UJIAN AKHIR SEMESTER

SAINS DATA GENOM

ANALISIS EKSPRESI GEN DATA *WHOLE BLOOD* ANAK DENGAN *SEPTIC SHOCK* MENGGUNAKAN METODE *SUPERVISED* DAN *UNSUPERVISED LEARNING*

Kelompok 7

Khalila Izzatunnisa	(2206051544)
Aulia Nisrina Rosanita	(2206051380)
Pinky Siwi Nastiti	(2206051430)
Cecylia Ongso Putri	(2206048902)

DEPARTEMEN MATEMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS INDONESIA

2025

DAFTAR ISI

I. Pendahuluan

1.1 Latar Belakang

Septic shock merupakan bentuk lanjut dari sepsis yang ditandai oleh hipotensi persisten dan gangguan perfusi jaringan, serta memiliki angka mortalitas yang tinggi, terutama pada pasien anak yang dirawat di Pediatric Intensive Care Unit (PICU). Di seluruh dunia, sepsis dan septic shock menjadi penyebab kematian utama pada bayi dan anak-anak, bahkan di negara maju sekalipun. Diagnosis klinis yang cepat dan akurat sangat penting untuk meningkatkan angka kelangsungan hidup, namun seringkali sulit dilakukan karena gejala awal yang tidak spesifik.

Perkembangan teknologi omik, seperti mikroarray dan RNA-Seq, memungkinkan dilakukan analisis ekspresi gen secara masif untuk mengidentifikasi perubahan molekuler pada pasien septic shock. Salah satu pendekatan penting dalam bioinformatika adalah analisis ekspresi gen diferensial (Differentially Expressed Genes/DEG), yang dapat digunakan untuk mengidentifikasi gen-gen yang memiliki ekspresi berbeda secara signifikan antara kelompok kontrol dan kelompok kasus (septic shock). Selain itu, pendekatan data mining seperti supervised dan unsupervised learning telah terbukti efektif dalam mengelompokkan data biologis dan membangun model prediktif berbasis data genetik.

Dalam studi ini, digunakan data ekspresi gen dari dataset GEO Series GDS4274, yang mencakup sampel darah lengkap dari 130 pasien anak (≤10 tahun) dalam 24 jam pertama masuk PICU karena septic shock, serta healthy control. Platform *microarray* yang digunakan adalah Affymetrix Human Genome U133 Plus 2.0 Array (GPL570), yang mencakup lebih dari 54.000 probe gen.

Penelitian ini bertujuan untuk menganalisis ekspresi gen pada anak-anak dengan *septic shock* menggunakan kombinasi pendekatan *supervised* learning (klasifikasi dengan *Logistic Regression, K-Nearest Neighbors,* dan *Support Vector Machine*) dan *unsupervised* learning (clustering dengan K-Means, *Hierarchical*, GMM, HDBSCAN, OPTICS serta *biclustering* dengan Spectral *Biclustering*). Diharapkan, hasil dari analisis ini dapat mengungkap gen-gen penting yang dapat berfungsi sebagai biomarker

potensial dan memperdalam pemahaman mengenai mekanisme molekuler *septic shock* pada pasien anak.

1.2 Masalah Penelitian

Berdasarkan latar belakang yang telah disampaikan, penelitian ini berfokus pada perumusan masalah sebagai berikut:

1) Bagaimana penerapan analisis DEG dalam mengidentifikasi gen-gen yang terekspresi secara berbeda secara signifikan antara kelompok *healthy control* dan *septic shock?*
2) Bagaimana penerapan dan performa model klasifikasi seperti Regresi Logistik, *K-Nearest Neighbors* (KNN), dan *Support Vector Machine* (SVM) dalam mengklasifikasikan kelompok *healthy control* dan *septic shock* berdasarkan ekspresi gen yang signifikan, dan gen apa saja yang memberikan kontribusi tertinggi terhadap hasil klasifikasi berdasarkan model terbaik yang diperoleh?
3) Bagaimana penerapan dan performa metode *clustering* seperti K-Means, Hierarchical Clustering, GMM, HDBSCAN, dan OPTICS dalam mengelompokkan gen-gen yang terekspresi secara berbeda secara signifikan?
4) Bagaimana penerapan dan performa metode Spectral *Biclustering* dalam mengidentifikasi pola ko-ekspresi antara gen dan subset sampel secara simultan?
5) Bagaimana penerapan analisis *Gene Ontology Enrichment* dalam mengungkap fungsi biologis gen-gen hasil *clustering* dan *biclustering* dalam tiga domain utama yaitu *biological process, molecular function, dan cellular component?*

1.3 Tujuan Penelitian

Sejalan dengan rumusan masalah yang telah ditetapkan, penelitian ini bertujuan untuk:

1) Menganalisis penerapan metode Differentially Expressed Genes (DEG) untuk mengidentifikasi gen-gen yang terekspresi secara berbeda secara signifikan antara kelompok *healthy control* dan *septic shock*.
2) Menerapkan dan mengevaluasi performa model klasifikasi seperti Regresi Logistik, *K-Nearest Neighbors* (KNN), dan *Support Vector Machine* (SVM) dalam membedakan kelompok *healthy control* dan *septic shock* berdasarkan

ekspresi gen yang signifikan, serta mengidentifikasi gen-gen yang memberikan kontribusi tertinggi terhadap hasil klasifikasi berdasarkan model terbaik.

3) Menerapkan dan mengevaluasi performa metode *clustering*, yaitu K-Means, Hierarchical Clustering, GMM, HDBSCAN, dan OPTICS dalam mengelompokkan gen-gen yang terekspresi secara berbeda secara signifikan.
4) Menerapkan dan mengevaluasi performa metode Spectral *Biclustering* dalam mengidentifikasi pola ko-ekspresi antara gen dan subset sampel secara simultan.
5) Menganalisis penerapan *Gene Ontology Enrichment* untuk mengungkap fungsi biologis dari gen-gen hasil clustering dan biclustering dalam tiga domain utama, yaitu *biological process*, *molecular function*, dan *cellular component*.

II. Tinjauan Pustaka

2.1 Analisis *Differentially Expresed Genes* (DEG)

Analisis *Differentially Expressed Genes* (DEG) merupakan metode statistik untuk mengidentifikasi gen-gen yang menunjukkan perbedaan signifikan dalam tingkat ekspresi antara dua kondisi biologis yang berbeda, misalnya antara kelompok pasien dan kontrol. DEG biasanya digunakan dalam studi transcriptomic seperti mikroarray dan RNA-Seq untuk menemukan biomarker potensial atau gen target terapi. Uji statistik yang umum digunakan dalam DEG adalah uji-t untuk dua kelompok, atau ANOVA untuk lebih dari dua kelompok. Namun, dalam konteks data ekspresi gen, metode statistik harus disesuaikan dengan karakteristik data yang memiliki jumlah fitur (gen) jauh lebih besar daripada jumlah sampel, serta variabilitas yang tinggi. Oleh karena itu, digunakan pendekatan Empirical Bayes seperti pada paket limma untuk mikroarray atau metode DESeq2/edgeR untuk RNA-Seq.

Selain nilai p, hasil uji statistik biasanya dikoreksi dengan metode *False Discovery Rate (FDR)* untuk menghindari kesalahan positif akibat multiple testing. Gen dikategorikan sebagai *upregulated* jika nilai log2 fold change (log2FC) positif dan *downregulated* jika negatif, dengan ambang batas signifikansi biasanya ditentukan oleh FDR < 0.05.

2.2 Analisis Klasifikasi

Analisis klasifikasi merupakan pendekatan supervised learning yang digunakan untuk membangun model prediktif berdasarkan data berlabel. Dalam konteks ekspresi gen, klasifikasi digunakan untuk membedakan kondisi biologis berdasarkan profil ekspresi gen. Berikut adalah beberapa algoritma klasifikasi yang digunakan dalam studi ini:

2.2.1 Model Klasifikasi

2.2.1.1 Regresi Logistik

Regresi logistik adalah metode statistik untuk memodelkan probabilitas suatu kejadian biner sebagai fungsi dari variabel independen. Model ini mengestimasi peluang dari suatu sampel masuk ke dalam kelas tertentu (misalnya sehat atau septic shock) berdasarkan kombinasi linier dari fitur-fitur input. Kelebihan regresi logistik adalah interpretabilitasnya yang baik dan efisien pada dataset yang tidak terlalu kompleks secara non-linear. Model dihitung dengan fungsi sigmoid:

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

2.2.1.2 *K-Nearest Neighbors* (K-NN)

KNN adalah algoritma non-parametrik yang mengklasifikasikan suatu data berdasarkan mayoritas label dari *k* tetangga terdekat dalam ruang fitur. Jarak antar sampel umumnya dihitung menggunakan metrik Euclidean atau Manhattan. Kelebihannya adalah kesederhanaan dan tidak memerlukan pelatihan model, namun performanya dapat dipengaruhi oleh pemilihan nilai *k* dan skala fitur.

2.2.1.3 *Support Vector Machine* (SVM)

SVM adalah algoritma klasifikasi yang berusaha mencari hyperplane optimal yang memisahkan kelas-kelas dalam ruang fitur dengan margin maksimal. Untuk data yang tidak dapat dipisahkan secara linear, SVM menggunakan kernel trick untuk memetakan data ke dimensi lebih tinggi. Kelebihan SVM adalah kemampuannya untuk bekerja dengan data berdimensi

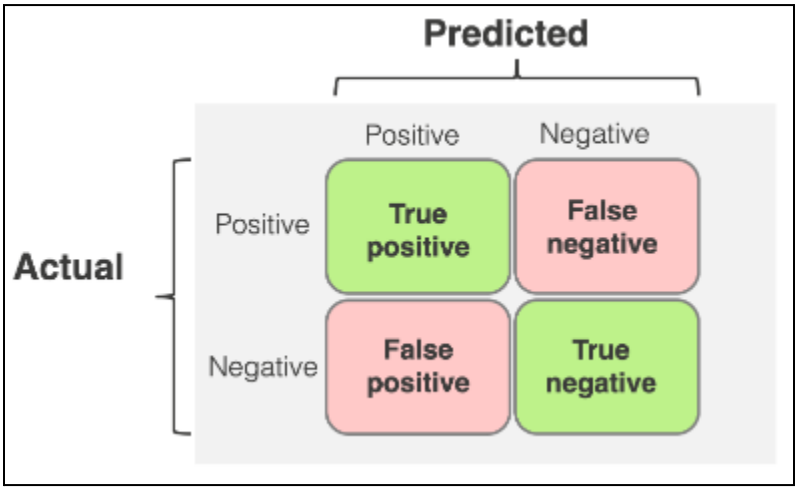
tinggi seperti ekspresi gen, dan relatif tahan terhadap overfitting jika parameter diatur dengan baik.

2.2.2 Metrik Evaluasi

Evaluasi model klasifikasi merupakan aspek krusial dalam proses supervised learning, karena menentukan seberapa baik model dapat membedakan antara kelas target. Dalam studi ini, empat metrik utama digunakan untuk menilai performa model klasifikasi, yaitu: *accuracy, precision, recall (sensitivity)*, dan F1-score. Pemilihan metrik ini mempertimbangkan ketidakseimbangan data, yaitu proporsi yang berbeda antara pasien *septic shock* dan *healthy control*.

2.2.2.1 *Confusion Matrix*

Confusion matrix memberikan ringkasan kinerja model klasifikasi dalam bentuk tabel 2x2, memisahkan prediksi ke dalam empat kategori (TP, TN, FP, FN). Ini memudahkan analisis mendalam atas jenis kesalahan yang dilakukan oleh model.



2.2.2.2 *Accuracy*

Accuracy atau akurasi mengukur proporsi prediksi yang benar dari seluruh prediksi yang dibuat. Meskipun akurasi sering digunakan sebagai metrik utama, pada dataset yang tidak seimbang (*class imbalance*), nilai akurasi bisa menyesatkan. Misalnya, jika 90% data termasuk dalam satu kelas, maka model yang selalu memprediksi kelas mayoritas bisa saja memiliki akurasi tinggi meskipun tidak bermanfaat.

Rumusnya adalah:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

2.2.2.3 *Precision*

Precision adalah proporsi prediksi positif yang benar. Artinya, dari semua prediksi yang menyatakan seseorang memiliki *septic shock*, berapa banyak yang benar-benar memiliki kondisi tersebut.

$$\text{Precision} = \frac{TP}{TP + FP}$$

2.2.2.4 *Recall (Sensitivity)*

Recall mengukur kemampuan model untuk menemukan semua sampel positif dari data. Dalam konteks ini, recall menjawab pertanyaan: dari seluruh pasien *septic shock* yang sebenarnya ada, berapa banyak yang berhasil dikenali oleh model.

$$\text{Recall} = \frac{TP}{TP + FN}$$

2.2.2.5 F1-*Score*

F1-*Score* adalah rata-rata harmonik dari *precision* dan *recall*. F1 memberikan keseimbangan antara keduanya dan berguna ketika ada *trade-off* antara keduanya, terutama pada data yang tidak seimbang.

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

2.3 Analisis *Clustering*

2.3.1 Metode *Clustering*

2.3.1.1 K-Means

K-Means adalah algoritma *clustering* yang paling populer dan banyak digunakan dalam berbagai bidang, termasuk bioinformatika. Metode ini bertujuan untuk membagi data ke dalam sejumlah klaster k yang telah ditentukan

sebelumnya dengan cara meminimalkan jarak antara data dalam satu klaster terhadap pusat klaster (centroid). Proses dimulai dengan inisialisasi centroid secara acak, kemudian data dikelompokkan berdasarkan jarak terdekat ke centroid tersebut. Selanjutnya, centroid diperbarui sebagai rata-rata dari anggota klaster yang baru terbentuk, dan proses ini diulang hingga konvergen. Meskipun efisien dan sederhana secara komputasi, K-Means memiliki beberapa kelemahan, seperti keharusan menentukan jumlah klaster di awal, sensitivitas terhadap nilai awal centroid, dan asumsi bahwa bentuk klaster bersifat isotropik (bulat). Dalam konteks ekspresi gen, K-Means cukup efektif jika digunakan pada data yang telah direduksi dimensinya, misalnya menggunakan PCA.

2.3.1.2 *Hierarchical Clustering*

Hierarchical clustering merupakan pendekatan yang membentuk struktur hierarkis dari data dalam bentuk pohon atau dendrogram. Terdapat dua pendekatan utama dalam metode ini, yaitu agglomerative (bottom-up) dan divisive (top-down). Pendekatan agglomerative memulai proses dengan menganggap setiap data sebagai klaster tersendiri, kemudian secara iteratif menggabungkannya berdasarkan kemiripan sampai seluruh data tergabung dalam satu klaster besar. Sebaliknya, pendekatan divisive memulai dari satu klaster besar dan memecahnya secara bertahap. Proses penggabungan atau pemisahan dilakukan berdasarkan metrik jarak seperti Euclidean atau Manhattan dan metode linkage seperti single, complete, atau average. Hierarchical clustering memiliki kelebihan karena tidak mengharuskan jumlah klaster ditentukan di awal dan mampu merepresentasikan struktur nested antar klaster. Namun, metode ini cukup mahal secara komputasi untuk dataset besar dan hasilnya sangat bergantung pada pilihan metrik dan linkage.

2.3.1.3 Gaussian Mixture Model

Gaussian Mixture Model (GMM) merupakan metode clustering berbasis probabilistik yang mengasumsikan bahwa data berasal dari kombinasi beberapa distribusi Gaussian. Setiap klaster dalam GMM direpresentasikan oleh distribusi Gaussian dengan parameter mean dan covariance tertentu. Untuk menentukan

parameter tersebut, GMM menggunakan algoritma Expectation-Maximization (EM), yang secara iteratif memperbarui estimasi probabilitas keanggotaan dan parameter distribusi. GMM lebih fleksibel dibanding K-Means karena tidak mengasumsikan bentuk klaster yang bulat dan memungkinkan klaster dengan ukuran serta orientasi berbeda. Selain itu, GMM tidak hanya memberikan label klaster, tetapi juga probabilitas keanggotaan masing-masing data terhadap semua klaster, yang berguna dalam interpretasi data biologis yang bersifat kompleks dan ambigu. Namun, GMM tetap memiliki kelemahan, seperti sensitivitas terhadap inisialisasi dan keharusan menentukan jumlah komponen klaster di awal.

2.3.1.4 HDBSCAN

HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*) adalah metode clustering berbasis densitas yang menggabungkan pendekatan DBSCAN dengan struktur hierarki. HDBSCAN bekerja dengan membangun dendrogram berdasarkan estimasi kepadatan lokal, kemudian mengekstrak klaster paling stabil dari struktur tersebut. Tidak seperti K-Means atau GMM, HDBSCAN tidak memerlukan jumlah klaster sebagai parameter masukan dan dapat mendeteksi klaster dengan bentuk serta ukuran yang beragam. Salah satu keunggulan utama HDBSCAN adalah kemampuannya dalam mengidentifikasi outlier secara eksplisit, yang sangat penting dalam analisis ekspresi gen untuk memisahkan sinyal biologis dari noise. Parameter utama yang digunakan adalah min_cluster_size, yang menentukan ukuran minimum klaster yang akan dianggap valid. Dengan kemampuan deteksi outlier dan fleksibilitas terhadap bentuk klaster, HDBSCAN sangat sesuai untuk data biologis yang kompleks dan heterogen.

2.3.1.5 OPTICS

OPTICS (*Ordering Points To Identify the Clustering Structure*) merupakan metode clustering berbasis densitas yang dikembangkan sebagai perbaikan atas DBSCAN. Alih-alih menghasilkan satu partisi tetap seperti DBSCAN, OPTICS menyusun titik-titik data dalam urutan berdasarkan *reachability distance*, yang mencerminkan seberapa mudah suatu titik dapat

dicapai dari klaster yang padat. Urutan ini kemudian divisualisasikan dalam bentuk *reachability plot* yang memungkinkan analisis struktur klaster dengan berbagai tingkat kepadatan tanpa memerlukan parameter jarak tetap seperti ε. Keunggulan OPTICS terletak pada kemampuannya untuk mengidentifikasi struktur klaster yang kompleks dan multiskala, termasuk klaster dengan kepadatan tidak seragam. Namun, interpretasi hasil clustering dari OPTICS cenderung lebih subjektif dan memerlukan visualisasi tambahan, serta membutuhkan waktu komputasi yang lebih tinggi dibandingkan metode lain. Dalam analisis ekspresi gen, OPTICS berguna untuk menyaring struktur klaster laten yang mungkin tidak terdeteksi oleh metode lain.

2.3.2 Metrik Evaluasi

2.3.2.1 *Silhouette Score*

Silhouette score adalah salah satu metrik yang digunakan untuk mengevaluasi kualitas hasil clustering. Metrik ini menilai seberapa baik suatu data cocok dengan klaster tempatnya berada dibandingkan dengan klaster lain yang paling dekat. Nilai *silhouette* berkisar antara -1 hingga 1, di mana nilai yang mendekati 1 menunjukkan bahwa data berada dalam klaster yang tepat, sedangkan nilai yang mendekati 0 menunjukkan bahwa data berada di batas antara dua klaster. Jika nilai negatif, berarti data mungkin lebih cocok berada di klaster lain. Rata-rata nilai *silhouette* dari seluruh data digunakan untuk mengukur performa keseluruhan dari hasil *clustering*. Dalam konteks analisis data ekspresi gen, silhouette score berguna untuk menilai apakah gen-gen yang telah dikelompokkan memiliki pola ekspresi yang serupa di dalam satu klaster dan berbeda dari klaster lain. Metrik ini juga membantu dalam memilih metode clustering yang paling sesuai dan menentukan jumlah klaster yang optimal.

2.4 Analisis *Biclustering*

Biclustering merupakan teknik eksplorasi data yang secara simultan mengelompokkan baris dan kolom dalam sebuah matriks data. Dalam konteks analisis ekspresi gen, biclustering sangat berguna karena memungkinkan identifikasi subset gen yang menunjukkan pola ekspresi serupa pada subset kondisi atau sampel tertentu.

Pendekatan ini lebih fleksibel dibanding clustering tradisional karena tidak mengharuskan keseluruhan gen atau sampel masuk dalam satu klaster besar, melainkan memungkinkan adanya tumpang tindih antar bicluster dan variasi pola ko-ekspresi lokal.

2.4.1 Spectral *Biclustering*

Salah satu metode yang banyak digunakan adalah Spectral *Biclustering*, yang memanfaatkan dekomposisi matriks berdasarkan teori spektral untuk menemukan struktur checkerboard dalam data ekspresi gen. Metode ini bekerja dengan memetakan data ke ruang berdimensi rendah menggunakan *singular value decomposition* (SVD) atau *eigenvalue decomposition*, kemudian menerapkan clustering dalam ruang tersebut. Keunggulan utama dari *biclustering* terletak pada kemampuannya mendeteksi interaksi antara gen dan kondisi tertentu, yang seringkali terlewatkan oleh metode clustering biasa. Dalam penelitian biomedis, biclustering membantu peneliti menemukan jalur biologis atau modul gen yang hanya aktif pada subkelompok pasien tertentu, sehingga sangat relevan untuk studi stratifikasi penyakit seperti *septic shock* pada anak.

2.4.2 Metrik Evaluasi

2.4.2.1 *Mean Squared Residue* (MSR)

Mean Squared Residue (MSR) adalah metrik evaluasi yang umum digunakan dalam analisis *biclustering* untuk menilai kualitas sebuah *bicluster*. MSR mengukur sejauh mana nilai-nilai dalam sebuah bicluster menyimpang dari pola ko-ekspresi yang diharapkan, baik secara baris (gen) maupun kolom (sampel). Semakin rendah nilai MSR, semakin homogen pola ekspresi dalam *bicluster* tersebut, yang berarti gen-gen dalam *bicluster* menunjukkan pola ekspresi yang serupa pada subset sampel tertentu. Dalam analisis data ekspresi gen, MSR membantu mengidentifikasi *bicluster* yang relevan secara biologis, karena bicluster dengan MSR rendah cenderung mencerminkan hubungan ko-ekspresi yang nyata antara gen dan kondisi biologis tertentu. Oleh karena itu, nilai MSR digunakan sebagai salah satu dasar dalam pemilihan konfigurasi biclustering yang paling optimal.

2.4.2.2 *Variance Accounted For* (VAF)

Variance Accounted For (VAF) adalah metrik yang digunakan untuk mengukur seberapa besar proporsi variasi data yang berhasil dijelaskan oleh hasil *biclustering*. Dalam konteks analisis *biclustering* pada data ekspresi gen, VAF menunjukkan sejauh mana pola ko-ekspresi yang terbentuk dalam *bicluster* dapat mewakili keseluruhan variasi dalam data. Semakin tinggi nilai VAF, semakin baik kemampuan *biclustering* dalam menangkap struktur atau pola penting dari data. Nilai ini sering digunakan bersama dengan MSR untuk mengevaluasi kualitas *bicluster*, di mana konfigurasi terbaik adalah yang memiliki nilai MSR serendah mungkin dan VAF setinggi mungkin. Dengan demikian, VAF menjadi indikator penting dalam menilai seberapa representatif dan bermakna hasil pengelompokan simultan antara gen dan sampel.

2.5 Analisis *Gene Ontology Enrichment*

Analisis *Gene Ontology (GO) enrichment* merupakan pendekatan bioinformatika yang digunakan untuk mengidentifikasi kategori fungsi genetik yang secara statistik diperkaya dalam suatu kumpulan gen, misalnya gen-gen yang telah teridentifikasi signifikan melalui analisis *differential expression*, *clustering*, atau *biclustering*. GO enrichment berfokus pada tiga domain utama dalam struktur ontologi gen: biological process, molecular function, dan cellular component. Dengan memetakan kumpulan gen tersebut ke dalam kategori GO, analisis ini bertujuan untuk mengetahui apakah terdapat akumulasi gen yang lebih banyak dari yang diharapkan secara acak dalam suatu kategori fungsi tertentu, seperti jalur inflamasi, respon imun, atau apoptosis yang sering terkait dengan kondisi seperti *septic shock*.

Proses *enrichment* ini dilakukan menggunakan uji statistik seperti Fisher's Exact Test atau hypergeometric test, yang kemudian dikoreksi untuk multiple testing, umumnya menggunakan metode False Discovery Rate (FDR). Beberapa tools yang umum digunakan dalam GO enrichment antara lain DAVID, Enrichr, dan GOstats. Hasil dari analisis ini memberikan wawasan tentang mekanisme biologis yang mendasari perubahan ekspresi gen, serta membantu mengarahkan penelitian lebih lanjut dalam pengembangan hipotesis, identifikasi biomarker, dan pemahaman kontekstual terhadap kondisi klinis.

Oleh karena itu, GO enrichment analysis menjadi komponen penting dalam interpretasi data omik secara sistematis dan biologis.

2.6 *Septic Shock*

Septic shock merupakan komplikasi serius dari sepsis yang ditandai dengan penurunan tekanan darah yang persisten dan tidak dapat dikoreksi hanya dengan pemberian cairan, serta disertai dengan gangguan fungsi organ. Pada pasien anak, *septic shock* menjadi salah satu penyebab utama kematian di unit perawatan intensif (*Pediatric Intensive Care Unit*/PICU). Diagnosis septic shock pada anak tidak hanya bergantung pada tanda klinis, tetapi juga dapat didukung melalui pendekatan molekuler seperti analisis ekspresi gen, yang memungkinkan identifikasi biomarker dan pemahaman terhadap mekanisme patofisiologi penyakit secara lebih dalam.

2.7 Dataset dan Penelitian Terkait

	Deskripsi Data
Kode	GDS4274
Judul	Stratification of pediatric septic shock patients: whole blood
Ringkasan	Analisis darah lengkap (whole blood) dari anak-anak (usia ≤10 tahun) dalam 24 jam pertama setelah masuk PICU (Pediatric Intensive Care Unit) karena septic shock.
Organisme	*Homo sapiens*
Platform	GPL570: [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array
Jumlah Sampel	130
Tanggal Publikasi	2011/01/13

Dataset yang digunakan dalam penelitian ini merupakan bagian dari studi oleh Wong et al. (2009), yang bertujuan untuk mengklasifikasikan pasien anak dengan septic shock berdasarkan profil ekspresi gen dari darah perifer. Dataset ini tersedia secara publik di Gene Expression Omnibus (GEO) dengan kode akses GDS4274, dan menggunakan platform Affymetrix Human Genome U133 Plus 2.0 Array (GPL570), yang mencakup lebih dari 54.000 probe gen yang merepresentasikan lebih dari 20.000 gen unik. Sampel terdiri dari 130 anak usia ≤10 tahun, yang dibedakan ke dalam dua kelompok: 98 pasien septic shock dan 32 healthy control, dengan pengambilan darah dilakukan dalam 24 jam pertama sejak masuk Pediatric Intensive Care Unit (PICU).

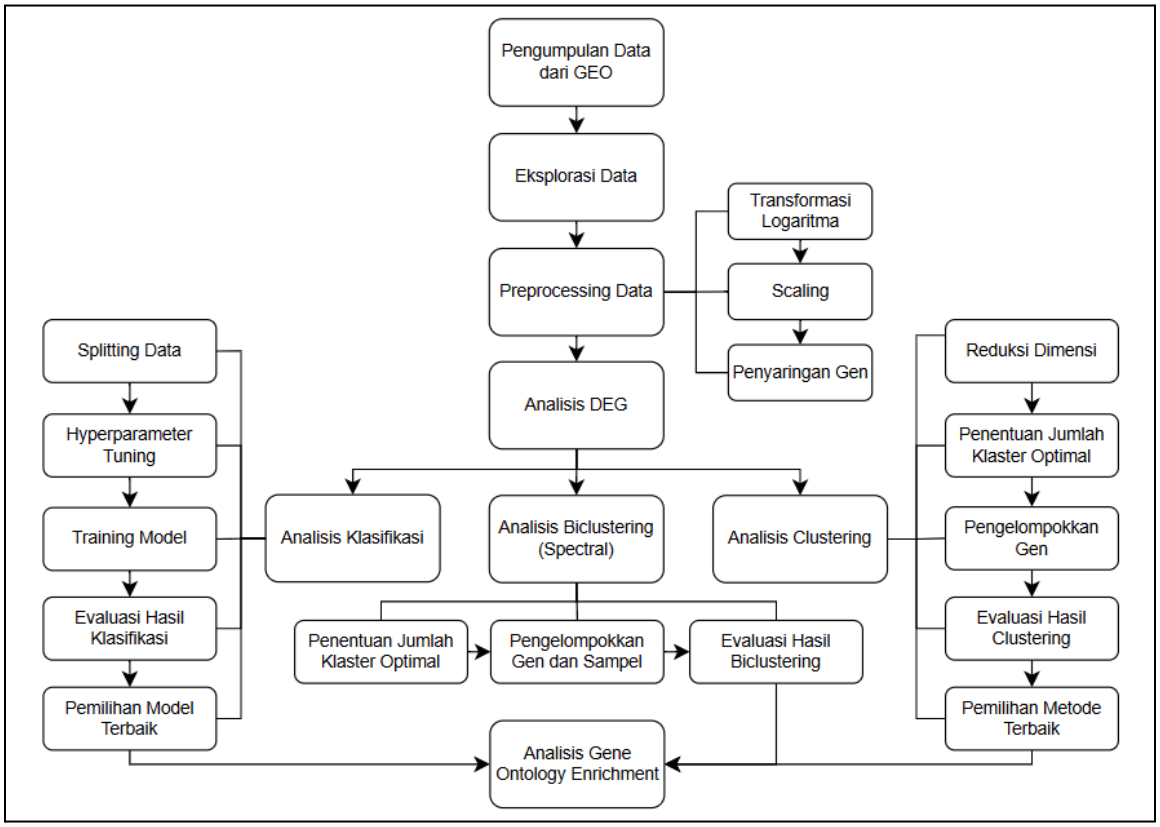
Proses pengambilan data ekspresi gen dimulai dengan pengumpulan sampel *whole blood*, kemudian dilakukan isolasi RNA total dari sel darah menggunakan prosedur standar laboratorium, seperti kit berbasis kolom atau metode ekstraksi fenol-kloroform. RNA yang telah diperoleh selanjutnya dikualifikasi dan dikuantifikasi menggunakan spektrofotometer (misalnya NanoDrop) dan analisis kualitas (misalnya Bioanalyzer). RNA berkualitas tinggi kemudian diubah menjadi cDNA (komplemen DNA) melalui proses reverse transcription.

Platform Affymetrix U133 Plus 2.0 menggunakan pendekatan mikroarray oligonukleotida, di mana probe pendek (25-mer) yang spesifik terhadap sekuens gen target diinkubasi dengan cDNA sampel yang telah diberi label fluoresen. Setelah hibridisasi, sinyal fluoresen dibaca oleh scanner dan dikonversi menjadi nilai intensitas ekspresi gen. Nilai-nilai ini selanjutnya ditransformasikan menjadi data numerik berbasis log dan dinormalisasi antar sampel menggunakan metode seperti MAS5 atau RMA (Robust Multiarray Average) untuk mengurangi efek batch dan bias teknis.

Dataset GDS4274 secara khusus telah melalui preprocessing awal oleh tim GEO, namun dalam penelitian ini dilakukan pemrosesan tambahan berupa transformasi logaritmik, standardisasi (z-score scaling), serta seleksi gen berdasarkan variansi tertinggi (top 5% berdasarkan persentil ke-95), untuk memastikan bahwa hanya gen-gen yang menunjukkan variabilitas ekspresi tinggi yang digunakan dalam analisis lanjutan. Proses

ini penting untuk mengurangi noise biologis dan teknis, serta meningkatkan efisiensi dan performa algoritma klasifikasi dan clustering yang digunakan dalam studi.

III. Metode Penelitian



3.1 Pengumpulan Data

Analisis diawali dengan pengumpulan data dari Gene Expression Omnibus (GEO), dengan menggunakan dataset GDS4274 yang memuat profil ekspresi gen darah lengkap (*whole blood*) pada anak-anak berusia ≤10 tahun dalam 24 jam pertama setelah masuk ke Pediatric Intensive Care Unit (PICU) akibat *septic shock.*

3.2 Eksplorasi Data

Langkah selanjutnya dimulai dengan ekstraksi data ekspresi gen dari dataset, yang kemudian dilanjutkan dengan eksplorasi data. Tahapan ini mencakup pemeriksaan distribusi *disease state* untuk mengetahui jumlah sampel pada masing-masing kelompok, yaitu *healthy control* dan *septic shock*, serta identifikasi jumlah gen yang tersedia dan analisis variansi gen.

3.3 *Preprocessing* Data

Data ekspresi gen selanjutnya diproses melalui tahapan transformasi logaritmik, penskalaan (*scaling*), dan penyaringan berdasarkan nilai variansi. Pada tahap penyaringan ini, hanya 5% gen dengan variansi tertinggi, berdasarkan persentil ke-95, yang dipertahankan untuk dianalisis lebih lanjut.

3.4 Analisis *Differentially Expressed Genes* (DEG)

Gen yang diperoleh dari hasil penyaringan kemudian dianalisis lebih lanjut menggunakan analisis *Differentially Expressed Genes* (DEG) untuk mengidentifikasi gen-gen yang menunjukkan perbedaan ekspresi signifikan antara kelompok *healthy control* dan *septic shock*. Gen dikategorikan sebagai *upregulated* jika memiliki nilai log2 *fold change* positif yang signifikan, sedangkan gen *downregulated* ditandai dengan nilai log2 *fold change* negatif yang signifikan. Signifikansi statistik ditentukan berdasarkan nilai *adjusted p-value* (misalnya FDR < 0.05). Hasil analisis ini divisualisasikan dalam bentuk *volcano plot* untuk menampilkan hubungan antara tingkat perubahan ekspresi (log2 fold change) dan signifikansinya (–log10 *p-value*), sehingga gen-gen yang paling berbeda ekspresinya dapat diidentifikasi secara visual.

3.5 Analisis Klasifikasi

Gen-gen yang dinyatakan signifikan dalam analisis DEG, yaitu yang memiliki nilai *adjusted p-value* kurang dari 0,05, akan digunakan sebagai fitur atau variabel input dalam analisis klasifikasi. Sebelum pelatihan model, data akan dibagi menjadi data latih dan data uji dengan proporsi 80:20 (*train-test split*) untuk memastikan evaluasi model yang objektif. Tahapan klasifikasi diawali dengan *hyperparameter tuning* untuk masing-masing model guna memperoleh konfigurasi parameter terbaik. Model klasifikasi yang digunakan mencakup Regresi Logistik, *K-Nearest Neighbors* (KNN), dan *Support Vector Machine* (SVM). Kinerja dari ketiga model akan dievaluasi menggunakan metrik *accuracy* untuk menentukan model dengan performa terbaik. Model yang terpilih selanjutnya akan dievaluasi lebih lanjut menggunakan *confusion matrix*, serta dihitung nilai *precision*, *recall*, dan *F1-score*. Analisis kontribusi fitur juga akan dilakukan pada model terbaik menggunakan nilai SHAP (*SHapley Additive exPlanations*) untuk

mengidentifikasi gen-gen yang memberikan kontribusi paling besar dalam proses klasifikasi antara *healthy control* dan *septic shock*.

3.6 Analisis *Clustering*

Gen-gen yang dinyatakan signifikan dalam analisis DEG, yaitu dengan nilai adjusted p-value kurang dari 0,05, selanjutnya akan dikelompokkan menggunakan beberapa metode clustering, yaitu K-Means, *Hierarchical Clustering*, dan Gaussian Mixture Model (GMM). Sebelum dilakukan pengelompokan, terlebih dahulu dilakukan Principal Component Analysis (PCA) untuk mereduksi dimensi data dan mengurangi noise, sehingga pola dalam data dapat lebih mudah dikenali oleh algoritma clustering. Setelah PCA dilakukan, langkah selanjutnya adalah menentukan jumlah klaster yang optimal untuk masing-masing metode. Untuk K-Means, jumlah klaster optimal akan ditentukan menggunakan metode elbow. Pada *Hierarchical Clustering*, penentuan jumlah klaster dilakukan berdasarkan pengamatan dendrogram. Sementara itu, untuk GMM, jumlah klaster terbaik ditentukan berdasarkan nilai AIC (*Akaike Information Criterion*) dan BIC (Bayesian *Information Criterion*).

Ketiga metode tersebut tidak memiliki kemampuan untuk mendeteksi *outlier*, sehingga seluruh gen akan dimasukkan ke dalam salah satu klaster. Jika hasil pengelompokan dari ketiga metode tersebut kurang memuaskan, maka akan digunakan metode *clustering* yang mampu mendeteksi *outlier*, yaitu HDBSCAN dan OPTICS. Untuk HDBSCAN, parameter utama yang akan diuji adalah *min_cluster_size*, sedangkan pada OPTICS, parameter *min_samples* akan divariasikan untuk menemukan struktur klaster yang paling sesuai.

Hasil *clustering* dari seluruh metode akan dievaluasi menggunakan metrik *silhouette score*. Khusus untuk HDBSCAN dan OPTICS, pemilihan metode terbaik tidak hanya mempertimbangkan nilai *silhouette score*, tetapi juga jumlah gen yang terdeteksi sebagai *outlier* serta proporsi anggota dalam setiap klaster. Metode yang terpilih sebagai metode *clustering* terbaik akan digunakan untuk visualisasi hasil pengelompokan gen dan ditampilkan dalam bentuk grafik *clustering* serta *heatmap* ekspresi gen, untuk memberikan gambaran visual terhadap pola-pola ekspresi antar klaster yang terbentuk.

3.7 Analisis *Biclustering*

Gen yang dinyatakan signifikan dalam analisis DEG dengan nilai *adjusted p-value* kurang dari 0,05, beserta seluruh sampel, akan dikelompokkan secara simultan menggunakan metode *Spectral Biclustering*. Metode ini memungkinkan identifikasi pola ko-ekspresi gen dalam subset tertentu dari sampel. Akan dilakukan uji coba dengan berbagai jumlah klaster, dan jumlah klaster yang menghasilkan nilai *mean squared residue* (MSR) terendah serta *variance accounted for* (VAF) tertinggi akan dipilih sebagai konfigurasi yang paling optimal. Untuk memvisualisasikan hasil *biclustering*, akan ditampilkan *heatmap* ekspresi gen yang disusun ulang berdasarkan hasil pengelompokan, sehingga pola-pola ekspresi dalam setiap bicluster dapat diamati secara lebih jelas.

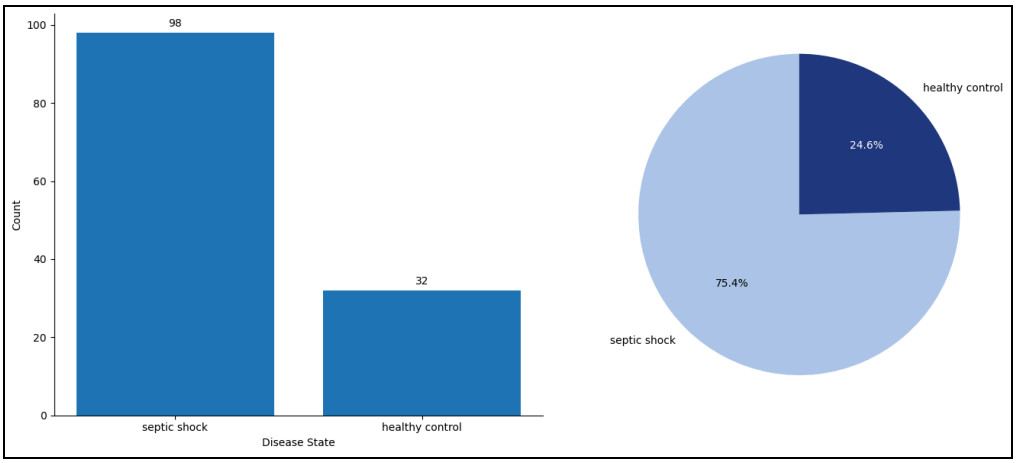
3.8 Analisis *Gene Ontology Enrichment*

Nama-nama gen yang tergabung dalam klaster hasil analisis *clustering* dan *biclustering* akan ditampilkan dalam tahap analisis *gene ontology*. Selain daftar gen, tahap ini juga akan menyajikan visualisasi hasil anotasi berupa plot untuk tiga kategori utama: *biological process* (BP), *molecular function* (MF), dan *cellular Component* (CC). Penyajian ini bertujuan untuk memberikan gambaran awal mengenai potensi fungsi biologis gen-gen tersebut serta keterkaitannya dengan kondisi *septic shock*, sebagai dasar untuk analisis lanjutan yang lebih mendalam.

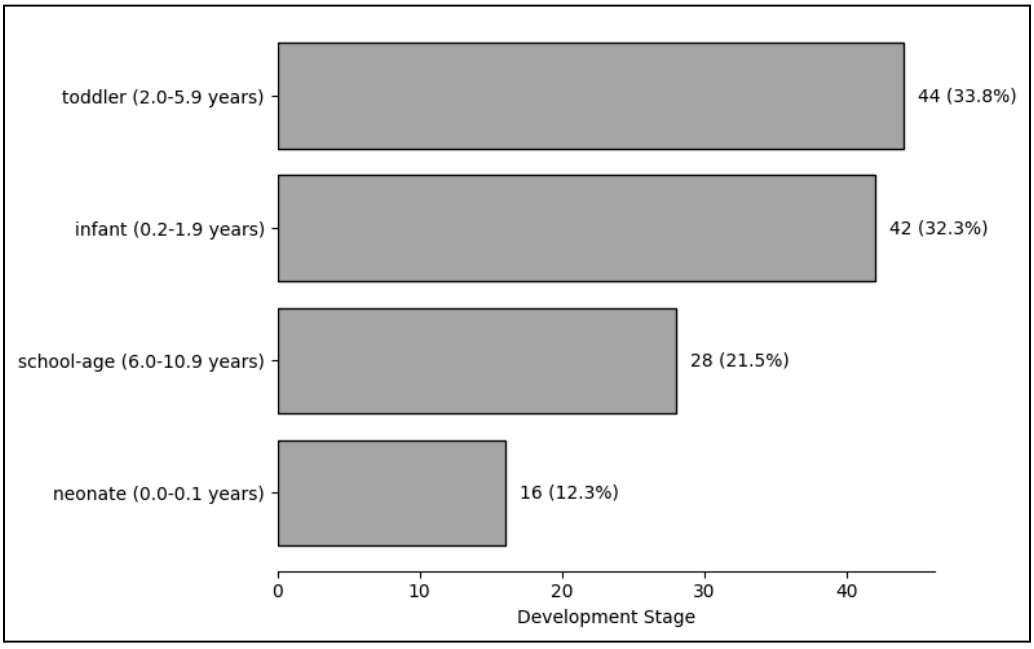
IV. Hasil dan Pembahasan

4.1 Eksplorasi dan *Preprocessing* Data

Dari 130 sampel, 75,4% (98) di antaranya merupakan kelompok sampel dengan *septic shock,* sedangkan 24,6% (32) lainnya merupakan kelompok sampel *healthy control*.

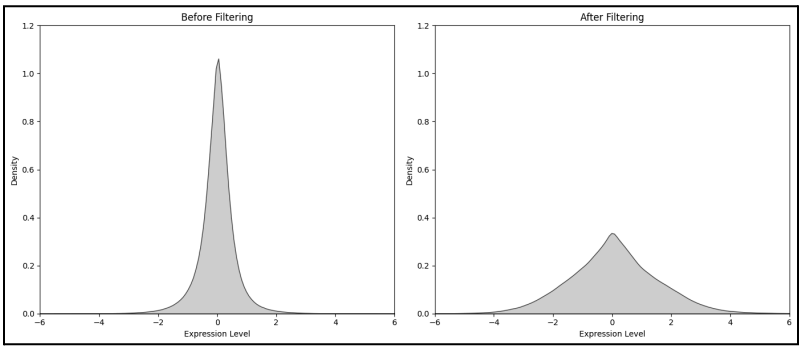


Berdasarkan tahap perkembangan, 33,8% (44) di antaranya merupakan kelompok *toddler* dengan rentang umur 2 hingga 5,9 tahun, 32,3% (42) diantaranya merupakan kelompok *infant* dengan rentang umur 2 bulan hingga 1,9 tahun, 21,5% (28) di antaranya merupakan kelompok *school-age* dengan rentang umur 6 hingga 10,9 tahun, dan 12,3% (16) sisanya merupakan kelompok *neonate* dengan rentang umur 0 hingga 1 bulan.



Jumlah gen dalam data sebanyak 54.675 gen (tanpa *missing value*) dengan rentang nilai variansi (setelah data ekspresi gen ditransformasi logaritma),

	Rentang Nilai Variansi	Jumlah Gen
Sebelum Penyaringan Gen	[0,0092, 15,4369]	54.675
Setelah Penyaringan Gen	[0,9376, 15,4369]	2.734



Melalui penyaringan gen berdasarkan persentil ke-95, gen yang diambil adalah 5% gen dengan variansi tertinggi, sehingga menyisakan 2.734 gen. Gen yang akan digunakan untuk analisis selanjutnya akan melalui tahap standardisasi.

4.2 Analisis *Differentially Expressed Genes* (DEG)

Dengan menggunakan uji-t, berdasarkan *adjusted p-value* dengan metode False Discovery Rate - Benjamini-Hochberg (FDR-BH),

H_0 : Tidak terdapat perbedaan yang signifikan dalam rata-rata ekspresi gen antara kelompok *septic shock* dan *healthy control*.

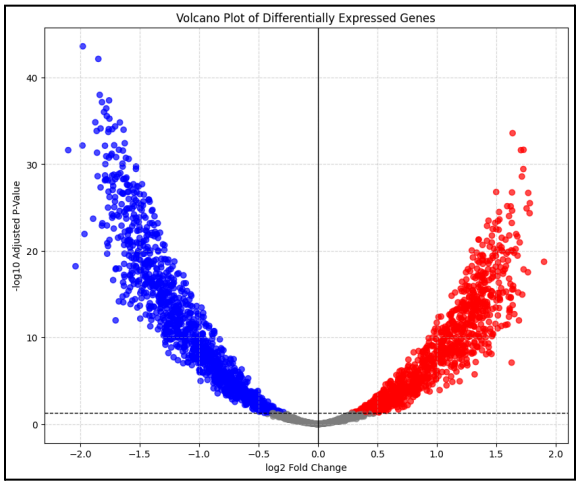
H_1 : Terdapat perbedaan yang signifikan dalam rata-rata ekspresi gen antara kelompok *septic shock* dan *healthy control*.

Gen dinyatakan berbeda dalam rata-rata ekspresi gen antara kelompok *septic shock* dan *healthy control* secara signifikan apabila nilai *adjusted p-value* lebih besar dari 0,05. Gen yang menunjukkan peningkatan tingkat ekspresi pada kondisi *septic shock* dibandingkan dengan *healthy control* disebut sebagai gen *upregulated*, yang ditandai dengan nilai *log fold change* (logFC) positif. Sementara itu, gen yang mengalami penurunan tingkat

ekspresi disebut sebagai gen *downregulated*, dengan nilai logFC negatif. Berdasarkan kriteria tersebut, ditampilkan sebagian gen yang signifikan:

	Jumlah	**ID Gen**	**Simbol Gen**	**Nama Gen**
Upregulated	959	1405_i_at	CCL5	C-C motif chemokine ligand 5
		1552316_a_at	GIMAP1	GTPase, IMAP family member 1
		1553132_a_at	TC2N	tandem C2 domains, nuclear
		1553177_at	SH2D1B	SH2 domain containing 1B
		1553589_a_at	PDZK1IP1	PDZK1 interacting protein 1
Downregulated	1.170	1552263_at	MAPK1	mitogen-activated protein kinase 1
		1552274_at	PXK	PX domain containing serine/threonine kinase like
		1552485_at	LACTB	lactamase beta
		1552553_a_at	NLRC4	NLR family CARD domain

		containing 4
1552670_a_at	PPP1R3B	protein phosphatase 1 regulatory subunit 3B



Sebanyak 959 gen yang signifikan merupakan gen *upregulated* dan 1.170 gen yang signifikan merupakan gen *downregulated.* Dari 2.734 gen yang digunakan dalam analisis DEG, hanya 2.129 gen signifikan yang akan digunakan untuk analisis selanjutnya, yaitu analisis klasifikasi.

4.3 Analisis Klasifikasi

Sebanyak 130 sampel akan dibagi dalam proporsi 80:20 menjadi 104 sampel data latih dan 26 sampel data uji. Model yang digunakan dalam analisis klasifikasi adalah model *Logistic Regression* (LR), *K-Nearest Neighbors* (K-NN), dan *Support Vector Machine* (SVM).

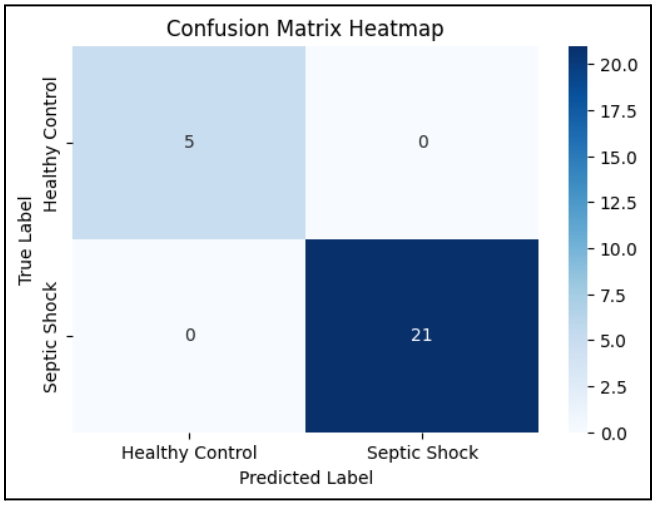
Model	Hyperparameter Terbaik	Akurasi Data Latih	Akurasi Data Uji
LR	{'solver': 'liblinear',	0,981	1,000

	'penalty': 'l1', 'max_iter': 500, 'C': 0.1}		
K-NN	{'weights': 'uniform', 'n_neighbors': 3, 'metric': 'manhattan'}	0,990	1,000
SVM	{'kernel': 'linear', 'gamma': 'scale', 'C': 0.1}	1,000	1,000

Berdasarkan akurasi, nilai akurasi tertinggi untuk data uji diperoleh oleh model SVM dengan akurasi data latih dan data uji sebesar 100%. Walaupun sama-sama memperoleh akurasi sebesar 100% pada data uji, model LR dan K-NN memiliki akurasi data latih yang lebih rendah, sehingga diputuskan model SVM sebagai model terbaik.

Karena hasil analisis klasifikasi tanpa *resampling* sudah memuaskan, maka *resampling* untuk mengatasi ketidakseimbangan kelas pada *disease state* tidak lagi diperlukan.

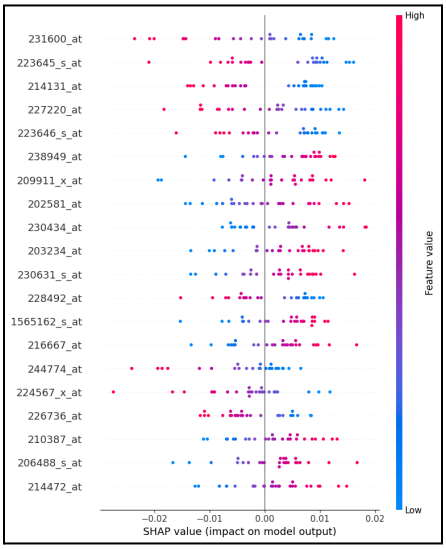
Berikut adalah *classification report* untuk model SVM:



Model mengklasifikasikan keseluruhan 5 pasien *healthy control* dan 21 pasien *septic shock* dengan benar.

Kelompok	Metrik	Nilai	Interpretasi
Healthy Control	*Precision*	1,00	Semua prediksi sebagai "*healthy control*" benar seluruhnya. Tidak ada *false positive.*
	Recall	1,00	Semua anggota kelompok "*healthy control*" berhasil dikenali.
	F1-Score	1,00	Model sangat baik dalam mengenali kelompok "*healthy control*".
Septic Shock	*Precision*	1,00	Semua prediksi sebagai "*septic shock*" benar seluruhnya. Tidak ada *false positive.*
	Recall	1,00	Semua anggota kelompok "*septic shock*" berhasil dikenali.
	F1-Score	1,00	Model sangat baik dalam mengenali kelompok "*septic shock*".

Untuk melihat gen yang berkontribusi tertinggi dalam klasifikasi berdasarkan model terbaik yaitu model SVM, akan digunakan metode seperti *SHapley Additive exPlanations* (SHAP).

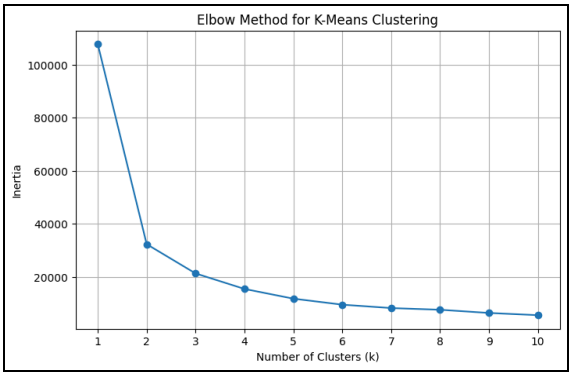


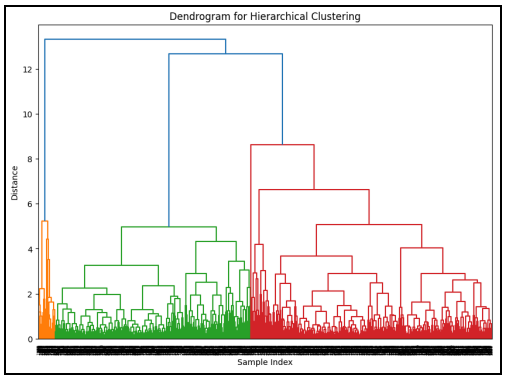
Berikut adalah sebagian gen yang berkontribusi tertinggi dalam klasifikasi menggunakan model SVM berdasarkan nilai rata-rata SHAP:

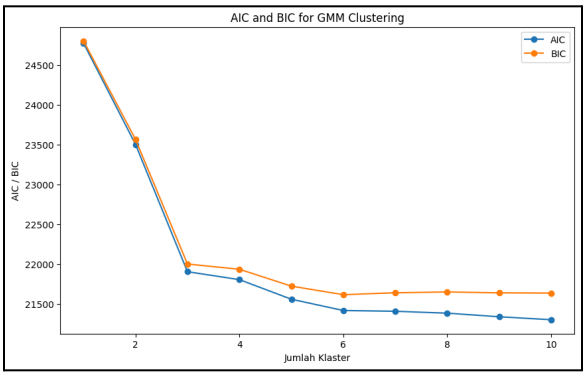
ID Gen	Simbol Gen	Nama Gen	Nilai Mean SHAP
231600_at	CLEC12B	C-type lectin domain family 12 member B	0,008695
223645_s_at	TXLNGY	taxilin gamma pseudogene, Y-linked	0,008544
214131_at	TXLNGY	taxilin gamma pseudogene, Y-linked	0,007876
227220_at	NFXL1	nuclear transcription factor, X-box binding like 1	0,007651
223646_s_at	TXLNGY	taxilin gamma pseudogene, Y-linked	0,007209

4.4 Analisis *Clustering*

Untuk analisis *clustering*, gen yang digunakan merupakan gen yang dinyatakan signifikan dalam analisis DEG, yaitu sebanyak 2.129 gen. Akan dilakukan PCA untuk mereduksi sampel sebelum dilakukan *clustering*. Metode *clustering* yang digunakan dalam analisis ini adalah K-Means, *Hierarchical Clustering* dengan *average link*, dan Gaussian Mixture Model. Sebelum menerapkan *clustering*, akan dilakukan pencarian jumlah klaster optimal dengan menggunakan metode Elbow untuk K-Means, visualisasi dendogram untuk *Hierarchical Clustering*, serta nilai AIC dan BIC untuk Gaussian Mixture Model. Berikut adalah ringkasan jumlah klaster optimal dan nilai *silhouette score* untuk setiap metode:







Metode	**Metode Penentuan Jumlah Klaster Optimal**	**Jumlah Klaster Optimal**	**Skor *Silhouette***
K-Means	Metode Elbow	2	0,627
Hierarchical	Visualisasi Dendogram	4	0,541
Gaussian Mixture Model	AIC	10	0,358
	BIC	6	0,423

Metode K-Means memperoleh skor *silhouette* tertinggi sebesar 0,627. Walaupun dapat dinyatakan cukup, pemisahan klaster belum optimal. Sedangkan dua metode lainnya menghasilkan skor *silhouette* yang lebih rendah menunjukkan *clustering* yang lemah. Jika divisualisasikan, akan terdapat lebih banyak data dengan klaster yang berbeda yang sangat berdekatan atau tumpang tindih sehingga batas antar klaster menjadi kurang jelas. Hal ini mungkin disebabkan oleh ketidakcocokkan data untuk *clustering*.

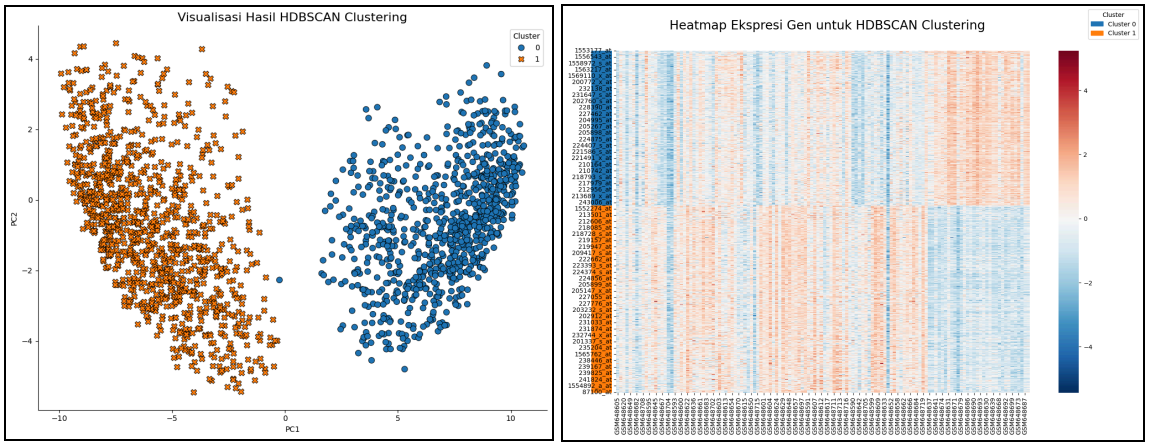
Oleh karena itu, akan diterapkan metode *clustering* yang dapat mendeteksi *outlier*. Gen yang terdeteksi sebagai *outlier* tidak akan ditampilkan dalam visualisasi sehingga batas antar klaster menjadi lebih jelas.

Metode	**Parameter**	**Nilai Parameter Terbaik**	**Jumlah Klaster**	**Skor *Silhouette***
HDBSCAN	min_cluster_size	50	2	0,711
OPTICS	min_samples	35	2	0,740

Metode OPTICS memberikan skor *silhouette* yang paling tinggi dibandingkan metode lainnya, yaitu sebesar 0,740.

Metode	**-1 (Outlier)**	**0**	**1**
HDBSCAN	236	856	1037
OPTICS	1180	914	35

Walaupun memberikan nilai *silhouette score* yang lebih tinggi, metode OPTICS mengidentifikasi jumlah gen yang berupa *outlier* lebih banyak dari metode HDBSCAN. Hal ini akan mengakibatkan banyak informasi akan terbuang karena gen yang akan diidentifikasi dan dianalisis lebih lanjut adalah gen-gen yang bukan *outlier*. Perbandingan proporsi jumlah anggota klaster yang dihasilkan oleh metode OPTICS juga cukup jauh atau tidak seimbang. Maka dari itu, kami memilih metode HDBSCAN sebagai metode terbaik yang menghasilkan klaster yang lebih proporsional dengan lebih sedikit gen yang teridentifikasi sebagai *outlier* (lebih sedikit informasi yang terbuang).



Berdasarkan hasil *clustering* menggunakan metode HDBSCAN, gen-gen dalam dataset berhasil dikelompokkan ke dalam dua klaster utama, yaitu Cluster 0 dan Cluster 1, tanpa menyertakan gen yang diklasifikasikan sebagai *outlier*. Visualisasi PCA menunjukkan gen-gen dalam Cluster 0 ditandai dengan titik berwarna biru, sementara Cluster 1 ditandai dengan tanda silang oranye. Pemisahan yang jelas antara dua klaster ini menunjukkan bahwa terdapat pola ekspresi yang berbeda secara signifikan antar kelompok gen.

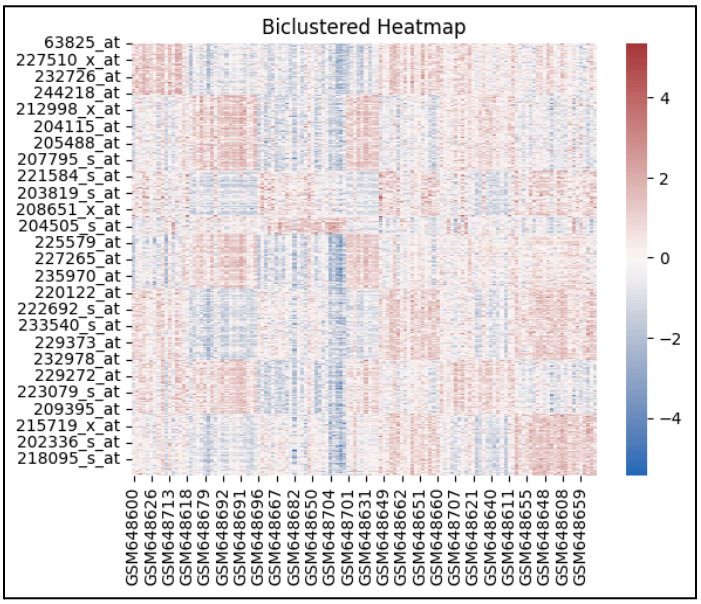
Selanjutnya, *heatmap* menyajikan pola ekspresi dari gen-gen yang telah dikelompokkan berdasarkan hasil *clustering* tersebut. Baris pada *heatmap* mewakili gen, sedangkan kolom mewakili sampel. Warna biru menunjukkan tingkat ekspresi yang lebih rendah, sedangkan warna merah menunjukkan tingkat ekspresi yang lebih tinggi. Terlihat bahwa gen-gen dalam Cluster 0 memiliki pola ekspresi yang cenderung mirip satu sama lain namun berbeda dengan gen-gen dalam Cluster 1, yang mengindikasikan bahwa pengelompokan ini berhasil memisahkan gen berdasarkan karakteristik ekspresi mereka di seluruh sampel.

4.5 Analisis *Biclustering*

Untuk analisis *biclustering*, gen yang digunakan merupakan gen yang dinyatakan signifikan dalam analisis DEG, yaitu sebanyak 2.129 gen. Dengan menggunakan Spectral Biclustering, jumlah *bicluster* yang optimal akan dipilih melalui metrik evaluasi seperti MSR dan VAF. Jumlah *bicluster* dengan rata-rata nilai MSR terendah dan rata-rata nilai VAF tertinggi akan dipilih sebagai jumlah *bicluster* optimal.

	MSR	VAF
2	0,501	0,277
3	0,560	0,353
4	0,465	0,408
5	0,431	0,466
6	0,376	0,514
7	0,355	0,556
8	0,346	0,581

Berdasarkan percobaan jumlah *bicluster* 2 hingga 8, jumlah *bicluster* 8 memperoleh nilai rata-rata MSR terendah dan nilai rata-rata VAF tertinggi.



Walaupun terlihat samar, hasil biclustering dengan jumlah 8 bicluster menunjukkan pola *checkerboard.* Berikut adalah proporsi jumlah gen dan sampel untuk bicluster,

	0	1	2	3	4	5	6	7

Gen	256	374	226	84	268	350	269	302
Sampel	14	21	20	5	9	17	21	23

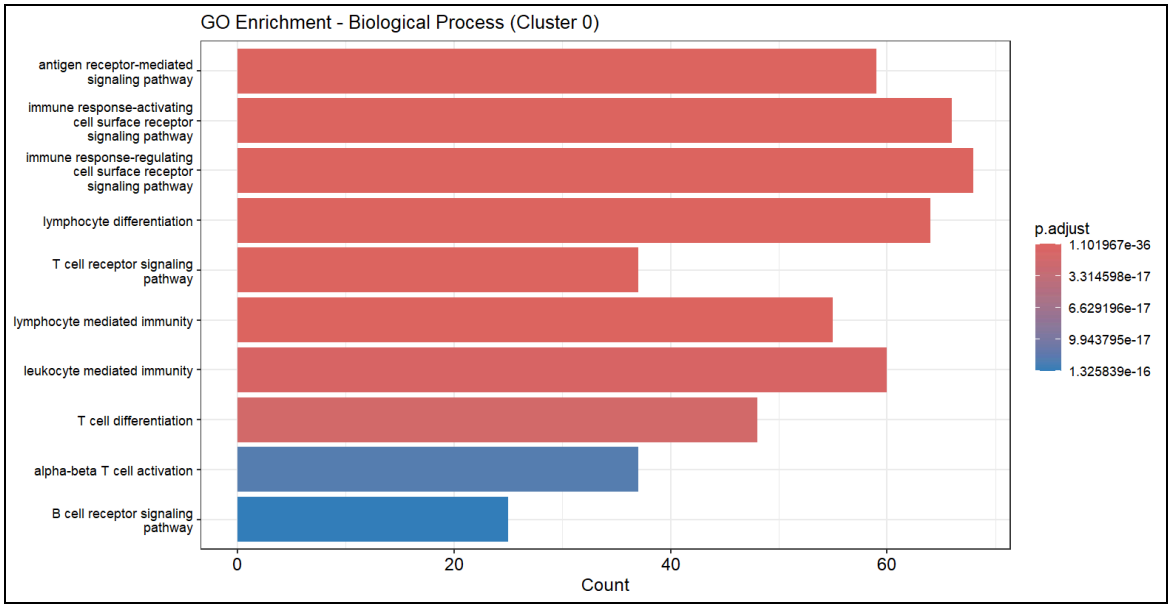
4.6 Analisis *Gene Ontology Enrichment*

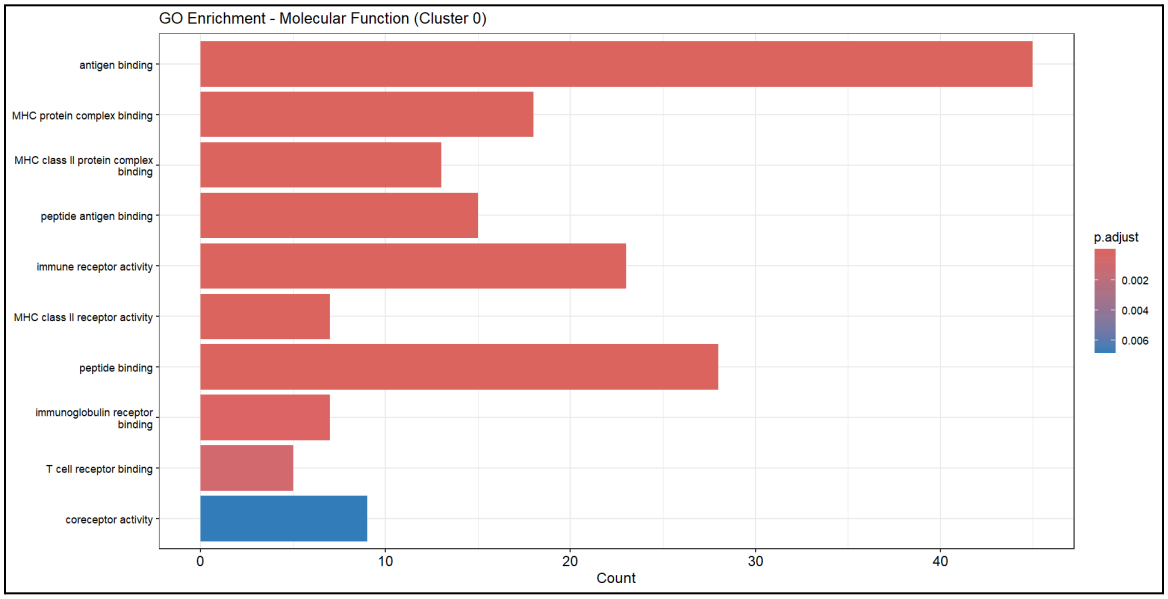
4.6.1 Analisis *Clustering*

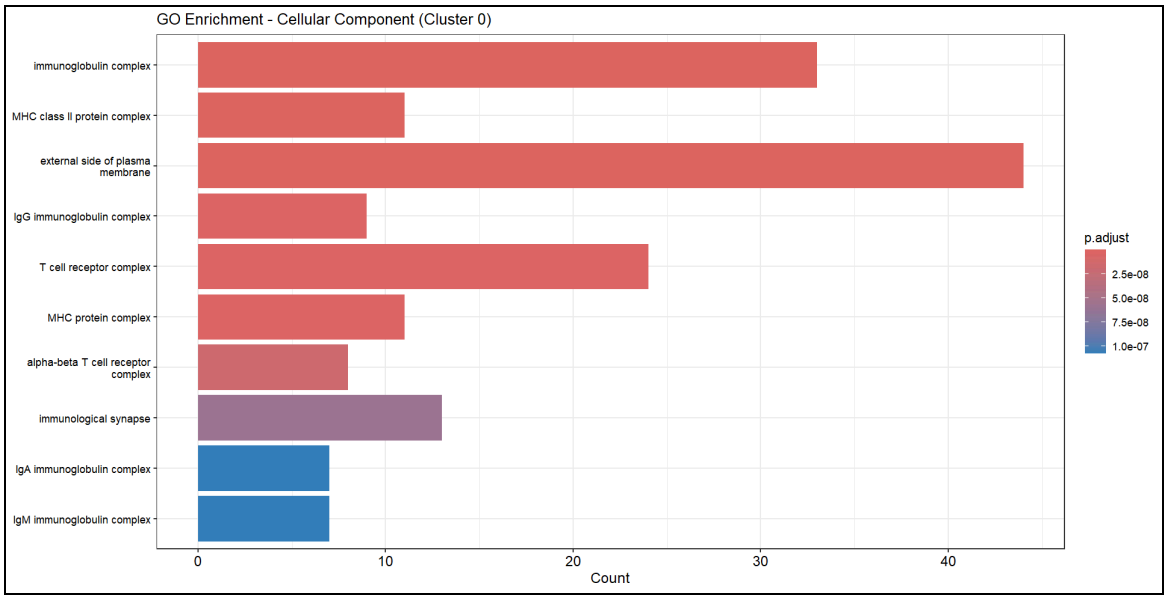
Cluster 0

Dengan jumlah gen sebanyak 856 gen dalam Cluster 0, berikut adalah sebagian nama-nama gen anggota cluster dan visualisasi dan interpretasi GO *Enrichment* untuk *biological process, molecular function,* dan *cellular component*,

ID Gen	Simbol Gen	Nama Gen
1405_i_at	CCL5	C-C motif chemokine ligand 5
1552316_a_at	GIMAP1	GTPase, IMAP family member 1
1553132_a_at	TC2N	tandem C2 domains, nuclear
1553177_at	SH2D1B	SH2 domain containing 1B
1553678_a_at	ITGB1	integrin subunit beta 1


GO Enrichment - Biological Process (Cluster 0)
antigen receptor-mediated signaling pathway
immune response-activating cell surface receptor signaling pathway
immune response-regulating cell surface receptor signaling pathway
lymphocyte differentiation
T cell receptor signaling pathway
lymphocyte mediated immunity
leukocyte mediated immunity
T cell differentiation
alpha-beta T cell activation
B cell receptor signaling pathway
0
20
40
60
Count
p.adjust
1.101967e-36
3.314598e-17
6.629196e-17
9.943795e-17
1.325839e-16


GO Enrichment - Molecular Function (Cluster 0)
antigen binding
MHC protein complex binding
MHC class II protein complex binding
peptide antigen binding
immune receptor activity
MHC class II receptor activity
peptide binding
immunoglobulin receptor binding
T cell receptor binding
coreceptor activity
0
10
20
30
40
Count
p.adjust
0.002
0.004
0.006



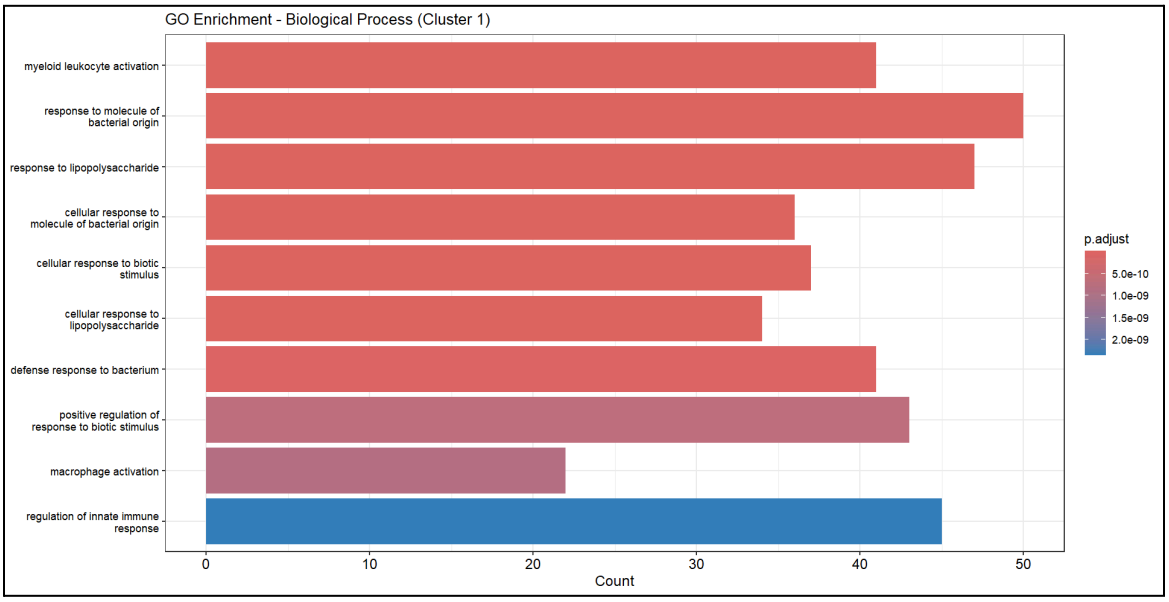
Pada Cluster 0, sebagian besar gen yang masuk ke dalam kelompok ini berperan dalam fungsi imun, komunikasi antar sel, dan regulasi ekspresi gen. Contohnya pada gen CCL5, gen ini dikenal sebagai *chemokine* yang merupakan protein kecil pemandu pergerakan sel-sel imun ke lokasi inflamasi atau infeksi. Ini menunjukkan bahwa *cluster* ini berkaitan erat dengan kondisi aktivasi sistem imun. Gen GIMAP1 juga termasuk dalam keluarga protein yang terlibat dalam kelangsungan hidup sel T dan B, dua jenis sel darah putih yang memainkan peran penting dalam sistem imun tubuh. Gen SH2D1B berperan dalam pensinyalan sel imun, sedangkan gen ITGB1 adalah subunit dari integrin, protein penting untuk adhesi sel dan komunikasi antar sel dengan matriks ekstraseluler.

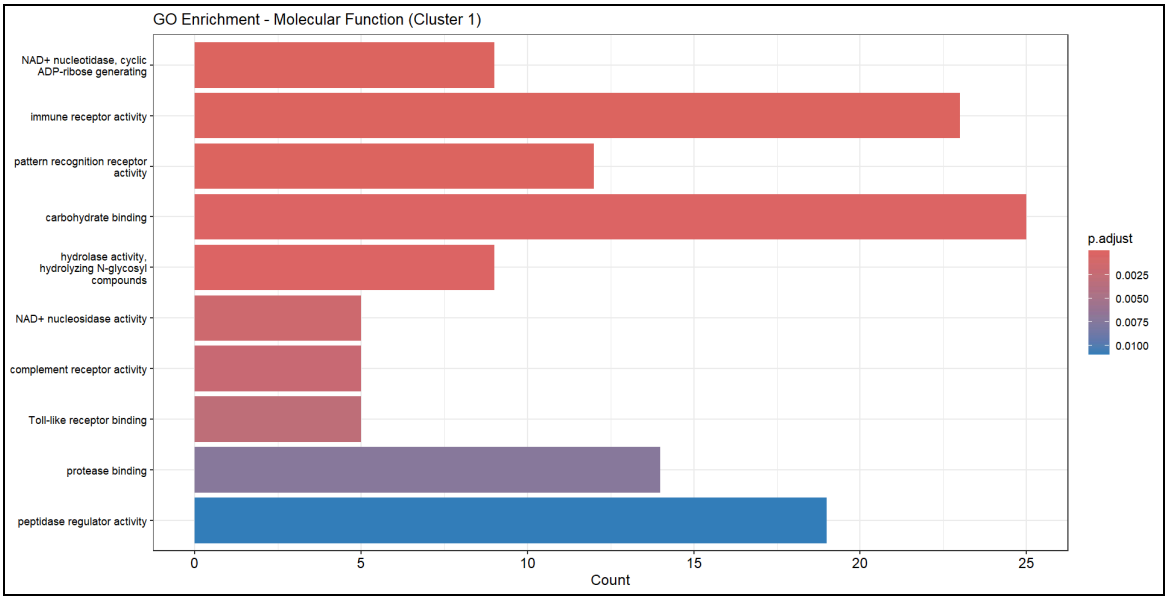
Analisis GO menunjukkan bahwa gen-gen dalam *cluster* ini berperan dalam proses imun adaptif, khususnya dalam jalur pensinyalan reseptor antigen, aktivasi dan diferensiasi limfosit, serta respon regulatif terhadap aktivasi reseptor permukaan sel. Secara fungsi molekuler, gen-gen ini terlibat dalam pengikatan antigen dan kompleks MHC kelas II, serta berperan sebagai reseptor imun. Dari sisi komponen seluler, mereka terlokalisasi pada kompleks imunoglobulin, kompleks MHC, dan sisi luar membran plasma, termasuk dalam kompleks reseptor sel T, yang menandai keterlibatannya dalam pengenalan dan penyajian antigen kepada sel T.

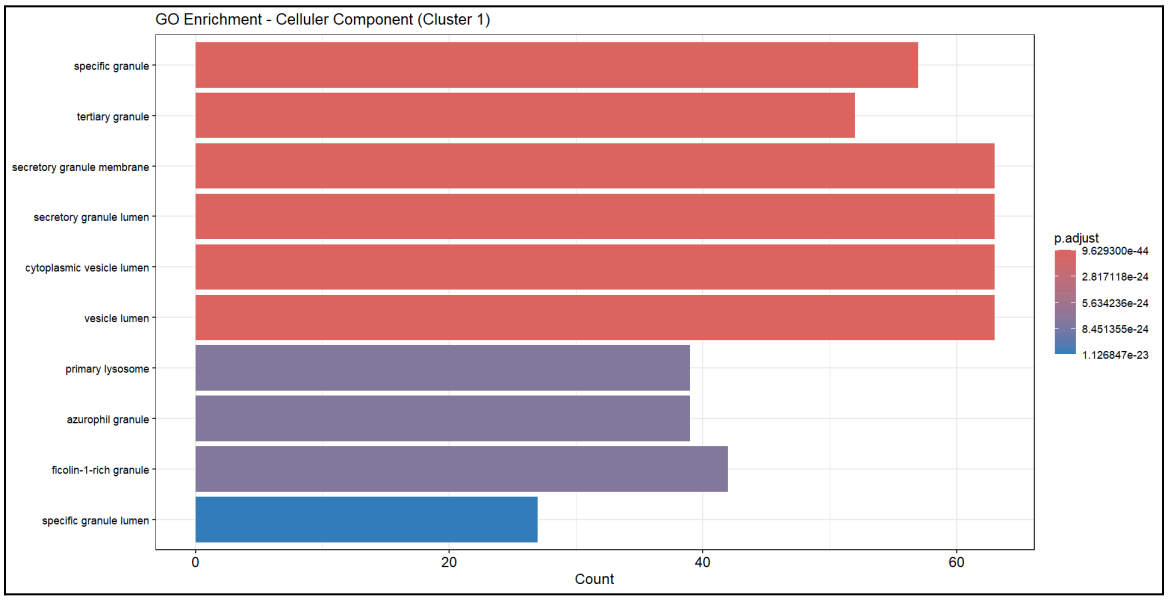
Cluster 1

Dengan jumlah gen sebanyak 1.037 gen dalam Cluster 1, berikut adalah sebagian nama-nama gen anggota cluster dan visualisasi dan interpretasi GO *Enrichment* untuk *biological process, molecular function,* dan *cellular component*,

ID Gen	Simbol Gen	Nama Gen
1552263_at	MAPK1	mitogen-activated protein kinase 1
1552274_at	PXK	PX domain containing serine/threonine kinase like
1552485_at	LACTB	lactamase beta
1552553_a_at	NLRC4	NLR family CARD domain containing 4
1552670_a_at	PPP1R3B	protein phosphatase 1 regulatory subunit 3B







Beberapa gen yang termasuk dalam Cluster 1 antara lain PXK (PX domain containing serine/threonine kinase like), LACTB (lactamase beta), NLRC4 (NLR family CARD domain containing 4), dan PPP1R3B (protein phosphatase 1 regulatory subunit 3B). PXK berperan dalam regulasi pensinyalan sel, sedangkan LACTB diduga terlibat dalam metabolisme mitokondria. NLRC4 merupakan sensor penting dalam sistem imun bawaan yang membentuk inflammasom untuk merespon infeksi bakteri. PPP1R3B terlibat dalam regulasi metabolisme glikogen, namun juga menunjukkan hubungan dengan aktivitas imun dalam konteks tertentu.

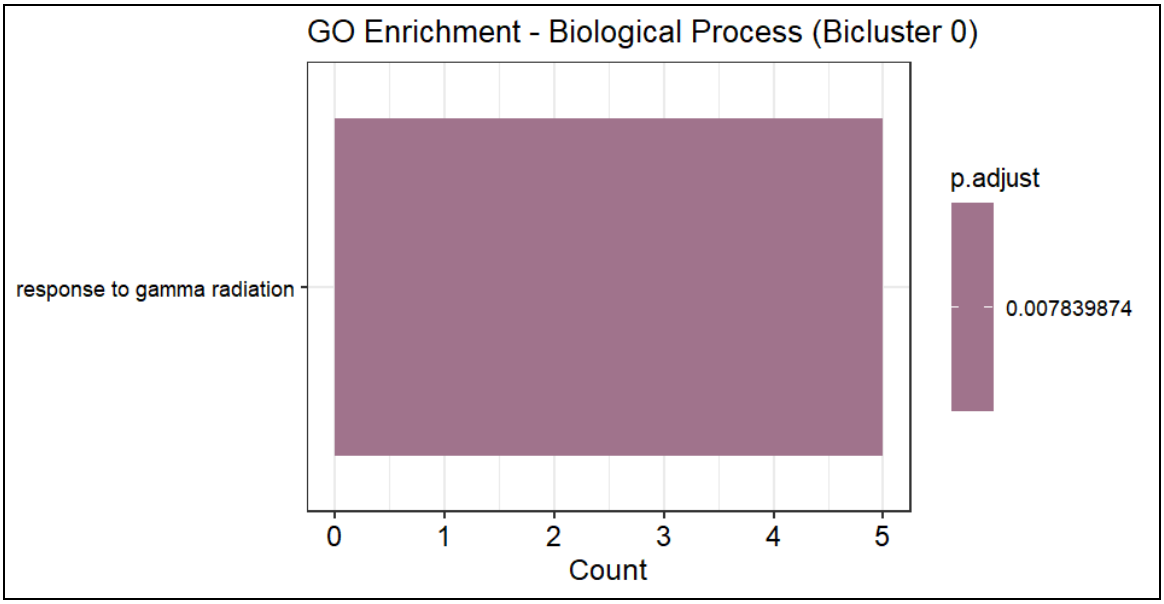
Analisis GO menunjukkan bahwa gen-gen ini aktif dalam proses biologis seperti aktivasi leukosit mieloid dan respon seluler terhadap molekul bakteri, terutama lipopolisakarida, menandakan keterlibatan dalam sistem imun bawaan. Secara fungsi molekuler, mereka berperan sebagai reseptor imun dan pengenal pola, serta memiliki aktivitas enzimatik seperti hidrolase dan NAD^+ nukleotidase. Dari sisi komponen seluler, gen-gen ini terutama berada di granula spesifik dan tersier, serta membran dan lumen granula sekretorik dan vesikel, yang berfungsi sebagai lokasi utama penyimpanan dan sekresi mediator imun.

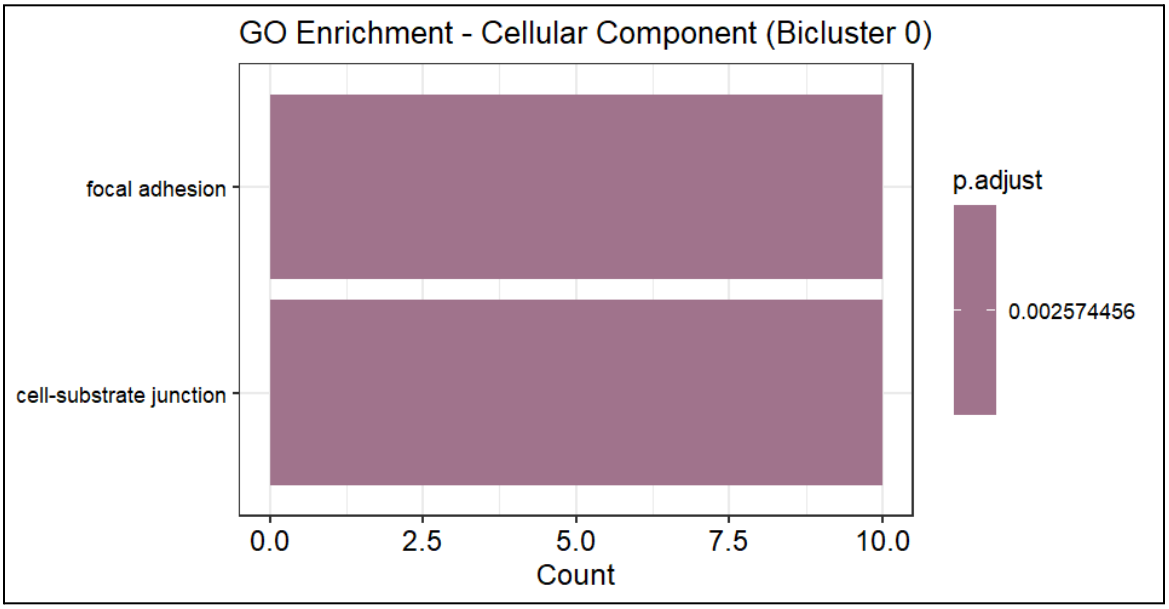
4.6.2 Analisis *Biclustering*

Bicluster 0

Dengan jumlah gen sebanyak 256 gen dalam Bicluster 0, berikut adalah sebagian nama-nama gen anggota bicluster dan visualisasi dan interpretasi GO *Enrichment* untuk *biological process, molecular function,* dan *cellular component*,

ID Gen	Simbol Gen	Nama Gen
1552274_at	PXK	PX domain containing serine/threonine kinase like
1553920_at	C9orf84	chromosome 9 open reading frame 84
1554638_at	ZFYVE16	zinc finger FYVE-type containing 16





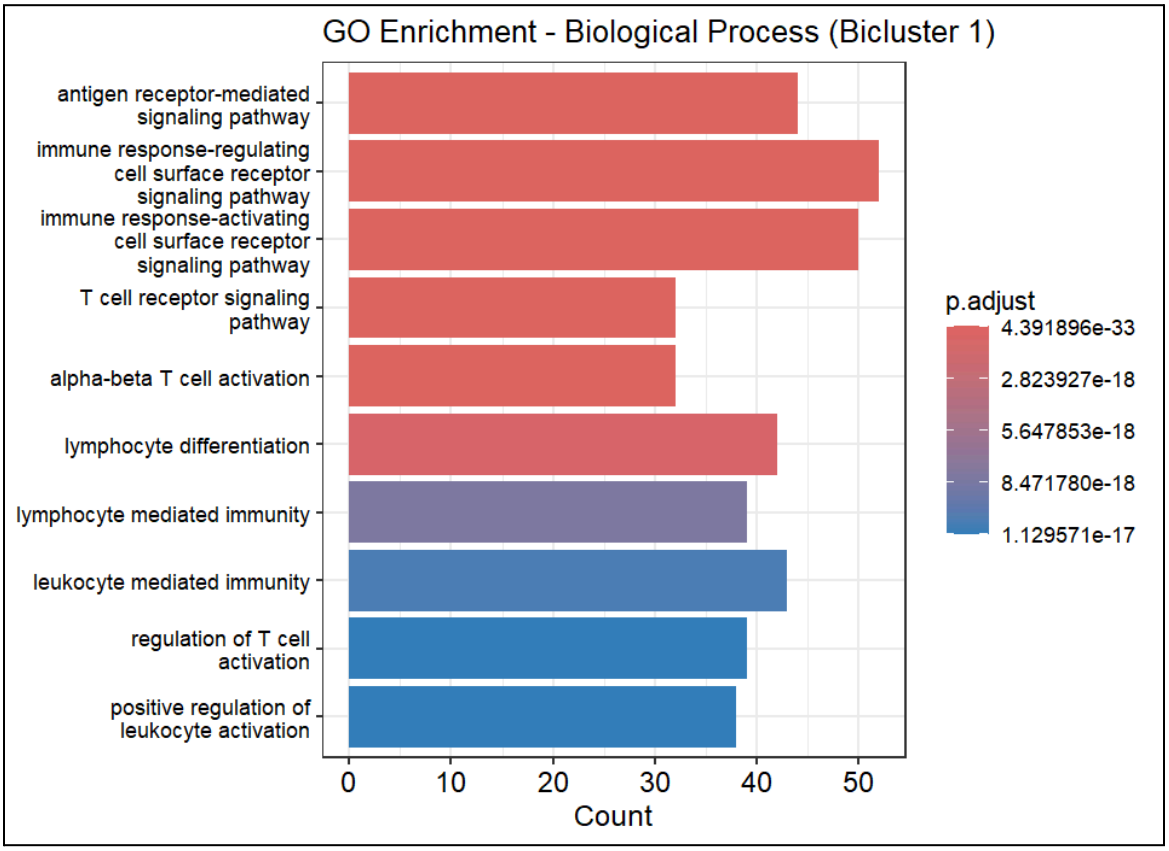
Contoh gen dalam Bicluster 0 seperti PXK, C9orf84, dan ZFYVE16 memberikan gambaran mengenai fungsi biologis kelompok ini. PXK diketahui berperan dalam proses transduksi sinyal dan pengaturan transporter membran, yang penting dalam respon sel terhadap rangsangan eksternal. C9orf84 masih belum banyak dikarakterisasi, namun keterlibatannya di berbagai jaringan menunjukkan potensi peran fungsional yang belum sepenuhnya diketahui. Sementara itu, ZFYVE16 terlibat dalam transport vesikular dan fungsi endosomal, yang berkaitan erat dengan dinamika membran dan adhesi sel.

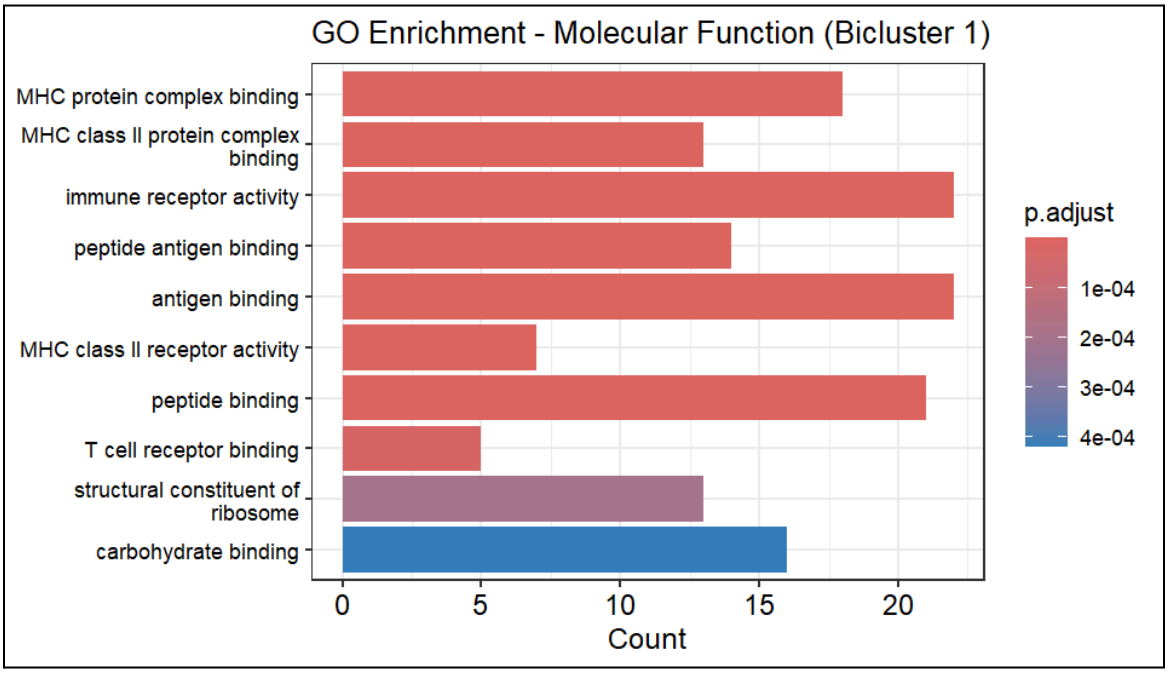
Hasil GO *enrichment* pada Bicluster 0 menunjukkan bahwa gen-gen ini secara signifikan terlibat dalam respon terhadap radiasi gamma (Biological Process), serta berada di lokasi seluler seperti focal adhesion dan cell-substrate junction (Cellular Component). Ini mengindikasikan bahwa bicluster ini mungkin berkaitan dengan respons terhadap stres lingkungan serta proses adhesi dan interaksi antara sel dan matriks ekstraseluler.

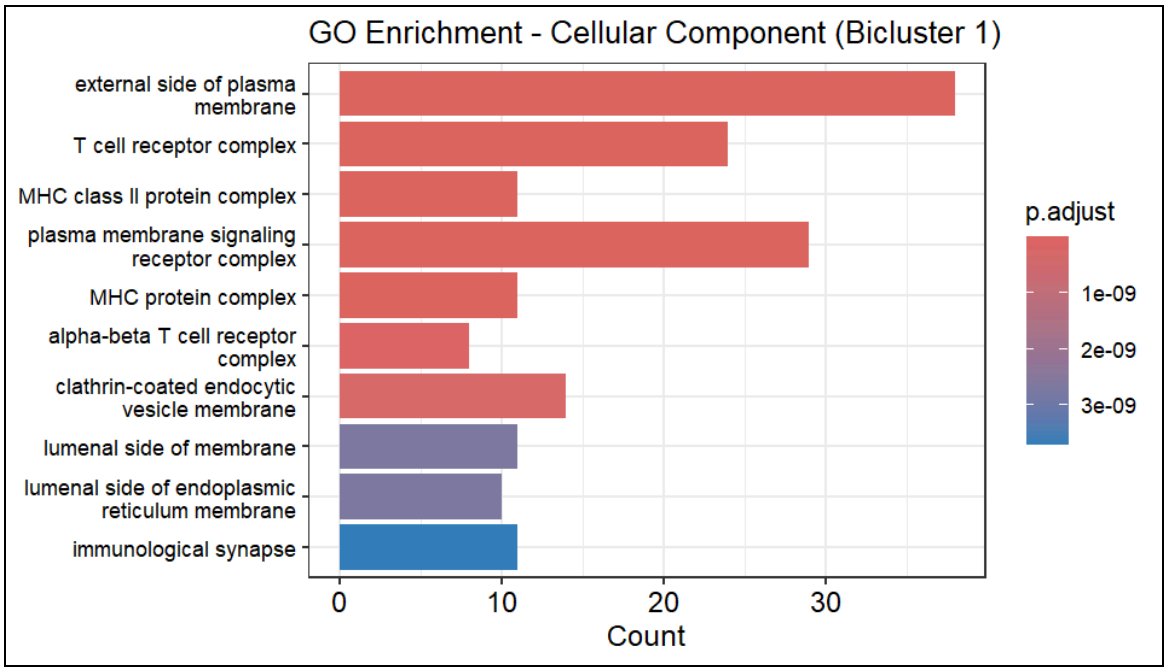
Bicluster 1

Dengan jumlah gen sebanyak 374 gen dalam Bicluster 1, berikut adalah sebagian nama-nama gen anggota bicluster dan visualisasi GO *Enrichment* untuk *biological process, molecular function,* dan *cellular component*,

ID Gen	Simbol Gen	Nama Gen
1405_i_at	CCL5	C-C motif chemokine ligand 5
1553132_a_at	TC2N	tandem C2 domains, nuclear
1553177_at	SH2D1B	SH2 domain containing 1B


GO Enrichment - Biological Process (Bicluster 1)
antigen receptor-mediated signaling pathway
immune response-regulating cell surface receptor signaling pathway
immune response-activating cell surface receptor signaling pathway
T cell receptor signaling pathway
alpha-beta T cell activation
lymphocyte differentiation
lymphocyte mediated immunity
leukocyte mediated immunity
regulation of T cell activation
positive regulation of leukocyte activation
0
10
20
30
40
50
Count
p.adjust
4.391896e-33
2.823927e-18
5.647853e-18
8.471780e-18
1.129571e-17


GO Enrichment - Molecular Function (Bicluster 1)
MHC protein complex binding
MHC class II protein complex binding
immune receptor activity
peptide antigen binding
antigen binding
MHC class II receptor activity
peptide binding
T cell receptor binding
structural constituent of ribosome
carbohydrate binding
0
5
10
15
20
Count
p.adjust
1e-04
2e-04
3e-04
4e-04



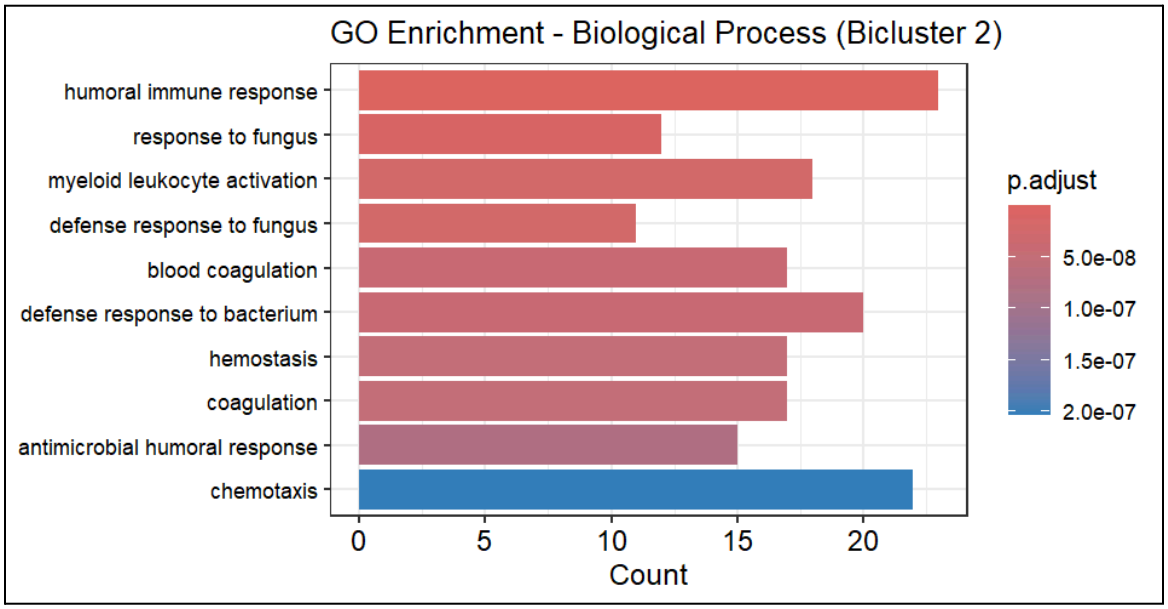
Beberapa gen yang termasuk dalam Bicluster 1 antara lain CCL5 (C-C motif chemokine ligand 5), TC2N (Tandem C2 Domains, Nuclear), dan SH2D1B (SH2 Domain Containing 1B). CCL5 merupakan kemokin yang berperan penting dalam kemotaksis sel imun, khususnya dalam merekrut sel T, eosinofil, dan basofil ke lokasi inflamasi. TC2N mengandung domain C2 ganda yang kemungkinan terlibat dalam proses pensinyalan intraseluler, meskipun fungsinya belum sepenuhnya dipahami. SH2D1B adalah protein adaptor yang berfungsi dalam regulasi sinyal imun, terutama pada sel NK dan sel T.

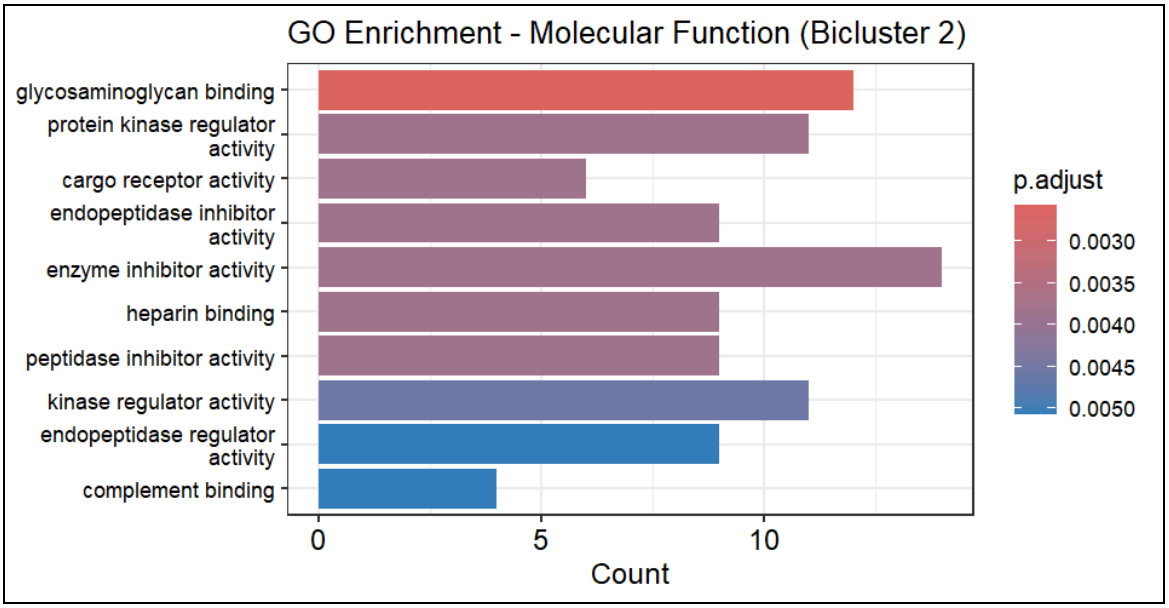
Analisis GO *enrichment* menunjukkan bahwa gen-gen ini aktif dalam proses-proses biologis yang berkaitan dengan aktivasi dan regulasi respon imun, khususnya melalui jalur pensinyalan reseptor antigen dan aktivasi sel T alpha-beta. Secara fungsi molekuler, mereka berkontribusi dalam pengikatan terhadap kompleks MHC kelas II, serta memiliki aktivitas reseptor imun dan pengikatan antigen. Dari aspek komponen seluler, gen-gen ini berlokalisasi di sisi luar membran plasma, termasuk dalam kompleks reseptor sel T dan kompleks protein MHC kelas II, yang mempertegas peran sentralnya dalam pengenalan antigen dan aktivasi sel imun adaptif.

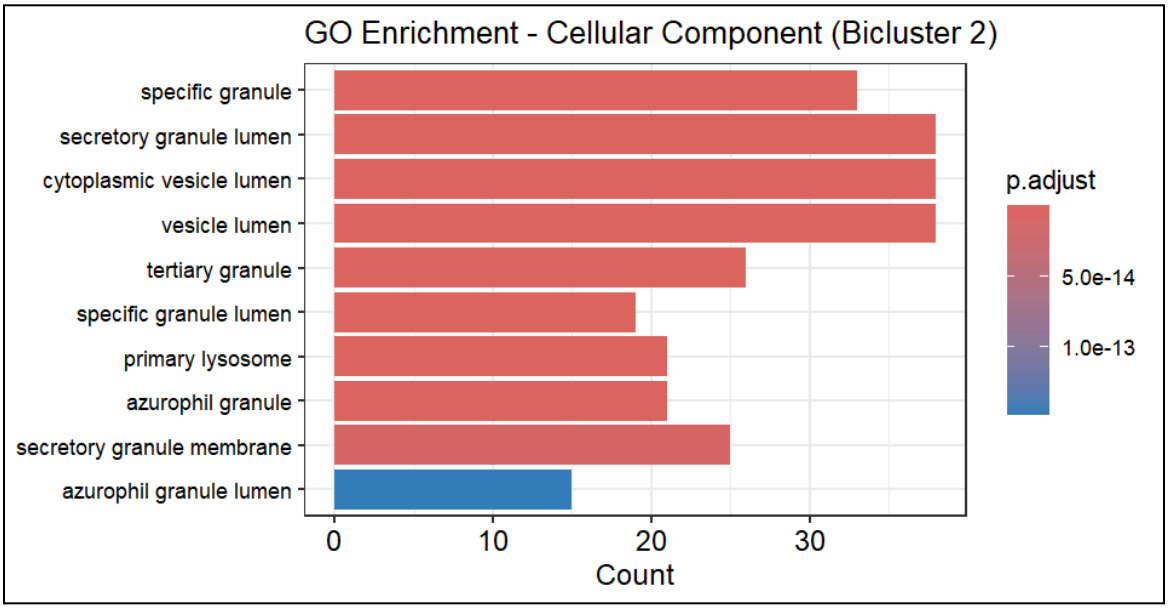
Bicluster 2

Dengan jumlah gen sebanyak 226 gen dalam Bicluster 2, berikut adalah sebagian nama-nama gen anggota bicluster dan visualisasi GO *Enrichment* untuk *biological process, molecular function,* dan *cellular component*,

ID Gen	**Simbol Gen**	**Nama Gen**
1552701_a_at	CARD16	caspase recruitment domain family member 16
1552772_at	CLEC4D	C-type lectin domain family 4 member D
1552772_at	ABCA13	ATP binding cassette subfamily A member 13







Beberapa contoh gen dari Bicluster 2 adalah CARD16, CLEC4D, dan ABCA13. CARD16 berperan dalam regulasi inflamasi melalui jalur sinyal inflammasome, sementara CLEC4D adalah reseptor pengenal pola yang berperan dalam respon imun terhadap patogen seperti jamur dan bakteri. ABCA13 termasuk dalam keluarga transporter ABC, yang meskipun fungsinya belum sepenuhnya dipahami, diduga terkait dengan transportasi lipid atau xenobiotik dan regulasi aktivitas seluler tertentu.

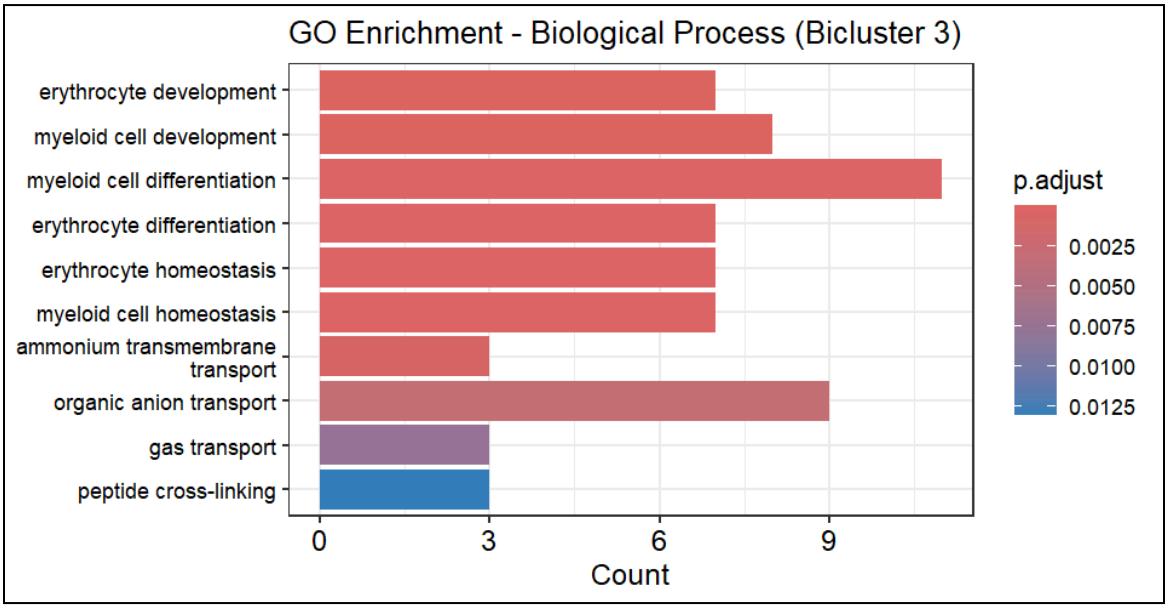
Hasil analisis GO *enrichment* pada Bicluster 2 menunjukkan keterlibatan gen dalam berbagai proses imun, terutama yang terkait dengan respon humoral, respon terhadap

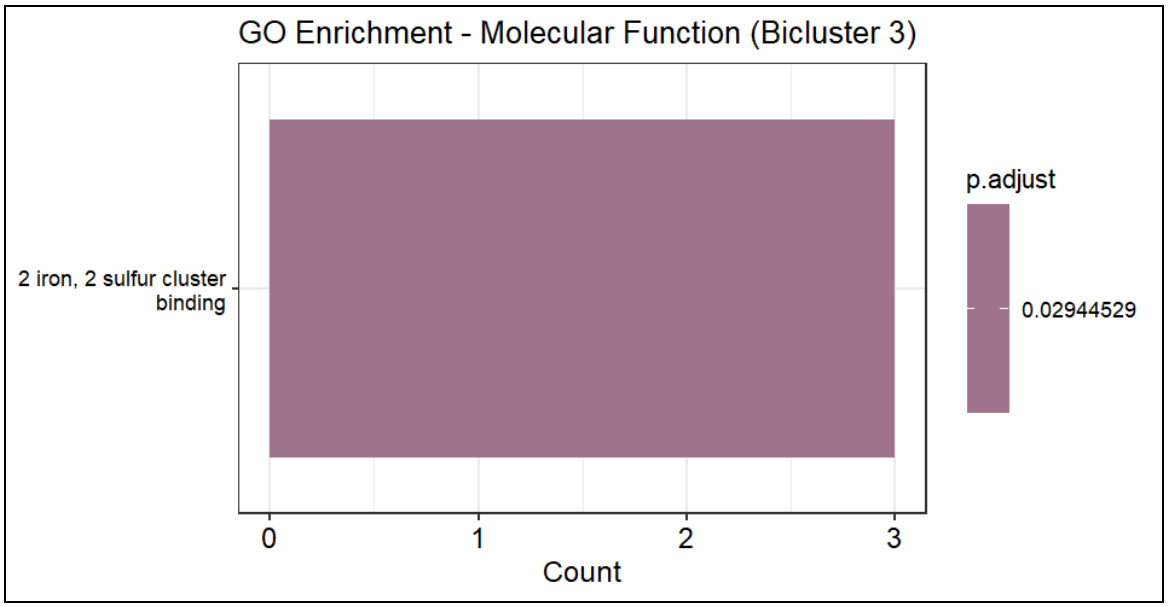
jamur dan bakteri, serta aktivasi leukosit (BP). Secara molekuler (MF), gen-gen tersebut banyak terlibat dalam aktivitas pengikatan dan penghambatan enzim, termasuk glycosaminoglycan binding dan enzyme inhibitor activity. Dari sisi komponen seluler (CC), gen-gen tersebut secara signifikan terlokalisasi pada struktur granula sekretorik, vesikel, dan lisosom, yang mendukung fungsi imun seperti sekresi molekul efektor dan degradasi patogen.

Bicluster 3

Dengan jumlah gen sebanyak 84 gen dalam Bicluster 3, berikut adalah sebagian nama-nama gen anggota bicluster dan visualisasi GO *Enrichment* untuk *biological process, molecular function,* dan *cellular component*,

ID Gen	Simbol Gen	Nama Gen
1553589_a_at	PDZK1IP1	PDZK1 interacting protein 1
1556283_s_at	FGFR1OP2	FGFR1 oncogene partner 2
201109_s_at	THBS1	thrombospondin 1





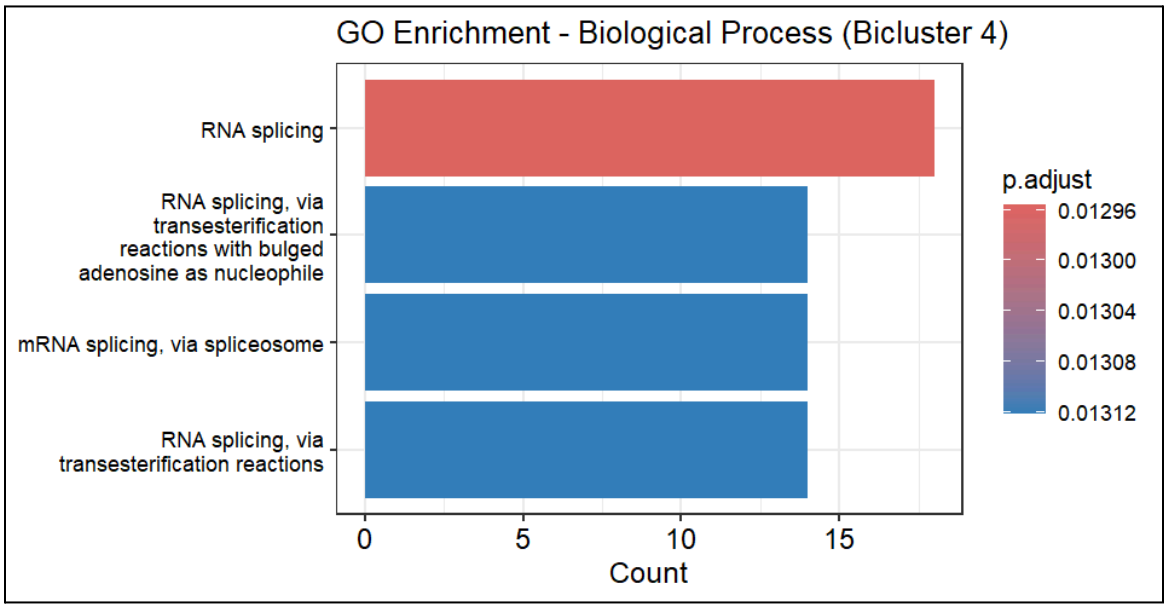
Beberapa gen yang tergolong dalam Bicluster 3 antara lain PDZK1IP1 (PDZK1 Interacting Protein 1), FGFR1OP2 (FGFR1 Oncogene Partner 2), dan THBS1 (Thrombospondin 1). PDZK1IP1 berperan dalam regulasi metabolisme dan proliferasi sel, sementara FGFR1OP2 terlibat dalam fusi kromosom yang berhubungan dengan kanker. THBS1 dikenal memiliki peran penting dalam adhesi sel, angiogenesis, dan regulasi respon imun.

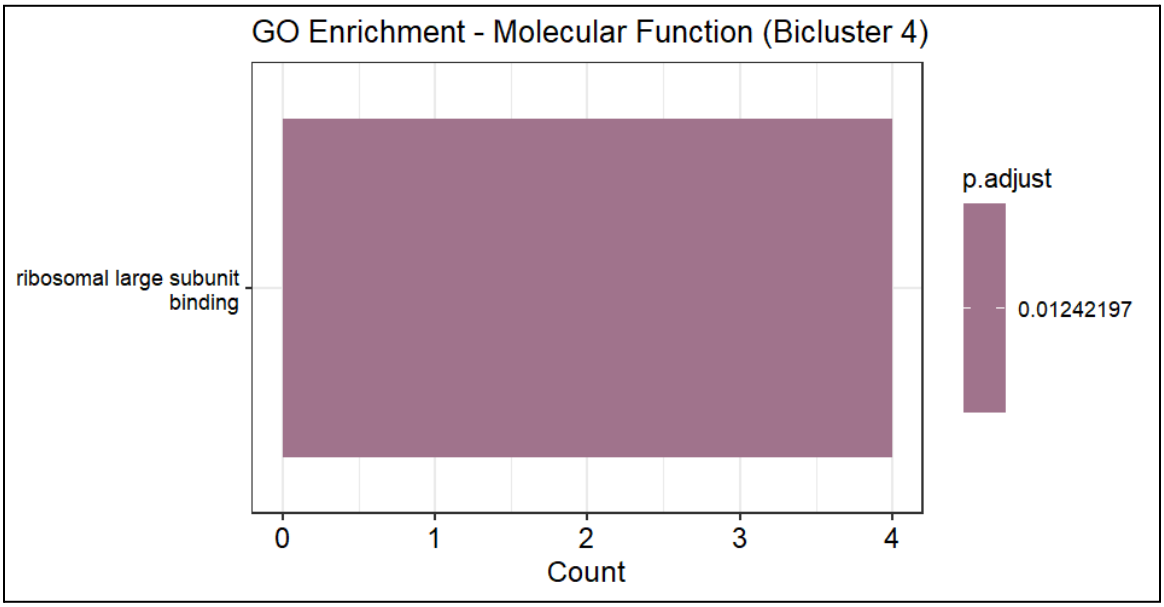
Berdasarkan hasil analisis GO *enrichment*, gen-gen dalam Bicluster 3 menunjukkan keterkaitan kuat dengan proses biologis seperti perkembangan eritrosit dan mieloid, serta diferensiasi dan homeostasis sel darah merah. Dari segi fungsi molekuler, hanya satu kategori yang diperkaya, yaitu aktivitas pengikatan gugus besi-sulfur [2Fe-2S], yang penting dalam aktivitas enzimatik dan transfer elektron. Hal ini mengindikasikan bahwa Bicluster 3 berhubungan erat dengan regulasi dan diferensiasi sel-sel darah, khususnya dalam konteks hematopoiesis.

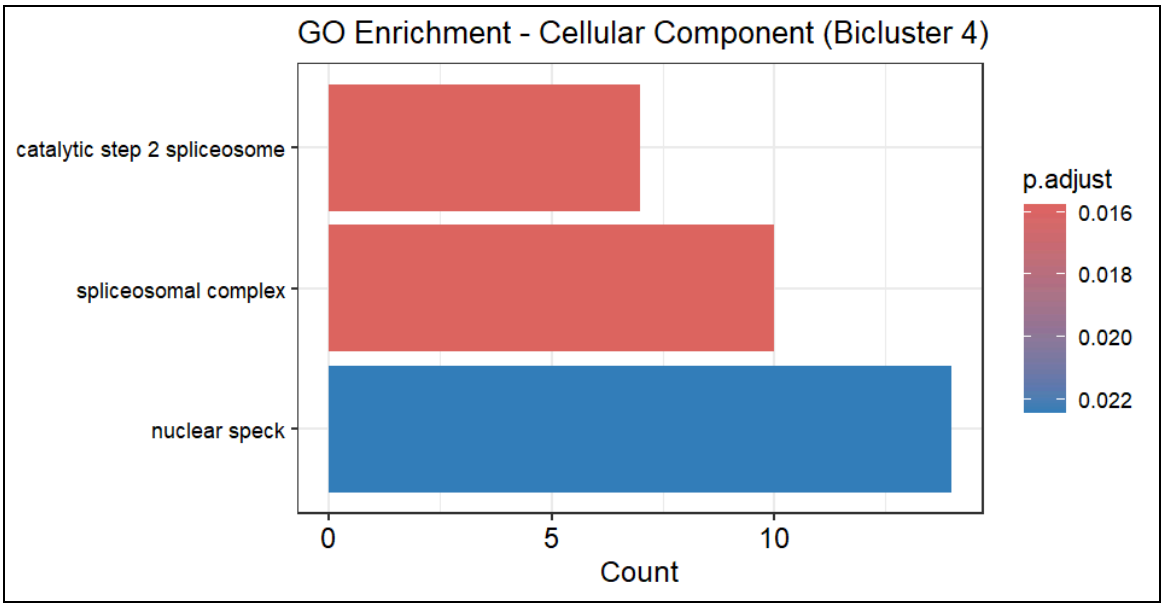
Bicluster 4

Dengan jumlah gen sebanyak 268 gen dalam Bicluster 4, berikut adalah sebagian nama-nama gen anggota bicluster dan visualisasi GO *Enrichment* untuk *biological process, molecular function,* dan *cellular component*,

ID Gen	Simbol Gen	Nama Gen
1552316_a_at	GIMAP1	GTPase, IMAP family member 1
1553678_a_at	ITGB1	integrin subunit beta 1
1553979_at	ZNF121	zinc finger protein 121







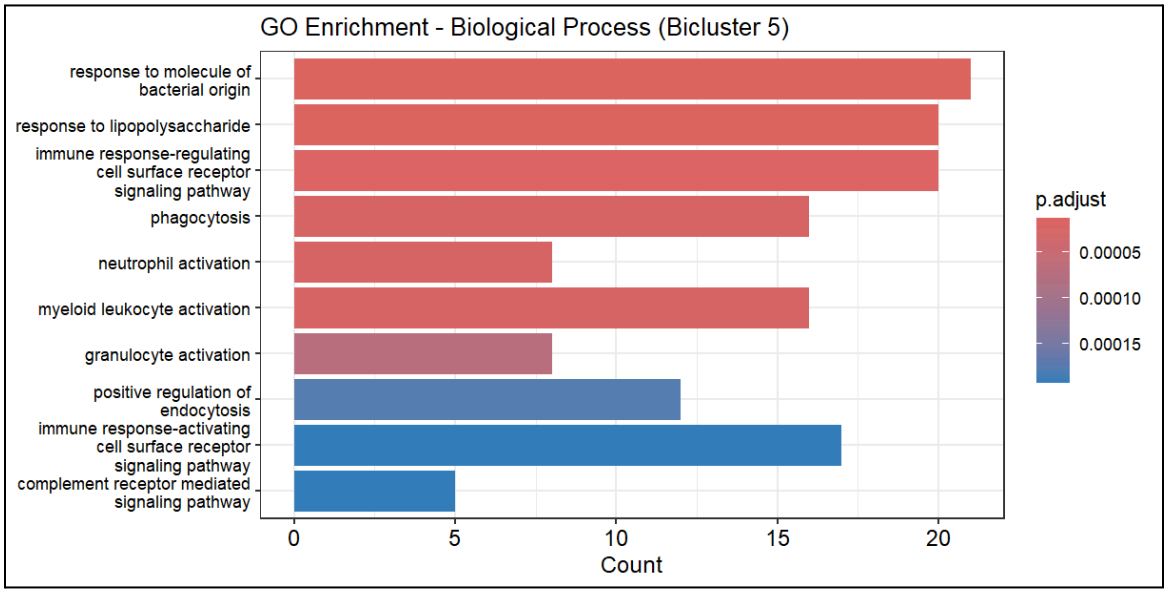
Beberapa gen yang tergolong dalam Bicluster 4 antara lain GIMAP1 (GTPase, IMAP Family Member 1), ITGB1 (Integrin Subunit Beta 1), dan ZNF121 (Zinc Finger Protein 121). GIMAP1 berperan dalam kelangsungan hidup dan perkembangan sel imun, ITGB1 terlibat dalam adhesi dan migrasi sel melalui interaksi dengan matriks ekstraseluler, sementara ZNF121 merupakan protein regulator transkripsi yang mengandung domain zinc finger.

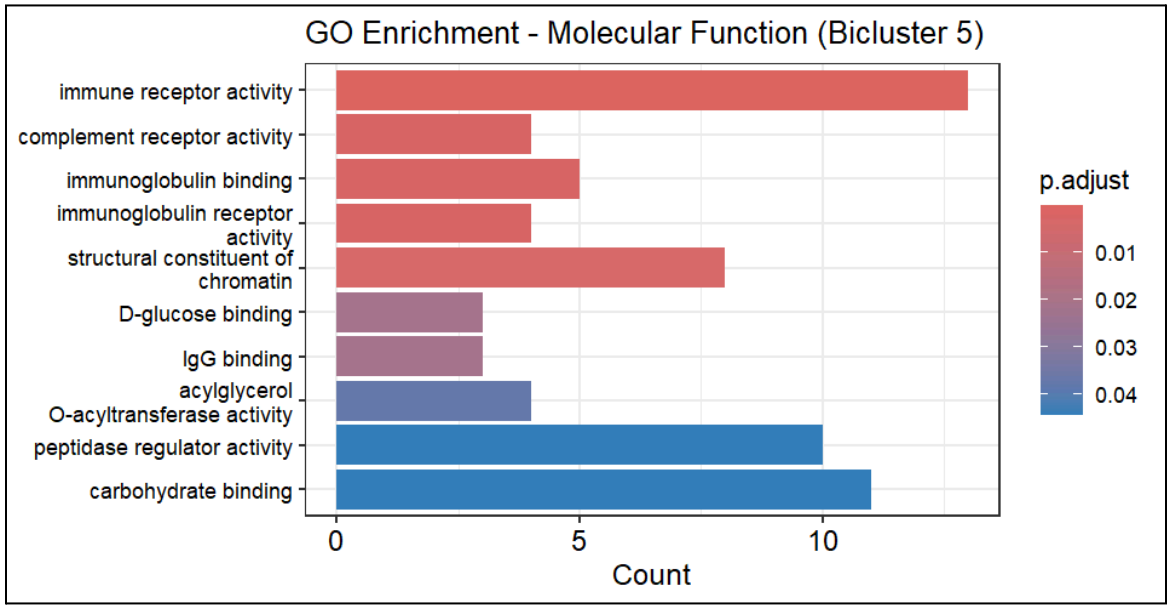
Analisis GO *enrichment* menunjukkan bahwa gen-gen dalam Bicluster 4 diperkaya dalam proses biologis yang berhubungan dengan *RNA splicing*, khususnya melalui jalur transesterifikasi yang melibatkan spliceosome. Fungsi molekuler yang dominan adalah kemampuan mengikat subunit besar ribosom, yang mencerminkan keterlibatan dalam regulasi ekspresi gen pada level pascatranskripsi. Dari sisi komponen seluler, gen-gen ini berasosiasi kuat dengan struktur spliceosome, terutama pada tahap katalitik kedua, serta berada di nuclear speck (wilayah nukleus yang menjadi tempat pemrosesan RNA).

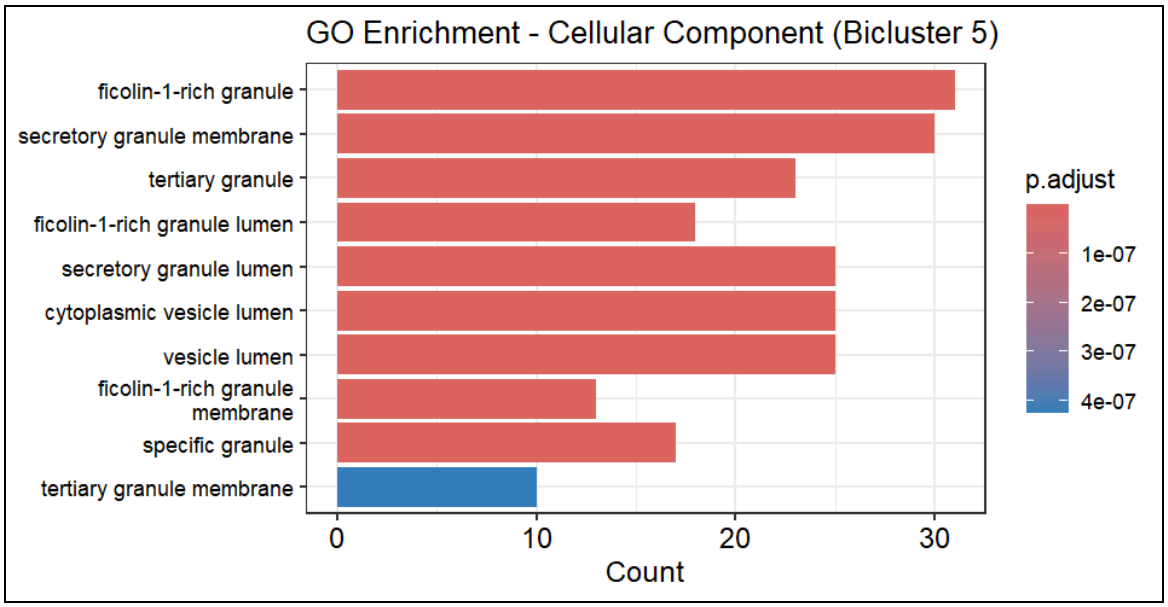
Bicluster 5

Dengan jumlah gen sebanyak 350 gen dalam Bicluster 5, berikut adalah sebagian nama-nama gen anggota bicluster dan visualisasi GO *Enrichment* untuk *biological process, molecular function,* dan *cellular component*,

ID Gen	Simbol Gen	Nama Gen
1552316_a_at	CSF3R	colony stimulating factor 3 receptor
1553723_at	ADGRG3	adhesion G protein-coupled receptor G3
1554178_a_at	FAM126B	family with sequence similarity 126 member B







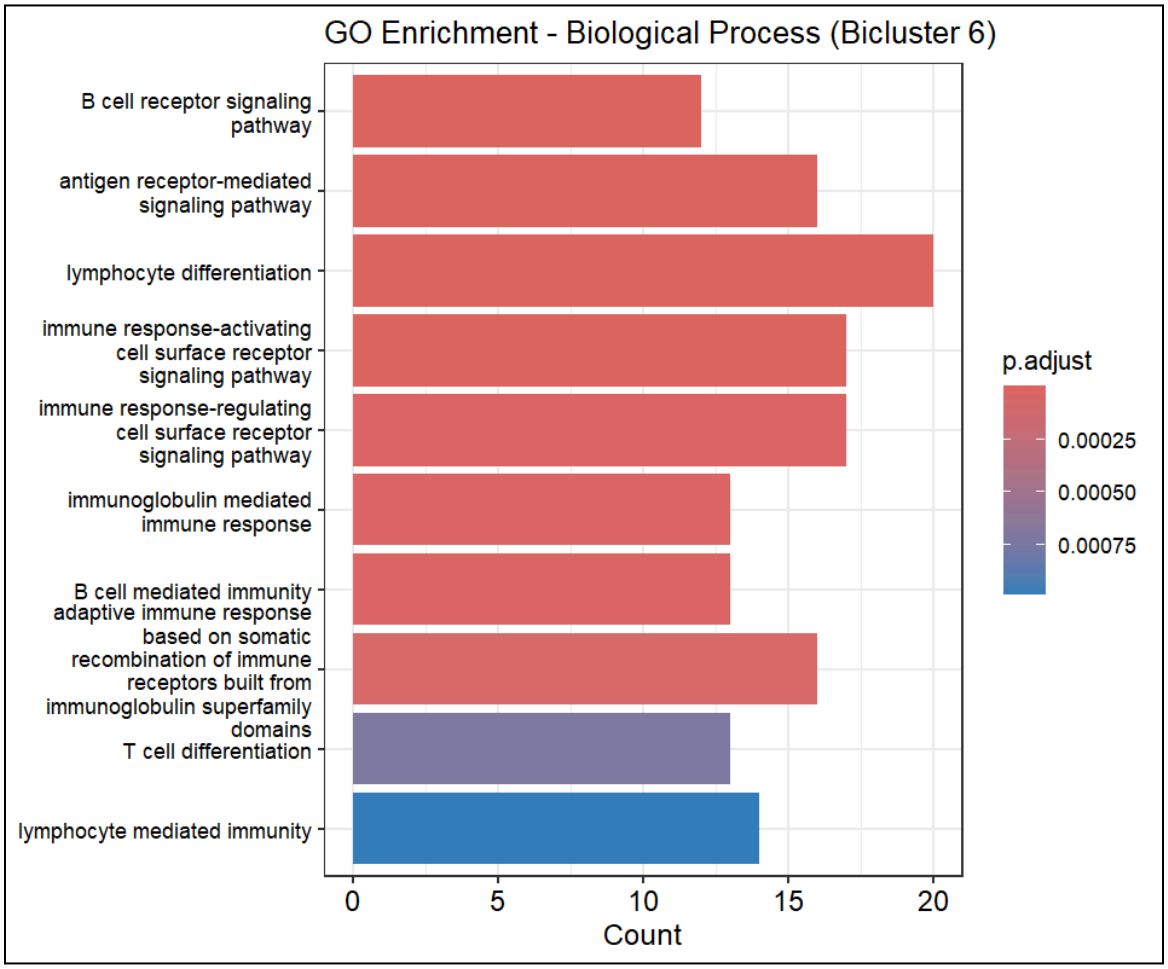
Beberapa gen yang termasuk dalam Bicluster 5 antara lain CSF3R (Colony Stimulating Factor 3 Receptor), ADGRG3 (Adhesion G Protein-Coupled Receptor G3), dan FAM126B (Family with Sequence Similarity 126 Member B). CSF3R memainkan peran penting dalam proliferasi dan diferensiasi granulosit, terutama neutrofil, yang penting dalam respon imun. ADGRG3 terlibat dalam transmisi sinyal imun dan adhesi sel, sementara fungsi FAM126B masih belum sepenuhnya dipahami, namun diduga berkontribusi dalam regulasi sel imun.

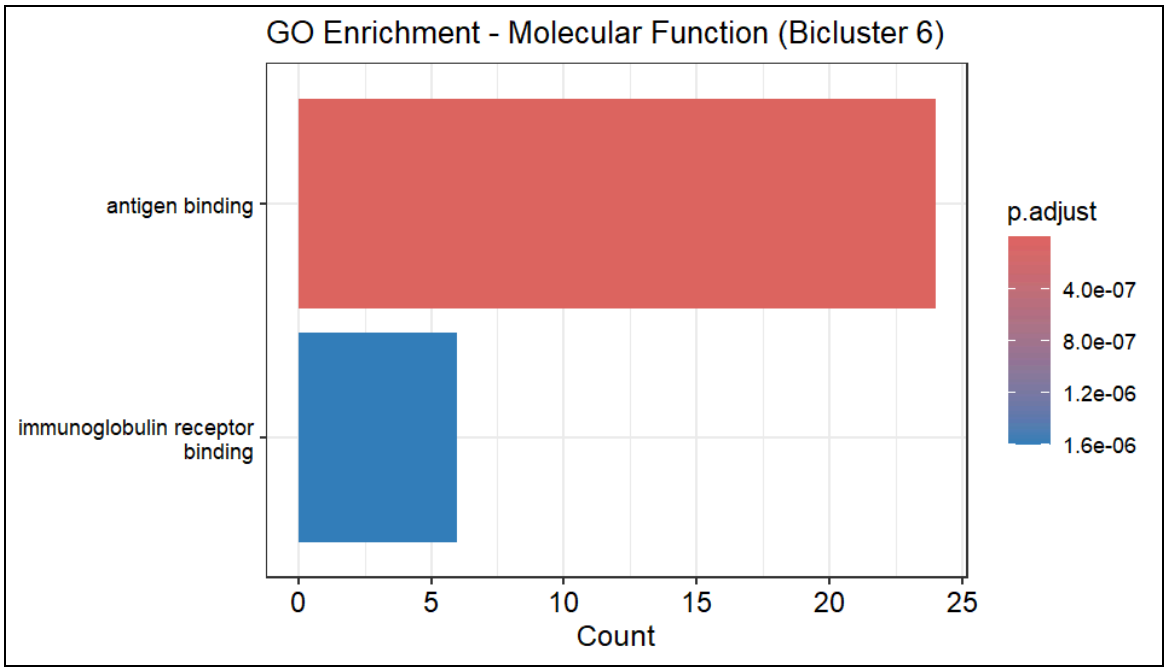
Analisis GO *enrichment* menunjukkan bahwa gen-gen ini terlibat dalam proses biologis terkait aktivasi dan respons imun terhadap molekul asal bakteri, termasuk lipopolisakarida, serta berperan dalam fagositosis dan aktivasi neutrofil. Secara fungsi molekuler, mereka berhubungan dengan aktivitas reseptor imun, pengikatan imunoglobulin, dan konstituen struktural kromatin. Dari aspek komponen seluler, gen-gen ini terkonsentrasi di organel granula seperti ficolin-1-rich granule dan secretory granule lumen, yang berperan penting dalam penyimpanan dan sekresi protein imun.

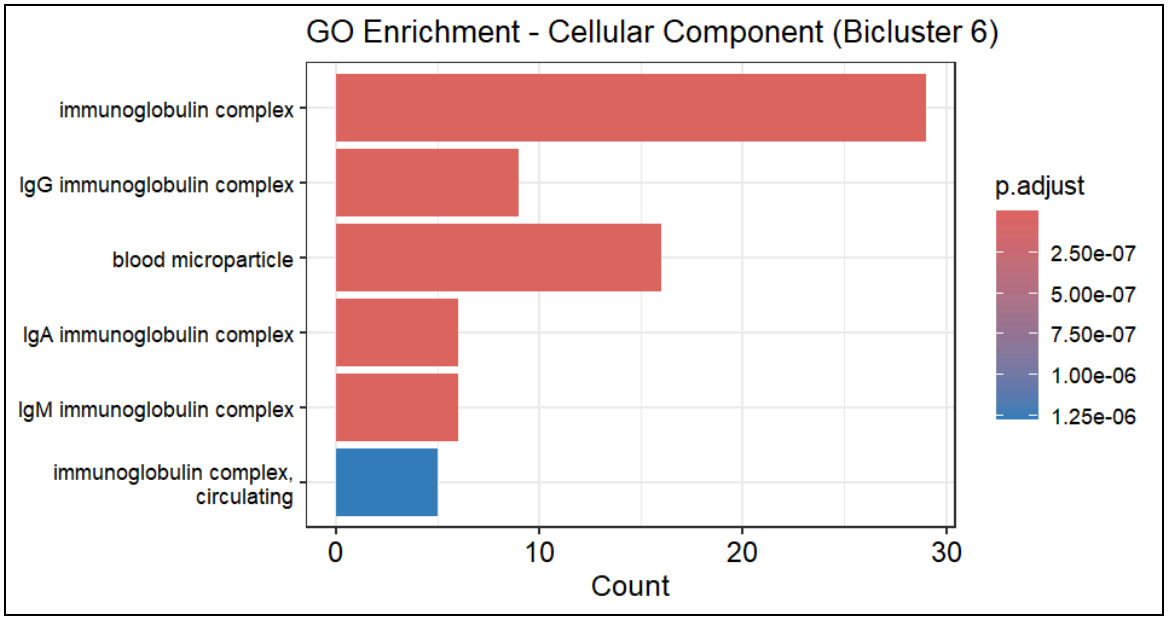
Bicluster 6

Dengan jumlah gen sebanyak 269 gen dalam Bicluster 6, berikut adalah sebagian nama-nama gen anggota bicluster dan visualisasi GO *Enrichment* untuk *biological process, molecular function,* dan *cellular component*,

ID Gen	Simbol Gen	Nama Gen
1554786_at	CASS4	Cas scaffolding protein family member 4
1557239_at	BBX	BBX, HMG-box containing
1557828_a_at	TMEM267	transmembrane protein 267







Beberapa gen yang termasuk dalam Bicluster 6 antara lain CASS4 (Cas scaffolding protein family member 4), BBX (HMG-box containing protein), dan TMEM267 (Transmembrane Protein 267). CASS4 diketahui terlibat dalam regulasi sinyal seluler melalui interaksinya dengan protein pensinyalan, sedangkan BBX merupakan faktor transkripsi yang mengandung domain HMG-box dan berperan dalam regulasi ekspresi gen. TMEM267 merupakan protein transmembran yang belum banyak dikarakterisasi,

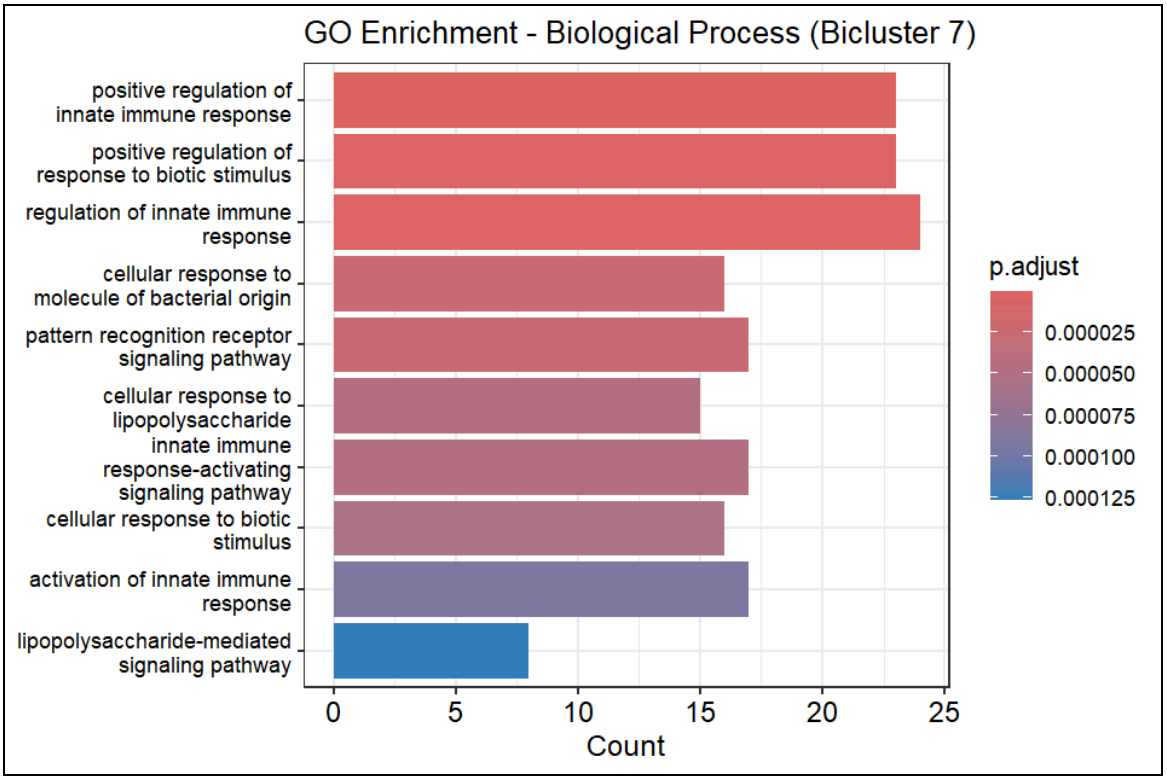
namun potensi keterlibatannya dalam jalur pensinyalan atau komunikasi antar sel tengah dieksplorasi.

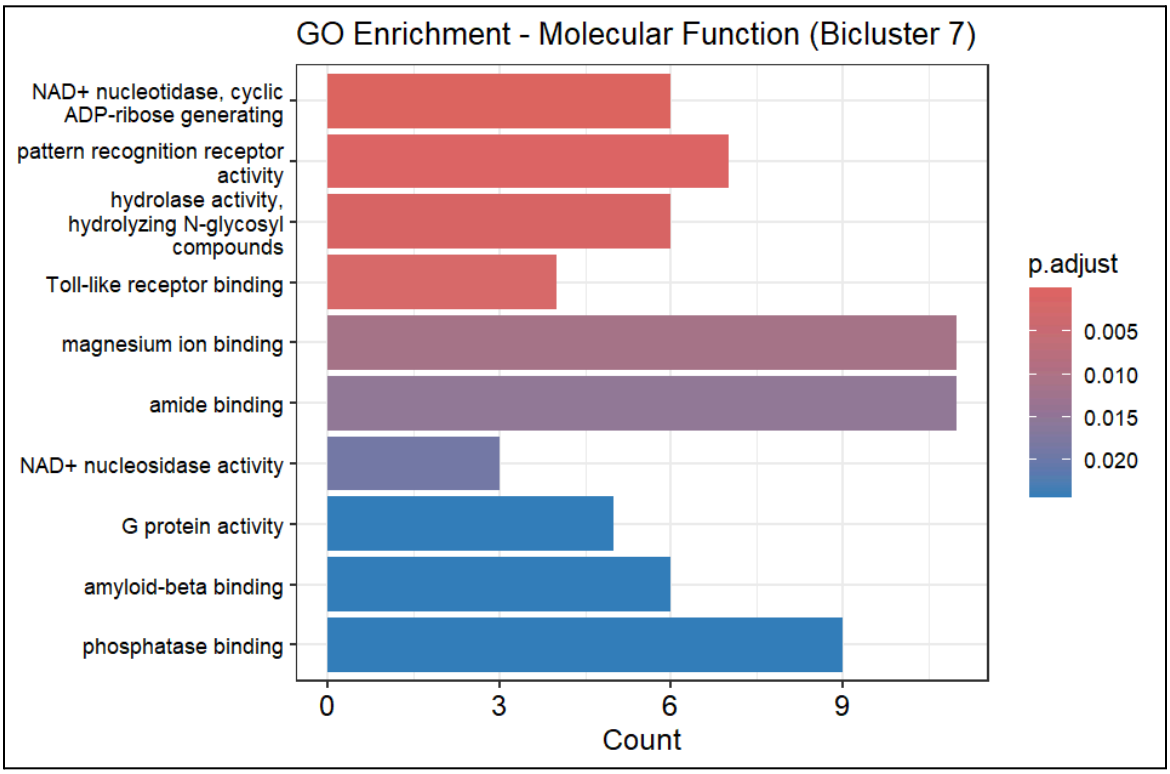
Analisis GO *enrichment* menunjukkan bahwa gen-gen ini berperan penting dalam proses biologis terkait sistem imun, seperti jalur pensinyalan reseptor sel B, diferensiasi limfosit, dan aktivasi reseptor permukaan sel terhadap respons imun. Dari sisi fungsi molekuler, gen-gen ini memiliki kemampuan untuk mengikat antigen dan reseptor imunoglobulin, yang mengindikasikan peran langsung dalam pengenalan patogen. Sementara itu, dari aspek komponen seluler, gen-gen ini terlokalisasi dalam berbagai kompleks imunoglobulin seperti IgG, IgA, dan IgM, serta berada pada struktur darah seperti blood microparticles, mendukung perannya dalam respon imun humoral.

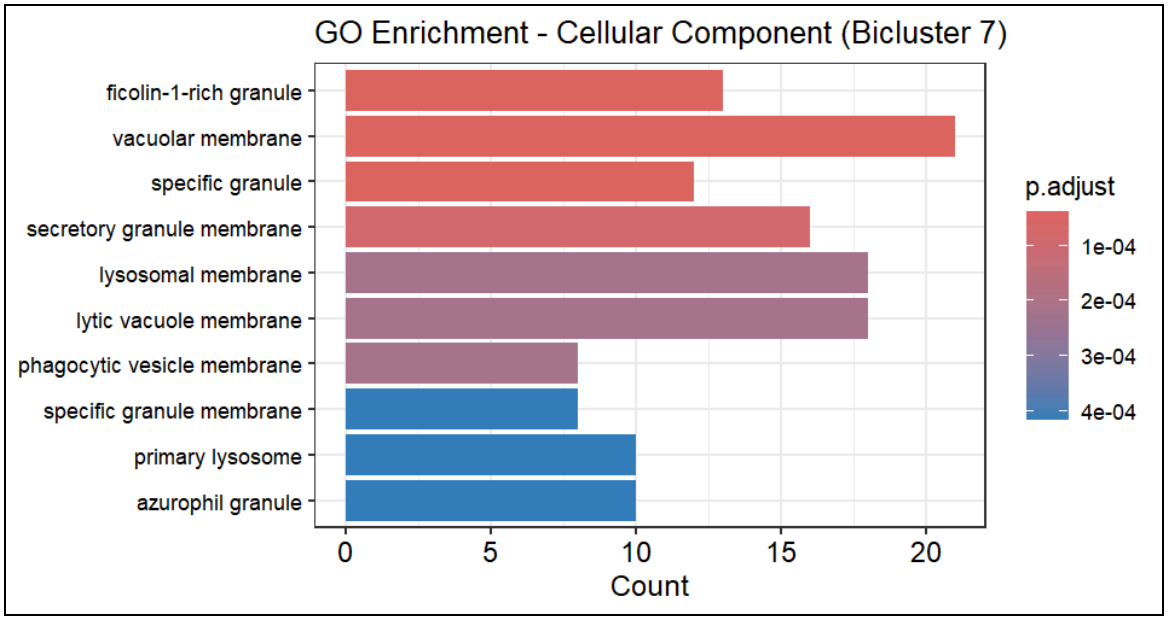
Bicluster 7

Dengan jumlah gen sebanyak 302 gen dalam Bicluster 7, berikut adalah sebagian nama-nama gen anggota bicluster dan visualisasi GO *Enrichment* untuk *biological process, molecular function,* dan *cellular component*,

ID Gen	Simbol Gen	Nama Gen
1552263_at	MAPK1	mitogen-activated protein kinase 1
1552485_at	LACTB	lactamase beta
1552553_a_at	NLRC4	NLR family CARD domain containing 4


GO Enrichment - Biological Process (Bicluster 7)
positive regulation of innate immune response
positive regulation of response to biotic stimulus
regulation of innate immune response
cellular response to molecule of bacterial origin
pattern recognition receptor signaling pathway
cellular response to lipopolysaccharide
innate immune response-activating signaling pathway
cellular response to biotic stimulus
activation of innate immune response
lipopolysaccharide-mediated signaling pathway
0
5
10
15
20
25
Count
p.adjust
0.000025
0.000050
0.000075
0.000100
0.000125


GO Enrichment - Molecular Function (Bicluster 7)
NAD+ nucleotidase, cyclic ADP-ribose generating
pattern recognition receptor activity
hydrolase activity, hydrolyzing N-glycosyl compounds
Toll-like receptor binding
magnesium ion binding
amide binding
NAD+ nucleosidase activity
G protein activity
amyloid-beta binding
phosphatase binding
0
3
6
9
Count
p.adjust
0.005
0.010
0.015
0.020



Beberapa gen dalam Bicluster 7 termasuk MAPK1 (Mitogen-Activated Protein Kinase 1), LACTB (Lactamase Beta), dan NLRC4 (NLR Family CARD Domain Containing 4). MAPK1 merupakan komponen penting dalam jalur transduksi sinyal yang mengatur proliferasi, diferensiasi, dan respon imun. LACTB adalah enzim mitokondria yang berperan dalam metabolisme sel dan regulasi diferensiasi, sedangkan NLRC4 merupakan bagian dari inflammasome yang berperan dalam deteksi patogen dan aktivasi respon imun bawaan.

Berdasarkan hasil GO *enrichment*, gen-gen pada Bicluster 7 terlibat dalam regulasi positif terhadap respon imun bawaan dan respon terhadap stimulus biotik, termasuk respon terhadap molekul bakteri. Dari sisi fungsi molekuler, mereka memiliki aktivitas pengenalan pola (pattern recognition), aktivitas enzimatik seperti *NAD+ nucleotidase*, serta kemampuan mengikat reseptor Toll-like dan ion seperti magnesium. Lokasi selulernya teridentifikasi di berbagai kompartemen yang berkaitan dengan sistem imun, seperti *granula sekretorik*, *membran lisosom*, dan *vakuola litik*, yang mendukung peran aktif dalam deteksi dan penghancuran patogen.

V. Penutup

5.1 Kesimpulan

Penelitian ini bertujuan untuk menganalisis ekspresi gen pada anak-anak yang mengalami *septic shock* menggunakan pendekatan *supervised* dan *unsupervised learning*. Analisis ini dilakukan pada data ekspresi gen dari dataset GDS4274 yang terdiri dari 130 sampel. Dengan 98 sampel *septic shock* dan 32 sampel *healthy control.* Proses analisis melibatkan berbagai tahapan, mulai dari eksplorasi dan *preprocessing* data, analisis DEG, klasifikasi, *clustering, biclustering,* hingga *gene ontology enrichment.*

Beberapa poin penting yang dapat disimpulkan dari hasil analisis adalah sebagai berikut:

1. *Differentially Expressed Genes* (DEG): Dari 2.734 gen yang telah diseleksi berdasarkan variansi tertinggi, sebanyak 2.129 gen menunjukkan perbedaan ekspresi signifikan antara kelompok *septic shock* dan *healthy control.* Terdapat 959 gen yang *upregulated* dan 1.170 gen yang *downregulated.*
2. Analisis Klasifikasi: *Model Support Vector Machine* (SVM) menghasilkan performa klasifikasi terbaik dengan akurasi sempurna (100%) pada data uji. Model ini juga memiliki nilai *precision, recall,* dan F1-*score* sebesar 1.00 untuk kedua kelas. Gen yang paling berkontribusi dalam proses klasifikasi diidentifikasi menggunakan metode SHAP, termasuk gen seperti CLEC12B, TXLNGY, dan NFXL1.
3. Analisis *Clustering:* Metode K-Means, *Hierarchical*, dan *Gaussian Mixture Model* memberikan hasil *clustering* yang kurang optimal, dengan *silhouette score* tertinggi sebesar 0.627. Oleh karena itu, diterapkan metode HDBSCAN dan OPTICS. Meskipun OPTICS memiliki *silhouette* score tertinggi (0.740), HDBSCAN dipilih karena menghasilkan klaster yang lebih proposional dan sedikit membuang informasi gen sebagai *outlier*.
4. Analisis *Biclustering:* Dengan metode *Spectral Biclustering,* jumlah *bicluster* optimal ditentukan sebanyak 8 *bicluster* berdasarkan nilai MSR dan VAF. Hasil *biclustering* menunjukkan adanya pola *checkerboard* yang mencerminkan ko-ekspresi gen dalam subset sampel tertentu.
5. Analisis *Gene Ontology* (GO) *Enrichment*: Gen-gen yang diidentifikasi melalui DEG, klasifikasi, *clustering*, dan *biclustering* dikompilasi untuk dianalisis lebih

lanjut melalui analisis GO *enrichment* guna memperoleh wawasan awal tentang keterkaitannya dengan fungsi biologis dan mekanisme patofisiologi *septic shock*. Khusus untuk hasil *clustering* dan *biclustering*, dilakukan analisis *enrichment* pada tiga domain utama Gene Ontology, yaitu *biological process* (BP), *molecular function* (MF), dan *cellular component* (CC), untuk mengungkap peran fungsional gen dalam proses biologis yang relevan.

5.2 Saran

Berdasarkan hasil penelitian ini, terdapat beberapa hal yang dapat disarankan untuk pengembangan studi di masa mendatang, yaitu:

1. Analisis *gene ontology* yang telah dilakukan secara deskriptif dapat diperluas dengan *pathway analysis* untuk mengidentifikasi jalur-jalur biologis yang berperan signifikan dalam patofisiologi *septic shock*. Pendekatan ini dapat memberikan pemahaman yang lebih mendalam terhadap fungsi biologis gen-gen yang teridentifikasi.
2. Penelitian lanjutan disarankan untuk memanfaatkan data omik lainnya, seperti proteomik atau metabolomik, guna melakukan validasi silang terhadap gen-gen yang ditemukan, sehingga meningkatkan keandalan dan relevansi biologis temuan.
3. Mengingat kompleksitas dan dimensi tinggi dari data ekspresi gen, penggunaan pendekatan *deep learning* dapat dipertimbangkan untuk menggantikan atau melengkapi metode *machine learning* konvensional. Pendekatan ini berpotensi menangkap pola-pola nonlinier yang kompleks dan meningkatkan performa model dalam klasifikasi atau *clustering*.
4. Peningkatan jumlah sampel serta integrasi data dari berbagai sumber dapat memperkuat generalisasi temuan dan mendukung identifikasi biomarker yang lebih konsisten dan aplikatif dalam konteks klinis.

DAFTAR PUSTAKA

Ashburner, M., Ball, C. A., Blake, J. A., et al. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1), 25–29. https://doi.org/10.1038/75556

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1), 44–57. https://doi.org/10.1038/nprot.2008.211

Kluger, Y., Basri, R., Chang, J. T., & Gerstein, M. (2003). Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Research*, 13(4), 703–716. https://doi.org/10.1101/gr.648603

Kourou, K., et al. (2015). Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 13, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Madeira, S. C., & Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1), 24–45. https://doi.org/10.1109/TCBB.2004.2

Mi, H., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols*, 8(8), 1551–1566. https://doi.org/10.1038/nprot.2013.092

Smyth, G. K. (2004). Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. Statistical Applications in Genetics and Molecular Biology, 3(1), Article3. https://doi.org/10.2202/1544-6115.1027

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437.

Tarca, A. L., et al. (2007). Machine learning and its applications to biology. PLoS Computational Biology, 3(6), e116. https://doi.org/10.1371/journal.pcbi.0030116

Weiss, S. L., et al. (2015). The Epidemiology of Global Pediatric Severe Sepsis. Pediatric Critical Care Medicine, 16(6), 428–436. https://doi.org/10.1097/PCC.0000000000000353

Wong, H. R., et al. (2009). Genomic expression profiling across the pediatric systemic inflammatory response syndrome, sepsis, and septic shock spectrum. Critical Care Medicine, 37(5), 1558–1566. https://doi.org/10.1097/CCM.0b013e31819c180c (Sumber data GEO: https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4274)

LAMPIRAN

Kode dengan menggunakan *software* Python dan R, serta *file* anotasi GPL570 dapat diunduh melalui tautan berikut: UAS Sains Data Genom (A) - Kelompok 7