SAINS DATA GENOM

Laporan Ujian Tengah Semester

Breast cancer gene expression - CuMiDa

DISUSUN OLEH:

Wahyu Dimasdi Putra (2106704736 / Genom B)

DEPARTEMEN MATEMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS INDONESIA

TAHUN AJARAN 2023/2024

A. Pendahuluan

Kanker payudara merupakan salah satu jenis kanker yang paling umum dan mematikan di seluruh dunia. Selain mempengaruhi kesehatan fisik, kanker payudara juga memiliki dampak yang signifikan pada tingkat sosial, psikologis, dan ekonomi individu yang terkena dampak dan masyarakat secara keseluruhan. Oleh karena itu, pemahaman mendalam tentang faktor-faktor yang berkontribusi terhadap perkembangan dan perkembangan kanker payudara menjadi sangat penting dalam upaya pencegahan, diagnosis dini, dan perawatan yang efektif.

Dalam era digital saat ini, data menjadi aset yang sangat berharga dalam ilmu biomedis dan penelitian medis. Data genetik, khususnya, telah membuka pintu bagi pemahaman lebih dalam tentang kanker payudara dan perkembangannya. Kaggle, salah satu platform data ilmiah terkemuka, telah menyediakan akses ke sejumlah besar dataset terkait kanker payudara yang disumbangkan oleh para peneliti dan ilmuwan dari seluruh dunia. Data ini mencakup ekspresi gen, informasi klinis, dan berbagai parameter lainnya yang relevan dengan perkembangan penyakit ini.

Tujuan laporan ini adalah untuk menjelaskan proses pengolahan dan analisis data genetik kanker payudara yang diperoleh dari Kaggle. Kami akan mengeksplorasi dataset ini dengan tujuan utama untuk mengidentifikasi faktor-faktor genetik yang mungkin berkontribusi terhadap perkembangan berbagai subtip kanker payudara. Selain itu, kami akan mencoba mengidentifikasi perbedaan genetik antara sel kanker payudara dan sel payudara normal, yang dapat membantu dalam memahami mekanisme dasar penyakit ini.

Laporan ini akan menggambarkan metode yang digunakan dalam analisis, hasil yang ditemukan, serta interpretasi dan implikasi potensial dari temuan tersebut. Harapannya, penelitian ini akan memberikan wawasan yang berharga tentang gen kanker payudara dan mendorong penelitian lebih lanjut dalam upaya melawan penyakit ini.

B. Metodologi

a. Pengumpulan Data

Pengaksesan Data Kaggle: Akses dataset yang relevan dari Kaggle yang berkaitan dengan gen kanker payudara. Pastikan dataset memiliki informasi ekspresi gen dan parameter lain yang diperlukan.

b. Analisis Data

- Identifikasi Grup: Jelaskan jenis analisis yang akan dilakukan, termasuk perbandingan antara grup kanker payudara dan sel sehat atau perbandingan antara subtip kanker payudara.
- Limma Analysis: Gunakan paket Limma untuk melakukan analisis yang memungkinkan perbandingan antara kelompok. Buat model statistik yang sesuai dengan desain eksperimen Anda.

- T-test: Gunakan uji t untuk mengidentifikasi perbedaan signifikan dalam ekspresi gen antara dua grup yang berbeda.
- F-test: Gunakan uji F (analisis varians) jika Anda ingin membandingkan lebih dari dua grup.

c. Visualisasi Data
- Volcano Plot: Buat plot gunung api (volcano plot) untuk mengidentifikasi gen dengan perbedaan signifikan dan nilai p-nilai yang rendah.
- Heatmap: Hasilkan heatmap yang menampilkan ekspresi gen dengan warna yang berbeda untuk menggambarkan perbedaan antara kelompok.
- Box Plot: Buat box plot untuk menunjukkan distribusi ekspresi gen dalam kelompok yang berbeda.

d. Interpretasi Hasil
- Identifikasi Gen: Identifikasi gen yang memiliki ekspresi yang signifikan dalam perbandingan antar kelompok atau subtype.
- Kesimpulan: Buat kesimpulan berdasarkan hasil analisis dan interpretasi, serta implikasi dari temuan ini dalam pemahaman kanker payudara.

C. Olah data

1. Load data

```
> df <- fread("C:/Users/wahyu/OneDrive/Documents/KULIAH/SEMESTER 5/Genom/Breast_GSE45827.csv")
> data <- as.data.frame(df)
```

samples	type	1007_s_at	1053_at	117_at	121_at	1255_g_at	1294_at	1316_at	1320_a
84	basal	9.850040	8.097927	6.424728	7.353027	3.029122	6.880079	4.963740	4.4(
85	basal	9.861357	8.212222	7.062593	7.685578	3.149468	7.542283	5.129607	4.5{
87	basal	10.103478	8.936137	5.735970	7.687822	3.125931	6.562369	4.813449	4.4:
90	basal	9.756875	7.357148	6.479183	6.986624	3.181638	7.802344	5.490982	4.5(
91	basal	9.408330	7.746404	6.693980	7.333426	3.169923	7.610457	5.372469	4.4:
92	basal	7.505488	8.802820	6.235074	7.202227	2.987976	7.985281	5.413368	4.4(
93	basal	10.422371	8.933601	5.630488	6.881770	3.097372	6.273211	5.414001	4.4:
94	basal	10.190705	7.813057	6.701297	6.921350	3.140037	7.524231	5.102223	4.8:
99	basal	10.256077	7.796936	6.725722	7.098550	3.139031	6.885392	5.171368	4.6{
101	basal	9.053138	8.043154	7.655891	6.900599	2.988920	7.669423	5.417630	4.4{
102	basal	9.775080	6.766403	6.539591	7.631263	3.130671	6.894172	5.430312	4.7:

2. Ambil 50% Data dengan Seed NPM Mahasiswa
- Ambil data

```
> #Ambil random 50% data dengan seed NPM = 2106704736
> set.seed(2106704736)
> cols_to_remove <- sample(3:54674, size = 54674 / 2)
> data <- data[ , -cols_to_remove]
```

samples	type	1007_s_at	117_at	1294_at	1405_i_at	1438_at	1487_at	1494_f_at	155
84	basal	9.850040	6.424728	6.880079	8.870780	7.317273	6.941792	4.753893	
85	basal	9.861357	7.062593	7.542283	7.767646	8.768129	7.567034	4.859823	
87	basal	10.103478	5.735970	6.562369	9.417956	7.945085	7.238284	4.872635	
90	basal	9.756875	6.479183	7.802344	9.022345	7.754670	7.078843	4.867320	
91	basal	9.408330	6.693980	7.610457	9.400056	9.052192	6.889370	4.939912	
92	basal	7.505488	6.235074	7.985281	7.439545	7.829680	7.237458	4.699347	
93	basal	10.422371	5.630488	6.273211	3.931658	10.581627	7.118396	4.927776	
94	basal	10.190705	6.701297	7.524231	9.072734	8.975758	6.639122	4.970631	
99	basal	10.256077	6.725722	6.885392	7.962521	9.070648	6.597565	4.427259	
101	basal	9.053138	7.655891	7.669423	10.033215	7.047227	7.150225	4.874087	
102	basal	9.775080	6.539591	6.894172	7.863688	8.295889	7.362132	5.052474	

- Dimensi data setelah diambil 50% random data dengan seed

```
> dim(data)
[1]   151 27340
```

Terlihat bahwa data sekarang hanya tersisa setengahnya yaitu sebanyak 27340 dari 54677 features.

3. Eksplorasi data
   - Dimensi data

```
> dim(data)
[1]   151 27340
```

Terlihat dimensi data sebesar 151x27340 dengan 151 adalah banyaknya sampel dan 27338 adalah jumlah fitur.

   - Data structure

```
'data.frame':   151 obs. of  27340 variables:
 $ samples                : int  84 85 87 90 91 92 93 94 99 101 ...
 $ type                   : chr  "basal" "basal" "basal" "basal" ..
 $ 1007_s_at              : num  9.85 9.86 10.1 9.76 9.41 ...
 $ 117_at                 : num  6.42 7.06 5.74 6.48 6.69 ...
 $ 1294_at                : num  6.88 7.54 6.56 7.8 7.61 ...
 $ 1405_i_at              : num  8.87 7.77 9.42 9.02 9.4 ...
 $ 1438_at                : num  7.32 8.77 7.95 7.75 9.05 ...
 $ 1487_at                : num  6.94 7.57 7.24 7.08 6.89 ...
 $ 1494_f_at              : num  4.75 4.86 4.87 4.87 4.94 ...
 $ 1552257_a_at           : num  9.53 8.97 9.13 7.21 9.24 ...
 $ 1552258_at             : num  4.42 4.5 4.36 4.68 4.69 ...
 $ 1552269_at             : num  3.71 3.94 4.26 3.31 4.39 ...
 $ 1552271_at             : num  5.08 5.23 5.51 5.17 5.27 ...
 $ 1552274_at             : num  6.74 6.47 7.1 6.73 6.98 ...
 $ 1552276_a_at           : num  5.52 5.31 5.14 6.09 5.19 ...
 $ 1552277_a_at           : num  8.11 7.51 7.07 7.16 8.23 ...
 $ 1552279_a_at           : num  5.79 5.9 6.23 5.73 6.17 ...
 $ 1552281_at             : num  5.69 5.58 6.01 5.65 5.69 ...
 $ 1552286_at             : num  4.91 5.87 5.72 6.15 5.21 ...
 $ 1552291_at             : num  5.94 6.29 7.26 6.3 6.22 ...
 $ 1552293_at             : num  4.19 4.06 4.33 4.19 4 ...
 $ 1552296_at             : num  3.28 3 4.03 3.25 3.63 ...
 $ 1552299_at             : num  4.35 5.17 4.2 5.05 3.49 ...
```

Dari structure data terlihat bahwa terdapat satu variabel type dengan data tipe karakteristik dan variabel lainnya bersifat numerik yang merupakan fitur.

- Variabel kategorik type

```
> data$type
  [1] "basal"     "basal"     "basal"     "basal"     "basal"     "basal"
  [7] "basal"     "basal"     "basal"     "basal"     "basal"     "basal"
 [13] "basal"     "basal"     "basal"     "basal"     "basal"     "basal"
 [19] "basal"     "basal"     "basal"     "basal"     "basal"     "basal"
 [25] "basal"     "basal"     "basal"     "basal"     "basal"     "basal"
 [31] "basal"     "basal"     "basal"     "basal"     "basal"     "basal"
 [37] "basal"     "basal"     "basal"     "basal"     "basal"     "HER"
 [43] "HER"       "HER"       "HER"       "HER"       "HER"       "HER"
 [49] "HER"       "HER"       "HER"       "HER"       "HER"       "HER"
 [55] "HER"       "HER"       "HER"       "HER"       "HER"       "HER"
 [61] "HER"       "HER"       "HER"       "HER"       "HER"       "HER"
 [67] "HER"       "HER"       "HER"       "HER"       "HER"       "cell_line"
 [73] "cell_line" "cell_line" "cell_line" "cell_line" "cell_line" "cell_line"
 [79] "cell_line" "cell_line" "cell_line" "cell_line" "cell_line" "cell_line"
 [85] "cell_line" "normal"    "normal"    "normal"    "normal"    "normal"
 [91] "normal"    "normal"    "luminal_A" "luminal_A" "luminal_A" "luminal_A"
 [97] "luminal_A" "luminal_A" "luminal_A" "luminal_A" "luminal_A" "luminal_A"
[103] "luminal_A" "luminal_A" "luminal_A" "luminal_A" "luminal_A" "luminal_A"
[109] "luminal_A" "luminal_A" "luminal_A" "luminal_A" "luminal_A" "luminal_A"
[115] "luminal_A" "luminal_A" "luminal_A" "luminal_A" "luminal_A" "luminal_A"
[121] "luminal_A" "luminal_B" "luminal_B" "luminal_B" "luminal_B" "luminal_B"
[127] "luminal_B" "luminal_B" "luminal_B" "luminal_B" "luminal_B" "luminal_B"
[133] "luminal_B" "luminal_B" "luminal_B" "luminal_B" "luminal_B" "luminal_B"
[139] "luminal_B" "luminal_B" "luminal_B" "luminal_B" "luminal_B" "luminal_B"
[145] "luminal_B" "luminal_B" "luminal_B" "luminal_B" "luminal_B" "luminal_B"
[151] "luminal_B"
> unique(data$type)
[1] "basal"     "HER"       "cell_line" "normal"    "luminal_A" "luminal_B"
```

Terlihat bahwa terdapat 6 kategorik untuk datanya yaitu basal, HER, cell_line, normal, luminal_A, dan luminal_B.

4. Expression Set

```
> #eskpression set
> AssayData <- data[, -c(1,2)]
> AssayData <- t(AssayData)
> rownames <- data$samples
> colnames(AssayData) <- rownames
> View(AssayData)
> class(AssayData)
[1] "matrix" "array"
> dim(AssayData)
[1] 27338   151
> colnames(AssayData)
  [1] "84"  "85"  "87"  "90"  "91"  "92"  "93"  "94"  "99"  "101" "102" "106" "108"
 [14] "109" "110" "111" "112" "113" "114" "115" "118" "119" "120" "124" "125" "128"
 [27] "130" "131" "135" "136" "138" "139" "142" "143" "144" "145" "146" "147" "148"
 [40] "149" "150" "86"  "88"  "89"  "95"  "96"  "97"  "98"  "100" "103" "104" "105"
 [53] "107" "116" "117" "121" "122" "123" "126" "127" "129" "132" "133" "134" "137"
 [66] "140" "141" "151" "152" "153" "154" "155" "156" "157" "158" "159" "160" "161"
 [79] "162" "163" "164" "165" "166" "167" "168" "171" "172" "173" "174" "175" "177"
 [92] "178" "180" "181" "184" "187" "190" "191" "194" "195" "199" "200" "203" "204"
[105] "205" "208" "209" "213" "216" "217" "220" "221" "222" "223" "224" "225" "228"
[118] "231" "232" "234" "235" "182" "183" "185" "186" "188" "189" "192" "193" "196"
[131] "197" "198" "201" "202" "206" "207" "210" "211" "212" "214" "215" "218" "219"
[144] "226" "227" "229" "230" "233" "236" "237" "238"
```

```
> phenodt <- data[, c(1,2)]
> rownames(phenodt) <- phenodt$samples
> phenodt$samples <- NULL
> summary(phenodt)
     type
 Length:151
 Class :character
 Mode  :character
> View(phenodt)
> phenodt <- AnnotatedDataFrame(phenodt)
> eset <- ExpressionSet(assayData = AssayData, phenoData = phenodt)
```

Dilakukan convert data ke Expression Set agar struktur data didesain khusus untuk menyimpan data ekspresi gen secara terorganisir sehingga memudahkan untuk mengakses, memanipulasi, dan menganalisis data ekspresi gen. Data ekspresi gen umumnya memiliki dua dimensi: sampel (kolom) dan gen (baris), dan ExpressionSet memodelkannya dengan baik.

5. Expression Data

```
> ## Expression data ##
> expdtgeo <- exprs(eset)
> dim(expdtgeo)
[1] 27338   151
> head(expdtgeo)
               84       85        87       90       91       92        93        94
1007_s_at 9.850040 9.861357 10.103478 9.756875 9.408330 7.505488 10.422371 10.190705
117_at    6.424728 7.062593  5.735970 6.479183 6.693980 6.235074  5.630488  6.701297
1294_at   6.880079 7.542283  6.562369 7.802344 7.610457 7.985281  6.273211  7.524231
1405_i_at 8.870780 7.767646  9.417956 9.022345 9.400056 7.439545  3.931658  9.072734
1438_at   7.317273 8.768129  7.945085 7.754670 9.052192 7.829680 10.581627  8.975758
1487_at   6.941792 7.567034  7.238284 7.078843 6.889370 7.237458  7.118396  6.639122
               99       101       102       106       108       109       110
1007_s_at 10.256077  9.053138 9.775080 10.651036 9.588651 10.216060 10.109957
117_at     6.725722  7.655891 6.539591  6.222672 7.530525  6.617939  6.441200
1294_at    6.885392  7.669423 6.894172  7.290754 7.488740  7.759867  7.288320
1405_i_at  7.962521 10.033215 7.863688  8.216927 8.033562 10.401497  8.045014
1438_at    9.070648  7.047227 8.295889  9.061330 5.799147  6.818021  7.302156
1487_at    6.597565  7.150225 7.362132  7.222223 7.184267  7.107183  6.841015
```

Selanjutnya mengambil dan menampilkan data ekspresi gen dalam sebuah objek yang disebut "eset" yang merupakan langkah penting dalam analisis data biologis seperti data mikroarray atau RNA-Seq. Data ekspresi gen ini adalah data yang mengukur tingkat ekspresi gen berdasarkan sampel. Didapat dimensi dataframe yang berisi jumlah baris (gen) dan jumlah kolom (sampel) dalam data frame. Pada kasus ini, terdapat 27338 gen dan 151 sampel.

D. Filtering gene

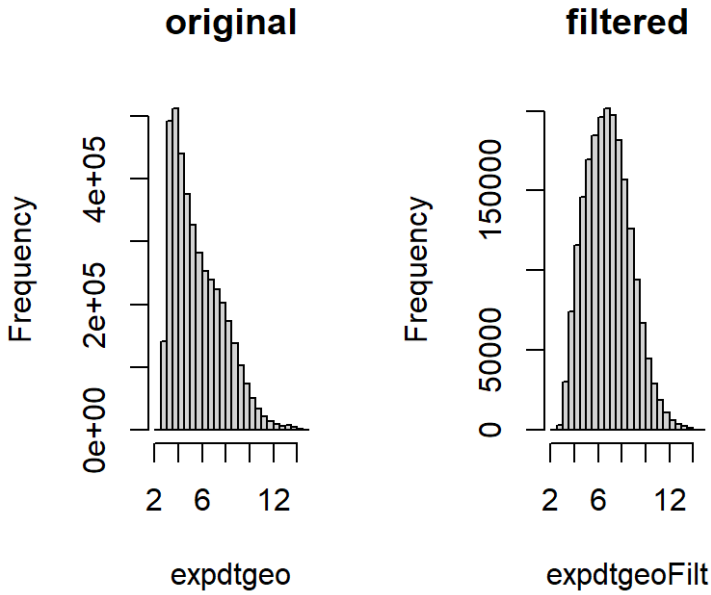
1. Hasil Filtering

```
> # Calculate the interquartile range (IQR) for each row
> iqr_values <- apply(expdtgeo, 1, IQR)
> # Filter rows based on IQR values
> expdtgeoFilt <- expdtgeo[iqr_values > quantile(iqr_values, 0.5), ]
> dim(expdtgeoFilt)
[1] 13669   151
> par(mfrow=c(1,2))
> hist(expdtgeo, main='original')
> hist(expdtgeoFilt, main='filtered')
```

Filtering gene dilakukan dengan interquartile range (IQR) untuk setiap barisnya. Filtering menggunakan IQR merupakan salah satu metode umum dalam analisis data ekspresi gen untuk mengidentifikasi gen yang memiliki variasi signifikan dalam ekspresi mereka di seluruh sampel. IQR adalah metode statistik yang digunakan untuk mengukur sebaran data dalam kumpulan data, terutama data yang tidak terdistribusi normal. Didapatkan hasil sebagai berikut.

2. Comparing before and after filtering

```
> par(mfrow=c(1,2))
> hist(expdtgeo, main='original')
> hist(expdtgeoFilt, main='filtered')
```



Selanjutnya, dibuat plot perbandingan pada data sebelum dan sesudah filtering. Hasil filtering secara keseluruhan telah mengurangi jumlah fitur dari 27338 menjadi 13669.

E. T test

Pertama dilakukan t test untuk model yang dibuat.

```
fit <- lmFit(eset, design)
fit <- eBayes(fit)
# Ambil statistik t untuk setiap gen
t_statistics <- fit$t
# Hitung p-value
p_values <- 2 * pt(-abs(t_statistics), df = fit$df.residual)
# Gabungkan statistik t dan p-value ke dalam satu data frame
results <- data.frame(T_Statistic = t_statistics, P_Value = p_values)
# Lihat hasilnya
head(results)
```

```
          T_Statistic.basal T_Statistic.cell_line T_Statistic.HER
1007_s_at         118.15515              68.17650       104.69184
117_at             78.85178              37.37100        64.25500
1294_at            86.35949              45.92564        72.99425
1405_i_at          35.34510              14.80597        28.13977
1438_at            57.48322              28.00571        44.24535
1487_at           127.40596              85.73108       110.29183
          T_Statistic.luminal_A T_Statistic.luminal_B T_Statistic.normal
1007_s_at             104.18888             108.48177           49.74018
117_at                 62.83462              67.14670           27.63165
1294_at                77.24554              77.81781           40.43866
1405_i_at              26.50496              28.25237           13.16135
1438_at                42.72786              43.75677           21.81695
1487_at               105.97729             111.37040           54.14697
          P_Value.basal P_Value.cell_line   P_Value.HER P_Value.luminal_A
1007_s_at 4.908220e-146     4.808390e-112 1.664664e-138     3.316141e-138
117_at    5.768982e-121      2.588562e-76 1.966762e-108     4.497050e-107
1294_at   1.420895e-126      2.769929e-88 3.185423e-116     1.064596e-119
1405_i_at  3.783569e-73      8.499750e-31  1.287370e-60      1.858761e-57
1438_at   1.104588e-101      2.312052e-60  4.308899e-86      4.748257e-84
1487_at   9.751390e-151     4.015047e-126 9.557324e-142     2.901841e-139
          P_Value.luminal_B P_Value.normal
1007_s_at     1.023296e-140   5.200367e-93
117_at        4.085284e-111   1.196850e-59
1294_at       3.743290e-120   7.540037e-81
1405_i_at      7.883967e-61   1.590203e-26
1438_at        1.928487e-85   1.197959e-47
1487_at       2.369212e-142   4.383938e-98
```

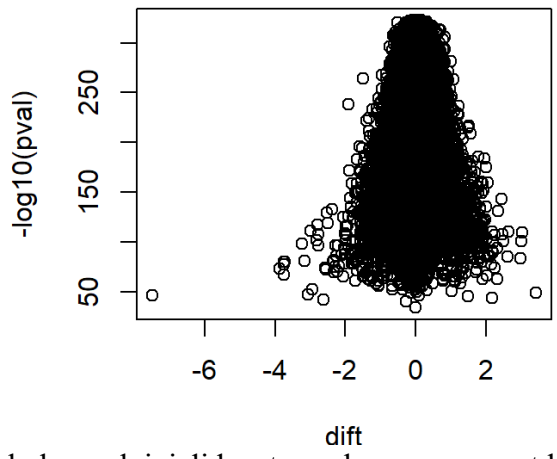
Dapat dihat untuk data teratasnya didapat nilai p-value yang beragam tetapi nilainya sangat kecil sehingga dapat dikatakan signifikan.

Selanjutnya akan dilakukan t-test untuk membandingkan 4 jenis cancernya (seharusnya anova tetapi tidak menemukan code yang sesuai sehingga dilakukan 1 per satu dengan t test)

```
> ### Assign Group Code ###
> vargrp <- phdtgeo
> table(vargrp)
vargrp
    basal cell_line       HER luminal_A luminal_B    normal
       41        14        30        29        30         7
```

- Basal dan bukan Basal

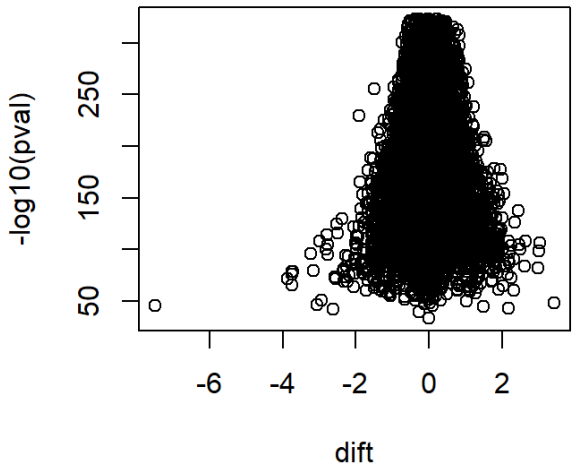
```
> t.test(group, expdtgeoFilt[1, ])$p.value
[1] 2.012228e-269
> pval <- apply(expdtgeoFilt, 1, function(x) t.test(group, x)$p.value)
> dift <- apply(expdtgeoFilt, 1, function(x) diff(t.test(x[1:4], x[5:8])$esti
mate))
> plot(dift, -log10(pval))
```



Untuk perbandingan kelompok ini didapat p-value yang sangat kecil yaitu 2.012228e-269 yang mengindikasi bahwa kelompok basal signifikan

- HER dan bukan HER

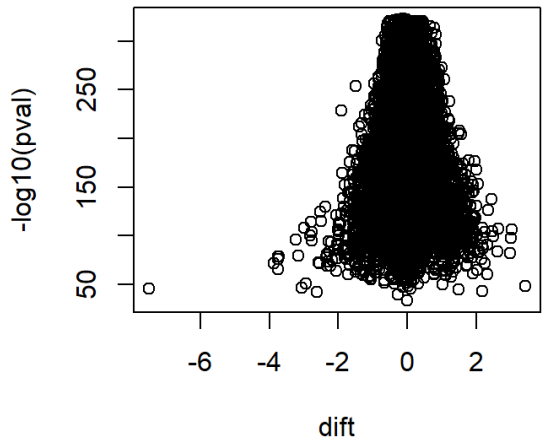
```
> t.test(group, expdtgeoFilt[1, ])$p.value
[1] 2.870182e-260
> pval <- apply(expdtgeoFilt, 1, function(x) t.test(group, x)$p.value)
> dift <- apply(expdtgeoFilt, 1, function(x) diff(t.test(x[1:4], x[5:8])$estimate))
> plot(dift, -log10(pval))
```



Untuk perbandingan kelompok ini didapat p-value yang sangat kecil yaitu 2.870182e-260 yang mengindikasi bahwa kelompok HER signifikan

- Luminal A dan bukan Luminal A

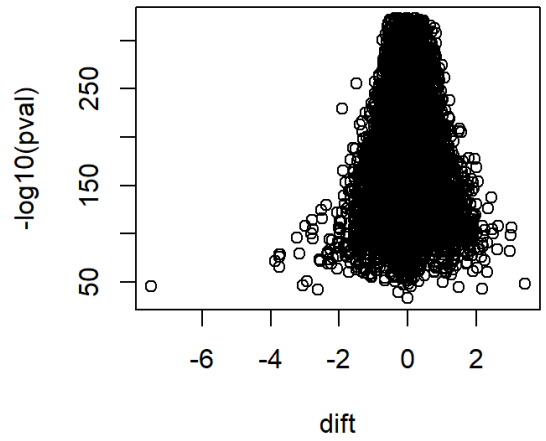
```
> t.test(group, expdtgeoFilt[1, ])$p.value
[1] 4.272104e-259
> pval <- apply(expdtgeoFilt, 1, function(x) t.test(group, x)$p.value)
> dift <- apply(expdtgeoFilt, 1, function(x) diff(t.test(x[1:4], x[5:8])$estimate))
```



Untuk perbandingan kelompok ini didapat p-value yang sangat kecil yaitu 4.272104e-259 yang mengindikasi bahwa kelompok Luminal A signifikan.

- Luminal B dan bukan Luminal B

```
> t.test(group, expdtgeoFilt[1, ])$p.value
[1] 2.870182e-260
> pval <- apply(expdtgeoFilt, 1, function(x) t.test(group, x)$p.value)
> dift <- apply(expdtgeoFilt, 1, function(x) diff(t.test(x[1:4], x[5:8])$estimate))
> plot(dift, -log10(pval))
```



Untuk perbandingan kelompok ini didapat p-value yang sangat kecil yaitu 2.870182e-260 yang mengindikasi bahwa kelompok Luminal B signifikan.

F. Limma Analysis

1. Preparation

```
> vargrp <- data[, 2]
> col_cancer_idx <- ifelse(vargrp %in% c('basal', 'HER', 'luminal_A', 'luminal_B'),
+                          0, 1)
> col_cancer <- vargrp[col_cancer_idx == 0]
> col_normal_idx <- ifelse(vargrp %in% c('normal', 'basal', 'HER', 'luminal_A', 'lu
minal_B'),
+                          0, 1)
> col_normal <- vargrp[col_normal_idx == 0]
> expdtgeo2 = expdtgeoFilt[, col_cancer_idx==0]
> expdtgeo3 = expdtgeoFilt[, col_normal_idx==0]
> dim(expdtgeo2)
[1] 13669   130
> dim(expdtgeo3)
[1] 13669   137
> group_cancer <- ifelse(col_cancer == "basal", 0,
+                        ifelse(col_cancer == "HER", 1,
+                               ifelse(col_cancer == "luminal_A", 2, 3)))
> group_cancer
  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [38] 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2
 [75] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3
[112] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

Dibuat kelompok untuk dilakukan perbandingan agar lebih mudah dan terstruktur.

2. Limma Analysis for Cancer Group

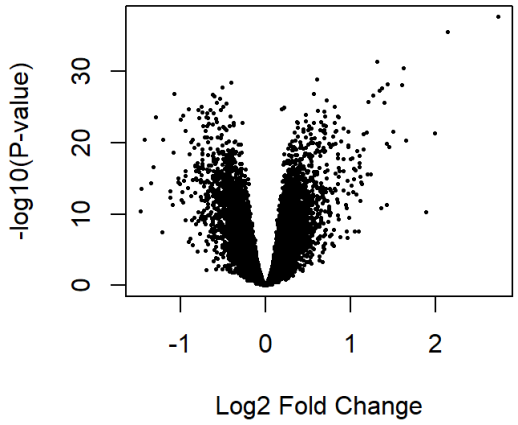
```
> #Limma
> design <- model.matrix(~group_cancer)
> fit <- eBayes(lmFit(expdtgeo2, design))
> fit
An object of class "MArrayLM"
$coefficients
          (Intercept) group_cancer
1007_s_at   10.066976   0.23742493
117_at       6.397457  -0.02479745
1294_at      7.126536   0.16282014
1405_i_at    8.280253  -0.21688654
1438_at      7.799478  -0.29685782
13664 more rows ...

$rank
[1] 2

$assign
[1] 0 1

$qr
$qr
  (Intercept) group_cancer
1 -11.4017543  -15.6116327
2   0.0877058   13.1254304
3   0.0877058    0.0959073
4   0.0877058    0.0959073
5   0.0877058    0.0959073
125 more rows ...

$qraux
[1] 1.087706 1.095907

$pivot
[1] 1 2

$tol
[1] 1e-07

$rank
[1] 2


$df.residual
[1] 128 128 128 128 128
13664 more elements ...
```

```
$sigma
1007_s_at    117_at   1294_at 1405_i_at   1438_at
0.4875963 0.5730573 0.5630307 1.4950210 0.9356038
13664 more elements ...

$cov.coefficients
             (Intercept) group_cancer
(Intercept)   0.018574745 -0.007947848
group_cancer -0.007947848  0.005804608

$stdev.unscaled
         (Intercept) group_cancer
1007_s_at   0.1362892   0.07618798
117_at      0.1362892   0.07618798
1294_at     0.1362892   0.07618798
1405_i_at   0.1362892   0.07618798
1438_at     0.1362892   0.07618798
13664 more rows ...

$pivot
[1] 1 2

$Amean
1007_s_at    117_at   1294_at 1405_i_at   1438_at
10.392065  6.363503  7.349474  7.983285  7.393011
13664 more elements ...
$method
[1] "ls"

$design
  (Intercept) group_cancer
1           1            0
2           1            0
3           1            0
4           1            0
5           1            0
125 more rows ...

$df.prior
[1] 4.943439

$s2.prior
[1] 0.2839017

$var.prior
[1] 56.35753  0.81380

$proportion
[1] 0.01

$s2.post
1007_s_at    117_at   1294_at 1405_i_at   1438_at
0.2394662 0.3267402 0.3157727 2.1625339 0.8533615
13664 more elements ...
$t
          (Intercept) group_cancer
1007_s_at   150.94386    6.3682151
117_at       82.11918   -0.5694025
1294_at      93.05287    3.8030709
1405_i_at    41.31433   -1.9358187
1438_at      61.94947   -4.2178925
13664 more rows ...

$df.total
[1] 132.9434 132.9434 132.9434 132.9434 132.9434
13664 more elements ...

$p.value
            (Intercept) group_cancer
1007_s_at 1.503670e-150 2.877652e-09
117_at    8.527631e-116 5.700439e-01
1294_at   6.867112e-123 2.169498e-04
1405_i_at  1.009092e-77 5.501240e-02
1438_at   6.132361e-100 4.531234e-05
13664 more rows ...
```

```
$lods
          (Intercept) group_cancer
1007_s_at    332.6021   10.6158163
117_at       254.5524   -6.9082314
1294_at      270.6958   -0.2054656
1405_i_at    167.0812   -5.2218990
1438_at      218.2494    1.2775523
13664 more rows ...

$F
[1] 29334.123  8055.863 11125.863  1917.513  4172.055
13664 more elements ...

$F.p.value
[1] 1.362998e-176 1.864663e-139 1.035353e-148  9.136264e-99 1.113267e-120
13664 more elements ...
```

Didapatkan output dari LIMMA, dapat dilihat statistik untuk t, F, dan p-valuenya terhadap model untuk grup kangker.

```
> volcanoplot(fit, coef=2)
```

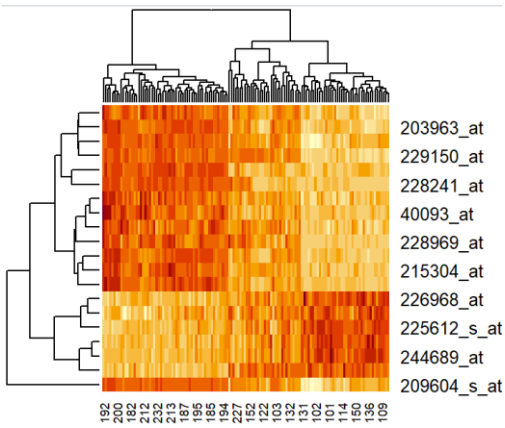


Top Result for Cancer Group

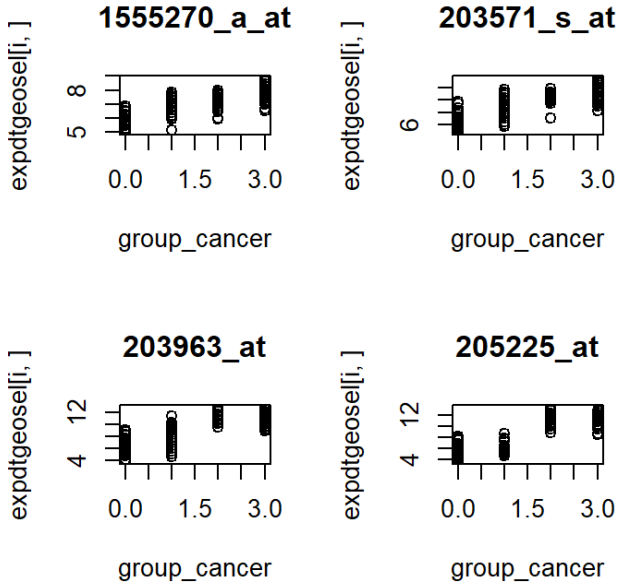
Berikut adalah 20 top result dari kelompok cancernya yang memiliki perbedaan cukup signifikan.

```
> #top Result
> topResult <- topTable(fit, coef=2, number=20)
> rownames(topResult)
 [1] "228241_at"    "205225_at"    "233388_at"    "203963_at"    "218394_at"
 [6] "228673_s_at"  "228969_at"    "229150_at"    "226968_at"    "203571_s_at"
[11] "210735_s_at"  "1555270_a_at" "225612_s_at"  "223438_s_at"  "209604_s_at"
[16] "210092_at"    "244689_at"    "40093_at"     "215304_at"    "209602_s_at"
> selected <- rownames(expdtgeo2) %in% rownames(topResult)
> expdtgeosel <- expdtgeo2[selected, ]
> heatmap(expdtgeosel)
```

Berikut adalah heatmap yang menggambarkan pola ekspresi ke 20 gen tersebut.

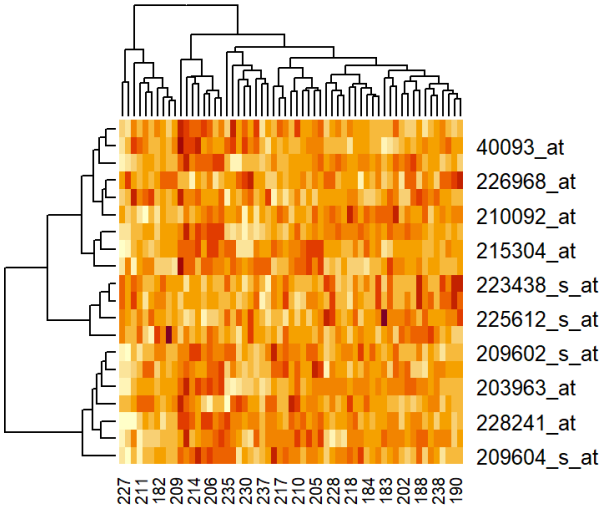



```
> par(mfrow=c(2,2))
> for (i in 1:4) plot(group_cancer, expdtgeosel[i,],
+                     main = rownames(expdtgeosel)[i])
```



Dari plot di atas terlihat bahwa gen-gen tersebut memiliki ekspresi yang sangat berbeda. Selanjutnya dilihat data untuk luminal A dan B untuk dibandingkan dengan yang sebelumnya.

```
> luminal_data = data[data[, 'type'] %in% c('luminal_A', 'luminal_B'),
+                     c('samples', rownames(topResult))]
> dim(luminal_data)
[1] 59 21
> selected_luminal <- rownames(expdtgeo2) %in% rownames(topResult)
> expdtgeosel_luminal <- expdtgeo2[selected_luminal, as.character(luminal_data[, 's
amples'])]
> heatmap(expdtgeosel_luminal)
```



3. Limma Analysis for Normal Group

   Pertama dibuat grup bentuk dikotomik untuk memudahkan proses analisis.

```
> group_normal <- ifelse(col_normal == "normal", 0, 1)
> group_normal
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0
 [75] 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[112] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

   Berikut adalah hasil analisis LIMMAnya

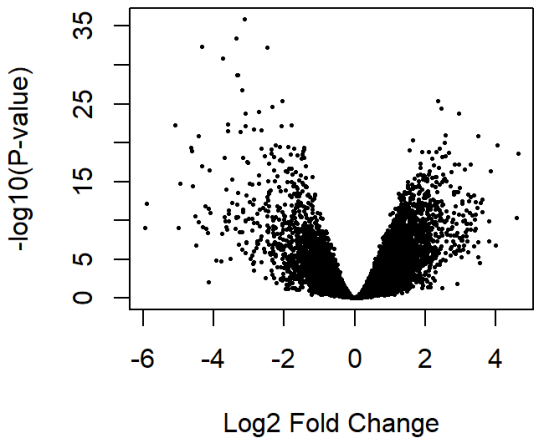
```
> design <- model.matrix(~group_normal)
> fit <- eBayes(lmFit(expdtgeo3, design))
> fit
An object of class "MArrayLM"
$coefficients
          (Intercept) group_normal
1007_s_at   10.219646   0.17241981
117_at       5.527662   0.83584104
1294_at      8.135556  -0.78608212
1405_i_at    7.615436   0.36784893
1438_at      7.336943   0.05606836
13664 more rows ...

$rank
[1] 2

$assign
[1] 0 1

$qr
$qr
   (Intercept) group_normal
1 -11.70469991 -11.10664955
2   0.08543577  -2.57727293
3   0.08543577   0.01826472
4   0.08543577   0.01826472
5   0.08543577   0.01826472
132 more rows ...

$qraux
[1] 1.085436 1.018265
```

```
$pivot
[1] 1 2

$tol
[1] 1e-07

$rank
[1] 2


$df.residual
[1] 135 135 135 135 135
13664 more elements ...

$sigma
1007_s_at    117_at   1294_at 1405_i_at   1438_at
0.5534755 0.5643170 0.5795877 1.4912309 0.9846985
13664 more elements ...

$cov.coefficients
             (Intercept) group_normal
(Intercept)    0.1428571   -0.1428571
group_normal  -0.1428571    0.1505495


$stdev.unscaled
          (Intercept) group_normal
1007_s_at   0.3779645     0.388007
117_at      0.3779645     0.388007
1294_at     0.3779645     0.388007
1405_i_at   0.3779645     0.388007
1438_at     0.3779645     0.388007
13664 more rows ...

$pivot
[1] 1 2

$Amean
1007_s_at    117_at   1294_at 1405_i_at   1438_at
10.383256  6.320796  7.389639  7.964490  7.390146
13664 more elements ...

$method
[1] "ls"

$design
  (Intercept) group_normal
1           1            1
2           1            1
3           1            1
4           1            1
5           1            1
132 more rows ...

$df.prior
[1] 4.758756
```

```
$s2.prior
[1] 0.3118214

$var.prior
[1] 51.31143 15.34825

$proportion
[1] 0.01

$s2.post
1007_s_at    117_at   1294_at 1405_i_at   1438_at
0.3065219 0.3182278 0.3351013 2.1586681 0.9472329
13664 more elements ...

$t
          (Intercept) group_normal
1007_s_at    48.83757    0.8026326
117_at       25.92516    3.8186954
1294_at      37.18332   -3.4997761
1405_i_at    13.71358    0.6452633
1438_at      19.94508    0.1484738
13664 more rows ...

$df.total
[1] 139.7588 139.7588 139.7588 139.7588 139.7588
13664 more elements ...
$p.value
           (Intercept) group_normal
1007_s_at 1.028246e-89 0.4235499033
117_at    2.974465e-55 0.0002010077
1294_at   2.368796e-74 0.0006251674
1405_i_at 1.184425e-27 0.5198144638
1438_at   1.022629e-42 0.8821828556
13664 more rows ...

$lods
          (Intercept) group_normal
1007_s_at   192.88323 -6.591715130
117_at      115.35538  0.001478462
1294_at     158.62746 -1.059546182
1405_i_at    52.20108 -6.704909269
1438_at      86.71788 -6.901240471
13664 more rows ...

$F
[1] 24093.612  8607.242 11168.608  2013.101  3949.491
13664 more elements ...

$F.p.value
[1] 3.885894e-178 4.685135e-147 6.622417e-155 9.425595e-104 1.060823e-123
13664 more elements ...
```

Didapatkan output dari LIMMA, dapat dilihat statistik untuk t, F, dan p-valuenya terhadap model untuk grup normal.

```
> volcanoplot(fit, coef=2)
```

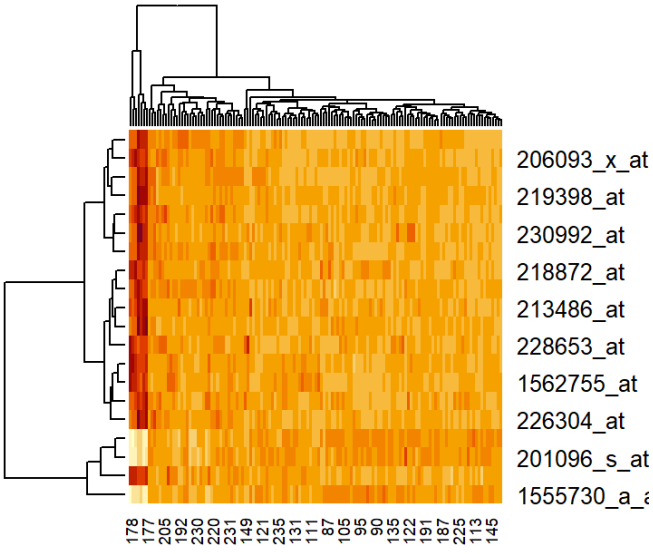


Top Result for Normal Group

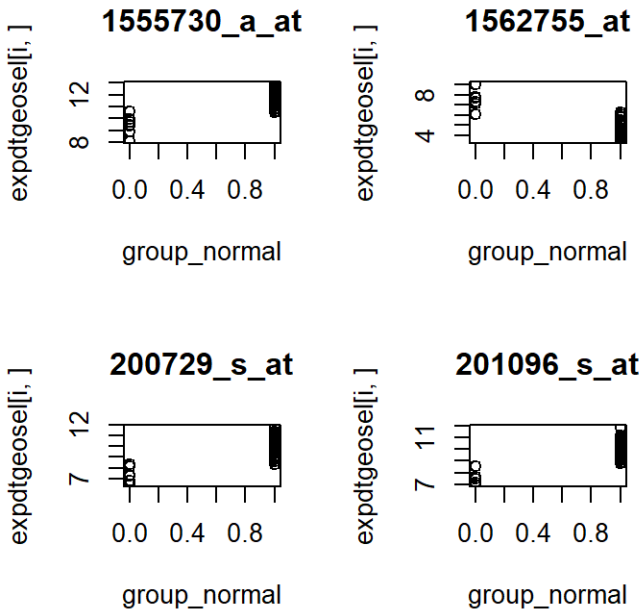
Berikut adalah 20 top result dari kelompok normalnya yang memiliki perbedaan cukup signifikan.

```
> topResult_normal <- topTable(fit, coef=2, number=20)
> rownames(topResult_normal)
 [1] "211565_at"    "239523_at"    "206093_x_at"  "218872_at"    "219398_at"
 [6] "1562755_at"   "213486_at"    "208609_s_at"  "1555730_a_at" "234675_x_at"
[11] "230992_at"    "201096_s_at"  "226304_at"    "216331_at"    "200729_s_at"
[16] "221928_at"    "205507_at"    "205913_at"    "228653_at"    "229015_at"
```

Berikut adalah heatmap yang menggambarkan pola ekspresi ke 20 gen tersebut.




```
> par(mfrow=c(2,2))
> for (i in 1:4) plot(group_normal, expdtgeosel[i,],
+                     main = rownames(expdtgeosel)[i])
```



Dari plot di atas terlihat bahwa gen-gen tersebut memiliki ekspresi yang sangat berbeda.

G. Kesimpulan

- Proses Filtering IQR: Proses filtering dengan menggunakan Interquartile Range (IQR) telah berhasil mengurangi jumlah fitur dari 27,338 menjadi 13,669. Ini menunjukkan bahwa metode IQR efektif dalam menghilangkan fitur-fitur yang kurang signifikan dalam dataset, menghasilkan dataset yang lebih terfokus.

- P-value yang Sangat Rendah: P-value yang dihasilkan dari berbagai uji statistik, seperti t-test dan F-test yang dilakukan dengan metode Limma, memiliki nilai yang sangat rendah. Nilai p-nilai yang sangat kecil ini mengindikasikan bahwa perbedaan antara kelompok yang dianalisis adalah signifikan secara statistik.

- Gen Terbaik untuk Kelompok Kanker: Berdasarkan analisis, telah mengidentifikasi gen terbaik yang berperan dalam kelompok kanker. Gen-gen ini adalah "228241_at," "205225_at," "233388_at," "203963_at," "218394_at," "228673_s_at," "228969_at," "229150_at," "226968_at," "203571_s_at," "210735_s_at," "1555270_a_at," "225612_s_at," "223438_s_at," "209604_s_at," "210092_at," "244689_at," "40093_at," "215304_at," dan "209602_s_at."

- Gen Terbaik untuk Kelompok Normal: Telah diidentifikasi gen terbaik yang berperan dalam kelompok normal. Gen-gen ini adalah "211565_at," "239523_at," "206093_x_at," "218872_at," "219398_at," "1562755_at," "213486_at," "208609_s_at," "1555730_a_at," "234675_x_at," "230992_at," "201096_s_at," "226304_at," "216331_at," "200729_s_at," "221928_at," "205507_at," "205913_at," "228653_at," dan "229015_at."

Referensi:

Dai, X., Cheng, H., Bai, Z., & Li, J. (2017). Breast Cancer Cell Line Classification and Its Relevance with Breast Tumor Subtyping. *Journal of Cancer*, *8*(16), 3131–3141. https://doi.org/10.7150/jca.18457

Understanding volcano plots. (2022, August 25). HTG Molecular. https://www.htgmolecular.com/blog/2022-08-25/understanding-volcano-plots

Breast cancer gene expression - CuMiDa. (2020, February 1). Kaggle. https://www.kaggle.com/datasets/brunogrisci/breast-cancer-gene-expression-cumida/code

Chaudhary, S. (2021, December 12). Why “1.5” in IQR Method of Outlier Detection? - towards data science. *Medium*. https://towardsdatascience.com/why-1-5-in-iqr-method-of-outlier-detection-5d07fdc82097

Lampiran R Code

```
library(Biobase)
library(GEOquery)
library(BiocManager)
library(BiocGenerics)
library(limma)
library(data.table)
df <- fread("C:/Users/wahyu/OneDrive/Documents/KULIAH/SEMESTER
5/Genom/Breast_GSE45827.csv")
data <- as.data.frame(df)
#head(data)

#Ambil random 50% data dengan seed NPM = 2106704736
set.seed(2106704736)
cols_to_remove <- sample(3:54674, size = 54674 / 2)
data <- data[ , -cols_to_remove]
dim(data)

#Eksplorasi data
dim(data)
str(data)
unique(data$type)

#eskpression set
AssayData <- data[, -c(1,2)]
AssayData <- t(AssayData)
rownames <- data$samples
colnames(AssayData) <- rownames
View(AssayData)
class(AssayData)
dim(AssayData)
colnames(AssayData)
phenodt <- data[, c(1,2)]
rownames(phenodt) <- phenodt$samples
phenodt$samples <- NULL
summary(phenodt)
View(phenodt)
phenodt <- AnnotatedDataFrame(phenodt)
eset <- ExpressionSet(assayData = AssayData, phenoData = phenodt)

### Phenotype Data ###
phdtgeo <- pData(eset)
```

```
head(phdtgeo)

## Expression data ##
expdtgeo <- exprs(eset)
dim(expdtgeo)
head(expdtgeo)


## gene filtering ##
library(dplyr)
# Calculate the interquartile range (IQR) for each row
iqr_values <- apply(expdtgeo, 1, IQR)
# Filter rows based on IQR values
expdtgeoFilt <- expdtgeo[iqr_values > quantile(iqr_values, 0.5), ]
dim(expdtgeoFilt)

par(mfrow=c(1,2))
hist(expdtgeo, main='original')
hist(expdtgeoFilt, main='filtered')

#t test
design <- model.matrix(~0 + factor(pData(eset)$type))
fit <- lmFit(eset, design)
fit <- eBayes(fit)
# Ambil statistik t untuk setiap gen
t_statistics <- fit$t
# Hitung p-value
p_values <- 2 * pt(-abs(t_statistics), df = fit$df.residual)
# Gabungkan statistik t dan p-value ke dalam satu data frame
results <- data.frame(T_Statistic = t_statistics, P_Value = p_values)
# Lihat hasilnya
head(results)


### Assign Group Code ###
vargrp <- phdtgeo
table(vargrp)

# t-test basal dan bukan basal
group <- ifelse(vargrp == "basal", 0, 1)
group
t.test(group, expdtgeoFilt[1, ])$p.value
```

```
boxplot(group, expdtgeoFilt[1, ])
pval <- apply(expdtgeoFilt, 1, function(x) t.test(group, x)$p.value)
dift <- apply(expdtgeoFilt, 1, function(x) diff(t.test(x[1:4],
x[5:8])$estimate))
plot(dift, -log10(pval))

# t-test HER dan bukan HER
group <- ifelse(vargrp == "HER", 0, 1)
group
t.test(group, expdtgeoFilt[1, ])$p.value
boxplot(group, expdtgeoFilt[1, ])
pval <- apply(expdtgeoFilt, 1, function(x) t.test(group, x)$p.value)
dift <- apply(expdtgeoFilt, 1, function(x) diff(t.test(x[1:4],
x[5:8])$estimate))
plot(dift, -log10(pval))

# t-test luminal_A dan bukan luminal_A
group <- ifelse(vargrp == "luminal_A", 0, 1)
group
t.test(group, expdtgeoFilt[1, ])$p.value
boxplot(group, expdtgeoFilt[1, ])
pval <- apply(expdtgeoFilt, 1, function(x) t.test(group, x)$p.value)
dift <- apply(expdtgeoFilt, 1, function(x) diff(t.test(x[1:4],
x[5:8])$estimate))
plot(dift, -log10(pval))

# t-test luminal_B dan bukan luminal_B
group <- ifelse(vargrp == "luminal_B", 0, 1)
group
t.test(group, expdtgeoFilt[1, ])$p.value
boxplot(group, expdtgeo[1, ])
pval <- apply(expdtgeoFilt, 1, function(x) t.test(group, x)$p.value)
dift <- apply(expdtgeoFilt, 1, function(x) diff(t.test(x[1:4],
x[5:8])$estimate))
plot(dift, -log10(pval))

### Limma Analysis ###
vargrp <- data[, 2]
col_cancer_idx <- ifelse(vargrp %in% c('basal', 'HER', 'luminal_A',
'luminal_B'),
                         0, 1)
col_cancer <- vargrp[col_cancer_idx == 0]
```

```
col_normal_idx <- ifelse(vargrp %in% c('normal', 'basal', 'HER',
'luminal_A', 'luminal_B'),
                         0, 1)
col_normal <- vargrp[col_normal_idx == 0]

expdtgeo2 = expdtgeoFilt[, col_cancer_idx==0]
expdtgeo3 = expdtgeoFilt[, col_normal_idx==0]

dim(expdtgeo2)
dim(expdtgeo3)

group_cancer <- ifelse(col_cancer == "basal", 0,
                       ifelse(col_cancer == "HER", 1,
                              ifelse(col_cancer == "luminal_A", 2,
3)))

group_cancer

#Limma
design <- model.matrix(~group_cancer)
fit <- eBayes(lmFit(expdtgeo2, design))
fit

volcanoplot(fit, coef=2)

#top Result
topResult <- topTable(fit, coef=2, number=20)
rownames(topResult)

selected <- rownames(expdtgeo2) %in% rownames(topResult)
expdtgeosel <- expdtgeo2[selected, ]
heatmap(expdtgeosel)

par(mfrow=c(2,2))
for (i in 1:4) plot(group_cancer, expdtgeosel[i,],
                    main = rownames(expdtgeosel)[i])

luminal_data = data[data[, 'type'] %in% c('luminal_A', 'luminal_B'),
                    c('samples', rownames(topResult))]
dim(luminal_data)

selected_luminal <- rownames(expdtgeo2) %in% rownames(topResult)
```

```
expdtgeosel_luminal <- expdtgeo2[selected_luminal,
as.character(luminal_data[, 'samples'])]
heatmap(expdtgeosel_luminal)

group_normal <- ifelse(col_normal == "normal", 0, 1)
group_normal
design <- model.matrix(~group_normal)
fit <- eBayes(lmFit(expdtgeo3, design))
fit
volcanoplot(fit, coef=2)

topResult_normal <- topTable(fit, coef=2, number=20)
rownames(topResult_normal)

selected <- rownames(expdtgeo3) %in% rownames(topResult_normal)
expdtgeosel <- expdtgeo3[selected, ]

heatmap(expdtgeosel)

par(mfrow=c(2,2))
for (i in 1:4) plot(group_normal, expdtgeosel[i,],
                    main = rownames(expdtgeosel)[i])
```