H1

简述基因芯片是如何测量基因表达量的?

基因探针（probe）针对不同基因与 probe 序列互补的 mRNA (cDNA)会被捕获并杂交于探针。表达量高的基因对应的 probe 会更亮。Probe 上的荧光强度代表基因表达量

简述基因芯片 data normalization 的必要性，并简述 quantile normalization 的过程及目的。

来自不同芯片的数据受到多种技术噪音的干扰，致使数据分布不同，没有直接可比性。quantile normalization 强制不同芯片的数据服从同一个分布，但是基因表达的排序与原来一致。

简述差异基因表达的目的，并简述两组差异表达分析的过程。

目的：找出在不同条件下（例如：实验组和对照组）表达差异显著的基因。分析过程：假定两组数据已经经过正确前期处理，两组数据均为矩阵，其中每一行代表一个基因，每一列代表一个样本；可以通过一定的计算方法（例如 t-test, ANOVA）算出同一个基因的不同分值，通过对比同一样本的分值，可实现差异表达分析。

什么是多重比较（multiple test)？为什么它在高通量数据分析中非常重要

多重比较是指在进行多次实验的时候，每次实验的假阳性结果的累积，导致整体假阳性比率增加。高通量数据中通常涉及很多变量或不同的实验条件，如果不控制多重比较，假阳性结果会大量增加，从而误导研究结论。为了减少多重比较带来的影响，可以通过一些方法来进行校正，例如：Bonferroni 校正，FDR 控制。

什么是批次效应（batch effect)？它在高通量数据分析中会带来什么问题?

批次效应指的是在不同的实验条件下（例如，不同的生产实验室，不同的日期等）生成的数据可能带来的误差或者其他变量，例如：相对不同生产批次的实验数据来说，同一批生产的实验数据可能更接近。批次效应可能在数据分析中带来潜在误差，使得数据本身由批次效应带来的误差远大于生物特殊性导致的误差，从而增大数据分析的难度和准确性。

H2

可以通过微阵列或 RNA 测序来测量基因表达。简要描述每种技术是如何测量基因表达的（10 分），讨论它们的差异并简要描述在差异表达测试程序中它们是如何建模的?（10 分）

微阵列：通过将荧光标记的 cDNA 与芯片上固定的 DNA 探针网格杂交。荧光强度表示表达水平。

RNA - Seq：使用下一代测序直接对 RNA 衍生的 cDNA 进行测序。基于映射到每个基因的读数数量来量化基因表达。

差异：微阵列受探针特异性和预定义转录本的限制。RNA - Seq 更灵敏，可检测新的转录本，并且具有更大的动态范围。 在差异表达测试中的建模方式：微阵列：通常使用线性模型（例如 LIMMA）对对数转换后的荧光强度进行建模。RNA - Seq：依赖于基于计数的模型（例如 DESeq2 或 edgeR 中的负二项分布）来解释测序数据中的过度离散。

为什么必须对微阵列或 RNA - seq 数据进行归一化?（20 分）

归一化可校正技术偏差并确保不同样本间表达水平的可比性。（在微阵列中：调整杂交效率或背景信号的差异。在 RNA - Seq 中：考虑文库大小和测序深度的差异。）

简要描述分位数归一化。（10 分）

分位数归一化强制使不同样本间的强度或计数分布相同。它对每个样本中的值进行排序，计算样本间的平均秩，并将平均秩分配给原始值。这减少了由技术噪声引起的变异性。

根据所附的 r 文件，回答注释部分中的问题。（共 50 分）

```r
# 加载包
library(oligo)
library(siggenes)
library(limma)
library(DESeq2)
library(edgeR)
library(ROCR)
library(EnhancedVolcano)

# 第一部分：微阵列分析
CELfiles = dir(pattern="CEL.gz$")

rawdata.u133a = read.celfiles(filenames= CELfiles[1:6])

dat = exprs(rawdata.u133a)

colnames(dat) = c(paste("Ref",1:3), paste("Brain", 1:3))

head(dat)
dim(dat)

normdata.u133a = rma(rawdata.u133a) # 请解释这一行所进行的操作，以及为什么需要这一步操作(10')
dat.u133a.norm = exprs(normdata.u133a)

boxplot(dat.u133a.norm)

design = cbind(mu=1,beta=c(rep(0,3),rep(1,3)))  # 请回答这一步design变量的维度（[行数 列数]），以及每一维度的含义(10')

DE.limma.u133a = lmFit(dat.u133a.norm, design=design)

DE.limma.u133a = eBayes(DE.limma.u133a)  # 请解释这一行所进行的操作，以及为什么需要这一步操作(10')

pval.u133a = DE.limma.u133a$p.value[,"beta"]

head(pval.u133a)

hist(pval.u133a)

result = data.frame(log2FoldChange = DE.limma.u133a$coefficients[,"beta"],
                    pvalue = DE.limma.u133a$p.value[,"beta"])

EnhancedVolcano(result,
                x = "log2FoldChange",
                y = "pvalue",
                lab = rownames(result))
```

```r
# 第二部分：RNA-seq差异表达分析：RNA-seq数据的差异表达分析（DESeq2和edgeR）
load("RNAseq.rda")
nreps = ncol(X)/2

# DESeq2:
condition = factor(c(rep(0,nreps),rep(1,nreps)))  # 定义实验组和对照组
dds = DESeqDataSetFromMatrix(X, DataFrame(condition), ~ condition)
dds = DESeq(dds)
result.DESeq = results(dds)

# 绘制火山图：请简要解释火山图展示的含义(10')
EnhancedVolcano(result.DESeq,
                x = "log2FoldChange",
                y = "pvalue",
                lab = 1:nrow(result.DESeq))
pval.DEseq = result.DESeq$pvalue

# edgeR:
d = DGEList(counts=X, group=condition)
d = calcNormFactors(d)
d = estimateCommonDisp(d)
d = estimateTagwiseDisp(d)
fit.edgeR = exactTest(d)

pval.edgeR = fit.edgeR$table$PValue

plot(pval.DEseq, pval.edgeR, pch=".", xlab="DEseq", ylab="edgeR")

ROC.DE <- function(DE.gs, pval) {
    pred = prediction(1-pval, DE.gs)
    perf = performance(pred,"tpr","fpr")
    perf
}
roc.DEseq = ROC.DE(flag, pval.DEseq)
roc.edgeR = ROC.DE(flag, pval.edgeR)

xlim = c(0,1)
plot(roc.DEseq, xlim=xlim)
plot(roc.edgeR, add=TRUE, col="red")
legend("bottomright", legend=c("DEseq", "edgeR"),col=c("black","red"), lty=1)
# 绘制ROC曲线，请通过ROC曲线简要评估和比较两种方法的性能(10')

```

1. RMA 归一化（rma）执行稳健多阵列平均归一化。

· 操作内容：背景校正、分位数归一化以及将探针强度汇总为表达值。

· 原因：确保不同阵列间数据的可比性并减少技术噪声。

2. design 矩阵维度为 [6×2]。

· 行数：6 个样本（3 个参考样本和 3 个脑样本）。

· 列数：mu：总体均值（截距）；beta：组效应（参考样本为 0，脑样本为 1）。

3. eBayes 函数计算调整后的 t 统计量和调整后的 p 值。

· 操作内容：使用经验贝叶斯方法缩小基因间的方差。

· 原因：稳定方差估计，特别是在样本量较小时，从而得到更可靠的结果。

4. 火山图显示效应大小（log2 倍变化）和统计显著性（p 值）之间的关系。

· x 轴：基因表达的 log2 倍变化。

· y 轴：负 log10 p 值。

· 目的：识别具有大变化和高显著性的基因。

5. ROC 曲线评估 DESeq2 和 edgeR 方法的性能。

· 真阳性率（TPR）：灵敏度（y 轴）。

· 假阳性率（FPR）：1 - 特异性（x 轴）。

· 比较：曲线越接近左上角，灵敏度和特异性越高。重叠的曲线表示性能相似。在这种情况下，两种方法性能相似。

H3

bulk 和单细胞测序数据的主要区别是什么？单细胞数据可以提供哪些额外信息?

Bulk 数据测量许多（数百万个）细胞的平均信号，而单细胞数据为每个单个细胞提供信号。单细胞数据具有更高的分辨率。它提供了细胞类型数量、细胞类型组成、细胞类型特异性信号等信息。

对于 scRNA - seq 技术，基于液滴的方法相对于基于平板的方法有什么优势?

基于液滴的方法提供更高的通量，这意味着它可以在一次运行中对更多的细胞进行测序。它有利于研究细胞亚群。

简要描述单细胞 RNA - seq（scRNA - seq）细胞聚类的过程。解释每个步骤的目的。

数据质量控制：去除质量差的细胞和基因。数据归一化：去除数据中的生物和技术伪影。特征选择：选择包含细胞类型信息的基因，提高数据的信噪比。降维：进一步增强数据中的信号。降低数据维度以提高计算效率。无监督聚类：使用现有方法根据所选特征、降维后的表达对细胞进行聚类。

细胞聚类和伪时间构建之间的异同点是什么?

细胞聚类假设细胞来自独立且可交换的组。这些组没有顺序。伪时间构建假设细胞组遵循时间连续体，代表某种生物或临床进展。

有监督细胞类型识别方法的优缺点是什么?

优点：更准确、更快。缺点：依赖参考数据；无法识别新的细胞类型。

tSNE 和 UMAP 的目的是什么?

它们是降维技术，用于在低维空间中可视化高维数据（如 scRNA - seq）。

根据实验室获得的结果，写一份简短的报告来呈现 scRNA - seq 分析结果。解释数据分析的步骤并讨论图表的含义。

H4

给出 DNA 甲基化的几个生物学功能示例。

DNA 甲基化在基因调控中起重要作用：例如，启动子区域的甲基化可以抑制基因表达。DNA 甲基化在发育中起着关键作用：例如，甲基化在细胞分裂过程中是可遗传的，并且可以帮助细胞在细胞 / 组织分化过程中建立身份（ppt 第 5 页）。

简要描述为什么 BS - seq 的序列比对更复杂。

由于亚硫酸氢盐转化，未甲基化的 C 将转化为 T。这会导致序列读取中的错配，因此这些读取不能直接与参考基因组比对。因此，BS - seq 数据的序列比对必须考虑到这一点进行设计。

简要描述差异甲基化分析的目标。

比较两组之间的甲基化水平，以发现甲基化的潜在影响（例如，在肿瘤组织和正常组织之间）。

简要描述 BS - seq 的数据。

在每个位置，我们有读取的总数，以及在这些总样本中甲基化读取的数量，然后我们可以获得特定位置的甲基化水平（ppt 第 16、17 页）。

在课堂上给出的差异甲基化示例中，为什么 Fisher 精确检验和 t 检验给出截然不同的结果?

Fisher 精确检验和 t 检验的观察单位不同：Fisher 精确检验的单位是读取，而 t 检验的单位是样本。

为什么对于生物变异性，离散度是比方差更好的度量?

方差不能独立描述数据中的个体差异，因为方差与均值往往相关，不能独立描述个体差异；随着散度的增加，随机变量的取值越来越分散；因此散度能更好地反映不同均值水平下的差异。

简要描述用于 BS - seq 数据的 β - 二项式模型。你需要提供模型，并解释它们如何捕获数据中的不同类型的变异。

给定 N（总数），M（甲基化读取数）遵循二项分布（M Bin (N, π)），其中 π 是真实甲基化水平，由于 π 可能因 DNA 的随机抽样而变化，β 分布用于模拟 π 的变化（即：π Beta (μ, ϕ)，由离散度参数 ϕ 建模），然后可以在不同重复中测试 μ。

H5

ST 技术的两种主要类型是什么？它们的优缺点是什么？（10 分）

基于图像的 ST 和基于测序的 ST。基于图像的 ST 优点：测序准确性高。基于图像的 ST 缺点：分析大量转录本可能导致光学拥挤，无法测量所有基因。基于测序的 ST 优点：能够在全基因组水平测量基因表达。基于测序的 ST 缺点：缺乏单细胞分辨率。

图是指由节点和边组成的结构。如何从 ST 数据构建图？具体来说，这个图中的节点是什么，我们如何确定节点之间是否有边？（10 分）

根据细胞的空间坐标构建图。节点：细胞；边：如果两个细胞在物理上彼此接近，则它们之间有边。

简要解释空间可变基因（SVGs）是什么，并描述 SVGs 和高变基因（HVGs）之间的区别。（10 分）

SVGs：在不同空间位置表现出显著表达水平变化的基因。HVGs 是在数据集中的细胞或样本间表现出显著表达变异性的基因。然而，如果 HVGs 的整体变异性较低，它们可能会遗漏具有重要空间模式的基因。SVGs 的表达水平与空间位置相关，这可能是由于细胞类型定位、细胞类型内的变异或独立于细胞类型的变异。

为什么整合基于测序的 ST 数据、基于成像的 ST 数据和 scRNA - seq 数据很重要？整合后，每种类型的数据可以获得哪些额外信息？（20 分）

基于成像的 ST 技术在基因数量上有限；基于测序的 ST 在空间分辨率上较低，无法达到单细胞水平；scRNA - seq 可以达到单细胞水平但缺乏空间信息。 整合所有这些的目的是为基于成像的数据进行基因插补，为基于测序的数据实现空间解卷积，为 scRNA - seq 数据实现空间重建。

简要描述在 SpatialDE 中如何利用高斯过程识别 SVGs。包括模型并解释它们如何捕获数据中的不同类型的变异。（20 分）

在 SpatialDE 中利用高斯过程时，均值是基因的平均表达水平；协方差矩阵对于给定基因有空间（Σ）和非空间（δ）成分，其在空间坐标 $X = \left(x_1, ..., x_{\{n\}}\right)$ 上的表达谱 $X = \left(y_1, ..., y_{\{n\}}\right)$ 遵循多元正态分布：

$$P\left(y|\mu, \sigma, \delta, \sum\right) = N\left(y|\mu \cdot 1, \sigma_s^2\left(\sum + \delta \cdot I\right)\right)$$

对于细胞 i 和细胞 j，空间成分（Σ）使用高斯核：

$$\sum_{\text{ij}} = k\left(x_i, x_j\right) = \exp\left(-\frac{|x_i - x_j|^2}{2 \cdot l^2}\right)$$

为了捕获数据中的不同类型的变异：通过模型比较评估空间方差成分的显著性。替代模型和零模型是嵌套的。