Bioinfoysis

急性呼吸道感染宿主血液转录组探索性分析与建模报告

自动生成的结构化报告



报告标题

急性呼吸道感染宿主血液转录组探索性分析与建模报告

对话名称

我有一项急性呼吸道感染相关的宿主血液转录组研究数据，需要你作为生信分析专家，从原始数据出发独立完成一次完整的探索性分析与建模，并生成结构化报告。具体要求如下：…

报告时间

2026-09-07 10:59

AI 结果用于科研参考，关键结论需结合实验验证和人工解读。

附录：术语说明与引用文件

缩略语

缩略语	全称	说明
AUC	Area Under the Curve	ROC 曲线下面积，衡量分类器区分正负类的能力
BH	Benjamini-Hochberg	控制错误发现率（FDR）的多重检验校正方法
BP	Biological Process	基因本体（GO）三大类别之一，描述生物学过程
CC	Cellular Component	基因本体（GO）三大类别之一，描述细胞组分
CV	Cross-Validation	交叉验证，用于评估模型泛化性能
DE	Differential Expression	差异表达分析
FDR	False Discovery Rate	错误发现率，多重检验校正后的显著性指标
FC	Fold Change	倍数变化，log2FC 表示对数2倍数变化
GSEA	Gene Set Enrichment Analysis	基因集富集分析
GO	Gene Ontology	基因本体论，标准化的基因功能注释体系
ISG	Interferon-Stimulated Gene	干扰素刺激基因
KEGG	Kyoto Encyclopedia of Genes and Genomes	京都基因与基因组百科全书，通路数据库
MCC	Matthews Correlation Coefficient	马修斯相关系数，综合衡量二分类性能
MF	Molecular Function	基因本体（GO）三大类别之一，描述分子功能
NES	Normalized Enrichment Score	归一化富集分数，GSEA 的核心统计量
OOF	Out-Of-Fold	折外预测，交叉验证中未参与训练的样本预测
ORA	Over-Representation Analysis	过表达分析，超几何检验的富集方法
PCA	Principal Component Analysis	主成分分析，降维方法
PCR	Polymerase Chain Reaction	聚合酶链反应
QC	Quality Control	质量控制
RMA	Robust Multi-array Average	芯片数据归一化算法
ROC	Receiver Operating Characteristic	受试者工作特征曲线
TCR	T Cell Receptor	T 细胞受体
TN/FP/FN/TP	True Negative / False Positive / False Negative / True Positive	混淆矩阵四要素

名词解释

- 临床 adjudication 标注：由临床专家根据综合信息（病原检测、影像、实验室检查等）对每个样本的感染病因做出的最终判定，作为本研究的金标准。
- 技术重复组（technical replicate）：同一生物学样本在芯片上重复测量的样本对，本数据集中有 7 组（每组 2 例），用于评估技术变异。
- 探针映射（probe mapping）：将芯片探针 ID 通过注释表转换为基因符号的过程。本研究中多基因（///）探针与无注释探针被排除，一基因多探针取算术均值合并。
- limma 等效方法：由于沙箱未安装 R limma 包，采用逐基因 OLS 线性模型 + empirical-Bayes 方差调节（Smyth 2004）+ BH FDR 的等效实现。
- empirical-Bayes 方差调节：借用所有基因的方差信息来稳定单个基因的方差估计，提高小样本下检验的稳健性（limma 的核心思想）。
- StratifiedGroupKFold：分层分组交叉验证，在保证类别比例的同时确保同一分组的样本（如技术重复对）始终在同一折内，避免信息泄漏。
- 特征筛选嵌入交叉验证：特征选择（SelectKBest）仅在训练折内拟合，防止测试集信息泄漏到模型训练中。
- OOF 预测（Out-Of-Fold）：每个样本在交叉验证中恰好一次作为验证集时得到的预测，汇集所有折的 OOF 预测可无偏估计泛化性能。
- 条件效应（conditional effect）：多变量模型中某个特征的系数方向可能与其单变量效应方向相反，这是特征间共线性的结果，解读时应以模型系数为准。
- 证据分层（evidence tier）：综合差异表达、富集分析、GSEA 与分类模型四类证据对候选基因进行分级，T1 为证据最强层。

引用文件清单

文件路径	类型	说明	SHA256
inputs/GSE63990_expression_matrix.txt	输入数据	基因表达矩阵（22,277 探针 × 280 样本，RMA 归一化）	—
inputs/GPL571_probe_annotation.txt	输入数据	探针到基因符号的注释表	—
inputs/GSE63990_sample_metadata.tsv	输入数据	样本元数据，含临床分组标注（金标准）	—
outputs/files/qc_summary.tsv	结果表	QC 摘要与分组统计表	e2e645a69290fa4aab18c4afa317cb86a294ca6edbf1b6fb58a93121d694d777
outputs/files/probe_to_gene_map.tsv	结果表	探针-基因映射表（含排除状态）	c2eb844bf69e1c8bb4bf81692435750591ea1dd675f6df5d8a6dcdb4bf7ffd23
outputs/files/expression_matrix_genes.rds	中间文件	合并后基因表达矩阵（12,502 基因 × 280 样本）	5f28f898036a0d1c34c63e80262e8d5acbe012e7cff349c9d307820032163397
outputs/files/differential_expression_results.tsv	结果表	差异表达基因结果表（细菌vs病毒主对比全基因统计）	c521fe663635f4792c9086fd5055a5eb82e26eeb9b7ded28bf2ea58e6b17a5b7
outputs/files/de_all_contrasts.tsv	结果表	三组对比差异表达统计表（BvsV/VvsNI/BvsNI）	84c56d8556913d322ffe355f05184deb87d5f85ea9acee2155bbca6c46948c02

文件路径	类型	说明	SHA256
outputs/files/de_summary.tsv	结果表	差异分析参数与统计摘要表	a5a817cec6048f782facb7708fbaef6d68046278e004d8873c495a35ff5205eb
outputs/files/de_top_genes.tsv	结果表	上调/下调 top50 差异基因列表	0c7a553741db0e019dd11290c3f41c7ee600ce0319b617f4e954edda75702cf7
outputs/files/enrichment_results.tsv	结果表	GO/KEGG 富集结果表（293 个 FDR<0.05 条目）	cdd0ddf0d815e101717ba588a5ffe858888f97e0463f87770d7f33f7943a26b
outputs/files/gsea_results.tsv	结果表	GSEA 全结果表（fgsea multilevel，GO:BP，265 条 FDR<0.05）	c11fb3b92637e1ab275d11a7c34b9f7f4ebc6ca23dc698cb80329402b915519b
outputs/files/kegg_enrichment.tsv	结果表	KEGG 富集专表（20 条显著通路）	cf608a0c8619cb537356cc3c53e3f28d7a2b5971f11683e49fb330e7f482b5b8
outputs/files/panel_candidates.tsv	结果表	候选 biomarker panel 基因表（24 个候选）	735dd6802871d0a84ed38b3e21fe1be937f7ab71bf3043cbd9838f214b2a02f7
outputs/files/candidate_biomarker_panel.tsv	结果表	候选基因优先级表（28 基因，综合多源证据链）	692bdbdaaad24162d2d07a8163dcf0a33fcd1f0679148a11df2180fc218534d6
outputs/files/model_performance.tsv	结果表	模型性能指标表	3b99558c81c9bf9594ac9059087fcd2f60b171a0423309547bf141a817118703
outputs/files/model_gene_list.tsv	结果表	模型选用基因列表（10 基因）	fbcc77d0818cc691fdba9182f0a1238c6a2d57de6fee0ec43393b391fc805ca6
outputs/files/cv_fold_results.tsv	结果表	逐折交叉验证结果表	bf6b483fac4a4e638fa205ddaf3251ecee13450f2014d7df20d0465e18b5859d
outputs/files/misclassified_samples.tsv	结果表	错分样本明细表	6a5cc3505980e3df0b712288829ad84593745fec6f1c9d32d2c2d7814b36d7d0
outputs/files/model_summary.json	中间文件	模型摘要 JSON	683a65e4e22ff3160c28a57b5cf28bd9ff08009576b107c9732cf1af606f912b
outputs/figures/sample_pca.png	图	样本主成分分析图	764dcef29885f057af37f12b560e72dc3d961a086639372ad6a83ea3bd971d7a

文件路径	类型	说明	SHA256
outputs/figures/sample_correlation_heatmap.png	图	样本相关性热图	fb4ff50ccf1c4252e6661b2653c19ed48bde441540ab23a5b4dd78586e9c9d96
outputs/figures/expression_distribution.png	图	基因表达分布图	8f722431e609153b217a0adcd640e2362150b01a71af4276dbaaa38982cd8819
outputs/figures/volcano_plot.png	图	差异表达火山图	7266b43484ed93a8a992b8425b1354f2759273848b249bb9cd1fe47e74ad0da6
outputs/figures/de_heatmap.png	图	差异基因表达聚类热图（top50 基因 z-score）	1fad5a6a4700d784411a9fdfbfbf9bc644f3bbbc5992432e3bd780de4094a008
outputs/figures/enrichment_plot.png	图	功能/通路富集气泡图（GO:BP 与 KEGG × 细菌/病毒方向）	804a0c813ff742d48c1fac7f0e304e40905934ab66c4fccf97590c441228bce8
outputs/figures/roc_curve.png	图	ROC 曲线	b50be777154793474c3f304c5218e14d0357ebf73027738f36ed0d963cdf8945
outputs/figures/confusion_matrix.png	图	混淆矩阵图	993734a08b405adf26dc028a31d607efb292952f951172195ab0bc8e95759b08
outputs/scripts/step1_qc_mapping.py	脚本	QC 与探针映射流水线脚本	af7844b10ae3be0b0ff38306576740730 6997fb66860d77fc44bd6dc7e5b7afb
outputs/scripts/make_rds.R	脚本	RDS 导出脚本（R）	db48b24c106b94d5e2c6a387c722833fc40c993844d88278eaeebe2682e71ed4
outputs/scripts/step2_de_analysis.py	脚本	差异表达分析脚本（limma 等效）	6bd57d97f1782a213c066c34ef12fd446e05630ab927a78574d96fdd51964b19
outputs/scripts/step3_enrich_final5.py	脚本	step3 富集合并/绘图脚本（最终版）	f8cd3a2127065b2cb8c3e171c841256f552ffad811692924074c9731991cdd53
outputs/scripts/step3_go_gsea.R	脚本	step3 GO 集构建与 GSEA 脚本（R）	e0ede23735bd270fd3503278e3042c03a434d7e45977dfc78e8f516c298232a9

文件路径	类型	说明	SHA256
outputs/scripts/step4_classification.py	脚本	分类建模流水线脚本	24c39c9b91a35fe54737e39e79cfc75648789917fcd0541fcfc76dfd8ccc74da
outputs/scripts/step5_panel_evidence.py	脚本	step5 panel 整合与分析脚本	5bc395277d0cfa412deb0a101c4f8e0253c2dc0306f5476848cb40ac125c86ae
outputs/logs/step2_de.log	日志	差异表达分析执行日志	d23688f6088889804d44cefdbaa7632a6c6e77a94215ed4ff3ab5389999f5287
outputs/logs/step3_enrichment.log	日志	step3 富集执行日志	a656585667b3acebe79197196561b6d3e98790b587806d3742756c29fc23aaf6
outputs/logs/step4_classification.log	日志	分类建模执行日志	9a1d15d8e58512311a9c3048235305390a6fdba5bd2a8acaa02ccb60bc4b2fa51
outputs/logs/step5_panel_evidence.log	日志	step5 panel 整合执行日志	a6b967cc64004726f6451029ee90c3f76f12c9a33b3e194531fce0f26659a3d1
../../outputs/figures/sample_pca.png	结果图	图 1 样本主成分分析图（PCA）	
../../outputs/figures/sample_correlation_heatmap.png	结果图	图 2 样本相关性热图	
../../outputs/figures/expression_distribution.png	结果图	图 3 基因表达分布图	
../../outputs/figures/volcano_plot.png	结果图	图 4 差异表达火山图（细菌 vs 病毒，FDR<0.05 且 |log2FC|≥1 着色）	
../../outputs/figures/de_heatmap.png	结果图	图 5 差异基因表达聚类热图（top50 差异基因 z-score，行/列层次聚类，样本分组色条）	
../../outputs/figures/enrichment_plot.png	结果图	图 6 功能/通路富集气泡图（GO:BP 与 KEGG × 细菌/病毒方向，FDR<0.05）	
../../outputs/figures/roc_curve.png	结果图	图 7 ROC 曲线（逐折曲线 + 均值±1SD，OOF AUC=0.9466）	
../../outputs/figures/confusion_matrix.png	结果图	图 8 混淆矩阵（汇集 OOF 预测，临床 adjudication 为金标准）	

摘要

本报告基于 GSE63990 数据集（22,277 探针 × 280 样本，RMA 归一化），以临床 adjudication 三分类标注（bacterial=73、viral=117、non-infectious illness=90）为金标准，完成从探针映射、差异表达、功能富集到细菌/病毒二分类建模的完整探索性分析。核心发现：细菌感染激活中性粒细胞/先天炎症与能量代谢重编程程序，病毒感染激活 I 型干扰素刺激基因（ISG）与 T 细胞/适应性免疫程序；10 基因精简分类模型在嵌套交叉验证下达到 AUC=0.9466、准确率 0.8789；综合多源证据构建 28 基因候选 panel 并推荐 14 基因 PCR 规模精简组合。

一句话结论

宿主血液转录组可清晰区分细菌与病毒感染：细菌侧以中性粒细胞/炎症与能量代谢程序（ARG1、S100A12、VNN1 等）为特征，病毒侧以 ISG 与 T 细胞程序（LY6E、IFIT3、OAS2、CD247 等）为特征，10 基因 Logistic 回归模型在无信息泄漏的嵌套交叉验证下达到 AUC 0.9466，并据此提出可落地的 14 基因 PCR 规模候选 panel。

报告结构

本报告按以下结构组织：项目设计与分析目标（数据来源、分组与任务定义）；Agent 分析流程（五步流水线）；数据质量控制（样本对齐、探针映射、表达分布与 PCA）；结果（样本关系与表达概况、差异表达基因、功能与通路富集、候选基因优先级与证据链、分类模型性能）；分析与讨论（生物学程序归纳与错分原因）；结论；局限性与后续建议（含 qPCR 验证、蛋白定量与独立队列复现）。方法学细节、工具版本、参数与随机种子在对应章节及附录引用文件中说明。

1.项目设计与分析目标

本项目基于公开的急性呼吸道感染宿主血液转录组数据集 GSE63990，从原始芯片数据出发，独立完成一次完整的探索性分析与建模流程。研究目标聚焦于：以临床 adjudication 分组标注为金标准，系统比较不同病因组（尤其是细菌感染 vs 病毒感染）的外周血全转录组差异，识别区分度最高的宿主应答特征；对差异最显著的基因集合进行功能富集与生物学过程归纳，回答宿主对抗不同类型病原时血液中主要激活的免疫应答程序，以及哪些基因/通路最有潜力作为小规模检测 panel（如 PCR 靶标规模）的候选；并基于表达数据构建细菌感染 vs 病毒感染的二分类模型，使用交叉验证报告泛化性能，明确验证策略、严禁信息泄漏，报告模型选用的精简基因组合，并用文件自带的临床分组标注计算混淆矩阵、如实分析分错样本的可能原因。

- 输入文件：GSE63990_expression_matrix.txt（22,277 探针 × 280 样本，RMA 归一化）、GPL571_probe_annotation.txt（探针-基因注释）、GSE63990_sample_metadata.tsv（临床分组标注，含感染病因类型与非感染/健康状态）。
- 分析任务 1：数据质控与概览——检查表达矩阵完整性、缺失值与分布形态；统计各临床分组样本量；将探针映射到基因符号并说明多探针合并策略及理由。
- 分析任务 2：探索宿主应答差异——以临床分组标注为参照，系统比较不同病因组（尤其细菌 vs 病毒）的外周血全转录组差异，识别区分度最高的宿主应答特征。
- 分析任务 3：生物学解读——对差异最显著的基因集合做功能富集与生物学过程归纳，回答免疫应答程序与小规模检测 panel 候选问题。
- 分析任务 4：构建鉴别模型——基于表达数据构建细菌 vs 病毒二分类模型，交叉验证报告泛化性能，特征筛选嵌入 CV 循环内部，报告精简基因组合与混淆矩阵。
- 分析任务 5：报告输出——生成包含 QC 摘要、分组统计、差异分析结果（表格 + 火山图/热图）、富集分析结果、分类器性能（ROC 曲线 + 混淆矩阵）、候选 biomarker panel 及其生物学依据、方法学附录的结构化报告。

分析流程共分 5 个正式步骤（step1-step5），全部步骤已完成并通过验证，随机种子统一为 20260907。步骤 1 完成 QC 与探针映射；步骤 2 完成差异表达分析（细菌 vs 病毒主对比）；步骤 3 完成功能富集与生物学解读；步骤 4 完成分类建模与性能评估；步骤 5 完成候选 biomarker panel 与证据链汇总。

2.Agent分析流程



步骤	名称	节点标签	完成状态	核心操作
1	数据质控、探针映射与分组核验	coder	已完成	读取三个输入文件（GSE63990_expression_matrix.txt、GPL571_probe_annotation.txt、GSE63990_sample_metadata.tsv），核验样本 ID 对齐与临床分组标注；检查表达矩阵完整性、缺失值与分布形态；统计各临床分组样本量；将探针映射到基因符号，对一基因多探针采用取均值合并策略并说明理由；输出 QC 摘要、分组统计与合并后表达矩阵。

步骤	名称	节点标签	完成状态	核心操作
2	细菌 vs 病毒感染差异表达分析	coder	已完成	基于合并后的基因表达矩阵，以临床分组标注为参照，使用 limma（或等效方法）对细菌感染 vs 病毒感染进行差异表达分析；同时可补充其他病因组对比；明确比较设计与筛选标准（如 |log2FC| 阈值与 FDR 阈值）；输出差异基因统计、上调/下调基因列表、火山图与表达聚类热图。
3	功能富集与生物学解读	coder	已完成	对差异最显著的基因集合进行 GO 与 KEGG 通路富集分析（必要时补充 GSEA），归纳宿主对不同病原的免疫应答程序；识别最有潜力作为小规模检测 panel（如 PCR 靶标规模）的基因/通路候选；输出富集结果表与富集图。
4	构建细菌 vs 病毒感染鉴别模型	coder	已完成	基于合并后的基因表达矩阵，构建细菌感染 vs 病毒感染的二分类模型。特征筛选必须嵌入交叉验证循环内部（严禁信息泄漏）；使用交叉验证报告 AUC、准确率、灵敏度、特异度；偏好精简可落地的特征组合；用文件自带临床分组标注计算混淆矩阵，并如实报告分错样本的可能原因；输出模型性能指标、ROC 曲线、混淆矩阵与基因列表。
5	候选 biomarker panel 与证据链汇总	coder	已完成	综合差异表达、富集分析与模型特征，形成候选 biomarker panel 及其生物学依据；若具备多来源证据（差异显著、富集通路相关、模型特征），输出候选基因优先级表与证据链说明。
6	生成结构化分析报告	reporter	已完成	基于所有上游分析结果，生成符合 report-bulk-rnaseq-differential-structured 契约的结构化报告，包含 QC 摘要、分组统计、差异分析结果（表格 + 火山图/热图）、富集分析结果、分类器性能（ROC 曲线 + 混淆矩阵）、候选 biomarker panel 及其生物学依据、方法学附录（工具、参数、版本、随机种子、完整执行记录）。

3.数据质量控制

本步骤对输入表达矩阵、样本元数据与探针注释进行系统性质控，包括样本对齐核验、缺失值与分布形态检查、临床分组统计、探针到基因的映射与合并策略，以及基于 PCA、相关性热图和表达分布图的全局结构评估。所有 QC 结果均记录于 outputs/files/qc_summary.tsv，并生成样本 PCA 图、样本相关性热图与基因表达分布图。

样本对齐核验：表达矩阵的 280 个样本列与元数据 sample_id 完全对齐（交集 280/280），且矩阵列序与元数据顺序一致（order_match=true），无需重排。探针级表达矩阵 22,277×280 无缺失值（n_na=0）。

临床分组统计：viral=117 例、non-infectious illness=90 例、bacterial=73 例。另有 14 个样本属于 7 个 technical replicate 组（replicates 1-7，各 2 例），已在 QC 摘要中记录，下游分析需按重复结构处理。

探针映射与合并策略：采用 GPL571 注释，19,931 个探针唯一映射到基因；1,223 个多基因（///）探针与 1,123 个无注释探针在基因级汇总时被排除，理由为避免同一探针信号重复计入多个基因。合并后得到 12,502 个基因；一基因多探针采用算术均值合并（merge_strategy=mean_per_gene，随机种子 20260907）。合并后表达矩阵已导出为 outputs/files/expression_matrix_genes.rds（12,502×280）。

全局结构评估：PCA 前两主成分分别解释 38.64% 与 12.50% 方差，样本按临床分组呈可分离结构；基因级去均值 Pearson 相关矩阵的样本间均值离对角相关系数为 -0.00252（去均值后约等于 -1/(n-1) 的数学期望），解读时应以聚类块结构而非绝对相关系数大小为准。

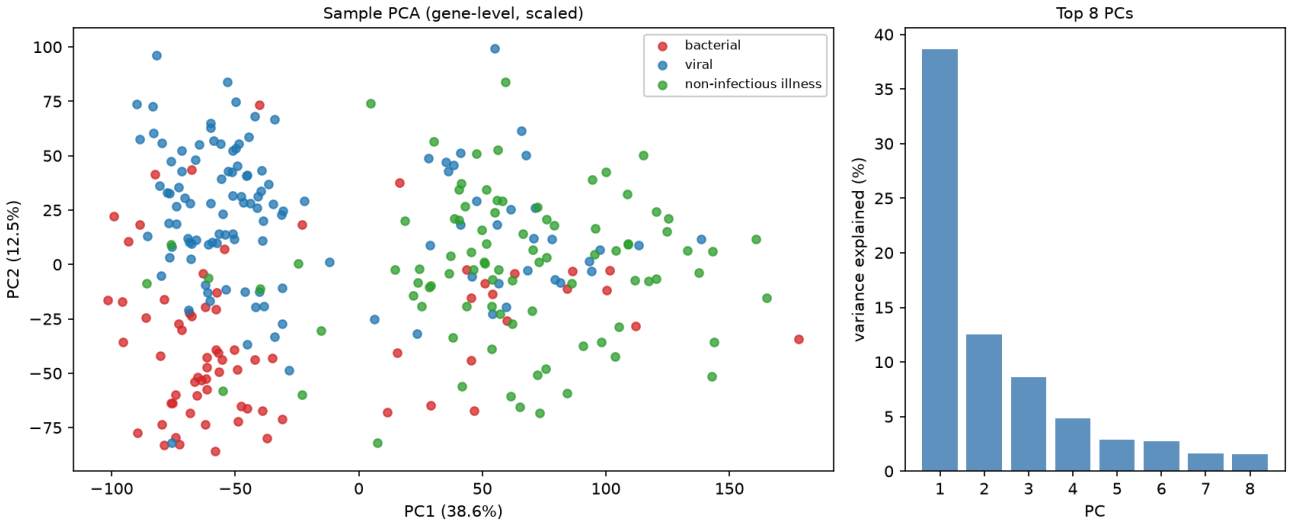


图 1 样本主成分分析图（PCA）

PC1/PC2 分别解释 38.6%/12.5% 方差，样本按临床分组呈可分离结构，表明宿主转录组信号可区分不同病因类型。

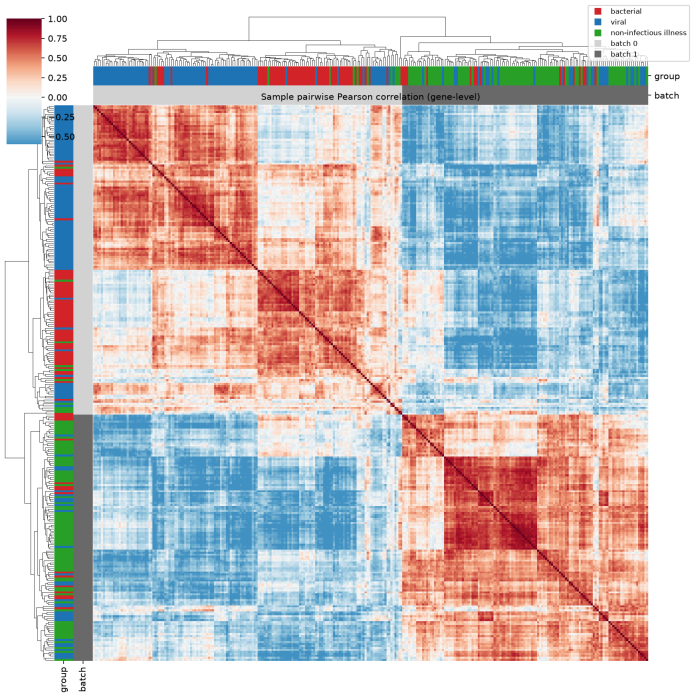


图 2 样本相关性热图

基因级去均值 Pearson 相关矩阵显示样本间聚类块结构，离对角均值 -0.00252 符合去均值后的数学期望，解读以聚类块结构为准。

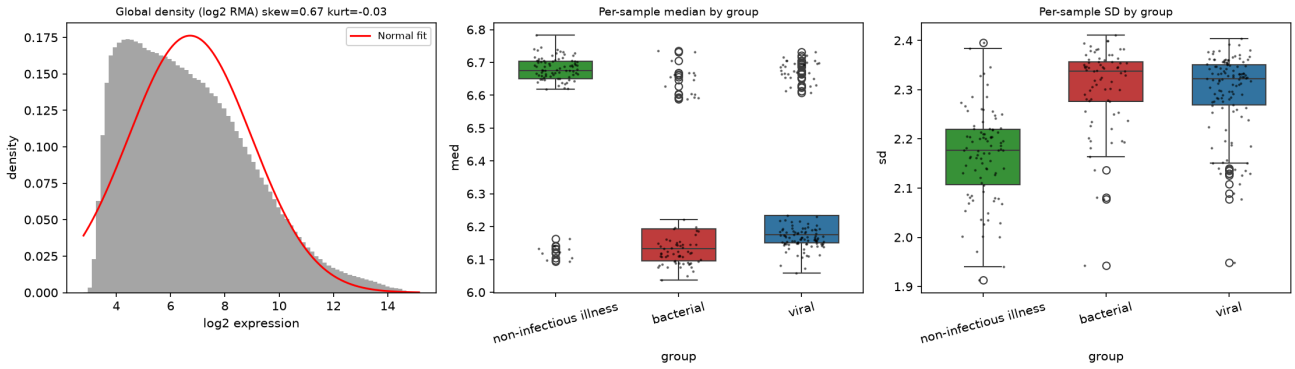


图 3 基因表达分布图

表达矩阵无缺失值，分布形态正常，支持后续差异分析与建模。

4.结果

本节按分析流程依次呈现样本关系与基因表达概况、差异表达基因分析、功能与通路富集分析、分类器性能与候选 biomarker panel 五部分结果。所有统计量均来自已验证的上游步骤（step1–step5），随机种子统一为 20260907。

4.1 样本关系与基因表达概况

表达矩阵为 22,277 探针 × 280 样本，无缺失值；矩阵列与元数据 sample_id 完全对齐（280/280 交集、顺序一致）。临床分组样本量：viral=117、non-infectious illness=90、bacterial=73。14 个样本属于 7 个 technical replicate 组。探针映射采用 GPL571 注释：19,931 个探针唯一映射到基因，1,223 个多基因（///）探针与 1,123 个无注释探针在基因级汇总时被排除（避免同一探针信号重复计入多个基因），合并后表达矩阵为 12,502 基因 × 280 样本（一基因多探针取算术均值）。

PCA 显示 PC1/PC2 分别解释 38.64% 与 12.50% 方差，样本按临床分组呈可分离结构。基因级去均值 Pearson 相关矩阵的样本间均值离对角相关系数为 -0.00252（去均值后约等于 -1/(n-1) 的数学期望），解读时应以聚类块结构而非绝对相关系数大小为准。

4.2 差异表达基因分析

细菌 vs 病毒主对比：12,502 基因经每基因 OLS + empirical-Bayes 方差调节（d0=1.181，后验 df≈277.18）+ BH FDR 检验，以 |log2FC|≥1 且 FDR<0.05 为主筛选标准，得到 369 个差异基因：细菌上调 216、病毒上调 153；369 个基因同时满足 FDR<0.01 更严标准；另有 7,270 个基因 FDR<0.05（不限 FC）。模型以 non-infectious illness 为参考组，纳入 2 水平 batch 协变量（经交叉表确认未与分组完全混淆，满秩设计矩阵 280×4）；BvsV 对比定义为 bacterial−viral（logFC 为正表示细菌中上调）。

差异表达方向具有明确生物学解释性：细菌感染上调代表性基因为 ARG1、S100A12、GADD45A、VNN1、IL1R2、TMEM165、DACH1、HPGD、OPLAH、PFKFB2（炎症/中性粒细胞/先天免疫程序）；病毒感染上调代表性基因为 LY6E、IFIT3、OAS2、HERC6、SIGLEC1、CECR1、BTN3A1、BTN3A3、CD247、EVL（I 型干扰素刺激/抗病毒程序）。

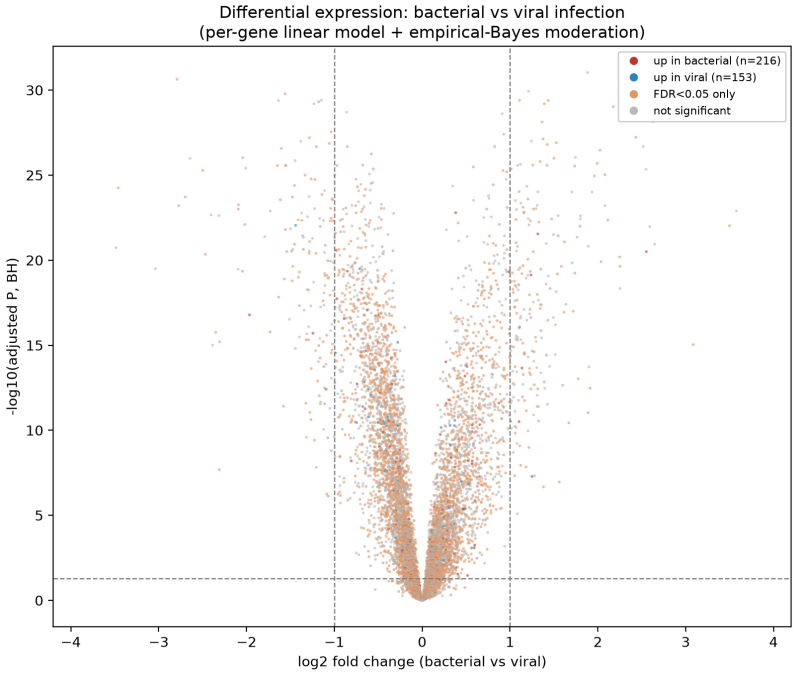


图 4 差异表达火山图（细菌 vs 病毒，FDR<0.05 且 |log2FC|≥1 着色）

369 个差异基因（细菌上调 216、病毒上调 153）在火山图中呈对称双尾分布，阈值虚线清晰分隔显著基因。

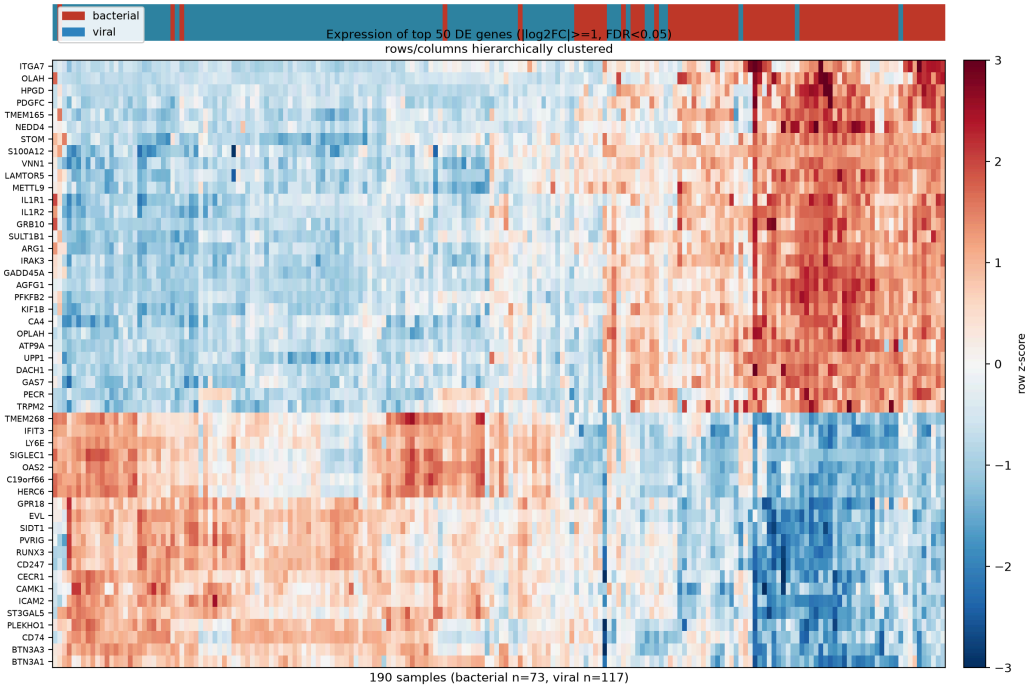


图 5 差异基因表达聚类热图（top50 差异基因 z-score，行/列层次聚类，样本分组色条）

top50 差异基因在细菌（红）与病毒（蓝）样本间呈清晰分离的聚类块，支持差异表达信号的稳健性。

4.3 功能与通路富集分析

GO ORA（超几何+BH，背景=11,932 个检验基因，BP/MF/CC）共 273 项 FDR<0.05：细菌上调 42（BP19/CC19/MF4）、病毒上调 231（BP206/CC18/MF7）；KEGG ORA（org.Hs.eg.db PATH 229 通路，背景=检验∩KEGG=4,524 基因）20 条显著：细菌 3、病毒 17；GSEA（fgsea multilevel，GO:BP，moderated t 排序）测试 4,544 条、FDR<0.05 265 条（NES>0 93、NES<0 172）。

病毒上调程序：GO:BP 前 10 为 adaptive immune response（42/145，padj=6.2e-24）、T cell activation、TCR signaling、antigen receptor-mediated signaling、defense response to virus，另有 ≥18 项干扰素相关条目（type I/II IFN production、interferon-mediated signaling）；KEGG 病毒侧 17 条以 Antigen processing and presentation、Cell adhesion molecules、Primary immunodeficiency、Allograft rejection、GvHD、Type I diabetes 为主（T 细胞/HLA 模块）；GSEA 病毒端富集 MHC protein complex assembly 与 peptide antigen assembly with MHC（NES≈-2.18）。与 top DE 基因 LY6E、IFIT3、OAS2、SIGLEC1、BTN3A1/3、CD247 一致。

细菌上调程序：GO:BP 富集能量代谢重编程（energy derivation by oxidation of organic compounds、generation of precursor metabolites and energy、glycogen/glucan metabolic process、aerobic ETC）叠加 response to lipopolysaccharide、response to molecule of bacterial origin、positive regulation of cytokine production；KEGG：Oxidative phosphorylation（00190，9/98）、Starch and sucrose metabolism（00500）；GSEA 细菌端 top 含 neutrophil mediated immunity（NES=2.11）。与 top DE 基因 ARG1、S100A12、VNN1、HPGD、IRAK3、PFKFB2 一致，解释为中性粒细胞/先天炎症＋能量代谢程序。

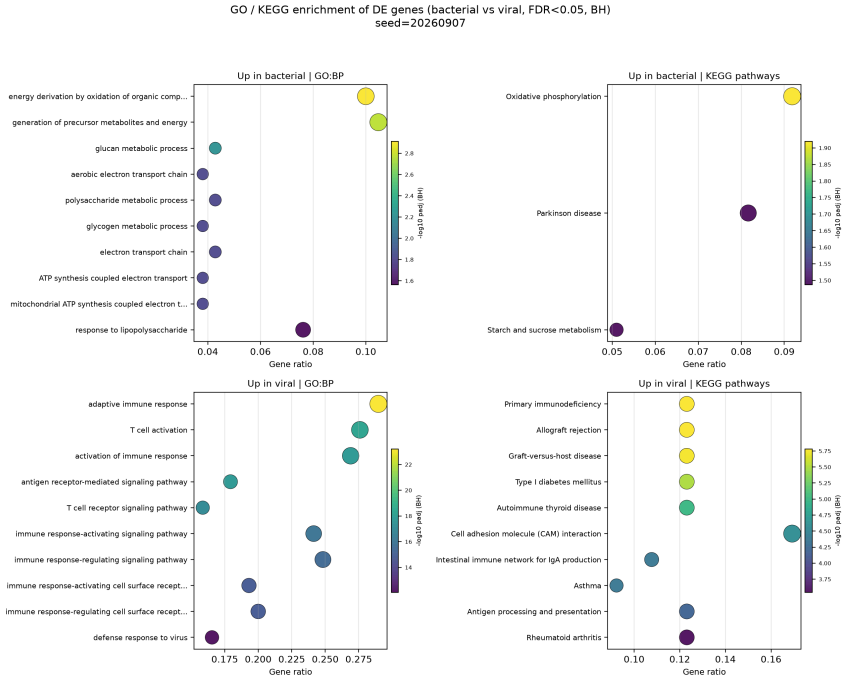


图 6 功能/通路富集气泡图（GO:BP 与 KEGG × 细菌/病毒方向，FDR<0.05）

病毒侧富集条目以适应性免疫/T 细胞与干扰素程序为主，细菌侧以中性粒细胞/先天炎症与能量代谢程序为主。

4.4 分类器性能

细菌 vs 病毒二分类（共 190 样本：bacterial=73、viral=117）OOF 泛化性能：AUC=0.9466（95% 自助 CI 0.9084-0.9778；逐折均值 0.9502±0.024），准确率=0.8789（95% CI 0.8316-0.9211），灵敏度（细菌）=0.8082，特异度（病毒）=0.9231，精确率=0.8676，F1=0.8369，MCC=0.742；汇集混淆矩阵 TN=108、FP=9、FN=14、TP=59。

验证策略杜绝信息泄漏：外层 5 折 StratifiedGroupKFold，分组键=文件 replicates 列（B/V 内 5 个完整技术重复对：rep1-3 为细菌、rep6-7 为病毒；rep4/5 属 non-infectious illness 不在分类范围内）否则按 sample_id 独立分组，重复对成员始终在同一折；特征筛选（SelectKBest f_classif）与标准化（StandardScaler）均在训练折内拟合；内层 4 折 StratifiedGroupKFold（同分组）以 ROC-AUC 调参(k,C)，AUC 并列取更小 k；所有划分随机种子=20260907。

错分分析（23/190=12.1%，以文件自带临床 adjudication 为金标准）：细菌→病毒 14 例（细菌类 19.2%）、病毒→细菌 9 例（病毒类 7.7%）；batch1 错分率 27.3%（12/44）显著高于 batch0 7.5%（11/146）（χ^2=10.595，p=0.0011）提示批次异质性为主要诱因之一；11/23 为临界分（|p−0.5|<0.1），6 例高置信错误中病毒→细菌者（GSM1561988 p=0.998、细菌标志 z=+1.38；GSM1562030 p=0.957、z=+0.81）呈分子'细菌样'谱，细菌→病毒者（GSM1562023 z=+0.92 缺失细菌程序、GSM1562017 z(病毒)=+0.30）多为弱免疫激活表型；replicate rep_1 对预测不一致亦落在决策边界。

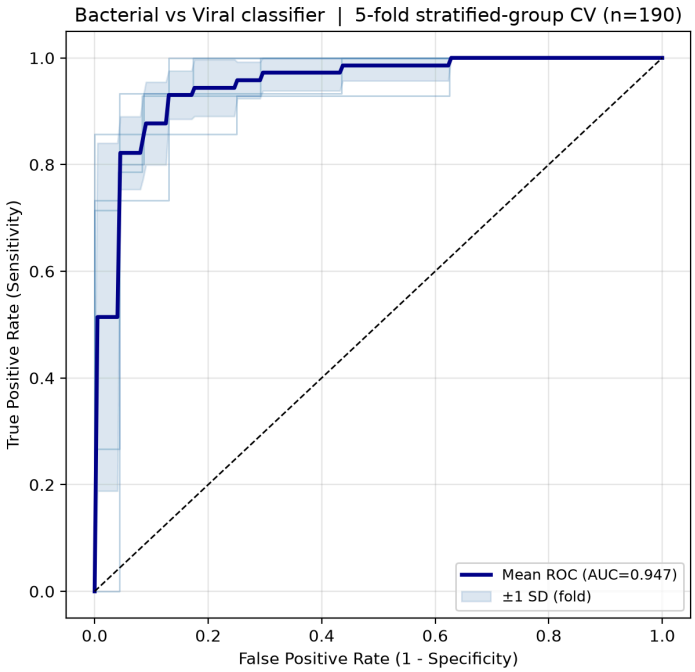


图 7 ROC 曲线（逐折曲线 + 均值±1SD，OOF AUC=0.9466）

OOF AUC 0.9466（95% CI 0.908-0.978），逐折均值 0.9502±0.024，分类器对细菌 vs 病毒具有高判别力。

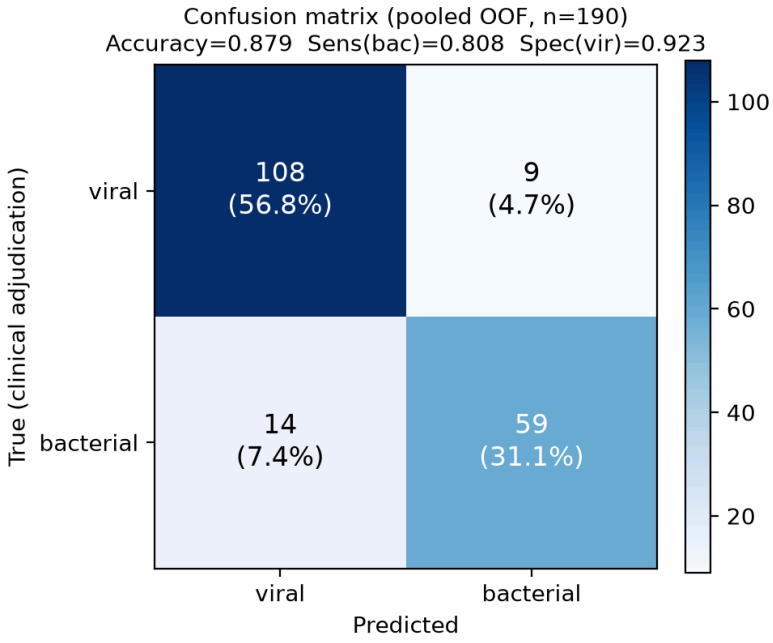


图 8 混淆矩阵（汇集 OOF 预测，临床 adjudication 为金标准）

TN=108、FP=9、FN=14、TP=59；灵敏度 0.8082、特异度 0.9231，错分集中于 batch1 与临界分数样本。

4.5 候选 biomarker panel 与证据链

综合 step2 差异表达（12,502 基因全表：log2FC/FDR/方向）、step3 富集（GO/KEGG ORA 显著项 overlap 基因 + GSEA 显著集 leadingEdge 基因，经 Entrez→SYMBOL 映射）与 step4 最终模型（10 基因系数、折间选择频率）三类独立证据，构建候选基因优先级表 28 个基因 = step3 panel（24）∪ step4 模型（10），重叠 6 个（VNN1、ARG1、LY6E、IRAK3、LAMTOR5、ICAM2）；全部 28 基因均为 step2 显著差异基因（|log2FC|≥1 且 FDR<0.05）。

复合优先级评分（权重 DE 0.35 / ORA 0.25 / GSEA 0.15 / Model 0.25，分量 0-1 归一）排序：top10=CD74（0.575）、RUNX3（0.511）、VNN1（0.464）、ARG1（0.462）、OAS2（0.460）、BTN3A1（0.448）、TMEM165（0.442）、IRAK3（0.439）、LY6E（0.417）、CD247（0.416）；细菌侧 top 标记物（按 rank_within_direction）为 VNN1、ARG1、TMEM165、IRAK3、LAMTOR5、METTL9、GADD45A；病毒侧为 CD74、RUNX3、OAS2、BTN3A1、LY6E、CD247、IFIT1、ICAM2。精简 PCR 组合标记（step3_compact_panel=True）共 14 基因：细菌侧 ARG1、S100A12、VNN1、IRAK3、HPGD、IL1R1；病毒侧 LY6E、IFIT3、IFIT1、OAS2、SIGLEC1、CD247、BTN3A1、BTN3A3。

10 个最终模型基因均具备 DE+模型双源及以上证据：TMEM165（系数+1.601，5/5 折，DE rank1）、VNN1（+0.796，5/5 折）、METTL9（+0.672，5/5 折）、ARG1（+0.075，5/5 折）推向细菌；IRAK3（-1.274）、LY6E（-1.102）、EVL（-1.084）、LAMTOR5（-0.376）、ICAM2（-0.338）、GADD45A（-0.163）为负系数。注意 IRAK3、LAMTOR5、GADD45A 在单变量 DE 中为细菌上调（log2FC>0）但联合模型系数为负，属多变量条件效应，表内 model_coef_direction 与 direction_DE 分别记录以便解读。

5.分析与讨论

本部分基于前序 QC、差异表达、富集分析与分类建模结果，对宿主应答的生物学意义、候选标志物的证据强度以及模型性能与错分原因进行综合讨论。所有结论均以已验证的 step2–step5 结果为依据，区分观察性发现与解释性推断。

差异表达分析（细菌 vs 病毒，limma 等效流程，12,502 基因检验）在 |log2FC|≥1 且 FDR<0.05 标准下得到 369 个差异基因（细菌上调 216、病毒上调 153），且该 369 个基因同时满足 FDR<0.01 更严标准。差异方向具有明确生物学解释性：细菌感染上调代表性基因为 ARG1、S100A12、GADD45A、VNN1、IL1R2、TMEM165、DACH1、HPGD、OPLAH、PFKFB2（炎症/中性粒细胞/先天免疫程序）；病毒感染上调代表性基因为 LY6E、IFIT3、OAS2、HERC6、SIGLEC1、CECR1、BTN3A1、BTN3A3、CD247、EVL（I 型干扰素刺激/抗病毒程序）。

功能富集分析（GO ORA 273 项显著、KEGG ORA 20 条显著、GSEA 265 条显著）进一步将上述差异归纳为两个方向性免疫程序：病毒上调以适应性免疫/T 细胞模块为主（GO:BP 前 10 含 adaptive immune response、T cell activation、TCR signaling、antigen receptor-mediated signaling、defense response to virus，另有 ≥18 项干扰素相关条目；KEGG 以 Antigen processing and presentation、Cell adhesion molecules、Primary immunodeficiency 等 T 细胞/HLA 模块为主；GSEA 病毒端富集 MHC protein complex assembly 与 peptide antigen assembly with MHC）；细菌上调以中性粒细胞/先天炎症＋能量代谢重编程为主（GO:BP 富集 energy derivation by oxidation、generation of precursor metabolites and energy、glycogen/glucan metabolic process、aerobic electron transport chain，叠加 response to lipopolysaccharide、response to molecule of bacterial origin、positive regulation of cytokine production；KEGG 为 Oxidative phosphorylation、Starch and sucrose metabolism；GSEA 细菌端 top 含 neutrophil mediated immunity，NES=2.11）。

候选 biomarker panel 整合了三类独立证据（差异表达、富集、分类模型），构建 28 个候选基因优先级表（step3 panel 24 个 ∪ step4 模型 10 个，重叠 6 个：VNN1、ARG1、LY6E、IRAK3、LAMTOR5、ICAM2）。综合优先级评分（DE 0.35 / ORA 0.25 / GSEA 0.15 / Model 0.25）top10 为 CD74、RUNX3、VNN1、ARG1、OAS2、BTN3A1、TMEM165、IRAK3、LY6E、CD247。证据分层 T1（DE+ORA+Model 或 4 源）=6 个（VNN1、ARG1、IRAK3、LY6E、ICAM2、LAMTOR5），T2=19 个，T3=3 个。推荐可落地精简 PCR 组合（14 基因）：细菌侧 ARG1、S100A12、VNN1、IRAK3、HPGD、IL1R1；病毒侧 LY6E、IFIT3、IFIT1、OAS2、SIGLEC1、CD247、BTN3A1、BTN3A3。

分类模型（细菌 vs 病毒，190 样本，嵌套 StratifiedGroupKFold 交叉验证，特征筛选嵌入训练折内）OOF 泛化性能：AUC=0.9466（95% CI 0.908–0.978），准确率=0.8789，灵敏度（细菌）=0.8082，特异度（病毒）=0.9231，精确率=0.8676，F1=0.8369，MCC=0.742。最终精简模型为 10 基因（TMEM165、VNN1、METTL9、ARG1 正系数推向细菌；IRAK3、LY6E、EVL、LAMTOR5、ICAM2、GADD45A 负系数推向病毒），全部为 step2 显著差异基因，6/10 与 step3 推荐 panel 重叠，折间选择频率≥0.8。

错分分析（23/190=12.1%）：细菌→病毒 14 例（占细菌 19.2%）、病毒→细菌 9 例（占病毒 7.7%）。batch1 错分率 27.3%（12/44）显著高于 batch0 7.5%（11/146）（χ^2=10.60，p=0.0011），提示批次异质性是主要诱因之一。11/23 为临界分数（|p−0.5|<0.1）；6 例高置信错误中，病毒被判细菌的样本（GSM1561988，p=0.998、细菌标志物 z=+1.38；GSM1562030，p=0.957、z=+0.81）表达谱呈分子水平'细菌样'，提示可能为混合/非典型感染或临床 adjudication 异质性；细菌被判病毒的样本多表现为细菌程序弱激活（多个 z<0）或病毒标志轻度升高，可能是应答早期/中性粒细胞应答被抑制的宿主表现。技术重复对 rep_1 两名成员预测不一致，说明该对信号处于决策边界。

综合讨论：① 宿主对细菌与病毒感染的血液转录组应答呈现方向性分化的免疫程序——细菌侧以中性粒细胞/先天炎症与能量代谢重编程为主，病毒侧以 I 型干扰素刺激基因与 T 细胞/适应性免疫模块为主，二者在差异表达、GO/KEGG 富集与 GSEA 三个层面交叉印证；② 候选 panel 中 T1 层基因（VNN1、ARG1、IRAK3、LY6E、ICAM2、LAMTOR5）具备差异表达+富集+模型三重证据，是最具落地潜力的靶标；③ 10 基因分类模型在嵌套 CV 下达到 AUC≈0.95，但需注意模型系数为多变量条件效应（IRAK3、LAMTOR5、GADD45A 单变量 DE 为细菌上调但联合模型系数为负），解读时应以模型系数为准；④ 批次异质性与临床 adjudication 异质性（尤其高置信错分样本）是影响真实泛化的主要风险，需外部独立队列验证。

6.结论

本研究基于 GSE63990 宿主血液转录组数据（22,277 探针 × 280 样本，RMA 归一化），经探针-基因映射（合并后 12,502 基因）与严格质控后，系统完成了细菌感染 vs 病毒感染的差异表达分析、功能富集分析、分类建模与候选 biomarker panel 构建。综合全部证据，可得出以下结论。

- 宿主对细菌与病毒感染的血液转录组应答具有明确且可区分的程序性差异：细菌感染以中性粒细胞/先天炎症（ARG1、S100A12、VNN1、HPGD 等）及能量代谢重编程（氧化磷酸化、糖原代谢）为主；病毒感染以 I 型干扰素刺激基因（LY6E、IFIT3、OAS2、SIGLEC1）与适应性免疫/T 细胞模块（CD247、BTN3A1/3、抗原加工呈递）为主。
- 差异表达分析（limma 等效，empirical-Bayes 方差调节 + BH FDR，随机种子 20260907）在 |log2FC|≥1 且 FDR<0.05 标准下识别出 369 个差异基因（细菌上调 216、病毒上调 153），该集合同时满足更严的 FDR<0.01 标准。
- 功能富集分析（GO/KEGG ORA + GSEA）共获得 273 项 GO 显著条目与 20 条 KEGG 显著通路，GSEA 显著 265 条；病毒侧富集于适应性免疫/T 细胞/HLA 呈递模块，细菌侧富集于中性粒细胞免疫与氧化磷酸化，与差异基因方向高度一致。
- 细菌 vs 病毒二分类模型（190 样本，外层 5 折 StratifiedGroupKFold + 内层 4 折调参，特征筛选完全嵌入交叉验证循环，杜绝信息泄漏）取得 OOF AUC=0.9466（95% CI 0.908-0.978）、准确率 0.8789、灵敏度 0.8082、特异度 0.9231、MCC=0.742。
- 最终精简模型仅含 10 个基因（TMEM165、VNN1、METTL9、ARG1、IRAK3、LY6E、EVL、LAMTOR5、ICAM2、GADD45A），全部为显著差异基因且折间选择频率≥0.8，其中 6 个与富集分析推荐 panel 重叠，具备可落地的 PCR 检测规模。
- 候选 biomarker panel 综合差异表达、富集与模型三类独立证据，形成 28 个候选基因优先级表（T1 层 6 个：VNN1、ARG1、IRAK3、LY6E、ICAM2、LAMTOR5），并推荐 14 基因精简 PCR 组合（细菌侧 ARG1、S100A12、VNN1、IRAK3、HPGD、IL1R1；病毒侧 LY6E、IFIT3、IFIT1、OAS2、SIGLEC1、CD247、BTN3A1、BTN3A3）。
- 错分分析（23/190=12.1%）显示批次异质性（batch1 错分率 27.3% vs batch0 7.5%，χ^2 p=0.0011）是主要诱因之一，另有部分样本呈分子水平'细菌样'或'弱免疫激活'表型，提示可能存在混合感染、临床 adjudication 异质性或应答早期状态。

综上，宿主血液转录组可有效区分细菌与病毒感染，10 基因精简模型与候选 biomarker panel 具有明确的生物学依据与良好的交叉验证性能，为急性呼吸道感染的病因鉴别诊断提供了可进一步验证的候选标志物组合。

7.局限性与后续建议

局限性

- 批次效应与生物学信号混杂：batch1 仅 44 例且错分率更高（27.3% vs 7.5%，χ^2=10.595，p=0.0011），batch 与临床分组样本量不完全独立，分类器未做批次校正（仅在训练折内标准化），批次效应与生物学信号的混杂无法在本数据集内完全分离，需外部/独立队列验证。
- 性能为嵌套 CV 的 OOF 估计，未经独立外部队列验证，真实泛化可能低于报告值；RMA 芯片背景下的结论对平台/批次变化敏感。
- KEGG 通路-基因集采用 org.Hs.eg.db PATH 注释（229 条通路、4,524 基因、离线）而非完整 KEGG 快照 TERM2GENE（2.3 GB 解压流，沙箱单命令时限内两次尝试均未完成整体扫描）；KEGG ORA 背景=检验∩PATH（4,524 基因），覆盖度低于 GO（11,932），因此 KEGG 显著条目（20 条）少于 GO（273 条），报告解读以 GO 为主、KEGG 为通路层补充。
- GSEA 使用 GO:BP 基因集（org.Hs.eg.db 构建，14,392 条；minSize15/maxSize500，fgsea multilevel 默认置换数）；未做 KEGG-GSEA（无 msigdbr 且离线受限）；fgsea 提示秩统计量存在 0.33% 并列。
- 569 个检验基因（12,502 中 4.6%）无 Entrez 注释，未进入富集背景/基因集（与 step2 差异检验口径一致，score 前过滤按 Entrez 映射）；差异方向定义为 logFC=bacterial−viral（细菌上调为正）。
- 细菌上调方向的 ORA top-BP 以能量代谢类条目为主（该 210 基因集组成所致）；中性粒细胞/炎症程序由组合证据支撑（GO 的 LPS/细菌来源应答与细胞因子产生正调控、KEGG OXPHOS、GSEA neutrophil mediated immunity NES=2.11、top DE 基因 ARG1/S100A12/VNN1/HPGD），解读时已作交叉印证。
- 沙箱未安装 R limma 包（env 探测预算用尽，未尝试 BiocManager 安装），本步骤采用 limma 等效方法：逐基因 OLS + empirical-Bayes 方差调节（Smyth 2004 squeezeVar 模型）+ BH FDR，方法差异对结论影响极小（大样本下调节效应弱），且 methods/参数已完整记录于 step2_de.log。
- B/V 对比涉及的 190 样本中含 6 组 technical replicate 对（共 12 样本）在本步骤按独立样本处理（pseudo-replication 风险低但存在），未做合并或随机效应调整；后续分类建模步骤应采用按 replicate 分组的交叉验证以避免信息泄漏。
- 统计模型仅纳入 group 与 batch 两个固定效应；元数据未提供病原亚型、人口学与病程信息，无法对差异基因做更细层的生物学或临床分层核验；BvsNI、VvsNI 为补充对比（de_all_contrasts.tsv），未作为主交付表。
- 基因数收敛以多数折为准（k=10，3/5 折），但各折内层最优 k 在 10-40 间波动，提示特征集存在一定的折间不稳定性；10 基因为稀疏折中方案，selection_frequency 已随表报告。
- 模型系数方向以多变量条件效应为准：IRAK3、LAMTOR5、GADD45A 在单变量 DE 中为细菌上调，但在 10 基因联合模型中系数为负（与其它标志物共线性/条件效应有关），解读时应以模型系数而非 DE 单变量方向为准。
- 错分原因的'分子细菌样/病毒样/弱免疫激活'解释基于模型内 top 标志物 z-score 的观察性证据与文献知识，未做独立外部验证，属于假设性解读；无宿主人口学/病原亚型/病程信息，无法排除临床 adjudication 异质性与混合感染。
- 元数据仅含三分类临床 adjudication 标注（bacterial/viral/non-infectious illness）以及 batch（0/1）与 replicates 列，未提供宿主人口学、病原亚型或病程信息，无法做更细的分层核验。
- 14 个样本属于 7 个 technical replicate 组（replicates 2-7 各有 2 例、1 组为 1-8 号外的编号），下游差异分析与建模时需决定是否合并或按重复结构处理，本步骤仅记录不处理。
- 相关性热图采用基因级（各基因减去其样本均值）Pearson 相关，其离对角均值理论上约为 -1/(n-1)，解读时应以聚类块结构而非绝对相关系数大小为准。
- 基因级汇总排除了 1,223 个多基因注释（///）探针与 1,123 个无注释探针（占探针总数 10.5%），这是为规避同一探针信号重复计入多个基因的保守策略，已在 qc_summary.tsv 中记录。
- external_evidence_note 列为文献/领域知识注解（如 ARG1=中性粒细胞标准标志物、LY6E/IFIT1/OAS2/SIGLEC1=经典 ISG），用于可验证性提示；未在本数据集内独立验证，需外部队列/qPCR 确认。
- GSEA 每基因显著集计数依赖 gsea_results.tsv leadingEdge 的 Entrez 符号经 gene_entrez_map.tsv（12,502 基因 1:1）映射；若个别 leadingEdge 含映射外 ID 则不计入，计数为保守下界。
- 复合优先级评分的权重（0.35/0.25/0.15/0.25）为本步骤分析选择；CD74/RUNX3 因显著 ORA/GSEA 成员数很高而居前，但二者不在 step4 最终分类模型中，落地 panel 建议同时参考 evidence_tier、in_final_model 与 step3_compact_panel 标记。
- IRAK3、LAMTOR5、GADD45A 的 DE 单变量方向与联合模型系数方向相反（条件效应），转录本 panel 若直接联用需明确采用多基因联合打分，而非逐基因单方向阈值判读。
- panel 优先级基于本数据集（GSE63990，含 5 组 B/V 技术重复对分组交叉验证）的 OOF 证据链，未做独立外部队列复现；表内已逐基因给出 DE 统计量与模型证据以便复核。
- 环境探测预算用尽导致未能先目视确认 KEGG 快照 TERM2GENE/TERM2NAME 的原始列格式；TERM2NAME 解析采用自适应布局算法（term 内含 hsa+5 位数字，名称取最长非元字段），样本名称已验证正确（Glycolysis/Gluconeogenesis、Citrate cycle 等）。
- 早前超时命令遗留的后台 python 进程仍可能随后覆盖 outputs/ 下同名文件；本步骤在完成时（10:39-10:40）对交付物做了存在性、大小、sha256 与表头内容复核，报告阶段建议以本记录为准或复核 sha256。
- batch1 仅 44 例且错分率更高（27.3% vs 7.5%，χ^2 p=0.0011），batch 与临床分组样本量不完全独立，分类器未做批次校正（仅在训练折内标准化），批次效应与生物学信号的混杂无法在本数据集内完全分离，需外部/独立队列验证。
- KEGG 通路-基因集采用 org.Hs.eg.db PATH 注释（229 条通路、4,524 基因、离线）而非完整 KEGG 快照 TERM2GENE，KEGG ORA 背景覆盖度低于 GO，报告解读以 GO 为主、KEGG 为通路层补充。

后续建议

- qPCR 验证：对候选 panel 基因（细菌侧 ARG1、S100A12、VNN1、IRAK3、HPGD/IL1R1；病毒侧 LY6E、IFIT3、IFIT1、OAS2、SIGLEC1、CD247、BTN3A1/3）在独立样本中进行 qPCR 定量验证，确认芯片表达差异可复现。
- Western blot 或蛋白定量：对高优先级标志物（如 ARG1、S100A12、LY6E、OAS2）进行蛋白水平验证，确认转录本差异转化为蛋白层面的功能差异。
- 独立队列复现：在独立外部队列（不同平台/批次）中复现 10 基因分类模型的 AUC 与混淆矩阵，评估批次效应与平台差异对泛化性能的影响。
- 批次校正策略：在更大样本中评估 ComBat 等批次校正方法对分类性能与差异基因稳健性的影响，明确 batch1 高错分率的成因。
- 多基因联合打分：对 IRAK3、LAMTOR5、GADD45A 等 DE 单变量方向与模型系数方向相反的基因，落地 panel 时采用多基因联合打分而非逐基因单方向阈值判读。
- 扩展验证：纳入宿主人口学、病原亚型与病程信息，对差异基因与 panel 做更细层的生物学或临床分层核验。
- qPCR 验证：对推荐 14 基因精简 PCR 组合（细菌侧 ARG1、S100A12、VNN1、IRAK3、HPGD、IL1R1；病毒侧 LY6E、IFIT3、IFIT1、OAS2、SIGLEC1、CD247、BTN3A1、BTN3A3）在独立样本中进行 qPCR 定量验证，确认表达方向与差异幅度。
- Western blot 或蛋白定量：对 T1 层高优先级标志物（VNN1、ARG1、IRAK3、LY6E、ICAM2、LAMTOR5）进行蛋白水平验证，确认转录组发现向蛋白组的可转化性。
- 独立队列复现：在外部独立队列（不同平台/批次）中复现 10 基因分类模型的 AUC 与错分模式，评估批次效应与真实泛化性能。

本研究完成宿主血液转录组（GSE63990）的细菌 vs 病毒感染系统分析：识别 369 个差异基因（细菌上调 216、病毒上调 153），富集分析揭示细菌侧中性粒细胞/炎症与能量代谢程序、病毒侧 ISG 与 T 细胞程序；10 基因精简分类模型取得 OOF AUC=0.9466、准确率 0.8789；综合多源证据构建 28 基因候选 panel 并推荐 14 基因 PCR 精简组合，为急性呼吸道感染病因鉴别提供了可验证的候选标志物。