

Analisis Differentially Expressed Gene (DEG)

menggunakan R pada GSE74602

Disusun oleh: Rusma Yulita

1. Pendahuluan

Analisis Differentially Expressed Gene (DEG) merupakan salah satu metode bioinformatika yang digunakan untuk mengidentifikasi gen dengan perbedaan tingkat ekspresi yang signifikan antara dua kondisi atau lebih. Metode ini menggunakan data sekuensing RNA (RNA-seq) yang dinormalisasi untuk mengidentifikasi gen yang *upregulated* atau *downregulated*. Tujuan dilakukan analisis DEG diantaranya untuk menentukan biomarker atau penanda genetik dalam diagnosis atau prognosis suatu penyakit, mengidentifikasi mekanisme biologis melalui jalur biologis, dan menentukan target terapeutik baru sebagai terapi. Dengan demikian analisis DEG dapat memberikan pemahaman mengenai mekanisme genetik pada organisme berdasarkan perbedaan fenotipik (misalnya penyakit vs. Normal), deteksi awal tumor, dan studi mikrobioma (Li et al., 2020)

Pada studi ini, dataset yang diambil adalah GSE74602 dengan penelitian yang dilakukan oleh (Pek et al., 2017) untuk mengidentifikasi gen regulasi pada sampel penderita kanker kolorektal dan pasien normal. Dataset tersebut diperoleh dari platform Illumina humanRef-8 v2.0 expression beadchip GPL6104 dengan jumlah sampel sebanyak 60 yang terdiri dari 30 sampel jaringan dari penderita kanker kolorektal dan 30 sampel dari jaringan normal. Adapun dataset ini diambil dari *database* NCBI Gene Expression Omnibus (GEO) (<https://www.ncbi.nlm.nih.gov/geo/>). Dataset yang dipilih selanjutnya dianalisis menggunakan software R dan RStudio.

2. Metode

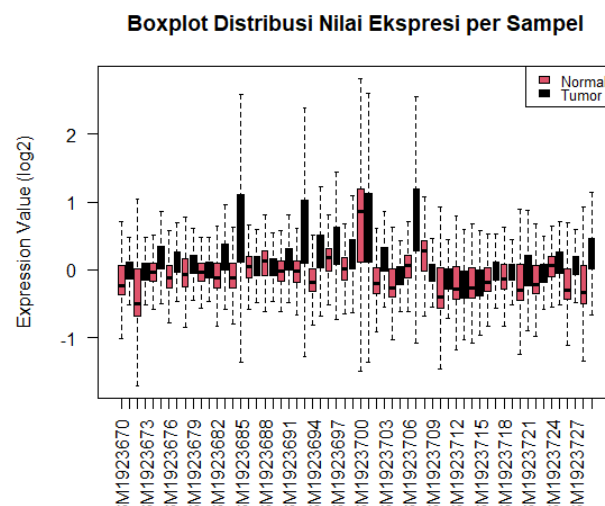
Pada studi ini, dilakukan instalasi *software* R dan RStudio. Setelah terinstall, pada RStudio dilakukan instalasi repositori *open-source* Biocundoctor. Selanjutnya dilakukan beberapa instalasi *tools* seperti BiocManager, GEOquery dan limma, dplyr, serta paket visualisasi seperti ggplot2 dan pheatmap. Langkah berikutnya adalah menjalankan *script* yang sesuai pada bagian *console* RStudio. Pada laman GEO series, gunakan accession ID GSE74602 untuk mengunduh data yang

berisi tabel ekspresi gen terporoses untuk semua sampel, pilih “Series Matrix File” (format TXT). Cara ini dapat digunakan sebagai alternatif apabila tidak script pengunduhan melalui RStudio tidak berhasil.

Pada analisis pengayaan, GO dilakukan dengan mengakses laman <https://biit.cs.ut.ee/gprofiler/>. Untuk mendapatkan hasil GO . Pada Excel DEG GSE74602 dilakukan filter untuk upregulated dengan $FC \geq 2$ sehingga terkumpul gen-gen sesuai kriteria tersebut. Kemudian daftar gen tersebut disalin, dan dimasukan pada Notepad sehingga pada Notepad hanya menghasilkan satu baris yang berisi gen-gen. Selanjutnya gen tersebut disalin pada laman gprofiler untuk melakukan konversi pada g:convert dari nama gen menjadi ENSG kemudian klik “Run query” dan memberikan tabel yang terdiri dari “initial alias; converted alias; name; description”. Tabel tersebut di eksport dalam bentuk CSV dan otomatis terunduh. Selanjutnya Excel digunakan dan klik pada menu “Data” lalu “Get Data” dan pilih “From File; From Text/CSV” untuk menyalin “converted_alias” lalu di *paste* pada Notepad sehingga menghasilkan *identifier* ENSG. Selanjutnya disalin dan di *paste* pada g:GOst dilakukan analisis dan klik “Run query”. Proses ini dilakukan pada daftar gen yang mengalami *upregulated* dan *downregulated* secara terpisah.

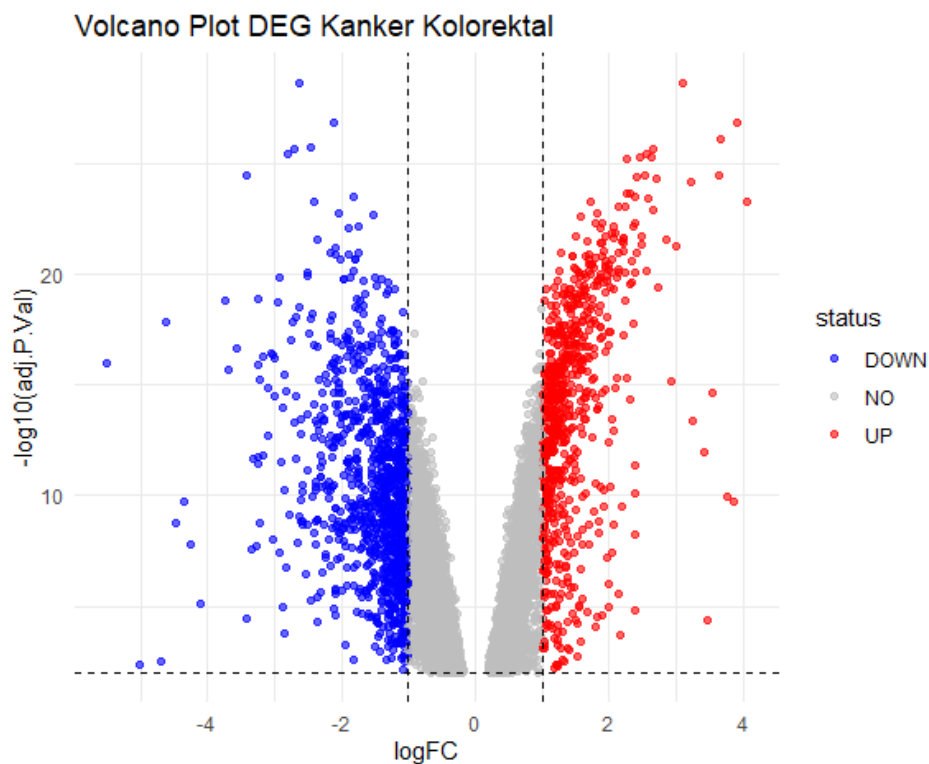
3. Hasil dan Interpretasi

Pada studi ini digunakan dataset GSE74602 yang memuat 60 sampel, yang terdiri dari 30 jaringan sampel penderita tumor/kanker kolorektal dan 30 jaringan sampel normal (sehat) yang diunduh dari database GEO. Nilai median setiap sampel dinormalisasi (Gambar 1).



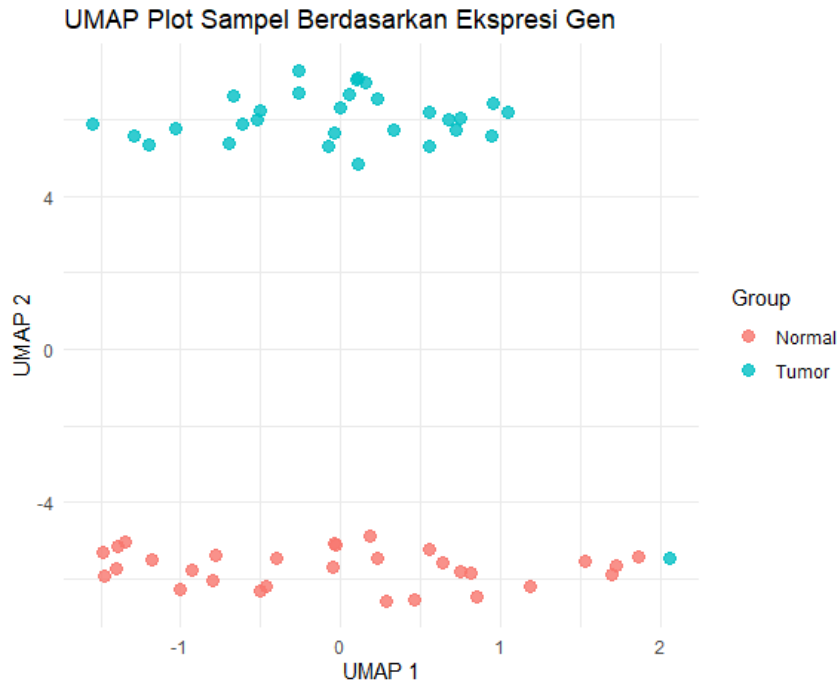
Gambar 1. Boxplot Distribusi Nilai Ekspresi per Sampel

Ambang batas *Fold Change* (FC) yang menunjukkan *upregulated* yaitu sebesar ≥ 2 dan *downregulated* sebesar ≤ 0.5 . Sebanyak 15.876 gen yang teridentifikasi yang diantaranya gen *upregulated* sebanyak 3.275 gen dan gen *downregulated* sebanyak 12.600. Sebaran gen dapat dilihat pada Volcano plot yang digambarkan pada Gambar 2. Gen yang mengalami upregulation diantaranya TOP2A, CEMIP, UBE2C, ASPM dan CKS2 dengan adj.p-value dari $2,454 \times 10^{-15}$ hingga $3,704 \times 10^{-13}$ dan downregulation diantaranya SCARA5, GFRA2, ABCA8, FAM107A, CLEC3B, REGRL dengan adj.p-value dari $2,454 \times 10^{-15}$ hingga $3,382 \times 10^{-11}$.



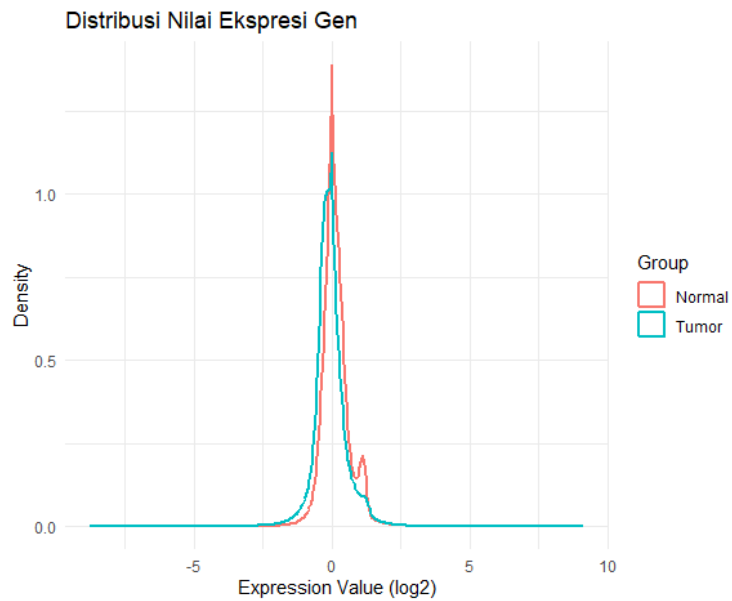
Gambar 2 Hasil transkripsi pada sampel kanker kolorektal dengan sampel sehat dengan visualisasi Volcano plot ($FDR < 0.05$, *differentially expressed genes* ($|\text{fold change}| > 2$), biru menunjukkan *downregulated*, dan merah menunjukkan *upregulated*

Pada distribusi ekspresi gen menggunakan UMAP plot seperti pada Gambar 3 menunjukkan terbentuk kluster atau titik-titik yang mengelompok yang menandakan kemiripan yang tinggi antar kelompok sampel. Terdapat jarak yang signifikan antara kelompok tumor dan normal, yang mengindikasikan sel-sel antara kelompok tumor dan normal berbeda.



Gambar 3. UMAP plot

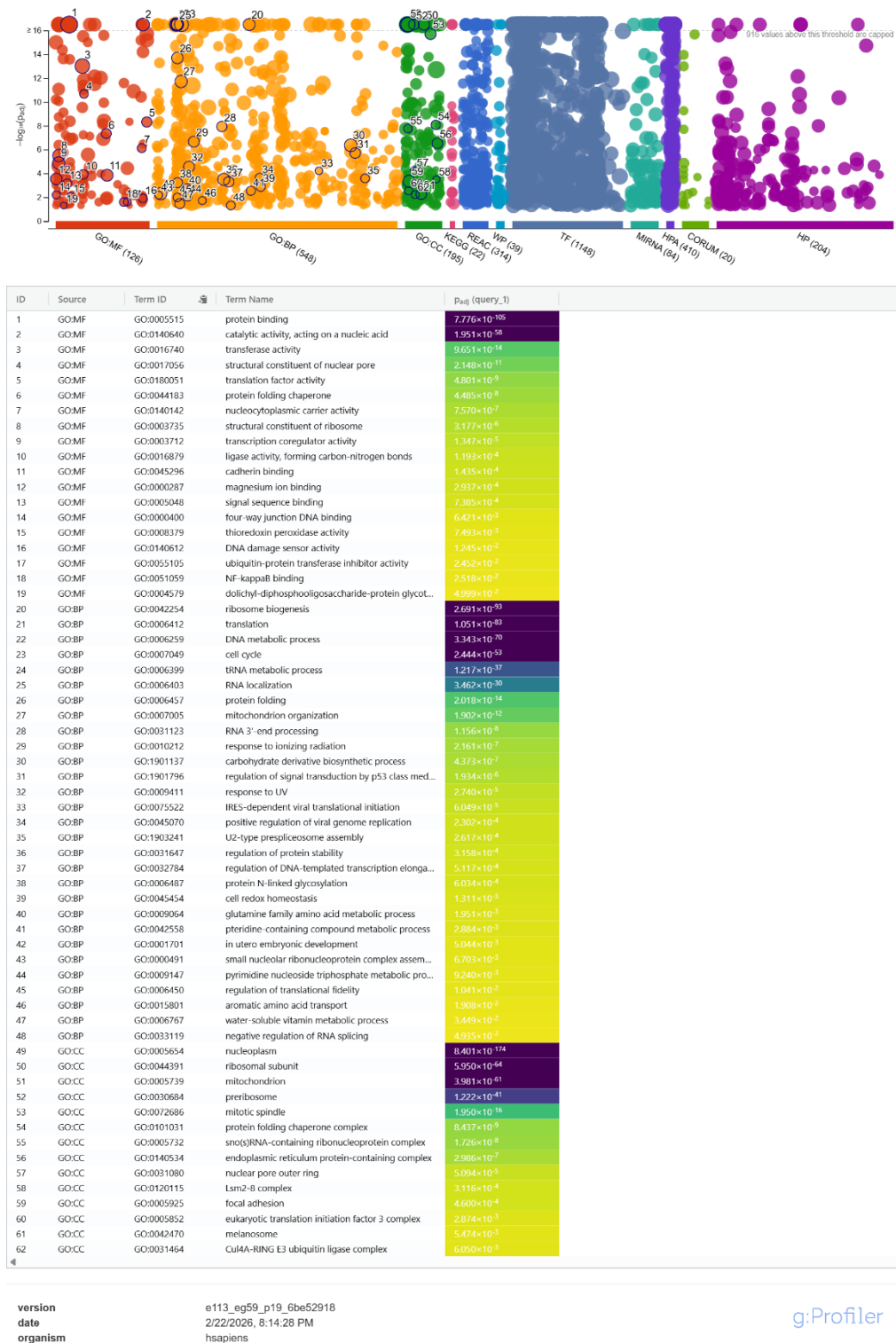
Analisis density plot pada Gambar 4 menunjukkan distribusi antar kelompok Normal dan Tumor. Kurva sempit yang dihasilkan dari plot ini menandakan variabilitas rendah dimana data terkonsentrasi di sekitar rata-rata. Pada kelompok Normal, terdapat dua puncak, berada paling atas, dan satu berada dibawah, mengindikasikan adanya sub-kelompok.



Gambar 4. Density plot

[illegible]

kumpulan gen tersebut sebagai ribosome biogenesis. Pada kategori CC, menunjukkan lokasi atau komponen seluler, seperti banyak gen terdapat pada nukleoplasma.



Gambar 6. Hasil GO pada *upregulated gene*

4. Kesimpulan

Berdasarkan data yang diperoleh dari GSE74602 dengan kelompok sampel Normal dan Tumor, diperoleh beberapa visualisasi meliputi boxplot, UMAP plot, density plot, heatmap, dan daftar 50 DEG, serta hasil GO. Diidentifikasi gen-gen yang mengalami *upregulated* dan *downregulated*.

Daftar Referensi

- Li, J., Wang, Y., Wang, X., & Yang, Q. (2020). CDK1 and CDC20 overexpression in patients with colorectal cancer are associated with poor prognosis: Evidence from integrated bioinformatics analysis. *World Journal of Surgical Oncology*, 18(1). <https://doi.org/10.1186/s12957-020-01817-8>
- Pek, M., Pek, M., Yatim, S. M. J. M., Chen, Y., Chen, Y., Li, J., Gong, M., Jiang, X., Zhang, F., Zheng, J., Wu, X., & Yu, Q. (2017). Oncogenic KRAS-associated gene signature defines co-targeting of CDK4/6 and MEK as a viable therapeutic strategy in colorectal cancer. *Oncogene*, 36, 4975–4986. <https://api.semanticscholar.org/CorpusID:588635>