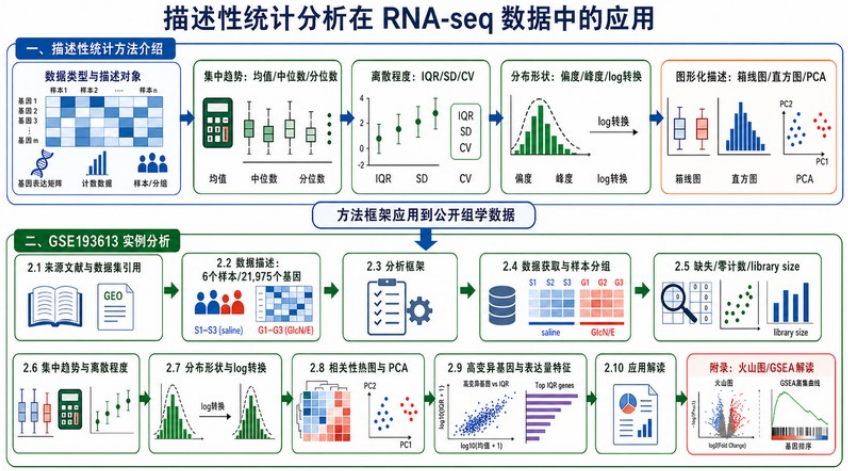
描述性统计分析在 RNA-seq 数据中的应用

方法梳理与 GSE193613 count 数据案例

用规模、中心、离散、分布和样本结构先建立数据背景，再进入差异分析解释。



案例数据：规模、分组与测序深度

count 矩阵包含 21,975 个基因和 6 个样本

saline 对照组：S1、S2、S3

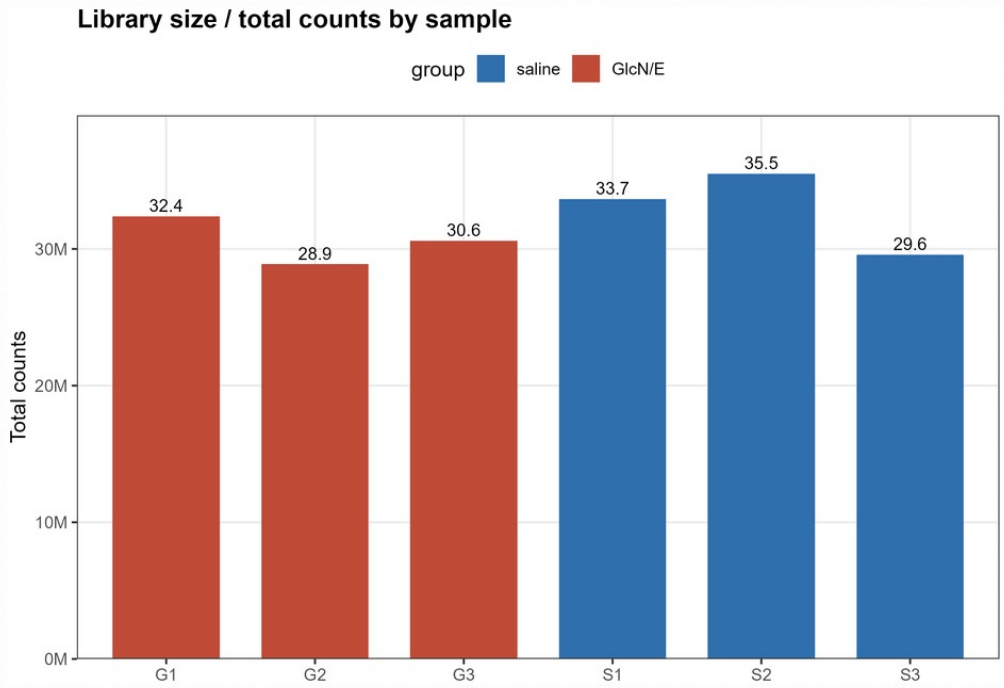
GlcN/E 给药组：G1、G2、G3

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

这些数值作为数据检查结果原样保留；raw count 不直接用于样本间强比较。



集中趋势与离散程度：均值被长尾拉高

全矩阵 raw count 均值 1,445.77，中位数 196

样本均值范围 1,314.61–1,615.95

样本中位数范围 181–214

均值/中位数比值约 7.05 至 7.62

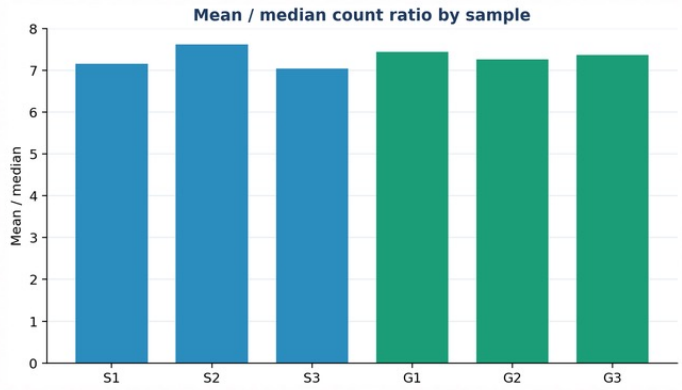
min = 0，max = 916,158，极差 916,158

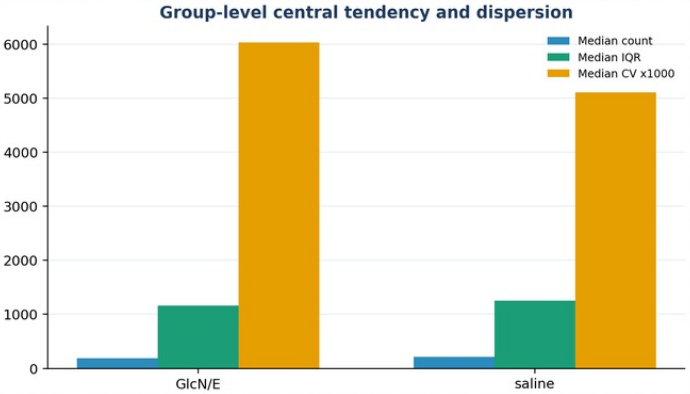
众数 0 出现 37,331 次，占全部观测值 28.31%

样本 SD 范围 6,821.84–9,843.64

CV 范围 4.52–6.68

均值不是错，但在 RNA-seq raw count 中必须和中位数、IQR、SD/CV 一起读。





分布与样本结构：
log 转换让主体更清楚

raw count：

P99 = 16,957.22，max = 916,158，偏度 43.49

log2(count + 1)：

P99 = 14.05，max = 19.81，偏度 -0.179

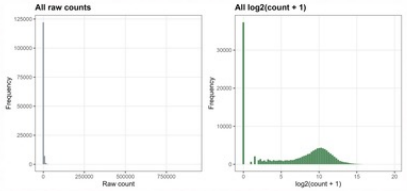
raw count 超额峰度 3,321.49；log2 后超额峰度 -1.48

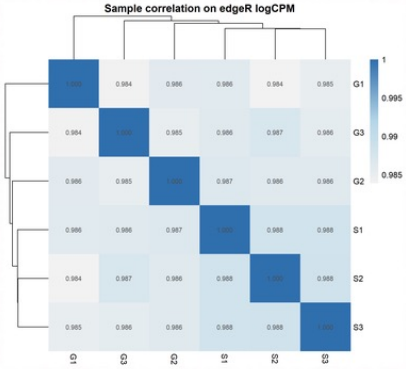
edgeR TMM 标准化后 logCPM：

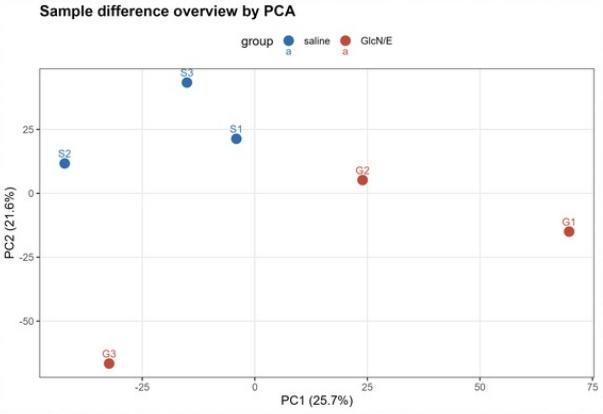
Pearson r 范围 0.9839–0.9884

PCA：PC1 解释 25.7%，PC2 解释 21.6%，合计 47.3%

PCA 用于结构概览和离群观察，不作为显著差异证据。







结论：描述性统计提供质量背景

无缺失值；零计数比例接近；library size 处于同一数量级

raw count 右偏长尾明显，

因此中位数、IQR、log 转换和 logCPM 更稳妥

IQR 排序最高基因：mt-Co1，IQR = 237,564.75

21,975 个转录本经筛选后 13,293 个进入分析；

DESeq2 得到 321 个差异转录本

saline 高表达 147 个，GlcN/E 高表达 174 个

GSEA：

IL-17 signaling pathway NES = -1.688，P = 1.484e-03

ECM-receptor interaction NES = 1.606，P = 5.401e-05

Dilated cardiomyopathy NES = 2.055，P = 3.925e-06