转录组数据分析

2025/6/11

汇报人：陈纹娜

第六小组：贺连琦 朱姿霖 李姝婵 陈纹娜

概述

北京大学 PEKING UNIVERSITY

- RNA 是遗传信息从 DNA 到蛋白质流动的中间阶段，包含许多可以通过 RNA 测序进行研究的表达和调控系统。
- 全基因组测序揭示了 DNA 水平的信息，而下一代 RNA 测序 （RNA-seq） 则表征了转录组，并通过定性和定量表达谱为基因组提供了额外的背景信息。
- 与基因组的固定状态相反，转录组在内容和数量的可变性为研究反应机制和控制系统提供了绝佳的机会。出于这个原因，许多 RNA-seq 项目比较了生物体在不同发育阶段、各种实验条件和多个时间点的转录本表达。
- 了解样本和条件之间的转录组水平变化可以解释基因组的功能区域，并进一步推断所得蛋白质如何发挥作用以支持生命。

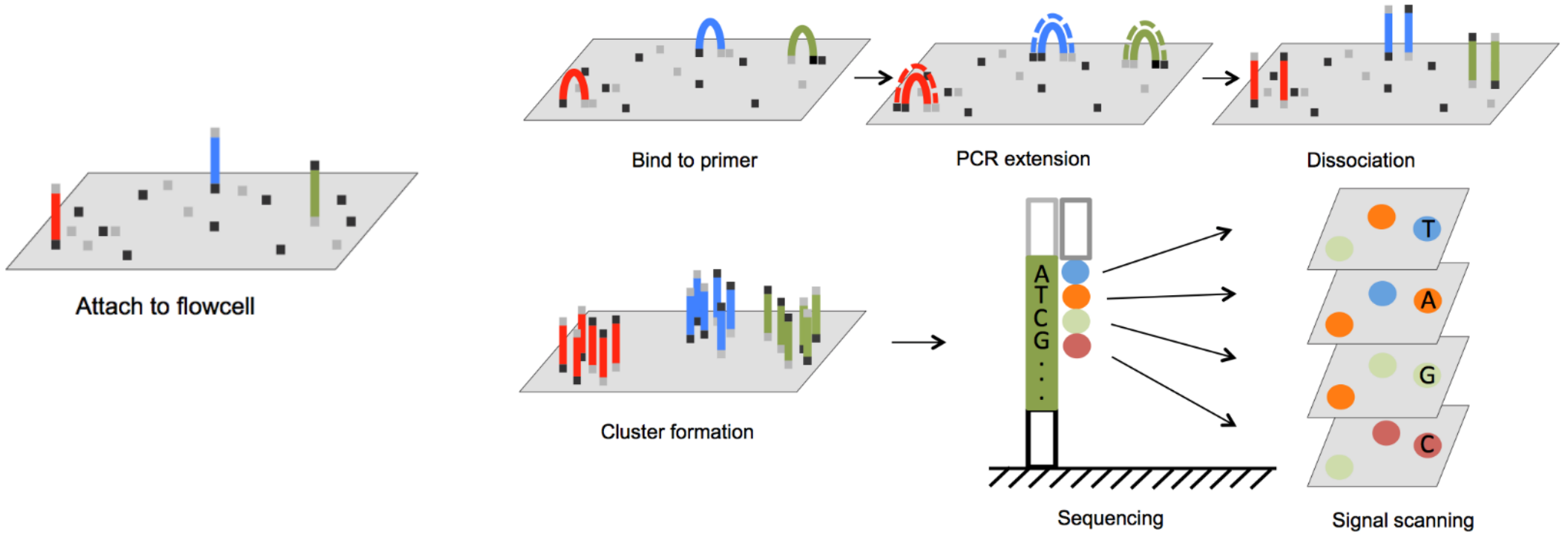
主要的RNA-Seq方法

北京大学 PEKING UNIVERSITY

- **mRNA测序：**灵敏而准确地定量基因表达，识别编码转录组中已知的和新的亚型，检测基因融合，测量等位基因特异性表达。
- **靶向RNA测序：**分析一组目标基因的表达。可以通过富集或基于扩增子的方法实现靶向RNA测序。
- **超低起始量，单细胞RNA-Seq：**使用深度RNA-Seq检测一个细胞在其周围环境中的信号和行为。该方法有利于生物学家研究分化、增殖、肿瘤发生等过程。
- **核糖体分析：**对核糖体保护的mRNA片段进行深度测序，获得特定时间细胞中活性核糖体的完整视图，并预测蛋白丰度。
- **总RNA测序：**准确检测基因和转录本丰度，并能检测编码RNA和多种形式的非编码RNA中已知的和新的特征。
- **小RNA测序：**分离并测序小RNA，例如microRNA，了解非编码RNA在基因沉默和转录后基因表达调控中的作用。
- **RNA外显子组捕获测序：**利用转录组编码区域的序列特异性捕获，实现经济有效的RNA外显子分析。非常适合低质量样本或有限的起始材料。

Bulk RNA-seq的操作及分析流程:

- **测序:** 对构建好的RNA文库进行高通量测序，通常使用Illumina等平台进行短读长度测序，生成原始测序数据（raw_data）。



Bulk RNA-seq的操作及分析流程：

- 在这里要注意，如果使用公共数据进行分析，从NCBI等数据库下载的原始数据很多为SRA格式，需要转换成fastq文件。常用工具为SRA Toolkit，*这时还需要对raw_data进行**质控***。
- 如果是自己测序，下机后得到了fastq的raw_data，通常测序公司在将数据返还给客户之前已经做了“clean”处理，即得到clean_data。
- **序列比对**： 将质控后的reads比对到参考基因组或转录组上，利用比对工具（如STAR、HISAT2）进行reads的比对和定位。

Bulk RNA-seq的操作及分析流程：

- 表达量计数： 根据比对结果，使用计数工具（如featureCounts、HTSeq）计算每个基因的表达量，通常以**FPKM（fragments per kilobase of transcript per million mapped reads）或TPM（transcripts per million）**为单位。

Why Normalize RNA-Seq Data?

Normalization is a crucial step in RNA-Seq analysis for the following reasons:

- **Sequencing depth:** Different RNA-Seq experiments produce varying numbers of reads, making direct comparisons between samples misleading.
- **Gene length:** Longer genes inherently generate more reads, irrespective of their actual expression level.
- **Bias reduction:** Normalization mitigates technical biases, enabling meaningful biological interpretation.

TPM (Transcripts Per Million)

TPM measures the proportion of reads mapped to a transcript, normalized by transcript length and sequencing depth. It is calculated as:

Key Features:

1. **Proportionality:** TPM values sum to 1,000,000 across all transcripts in a sample, making it easier to compare between samples.
2. **Intuitive interpretation:** TPM values directly represent the abundance of transcripts in a sample.
3. **Preferred for comparisons:** TPM facilitates between-sample comparisons better than FPKM.

FPKM (Fragments Per Kilobase Million)

FPKM normalizes read counts by transcript length and sequencing depth, but without enforcing proportionality like TPM. It is defined as:

Key Features:

1. **Historical significance:** FPKM was one of the first normalization methods used for RNA-Seq.
2. **Single-end vs. paired-end:** In paired-end sequencing, FPKM becomes RPKM (Reads Per Kilobase Million).
3. **Limited utility:** FPKM values are not as robust as TPM for cross-sample comparisons due to lack of proportionality.

3.1 前期准备与环境搭建

北京大学 PEKING UNIVERSITY

➢ 基础环境搭建

1. Conda是什么：Conda是跨平台的“软件管家”+“环境隔离器”，专门解决Python/R等语言的依赖冲突问题

核心功能：

✅ **包管理**：一键安装/卸载软件（如Python库、R工具），自动解决依赖关系（比如装A库时自动安装它需要的B库）。

✅ **环境隔离**：创建多个“独立房间”（虚拟环境），每个房间可装不同软件版本，互不干扰

2. 为什么使用Conda?

- 不同项目需不同Python版本——创建独立环境，自由切换Python 3.6/3.9/3.11等，避免重装系统或手动改配置
- 安装库时提示“依赖冲突”——自动识别兼容版本
- 想复现别人的实验环境——导出环境配置文件，一键还原所有依赖
- Windows装库总报错（缺DLL等）——Conda预编译包适配系统，
- 同时用Python/R/Julia等多语言——统一管理不同语言的工具链，不用学多套包管理命令

3.1 前期准备与环境搭建

北京大学 PEKING UNIVERSITY

➢ 使用conda安装分析工具

分析阶段	安装工具	安装命令	核心功能	备注
数据获取	sra-tools	conda install -c bioconda sra-tools	从NCBI下载SRA数据并转FASTQ	
	aspera-cli	conda install -c hcc aspera-cli	高速下载SRA数据（替代wget）	
质量控制	FastQC	conda install -c bioconda fastqc	原始数据质量评估（生成HTML报告）	
	Trim Galore	conda install -c bioconda trim-galore	自动切除接头与低质量序列（整合Cutadapt）	
序列比对	STAR	conda install -c bioconda star	高效RNA-seq比对（支持剪接比对）	
重复序列标记	Picard	conda install -c bioconda picard	标记PCR重复序列	需Java环境
定量分析	featureCounts	conda install -c bioconda subread	基因/转录本计数（速度快精度高）	
	HTSeq	conda install -c bioconda htseq	基于比对结果的基因计数	

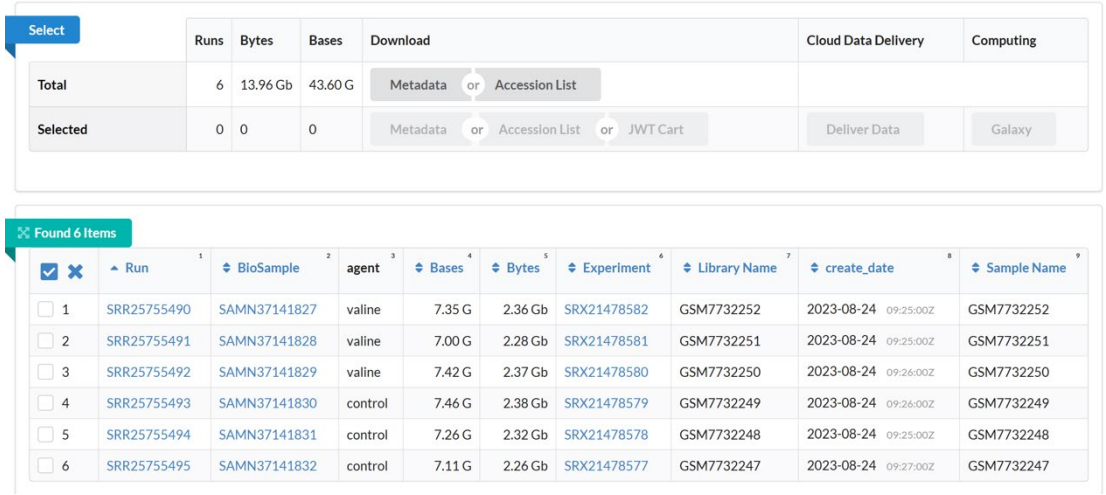
```
conda install -c conda-forge openjdk=11 # 安装Java export
JAVA_HOME=$CONDA_PREFIX # 设置临时环境变量
```

3.1 前期准备与环境搭建

北京大学 PEKING UNIVERSITY

➢ **数据获取：**从 NCBI SRA 数据库获取 GSE241623 的原始测序数据（fastq 格式）

✓ **获取数据源 Accession List**



Select

	Runs	Bytes	Bases	Download	Cloud Data Delivery	Computing
Total	6	13.96 Gb	43.60 G	Metadata or Accession List		
Selected	0	0	0	Metadata or Accession List or JWT Cart	Deliver Data	Galaxy

Found 6 Items

	Run	BioSample	agent	Bases	Bytes	Experiment	Library Name	create_date	Sample Name
1	SRR25755490	SAMN37141827	valine	7.35 G	2.36 Gb	SRX21478582	GSM7732252	2023-08-24 09:25:00Z	GSM7732252
2	SRR25755491	SAMN37141828	valine	7.00 G	2.28 Gb	SRX21478581	GSM7732251	2023-08-24 09:25:00Z	GSM7732251
3	SRR25755492	SAMN37141829	valine	7.42 G	2.37 Gb	SRX21478580	GSM7732250	2023-08-24 09:26:00Z	GSM7732250
4	SRR25755493	SAMN37141830	control	7.46 G	2.38 Gb	SRX21478579	GSM7732249	2023-08-24 09:26:00Z	GSM7732249
5	SRR25755494	SAMN37141831	control	7.26 G	2.32 Gb	SRX21478578	GSM7732248	2023-08-24 09:25:00Z	GSM7732248
6	SRR25755495	SAMN37141832	control	7.11 G	2.26 Gb	SRX21478577	GSM7732247	2023-08-24 09:27:00Z	GSM7732247

SRR_Acc_List.txt

这个文件包含一个或多个 SRA 运行编号 (SRR numbers)，每行一个

```
SRR25755495
SRR25755496
...
```

用途： 这个列表文件告诉后续工具 prefetch 需要下载哪些 SRA 文件

1. **操作：** 在 GEO 数据库 (*https://www.ncbi.nlm.nih.govgeo/query/acc.cgi?acc=GSE241623*) 找到 GSE241623 数据集。点击网页上的 “**Run Selector**” 链接。
2. **操作：** 在 Run Selector 页面，点击 “**Accession List**” 按钮进行下载。
3. **结果：** 得到一个名为 SRR_Acc_List.txt 的文本文件。

3.1 前期准备与环境搭建

北京大学 PEKING UNIVERSITY

➢ **数据获取：**从 NCBI SRA 数据库获取 GSE241623 的原始测序数据（fastq 格式）

✓ **将 SRA (.sra) 文件转换为 Fastq (.fastq.gz) 文件**

方案二：使用 fasterq-dump (更快但需额外压缩步骤)

```
fasterq-dump -p -e 30 --split-3 -O ./ SRR25755495.sra
gzip *
```

1.fasterq-dump——对于 output/SRR25755495.sra：
- 生成 SRR25755495_1.fastq 和 SRR25755495_2.fastq（注意：没有 .gz 压缩！）
- (+ 可能的 SRR25755495.fastq.gz 存放未配对 reads)

2.**gzip ***——文件名变为 .fastq.gz

1. fastq-dump：NCBI 开发的 fastq-dump 的替代品，用 C 语言重写并优化算法。其主要优势就是速度显著更快
2. -p：显示进度 (Progress) 信息。
3. -e 30：-e (Threads) 设置处理使用的**线程数**。这里设为 30。这个值需要根据服务器的 CPU 核心数进行调整，设置得越高（在核心数允许范围内），转换速度通常越快
4. --split-3：功能与 fastq-dump 中的 --split-3 **完全相同**。按双端拆分读段，并处理未配对 reads。
5. -O ./：指定输出目录为当前目录 (./)。
6. SRR25755495.sra：指定要转换的 **单个** .sra 文件（ fasterq-dump **无法像方案一那样使用通配符 *.sra 一次性处理多个文件**。你必须为每个 .sra文件单独运行一次命令，或者写一个循环脚本来处理列表中的所有文件）、
7. **gzip ***：由于 fasterq-dump **本身不提供内置压缩功能** (-p 不是压缩)，我们必须手动压缩所有生成的 .fastq 文件。*：通配符，匹配**当前目录下所有文件。结果：** 所有 .fastq 文件（上一步 fasterq-dump 生成的）被逐一压缩，文件名变为 .fastq.gz

3.2 上游数据分析

➢ 质控

序列过滤（Trim Galore）

```
trim_galore -q 20 \                    #切除质量值 < Q20（错误率>1%）的碱基
            --length 45 \              #丢弃长度 < 45bp的reads
            --max_n 3 \                #丢弃含 >3个N碱基的reads
            --stringency 3 \           #仅当接头重叠 ≥3bp才切除
            --fastqc \                 #过滤后自动再次运行FastQC
            -o ${QC_file} \
            --paired \                 #双端模式同步处理R1/R2
            ${rawdata}/${sample}_R1.fastq.gz ${rawdata}/${sample}_R2.fastq.gz \
            >${QC_file}/${sample}.QC.log
```

Measure	Value
Filename	SRR25755490_R1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	24487172
Total Bases	3.6 Gbp
Sequences flagged as poor quality	0
Sequence length	150
%GC	50

Total Sequences (总序列数)	24,187,172	**满足哺乳动物 RNA-seq 深度要求**
Total Bases (总碱基数)	3.6Gbp	**3.66 × 10^9 碱基** = 3.6 Gbp，
Sequences flagged as poor quality (低质量序列)	0	**无低质量序列**，表明过滤有效 (Trim Galore 作用显著)
Sequence length (读长)	150 bp	(标准 Illumina NextSeq 平台配置)
GC (GC含量)	50%	**核心质控指标**，小鼠预期 GC 含量 45-50% ✅ 无污染 (细菌/引物二聚体 GC 异常)

3.2 上游数据分析

北京大学 PEKING UNIVERSITY

➢ 测序数据比对

```
STAR --runMode alignReads \                    # 设置比对模式
    --runThreadN 15 \                          # 计算资源分配
    --alignIntronMax 50000 \                   # 内含子最大长度
    --genomeDir ${REF} \                       # 参考基因组索引路径
    --readFilesIn ${QC_file}/${sample}_R1_val_1.fq.gz ${QC_file}/ ${sample}_R2_val_2.fq.gz \  # 输入文件
    --readFilesCommand zcat \                  # 解压方式
    --outReadsUnmapped Fastx \                 # 未比对reads输出格式
    --outFilterMismatchNoverReadLmax 0.1 \     # 容错率参数
    --outSAMattributes NH HI AS NM MD \        # 输出属性配置
    --outFileNamePrefix ${STAR_mapping}/${sample}_ \  # 输出前缀
    --outSAMtype BAM SortedByCoordinate \      # 输出格式与排序
    --outBAMsortingThreadN 10 \                # 排序线程数
    --quantMode TranscriptomeSAM GeneCounts # 定量模式
```

1. **--alignIntronMax 50000**：允许的最大内含子长度-覆盖哺乳动物99%的基因（小鼠最大内含子~30kb）
2. **--outFilterMismatchNoverReadLmax 0.1**：允许的错配比例-150bp允许≤15个错配，适应测序错误和多态性位点
   - Illumina测序平均错误率约0.1-1% (Q20-Q30)
   - 小鼠基因组SNP密度约1/1000bp
   - 总容错阈值 ≈ 技术错误 + 生物变异 < 10%
3. **--outSAMattributes NH HI AS NM MD**：

NH	比对位置数	多重比对过滤
HI	比对序号	区分主/次要比对
AS	比对得分	最佳比对筛选
NM	编辑距离	错配碱基统计
MD	错配位置	SNP calling基础

3.2 上游数据分析

北京大学 PEKING UNIVERSITY

➢ 基因表达定量

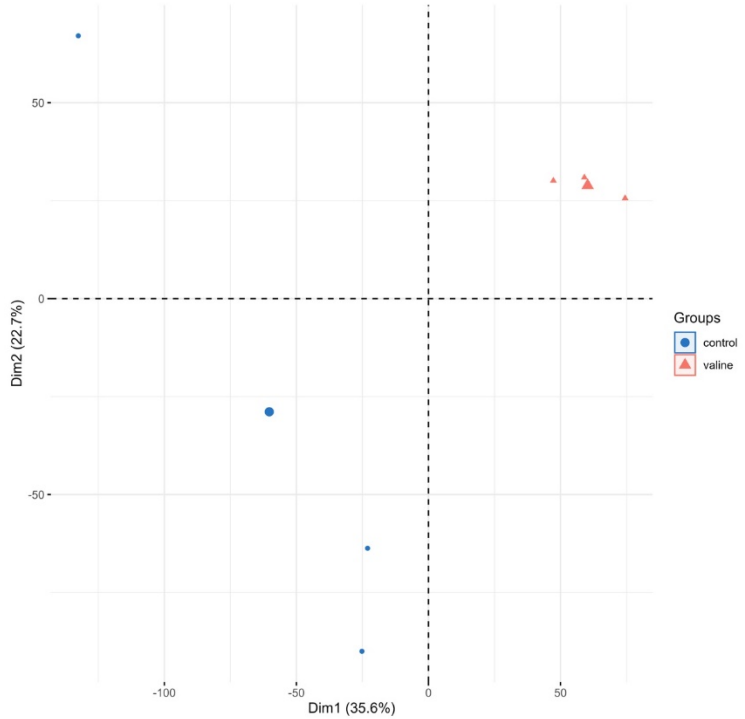
```
mkdir -p ./${sample}/featureCounts  #创建输出目录
featureCounts=./${sample}/featureCounts #定义路径变量
featureCounts    -t exon \                  # 计数目标为外显子
                 -g gene_id \           # 计数基因ID
                 --primary \            # 只保留主要比对
                 -J \                    # 统计可变剪切
                 -p \                    # 双端测序模式
                 --countReadPairs \ # 按read pair计数
                 -T 10 \                 # 10线程
                 -a ${GTF} \            # GTF注释文件
                 -R BAM \               # 输出read归属
                 -o ${featureCounts}/${sample}_primary_gene \ # 输出文件
                 ${dedup}/${sample}_Aligned.redup.bam          # 输入去重BAM
```

3.3 下游数据分析

➢ PCA分析

```
exprSet[1:6,1:6]                    #对表达式矩阵exprSet进行索引操作，提取前6行和前6列的数据，以验证数据质量

#函数draw_pca（来自前面的包tinyarray）会执行主成分分析（PCA），并根据分组信息绘制PCA图
pcagp <- draw_pca(exprSet, Group1)
ggsave("./pcagp.jpg", plot=pcagp, width=8, height=8, dpi=500)
```



PCA（Principal Component Analysis） 是一种维度缩减技术，通过线性变换将高维数据投影到低维空间，同时保留数据的主要变异特征。

PCA能：
1.提取最关键的变化模式
2.将样本间的复杂关系简化为2-3个维度直观展示
3.揭示隐藏的生物学模式

3.2 上游数据分析

➢ 测序数据比对

```
STAR --runMode alignReads \                    # 设置比对模式
    --runThreadN 15 \                          # 计算资源分配
    --alignIntronMax 50000 \                   # 内含子最大长度
    --genomeDir ${REF} \                       # 参考基因组索引路径
    --readFilesIn ${QC_file}/${sample}_R1_val_1.fq.gz ${QC_file}/ ${sample}_R2_val_2.fq.gz \  # 输入文件
    --readFilesCommand zcat \                  # 解压方式
    --outReadsUnmapped Fastx \                 # 未比对reads输出格式
    --outFilterMismatchNoverReadLmax 0.1 \     # 容错率参数
    --outSAMattributes NH HI AS NM MD \        # 输出属性配置
    --outFileNamePrefix ${STAR_mapping}/${sample}_ \  # 输出前缀
    --outSAMtype BAM SortedByCoordinate \      # 输出格式与排序
    --outBAMsortingThreadN 10 \                # 排序线程数
    --quantMode TranscriptomeSAM GeneCounts # 定量模式
```

1. **--alignIntronMax 50000**：允许的最大内含子长度-覆盖哺乳动物99%的基因（小鼠最大内含子~30kb）
2. **--outFilterMismatchNoverReadLmax 0.1**：允许的错配比例-150bp允许≤15个错配，适应测序错误和多态性位点
   - Illumina测序平均错误率约0.1-1% (Q20-Q30)
   - 小鼠基因组SNP密度约1/1000bp
   - 总容错阈值 ≈ 技术错误 + 生物变异 < 10%
3. **--outSAMattributes NH HI AS NM MD**：

NH	比对位置数	多重比对过滤
HI	比对序号	区分主/次要比对
AS	比对得分	最佳比对筛选
NM	编辑距离	错配碱基统计
MD	错配位置	SNP calling基础

3.2 上游数据分析

➢ 基因组及注释数据

•.fa/.fasta：基因组序列文件
•.gtf：基因结构注释文件

下载小鼠参考基因组，使用STAR软件构建相关的索引信息

```
wget -c
https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/release_M37/gencode.vM37.primary_assembly.annotation.gtf.gz
wget –c
https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/release_M37/GRCm39.primary_assembly.genome.fa.gz
```

解压下载文件

```
gunzip gencode.vM37.*.gz
gunzip GRCm39.*.gz
```

STAR基因组索引构建

```
STAR --runMode genomeGenerate \      # 设置工作模式为索引生成
        --runThreadN 30 \            # 使用30个CPU线程并行加速
        --genomeDir /data/.../mm38_star \ # 索引输出目录
        --genomeFastaFiles ${REF}/GRCm38.primary_assembly.genome.fa \  # 基因组FASTA文件
        --sjdbGTFfile ${REF}/gencode.vM25...gtf \     # 基因注释GTF文件
        --sjdbOverhang 149           # 指定读长减1（适配150bp测序）
```

3.2 上游数据分析

北京大学 PEKING UNIVERSITY

➢ 基因表达定量

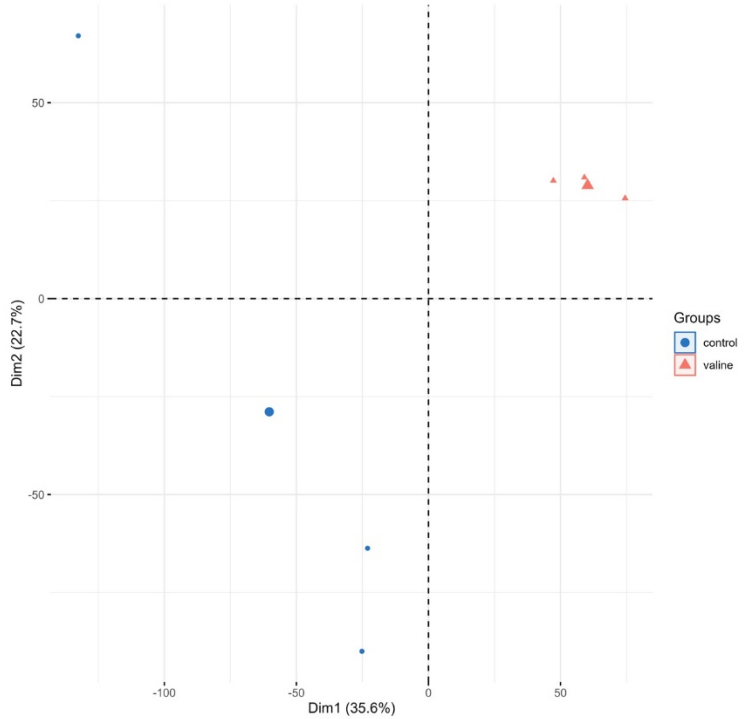
```
mkdir -p ./${sample}/featureCounts  #创建输出目录
featureCounts=./${sample}/featureCounts #定义路径变量
featureCounts    -t exon \                  # 计数目标为外显子
                 -g gene_id \           # 计数基因ID
                 --primary \            # 只保留主要比对
                 -J \                       # 统计可变剪切
                 -p \                       # 双端测序模式
                 --countReadPairs \ # 按read pair计数
                 -T 10 \                    # 10线程
                 -a ${GTF} \            # GTF注释文件
                 -R BAM \               # 输出read归属
                 -o ${featureCounts}/${sample}_primary_gene \ # 输出文件
                 ${dedup}/${sample}_Aligned.redup.bam          # 输入去重BAM
```

3.3 下游数据分析

➢ PCA分析

```
exprSet[1:6,1:6]                    #对表达式矩阵exprSet进行索引操作，提取前6行和前6列的数据，以验证数据质量

#函数draw_pca（来自前面的包tinyarray）会执行主成分分析（PCA），并根据分组信息绘制PCA图
pcagp <- draw_pca(exprSet, Group1)
ggsave("./pcagp.jpg", plot=pcagp, width=8, height=8, dpi=500)
```



PCA（Principal Component Analysis） 是一种维度缩减技术，通过线性变换将高维数据投影到低维空间，同时保留数据的主要变异特征。

PCA能：
1.提取最关键的变化模式
2.将样本间的复杂关系简化为2-3个维度直观展示
3.揭示隐藏的生物学模式

3.3 下游数据分析

➢ limma差异分析

1.样本分组与实验设计

```
design <-
model.matrix(~Group1)
```

2.数据标准化与质量控制

```
dge_limma <- DGEList(counts=filter_count)     #创建DGEList对象
dge_limma_1 <- calcNormFactors(dge_limma)   #TMM标准化
```

$$\mathrm{TMM}_s = \frac{\sum_{g \in G} w_g^s \log_2\left(\frac{y_g^s}{y_g^r}\right)}{\sum w_g^s}$$

- G：表达稳定的管家基因集
- w_g^s：Trimmed Mean加权
- **作用**：消除样本间RNA组成差异

3. MDS可视化

```
plotMDS(dge_limma_1, col=as.numeric(Group1)，main = "TMM归一化后的MDS图")
```

4. limma-voom差异分析核心

```
v <- voom(dge_limma_1, design, plot=TRUE，normalize.method = "quantile")
fit_limma <- lmFit(v, design)   #线性建模
fit_limma <- eBayes(fit_limma)   #经验贝叶斯收缩
plotSA(fit_limma, main="Final model: Mean-variance trend")
```

- 拟合均值-方差关系
- 计算观测权重（降低高离散基因影响）
- 输出连续表达值（兼容线性模型）

5. 差异结果提取

```
DEG_limma_voom <- topTable(fit_limma, coef=2, number=Inf)   #结果表格输出
logFC_t=1
P.Value_t = 0.05
DEG_limma_voom <- mutate(DEG_limma_voom, change = ifelse(k1,"down", ifelse(k2,"up","stable")))
```

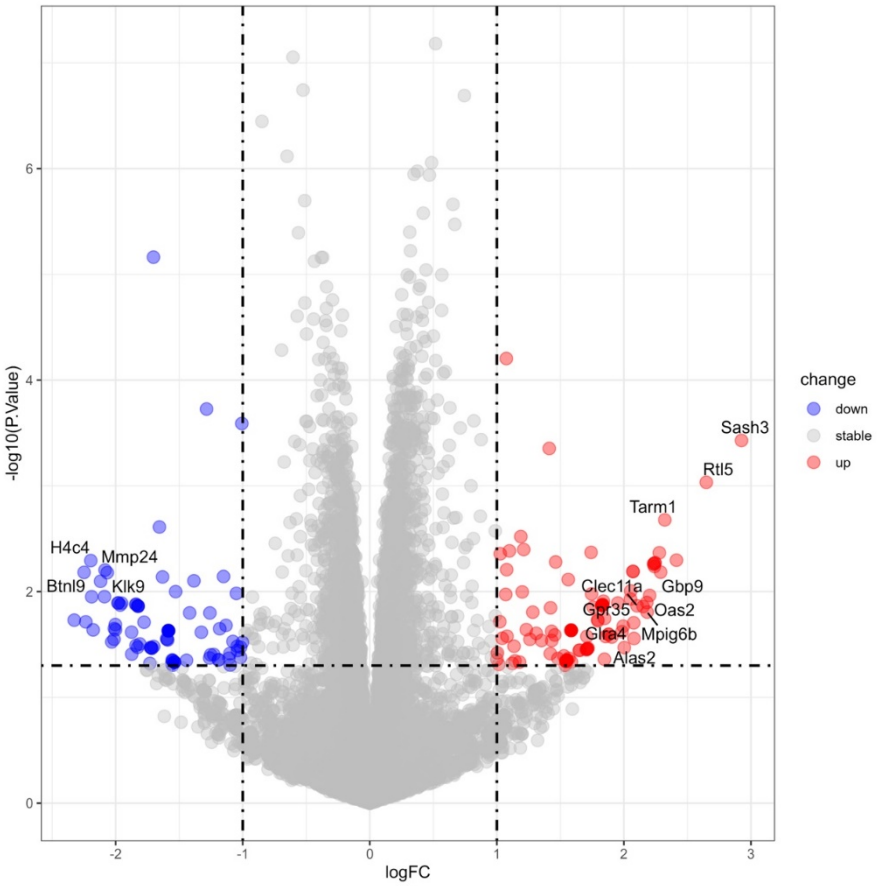
阈值：|logFC| > 1 & P.Val < 0.05

北京大学 PEKING UNIVERSITY

➢ limma差异分析

6.火山图绘制

```
DEG_limma_voom$logFC <- as.numeric(DEG_limma_voom$logFC)
up_data_logFC <- filter(DEG_limma_voom, change ==  'up' ) %>%
# 从DEG_limma_voom中筛选出上调的基因
  top_n(20, logFC)
down_data_logFC <- filter(DEG_limma_voom, change == 'down') %>%
# 从DEG_limma_voom中筛选出上调的基因
    top_n(20, -logFC)
# 选择-LogFC值最大的前20个基因
library(ggrepel)
DEG_limma1 <- DEG_limma +  # 基于普通火山图
  geom_text_repel(data = up_data_logFC,
                            aes(x = logFC, y = -log10(P.Value), label = rownames(up_data_logFC))) +  #
                            添加上调基因的标签
  geom_text_repel(data = down_data_logFC,
                            aes(x = logFC, y = -log10(P.Value), label = rownames(down_data_logFC)))
                            # 添加下调基因的标签
```



3.3 下游数据分析

➢ GO分析

1.加载必要包

```
library(clusterProfiler)
library(ggthemes)
……
```

2.上调基因GO富集分析

```
up_data <- filter(DEG_limma_voom, change == 'up  ') #提取上调基因
up_data$symbol <- rownames(up_data) #添加基因符号列
up_data_ENTR <- bitr(up_data$symbol,
                     fromType = "SYMBOL",
                     toType =  "ENTREZID"  ,
                     OrgDb = org.Mm.eg.db) #SYMBOL转ENTREZID
up_data <- inner_join(up_data, up_data_ENTR, by=c( "symbol"  =  "SYMBOL" ))
#关键转换结果
ego_up <- enrichGO(gene = up_data$ENTREZID,
          OrgDb= org.Mm.eg.db,
          ont = "ALL",
          readable = TRUE)        #GO富集分析
```

超几何分布检验（Fisher精确检验变体）

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

- N：背景基因总数
- M：通路中基因数
- n：输入基因数
- k：输入基因中属于通路的基因数

3.可视化

```
ego_up_limma_plot <- barplot(ego_up,
                             split =  "ONTOLOGY" ,
                             font.size = 10,
                             showCategory = 5) +
facet_grid(ONTOLOGY ~ ., space = "free_y", scales = "free_y")
```