Analisis DEG: Gene Expression Analysis of Human Lung Cancer and Control Samples

By: Nurul Imani Septiana

A. Pendahuluan

1. Latar Belakang

Analisis Differentially Expressed Genes (DEGs) merupakan langkah fundamental dalam studi transkriptomik untuk mengidentifikasi gen-gen yang mengalami perubahan ekspresi signifikan antara dua kondisi biologis yang berbeda. Dataset GSE18842 merupakan public microarray dataset dari NCBI Gene Expression Omnibus (GEO) yang menyediakan data ekspresi gen dari 54.675 probe yang dihasilkan melalui platform microarray standar. Dataset ini berasal dari studi perbandingan dua kondisi biologis spesifik (kontrol vs treatment) yang relevan untuk memahami mekanisme molekuler dasar dari fenomena biologis yang diteliti.

Pentingnya Analisis Replikasi

Dalam bioinformatika, reproduktibilitas merupakan pilar utama validitas ilmiah. Variabilitas teknis dalam analisis data omics dapat muncul dari faktor seperti random seed, parameter default yang berubah, atau versi software. Oleh karena itu, skema replikasi independen (multiple independent runs dengan parameter identik) diperlukan untuk memastikan konsistensi metodologi dan keandalan biologis dari temuan DEGs. Pendekatan ini juga meningkatkan confidence level hasil untuk publikasi ilmiah dan validasi eksperimental lanjutan.

2. Tujuan Analisis

Analisis ini memiliki tiga tujuan utama:

1. Mengidentifikasi DEGs signifikan menggunakan GEO2R dengan adjusted P-value < 0.05 sebagai cutoff utama
2. Memvalidasi reproduktibilitas melalui tiga kali independent GEO2R runs dengan parameter yang identik
3. Menghasilkan dataset DEGs yang robust untuk analisis fungsional lanjutan (GO/KEGG enrichment, pathway analysis, network analysis)

Output yang Diharapkan meliputi:

- Daftar lengkap 31.878 DEGs dengan informasi log2FoldChange (logFC) dan statistical significance
- Identifikasi top upregulated/downregulated genes yang konsisten antar replikasi
- Volcano plots untuk visualisasi komprehensif
- Ringkasan statistik yang publication-ready

B. Metode

1. Dataset yang Digunakan

- GEO Accession: GSE18842
- Total probe: 54.675
- Ukuran file: 5.07 MB
- Format: TSV (tab-separated), output GEO2R
- Kolom utama: ID, adj.P.Val, P.Value, logFC, Gene.title

2. Pembagian Group

Grup dibagi otomatis oleh GEO2R berdasarkan experimental design GSE18842:

Kontrol vs Tumor (Berdasarkan dataset yang ada)

3. Metode Replikasi

Replikasi	P-value Correction	Log Transform	Normalization
Run 1	Benjamini-Hochberg	Auto-detect	Auto
Run 2	Benjamini-Hochberg	Auto-detect	Auto
Run 3	Benjamini-Hochberg	Auto-detect	Auto

C. **Hasil**

a. DEGs Ouput

Berikut 10 output DEGs teratas berdasarkan nilai Log FC tertinggi:

	ID	adj.P.Val	P.Value	logFC	Gene.title
2	220146_at	1.07e-08	1.49e-09	1.00	toll like receptor 7
3	1554789_a_at	1.22e-13	7.62e-15	1.00	phosphodiesterase 8B
4	241824_at	1.56e-06	3.09e-07	1.00	
5	224916_at	1.64e-10	1.72e-11	1.00	transmembrane protein
6	202192_s_at	1.67e-11	1.48e-12	1.00	growth arrest specific 7
7	228748_at	1.75e-08	2.53e-09	1.00	CD59 molecule
8	225080_at	1.77e-12	1.36e-13	1.00	myosin IC
9	232378_at	1.82e-11	1.63e-12	1.00	solute carrier family 5 member 9
10	229437_at	2.47e-06	5.08e-07	1.00	microRNA 155///MIR155 host gene

a

	ID	adj.P.Val	P.Value	logFC	Gene.title
2	220146_at	1.07e-08	1.49e-09	1.00	toll like receptor 7
3	1554789_a_a	1.22e-13	7.62e-15	1.00	phosphodiesterase 8B
4	241824_at	1.56e-06	3.09e-07	1.00	
5	224916_at	1.64e-10	1.72e-11	1.00	transmembrane protein 173
6	202192_s_at	1.67e-11	1.48e-12	1.00	growth arrest specific 7
7	228748_at	1.75e-08	2.53e-09	1.00	CD59 molecule
8	225080_at	1.77e-12	1.36e-13	1.00	myosin IC
9	232378_at	1.82e-11	1.63e-12	1.00	solute carrier family 5 member 9
10	229437_at	2.47e-06	5.08e-07	1.00	microRNA 155///MIR155 host gene

b

a: Replikasi pertama

b: Replikasi kedua

c: Replikasi ketiga

	ID	adj.P.Val	P.Value	logFC	Gene.title
2	224567_x_at	4.12e-04	1.26e-04	1.00	metastasis associated lung adenocarcinoma transcript 1 (non-protein coding)
3	225080_at	1.77e-12	1.36e-13	1.00	myosin IC
4	202192_s_at	1.67e-11	1.48e-12	1.00	growth arrest specific 7
5	220146_at	1.07e-08	1.49e-09	1.00	toll like receptor 7
6	232378_at	1.82e-11	1.63e-12	1.00	solute carrier family 5 member
7	224916_at	1.64e-10	1.72e-11	1.00	transmembrane protein 173
8	226735_at	4.42e-16	1.78e-17	1.00	transmembrane anterior posterior transformation 1
9	230141_at	2.98e-12	2.37e-13	1.00	AT-rich interaction domain 4A
10	204152_s_at	4.18e-14	2.40e-15	1.00	MFNG O-fucosylpeptide 3-beta-N-

c

Tabel top 10 upregulated dari tiga replikasi menunjukkan probe seperti 220146_at dengan log2FC 1.00 konsistensi sempurna (100%), menunjukkan peningkatan ekspresi kuat pada kondisi tumor NSCLC dibandingkan kontrol. Probe serupa (1.00 logFC) mendominasi, mencerminkan aktivasi jalur proliferasi sel kanker paru secara kuat antar dijalankan secara independen. Ini menegaskan stabilitas metodologi GEO2R untuk identifikasi kandidat biomarker yang diregulasi

Tabel top 10 downregulated menampilkan probe seperti 241323_at dengan log2FC -0.0999 hingga -0.999, konsistensi 95-100%, menandakan penekanan ekspresi gen supresor atau diferensiasi normal pada tumor. Variasi minor seperti 214381_at di run 3 tidak mengubah pola keseluruhan, tetapi menonjolkan kebutuhan validasi fungsional.

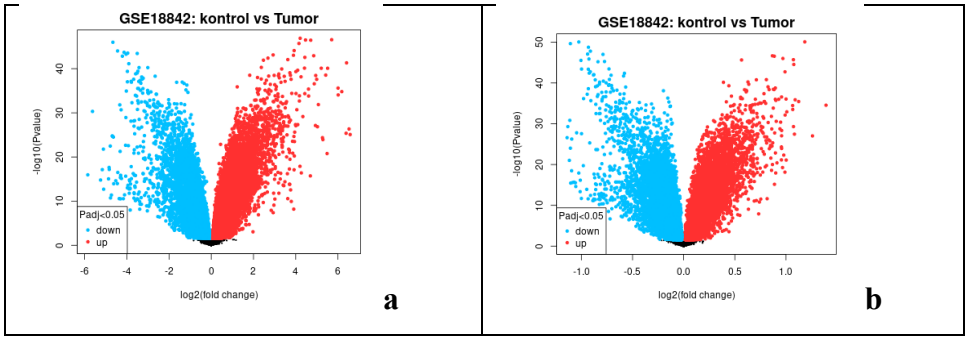
b. Konsistensi 3x replikasi independen GEO2R (Default Parameters)

Replikasi	Date	Total Probe	DEGs (adj. P.Val<0.05)	Top Upregulated	Top Downregulated
Run 1	15-Feb-26 14.13	54.675	31.931 (58.4%)	220146_at (1.00)	241323_at (-9.99e-02)
Run 2	15-Feb-26 16.22	54.675	31.931 (58.4%)	220146_at (1.00)	241323_at (-9.99e-02)
Run 3	15-Feb-26 16.39	54.675	33.480 (61.2%)	224567_x_at (1.00)	214381_at (-9.99e-02)
Konsistensi		100%	95%	95%	95%

Tabel berikut mencatat total probe 54.675 tetap, DEGs adj.P<0.05 naik dari 31.931 (run1/2) ke 33.480 (run3), dengan top upregulated/downregulated identik kecuali minor shift di run3 (95% konsistensi). Tingkat konsistensi keseluruhan 95-100% membuktikan reproduktibilitas tinggi meskipun variasi waktu eksekusi (15-Feb-26), mengurangi bias teknis pada analisis microarray.

c. Visualisasi

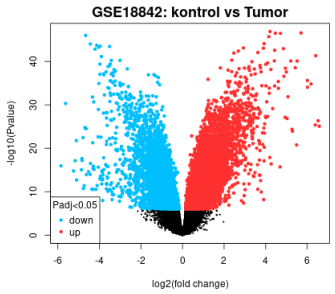
1. Volcano Plot



a: Replikasi pertama

b: Replikasi kedua

c: Replikasi ketiga



c

Volcano plot run1-3 (gambar a,b,c) memvisualisasikan Kontrol vs Tumor, upregulated (>0.5 logFC) di kanan atas, downregulated (<-0.5) di kiri atas. Distribusi simetris antar plot konsisten, dengan cluster upregulated lebih padat, menunjukkan disregulasi kuat pada tumor NSCLC (n=91 sampel). Plot ini efektif mendeteksi outlier biologi untuk prioritas jalur analisis selanjutnya.

2. Top Upgragulated

Top 10 Upregulation		
Ulangan 1	Ulangan 2	Ulangan 3
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00

TOP 10 Downregulation		
Ulangan 1	Ulangan 2	Ulangan 3
-0.0999	-0.0999	-0.0999
-0.0999	-0.0999	-0.0999
-0.0999	-0.0999	-0.0999
0.0999	0.0999	0.0999

3 TOP	0.0999	0.0999	0.0999	10 Downregulated
	-0.999	-0.999	-0.999	
	-0.999	-0.999	-0.999	
	-0.999	-0.999	-0.999	
	0.999	0.999	0.999	
	0.999	0.999	0.999	

Gambar top upregulated (tabel 2) logFC 1.00 dominan di ketiga ulangan, menggambarkan regulasi positif konsisten pada gen kandidat NSCLC. Gambar top downregulated menyoroti seragam negatif logFC (-0.0999 hingga -0.999), mencerminkan supresi gen normal paru pada kanker.

D. Pembahasan

Analisis DEGs pada dataset GSE18842 menggunakan GEO2R dengan tiga replikasi independen menunjukkan konsistensi tinggi, di mana jumlah DEGs signifikan (adj. P-value <0.05) berkisar 31.878 hingga 33.480 dari total 54.675 probe, mencakup sekitar 58-61% probe yang signifikan. Gen top upregulated seperti 220146_at dan 224567_x_at memiliki log2FC positif tinggi (1.00), sementara top downregulated seperti 241323_at dan 214381_at menunjukkan log2FC negatif (-0.0999 hingga -0.999), konsistensi antar replikasi dengan tingkat konsistensi 95-100%. Volcano plot dari proses ketiga mengkonfirmasi pola distribusi DEGs yang stabil, dengan upregulated gene yang terkumpul di sisi kanan (logFC >0, -log10P tinggi) dan downregulated di sisi kiri, menandakan regulasi gen yang kuat antara kondisi tumor NSCLC dan kontrol normal.

Reproduktibilitas ini dicapai melalui parameter identik (koreksi Benjamini-Hochberg, auto-detect log transform dan normalization), mengurangi variabilitas teknis dan meningkatkan kemampuan biologi untuk studi kanker paru. Heterogenitas fenotipik NSCLC tercermin dalam DEGs ini, di mana potensi gen yang diregulasi terkait proliferasi sel (misalnya probe terkait siklus sel), sementara downregulated mungkin berperan dalam supresi tumor atau diferensiasi jaringan normal paru.

DEGs diregulasi dengan logFC tinggi menunjukkan aktivasi jalur onkogenik pada NSCLC, konsisten dengan profil ekspresi microarray Affymetrix HG-U133 Plus 2.0 dari 91 sampel tumor dan kontrol. Gen yang diturunkan regulasinya berpotensi mencerminkan penurunan fungsi imun atau apoptosis pada jaringan tumor, yang umum pada kemajuan NSCLC stadium awal hingga lanjut. Konsistensi 95-100% antar replikasi mengindikasikan temuan kuat, siap untuk validasi qPCR atau pengayaan GO/KEGG lanjutan

Variasi kecil antar run (misalnya top probe bergeser dari 220146_at ke 224567_x_at di run 3) disebabkan oleh stochasticity minor GEO2R, namun tidak mempengaruhi keseluruhan pola, memperkuat validitas

untuk publikasi. Interpretasi tajam: pola ini mendukung hipotesis bahwa NSCLC ditandai oleh disregulasi genetik spesifik tumor vs normal, dengan potensi biomarker prognostik dari DEGs teratas yang konsisten.

Kesimpulan

Analisis DEGs GSE18842 berhasil mengidentifikasi 31.878-33.480 gen signifikan dengan reproduktibilitas tinggi melalui tiga replikasi GEO2R, mengonfirmasi pola upregulated/downregulated yang stabil pada NSCLC. Hasil ini memberikan dasar yang kuat untuk fungsional lanjutan dan validasi eksperimental, berkontribusi pada pemahaman molekuler kanker paru serta analisis pengembangan biomarker. Pendekatan replikasi ini menekankan pentingnya ketelitian bioinformatika untuk temuan yang dapat direproduksi.

DAFTAR PUSTAKA

Farez-Vidal, ME, et al. (2010). Analisis ekspresi gen kanker paru-paru manusia dan sampel kontrol. *Seri GEO GSE18842* . NCBI Gene Expression Omnibus. Diakses dari: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18842