



艾基生物 转录组项目结题报告

一、项目信息

- 二、概述
- 三、工作流程
- 四、结果展示及说明
- 五、参考文献

一、项目信息

合同信息	合同内容
合同编号	XXXX
客户(单位)	XXX
项目名称	6个人源有参转录组
报告时间	2023-02-02

IGEbio
广州艾基生物技术有限公司

广州艾基生物技术有限公司

二、概述

转录组是特定细胞在某一功能状态下所有RNA的总和，包括编码的mRNA和非编码RNA。研究不同环境下细胞内转录组的动态变化有助于我们了解环境、机体、细胞对自身的调控机制。转录组测序(Transcriptome Sequencing)基于 Illumina novaseq 高通量测序平台,对特定组织或细胞在某个时期转录出来的所有mRNA的集合进行测序及分析，拥有精确到单个核苷酸的分辨率。通过将测序获得的reads与参考基因组比对而定量基因表达水平，进而识别不同样品（或样品组）之间显著差异表达的基因或转录本；通过对受关注基因或转录本的功能注释和功能富集分析，为后续的生物学研究提供分子水平的依据。

IGEbio
广州艾基生物技术有限公司

广州艾基生物技术有限公司

三、工作流程

1. 实验流程

从RNA样品到最终数据获得，样品检测、建库、测序每一个环节都会对数据质量和数量产生影响，而数据质量又会直接影响后续信息分析的结果。为了从源头上保证测序数据的准确性、可靠性，广州艾基对样品检测、建库、测序每一个生产步骤都严格把控，从根本上确保了高质量数据的产出。实验建库测序流程如图所示。





广州艾基生物技术有限公司
GUANGZHOU IGE BIOTECHNOLOGY LTD

一、项目信息

二、概述

三、工作流程

四、结果展示及说明

五、参考文献

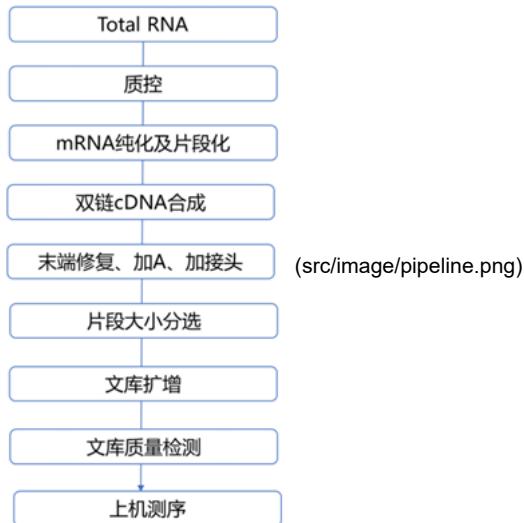


图 1 实验流程图

1.1. RNA样品检测

从组织样品中提取total RNA,对RNA样品进行检测,主要包括3种方法:

- (1) Nanodrop对RNA进行初步的浓度和纯度检测
 - (2) Qubit对RNA浓度进行精确定量
 - (3) Agilent 2100精确检测RNA的完整性 (RIN值)

1.2. 文库构建及库检

样品检测合格后，用mRNA Capture Beads富集真核生物mRNA（若为原核生物，则通过试剂盒去除rRNA来富集mRNA），随后采用高温和金属离子作用实现mRNA的片段化。以mRNA为模板，用六碱基随机引物(random hexamers)合成一链cDNA，随后进行二链cDNA的合成反应，再用DNA Clean Beads纯化双链cDNA。纯化的双链cDNA先进行末端修复、加A尾并连接测序接头，再用DNA Clean Beads进行片段大小分选。最后进行PCR扩增，并用DNA Clean Beads纯化PCR产物，得到最终的文库。文库构建完成后先使用Qubit 3.0对浓度进行初步测定，随后使用Agilent 2100 Bioanalyzer对文库插入片段进行检测，该项通过预期后使用ABI Step One Plus Real-Time PCR system 对文库有效浓度进行准确定量：



一、项目信息

- 二、概述
- 三、工作流程
- 四、结果展示及说明
- 五、参考文献

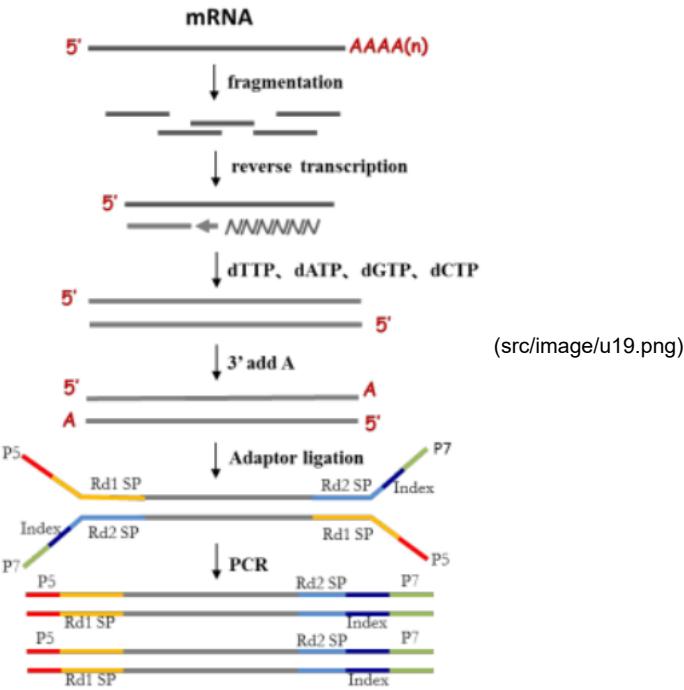


图 2 建库流程图

1.3. 上机测序

文库质控合格后，把不同文库按照有效浓度及目标下机数据量的需求pooling后进行Illumina PE150测序。PE150 (Pair end 150 bp) 指高通量双端测序，每端各测150 bp。在构建的小片段文库中，Insert cDNA，即插入片段是直接测序的单位。双端测序是将每条插入片段的两端进行测序的方法，由于插入片段的长度分布已知，双端测序时不仅可以获得片段两端的序列，也可获得这两段序列之间的长度，从而便于后续组装和比对。

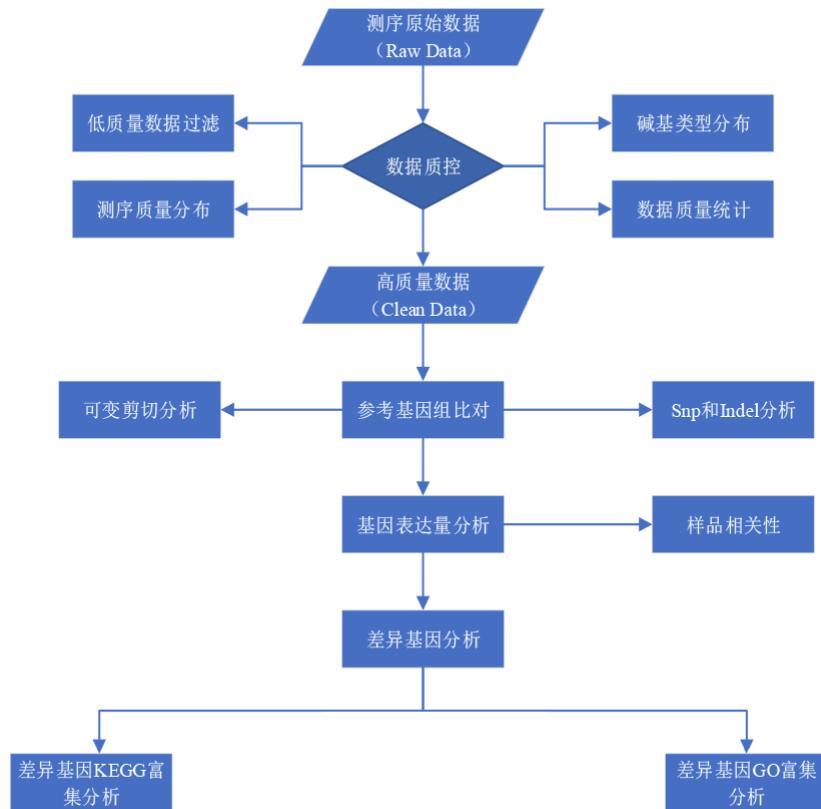
2. 数据分析流程

将下机数据进行过滤得到 Clean reads，将 Clean reads 与参考基因组序列进行比对，根据基因组的注释文件和reads比对到基因组的位置统计每个基因包含的reads数即表达量，然后进行统计检验获得差异表达基因，最后对差异表达基因进行功能注释及GO、KEGG富集检验，获得差异表达基因所富集的GO、KEGG集合。转录组生物信息分析流程示意图 如下图所示。



广州艾基生物
GUANGZHOU IGE BIO

- 一、项目信息
 - 二、概述
 - 三、工作流程
 - 四、结果展示及说明
 - 五、参考文献



(src/image/BioinfoPipeline.png)

图 3 数据分析流程图

3. 项目样本信息

样本 组间比较

表 1 样本基本信息

Search:

Index	Sample initial name	Sample name
1	U3si-SM-1	U3si-SM-1
2	U3si-SM-2	U3si-SM-2
3	T24si-SM-1	T24si-SM-1
4	T24si-SM-2	T24si-SM-2
5	T24si-Ctrl	T24si-Ctrl
6	U3si-Ctrl	U3si-Ctrl

Show 10 entries

Showing 1 to 6 of 6 entries

Previous 1 Next

注: Sample initial name: 样本名; Sample name: 现用样本名; Group name: 组名。



一、项目信息

二、概述

三、工作流程

四、结果展示及说明

五、参考文献

四、结果展示及说明

1. 原始序列数据

高通量测序（如 Illumina HiSeqTM2000/ Miseq 等测序平台）测序得到的原始图像数据文件经碱基识别（Base Calling）分析转化为原始测序序列（Sequenced Reads），我们称之为 Raw Data 或 Raw Reads，结果以 FASTQ 文件格式存储，其中包含测序序列（reads）的序列信息以及其对应的测序质量信息。

```
1 @ERR000589.41 EAS139_45:5:1:2:111/1
2 CTTTCCTCCCTGCTTCCTGGCCCCACCATTCAGGAAACATCTTGTCA
3 +
4 3IIIIIIIIIIII>IIIIFF9BG08E00I%IG+&?(4)%00646.C1#&( 
5 @ERR000589.42 EAS139_45:5:1:2:1293/1
6 AGTTGTTAAAATCCAAGCCAATTAAAGATAGTCTTATCTTTTAAAAGAAAT
7 +
8 IIIIIGII.AIIII=?I9G-/II=+I=4?761BA2C9I+5A711+&>1$/I
```

(src/image/FASTQ-file-format-example.png)

图 4 FASTQ-file-format-example

上述文件中第一行以“@”开头，随后为Illumina测序标识符(Sequence Identifiers)和描述文字；第二行是测序片段的碱基序列；第三行以“+”开头，随后为Illumina测序标识符(也可为空)；第四行是测序片段每个碱基相对应的测序质量值，该行中每个字符对应的ASCII值减去33或64，即为该碱基的测序质量值。

2. 数据质量控制

每个碱基测序错误率是通过测序Phred数值 (Phred score, Qphred) 通过公式 $Q_{\text{phred}} = -10 \log_{10}(e)$ 转化得到，而Phred 数值是在碱基识别 (Base Calling) 过程中通过一种预测碱基判别发生错误概率模型计算得到的，对应关系如下表所显示。质量值Q与错误率，准确率的对应关系见下表。

Phred分值	不正确的碱基识别	碱基正确识别率	Q-sorce
10	1/10	90%	Q10
20	1/100	99%	Q20
30	1/1000	99.9%	Q30
40	1/10000	99.99%	Q40

测序错误率与碱基质量有关，受测序仪本身、测序试剂、样品等多个因素共同影响。对于 Illumina高通量测序平台，测序错误率分布具有两个特点：

(1) 测序错误率会随着测序序列 (Sequenced Reads) 长度的增加而升高，这是由于测序过程中化学试剂的消耗而导致的，并且为 Illumina 高通量测序平台都具有的特征。

(2) 前几个碱基的位置也会发生较高的测序错误率，这是由于边合成边测序过程初始，测序仪荧光感光元件对焦速度较慢，获取的荧光图像质量较低，导致碱基识别错误率较高。。

测序错误率分布检查用于检测在测序长度范围内，有无测序错误率异常的碱基位置。一般情况下，每个碱基位置的测序错误率都应该低于1%。

3. 数据评估



测序错误率与碱基质量有关，受测序仪本身、测序试剂、样品等多个因素共同影响。通常测序序列 (Sequenced Reads) 5'端前几个碱基的错误率相对较高，随着序列的延伸，3'端碱基错误率会逐渐降低，这是由高通量测序的技术特点决定的^[1]。项目结果见图。

原始数据使用软件 fastp^[2]过滤，得到高质量的数据 (clean data)。

对原始数据和 clean 数据，使用软件 fastqc^[3]进行质控处理。

一、项目信息

二、概述

三、工作流程

四、结果展示及说明

五、参考文献

3.1. 碱基分布图

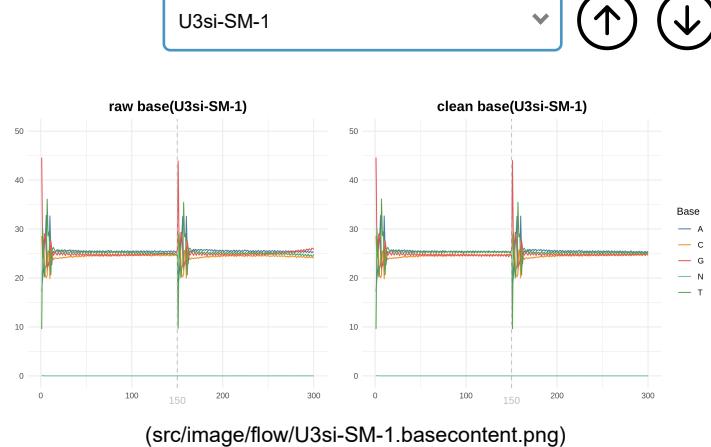


图 5 数据碱基分布图

横轴为序列每个位置的碱基，纵轴表示百分比，图中5条线分别代表A, C, T, G, N在每个位置上的平均含量。由于随机引物扩增偏好，以及测序仪开始状态不稳定使得前边一些碱基的碱基含量不平衡。

左图为 raw data, 右图为样品 clean data, 分割线左边为 read 1, 右边为 read 2。

3.2. 碱基错误率分布图



图 6 碱基错误率分布图

横轴为序列每个位置的碱基，纵轴表示错误率。由于整个测序过程耗时较长，荧光噪音以及酶活性及试剂的有效性会随着时间的延长而降低导致错误率逐渐增高。

左图为 raw data, 右图为样品 clean data, 分割线左边为 read 1, 右边为 read 2。

3.3. 数据统计

对原始数据和过滤处理得到高质量的数据 (clean reads) 进行数据统计，得到数据基本信息。

查看详细结果请点击: [Filter_Report.html](#) (src/html/Filter_Report.html)

表 3 质量统计表



广州艾基生物技术有限公司 Sample
GUANGZHOU IGE BIOTECHNOLOGY LTD

- 一、项目信息
- 二、概述
- 三、工作流程
- 四、结果展示
- 五、参考文献

Sample		RawReads	RawBases	CleanReads		C
U3si-SM-1	46,665,222	6,999,783,300	46,092,074	6,849,538,197	97.27%	92.33
U3si-SM-2	46,663,520	6,999,528,000	46,033,152	6,844,187,485	97.08%	91.91
T24si-SM-1	46,671,204	7,000,680,600	46,178,578	6,854,274,775	97.46%	92.70
T24si-SM-2	46,666,666	6,999,999,900	46,077,404	6,845,700,754	97.27%	92.33
T24si-Ctrl	42,364,560	6,354,684,000	41,874,546	6,222,474,774	97.36%	92.52
U3si-Ctrl	46,663,010	6,999,451,500	46,181,008	6,843,004,666	97.68%	93.19

Show 10 ✓ entries

Showing 1 to 6 of 6 entries

Previous 1 Next

统计说明：

1. **Sample** : 样品名称;
 2. **Rawreads** : 原始测序 reads 数量;
 3. **RawBases** : 原始测序数据的总碱基数;
 4. **CleanReads** : 将 Raw Reads 过滤得到的 reads 数量;
 5. **CleanBases** : 过滤得到的数据的总碱基数;
 6. **Q20** : 测序错误率小于 1% 的碱基数目占总碱基数比例;
 7. **Q30** : 测序错误率小于 0.1% 的碱基数目占总碱基数比例;
 8. **GC** : 碱基 G 和 C 的数量占总的碱基数量的百分比;

4. 参考序列比对分析

将 Clean reads 数据使用 Hisat2^[4] 软件 比对到参考基因组。比对后的统计结果见下表。

查看详细结果请点击: [Statistic Mapping.html](#)

(src/html/Statistic_Mapping.html)

4.1. 比对统计

表 4 比对结果统计表

Copy	CSV	Excel	PDF	Column visibility	Search:	
SampleID	Input reads	Mapped reads	non-unique	unique	Read-1	
U3si-SM-1	46092074	44013996	1496971	42517025	21503251	21
U3si-SM-2	46033152	43870572	1574488	42296084	21433873	20
T24si-SM-1	46178578	44288407	1500956	42787451	21583593	21
T24si-SM-2	46077404	44023572	1532423	42491149	21482550	21
T24si-Ctrl	41874546	39981229	1541551	38439678	19419711	19
U3si-Ctrl	46181008	43919401	1558664	42360737	21331575	21

Show 10 entries

Showing 1 to 6 of 6 entries

Previous 1 Next

**一、项目信息**

- 二、概述
- 三、工作流程
- 四、结果展示及说明
- 五、参考文献

统计说明：

1. **SampleID**：样本名称；
2. **Input reads**：经过过滤后的读段的数量统计(clean data)；
3. **Mapped reads**：能比对到参考基因组上的读段的数量统计；
4. **non-unique**：在参考基因组上有多个比对位置的读段的数量统计；
5. **Unique**：在参考基因组上有唯一比对位置的读段的数量统计；
6. **Read-1, Read-2**：双端测序读段中左右两端能比对到参考基因组上的reads数量的分别统计；
Read-1与Read-2应该大体相同；
7. **Reads map to '+', Reads map to '-'**：比对到参考基因组正链和负链的读段的数量统计；
8. **Concordant pair alignment rate**：比对后具有正确的相对方向和距离的read pairs (读段对)数量占所有read pairs数量的百分比；
9. **Mapped rate**：reads整体比对率；

4.2. 转录组质量评估

利用mapped reads使用RseQC软件^[5]对本次转录组测序进行整体质量评估，主要包括Reads全基因组覆盖度分布,测序饱和度(表达水平和junction site)、测序覆盖度以及不同区域Reads分布。

全基因组覆盖度分布	测序饱和度 (表达水平)	测序饱和度 (junction site)	测序均一性
基因元件分布			

如果物种的基因组已拼接到染色体水平，选取全部染色体（当该物种的基因组拼接为 scaffold 水平，选取20条 scaffold），对 Total mapped reads 的比对到基因组上的各个染色体的深度进行统计，如下图所示，具体作图的方法为用滑动窗口(window size) 为5K，计算窗口平均覆盖深度，并通过log2转换，得到最终的作图数值，由于RNA-seq存在过表达基因，将覆盖深度的分位数95%限定为最高阈值。

U3si-SM-1
▼
↑
↓



广州艾基生物技术有
GUANGZHOU IGE BIOTECHNC

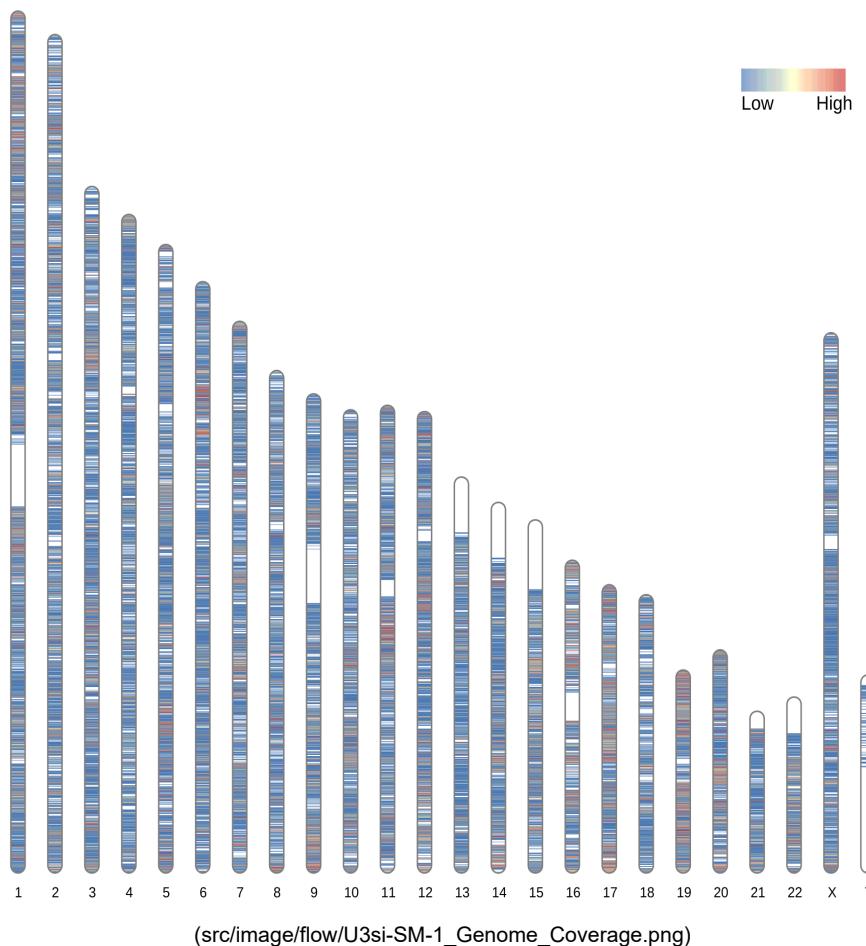


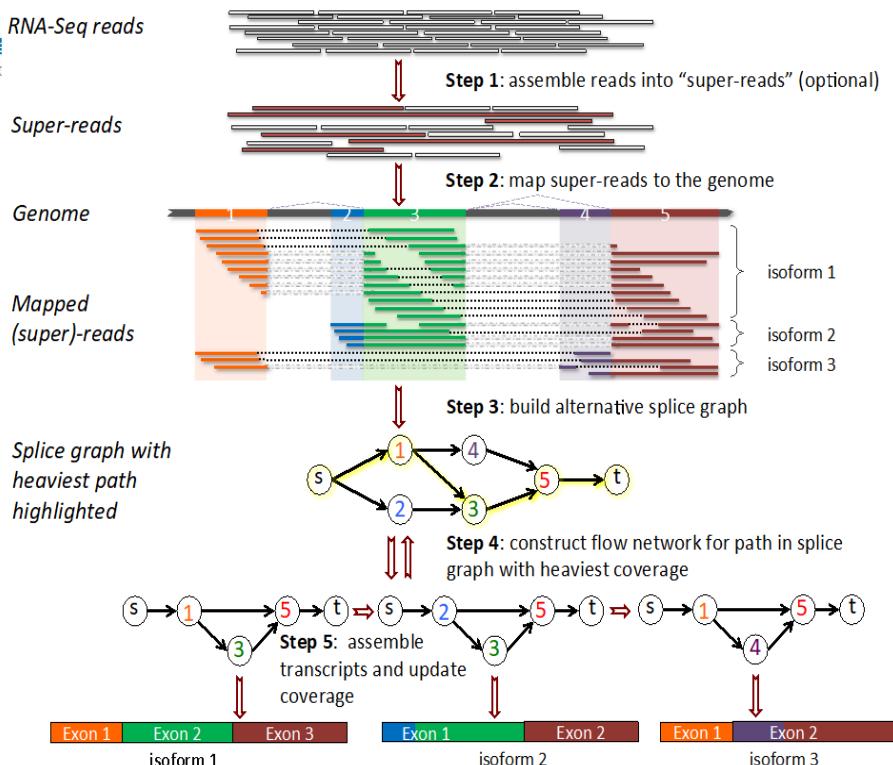
图 7 Reads 在染色体上的覆盖深度分布图

每条边角点（ (x_i, y_i) ）偏红色区域为平均深度高的区域，偏蓝色为低的区域。没有覆盖的区域为白色。

5 新基因和转录本预测

5.1. 拼接原理

对于非模式物种，有大量的新基因有待挖掘。我们采用StringTie^[6]软件进行新基因的拼接分析。StringTie是集转录本组装、转录本定量为一体的软件。首先StringTie针对低表达的基因组装的准确度和敏感度都要高于其他拼接软件；其次，就拼接效果而言，StringTie预测得到的新基因数目最多且在其他软件中都有重叠；最后，StringTie在基因重构方面对低冗余、高exon数目、多重转录本的基因更有效。拼接原理如下所示。



(src/image/stringtie.png)
图 12 stringtie 拼接原理图

5.2. 新基因和转录本预测原理

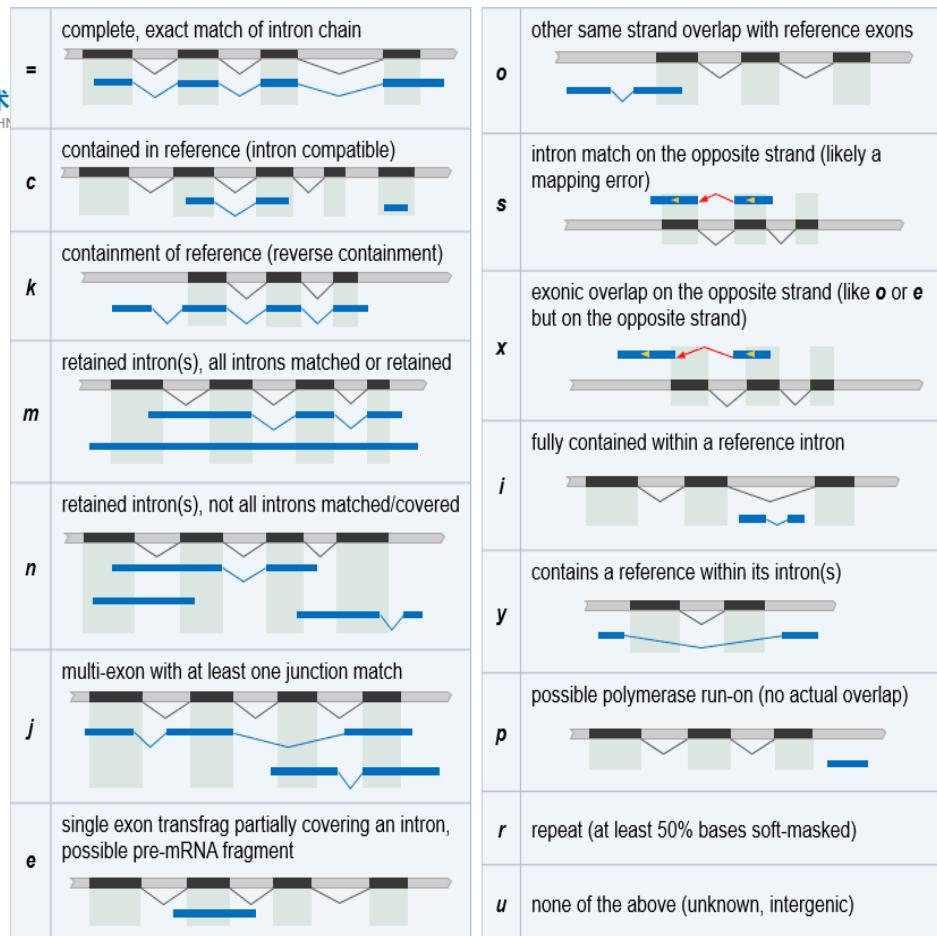
由于在组装的过程中产生了大量的新的转录本信息，使用 gffcompare^[7]与已知的转录本注释文件进行比较，通过软件给出的class codes着手，该信息记录了每个转录本相对于已知转录本的位置信息。取‘x’, ‘i’, ‘j’, ‘u’, ‘o’作为新转录本的筛选标准，具体class codes见下图解释。



广州艾基生物技术
GUANGZHOU IGE BIOTECHN

一、项目信息

- 二、概述
 - 三、工作流程
 - 四、结果展示及说明
 - 五、参考文献



(src/image/gffcompare_codes.png)

注： =：认为是与参考注释一致的已知转录本； c：拼装出来的转录本包含于已知转录本中； j：潜在的新转录本或转录片段，至少有一个junction位点与参考转录本一致； e：潜在的mRNA前体的片段； i：完全落入参考转录本内含子区的转录本片段； o：与参考转录本的外显子有一定的交集； p：可能是聚合酶延长产生的转录片段； r：拼接得到的转录本中50%的碱基处于DNA重复区域； u：未知的，基因区间区的转录本； x：与参考转录本外显子所处链的反义链有交集； s：转录片段的内含子与参考转录本内含子所处链的反义链有交集（可能产生于read mapping 错误）； :: 属于复合类型 (tracking file only, multiple classifications)； 其中'x','i','j','u','o'几种类型表示潜在的新转录本（只有code为'u'的转录本或code带'x'转录本所对应的基因定义为新基因）。

5.3. 新基因和转录本展示

本次结果中预测得到 15043 个新转录本和 419 个新基因,新基因^[8]和转录本预测结果如下所示。

表 5 新基因和转录本 (前50)

Trans	pos	class_code	ref_trans	gene_id
MSTRG.316.5	1:16104-199872:-	j	ENST00000488147	MSTRG.316 ENSG0
MSTRG.316.6	1:16104-199872:-	j	ENST00000488147	MSTRG.316 ENSG0
MSTRG.316.4	1:15961-20959:-	j	ENST00000488147	MSTRG.316 ENSG0
MSTRG.316.3	1:15930-199872:-	j	ENST00000488147	MSTRG.316 ENSG0
MSTRG.316.2	1:14054-199872:-	j	ENST00000488147	MSTRG.316 ENSG0
MSTRG.316.1	1:13482-29354:-	j	ENST00000488147	MSTRG.316 ENSG0
MSTRG.317.8	1:185278-199872:-	j	ENST00000623083	MSTRG.317 ENSG0
MSTRG.317.9	1:185278-199872:-	j	ENST00000623083	MSTRG.317 ENSG0

广州艾基生物技术有限公司 GUANGZHOU IGE BIOTECHNOLOGY LTD	
	Show 10 entries
Showing 1 to 10 of 49 entries	Previous
	1 2 3 4 5 Next
一、项目信息	
二、概述	统计说明：
三、工作流程	1. Trans ：新预测转录本id；
四、结果展示及说明	2. pos ：新预测转录本位置信息，格式：染色体:起始位置-终止位置:正负链；
五、参考文献	3. class_code ：据拼接转录本与已知转录本overlap关系,具体见上方解释；
	4. ref_trans ：参考转录本；
	5. gene_id ：新预测转录本对应的基因 ID；
	6. gene_pos ：新预测转录本对应的基因位置信息；
	7. ref_gene ：参考转录本对应的参考基因；

6. 基因表达水平分析

对于RNA-seq而言，由于技术误差、测序深度、基因长度、RNA组分不同，为了能够比较不同的样本，比较不同的基因的表达量，以及使表达水平分布符合统计方法的基本假设，就需要对原始数据进行标准化。

6.1. 基因表达结果

Salmon是不基于比对计数而直接对转录本进行定量的工具，适用于转录组、宏基因组等的分析，我们使用Salmon^[9] (<https://combine-lab.github.io/salmon/>) 对转录本/基因层面表达水平进行定量分析，以便后续分析不同样本间差异表达情况，并可通过结合功能信息，揭示基因的调控机制。

衡量表达水平的标准：

1) TPM (Transcripts Per Million reads) : 即每百万读段中来自于某转录本的读段数。TPM在标准化的过程中考虑到了测度深度和基因长度的问题，可以更直观地进行样品间和样品内的表达量比较。

TPM具体的计算公式如下：

$$\text{TPM} = A \times \frac{1}{\sum(A)} \times 10^6$$

图 14 TPM计算公式

2) TMM (Trimmed mean of Mvalues) : M-值的加权截尾均值^[10]。这个方法假设大部分基因的表达是没有差异的。简单来说就是先找到一个参考样品，每个样品对该参考样品计算一个文库标准化因子（NormFactor）（计算每个基因相对于参考基因表达倍数的log2值，叫M值，截掉最高最低的30%的M值，剩下的M值计算加权平均值作为标准化因子），从而得到标准化文库大小，这里最终得到的文库大小是消除了RNA组分而导致基因count偏倚情况。注：由于TMM值不考虑基因长度，所有TMM只能在样品之间进行比较。

后续展示通通以TMM值为例!

$$TMM \text{ Normalized Counts} = \frac{\text{Raw Count} * 10^6}{\text{Library size} * \text{NormFactor}} \quad (\text{src}/\text{image}/\text{TMM_cal.png})$$

图 15 TMM计算公式

这里展示部分基因/转录本展示其在所有样品中的**TMM值**, 见下表!



Gene Transcript

表 6 基因表达水平(TMM)统计表(前50)

Column visibility

Search: **一、项目信息**

二、概述

三、工作流程

四、结果展示及说明

五、参考文献

gene_id	U3si-SM-1	U3si-SM-2	T24si-SM-1	T24si-SM-2	T24si-C
ENSG00000227629	0	0	0	0	0
ENSG00000224240	0	0	0	0	0
ENSG00000215506	0	0	0	0	0
ENSG00000233843	0	0	0	0	0
ENSG00000234744	0	0	0	0	0
ENSG00000172283	0	0	0	0	0
ENSG00000237546	0	0	0	0	0
ENSG00000230854	0	0	0	0	0
ENSG00000235511	0	0	0	0	0

Show entries

Showing 1 to 10 of 49 entries

Previous 2 3 4 5 Next**6.2. 样本表达量分布图**

小提琴图的目的是比较不同组的基因表达水平的密度估计 (density estimates)。小提琴图的图形越宽处表示处于该表达水平的基因越多。小提琴图比箱线图更优之处在于它体现了数据的完整的概率分布，可从中看到是否有多峰及其幅度。

Gene (TPM) Gene (TMM)

图中纵坐标是基因的log2(TPM)值。横坐标表示各样品。小提琴图的中央是箱线图，中间横线是中位数，上下沿表示四分位数。

注：当下图的样品TPM表达量中小提琴图出现偏倚时，基本是RNA组分导致，建议不再用TPM，而是用TMM做后续的基因表达的衡量指标。



广州艾基生物技术
GUANGZHOU IGE BIOTECHN

- 一、项目信息
 - 二、概述
 - 三、工作流程
 - 四、结果展示及说明
 - 五、参考文献

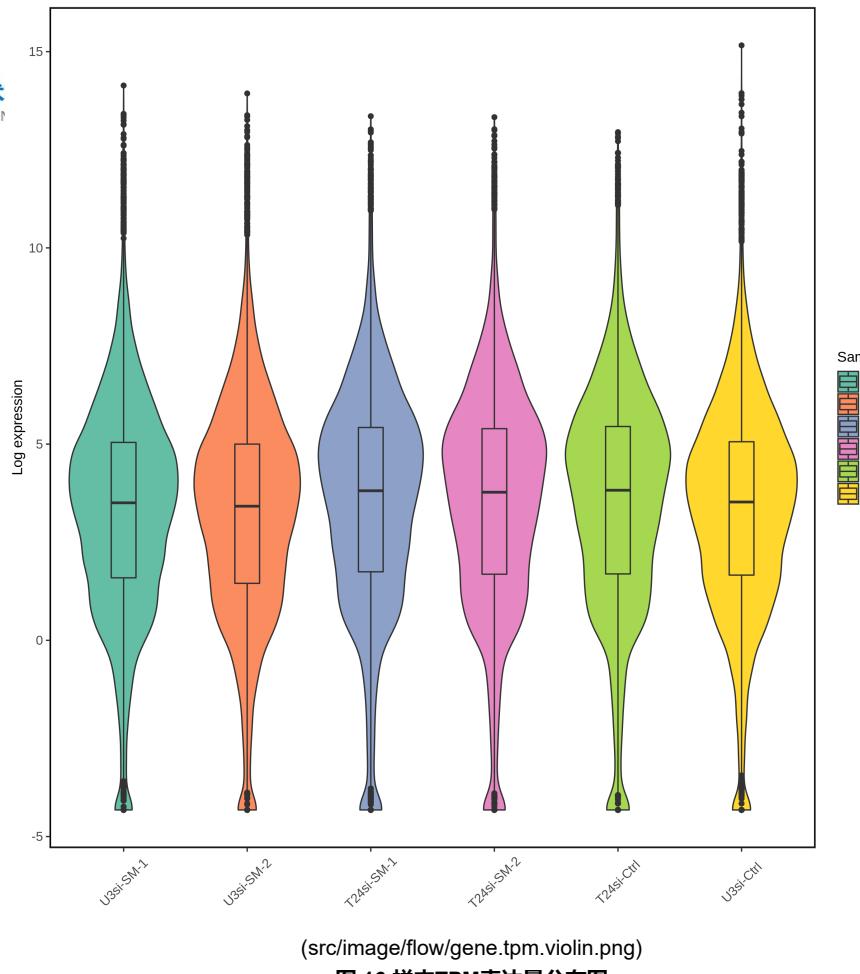


图 16 样本TPM表达量分布图

6.3. 样本PCA图

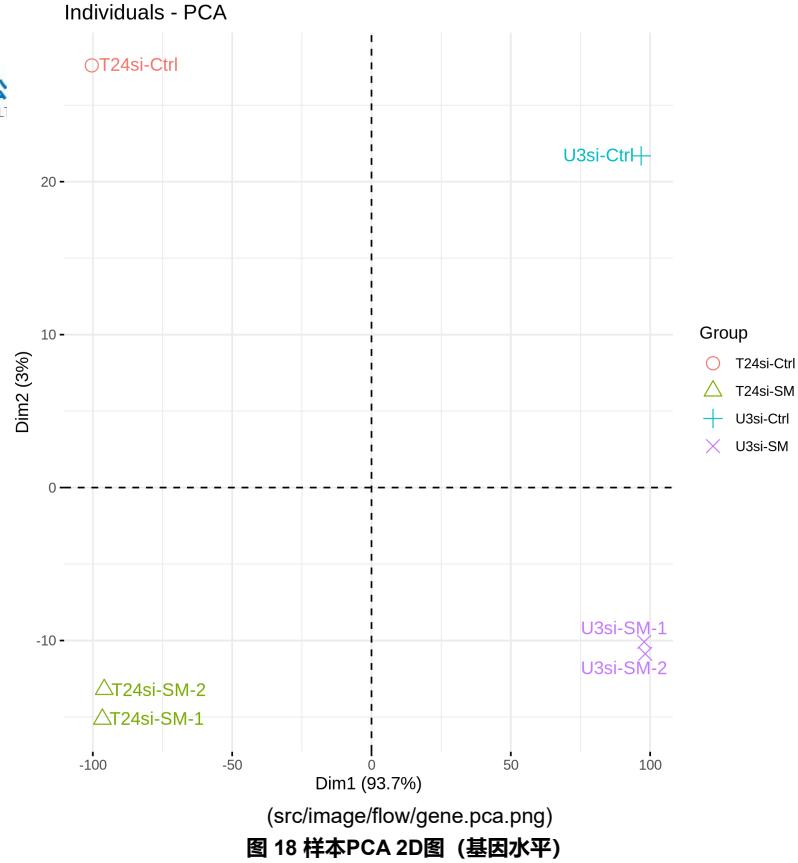
在高通量测序的分析中，样品之间的关系由众多变量决定。主成分分析是设法将原来众多具有一定相关性的指标 (reads 的分布特征)，重新组合成一组新的互相无关的综合指标，从而降低问题的复杂性，来研究样品间的主成分关系。二维 PCA 分析结果中，会展示主成分 1(PC1) 和主成分 2(PC2) 分别作为 X 轴和 Y 轴的散点图，每个点代表 1 个样本。

图中，如果两个样本距离越远，则说明两个样本 reads 分布的差异越大。反之，则说明相应样本 reads 整体分布模式越接近。所以，PCA 分析常用于评估样本重复性的好坏。理想情况下，生物学重复的样本应该聚类在一起，而处理组间应该可以清晰区分开。

Gene(2D)	Gene(3D-static)	Gene(3D-dynamic)	Transcript(2D)	Transcript(3D-static)
		Transcript(3D-dynamic)		

**一、项目信息**

- 二、概述
- 三、工作流程
- 四、结果展示及说明
- 五、参考文献



7. 差异表达分析

差异表达分析是指找出在不同的样品组间表达量差异显著的基因/转录本，分组可能是不同的生物学状态，例如药物处理与对照、疾病个体与健康个体、不同组织、不同发育阶段等等。

7.1. 差异表达分析结果

生物学角度看，一个基因必须有一定的表达量，才能够转录、翻译蛋白质并有生物学意义；从统计角度看，一个过低表达的基因不能够获得显著性差异，不具有统计学意义

低表达基因过滤：以一个差异比较包含的样品为基础，**CPM > 0.5样品至少占样品半数，或者 CPM > 0.5 的样品占某个重复组样品三分之二**（说明：CPM全称为count-per-million，使用CPM可以消除测序量差异，一个基因的count数，至少在某些样本中应达到10-15，其CPM 0.5正好表示在6G数据量时count值为10）

差异表达分析使用 DESeq2^[11](有重复样品) 或 edgeR^[12](无重复样品)。从统计学意义的角度上考虑，用p value或矫正后的p value (padj) 判断显著性水平。Padj是当假阳性率较高时利用BH的方法对p-value进行多重检验校正得到的值。默认使用pvalue < 0.05, |log2FoldChange| > 1作为差异显著性标准。

这里展示基因和转录本两个水平的部分差异分析结果，见下表！

Gene Transcript

表 8 差异表达基因(前50)

Copy CSV Excel PDF Column visibility
Search:

id	T24si-SM-1	T24si-SM-2	T24si-Ctrl	log2FoldCI
----	------------	------------	------------	------------

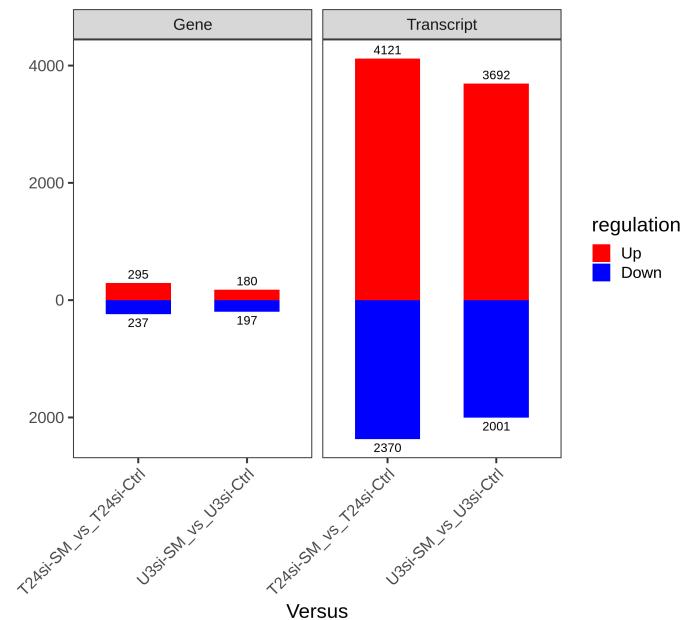
3. DET(All|Up|Down) : 所有差异转录本|上调转录本|下调转录本;



广州艾基生物技术有限公司
GUANGZHOU IGE BIOTECHNOLOGY

一、项目信息

- 二、概述
- 三、工作流程
- 四、结果展示及说明
- 五、参考文献



(src/image/flow/DE_Statistic_bar.png)

图 24 差异基因/转录本统计图

7.3. 差异基因/转录本火山图

火山图可直观显示表达差异显著性基因/转录本的整体分布情况。

Gene Transcript

T24si-SM_vs_T24si-Ctrl

▼





- 一、项目信息
 - 二、概述
 - 三、工作流程
 - 四、结果展示及说明
 - 五、参考文献

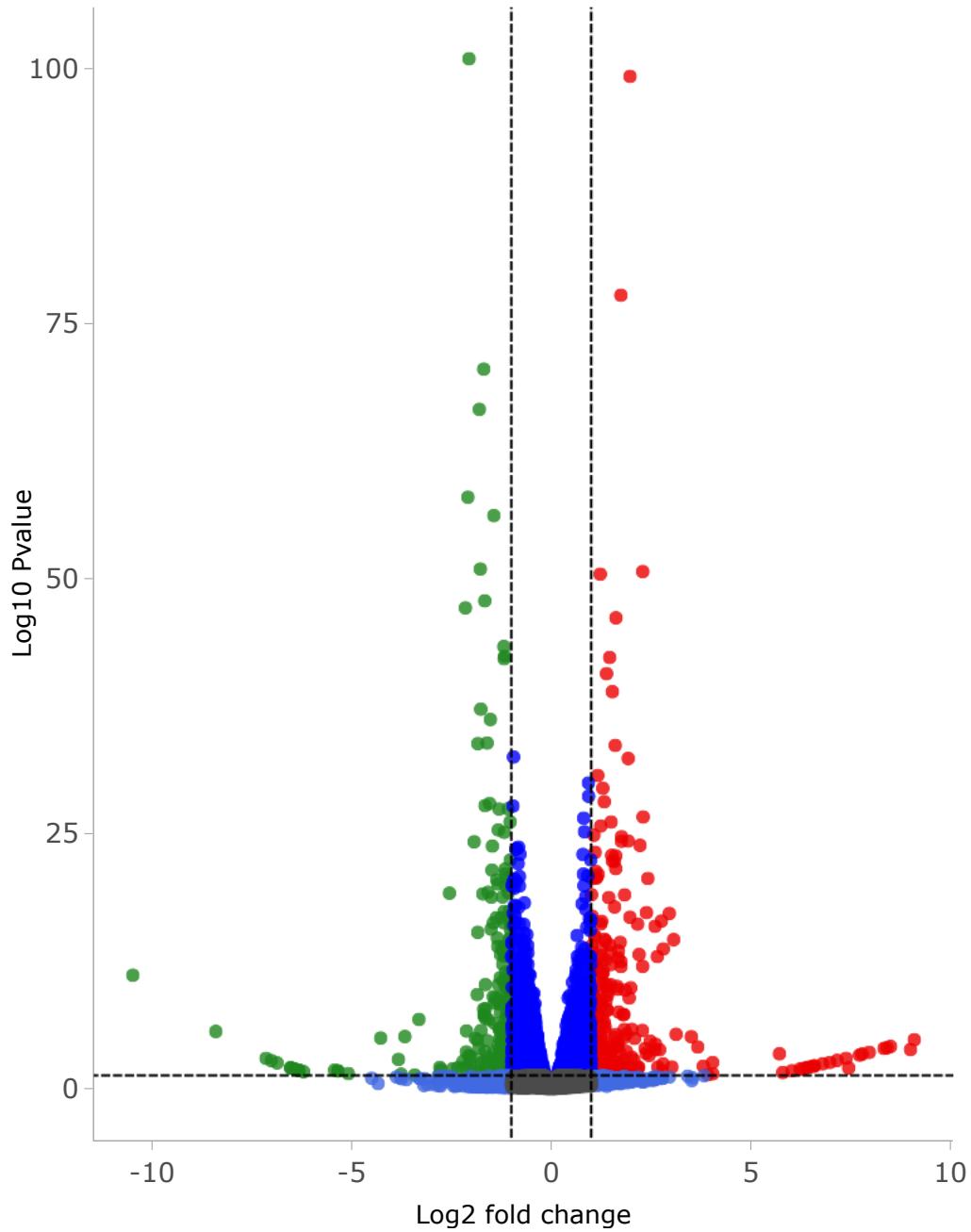


图 25 差异基因火山图

横坐标表示基因/转录本在不同样本或比较组合间的表达倍数变化($\log_2\text{FoldChange}$)，横坐标的绝对值越大表明两个比较组合之间的表达变化倍数越大；纵坐标表示表达差异的显著性水平；表达上调基因/转录本用红色点表示，下调基因/转录本用绿色点表示，蓝色点为通过 p 值阈值不通过 $\log_2\text{FoldChange}$ 阈值的基因/转录本，浅蓝色点为通过 $\log_2\text{FoldChange}$ 阈值而不通过 p 值阈值的基因/转录本，灰色表示两者都不通过的基因/转录本。

7.4. 差异基因/转录本聚类热图

聚类图是差异表达基因的另一种展示方式，将表达模式相近的基因聚在一起，这些基因可能具有共同的功能或参与到共同的代谢途径及信号通路中。采用主流的层次聚类法，将TMM值z-score标准化并进行聚类。聚类结果用热图（heatmap）表示。

Gene Transcript

T24si-SM vs T24si-Ctrl



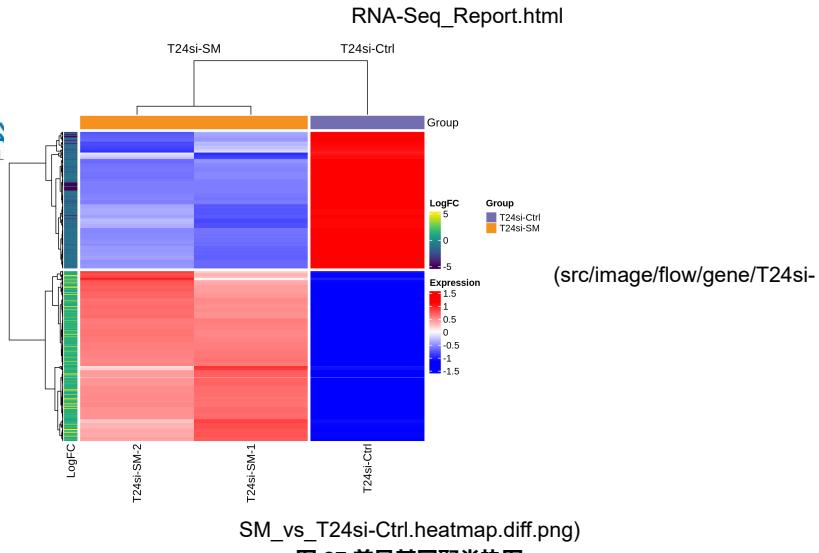
**一、项目信息**

二、概述

三、工作流程

四、结果展示及说明

五、参考文献

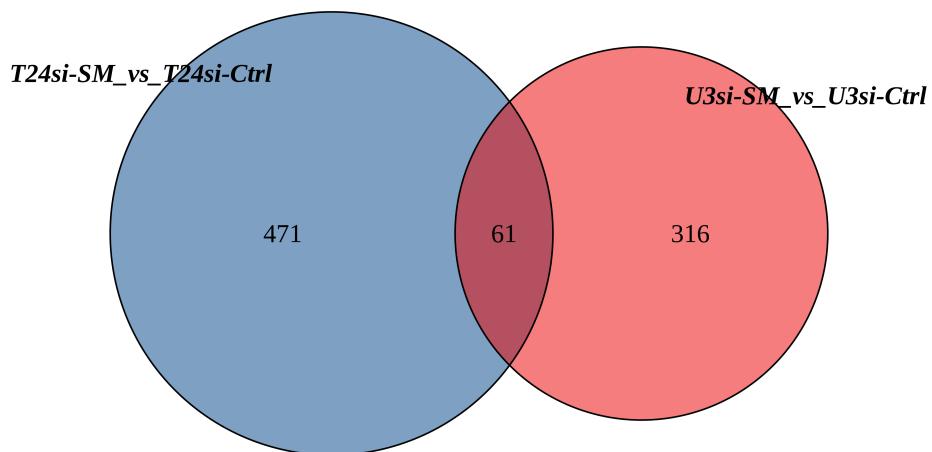
**图 27 差异基因聚类热图**

横坐标为样本，纵坐标为差异基因/转录本，左侧根据表达相似程度对基因/转录本进行聚类，由蓝至红表达量逐渐上调。

7.5. 差异表达基因/转录本韦恩图

以各组比较得到的差异表达基因作出的韦恩图如下所示。韦恩图中最少包含2组、最多包含5组比较。若超过5组比较，以Upset图展示。

Gene **Transcript**



(src/image/flow/DEG_Vennplot.png)

图 29 差异表达基因 venn/upset 图**8. 差异表达基因Gene Ontology 分析**



一、项目信息

二、概述

三、工作流程

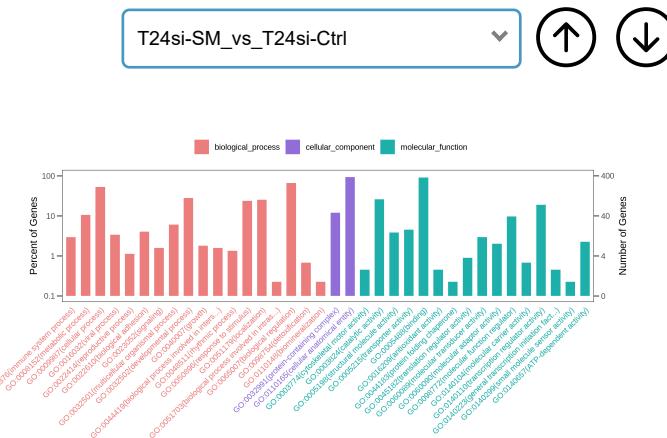
四、结果展示及说明

五、参考文献

8.1. 差异基因 GO 注释

Gene Ontology^[13]可分为分子功能（Molecular Function），生物过程（Biological Process）和细胞组成（Cellular Component）三个部分，它常用于提供基因功能分类标签和基因功能研究的背景知识。通过物种和基因信息，用Gene Ontology数据库进行查找，从而得到基因的GO注释信息（功能信息）。

注：报告中只展示所有差异基因的GO分析，最终结果中还包括上调、下调基因、差异转录本以及上调、下调转录本GO分析



(src/image/flow/T24si-SM_vs_T24si-Ctrl.DEG.go.anno.png)

图 31 差异基因GO level2分类图

横坐标为GO三个大类的下一层级的GO term，纵坐标为注释到该term下的子term的差异基因数目，3种不同分类用不同颜色表示。

8.2. 差异基因 GO 富集

Gene Ontology（简称 GO, www.geneontology.org）是基因功能国际标准分类体系。根据实验目的筛选特异基因后，研究特异基因在 Gene Ontology 中的分布状况将阐明实验中样本差异在基因功能上的体现。^[14]

8.2.1 差异基因 GO 富集结果

表 11 差异基因 GO 富集结果 (前50)

Copy	CSV	Excel	PDF	Column visibility	Search:
ID	Ontology		Description		GeneRatio
GO:0000278	biological_process		mitotic cell cycle		12/443
GO:0005930	cellular_component		axoneme		10/443
GO:0003341	biological_process		cilium movement		6/443
GO:0070593	biological_process		dendrite self-avoidance		4/443
GO:0017177	cellular_component		glucosidase II complex		2/443
GO:0070895	biological_process		negative regulation of transposon integration		2/443
GO:0042088	biological_process		T-helper 1 type immune response		3/443



Show 10 entries

Showing 1 to 10 of 49 entries

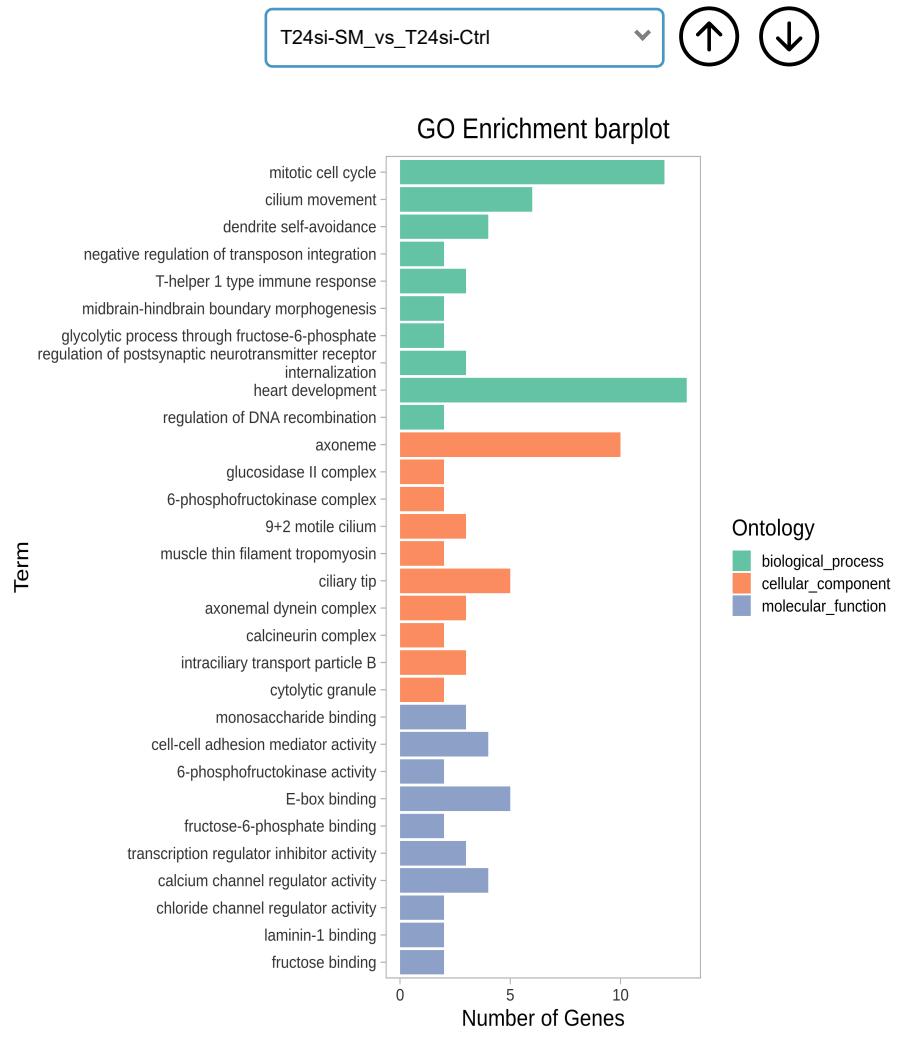
Previous 1 2 3 4 5 Next

统计说明：

1. **ID** : GO 条目名称;
2. **Description** : GO 的描述信息;
3. **GeneRatio** : 属于该GO分类下的的差异基因与所有GO分类下的差异基因比值;
4. **BgRatio** : 属于该GO分类下基因数目与所有GO分类下的基因比值;
5. **pvalue** : 富集分析统计学显著水平, 一般情况下, p-value < 0.05 该功能为富集项;
6. **Count** : 属于该GO条目的差异基因的数量;
7. **genelD** : 属于该GO条目的差异基因ID;
8. **Input Hyperlink** : 该GO分类的链接;
9. **Classification** : BP(Biological process)生物过程, CC(cellular component)细胞组分, MF(molecular function)分子功能;
10. **Count** : 属于该GO条目的差异基因的数量。

8.2.2 差异基因 GO 富集柱状图

将所有 GO 富集结果, 分别对细胞组分 (cellular component), 分子功能 (molecular function), 生物过程 (biological process) 按照 pvalue 从小到大排序各取前10个 GO Term绘制柱状图。



(src/image/flow/gene/T24si-SM_vs_T24si-Ctrl.diff.GO_barplot.png)

图 32 差异基因 GO 富集柱状图



广州艾基生物技术有限公司
GUANGZHOU IGE BIOTECHNOLOGY LTD.

将所有 GO 富集结果，按照 pvalue 从小到大排序，取前 20 个 GO Term 绘制气泡图

一、项目信息

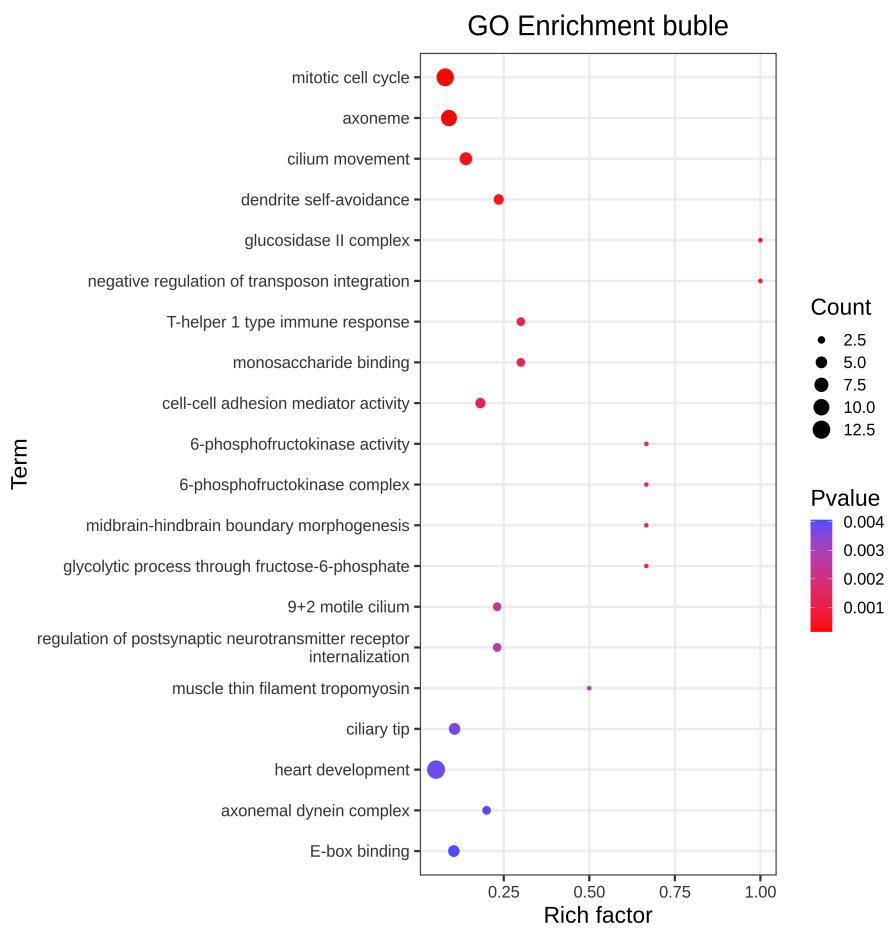
二、概述

三、工作流程

四、结果展示及说明

五、参考文献

T24si-SM_vs_T24si-Ctrl



(src/image/flow/gene/T24si-SM_vs_T24si-Ctrl.diff.GO_buble.png)

图 33 差异基因 GO 富集气泡图

纵轴表示Term名称，横轴表示Rich factor，点的大小表示此Term中差异基因个数多少，而点的颜色对应于不同的pvalue范围。

8.2.4 差异基因 GO 有向无环图

由于GO term间具有包含关系，GO term之间可以构建复杂的结构网络。将所有GO富集结果，分别对细胞组分（cellular component），分子功能（molecular function），生物过程（biological process）按照pvalue从小到大排序各取前10个GO Term绘制有向无环图（DAG）。

biological process **molecular function** **cellular component**

molecular function

cellular component

T24si-SM_vs_T24si-Ctrl





一、项目信息

- 二、概述
- 三、工作流程
- 四、结果展示及说明
- 五、参考文献



(src/image/flow/gene/T24si-

SM_vs_T24si-Ctrl.diff.GO_DAG_BP.png)

图 34 DAG (生物过程)

图中分枝代表包含关系，从上至下所定义的功能范围越来越小，每个节点代表一个GO Term，节点颜色越深， p 值越小。其中方型表示富集程度前10的GO Term，圆形代表未在前10的GO Term，每个GO Term标注了对应的GO ID、描述信息、富集分析检验的 p 值、以及该GO term下差异基因数目/term下基因总数。

9. 差异表达基因 KEGG 分析

- 1) 基因的KEGG^[15]生物通路注释
- 2) 显著性计算方法---超几何分布计算P值
- 3) 显著性校正方法---FDR 校正 (Q-value)

9.1. 差异表达基因 KEGG 注释

Kyoto encyclopedia of genes and genomes (KEGG) (<http://www.genome.jp/>)是一个包含生化反应、信号通路、代谢通路和生物学过程的数据库。对差异表达基因集进行基于KEGG数据库的生物通路富集分析，提取出差异表达基因显著富集的通路（pathway），有利于下游实验的开展。

注：报告中只展示所有差异基因的KEGG分析，最终结果中还包括上调、下调基因、差异转录本以及上调、下调转录本KEGG分析

T24si-SM_vs_T24si-Ctrl

▼





一、项目信息

- 二、概述
 - 三、工作流程
 - 四、结果展示及说明
 - 五、参考文献

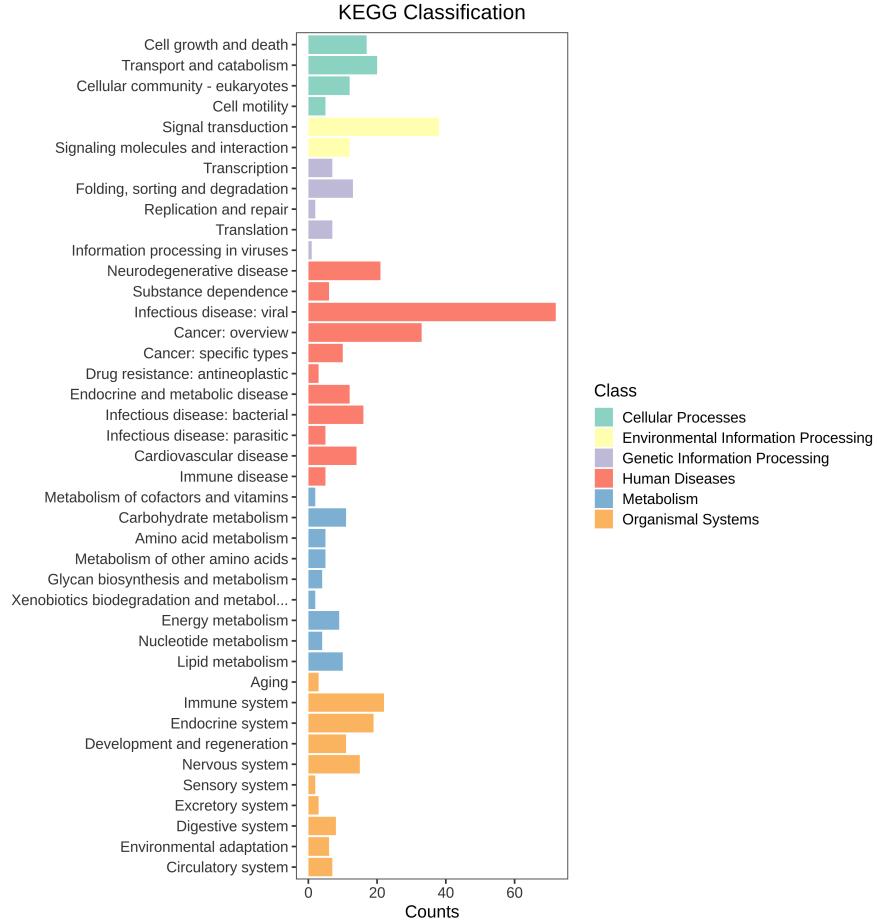


图 37 差异基因KEGG level2分类图

横坐标为注释到该term下（包括该term的子term）的差异基因数目，纵坐标为KEGG大类的下一层级的KEGG term。不同分类用不同颜色表示。

9.2. 差异表达基因 KEGG 富集

在生物体内，不同基因相互协调行使其生物学功能，通过 Pathway 显著性富集能确定特异基因参与的最主要生化代谢途径和信号转导途径。KEGG (Kyoto Encyclopedia of Genes and Genomes, <https://www.kegg.jp/> (<https://www.kegg.jp/>)) 是有关 pathway 的主要公共数据库。Pathway 显著性富集分析以 KEGG Pathway 为单位，应用超几何检验(hypergeometric test)，找出与整个基因组背景相比，在特异基因中显著性富集的 pathway，该分析的计算公式：

$$P = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i} \binom{N-M}{n-i}}{\binom{N}{n}} \quad (\text{src}/\text{image}/\text{Hyper.png})$$

图 38 超几何检验公式

N为背景基因数目, n为背景基因集中该通路的基因数目; M为差异基因数目,m为差异基因集中该通路的基因数目。

pvalue < 0.05 的 pathway 定义为显著富集的 pathway。其中**Rich factor=m/n**

9.2.1 差异表达基因 KEGG 富集结果

表 12 差异基因 KEGG 富集结果 (前50)

<input type="checkbox"/>	ID	Description	GeneRatio	BgRatio	pvalue
	hsa05168	Herpes simplex virus 1 infection	53/189	450/7528	1.504921832978301
	hsa04668	TNF signalling pathway	7/189	106/7528	0.01699193024198



广州艾基生物技术有限公司 hsa00565
GUANGZHOU IGE BIOTECHNOLOGY LTD

一、项目信息

二、概述

三、工作流程

四、结果展示及说明

五、参考文献

hsa01200	Carbon metabolism	7/189	112/7528	0.02230687962047
hsa00565 TECHNOLOGY LTD	Ether lipid metabolism	4/189	47/7528	0.0296529863651
hsa04370	VEGF signaling pathway	4/189	57/7528	0.05426212514929
hsa05014	Amyotrophic lateral sclerosis	14/189	349/7528	0.0556427774710
hsa00564	Glycerophospholipid metabolism	5/189	91/7528	0.0782300906102
hsa04662	B cell receptor signaling pathway	4/189	65/7528	0.0799466358075
hsa04659	Th17 cell differentiation	5/189	92/7528	0.0811287697967

Show 10 entries

Showing 1 to 10 of 49 entries

Previous 1 2 3 4 5 Next ▾

统计说明：

1. **ID** : KEGG通路编号;
 2. **Description** : KEGG通路编号对应的功能描述;
 3. **GeneRatio** : 属于该通路下的的差异基因与所有KEGG通路下的差异基因比值;
 4. **BgRatio** : 属于该通路下基因数目与所有KEGG通路下的基因比值;
 5. **pvalue** : 富集分析统计学显著水平, 一般情况下, $p\text{-value} < 0.05$ 该功能为富集项;
 6. **qvalue** : q统计值;
 7. **geneID** : 属于该通路的差异基因ID;
 8. **Count** : 属于该通路的差异基因的数量。

9.2.2 差异表达基因 KEGG 富集柱状图

将所有 KEGG 富集结果，按照 pvalue 从小到大排序，取前 20个 KEGG 富集功能绘制柱状图

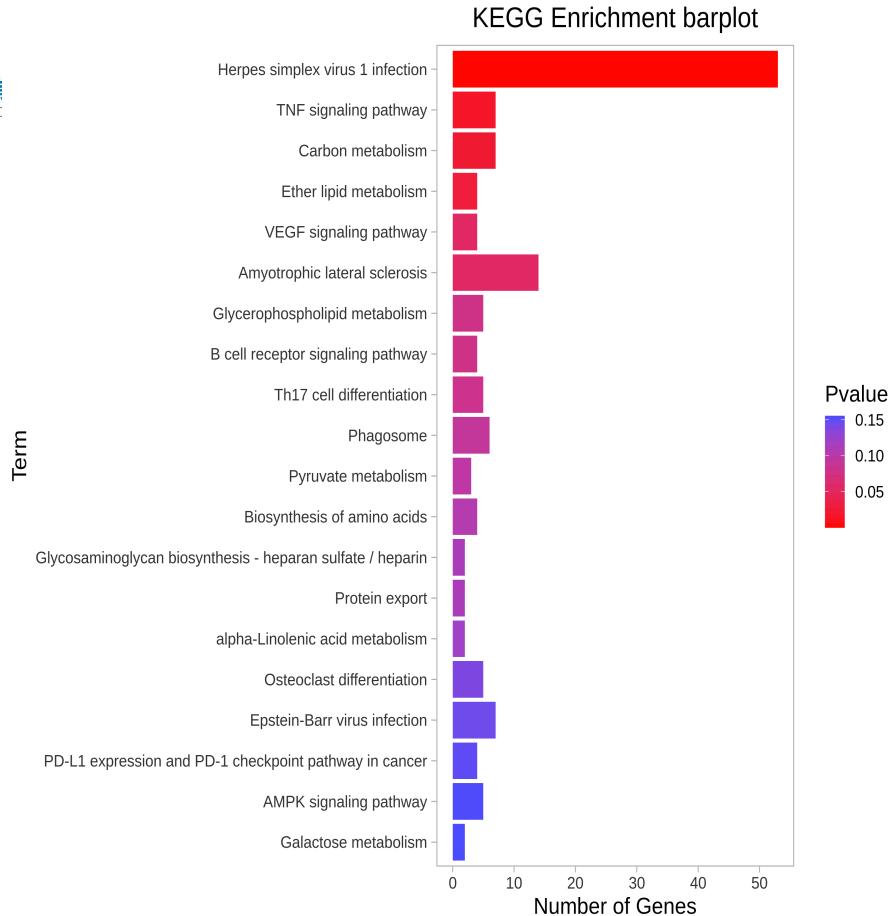
T24si-SM_vs_T24si-Ctrl





广州艾基生物技术有
GUANGZHOU IGE BIOTECHNC

- 一、项目信息
 - 二、概述
 - 三、工作流程
 - 四、结果展示及说明
 - 五、参考文献



(src/image/flow/gene/T24si-SM_vs_T24si-Ctrl.diff.KEGG_barplot.png)

图 39 差异基因KEGG 富集柱状图

9.2.3 差异表达基因 KEGG 富集气泡图

将所有 KEGG 富集结果，按照 pvalue 从小到大排序，取前 20个 KEGG 富集功能绘制气泡图

T24si-SM_vs_T24si-Ctrl

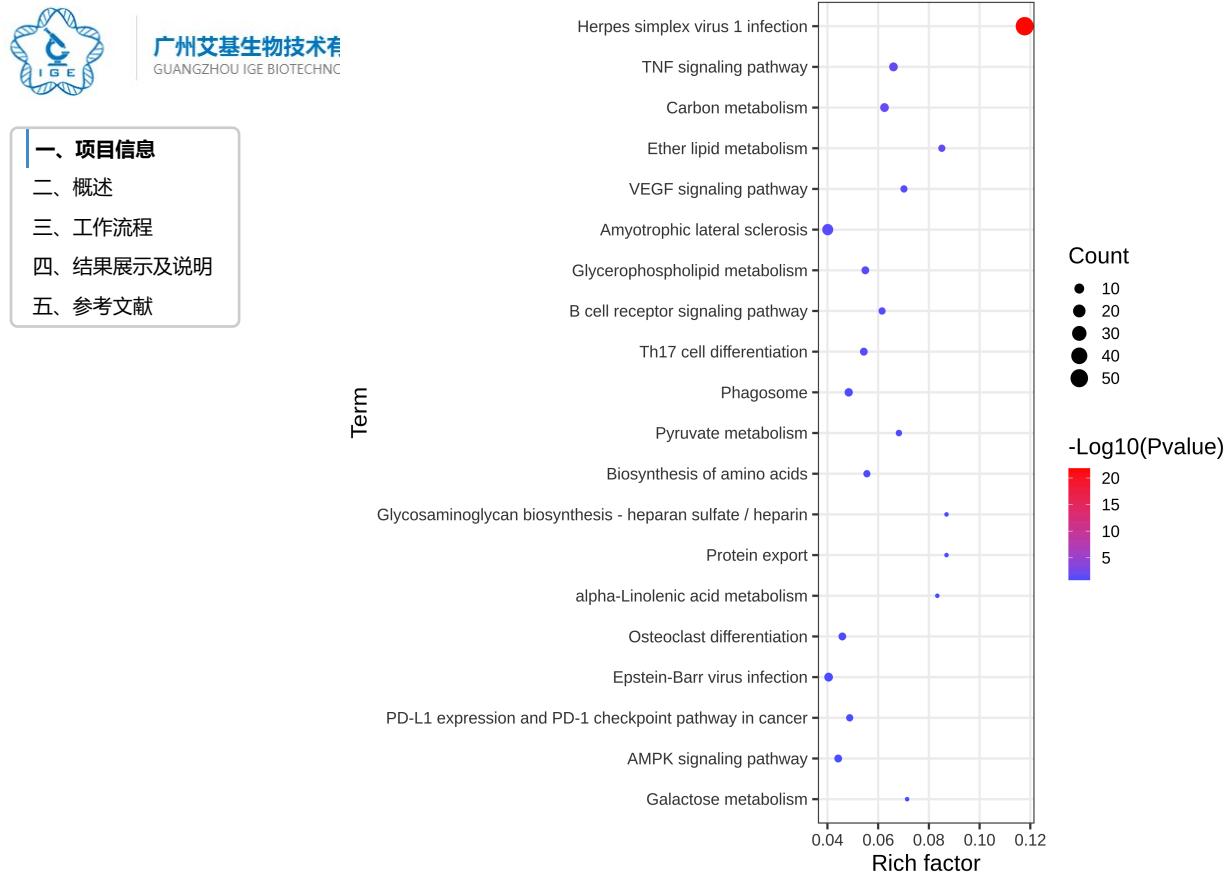


图 12 差异基因KEGG 富集气泡图

纵轴表示Term名称，横轴表示Rich factor，点的大小表示此Term中差异基因基因个数多少，而点的颜色对应于不同的pvalue范围。

9.2.4 差异表达基因 KEGG 富集通路图

为了便于查看差异基因在通路图中的分布情况，我们将差异基因标注到通路图中，查看方法如下：拿到全部分析结果后，打开结果文件中 KEGG_Enrichment 目录下，不同比较组的子目录下的 html 文件，点击不同的通路名即可看到通路图，包含差异转录本对应基因的 KO 节点边框标为红色。鼠标悬停于标记的 KO 节点，弹出差异基因细节框。以上步骤可脱机实现，如连接互联网，点击各个节点，可以连接到 KEGG 官方数据库中各个 KO 的具体信息页。



一、项目信息

- 二、概述
 - 三、工作流程
 - 四、结果展示及说明
 - 五、参考文献

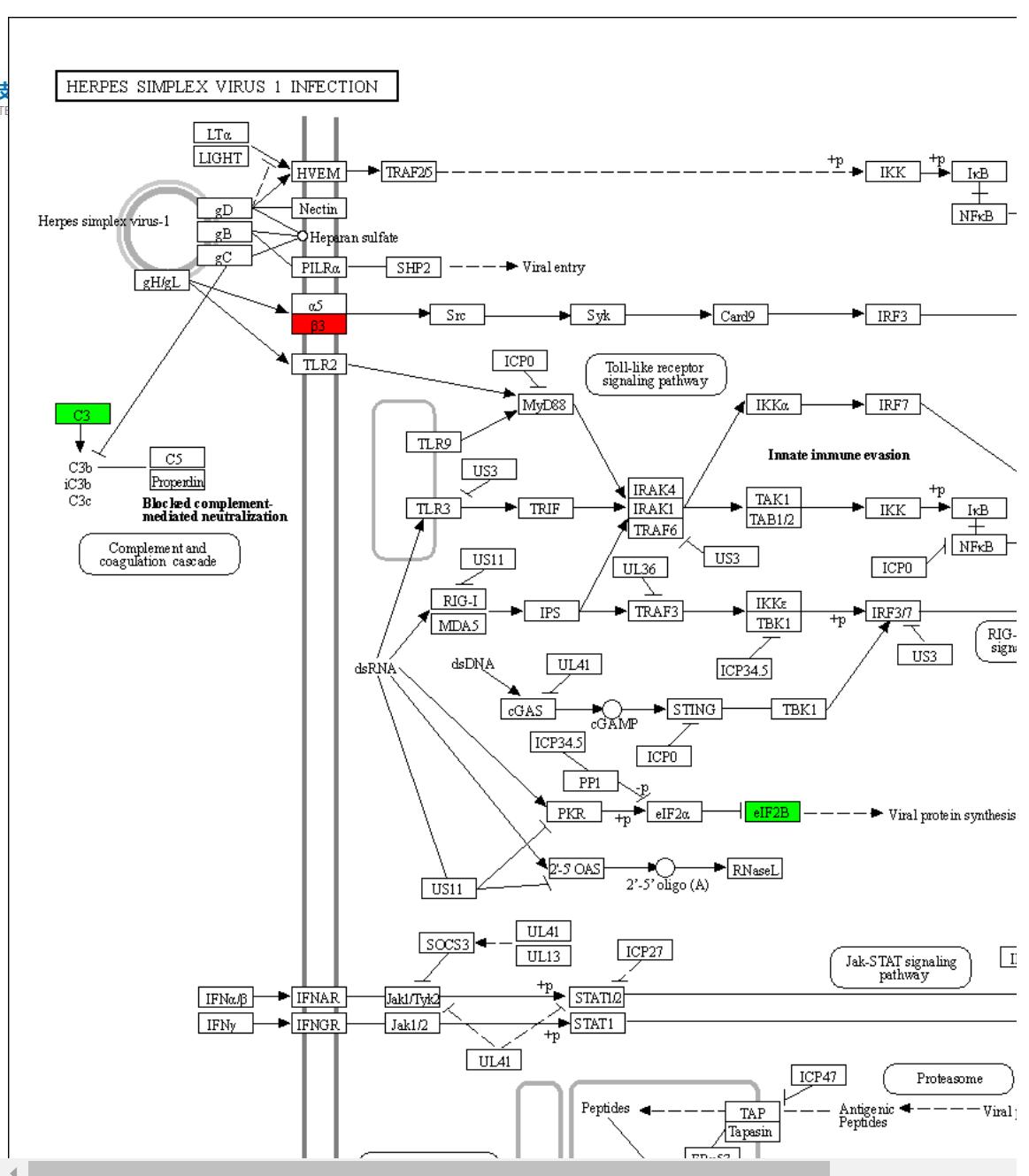


图 41 差异基因 KEGG 富集通路图

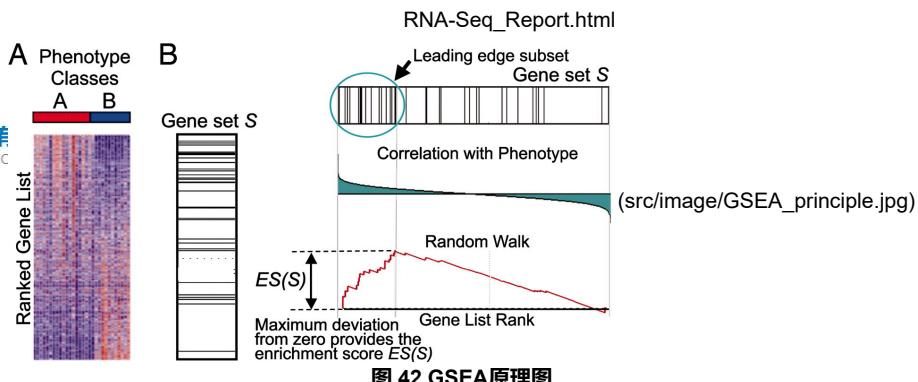
10. GSEA 基因集富集分析

基因集富集分析GSEA使用的基因数据源是我们实验组和对照组检测到的**所有基因**（无论是否被人为阈值判断为差异基因），将所有基因与预定义的GSEA基因集（类似pathway、GO这样的基因与功能对应关系）进行比较、富集，从而判断基因对表型(功能)的贡献。因此，从基因集的富集角度出发，GSEA不局限于是否为差异基因，理论上更容易发现一些对生物通路/功能有细微变化(基因倍数变化小)的影响。

GSEA原理见下图，若按照差异表达倍数从正到负排序后（Gene set S），我们参与某条通路的差异基因密集排列在排序表顶端（Leading edge subset），及显著上调，因此我们认为这条通路下的基因表达水平在实验处理后是显著上调，可能被激活。



- 一、项目信息
 - 二、概述
 - 三、工作流程
 - 四、结果展示及说明
 - 五、参考文献



10.1. GSEA 分析结果

我们对GO, KEGG和MsigDB (分子特征数据库) (<https://www.gsea-msigdb.org/gsea/msigdb/index.jsp>)进行GSEA分析,过滤其基因数小于10, 大于500的基因集, 保留**pvalue < 0.05**作为统计上显著性结果。

另：分析时只使用MSigDB 的两个典型的基因集合,MSigDB数据库分析目前只针对人和小鼠。

- H(hallmark gene sets)
 - C2:CP(curated gene sets, canonical pathways,BIOCATRA、PID、REACTOME、WIKIPATHWAYS)

这里展示部分分析结果，见下表！

表 13 GO GSEA分析结果(前50)

ID	Ontology	Description	setSize
GO:0005929	cellular_component	cilium	262
GO:0030030	biological_process	cell projection organization	169
GO:0097542	cellular_component	ciliary tip	43
GO:0060271	biological_process	cilium assembly	199
GO:0005930	cellular_component	axoneme	75
GO:0036064	cellular_component	ciliary basal body	136
GO:0042073	biological_process	intraciliary transport	29
GO:0035735	biological_process	intraciliary transport involved in cilium assembly	39
GO:0031514	cellular_component	motile cilium	88

统计说明：

1. **ID** : GO 编号;
 2. **Ontology** : GO 所属三大分类之一;
 3. **Description** : GO 功能描述;
 4. **setSize** : 基因集大小;
 5. **enrichmentScore** : 富集得分;
 6. **NES** : 标准富集得分;
 7. **pvalue** : 富集分析统计学显著水平, 一般情况下, $p\text{-value} < 0.05$ 该功能为富集项;
 8. **p.adjust** : 校正后的p值;



广州艾基生物技术有限公司
GUANGZHOU IGE BIOTECHNOLOGY LTD.

10. rank : 当ES值最大时

Leading_edge：该处有3个统计值，tags=xx% 表示核心基因占该基因集中基因总数的百分比；list=xx% 表示排序后的基因中ES peak左边的基因（es正值）或右边的基因（es负值）占所有

12. **core_enrichment** : 核心基因, 贡献大的基因子集 (ES为正值, ES peak 左边的基因或者ES为负值, ES peak 右边的基因)

一、项目信息

二、概述

三、工作流程

四、结果展示及说明

五、参考文献

10.2. GSEA lollipop和ridge分布图

这里对 GSEA 分析结果中激活及抑制最显著的前 20 个通路进行可视化，示例如下！

GO KEGG MsigDB

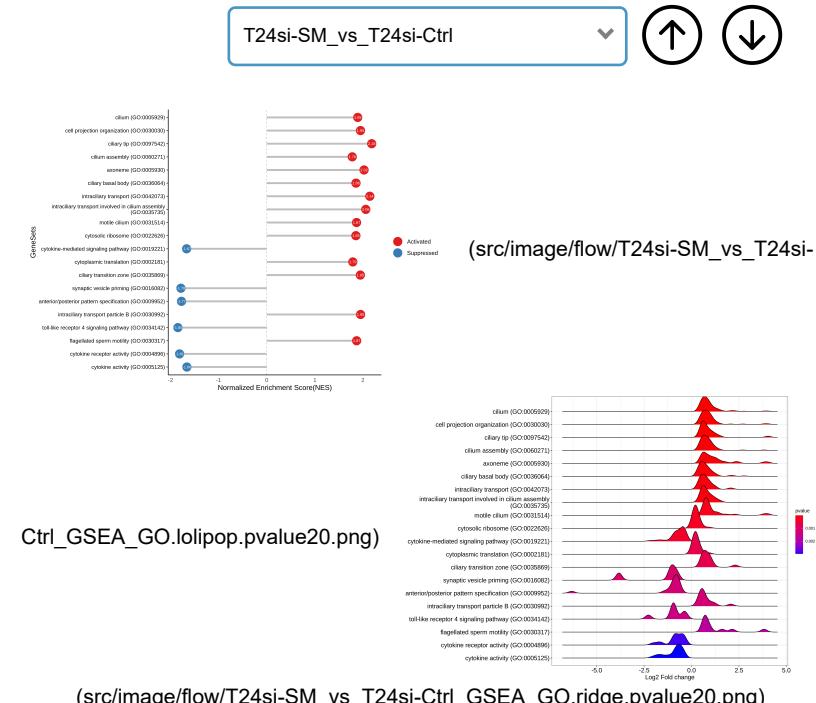


圖 43 GSEA GO lollipop&ridge 圖

左边为棒棒糖 (ollipop) 图，横坐标为标准化的富集得分(NES)，纵轴为显著的基因集，红色点为激活，蓝色为抑制。

右边为山脊(ridge)分布图，横轴为基因表达(此处为Log2 fold change)，颜色表示p值，越小颜色越偏红，这里使用的是基因集中core_enrichment(基因集中贡献大的基因子集)，图中可以看出核心基因的主要分布，从而推断通路上下调。

10.3. GSEA 富集图

结果对每个数据集取前10个通路进行富集图绘制，富集图示例如下！

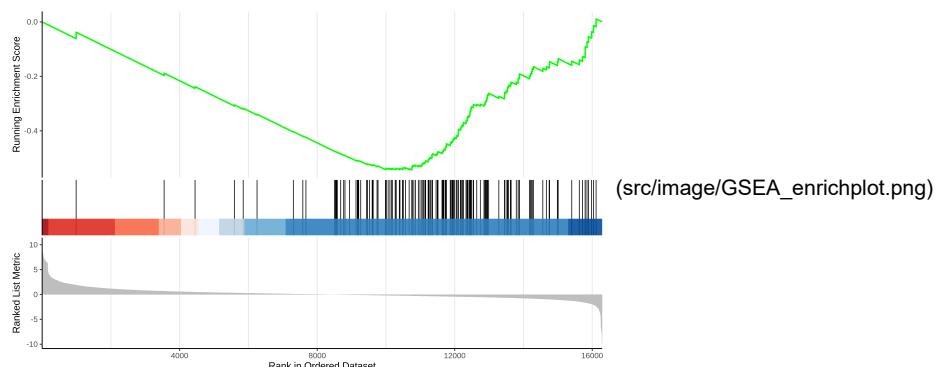


图 46 GSEA 富集示例图

图中分三部分：最上面是富集得分折线图（line chart），中间为基因排秩毛毯图（rug plot），最下面是基因log2fc排序的条形图。



广州艾基生物技术有限公司

GUANGZHOU IGE BIOTECHNOLOGY LTD

11. 样品间相关性分析

一、项目信息

二、概述

三、工作流程

四、结果展示及说明

五、参考文献

样品间基因表达水平相关性是检验实验可靠性和样品选择是否合理的重要指标，相关性系数越接近1，表明样品之间的表达模式的相似度越高，当样品中有生物学重复时，通常生物体重复间的相关性系数要求较高。

结果判定标准：相关系数 $R < 0.3$ 表示相关程度低，相关系数 $0.3 \leq R < 0.5$ 表示相关程度普通，相关系数 $0.5 \leq R < 0.7$ 表示相关程度显著，相关系数 $0.7 \leq R < 0.9$ 表示相关程度高，相关系数 $0.9 \leq R < 1.0$ 表示相关程度极高。Encode 计划建议 R 大于 0.92（理想的取样和实验条件下）。具体的项目操作中，我们建议重复性样品的 R 至少要大于 0.8。

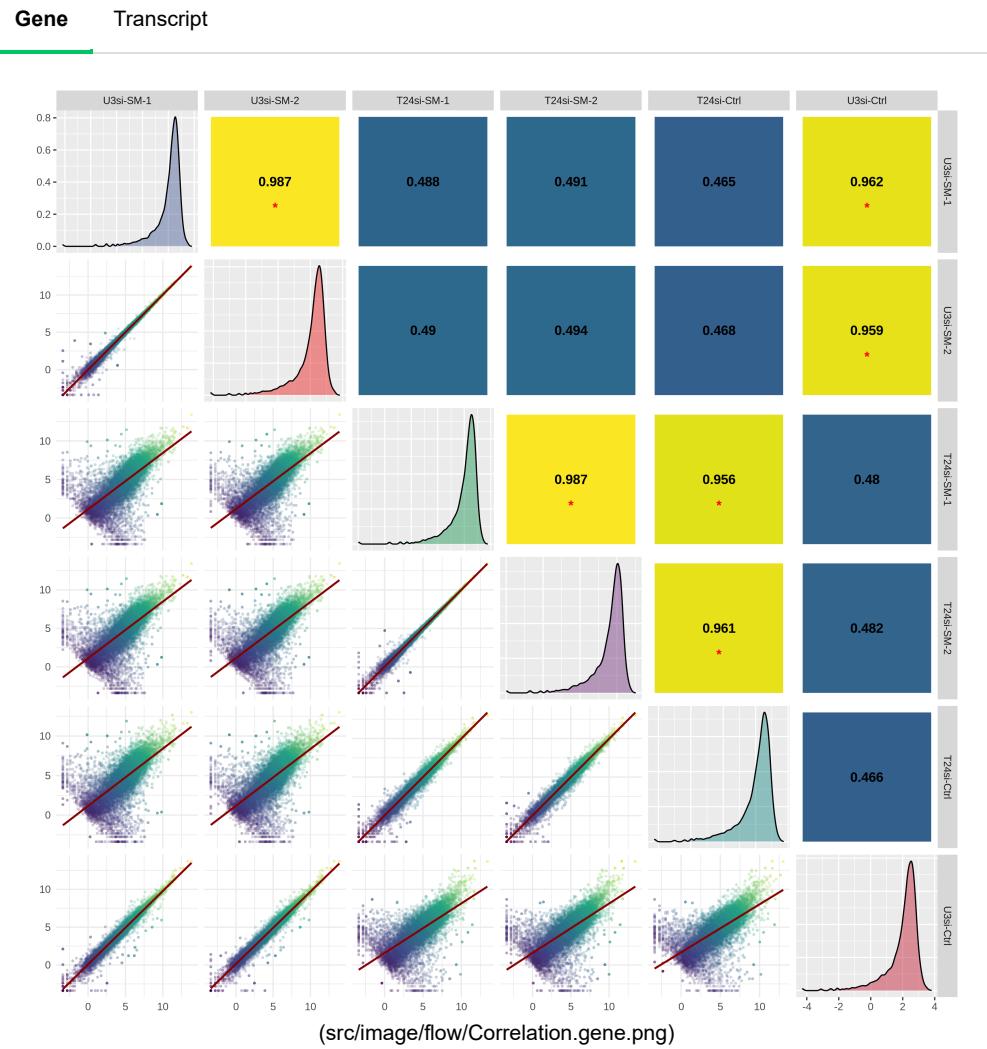


图 47 样品间相关性图(基因水平)

对角线是每个样品的log2(TMM)密度图，对角线上方是Pearson相关系数的平方(R^2)热图，其中*表示 R^2 大于0.95，以此类推，** 0.99，*** 0.999。

对角线下方为样品两两间log2(TMM)散点图，偏离对角线的点越多，说明样品表达量相关性越低。

12. 差异基因蛋白互作网络分析

分析采用STRING^[16] (<https://string-db.org/>) 数据库，STRING数据库有多种数据来源并打分，最终以combined score (score ≥ 400) 进行筛选。

使用igraph^[17]进行网络图绘制，若互作网络关系对超过500，则展示得分最大的前500个关系对，展示如下。



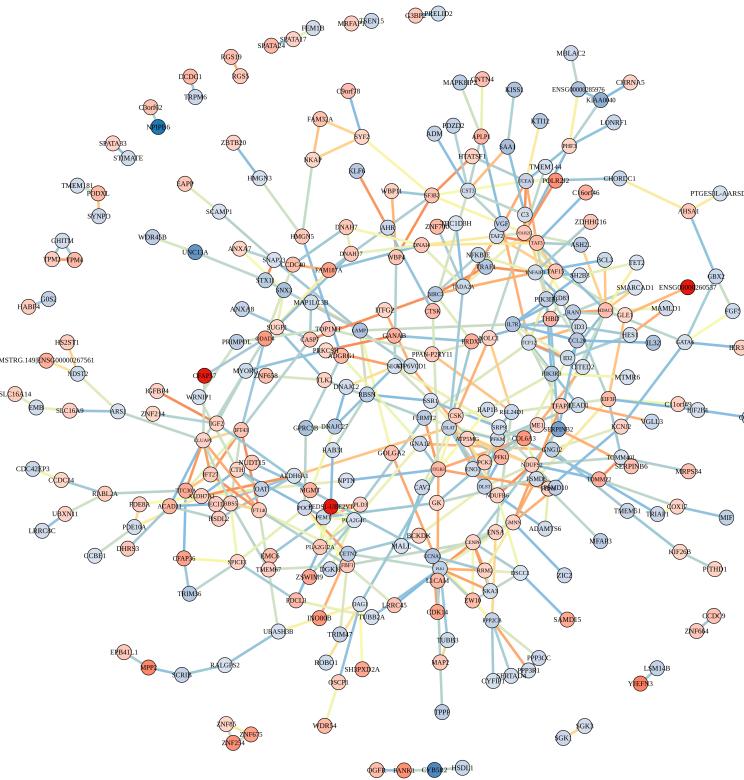
一、项目信息

二、概述

三、工作流程

四、结果展示及说明

五、参考文献



(src/image/flow/T24si-SM_vs_T24si-Ctrl_protein.link.network.png)

图 49 差异基因蛋白网络互作图

每个节点表示差异基因，名称用基因symbol表示，节点红色代表上调基因，蓝色代表下调基因，颜色的深浅与差异倍数正相关，连线表示蛋白互作，连线颜色跟互作得分（combined score）相关，得分低-高对应颜色蓝-黄-红。

IGEbio
广州艾基生物技术有限公司

广州艾基生物技术有限公司

13. 可变剪切分析

可变剪切 (Alternative Splicing, AS)，是大多数真核生物细胞中普遍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 前体后内含子会被 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>)^[20] 是一款适用于 RNA-seq 数据的可变剪切分析软件，它不仅可以对可变剪切事件进行分类，还可以进行不同样本间可变剪切事件的差异分析。rMATS 软件对可变剪切事件分类如下图所示：



一、项目信息

二、概述

三、工作流程

四、结果展示及说明

五、参考文献

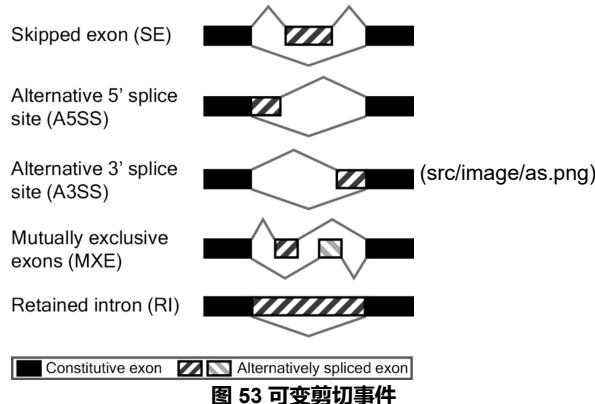


图 53 可变剪切事件

5类可变剪切事件定义如下：

1. **SE** : Skipped exon 外显子跳跃;
 2. **A5SS** : Alternative 5' splice site 第一个外显子可变剪切;
 3. **A3SS** : Alternative 3' splice site 最后一个外显子可变剪切;
 4. **MXE** : Mutually exclusive exon 互斥外显子;
 5. **RI** : Retained intron 内含子滞留;

14. 差异可变剪接事件统计

以每个进行差异可变剪切分析的比较组为单位，统计发生的可变剪切事件的种类及数量。

表 19 可变剪切统计结果

Group	SE	RI
T24si-SM_vs_T24si-Ctrl	1459(34803)	181(4688)
U3si-SM_vs_U3si-Ctrl	1333(39647)	148(4719)

统计说明：

1. **Group** : 比较组;
 2. ... : 发生的差异可变剪切事件以及对应的数量, 括号内为总数目, 括号外为差异数目;

14.1. 差异可变剪接事件分析

可变剪接差异分析包括可变剪接事件定量及表达差异显著性分析。每个可变剪接事件对应两个 Isoform，分别为 Exon Inclusion Isoform 和 Exon Skipping Isoform。

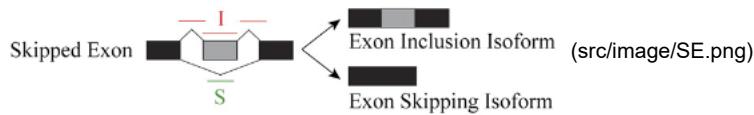


图 54 Exon Inclusion/Skipping Isoform 图解

不同的剪切事件对应的上下游外显子可能不太一样，对于SE事件来说，exonStart_0base、exonEnd、upstreamES、upstreamEE、downstreamES、downstreamEE分别对应下图位置。



广州艾基生物技术有限公司
GUANGZHOU IGE BIOTECHNOLOGY LTD

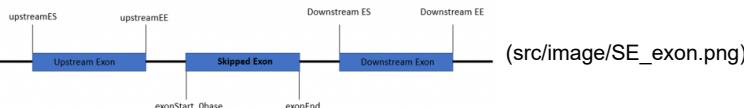


图 55 SE Exon图解

一、项目信息

二、概述

三、工作流程

四、结果展示及说明

五、参考文献

分别对两种 Isoform 进行表达量统计，并除以其有效长度，得到校正后表达量，然后计算Exon Inclusion Isoform在两个Isoform总表达量的比值，最后进行差异显著性分析，结果如下表所示。

表 20 差异可变剪切SE结果 (前50)

	ID	GeneID	geneSymbol	chr	strand	exonStart_Obase	exon
4	ENSG00000155313	USP25	chr21	+		15762890	15
5	ENSG00000155313	USP25	chr21	+		15777903	15
6	ENSG00000155313	USP25	chr21	+		15799756	15
7	ENSG00000155313	USP25	chr21	+		15808808	15
8	ENSG00000155313	USP25	chr21	+		15811136	15
9	ENSG00000155313	USP25	chr21	+		15818697	15
11	ENSG00000155313	USP25	chr21	+		15833347	15
12	ENSG00000155313	USP25	chr21	+		15833347	15
13	ENSG00000155313	USP25	chr21	+		15842397	15

Show 10 entries

Showing 1 to 10 of 49 entries

Previous 1 2 3 4 5 Next

统计说明：

1. **ID**：剪切事件唯一的编号，从0开始；
2. **GeneID**：可变剪接事件所在基因编号；
3. **geneSymbol**：可变剪接事件所在基因名称，若无对应symbol，则显示NA；
4. **chr**：可变剪接事件所在染色体；
5. **strand**：可变剪接事件所在链的方向；
6. **exonStart_Obase**：发生该可变剪接事件的外显子起始位置；
7. **exonEnd**：发生该可变剪接事件的外显子终止位置；
8. **upstreamES**：发生该可变剪接事件上游exon起始位置；
9. **upstreamEE**：发生该可变剪接事件上游exon终止位置；
10. **downstreamES**：发生该可变剪接事件下游exon起始位置；
11. **downstreamEE**：发生该可变剪接事件下游exon终止位置；
12. **IC_SAMPLE_1**：可变剪接事件Exon Inclusion Isoform在组1（处理组）中的表达量，多个样品以逗号分隔；
13. **SC_SAMPLE_1**：可变剪接事件Exon Skipping Isoform在组1（处理组）中的表达量，多个样品以逗号分隔；
14. **IC_SAMPLE_2**：可变剪接事件Exon Inclusion Isoform在组2（对照组）中的表达量，多个样品以逗号分隔；
15. **SC_SAMPLE_2**：可变剪接事件Exon Skipping Isoform在组2（对照组）中的表达量，多个样品以逗号分隔；
16. **IncFormLen**：可变剪接事件Exon Inclusion Isoform的有效长度；
17. **SkipFormLen**：可变剪接事件Exon Skipping Isoform的有效长度；
18. **PValue**：可变剪接事件表达差异显著性p值；
19. **FDR**：可变剪接事件表达差异显著性FDR值；



广州艾基生物技术有限公司
GUANGZHOU IGE BIOTECHNOLOGY LTD

20. **IncLevel1**：组1可变剪接事件Exon Inclusion Isoform在两个Isoform总表达量的比值；
21. **IncLevel2**：组2可变剪接事件Exon Inclusion Isoform在两个Isoform总表达量的比值；

一、项目信息

二、概述

三、工作流程

四、结果展示及说明

五、参考文献

对上述剪切事件进行筛选, **FDR < 0.05** 以及 **IncLevelDifference > 0.2** 作为显著性差异可变剪切事件。

表 21 每组差异可变剪切统计结果

Copy

CSV

Excel

PDF

Column visibility

Search:

EventType	EventTypeDescription	TotalEvents	totalSignificantEvents	Sig
SE	skipped exon	34803	1459	
RI	retained intron	4688	181	
A3SS	alternative 5' splice sites	5516	203	
A5SS	alternative 3' splice sites	3881	160	
MXE	mutually exclusive exons	4390	131	

Show 10 entries

Showing 1 to 5 of 5 entries

Previous

1

Next

统计说明：

1. **EventType** : AS事件;
 2. **EventTypeDescription** : AS事件类型描述;
 3. **TotalEvents** : 所有的AS事件数量;
 4. **totalSignificantEvents** : 显著差异AS事件数量;
 5. **SigEventsSample1HigherInclusion** : 表示在处理组中具有更高的剪切包含水平的可变剪切事件，例如当事件为SE时，处理组中包含外显子的事件比例更高，而对照组跳过外显子的事件比例更高
 6. **SigEventsSample2HigherInclusion** : 表示在对照组中具有更高的剪切包含水平的可变剪切事件;

14.2. 差异可变剪接事件可视化

使用rmats2sashimiplot对差异显著性的可变剪接事件进行可视化展示，结果文件中每个类型取显著的前20个事件可视化展示，若不足20个则展示全部。

注：结果中的剪切事件的可视化图中数字可能与结果表中每个样品对应的数值有小部分差别，这是因为rmats2sashimiplot会根据rmats结果重新比对计算每个事件支持reads数目。后续任取一方的结果都可以。

7:100119204:100119339:-@7:100118518:100118649:-@7:100114054:100114268:-



广州艾基生物技术
GUANGZHOU IGE BIOTECH

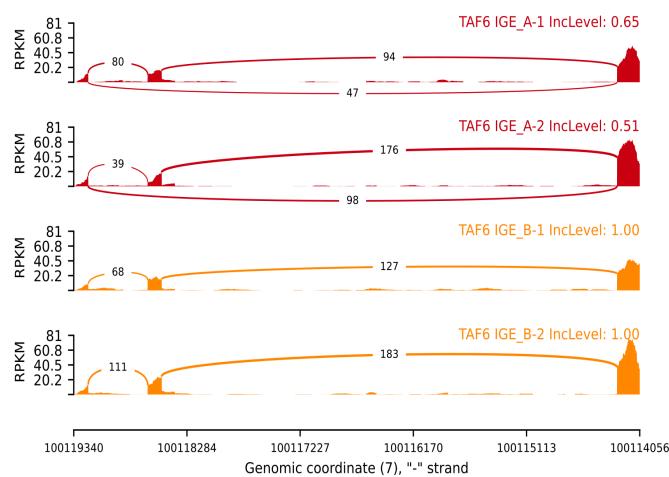
一、项目信息

二、概述

三、工作流程

四、结果展示及说明

五、参考文献



(src/image/es_example.png)

图 56 SE 可变剪切事件示例

上述示意图表示的是一个外显子跳跃的可变剪切事件

图中标题为可变剪接事件三个外显子所在的染色体坐标及正负链信息

跨外显子比对的 reads 使用连接外显子 junction 边界的弧线表示,同时弧线上的数字指出了 junction reads 的数目

右上方标注了各个样本可变剪接事件所在的基因及 IncLevel 值

最下方是可变剪切isofrom的示意图, 分别对应inclusion isofrom 和 skipping isofrom

采用RPKM值量化表示样本中对应的测序深度分布, 不同分组样本用不同颜色表示。

IGEbio
广州艾基生物技术有限公司

广州艾基生物技术有限公司

五、参考文献

- [1] S. A. Hardwick et al., “Spliced synthetic genes as internal controls in RNA sequencing experiments,” Nat. Methods, vol. 13, no. 9, pp. 792–798, 2016.
- [2] Shifu Chen, Yanqing Zhou, Yaru Chen, Jia Gu; fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics, Volume 34, Issue 17, 1 September 2018, Pages i884–i890, <https://doi.org/10.1093/bioinformatics/bty560>
- [3] Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.
- [4] D. Kim, B. Langmead, and S. L. Salzberg1, “HISAT: a fast spliced aligner with low memory requirements Daehwan HHS Public Access,” Nat. Methods, vol. 12, no. 4, pp. 357–360, 2015.
- [5] Wang, L., Wang, S., & Li, W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics (Oxford, England), 28(16), 2184–2185. <http://doi.org/10.1093/bioinformatics/bts356>
- [6] 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.



广州艾基生物技术有限公司
GUANGZHOU IGE BIOTECHNOLOGY LTD
10.12688/f1000research.23297.1

一、项目信息

二、概述

三、工作流程

四、结果展示及说明

五、参考文献

- [7] Pertea G and Pertea M. GFF Utilities: GffRead and GffCompare. F1000Research 2020, 9:304 DOI: 10.12688/f1000research.23297.1
- [8] J. Harrow et al., “GENCODE: The reference human genome annotation for the ENCODE project,” Genome Res., vol. 22, no. 9, pp. 1760–1774, 2012.
- [9] Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods.
- [10] Robinson, M.D., Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11, R25 (2010). <https://doi.org/10.1186/gb-2010-11-3-r25>
- [11] Love MI, Huber W, Anders S (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” Genome Biology, 15, 550. doi: 10.1186/s13059-014-0550-8.
- [12] Mark D. Robinson, D.J.M., Gordon K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 2010. 26(1): p. 139-140.
- [13] Consortium, T.G.O., Gene Ontology Consortium: going forward. Nucleic Acids Research, 2015. 43: p. D1049.
- [14] Yu G, Wang L G, Han Y, et al. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters[J]. OMICS-A JOURNAL OF INTEGRATIVE BIOLOGY, 2012, 16(5):284-287.
- [15] Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research, 28(1), 27-30.
- [16] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Mering CV. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019 Jan 8;47(D1):D607-D613. doi: 10.1093/nar/gky1131. PMID: 30476243; PMCID: PMC6323986.
- [17] Csardi G, Nepusz T (2006). “The igraph software package for complex network research.” InterJournal, Complex Systems, 1695. <https://igraph.org>.