mRNA-Seq 差异表达一般流程及高阶分析

General Pipeline and Advanced Analysis of mRNA-Seq Differential Expression

谢忱*

Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, Ploen, Germany

*通讯作者邮箱: xie@evolbio.mpg.de

引用格式: 谢忱. (2021). mRNA-Seq 差异表达一般流程及高阶分析. *Bio-101* e1010645. Doi: 10.21769/BioProtoc. e1010645.

How to cite: Xie, C. (2021). General pipeline and advanced analysis of mRNA-Seq differential expression. *Bio-101* e1010645. Doi: 10.21769/BioProtoc.1010645. (in Chinese)

摘要: 利用 mRNA-Seq 技术测得两组或多组样本的转录组数据, 并通过差异表达分析和对所发现的差异表达基因集合进行功能富集分析以推断生物学功能, 是包括动物进化研究领域在内的诸多生物学领域中被普遍应用的重要研究思路。然而, 在实验设计方面所涉及的样本量和测序深度等重要参数的选择通常并没有被重视。例如绝大多数研究的样本量为 3-5, 仅仅达到可进行统计分析的最低标准。此外, 由于实验前未知原因或实验中无法控制的变量, 可能会导致得到的转录组数据出现分层。针对这种情况的数据分析可以厘清导致分层的原因并在此基础上进行差异表达分析。这对野生动物样本研究可能会更有意义, 因为野外采集的样本比实验室中收集的样本更难控制各类条件。本文首先介绍 mRNA-Seq 差异表达目前被普遍应用的一般流程, 之后会针对上述两个问题介绍功效分析和分层分析的方法, 以期会有助于包括动物进化领域在内的生物学研究。

关键词: 转录组, mRNA-Seq, 差异表达分析, 功能富集分析, 功效分析, 分层分析

研究背景: 利用 mRNA-Seq 技术测得两组或多组样本的转录组数据, 并通过差异表达分析和对所发现的差异表达基因集合进行功能富集分析以推断其生物学功能, 是包括动物进化研究领域在内的诸多生物学领域中被普遍应用的重要研究思路。针对此类分析流程的方法介绍目前已有许多, 包括中文文章 (赵 婧和刘 南, 2019)。

然而，mRNA-Seq 的实验设计往往没有被重视，与研究结果准确性息息相关的参数没有被科学严谨的选择，例如样本量和测序深度。功效分析可以帮助研究者了解既定实验在真实存在组间差异时可以统计检验出显著性差异的几率，从而有助于实验设计和对实验结果的解释。此外，实验得到的转录组数据可能由于实验前未知原因或实验中无法控制的变量出现分层的情况。这种情况更可能出现在野生动物样本的数据中，因为野外采集的样本比实验室中收集的样本更难控制各类条件。针对这种情况进行分层分析可以厘清导致分层的原因并在此基础上对数据进行分组，再基于分组后的数据进行差异表达分析。

本文首先介绍 mRNA-Seq 转录组差异表达分析目前被普遍应用的一般流程，其中侧重于介绍之前中文文章 (赵 婧和刘 南, 2019) 中未提及或简略提及的内容，与其互为补充。之后会针对上述两个问题介绍功效分析和分层分析的方法，以期会有助于包括动物进化领域在内的生物学研究。本文基于作者本人以第一作者于 2019 年发表于 *eLife* 期刊 (Xie *et al.*, 2019)，以及第一作者兼通讯作者于 2020 年发表于 *PLoS Computational Biology* 期刊 (Xie *et al.*, 2020) 的两篇开放获取 (open access) 文章中的相关内容。

仪器设备

1. Linux 系统服务器 (理论上几乎当前任何配置的服务器均可以完成分析，但是读者需根据实际数据量使用相应性能的服务器；作者进行上文提到文章中的数据分析所使用的服务器配置相对比较高，即具备 20 个计算节点且每个节点有 512GB 内存和 32 个 2.1GHz 的 CPU 核，这样只利用部分计算资源就在很短的时间内完成了分析)。
2. 个人电脑，系统不限，用于运行 R 软件。此非必须，可在服务器上运行 R 软件。

软件版本信息及下载地址

此处所列软件及其版本为下文实验步骤中及本文所基于的两篇文章中所使用的。其它相关软件及其下载地址见实验步骤中所涉及处。建议读者在实际研究工作中使用所有软件的最新版本。

1. FastQC (0.11.4, "<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>").

2. Trimmomatic (Bolger *et al.*, 2014) (0.35, "<http://www.usadellab.org/cms/?page=trimmomatic>").
3. HISAT2 (Kim *et al.*, 2019) (2.0.4, "<http://daehwankimlab.github.io/hisat2/>").
4. SAMtools (Li and Durbin, 2009) (1.3.1, "<http://www.htslib.org/>").
5. HTSeq (Anders *et al.*, 2015) (0.6.1p1, "<https://htseq.readthedocs.io/en/master/>").
6. R (3.3.1, "<https://www.r-project.org/>").
7. DESeq2 (Love *et al.*, 2014) (1.14.1, "<https://bioconductor.org/packages/release/bioc/html/DESeq2.html>").
8. GOseq (Young *et al.*, 2010) (1.38.0, "<https://bioconductor.org/packages/release/bioc/html/goseq.html>").
9. RnaSeqSampleSize (Zhao *et al.*, 2018) (1.6.0, "<https://bioconductor.org/packages/release/bioc/html/RnaSeqSampleSize.html>").

实验步骤

一、本文相关数据获取

1. 测序数据

本文所基于的原始测序数据 (FASTQ 格式) 可从 European Nucleotide Archive (ENA) 数据库 ("<https://www.ebi.ac.uk/ena/browser/home>") 下载 (PRJEB28348)。对数据的详细描述见本文所基于的两篇文章 (Xie *et al.*, 2019 and 2020)。

注 1: 已发表文章中的测序数据大多会被上传至上述的 ENA 数据库, 或 NCBI 的 Sequence Read Archive (SRA) 数据库 ("<https://www.ncbi.nlm.nih.gov/sra>").

注 2: 研究者通常得到的 Illumina 测序原始数据为 FASTQ 格式, 但是有时也会得到 BCL 格式的文件。这时需要利用工具 *bcl2fastq* ("https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html") 将其转为 FASTQ 格式以进行后续分析。

2. 基因组序列和基因注释数据

本文所基于的小鼠基因组序列和基因注释数据可从 Ensembl (Version 86) 下载 ("http://ftp.ensembl.org/pub/release-86/fasta/mus_musculus/dna/Mus_musculus.GRCm38.dna.primary_assembly.fa.gz") 和 "

[86/gtf/mus_musculus/Mus_musculus.GRCm38.86.gtf.gz">86/gtf/mus_musculus/Mus_musculus.GRCm38.86.gtf.gz](#)")。建议读者在实际研究工作中使用最新版本数据。对于小鼠数据，目前最新版的基因组为 GRCm39/mm39。

二、mRNA-Seq 转录组差异表达分析目前被普遍应用的一般流程

1. 测序数据质量评估

通常利用 **FastQC** 工具对测序数据进行质量评估。Linux 命令见图 1。

```
fastqc \
  -t <threads> \ #线程数
  -o <output_dir> \ #输出文件夹
  <input_file_1> <input_file_2> #输入文件，FASTQ 格式
```

图 1. FastQC 命令

结果文件中包含一系列质量评估项目，其中比较重要的包括测序数据的基本信息(Basic Statistics)、直观了解整体测得碱基质量(Per Base Sequence Quality)和 "Adapter Content" (是否包含 adapter)，详细说明见 "<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/>"。

2. 原始测序数据预处理

原始测序数据中可能存在低质量序列，以及一些 read 含有 adapter，需要进行预处理以便后续分析。推荐使用 **Trimmomatic** 工具，与其它类似工具相比，其除了具有运行速度快的优点，对于双端 (paired-end) 测序的数据可以利用被称为 "palindrome" 的方法更加有效地处理只包含部分 adapter 序列的 read。Linux 命令见图 2。

```
java -jar trimmomatic-0.35.jar \
  PE \ #此例子中数据为双端测序
  -threads <threads> \ #线程数
  -phred33 \ #碱基质量编码方式
  <input_file_1> <input_file_2> \ #输入文件，一对 FASTQ 格式文件
  <paired_output_1> <unpaired_output_1> <paired_output_2> <unpaired output_2> \ #输出文件，FASTQ 格式，均可用于后续序列比对；或只保留配对文件用于比对
  ILLUMINACLIP:<adapters_PE>:2:30:10 \ #去除 adapter
  LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 #去除低质量序列
```

图 2. Trimmomatic 命令

注：对原始测序数据预处理后可再次利用 *FastQC* 检查数据质量。

3. 将测序得到的 read 比对到参考基因组上

目前 mRNA-Seq 转录组差异表达分析通常会将经过预处理的 read 比对 (mapping) 到参考基因组上，并以此为基础进行后续分析。对于没有合适参考基因组，以及不通过序列比对进行差异表达分析这两种情况，均有相应的分析方法，但不在本文讨论范围之内。目前比较流行的两个工具为 HISAT2 和 STAR (Dobin *et al.*, 2013) ("<https://github.com/alexdobin/STAR>"), 各具优点。本文以 HISAT2 为例，而上文提到的中文方法介绍 (赵 婧和刘 南, 2019) 中则以 STAR 为例。

首先利用 hisat2-build 建立参考基因组序列索引。HISAT2 的一个优势是可以利用已知基因注释信息或 SNP 注释信息等提高序列比对能力。此例子中利用了小鼠已有基因注释信息。随后利用 hisat2 进行序列比对。比较重要的参数包括"--rna-strandness" (测序建库是否为 stranded, 以及相应方法) 和"--dta" (输出结果便于 StringTie 进行转录本组装)。Linux 命令见图 3。

```
extract_splice_sites.py \
    <gtf_file> \ #基因注释文件
    > <splice_site_file> #hisat2-build 输入文件
extract_exons.py \
    <gtf_file> \ #基因注释文件
    > <exon_file> #hisat2-build 输入文件
hisat2-build \
    -p <threads> \ #线程数
    --ss <splice_site_file> \ #剪切位点文件，由 HISAT2 脚本生成
    --exon <exon_file> \ #外显子文件，由 HISAT2 脚本生成
    <reference_file> \ #参考基因组序列文件
    <ht2_index_base> #输出文件前缀
hisat2 \
    -x <ht2_index_base> \ #参考基因组序列索引文件前缀
    -1 <input_file_1> \ #此例子为双端测序，FASTQ 文件 1
    -2 <input_file_2> \ #相应的 FASTQ 文件 2，可以同时提供多对文件
    --rna-strandness RF \ #此例子为 stranded 测序建库
    --dta \ #输出结果便于 StringTie 进行转录本组装
    -p <threads> \ #线程数
    | samtools view -Su - | samtools sort -@ <threads> - <sorted_bam_file> #利用 SAMtools 将结果
    转为 bam 格式并排序输出
```

图 3. HISAT2 命令

注 1：在本例子中，由于 (1) 测序数据来源于实验小鼠系 C57BL/6N，其基因组与小鼠参考基因组极为接近；(2) 小鼠基因组已有基因注释已经相对完备；以及 (3) 此分析想要得到的重点结果为已有功能注释的基因的差异表达情况，因此没有再利用测序数据组装新的转录本。如需组装新的转录本，可利用 StringTie (Pertea et al., 2015) ("<https://ccb.jhu.edu/software/stringtie/>") 等工具。

注 2：如上所述，由于此例子中测序数据来源于实验小鼠系 C57BL/6N，其基因组与小鼠参考基因组极为接近，因此几乎使用了 HISAT2 所有默认参数值。但是对于样本进化距离与参考基因组相差较远的情况，或者有其他特殊分析需要时，读者需根据具体情况相应调整参数。针对样本进化距离与参考基因组相差较远的这一情况，推荐尝试修改"Scoring options"中各参数。

注 3: 对于 mRNA-Seq 数据, 在 read 比对后, 可以利用 RSeQC (Wang *et al.*, 2012) 等工具进行 mRNA-Seq 特异的数据质量评估。

4. 统计比对到各基因的 read/fragment 数目

在进行差异表达分析前, 需要在每个样本中统计比对到各个基因 (或其它基因组区域) 的 read 数目。如果是双端测序, 则需要统计 read pair, 即 fragment 的数目。本文中使用的工具为 HTSeq 中的 htseq-count, 但更推荐读者使用 featureCounts (Liao *et al.*, 2014) ("<http://subread.sourceforge.net/>")。这两个工具与其它类似工具的统计结果并没有非常大的差异, 但是 featureCounts 运行速度更快且可选参数更多。运行 htseq-count 的 Linux 命令见图 4。

```
htseq-count \
  -f bam \ #输入数据格式
  -s reverse \ #此例子为 stranded 测序建库
  -a 60 \ #最小 mapping quality
  -m intersection-strict \ #对映射到多个基因的 read 的处理方式
  -q \ #安静运行模式
  <bam_file> \ #按 read name 排序的 bam 文件
  <gtf_file> \ #基因注释文件
  > <fragment_count_file> #输出文件
```

图 4. HTSeq 命令

5. 差异表达分析

此步骤是 mRNA-Seq 转录组差异表达分析流程的核心。目前被普遍认可的分析方法是利用每个样本的 "raw read/fragment gene counts" 作为输入数据, 利用其服从负二项分布为核心思想进行统计检验。读者如感兴趣可以通过阅读早期 mRNA-Seq 分析相关的一系列经典文章 (Marioni *et al.*, 2008; Love *et al.*, 2014; Anders and Huber, 2010; Robinson *et al.*, 2010) 深入了解。

目前比较流行的软件包包括 DESeq2、edgeR (Robinson *et al.*, 2010) ("<https://bioconductor.org/packages/release/bioc/html/edgeR.html>")、limma (Law *et al.*, 2014) ("<https://bioconductor.org/packages/release/bioc/html/limma.html>")

等，其中前两者是利用负二项分布思想。正常情况下，对于同一组数据这些不同的方法应该得到类似的差异表达基因集合，如文章 (Xie *et al.*, 2020) 中图 3A 所示。本文以 DESeq2 为例介绍分析方法。下面的例子为文章 (Xie *et al.*, 2020) 中基因 *A930004D18Rik* 敲除小鼠系，胚胎期雌性敲除组 (12 个胚胎) 和野生组 (14 个胚胎) 的 mRNA-Seq 数据。R 代码见图 5。

```
library(DESeq2)

#"sample.info" is a "data.frame" with sample code ("code"), file name
#"name"), and genotype ("genotype": "KO" or "WT") information.
#"file.dir" is a folder with all the count files.

#创建 DESeqDataSet
sample.table <- data.frame(sampleName = sample.info[, "code"],
                           fileName = sample.info[, "name"],
                           genotype = sample.info[, "genotype"])
sample.dds <- DESeqDataSetFromHTSeqCount(sampleTable = sample.table,
                                         directory = file.dir,
                                         design = ~ genotype)
sample.dds$genotype <- relevel(sample.dds$genotype, ref = "WT")

#差异表达分析
sample.dds <- DESeq(sample.dds)
sample.res <- results(sample.dds)
sample.res <- sample.res[order(sample.res$padj), ]
```

图 5. DESeq2 差异表达分析代码

6. 功能富集分析

在差异表达分析后，依据所选择的显著性水平标准 (通常为 Adjusted P-Value ≤ 0.05)，得到一个差异表达基因集合。为了理解这一基因集合的生物学功能意义，需要针对其进行功能富集分析。

在差异表达分析过程中，不同基因由于各种原因 (详见功效分析) 被检验为统计显著的功效是不同的。例如对于两个其它情况相同的基因，有更多 read/fragment 数目的基因更倾向于被检验为差异表达基因。因此，在功能富集分析环节需要针对这种基因间功效偏差进行校正，以避免导致得到的富集的功能条目 (Gene Ontology

term 或 pathway) 仅仅是因为其中包含更多有高 read/fragment 数目的基因，进而错误地推断了生物学功能意义。基于基因集合的功能富集分析工具有很多，然而绝大多数没有考虑到 mRNA-Seq 中存在的功效偏差的问题，而 GOseq 软件包则解决了这一问题。上文例子 (图 5) 的 Gene Ontology (GO) 富集分析方法见下 (图 6)。GOseq 软件包除了可以进行 GO 富集分析，还可以利用如 KEGG pathway 等其它功能注释数据进行分析。

```
library(goseq)

#"sample.res" is the last "sample.res" in Figure 5.

#功能富集分析准备工作：确定背景基因集和显著差异表达基因集，以及 count 偏见的校正
sample.res.bg <- sample.res[is.na(sample.res[, "padj"]) == F, ]
sample.res.genes <- as.integer(sample.res.bg$padj <= 0.05)
names(sample.res.genes) <- rownames(sample.res.bg)
sample.res.countbias <- sample.res.bg$baseMean
sample.res.pwf <- nullp(sample.res.genes,
                        bias.data = sample.res.countbias)

#GO 富集分析
sample.res.go <- goseq(sample.res.pwf, "mm10", "ensGene")
sample.res.go$BH <- p.adjust(sample.res.go$over_represented_pvalue,
                             method = "BH")
```

图 6. GOseq 代码

三、功效分析

功效分析可以帮助研究者了解既定实验在真实存在组间差异时可以统计检验出显著性差异的几率，从而有助于实验设计和对实验结果的解释。针对传统生物学实验的功效分析已被研究者普遍重视，尤其是样本量对功效的影响往往是研究者设计实验时所考虑的重点问题之一。然而，对于如 mRNA-Seq 等高通量实验，研究者往往只设置样本量为 3-5，满足允许进行统计检验的最低标准。原因之一是高通量实验价格昂贵；而以 Illumina 方法为代表的测序价格逐年下降，使得这一障碍变得越来越容易跨越。即使研究者受研究经费的限制无法通过提高样本量等方式达到理想的功效，进行功效分析也可以帮助正确地理解研究结果。下面从理论分析和实际数据分析两方面介绍功效分析；前

者有助于在实验前设计实验；后者对于预实验的分析有助于设计最终实验，而对最终实验的分析有助于理解实验结果。

1. 理论分析

对于以上文中利用 DESeq2 为代表的基于 read/fragment 数目服从负二项分布思想的 mRNA-Seq 差异表达分析，我们可以利用 RNASeqSampleSize 进行功效分析。对于一个既定的基因，在一个既定的实验中其被检验为差异表达的功效与如下几个因素相关：(1) 各样本，尤其是组间样本，测序深度的均一性（由于通常情况下各样本测序深度大致相当，我们在分析中通常将其设为 $w = 1$ ）；(2) 显著性水平（在此例子中我们设置 Adjusted P-Value 为 0.05，总计 15,000 个基因被表达而被统计检验，因此 Bonferroni 校正后的显著性水平 $\alpha = 3.3 \times 10^{-6}$ ，这里进行 Bonferroni 校正的原因是对单一基因的理论分析 FDR 校正方法不适用，实际 mRNA-Seq 差异表达分析中往往由于 Bonferroni 校正过于严格而使用 FDR 方法）；(3) 样本量 (n)；(4) 组间倍数改变 (fold change, ρ)；(5) 样本间 read/fragment 数目的平均值 (λ_0)；(6) dispersion (负二项分布的参数，对于由生物学差异或技术差异导致的样本间方差在考虑到 read/fragment 数目的影响的标准化数值， ϕ_0)。下面的 R 代码 (图 7) 给出了一个计算功效的具体例子，对大量参数组合的系统分析结果可见文章 (Xie *et al.*, 2020) 中图 2A 和补充表 S4。研究者可以通过提高样本量和测序深度 (read/fragment 数目的平均值)，以及控制生物学差异和技术差异，从而达到目标功效。读者在理论分析时可根据研究需求尝试参数选择，其中基因 read/fragment 数目的平均值，可以由先验的基因表达水平 (如 FPKM) 和基因长度，以及平均样本 read/fragment 总数反推；dispersion 最好依据目标样本的预实验结果确定 (见下面实际数据分析中描述)，或者利用其它类似样本的 mRNA-Seq 数据的估计值，7 种不同样本的小鼠 mRNA-Seq 数据的 dispersion 估计值见文章 (Xie *et al.*, 2020) 中图 3C 和补充表 S7。

```
library(RnaSeqSampleSize)

#影响功效的参数，数值仅代表此例子，供读者参考
w <- 1          #ratio of normalization factors
alpha <- 0.05 / 15000 #significance level
n <- 10         #sample size
rho <- 1.5      #fold change
lambda0 <- 512  #average read/fragment count
phi0 <- 0.008   #dispersion

#计算功效
est_power(n = n, w = w, rho = rho, lambda0 = lambda0,
          phi0 = phi0, alpha = alpha) #power is 0.996 in this example
```

图 7. RnaSeqSampleSize 代码

2. 实际数据分析

由于不同实验产生的样本基因表达谱和样本间表达差异不同，因此最好的策略是先用低样本量和低测序深度进行预实验。对预实验结果的分析可以得到各个基因的组间倍数改变 (fold change)、平均 read/fragment 数目和 dispersion 等参数，再利用这些参数进行功效分析以确定最终实验的样本量和测序深度。上述图 5 中的例子中，DESeq2 默认的输出结果中并不含有 dispersion 值，需用"dispersions"函数获取 (图 8)。

```
library(DESeq2)

#"sample.dds" is the last "sample.dds" in Figure 5.

#获取 dispersion 值
dispersions(sample.dds)
```

图 8. DESeq2 获取 dispersion 值代码

对于最终 mRNA-Seq 实验进行功效分析，有助于理解实验结果，即最终实验的样本量和测序深度是否有足够的功效以找到大部分真实差异表达的基因，以及是否

有足够的功效以找到主要的富集的功能条目 (Gene Ontology term 或 pathway)。例如文章 (Xie *et al.*, 2020) 中通过二次取样分析 (subsampling analysis) 发现, 对于基因 *A930004D18Rik* 敲除小鼠系, 胚胎期雌性敲除组 (12 个胚胎) 和野生组 (14 个胚胎) 的 mRNA-Seq 数据, 测序深度为平均 6 千万个 read (3 千万个 fragment) 足以找到大部分真实差异表达的基因 (文章 (Xie *et al.*, 2020) 图 2C); 但样本量为 10 不足以找到大部分真实差异表达的基因 (文章 (Xie *et al.*, 2020) 图 2B)。

四、分层分析

由于实验前未知原因或实验中无法控制的变量, 可能会导致得到的转录组数据出现分层的情况, 比如下例中转录组样本位于动情周期的不同阶段。针对这类情况进行合适的数据分析, 如主成分分析、聚类分析和标志基因分析, 可以厘清导致分层的原因。在此基础上可以先将所有样本进行分组, 并在各组中分别进行差异表达分析。本文使用文章 (Xie *et al.*, 2019) 中基因 *Gm13030/Shj* 敲除小鼠系, 成年雌性敲除组 (12 只小鼠) 和野生组 (12 只小鼠) 输卵管的 mRNA-Seq 数据为例说明分层分析。在最初的差异表达分析中, 没有发现任何统计显著的基因。其中一个可能的原因是成年雌性小鼠在动情周期不同阶段具有不同的表达谱和不同的基因差异表达现象。因此, 在差异表达分析前首先需要厘清动情周期各阶段。我们通过主成分分析 (本例中区分来自主成分 1) 和层次聚类分析将所有样本分为三组; 并通过对标志基因 (marker gene), 即孕激素受体和雌激素受体 (*Pgr*、*Esr1* 和 *Gper1*) 的表达水平分析, 确证了这三组对应于动情周期的三个不同阶段。主成分分析、层次聚类分析和标志基因分析方法见下面 R 代码 (图 9), 结果见文章 (Xie *et al.*, 2019) 中图 4。对这三个阶段的样本分别进行差异表达分析, 发现在其中的一个阶段有 21 个统计显著的差异表达基因, 但在另两个阶段没有任何差异表达基因。

```
library(DESeq2)

# "sample.info" is a "data.frame" with sample code ("code"), file name
# ("name"), and genotype ("genotype") information.
# "file.dir" is a folder with all the count files.

# 创建 DESeqDataSet
sample.table <- data.frame(sampleName = sample.info[, "code"],
                           fileName = sample.info[, "name"],
                           genotype = sample.info[, "genotype"])
sample.dds <- DESeqDataSetFromHTSeqCount(sampleTable = sample.table,
                                         directory = file.dir,
                                         design = ~ genotype)

# 标准化及"log2 scale"转化，此处为 variance stabilizing transformation
sample.vsd <- varianceStabilizingTransformation(sample.dds)

# 主成分分析
sample.PCA <- plotPCA(sample.vsd,
                     intgroup = c("genotype"),
                     ntop = length(sample.vsd),
                     returnData = T)
sample.PC1 <- sample.PCA[, c("PC1", "genotype", "name")]

# 层次聚类分析
sample.dists <- dist(t(assay(sample.vsd)))
sample.hc <- hclust(sample.dists)

# 标志基因分析
sample.dds <- DESeqDataSetFromHTSeqCount(sampleTable = sample.table,
                                         directory = file.dir,
                                         design = ~ 1)
sample.dds <- DESeq(sample.dds)
Pgr.count <- plotCounts(sample.dds, gene = "ENSMUSG00000031870",
                       intgroup = c("genotype"), returnData = T)
```

图 9. 主成分分析、聚类分析和标志基因分析代码

注：差异表达分析基于原始 *read/fragment* 数目，并在统计检验模型中统一处理各类因素；但主成分和聚类分析基于标准化及“log2 scale”转化后的 *read/fragment* 数目以消除样本间差异以及减小不同基因长度和表达水平间差异。

小结与建议

本文介绍了 mRNA-Seq 转录组差异表达分析目前被普遍应用的一般流程，以及介绍了功效分析和分层分析的方法。建议读者在 mRNA-Seq 转录组实验设计阶段进行初步功效分析。如果可能的话，在最终实验前安排低样本量和低测序深度的预实验，以进行功效分析找到最终实验的最佳方案；同时也初步分析是否存在样本分层现象。在得到最终实验数据后，优先分析样本是否存在分层现象，如果存在分层则要在区分各层后分别进行差异表达分析；此外，对最终实验的功效分析有助于对结果的理解。希望本文会有助于包括动物进化领域在内的生物学研究。

竞争性利益声明

作者声明不存在竞争性利益。

伦理学声明

本文无需伦理学声明。

致谢

感谢中国科学院动物研究所朱朝东教授邀请撰文。感谢本文所基于的两篇文章的其他作者的支持，尤其是 Max Planck Institute for Evolutionary Biology 的 Diethard Tautz 教授的支持。

参考文献

1. 赵婧, 刘南. (2019). [mRNA-seq 分析](#). Bio-101: e1010251. DOI: 10.21769/BioProtoc.1010251.
2. Anders, S and Huber, W. (2010). [Differential expression analysis for sequence count data](#). *Genome Biol* 11(10): R106.

3. Anders, S., Pyl, P. T. and Huber, W. (2015). [HTSeq--a Python framework to work with high-throughput sequencing data](#). *Bioinformatics* 31(2): 166-169.
4. Bolger, A. M., Lohse, M. and Usadel, B. (2014). [Trimmomatic: a flexible trimmer for Illumina sequence data](#). *Bioinformatics* 30(15): 2114-2120.
5. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). [STAR: ultrafast universal RNA-seq aligner](#). *Bioinformatics* 29(1): 15-21.
6. Kim, D., Paggi, J. M., Park, C., Bennett, C. and Salzberg, S. L. (2019). [Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype](#). *Nat Biotechnol* 37(8): 907-915.
7. Law, C. W., Chen, Y., Shi, W. and Smyth, G. K. (2014). [voom: Precision weights unlock linear model analysis tools for RNA-seq read counts](#). *Genome Biol* 15(2): R29.
8. Li, H. and Durbin, R. (2009). [Fast and accurate short read alignment with Burrows-Wheeler transform](#). *Bioinformatics* 25(14): 1754-1760.
9. Liao, Y., Smyth, G. K. and Shi, W. (2014). [featureCounts: an efficient general purpose program for assigning sequence reads to genomic features](#). *Bioinformatics* 30(7): 923-930.
10. Love, M. I., Huber, W. and Anders, S. (2014). [Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2](#). *Genome Biol* 15(12): 550.
11. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. and Gilad, Y. (2008). [RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays](#). *Genome Res* 18(9): 1509-1517.
12. Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T. and Salzberg, S. L. (2015). [StringTie enables improved reconstruction of a transcriptome from RNA-seq reads](#). *Nat Biotechnol* 33(3): 290-295.
13. Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010). [edgeR: a Bioconductor package for differential expression analysis of digital gene expression data](#). *Bioinformatics* 26(1): 139-140.
14. Xie, C., Bekpen, C., Kunzel, S., Keshavarz, M., Krebs-Wheaton, R., Skrabar, N., Ullrich, K. K. and Tautz, D. (2019). [A de novo evolved gene in the house mouse regulates female pregnancy cycles](#). *eLife* 8: e44392.

15. Xie, C., Bekpen, C., Kunzel, S., Keshavarz, M., Krebs-Wheaton, R., Skrabar, N., Ullrich, K. K., Zhang, W. and Tautz, D. (2020). [Dedicated transcriptomics combined with power analysis lead to functional understanding of genes with weak phenotypic changes in knockout lines](#). *PLoS Comput Biol* 16(11): e1008354.
16. Wang, L., Wang, S. and Li, W. (2012). [RSeQC: quality control of RNA-seq experiments](#). *Bioinformatics* 28(16): 2184-2185.
17. Young, M. D., Wakefield, M. J., Smyth, G. K. and Oshlack, A. (2010). [Gene ontology analysis for RNA-seq: accounting for selection bias](#). *Genome Biol* 11(2): R14.
18. Zhao, S., Li, C. I., Guo, Y., Sheng, Q. and Shyr, Y. (2018). [RnaSeqSampleSize: real data based sample size estimation for RNA sequencing](#). *BMC Bioinformatics* 19(1): 191.