LAPORAN SAINS DATA GENOM

Tugas UTS

Ekspresi gen kanker payudara

Dosen pengampu
Prof. Setia Pramana, Ph.D.

Oleh

Tulus Setiawan 2006568802

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS INDONESIA
DEPOK

2023

PENDAHULUAN

1.1. Latar Belakang Masalah

Ekspresi gen diferensial adalah fenomena biologis yang mendasar dalam memahami bagaimana organisme baik manusia, hewan, maupun tumbuhan berfungsi dan bereaksi terhadap lingkungan. Proses ini melibatkan aktivasi atau nonaktifasi gen dalam sel, yang mengakibatkan perubahan dalam fungsi sel. Dalam konteks kesehatan manusia, penelitian tentang ekspresi gen diferensial telah membantu kita memahami bagaimana penyakit seperti kanker berkembang dan bagaimana kita dapat merancang terapi yang lebih efektif.

Kanker adalah penyakit kompleks yang melibatkan perubahan pada banyak gen dalam sel. Beberapa gen mungkin menjadi hiperaktif, sementara yang lain mungkin menjadi kurang aktif atau dinonaktifkan sama sekali. Memahami pola ekspresi gen ini membantu kita mengidentifikasi gen mana yang berperan dalam perkembangan kanker dan bagaimana mereka berinteraksi satu sama lain dan dengan lingkungan sel. Penelitian lebih lanjut diperlukan untuk memahami pola ekspresi gen ini dan bagaimana mereka dapat digunakan untuk merancang terapi kanker yang lebih efektif.

Beberapa subtipe kanker yang terdapat pada kanker payudara diantaranya, Basal, HER, Luminal A, dan Luminal B. Basal sendiri memiliki sifat yang cenderung agresif dan memiliki prognosis yang lebih buruk. Kemudian untuk HER memiliki sifat yang terkadang memiliki pertumbuhan cepat dan memerlukan perawatan khusus. Selain itu, Luminal A memiliki prognosis baik, dan sifat yang pertumbuhannya cenderung lambat. Sedangkan Luminal B memiliki sifat pertumbuhan yang lebih cepat daripada Luminal A

Semua tipe ini mempengaruhi pilihan pengobatan dan prognosis pasien dengan kanker payudara. Penelitian terus dilakukan untuk memahami lebih lanjut peran gen-gen ini dalam perkembangan kanker payudara.

1.2. Rumusan Masalah

1. Manakah gen yang berekpresi berbeda antara empat subtipe kanker, yaitu Luminal A, Luminal B, Basal, dan HER?

2. Bagaimana perbedaan ekspresi gen antara sel sehat dan sel kanker?

1.3. Tujuan Penulisan

1. Mengidentifikasi gen yang berekpresi berbeda antara empat subtipe kanker, yaitu Luminal A, Luminal B, Basal, dan HER.
2. Menginvestigasi perbedaan ekspresi gen antara sel sehat dan sel kanker.

METODE

2.1. Gene Filtering

Metode penyaringan gen atau *gene filtering* adalah teknik yang bertujuan untuk mengurangi noise dan dimensionalitas data ekspresi gen dengan cara menghilangkan gen yang tidak informatif atau tidak relevan untuk analisis. Metode penyaringan gen dapat meningkatkan kekuatan dan akurasi dalam mendeteksi gen yang diekspresikan secara berbeda, mengelompokkan jenis sel, atau mengidentifikasi biomarker dan terapi tertarget. Ada beberapa pendekatan yang berbeda untuk metode penyaringan gen, seperti menggunakan ambang batas tetap atau adaptif untuk rata-rata atau varian ekspresi gen, menggunakan analisis komponen utama, atau menggunakan algoritme pengoptimalan. Metode penyaringan gen dapat diterapkan pada berbagai jenis data ekspresi gen, seperti microarray, RT-PCR, atau RNA-seq. Penelitian lebih lanjut terus dilakukan untuk memperbaiki dan mengoptimalkan teknik ini dalam analisis biologis.

2.2. Differential Expression

Metode ekspresi diferensial dalam ekspresi gen adalah teknik yang bertujuan untuk mengidentifikasi gen yang menunjukkan perubahan tingkat ekspresi antara kondisi atau kelompok yang berbeda, seperti sehat versus sakit, diobati versus tidak diobati, atau tahap perkembangan yang berbeda. Metode ekspresi diferensial dapat membantu untuk menemukan mekanisme biologis dan jalur yang terlibat dalam variasi fenotipik atau respon terhadap rangsangan. Metode ekspresi diferensial juga dapat membantu menemukan biomarker potensial atau target terapeutik untuk penyakit.

2.2.1. T-test

t-test adalah metode statistik yang digunakan untuk membandingkan rata-rata dua kelompok sampel yang independen. t-test mengasumsikan bahwa data mengikuti

distribusi normal dan memiliki varians yang sama. t-test menghasilkan nilai p yang menunjukkan probabilitas bahwa perbedaan rata-rata yang diamati terjadi secara kebetulan atau karena faktor acak. Nilai p yang kecil menunjukkan bahwa perbedaan tersebut bermakna secara statistik.

Untuk melakukan t test pada data gen, kita perlu melakukan langkah-langkah berikut:

- Mengelompokkan data gen berdasarkan kondisi, misalnya kelompok sehat dan kelompok kanker.
- Menghitung rata-rata dan standar deviasi tingkat ekspresi untuk setiap gen di setiap kelompok.
- Menghitung nilai t dengan rumus berikut:

  $$t = \frac{\hat{x}_1 - \hat{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

  dimana, $\hat{x}_1$ dan $\hat{x}_2$ adalah rata-rata sampel dari masing-masing kelompok, s_p adalah standar deviasi gabungan dari kedua kelompok, serta n_1 dan n_2 adalah ukuran sampel dari masing-masing kelompok.

Setelah mendapatkan nilai t, kita perlu menentukan nilai kritis t yang sesuai dengan tingkat signifikansi dan derajat kebebasan yang kita pilih. Jika nilai t absolut lebih besar dari nilai kritis t, maka kita menolak hipotesis nol dan menerima hipotesis alternatif. Jika tidak, kita gagal menolak hipotesis nol.

2.2.2. Limma

Metode differential expression gen dengan limma adalah metode yang menggunakan model linear untuk mengestimasi perbedaan ekspresi gen antara dua atau lebih kelompok perlakuan. Metode ini memiliki beberapa keunggulan, seperti:

1. Dapat menangani data dengan variasi heterogen, misalnya akibat faktor teknis atau biologis.
2. Dapat menguji hipotesis kompleks, seperti interaksi antara perlakuan dan kondisi eksperimental.
3. Dapat mengontrol tingkat kesalahan tipe I dengan menggunakan koreksi multiple testing, seperti Benjamini-Hochberg atau Bonferroni.

4. Dapat menghasilkan estimasi interval kepercayaan dan p-value yang lebih akurat daripada metode lain, seperti t-test atau ANOVA.

Untuk menggunakan metode limma, langkah-langkah yang perlu dilakukan adalah:

1. Melakukan normalisasi data ekspresi gen untuk menghilangkan bias sistematis dan meningkatkan kualitas data.
2. Membuat matriks desain yang merepresentasikan perlakuan yang diberikan pada setiap sampel.
3. Membuat matriks kontras yang merepresentasikan perbandingan yang ingin diuji, misalnya normal vs basal.
4. Membangun model linear untuk setiap gen.
5. Mengestimasi koefisien dan statistik uji untuk setiap kontras.
6. Memilih gen yang secara signifikan berbeda secara ekspresi dengan menggunakan kriteria tertentu, misalnya p-value < 0.05 atau log fold change > 2.

Hasil

Dataset bersumber dari Kaggle, dengan judul dataset Breast cancer gene expression (CuMiDa).

3.1. Tahapan Load Dataset

Di tahap ini, dilakukan load dataset sehingga didapat informasi umum dari data, yaitu data terdiri dari 151 sampel, 54675 gen, dan 6 kelas. Tetapi, karena hanya akan menggunakan sampel dengan tipe normal, basal, HER, luminal_A, dan luminal_B maka dilakukan subset dari dari keseluruhan data sehingga hanya menggunakan 137 sampel.

A data.frame: 6 × 54677

samples	type	X1007_s_at	X1053_at	X117_at	X121_at	X1255_g_at	X1294_at	X1316_at	X1320_at	⋯	AFFX.r2.Ec.bioD.3_at	AFFX.r2.Ec.bioD.5_at	AFFX.r2.
<int>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	⋯	<dbl>	<dbl>	
229	luminal_B	10.35583	7.158921	8.204383	7.055268	3.127653	7.647573	5.406018	4.842870	⋯	12.57875	12.08395	
230	luminal_B	10.39264	7.334408	6.848586	7.020486	3.228065	7.815439	5.448470	4.496955	⋯	12.63856	12.12213	
233	luminal_B	10.93088	8.415294	5.906827	7.753572	3.270557	7.367931	5.906849	5.194349	⋯	12.52351	11.97797	
236	luminal_B	11.02710	7.180876	6.304736	7.641197	3.206950	8.569296	5.823146	4.617309	⋯	12.25677	11.66113	
237	luminal_B	10.44439	7.525153	5.964460	7.825939	3.384147	7.268454	5.245072	5.088004	⋯	12.32190	11.72769	
238	luminal_B	11.34582	7.379299	5.891172	7.394586	3.183420	7.792885	5.355978	4.457914	⋯	12.12611	11.47889	

Gambar 1: Tampilan contoh 6 baris dari data

3.2. Tahapan Eksplorasi Data

Dataset memiliki data gen yang cukup besar, oleh karena itu dilakukan proses sampling dengan mengambil sebanyak 50% dari total data gen supaya tidak memakan waktu komputasi yang lama. Proses sampling menggunakan random seed 2006568802, yaitu NPM dari penulis. Diperoleh gen sebanyak 27338 setelah proses sampling.

Kemudian, dilakukan proses normalisasi data menggunakan penyetaraan log2. Tujuan dilakukan normalisasi data adalah untuk komponen tahap awal proses data untuk membuat data dapat dibandingkan dalam rentang yang sama.

Menggunakan normalisasi data, didapat data sesudah dan sebelum normalisasi memiliki nilai minimum dan maksimum terlampir dalam output berikut:

```
Nilai maksimal-minimal data sebelum normalisasi
[1]  2.1711 14.9701
```

Gambar 2: tampilan nilai minimum dan maksimum sebelum dinormalisasi

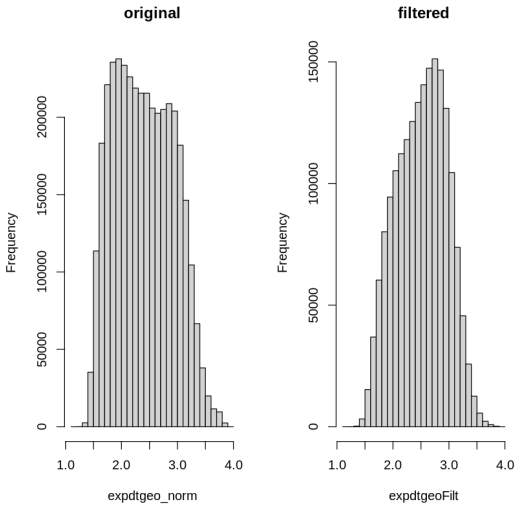
```
Nilai maksimal-minimal data setelah normalisasi
[1] 1.118426 3.904012
```

Gambar 3: tampilan nilai minimum dan maksimum setelah dinormalisasi

3.2. Tahapan Gene Filtering

Pada tahapan penyaringan gen atau *gene filtering* tidak dapat menggunakan modul atau *library genefilter* secara langsung pada R. Sehingga, pada tahap ini hanya dilakukan filtering pada gen-gen yang memiliki variasi yang rendah di antara sampel-sampel, dengan cara menghapus gen yang memiliki IQR lebih rendah dari median IQR seluruh gen. Proses ini menyisakan gen sebanyak 13669 gen.

Berikut plot histogram sebelum dan sesudah tahapan gene filtering



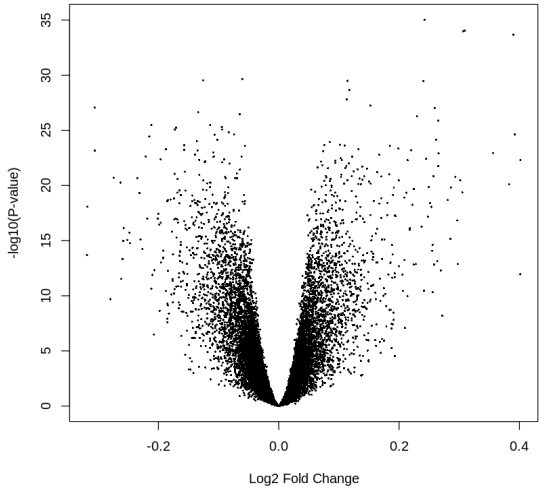
Gambar 4: tampilan plot histogram yang memperlihatkan sebaran data sebelum dan sesudah tahapan gene filtering

3.3. Tahapan Differential Expression

3.3.1. Differential Expression pada 4 Subtipe Kanker

Pada tahap ini dilakukan differential gene expression pada 4 subtipe kanker, yaitu basal, HER, luminal_A, dan luminal_B.

Setelah dilakukannya proses limma, menggunakan Volcano plot didapat tampilan signifikansi statistik (-log10 P value) sebagai sumbu y, dan besarnya perubahan log2 fold yang berguna untuk memvisualisasikan gen yang diekspresikan secara berbeda.



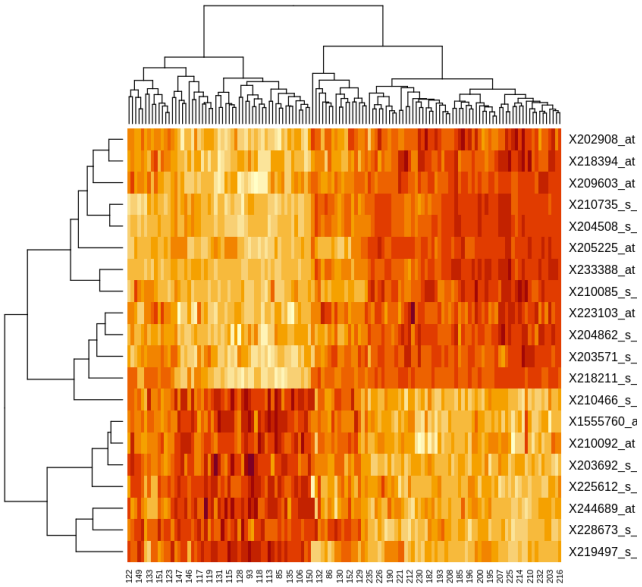
Gambar 5: tampilan volcano plot menggunakan metode limma dengan coef =2

Menggunakan parameter koefisien 2 dan display 20 gen tertinggi, metode limma akan menghasilkan nilai statistik (F) dan p-value dari masing-masing gen. Berikut 20 ekspresi gen yang berbeda diantara 4 subtipe kanker, yaitu basal, HER, luminal_A, dan luminal_B, antara lain:

```
'X210085_s_at' · 'X233388_at' · 'X204508_s_at' · 'X205225_at' · 'X210466_s_at' · 'X228673_s_at' · 'X204862_s_at' · 'X218211_s_at' · 'X218394_at' · 'X202908_at' · 'X223103_at' ·
'X219497_s_at' · 'X210735_s_at' · 'X244689_at' · 'X1555760_a_at' · 'X203571_s_at' · 'X209603_at' · 'X210092_at' · 'X225612_s_at' · 'X203692_s_at'
```

Gambar 6: tampilan 20 gen teratas

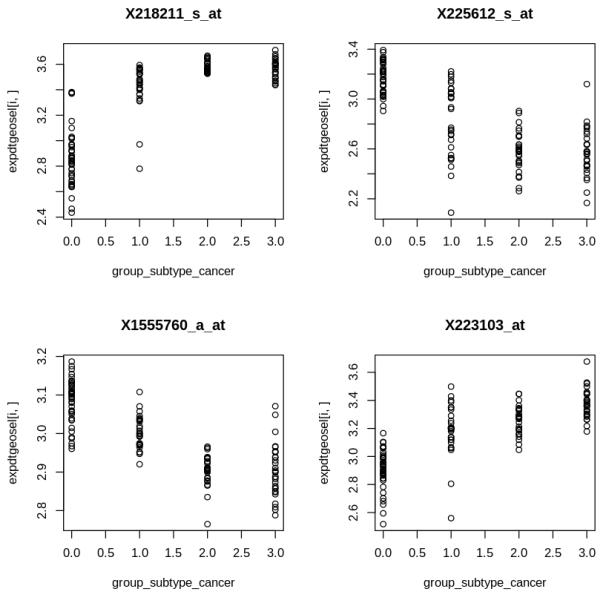
Dengan menggunakan *plot heatmap*, diperoleh pola ekspresi 20 gen tertinggi adalah sebagai berikut.



Gambar 7: tampilan grafik heatmap 20 gen tertinggi

Heatmap tersebut menggambarkan ekspresi dari tiap gen (baris) dan tiap sampel (kolom). Warna merah menandakan gen tersebut memiliki ekspresi lebih tinggi (over expression). Namun, perbedaan tidak cukup terlihat jelas, sehingga perlu ditinjau lebih lanjut.

Dengan menggunakan scatter plot pada pola ekspresi 4 gen tertinggi, terlihat jelas perbedaan antara keempat, yaitu sebagai berikut.



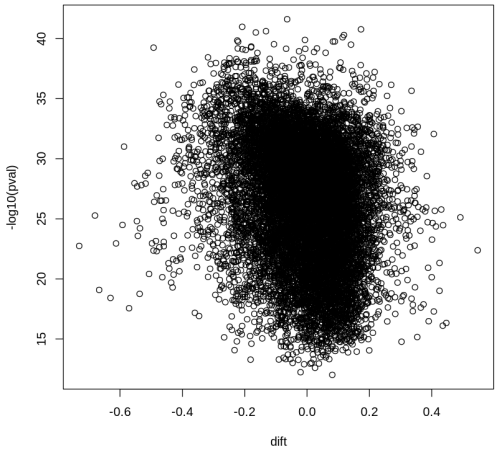
Gambar 8: tampilan grafik scatterplot pada 4 gen tertinggi

Dengan keterangan untuk axis x pada scatter plot di atas adalah 0.0 untuk basal, 1.0 untuk HER, 2.0 untuk luminal_A, dan 3.0 untuk luminal_B.

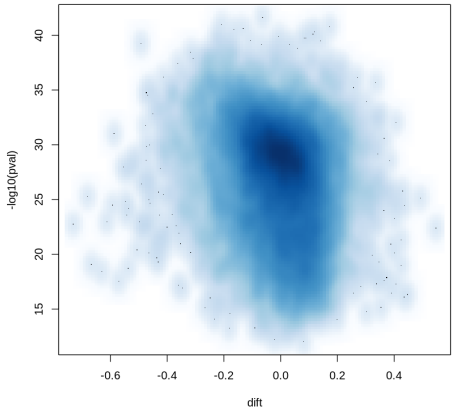
3.3.2. Differential Expression pada Sel Sehat dan Sel Kanker

Pada tahap ini dilakukan differential gene expression pada dua grup, yaitu grup normal (sehat) dan grup sel kanker. Grup sel kanker adalah gabungan dari grup basal, HER, luminal_A, dan luminal_B. Akan ditinjau juga differential gene expression antara grup normal dengan masing-masing dari keempat subtipe kanker.

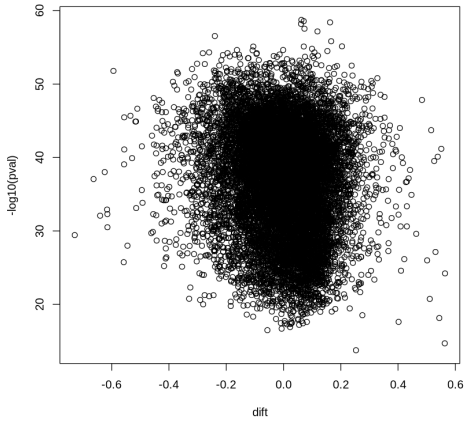
Menggunakan uji t-test, didapat perbedaan rata-rata dari grup normal dan tiap tipe kanker, sumbu y menunjukkan -log10 dari p-value.



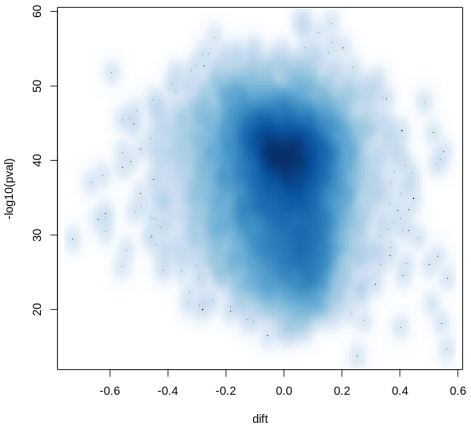
Gambar 9: tampilan plot t-test dengan standard p-value untuk grup normal dan HER



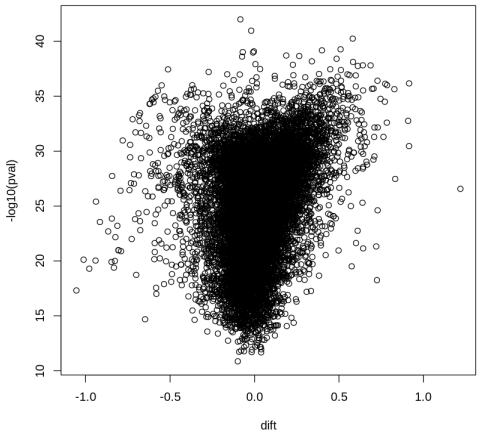
Gambar 10: tampilan plot t-test dengan adjusted p-value metode bonferonni untuk grup normal dan HER



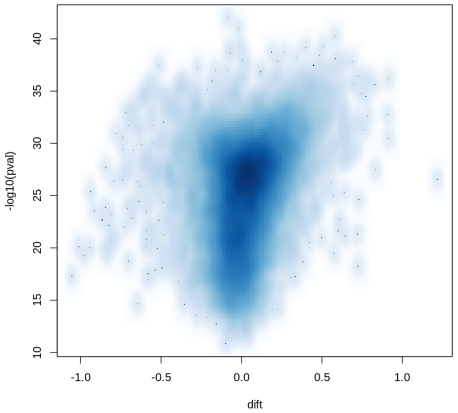
Gambar 11: tampilan plot t-test dengan standard p-value untuk grup normal dan basal



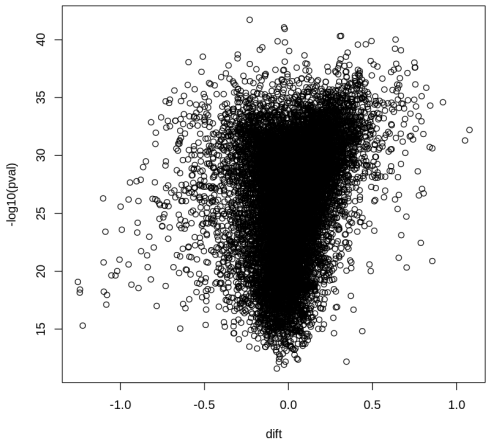
Gambar 12: tampilan plot t-test dengan adjusted p-value metode bonferonni untuk grup normal dan basal



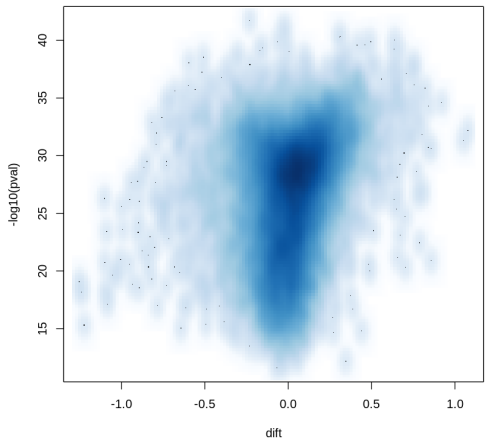
Gambar 13: tampilan plot t-test dengan standard p-value untuk grup normal dan luminal_A



Gambar 14: tampilan plot t-test dengan adjusted p-value metode bonferonni untuk grup normal dan luminal_A



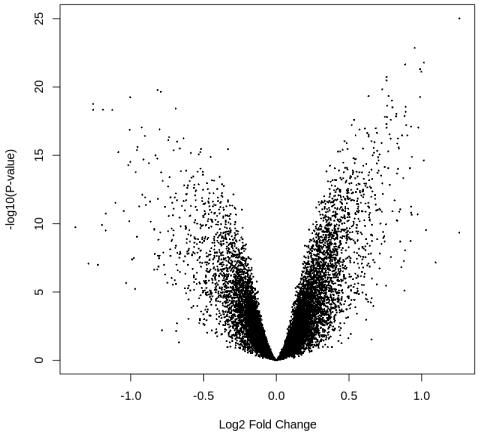
Gambar 15: tampilan plot t-test dengan standard p-value untuk grup normal dan luminal_B



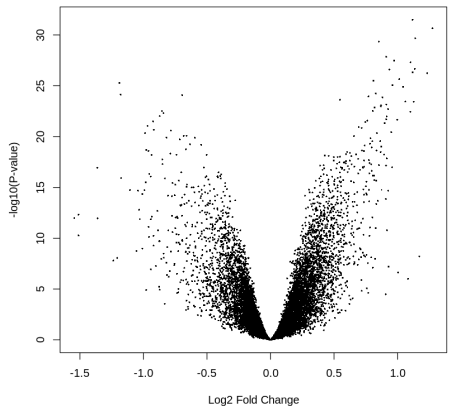
Gambar 16: tampilan plot t-test dengan adjusted p-value metode bonferonni untuk grup normal dan luminal_B

Titik yang berada di sekitar daerah 0 pada sumbu x memiliki interpretasi bahwa tidak ada perbedaan yang signifikan antara gen dari grup normal dan kanker. Apabila titik semakin ke atas maka tingkat signifikansi semakin tinggi.

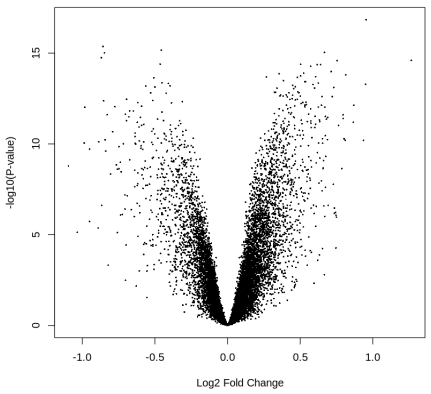
Menggunakan Volcano plot dengan metode limma, didapat tampilan signifikansi statistik (-log10 P value) sebagai sumbu y, dan besarnya perubahan log2 fold yang berguna untuk memvisualisasikan gen yang diekspresikan secara berbeda.



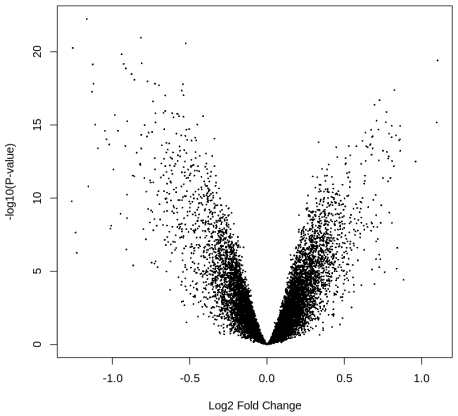
Gambar 17: tampilan volcano plot menggunakan metode limma untuk grup normal dan HER



Gambar 18: tampilan volcano plot menggunakan metode limma untuk grup normal dan basal

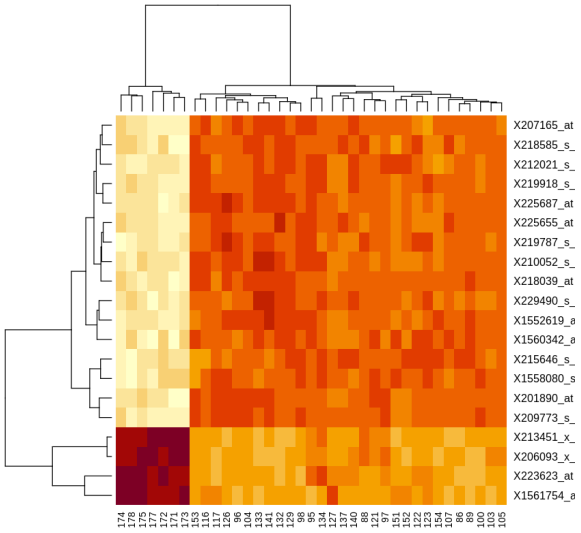


Gambar 19: tampilan volcano plot menggunakan metode limma untuk grup normal dan luminal_A

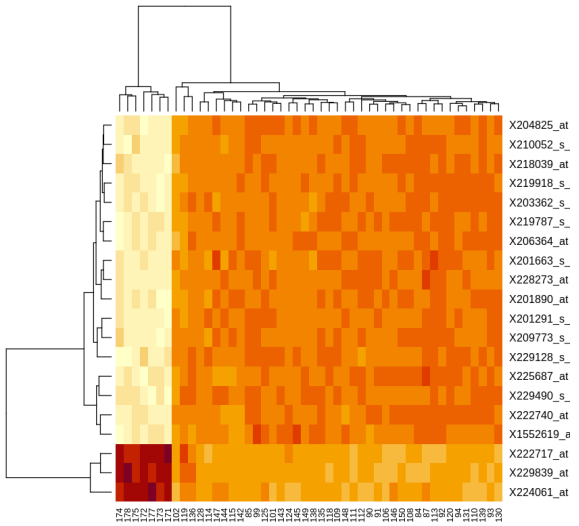


Gambar 20: tampilan volcano plot menggunakan metode limma untuk grup normal dan luminal_B

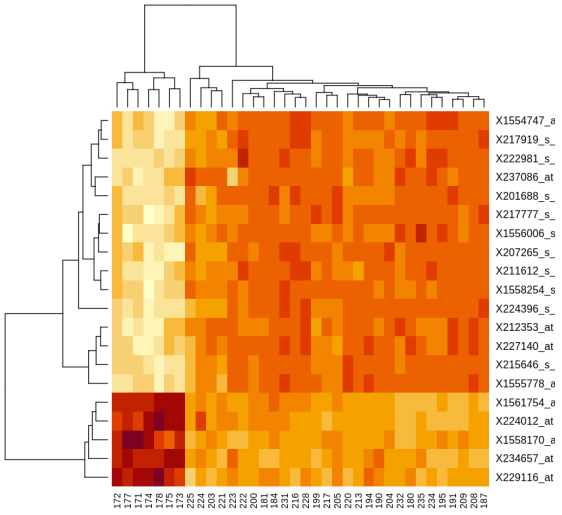
Adapun dengan menggunakan parameter koefisien 2, yang mewakili perubahan log-fold dari variabel yang sesuai dalam model, didapat hasil pengujian grup sehat dan kanker yang jelas terpisah. Dengan menggunakan plot heatmap, diperoleh pola ekspresi 20 gen tertinggi adalah sebagai berikut.



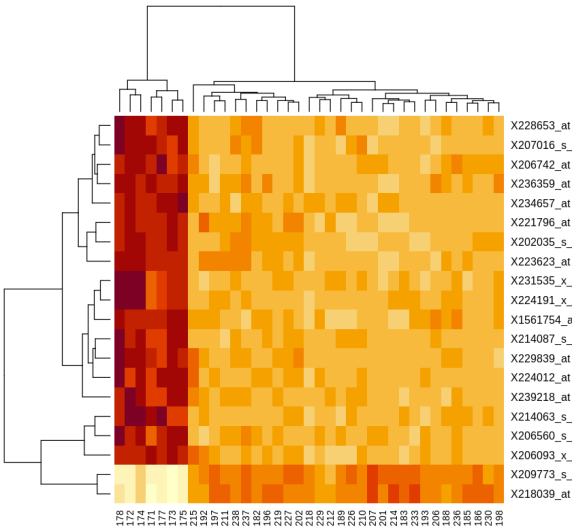
Gambar 21: tampilan grafik heatmap 20 gen tertinggi untuk grup normal dan HER



Gambar 22: tampilan grafik heatmap 20 gen tertinggi untuk grup normal dan basal



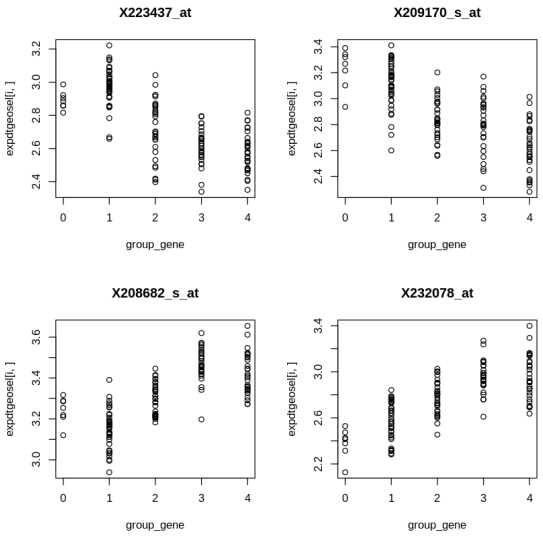
Gambar 23: tampilan grafik heatmap 20 gen tertinggi untuk grup normal dan luminal_A



Gambar 24: tampilan grafik heatmap 20 gen tertinggi untuk grup normal dan luminal_B

Heatmap tersebut menggambarkan ekspresi dari tiap gen (baris) dan tiap sampel (kolom). Warna merah menandakan gen tersebut memiliki ekspresi lebih tinggi (over expression). Terlihat pengelompokan yang cukup jelas antara grup normal dan masing-masing dari subtipe sel kanker. 20 gen dengan perbedaan ekspresi tertinggi berbeda-beda antara sel sehat dengan masing-masing subtipe sel kanker. Nama gen dapat dilihat pada sumbu-y heatmap.

Dengan menggunakan scatter plot pada pola ekspresi 4 gen tertinggi, terlihat perbedaan antara grup normal dan kanker, yaitu sebagai berikut.



Gambar 25: Scatter plot pada 4 gen tertinggi antara grup normal dan kanker

Dengan keterangan untuk axis x pada scatter plot di atas adalah 0 untuk normal, 1 untuk basal, 2 untuk HER, 3 untuk luminal_A, dan 4 untuk luminal_B.

Kesimpulan

Berdasarkan analisis yang telah dilakukan, berikut adalah kesimpulan yang dapat diambil:

1. Melalui analisis ekspresi gen menggunakan metode limma, berhasil diidentifikasi 20 gen yang memiliki perbedaan signifikan dalam ekspresi antara empat subtipe kanker: Luminal A, Luminal B, Basal, dan HER. 20 gen yang memiliki perbedaan ekspresi tertinggi adalah X210085_s_at, X233388_at, X204508_s_at, X205225_at, X210466_s_at, X228673_s_at, X204862_s_at, X218211_s_at, X218394_at, X202908_at, X223103_at, X219497_s_at, X210735_s_at, X244689_at, X1555760_a_at, X203571_s_at, X209603_at, X210092_at, X225612_s_at, X203692_s_at. Gen-gen ini dapat menjadi calon biomarker untuk membedakan subtipe kanker tersebut. Hasil analisis juga menunjukkan adanya perbedaan ekspresi yang konsisten antara subtipe kanker, yang dapat membantu dalam pemahaman lebih lanjut tentang perbedaan biologis di antara mereka. 2. Hasil analisis menunjukkan bahwa terdapat perbedaan ekspresi gen yang signifikan antara sel sehat dan sel kanker. Gen-gen yang diekspresikan secara berbeda antara kedua kelompok ini dapat berpotensi sebagai biomarker atau target potensial untuk terapi kanker. Analisis ini juga menunjukkan bahwa pengelompokan sel sehat dan sel kanker dapat dibedakan dengan jelas berdasarkan pola ekspresi gen. Hasil analisis ini dapat menjadi landasan untuk penelitian

lebih lanjut dalam pengembangan terapi yang lebih tepat sasaran dan pemahaman lebih dalam tentang peran gen dalam perkembangan kanker

Lampiran

Berikut adalah lampiran kode R yang digunakan, lengkap dengan penjelasannya sesuai alur. https://colab.research.google.com/drive/1K73YEKKCwvXMzGIRhs3t6ZUxBP7gY18a?usp=sharing