

# 美吉生物结题报告

医学版有参分析项目

客户： 李蒙

2025 年 01 月 13 日

# 目录

一 项目信息

二 工作流程

- 2.1 测序实验流程
- 2.2 信息分析流程

三 分析报告

- 3.1 项目结果速览
- 3.2 差异基因数据挖掘
  - 3.2.1 表达差异统计
  - 3.2.2 差异基因Venn分析
  - 3.2.3 差异基因聚类分析
  - 3.2.4 差异基因功能注释分析
    - 3.2.4.1 GO注释
    - 3.2.4.2 KEGG注释
    - 3.2.4.3 Reactome注释
  - 3.2.5 差异基因功能富集分析
    - 3.2.5.1 GO富集分析
    - 3.2.5.2 KEGG富集分析
    - 3.2.5.3 Reactome富集分析
- 3.3 所有基因数据挖掘
  - 3.3.1 功能注释与表达量查询
  - 3.3.2 样本间相关性分析
  - 3.3.3 样本间PCA分析
  - 3.3.4 样本间Venn分析
- 3.4 基因结构数据挖掘
  - 3.4.1 差异可变剪切
    - 3.4.1.1 差异可变剪切事件统计
- 3.5 基础分析
  - 3.5.1 测序数据质控
  - 3.5.2 序列比对分析
    - 3.5.2.1 比对结果统计
    - 3.5.2.2 转录组质量评估
  - 3.5.3 表达量分布
- 3.6 项目背景
  - 3.6.1 基本信息
    - 3.6.1.1 项目样本信息
    - 3.6.1.2 分析软件信息

四 备注

参考文献

一、项目信息

项目名称			
医学版有参转录组分析项目			
合同编号			
MJ20241210177			
项目样本信息			
样本来源	小鼠		
样本类型	脊髓		
备注			
客户信息			
单位名称	华中农业大学		
单位地址	湖北省武汉市洪山区狮子山街1号		
课题组	丁一	电话	18613711600
		邮箱	86556439@qq.com
项目联系人	李蒙	电话	18613711600
		邮箱	86556439@qq.com
美吉联系人信息			
销售员	贾含笑	电话	18538949930
		邮箱	hanxiao.jia@majorbio.com
技术支持	虞阿巧	电话	021-20725056
		邮箱	rna@majorbio.com
项目总监审批			
<div>签名：_____</div> <div>____年__月__日</div>			

## 二、工作流程

### 2.1测序实验流程

真核mRNA测序是基于二代高通量测序平台，对真核生物特定组织或细胞在某个时期转录出来的所有mRNA进行测序，文库构建流程图如下图所示：



#### ① 提取total RNA

从组织样品中提取total RNA，利用Nanodrop2000对所提RNA的浓度和纯度进行检测，琼脂糖凝胶电泳检测RNA完整性，Agilent5300测定RQN值。单次建库要求RNA总量1 μg,浓度≥ 30 ng/μL, RQN > 6.5, OD260/280介于1.8~2.2之间。

#### ② Oligo dT富集mRNA

真核生物mRNA 3'末端具有polyA尾的结构，利用带有Oligo(dT)的磁珠与polyA进行A-T 碱基配对，可以从总RNA中分离出mRNA，用于分析转录组信息。

#### ③ 片段化mRNA

二代高通量测序平台是针对短序列片段进行测序，富集得到的mRNA是完整的RNA序列，平均长度达几kb，因此需要对其进行随机打断。加入fragmentation buffer，选择合适条件，可以将mRNA随机断裂成300bp左右的小片段。

#### ④ 反转合成cDNA

在逆转录酶的作用下，利用随机引物，以mRNA为模板反转合成一链cDNA，随后进行二链合成，形成稳定的双链结构。

#### ⑤ 连接adapter

双链的cDNA结构为粘性末端，加入End Repair Mix将其补成平末端，随后在3'末端加上一个A碱基，方便后面加入adapter序列。

#### ⑥ 片段筛选和文库富集

对连接adapter后的产物进行纯化和片段分选，用分选产物进行PCR扩增，纯化得到最终的文库。

#### ⑦ NovaSeq X Plus平台上机测序（根据实际情况选择）

- 1) Qubit 4.0定量，按数据比例混合上机；
- 2) cBot上进行桥式PCR扩增，生成clusters；
- 3) 上机测序。

#### ⑦ DNBSEQ-T7 平台上机测序（根据实际情况选择）

- 1) 文库DNA单链环化；
- 2) 去除未环化序列，并进行纯化；
- 3) 制备DNA纳米球（DNA nanoball, DNB）；
- 4) 上机测序。

### 2.2信息分析流程

RNA-seq的核心是基因表达差异的显著性分析，使用统计学方法，比较两个条件或多个条件下的基因表达差异，从中找出与条件相关的特异性基因，然后进一步分析这些特异性基因的生物学意义，分析过程包括质控，比对，定量，差异显著性分析，功能富集六个环节，如下图所示。另外可变剪接，变异位点也是RNA-seq的重要分析内容。同时，根据不同的研究需求，推出转录组个性化分析内容，如基因共表达网络构建（WGCNA）、基因集富集分析（GSEA）、蛋白互作网络分析等。信息分析流程如下图所示：



在医学有参转录组的研究中，常规的研究思路为：

#### 1) 疾病 vs 对照

对于典型“2×n”类的实验设计，即2种处理：疾病组+对照组，我们通常可利用差异分析获得两组之间的差异基因，这些基因也是后续开展数据分析及分子实验研究的重点关注对象。可以从差异分析入手，再到功能富集分析及基因结构分析进行数据的深入挖掘。

#### 2) 梯度处理

若实验设计有时间梯度或浓度梯度上的变化，那么可利用趋势分析将上千个 DEGs 进行模块的划分，研究表达量变化趋势。趋势分析是梯度设计类文章的核心分析点。对于多组梯度的样本，利用趋势分析对基因进行归类，有助于我们快速找到目标基因和目标通路，迅速对结果进行解读。

#### 3) 大样本型

若涉及大样本量取样，可以考虑进行加权基因共表达网络分析（Weighted Gene Co-expression Network Analysis, WGCNA）分析。WGCNA工具可以助力我们从大样本的转录组数据中挖掘与研究目的相关的候选基因或挖掘与表型数据相关的候选基因，广泛应用于医学的各个研究领域，也可发掘新的调控关系网络，从整体视角的分析思路去解决更多生物学问题。WGCNA还可以构建基因共表达网络，找出位于基因网络中心的核心基因（hub gene），往往更具有生物学意义。

## 三、分析报告

### 温馨提示：

亲爱的老师您好，欢迎并感谢您选择医学有参转录组测序服务。您的数据已经全部分析完成，现在您可以对测序数据进行进一步的分析和挖掘。请您耐心阅读报告，我们会用一些特殊标注提醒您每个模块的重点分析内容，助力您全面快速获取您关注的特定组织或细胞在某个特定状态下转录的所有mRNA序列和丰度信息。同时，云平台客户中心包含医学有参转录组测序报告的视频解读和云平台操作指南，如有需要可前往查看。

1) 医学有参转录组测序报告的视频解读链接：<https://edu.majorbio.com/course/lesson/3537>;

2) 云平台操作指南：<https://www.majorbio.com/wenku/info/1366>;

3) 中英文写作方法材料：<https://www.majorbio.com/wenku/info/1529>。

如您想要获取更多组学信息，请及时与我们联系！

### 3.1 项目结果速览

1、测序数据统计：完成12个样品的转录组测序，共获得87.27 Gb Clean Data，各样品Clean Data均达到5.93 Gb以上，Q30碱基百分比在96.96%以上；

2、与参考基因组比对：分别将各样品的Clean Reads与指定的参考基因组进行序列比对，比对率从99.11%到99.29%不等；

3、表达量分析：本次分析共检测到表达基因共32131个；表达转录本共99730个；

4、表达量差异分析：基于表达量定量结果，进行组间差异基因分析，获得两组间发生差异表达的基因，差异分析软件为：DESeq2，筛选阈值为： $|\log_2FC| \geq 1$  &  $p\text{-adjust} < 0.05$ ，具体结果如下：

diff_group	total DE	up	down
CFA_vs_SAL	425	407	18
DEX_EA_vs_CFA	65	18	47
DEX_EA_vs_SAL	15	11	4

☆Tips：这部分内容比较重要，老师通过这部分内容快速了解本次项目的数据基本情况以及差异组别获得的差异基因数目等。

### 3.2 差异基因数据挖掘

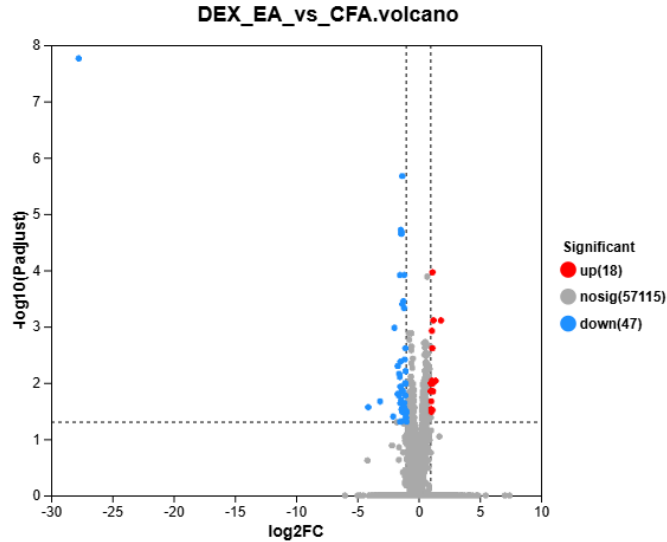
#### 3.2.1 表达差异统计

医学应用：筛选比较组间的 **显著差异基因/转录本**，进而研究差异基因/转录本相关功能。

☆Tips:

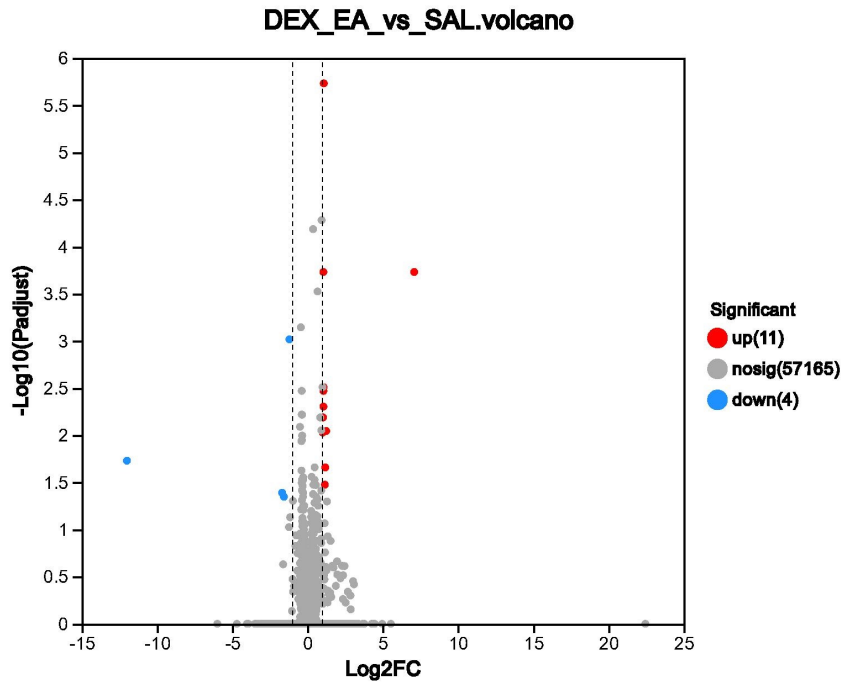
- ①除了筛选得到的显著差异基因，老师也可通过表格筛选功能查找关注的目标基因，进而创建个性化的基因集，在【目标基因集分析模块】进行功能注释、富集、GSEA、PPI网络等个性化分析；
- ②显著差异且在样本间稳定高表达的基因，很有可能会参与调控信号通路，行使极其重要的生物学功能，亦或是疾病的 biomarkers和预后指标；
- ③差异统计图、火山图、MA图等为文献中的常用出图，直观展示组间差异情况，老师们可按需下载。

表达量差异火山图



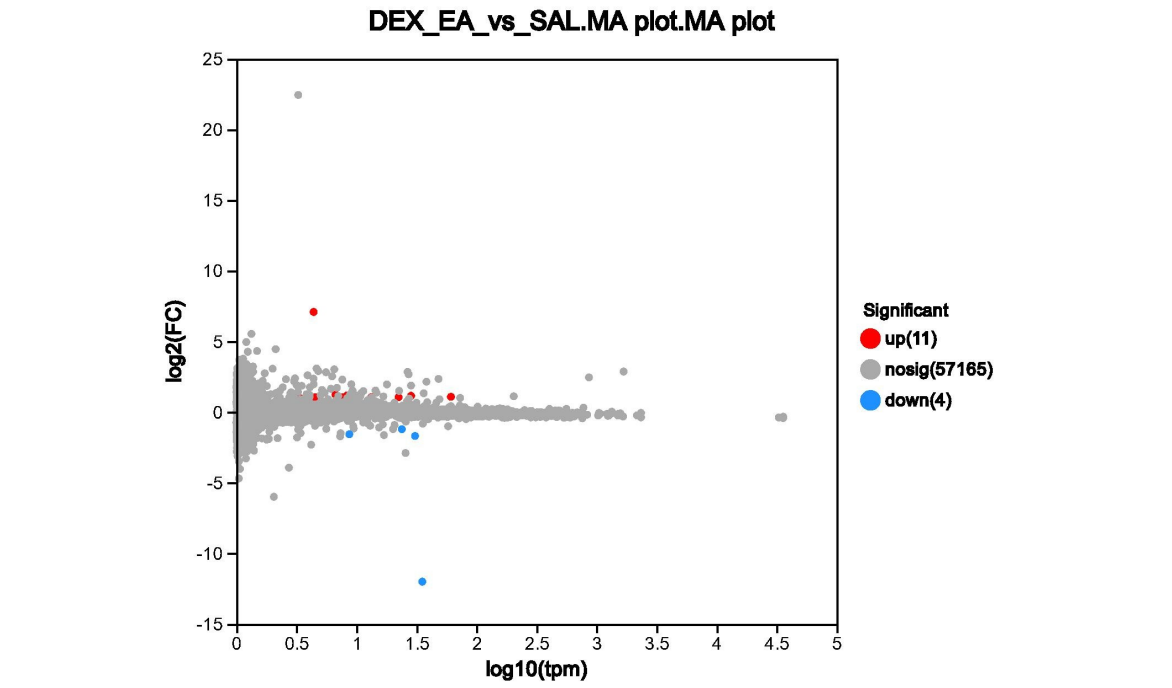
注：横坐标为基因/转录本在两组别间表达差异的倍数变化值，即处理组的表达量除以对照组的表达量得到的数值；纵坐标为基因/转录本表达量变化差异的统计学检验值，即Padjust。P值越小则表达差异越显著，横纵坐标的数值均进行了对数化处理。图中每个点代表一个特定的基因/转录本，默认红色点表示显著上调的基因/转录本，蓝色点表示显著下调的基因/转录本，灰色点为非显著差异基因/转录本。将所有基因/转录本映射上去之后，可以获知，在左边的点为表达差异下调的基因/转录本，右边的点为表达差异上调的基因/转录本，越靠两边和上边的点表达差异越显著。

表达量差异火山图



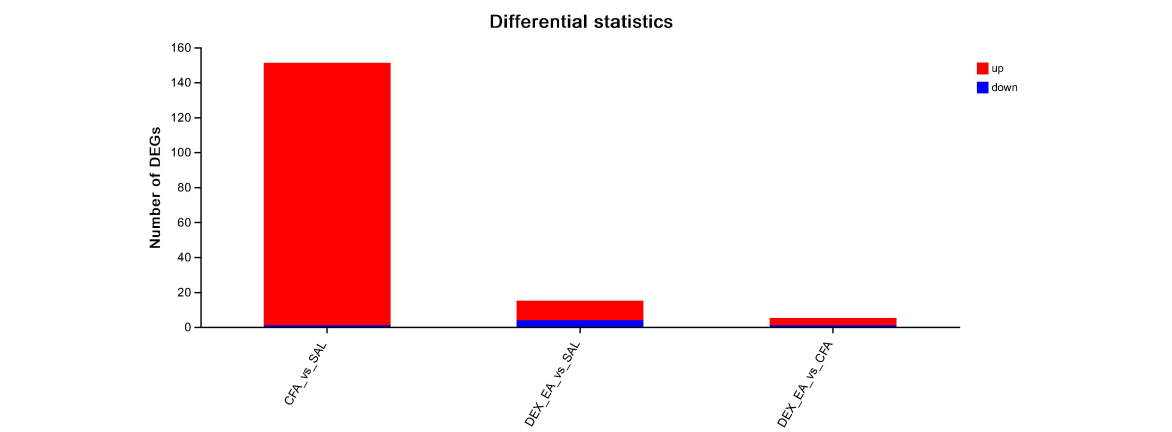
注：横坐标为基因/转录本在两组别间表达差异的倍数变化值，即处理组的表达量除以对照组的表达量得到的数值；纵坐标为基因/转录本表达量变化差异的统计学检验值，即Padjust。P值越小则表达差异越显著，横纵坐标的数值均进行了对数化处理。图中每个点代表一个特定的基因/转录本，默认红色点表示显著上调的基因/转录本，蓝色点表示显著下调的基因/转录本，灰色点为非显著差异基因/转录本。将所有基因/转录本映射上去之后，可以获知，在左边的点为表达差异下调的基因/转录本，右边的点为表达差异上调的基因/转录本，越靠两边和上边的点表达差异越显著。

表达量差异MA图



注：横坐标为利用差异对比组的表达量值进行计算，纵坐标为差异表达分析获得的差异对比组之间基因表达变化，横纵坐标的数值均进行了对数化处理。图中每个点代表一个特定的基因/转录本，默认红色点表示显著上调的基因/转录本，蓝色点表示显著下调的基因/转录本，灰色点为非显著差异基因/转录本。将所有基因/转录本映射上去之后，可以获知，在上边的点为表达差异上调的基因/转录本，下边的点为表达差异下调的基因/转录本，越靠上边和下边的点表达差异越显著。

表达量差异统计图-堆积图



注：横坐标代表不同的差异比较组别，纵坐标代表对应的上下调基因/转录本数目。默认红色代表上调，绿色代表下调。

差异基因数目统计表

diff_group	total DEG	up	down
CFA_vs_SAL	425	407	18
DEX_EA_vs_CFA	65	18	47
DEX_EA_vs_SAL	15	11	4

注：（1）diff\_group：差异比较组别；（2）total DEG/DET：差异基因/转录本数目；（3）up：上调差异基因/转录本数目；（4）down：下调差异基因/转录本数目。

### 3.2.2 差异基因Venn分析

医学应用：Venn分析可以进□组间差异基因的联合筛选，展示了 **各基因集之间的重叠关系**，可以选择感兴趣的区域基因（**共性、特性**）构建基因集合进行分析。

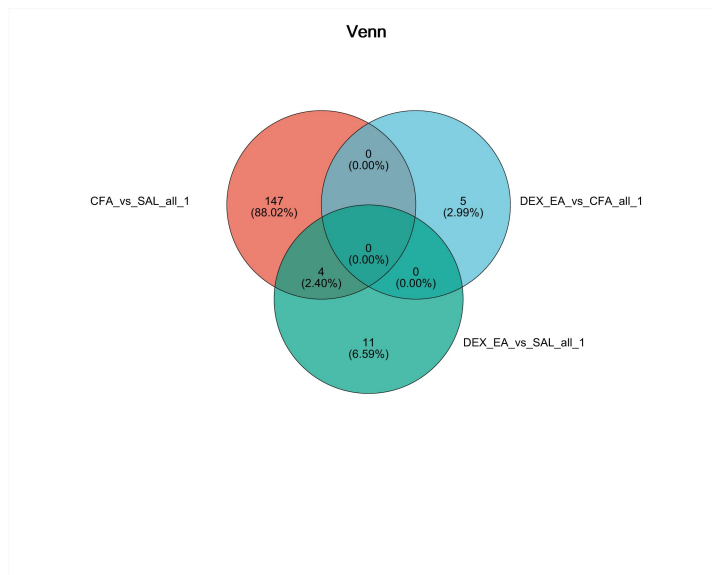
☆Tips:

①选择**共性、特性**部分构建基因集，可在【目标基因集分析模块】进行功能注释、富集、GSEA、PPI网络等个性化分析；

②文章写作中，Venn分析结果可直观展示基因/转录本在各个差异基因集的数量分布状况；

③该部分可用于进一步缩小范围，筛选组间关键的差异基因。

差异基因Venn图



注：不同颜色的圆圈代表不同的基因集，数值代表不同基因集间共有和特有的基因/转录本数目。百分比是以所选基因集的并集作为分母计算。  
【当参数设置处基因集≤5个的时候：圆圈内部所有数字之和代表该基因集基因/转录本个数的总和，圆圈的交叉区域代表各基因集中共有基因/转录本个数。当参数设置处基因集>6个的时候：花瓣图展示两部分内容，所有基因集共有的基因/转录本数目（中心区域）和各基因集特有的基因/转录本数目（花瓣区域）。】【点击数字可创建基因集】

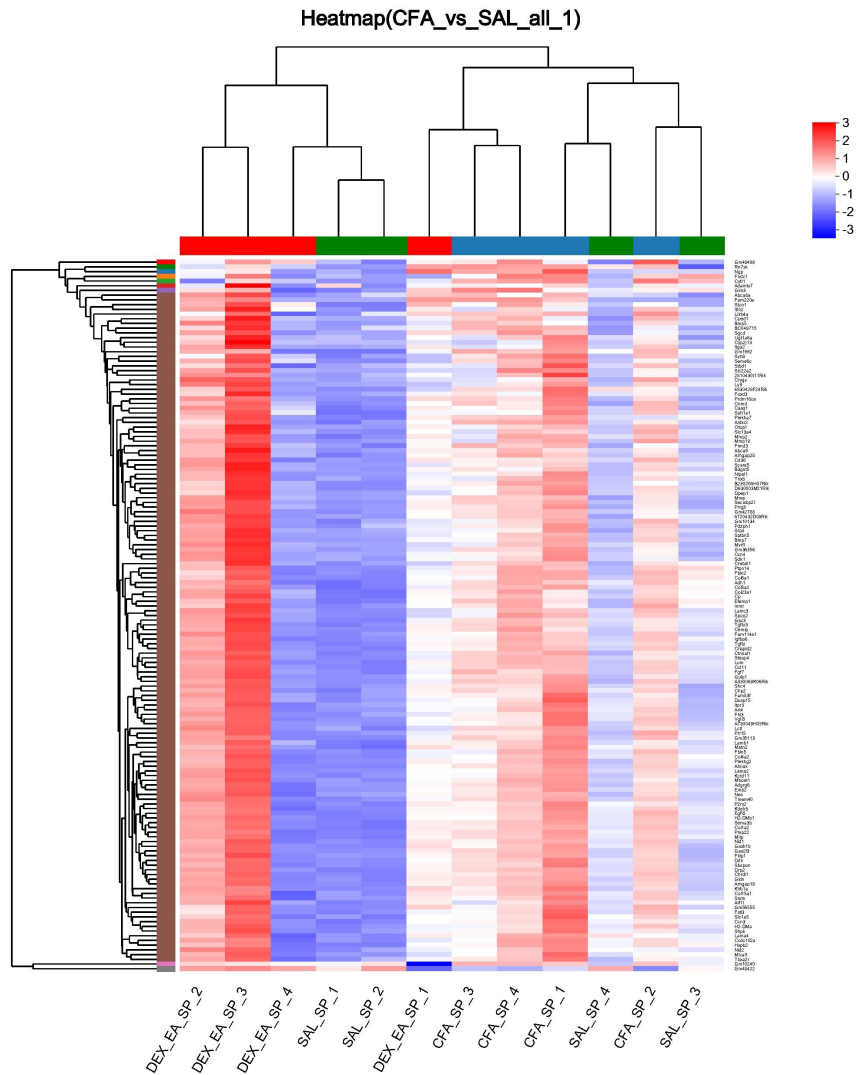
### 3.2.3 差异基因聚类分析

医学应用：对选择的 **基因集在样本间的表达模式** 进行聚类分析，并提供热图可视化结果。

☆Tips:

- ①文章**写作常用图形**，能够直观看出基因/转录本在不同组别间的差异表达情况；
- ②热图一侧的聚类树，能够直观看出 **基因/转录本间的表达相似性**，同一分支下的基因表达相似性高；
- ③聚类结果中，**相同子聚类（同一分支下）的基因/转录本往往具有功能相关性**，可在【子聚类趋势图】中选择子聚类基因，构建基因集在【目标基因集分析模块】进行功能分析。

[差异基因聚类热图](#)



注：图中每列表示一个样本，每行表示一个基因，图中的颜色表示基因/转录本在该样本中的表达量大小，默认红色代表该基因/转录本在该样本中表达量较高，蓝色代表表达量较低，具体表达量大小变化趋势请见左上方颜色条下的数字标注。左侧为基因/转录本聚类树状图，右侧为基因/转录本名称，两个基因/转录本分支离得越近，说明其表达量越接近；上方为样本聚类的树状图，下方为样本名称，两个样本分支离得越近，说明这两个样本所有基因/转录本的表达模式越接近，即基因/转录本表达量变化趋势越接近。

趋势分析统计表

Cluster id	Members
1	142
2	1
3	1
4	1
5	1
6	1
7	1
8	1
9	1
10	1

注：（1）Cluster id：Cluster编号；（2）Members：不同cluster对应的基因/转录本数目。【点击cluster数字可切换展示不同cluster的子聚类趋势图】

### 3.2.4 差异基因功能注释分析

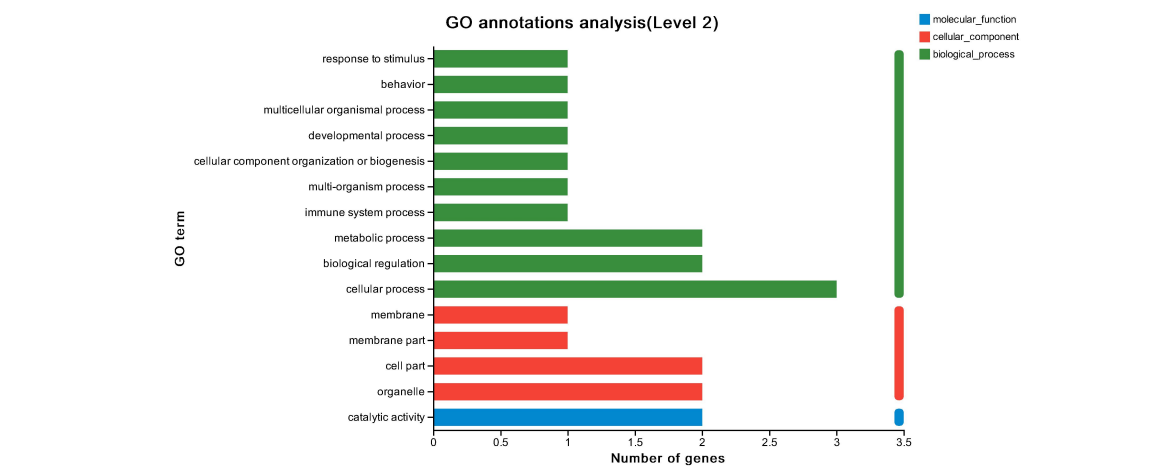
医学应用：富集分析是用于确定 **差异基因是否集中在特定的功能或通路中**，以便推断这些功能或通路在实验条件下是否显著富集（将筛选得到的基因进行功能分类，同时筛选显著富集的功能通路）。

☆Tips:

- ①注释分析得到的是大量通路和功能节点，富集则是分析差异基因是否大概率出现在这些通路或功能节点中。
- ②功能富集的结果可以提供关于基因的 **功能、生物过程、细胞定位、调控机制** 等方面的信息，从而帮助研究者理解基因在生物体中的功能和相关的生理过程。
- ③富集分析使用超几何分布算法获得基因集中的基因显著富集的功能，默认采用BH方法对P值进行校正，当经过校正的P值（Padjust）< 0.05时，认为此功能存在显著富集情况，其中DO和DisGeNET只能针对研究物种是人的项目进行分析；
- ④示例：选择组间显著差异的上/下调基因，进行功能富集分析，发现显著富集到了一系列特定生物学功能通路，从而从分子机制上来解释所观察到的生物学现象。

#### 3.2.4.1 GO注释

差异基因GO分类统计柱形图（条带式）



注：图中纵坐标表示GO的二级分类术语，横坐标表示比对该二级分类的基因/转录本数量，三个颜色表示三大分类。【可通过图表工具进行图形转置切换。】

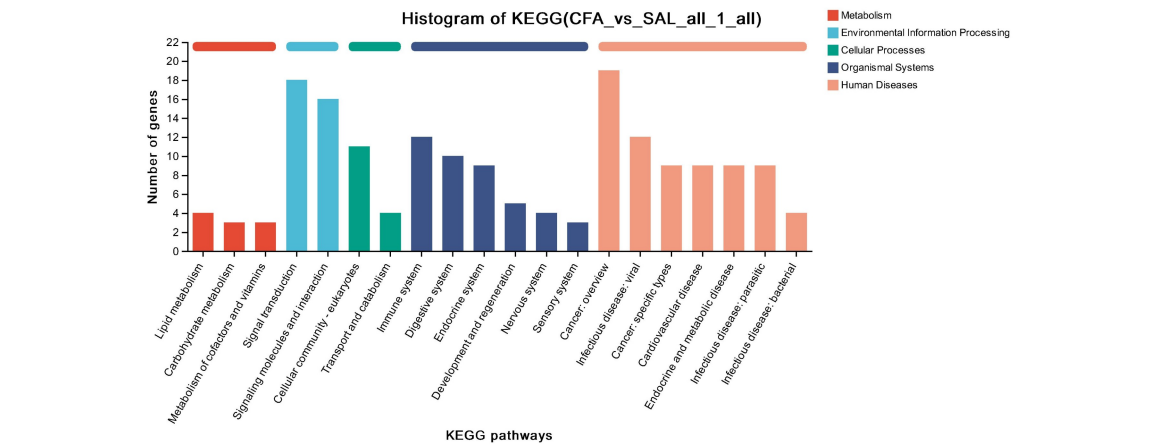
差异基因GO分类统计表

Diff_GOClass_20241229_161735558				
GO ID	Term description	Term type	DEX_EA_vs_CFA_all_1 num	DEX_EA_vs_CFA_all_1 perc...
GO:0002376	immune system process	biological_process	1	1/ 5
GO:0065007	biological regulation	biological_process	2	2/ 5
GO:0008152	metabolic process	biological_process	2	2/ 5
GO:0051704	multi-organism process	biological_process	1	1/ 5
GO:0071840	cellular component organ...	biological_process	1	1/ 5
GO:0009987	cellular process	biological_process	3	3/ 5
GO:0032502	developmental process	biological_process	1	1/ 5
GO:0032501	multicellular organismal p...	biological_process	1	1/ 5
GO:0007610	behavior	biological_process	1	1/ 5
GO:0050896	response to stimulus	biological_process	1	1/ 5

注：（1）GO：注释到的GO编号；（2）Term description：GO二级分类术语；（3）Trem type：GO一级分类名称；（4）Number：注释到该GO二级分类功能的基因/转录本数目；（5）Percent：注释到该GO二级分类功能的基因/转录本数目占基因集总数目的百分比。【每个基因/转录本具有多种GO功能，因此，所有百分比加在一起的数字会大于1。】

3.2.4.2 KEGG注释

差异基因KEGG分类统计柱状图（条带式）



注：纵坐标为KEGG代谢通路的名称；横坐标为注释到该通路下的基因/转录本的数量。KEGG代谢通路的七大类：代谢（Metabolism），遗传信息处理（Genetic Information Processing），环境信息处理（Environmental Information Processing），细胞过程（Cellular Processes），生物体系统（Organismal Systems），人类疾病（Human Diseases），药物开发（Drug Development）。（注：根据具体物种情况，获得的分类数目不同）【可通过图表工具进行图形转置切换】

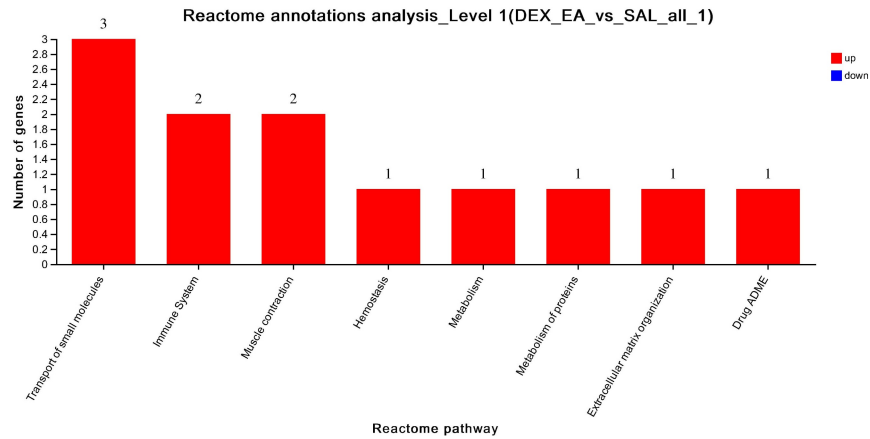
差异基因KEGG分类注释统计表

undefined				
注：（1）First category：KEGG代谢通路的7个分支：代谢（Metabolism），遗传信息处理（Genetic Information Processing），环境信息处理（Environmental Information Processing），细胞过程（Cellular Processes），生物体系统（Organismal Systems），人类疾病（Human Diseases），药物开发（Drug Development）；（2）Second category：KEGG代谢通路的名称；（3）Pathway id：代谢通路编号；（4）Pathway description：KEGG通路具体描述；（5）Number：基因集中注释到该通路下的基因/转录本数量，up代表上调				

, down代表下调。

### 3.2.4.3 Reactome注释

差异基因Reactome分类统计图-柱状图



注：纵坐标为Reactome代谢通路的名称；横坐标为注释到该通路下的基因数量。（注：根据具体物种情况，获得的分类数目不同）【可通过图表工具进行图形转置切换】

差异基因Reactome分类统计表

Diff_ReactineClass_20241229_161737675					
Category	Pathway ID	Pathway description	DEX_EA_vs_SAL_all_1...	DEX_EA_vs_SAL_all_1...	DEX_EA_vs_SAL_all_1...
Extracellular matrix o...	R-MMU-1474244	Extracellular matrix o...	1	1	0
Immune System	R-MMU-6798695	Neutrophil degranula...	2	2	0
Transport of small mo...	R-MMU-174824	Plasma lipoprotein as...	1	1	0
Hemostasis	R-MMU-76002	Platelet activation, sig...	1	1	0
Hemostasis	R-MMU-76005	Response to elevated...	1	1	0
Muscle contraction	R-MMU-5576891	Cardiac conduction	2	2	0
Hemostasis	R-MMU-114608	Platelet degranulation	1	1	0
Muscle contraction	R-MMU-5578775	Ion homeostasis	1	1	0
Immune System	R-MMU-168249	Innate Immune System	2	2	0
Transport of small mo...	R-MMU-983712	Ion channel transport	1	1	0

注：（1）Category：Reactome代谢通路的一级分类；（2）Pathway id：代谢通路编号；（3）Pathway description：Reactome通路具体描述；（4）Number：基因集中注释到该通路下的基因数量，up代表上调，down代表下调。

### 3.2.5 差异基因功能富集分析

医学应用：富集分析是用于确定 **差异基因是否集中在特定的功能或通路中**，以便推断这些功能或通路在实验条件下是否显著富集（将筛选得到的基因进行功能分类，同时筛选显著富集的功能通路）。

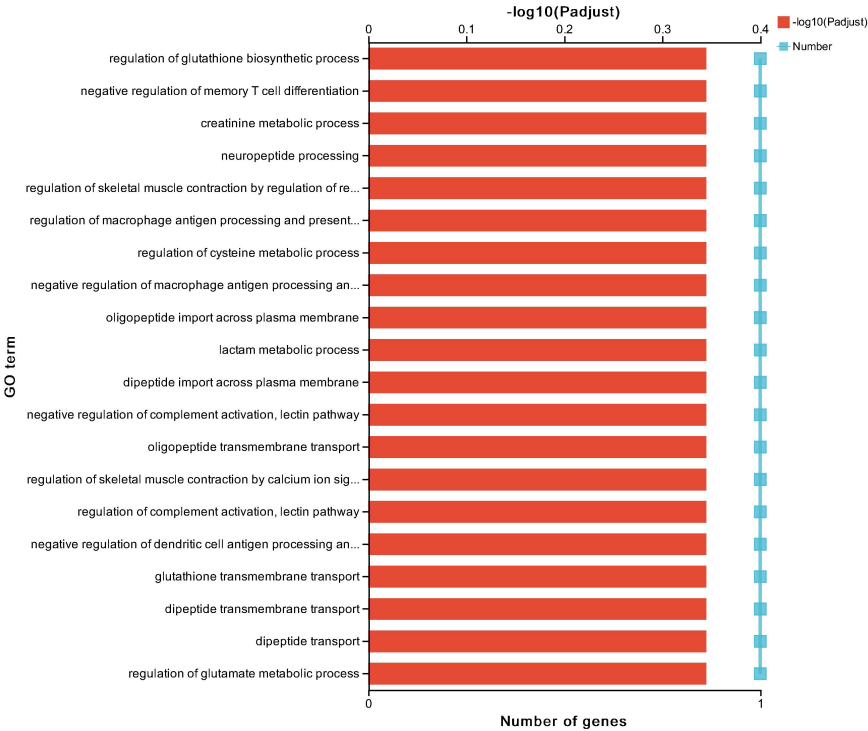
☆Tips:

- ①注释分析得到的是大量通路和功能节点，富集则是分析差异基因是否大概率出现在这些通路或功能节点中。
- ②功能富集的结果可以提供关于基因的 **功能、生物过程、细胞定位、调控机制** 等方面的信息，从而帮助研究者理解基因在生物体中的功能和相关的生理过程。
- ③富集分析使用超几何分布算法获得基因集中的基因显著富集的功能，默认采用BH方法对P值进行校正，当经过校正的P值（Padjust）< 0.05时，认为此功能存在显著富集情况，其中DO和DisGeNET只能针对研究物种是人的项目进行分析；
- ④示例：选择组间显著差异的上/下调基因，进行功能富集分析，发现显著富集到了一系列特定生物学功能通路，从而从分子机制上来解释所观察到的生物学现象。

#### 3.2.5.1 GO富集分析

差异基因GO富集分析结果图-带折线柱形图

GO enrichment analysis(DEX\_EA\_vs\_SAL\_all\_1)



注：纵坐标表示GO term，上方横坐标表示比对上该GO term的基因/转录本数量，对应的是折线上的不同的点；下方横坐标表示富集的显著性水平，对应的是柱子的高度，其中，Padjust越小， $-\log_{10}(\text{Padjust})$  值越大，该GO term越显著富集。【注：默认按Padjust从小到大排序显示top20 的富集结果，可图表工具修改参数调整绘图形式。】

差异基因GO富集分析统计表

Diff_GOEnrich_20241229_161736737								
Number	GO ID	Term type	Description	Ratio_in_study	Ratio_in_pop	Rich factor	Pvalue	Padjust
1	GO:1903786	BP	regulation of ...	1/ 14	1/ 43839	1	0.00031935035...	0.45187760425...
1	GO:0043381	BP	negative regu...	1/ 14	1/ 43839	1	0.00031935035...	0.45187760425...
1	GO:0046449	BP	creatinine me...	1/ 14	1/ 43839	1	0.00031935035...	0.45187760425...
1	GO:0061837	BP	neuropeptide...	1/ 14	1/ 43839	1	0.00031935035...	0.45187760425...
1	GO:0014809	BP	regulation of ...	1/ 14	1/ 43839	1	0.00031935035...	0.45187760425...
1	GO:0002616	BP	regulation of ...	1/ 14	1/ 43839	1	0.00031935035...	0.45187760425...
1	GO:1901494	BP	regulation of ...	1/ 14	1/ 43839	1	0.00031935035...	0.45187760425...
1	GO:0002617	BP	negative regu...	1/ 14	1/ 43839	1	0.00031935035...	0.45187760425...
1	GO:0140205	BP	oligopeptide i...	1/ 14	2/ 43839	0.5	0.00063860599...	0.45187760425...
1	GO:0072338	BP	lactam metab...	1/ 14	2/ 43839	0.5	0.00063860599...	0.45187760425...

注：（1）Number：富集到该GO term的基因或转录本数目；（2）GO id：GO term对应的编号；（3）Term type：GO三大分类（即BP、CC、MF）；（4）Term description：GO功能描述；（5）Ratio\_in\_study：目标基因集含有GO term注释的总基因中落到该GO term的比例，分子为目标基因集中注释到该GO term的基因数目，分母为目标基因集中具有GO term注释的总基因数目；（6）Ratio\_in\_pop：该GO在背景基因（测序得到的所有基因）中占有的比例，分子为所有背景基因中富集到该GO的数目，分母为所有背景基因中具有GO注释的总数目；（7）Rich factor：富集因子，计算公式为：目标基因集中属于这个Term的基因的数量÷背景基因集中这个Term所有基因的数量，即Ratio\_in\_study的分子÷Ratio\_in\_pop的分子；（8）Pvalue：未经校正的P值，P值代表富集出来的结果是否具有统计学上的显著意义，P值越小，在统计学上就越有显著意义；（9）Padjust：校正后的P值，默认采用BH方法对P值进行校正。

3.2.5.2 KEGG富集分析

差异基因KEGG富集分析结果图-气泡图

KEGG pathway

Rich Factor

Padjust

Number

KEGG pathway	Rich Factor	Padjust	Number
Viral life cycle - HIV-1	~0.016	~0.015	1
Nicotinate and nicotinamide metabolism	~0.023	~0.020	1
Human immunodeficiency virus 1 infection	~0.004	~0.025	1

### 差异基因KEGG富集分析统计表

注：（1）Number：富集到该通路的基因或转录本数目；（2）Pathway id：通路编号；（3）Pathway description：KEGG通路具体描述；（4）Database：KEGG数据库；（5）Ratio\_in\_study：目标基因集含有KEGG注释的基因集中落到该KEGG pathway的比例，分子为目标基因集中注释到该KEGG pathway的基因数目，分母为目标基因集中具有KEGG注释的总基因数目；（6）Ratio\_in\_pop：该KEGG pathway在背景基因（测序得到的所有基因）中占有的比例，分子为所有背景基因中富集到该KEGG pathway的数量，分母为所有背景基因中具有KEGG pathway的总数目；（7）Rich factor：富集因子，计算公式为：目标基因集中属于这个Term的基因的数量-背景基因集中这个Term所有背景基因的数量，即Ratio\_in\_study的分子+Ratio\_in\_pop的分子；（8）Pvalue：未经校正的P值，P值代表富集出来的结果是否具有统计学上的显著意义，P值越小，在统计学上就越有显著意义，一般P值小于0.05认为该功能为显著富集项；（9）Padjust：校正后的P值，默认采用BH方法对P值进行校正。（10）First category：KEGG代谢通路的7个分支：代谢（Metabolism），遗传信息处理（Genetic Information Processing），环境信息处理（Environmental Information Processing），细胞过程（Cellular Processes），生物体系统（Organismal Systems），人类疾病（Human Diseases），药物开发（Drug Development）；（11）Second category：KEGG代谢通路的名字。

### 差异基因Reactome富集分析结果图-气泡图

Reactome pathway

Citric acid cycle (TCA cycle)

Pyruvate metabolism and Citric Acid (TCA) cycle

The citric acid (TCA) cycle and respiratory electron t...

Metabolism

Rich factor

Number

● 1  
● 2  
● 3  
● 4

Padjust

0.20  
0.15  
0.10  
0.05  
0.00

### 差异基因Reactome富集分析统计表

Number	Category	Pathway ID	Pathway desc...	Ratio_in_study	Ratio_in_pop	Rich factor	Pvalue	Padjust
1	Metabolism	R-MMU-71403	Citric acid cycl...	1/ 1	22/ 8793	0.04545454545...	0.00250199021...	0.01000796087...
1	Metabolism	R-MMU-71406	Pyruvate met...	1/ 1	47/ 8793	0.02127659574...	0.00534516092...	0.01069032184...
1	Metabolism	R-MMU-14285...	The citric acid ...	1/ 1	167/ 8793	0.00598802395...	0.01899238030...	0.02532317373...
1	Metabolism	R-MMU-14307...	Metabolism	1/ 1	1703/ 8793	0.00058719906...	0.19367678835...	0.19367678835...

注：（1）Number：富集到该通路的基因或转录本数目；（2）Category：Reactome代谢通路的一级分类；（3）Pathway id：通路编号；（4）Pathway description：Reactome通路具体描述；（5）Ratio\_in\_study：目标基因集含有Reactome注释的总基因中落到该Reactome pathway的比例，分子为目标基因集中注释到该Reactome pathway的基因数目，分母为目标基因集中具有Reactome注释的总基因数目；（6）Ratio\_in\_pop：该Reactome pathway在背景基因（测序得到的所有基因）中占有的比例，分子为所有背景基因中富集到该Reactome pathway的数目，分母为所有背景基因中具有Reactome pathway的总数目；（7）Rich factor：富集因子，计算公式为：目标基因集中属于这个Term的基因的数量÷背景基因集中这个Term所有基因的数量，即Ratio\_in\_study的分子÷Ratio\_in\_pop的分子；（8）Pvalue：未经校正的P值，P值代表富集出来的结果是否具有统计学上的显著意义，P值越小，在统计学上就越有显著意义，一般P值小于0.05认为该功能为显著富集项；（9）Padjust：校正后的P值，默认采用BH方法对P值进行校正。

3.3 所有基因数据挖掘

3.3.1 功能注释与表达量查询

将基因/转录本在GO、KEGG、NR、EggNOG、Uniprot、Pfam、Reactome、DO和DisGeNet等九大数据库的功能注释进行汇总，并提供多种检索方式，实现查询特定功能对应的基因/转录本信息，或根据基因信息检索其功能，以便迅速锁定所需信息。

☆Tips：这部分内容提供所有基因的数据挖掘，若老师在关注差异基因的同时，也关注所有基因的表达情况，可以通过该总表筛选出与研究相关的基因、通路等信息。

功能注释信息表

Exp_G_RSEM_TPM_20241229_161644									
Gene id	Gene name	Gene descri...	Biotype	GO id	GO term	GO descripti...	KO id	KO name	KEGG pathw...
ENSMUSG00...	Gnai3	guanine nuc...	protein_cod...	GO:2001234...	biological_p...	negative reg...	K04630	GNAI	mmu04730;...
ENSMUSG00...	Pbsn	probasin [So...	protein_cod...	GO:0036094...	molecular_f...	small molec...	-----	-----	-----
ENSMUSG00...	Cdc45	cell division ...	protein_cod...	GO:0043138...	molecular_f...	3'-5' DNA he...	K06628	CDC45	mmu04110
ENSMUSG00...	H19	H19, imprint...	lncRNA	GO:0035195...	biological_p...	gene silenci...	-----	-----	-----
ENSMUSG00...	Scml2	Scm polycom...	protein_cod...	GO:0036353...	biological_p...	histone H2A...	K11465	SCML2	mmu03083
ENSMUSG00...	Apoh	apolipoprot...	protein_cod...	GO:0005543...	molecular_f...	phospholipi...	K17305	APOH, B2G1	mmu04979
ENSMUSG00...	Narf	nuclear prel...	protein_cod...	GO:0005634...	cellular_com...	nucleus;;lam...	-----	-----	-----
ENSMUSG00...	Cav2	caveolin 2 [S...	protein_cod...	GO:0016020...	cellular_com...	membrane;...	K12958	CAV2	mmu05020;...
ENSMUSG00...	Klf6	Kruppel-like...	protein_cod...	GO:0001650...	cellular_com...	fibrillar cent...	K09207	KLF6_7	-----
ENSMUSG00...	Scmh1	sex comb on...	protein_cod...	GO:0009952...	biological_p...	anterior/po...	K11461	SCMH1	mmu03083

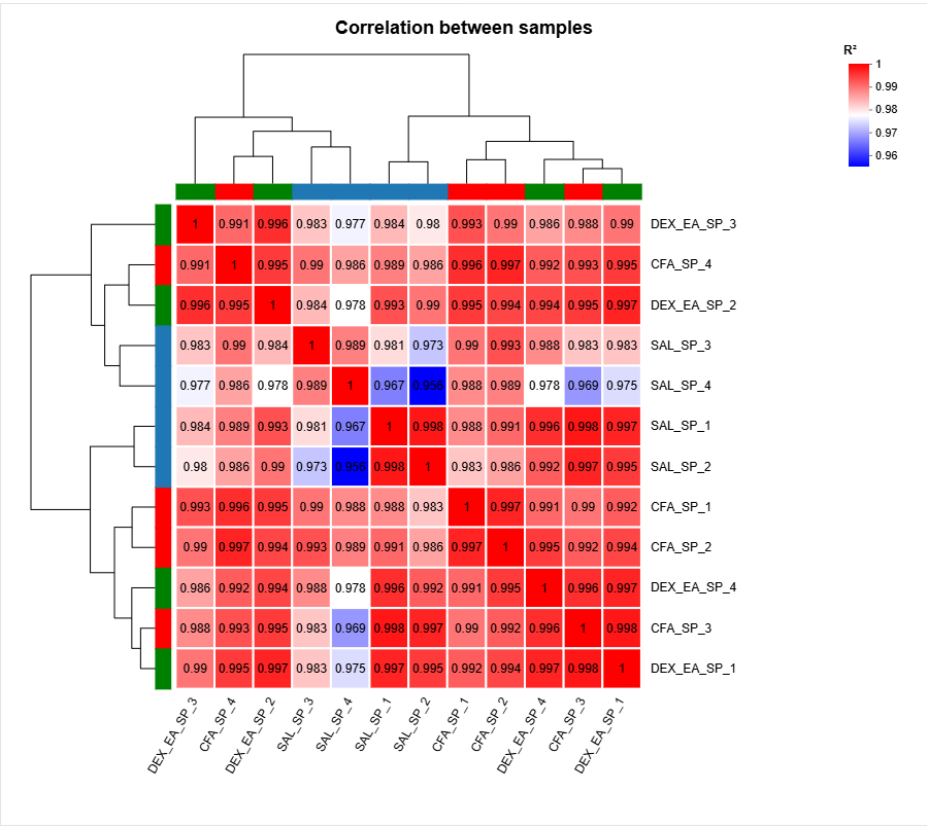
注：（1）Gene/Transcript id：基因/转录本编号；（2）Gene name：基因名称；（3）Gene description：基因描述信息；（4）Biotype：指基因的特征类型，比如蛋白编码基因、外显子、内含子、lncRNA等；类型信息来源于参考基因组注释，“NA”表示无信息，详细说明参看：<http://asia.ensembl.org/info/genome/genebuild/biotypes.html>；（5）其他列为基因/转录本在各大数据库的注释情况。【默认只展示GO和KEGG2个数据库的注释信息和所有样本的表达量信息，可通过筛选表格对其他数据库或者目标样本的表达量进行勾选展示】表格中蓝色字体均可以点击链接到基因详情页或者官网！

3.3.2 样本间相关性分析

生物学重复样本之间的相关性分析，一方面检验生物学重复之间的变异是否符合实验设计的预期，另一方面为差异基因分析提供基本参考。相关系数越接近于1，表明基因/转录本在样本间的表达量相似度越高，即样本间相关性越好。

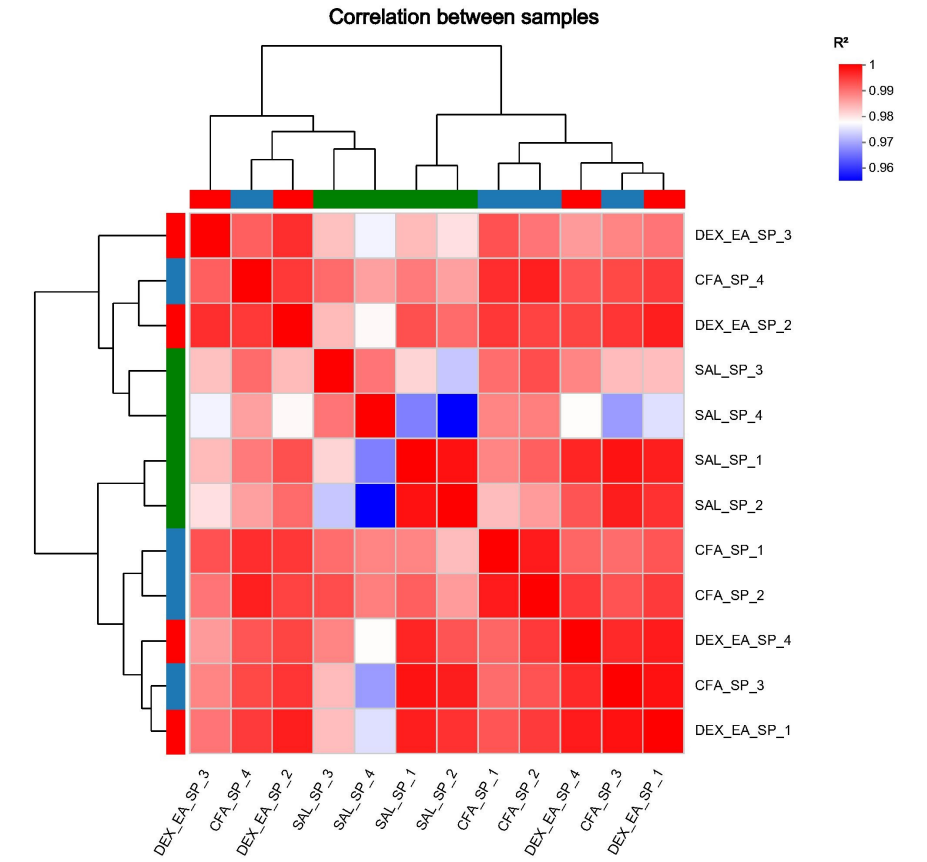
☆Tips：此部分可以判断样本重复性是否符合预期，若有偏离样本，可在分析环节将其剔除后进行后续分析。

样本间相关性热图-经典热图



注：图中右侧和下侧为本项目的样本/组别名称【参数设置处“以分组的均值分析”选择“是”的时候可以对样本分组进行相关性分析】。左侧和上侧为各样本/组别聚类情况，不同颜色的方块代表两个样本的相关性高低，数值越大，表示两个样本/组别的相关性越大，越相近，说明样本间的生物学重复越好。

样本间相关性热图-经典热图



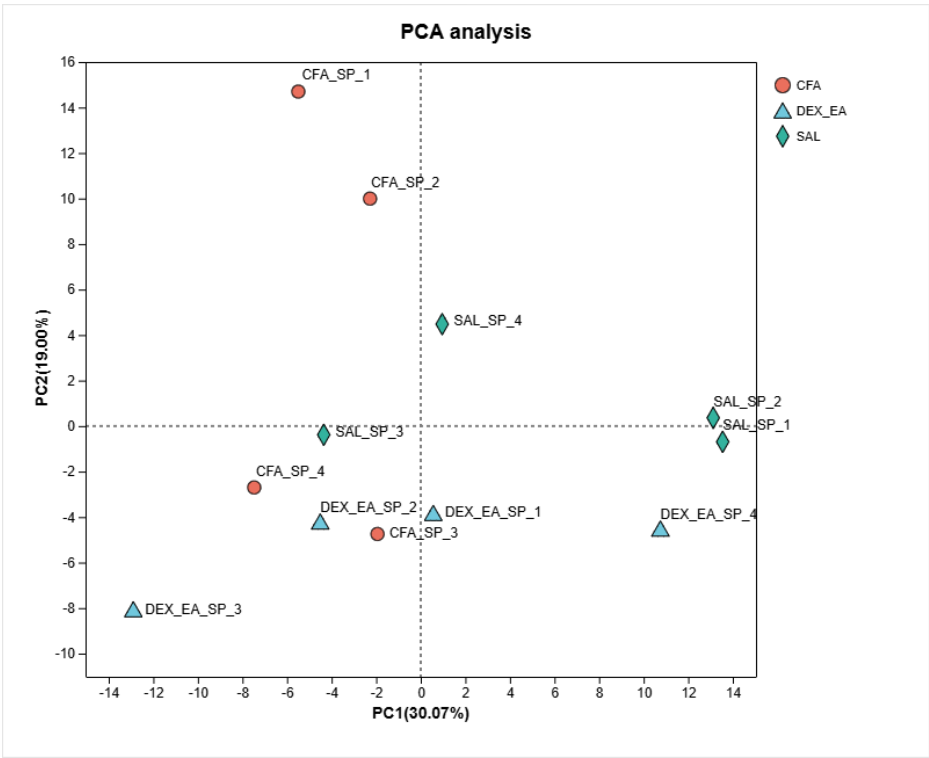
注：图中右侧和下侧为本项目的样本/组别名称【参数设置处“以分组的均值分析”选择“是”的时候可以对样本分组进行相关性分析】。左侧和上侧为各样本/组别聚类情况，不同颜色的方块代表两个样本的相关性高低，数值越大，表示两个样本/组别的相关性越大，越相近，说明样本间的生物学重复越好。

### 3.3.3 样本间PCA分析

主成分分析（PCA）可以降低数据的复杂性，深入挖掘样品之间的关系和变异大小。基本原理是，利用数学的方法，将原来变量重新组合成一组新的互相无关的几个综合变量（即主成分），对所有因素按重要性排序，通常靠后的微小因素被忽略掉，从而起到简化数据的作用。实际项目中，可以通过PCA找出离群样品、判别相似性高的样品簇等。

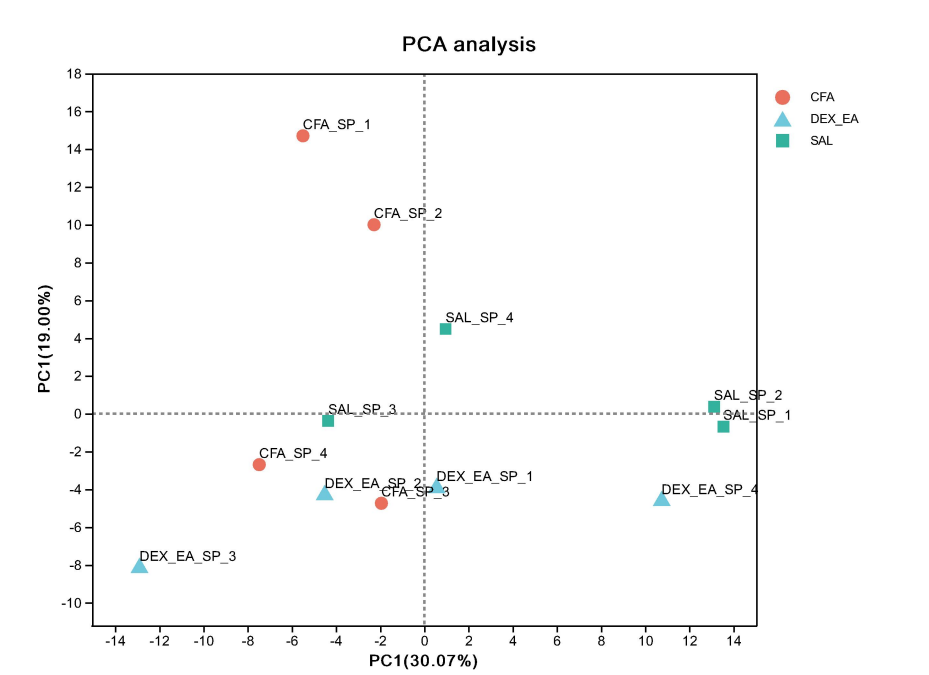
☆Tips: 在医学领域中，我们可以用PCA图来进行疾病危险因素分析，推断肿瘤亚群之间的进化关系.....还用它来观察样本的分组、趋势、剔除异常数据，在文献中出现率还是很高的！

PCA图



注：样本通过降维分析后，在主成分上有相对坐标点。各个样本点的距离代表了样本间的距离，距离越近表明样本间相似性越高，说明样本间的生物学重复越好。横坐标表示二维图中主成分1（PC1）对区分样本的贡献度，纵坐标表示二维图中主成分2（PC2）对区分样本的贡献度。【点击“切换图形”可以切换三维PCA图展现形式，Z轴表示三维图中主成分3（PC3）对区分样本的贡献度。】

PCA图



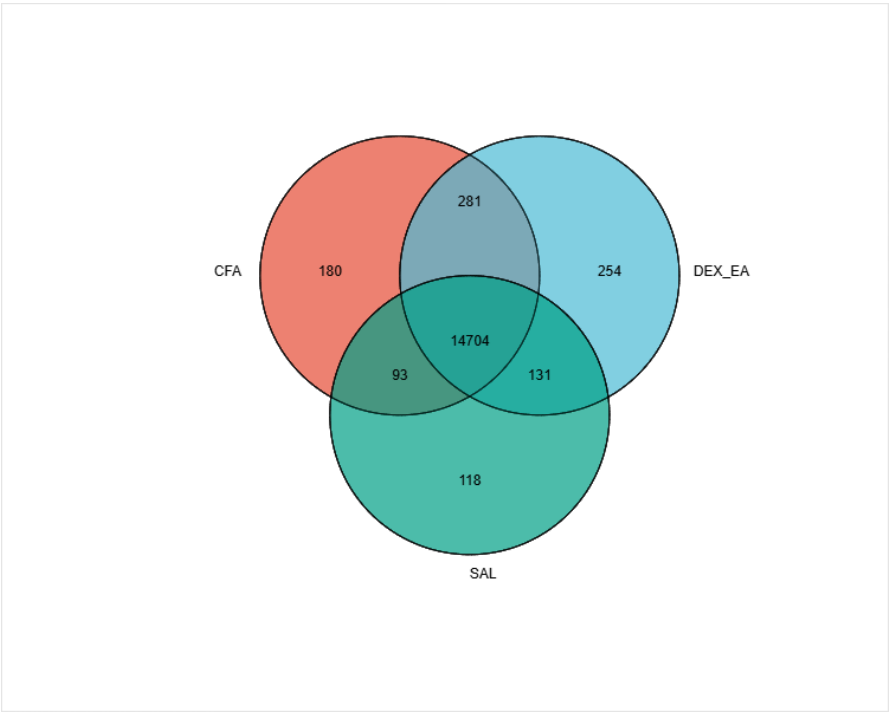
注：样本通过降维分析后，在主成分上有相对坐标点。各个样本点的距离代表了样本间的距离，距离越近表明样本间相似性越高，说明样本间的生物学重复越好。横坐标表示二维图中主成分1（PC1）对区分样本的贡献度，纵坐标表示二维图中主成分2（PC2）对区分样本的贡献度。【点击“切换图形”可以切换三维PCA图展现形式，Z轴表示三维图中主成分3（PC3）对区分样本的贡献度。】

### 3.3.4 样本间Venn分析

根据基因/转录本在不同样本间的表达情况，进行样本间Venn分析，获得样本间或组间的共表达和特表达基因/转录本。

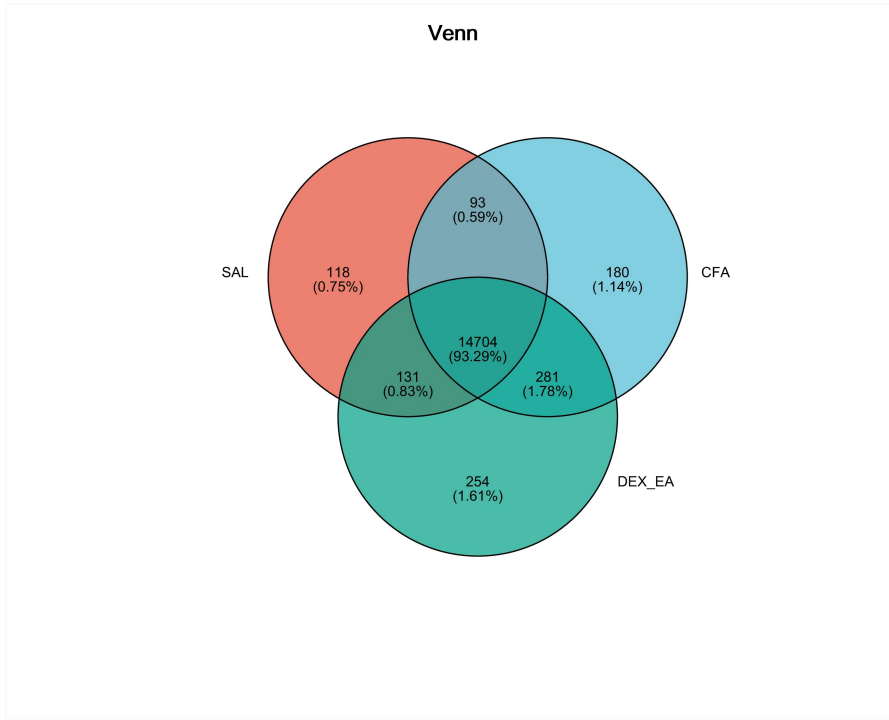
☆Tips: 在转录组数据分析过程中，经常需要针对某几组样本进行共有和特有基因的可视化展示，基于此需求，通常可以选择Venn分析进行可视化展示。

Venn图



注：不同颜色的圆圈代表一个样本/组别中基于表达量筛选的基因/转录本，数值代表不同样本/组别间共有和特有的基因/转录本数目。  
【当参数设置处样本/组别≤5个的时候：圆圈内部所有数字之和代表该样本/组别中基因/转录本个数的总和，圆圈的交叉区域代表各样本/组别中共有基因/转录本个数。当参数设置处样本/组别 > 6个的时候：花瓣图展示两部分内容，所有样本/组别共有的基因/转录本数目（中心区域）和各样本/组别特有的基因/转录本数目（花瓣区域）。】 【点击数字可创建基因集】

Venn图



注：不同颜色的圆圈代表一个样本/组别中基于表达量筛选的基因/转录本，数值代表不同样本/组别间共有和特有的基因/转录本数目。  
【当参数设置处样本/组别≤5个的时候：圆圈内部所有数字之和代表该样本/组别中基因/转录本个数的总和，圆圈的交叉区域代表各样本/组别中共有基因/转录本个数。当参数设置处样本/组别 > 6个的时候：花瓣图展示两部分内容，所有样本/组别共有的基因/转录本数目（中心区域）和各样本/组别特有的基因/转录本数目（花瓣区域）。】 【点击数字可创建基因集】

### 3.4 基因结构数据挖掘

#### 3.4.1 差异可变剪切

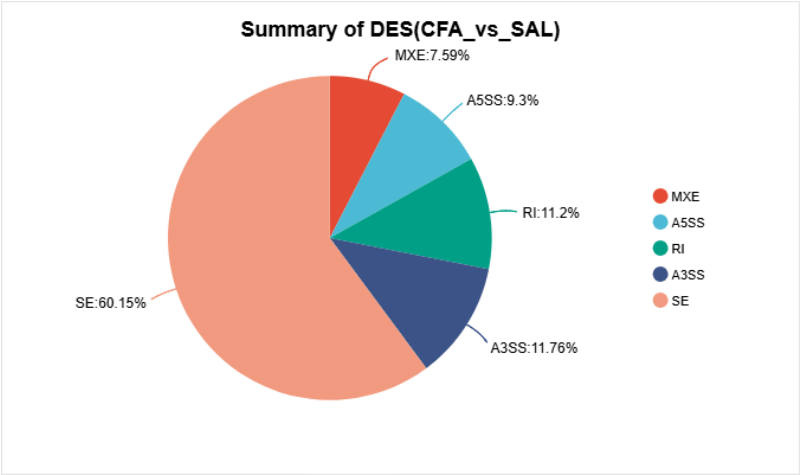
可变剪切 (Alternative Splicing, AS)，是大多数真核生物细胞中普遍存在的一种基因表达方式。真核细胞的基因序列包含内含子 (intron) 与外显子 (exon)，在基因转录成mRNA前体内内含子会被RNA剪切体移除，而外显子则保留于成熟mRNA中。一条未经剪切的RNA，可以具有多种外显子剪切形式，因此使得一个基因在不同时间、不同环境中可以翻译出不同的蛋白质，进而增加其生理状况下系统的复杂性或适应性。rMATS (<http://rnaseq-mats.sourceforge.net/index.html>) 是针对RNA-seq数据开发的用于分析可变剪切的软件，不仅可以对可变剪切事件进行分类，也可以对不同样本间可变剪切事件进行差异分析。rMATS是MATS软件的升级版，升级后的版本可以处理

双端或单端测序的生物学重复样本。

☆Tips：可变剪切是重要的基因表达的转录后调控，对于mRNA 的加工有很大影响。可变剪切缺陷，包括基因改变和/或前体mRNA 和反式作用因子的表达变化，都会导致许多癌症。在医学研究中，可变剪切在肿瘤发生、发展中起重要作用。对比正常组织，肿瘤样本存在广泛的可变剪切事件。异常的可变剪切体可通过抑制凋亡、促进耐药、促进DNA损伤修复、推进细胞周期循环等多种途径促进肿瘤的发生发展。[转录组差异可变剪切分析结果解读及数据挖掘思路：https://www.majorbio.com/wenku/info/1587](https://www.majorbio.com/wenku/info/1587)

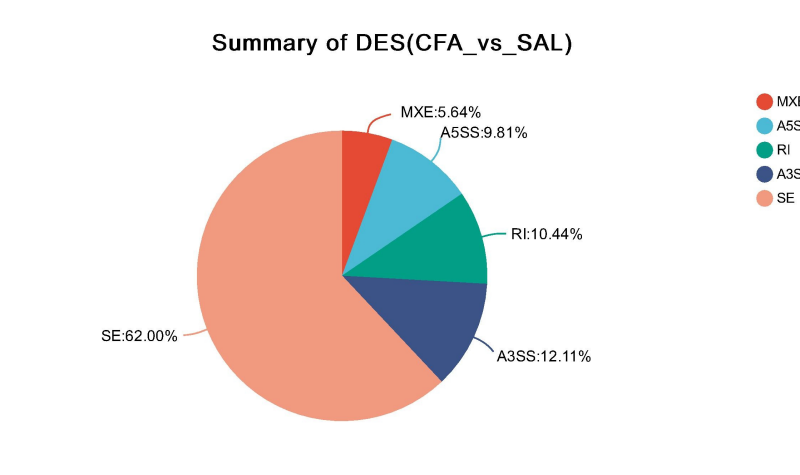
3.4.1.1 差异可变剪切事件统计

差异可变剪切事件统计图



注：展示组内各类差异可变剪切事件数目分布情况。（1）SE：外显子跳跃；（2）A5SS：第一个外显子发生可变剪切；（3）A3SS：最后一个外显子发生可变剪切；（4）MXE：外显子选择性跳跃；（5）RI：内含子滞留。

差异可变剪切事件统计图



注：展示组内各类差异可变剪切事件数目分布情况。（1）SE：外显子跳跃；（2）A5SS：第一个外显子发生可变剪切；（3）A3SS：最后一个外显子发生可变剪切；（4）MXE：外显子选择性跳跃；（5）RI：内含子滞留。

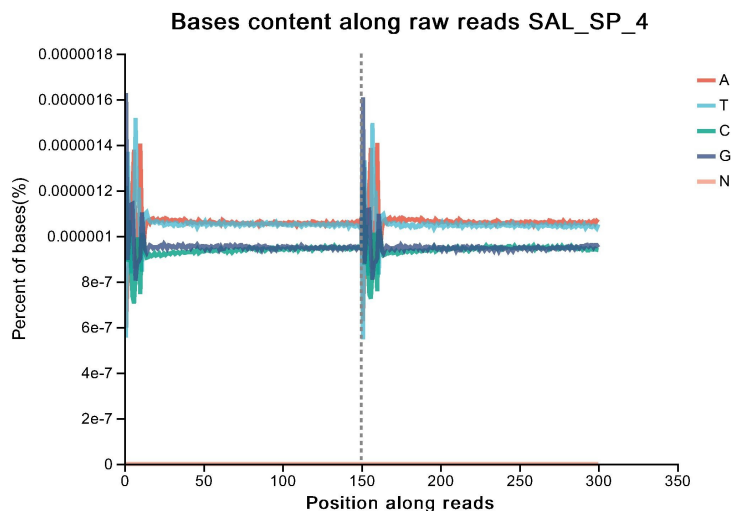
3.5 基础分析

3.5.1 测序数据质控

本项目采用高通量二代测序平台完成的转录组测序。为方便测序数据的分析、发布和共享，该平台通过将测序图像信号经CASAVA碱基识别（Base Calling）转换成文字信号，并将其以 fastq格式储存起来作为原始数据。根据index序列区分各个样本的数据，以便进行后续分析。在fastq文件中每条序列由4行数据组成，其中第一行和第三行为读段识别码（第一行以“@”开头，第三行以“+”开头），第二行为碱基序列，而第四行是第二行序列的各碱基所对应的测序质量值。高通量二代测序单次运行能产生数十亿级的reads，如此海量的数据无法逐个展示每条read的质量情况；因此我们运用统计学的方法，对所测序列进行统计和质控，可以从宏观上直观地反映出样本的文库构建质量和测序质量。测序数据质控共包括三部分内容：1. 测序数据统计；2. 原始数据统计；3. 质控数据统计。

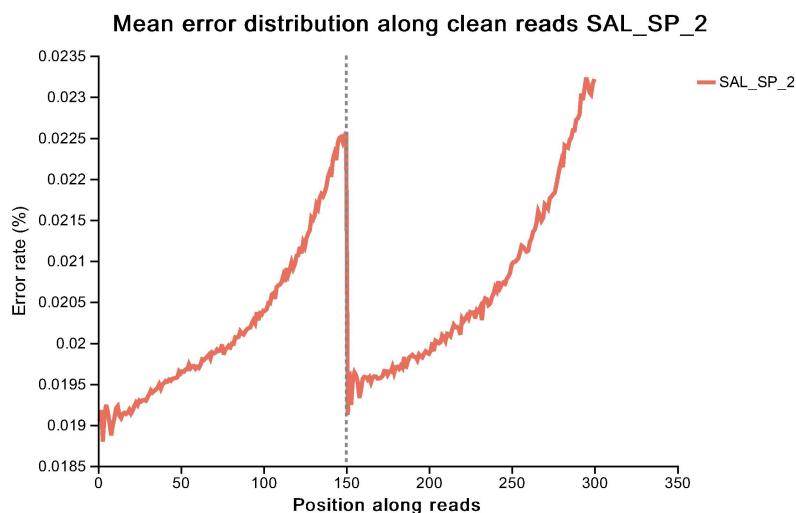
☆Tips：在文章发表时raw/clean data一般会被要求上传至公共数据库，如NCBI的GEO或SRA数据库，因此请妥善保存raw/clean data数据；另外如果需要上传数据可以咨询售后技术支持或参考医学有参云平台项目概览页面（链接是<https://www.majorbio.com/wenku/info/478>），按照操作指南上传raw/clean data数据。

碱基含量分布图-质控数据



注：横坐标是reads的碱基坐标，纵坐标是所有reads在该测序位置（如第一个测序碱基）A、C、G、T、N碱基分别占的百分比，其中N指的是模糊碱基，即不能准确判断为A、T、G、C的碱基，不同碱基用不同颜色表示。

碱基错误率分布图-质控数据



注：展示所有测序reads的各碱基测序错误率均值分布，一般在0.1%以下。横坐标是reads的碱基坐标，表示reads上从5'到3'端依次碱基的排列；纵坐标是所有reads在该位点处碱基的平均错误率。

测序数据统计表

Sample	Raw reads	Raw bases	Clean reads	Clean bases	Error rate(%)	Q20(%)	Q30(%)	GC content(%)
SAL_SP_2	51816036	7772405400	51815620	7740608729	0.0204	99.17	97.35	46.32
SAL_SP_1	45652100	6847815000	45651712	6822276213	0.0208	99.07	97.06	46.39
SAL_SP_3	51540864	7731129600	51540294	7697491688	0.0204	99.15	97.3	46.7
SAL_SP_4	49981048	7497157200	49980612	7462714735	0.0207	99.1	97.14	47.33
CFA_SP_1	46566570	6984985500	46566258	6952067076	0.021	99.03	96.96	46.85
CFA_SP_2	45414124	6812118600	45413726	6780963348	0.0207	99.1	97.14	46.55
CFA_SP_3	47732398	7159859700	47731974	7130280991	0.0205	99.14	97.28	46.59
CFA_SP_4	49409820	7411473000	49409478	7382278122	0.0207	99.09	97.13	47.21
DEX_EA_SP_1	55710842	8356626300	55710468	8321540308	0.0204	99.16	97.34	47.02
DEX_EA_SP_2	50431212	7564681800	50430700	7538027369	0.0206	99.12	97.2	47.21
DEX_EA_SP_3	50345208	7551781200	50344900	7520846667	0.0209	99.05	97	47.55
DEX_EA_SP_4	39662970	5949445500	39662558	5925146902	0.0207	99.11	97.17	46.5

注：（1）Sample：样品名称，点击样本名前面的“R”或“C”，可分别查看原始数据（Raw data）和质控数据（Clean data）的质量，包括碱基错误率和碱基含量分布情况；（2）Raw reads：原始测序数据的总条目数（reads，代表测序读段，一个reads即为一行）；（3）Raw bases：原始测序总数据量（即Raw reads数目乘以reads读长）；（4）Clean reads：质控后测序数据的总条目数；（5）Clean bases：质控后测序总数据量（即Clean reads数目乘以reads长度）；（6）Error rate（%）：质控数据对应的测序碱基平均错误率，一般在0.1%以下；（7）Q20（%）、Q30（%）：对质控后测序数据进行质量评估，Q20、Q30分别指测序质量在99%和99.9%以上的碱基占总碱基的百分比，一般Q20在85%以上，Q30在80%以上；（8）GC content（%）：质控数据对应的G和C碱基总和占总碱基的百分比。

3.5.2 序列比对分析

3.5.2.1 比对结果统计

将质控后的原始数据，即clean data（reads），与参考基因组比对，获取用于后续分析的有效数据（mapped reads），同时对该次转录组测序的比对结果进行质量评估。

☆Tips：通常情况下，测序reads中超过80%均可在参考基因组中找到其响应的定位，这其中具有多个定位的reads的比例通常不会超过10%。如果mapping到参考基因组上的reads比例较低，则可能是由于参考基因组本身不适合或在前期实验过程中出现污染。

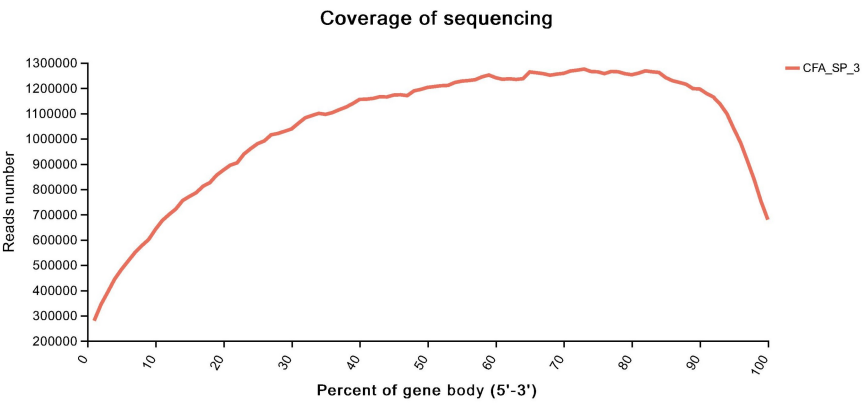
比对结果统计表

Sample	Total reads	Total mapped	Multiple mapped	Unique mapped
SAL_SP_2	51815620	51446755(99.29%)	4610100(8.9%)	46836655(90.39%)
SAL_SP_1	45651712	45319718(99.27%)	4023935(8.81%)	41295783(90.46%)
SAL_SP_3	51540294	51126616(99.2%)	4110585(7.98%)	47016031(91.22%)
SAL_SP_4	49980612	49562231(99.16%)	4127462(8.26%)	45434769(90.9%)
CFA_SP_1	46566258	46155789(99.12%)	4351429(9.34%)	41804360(89.77%)
CFA_SP_2	45413726	45009810(99.11%)	4314230(9.5%)	40695580(89.61%)
CFA_SP_3	47731974	47313508(99.12%)	3690210(7.73%)	43623298(91.39%)
CFA_SP_4	49409478	49001054(99.17%)	3746540(7.58%)	45254514(91.59%)
DEX_EA_SP_1	55710468	55258339(99.19%)	4353957(7.82%)	50904382(91.37%)
DEX_EA_SP_2	50430700	50074980(99.29%)	3695555(7.33%)	46379425(91.97%)
DEX_EA_SP_3	50344900	49967539(99.25%)	3224264(6.4%)	46743275(92.85%)
DEX_EA_SP_4	39662558	39363270(99.25%)	3209629(8.09%)	36153641(91.15%)

注：（1）Sample：样本名称；（2）Total reads：测序序列经过过滤后的序列数量统计（即Clean reads）；（3）Total mapped：能定位到基因组上的Clean reads数目及百分比；（4）Multiple mapped：在参考序列上有多个比对位置的Clean reads数目及百分比；（5）Unique mapped：在参考序列上有唯一比对位置的Clean reads数目及百分比。

3.5.2.2 转录组质量评估

测序覆盖度分布图

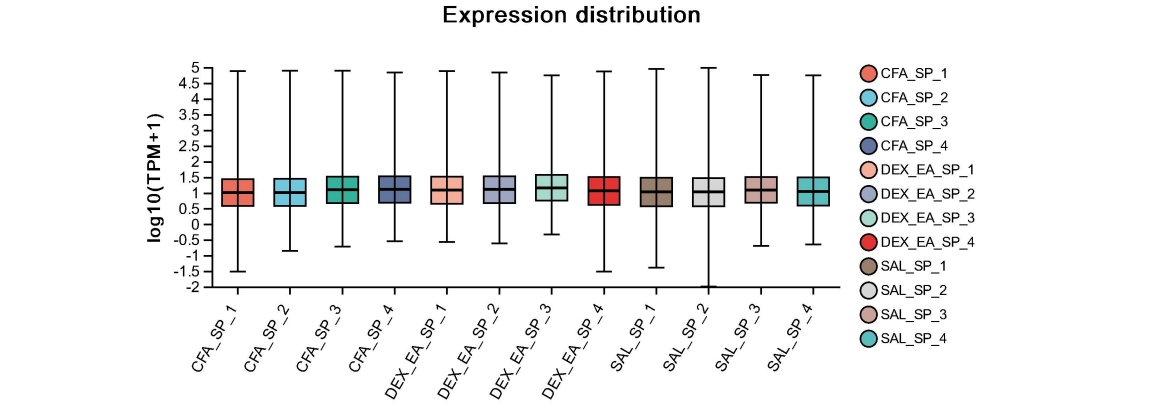


注：横坐标为单个基因的碱基长度占总碱基长度的百分比，0表示基因的5'端，100表示基因的3'端；纵坐标为比对到所有基因横轴位置上相应区间内的序列条数总和。图中体现了所有基因覆盖情况的叠加结果，曲线中每个点的纵坐标表示所有基因在该相对比例位置上所有序列的数量；曲线反映了测序所得序列是否在基因上均匀分布。若无明显偏向峰，则说明测序无偏向性。

3.5.3 表达量分布

使用软件分别对基因/转录本的整体表达水平进行定量分析，以便后续分析不同样本间基因/转录本的差异表达情况，并通过盒形图和小提琴图展示不同样本间的基因表达分布情况。此外也可通过结合序列功能信息，揭示基因的调控机制。

表达量分布盒形图



### 3.6 项目背景

#### 3.6.1 基本信息

展示该项目的样本基本信息，包括样品描述、样本名称以及组名；此外也提供了真核有参医学版转录组生信分析中用到的软件名称/数据库、版本及其对应使用的分析项，并给出各软件的来源。

☆Tips：此部分提供样本基本信息，可在此处进行样本名称更改，分析等相关软件可用于文章写作中使用。

##### 3.6.1.1 项目样本信息

项目样本信息表

Sample description	Sample productive name	Sample initial name	Sample analysis name	Group name
样本信息表	SAL_SP_2	SAL_SP_2	SAL_SP_2	SAL
样本信息表	SAL_SP_2	SAL_SP_1	SAL_SP_1	SAL
样本信息表	SAL_SP_3	SAL_SP_3	SAL_SP_3	SAL
样本信息表	SAL_SP_4	SAL_SP_4	SAL_SP_4	SAL
样本信息表	CFA_SP_1	CFA_SP_1	CFA_SP_1	CFA
样本信息表	CFA_SP_2	CFA_SP_2	CFA_SP_2	CFA
样本信息表	CFA_SP_3	CFA_SP_3	CFA_SP_3	CFA
样本信息表	CFA_SP_4	CFA_SP_4	CFA_SP_4	CFA
样本信息表	DEX_EA_SP_1	DEX_EA_SP_1	DEX_EA_SP_1	DEX_EA
样本信息表	DEX_EA_SP_2	DEX_EA_SP_2	DEX_EA_SP_2	DEX_EA
样本信息表	DEX_EA_SP_3	DEX_EA_SP_3	DEX_EA_SP_3	DEX_EA
样本信息表	DEX_EA_SP_4	DEX_EA_SP_4	DEX_EA_SP_4	DEX_EA

注：展示本项目的样本基本信息，包括Sample description（样品描述）、Sample initial/analysis name（样本初始/现用名称）以及Group name（组名）。“样本描述”和“样本现用名称”可通过【批量编辑】进行再次编辑，修改后，页面展示修改后的样本名。

##### 3.6.1.2 分析软件信息

分析软件信息表

Soft/Database	Version	Analysis	Source
NR 数据库	Version 2023.07	基因注释	<a href="https://www.ncbi.nlm.nih.gov/p...">https://www.ncbi.nlm.nih.gov/p...</a>
GO 数据库	Version 2023.07	基因注释	<a href="http://www.geneontology.org/">http://www.geneontology.org/</a>
NCBI 物种分类数据库	Version 2023.11	基因注释	<a href="ftp://ftp.ncbi.nlm.nih.gov/pub/ta...">ftp://ftp.ncbi.nlm.nih.gov/pub/ta...</a>
Swiss-prot 数据库	Version 2023.11	基因注释	<a href="ftp://ftp.uniprot.org/pub/databa...">ftp://ftp.uniprot.org/pub/databa...</a>
Rfam 数据库	Version 14.10	核糖体比例评估	<a href="http://rfam.janelia.org/">http://rfam.janelia.org/</a>
eggNOG 数据库	Version 2020.06	基因注释	<a href="http://eggnogdb.embl.de/#/app...">http://eggnogdb.embl.de/#/app...</a>
KEGG 数据库	Version 2023.09	基因注释	<a href="http://www.genome.jp/kegg/">http://www.genome.jp/kegg/</a>
Pfam 数据库	Version 36.0	基因注释	<a href="http://pfam.xfam.org/">http://pfam.xfam.org/</a>
Uniprot 数据库	Version 2023.11	基因注释	<a href="ftp://ftp.uniprot.org/pub/databa...">ftp://ftp.uniprot.org/pub/databa...</a>
STRING 数据库	v12.0	靶基因蛋白互作网络	<a href="https://string-db.org/">https://string-db.org/</a>
Reactome 数据库	Version 86	Reactome生物通路分析	<a href="https://reactome.org/">https://reactome.org/</a>
DO 数据库	Version 2023.09	基因注释	<a href="https://disease-ontology.org/">https://disease-ontology.org/</a>
PIR idmapping 数据库	Version 2023.11	基因注释	<a href="ftp://ftp.pir.georgetown.edu/da...">ftp://ftp.pir.georgetown.edu/da...</a>
GATK	Version 3.8	SNP/Indel分析	<a href="https://software.broadinstitute.o...">https://software.broadinstitute.o...</a>

cufflinks	Version 2.2.1	转录本组装	<a href="http://cole-trapnell-lab.github.io...">http://cole-trapnell-lab.github.io...</a>
STAR	Version 2.7.1a	序列比对分析	<a href="http://code.google.com/p/rna-s...">http://code.google.com/p/rna-s...</a>
sentieon	Version sentieon-genomics-20230...	SNP/Indel分析	<a href="https://support.sentieon.com/qu...">https://support.sentieon.com/qu...</a>
edgeR	Version 4.0.2	表达量差异分析	<a href="https://bioconductor.org/packag...">https://bioconductor.org/packag...</a>
BLAST+	Version 2.9.0	多类数据库注释比较分析等	<a href="ftp://ftp.ncbi.nlm.nih.gov/blast/...">ftp://ftp.ncbi.nlm.nih.gov/blast/...</a>
Mfuzz	Version 2.6.0	Mfuzz时序分析	<a href="https://www.bioconductor.org/p...">https://www.bioconductor.org/p...</a>
goatools	Version 1.4.4	基因集分析（GO富集）	<a href="https://pypi.org/project/goatool...">https://pypi.org/project/goatool...</a>
e1071	Version 1.7-13	支持向量机	<a href="https://cran.r-project.org/web/p...">https://cran.r-project.org/web/p...</a>
Diamond	Version 2.1.9	多类数据库注释比较分析等	<a href="https://github.com/bbuchfink/di...">https://github.com/bbuchfink/di...</a>
NOISeq	Version 2.46.0	表达量差异分析	<a href="https://www.bioconductor.org/p...">https://www.bioconductor.org/p...</a>
star_fusion	Version 1.8.1	基因融合鉴定	<a href="https://github.com/STAR-Fusion/...">https://github.com/STAR-Fusion/...</a>
WGCNA	Version 1.63	WGCNA分析	<a href="https://horvath.genetics.ucla.edu...">https://horvath.genetics.ucla.edu...</a>
caret	Version 6.0-94	数据建模-分类与回归	<a href="https://cran.r-project.org/web/p...">https://cran.r-project.org/web/p...</a>
DESeq2	Version 1.42.0	表达量差异分析	<a href="https://bioconductor.org/packag...">https://bioconductor.org/packag...</a>
TopHat	Version v2.1.1	序列比对分析	<a href="http://ccb.jhu.edu/software/toph...">http://ccb.jhu.edu/software/toph...</a>
fastp	Version 0.23.4	测序数据质控	<a href="https://github.com/OpenGene/f...">https://github.com/OpenGene/f...</a>
hisat2	Version 2.2.1	序列比对分析	<a href="https://daehwankimlab.github.io...">https://daehwankimlab.github.io...</a>
fastx_toolkit	Version 0.0.14	测序数据质控	<a href="http://hannonlab.cshl.edu/fastx...">http://hannonlab.cshl.edu/fastx...</a>
Sickle	--	测序数据质控	<a href="https://github.com/najoshi/sickle">https://github.com/najoshi/sickle</a>
stringtie	Version 2.2.1	转录本组装	<a href="https://ccb.jhu.edu/software/str...">https://ccb.jhu.edu/software/str...</a>
kallisto	Version 0.46.2	表达量分析	<a href="https://pachterlab.github.io/kall...">https://pachterlab.github.io/kall...</a>
Limma	Version 3.58.1	表达量差异分析	<a href="https://www.bioconductor.org/p...">https://www.bioconductor.org/p...</a>
sva	Version 3.50.0	移除批次效应影响	<a href="http://www.bioconductor.org/pa...">http://www.bioconductor.org/pa...</a>
SeqPrep	--	测序数据质控	<a href="https://github.com/jstjohn/SeqPr...">https://github.com/jstjohn/SeqPr...</a>
GSEA	Version 4.3.2	GSEA分析	<a href="http://software.broadinstitute.or...">http://software.broadinstitute.or...</a>
MSigDB 数据库	Version 2023.2	GSEA分析	<a href="https://docs.gsea-msigdb.org/#M...">https://docs.gsea-msigdb.org/#M...</a>
survminer	Version 0.4.9	lasso_cox	<a href="https://cran.r-project.org/web/p...">https://cran.r-project.org/web/p...</a>
bedtools	Version 2.27.1	序列提取	<a href="https://bedtools.readthedocs.io/...">https://bedtools.readthedocs.io/...</a>
useful	Version 1.2.6	支持向量机	<a href="https://cran.r-project.org/web/p...">https://cran.r-project.org/web/p...</a>
RSEM	Version 1.3.3	表达量分析	<a href="http://deweylab.biostat.wisc.edu...">http://deweylab.biostat.wisc.edu...</a>
rmats	Version 4.0.2	可变剪切分析	<a href="http://rnaseq-mats.sourceforge....">http://rnaseq-mats.sourceforge....</a>
Bowtie2	Version 2.5.4	序列比对	<a href="https://sourceforge.net/projects/...">https://sourceforge.net/projects/...</a>
STEM	Version 1.3.11	时序表达趋势分析	<a href="http://www.cs.cmu.edu/~jernst/...">http://www.cs.cmu.edu/~jernst/...</a>
kernlab	Version 0.9-32	支持向量机	<a href="https://cran.r-project.org/web/p...">https://cran.r-project.org/web/p...</a>
samtools	Version 1.19.2	SNP/Indel分析	<a href="https://github.com/samtools/sam...">https://github.com/samtools/sam...</a>
Salmon	Version 1.10.3	表达量分析	<a href="https://github.com/COMBINE-lab...">https://github.com/COMBINE-lab...</a>
lpath 数据库	Version 3	iPath代谢通路分析	<a href="https://pathways.embl.de/">https://pathways.embl.de/</a>
glmnet	Version 4.1-4	数据建模-分类与回归	<a href="https://cran.r-project.org/web/p...">https://cran.r-project.org/web/p...</a>
randomForest	Version 4.7-1.1	随机森林	<a href="https://cran.r-project.org/web/p...">https://cran.r-project.org/web/p...</a>
DEGSeq	Version 1.56.1	表达量差异分析	<a href="https://bioconductor.org/packag...">https://bioconductor.org/packag...</a>
maSigPro	Version 1.56.0	时序差异聚类	<a href="http://www.bioconductor.org/pa...">http://www.bioconductor.org/pa...</a>
TransDecoder	Version 5.5.0	多类数据库注释比较分析等	<a href="http://transdecoder.github.io/">http://transdecoder.github.io/</a>
HMMER	Version 3.2.1	功能注释	<a href="ftp://selab.janelia.org/pub/softw...">ftp://selab.janelia.org/pub/softw...</a>
rms	Version 6.7-0	数据建模-分类与回归	<a href="https://cran.r-project.org/web/p...">https://cran.r-project.org/web/p...</a>
pROC	Version 1.18.4	数据建模-分类与回归	<a href="https://cran.r-project.org/web/p...">https://cran.r-project.org/web/p...</a>
timeROC	Version 0.4	lasso_cox	<a href="https://cran.r-project.org/web/p...">https://cran.r-project.org/web/p...</a>
AnimalTFDB	Version 4.0	转录因子预测	<a href="http://bioinfo.life.hust.edu.cn/An...">http://bioinfo.life.hust.edu.cn/An...</a>
JASPAR	Version 2024	转录因子预测	<a href="https://jaspar.elixir.no/">https://jaspar.elixir.no/</a>
mlbench	Version 2.1-3.1	支持向量机	<a href="https://cran.r-project.org/web/p...">https://cran.r-project.org/web/p...</a>
DisGeNET 数据库	Version 7.0	基因注释	<a href="https://www.disgenet.org/">https://www.disgenet.org/</a>
ASprofile	Version 1.0	可变剪切分析	<a href="http://ccb.jhu.edu/software/ASP...">http://ccb.jhu.edu/software/ASP...</a>
survival	Version 3.5-7	lasso_cox	<a href="https://cran.r-project.org/web/p...">https://cran.r-project.org/web/p...</a>
GSVA	Version 1.52.3	GSVA分析	<a href="https://www.bioconductor.org/p...">https://www.bioconductor.org/p...</a>

注：展示真核有参医学版转录组生信分析中用到的软件名称/数据库、版本及其对应使用的分析项，并给出各软件的来源。【点击链接可跳转到相关网页】

## 四、备注

### 4.1 结果文件查看说明

结果文件格式	结果文件解释说明	结果文件查看方式
PNG	结果图像文件，位图，无损压缩。	unix/Linux/Mac用户使用display命令打开。
		windows用户可以使用图片浏览器打开，如photoshop等。
JPEG	结果图像文件，位图，有损压缩。	unix/Linux/Mac用户使用display命令打开。
		windows用户可以使用图片浏览器打开，如photoshop等。
SVG	结果图像文件，矢量图，可放大、缩小不失真，方便用户查看和编辑处理，可使用Adobe Illustrator进行编辑，用于文章发表等。	unix/Linux/Mac用户使用display命令打开。
		windows用户可以使用图片浏览器打开，如photoshop等。
xls, CSV	结果数据表格结果，文件以制表符（Tab）分隔	unix/Linux/Mac用户使用 less 或 more 命令查看；
		windows用户使用高级文本编辑器 Editplus/Notepad++等，也可以用Microsoft Excel打开。
PDF	结果图像文件，矢量图，可放大、缩小不失真，方便用户查看和编辑处理，可使用Adobe Illustrator进行编辑，用于文章发表等。	unix/Linux用户使用evince命令打开。
		windows/Mac用户可以使用Adobe Reader/福昕阅读器/网页浏览器等打开。

参考文献

[1] Kim D, Pertea G, Trapnell C, et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions[J]. Genome biology, 2013, 14(4): R: (TopHat2)

[2] Kim D, Langmead B, Salzberg S L. HISAT: a fast spliced aligner with low memory requirements[J]. Nature methods, 2015, 12(4): 357-360.(HISAT)

[3] Trapnell C, Williams B A, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation[J]. Nature biotechnology, 2010, 28(5): 511-515.(Cufflinks)

[4] Pertea M, Pertea G M, Antonescu C M, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads[J]. Nature biotechnology, 2015, 33(3): 290-295.(StringTie)

[5] Buchfink B, Xie C, Huson D H. Fast and sensitive protein alignment using DIAMOND[J]. Nature methods, 2015, 12(1): 59-60.(DIAMOND)

[6] Finn R D, Clements J, Eddy S R. HMMER web server: interactive sequence similarity searching[J]. Nucleic acids research, 2011, 39(suppl\_2): W29-W37. (HMMER)

[7] Liao Y, Smyth G K, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features[J]. Bioinformatics, 2013, 30(7): 923-930.(FeatureCounts)

[8] Li B, Dewey C N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome[J]. BMC bioinformatics, 2011, 12(1): 323. (RSEM)

[9] Love M I, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2[J]. Genome biology, 2014, 15(12): 550. (DESeq2)

[10] Wang L, Feng Z, Wang X, et al. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data[J]. Bioinformatics, 2009, 26(1): 136-138.(DEGseq)

[11] Robinson M D, McCarthy D J, Smyth G K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data[J]. Bioinformatics, 2010, 26(1): 13-140.(edgeR)

[12] Tang H, Klopfenstein D, Pedersen B, et al. GOATOOLS: tools for gene ontology[J]. Zenodo., 2015.(goatools)

[13] McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data[J]. Genome research, 2010, 20(9): 1297-1303.(GATK)

[14] Shen S, Park J W, Lu Z, et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data[J]. Proceedings of the National Academy of Sciences, 2014, 111(51): E5593-E5601.(rMATS)

[15] Conesa A, Götz S, García-Gómez J M, et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research[J]. Bioinformatics, 2005, 21(18): 3674-3676.(BLAST2GO)

[16] Mao X, Cai T, Olyarchuk J G, et al. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary[J]. Bioinformatics, 2007, 23(18): 3793-3800.(KOBAS)