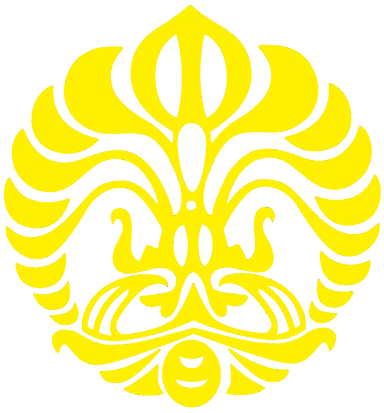
UTS SAINS DATA GENOM



Disusun oleh:

Diki Wahyudi 2106709131

PROGRAM STUDI STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS INDONESIA

OKTOBER 2023

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

PENDAHULUAN

Kanker Payudara (*Breast Cancer*) adalah penyakit di mana sel-sel di payudara tumbuh tidak terkendali. Terdapat berbagai jenis kanker payudara. Jenis kanker payudara bergantung pada sel mana di payudara yang berubah menjadi kanker. Kebanyakan kanker payudara dimulai di saluran atau lobulus. Kanker payudara dapat menyebar ke luar payudara melalui pembuluh darah dan pembuluh getah bening. Ketika kanker payudara menyebar ke bagian tubuh lain, kanker tersebut dikatakan telah bermetastasis. Jenis kanker payudara yang paling umum yaitu sebagai berikut.

1. **Karsinoma duktal invasif**. Sel-sel kanker dimulai di saluran dan kemudian tumbuh di luar saluran ke bagian lain dari jaringan payudara. Sel kanker invasif juga dapat menyebar atau bermetastasis ke bagian tubuh lain.
2. **Karsinoma lobular invasif**. Sel kanker dimulai di lobulus dan kemudian menyebar dari lobulus ke jaringan payudara di dekatnya. Sel kanker invasif ini juga dapat menyebar ke bagian tubuh lain.

Ada beberapa jenis kanker payudara lain yang kurang umum, seperti penyakit Paget, meduler (*medullary*), musinous (*mucinous*), dan inflamasi kanker payudara/*inflammatory breast cancer* (IBC). *Ductal carcinoma in situ* (DCIS) adalah penyakit payudara yang dapat menyebabkan kanker payudara invasif. Sel kanker tersebut hanya berada pada lapisan saluran dan belum menyebar ke jaringan lain di payudara [6].

Kanker payudara, yang menyerang sekitar satu dari sembilan wanita di seluruh dunia, merupakan salah satu kanker yang paling luas penyebarannya di antara tumor ganas wanita [2]. Diketahui bahwa kanker payudara memiliki heterogenitas yang tinggi pada tingkat molekuler. Menurut model pengetikan molekuler PAM50, terdapat subtipe luminal A, subtipe luminal B, subtipe positif reseptor faktor pertumbuhan epidermal manusia 2 (her2), dan subtipe kanker payudara mirip basal [10, 17]. Kanker payudara adalah penyakit kompleks dengan karakteristik genetik dan molekuler yang berubah. Memahami sifat perubahan tersebut dapat memberikan peluang untuk pendekatan pengobatan individual. Saat ini, analisis keseluruhan genom adalah salah satu cara paling efisien untuk mempelajari suatu penyakit. Banyak peneliti fokus pada penanda prognostik atau terapeutik untuk identifikasi menggunakan analisis ekspresi diferensial.

METODE

Import Library

```
library(limma)
library(ggplot2)
library(dplyr)
library(data.table)
library(knitr)
library(kableExtra)
library(genefilter)
library("annotate")
library("hgu133plus2.db")
library(GO.db)
library(magrittr)
library(ggpubr)
```

Import Data

Dataset yang akan digunakan pada analisis ini diambil dari Kaggle, yaitu GSE45827. Dataset GSE45827 berisi tentang ekspresi gen kanker payudara dari CuMiDa, terdiri atas 6 kelas, 54676 gen, dan 151 sampel. Ada 5 jenis kanker payudara (ditambah jaringan sehat) yang terdapat dalam dataset ini (kolom `type`).

```
df <- fread("D:/Materi Kuliah UI/Sains Data Genom/Tugas Sains Data Genom/Breast_GSE45827.csv",
            header = TRUE)
df <- data.frame(df)
```

Karena pada keterangan soal hanya perlu sampel sebesar 50% dari total gen, maka akan dipilih gen-gen tersebut secara *random* dengan *seed* 2106709131.

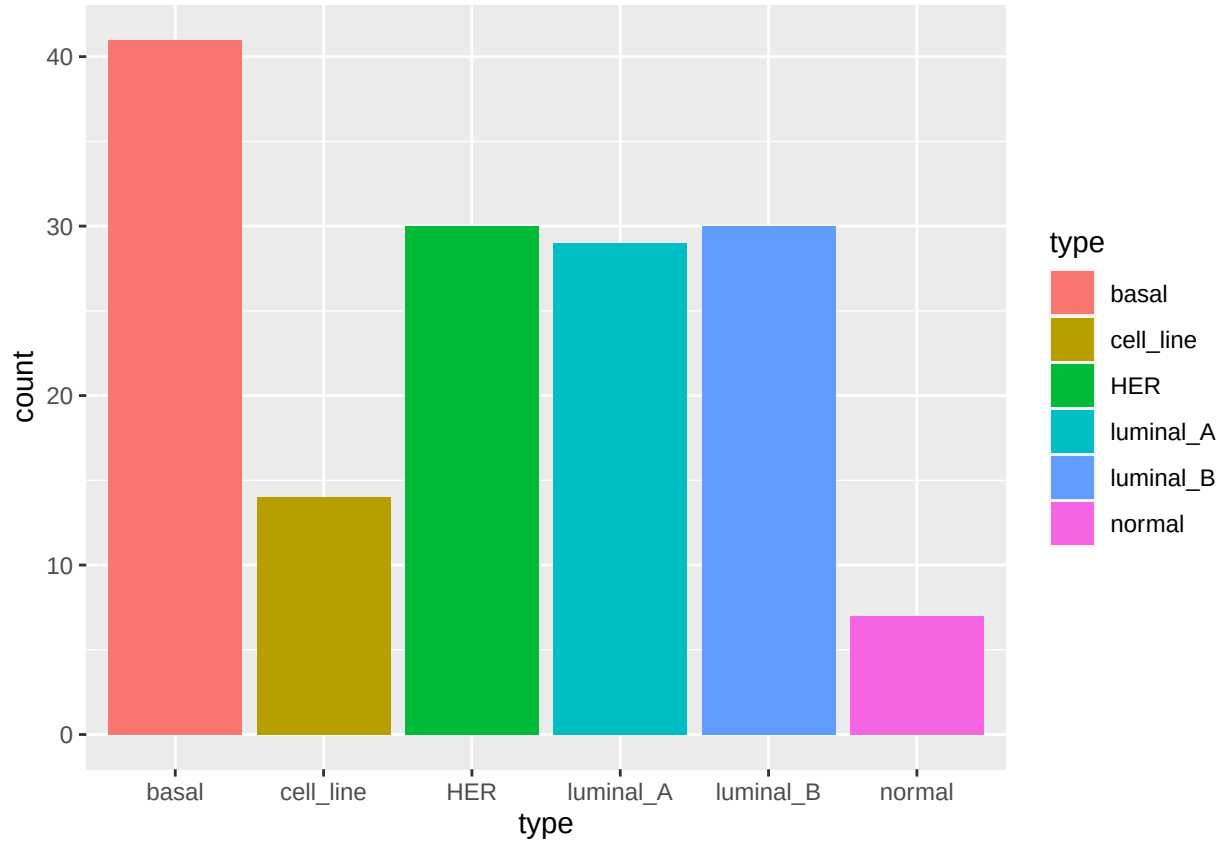
```
set.seed(2106709131)
df <- df[, c(1:2, sample(3:ncol(df), size = (ncol(df) - 2)*0.5))]
colnames(df) <- gsub("X", "", colnames(df))
```

Akan dicek berapa jumlah anggota per grup *type* kanker payudara yang ada dalam data tersebut.

```
table(df$type)
```

```
    basal cell_line       HER luminal_A luminal_B    normal
       41        14        30        29        30         7
```

```
ggplot(df, aes(x = type, fill = type)) + geom_bar()
```



Gambar 1: *Bar Plot* Jumlah Kategori *Type*

Subtipe molekular basal merupakan subtipe terbanyak pada data ini. Berikut ini merupakan keterangan tentang tipe-tipe kanker payudara tersebut [5].

- *Luminal A breast cancer*

Kanker payudara Luminal A adalah reseptor estrogen positif dan reseptor progesteron positif, HER2 negatif, dan memiliki kadar protein Ki-67 yang rendah, yang membantu mengontrol seberapa cepat sel kanker tumbuh. Kanker luminal A cenderung tumbuh lebih lambat dibandingkan kanker lainnya, memiliki tingkat keparahan yang lebih rendah, dan memiliki prognosis yang baik.

- *Luminal B breast cancer*

Kanker payudara luminal B adalah reseptor estrogen positif dan HER2 negatif, dan juga memiliki tingkat Ki-67 yang tinggi (yang menunjukkan pertumbuhan sel kanker yang lebih cepat) atau negatif terhadap reseptor progesteron.

- *Luminal B-like breast cancer*

Kanker payudara mirip luminal B adalah reseptor estrogen positif dan HER2 positif serta memiliki tingkat Ki-67 apa pun dan mungkin positif reseptor progesteron atau negatif reseptor progesteron. Kanker luminal B cenderung tumbuh lebih cepat dibandingkan kanker luminal A dan mempunyai prognosis yang sedikit lebih buruk.

- HER2-*enriched breast cancer*

Kanker payudara yang diperkaya HER2 adalah reseptor estrogen negatif dan reseptor progesteron negatif dan HER2 positif. Kanker yang diperkaya HER2 cenderung tumbuh lebih cepat dibandingkan kanker luminal dan mempunyai prognosis yang lebih buruk, namun biasanya berhasil diobati dengan obat terapi bertarget yang ditujukan pada protein HER2.

- *Triple-negative* atau *basal-like breast cancer*

Kanker payudara triple-negatif atau mirip basal adalah reseptor estrogen-negatif, reseptor progesteron-negatif, dan HER2-negatif. Kanker payudara triple-negatif lebih sering terjadi pada:

- orang dengan mutasi BRCA1;
- wanita yang lebih muda;
- perempuan dengan ras hitam.

Kanker payudara *triple-negatif* dianggap lebih agresif dibandingkan kanker payudara luminal A atau luminal B.

```
knitr::kable(df[1:10, 1:10], format = "latex", booktabs = TRUE,
             align = rep("c", 10), caption = "\\textit{Head} dari Dataframe Asli") %>%
  kableExtra::kable_styling(latex_options = c("scale_down", "HOLD_position"))
```

Tabel 1: *Head* dari Dataframe Asli

samples	type	213905_x_at	1560980_a_at	231355_at	234287_at	235625_at	223598_at	1559582_at	204341_at
84	basal	9.722219	2.753460	5.317721	5.419051	5.024762	8.432204	6.563189	8.853595
85	basal	8.359922	2.973640	5.265508	5.173865	5.256550	8.717622	6.667046	7.157017
87	basal	9.104035	2.669945	5.337311	5.168241	4.779537	7.751498	7.045690	8.001522
90	basal	8.487306	2.697166	5.031752	4.983365	4.302732	8.793953	7.434245	7.958039
91	basal	10.254484	2.737279	5.445959	5.297576	4.599508	8.459872	7.396452	9.279979
92	basal	9.513220	2.688378	5.509122	4.544255	5.394276	9.131923	7.546461	6.561909
93	basal	8.381072	2.684196	5.178012	5.537809	5.385988	8.415015	7.041150	6.623454
94	basal	9.982283	2.780352	4.946612	5.243677	4.664045	8.880032	7.192393	6.740143
99	basal	9.339658	2.680379	4.858403	5.148114	5.300146	8.614927	6.425827	6.234407
101	basal	7.250472	2.673608	5.180963	5.484878	4.725753	9.104857	7.039153	7.221436

Terlihat bahwa bentuk dataframe belum sesuai dengan format data untuk analisis *differentially expressed genes*. Oleh karena itu, data tersebut akan diubah ke dalam format yang sesuai.

```
row_name <- colnames(df[, -c(1, 2)])
column_name <- df$samples
type_c <- df$type
df <- t(df[, -c(1, 2)])
rownames(df) <- row_name
colnames(df) <- column_name
knitr::kable(df[1:8, 1:8], format = "latex", booktabs = TRUE,
             align = rep("c", 8), caption = "\\textit{Head} dari df") %>%
  kableExtra::kable_styling(latex_options = c("scale_down", "HOLD_position")) %>%
  kable_styling(position = "center")
```

Tabel 2: *Head* dari df

	84	85	87	90	91	92	93	94
213905_x_at	9.722219	8.359922	9.104035	8.487306	10.254484	9.513220	8.381072	9.982283
1560980_a_at	2.753460	2.973640	2.669945	2.697166	2.737279	2.688378	2.684196	2.780352
231355_at	5.317721	5.265508	5.337311	5.031752	5.445959	5.509122	5.178012	4.946612
234287_at	5.419051	5.173865	5.168241	4.983365	5.297576	4.544255	5.537809	5.243677
235625_at	5.024762	5.256550	4.779537	4.302732	4.599508	5.394276	5.385988	4.664045
223598_at	8.432204	8.717622	7.751498	8.793953	8.459872	9.131923	8.415015	8.880032
1559582_at	6.563189	6.667046	7.045690	7.434245	7.396452	7.546461	7.041150	7.192393
204341_at	8.853595	7.157017	8.001522	7.958039	9.279979	6.561909	6.623454	6.740143

$\log_2$ ***Transform***

Akan dicek apakah data perlu ditransformasi atau tidak.

```
qx <- as.numeric(quantile(df, c(0.0, 0.25, 0.5, 0.75, 0.99, 1.0), na.rm = TRUE))
LogC <- (qx[5]>100) || (qx[6] - qx[1]>50 && qx[2]>0)
if (LogC){
  df[which(df<=0)] <- NaN
  df <- log2(df)
}
df[1:10, 1:10]
```

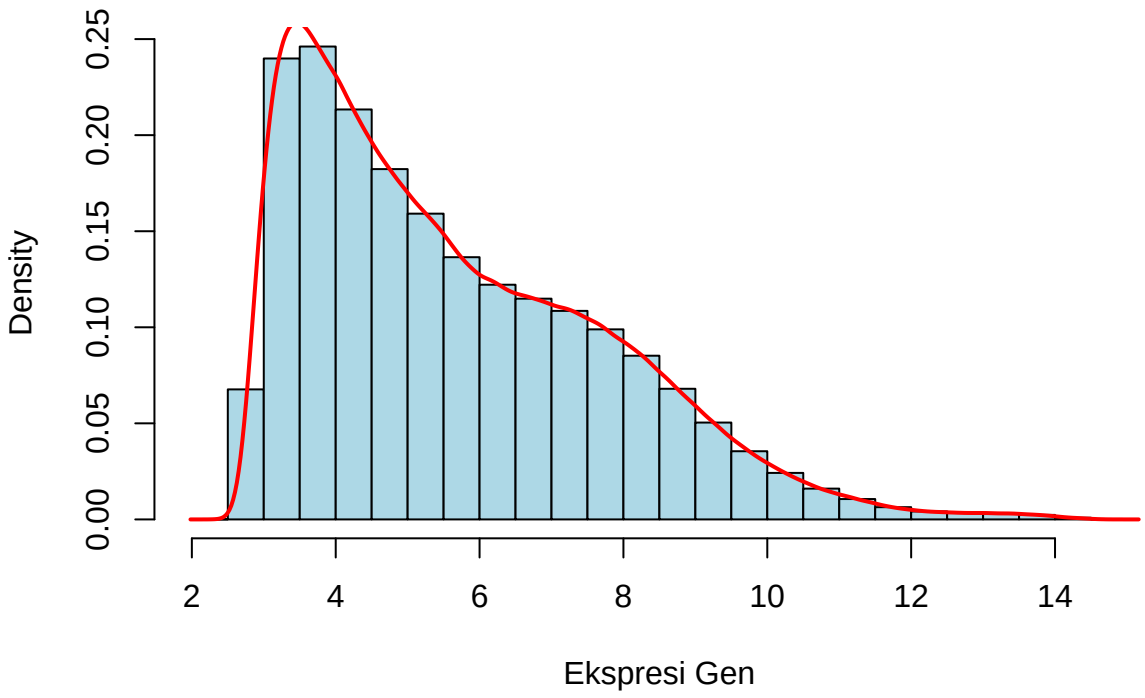
```
                  84       85       87       90        91       92       93
213905_x_at  9.722219 8.359922 9.104035 8.487306 10.254484 9.513220 8.381072
1560980_a_at 2.753460 2.973640 2.669945 2.697166  2.737279 2.688378 2.684196
231355_at    5.317721 5.265508 5.337311 5.031752  5.445959 5.509122 5.178012
234287_at    5.419051 5.173865 5.168241 4.983365  5.297576 4.544255 5.537809
235625_at    5.024762 5.256550 4.779537 4.302732  4.599508 5.394276 5.385988
223598_at    8.432204 8.717622 7.751498 8.793953  8.459872 9.131923 8.415015
1559582_at   6.563189 6.667046 7.045690 7.434245  7.396452 7.546461 7.041150
204341_at    8.853595 7.157017 8.001522 7.958039  9.279979 6.561909 6.623454
225216_at    7.679953 8.167684 8.895751 8.228383  8.482583 7.560949 8.007203
235768_at    3.700270 3.408611 3.540426 3.603497  3.133892 3.780202 3.306009
                  94       99      101
213905_x_at  9.982283 9.339658 7.250472
1560980_a_at 2.780352 2.680379 2.673608
231355_at    4.946612 4.858402 5.180963
234287_at    5.243677 5.148114 5.484878
235625_at    4.664045 5.300146 4.725753
223598_at    8.880032 8.614927 9.104857
1559582_at   7.192393 6.425827 7.039153
204341_at    6.740143 6.234407 7.221436
225216_at    8.067310 7.473952 7.509730
235768_at    3.609544 3.550193 3.721329
```

```
LogC
```

```
[1] FALSE
```

Terlihat bahwa data tersebut tidak berubah (tidak ditransformasi) karena `LogC = FALSE`.

```
hist(df, col = "lightblue", prob = TRUE, xlab = "Ekspresi Gen", main = "")
lines(density(df), lwd = 2, col = "red")
```



Gambar 2: *Density Plot* dari Ekspresi Gen

Karena deskripsi data tidak dapat dilihat detailnya, maka diasumsikan data telah ditransformsikan ke dalam bentuk $\log_2$.

Gene Filtering

Selanjutnya, akan dilakukan *gene filtering*. Hal tersebut dilakukan untuk mengeluarkan gen-gen yang tidak banyak bervariasi antarsampel dan memiliki ekspresi yang kecil di seluruh sampel. Proses ini dilakukan agar dapat mengurangi terjadinya *false positif* (kesalahan tipe I yaitu $\alpha = \Pr(\text{Menolak } H_0 | H_0 \text{ benar})$) yang akan meningkatkan *power* (peluang hasil uji statistik untuk bebas dari kesalahan statistik tipe II) dari analisis data ini. *Gene filtering* akan dilakukan sesuai prosedur pada sumber [16].

```r
f1 <- pOverA(0.25, log2(100))
f2 <- function(x)(IQR(x)>0.5)
ff <- filterfun(f1, f2)
selected <- genefilter(df, ff)
# Sebelum filtering
df2 <- df
dim(df2)
```

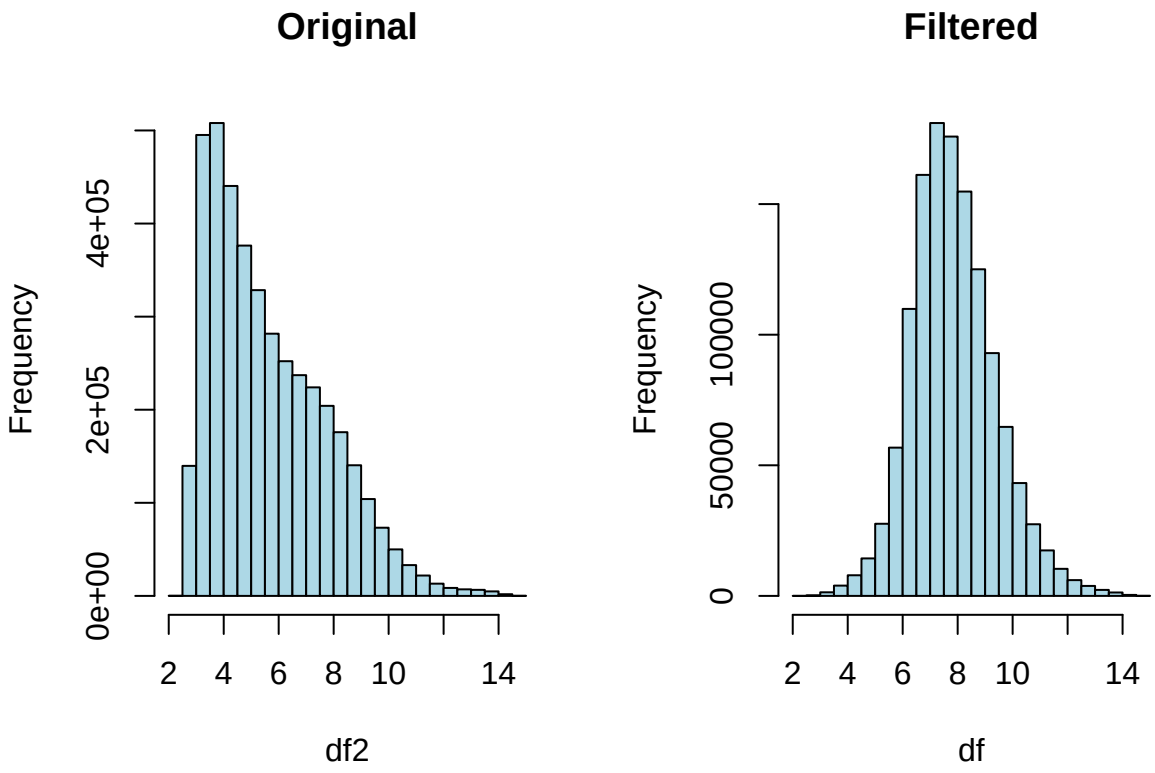
```
[1] 27337   151
```

```r
# Sesudah filtering
df <- df[selected, ]
dim(df)
```

```
[1] 8543  151
```

Setelah tahap tersebut, akan dilakukan *plotting* untuk melihat perbedaan ekspresi gen sebelum dan sesudah *filtering*.

```r
par(mfrow = c(1, 2))
hist(df2, main = "Original", col = "lightblue")
hist(df, main = "Filtered", col = "lightblue")
```



Gambar 3: Grafik Perbandingan Ekspresi Gen sebelum dan sesudah *Filtering*

```
par(mfrow = c(1, 1))
```

Dari hasil di atas, didapatkan informasi bahwa sebelum dilakukan *filtering* terdapat 27337 *features*/*gene* dan setelah *filtering* tersisa 8543 gen. Gen yang disaring merupakan gen yang memiliki ekspresi rendah seperti terlihat pada histogram di atas.

LIMMA (*Linear Models for Microarray Data*)

LIMMA adalah *library* untuk analisis data ekspresi gen microarray, khususnya digunakan pada model linier untuk menganalisis eksperimen yang dirancang dari *differential expression*. LIMMA memberikan kemampuan untuk menganalisis perbandingan antara banyak target RNA secara bersamaan dalam eksperimen yang dirancang rumit dan *random*. Metode empiris Bayesian digunakan untuk memberikan hasil yang stabil meskipun jumlah arraynya sedikit. Model linier dan fungsi ekspresi diferensial berlaku untuk semua teknologi ekspresi gen, termasuk microarray, RNA-seq, dan PCR kuantitatif [11].

Proses selanjutnya akan menggunakan metode LIMMA untuk mengetahui ekspresi gen mana yang berbeda, sesuai kriteria soal.

Bagian 1

Pertama, akan dicari gen yang berekspresi berbeda antara empat subtipe kanker: Luminal A, Luminal B, Basal, dan HER.

```
unique(type_c)
```

```
[1] "basal"     "HER"       "cell_line" "normal"    "luminal_A" "luminal_B"
```

```
table(type_c)
```

```
type_c
    basal cell_line       HER luminal_A luminal_B    normal
       41        14        30        29        30         7
```

```r
kanker <- c("luminal_A", "luminal_B", "basal", "HER")
exp1 <- df[, which(type_c %in% kanker)]
dim(exp1)
```

```
[1] 8543  130
```

```r
knitr::kable(exp1[1:10, 1:10], format = "latex", booktabs = TRUE,
             align = rep("c", 10), caption = "\\textit{Head} dari exp1") %>%
  kableExtra::kable_styling(latex_options = c("scale_down", "HOLD_position")) %>%
  kable_styling(position = "center")
```

Tabel 3: *Head* dari exp1

	84	85	87	90	91	92	93	94	99	101
213905_x_at	9.722219	8.359922	9.104035	8.487306	10.254484	9.513220	8.381072	9.982283	9.339658	7.250472
223598_at	8.432204	8.717622	7.751498	8.793953	8.459872	9.131923	8.415015	8.880032	8.614927	9.104857
1559582_at	6.563189	6.667046	7.045690	7.434245	7.396452	7.546461	7.041150	7.192393	6.425827	7.039153
204341_at	8.853595	7.157017	8.001522	7.958039	9.279979	6.561909	6.623454	6.740143	6.234407	7.221436
225216_at	7.679953	8.167684	8.895751	8.228383	8.482583	7.560949	8.007203	8.067310	7.473952	7.509730
227262_at	7.768290	8.849295	7.454536	7.567084	8.092288	6.892589	10.017969	8.461030	8.999427	8.993862
201950_x_at	9.589339	9.895621	10.138898	10.210103	10.034316	9.712977	9.354254	10.229080	9.957247	10.023711
226104_at	7.173663	7.105331	7.337900	7.391662	7.918452	6.512620	7.381303	7.573594	7.095866	7.063872
242932_at	6.874082	5.813877	5.625261	5.897511	7.078919	6.633802	6.187999	5.581978	6.189158	5.971807
225671_at	5.704296	6.688602	5.459599	5.900380	6.500822	5.695091	5.686979	6.379044	6.587404	6.716904

Akan dibuat matriks model (*design matrix*) berdasarkan variabel `type1`. Dalam model matriks, akan digunakan `+ 0` dalam model yang menyetel intersep ke 0 sehingga efek `type1` menangkap ekspresi untuk grup tersebut, bukan perbedaan dari grup terhadap *base level*.

```r
# Create the design matrix
type1 <- type_c[which(type_c %in% kanker)]
des_mat <- model.matrix(~ type1 + 0)
head(des_mat, 15)
```

```
   type1basal type1HER type1luminal_A type1luminal_B
1           1        0              0              0
2           1        0              0              0
3           1        0              0              0
4           1        0              0              0
```

5	1	0	0	0
6	1	0	0	0
7	1	0	0	0
8	1	0	0	0
9	1	0	0	0
10	1	0	0	0
11	1	0	0	0
12	1	0	0	0
13	1	0	0	0
14	1	0	0	0
15	1	0	0	0

Selanjutnya, akan dilakukan *fitting* model *differential expression* pada data. Model limma untuk data ini yaitu

$$\boldsymbol{Y}_j^T = (Y_{1j}, ..., Y_{N_j})$$

$$E(\boldsymbol{Y}_j) = \boldsymbol{X}\boldsymbol{\beta}_j$$

di mana $\boldsymbol{Y}_j$: ekspresi gen, $\boldsymbol{X}$: matriks model (*design matrix*) yang *full rank* (misalnya kondisi grup), dan $\boldsymbol{\beta}_j$: vektor efek dari kolom di matriks $\boldsymbol{X}$, $\boldsymbol{\beta}_j^T = (\beta_{j1}, \beta_{j2}, \beta_{j3}, \beta_{j4})$ dengan β_{jk} merupakan ekspektasi dari level ekspresi dari gen j dalam grup k. Keterangan level: Basal = 1, HER = 2, Luminal A = 3, dan Luminal B = 4.

```
# Apply linear model to data
fit <- lmFit(exp1, design = des_mat)
# Apply empirical Bayes to smooth standard errors
fit <- eBayes(fit)
```

Setelah *fitting* model, ingin diselidiki perbedaan di antara semua kelompok dengan menggunakan *contrast* sebagai berikut.

$$\boldsymbol{C}_1 = \begin{bmatrix} 0 & -1 & 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 & -1 & -1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

$$
\begin{aligned}
\boldsymbol{T}_j &= \boldsymbol{C}_1^T \boldsymbol{\beta}_j \\
&= \begin{bmatrix} 0 & 0 & 1 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 0 & 1 \\ 1 & -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_{j1} \\ \beta_{j2} \\ \beta_{j3} \\ \beta_{j4} \end{bmatrix} \\
&= \begin{bmatrix} \beta_{j3} - \beta_{j4} \\ \beta_{j3} - \beta_{j1} \\ \beta_{j3} - \beta_{j2} \\ \beta_{j4} - \beta_{j1} \\ \beta_{j4} - \beta_{j2} \\ \beta_{j1} - \beta_{j2} \end{bmatrix}
\end{aligned} \tag{1}
$$

Keterangan: $\boldsymbol{T}_j$ merupakan ekspektasi dari perbedaan dalam level ekspresi dari gen j dengan perbandingan Luminal A versus Luminal B, Luminal A versus Basal, Luminal A versus HER, Luminal B versus Basal, Luminal B versus HER, dan Basal versus HER.

```
contrast_matrix <- makeContrasts(
  "luminal_A.vs.luminal_B" = type1luminal_A - type1luminal_B,
  "luminal_A.vs.basal" = type1luminal_A - type1basal,
  "luminal_A.vs.HER" = type1luminal_A - type1HER,
  "luminal_B.vs.basal" = type1luminal_B - type1basal,
  "luminal_B.vs.HER" = type1luminal_B - type1HER,
  "basal.vs.HER" = type1basal - type1HER,
  levels = colnames(des_mat)
)
contrast_matrix
```

```
               Contrasts
Levels          luminal_A.vs.luminal_B luminal_A.vs.basal luminal_A.vs.HER
  type1basal                         0                -1                0
```

```
  type1HER                          0                 0              -1
  type1luminal_A                    1                 1               1
  type1luminal_B                   -1                 0               0
               Contrasts
Levels          luminal_B.vs.basal luminal_B.vs.HER basal.vs.HER
  type1basal                    -1                0           1
  type1HER                       0               -1          -1
  type1luminal_A                 0                0           0
  type1luminal_B                 1                1           0
```

```
fit <- contrasts.fit(fit, contrast_matrix)
```

Beberapa koreksi pengujian diperlukan setiap kali beberapa pengujian hipotesis (*multiple hypothesis tests*) dilakukan, untuk meminimalkan jumlah *false positif* yang diperoleh. Dalam analisis ini, akan digunakan metode *False Discovery Rate* (FDR) untuk melakukan koreksi *multiple hypothesis tests*, dan menetapkan batas signifikansi pada 0.05. Ini berarti bahwa hanya gen dengan nilai $p-value$ yang disesuaikan dengan FDR < 0.05 dan perubahan $\log_2$ absolut sebesar 1 atau lebih yang akan dianggap *berbeda secara signifikan.*

```
# Identifying differentially expressed genes
results <- decideTests(fit, p.value = 0.05, adjust.method = "fdr")
summary(results)
```

```
       luminal_A.vs.luminal_B luminal_A.vs.basal luminal_A.vs.HER
Down                     1313               2928             2179
NotSig                   6064               2821             4311
Up                       1166               2794             2053
       luminal_B.vs.basal luminal_B.vs.HER basal.vs.HER
Down                 2656             1532         1680
NotSig               3255             5509         5229
Up                   2632             1502         1634
```

Didapatkan informasi bahwa subtipe kanker yang paling banyak berbeda secara signifikan yaitu luminal A dan basal, dengan jumlah yang signifikan berbeda sebanyak $2928 + 2794 = 5722$.

Selanjutnya, akan dibuat tabel hasil berdasarkan model yang dilengkapi kontras. Langkah ini akan

menerapkan koreksi *multiple hypothesis tests* Benjamini-Hochberg. Default fungsi `topTable()` adalah menggunakan metode koreksi Benjamini-Hochberg.

```
# Re-smooth the Bayes
contrasts_fit <- eBayes(fit)
# Apply multiple testing correction and obtain stats
stats_df <- topTable(contrasts_fit, number = nrow(exp1)) %>%
  tibble::rownames_to_column("Gene")
knitr::kable(head(stats_df, 15), format = "latex", booktabs = TRUE,
             align = rep("c", 11), caption = "\\textit{Head} dari stats\\_df") %>%
  kableExtra::kable_styling(latex_options = c("scale_down", "HOLD_position"))
```

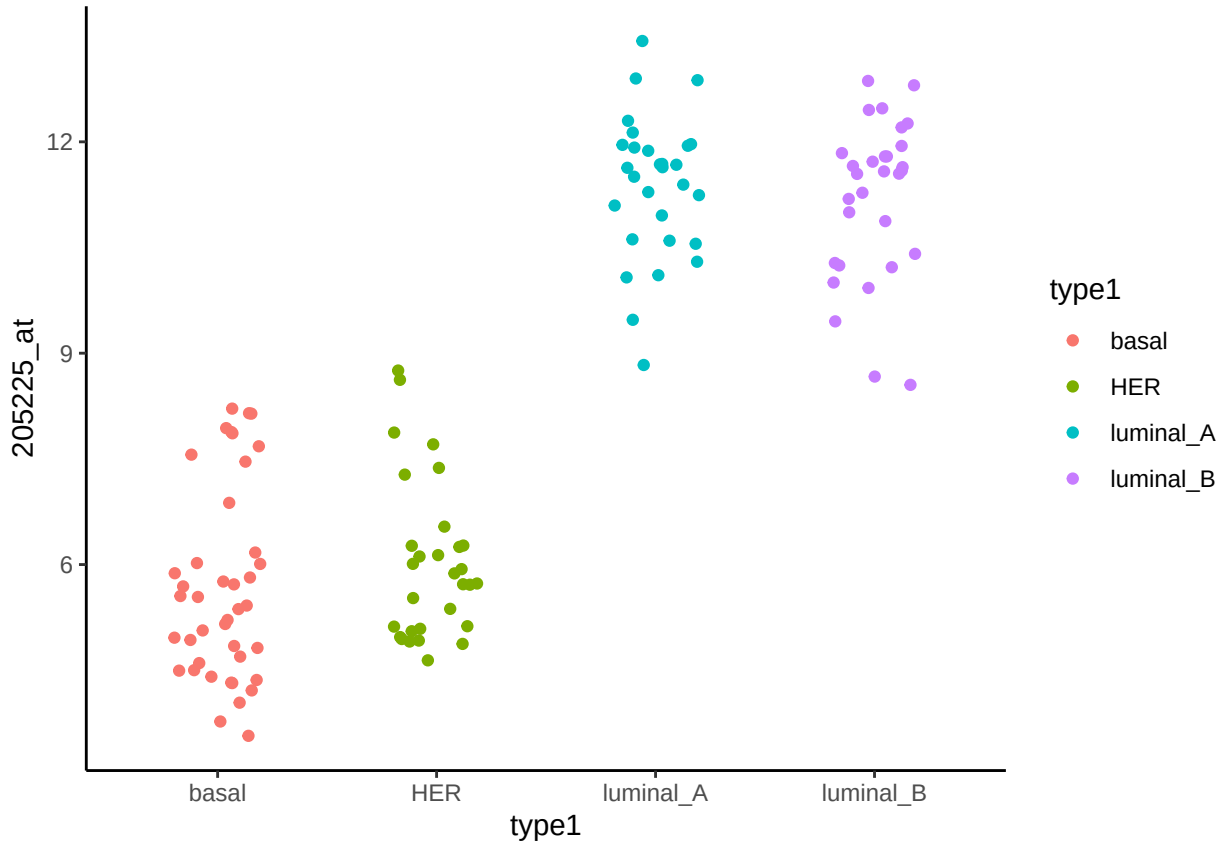
Tabel 4: *Head* dari stats_df

Gene	luminal_A.vs.luminal_B	luminal_A.vs.basal	luminal_A.vs.HER	luminal_B.vs.basal	luminal_B.vs.HER	basal.vs.HER	AveExpr	F	P.Value	adj.P.Val
205225_at	0.1733525	5.6844624	5.3436773	5.511110	5.1703248	-0.3407850	8.301000	235.4761	0	0
229150_at	0.3749415	5.0136527	1.1809771	4.638711	0.8060357	-3.8326756	8.816955	182.9867	0	0
228241_at	-0.0416321	7.2544647	5.6391547	7.296097	5.6807869	-1.6153100	8.065677	179.3168	0	0
214164_x_at	0.2843417	3.9586439	2.8131712	3.674302	2.5288295	-1.1454727	8.813822	153.3942	0	0
215867_x_at	0.3090312	4.2787194	2.9607108	3.969688	2.6516796	-1.3180086	8.732381	145.1785	0	0
1552619_a_at	-1.3781923	-3.3349617	-3.0315660	-1.956769	-1.6533737	0.3033957	6.711249	144.9683	0	0
237086_at	0.5879930	5.4431323	0.8208998	4.855139	0.2329069	-4.6222325	7.549293	138.4942	0	0
213226_at	-1.3892984	-2.8945431	-1.9486918	-1.505245	-0.5593934	0.9458513	7.661218	137.7692	0	0
209173_at	-0.7639931	6.1025074	0.7403146	6.866500	1.5043077	-5.3621928	9.612653	137.2913	0	0
210930_s_at	-2.2042778	-0.1865698	-4.6367087	2.017708	-2.4324309	-4.4501389	6.022395	137.2760	0	0
212956_at	0.6119468	3.4009465	2.1673454	2.789000	1.5553986	-1.2336011	10.011801	135.2714	0	0
209642_at	-1.6892254	-3.3829006	-2.2200931	-1.693675	-0.5308677	1.1628075	6.660908	134.1884	0	0
204962_s_at	-1.5566136	-3.4310538	-2.5483047	-1.874440	-0.9916911	0.8827490	7.494318	133.7304	0	0
204667_at	0.1565362	4.5269491	0.7665398	4.370413	0.6100036	-3.7604093	7.316721	132.0662	0	0
209408_at	-1.2920022	-2.8846903	-1.9904305	-1.592688	-0.6984283	0.8942598	7.887442	131.2005	0	0

Untuk menguji apakah hasil tersebut masuk akal, dapat dibuat plot dari salah satu gen teratas. Akan diekstrak data untuk gen `205225_at`, kemudian akan dibuat dataframe untuk tujuan visualisasi. Berdasarkan hasil di `stats_df`, diperkirakan keempat subtipe kanker tersebut berbeda secara signifikan.

```
top_gene_df <- data.frame(X205225_at = exp1["205225_at", ], type1)
```

```
ggplot(top_gene_df, aes(x = type1, y = X205225_at, color = type1)) +
  labs(y = "205225_at") +
  geom_jitter(width = 0.2, height = 0) + # Make this a jitter plot
  theme_classic() # This makes some aesthetic changes
```



Gambar 4: *Scatter Plot* Ekspresi Gen 205225_at Berdasarkan Subtipe Kanker

Hasil visualisasi tersebut sejalan dengan `stat_df` sebelumnya di mana keempat subtipe kanker berbeda secara signifikan.

Selanjutnya, akan dibuat *volcano plot* dari data tersebut.

```
# Let's extract the contrast p-values for each and transform them with -log10()
contrast_p_vals_df <- -log10(contrasts_fit$p.value) %>%
  # Make this into a dataframe
  as.data.frame() %>%
  # Store genes as their own column
  tibble::rownames_to_column("Gene") %>%
  # Make this into long format
  tidyr::pivot_longer(dplyr::contains("vs"),
                      names_to = "contrast",
                      values_to = "neg_log10_p_val")
# Let's extract the fold changes from stats_df
log_fc_df <- stats_df %>%
  # Only want to keep the `Gene` column as well
  dplyr::select("Gene", dplyr::contains("vs")) %>%
  # Make this a longer format
  tidyr::pivot_longer(dplyr::contains("vs"),
                      names_to = "contrast",
                      values_to = "logFoldChange")
plot_df <- log_fc_df %>%
  dplyr::inner_join(contrast_p_vals_df,
                    by = c("Gene", "contrast"),
                    # This argument will add the given suffixes to the column names
                    # from the respective dataframes, helping us keep track of which columns
                    # hold which types of values
                    suffix = c("_log_fc", "_p_val"))
# Print out what this looks like
knitr::kable(head(plot_df), format = "latex", booktabs = TRUE,
             align = rep("c", 4), caption = "\\textit{Head} dari plot\\_df") %>%
  kableExtra::kable_styling(latex_options = c("scale_down", "HOLD_position")) %>%
```

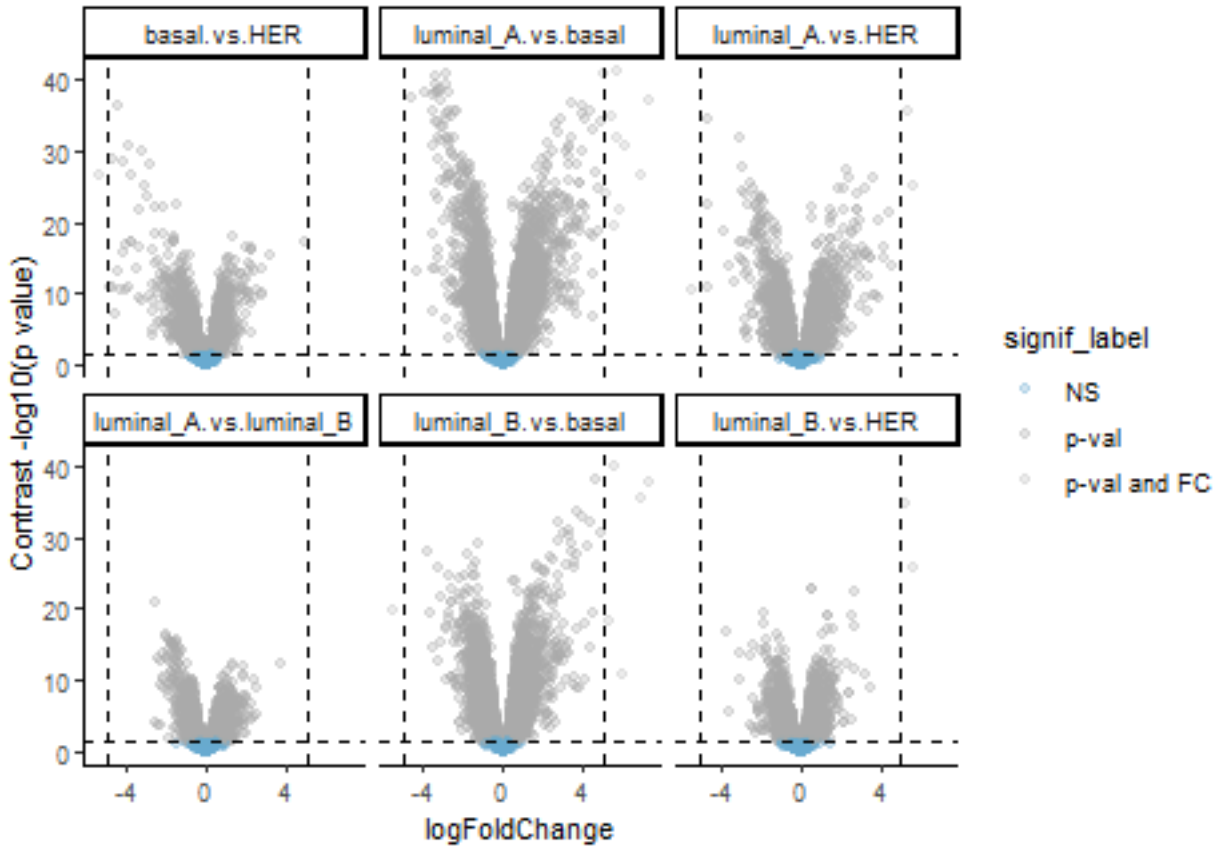
```
  kable_styling(position = "center")
```

Tabel 5: *Head* dari plot_df

Gene	contrast	logFoldChange	neg_log10_p_val
205225_at	luminal_A.vs.luminal_B	0.1733525	0.2448637
205225_at	luminal_A.vs.basal	5.6844624	41.0896040
205225_at	luminal_A.vs.HER	5.3436773	35.5786913
205225_at	luminal_B.vs.basal	5.5111099	40.1859310
205225_at	luminal_B.vs.HER	5.1703248	34.6065840
205225_at	basal.vs.HER	-0.3407850	0.6459322

```
# Convert p-value cutoff to negative log 10 scale
p_val_cutoff <- -log10(0.05)
# Absolute value cutoff for fold changes
abs_fc_cutoff <- 5
plot_df <- plot_df %>%
  dplyr::mutate(signif_label = dplyr::case_when(
    abs(logFoldChange)>abs_fc_cutoff & neg_log10_p_val>p_val_cutoff ~ "p-val and FC",
    abs(logFoldChange)>abs_fc_cutoff ~ "FC",
    neg_log10_p_val>p_val_cutoff ~ "p-val",
    TRUE ~ "NS"))
volcanoes_plot <- ggplot(plot_df,
  aes(
    x = logFoldChange, # Fold change as x value
    y = neg_log10_p_val, # -log10(p-value) for the contrasts
    color = signif_label # Color code by significance cutoffs variable
  )) +
  # Make a scatter plot with points that are 30% opaque using `alpha`
  geom_point(alpha = 0.3) +
  # Draw our `p_val_cutoff` for line here
  geom_hline(yintercept = p_val_cutoff, linetype = "dashed") +
  # Using our `abs_fc_cutoff` for our lines here
  geom_vline(xintercept = c(-abs_fc_cutoff, abs_fc_cutoff), linetype = "dashed") +
```

```
  # Specify color
  scale_colour_manual(values = c("#67a9cf", "darkgray", "gray", "#a1d76a")) +
  # Let's be more specific about what this p-value is in our y axis label
  ylab("Contrast -log10(p value)") +
  # This makes separate plots for each contrast
  facet_wrap(~ contrast) +
  theme(text = element_text(size = 7)) + theme_classic()
# Print out the plot
volcanoes_plot
```



Gambar 5: *Volcano Plot* dari Masing-Masing *Contrast*

Akan dicari top 50 dari gen-gen yang berbeda antara keempat subtipe kanker tersebut menggunakan fungsi `topTable`, sama seperti yang telah dilakukan sebelumnya.

```
topResult1 <- topTable(contrasts_fit, number = 50)
```

Selanjutnya, akan ditampilkan pola ekspresi dari 50 gen tersebut dengan menggunakan *heat map* dan *box plot*.

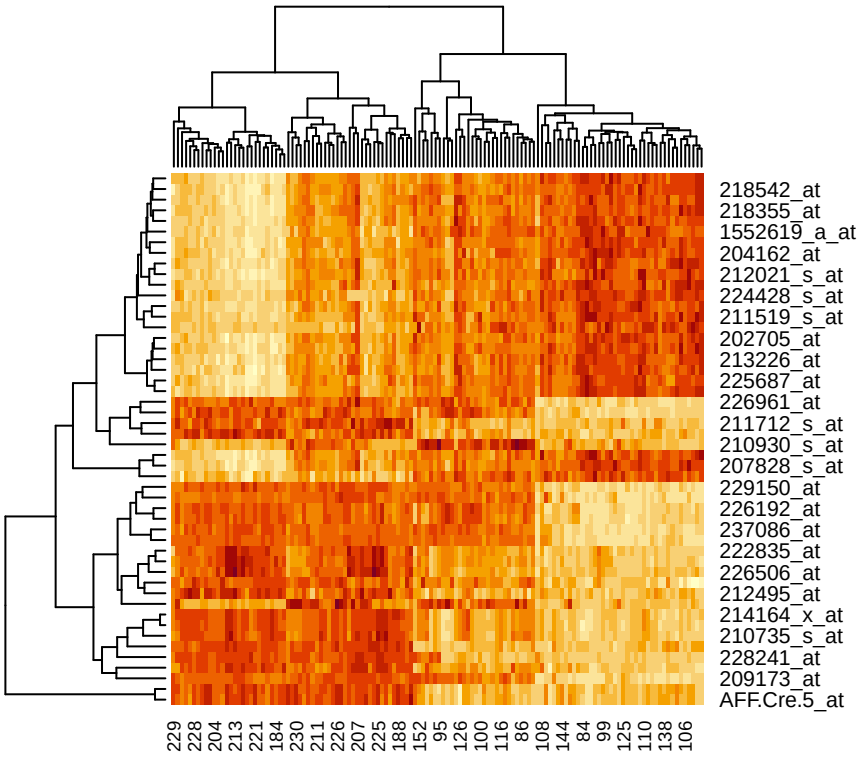
```
# Selected genes
rownames(topResult1)
```

```
 [1] "205225_at"          "229150_at"          "228241_at"
 [4] "214164_x_at"        "215867_x_at"        "1552619_a_at"
 [7] "237086_at"          "213226_at"          "209173_at"
[10] "210930_s_at"        "212956_at"          "209642_at"
[13] "204962_s_at"        "204667_at"          "209408_at"
[16] "226192_at"          "224428_s_at"        "226961_at"
[19] "211519_s_at"        "226197_at"          "218542_at"
[22] "204822_at"          "225687_at"          "203418_at"
[25] "1558448_a_at"       "215304_at"          "207828_s_at"
[28] "221811_at"          "202705_at"          "208433_s_at"
[31] "226506_at"          "211712_s_at"        "212021_s_at"
[34] "205046_at"          "212495_at"          "202580_x_at"
[37] "224753_at"          "218355_at"          "210735_s_at"
[40] "205967_at"          "222835_at"          "222457_s_at"
[43] "211621_at"          "AFF.r2.P1.cre.5_at" "203554_x_at"
[46] "204162_at"          "203764_at"          "203362_s_at"
[49] "220192_x_at"        "AFF.Cre.5_at"
```

```
# Extract selected genes names
selected1 <- rownames(exp1) %in% rownames(topResult1)
# Extract the expression of the selected genes
exptop50_1 <- exp1[selected1, ]
```

```
# Heat map of the top genes

heatmap(exptop50_1)
```



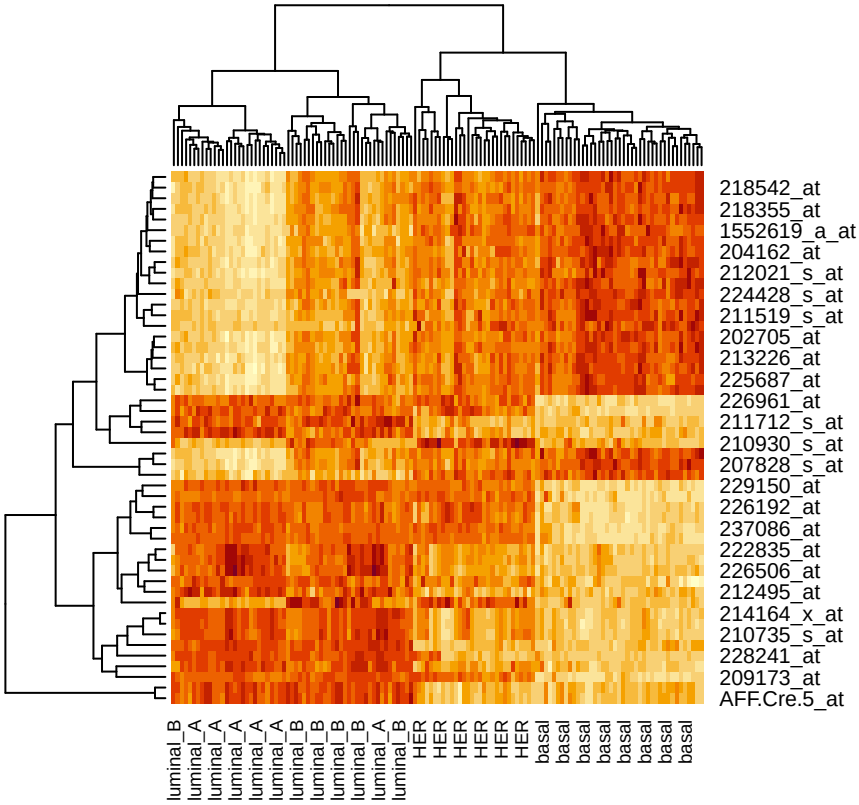
Gambar 6: *Heat Map* dari Ekspresi Gen Dataframe exp1

```
# Heat map dari kategori grup

exptop50_12 <- exptop50_1

colnames(exptop50_12) <- type1
```

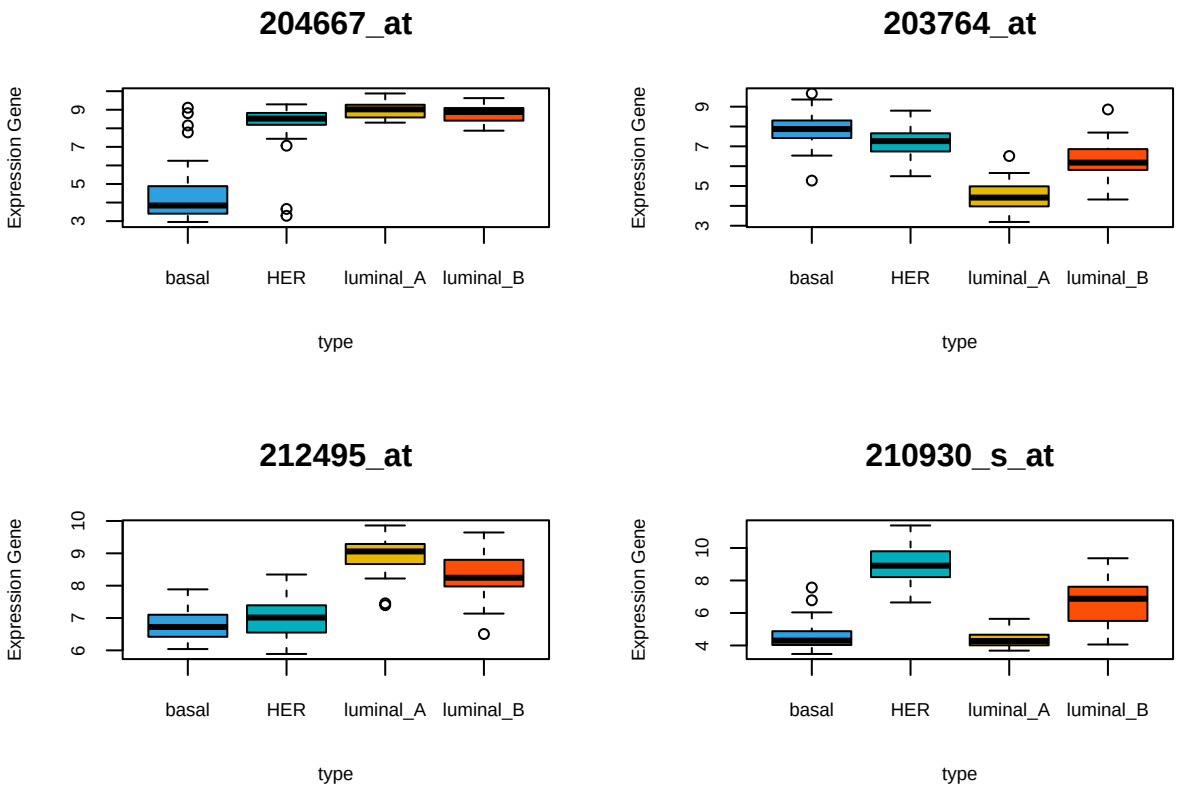
```
heatmap(exptop50_12)
```



Gambar 7: *Heat Map* dari Ekspresi Gen Dataframe exp1 dengan Label Grup

Terlihat bahwa ekpresi gen dari subtipe kanker payudara Luminal A berbeda dengan Basal.

```
# Boxplot for the top 4 genes
par(mfrow = c(2, 2))
for(i in 1:4){
  df_bp1 <- data.frame(y = exptop50_1[i, ], type1)
  df_bp1$type1 <- factor(df_bp1$type1)
  boxplot(df_bp1$y ~ df_bp1$type1, xlab = "type", ylab = "Expression Gene",
          col = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
          cex.axis = 0.7, cex.lab = 0.7, main = rownames(exptop50_1)[i])
}
```



Gambar 8: *Box Plot* dari Top 4 Gen *Differentialy Expressed* Dataframe exp1

```
par(mfrow = c(1, 1))
```

Dari *box plot* di atas, terlihat bahwa setiap gen memiliki karakteristik yang berbeda terhadap keempat subtipe kanker payudara.

Bagian 2

Kedua, akan dicari gen yang berekpresi berbeda antara sel sehat sehat dengan sel kanker. Lima subtipe kanker pada data ini yaitu Luminal A, Luminal B, Basal, HER, dan Cell Line akan digolongkan sebagai sel kanker, sedangkan sel normal akan digolongkan sebagai sel sehat.

```
exp2 <- df
knitr::kable(exp2[1:10, 1:10], format = "latex", booktabs = TRUE,
             align = rep("c", 10), caption = "\\textit{Head} dari exp2") %>%
  kableExtra::kable_styling(latex_options = c("scale_down", "HOLD_position")) %>%
  kable_styling(position = "center")
```

Tabel 6: *Head* dari exp2

	84	85	87	90	91	92	93	94	99	101
213905_x_at	9.722219	8.359922	9.104035	8.487306	10.254484	9.513220	8.381072	9.982283	9.339658	7.250472
223598_at	8.432204	8.717622	7.751498	8.793953	8.459872	9.131923	8.415015	8.880032	8.614927	9.104857
1559582_at	6.563189	6.667046	7.045690	7.434245	7.396452	7.546461	7.041150	7.192393	6.425827	7.039153
204341_at	8.853595	7.157017	8.001522	7.958039	9.279979	6.561909	6.623454	6.740143	6.234407	7.221436
225216_at	7.679953	8.167684	8.895751	8.228383	8.482583	7.560949	8.007203	8.067310	7.473952	7.509730
227262_at	7.768290	8.849295	7.454536	7.567084	8.092288	6.892589	10.017969	8.461030	8.999427	8.993862
201950_x_at	9.589339	9.895621	10.138898	10.210103	10.034316	9.712977	9.354254	10.229080	9.957247	10.023711
226104_at	7.173663	7.105331	7.337900	7.391662	7.918452	6.512620	7.381303	7.573594	7.095866	7.063872
242932_at	6.874082	5.813877	5.625261	5.897511	7.078919	6.633802	6.187999	5.581978	6.189158	5.971807
225671_at	5.704296	6.688602	5.459599	5.900380	6.500822	5.695091	5.686979	6.379044	6.587404	6.716904

Akan dibuat matriks model (*design matrix*) berdasarkan variabel `type2`. Dalam model matriks, akan digunakan `+ 0` dalam model yang menyetel intersep ke 0 sehingga efek `type2` menangkap ekspresi untuk grup tersebut, bukan perbedaan dari grup terhadap *base level*.

```
kanker2 <- c("luminal_A", "luminal_B", "basal", "HER", "cell_line")
# Create the design matrix
type2 <- replace(type_c, which(type_c %in% kanker2), "cancer")
des_mat2 <- model.matrix(~ type2 + 0)
head(des_mat2)
```

```
  type2cancer type2normal
1           1           0
2           1           0
3           1           0
4           1           0
5           1           0
6           1           0
```

Selanjutnya, akan dilakukan *fitting* model *differential expression* pada data. Model limma untuk data ini yaitu

$$\boldsymbol{Y}_j^T = (Y_{1j}, ..., Y_{N_j})$$

$$E(\boldsymbol{Y}_j) = \boldsymbol{X}\boldsymbol{\beta}_j$$

di mana $\boldsymbol{Y}_j$: ekspresi gen, $\boldsymbol{X}$: matriks model (*design matrix*) yang *full rank* (misalnya kondisi grup), dan $\boldsymbol{\beta}_j$: vektor efek dari kolom di matriks $\boldsymbol{X}$, $\boldsymbol{\beta}_j^T = (\beta_{j1}, \beta_{j2})$ dengan β_{jk} merupakan ekspektasi dari level ekspresi dari gen j dalam grup k. Keterangan level: Kanker = 1 dan Normal = 2.

```
# Apply linear model to data
fit2 <- lmFit(exp2, design = des_mat2)
# Apply empirical Bayes to smooth standard errors
fit2 <- eBayes(fit2)
```

Setelah *fitting* model, ingin diselidiki perbedaan di antara kelompok sehat dengan kanker menggunakan *contrast* sebagai berikut.

$$\boldsymbol{C}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$\begin{aligned} \boldsymbol{T}_j &= \boldsymbol{C}_2^T \boldsymbol{\beta}_j \\ &= \begin{bmatrix} -1 & 1 \end{bmatrix} \begin{bmatrix} \beta_{j1} \\ \beta_{j2} \end{bmatrix} \\ &= \begin{bmatrix} \beta_{j2} - \beta_{j1} \end{bmatrix} \end{aligned} \tag{2}$$

Keterangan: $\boldsymbol{T}_j$ merupakan ekspektasi dari perbedaan dalam level ekspresi dari gen j dengan perbandingan Normal versus Kanker.

```
contrast_matrix2 <- makeContrasts(
  "normal-vs-cancer" = type2normal - type2cancer,
  levels = colnames(des_mat2)
)
contrast_matrix2
```

```
             Contrasts
Levels        normal-vs-cancer
  type2cancer               -1
  type2normal                1
```

```
fit2 <- contrasts.fit(fit2, contrast_matrix2)
```

Jumlah gen yang diekspresikan berbeda secara signifikan yaitu sebagai berikut.

```
# Identifying differentially expressed genes
results2 <- decideTests(fit2, p.value = 0.05, adjust.method = "fdr")
summary(results2)
```

```
       normal-vs-cancer
Down               2889
NotSig             4531
Up                 1123
```

Secara keseluruhan, didapatkan jumlah gen DE yang bagus untuk perbandingan antara sel sehat dan sel kanker. Pada *output* tersebut, `Up` mengacu pada gen dengan peningkatan ekspresi pada sel sehat relatif terhadap sel kanker, sedangkan `Down` mengacu pada gen dengan penurunan ekspresi pada sel sehat relatif terhadap sel kanker. Terdapat 1123 gen yang memiliki peningkatan ekspresi dan 2889 gen yang memiliki penurunan ekspresi pada sel sehat relatif terhadap sel kanker.

Gen yang diekspresikan secara berbeda dari analisis limma di atas dapat dianalisis lebih lanjut menggunakan fungsi `topTreat` dari `limma`. Fungsi tersebut memberikan *output* tabel seperti Spreadsheet di

mana setiap baris adalah gen dan kolom berisi informasi seperti $p - value$ (`P.Value`), $p - value$ yang disesuaikan dengan FDR (`adj.P.Val`), t-statistic (`t`) , perubahan kali lipat $\log_2$ (`logFC`), dan lain-lain.

```
# Re-smooth the Bayes
fit2 <- eBayes(fit2)
allDEresults <- topTreat(fit2, coef = "normal-vs-cancer",
                         number = Inf, adjust.method = "fdr") %>% as.data.frame()
allDEresults <- allDEresults %>%
  dplyr::mutate(isSignificant = case_when(
    adj.P.Val<0.05 & abs(logFC)>1 ~ TRUE,
    TRUE ~ FALSE # If conditions in the line above are not met, gene is not DE
    ))
```

Cuplikan beberapa gen yang diekspresikan secara berbeda ditunjukkan di bawah ini.

```
sigDEresults <- allDEresults %>%
  dplyr::filter(isSignificant==TRUE)
knitr::kable(head(sigDEresults, 15), format = "latex", booktabs = TRUE,
             align = rep("c", 7), caption = "\\textit{Head} dari sigDEresults") %>%
  kableExtra::kable_styling(latex_options = c("scale_down", "HOLD_position"))
```

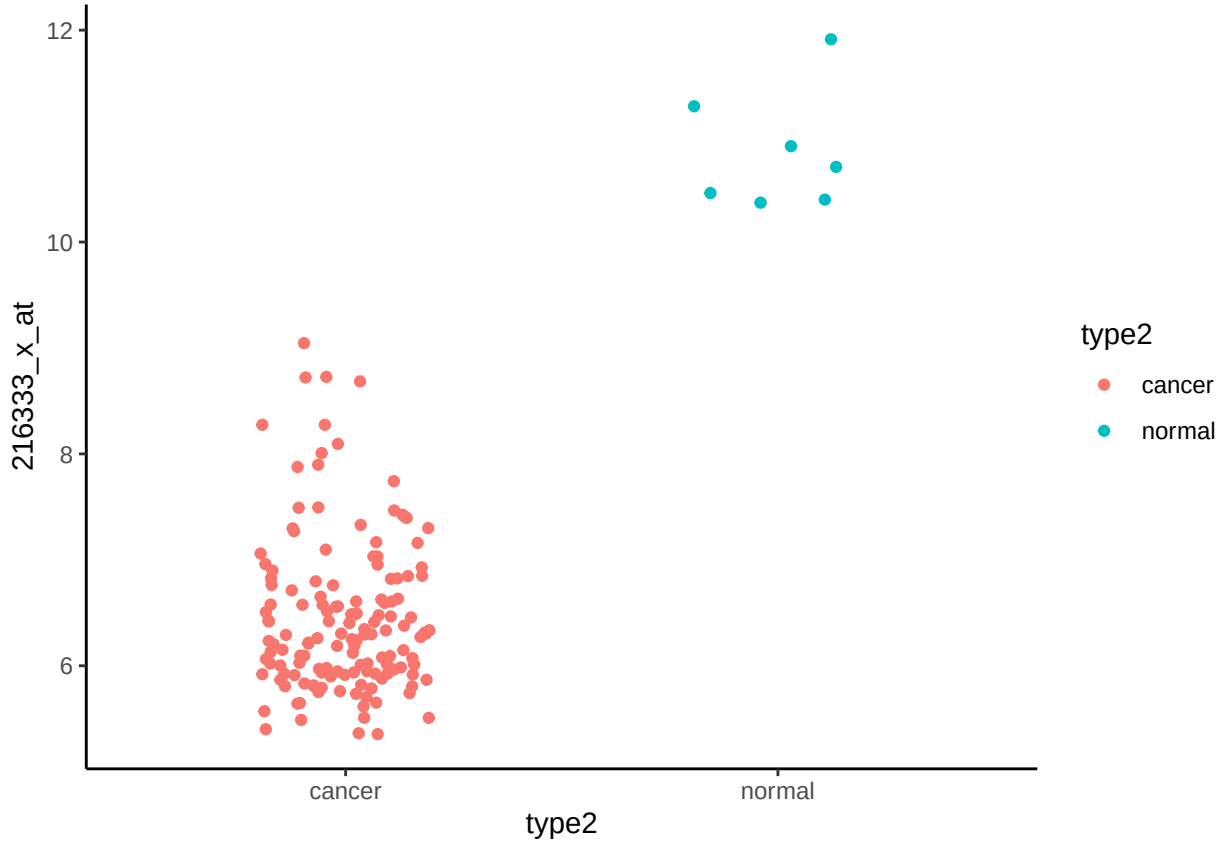
Tabel 7: *Head* dari sigDEresults

	logFC	AveExpr	t	P.Value	adj.P.Val	B	isSignificant
216333_x_at	4.402458	6.664851	15.67958	0	0	65.60897	TRUE
206323_x_at	2.216380	10.040274	13.25372	0	0	51.07976	TRUE
200729_s_at	-2.965775	10.346985	-12.57790	0	0	46.97048	TRUE
201096_s_at	-2.487454	10.060334	-12.52935	0	0	46.67480	TRUE
221928_at	3.658338	6.513415	12.21009	0	0	44.72976	TRUE
1555948_s_at	-2.516477	9.062948	-12.19279	0	0	44.62433	TRUE
234675_x_at	2.155113	9.508980	11.30056	0	0	39.19129	TRUE
208750_s_at	-3.285110	10.124334	-11.24119	0	0	38.83051	TRUE
207791_s_at	-2.573993	9.577408	-11.19151	0	0	38.52869	TRUE
1568954_s_at	-2.644354	7.189088	-10.96815	0	0	37.17323	TRUE
200712_s_at	-2.610244	8.583245	-10.94047	0	0	37.00541	TRUE
49452_at	4.757725	6.307028	10.86942	0	0	36.57494	TRUE
AFF.r2.Ec.bioB.M_at	1.490313	8.742476	10.85747	0	0	36.50260	TRUE
213872_at	-4.708092	10.119181	-10.79230	0	0	36.10808	TRUE
205200_at	4.250123	6.260898	10.79115	0	0	36.10112	TRUE

Untuk menguji apakah hasil tersebut masuk akal, dapat dibuat plot dari salah satu gen teratas. Akan diekstrak data untuk gen `216333_x_at`, kemudian akan dibuat dataframe untuk tujuan visualisasi. Berdasarkan hasil di `sigDEresults`, diperkirakan ekpresi gen dari sel normal dan sel kanker tersebut berbeda secara signifikan.

```
top_gene_df2 <- data.frame(X216333_x_at = exp2["216333_x_at", ], type2)
```

```
ggplot(top_gene_df2, aes(x = type2, y = X216333_x_at, color = type2)) +
  labs(y = "216333_x_at") +
  geom_jitter(width = 0.2, height = 0) + # Make this a jitter plot
  theme_classic() # This makes some aesthetic changes
```

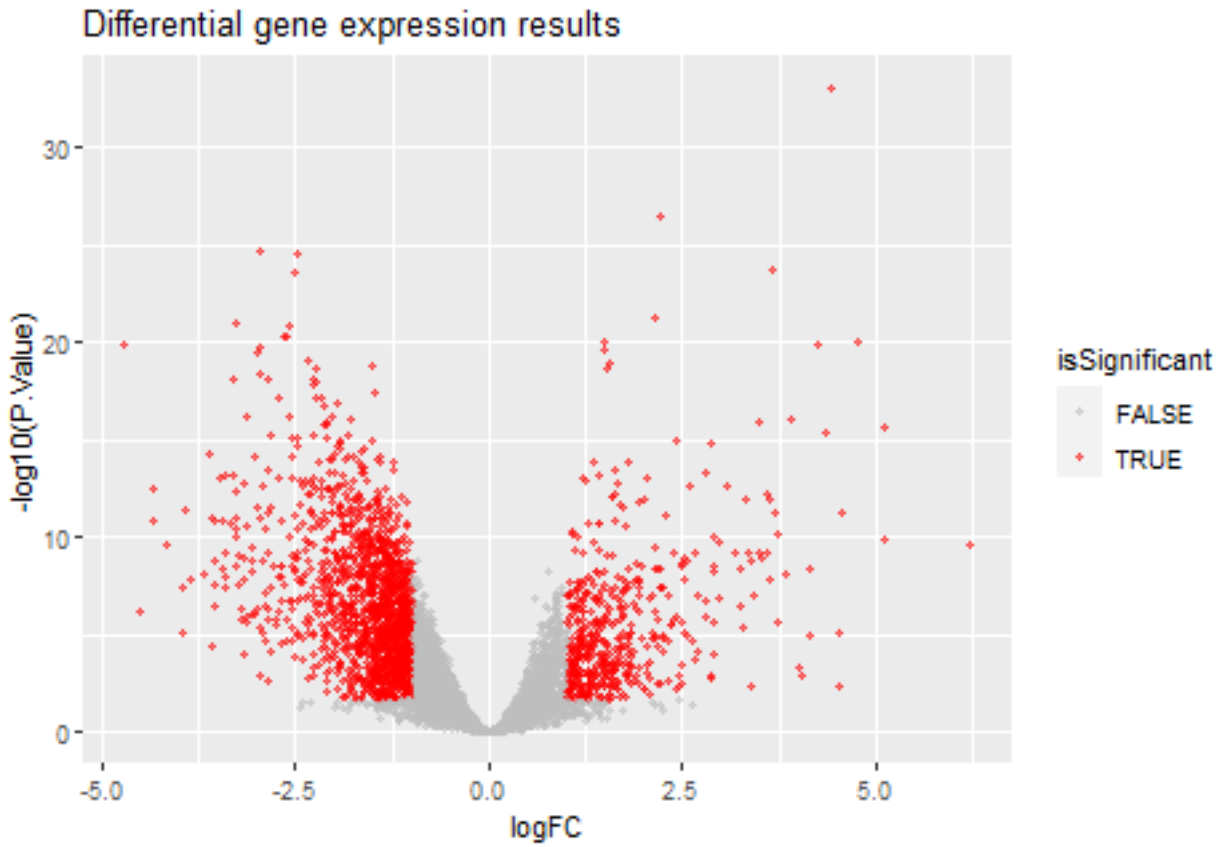


Gambar 9: *Scatter Plot* Ekspresi Gen 216333_x_at antara Sel Sehat versus Sel Kanker

Hasil visualisasi tersebut sejalan dengan `sigDEresults` sebelumnya di mana sel normal dan sel kanker berbeda secara signifikan.

Volcano plot untuk data ini yaitu sebagai berikut.

```
volcano_plot <- allDEresults %>%
  ggplot(aes(x = logFC, y = -log10(P.Value), colour = isSignificant)) +
  geom_point(size = 1, alpha = 0.5) +
  scale_colour_manual(values = c("grey", "red")) +
  ggtitle("Differential gene expression results")
volcano_plot
```



Gambar 10: *Volcano Plot* dari Sel Sehat versus Sel Kanker

Akan dicari gen yang diekspresikan paling berbeda sebanyak 50 gen teratas.

```
topResult2 <- topTable(fit2, coef = "normal-vs-cancer", number = 50)
```

Selanjutnya, akan ditampilkan pola ekspresi dari 50 gen tersebut dengan menggunakan *heat map* dan *box plot*.

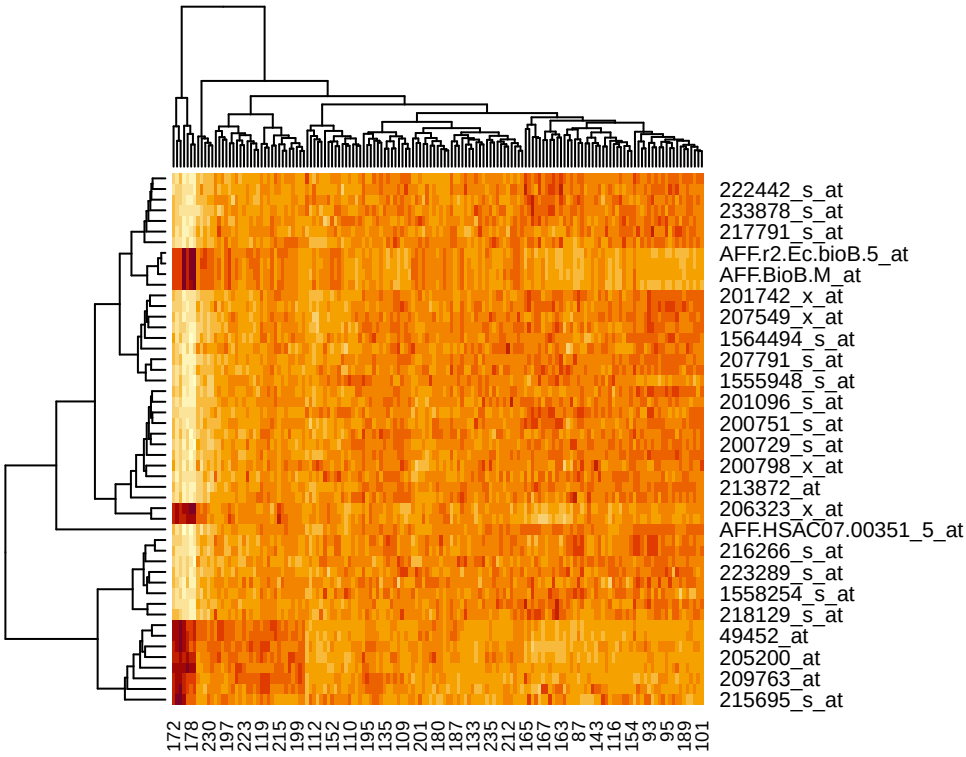
```
# Selected genes
rownames(topResult2)
```

```
 [1] "216333_x_at"           "206323_x_at"           "200729_s_at"
 [4] "201096_s_at"           "221928_at"             "1555948_s_at"
 [7] "234675_x_at"           "208750_s_at"           "207791_s_at"
[10] "1568954_s_at"          "200712_s_at"           "49452_at"
[13] "AFF.r2.Ec.bioB.M_at"   "213872_at"             "205200_at"
[16] "AFF.HSAC07.00351_5_at" "AFF.BioB.M_at"         "217777_s_at"
[19] "202583_s_at"           "AFF.r2.Ec.bioB.5_at"   "200728_at"
[22] "AFF.BioB.5_at"         "200751_s_at"           "216266_s_at"
[25] "222442_s_at"           "203007_x_at"           "1564494_s_at"
[28] "222399_s_at"           "217140_s_at"           "201726_at"
[31] "208622_s_at"           "1558254_s_at"          "201083_s_at"
[34] "217791_s_at"           "223289_s_at"           "207549_x_at"
[37] "1554747_a_at"          "202955_s_at"           "202817_s_at"
[40] "204894_s_at"           "1555278_a_at"          "215695_s_at"
[43] "218129_s_at"           "201523_x_at"           "209763_at"
[46] "213524_s_at"           "201742_x_at"           "200798_x_at"
[49] "226400_at"             "233878_s_at"
```

```
# Extract selected genes names
selected2 <- rownames(exp2) %in% rownames(topResult2)
# Extract the expression of the selected genes
exptop50_2 <- exp2[selected2, ]
```

```
# Heat map of the top genes

heatmap(exptop50_2)
```



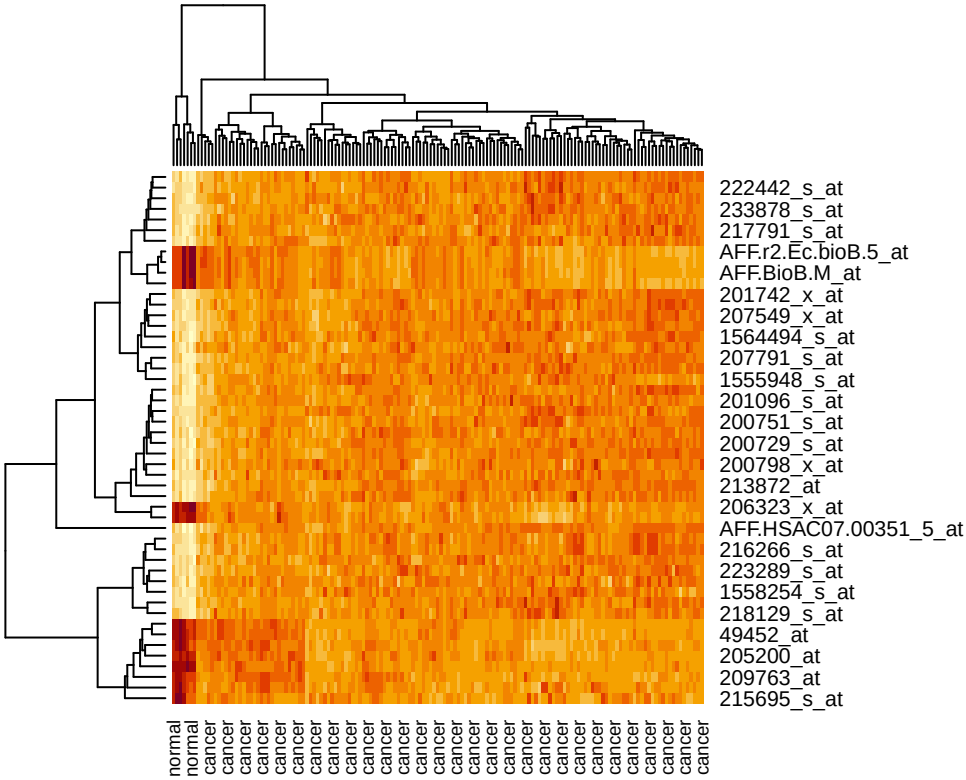
Gambar 11: *Heat Map* dari Ekspresi Gen Dataframe exp2

```
# Heat map dari kategori grup

exptop50_22 <- exptop50_2

colnames(exptop50_22) <- type2
```

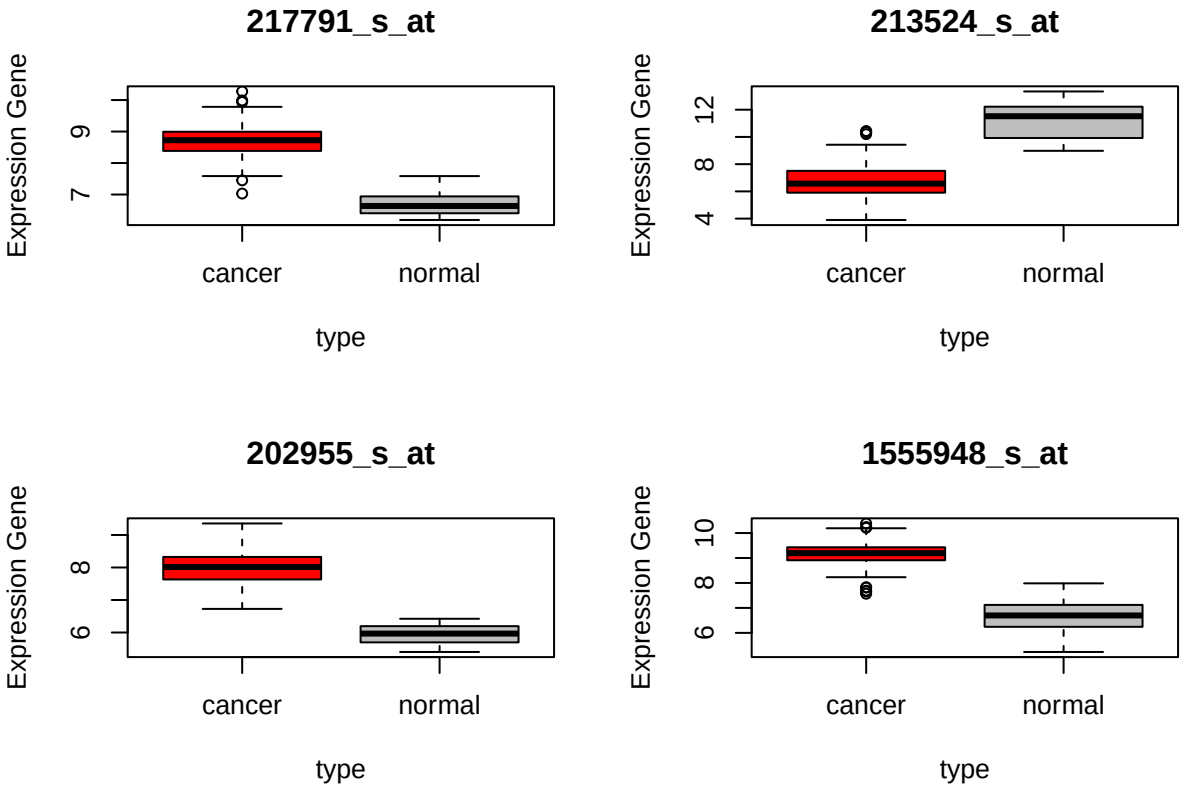
```
heatmap(exptop50_22)
```



Gambar 12: *Heat Map* dari Ekspresi Gen Dataframe exp2 dengan Label Grup

Heat map di atas menggambarkan ekspresi dari tiap gen (baris) dan tiap sampel (kolom). Warna merah menandakan gen tersebut memiliki ekspresi lebih tinggi (*overexpression*). Terlihat pengelompokkan yang relatif "jelas" antara sampel grup dengan sel sehat dan sampel grup dengan sel kanker.

```
# Boxplot for the top 4 genes
par(mfrow = c(2, 2))
for(i in 1:4){
  df_bp2 <- data.frame(y = exptop50_2[i, ], type2)
  df_bp2$type2 <- factor(df_bp2$type2)
  boxplot(df_bp2$y ~ df_bp2$type2, xlab = "type", ylab = "Expression Gene",
          col = c("red", "grey"), main = rownames(exptop50_2)[i])
}
```



Gambar 13: *Box Plot* dari Top 4 Gen *Differentialy Expressed*

```r
par(mfrow = c(1, 1))
```

Terlihat bahwa gen-gen tersebut memiliki ekspresi yang sangat berbeda antara sel sehat dengan sel kanker.

HASIL

Setelah melakukan analisis *differentially expressed genes*, proses selanjutnya adalah mencari nama gen dan deskripsi dari gen yang telah dipilih pada proses sebelumnya. Karena *link* https://sbcb.inf.ufrgs.br/carbm/static/cumida/Genes/Breast/GSE45827/Breast_GSE45827.csv tidak bisa dibuka, detail tentang nama gen dan deskripsinya tidak dapat diketahui. Namun, ditemukan *website* yang mengandung informasi tentang dataset `GSE4582` yaitu pada *link* berikut. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45827. Dalam *website* tersebut, didapatkan informasi bahwa chip menggunakan platform GPL570 dan anotasi yang digunakan yaitu HG-U133_Plus_2. Diasumsikan bahwa data pada Kaggle [9] memiliki informasi yang sama dengan data pada *website* NCBI tersebut.

Bagian 1

Selanjutnya, akan dicari nama gen dan deskripsi dari gen yang telah dipilih pada bagian sebelumnya.

```
# See gene name and description
GeneSelected1 <- select(hgu133plus2.db, rownames(topResult1),
                        c("SYMBOL", "ENTREZID", "GENENAME"))
ids1 <- rownames(topResult1)
GeneSelected1 <- select(hgu133plus2.db, ids1, c("SYMBOL", "ENTREZID", "GENENAME", "GO"))
knitr::kable(head(GeneSelected1), format = "latex", booktabs = TRUE,
             align = rep("c", 7), caption = "\\textit{Head} dari GeneSelected1") %>%
  kableExtra::kable_styling(latex_options = c("scale_down", "HOLD_position"))
```

Tabel 8: *Head* dari GeneSelected1

PROBEID	SYMBOL	ENTREZID	GENENAME	GO	EVIDENCE	ONTOLOGY
205225_at	ESR1	2099	estrogen receptor 1	GO:0000122	IMP	BP
205225_at	ESR1	2099	estrogen receptor 1	GO:0000785	IDA	CC
205225_at	ESR1	2099	estrogen receptor 1	GO:0000785	ISA	CC
205225_at	ESR1	2099	estrogen receptor 1	GO:0000791	IDA	CC
205225_at	ESR1	2099	estrogen receptor 1	GO:0000978	IBA	MF
205225_at	ESR1	2099	estrogen receptor 1	GO:0000978	IDA	MF

Setelah mengetahui nama gen, selanjutnya ingin diketahui model dan fungsi dari gen tersebut dengan menggunakan Gene Ontology. Terdapat tiga aspek fungsi gen yaitu

1. *molecular function* (aktivitas molekuler dari hasil gen);

2. *cellular component*, di mana hasil gen tersebut aktif;
3. *biological process*, informasi dimana proses dan *pathways* biologi dari gen tersebut ikut serta.

Untuk melakukan hal tersebut, diperlukan *package* `GO.db` yang akan menghubungkan nama-nama gen dengan Gene Ontology.

```
# Gene ontology for the top genes
GOselected1 <- select(GO.db, GeneSelected1$GO, c("TERM", "GOID"))
# Combine the result
finalres1 <- cbind(GeneSelected1, GOselected1)
knitr::kable(head(finalres1), format = "latex", booktabs = TRUE,
             align = rep("c", 9), caption = "\\textit{Head} dari finalres1") %>%
  kableExtra::kable_styling(latex_options = c("scale_down", "HOLD_position"))
```

Tabel 9: *Head* dari finalres1

PROBEID	SYMBOL	ENTREZID	GENENAME	GO	EVIDENCE	ONTOLOGY	GOID	TERM
205225_at	ESR1	2099	estrogen receptor 1	GO:0000122	IMP	BP	GO:0000122	negative regulation of transcription by RNA polymerase II
205225_at	ESR1	2099	estrogen receptor 1	GO:0000785	IDA	CC	GO:0000785	chromatin
205225_at	ESR1	2099	estrogen receptor 1	GO:0000785	ISA	CC	GO:0000785	chromatin
205225_at	ESR1	2099	estrogen receptor 1	GO:0000791	IDA	CC	GO:0000791	euchromatin
205225_at	ESR1	2099	estrogen receptor 1	GO:0000978	IBA	MF	GO:0000978	RNA polymerase II cis-regulatory region sequence-specific DNA binding
205225_at	ESR1	2099	estrogen receptor 1	GO:0000978	IDA	MF	GO:0000978	RNA polymerase II cis-regulatory region sequence-specific DNA binding

```
write.csv2(finalres1, file = "D:/Materi Kuliah UI/Sains Data Genom/Tugas Sains Data Genom/gen_de1
```

Untuk versi lengkapnya, dapat dilihat pada *link* berikut. gen_de1

Dari top 50 gen tersebut, akan dipilih dua gen untuk dicari deskripsi dan kaitannya dengan kanker payudara.

1. *Estrogen Receptor* 1 (ESR1)

Gen ini mengkodekan reseptor estrogen dan faktor transkripsi yang diaktifkan ligan. Protein kanonik mengandung domain *N-terminal ligand-independent transactivation*, domain *central DNA binding*, domain *hinge* dan domain *C-terminal ligand-dependent transactivation*. Protein terlokalisasi di nukleus, di mana protein tersebut dapat membentuk homodimer atau heterodimer dengan reseptor estrogen 2. Protein yang dikodekan oleh gen ini mengatur transkripsi banyak gen yang diinduksi estrogen yang berperan dalam pertumbuhan, metabolisme, perkembangan seksual, kehamilan, dan fungsi reproduksi lainnya dan diekspresikan di banyak jaringan nonreproduksi. Reseptor yang dikodekan oleh gen ini memiliki peran

penting dalam kanker payudara, kanker endometrium, dan osteoporosis. Gen ini dilaporkan memiliki lusinan varian transkrip karena penggunaan promotor alternatif dan penyambungan alternatif, namun sifat keseluruhan dari banyak varian ini masih belum pasti [13].

Laporan patologi yang dilakukan akan mencakup hasil uji reseptor hormon, tes yang memberi tahu kita apakah sel kanker payudara memiliki reseptor untuk hormon estrogen dan progesteron. Reseptor hormon adalah protein, yang ditemukan di dalam dan pada sel payudara, yang menangkap sinyal dari hormon yang memerintahkan sel untuk tumbuh. Kanker payudara bersifat reseptor estrogen positif jika memiliki reseptor untuk estrogen. Hal ini menunjukkan bahwa sel-sel kanker, seperti sel-sel payudara normal, mungkin menerima sinyal dari estrogen yang memberitahu sel-sel tersebut untuk tumbuh. Kanker merupakan reseptor progesteron positif jika memiliki reseptor progesteron. Hal tersebut berarti bahwa sel kanker mungkin menerima sinyal dari progesteron yang memerintahkan mereka untuk tumbuh. Sekitar dua dari setiap tiga kanker payudara dites positif terhadap reseptor hormon. Pengujian reseptor hormon penting karena hasilnya dapat membantu penderita kanker dan dokter dalam memutuskan apakah kanker kemungkinan akan merespon obat-obatan terapi hormonal.

Penting untuk diketahui bahwa beberapa kanker payudara dengan reseptor hormon positif dapat kehilangan reseptornya seiring berjalannya waktu. Hal sebaliknya juga terjadi, yaitu kanker dengan reseptor hormon negatif dapat mengembangkan reseptor hormon. Jika kanker payudara muncul kembali setelah pengobatan, ada baiknya penderita kanker bertanya kepada dokter tentang biopsi lain untuk menguji reseptor hormon pada kanker. Jika sel kanker tidak lagi memiliki reseptor, terapi hormonal tidak mungkin membantu mengobati kanker. Jika sel telah mengembangkan reseptor hormon, terapi hormonal mungkin membantu dalam proses penyembuhan [4].

2. *Carbonic Anhydrase* 12 (CA12)

Carbonic Anhydrase (CA) adalah keluarga besar metaloenzim seng yang mengkatalisis hidrasi karbon dioksida yang dapat dibalik. Mereka berpartisipasi dalam berbagai proses biologis, termasuk respirasi, kalsifikasi, keseimbangan asam-basa, resorpsi tulang, dan pembentukan aqueous humor, cairan serebrospinal, air liur, dan asam lambung. Produk gen ini adalah protein membran tipe I yang banyak diekspresikan di jaringan normal, seperti ginjal, usus besar, dan pankreas, dan ditemukan diekspresikan berlebih pada 10% karsinoma ginjal sel jernih. Tiga varian transkrip yang mengkode isoform berbeda telah diidentifikasi untuk gen ini [12].

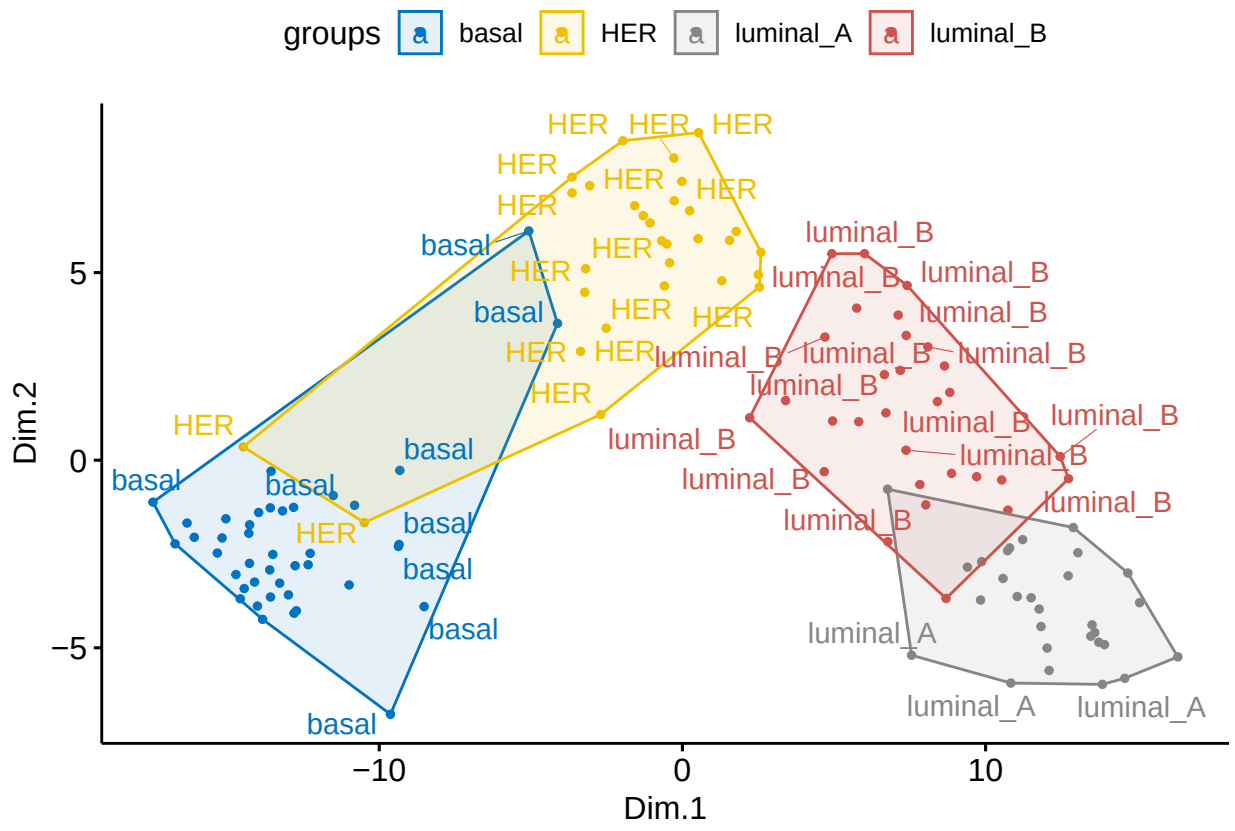
Penelitian intensif terkonsentrasi pada peran CA IX pada tumor hipoksia, dengan CA XII yang kurang diteliti namun juga dianggap sebagai pemicu tumor hipoksia. CA XII pertama kali diidentifikasi pada

kanker sel ginjal sebagai bagian dari penelitian yang bertujuan untuk mengidentifikasi antigen tumor baru melalui penerapan SEREX, identifikasi serologis antigen dengan kloning ekspresi rekombinan. Gen yang mengkode CA XII, CA12, merupakan target dari *hypoxia-inducible factor 1* (HIF-1), dan ciri umum dari banyak karsinoma ginjal adalah mutasi pada gen penekan tumor *von Hippel-Lindau* (VHL), sebuah pengatur HIF-1. Berbeda dengan CA IX, ekspresi CA XII diinduksi oleh estrogen dan CA XII sangat diekspresikan pada kanker payudara reseptor estrogen positif. Penelitian lain juga menunjukkan perbedaan antara ekspresi dan fungsi CA IX dan CA XII pada berbagai jenis kanker, termasuk keterlibatan CA XII dalam mendorong invasi dan migrasi sel tumor. Peran CAXII dalam *tumorigenesis* mungkin bergantung pada subtipe sel kanker, *hypoxic microenvironment*, perkembangan penyakit, dan fenotip resisten obat yang baru-baru ini dijelaskan [18].

Selanjutnya, akan dibuat *muldimensional scaling plot* dari top 50 gen tersebut untuk mengetahui persebaran subtipe kanker berdasarkan dimensi yang lebih rendah yaitu dimensi 2.

```
mds1 <- t(exptop50_12) %>% dist() %>% cmdscale() %>% as_tibble
colnames(mds1) <- c("Dim.1", "Dim.2")
mds1 <- mds1 %>% mutate(groups = colnames(exptop50_12))
```

```
# Plot MDS
ggscatter(mds1, x = "Dim.1", y = "Dim.2", label = colnames(exptop50_12),
          color = "groups", palette = "jco", size = 1, ellipse = TRUE,
          ellipse.type = "convex", repel = TRUE)
```



Gambar 14: *Multidimensional Scaling Plot* dari exptop50_12

Makin dekat suatu label, makin sama sampelnya. Meskipun terdapat beberapa label yang tumpang tindih, secara keseluruhan terlihat bahwa subtipe kanker mengelompok dengan jenis yang sama.

Bagian 2

Pada bagian ini, akan dicari nama gen dan deskripsi gen pada top 50 gen yang berbeda dari analisis limma sebelumnya.

```
# See gene name and description
GeneSelected2 <- select(hgu133plus2.db, rownames(topResult2),
                        c("SYMBOL", "ENTREZID", "GENENAME"))
```

```
ids2 <- rownames(topResult2)
GeneSelected2 <- select(hgu133plus2.db, ids2, c("SYMBOL", "ENTREZID", "GENENAME", "GO"))
knitr::kable(head(GeneSelected2), format = "latex", booktabs = TRUE,
             align = rep("c", 7), caption = "\\textit{Head} dari GeneSelected2") %>%
  kableExtra::kable_styling(latex_options = c("scale_down", "HOLD_position"))
```

Tabel 10: *Head* dari GeneSelected2

PROBEID	SYMBOL	ENTREZID	GENENAME	GO	EVIDENCE	ONTOLOGY
216333_x_at	TNXB	7148	tenascin XB	GO:0005178	ISS	MF
216333_x_at	TNXB	7148	tenascin XB	GO:0005201	ISS	MF
216333_x_at	TNXB	7148	tenascin XB	GO:0005201	RCA	MF
216333_x_at	TNXB	7148	tenascin XB	GO:0005515	IPI	MF
216333_x_at	TNXB	7148	tenascin XB	GO:0005518	IEA	MF
216333_x_at	TNXB	7148	tenascin XB	GO:0005576	HDA	CC

Selanjutnya, akan dilakukan *gene ontology*.

```
# Gene ontology for the top genes
GOselected2 <- select(GO.db, GeneSelected2$GO, c("TERM", "GOID"))
# Combine the result
finalres2 <- cbind(GeneSelected2, GOselected2)
knitr::kable(head(finalres2), format = "latex", booktabs = TRUE,
             align = rep("c", 9), caption = "\\textit{Head} dari finalres2") %>%
  kableExtra::kable_styling(latex_options = c("scale_down", "HOLD_position"))
```

Tabel 11: *Head* dari finalres2

PROBEID	SYMBOL	ENTREZID	GENENAME	GO	EVIDENCE	ONTOLOGY	GOID	TERM
216333_x_at	TNXB	7148	tenascin XB	GO:0005178	ISS	MF	GO:0005178	integrin binding
216333_x_at	TNXB	7148	tenascin XB	GO:0005201	ISS	MF	GO:0005201	extracellular matrix structural constituent
216333_x_at	TNXB	7148	tenascin XB	GO:0005201	RCA	MF	GO:0005201	extracellular matrix structural constituent
216333_x_at	TNXB	7148	tenascin XB	GO:0005515	IPI	MF	GO:0005515	protein binding
216333_x_at	TNXB	7148	tenascin XB	GO:0005518	IEA	MF	GO:0005518	collagen binding
216333_x_at	TNXB	7148	tenascin XB	GO:0005576	HDA	CC	GO:0005576	extracellular region

```
write.csv2(finalres2, file = "D:/Materi Kuliah UI/Sains Data Genom/Tugas Sains Data Genom/gen_de2
```

Untuk versi lengkapnya, dapat dilihat pada *link* berikut. gen_de2

Dari top 50 gen tersebut, akan dipilih dua gen untuk dicari deskripsi dan kaitannya dengan kanker payudara.

1. *Tenascin* XB (TNXB)

Gen ini mengkodekan anggota keluarga *tenascin* dari *extracellular matrix glycoproteins*. *Tenascins* memiliki efek antiperekat, berbeda dengan fibronektin yang bersifat perekat. Protein ini diperkirakan berfungsi dalam pematangan matriks selama penyembuhan luka, dan kekurangannya telah dikaitkan dengan jaringan ikat *disorder Ehlers-Danlos syndrome*. Gen ini melokalisasi ke wilayah kelas III kompleks histokompatibilitas utama (MHC) pada kromosom 6. Gen ini adalah salah satu dari empat gen dalam *cluster* ini yang telah diduplikasi. Salinan duplikat gen ini tidak lengkap dan merupakan pseudogen yang ditranskripsi tetapi tidak mengkode protein. Struktur gen ini tidak biasa karena tumpang tindih dengan gen CREBL1 dan CYP21A2 pada ujung 5' dan 3'. Berbagai varian transkrip yang mengkode isoform berbeda telah ditemukan untuk gen ini [15]. Penelitian pada [8] mendapatkan bahwa *tenascin* sangat diekspresikan pada *breast invasive carcinoma* (TCGA-BRCA).

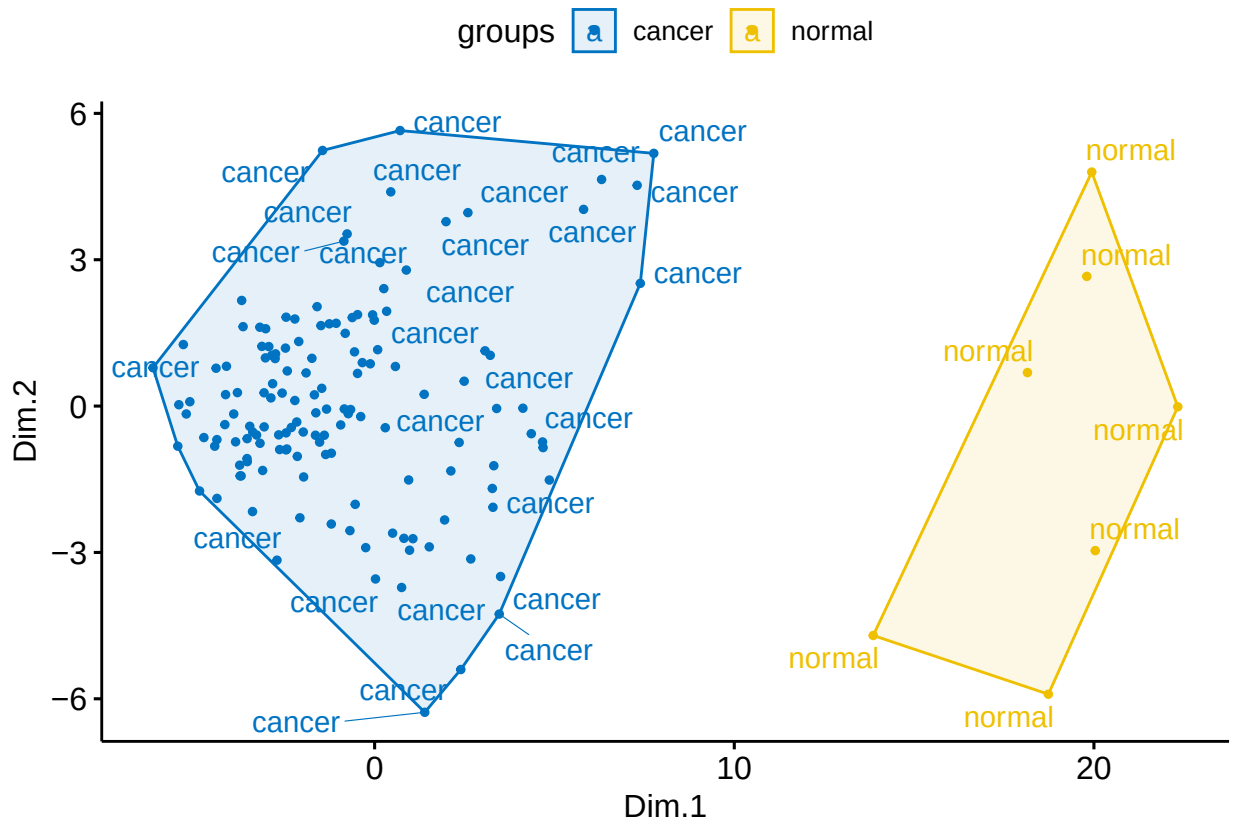
2. *Oligophrenin* 1 (OPHN1)

Gen ini mengkode protein *Rho-GTPase-activating* yang mendorong hidrolisis GTP anggota subfamili Rho. Protein Rho adalah mediator penting transduksi sinyal intraseluler, yang mempengaruhi migrasi sel dan morfogenesis sel. Mutasi pada gen ini bertanggung jawab atas kecacatan *OPHN1-related X-linked cognitive* dengan *cerebellar hypoplasia* dan *distinctive facial dysmorhphism* [14]. Berdasarkan [7], gen ini diekspresikan secara berlebihan (*overexpression*) dalam sel *peripheral blood mononuclear*, payudara, plasenta, dan otak janin.

Selanjutnya, akan dibuat *muldimensional scaling plot* dari top 50 gen tersebut untuk mengetahui persebaran sel kanker dengan sel normal berdasarkan dimensi yang lebih rendah yaitu dimensi 2.

```
mds2 <- t(exptop50_22) %>% dist() %>% cmdscale() %>% as_tibble
colnames(mds2) <- c("Dim.1", "Dim.2")
mds2 <- mds2 %>% mutate(groups = colnames(exptop50_22))
```

```
# Plot MDS
ggscatter(mds2, x = "Dim.1", y = "Dim.2", label = colnames(exptop50_22),
          color = "groups", palette = "jco", size = 1, ellipse = TRUE,
          ellipse.type = "convex", repel = TRUE)
```



Gambar 15: *Multidimensional Scaling Plot* dari exptop50_22

Terlihat bahwa sel sehat/normal terpisah dengan sel kanker yang menandakan bahwa top 50 gen tersebut memang benar berbeda secara signifikan.

KESIMPULAN

Bagian 1

Analisis *differentially expressed genes* telah dilakukan dengan menggunakan metode limma. Dengan menggunakan *contrast* pada persamaan (1), didapatkan gen-gen yang signifikan berbeda seperti terlihat pada tabel berikut.

Tabel 12: *Differentially Expressed Genes* dari Dataframe exp1

	Luminal A vs Luminal B	Luminal A vs Basal	Luminal A vs HER	Luminal B vs Basal	Luminal B vs HER	Basal vs HER
Down	1313	2928	2179	2656	1532	1680
Not Significance	6064	2821	4311	3255	5509	5229
Up	1166	2794	2053	2632	1502	1634

Dari tabel tersebut, didapatkan informasi bahwa subtipe kanker yang paling banyak berbeda secara signifikan yaitu luminal A dan basal, dengan jumlah yang signifikan berbeda sebanyak 2928 + 2794 = 5722.

Berikut ini merupakan top 50 gen yang berbeda secara signifikan berdasarkan analisis limma sebelumnya.

```
rownames(topResult1)
```

```
 [1] "205225_at"          "229150_at"          "228241_at"
 [4] "214164_x_at"        "215867_x_at"        "1552619_a_at"
 [7] "237086_at"          "213226_at"          "209173_at"
[10] "210930_s_at"        "212956_at"          "209642_at"
[13] "204962_s_at"        "204667_at"          "209408_at"
[16] "226192_at"          "224428_s_at"        "226961_at"
[19] "211519_s_at"        "226197_at"          "218542_at"
[22] "204822_at"          "225687_at"          "203418_at"
[25] "1558448_a_at"       "215304_at"          "207828_s_at"
[28] "221811_at"          "202705_at"          "208433_s_at"
[31] "226506_at"          "211712_s_at"        "212021_s_at"
[34] "205046_at"          "212495_at"          "202580_x_at"
[37] "224753_at"          "218355_at"          "210735_s_at"
[40] "205967_at"          "222835_at"          "222457_s_at"
[43] "211621_at"          "AFF.r2.P1.cre.5_at" "203554_x_at"
[46] "204162_at"          "203764_at"          "203362_s_at"
```

```
[49] "220192_x_at"        "AFF.Cre.5_at"
```

Gen tersebut mempunyai nama sebagai berikut.

```
unique(finalres1$GENENAME)
```

```
 [1] "estrogen receptor 1"
 [2] NA
 [3] "anterior gradient 3, protein disulphide isomerase family member"
 [4] "carbonic anhydrase 12"
 [5] "anillin, actin binding protein"
 [6] "forkhead box A1"
 [7] "cyclin A2"
 [8] "anterior gradient 2, protein disulphide isomerase family member"
 [9] "erb-b2 receptor tyrosine kinase 2"
[10] "TBC1 domain family member 9"
[11] "BUB1 mitotic checkpoint serine/threonine kinase"
[12] "centromere protein A"
[13] "kinesin family member 2C"
[14] "androgen receptor"
[15] "cell division cycle associated 7"
[16] "proline rich 15"
[17] "centrosomal protein 55"
[18] "TTK protein kinase"
[19] "family with sequence similarity 83 member D"
[20] "centromere protein F"
[21] "post-GPI attachment to proteins phospholipase 3"
[22] "cyclin B2"
[23] "LDL receptor related protein 8"
[24] "thrombospondin type 1 domain containing 4"
[25] "annexin A9"
[26] "marker of proliferation Ki-67"
[27] "centromere protein E"
[28] "lysine demethylase 4B"
```

```
[29] "forkhead box M1"
[30] "cell division cycle associated 5"
[31] "kinesin family member 4A"
[32] "H4 clustered histone 3"
[33] "LIM domain and actin binding 1"
[34] "PTTG1 regulator of sister chromatid separation, securin"
[35] "NDC80 kinetochore complex component"
[36] "DLG associated protein 5"
[37] "mitotic arrest deficient 2 like 1"
[38] "SAM pointed domain containing ETS transcription factor"
```

Gen-gen tersebut signifikan berbeda pada analisis limma sebelumnya. Gen tersebut direkomendasikan sebagai gen yang berekpresi berbeda di antara empat subtipe kancer payudara yaitu Luminal A, Luminal B, Basal, dan HER.

Bagian 2

Analisis *differentially expressed genes* telah dilakukan dengan menggunakan metode limma. Dengan menggunakan *contrast* pada persamaan (2), didapatkan gen-gen yang signifikan berbeda seperti terlihat pada tabel berikut.

Tabel 13: *Differentially Expressed Genes* dari Dataframe exp2

	Normal vs Kanker
Down	2889
Not Significance	4531
Up	1123

Dari tabel tersebut, didapatkan informasi bahwa terdapat 1123 gen yang memiliki peningkatan ekspresi dan 2889 gen yang memiliki penurunan ekspresi pada sel sehat relatif terhadap sel kanker.

Berikut ini merupakan top 50 gen yang berbeda secara signifikan berdasarkan analisis limma sebelumnya.

```
rownames(topResult2)
```

```
 [1] "216333_x_at"          "206323_x_at"          "200729_s_at"
 [4] "201096_s_at"          "221928_at"            "1555948_s_at"
 [7] "234675_x_at"          "208750_s_at"          "207791_s_at"
```

```
[10] "1568954_s_at"          "200712_s_at"           "49452_at"
[13] "AFF.r2.Ec.bioB.M_at"   "213872_at"             "205200_at"
[16] "AFF.HSAC07.00351_5_at" "AFF.BioB.M_at"         "217777_s_at"
[19] "202583_s_at"           "AFF.r2.Ec.bioB.5_at"   "200728_at"
[22] "AFF.BioB.5_at"         "200751_s_at"           "216266_s_at"
[25] "222442_s_at"           "203007_x_at"           "1564494_s_at"
[28] "222399_s_at"           "217140_s_at"           "201726_at"
[31] "208622_s_at"           "1558254_s_at"          "201083_s_at"
[34] "217791_s_at"           "223289_s_at"           "207549_x_at"
[37] "1554747_a_at"          "202955_s_at"           "202817_s_at"
[40] "204894_s_at"           "1555278_a_at"          "215695_s_at"
[43] "218129_s_at"           "201523_x_at"           "209763_at"
[46] "213524_s_at"           "201742_x_at"           "200798_x_at"
[49] "226400_at"             "233878_s_at"
```

Gen tersebut mempunyai nama sebagai berikut.

```
unique(finalres2$GENENAME)
```

```
 [1] "tenascin XB"
 [2] "tenascin XA (pseudogene)"
 [3] "oligophrenin 1"
 [4] "actin related protein 2"
 [5] "ADP ribosylation factor 4"
 [6] "acetyl-CoA carboxylase beta"
 [7] "family with sequence similarity 120A"
 [8] NA
 [9] "ADP ribosylation factor 1"
[10] "RAB1A, member RAS oncogene family"
[11] "HUWE1 associated protein modifying stress responses"
[12] "microtubule associated protein RP/EB family member 1"
[13] "C-type lectin domain family 3 member B"
[14] "exosome component 7"
[15] "3-hydroxyacyl-CoA dehydratase 3"
```

```
[16] "RAN binding protein 9"
[17] "heterogeneous nuclear ribonucleoprotein C"
[18] "ADP ribosylation factor guanine nucleotide exchange factor 1"
[19] "ADP ribosylation factor like GTPase 8B"
[20] "lysophospholipase 1"
[21] "prolyl 4-hydroxylase subunit beta"
[22] "transmembrane 9 superfamily member 3"
[23] "voltage dependent anion channel 1"
[24] "ELAV like RNA binding protein 1"
[25] "ezrin"
[26] "SRSF protein kinase 2"
[27] "BCL2 associated transcription factor 1"
[28] "aldehyde dehydrogenase 18 family member A1"
[29] "ubiquitin specific peptidase 38"
[30] "CD46 molecule"
[31] "septin 2"
[32] "SS18 subunit of BAF chromatin remodeling complex"
[33] "amine oxidase copper containing 3"
[34] "cytoskeleton associated protein 5"
[35] "glycogenin 2"
[36] "nuclear transcription factor Y subunit beta"
[37] "ubiquitin conjugating enzyme E2 N"
[38] "chordin like 1"
[39] "G0/G1 switch 2"
[40] "serine and arginine rich splicing factor 1"
[41] "MCL1 apoptosis regulator, BCL2 family member"
[42] "cell division cycle 42"
[43] "5'-3' exoribonuclease 2"
```

Gen-gen tersebut signifikan berbeda pada analisis limma sebelumnya. Gen tersebut direkomendasikan sebagai gen yang berekpresi berbeda di antara sel sehat dan sel kanker.

REFERENSI

[1] alexslemonade.github.io. (2020). *Differential Expression - Several Groups - Microarray*. https://alexslemonade.github.io/refinebio-examples/02-microarray/differential-expression_microarray_02_several-groups.html

[2] Andrew, M. G., dkk. (2015). *Effects of Age on the Detection and Management of Breast Cancer*. Cancers 7, 908–929. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4491690/

[3] biocellgen-public.svi.edu.au. (2021). *Differential Gene Expression Analysis*. https://biocellgen-public.svi.edu.au/sahmri-bulk-rnaseq/de.html#Gene_Ontology_(GO)_Enrichment_Analysis

[4] breastcancer.org. (n.d.). *Breast Cancer Hormone Receptor Status*. https://www.breastcancer.org/pathology-report/hormone-receptor-status#section-hormone-receptor s-estrogen-and-progesterone

[5] breastcancer.org. (n.d.). *Molecular Subtypes of Breast Cancer*. https://www.breastcancer.org/types/molecular-subtypes

[6] cdc.gov. (2023). *What is Breast Cancer?*. https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm

[7] genecards.org. (2023). *Gene - Oligophrenin 1*. https://www.genecards.org/cgi-bin/carddisp.pl?gene=OPHN1

[8] journals.plos.org. (n.d.). *Down-regulation of Tenascin-C Inhibits Breast Cancer Cells Development by Cell Growth, Migration, and Adhesion Impairment*. https://journals.plos.org/plosone/article/figure?id=10.1371/journal.pone.0237889.g001

[9] kaggle.com. (2019). *Breast Cancer Gene Expression - CuMiDa*. https://www.kaggle.com/datasets/brunogrisci/breast-cancer-gene-expression-cumida/

[10] Liu, M. C., dkk. (2015). *PAM50 Gene Signatures and Breast Cancer Prognosis with Adjuvant Anthracycline- and Taxane-Based Chemotherapy*: Correlative Analysis of C9741 (Alliance). NPJ Breast Cancer 2, 15023 10.1038/npjbcancer.2015.23. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5501351/

[11] mit.edu. (n.d.). *Introduction to the LIMMA Package*. https://web.mit.edu/~r/current/arch/i386_linux26/lib/R/library/limma/html/01Introduction.html#

[12] ncbi.nlm.nih.gov. (2023). *Carbonic Anhydrase 12 [Homo sapiens (Human)]*. https://www.ncbi.nlm.nih.gov/gene/771

[13] ncbi.nlm.nih.gov. (2023). *Estrogen Receptor 1 [Homo sapiens (Human)]*. https://www.ncbi.nlm.nih.gov/gene/2099

[14] ncbi.nlm.nih.gov. (2023). *Oligophrenin 1 [Homo sapiens (Human)]*. https://www.ncbi.nlm.nih.gov/gene/4983

[15] ncbi.nlm.nih.gov. (2023). *Tenascin XB [Homo sapiens (Human)]*. https://www.ncbi.nlm.nih.gov/gene/7148

[16] Nygard, S. (2010). *Microarray Data Analysis Using R/Bioconductor*. https://irefindex.vib.be/wiki/images/2/25/R-lab-sn.pdf

[17] Parker, J. S., dkk. (2009). *Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes*. J. Clin. Oncol. 27, 1160–1167 10.1200/JCO.2008.18.1370. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2667820/

[18] Tonissen, K. F., & Poulsen, SA. (2021). *Carbonic Anhydrase XII Inhibition Overcomes P-glycoprotein-mediated Drug Resistance: a Potential New Combination Therapy in Cancer*. Cancer Drug Resistance. 4, no.2: 343-55. http://dx.doi.org/10.20517/cdr.2020.110

[19] Wang, Y., dkk. (2021). *Identifying Breast Cancer Subtypes Associated Modules and Biomarkers by Integrated Bioinformatics Analysis*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7796196/