

RNA-seq

数据分析标准报告

项目编号：

项目名称：RNA-seq分析(参考物种mm10)

项目日期：

上海达澈生物科技有限公司　监制

2020年11月

目 录

[第1章 前 言 1](#_Toc50386424)

[第2章 方法流程 2](#_Toc50386425)

**[2.1 文库构建方法](#_Toc50386426)** [2](#_Toc50386426)

**[2.2 Overview of RNA-seq technology](#_Toc50386427)** [2](#_Toc50386427)

**[2.3 数据分析流程](#_Toc50386428)** [2](#_Toc50386428)

**[2.4 Overview of RNA-seq Data analysis](#_Toc50386429)** [2](#_Toc50386429)

**[2.5 引用说明](#_Toc50386430)** [2](#_Toc50386430)

[第3章 结果与说明 3](#_Toc50386431)

**[3.1 材料方法](#_Toc50386432)** [3](#_Toc50386432)

**[3.2 数据质控](#_Toc50386433)** [3](#_Toc50386433)

**[3.3 RNA-seq主要结果展示](#_Toc50386434)** [4](#_Toc50386434)

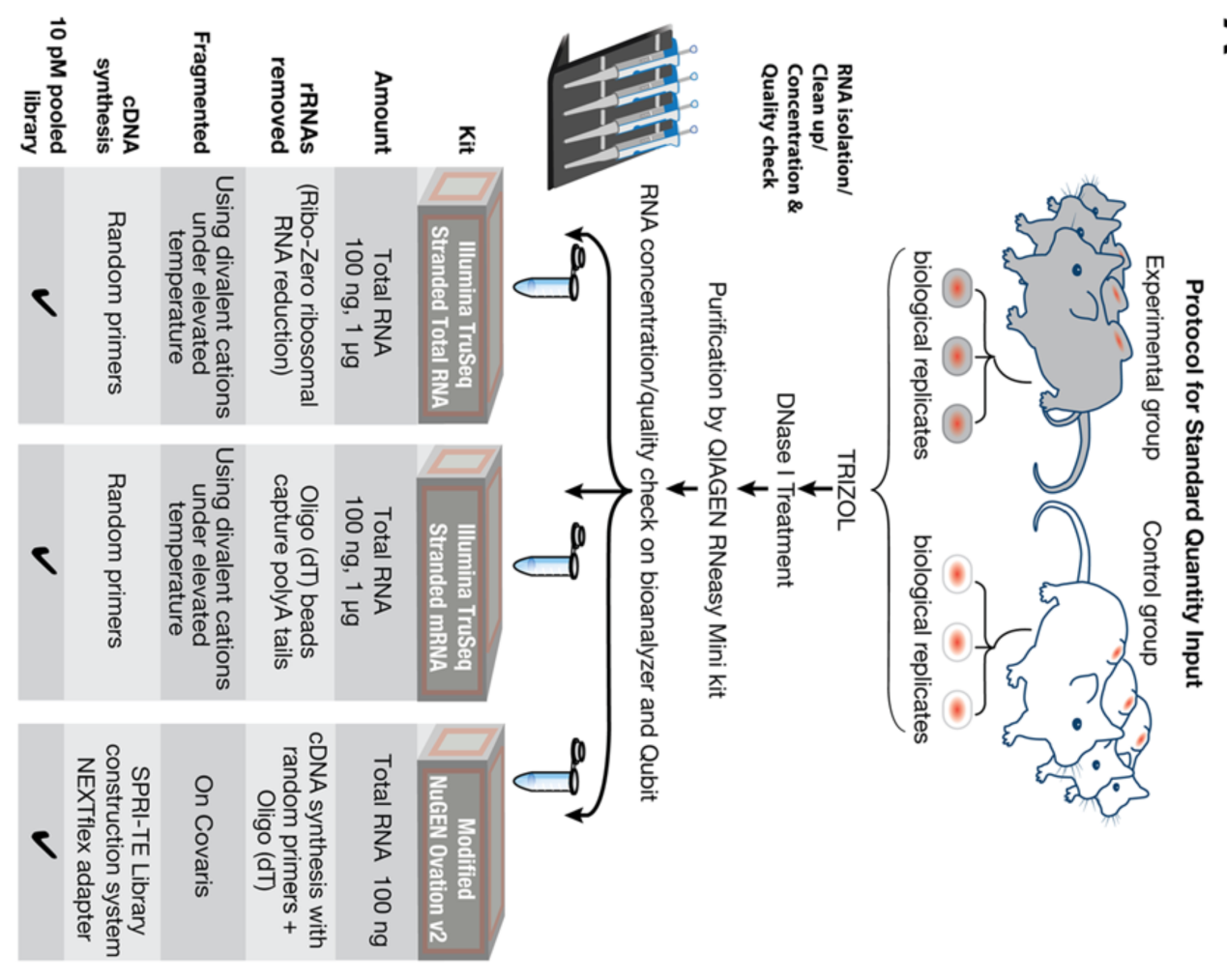
**[3.4 使用软件列表](#_Toc50386435)** [21](#_Toc50386435)

**[3.5 参考文献](#_Toc50386436)** [21](#_Toc50386436)

[关于 达澈 22](#_Toc50386437)

# **第1章 前 言**

RNA-Seq技术，全称为：RNA sequencing，是通过高通量测序手段，即二代测序技术（NGS），来检测某一特定时间点，生物样品中RNA的表达情况及定量的方法，可以用来研究分析细胞转录组在不同组织样品间基因表达差异情况、在时间轴上不同时间点连续的变化，以及转录本的可变剪切和SNP位点等。除了mRNA转录本信息，RNA-Seq技术可以用来研究不同类群的RNA，包括miRNA，tRNA，rRNA，lncRNA等(Chao et al., 2019)。RNA-Seq的最新进展还包括单细胞转录组学（single cell sequencing）和固定组织的原位测序（in situ sequencing of fixed tissue）。



# **第2章** 方法流程

**2.1 文库构建方法**

关于RNA-seq的文库构建，主要分为以下几步：

1. 抽提RNA；
2. 反转录cDNA；
3. 磁珠纯化cDNA
4. 末端补平；
5. 连接接头；
6. PCR扩增;
7. 磁珠纯化DNA；
8. 上机测序。

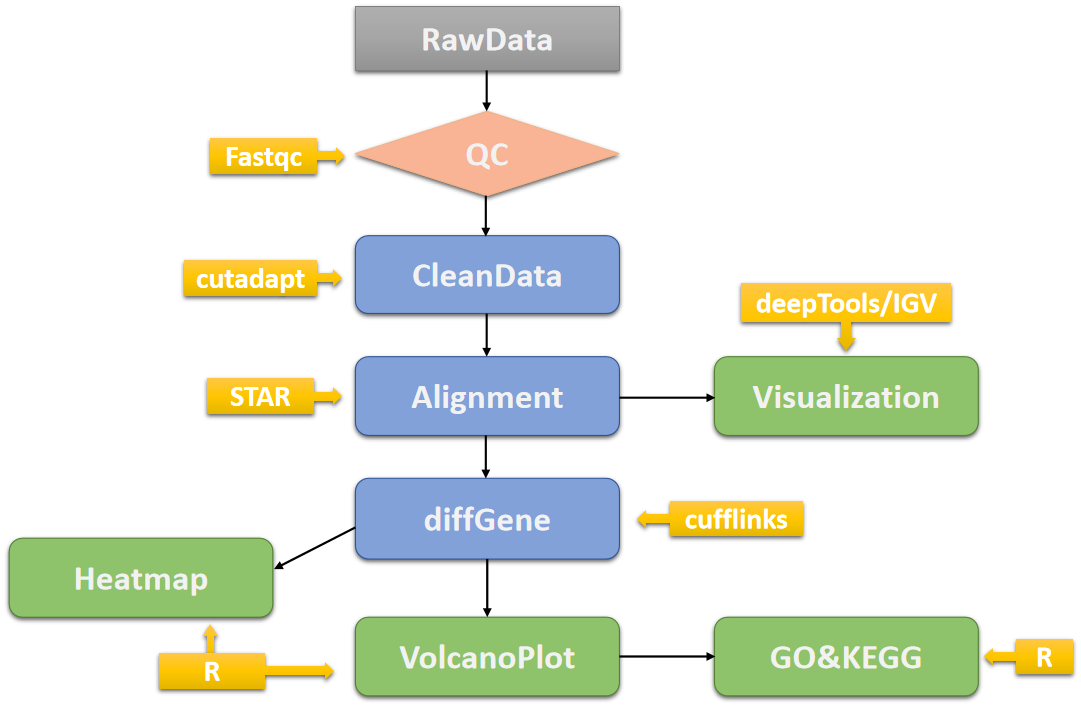
**2.2 Overview of RNA-seq technology**

Briefly, the process involves the following steps:

1. Total RNA extraction.
2. Perform first strand cDNA synthesis，Proceed immediately to second strand cDNA synthesis.
3. Purifiy double-stranded cDNA using DNA clean beads.
4. End prep of cDNA library.
5. Adaptor ligation.
6. Purifiy the Adaptor Ligated DNA Using Beads.
7. PCR Enrichment of Adaptor Ligated DNA.
8. Purification of the PCR reaction using beads.

**2.3 数据分析流程**

对于RNA-Seq数据，首先利用fastqc软件对原始数据进行质量评估；再使用STAR软件将原始数据比对到参考基因组；然后用cuffdiff软件分别对每例样本进行表达量计算，并对基因进行注释，差异基因分析（比较不同组之间的差别）。随后通过R包（如DEseq2，edgeR等），对差异基因数据进行进一步分析，并对靶基因进行GO和KEGG富集分析；



**2.4 Overview of RNA-seq Data analysis**

FastQC initially processed raw sequence reads for quality control, and then adapter sequences and poor quality reads were removed. Quality filtered reads were then mapped to reference genome using STAR, and only uniquely mapped reads were kept. Sam files were converted to Bam format using Samtools. Cufflinks softerware was used to computing gene expression level and difference. R packages such as DEseq2, edgeR were futher used to perform data.

**2.5 引用说明**

达澈预祝您的研究成果发表见刊，材料方法（METHODS）中数据分析部分引用时，请注明公司全称“上海达澈生物科技有限公司”，参考如下：

RNA-seq data analyses were performed by DIATRE Biotechnology, Shanghai, China.

***OR***

DIATRE Biotechnology, Shanghai, China, performed RNA-seq data analyses.

# **第3章** 结果与说明

**3.1 材料方法**

测序平台：Illumina Nova Seq 测序模式：PE150

样本属性：小鼠 参考物种：mm10

**3.2 数据质控**

测序产生的原始数据（raw data）需要进行预处理，我公司使用cutadapt过滤掉不合格的序列后得到有效数据（clean data），结果见下表3-1。

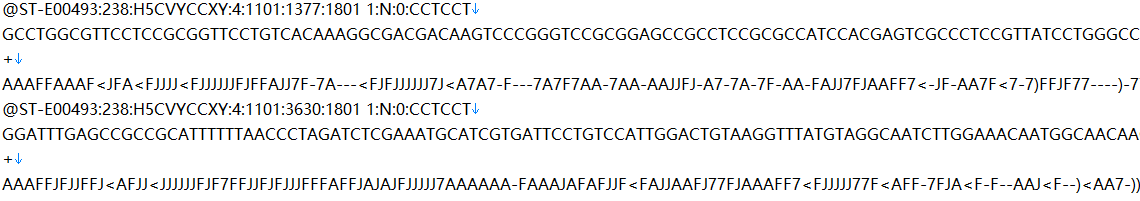
**表格 3-1 测序序列统计**

|  |  |  |
| --- | --- | --- |
| **Sample\_ID** | **Rawdata\_Reads** | **Cleandata\_Reads** |
| **M14-1** | 18441322 | 18421701 |
| **M14-2** | 25280190 | 25262904 |
| **M14-4** | 27157804 | 27143443 |
| **M14-5** | 25381065 | 25365910 |
| **M14-9** | 17883390 | 17871739 |
| **M14-10** | 27629565 | 27608473 |
| **M14-12** | 23982706 | 23970878 |
| **M14-13** | 25744366 | 25730037 |
| **M14-14** | 25664464 | 25652861 |
| **M14-15** | 25176886 | 25143678 |

**备注：**

1. Sample\_ID：样本名 R1、R2为双端原始数据。
2. Rawdata Reads：原始下机数据的reads数
3. Cleandata Reads：有效数据的reads数，去除adapter

测序结果以fastq格式储存，它包括序列的碱基组成信息以及其对应的序列质量信息，双端测序(pair-end)会分为两个Reads文件：\_R1,\_R2两个文件。



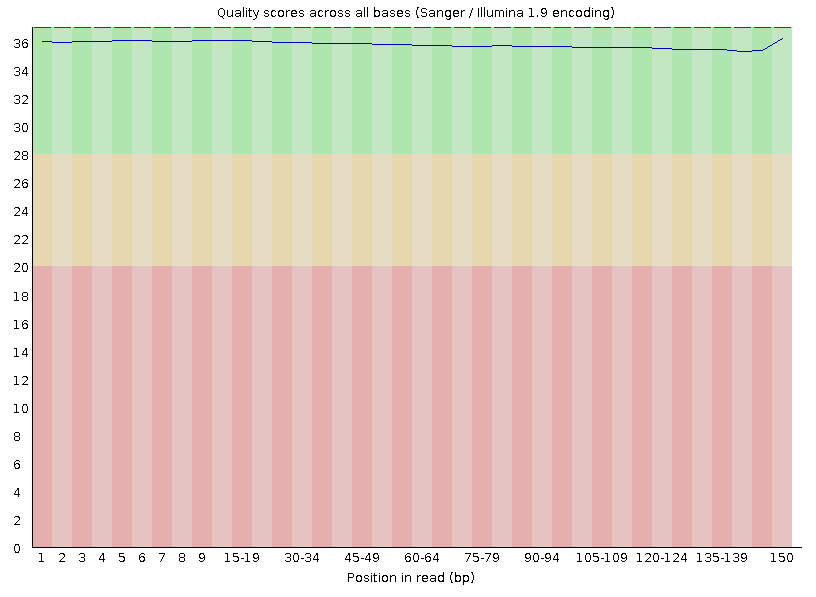
第一行：reads名称，以@开头，后为Illumina测序标识别符和描述信息；

第二行：碱基序列；

第三行：以+开头，存储与第一行相同的信息或为缺省值；

第四行：碱基的测序质量值，该行字符为第二行对应碱基的质量值加上33后转换为ASCII码，逆向转化即可直观得到每个碱基的质量信息。

我们使用fastqc软件对测序数据进行质控，输出的主要结果如图3-1。横坐标代表序列的碱基位置，纵坐标代表该碱基位置的质量分数分布，分数越高（绿色区域）表示测序质量越好，可靠性更高。详见附件“\phase1-fastqc”。

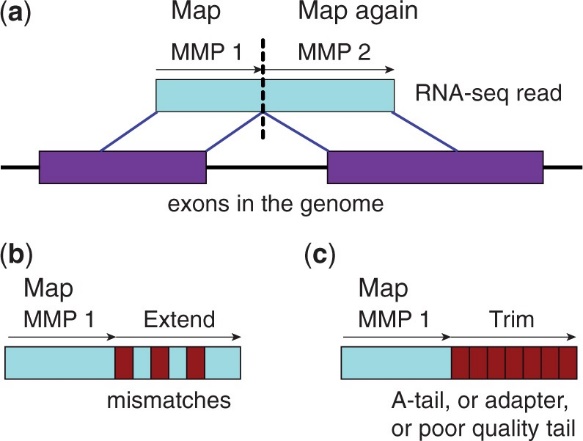


**图 3-1 Quality scores Per base sequence quality**

**3.3 RNA-seq主要结果展示**

**3.3.1 参考序列比对(Reads Mapping)**

将Reads与参考序列比对，并对比对情况进行统计，整体评估测序数据，评估参考基因组是否合适。Reads mapping主要检验特异性回帖比例（unique mapped reads），特异性回帖率越高代表数据质量越好。分析采用主流软件STAR，进行两步分析：将读长完整的回贴到参考基因组；将第一步不能回贴的RNA读长拆分，进行回贴。



**图 3-2 数据回帖分析流程**

**表3-2 参考序列比对结果**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sample** | **M14-1** | **M14-2** | **M14-4** | **M14-5** | **M14-9** |
| Number of input reads | 18421701 | 25262904 | 27143443 | 25365910 | 17871739 |
| Average input read length | 282 | 283 | 287 | 285 | 288 |
| UNIQUE READS: | | | | | |
| Uniquely mapped reads number | 14347214 | 18992890 | 20689644 | 19603425 | 14233023 |
| Uniquely mapped reads % | 77.88% | 75.18% | 76.22% | 77.28% | 79.64% |
| Average mapped length | 278.93 | 280.05 | 284.48 | 283.02 | 284.77 |
| Number of splices: Total | 14244772 | 17232504 | 18622428 | 13065461 | 12155439 |
| Sample | M14-10 | M14-12 | M14-13 | M14-14 | M14-15 |
| Number of input reads | 27608473 | 23970878 | 25730037 | 25652861 | 25143678 |
| Average input read length | 287 | 260 | 276 | 272 | 275 |
| UNIQUE READS: | | | | | |
| Uniquely mapped reads number | 20740891 | 19839793 | 21399875 | 20831422 | 18633219 |
| Uniquely mapped reads % | 75.13% | 82.77% | 83.17% | 81.21% | 74.11% |
| Average mapped length | 285.47 | 258.47 | 273.07 | 270.28 | 273.28 |
| Number of splices: Total | 17655147 | 14966179 | 19187941 | 16237723 | 12093382 |

**备注：**

1. Total reads：处理RNA-seq原始数据，参考序列比对分析中输入序列数目统计；
2. Mapped reads：成功比对至参考基因组的reads统计；
3. 所有基因回帖率完整列表见附件“\phase2-mapping\_stats”

**3.3.2 基因表达值提取**

根据回贴定位到基因区域的Reads数目来评估相应基因的表达水平。Reads数目直接反应了基因的表达量，因此需要统计基因上回贴到的Reads数。

**表3-3 基因表达水平统计**

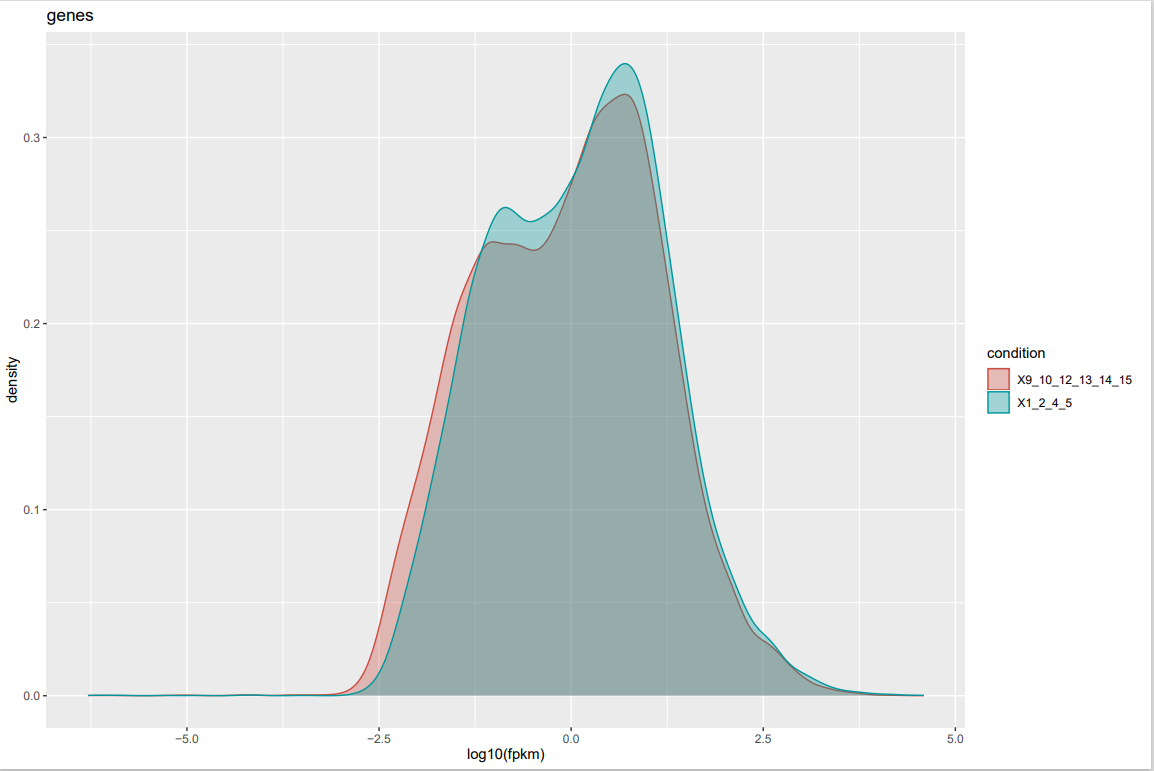
|  |  |  |  |
| --- | --- | --- | --- |
| **Gene ID** | **9** | **10** | **...** |
| ENSMUSG00000000001 | 18.3437 | 16.397 | ... |
| ENSMUSG00000000003 | 0 | 0 | ... |
| ENSMUSG00000000028 | 0.971116 | 0.768537 | ... |
| ENSMUSG00000000037 | 0 | 0 | ... |
| ENSMUSG00000000049 | 1549.68 | 1710.8 | ... |
| ENSMUSG00000000056 | 2.67954 | 2.98346 | ... |
| ENSMUSG00000000058 | 1.56735 | 0.667074 | ... |
| ... | ... | ... | ... |

**备注：**

1. 列为样本统计信息，行为基因编号，表格中样本所在列为基因在对应样本中的表达值；
2. Gene ID: 基因名称，默认格式为gene symbol ID；
3. 差异表达基因分析，选用Cufflinks软件；
4. 所有基因表达值，详见附件“phase3-Gene fpkm”

**3.3.3 基因表达水平分布分析**

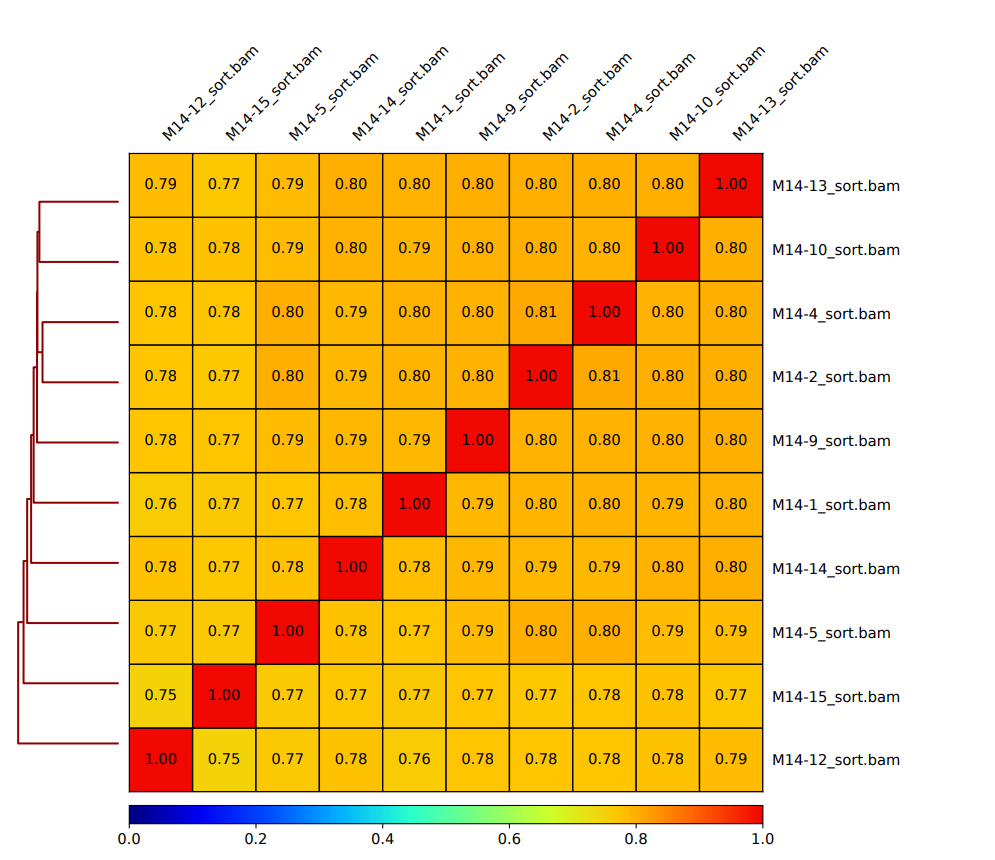
图3-3横坐标为log10（fpkm），纵坐标为基因丰度。统计发现，实验组与对照组大部分基因的表达分布基本相同。（FPKM，Fragments per kilobase of exon per million reads mapped）,分布图参见附件“\phase3-Gene fpkm”。



**图 3-3 基因表达水平分布图**

**3.3.4 Spearman相关系数分析(有重复组才能做)**

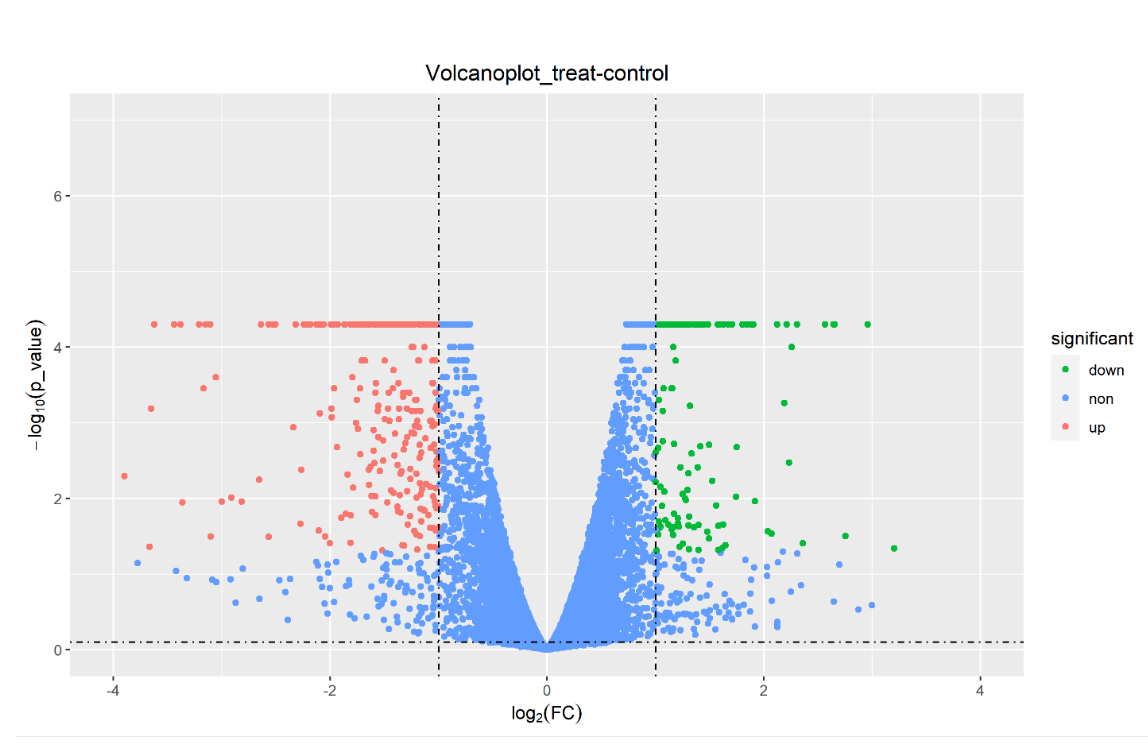
相关性系数（Spearman correlation coefficient）是用来反映两个变量线性相关程度的统计量。相关系数的绝对值越大，相关性越强：相关系数越接近于1或-1，相关度越强，相关系数越接近于0，相关度越弱。分布图可参见附件“\phase3-Gene fpkm”。



**图 3-4 组间spearman相关系数**

**3.3.5 基因表达变化可视化**

两组间基因变化分三种：上调、下调、无明显变化，可将回帖所得的总体转录组数据进行可视化。

**图 3-5 基因附近的平均信号分布**

如

图3-5所示，图中纵坐标为-log10（p value），用于展示p值。，蓝色点聚集的区域表示无统计学差异，红色部分表示有统计学差异；图中横坐标为log2（fold change），用于展示基因表达量的情况，0点左侧负值区域表示相对于对照，实验组中下调的基因分布，，右侧正值区域表示表达上调的基因分布情况。详细的分布图参见附件“\phase5-heatmap&volcanoplot”。

**3.3.6 筛选差异表达基因**

获得基因表达值后，筛选差异基因是整个RNA-seq分析的关键。目前针对RNA-seq数据，进行差异基因的筛选软件较多，如edgeR、Limma、DEseq等。

软件的选择和参数的调整均会对结果产生显著影响，因此本节分析的软件选取和参数设置，会根据客户样品属性，数据特点，选择合适的软件进行分析。

差异表达基因结果：

**表3-4 上调差异基因统计**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gene\_ID | Gene | 9\_10\_12\_13\_14\_15 | 1\_2\_4\_5 | Log2FC |
| ENSMUSG00000000078 | Klf6 | 4.62086 | 1.55344 | -1.57269 |
| ENSMUSG00000000303 | Cdh1 | 9.62333 | 4.56747 | -1.07514 |
| ENSMUSG00000001403 | Ube2c | 2.50404 | 0.859792 | -1.5422 |
| ENSMUSG00000001506 | Col1a1 | 4.49423 | 1.0541 | -2.09206 |
| ENSMUSG00000002076 | Hsf2bp | 1.44904 | 0.230035 | -2.65517 |
| ENSMUSG00000002257 | Def6 | 1.31092 | 0.48382 | -1.43804 |
| ... | ... | ... | ... | ... |

**备注：**表达值为FPKM值。

**表3-5 下调差异基因统计**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| gene\_id | gene | 9\_10\_12\_13\_14\_15 | 1\_2\_4\_5 | Log2FC |
| ENSMUSG00000001763 | Tspan33 | 5.44682 | 11.0711 | 1.02331 |
| ENSMUSG00000002265 | Peg3 | 1.99366 | 5.94445 | 1.57612 |
| ENSMUSG00000002324 | Rec8 | 0.848075 | 2.13979 | 1.3352 |
| ENSMUSG00000005373 | Mlxipl | 74.6797 | 164.605 | 1.14022 |
| ENSMUSG00000007617 | Homer1 | 1.96549 | 4.55381 | 1.21218 |
| ... | ... | ... | ... | ... |

**备注：**

1. 差异表达基因筛选条件：cutoff: p-value < 0.05；
2. 完整列表见附件“phase6-diffGene\_up&down”
3. 表达值为FPKM值。

**3.3.7 差异表达基因结果统计**

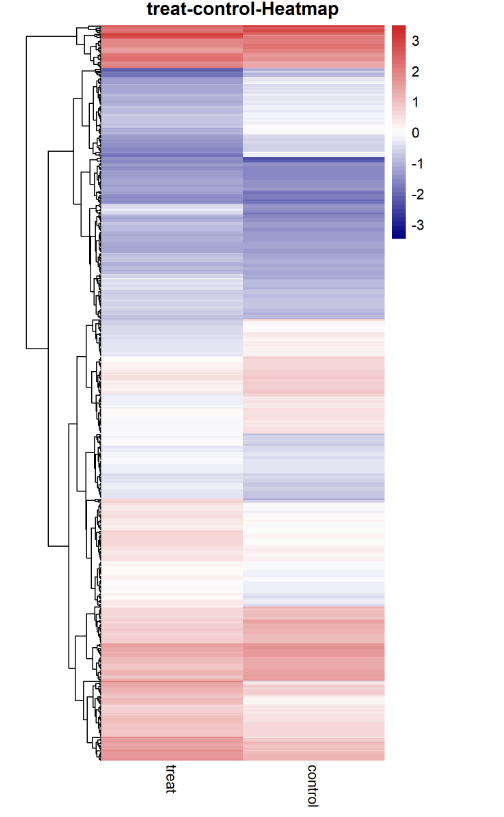
根据实验设计，与对照组相比，找出实验组中表达具有显著性差异基因的数据，并进行统计，其中上调基因288个，下调基因129个，如图3-6.

**图 3-6 差异表达基因数目统计**

**备注：**

1. 分组比较方式为“Treat vs Control”；
2. 根据表达量变化，分为两类：Up（上调基因）、Down（下调基因）；
3. 筛选参数的q值 < 0.05 ， |log2FoldChange|>1，即2倍FPKM差异的基因作为显著差异基因；
4. 原始文档见附件“\phase6-diffGene\_up&down”

**3.3.8 差异基因Heatmap**



**图 3-7 差异表达基因结果热图**

**备注：**

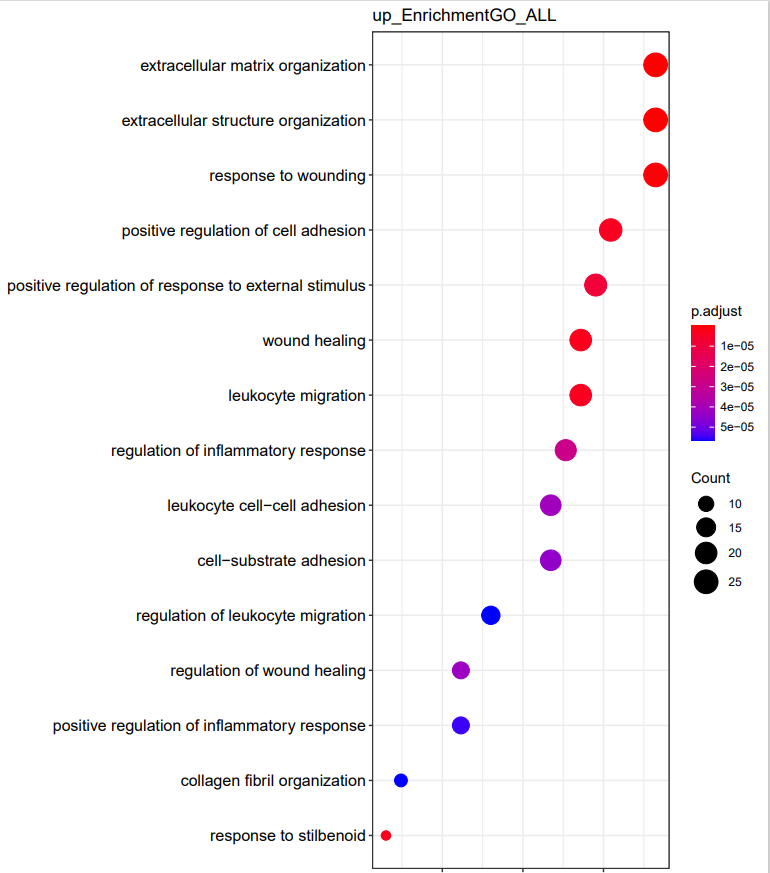
1. 差异基因热图，用于反映两组样品间基因的表达差异程度；
2. 筛选参数：q < 0.05 ，|log2FoldChange| > 1 （即2倍差异）；
3. 原始文档见附件“\phase5-heatmap&volcanoplot”

**3.3.9 差异表达GO富集分析**

GO（Gene ontology）是国际通用的基因功能分类体系，按照基因的细胞成分（cellular component），分子功能（molecule function）以及生物过程（biological process）分为3类。在得到差异基因的集合后，进行GO富集能够看到不同样本在这三大类中的基因分布情况，也可以用于对目标基因的聚集。通过计算数据集的超几何分布，得出基因富集对应GO terms的排序。具体结果详见附件“\9.GO&KEGG”。

**表 3-6 上调差异基因的GO富集分析**

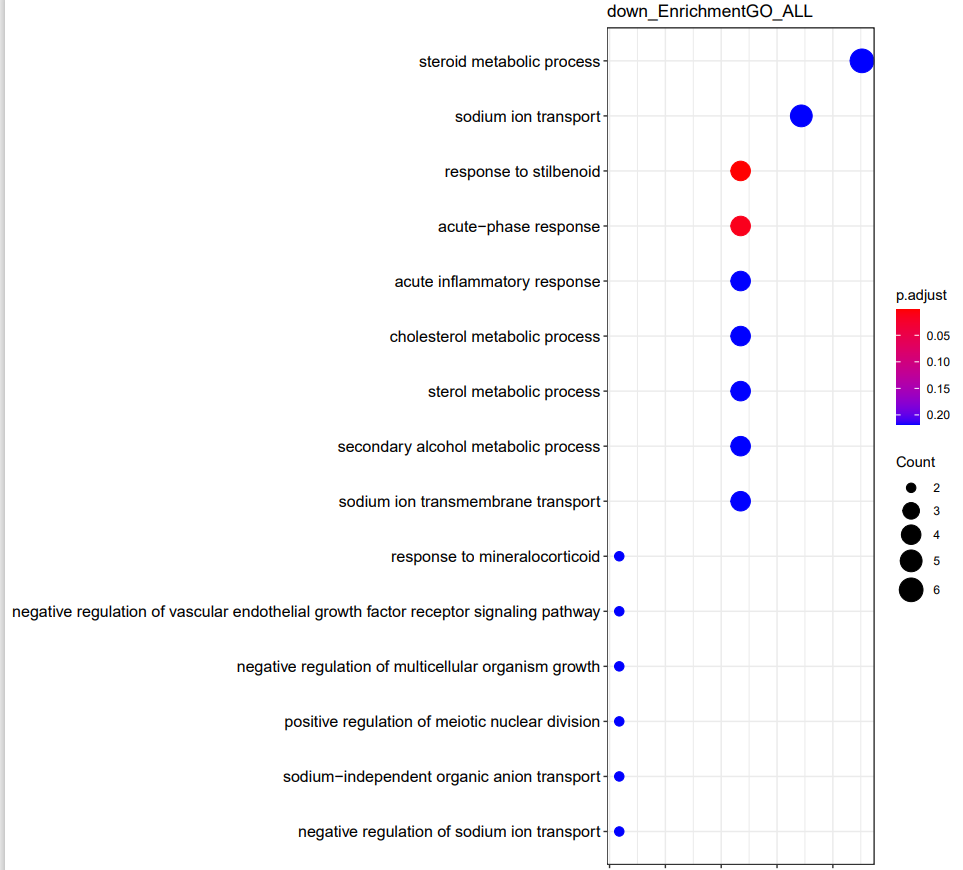
|  |  |  |
| --- | --- | --- |
| GO term | Description | Count |
| GO:0030198 | extracellular matrix organization | 25 |
| GO:0043062 | extracellular structure organization | 25 |
| GO:0009611 | response to wounding | 25 |
| GO:0035634 | response to stilbenoid | 7 |
| GO:0042060 | wound healing | 20 |
| ... | ... | ... |



**图 3-8 上调差异基因的GO富集分析**

**表 3-7 下调差异基因的GO富集分析**

|  |  |  |
| --- | --- | --- |
| GO term | Description | Count |
| GO:0035634 | response to stilbenoid | 4 |
| GO:0006953 | acute-phase response | 4 |
| GO:0051385 | response to mineralocorticoid | 2 |
| GO:0002526 | acute inflammatory response | 4 |
| GO:0030948 | negative regulation of vascular endothelial growth factor receptor signaling pathway | 2 |
| ... | ... | ... |

**图 3-9 下调差异基因的GO富集分析**

**备注：**

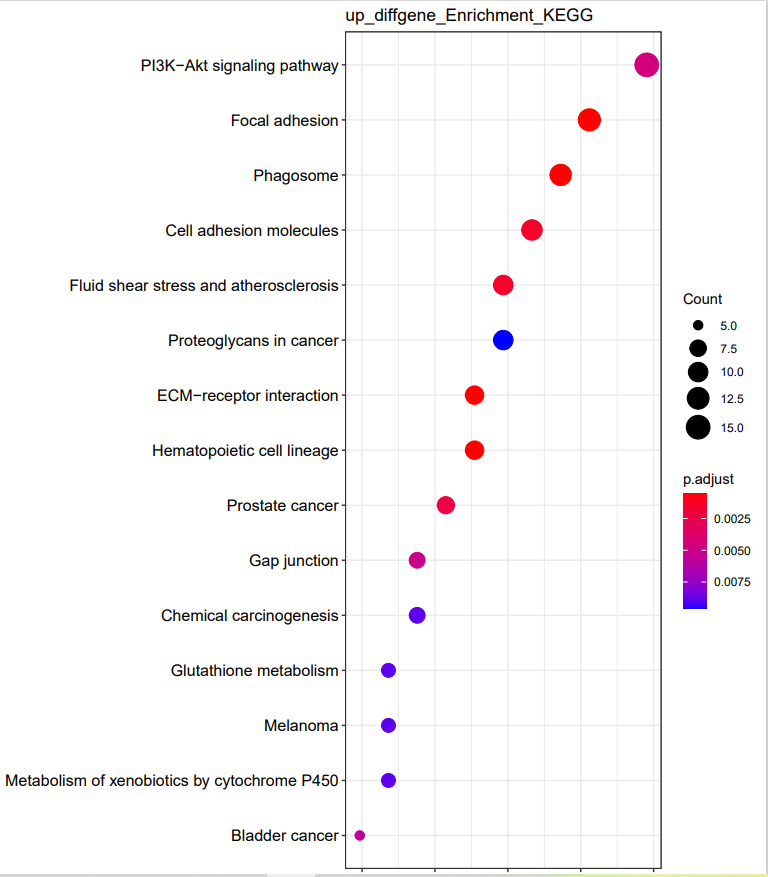
1. GOBPID: gene ontology biological process ID，GO数据库中生物学过程分类编号信息；
2. P value：富集分析显著性统计参数；
3. Count: 横坐标为基因Count数，即差异表达基因中隶属于该生物学过程的基因个数；
4. Description：生物学过程名称；
5. GO富集分析按照显著性统计参数pvalue排序；
6. GO富集分析显著性筛选阀值：cutoff: pvalue < 0.05
7. 上、下调差异基因GO富集分析完整列表见附件“phase7-GO&KEGG”

**3.3.10 差异基因的KEGG pathway富集**

研究发现不同基因间存在相互作用的关系，这些基因相互协调，发挥各自的生物学功能。针对该现象，科学家们分门别类的制定出各种pathway，较为著名是KEGG（Kyoto Encyclopedia of Genes and Genomes）数据库。利用pathway显著性富集，确定所属的pathway，进而找出可能相互作用的基因，挖掘现象背后的分子机制。

**表 3-8 上调差异基因的KEGG富集分析表**

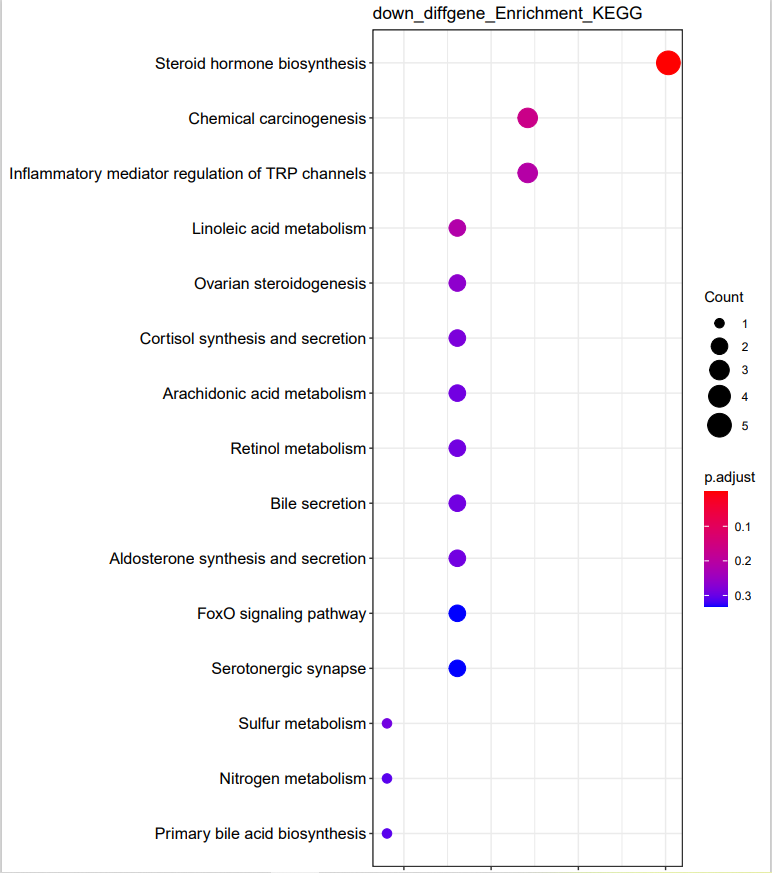
|  |  |  |
| --- | --- | --- |
| Pathway accession | Description | Count |
| mmu04512 | ECM-receptor interaction | 9 |
| mmu04510 | Focal adhesion | 13 |
| mmu04640 | Hematopoietic cell lineage | 9 |
| mmu04145 | Phagosome | 12 |
| mmu04514 | Cell adhesion molecules | 11 |
| ... | ... | ... |



**图 3-10 上调差异基因的KEGG富集分析图**

**表 3-9 下调差异基因的KEGG富集分析表**

|  |  |  |
| --- | --- | --- |
| Pathway accession | Description | Count |
| mmu00140 | Steroid hormone biosynthesis | 5 |
| mmu05204 | Chemical carcinogenesis | 3 |
| mmu04750 | Inflammatory mediator regulation of TRP channels | 3 |
| mmu00591 | Linoleic acid metabolism | 2 |
| mmu04913 | Ovarian steroidogenesis | 2 |
| ... | ... | ... |



**图 3-11 下调差异基因的KEGG富集分析图**

**备注：**

1. KEGGID: KEGG pathway ID，KEGG数据库中生物学通路编号信息；
2. pvalue: 富集分析显著性统计参数；
3. GeneRatio: 差异表达基因中隶属于该生物学通路的基因统计比例；
4. Description：生物学通路名称；
5. KEGG pathway富集分析按照显著性统计参数pvalue排序；
6. KEGG pathway富集分析显著性筛选阀值：cutoff: pvalue < 0.05；
7. 上、下调差异基因KEGG pathway富集分析完整列表，见附件“\phase7-GO&KEGG”

**3.4 使用软件列表**

**表格 3-10 RNA-seq数据分析使用软件列表**

|  |  |  |
| --- | --- | --- |
| **Software** | **Version** | **Main function** |
| **Fastqc** | 0.11.8 | Quality control |
| **STAR** | 1.5.1 | Alignment |
| **Samtools** | 1.8 | Bam files |
| **Cufflinks** | 2.17.0 | Diff files |
| **R** | 4.0 | Data analysis |

**3.5 参考文献**

Chao, H. P., Chen, Y., Takata, Y., Tomida, M. W., Lin, K., Kirk, J. S., . . . Shen, J. (2019). Systematic evaluation of RNA-Seq preparation protocol performance. BMC Genomics, 20(1), 571. doi:10.1186/s12864-019-5953-1

# 关于 达澈

上海达澈生物科技有限公司，专注提供领先的科研产品和技术服务，致力于成为全球领先的表观组学产品和服务的提供者。

我们的NGS产品线覆盖如下：

基因组：Circle-seq（eccDNA），WGBS，Hi-C（三维基因组）；

转录组：miRNA、mRNA、lncRNA、circRNA测序；

表观组：CUT&Tag，ATAC-seq，ChIP-seq；

RNA互作组：RIC-seq；

蛋白和RNA互作：RIP-seq和CLIP-seq；

翻译组：Ribo-seq；

RNA修饰类：m1A，m6A，m7G，m5C，ac4C，Ψ假尿嘧啶测序、

2'-O-RNA测序

关注我们的公众号，更多精彩内容等待您的发掘！

