描述性统计分析在 RNA-seq 数据中的应用

选题范围：描述性统计分析。

报告中展示了两部分内容，先梳理描述性统计的常用方法，再基于公开 RNA-seq 数据集 GSE193613 完成数据检查、统计描述、图形诊断和生物医学解释。

案例数据用于展示描述性统计在组学数据分析前期的作用；附录中的标准分析流程：差异表达、KEGG 与 GSEA 结果作为机制解释延伸。

目录

1. 描述性统计分析方法

描述性统计分析的目标是用少量指标和图形概括数据的主要特征，包括数据规模、集中位置、离散程度、分布形状和样本间结构。它不以显著性检验为核心，而是为后续建模、参数估计或假设检验提供数据认识。

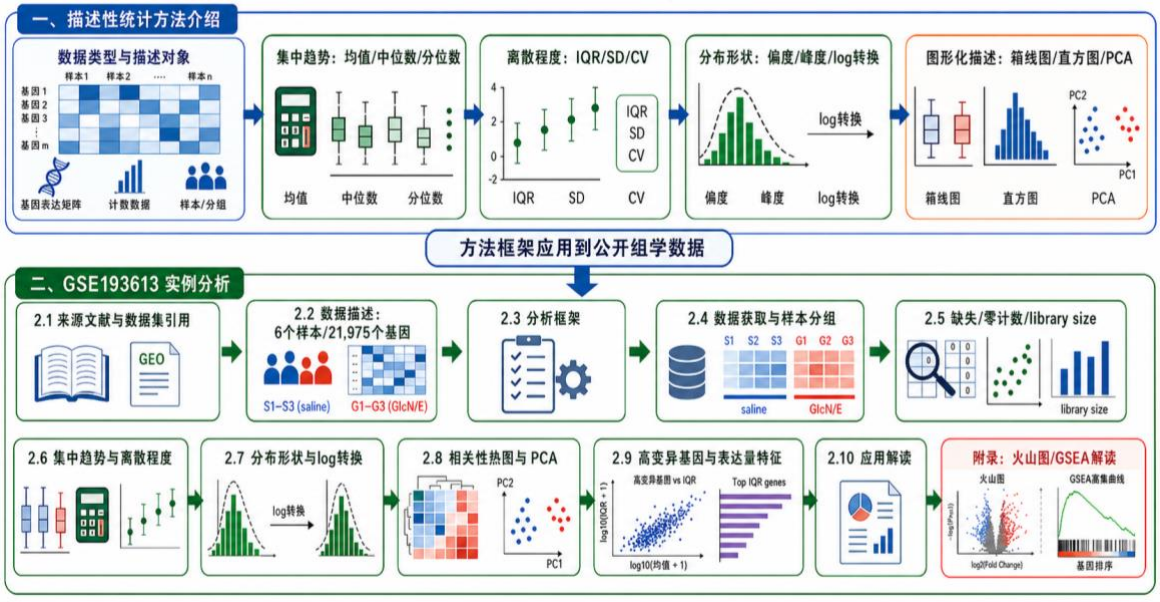

图 1 描述性统计分析与 RNA-seq 案例解释框架

1.1 数据类型与描述对象

生物医学数据通常同时包含分类变量和数值变量。分类变量如处理组别、疾病状态和样本来源，常用频数和比例描述；数值变量如 RNA-seq count、logCPM、临床指标和蛋白表达量，常用均值、中位数、分位数和离散指标描述。

组学数据的特点是维度高、样本数相对少、分布偏态明显，因此描述性统计不能只报告单一均值，还需要结合零值比例、四分位数、变换后分布和可视化结果。

1.2 集中趋势：均值、中位数与分位数

均值反映所有观测值的平均水平，但容易被极端值影响；中位数反映排序后位于中间的水平，对长尾分布更稳健；四分位数进一步描述数据主体落在哪个区间。RNA-seq count 数据中少数高表达基因可能拉高均值，因此中位数和四分位数通常更能代表大多数基因的表达状态。

```
mean_count = mean(count)
median_count = median(count)
q1 = quantile(count, 0.25)
q3 = quantile(count, 0.75)
```

1.3 离散程度：极差、IQR、SD 与 CV

离散程度用于描述数据围绕中心位置的波动幅度。SD 表示观测值相对均值的绝对波动；IQR 表示中间 50% 数据的跨度，对极端值更稳健；CV 为 SD 与均值的比值，适合比较不同均值水平下的相对波动。

在表达矩阵中，IQR 可用于识别跨样本波动较大的基因，CV 可补充观察低表达或相对波动明显的基因，但 CV 在均值很低时容易被放大，需要结合平均表达量解释。

```
iqr = IQR(count)
sd = sd(count)
cv = sd / mean_count
```

1.4 分布形状：偏度、峰度与变换

偏度用于描述分布是否对称，正偏说明右尾较长；峰度用于描述尾部和峰部是否比正态分布更极端。RNA-seq raw count 通常呈强右偏长尾，大量基因低表达或未检出，少数基因 count 极高。

log2(count + 1) 转换常用于压缩极端高值，使主体分布更容易观察。该转换只改变展示尺度，不等同于差异表达检验。

```
log2_count_plus1 = log2(count + 1)
skewness = mean((count - mean_count)^3) / sd^3
```

1.5 图形化描述与多维数据概览

箱线图适合展示中位数、IQR 和极端值；直方图适合展示整体分布形状；相关性热图可观察样本表达轮廓是否相近；PCA 将高维表达矩阵投影到主成分坐标，用于观察样本整体结构和潜在离群点。

对于 RNA-seq 数据，相关性和 PCA 一般建议基于标准化后的 logCPM 或类似尺度，而不是直接使用 raw count。

2. 实际案例分析：GSE193613 RNA-seq count 数据

2.1 来源文献与数据集引用

案例数据来源于 GEO 数据集 GSE193613。该数据集对应 Zhou 等发表在 Clinical and Translational Medicine 的研究，题为“Glucosamine facilitates cardiac ischemic recovery via recruiting Ly6Clow monocytes in a STAT1 and O-GlcNAcylation-dependent fashion”（PMID: 35343077，DOI: 10.1002/ctm2.762）。

GEO 页面记录该数据为 Mus musculus 高通量测序表达谱，整体设计为心梗后第 3 天 ischemic zone 组织中 saline 与 GlcN/E 处理组的比较，补充文件提供 GSE193613_gene_count.txt.gz 作为 processed count 矩阵。

2.2 数据描述

count 矩阵包含 21975 个基因和 6 个样本，其中 saline 对照组 3 个样本（S1、S2、S3），GlcN/E 给药组 3 个样本（G1、G2、G3）。共有 4718 个基因在六个样本中均为 0，17257 个基因至少在一个样本中有表达。

项目	结果
基因总数	21,975
样本总数	6（saline 3 个，GlcN/E 3 个）
全零基因	4,718
至少一个样本检出基因	17,257
library size 范围	28.89M 至 35.51M

2.3 数据读取与样本分组

数据读取后将第一列作为基因 ID，其余列转为数值矩阵。样本名按 S 或 G 前缀归入 saline 或 GlcN/E 组。该步骤保证后续统计量的行列含义清晰。

```
counts_dt <- data.table::fread(count_file)
count_mat <- as.matrix(counts_dt[, -1])
storage.mode(count_mat) <- "numeric"
group = if_else(grepl("^S", sample_names), "saline", "GlcN/E")
```

2.4 数据检查：缺失、零计数与 library size

六个样本缺失值均为 0，逐样本零计数比例集中在 27.90% 至 28.60%。library size 位于 28.89M 至 35.51M，说明样本测序深度处于同一数量级，但仍需避免直接用 raw count 进行样本间强比较。

```
library_size = sum(count)
zero_genes = sum(count == 0)
zero_fraction = mean(count == 0)
detected_genes = sum(count > 0)
```

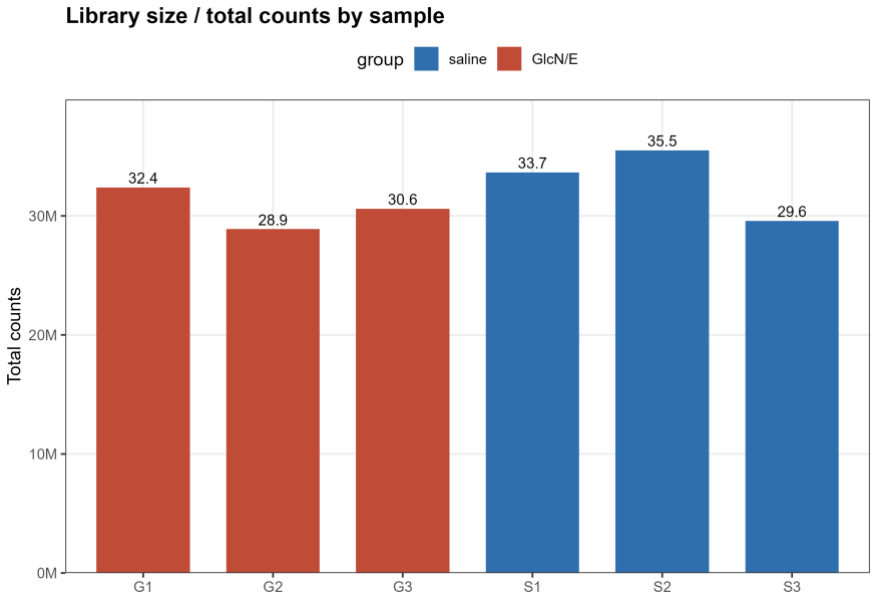

图 2 每个样本 library size / total counts 比较

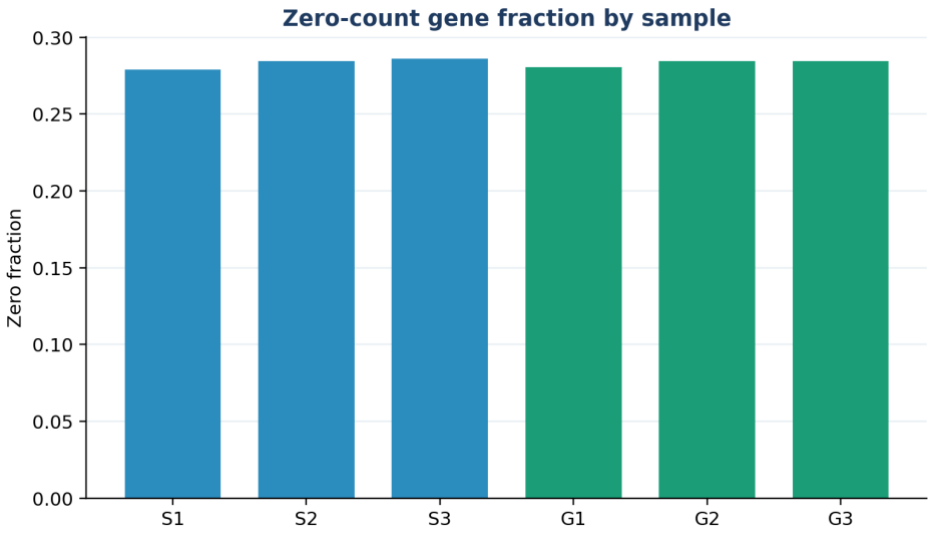

图 3 每个样本零计数基因比例

2.5 集中趋势与离散程度

全矩阵 raw count 均值为 1,445.77，样本均值范围为 1,314.61–1,615.95，全矩阵中位数为 196，样本中位数范围为 181–214，saline 组样本中位数的组内中位值为 212，GlcN/E 组为 189，样本均值均明显高于中位数，均值/中位数比值约为 7.05 至 7.62，全矩阵 min=0，max=916,158，极差为 916,158；样本极差范围为 503,82–916,158，提示 raw count 受极端高表达基因影响，说明少数高表达基因拉高了均值。全矩阵众数为 0，出现 37,331 次，占全部观测值的 28.31%。0 是最常见取值，说明未检出或低表达基因占有较大比例。样本 SD 范围为 6,821.84–9,843.64。绝对离散程度较高，符合 RNA-seq count 长尾特征。样本 CV 范围为 4.52–6.68，相对波动明显。两组中位 IQR 分别为 1251 和 1161。

```
mean_count = mean(count)
median_count = median(count)
iqr = IQR(count)
sd = sd(count)
cv = sd / mean_count
```

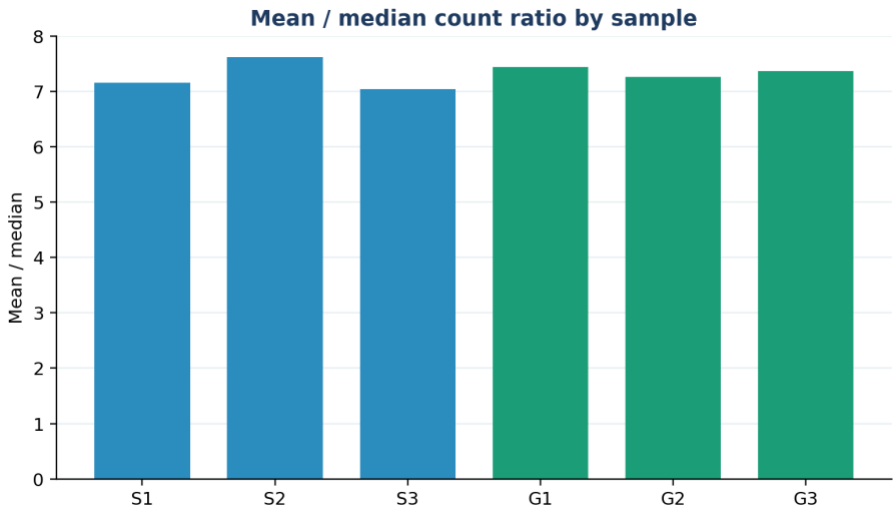

图 4 每个样本 mean / median count 比值

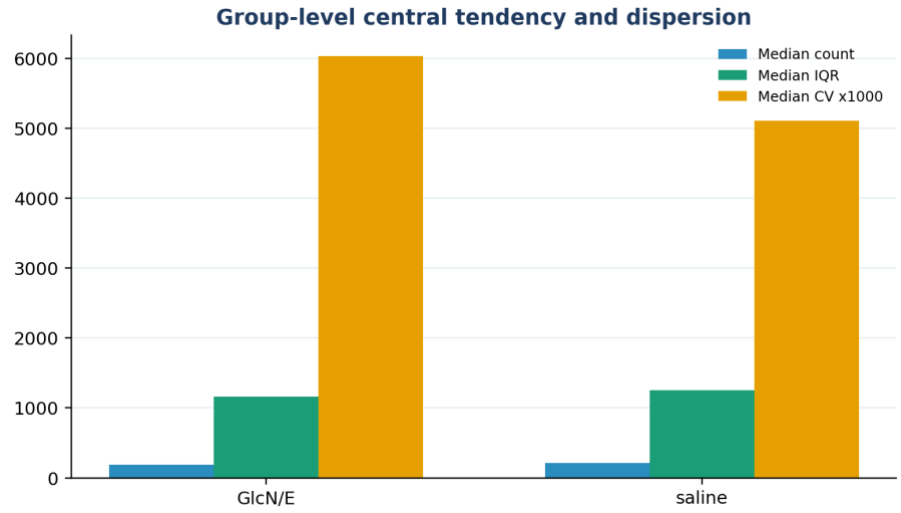

图 5 组层面集中趋势与离散程度汇总

2.6 分布形状与 log 转换

全矩阵 raw count 中位数为 196，P99 为 16957.22，最大值达到 916158，偏度为 43.49，说明原始计数强烈右偏。log2(count + 1) 后 P99 降至 14.05，最大值为 19.81，偏度接近 0（-0.179），可视化中主体分布更清楚，log 转换后分布更接近对称。raw count 超额峰度为 3,321.49；log2(count+1) 后超额峰度为 -1.48。raw count 厚尾和极端值非常明显，log 转换压缩尾部。

```
raw_long <- raw_long %>%
  mutate(log2_count_plus1 = log2(count + 1))
skewness = mean((count - mean_count)^3) / sd^3
```

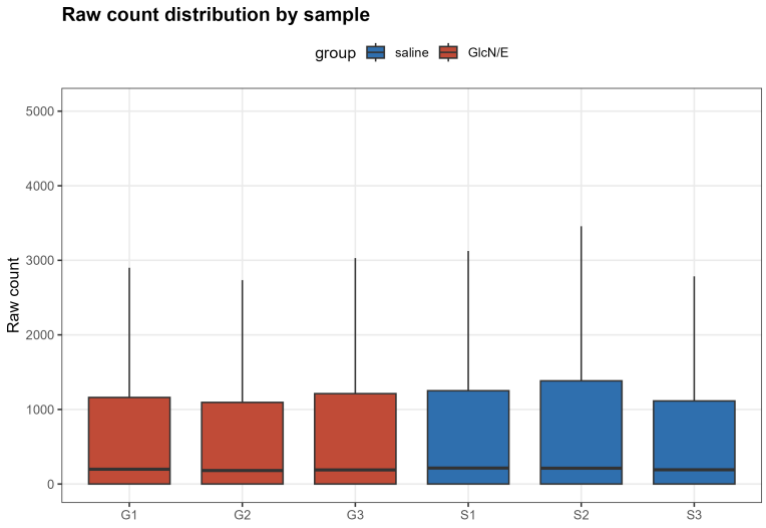

图 6 原始 count 箱线图

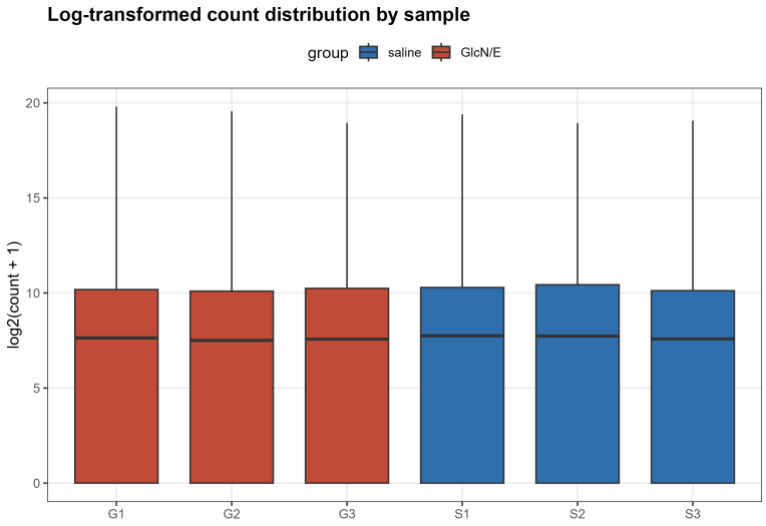

图 7 log2(count + 1) 后的样本箱线图

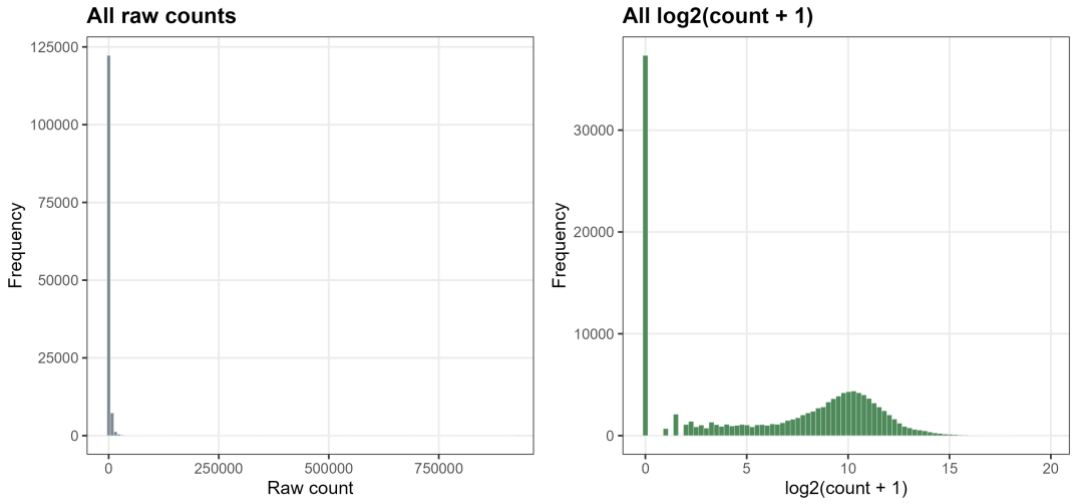

图 8 原始 count 与 log2(count + 1) 直方图对比

2.7 样本相关性与 PCA 概览

相关性热图和 PCA 均基于 edgeR TMM 标准化后的 logCPM。TMM 用于校正样本间测序深度和组成差异，logCPM 则使高维表达矩阵更适合进行样本相似性和主成分观察。

```
dge <- edgeR::DGEList(counts = count_mat, group = group_factor)
dge <- edgeR::calcNormFactors(dge, method = "TMM")
```

```
log_cpm <- edgeR::cpm(dge, log = TRUE, prior.count = 1)
cor_mat <- cor(log_cpm, method = "pearson")
pca <- prcomp(t(log_cpm), center = TRUE, scale. = FALSE)
```

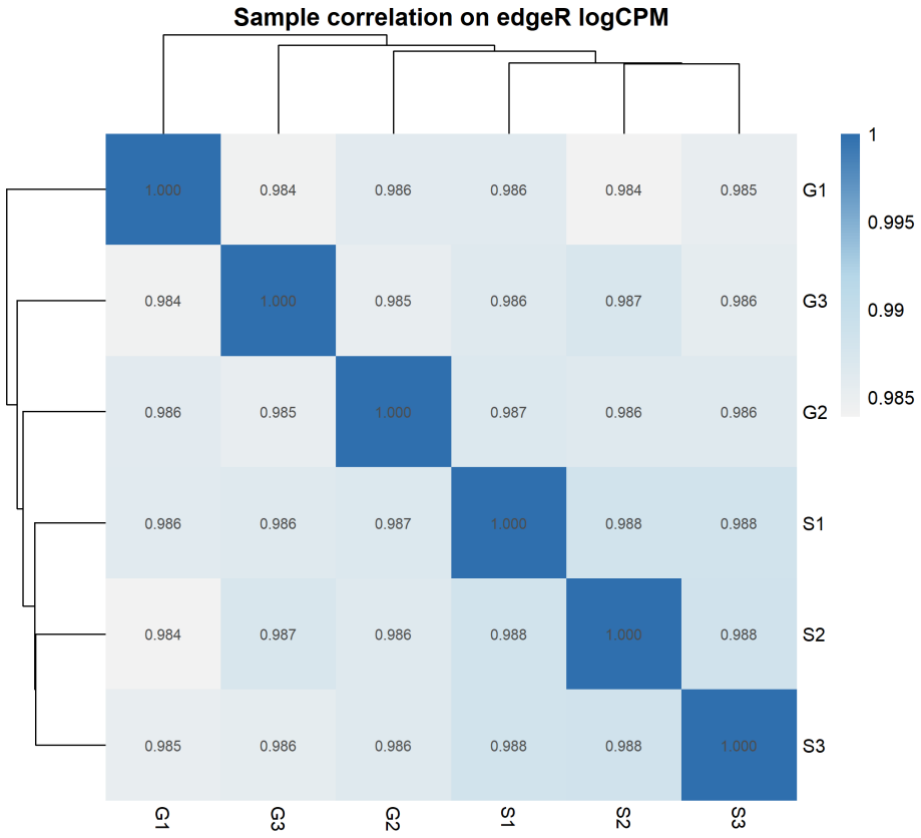

图 9 基于 edgeR logCPM 的样本 Pearson 相关性热图

基于 edgeR logCPM 的样本两两 Pearson r 范围为 0.9839–0.9884。样本相关性整体较高，说明同一组织来源的全局表达轮廓接近，适合进一步做 PCA 结构概览。PC1 解释 25.7% 变异，PC2 解释 21.6% 变异，PC1 + PC2 合计 47.3%。PCA 用于观察整体结构和离群情况，不作为显著差异证据。

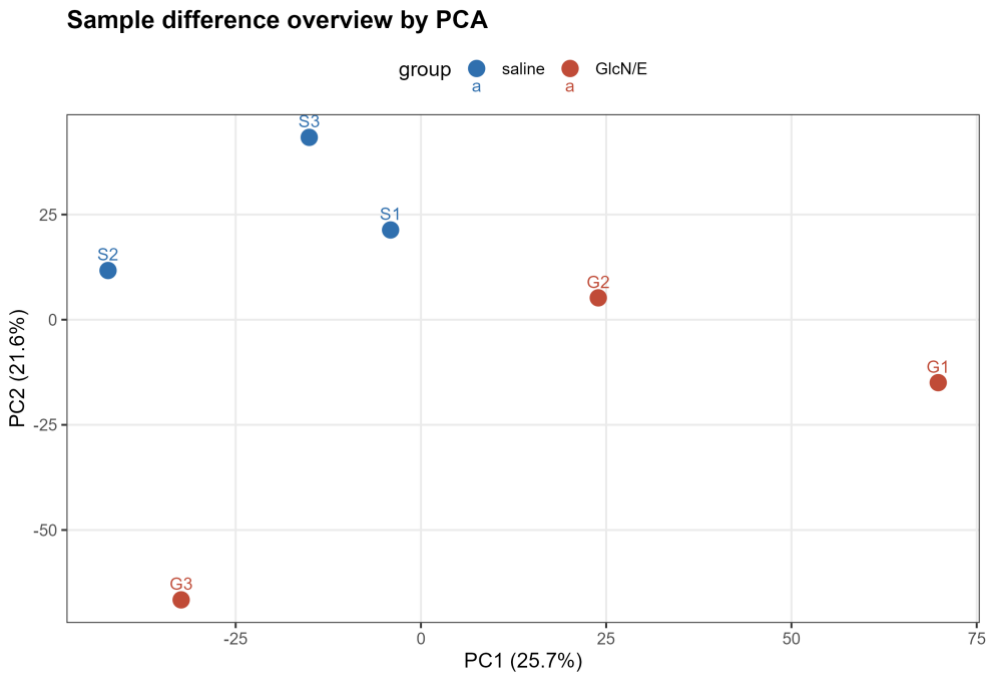

图 10 基于 edgeR logCPM 的 PCA 样本差异概览

2.8 基因层面表达异质性

基因层面按 IQR 对表达波动排序，最高的是 mt-Co1，IQR 为 237564.75。Top 基因包括 mt-Co1、mt-Cytb、Myh6、Eef1a1、Actb 和 Ftl1，提示高变异基因主要集中于线粒体能量代谢、心肌结构和蛋白合成相关方向。

```
gene_stats <- tibble(
  mean_all = matrixStats::rowMeans2(count_mat),
  iqr_all = q3_all - q1_all,
  sd_all = matrixStats::rowSds(count_mat),
  cv_all = if_else(mean_all > 0, sd_all / mean_all, NA_real_)
)
top_high_var_iqr <- gene_stats %>% arrange(desc(iqr_all)) %>% slice_head(n = 30)
```

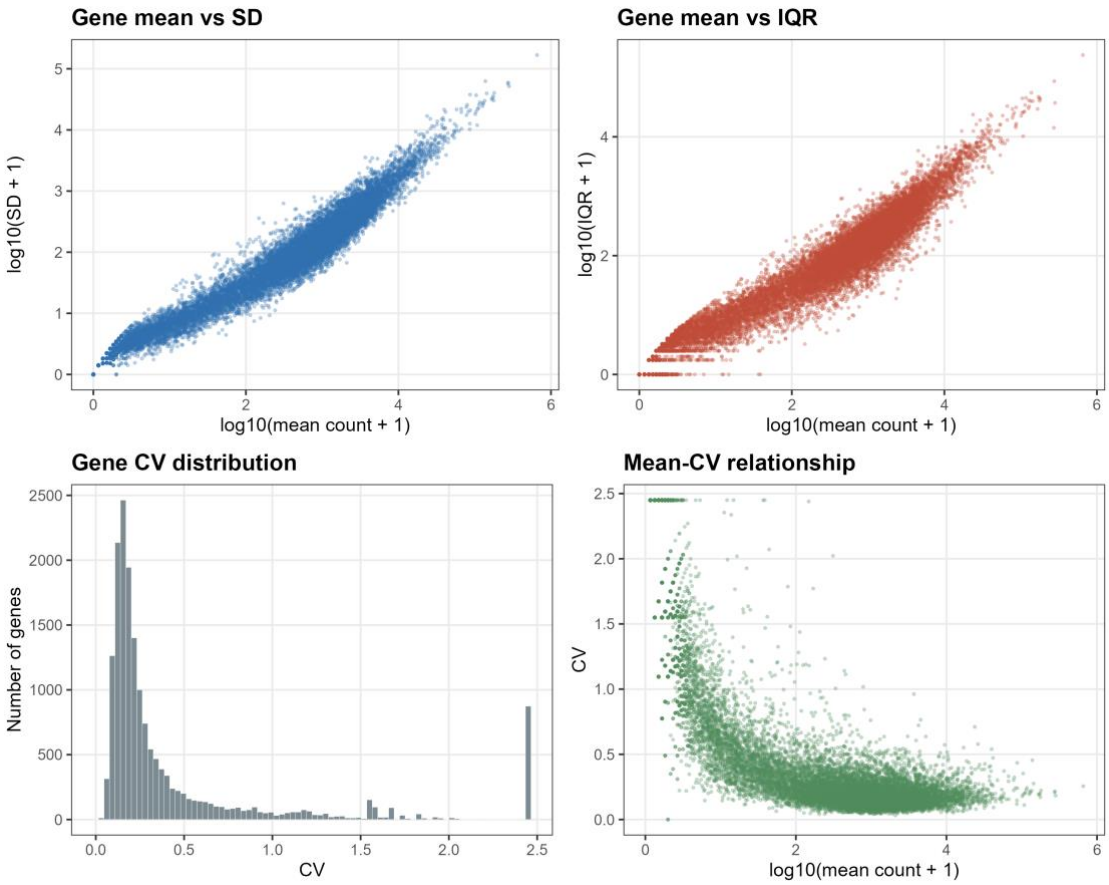

图 11 基因均值与 SD、IQR、CV 的关系

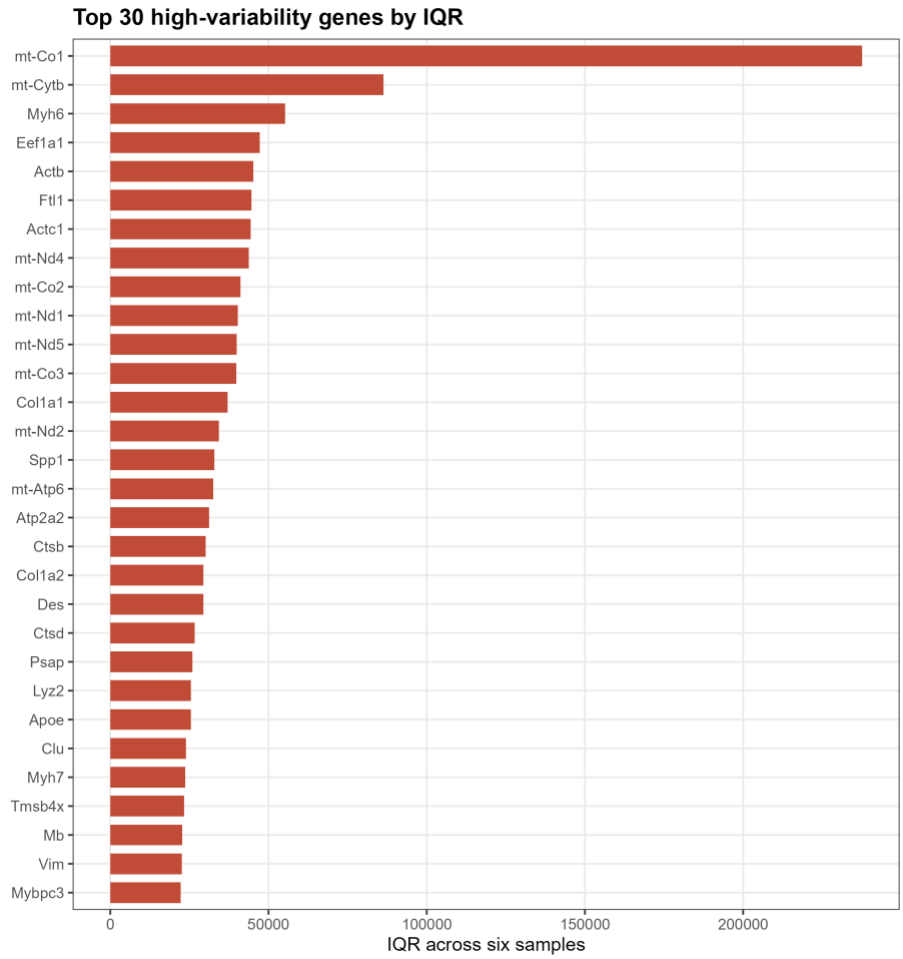

图 12 按 IQR 排序的高变异基因 Top 30

描述性组均值差距中，Prn 在 GlcN/E 组均值较高，log2 组差约为 7.20。该列表只用于形成后续差异分析候选，不报告 p 值或 FDR。

2.9 应用解释

描述性统计结果表明，该 count 矩阵无缺失值、零计数比例接近、library size 处于同一数量级，适合开展后续标准化和探索性可视化。raw count 具有典型 RNA-seq 右偏长尾特征，因此用中位数、IQR、log 转换和 logCPM 视角比单纯均值更稳妥。

从生物医学角度看，高变异基因中出现线粒体编码基因和心肌结构相关基因，与心梗后组织损伤、能量代谢重塑和炎症微环境变化相符。描述性统计不能替代 DESeq2，但可以为差异表达和通路解释提供数据质量背景与候选方向。

3. 附录：DESeq2、KEGG 与 GSEA 结果解读

附录图展示同一数据集的标准 RNA-seq 下游推断分析。图 A 显示 21975 个转录本经筛选后，13293 个进入分析，并由 DESeq2 得到 321 个差异转录本；图 B 的热图显示差异转录本在 saline 与 GlcN/E 样本间形成方向性表达模式；图 C 火山图显示 saline 高表达 147 个、GlcN/E 高表达 174 个。

KEGG pathway annotation 提示 immune system、infectious disease、cardiovascular disease 等类别可作为解释方向。结合原研究主题，GlcN/E 处理可能通过调节免疫细胞募集、炎症反应和心肌修复相关过程影响心梗后恢复。

GSEA 中 IL-17 signaling pathway 的 NES 为 -1.688，P = 1.484e-03，按图中色带方向更偏向 saline 端，提示 GlcN/E 处理后 IL-17 相关炎症信号可能相对降低。ECM-receptor interaction 的 NES 为 1.606，P = 5.401e-05，偏向 GlcN/E 端，提示细胞外基质、细胞黏附和组织重塑相关基因集在处理组更活跃。Dilated cardiomyopathy 的 NES 为 2.055，P = 3.925e-06，也偏向 GlcN/E 端，提示心肌结构和收缩功能相关表达程序发生变化。

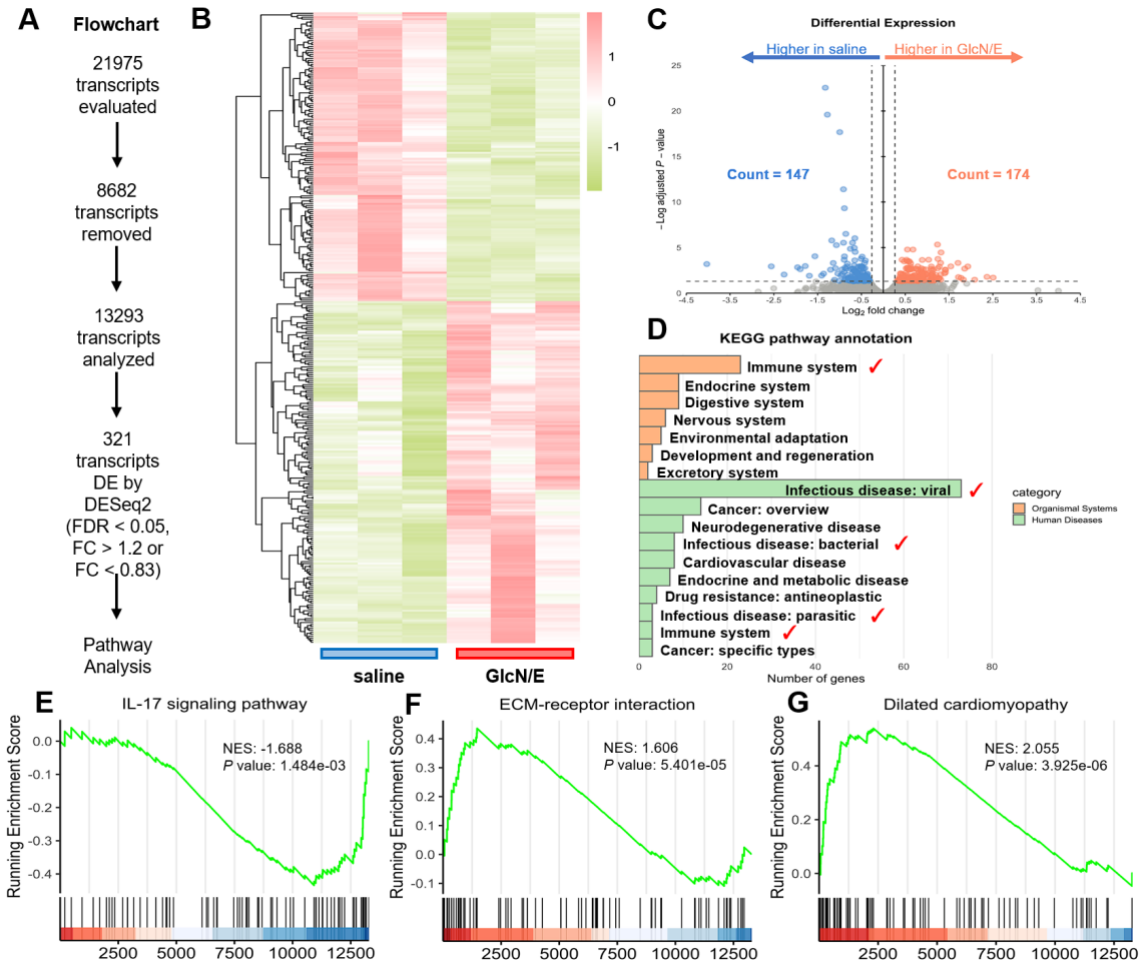
A
Flowchart
21975 transcripts evaluated
8682 transcripts removed
13293 transcripts analyzed
321 transcripts DE by DESeq2 (FDR < 0.05, FC > 1.2 or FC < 0.83)
Pathway Analysis
B
saline
GlcN/E
C
Differential Expression
Higher in saline
Higher in GlcN/E
Count = 147
Count = 174
D
KEGG pathway annotation
Immune system
Endocrine system
Digestive system
Nervous system
Environmental adaptation
Development and regeneration
Excretory system
Infectious disease: viral
Cancer: overview
Neurodegenerative disease
Infectious disease: bacterial
Cardiovascular disease
Endocrine and metabolic disease
Drug resistance: antineoplastic
Infectious disease: parasitic
Immune system
Cancer: specific types
category
Organismal Systems
Human Diseases
Number of genes
E
IL-17 signaling pathway
NES: -1.688
P value: 1.484e-03
F
ECM-receptor interaction
NES: 1.606
P value: 5.401e-05
G
Dilated cardiomyopathy
NES: 2.055
P value: 3.925e-06
Running Enrichment Score

图 13 标准分析大图

参考文献

1. Zhou W, Jiang X, Tang Q, Ding L, Xiao W, Li J, Wu Y, Ruan H-B, Shen Z, Chen W. Glucosamine facilitates cardiac ischemic recovery via recruiting Ly6Clow monocytes in a STAT1 and O-GlcNAcylation-dependent fashion. Clinical and Translational Medicine. 2022;12(3):e762. doi:10.1002/ctm2.762. PMID:35343077.

2. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-140.