181 – Widya Arum Prastanti

DASAR ANALISIS DIFFERENTIALLY EXPRESSED GENES (DEGs) MENGGUNAKAN BAHASA R PADA DATASET KANKER PAYUDARA GSE15852

PENDAHULUAN

Kanker payudara merupakan salah satu jenis kanker dengan angka kejadian tertinggi pada perempuan di seluruh dunia dan menjadi penyebab utama kematian akibat kanker pada wanita. Penyakit ini berkembang akibat akumulasi perubahan molekuler yang memengaruhi regulasi proliferasi sel, mekanisme apoptosis, diferensiasi, serta jalur pensinyalan yang mengatur pertumbuhan jaringan. Perubahan ekspresi gen memainkan peran sentral dalam proses inisiasi dan progresi kanker payudara. Perkembangan teknologi high-throughput genomics seperti microarray memungkinkan analisis ekspresi ribuan gen secara simultan. Pendekatan ini memungkinkan identifikasi Differentially Expressed Genes (DEGs), yaitu gen-gen yang mengalami perubahan ekspresi signifikan antara jaringan kanker dan jaringan normal. Identifikasi DEGs memberikan wawasan penting mengenai mekanisme molekuler yang mendasari patogenesis kanker serta berpotensi mengungkap biomarker diagnostik maupun target terapi.

Dalam analisis bioinformatika, metode statistik limma (Linear Models for Microarray Data) banyak digunakan untuk menganalisis data microarray karena kemampuannya dalam memodelkan variabilitas biologis dan melakukan koreksi multiple testing menggunakan pendekatan False Discovery Rate (FDR). Selain identifikasi gen yang mengalami upregulation dan downregulation, interpretasi biologis lebih lanjut dilakukan melalui analisis functional enrichment, termasuk Gene Ontology (GO) dan Kyoto Encyclopedia of Genes and Genomes (KEGG), untuk memahami keterlibatan gen dalam proses biologis dan jalur molekuler tertentu. Pada studi ini, dilakukan analisis ekspresi gen menggunakan dataset publik GSE15852 dari Gene Expression Omnibus (GEO) untuk mengidentifikasi gen yang mengalami perubahan ekspresi signifikan pada kanker payudara dibandingkan jaringan normal. Analisis dilanjutkan dengan visualisasi hasil serta enrichment analysis untuk mengevaluasi implikasi biologis dari gen-gen yang teridentifikasi sebagai DEGs.

METODE

2.1 Sumber Data

Dataset yang digunakan dalam penelitian ini adalah GSE15852, yang diperoleh dari database publik Gene Expression Omnibus (GEO). Dataset ini merupakan data ekspresi gen berbasis microarray yang membandingkan jaringan kanker payudara dengan jaringan normal. Data diunduh secara langsung menggunakan package GEOquery dalam lingkungan R.

2.2 Lingkungan Analisis

Seluruh analisis dilakukan menggunakan perangkat lunak R dan RStudio. Package yang digunakan meliputi:

- GEOquery untuk pengambilan data dari GEO
- limma untuk analisis Differential Gene Expression
- dplyr untuk manipulasi data
- ggplot2 untuk visualisasi
- pheatmap untuk pembuatan heatmap
- umap untuk reduksi dimensi
- clusterProfiler dan org.Hs.eg.db untuk analisis enrichment (GO dan KEGG)

2.3 Preprocessing Data

Data ekspresi gen diekstraksi dari objek ExpressionSet menggunakan fungsi exprs(). Distribusi nilai ekspresi diperiksa menggunakan boxplot dan density plot untuk mengevaluasi kualitas data serta memastikan konsistensi antar sampel. Transformasi log2 dilakukan apabila diperlukan untuk:

1. Menstabilkan varians
2. Memenuhi asumsi model linear
3. Memudahkan interpretasi log fold change

2.4 Analisis Differential Expression Gene

Kelompok sampel (kanker dan normal) didefinisikan berdasarkan metadata pada dataset. Design matrix dibuat menggunakan fungsi model.matrix(~0 + group) untuk memodelkan perbedaan ekspresi antar kelompok tanpa intercept. Analisis dilakukan menggunakan metode limma dengan tahapan:

1. lmFit() untuk membangun model linear
2. makeContrasts() untuk mendefinisikan perbandingan biologis
3. contrasts.fit() untuk menerapkan kontras
4. eBayes() untuk stabilisasi varians menggunakan pendekatan empirical Bayes
5. topTable() untuk memperoleh gen yang signifikan

Kriteria gen signifikan ditentukan berdasarkan:

- Adjusted p-value (FDR) < 0.05

  |logFC| > 1

Gen dengan logFC positif dikategorikan sebagai upregulated, sedangkan gen dengan logFC negatif dikategorikan sebagai downregulated.

2.5 Visualisasi Data

Beberapa visualisasi dilakukan untuk mendukung interpretasi hasil:

- Boxplot untuk mengevaluasi distribusi ekspresi antar sampel
- Density plot untuk melihat sebaran global nilai ekspresi
- UMAP plot untuk reduksi dimensi dan melihat pemisahan sampel
- Volcano plot untuk menampilkan hubungan antara log fold change dan signifikansi statistik
- Heatmap untuk menampilkan pola ekspresi 50 gen paling signifikan

2.6 Functional Enrichment Analysis

Gen signifikan dikonversi dari gene symbol menjadi Entrez ID menggunakan package org.Hs.eg.db. Analisis enrichment dilakukan menggunakan package clusterProfiler untuk:

- Gene Ontology (GO) pada kategori Biological Process (BP)
- KEGG pathway analysis

Signifikansi enrichment ditentukan berdasarkan adjusted p-value < 0.05. Hasil enrichment divisualisasikan menggunakan dotplot untuk menampilkan kategori biologis yang paling signifikan.

HASIL DAN INTERPRETASI

3.1 Identifikasi Differentially Expressed Genes (DEGs)

Analisis differential gene expression dilakukan menggunakan metode linear modeling melalui package limma dengan kriteria signifikansi adjusted p-value (FDR) < 0.05 dan |logFC| > 1. Berdasarkan hasil analisis pada dataset GSE15852, diperoleh sejumlah gen yang menunjukkan perbedaan ekspresi signifikan antara jaringan kanker payudara dan jaringan normal. Nilai adjusted p-value yang sangat kecil (hingga orde 10^{-14}) menunjukkan tingkat signifikansi statistik yang sangat tinggi serta konsistensi perbedaan ekspresi antar kelompok sampel.

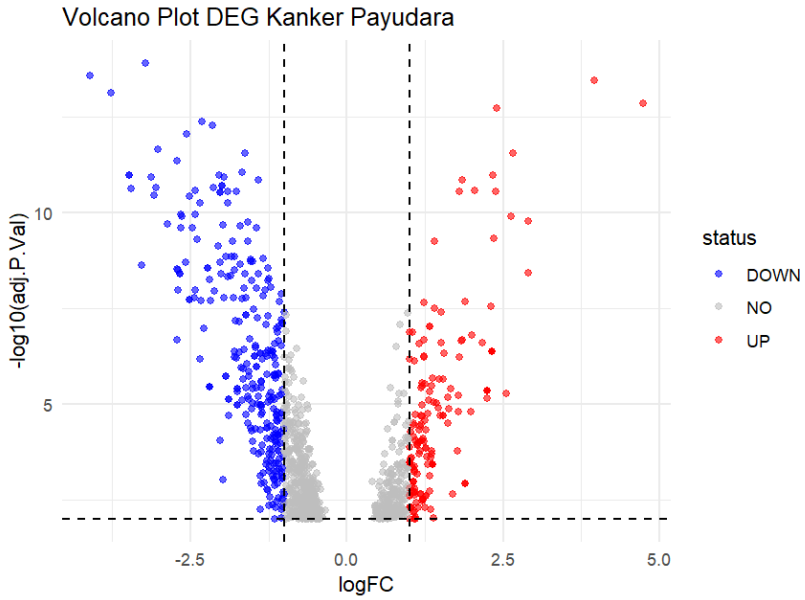
Beberapa gen yang mengalami peningkatan ekspresi paling signifikan (upregulated genes) antara lain KRT19 (logFC ≈ 4.74), CD24 (logFC ≈ 3.96), EPCAM (logFC ≈ 2.39), TACSTD2, dan KRT18. Gen-gen tersebut dikenal berperan dalam adhesi sel epitel, struktur sitoskeleton, serta karakteristik molekuler sel tumor. Nilai logFC yang tinggi

menunjukkan bahwa perubahan ekspresi tidak hanya signifikan secara statistik, tetapi juga bermakna secara biologis. Peningkatan ekspresi gen epitel seperti KRT19 dan EPCAM mencerminkan aktivasi jalur proliferatif dan peningkatan sifat invasif yang umum ditemukan pada kanker payudara.

Sebaliknya, sejumlah gen menunjukkan penurunan ekspresi signifikan (downregulated genes), di antaranya PPP1R1A (logFC ≈ −3.21), RBP4 (logFC ≈ −4.10), PDE3B, ACACB, ANGPTL4, serta beberapa gen yang terlibat dalam metabolisme lipid seperti LPL dan ADIPOQ. Sebagian besar gen tersebut berkaitan dengan regulasi metabolisme lipid dan fungsi jaringan adiposa. Penurunan ekspresi gen-gen metabolik ini mengindikasikan adanya perubahan metabolisme sel (metabolic reprogramming), yang merupakan salah satu karakteristik utama sel kanker dalam mendukung pertumbuhan dan kelangsungan hidup tumor.

Secara keseluruhan, distribusi nilai logFC dan adjusted p-value menunjukkan pola perubahan ekspresi gen yang sistematis dan terstruktur. Terlihat adanya aktivasi gen-gen yang mendukung proliferasi dan karakteristik sel tumor, disertai dengan supresi gen-gen yang berperan dalam fungsi fisiologis jaringan normal. Pola ini mencerminkan perubahan molekuler kompleks pada kanker payudara serta menunjukkan bahwa dataset GSE15852 memiliki kekuatan diskriminatif yang baik dalam membedakan profil ekspresi antara jaringan kanker dan jaringan normal.

3.2 Volcano Plot Differentially Expressed Genes



Gambar 3.2 Volcano Plot Differentially Expressed Genes pada Kanker Payudara

Hasil analisis differential expression gene divisualisasikan menggunakan volcano plot untuk menggambarkan hubungan antara besar perubahan ekspresi gen (log fold change/logFC) dan tingkat signifikansi statistik (-log10 adjusted p-value). Visualisasi ini memungkinkan identifikasi gen yang mengalami perubahan ekspresi signifikan sekaligus memiliki magnitude perubahan yang besar antara jaringan kanker payudara dan jaringan normal.

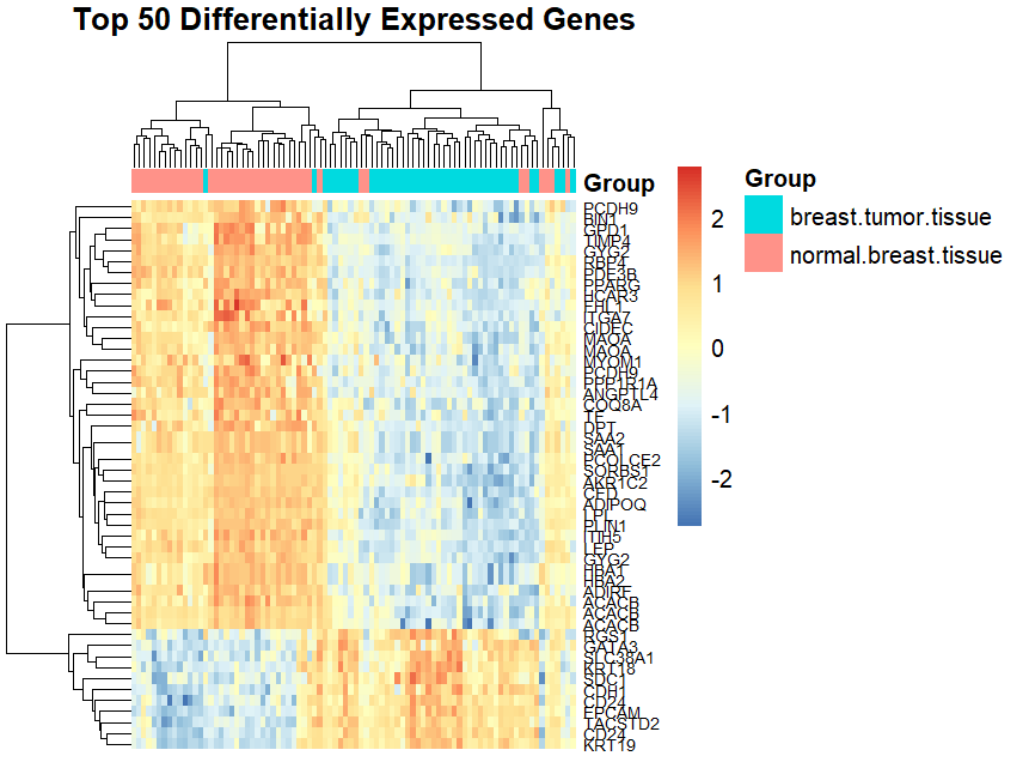
Pada plot terlihat tiga kelompok utama gen, yaitu gen yang mengalami upregulation (ditandai warna merah), gen yang mengalami downregulation (warna biru), dan gen yang tidak signifikan (warna abu-abu). Garis vertikal putus-putus menunjukkan batas $|logFC| > 1$, sedangkan garis horizontal menunjukkan batas signifikansi statistik dengan adjusted p-value < 0.05. Gen yang berada di luar kedua batas tersebut dikategorikan sebagai Differentially Expressed Genes (DEGs).

Distribusi titik pada sisi kanan plot menunjukkan sejumlah gen dengan nilai logFC positif tinggi, bahkan mencapai lebih dari 4, disertai nilai -log10(adj.P.Val) yang juga tinggi. Hal ini menunjukkan adanya peningkatan ekspresi gen yang sangat signifikan secara statistik pada jaringan kanker payudara dibandingkan jaringan normal. Secara biologis, peningkatan ekspresi ini umumnya berkaitan dengan aktivasi jalur proliferasi sel, peningkatan adhesi sel epitel, serta karakteristik molekuler khas sel tumor.

Sebaliknya, pada sisi kiri plot terlihat banyak gen dengan logFC negatif yang signifikan, menunjukkan terjadinya penurunan ekspresi pada jaringan kanker. Beberapa gen memiliki nilai -log10(adj.P.Val) yang tinggi, menandakan bahwa perbedaan ekspresi tersebut sangat kuat dan konsisten secara statistik. Pola ini mengindikasikan adanya supresi terhadap fungsi biologis tertentu, yang kemungkinan berkaitan dengan metabolisme normal jaringan dan fungsi diferensiasi sel.

Secara keseluruhan, volcano plot menunjukkan bahwa perubahan ekspresi gen pada dataset ini bersifat sistematis dan memiliki kekuatan sinyal biologis yang tinggi. Banyak gen yang melewati ambang batas signifikansi dan magnitude perubahan, sehingga mendukung validitas analisis differential expression yang dilakukan. Pola ini menegaskan bahwa kanker payudara dalam dataset GSE memiliki perbedaan molekuler yang jelas dibandingkan jaringan normal.

3.3 Heatmap 50 Differentially Expressed Genes (DEGs) Teratas



Gambar 3. Heatmap 50 Differentially Expressed Genes (DEGs) Teratas antara jaringan kanker payudara dan jaringan normal

Dalam rangka memperjelas pola ekspresi gen yang paling signifikan, dilakukan visualisasi terhadap 50 Differentially Expressed Genes (DEGs) teratas menggunakan heatmap. Gen-gen ini dipilih berdasarkan nilai adjusted p-value paling signifikan dan magnitude perubahan ekspresi (|logFC|) tertinggi.

Heatmap menampilkan pola ekspresi relatif (z-score) dari masing-masing gen pada seluruh sampel, yang telah dikelompokkan berdasarkan jenis jaringan, yaitu breast tumor tissue dan normal breast tissue. Skala warna menunjukkan tingkat ekspresi gen, di mana warna merah merepresentasikan ekspresi tinggi (upregulation), warna biru menunjukkan ekspresi rendah (downregulation), dan warna kuning menunjukkan ekspresi sedang.

Berdasarkan visualisasi, terlihat adanya pemisahan yang jelas antara sampel jaringan tumor dan jaringan normal pada dendrogram bagian atas (cluster kolom). Sebagian besar sampel tumor membentuk satu klaster utama yang terpisah dari sampel normal, menunjukkan bahwa 50 gen teratas ini memiliki kemampuan diskriminatif yang kuat dalam membedakan kedua kondisi biologis tersebut.

Pada bagian atas heatmap terlihat sekelompok gen yang menunjukkan ekspresi tinggi pada jaringan normal namun rendah pada jaringan tumor (ditandai warna kuning-oranye pada normal dan biru pada tumor). Pola ini menunjukkan adanya gen-gen yang mengalami

downregulation pada kanker payudara, yang kemungkinan berperan dalam fungsi fisiologis normal jaringan seperti regulasi metabolisme, diferensiasi sel, atau homeostasis jaringan. Sebaliknya, pada bagian bawah heatmap tampak kelompok gen dengan pola ekspresi tinggi pada jaringan tumor dan rendah pada jaringan normal. Gen-gen ini merupakan kandidat gen yang mengalami upregulation pada kanker payudara dan berpotensi terlibat dalam proses proliferasi sel, adhesi sel epitel, invasi tumor, serta mekanisme progresi kanker.

Konsistensi pola ekspresi antar sampel dalam masing-masing kelompok menunjukkan bahwa sinyal biologis yang diperoleh cukup kuat dan bukan merupakan variasi acak. Dengan demikian, 50 DEGs teratas ini tidak hanya signifikan secara statistik, tetapi juga memiliki relevansi biologis yang jelas dalam membedakan jaringan kanker payudara dari jaringan normal.

3.4 Interpretasi Biologis Awal dari Daftar Gen

Analisis terhadap daftar gen yang mengalami perubahan ekspresi signifikan menunjukkan adanya pola biologis yang konsisten dengan karakteristik molekuler kanker payudara. Gen-gen yang mengalami upregulation didominasi oleh gen yang berperan dalam struktur sel epitel, adhesi sel, dan proliferasi. Misalnya, peningkatan ekspresi KRT19, KRT18, EPCAM, CD24, dan TACSTD2 mengindikasikan aktivasi sifat epitelial dan peningkatan aktivitas proliferatif pada jaringan tumor. Gen-gen tersebut sering dikaitkan dengan pertumbuhan sel yang tidak terkontrol serta peningkatan kemampuan invasi dan metastasis.

Selain itu, beberapa gen seperti GATA3 juga menunjukkan peningkatan ekspresi yang signifikan. GATA3 diketahui berperan dalam diferensiasi sel epitel payudara dan sering digunakan sebagai marker molekuler pada kanker payudara tertentu. Peningkatan ekspresi gen-gen ini menunjukkan adanya perubahan regulasi transkripsi yang mendukung perkembangan tumor.

Sebaliknya, gen-gen yang mengalami downregulation sebagian besar berkaitan dengan metabolisme lipid dan fungsi jaringan adiposa, seperti RBP4, ACACB, LPL, ADIPOQ, dan PPARG. Penurunan ekspresi gen-gen tersebut menunjukkan adanya gangguan regulasi metabolisme normal jaringan payudara. Mengingat jaringan payudara memiliki komponen adiposa yang signifikan, supresi gen metabolik ini mencerminkan perubahan lingkungan mikro jaringan serta pergeseran metabolisme sel kanker untuk mendukung pertumbuhan dan kebutuhan energi yang meningkat.

Secara umum, daftar gen yang diperoleh menunjukkan dua kecenderungan utama, yaitu peningkatan ekspresi gen yang mendukung proliferasi dan karakteristik sel tumor, serta penurunan ekspresi gen yang berperan dalam metabolisme dan homeostasis jaringan normal. Pola ini memberikan gambaran awal bahwa kanker payudara pada dataset GSE15852 mengalami reprogramming molekuler yang melibatkan aktivasi jalur proliferatif dan disrupsi fungsi metabolik.

3.5 Functional Enrichment Analysis (GO & KEGG)

3.5.1 Gene Ontology (GO) Enrichment Analysis

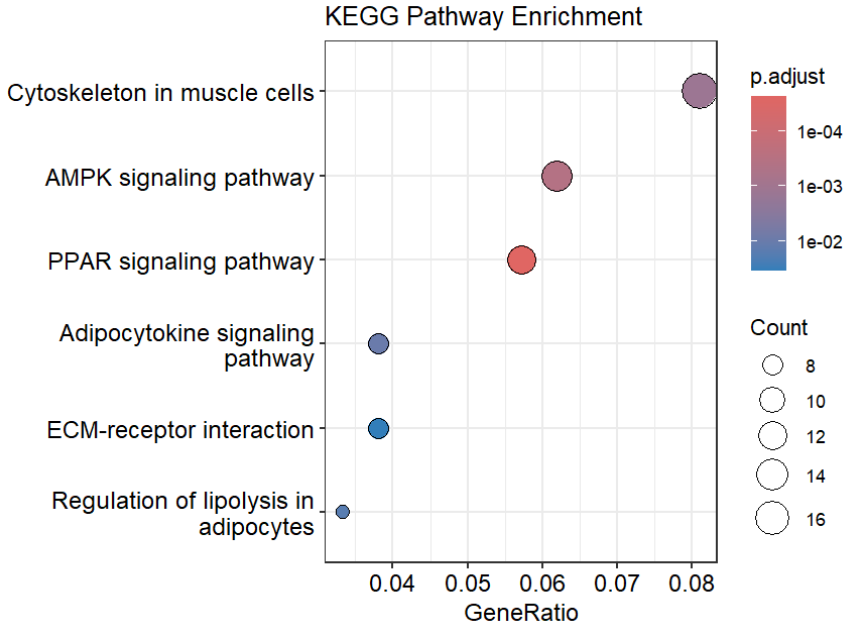
Analisis Gene Ontology (GO) dilakukan menggunakan kategori Biological Process (BP) dengan metode koreksi Benjamini-Hochberg (BH) dan cutoff adjusted p-value < 0.05 serta q-value < 0.2. Analisis dilakukan pada 365 gen signifikan hasil differential expression dengan organisme Homo sapiens.

Hasil analisis menunjukkan terdapat 4.102 istilah GO yang teridentifikasi, dengan 54 istilah signifikan setelah penyaringan berdasarkan p-value dan q-value. Hal ini menunjukkan bahwa gen-gen yang mengalami perubahan ekspresi pada kanker payudara terlibat dalam berbagai proses biologis yang luas dan kompleks. Secara umum, hasil enrichment GO mengindikasikan keterlibatan gen dalam proses-proses yang berkaitan dengan:

- Regulasi metabolisme lipid
- Respons terhadap stimulus hormonal
- Regulasi proliferasi dan diferensiasi sel
- Organisasi struktur sel dan komponen ekstraseluler
- Proses inflamasi dan respons imun

Keterlibatan proses-proses tersebut sejalan dengan karakteristik biologis kanker payudara, yang diketahui mengalami perubahan metabolik (metabolic reprogramming), disregulasi siklus sel, serta perubahan interaksi sel dengan lingkungan mikro (tumor microenvironment). Banyaknya istilah GO yang signifikan menunjukkan bahwa perubahan ekspresi gen pada dataset ini tidak bersifat acak, melainkan mencerminkan perubahan sistemik pada tingkat molekuler.

3.5.2 KEGG Pathway Enrichment Analysis



Gambar 5. Hasil analisis KEGG pathway enrichment pada gen yang mengalami differential expression.

Selain GO, dilakukan pula analisis pathway menggunakan database KEGG untuk mengidentifikasi jalur molekuler yang diperkaya secara signifikan. Berdasarkan hasil visualisasi dot plot KEGG (Gambar 5), beberapa pathway yang signifikan antara lain:

- Cytoskeleton in muscle cells
- AMPK signaling pathway
- PPAR signaling pathway
- Adipocytokine signaling pathway
- ECM-receptor interaction
- Regulation of lipolysis in adipocytes

Pathway dengan GeneRatio tertinggi adalah Cytoskeleton in muscle cells, menunjukkan proporsi gen yang relatif besar terlibat dalam pengaturan struktur sitoskeleton. Perubahan pada sitoskeleton berkaitan erat dengan kemampuan migrasi dan invasi sel kanker. Pathway AMPK signaling pathway dan PPAR signaling pathway juga menunjukkan signifikansi tinggi (p.adjust rendah), yang mengindikasikan adanya perubahan regulasi metabolisme energi dan lipid pada kanker payudara. Jalur AMPK berperan dalam homeostasis energi sel, sedangkan PPAR terlibat dalam metabolisme lipid dan diferensiasi sel. Aktivasi atau disfungsi jalur-jalur ini sering dikaitkan dengan reprogramming metabolik pada sel tumor. Selain itu, enrichment pada ECM-receptor interaction menunjukkan adanya perubahan interaksi antara sel kanker dan matriks

ekstraseluler, yang merupakan mekanisme penting dalam proses invasi dan metastasis. Sementara itu, pathway terkait adiposit dan adipocytokine signaling mencerminkan hubungan antara jaringan lemak dan perkembangan kanker payudara, yang memang memiliki keterkaitan biologis kuat.

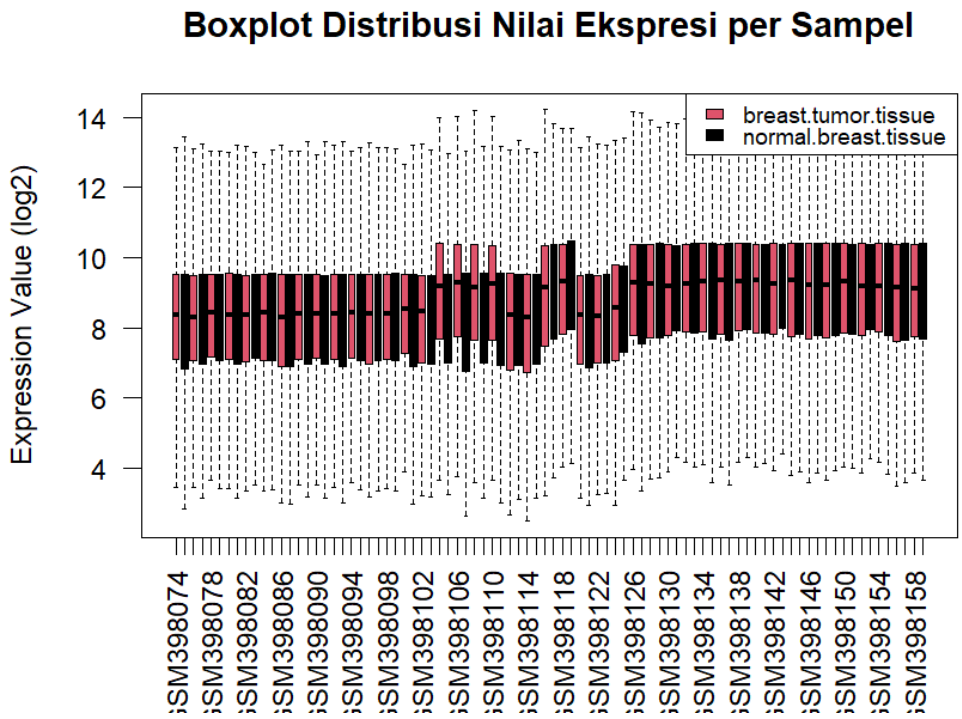
Secara keseluruhan, hasil enrichment KEGG memperkuat temuan differential expression sebelumnya, bahwa perubahan gen pada dataset ini terutama berkaitan dengan regulasi metabolisme, struktur sel, serta interaksi sel dengan lingkungan mikro tumor.

KESIMPULAN

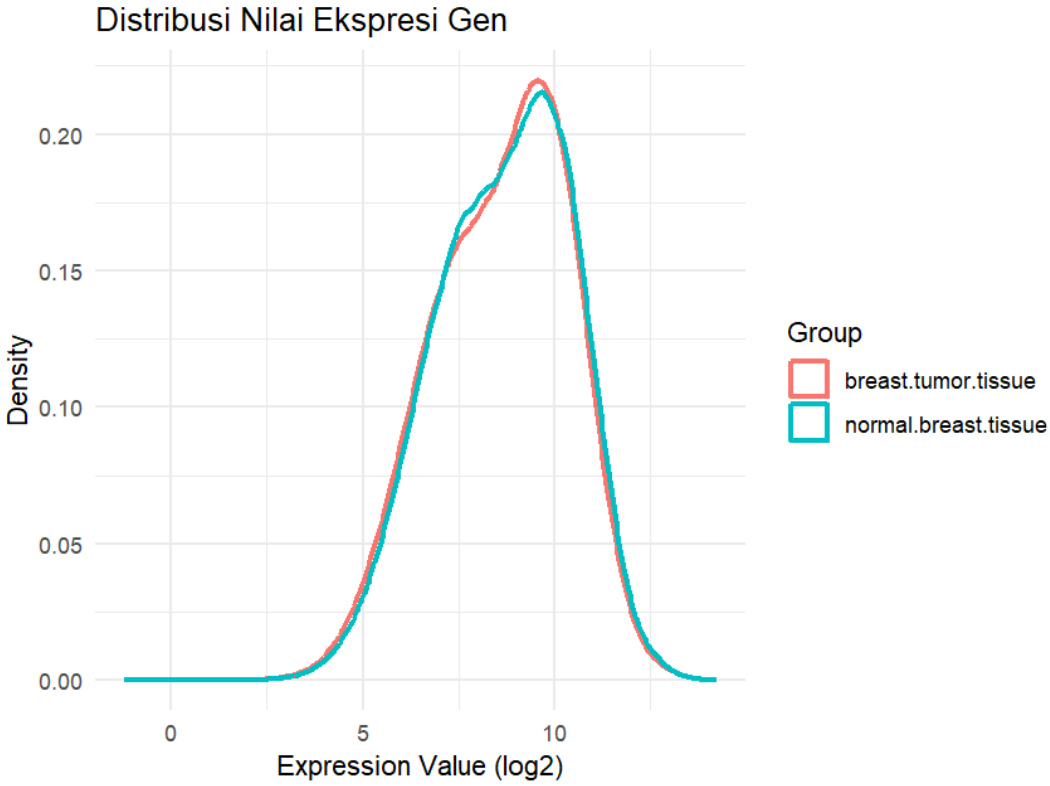
Analisis differential gene expression pada dataset GSE15852 berhasil mengidentifikasi gen-gen yang mengalami perubahan ekspresi signifikan antara jaringan kanker payudara dan jaringan normal. Gen yang mengalami upregulation umumnya berkaitan dengan proliferasi dan karakteristik epitel sel tumor, sedangkan gen yang mengalami downregulation didominasi oleh gen yang terlibat dalam metabolisme lipid dan fungsi jaringan adiposa. Hasil enrichment GO dan KEGG menunjukkan bahwa perubahan ekspresi gen terutama terkait dengan regulasi metabolisme energi, jalur sinyal proliferatif, organisasi sitoskeleton, serta interaksi sel dengan matriks ekstraseluler. Temuan ini menunjukkan adanya reprogramming molekuler yang mendukung pertumbuhan dan progresi tumor. Secara keseluruhan, analisis ini menegaskan bahwa dataset GSE15852 memiliki perbedaan molekuler yang jelas antara jaringan kanker dan normal serta memberikan gambaran biologis yang konsisten mengenai mekanisme utama dalam patogenesis kanker payudara.

LAMPIRAN

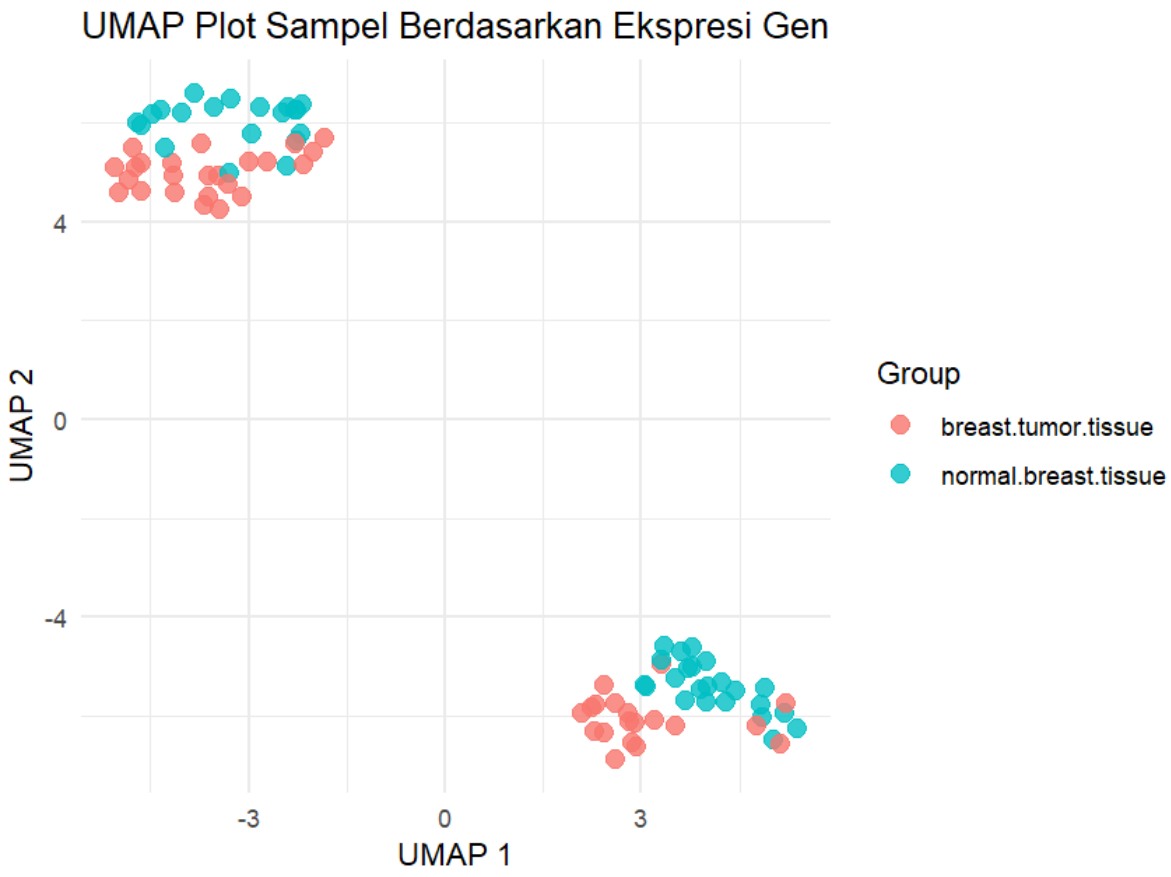
5.1 Visualisasi Boxplot



5.2 Visualisasi Distribusi Nilai Ekspresi Gen



5.3 Visualisasi UMAP Plot


UMAP Plot Sampel Berdasarkan Ekspresi Gen
UMAP 2
4
0
-4
-3
0
3
UMAP 1
Group
breast.tumor.tissue
normal.breast.tissue