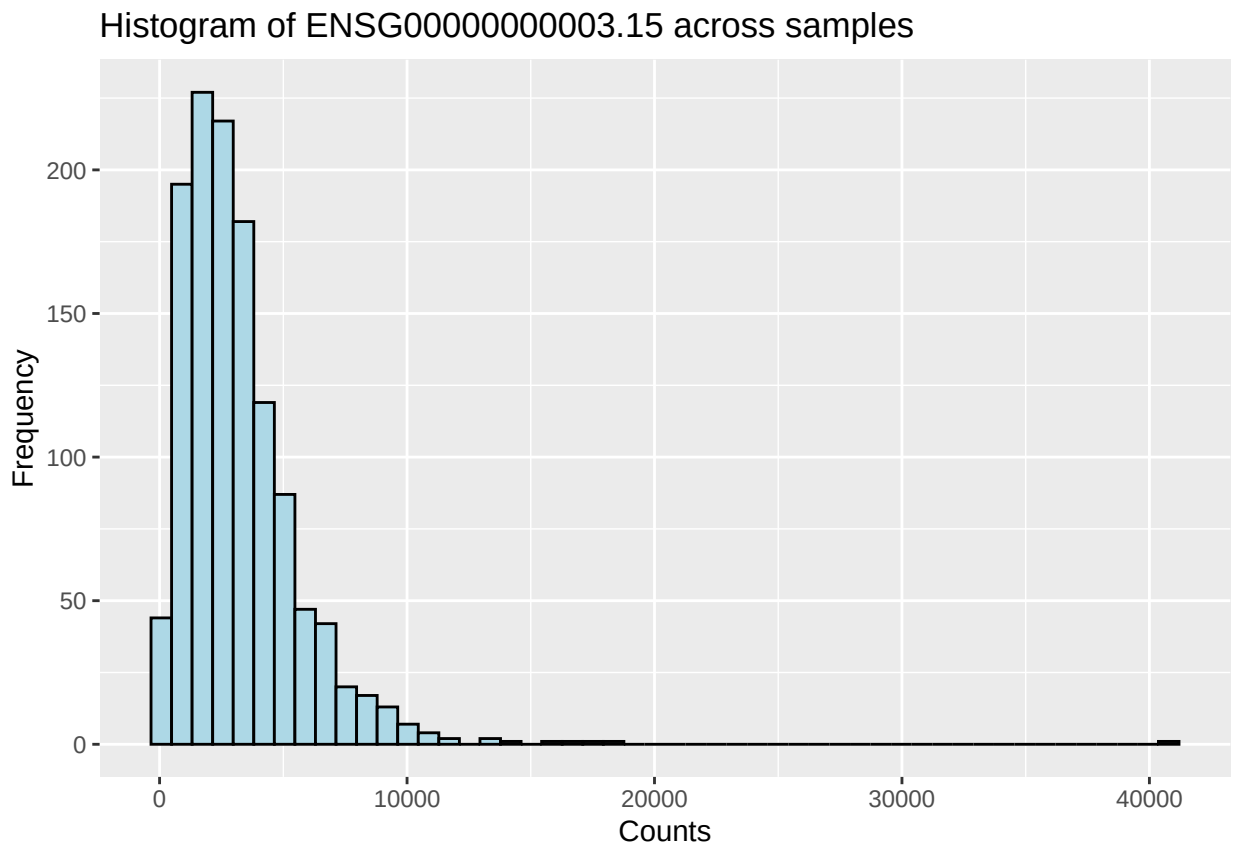
```
gene2_name
```

```
## [1] "ENSG00000000005.6"
```

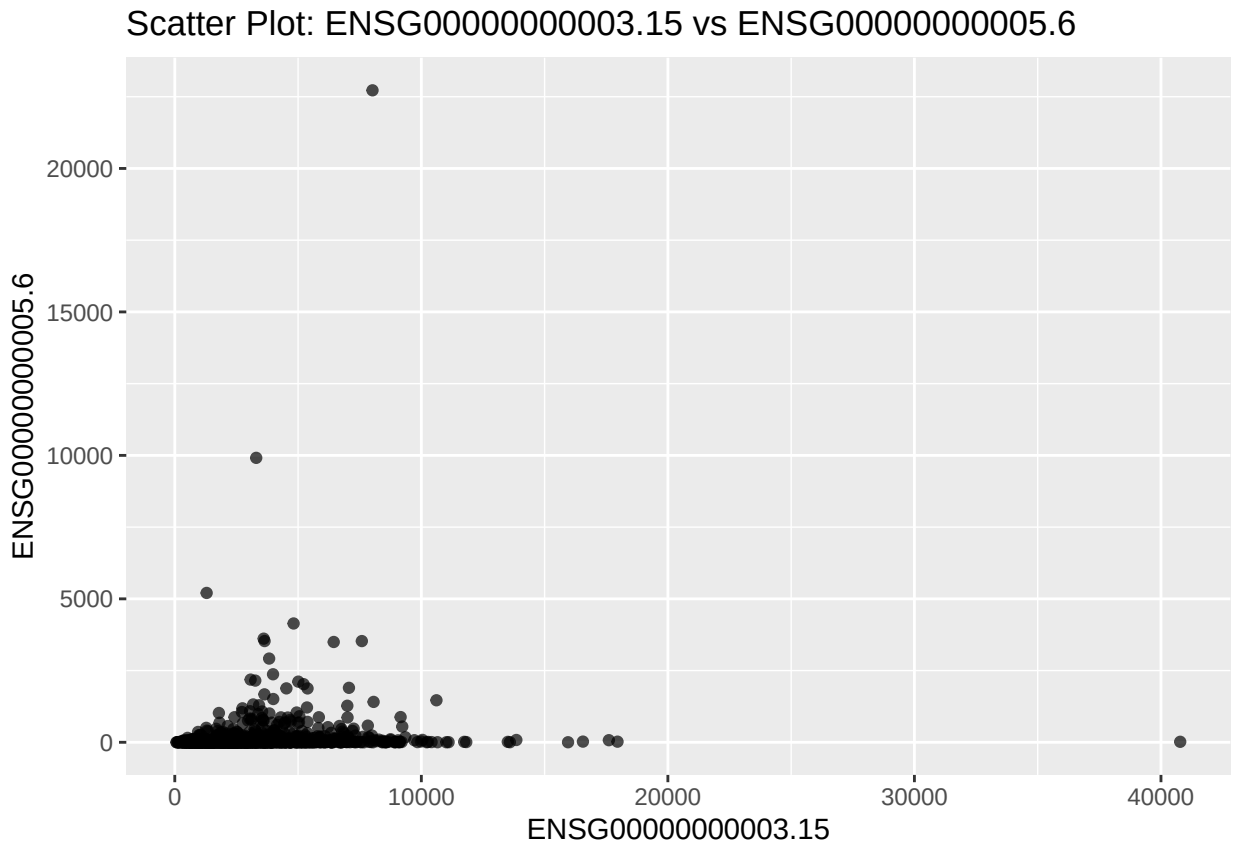
```
summary_stats <- data.frame(
  gene = gene1_name,
  mean = mean(gene1_df$data),
  median = median(gene1_df$data),
  sd = sd(gene1_df$data),
  min = min(gene1_df$data),
  max = max(gene1_df$data)
)

summary_stats
```

```
##                 gene     mean median       sd min   max
## 1 ENSG00000000003.15 3208.132   2700 2510.336  69 40780
```

```
library(ggplot2)

ggplot(gene1_df, aes(x = data)) +
  geom_histogram(bins = 50, fill = "lightblue", color = "black") +
  labs(
    title = paste("Histogram of", gene1_name, "across samples"),
    x = "Counts", y = "Frequency"
  )
```




```
scatter_df <- data.frame(
  gene1 = gene1_df$data,
  gene2 = gene2_df$data,
  sample = gene1_df$sample
)

ggplot(scatter_df, aes(x = gene1, y = gene2)) +
  geom_point(alpha = 0.7) +
  labs(
    title = paste("Scatter Plot:", gene1_name, "vs", gene2_name),
    x = gene1_name, y = gene2_name
  )
```




```
colnames(data)[1:10]
```

```
##  [1] ""                            "TCGA-GM-A2DL-01A-11R-A18M-07"
##  [3] "TCGA-AC-A2QI-01A-12R-A19W-07" "TCGA-A8-A06R-01A-11R-A00Z-07"
##  [5] "TCGA-EW-A1PD-01A-11R-A144-07" "TCGA-AO-A12D-01A-11R-A115-07"
##  [7] "TCGA-AR-A24N-01A-11R-A169-07" "TCGA-AR-A24U-01A-11R-A169-07"
##  [9] "TCGA-D8-A1JU-01A-11R-A13Q-07" "TCGA-A8-A0AD-01A-11R-A056-07"
```

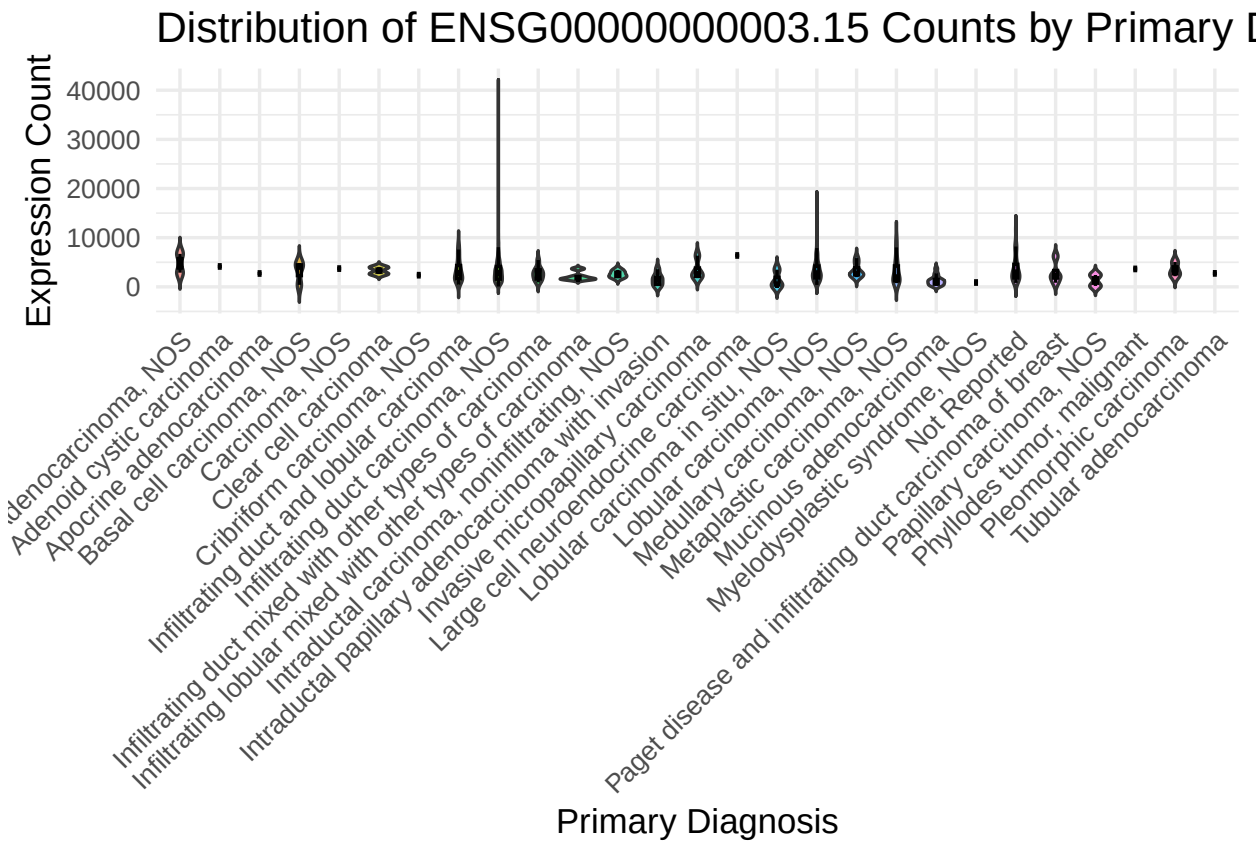
```
library(dplyr)
```

```
##
## 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
```

```
library(tidyr)
library(ggplot2)
```

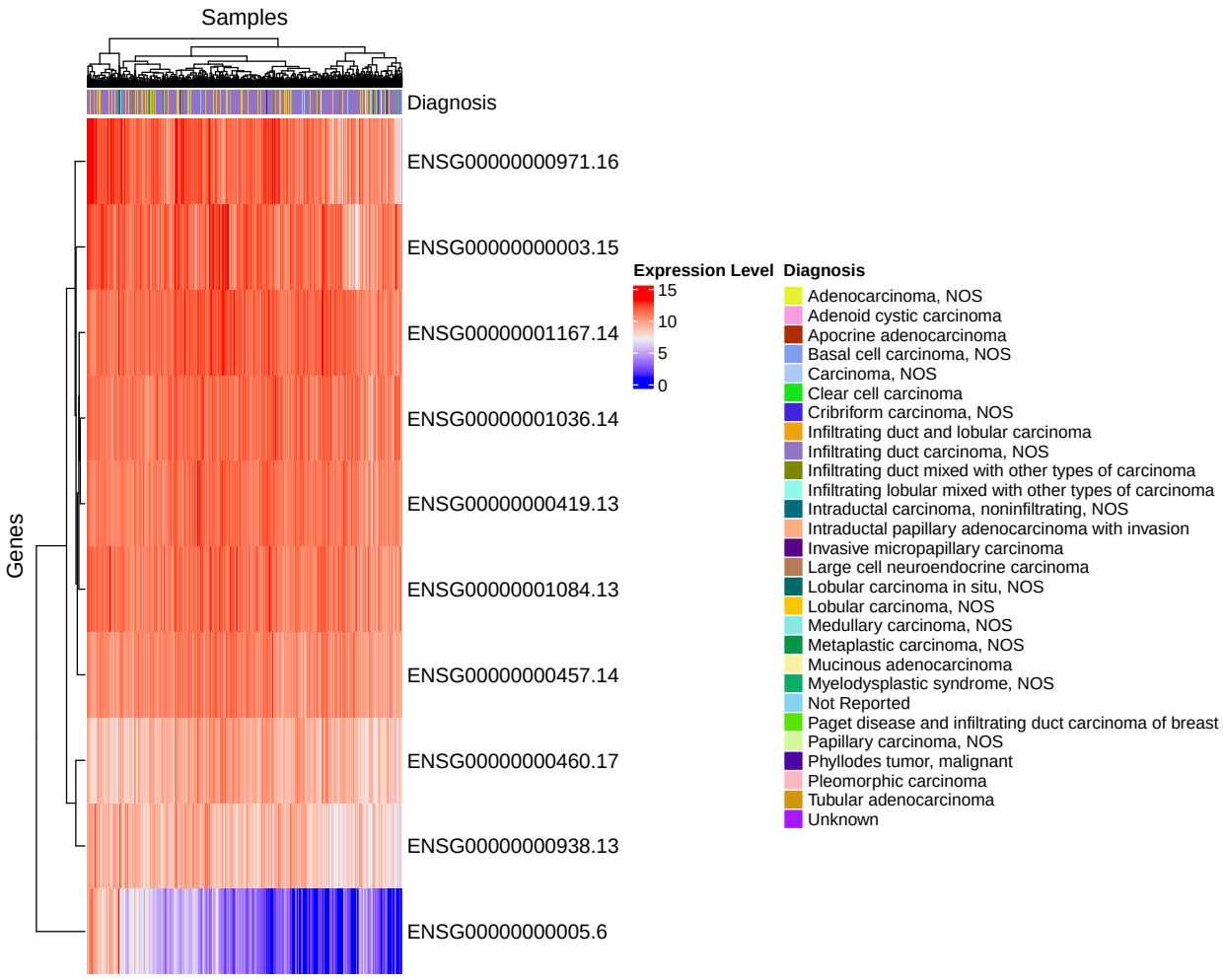
```r
if (colnames(data)[1] == "") {
  colnames(data)[1] <- "gene_id"
}

gene_col <- colnames(data)[1]

gene1_name <- data[[gene_col]][1]

data_long <- data %>%
  pivot_longer(
    cols = setdiff(colnames(data), gene_col),  # everything except gene_id
    names_to = "sample",
    values_to = "count"
  )

gene1_df <- data_long %>%
  filter(.data[[gene_col]] == gene1_name)

violin_df <- gene1_df %>%
  left_join(meta_data, by = c("sample" = "barcode")) %>%
  filter(!is.na(primary_diagnosis))

ggplot(violin_df, aes(x = primary_diagnosis, y = count, fill = primary_diagnosis)) +
  geom_violin(trim = FALSE, alpha = 0.7) +
  geom_boxplot(width = 0.1, outlier.shape = NA, alpha = 0.4, color = "black") +
  labs(
    title = paste("Distribution of", gene1_name, "Counts by Primary Diagnosis"),
    x = "Primary Diagnosis",
    y = "Expression Count"
  ) +
  theme_minimal(base_size = 13) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "none"
  )
```

```
## Warning: Groups with fewer than two datapoints have been dropped.
## i Set `drop = FALSE` to consider such groups for position adjustment purposes.
## Groups with fewer than two datapoints have been dropped.
## i Set `drop = FALSE` to consider such groups for position adjustment purposes.
## Groups with fewer than two datapoints have been dropped.
## i Set `drop = FALSE` to consider such groups for position adjustment purposes.
## Groups with fewer than two datapoints have been dropped.
## i Set `drop = FALSE` to consider such groups for position adjustment purposes.
## Groups with fewer than two datapoints have been dropped.
## i Set `drop = FALSE` to consider such groups for position adjustment purposes.
## Groups with fewer than two datapoints have been dropped.
## i Set `drop = FALSE` to consider such groups for position adjustment purposes.
## Groups with fewer than two datapoints have been dropped.
## i Set `drop = FALSE` to consider such groups for position adjustment purposes.
## Groups with fewer than two datapoints have been dropped.
## i Set `drop = FALSE` to consider such groups for position adjustment purposes.
```




```
## Loading required package: grid

## ========================================
## ComplexHeatmap version 2.22.0
## Bioconductor page: http://bioconductor.org/packages/ComplexHeatmap/
## Github page: https://github.com/jokergoo/ComplexHeatmap
## Documentation: http://jokergoo.github.io/ComplexHeatmap-reference
##
## If you use it in published research, please cite either one:
## - Gu, Z. Complex Heatmap Visualization. iMeta 2022.
## - Gu, Z. Complex heatmaps reveal patterns and correlations in multidimensional
##     genomic data. Bioinformatics 2016.
##
##
## The new InteractiveComplexHeatmap package can directly export static
## complex heatmaps into an interactive Shiny app with zero effort. Have a try!
##
## This message can be suppressed by:
##   suppressPackageStartupMessages(library(ComplexHeatmap))
## ========================================

## ========================================
## circlize version 0.4.16
## CRAN page: https://cran.r-project.org/package=circlize
## Github page: https://github.com/jokergoo/circlize
```

```
## Documentation: https://jokergoo.github.io/circlize_book/book/
##
## If you use it in published research, please cite:
## Gu, Z. circlize implements and enhances circular visualization
##   in R. Bioinformatics 2014.
##
## This message can be suppressed by:
##   suppressPackageStartupMessages(library(circlize))
## ========================================
```




```r
covariate1 <- "primary_diagnosis"
covariate2 <- "age_at_diagnosis"

table1 <- table(meta_data[[covariate1]], useNA = "ifany")
table2 <- data.frame(
  Covariate = c(covariate2),
  Mean = mean(meta_data[[covariate2]], na.rm = TRUE),
  Median = median(meta_data[[covariate2]], na.rm = TRUE),
  SD = sd(meta_data[[covariate2]], na.rm = TRUE),
  Min = min(meta_data[[covariate2]], na.rm = TRUE),
  Max = max(meta_data[[covariate2]], na.rm = TRUE)
)
```

```
# got code for formatting table cleanly from openai
knitr::kable(as.data.frame(table(meta_data$primary_diagnosis)),
             caption = "Primary Diagnosis Counts")
```

Table 1: Primary Diagnosis Counts

Var1	Freq
Adenocarcinoma, NOS	2
Adenoid cystic carcinoma	1
Apocrine adenocarcinoma	1
Basal cell carcinoma, NOS	3
Carcinoma, NOS	1
Clear cell carcinoma	2
Cribriform carcinoma, NOS	1
Infiltrating duct and lobular carcinoma	37
Infiltrating duct carcinoma, NOS	857
Infiltrating duct mixed with other types of carcinoma	20
Infiltrating lobular mixed with other types of carcinoma	5
Intraductal carcinoma, noninfiltrating, NOS	4
Intraductal papillary adenocarcinoma with invasion	6
Invasive micropapillary carcinoma	4
Large cell neuroendocrine carcinoma	1
Lobular carcinoma in situ, NOS	3
Lobular carcinoma, NOS	195
Medullary carcinoma, NOS	7
Metaplastic carcinoma, NOS	16
Mucinous adenocarcinoma	15
Myelodysplastic syndrome, NOS	1
Not Reported	37
Paget disease and infiltrating duct carcinoma of breast	4
Papillary carcinoma, NOS	2
Phyllodes tumor, malignant	1
Pleomorphic carcinoma	3
Tubular adenocarcinoma	1

```
table2
```

```
##          Covariate     Mean Median       SD  Min   Max
## 1 age_at_diagnosis 21530.01  21472 4815.423 9840 32872
```

```
library(corrplot)

colnames(data)[1] <- "Gene"

expr_only <- subset_data[, -1]
expr_only <- as.data.frame(lapply(expr_only, function(x) as.numeric(as.character(x))))

gene_expr <- t(expr_only)
colnames(gene_expr) <- data$Gene[data$Gene %in% selected_genes]
```

```r
cor_matrix <- cor(gene_expr, use = "pairwise.complete.obs", method = "pearson")

# Visualize correlation matrix
corrplot(cor_matrix,
         method = "color",
         tl.col = "black",
         tl.srt = 45,
         addCoef.col = "black",
         number.cex = 0.7,
         col = colorRampPalette(c("blue", "white", "red"))(200),
         title = "Correlation Matrix of Selected Breast Cancer Genes",
         mar = c(0,0,2,0))
```