

CATEGORIES

分析模块导图

任务参数

数据质控

数据比对

高级质控

可变剪接

表达定量

差异分析

GO富集分析

KEGG富集分析

PPI互作网络分析

参考文献

真核有参转录组测序

④ 分析模块导图



⑤ 任务参数

```
开始于Sun Dec 13 10:31:48 2020, 结束于Sun Dec 13 11:17:11 2020
任务名称: ref_report_check_02
所属项目: Update2020
任务目录: /project/home/update2019126com/final_check_202008/ref_test_03
链特异性文库: RF
测序类型: PE
物种: /database/Reference/oryza_sativa
参考文件:
    ref: /database/Reference/oryza_sativa/Oryza_sativa.IRGSP-1.0.dna_sm.toplevel.fa
    fai: /database/Reference/oryza_sativa/Oryza_sativa.IRGSP-1.0.dna_sm.toplevel.fa.fai
    gtf: /database/Reference/oryza_sativa/Oryza_sativa.IRGSP-1.0.38_exon.gtf
    genename: /database/Reference/oryza_sativa/geneName.txt
    go: /database/Reference/oryza_sativa/go.txt
    kegg: osa
    ppi: 39947
数据目录:
    wll3: /project/home/update2019126com/Public/rna_demo_data/wll3_1.fq.gz, /project/home/update2019126com/Public/rna_demo_data/wll3_2.fq.gz
    ck1: /project/home/update2019126com/Public/rna_demo_data/ck1_1.fq.gz, /project/home/update2019126com/Public/rna_demo_data/ck1_2.fq.gz
    ck3: /project/home/update2019126com/Public/rna_demo_data/ck3_1.fq.gz, /project/home/update2019126com/Public/rna_demo_data/ck3_2.fq.gz
    wll2: /project/home/update2019126com/Public/rna_demo_data/wll2_1.fq.gz, /project/home/update2019126com/Public/rna_demo_data/wll2_2.fq.gz
样本: wll3, ck1, ck3, wll2
分组:
    ck: ck1, ck3
    wll: wll2, wll3
比较:
    ckVSll: ck, wll
其他参数:
    phred: phred33
    diffTool: DESeq2
    foldchange: 2
    pType: padj
    pValue: 0.05
```

⑥ 基本数据QC

1. 测序质量分析

Illumina高通量测序平台 (HiSeq/MiSeq/NovaSeq) 的碱基质量值用Phred quality score表示。测序仪在碱基识别时，会给出每个碱基错误识别的概率P，碱基的Phred quality score则为 $-10\log_{10}P$ ，如某个碱基的错误识别的概率为0.001，则其质量值为30，对应关系如下：

Phred质量值	碱基错误识别概率	碱基正确识别概率	通常简称
10	1/10	90%	Q10
20	1/100	99%	Q20
30	1/1000	99.9%	Q30
40	1/10000	99.99%	Q40

由于测序平台和测序原理本身的限制，测序质量值会随着Reads从5'到3'端逐渐降低。主要是由于酶和试剂的消耗导致的，另外，phasing (即一轮循环中没有加入碱基) 和prechasing (即一轮循环中加入2个以上碱基) 也是一个重要因素。通常5'端最开始几个碱基质量一般也比较低。

CATEGORIES

分析模块导图

任务参数

数据质控

数据比对

高级质控

可变剪接

表达定量

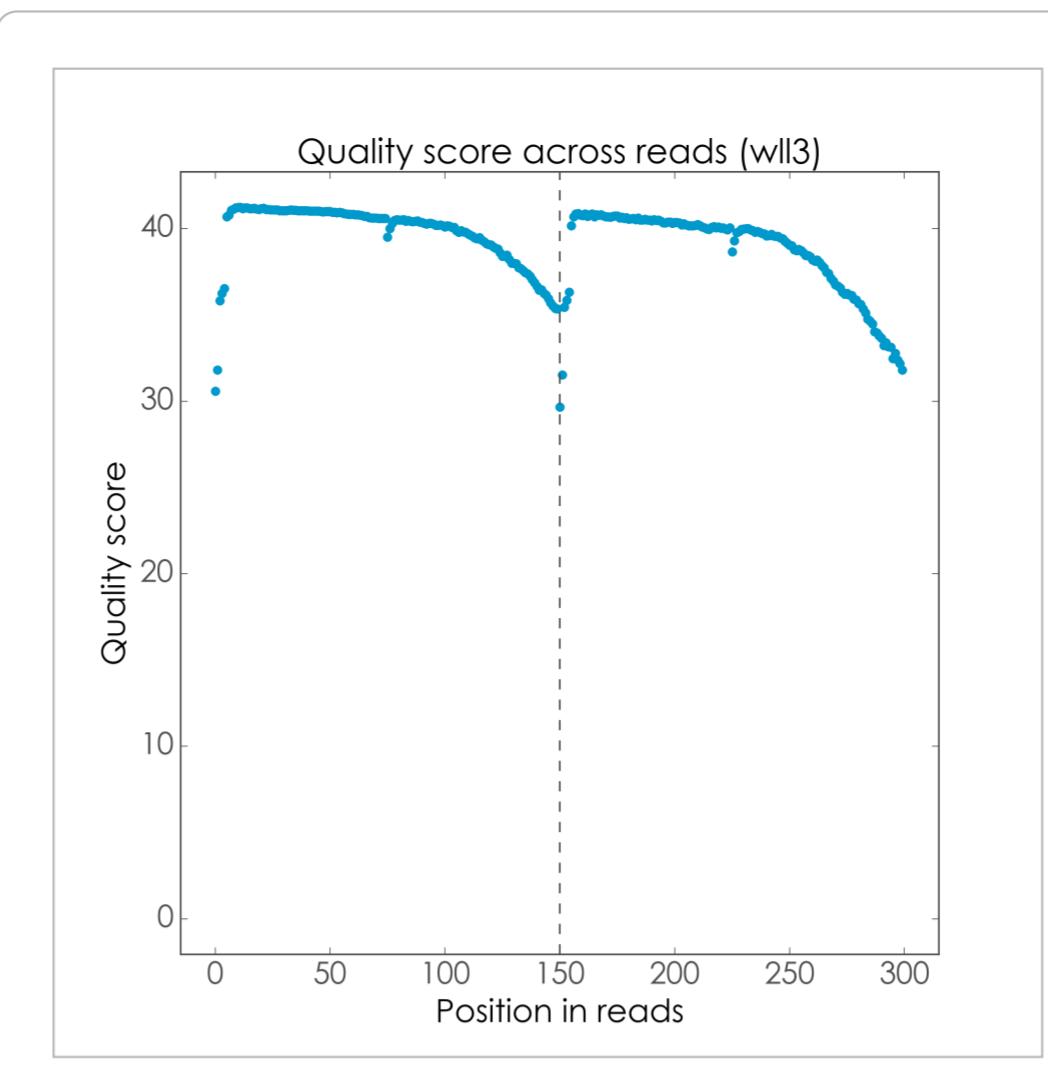
差异分析

GO富集分析

KEGG富集分析

PPI互作网络分析

参考文献

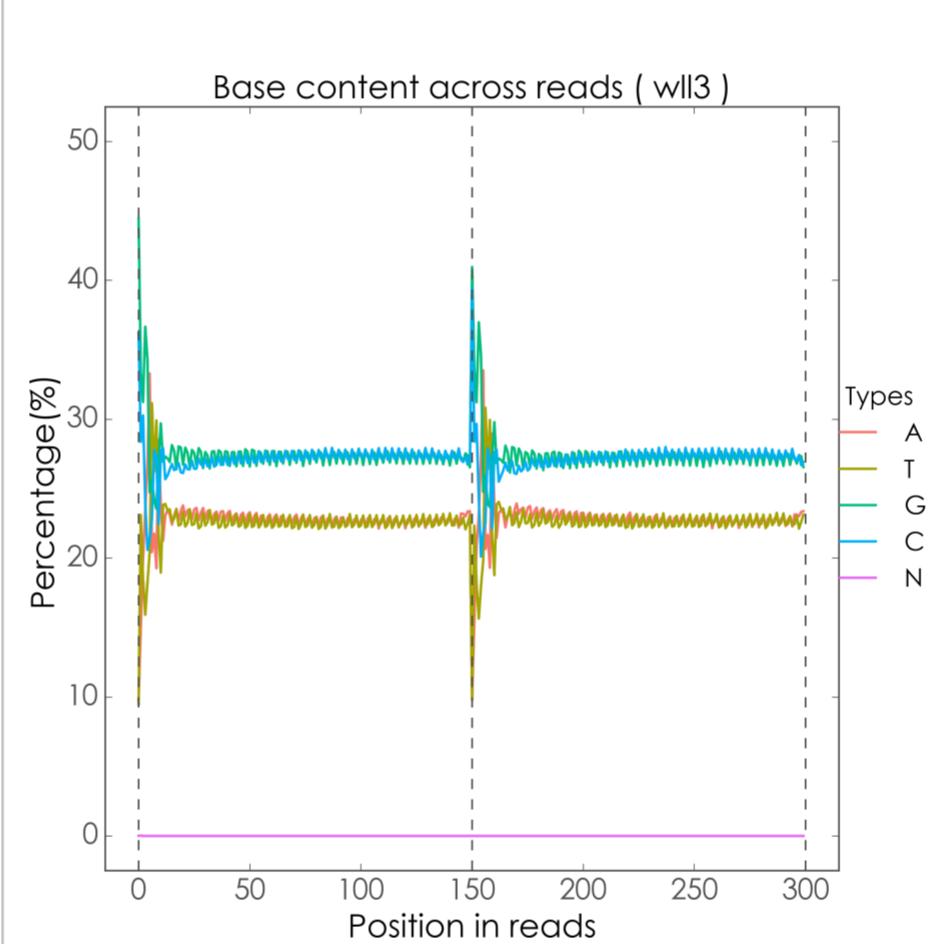


1/4 测序数据质量值分布

2.碱基含量分布

理想情况下，根据碱基互补配对原则，测序数据的GC (AT) 碱基含量理论上应该相等。实际测序过程中，由于测序样品本身特性，文库构建过程和测序过程可能引入带GC偏好，出现GC分离现象。为此我们在进行信息分析之前需要检查测序数据是否有GC分离现象。

在转录组建库过程中对mRNA进行反转录时使用的6 bp的随机引物，会引起扩增前几个位置的核苷酸组成存在一定的偏好性，由此引发read前几个碱基GC含量分布不均，这种不均属于正常情况。



1/8 碱基含量GC分布图，柱状图为前15个碱基GC分布图

3.数据过滤

下机数据 (Raw Data即Raw Reads) 通常会含有少量的接头污染及低质量的Reads，如果不对其进行过滤处理会对后续分析造成影响，为此我们对这部分Reads进行了过滤，数据过滤标准如下：

- 1) 使用cutadapt (<http://cutadapt.readthedocs.io/en/stable/>) (Martin M. , 2011; 版本: v1.10; 参数: -e 0.1 -O 6) 过滤带有测序接头 (adapter) 的Reads；
- 2) 过滤N (不确定碱基) 含量比例大于15%的Reads；
- 3) 过滤低质量碱基 (Q<20) 含量大于10%的Reads；
- 4) 注意：本流程中在质控报告部分展示的结果并没有过滤rRNA，之后的模块分析为过滤了rRNA的。

CATEGORIES

分析模块导图

任务参数

数据质控

数据比对

高级质控

可变剪接

表达定量

差异分析

GO富集分析

KEGG富集分析

PPI互作网络分析

参考文献

4. 测序数据统计

所有样本的测序数据量及质量统计表如下：

	Sample_name	Clean_reads	Clean_bases(Gb)	N_reads	low_quality_reads	adaptor_reads	Error_rate(%)
数据比对	wll3	3616818	0.54	0	572462	610918	0.02
高级质控	ck1	3172956	0.48	0	420318	406726	0.02
可变剪接	ck3	3769830	0.57	0	517896	512472	0.02
表达定量	wll2	3618974	0.54	0	571736	610156	0.02



- (1) Sample name: 样品名
 (2) Clean Reads: Raw Reads过滤含接头及低质量测序数据后的reads数
 (3) Clean Bases: Clean Reads的条数乘以测序数据读长, 即过滤后的总碱基量
 (4) N_reads: N含量过高的reads数
 (5) low_quality_reads: 低质量的reads数
 (6) adaptor_reads: 带接头的的reads数
 (7) Error Rate: 碱基的平均错误率
 (8) Q20、Q30: Phred 质量值大于20、30的碱基占总碱基的比例
 (9) GC Content: GC含量, 测序数据中GC碱基数占总碱基数的比例

④ 数据比对**1. 测序质量分析**使用Hisat2 (<http://ccb.jhu.edu/software/hisat2/index.shtml>) , 将reads比对到参考序列上。

所有样本与参考序列的比对统计如下：

Sample name	wll3	ck1	ck3	wll2
Clean reads	3,616,818	3,172,956	3,769,830	3,618,974
Optical/PCR duplicate	279,742 (7.73%)	380,402 (11.99%)	407,993 (10.82%)	280,171 (7.74%)
Total mapped	3,498,409 (96.73%)	3,080,398 (97.08%)	3,657,640 (97.02%)	3,500,707 (96.73%)
Unmapped	118,409 (3.27%)	92,558 (2.92%)	112,190 (2.98%)	118,267 (3.27%)
Multiple mapped	82,380 (2.28%)	62,430 (1.97%)	77,195 (2.05%)	82,436 (2.28%)
Uniquely mapped	3,416,029 (94.45%)	3,017,968 (95.12%)	3,580,445 (94.98%)	3,418,271 (94.45%)
Read-1	1,707,410 (47.21%)	1,508,410 (47.54%)	1,789,500 (47.47%)	1,708,513 (47.21%)
Read-2	1,708,619 (47.24%)	1,509,558 (47.58%)	1,790,945 (47.51%)	1,709,758 (47.24%)
Reads map to '+'	1,706,186 (47.17%)	1,508,036 (47.53%)	1,789,038 (47.46%)	1,707,281 (47.18%)
Reads map to '-'	1,709,843 (47.27%)	1,509,932 (47.59%)	1,791,407 (47.52%)	1,710,990 (47.28%)
Non-splice reads	2,151,214 (59.48%)	1,927,030 (60.73%)	2,278,680 (60.45%)	2,152,804 (59.49%)
Splice reads	1,264,815 (34.97%)	1,090,938 (34.38%)	1,301,765 (34.53%)	1,265,467 (34.97%)
Reads mapped in proper pairs	3,302,186 (91.30%)	2,930,496 (92.36%)	3,473,922 (92.15%)	3,304,116 (91.30%)
Proper-paired reads map to different chrom	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
lincRNA	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)
circRNA	17(0.00%)	14(0.00%)	19(0.00%)	18(0.00%)
rRNA	2382(0.07%)	2928(0.09%)	3314(0.09%)	2386(0.07%)

- (1) Sample name: 样本名
 (2) Clean reads: 经过质控后的总Reads数 (Clean reads)
 (3) Optical/PCR duplicate: Optical/PCR duplicate Reads数
 (4) Total mapped: 总的比对到参考基因组的reads数
 (5) Unmapped reads: 没有比对到参考基因组的reads数
 (6) Multiple mapped : 能够比对到参考序列多个位置的Reads数 (这部分数据的比例一般会小于10%)
 (7) Uniquely mapped : 比对到参考序列唯一位置的Reads数
 (8) Read-1, Read-2: Read1和Read2分别比对到参考序列上Reads数
 (9) Reads map to '+', Reads map to '-': 分别比对到序列上正链和负链的Reads数
 (10) Splice reads: 同一条Read分段比对到不同外显子的总数 (也称为Junction reads), Non-splice reads为同一Read只比对到一个外显子的总数, Splice reads的比例在很大程度上取决于测序的读长
 (11) Reads mapped in proper pairs: reads对满足插入片段大小, 同时read1和read2分别比对到不同链的数目
 (12) Proper-paired reads map to different chrom: (10)中read1和read2比对到不同染色体的reads对数目
 (13) lincRNA: 参考基因组中注释为lincRNA的reads数
 (14) circRNA: 参考基因组中注释为circRNA的reads数
 (15) rRNA: 参考基因组中注释为rRNA的reads数

⑤ 高级数据 QC

CATEGORIES

分析模块导图

任务参数

数据质控

数据比对

高级质控

可变剪接

表达定量

差异分析

GO富集分析

KEGG富集分析

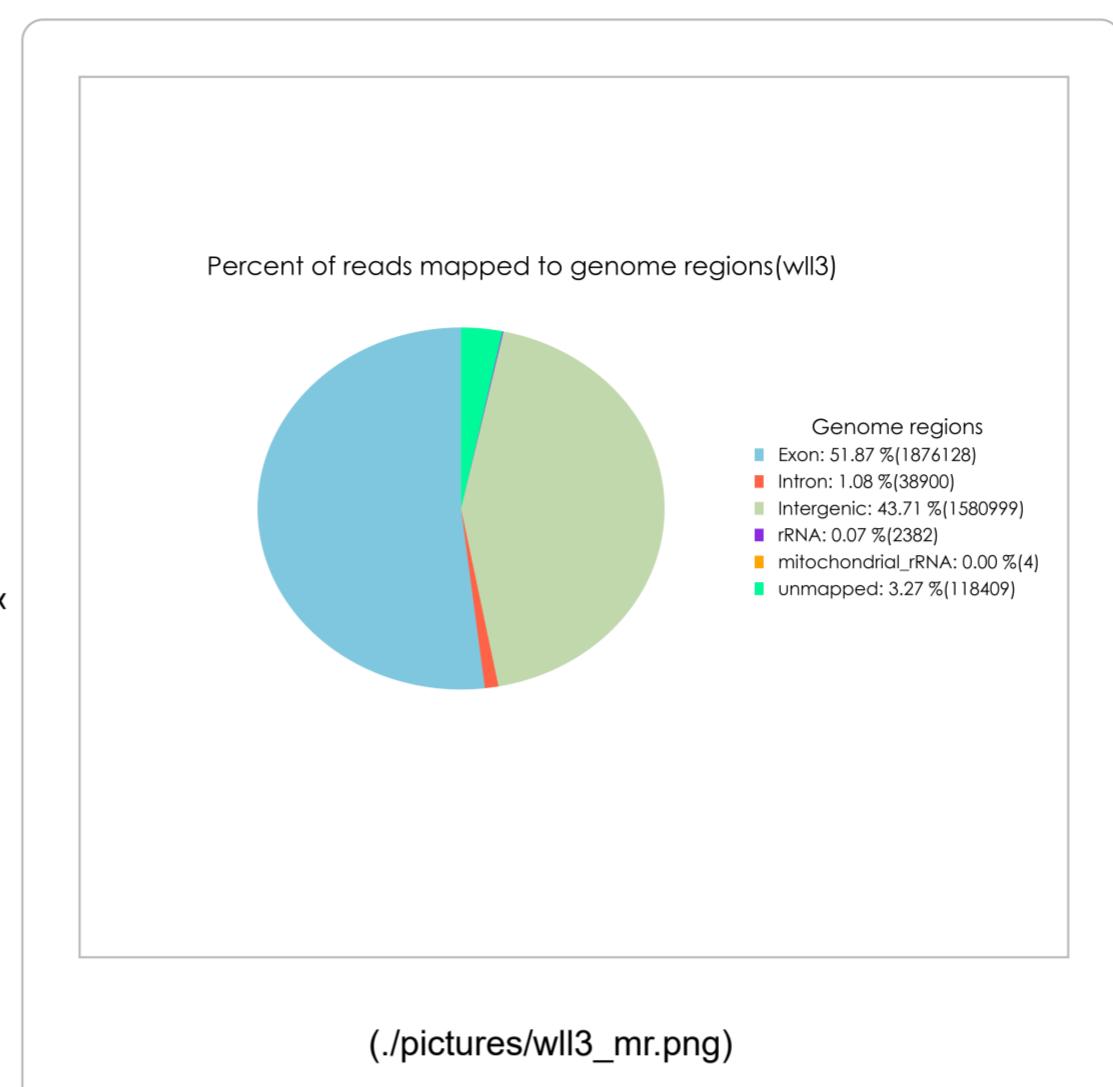
PPI互作网络分析

参考文献

1.比对区域分布

根据Reads比对到参考序列的Exon(外显子)、Intron(内含子)和Intergenic (基因间区)三个区域以及Reads是否属于rRNA (包括线粒体rRNA)，是否为没有比对到参考基因组的情况对其进行统计分析。

在参考序列注释较为完全的物种中，比对到Exon(外显子)的Reads含量最高，比对到Intron(内含子)区域的Reads是由于pre-mRNA的残留及可变剪切过程中发生的内含子滞留事件导致的，而比对到Intergenic(基因间区)的Reads是因为参考序列注释不完全。

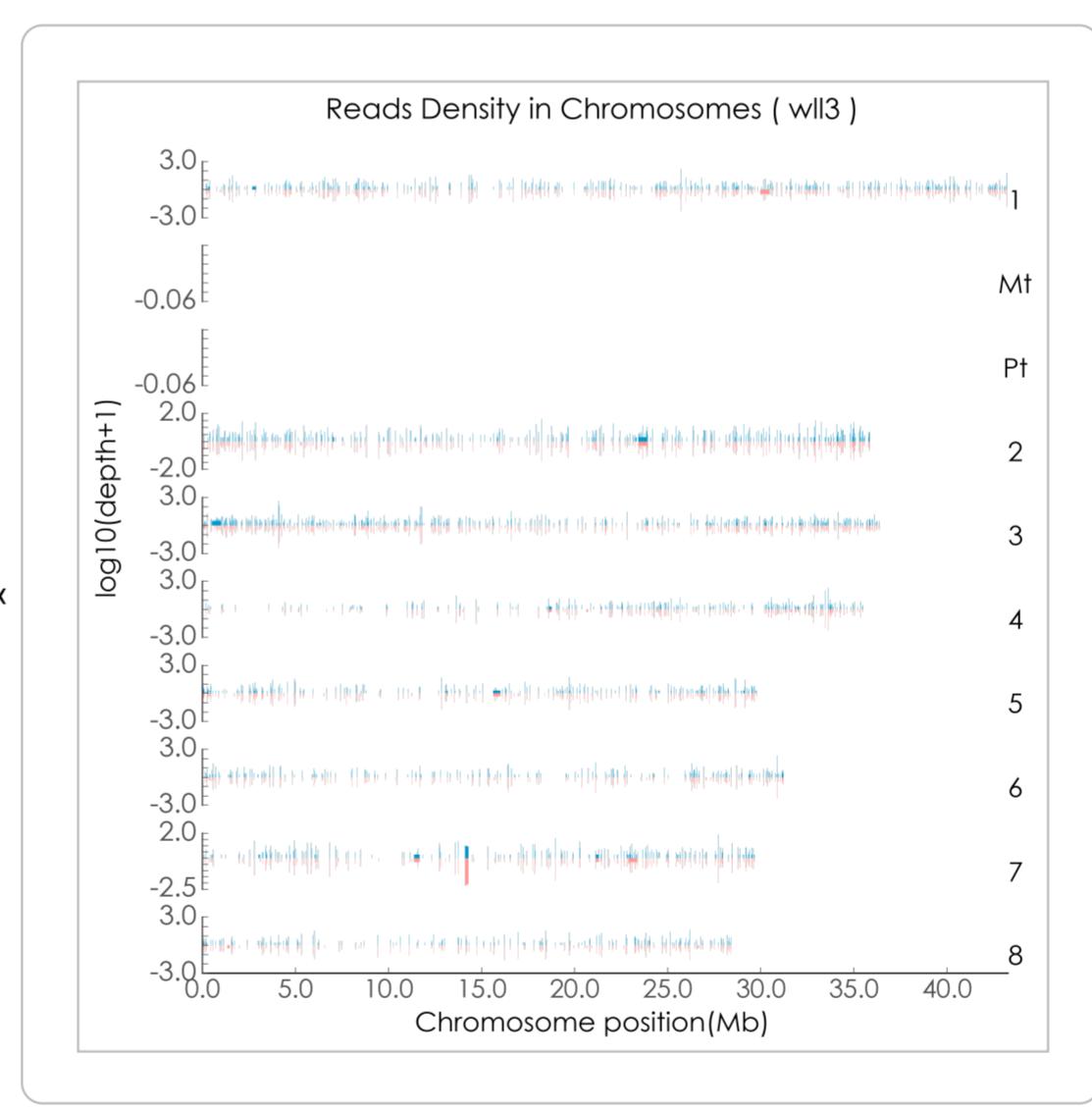


1/12 Reads在参考序列不同区域的分布情况/文库插入片段大小分布

图/bodycoverage分布图

2.Reads分布密度

对比对到参考序列上的Reads在各条染色体(分正负链)上的分布密度情况进行统计。如下图所示，设置滑动窗口为10Kb，并计算窗口内比对到每个碱基Reads的中位数的对数值(\log_2)。一般情况下，比对到该染色体内的Reads总数与染色体长度成正比(Marquez et al., 2012)。从比对到染色体上的Reads数与染色体长度的关系图中，可以更加直观看出染色体长度和Reads总数的关系。



1/8 Reads在染色体上的密度分布图

3.表达水平的饱和曲线检查

当前数据量是否能够准确定量基因表达水平可以通过定量饱和曲线来评估。一般来说，基因表达量越高，准确定量所需要的数据量就越小；基因表达量越小，则准确对其定量所需要的数据量就越大。抽取5%、10%、...、90%、95%的测序数据分别进行基因表达定量分析，将这些数据量下某个基因的定量结果与100%数据量的结果比较，如果差异小于15%，则认为这个基因当前定量准确。

定量饱和曲线检查及bodycoverage分布图

CATEGORIES

分析模块导图

任务参数

数据质控

数据比对

高级质控

可变剪接

表达定量

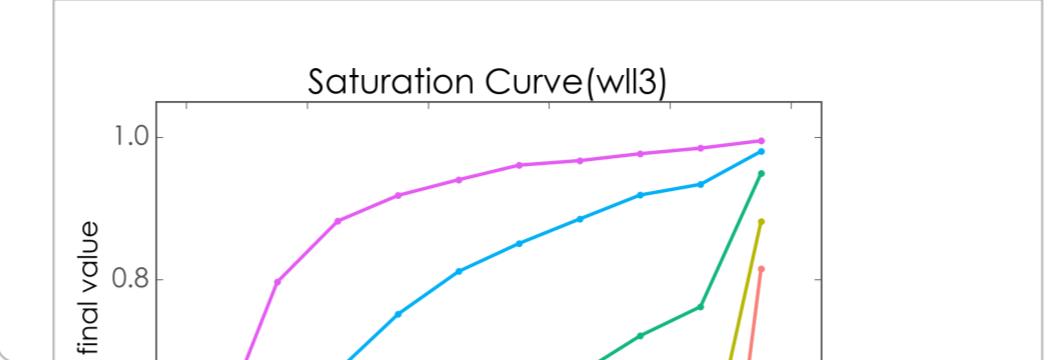
差异分析

GO富集分析

KEGG富集分析

PPI互作网络分析

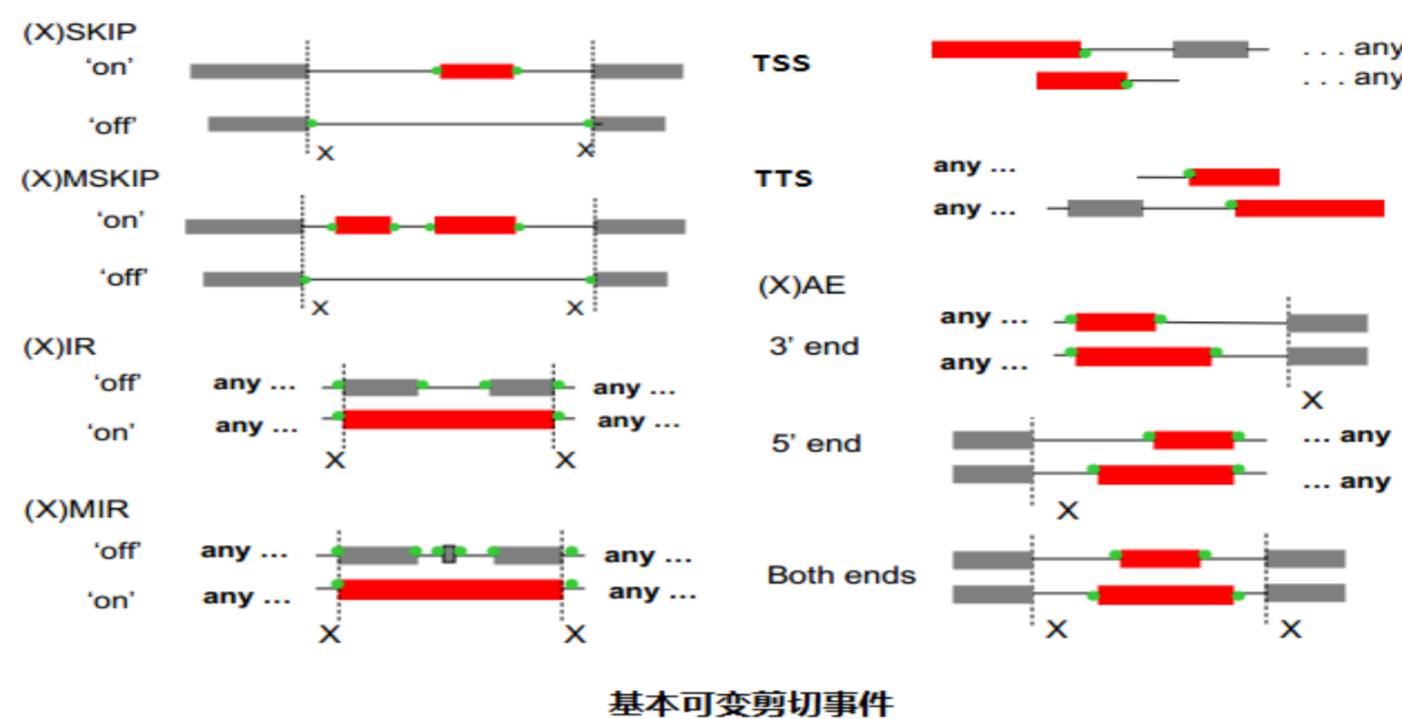
参考文献



1/4 定量饱和曲线检查

可变剪接分析

用ASprofile (<http://ccb.jhu.edu/software/ASprofile/>)软件对每个样品的可变剪接事件进行分类统计。ASprofile可变剪接分析流程及对可变剪接事件的分类如下图所示：

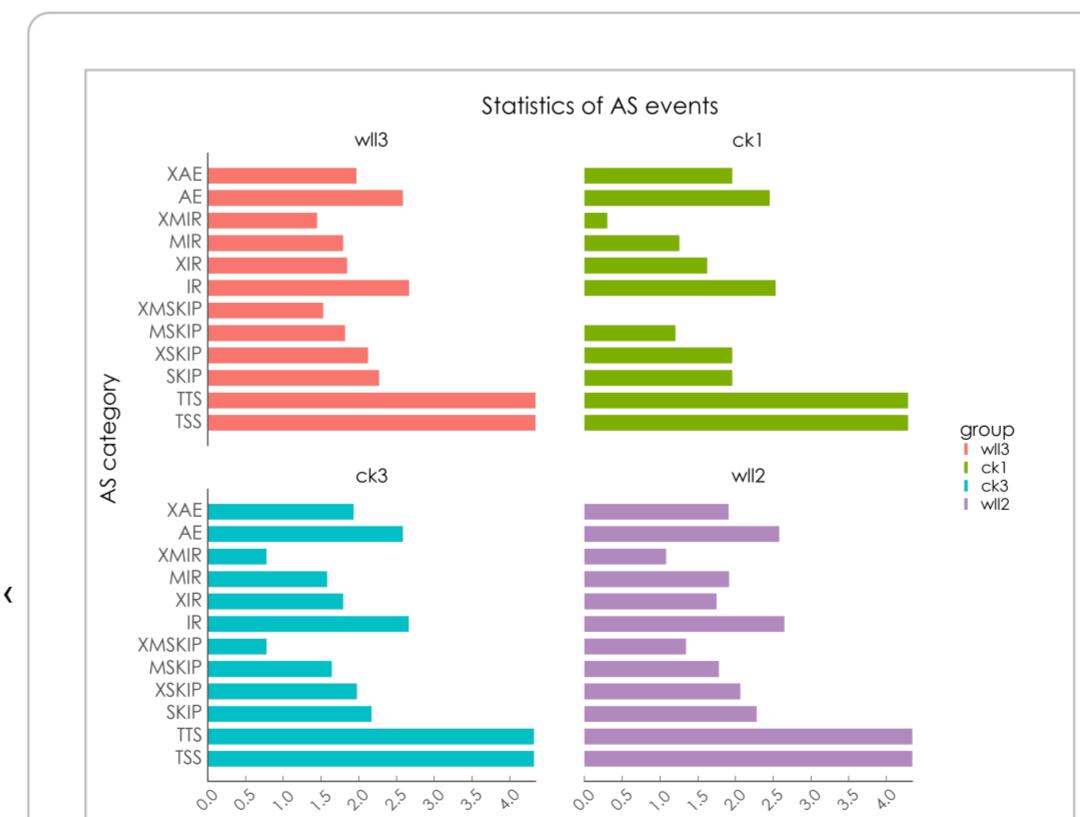


基本可变剪切事件

ASprofile对可变剪接事件分类的定义如下：

- 1) TSS: Alternative 5' first exon (transcription start site) 第一个外显子可变剪接
- 2) TTS: Alternative 3' last exon (transcription terminal site) 最后一个外显子可变剪接
- 3) SKIP: Skipped exon (SKIP_ON,SKIP_OFF pair) 单外显子跳跃
- 4) XSKIP: Approximate SKIP (XSKIP_ON,XSKIP_OFF pair) 单外显子跳跃（模糊边界）
- 5) MSKIP: Multi-exon SKIP (MSKIP_ON,MSKIP_OFF pair) 多外显子跳跃
- 6) XMSKIP: Approximate MSKIP (XMSKIP_ON,XMSKIP_OFF pair) 多外显子跳跃（模糊边界）
- 7) IR: Intron retention (IR_ON, IR_OFF pair) 单内含子滞留
- 8) XIR: Approximate IR (XIR_ON, XIR_OFF pair) 单内含子滞留（模糊边界）
- 9) MIR: Multi-IR (MIR_ON, MIR_OFF pair) 多内含子滞留
- 10) XMIR: Approximate MIR (XMIR_ON, XMIR_OFF pair) 多内含子滞留（模糊边界）
- 11) AE: Alternative exon ends (5', 3', or both) 可变 5'或3'端剪接

1. 可变剪接类别和数量



(./pictures/as_event.png)

1/1 可变剪接分类和数量统计

2. 可变剪接结构和表达量

可变剪接结构和表达量统计部分结果如下：

CATEGORIES	Event_id	Event_type	Gene_id	Chrom	Event_start	Event_end	Event_pattern	Strand	Fpkm
	1000001	TSS	STRG.1	1	3237	3255	3255	+	-1.0000000000
	1000002	TTS	STRG.1	1	4357	4456	4357	+	-1.0000000000
	1000003	TSS	STRG.10	1	27143	27292	27292	+	-1.0000000000
分析模块导图	1000004	TTS	STRG.10	1	28365	28644	28365	+	-1.0000000000
任务参数	1000005	TSS	STRG.100	1	549157	549399	549399	+	-1.0000000000

数据质控

- ① Event id: 可变剪接事件编号
- ② Event type: 可变剪接事件类型 (TSS, TTS, SKIP_{ON,OFF}, XSKIP_{ON,OFF}, MSKIP_{ON,OFF}, XMSKIP_{ON,OFF}, IR_{ON ,OFF}, XIR_{ON,OFF}, AE, XAE)

高级质控

- ③ Gene ID: cufflink组装结果中的基因编号
- ④ Chrom: 染色体编号

可变剪接

- ⑤ Event start: 可变剪接事件起始位置
- ⑥ Event end: 可变剪接事件结束位置

表达定量

- ⑦ Event pattern: 可变剪接事件特征
- ⑧ Strand: 基因正负链信息

差异分析

- ⑨ FPKM: 此可变剪接类型该基因表达量
- ⑩ Ref id: 此基因在参考注释文件中的编号

GO富集分析

基因表达分析

KEGG富集分析

1. 表达水平统计分析

衡量一个基因表达水平的最直接指标为其转录本的丰度：转录本丰度越高，则基因表达水平越高。由于实验中对转录本的扩增是随机的，因此转录本的丰度体现为其测序数据量，即比对到该转录本对应基因的Reads数量。为了使不同长度基因、不同实验、不同测序数据量的样本间具有可比性，人们引入了FPKM的概念。FPKM (Fragments Per Kilo bases per Million fragments mapped)是每百万比对到参考序列的Fragments中来自某一基因每千碱基长度的Fragments数目。FPKM同时考虑了测序深度和基因长度对Fragments计数的影响，是目前最为常用的基因表达水平衡量方法 (Mortazavi *et al.*, 2008)

采用HTSeq软件 (<http://www-huber.embl.de/users/anders/HTSeq/doc/count.html>)中htseq-count (版本: 0.6.0; 参数: 默认) 对各样品进行基因表达水平分析，模型为union。结果文件分别统计了不同表达水平下基因的数量以及单个基因的表达水平。一般情况下，使用FPKM值为0.1或者1作为判断基因是否表达的阈值。

不同表达水平区间的基因数量统计表如下：

FPKM Interval	wll3	ck1	ck3	wll2
0.0	65,245(72.36%)	66,614(73.88%)	65,571(72.73%)	65,297(72.42%)
>0.0	24,917(27.64%)	23,548(26.12%)	24,591(27.27%)	24,865(27.58%)
0.0~1.0	1,857(2.06%)	1,818(2.02%)	2,122(2.35%)	1,836(2.04%)
1.0~3.0	4,058(4.50%)	4,225(4.69%)	4,320(4.79%)	4,035(4.48%)
3.0~15.0	9,254(10.26%)	9,426(10.45%)	9,643(10.70%)	9,270(10.28%)
15.0~60.0	6,954(7.71%)	5,754(6.38%)	6,102(6.77%)	6,926(7.68%)
>60.0	2,794(3.10%)	2,325(2.58%)	2,404(2.67%)	2,798(3.10%)

基因表达水平统计表如下：

genelD	ck	wll
EPIOSAG00000042159	0.0	0.0
EPIOSAG00000042158	0.0	0.0
EPIOSAG00000042157	0.0	0.0
EPIOSAG00000042156	0.0	0.0
EPIOSAG00000042155	3.20760632776	20.6655541236

2. RNA-Seq相关性检查

生物学实验一般都需要设置重复对照，目的是为了证明实验的可重复性及评估结果的可靠性。相关系数越接近1，表明样品之间表达模式的相似度越高。Encode计划建议皮尔逊相关系数的平方 (R^2) 大于0.92 (理想的取样和实验条件下，RNA Standards v1.0 (https://www.encodeproject.org/documents/91494746-0ffe-4931-b219-a09802ce1cfa/@@download/attachment/RNA_standards_v1_2011_May.pdf))。而实际项目操作中，要求 R^2 至少要大于0.8，否则需要对样品做出合理的解释，或重新进行实验。

CATEGORIES

分析模块导图

任务参数

数据质控

数据比对

高级质控

可变剪接

表达定量

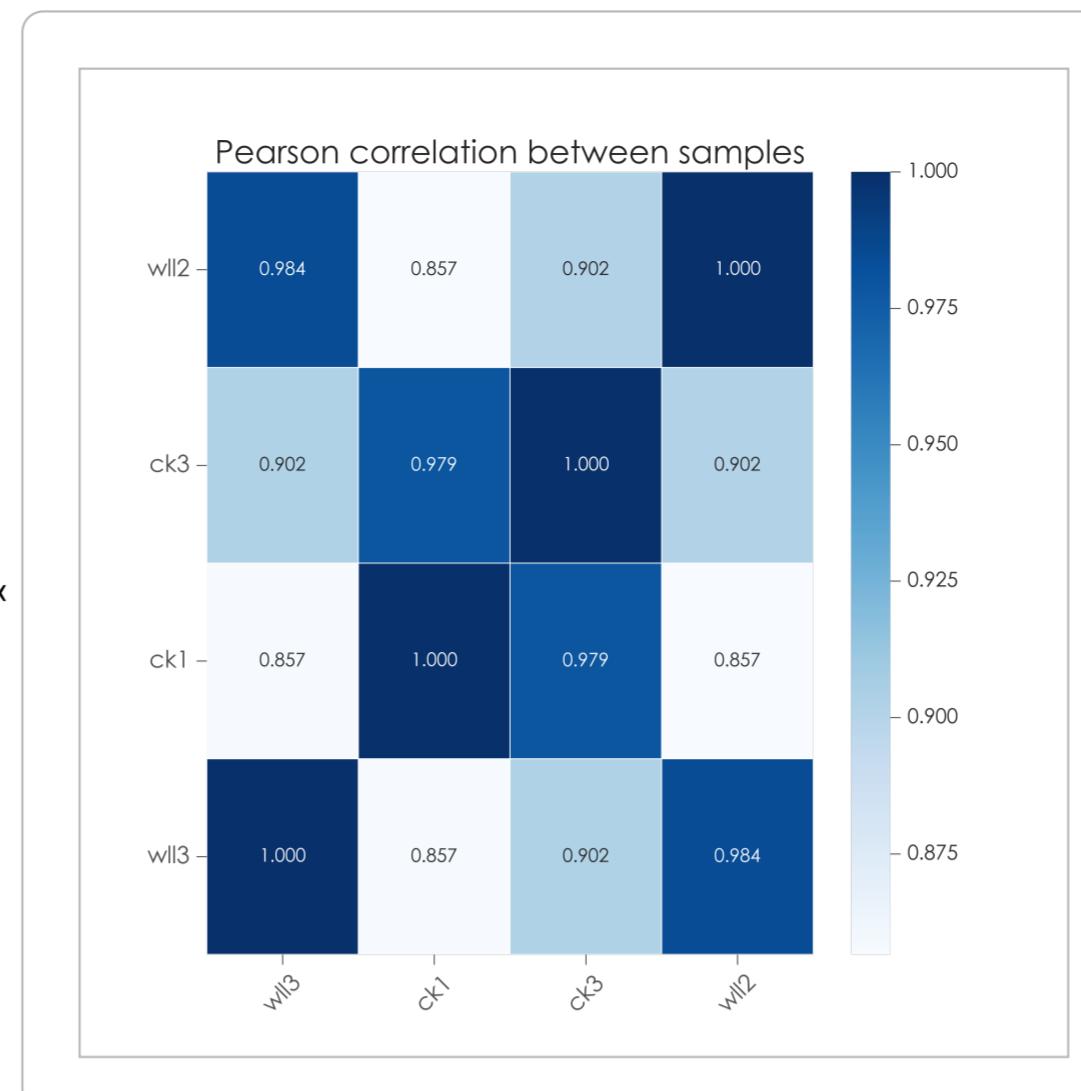
差异分析

GO富集分析

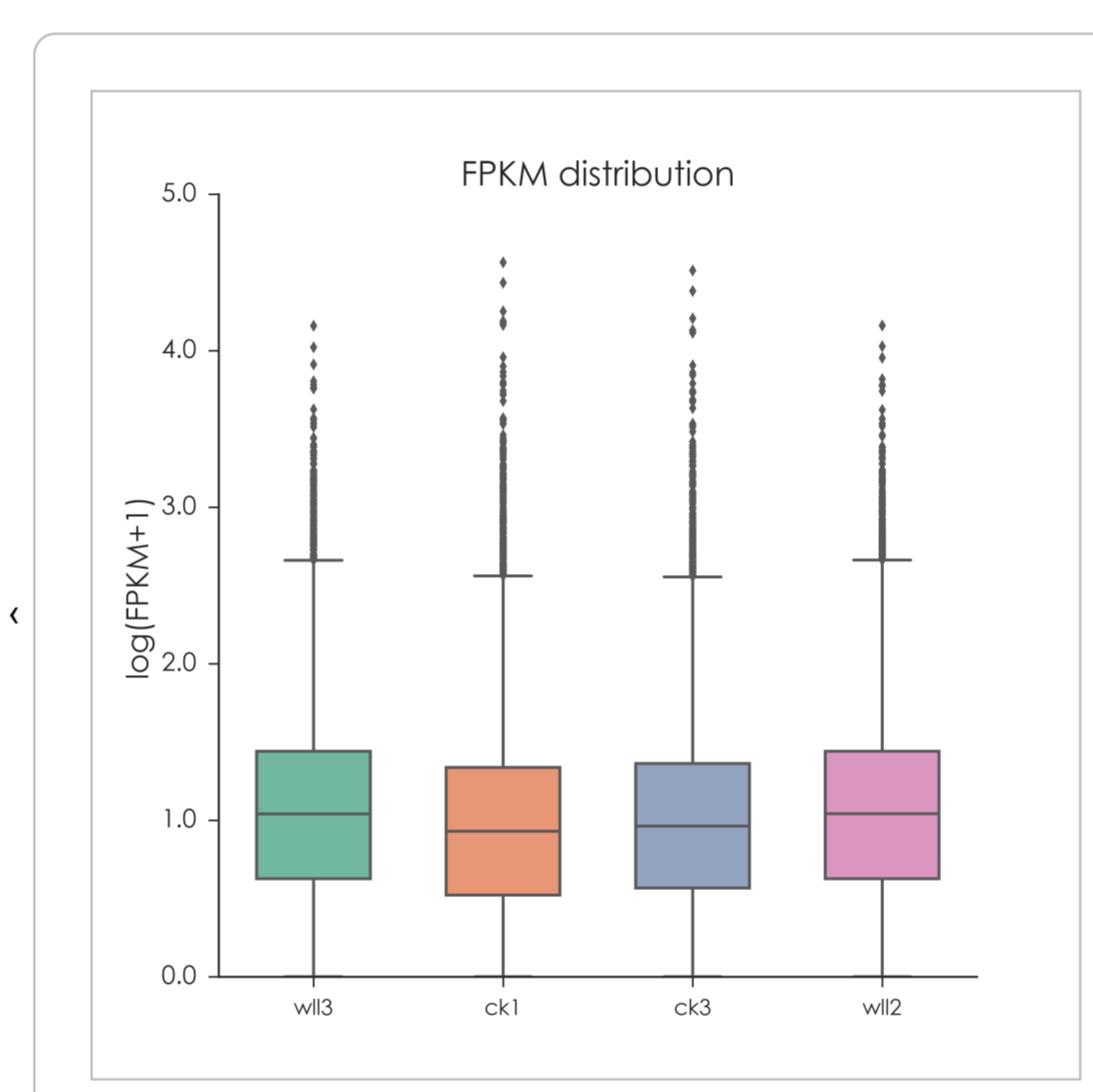
KEGG富集分析

PPI互作网络分析

参考文献



1/7 RNA-Seq相关性检查



1/3 样本表达水平对比

基因表达分析

1. 差异表达基因筛选

根据基因表达水平分析中得到的readcount数据来进行基因差异表达分析。对于有生物学重复的样品，优先使用DESeq2 (<http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html>) 进行差异表达分析。DESeq2使用的方法是基于负二项分布模型的算法，即第 i 个基因在第 j 个样本中的 read count 值为 K_{ij} ，则有： $K_{ij} \sim NB(\mu_{ij}, \sigma_{ij}^2)$ 。对于无生物学重复的样品，先用edgeR (<http://www.bioconductor.org/packages/release/bioc/html/edgeR.html>) (Robinson et.al., 2010; 版本: 3.12.1) 的TMM对readcount数据进行标准化处理后，再用DEGseq (<http://www.bioconductor.org/packages/release/bioc/html/DEGseq.html>) (Wang et.al.,; 版本: 1.24.0) 进行差异表达分析。

对于无生物学重复的实验，为避免引入实验误差，应该对结果进行严格控制，对差异基因进行筛选的阈值一般为： $|log_2(FoldChange)| > 1$ 且 $p\text{-Adjusted} < 0.005$ 。对于有生物学重复的实验，由于DESeq已经进行了实验误差的控制，我们对差异基因筛选的标准一般为： $p\text{-Adjusted} < 0.05$ 。

基因表达对比列表如下：

gene_id	wll3	ck1	ck3	wll2	log2FoldChange	pval	padj	bDiff
Os12g0292301	1095	3785	3864	1091	2.08516438389	1.43739253767e-62	2.07113890754e-58	True
Os01g0639900	1572	5268	5386	1579	2.03684939244	5.74517028276e-62	4.13910793022e-58	True
Os12g0292400	1027	3728	3787	1097	2.09949179943	6.12061200025e-58	2.93972994372e-54	True
Os02g0103850	419	1490	1520	440	2.08250901015	8.91677833385e-52	3.21204647531e-48	True
Os02g0103800	457	1447	1486	461	1.95171492302	6.51616090106e-49	1.87782724847e-45	True

① gene_id: 基因编号

CATEGORIES

- ② sample1: 样品的readcount值
- ③ sample2: 样品的readcount值
- ④ log2FoldChange: $\log_2(\text{Sample1}/\text{Sample2})$
- ⑤ pval: 统计学差异显著性水平, 即错误拒绝H0的概率
- ⑥ padj(qvalue): 多重假设检验校正后的p-Value
- ⑦ bDiff: 差异标记, True为差异, False为非差异

分析模块导图

任务参数

数据质控

数据比对

高级质控

可变剪接

表达定量

差异分析

GO富集分析

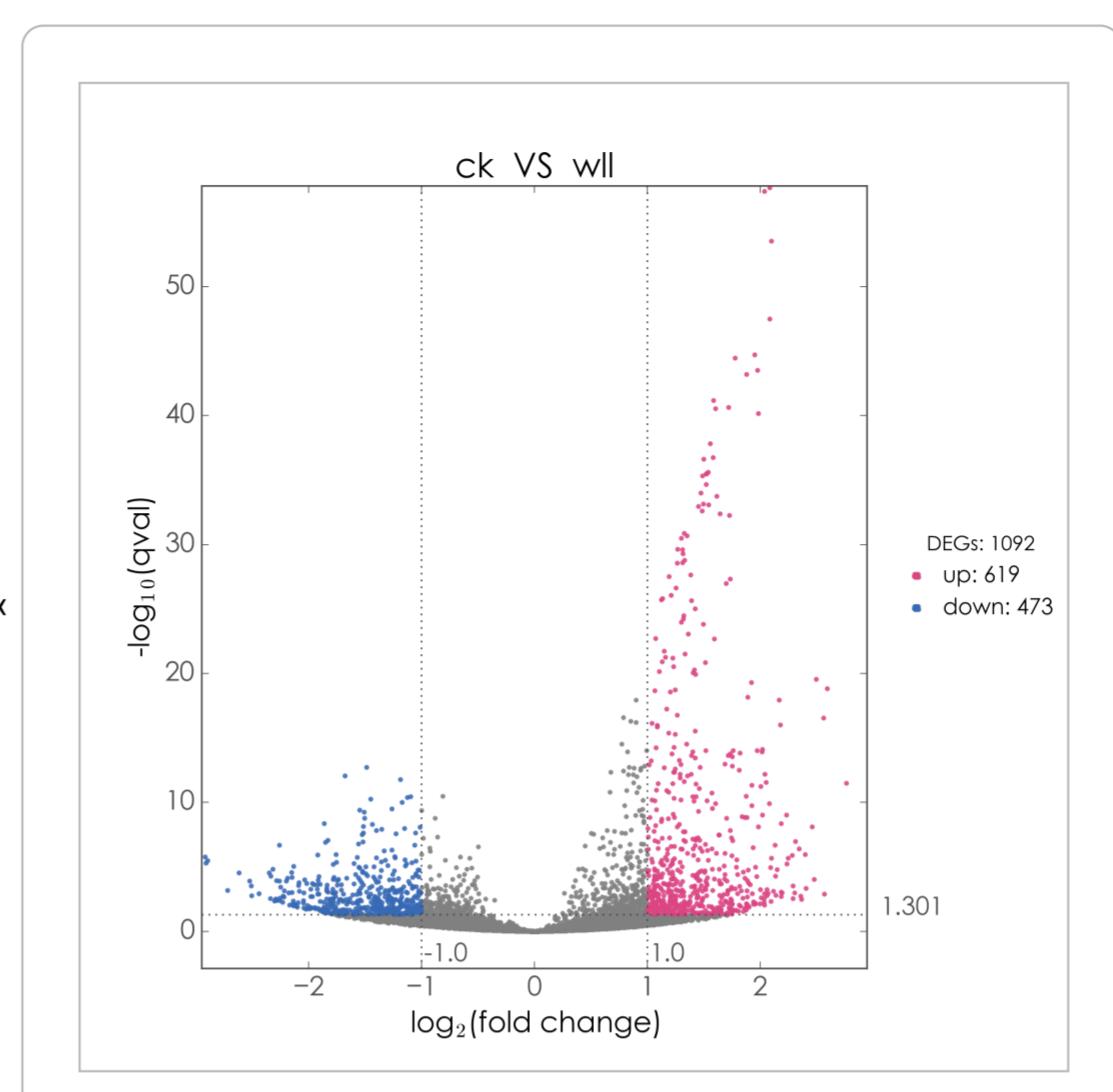
KEGG富集分析

PPI互作网络分析

参考文献

2. 差异表达基因火山图

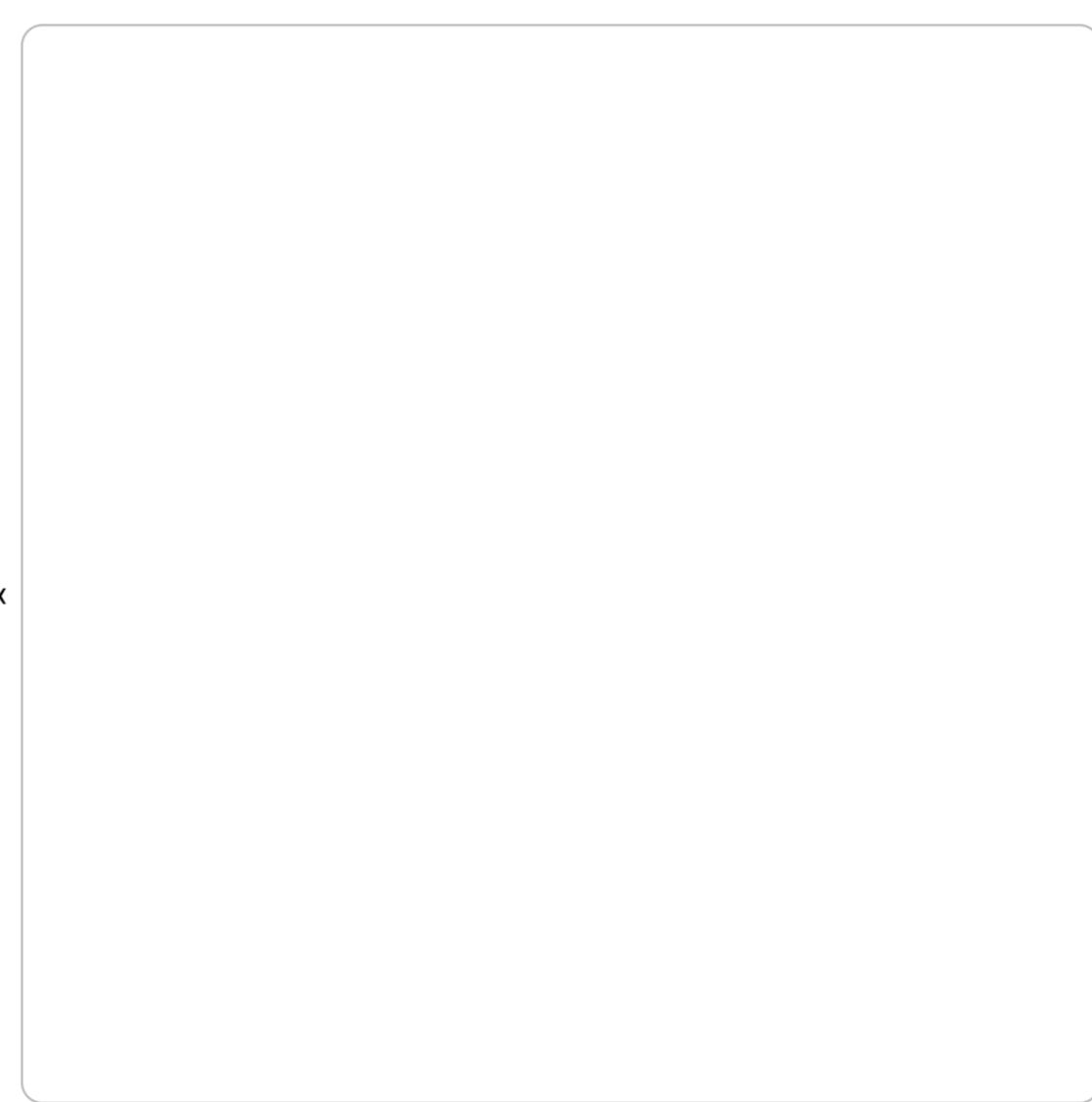
用火山图可以展示差异基因的整体分布情况。



1/1 差异基因火山图

3. 差异表达韦恩图

组间两两差异基因韦恩图以及所有组间韦恩图。



1/0 组间两两差异基因韦恩图以及所有组间韦恩图。

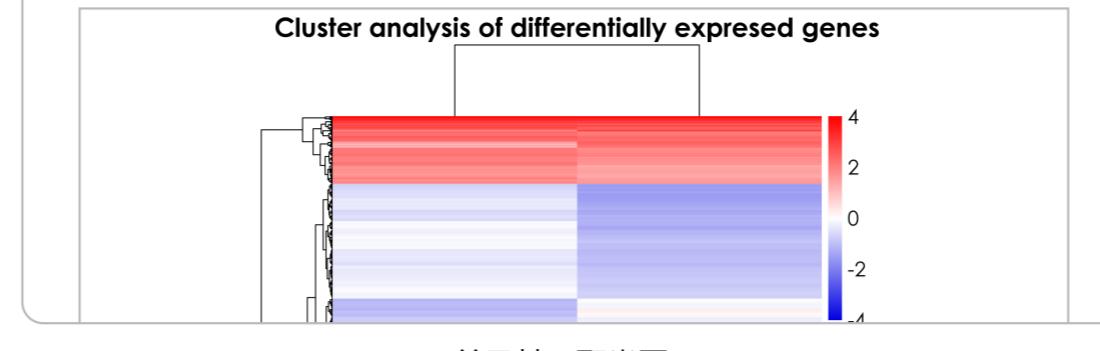
4. 差异基因表达聚类

差异基因表达聚类分析可用于评估差异基因在不同实验条件下的表达模式; 通过将表达模式相同或相近的基因聚集成类, 从而识别未知基因的功能或已知基因的未知功能。以不同实验条件下的差异基因的FPKM值为表达水平, 做层次聚类(hierarchical clustering)分析, 不同颜色的区域代表不同的聚类分组信息, 颜色相近聚类区内的基因表达模式相近, 说明这些基因可能具有相似的功能或参与调控同一的代谢通路。

除了对差异基因表达量FPKM进行层次聚类分析, 还分别用H-cluster、K-means和SOM等三种方法对差异基因的相对表达水平值 $\log_2(\text{ratios})$ 进行聚类。不同的聚类算法分别将差异基因分为若干cluster, 同一cluster中的基因在不同的处理条件下具有相似的表达水平变化趋势。

CATEGORIES

分析模块导图



任务参数

数据质控

数据比对

高级质控

可变剪接

表达定量

差异分析

GO富集分析

KEGG富集分析

PPI互作网络分析

参考文献

差异表达基因GO富集分析

GO (Gene Ontology, <http://www.geneontology.org>) 是一个国际标准化的基因功能分类体系。旨在建立一个适用于各物种的，对基因和蛋白质功能进行限定和描述的，并能随着研究不断深入而更新的语言词汇标准。GO分为分子功能 (Molecular Function)、生物过程 (biological process)、和细胞组成 (cellular component) 三个ontology。GO的基本单位为term，每个term对应一个功能或属性。

GO富集分析采用的软件为Goseq (<http://www.bioconductor.org/packages/release/bioc/html/goseq.html>) (版本：1.22.0, Young et al., 2010)，此方法基于 Wallenius non-central 超几何分布。相对于普通的超几何分布(Hyper-geometric distribution)的算法，该方法认为从某个类别中抽取个体的概率与从该类别之外抽取一个个体的概率是不同的，而这种概率的不同是通过对基因长度的偏好性进行估计得到的，从而能更为准确地计算出差异数所富集的GO term的概率。

1. 差异基因GO富集列表

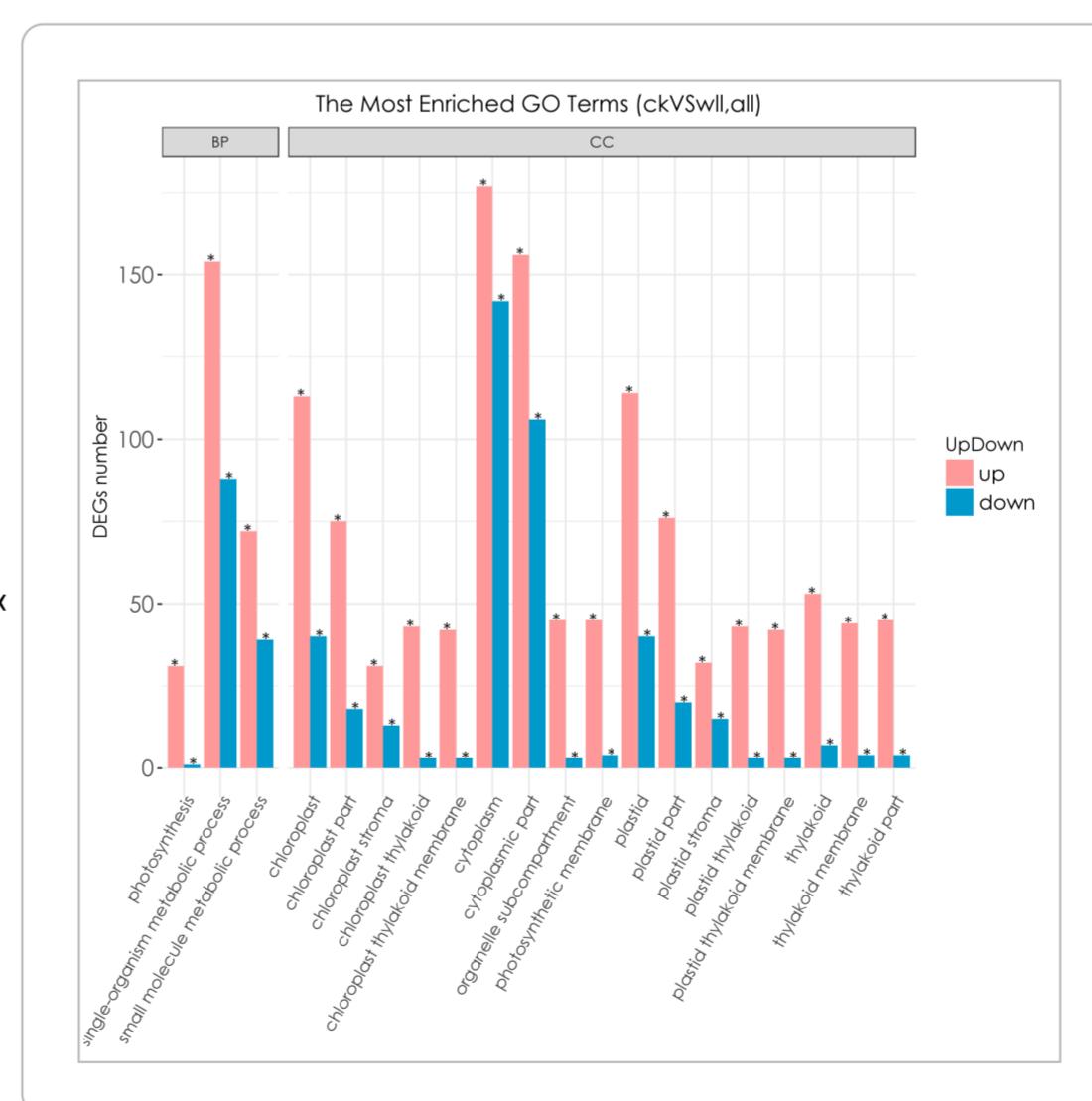
差异基因GO富集列表如下：

category	over_represented_pvalue	under_represented_pvalue	numDEInCat	numInCat	term	ontolo
GO:0009507	3.6451e-44	1	153	1261	chloroplast	CC
GO:0009536	1.8961e-43	1	154	1293	plastid	CC
GO:0044435	1.5229e-32	1	96	689	plastid part	CC
GO:0044434	7.0796e-31	1	93	681	chloroplast part	CC
GO:0009579	9.9921e-27	1	60	355	thylakoid	CC

- ① Category: Gene Ontology数据库中唯一的编号
- ② Over represented pvalue: 富集分析统计学显著水平
- ③ NumDEInCat: 与该GO相关的差异表达基因数
- ④ NumInCat: GO注释中与该GO相关的基因数目
- ⑤ Term: Gene Ontology功能的描述信息
- ⑥ Ontology: GO的类别(CC: 细胞组分; BP: 生物学过程; MF: 分子功能)
- ⑦ Correct: 对多重假设检验校正后的P-Value

2. 差异基因GO富集柱状图

差异表达基因GO富集柱状图能够直观的反映出差异表达基因在生物过程(biological process, BP)、细胞组分(cellular component, CC)和分子功能(molecular function, MF)的GO term上富集的分布情况。前30个富集最显著的GO term会在如下柱形图中展示，如果不足30个，则全部展示：



1/5 GO富集柱状图

3. 差异基因GO富集DAG图

差异表达基因GO富集分析结果可以用有向无环图(Directed Acyclic Graph, DAG)进行图形化展示。图中的分支代表包含关系，从上至下所定义的功能范围越来越小，有向无环图的主节点通常是GO富集分析结果的前10位，并通过包含关系，将相关联的GO Term一起展示，颜色的深浅代表富集程度。生物过程(biological process, BP)、分子功能(molecular function, MF)和细胞组分(cellular component, CC)的DAG图分别绘制。

CATEGORIES

分析模块导图

任务参数

数据质控

数据比对

高级质控

可变剪接

表达定量

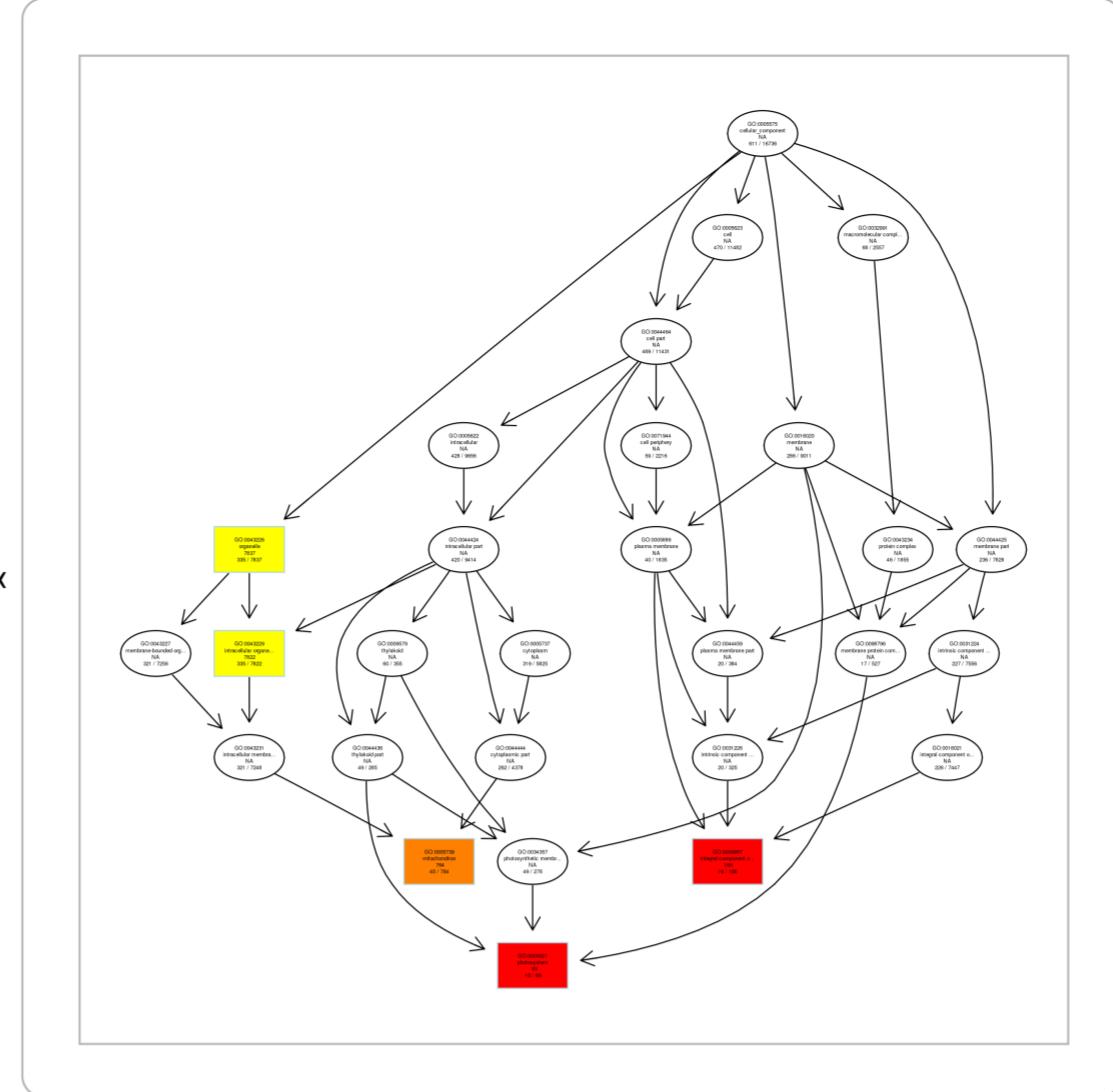
差异分析

GO富集分析

KEGG富集分析

PPI互作网络分析

参考文献



1/3 GO富集有向无环图

差异表达基因KEGG富集分析

KEGG (Kyoto Encyclopedia of Genes and Genomes, <http://www.kegg.jp> (<http://www.kegg.jp>)) 是系统分析基因功能、基因组信息数据库，它有助于研究者把基因及表达信息作为一个整体网络进行研究。作为Pathway相关的主要公共数据库(Kanehisa et al., 2008)，KEGG提供的整合代谢途径(pathway)查询十分出色，包括碳水化合物、核苷、氨基酸等的代谢及有机物的生物降解。不仅提供了所有可能的代谢途径，而且对催化各步反应的酶进行了全面的注解，包含有氨基酸序列、PDB库的链接等等，是进行生物体内代谢分析、代谢网络研究的强有力工具。

在生物体内，不同基因相互协调行使其生物学功能，通过Pathway显著性富集能确定差异表达基因参与的最主要生化代谢途径和信号转导途径。Pathway显著性富集分析以KEGG 数据库中Pathway为单位，应用超几何检验，找出与整个转录组背景相比，在差异表达基因中显著性富集的Pathway。

1. 差异表达基因KEGG富集列表

差异基因KEGG富集列表如下：

Term	Database	ID	Input number	Background number	P-Value
Valine, leucine and isoleucine degradation	KEGG PATHWAY	osa00280	12	40	9.61826927557e-12
Biosynthesis of secondary metabolites	KEGG PATHWAY	osa01110	36	848	3.77868781789e-08
Protein processing in endoplasmic reticulum	KEGG PATHWAY	osa04141	16	189	6.52070859363e-08
Metabolic pathways	KEGG PATHWAY	osa01100	47	1573	6.2254494808e-06
Histidine metabolism	KEGG PATHWAY	osa00340	5	16	1.07596364834e-05
Valine, leucine and isoleucine biosynthesis	KEGG PATHWAY	osa00290	4	14	0.000115545526014

- ①Term: KEGG中通路描述
- ②Database: 数据库
- ③ID: KEGG ID
- ④Input number: 输入基因属于该Term的总数
- ⑤Background number: 背景基因属于该Term的总数
- ⑥P-Value: 超几何检验p值
- ⑦Corrected P-Value: 对多重假设检验校正后的P-Value

2. 差异表达基因KEGG富集散点图

差异表达基因KEGG富集分析结果可以通过散点图进行图形化展示。Rich factor、Qvalue和富集到此通路上的基因个数被用来衡量KEGG富集程度。其中Rich factor指pathway中富集到的差异表达基因数量与注释基因数量的比值。Rich factor值越大，说明富集的程度越大。Qvalue (取值范围0-1) 是通过多重假设检验校正之后的Pvalue，Qvalue的值越接近于零，说明富集越显著。下图展示了20条富集最显著的pathway，若富集的pathway条目不足20条，则全部展示。

CATEGORIES

分析模块导图

3. KEGG富集通路图

任务参数

将差异表达基因标注到通路图中可以方便地查看差异基因在通路图中的分布情况。查看方法见结果目录中的说明文档。

数据质控

数据比对

高级质控

可变剪接

表达定量

差异分析

GO富集分析

KEGG富集分析

PPI互作网络分析

参考文献

Statistics of Pathway Enrichment

Galactose metabolism

Histidine metabolism

qvalue

1/3 差异基因KEGG富集散点图

差异基因KEGG富集散点图

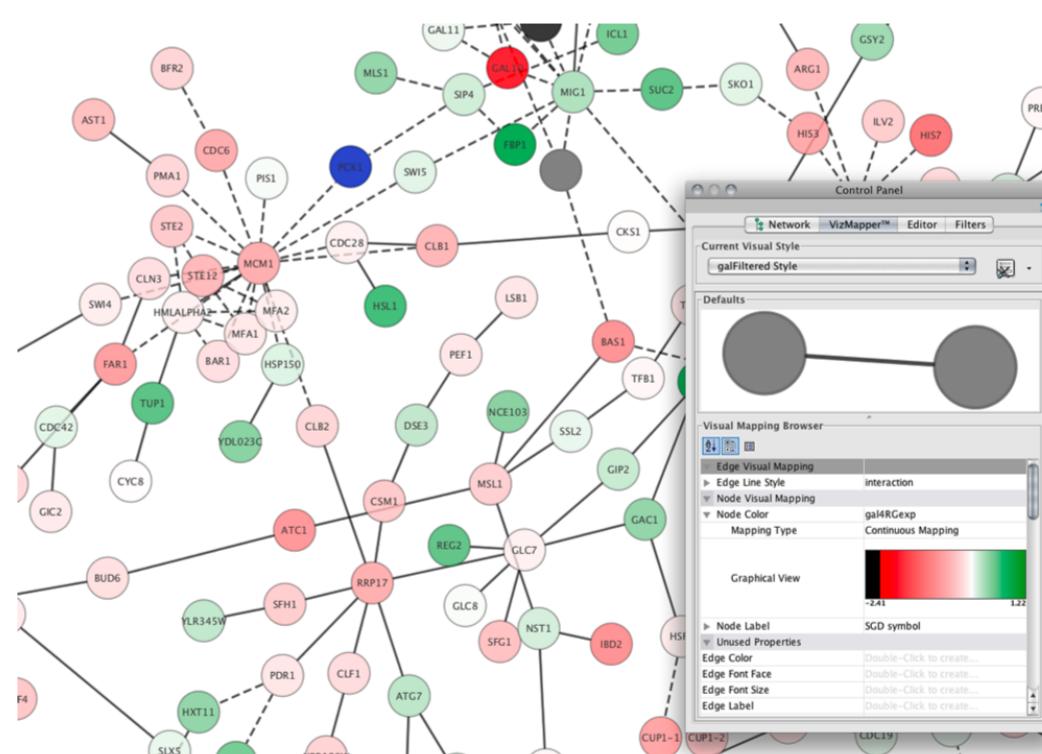
差异基因KEGG富集散点图

1/0 差异基因KEGG富集通路图

蛋白互作分析

STRING (<http://string-db.org>) 是一个强大的蛋白质互作数据库。对于数据库中包含的物种，我们直接将差异表达基因的互作关系从数据库中提取出来构建互作网络。而对于数据库中没有的物种，可以将差异表达基因序列通过blastx与数据库包含的参考物种的蛋白质序列比对，利用比对上的蛋白质互作关系构建互作网络。

将得到的差异表达基因互作网络数据文件导入Cytoscape软件，既可以可视化编辑。Cytoscape软件的使用方法可以查看其使用说明文档。使用Cytoscape软件查看基因互作网络的效果如下：



参考文献

Anders, S.(2010). HTSeq: Analysing high-throughput sequencing data with Python.(HTSeq)

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.*(DESeq)

Anders, S. and Huber, W. (2012). Differential expression of RNA-Seq data at the gene level-the DESeq package.(DESeq)

Kanehisa, M., M. Araki, et al. (2008). KEGG for linking genomes to life and the environment. *Nucleic acids research*.(KEGG)

Kim, D., G. Pertea, et al. (2012).TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.(TopHat2)

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*(Bowtie)

Langmead, B. and S. L. Salzberg (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*.(Bowtie 2)

Mao, X., Cai, T., Olyarchuk, J.G., Wei, L. (2005). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics*.(KOBAS)

Marioni, J. C., C. E. Mason, et al. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*.

Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research.*(GATK)

Mortazavi, A., B. A. Williams, et al. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods.*

Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.(edgeR)*

CATEGORIES

- 分析模块导图
- 任务参数
- 数据质控
- 数据比对
- 高级质控
- 可变剪接
- 表达定量
- 差异分析
- GO富集分析
- KEGG富集分析
- PPI互作网络分析
- 参考文献

© 2015-2017 拓美科 8omics.com 版权所有 | 咨询热线：010-53659701