

十一、其他招标文件要求的内容

结题报告模板



解析基因序列·诠释生命密码·改善人类生活

派森诺 | Personalbio

项目结题报告书

仅供招标投标项目使用

项目名称：真核有参转录组测序分析

客户单位：

制作日期：



上海派森诺生物科技股份有限公司
SHANGHAI PERSONALBIO TECHNOLOGY CO.,LTD.

真核有参转录组

1 项目整体流程概况

1.1 项目信息

Table 1: 项目概括信息

项目编号	XXXXXXXXXXXXXX
开题单号	XXXXXXXXXXXXXX
项目类型	有参考基因组转录组测序
物种名	小鼠 (<i>Mus musculus</i>)
样本形式	原样
样本数目	12
测序平台	DNBSEQ
数据量	6G/样本
分析项目	标准生物信息分析
技术支持	transsupport@personalbio.cn

1.2 实验流程

派森诺从样本提取到上机测序，有严格的样本检测、质控流程，通过各环节对样本质量控制，确保数据的真实可信。操作流程如下：



实验流程图

本项目通过Oligo (dT) 磁珠富集总RNA中带有polyA结构的mRNA, 采用离子打断的方式, 将RNA打断到长度300bp左右的片段。选择长度为300bp的片段, 这是因为接头长度是固定的, 如被打断的片段长度较短, 将导致接头序列的比例偏高, 从而降低了有效数据的比例; 如被打断的片段长度较长, 则不利于上机测序过程中簇的生成。以RNA为模板, 用6碱基随机引物和逆转录酶合成cDNA第一链, 并以第一链cDNA为模板进行第二链cDNA的合成。

文库构建完成后, 采用PCR扩增进行文库片段富集, 之后根据片段大小进行文库选择, 文库大小在450bp。接着, 通过Agilent 2100 Bioanalyzer对文库进行质检, 再对文库总浓度及文库有效浓度进行检测。然后根据文库的有效浓度以及文库所需数据量, 将含有不同Index序列 (各样本加上不同的Index, 最后根据Index区分各样本的下机数据) 的文库按比例进行混合。混合文库统一稀释到2nM, 通过碱变性, 形成单链文库。

样品经过RNA抽提、纯化、建库之后, 采用第二代测序技术 (Next-Generation Sequencing, NGS), 基于测序平台, 对这些文库进行双末端 (Paired-end, PE) 测序。

1.3 分析流程

首先对原始下机数据 (Raw Data) 进行过滤, 将过滤后得到的高质量序列 (Clean Data) 比对到该物种的参考基因组上。根据比对结果, 计算每个基因的表达量。在此基础上, 进一步对样品进行表达差异分析、富集分析和聚类分析。对比对上的Reads进行拼接, 还原出转录本序列。



1.4 参考基因组

本次项目使用的基因组信息如下表所示。

Table 2: 参考基因组信息

Genome	Mus_musculus.GRCm39.dna.primary_assembly.fa
Path	http://www.ensembl.org/Mus_musculus/Info/Annotation
Genebuild by	Ensembl
Database version	108.39
Base Pairs	2728222451

参考基因组注释情况统计见下表, 如果想了解每个数据库的介绍, 请查看附录-数据库介绍。

Table 3: 参考基因组注释信息

Database	Annotated	Percent
Ensembl	21959	100.00
GO	21567	98.21
KEGG	16519	75.22

Database	Annotated	Percent
EC	3785	17.23
eggNOG	20031	91.22
UniProtAC	20126	91.65
Entrez_geneID	20934	95.33

注：Database：数据库名称；

Annotated：该物种注释文件中的基因在以上数据库中注释的数目；

Percent：该注释文件中的基因在以上数据库中注释的比例

1.5 文库基本情况

每一个文库的基本情况见下表

Table 4: 文库基本情况

Sample	Lib. Name	Lib. Insert Size	Sequencing Platform	Sequencing Mode
TT001	TT001	380bp	NovaSeq	Paired-end, 2×150bp
TT002	TT002	380bp	NovaSeq	Paired-end, 2×150bp
TT003	TT003	380bp	NovaSeq	Paired-end, 2×150bp
WW001	WW001	380bp	NovaSeq	Paired-end, 2×150bp
WW002	WW002	380bp	NovaSeq	Paired-end, 2×150bp
WW003	WW003	380bp	NovaSeq	Paired-end, 2×150bp
SA001	SA001	380bp	NovaSeq	Paired-end, 2×150bp
SA002	SA002	380bp	NovaSeq	Paired-end, 2×150bp
SA003	SA003	380bp	NovaSeq	Paired-end, 2×150bp
EB001	EB001	380bp	NovaSeq	Paired-end, 2×150bp
EB002	EB002	380bp	NovaSeq	Paired-end, 2×150bp
EB003	EB003	380bp	NovaSeq	Paired-end, 2×150bp

注：Sample：样品名称；

Lib. Name：文库名

Lib. Insert Size：文库插入片段长度；

Sequencing platform：测序平台；

Sequencing Mode：测序模式。

2 原始数据处理与质控

2.1 数据整理

样品经过上机测序，得到图像文件，由测序平台自带软件进行转化，生成FASTQ的原始数据（Raw Data），即下机数据。我们对每个样品的下机数据（Raw Data）分别进行统计，包括样品名、测序数据量、Q30、模糊碱基所占百分比、GC含量、以及Q20（%）和Q30（%）。

Table 5: 下机数据统计

Sample	Raw Reads No	Raw Bases(bp)	Q30(bp)	GC(%)	N(%)	Q20(%)	Q30(%)
SA001	39131684	5908884284	5682167526	49.41	0.004079	98.64	96.16
SA002	47453592	7165492392	6910961928	49.66	0.004088	98.76	96.45
SA003	41535444	6271852044	6022781514	49.52	0.004090	98.59	96.03

Sample	Raw Reads No	Raw Bases(bp)	Q30(bp)	GC(%)	N(%)	Q20(%)	Q30(%)
EB001	43153446	6516170346	6250684909	49.57	0.003349	98.59	95.93
EB002	38021682	5741273982	5520388920	49.68	0.004075	98.64	96.15
EB003	44209370	6675614870	6409173436	49.86	0.004164	98.59	96.01
TT001	48757986	7362455886	7093621662	46.57	0.005327	98.70	96.35
TT002	51466798	7771486498	7466099093	46.85	0.005424	98.58	96.07
TT003	44739992	6755738792	6503707799	46.63	0.005419	98.67	96.27
WW001	58659476	8857580876	8539203592	46.97	0.005404	98.72	96.41
WW002	62872978	9493819678	9137821381	47.00	0.005047	98.66	96.25
WW003	43955736	6637316136	6380243492	46.90	0.005432	98.61	96.13

注: Sample: 样品名;
Raw_Read_No: Reads总数;
Raw_Bases (bp) : 碱基总数;
Q30 (bp) : 碱基识别准确率在99.9%以上的碱基总数;
GC (%) : GC含量;
N (%) : 模糊碱基所占百分比;
Q20 (%) : 碱基识别准确率在99%以上的碱基所占百分比;
Q30 (%) : 碱基识别准确率在99.9%以上的碱基所占百分比。

结果文件

2.2 数据过滤

测序数据包含一些带接头、低质量的Reads, 这些序列会对后续的信息分析造成很大的干扰, 因此需要对测序数据进行进一步过滤。数据过滤的标准主要包括:

- 1) 采用Fastp去除3'端带头的序列;
- 2) 去除平均质量分数低于Q20的Reads。

Table 6: 数据过滤统计

Sample	Clean Reads No	Clean Data(bp)	Clean Reads(%)	Clean Data(%)
SA001	38526726	5804383196	98.45	98.23
SA002	46830014	7054814509	98.69	98.46
SA003	40866326	6156683319	98.39	98.16
EB001	42585694	6415566728	98.68	98.46
EB002	37429916	5637926077	98.44	98.20
EB003	43505370	6553912051	98.41	98.18
TT001	48114530	7249697964	98.68	98.47
TT002	50659064	7633413418	98.43	98.22
TT003	44153542	6653511853	98.69	98.49
WW001	57923458	8727140832	98.75	98.53
WW002	62001026	9342505572	98.61	98.41
WW003	43328228	6528112844	98.57	98.35

注: Sample: 样品名;
Clean Read No: 高质量序列read数;

Clean Data (bp) : 高质量序列碱基数;
Clean Reads (%) : 高质量序列reads占测序reads的百分比;
Clean Data (%) : 高质量序列碱基占测序碱基的百分比。

结果文件

2.3 碱基质量分布

测序错误率受测序仪本身、测序试剂、样品等多个因素共同影响。对于RNASeq技术，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的长度增加而升高，这是由测序过程中化学试剂的消耗导致的，是高通量测序平台都具有的特征；
- 2) 前6个碱基的位置（即建库过程中反转录所需要的随机引物的长度）也会发生较高的测序错误率，这种错误是由随机引物和RNA模板的不完全结合引起的。

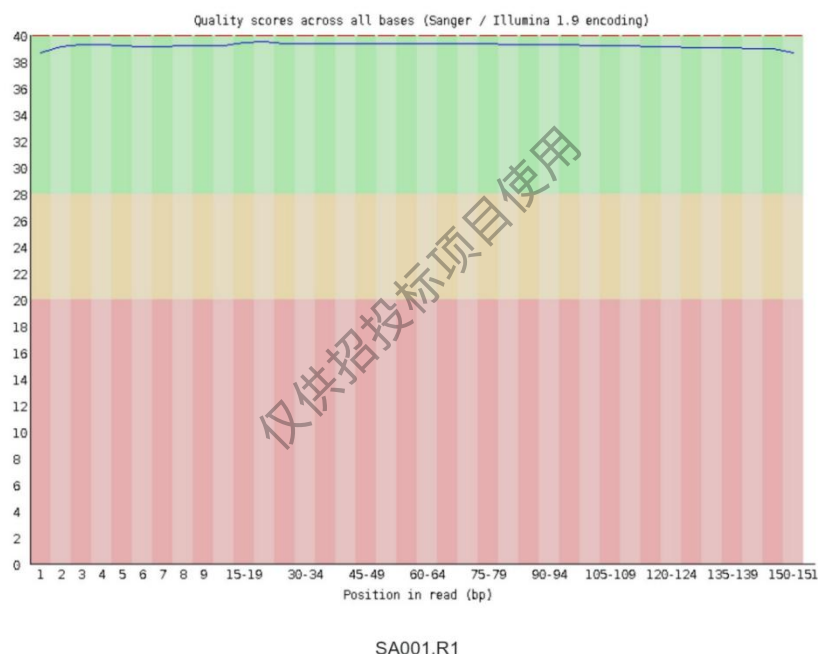


Fig 3: 单碱基质量分布图

注：横坐标是Reads中碱基位置（5'→3'），纵坐标是对应位点碱基Q值。

结果文件

2.4 碱基含量分布

碱基含量分布一般用于检测有无AT、GC分离现象。对于RNASeq来说，鉴于序列打断的随机性和G/C、A/T含量分别相等的原则，理论上每个测序循环中的GC含量相等、AT含量相等（如果是链特异性建库，可能会出现AT分离和/或GC分离），且在整个测序过程基本稳定不变，呈水平线。但在现有的高通量测序技术中，反转录合成cDNA时所用

的6bp的随机引物会引起前几个位置的核苷酸组成存在一定的偏好性，这种波动属于正常情况。

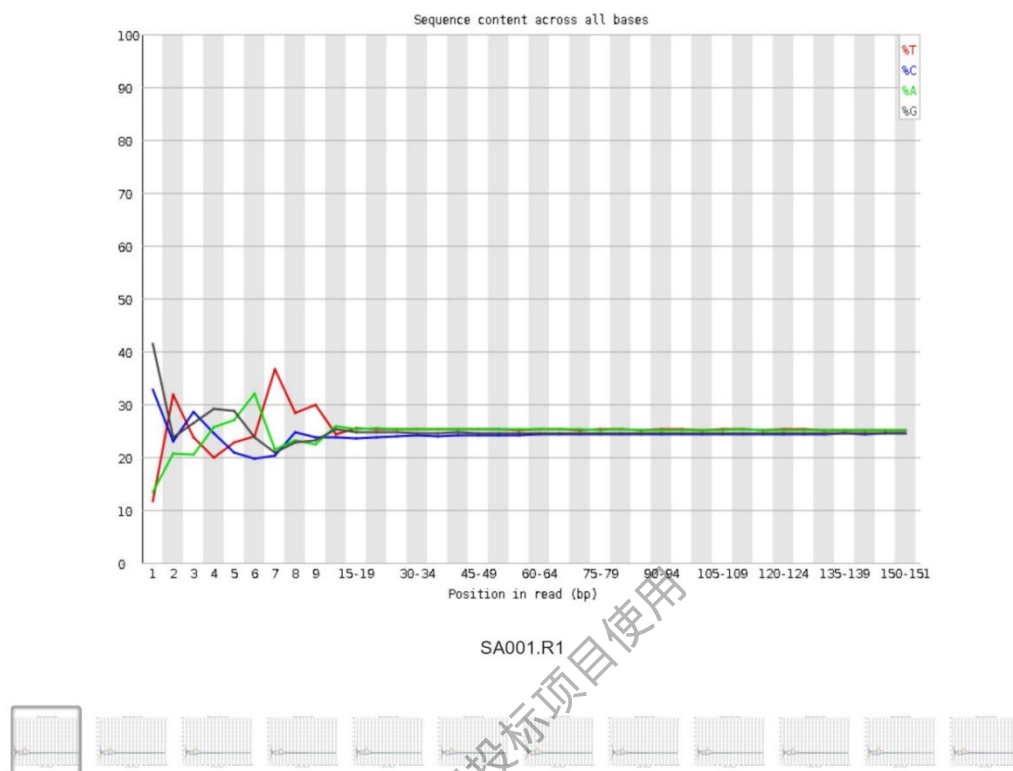


Fig 4: 碱基含量分布图

注：横坐标是Reads中碱基位置（5'→3'），纵坐标是该位点某碱基所占的比例统计。

结果文件

2.5 Reads平均质量分布

Reads平均质量分布主要用来检测测序数据的平均质量分布情况。峰尖代表主体Reads测序质量，峰宽表示整体测序质量分布。

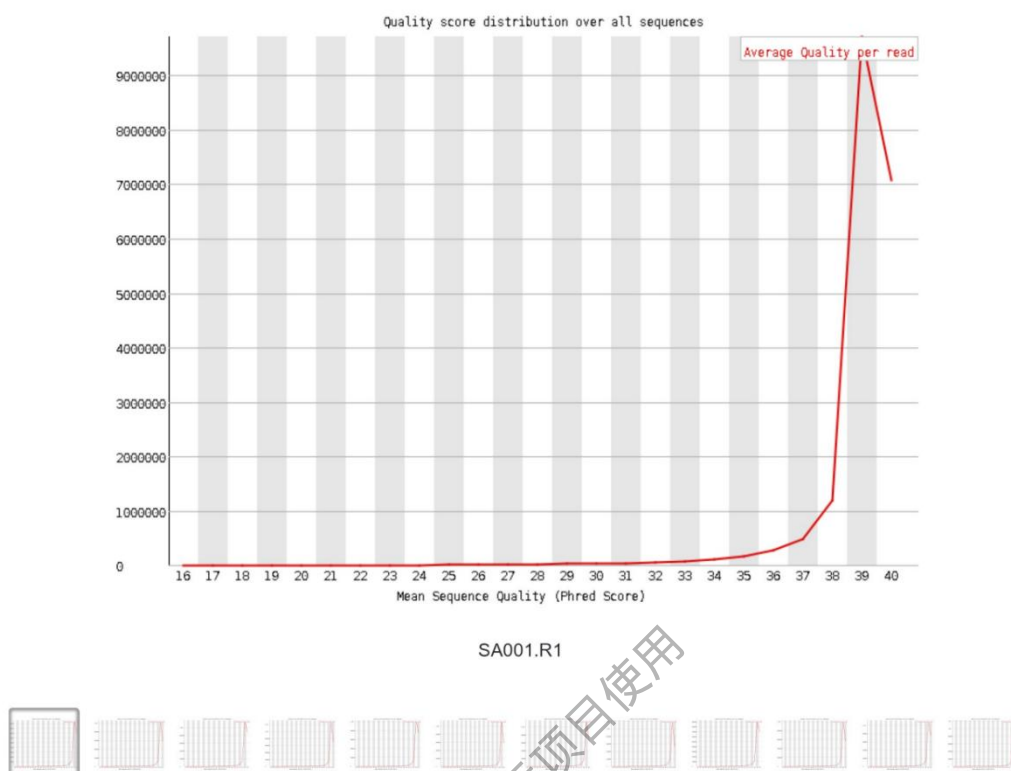


Fig 5: Reads平均质量分布图

注：Reads质量分布图，横坐标表示Reads的平均质量，纵坐标为对应平均质量值的Reads数目。

结果文件

3 比对分析

3.1 比对结果统计

使用TopHat2的升级版HISAT2 (<http://ccb.jhu.edu/software/hisat2/index.shtml>) 软件将过滤后的Reads比对到参考基因组上。HISAT2使用改进的BWT算法具有更快的速度并且资源占用较少。HISAT2比对时，对于非链特异性文库使用默认参数，链特异性文库需要指定文库类型（即first使用- ma-strandness RF，second使用- ma-strandness FR）。Mapping比例较低时的原因可能是：

- 1) 参考基因组组装不好，或者所测物种与参考基因组的亲缘关系较远；
- 2) 样本存在污染；
- 3) 样品的特殊前处理或者相对于参考基因组此样品本身的变异太大，导致Mapping Rate相对较低。

Table 7: 数据过滤统计

Sample	Clean_Reads	Total_Mapped	Multiple_Mapped	Uniquely_Mapped	Map_Events	Mapped_to_Gene	Mapp
SA001	38526726	38197767 (99.15%)	1539287 (4.03%)	36658480 (95.97%)	36658480	35485045 (96.80%)	111

Sample	Clean_Reads	Total_Mapped	Multiple_Mapped	Uniquely_Mapped	Map_Events	Mapped_to_Gene	Mapped_to_InterGene	Mapped_to_exon
SA002	46830014	46450237 (99.19%)	1860751 (4.01%)	44589486 (95.99%)	44589486	43190745 (96.86%)	135	135
SA003	40866326	40516548 (99.14%)	1632682 (4.03%)	38883866 (95.97%)	38883866	37588057 (96.67%)	125	125
EB001	42585694	42216497 (99.13%)	1720106 (4.07%)	40496391 (95.93%)	40496391	39177325 (96.74%)	135	135
EB002	37429916	37117051 (99.16%)	1564735 (4.22%)	35552316 (95.78%)	35552316	34575587 (97.25%)	97	97
EB003	43505370	43138081 (99.16%)	1773071 (4.11%)	41365010 (95.89%)	41365010	40217198 (97.23%)	114	114
TT001	48114530	47651994 (99.04%)	3375177 (7.08%)	44276817 (92.92%)	44276817	43380813 (97.98%)	89	89
TT002	50659064	50154988 (99.00%)	3348276 (6.68%)	46806712 (93.32%)	46806712	45840677 (97.94%)	96	96
TT003	44153542	43727236 (99.03%)	3048586 (6.97%)	40678650 (93.03%)	40678650	39830771 (97.92%)	84	84
WW001	57923458	57348646 (99.01%)	3642085 (6.35%)	53706561 (93.65%)	53706561	52618657 (97.97%)	105	105
WW002	62001026	61423072 (99.07%)	3937364 (6.41%)	57485708 (93.59%)	57485708	56391896 (98.10%)	105	105
WW003	43328228	42907786 (99.03%)	2935753 (6.84%)	39972033 (93.16%)	39972033	39200176 (98.07%)	77	77

注：Sample：样品；

Clean_Reads：用于比对的序列总数；

Total Mapped：比对上参考基因组的序列总数，百分比为Total Mapped / Clean Reads；

Multiple Mapped：比对到多个位置的序列总数，百分比为Multiple Mapped / Total Mapped；

Uniquely Mapped：只比对到一个位置的序列总数，百分比为Uniquely Mapped / Total Mapped；

Mapped to Gene：比对到基因区域的Reads总数，百分比为Mapped to Gene / Map Events；

Mapped to InterGene：比对到基因间区的Reads总数，百分比为Mapped to InterGene / Map Events；

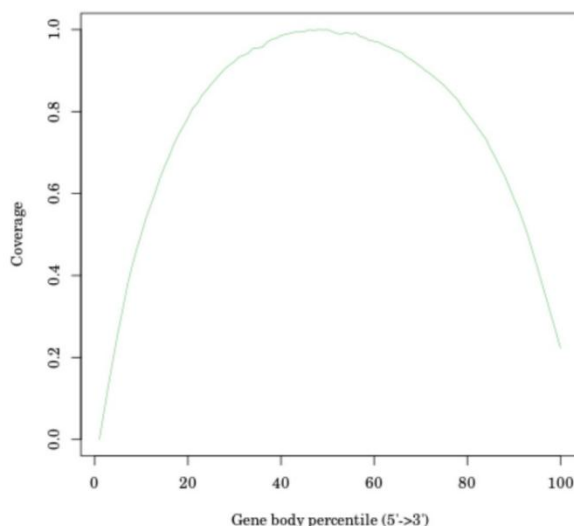
Mapped to exon：比对到外显子区域的Reads总数，百分比为Mapped to exon / Mapped to Gene。

结果文件

3.2 比对结果质控

3.2.1 基因覆盖均一度

测序Reads在基因上覆盖度的分布情况，展示每个样品所有基因的5'到3'区域上序列覆盖情况，用于评估测序结果的均一性（或是否有偏向性）。理想条件下，Reads在所有表达的基因上的分布应该呈现均一化分布。



SA001

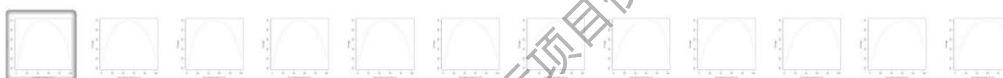


Fig 6: 基因覆盖均匀度图

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的横轴位置上相应区间内的序列条数的总和并归一化后的数值。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。

结果文件

3.2.2 饱和度分析

使用RSeQC分析表达量饱和度，即：对测序结果按5%，10%，15%...100%的比例进行抽样，对不同的抽样比例分别计算所有基因表达量（即每个基因计算20次），然后与实际表达量（假设100%抽样情况下得出的为实际表达量）进行比较，获得相对误差。

饱和度分析主要目的是评估所测数据量是否足够用于正确计算基因表达量。理论上在较少数据量情况下基因的表达量与实际表达量偏差较大，而当数据量达到饱和阈值后，数据量再增长，基因表达量也近乎不变，此时基因表达量不再受数据量的影响。

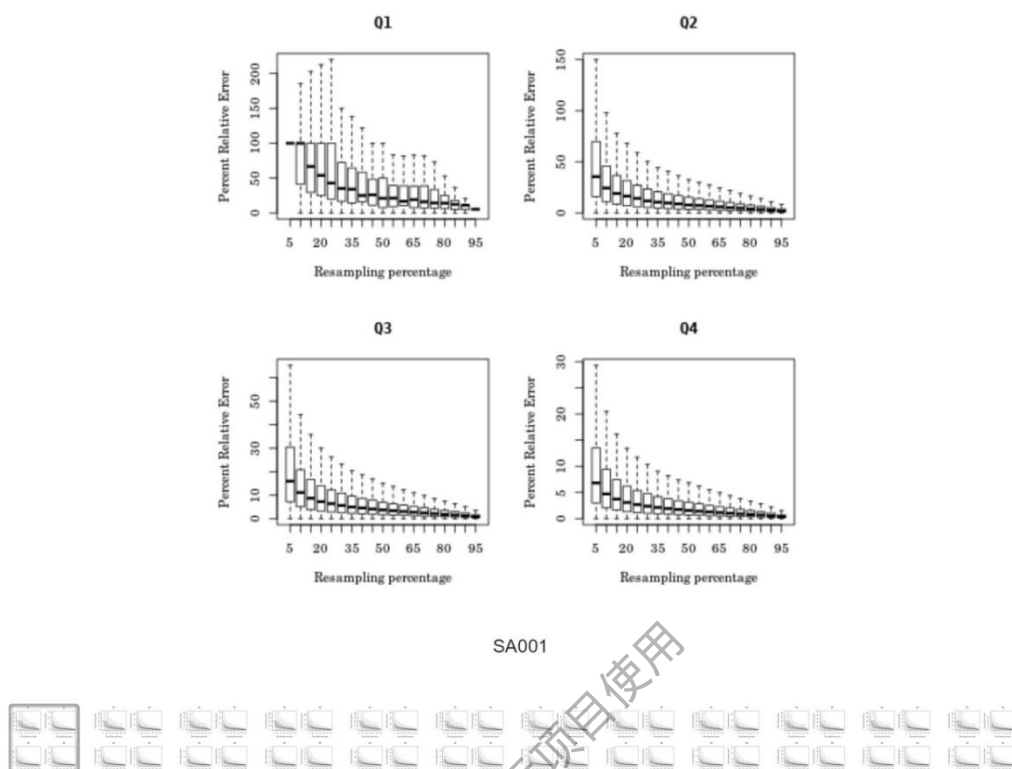


Fig. 7. 饱和度分析图

注：横坐标是重采样的比率，纵坐标为该比率下基因的表达量与实际表达量的相对误差。基因按表达量水平被划入四个组中，Q1：表达量水平居于倒数25%的基因；

Q2：表达量水平居于倒数25%-50%的基因；

Q3：表达量水平居于前25%-50%的基因；

Q4：表达量水平居于前25%的基因。

结果文件

3.2.3 比对区域分布统计

将比对到基因组上的Reads分布情况进行统计，定位区域分为CDS（编码区）、Intron（内含子）、Intergenic（基因间区）、UTR（5'和3'非翻译区）、TSS_up（转录起始位点上游）和TES_down（转录终止位点下游）

在基因组注释较为完全的物种中，通常比对到CDS（编码区）的reads含量最高。比对到Intron（内含子）区域的reads来源于pre-mRNA的残留或由可变剪切过程中发生的内含子保留事件导致的，而比对到Intergenic（基因间区）的Reads则可能转录自新基因或新的非编码RNA。

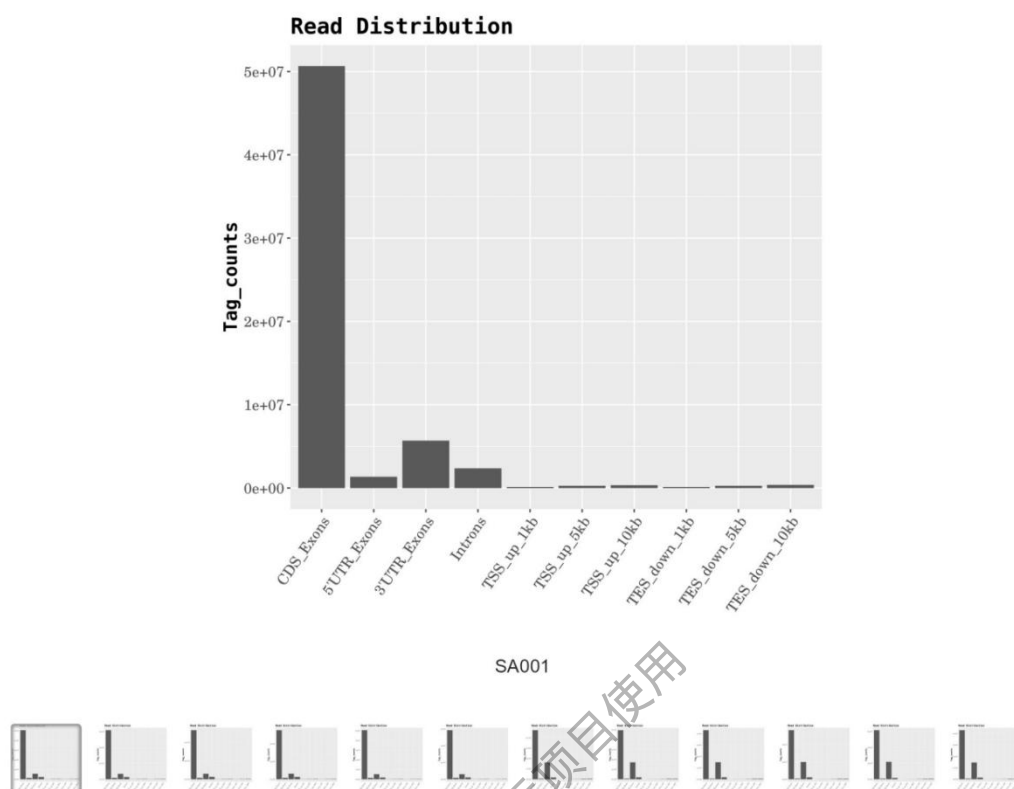


Fig 8: 比对结果统计（基因区域）

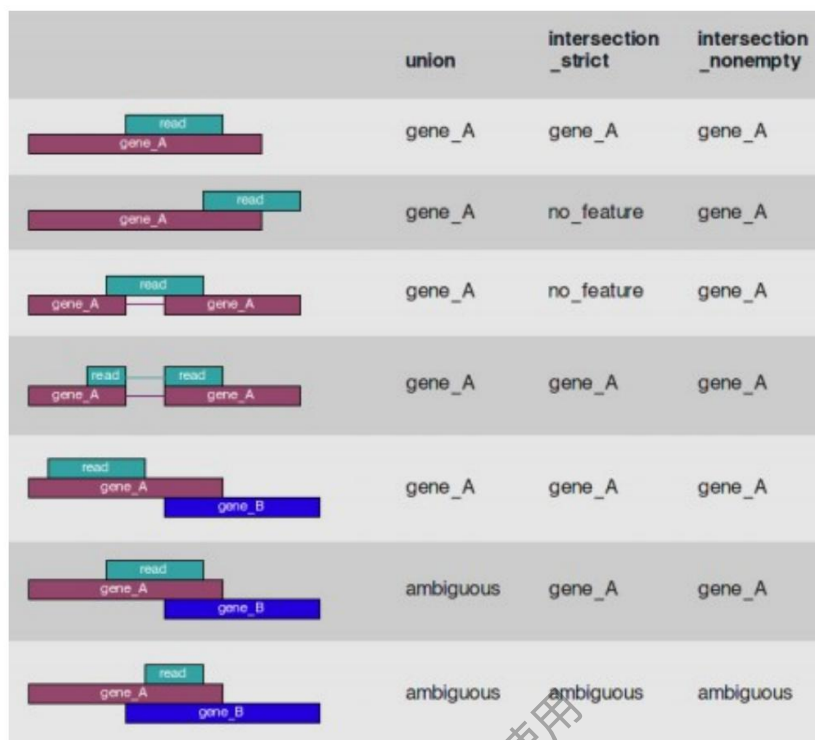
注：横坐标是基因的不同注释区域，纵坐标是比对到不同区域的Tags数目（Tags是将reads与参考基因组或转录组上的特定位置进行关联的序列标记）。

结果文件

4 表达量分析

4.1 表达量分析

使用HTSeq统计比对到每一个基因上Read Count值，作为基因的原始表达量。统计方法如下：首先读取基因结构注释信息（GTF文件），然后将比对结果与基因结构进行比较并统计结果。HTSeq有三种统计方案，如下图，其区别在于当一个Read仅有一部分覆盖在基因区域上或有一部分覆盖在基因的内含子区域时，Union方案和Intersection_nonempty方案认定Read属于该基因，而Intersection_strict方案认定Read不属于任何基因；当一个Read全部覆盖在一个基因上，并且部分覆盖在另一个基因上时，Union方案认定Read同属于两个基因，Intersection_strict方案和Intersection_nonempty方案则认定Read属于第一个基因。如无特殊要求，应按Union方案统计，该方案较为稳健（如果是链特异性建库，则还需要判别是否和注释中的Feature方向一致）。



Reads计数与基因的真实表达水平，以及基因的长度和测序深度成正相关。为了使不同基因、不同样本间的基因表达水平具有可比性，我们采用FPKM（fragments per kilobase of transcript per million fragments mapped）或TPM（transcripts per million）对表达量进行标准化（Normalization），标准分析中默认提供FPKM标准化结果[Zhao et al. 2020]。

Table 8: 表达量分析表

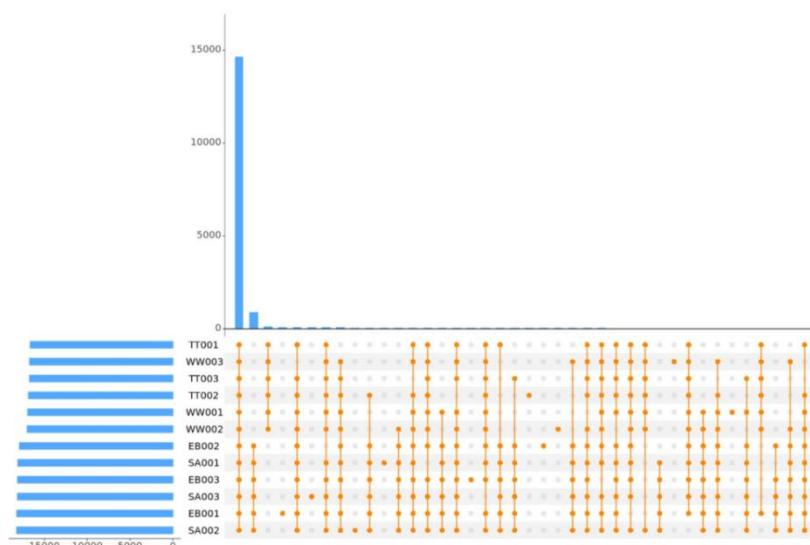
Gene_ID	SA003:read		SA002:read		SA001:read		EBI
	count	SA003:fpkm	count	SA002:fpkm	count	SA001:fpkm	
ENSMUSG00000000001	1205	23.3561125	1281	21.5011606	1190	24.3392834	

注：Gene_ID：基因ID；

Read count：比对到一个基因上的reads数目；

FPKM：使用FPKM标准化后的表达量；

Chromosome列-最后一列：基因的注释信息，包括在染色体上的位置注释、在各数据库的功能注释等。



每个样本鉴定得到的基因Upset图

注：number in each set表示每个样本表达的全部基因的数目；
nnumber of each intersection表示点或连线对应的基因数目；
横坐标点的连线表示对应样本共有表达的基因的数目，单点则表示对应样本中特有表达的基因的数目。

结果文件

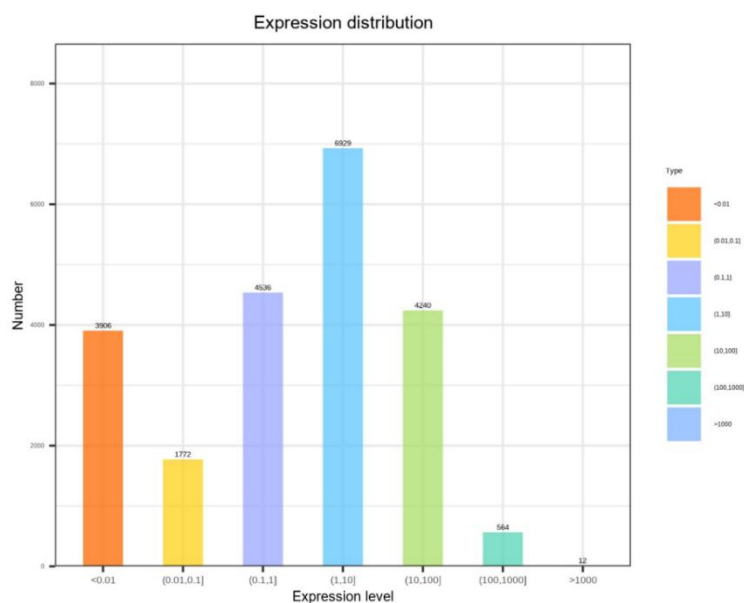
4.1.1 表达量区间统计

根据表达量的计算结果表格，将表达量分为不同的区间，对各样本在不同表达量区间内的基因的数目进行统计。

Table 9. 表达量区间统计表

Type	SA001	SA002	SA003	EB001	EB002	EB003	TT001	TT002	TT003	WW001	WW002	WW003
<0.01	3906	3812	3872	3798	4095	3891	5371	5167	5271	5094	5070	5279
(0.01,0.1]	1772	1963	1801	1805	1798	1896	1590	1641	1495	1630	1608	1483
(0.1,1]	4536	4538	4534	4535	4598	4545	3125	3243	3255	3201	3268	3226
(1,10]	6929	6899	6949	7027	6797	6925	7487	7457	7502	7523	7539	7533
(10,100]	4240	4149	4232	4212	4073	4106	4079	4132	4107	4178	4139	4109
(100,1000]	564	586	558	571	586	585	296	310	321	324	326	318
>1000	12	12	13	11	12	11	11	9	8	9	9	11

注：第一列为表达量值范围，其余为样品名。



SA001



Fig 10: 表达量区间统计图

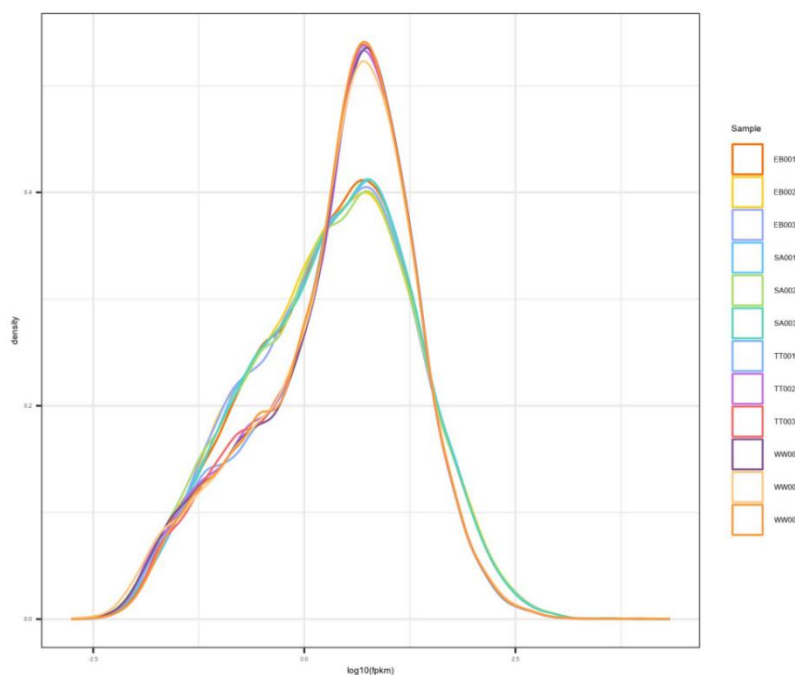
注：横坐标表示表达量值范围，纵坐标表示该表达量区间基因的个数。

结果文件

4.2 表达量分布

4.2.1 表达量密度分布

FPKM密度分布能整体地考察样品所有基因的表达量模式，一般来说中等表达的基因占绝大多数，低表达和高表达的基因占一小部分。



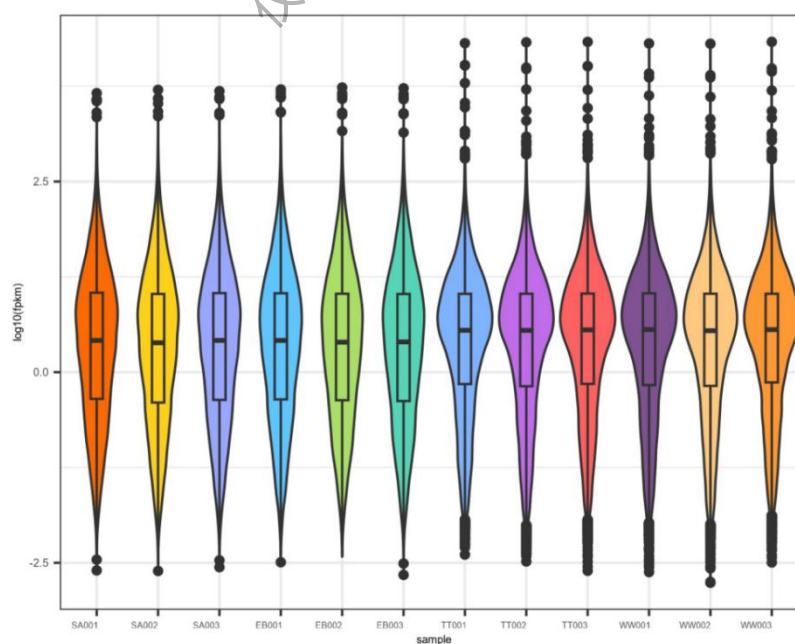
表达量密度分布图

注：横坐标为基因的log10 (FPKM) 值，纵坐标为对应表达量的基因分布密度。

结果文件

4.2.2 表达量小提琴图

FPKM密度分布能整体地考察样品所有基因的表达量模式，一般来说中等表达的基因占绝大多数，低表达和高表达的基因占一小部分。



表达量小提琴图

注：横坐标为不同样本，纵坐标为基因的log10 (FPKM) 值，盒型中间的横线是中位数，盒型的上下边缘为75%，上下限为90%。外部形状为核密度估计。

[结果文件](#)

4.2.3 样本已知基因类型分布

根据表达量的统计结果，对基因组的注释文件中的不同基因类型的reads数量进行统计。

Table 10: 样本已知基因类型分布

Sample	SA003	SA002	SA001	EB003	EB002	EB001	TT001
unitary_pseudogene	198	196	155	145	196	176	56
transcribed_processed_pseudogene	52959	63009	50630	58302	51498	55818	1004
ribozyme	4	1	4	1	0	0	0
Mt_tRNA	252	225	231	256	214	253	288
miRNA	1011	973	901	781	651	864	382
transcribed_unitary_pseudogene	11239	13247	11200	12734	11246	12020	890
IG_V_gene	243	273	192	274	243	275	27
snoRNA	328	337	288	236	214	350	345
TEC	27661	31470	25533	27162	22605	27991	1671
TR_J_pseudogene	0	0	0	0	0	0	0
unprocessed_pseudogene	14690	16382	13610	16085	14352	16208	2053
TR_D_gene	0	0	0	0	0	0	0
rRNA	8721	10686	10112	18127	12281	15257	1722

第一列为不同的基因类型，类型来源于基因组的注释文件，其余列为样品名称。

[结果文件](#)

4.3 相关性分析

样品间基因表达水平相关性是检验实验可靠性和样本选择是否合理的重要指标，在做差异表达分析之前，应先检查样品间基因的表达水平相关性。

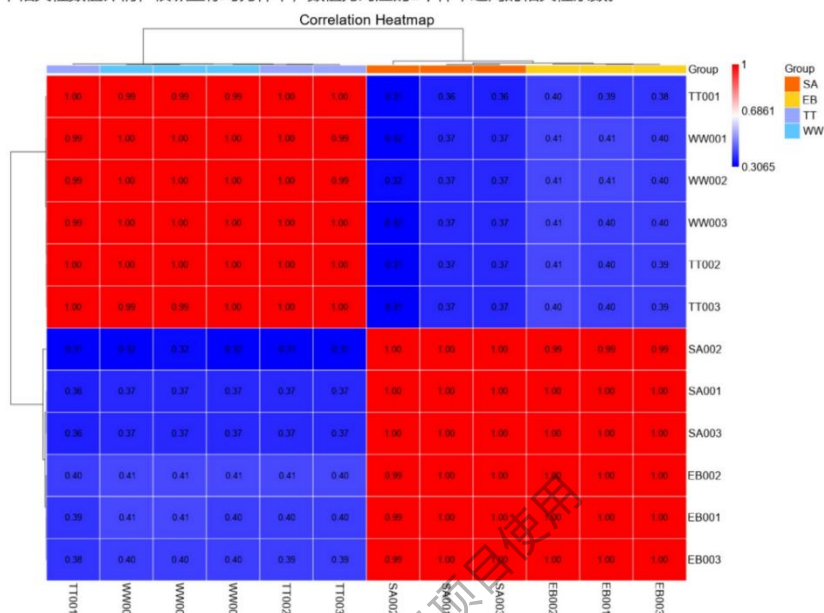
我们默认用皮尔逊相关系数表示样品间基因的表达水平相关性，相关系数越接近1，表明样品之间表达模式的相似度越高。一般来说，相关系数在0.8-1之间属于极强相关。

Table 11: 相关性分析表

sample	SA001	SA002	SA003	EB001	EB002	EB003	TT001	TT002	TT003	WW
SA001	1.0000000	0.9970614	0.9992841	0.9960777	0.9967500	0.9966899	0.3611694	0.3683033	0.3671149	0.373
SA002	0.9970614	1.0000000	0.9973762	0.9930229	0.9931286	0.9943419	0.3063964	0.3140679	0.3128087	0.315
SA003	0.9992841	0.9973762	1.0000000	0.9972762	0.9974339	0.9975002	0.3602794	0.3674556	0.3663305	0.373
EB001	0.9960777	0.9930229	0.9972762	1.0000000	0.9981515	0.9986452	0.3880799	0.3984414	0.3963694	0.406
EB002	0.9967500	0.9931286	0.9974339	0.9981515	1.0000000	0.9990907	0.3968838	0.4052246	0.4036432	0.410
EB003	0.9966899	0.9943419	0.9975002	0.9986452	0.9990907	1.0000000	0.3844206	0.3937345	0.3919114	0.400
TT001	0.3611694	0.3063964	0.3602794	0.3880799	0.3968838	0.3844206	1.0000000	0.9963502	0.9969863	0.987
TT002	0.3683033	0.3140679	0.3674556	0.3984414	0.4052246	0.3937345	0.9963502	1.0000000	0.9995010	0.996
TT003	0.3671149	0.3128087	0.3663305	0.3963694	0.4036432	0.3919114	0.9969863	0.9995010	1.0000000	0.993

sample	SA001	SA002	SA003	EB001	EB002	EB003	TT001	TT002	TT003	WW
WW001	0.3732055	0.3196400	0.3721164	0.4061838	0.4107123	0.4004138	0.9878487	0.9964271	0.9939643	1.000
WW002	0.3745438	0.3210438	0.3734303	0.4077242	0.4118680	0.4017120	0.9862980	0.9956766	0.9933591	0.996
WW003	0.3700681	0.3159317	0.3690813	0.4010569	0.4071832	0.3960731	0.9947482	0.9991195	0.9978985	0.996

注：各样本相关性数值详情，纵横坐标均为样本，数值为对应的2个样本之间的相关性系数。



相关性分析图

注：双向聚类，左侧和上侧聚类树为样本聚类情况，图中右侧和下侧为样本名称，方格上的数值为对应的横向与纵向两个样本的相关性系数，不同颜色的方格代表2个样本的相关性高低情况。

结果文件

4.4 PCA分析

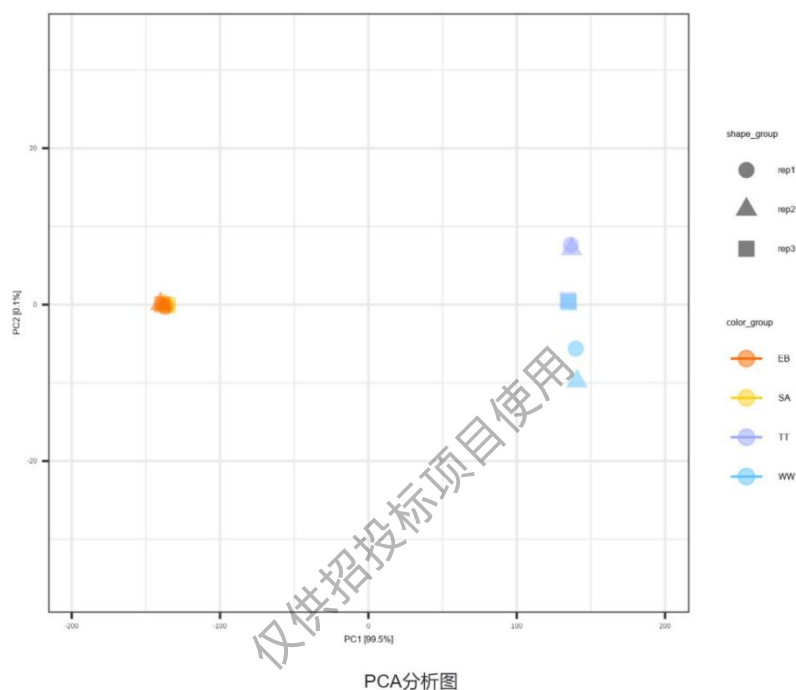
PCA主成分分析（Principal Components Analysis），通过线性变换，将高维数据降低至二维或三维，同时保持各方差贡献最大的特征，即降低数据复杂度。当有多个样品时，根据基因表达值对各样品进行PCA主成分分析。PCA分析可以把相似的样本聚到一起，距离越近表明样本间相似性越高。

Table 12: PCA分析表

ID	Group	PC1	PC2	PC3
SA001	SA	-135.0209	0.0147	0.0095
SA002	SA	-137.2727	-0.0043	-0.0980
SA003	SA	-135.2679	-0.0834	0.0007
EB001	EB	-136.9336	-0.2784	0.0335
EB002	EB	-140.0408	0.0962	0.0940
EB003	EB	-138.8964	0.0450	-0.0828
TT001	TT	136.4884	7.6362	3.7893
TT002	TT	136.9395	7.1296	-3.3227
TT003	TT	134.4801	0.5608	-2.0784

ID	Group	PC1	PC2	PC3
WW001	WW	139.8021	-5.6245	-9.9519
WW002	WW	140.5815	-9.7832	6.2964
WW003	WW	135.1406	0.2913	5.3104

注：ID：样品名；
Group：组名；
PC1：第一主成分；
PC2：第二主成分；
PC3：第三主成分。



注：PCA：x轴为第一主成分，y轴为第二主成分。在图中不同的颜色表示不同的分组；

[结果文件](#)

5 表达差异分析

5.1 差异表达分析

采用Deseq2对基因表达进行差异分析，并提供多种文章常用差异分析软件可供分析选择。默认筛选差异表达基因条件为：表达差异倍数 $|\log_2\text{FoldChange}| > 1$ ，显著性 $P\text{-value} < 0.05$ 。

5.2 差异表达结果统计

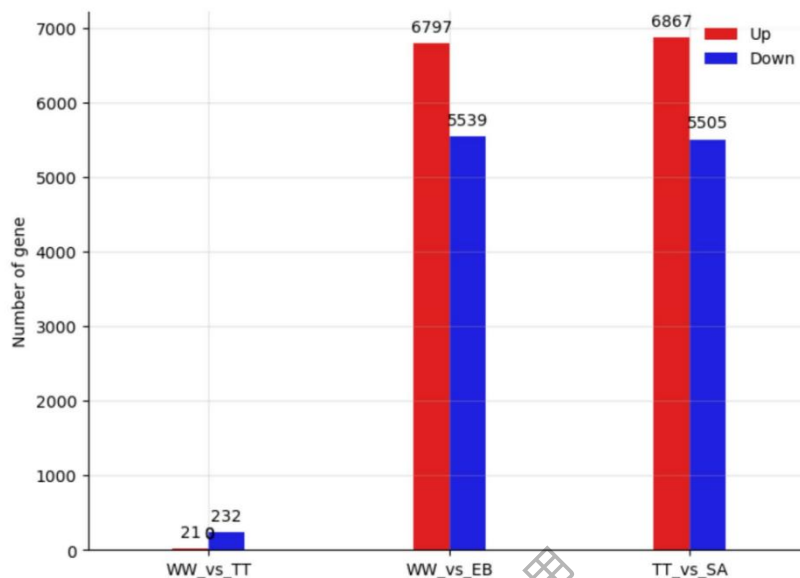
对差异表达分析结果中显著差异基因集进行统计，并对不同比较组之间的差异基因做柱状图，统计出每个比较组的上调差异基因和下调差异基因数量。

Table 13: 差异表达分析结果

Control_vs_Treat	Up Regulation	Down Regulation	Total
WW_vs_TT	21	232	253
WW_vs_EB	6797	5539	12336

Control_vs_Treat	Up Regulation	Down Regulation	Total
TT_vs_SA	6867	5505	12372

注：各样本相关性数值详情，横纵坐标均为样本，数值为对应的2个样本之间的相关性系数。



差异表达分析结果统计

注：横坐标表示各个比较组；

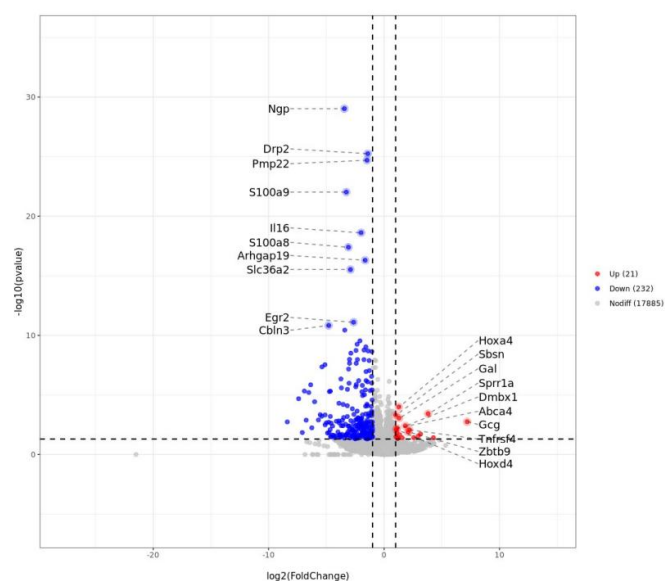
纵坐标表示比较组的差异基因数量；

up表示显著上调的基因数量，down表示显著下调的基因数量。

结果文件

5.3 火山图

基于差异表达分析的结果绘制火山图与MA图。火山图是基于表达倍数差异和显著性结果，展示基因的分布情况。左侧为Case相比于Control下调基因，右侧为Case相比于Control上调基因，可根据需要在火山图上对部分基因进行标注。MA图则是基于基因的表达量均值和表达差异倍数，展示基因分布情况。使用颜色对上调基因、下调基因及没有显著差异表达的基因进行标注。



WW_vs_TT

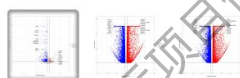


Fig 17: 火山图

注：横坐标为 $\log_2(\text{FoldChange})$ ，纵坐标为 $-\log_{10}(\text{pvalue})$ 。图中两条竖虚线为表达差异倍数的阈值；
横虚线为显著性水平阈值。颜色表示基因是up（上调）、down（下调）或none（非显著差异表达），默认标注上下调top10基因（显著性排序）。

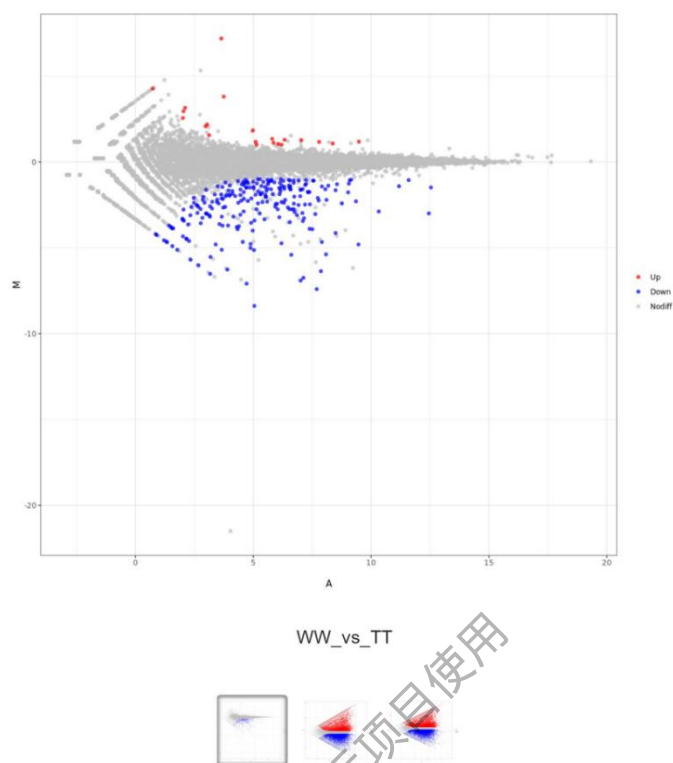


Fig 18: MA图:

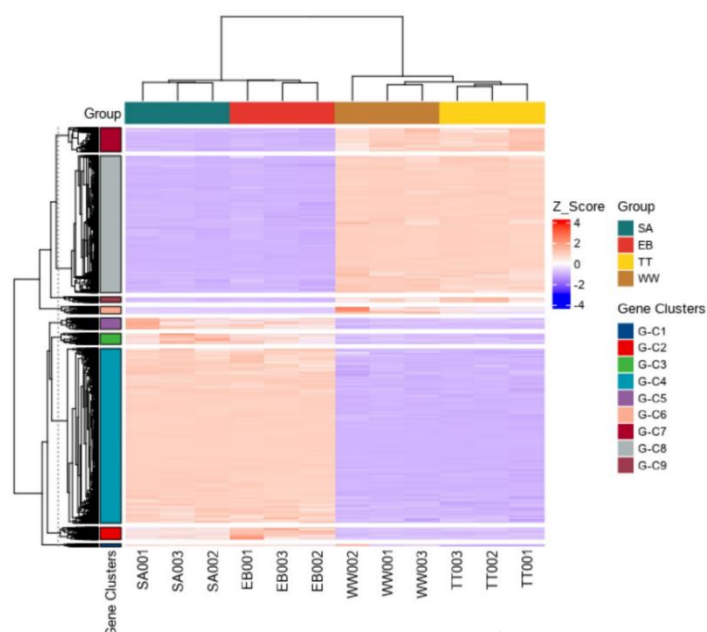
注：横坐标为两样品基因表达量均值，纵坐标为 \log_2 Fold Change。颜色表示基因是上调、下调或非显著差异表达。

结果文件

5.4 聚类分析

聚类分析用于判断差异表达基因在不同实验条件下的表达模式；在样品间表达量相关性高的基因被归为一类，通常这些基因在某些生物学过程，或者某个代谢、信号通路中存在实际的联系。因此通过表达量聚类我们可以发现基因间未知的生物学联系。

我们使用R语言Pheatmap和ComplexHeatmap软件包对所有比较组的差异基因的并集和样品进行双向聚类分析，根据同一基因在不同样品中的表达水平和同一样品中不同基因的表达模式进行聚类，默认采用Euclidean方法计算距离，层次聚类最长距离法（Complete Linkage）进行聚类。



差异表达基因聚类热图

注：热图主体区域：横向表示基因，每一列为一个样品；
颜色越红，代表基因在该样本中表达量越高，反之颜色越蓝，代表基因在该样本中表达量越低；
样本聚类树：样本聚类情况，表达模式相近的样本聚到一起；
基因聚类树：基因聚类情况，表达模式相近的基因聚到一起；
Group：样本分组信息，不同颜色代表不同组别；
Gene Clusters：基因聚类分类标签，表达模式相近的基因聚类为一个cluster；
基因标签：按照所选标签内容，对部分基因进行标记；
Sample Clusters：样本聚类分块，聚类到一起的样本使用颜色进行划分；。

结果文件

5.5 趋势分析

趋势分析，是基于双向聚类热图的分析结果，进一步根据基因表达模式的相似性将其划分成不同的cluster。适用于2个以上的样本，分析各样品间mRNA表达丰度的不同变化模式，将相同表达趋势的mRNA划分为一簇，并对簇基因作表达模式图，直观地展示不同类型基因在样品间的表达丰度变化情况。因此可以用于缩小分析范围，聚焦关键基因。

Table 14: 趋势分析表

GeneID	SA001	SA002	SA003	EB001	EB002	EB003	TT001
ENSMUSG000000040694	0.4354467	0.7181234	0.5173338	0.5133849	0.4685804	-0.6624216	-0.4329502
ENSMUSG000000005917	-0.0957510	0.3295880	0.0277376	1.0885695	0.9779122	-0.3704772	-0.8778990
ENSMUSG000000044086	0.8855914	0.5731980	-0.3626201	0.3784233	1.2640060	0.2094882	-0.9924187
ENSMUSG000000019803	0.4220059	0.2105849	1.7006053	0.0444730	-0.7907842	0.0130287	-0.5478298
ENSMUSG000000035296	-0.7264453	0.2160911	1.1045370	0.0628571	-0.1360094	0.3409601	-1.1062860

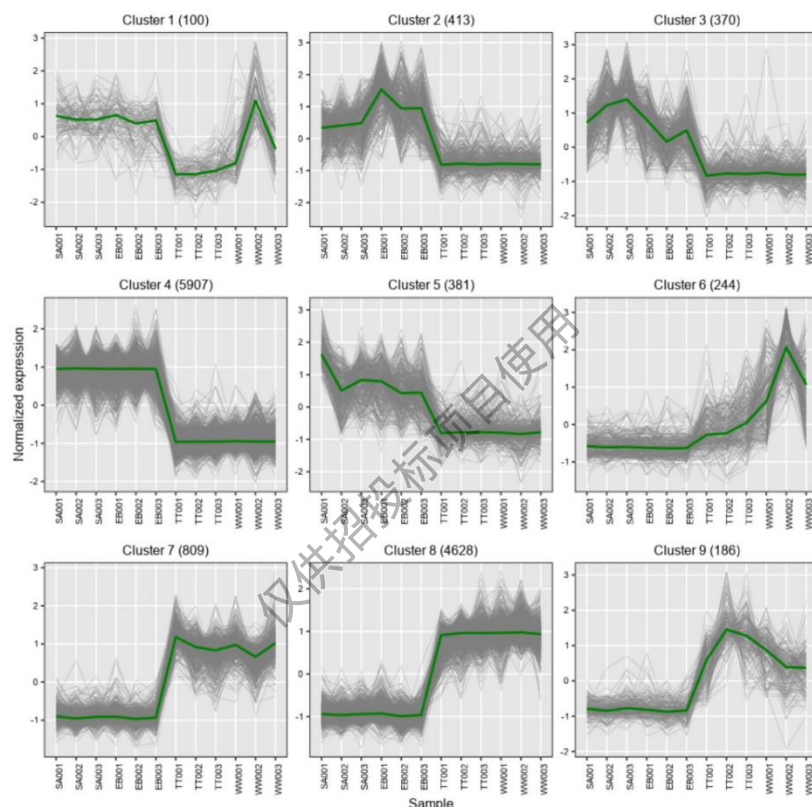
趋势分析表

GeneID	SA001	SA002	SA003	EB001	EB002	EB003	TT001
ENSMUSG00000023903	-0.4120653	0.2771876	0.4865640	-0.4299575	-0.0907921	0.2678836	-0.7496176
ENSMUSG00000007122	0.1414181	0.2429858	0.0784915	0.2314960	-0.4969801	-0.1993007	-0.8675474
ENSMUSG00000024049	0.2874066	0.2182545	0.2606156	0.2067502	-0.5461240	-0.4695038	-0.6797710
ENSMUSG00000061462	0.2572570	0.3619045	0.2122194	0.1127289	-0.0016604	0.0226085	-1.1500066
ENSMUSG00000026208	0.3681754	0.0129236	0.0209714	0.4335779	0.0770872	0.1218350	-1.2217406

注: cluster: 按照表达模式进行分簇后的cluster编号;

ID: Gene ID;

样品名: 基因在该样本中的表达量进行Z-score均一化之后的值, 是作图的直接数值。



趋势分析图

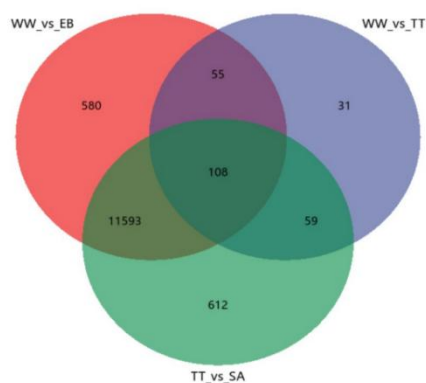
注: 图中背景线展示每个Cluster中基因的表达模式, 中间线表示Cluster中的所有基因在样品中表达量的平均值。

结果文件

5.6 差异基因韦恩图

根据差异分析的结果, 统计各比较组之间的共有特有差异基因数量

venn图 (仅提供2-6个比较的venn图) 通过比较组的重叠关系, 展示了各比较组间重叠的差异基因的个数



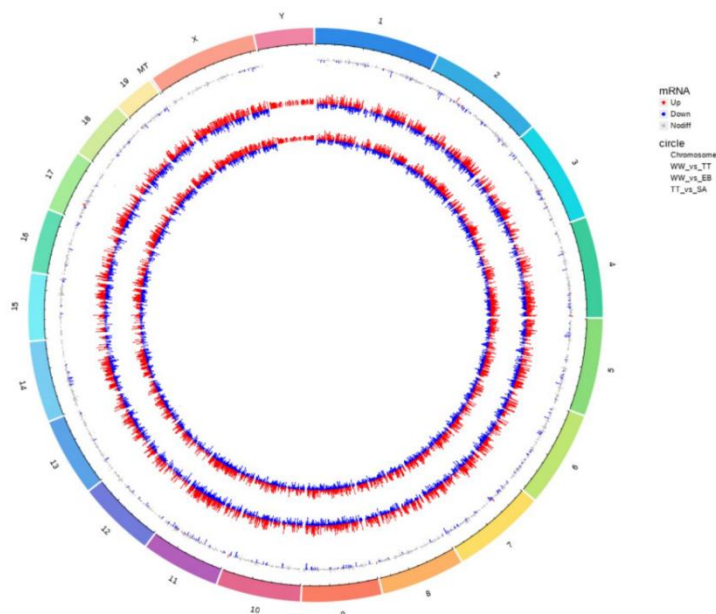
venn图

注：不同的颜色表示不同的比较组。各个圈中的数字之和代表该比较组合的差异基因总个数，圆圈重合的部分表示比较组之间共有的差异基因数。

结果文件

5.7 基因组圈图

使用R语言CircIze包，根据基因组信息和基因差异表达分析结果，在基因组上标记差异表达的基因，来绘制基因组圈图，它可以反应不同比较组之间的差异基因在染色体上的分布情况。



差异表达基因组圈图

注：最外圈是染色体条带，从外向内为不同比较组的差异表达分析结果。红色和蓝色分别为上调和下调基因的log2FoldChange值的柱状图，灰色为无差异表达基因的log2FoldChange值的散点图。

结果文件

5.8 蛋白网络互作分析

蛋白互作网络 (protein protein interaction network, PPI network) 是揭示基因之间互作关系的分析。分析使用 STRING (<https://string-db.org>) 数据库进行互作关系的预测。STRING数据库 (Search3 Tool for the Retrieval of Interacting Genes/Proteins) 是EMBL开发的蛋白质互作数据库，该数据库从最有力的实验证据到数据挖掘、同源预测的蛋白质互作关系都有收录。

PPI分析可以对目标基因集进行互作关系的探索，从基因集中筛选关键基因，进一步缩小目标的范围，是数据挖掘的重要组成。我们依据STRING数据库进行蛋白互作分析，以此揭示目的基因之间的作用关系。当STRING数据库中收录该物种的PPI信息时，我们根据基因差异表达分析结果，直接数据库中筛选含有差异基因并且Score>0.95的PPI作用对。当网络过大或过小时，可调节Score值。

当STRING数据库中没有该物种的PPI信息时，我们选择相近物种与该物种的蛋白质序列进行比对，进而得到该物种的蛋白质之间的相互关系。然后再筛选含有差异基因并且Score>0.95的PPI作用对。最终获得所有目的基因之间互关系 (*PPI.network.txt)，同时可以使用结果文件通过Cytoscape进行做图。

Table 15: 蛋白网络互作分析表

Node1	Node2	Score
Top2a	Mki67	0.985
Drp2	Prx	0.982
Mki67	Bub1b	0.951
Cdh1	S100a8	0.986

Node1	Node2	Score
Neb	Ttn	0.999
Hbb-bs	Hba-a1	0.991
Csrp3	Ttn	0.970
Trpa1	Trpm8	0.958
Trdn	Hrc	0.997
Cacna1s	Ryr1	0.998
Hba-a1	Hbb-bt	0.989
Casq1	Ryr1	0.957
Top2a	Bub1b	0.974
Gata3	Foxa1	0.970
S100a8	S100a9	0.999
Slc4a1	Gypa	0.996
Trdn	Ryr1	0.997
Casq1	Trdn	0.995
Ttn	Obscn	0.976

注：节点属性表 (attribute) Gene_ID：网络中节点名称；

Regulation：该基因在该比较组中上下调类型；

Type：基因类型；

Degree：节点大小。互作关系表 (network) Node：网络中互作的节点名称；

Score：综合计算后蛋白互作可信度。

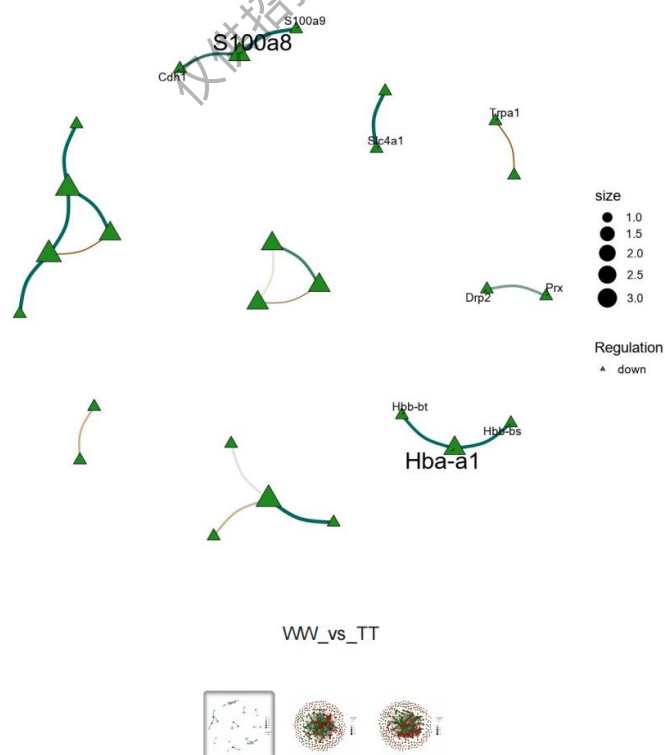


Fig 23: 蛋白互作网络图

注：点为基因（对应的蛋白质）。连线代表具有互作关系；
点的大小表示互作的节点多少。其中up表示上调基因，down表示下调基因。

[结果文件](#)

5.9 外显子差异分析

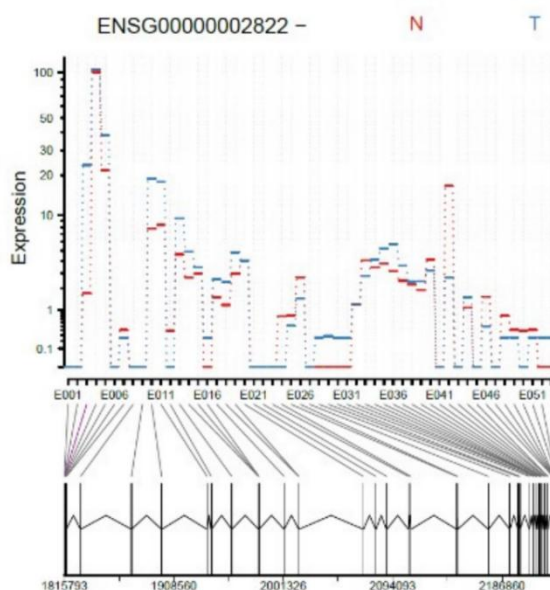
使用DEXSeq包来分析RNA-seq实验数据中外显子使用（exon usage）差异，这里外显子使用差异指的是由于实验条件导致的外显子使用上的差异（仅能对生物学重复 >2 的组别进行分析）。DEXSeq内部调用DESeq2软件进行分析，其分析原理与DESeq2软件相同。

DEU（differential exon usage）基本思想是，针对每个样本的外显子（或部分外显子），我们统计比对到该外显子的read数和比对到同一基因（包含多个转录本）其他外显子的read数，然后计算这两个统计数的比值，最后根据不同实验条件下的比值变化情况，推导出相对的exon usage改变。对于基因内的一个外显子，该exon usage同步于该exon被剪接到转录组（可变剪接）的比例，它也包含了发生在转录本5'端和3'端的可导致转录本边界的差异性的exon usage的可变剪接。因此，DEU（By differential exon usage）相比于可变剪接更直观。

Table 16: DEXSeq分析结果示例

groupID	featureID	exonBaseMean	dispersion	stat	pvalue	padj	ms_Mat	ms_Mei	log2fold_ms
G0001	E001	15.066965	0.0121391	0.3387443	0.5605550	0.9518734	6.967857	6.407635	-0.24
G0002	E002	18.604616	0.0044836	0.0423581	0.8369386	1.0000000	7.297135	7.398996	0.04
G0003	E003	14.543196	0.0094122	0.0125428	0.9108276	1.0000000	6.602671	6.472075	-0.05
G0004	E004	7.862697	0.0062393	9.7085328	0.0018341	0.0538519	2.697596	5.442162	2.02
G0005	E005	10.430123	0.0074573	0.0171581	0.8957840	1.0000000	5.611870	5.443038	-0.08

注：groupID: 基因ID；
featureID: 外显子位数；
exonBaseMean: 平均表达量；
dispersion: 离散；
stat: 统计量；
pvalue: 显著性p值；
padj: BH校正p值；
log2fold: 组间差异倍数取log值；
genomicData.seqnames: 在基因组上的位置。



外显子使用差异示意图

注：横坐标为外显子位置信息；
纵坐标表示两组样本各自在外显子上的表达值；
图的右上方展示两两比较的两组的颜色；
横坐标位置区域的紫红色表示该exon在两组中的exon usage有显著差异。

结果文件

6 富集分析

6.1 GO富集分析

GO是基因本体论联合会建立的数据库 (Gene Ontology, <http://geneontology.org/>)。GO的产生主要是为了解决同一基因在不同数据库定义的混乱性以及不同物种的同一基因在功能定义上的混乱性。它是一个国际化的基因功能分类体系，提供了一套动态更新的标准词汇表 (Controlled Vocabulary) 来全面描述生物体中基因和基因产物的属性。GO涵盖三个方面，分别描述基因的分子功能 (Molecular Function)、细胞的组件作用 (Cellular Component)、参与的生物学过程 (Biological Process)。基因或蛋白质可以通过ID对应或者序列注释的方法找到与之对应的GO编号，而GO编号可用于对应到GO Term，即功能类别或者细胞定位。

GO的基本单元是Term，每个Term有一个唯一的标示符 (由“GO:”加上7个数字组成，例如GO: 0072669)；每类Ontology的Term通过它们之间的联系 (is_a, part_of, regulate) 构成一个有向无环的拓扑结构。GOSlim是缩减版的GO术语，它提供了GO注释的概述性结果。

我们使用clusterProfiler进行GO富集分析，分析时利用GO term注释的差异基因对每个term的基因列表和基因数目进行计算，然后通过超几何分布方法计算P-value (显著富集的标准为P-value<0.05)，找出与整个基因组背景相比，差异基因显著富集的GO term，从而确定差异基因行使的主要生物学功能。

Table 17: Go富集分析表

GO_ID	Category	Term	Up	Down	DEG	Total
GO:0030017	CC	sarcomere	0	22	22	218

Go富集分析表

GO_ID	Category	Term	Up	Down	DEG	Total
GO:0030016	CC	myofibril	0	22	22	240
GO:0043292	CC	contractile fiber	0	22	22	250
GO:0003012	BP	muscle system process	1	26	27	424
GO:0006936	BP	muscle contraction	1	21	22	312
GO:0032501	BP	multicellular organismal process	14	138	152	8630
GO:0055001	BP	muscle cell development	0	18	18	207
GO:0009653	BP	anatomical structure morphogenesis	5	66	71	2761
GO:0009605	BP	response to external stimulus	3	69	72	2871
GO:0044057	BP	regulation of system process	2	28	30	642

注: GO.ID: GO Term编号;

Category: GO Term所处的分类;

Term: GO条目名称;

Up/Down: 富集到该GO条目的上下调基因数量;

DEG: 富集到该GO条目的差异基因的总>数;

Total: 注释到该GO条目的总基因数;

Pvalue: 富集显著性P值;

adjustPvalue: P值校正后;

Up_gene/Down_gene: 富集到该GO条目的上下调基因列表GeneRatio: 注释到GO term上的差异基因数与差异基因总数的比值;

BgRatio: 注释到GO term上的背景基因数与背景基因总数的比值;

RichFactor: 富集因子, 富集在GO term上的差异基因的数目与注释到GO term上的背景基因数的比值。

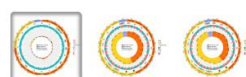
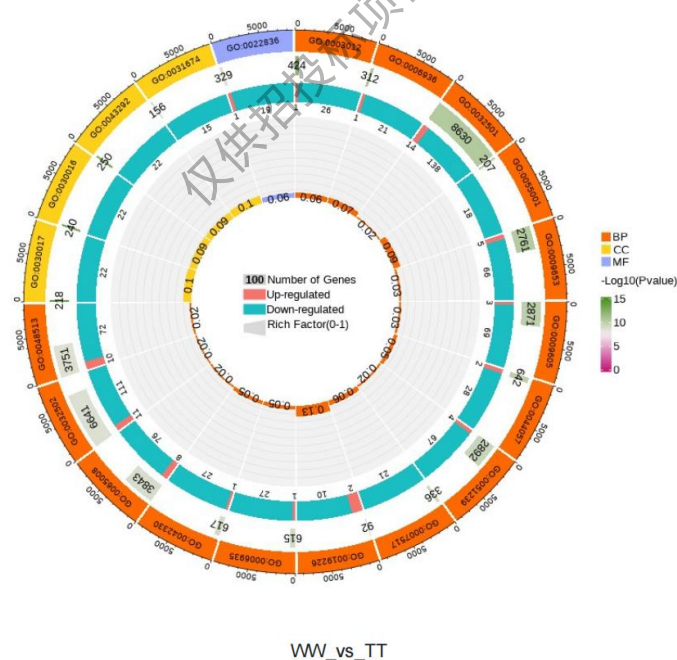


Fig 25: Go富集分析圈图

注: 从外到内共四圈。第一圈: Term的分类, 圈外为基因数目的坐标尺。不同的颜色代表不同的分类;第二圈: 背景基因中该term注释

的基因数目以及富集的显著性p值，条形长短代表基因数目，颜色代表富集的显著性；

第三圈：上下调基因比例条形图，红色代表上调基因数目，绿色代表下调基因数目；

下方显示具体的数值；

当输入的差异基因数量只有一列（未区分上下调）时，第三圈显示差异基因的总数目；

第四圈：各分类的Rich Factor值（富集到该GO Term的差异基因个数/注释到该GO Term的总基因数），背景辅助线每个小格表示0.1；

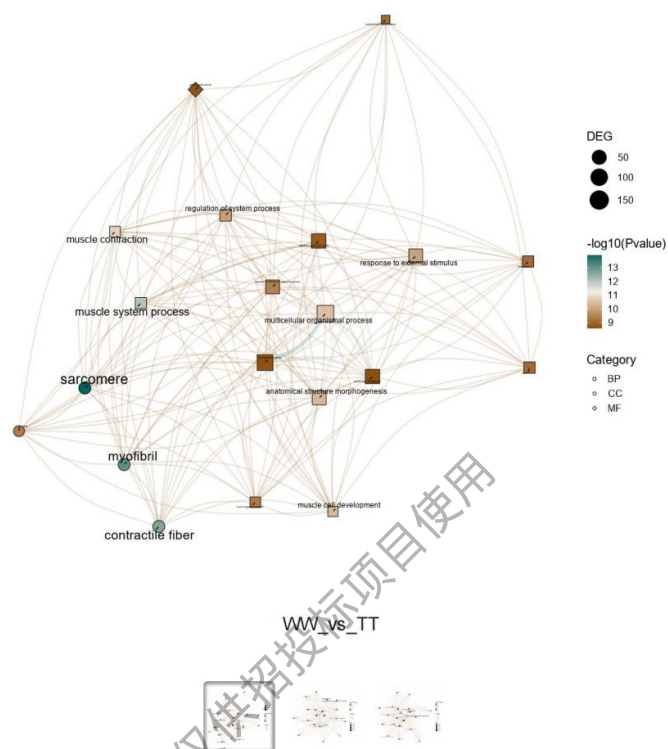
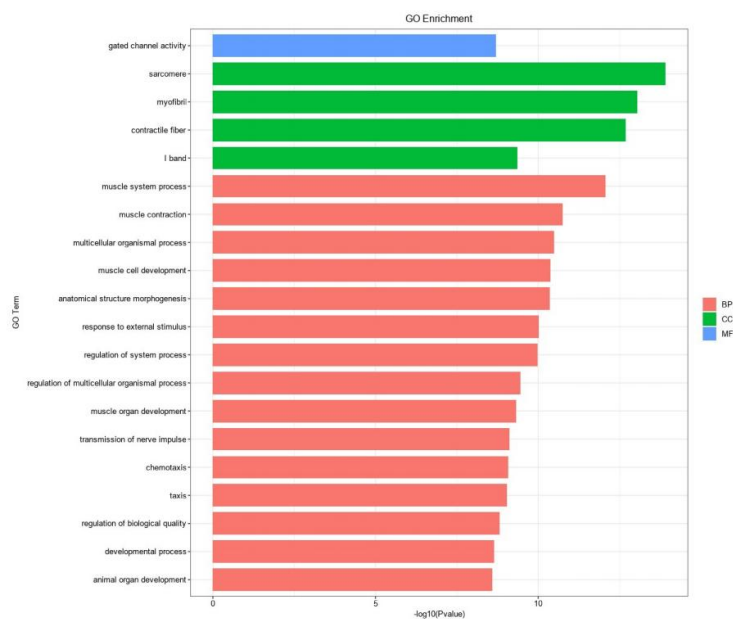


Fig 26: Go富集分析网络图

注：网络图圆圈表示GO Term名称，圆圈大小表示该Term上富集的差异基因数量多少，颜色的深浅表示显著性水平高低，连线粗细表示两Term中共有的差异基因数量，线条越粗表示共有差异基因越多。



WW_vs_TT



Fig 27: Go富集分析柱形图

注：柱状图：纵坐标为GO Term，横坐标默认为GO Term富集的 $-\log_{10}(p\text{-value})$ ，可选择p-adjust或差异基因（DEG）数目进行展示；

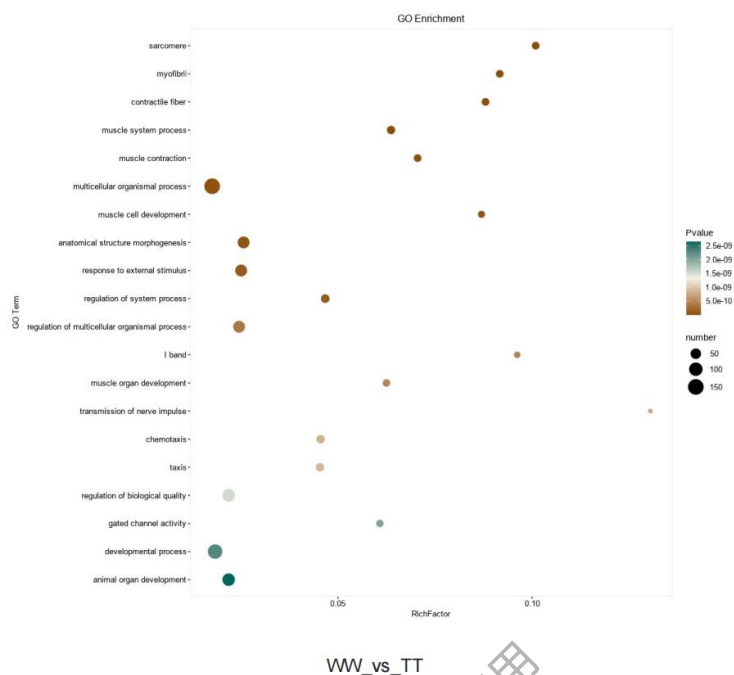
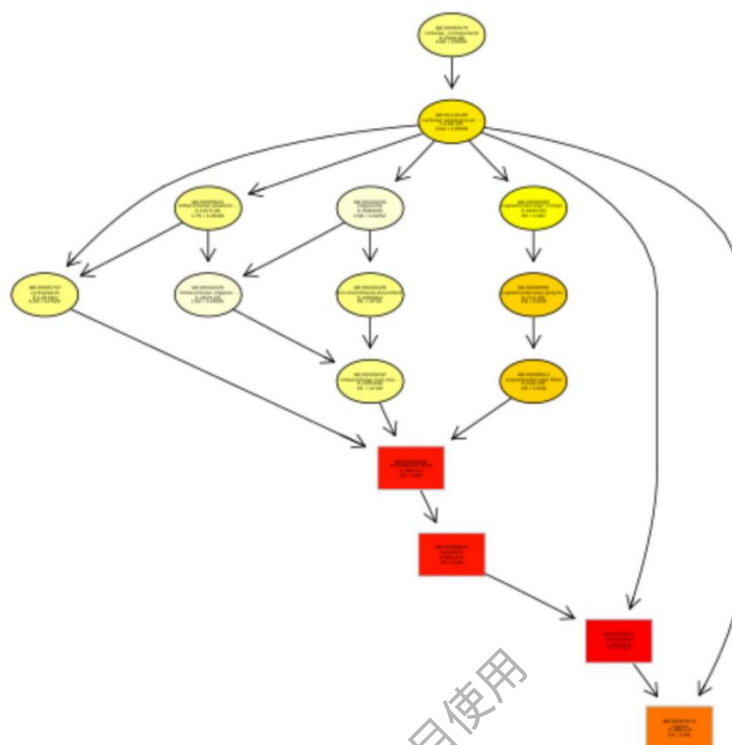


Fig 28: Go富集分析因子图

注：因子图：横坐标为富集因子（rich factor）（富集到该GO Term的差异基因个数/注释到该GO Term的总基因数），纵坐标为GO Term，图中点的大小表示相应Term中富集到的差异（上调或者下调，与分析时选择的基因集有关）基因数目，颜色的深浅表示显著性水平高低；



Go DAG图

结果文件

6.2 KEGG富集分析

KEGG, 京都基因与基因组百科全书 (Kyoto Encyclopedia of Genes and Genomes, (<http://www.kegg.jp/>)) 是一个整合了基因组、化学和系统功能信息的数据库。把从已经完整测序的基因组中得到的基因目录与更高级别的细胞、物种和生态系统水平的系统功能关联起来是KEGG数据库的特色之一。KEGG注释主要包括: (1) KO (KEGG Ortholog) 注释, 即将分子网络的相关信息跨物种注释; (2) KEGG Pathway注释, 即代谢通路注释, 获得物种内分子间相互作用和反应的网络。

我们使用ClusterProfiler进行KEGG富集分析, 分析时利用KEGG pathway注释的差异基因对每个pathway的基因列表和基因数目进行计算, 然后通过超几何分布方法计算P-value (显著富集的标准为P-value < 0.05), 找出与整个基因组背景相比, 差异基因显著富集的KEGG pathway, 从而确定差异基因行使的主要生物学功能。

Table 18: KEGG富集分析表

PathwayID	Pathway	level1	level2	Up	Down	DEG
mmu05144	Malaria	Human Diseases	Infectious disease: parasitic	1	8	9
mmu04080	Neuroactive ligand-receptor interaction	Environmental Information Processing	Signaling molecules and interaction	4	12	16
mmu04020	Calcium signaling pathway	Environmental Information Processing	Signal transduction	1	9	10

KEGG富集分析表

PathwayID	Pathway	level1	level2	Up	Down	DEG
mmu00910	Nitrogen metabolism	Metabolism	Energy metabolism	0	3	3
mmu04260	Cardiac muscle contraction	Organismal Systems	Circulatory system	0	5	5
mmu05202	Transcriptional misregulation in cancer	Human Diseases	Cancer: overview	0	8	8
mmu04657	IL-17 signaling pathway	Organismal Systems	Immune system	1	4	5
mmu04922	Glucagon signaling pathway	Organismal Systems	Endocrine system	1	4	5

注：PathwayID：KEGG Pathway编号；

Pathway：Pathway名称；

level1/level2：Pathway在level1或level2水平的分类；

Up/Down：富集到该pathway的上/下调基因数量；

DEG：富集到该Pathway的差异基因的总数；

Total：注释到该Pathway的总基因数；

Pvalue：富集显著性P值；

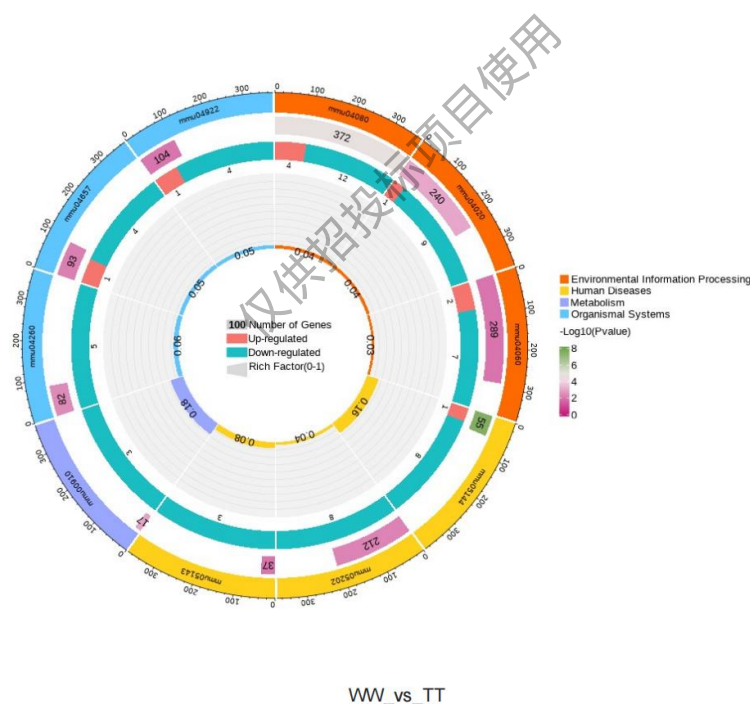
adjustPvalue：P值校正值；

Up_gene/Down_gene：富集到该pathway的上/下调基因数量；

pathway的上/下调基因列表；

GeneRatio：注释到pathway上的差异基因数与差异基因总数的比值；

BgRatio：注释到pathway上的背景基因数与背景基因总数的比值RichFactor：富集因子，富集在pathway上的差异基因的数目与注释到pathway上的背景基因数的比值。



WW_vs_TT



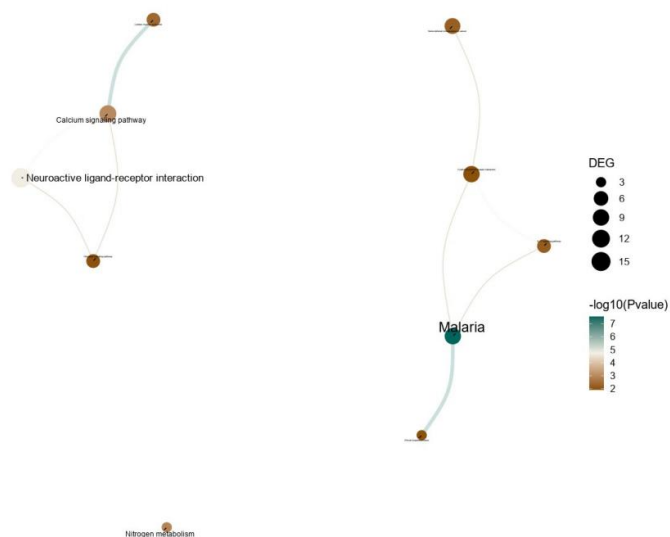
Fig 30: KEGG富集分析圈图

注：从外到内共四圈。第一圈：pathway所属的level1，圈外为基因数目的坐标尺。不同的颜色代表不同的level1；

第二圈：背景基因中该pathway注释的基因数目以及富集的显著性p值。基因越多条形越长，值越小颜色越红；

第三圈：上下调基因比例条形图，红色代表上调基因数目，蓝色代表下调基因数目；下方显示具体的数值；当输入的差异基因数量只有一列（未区分上下调）时，第三圈显示差异基因的总数目；

第四圈：各分类的Rich Factor值（富集到该pathway的差异基因个数/注释到该pathway的总基因数），背景辅助线每个小格表示0.1；



WW_vs_TT



Fig.31: KEGG富集分析网络图

注：圆圈表示pathway名称，圆圈大小表示该pathway上富集的差异基因数量多少，颜色的深浅表示显著性水平高低，连线粗细表示两pathway中共有的差异基因数量，线条越粗表示共有差异基因越多。

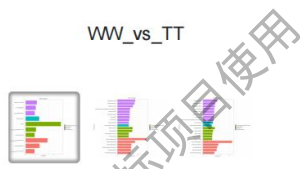
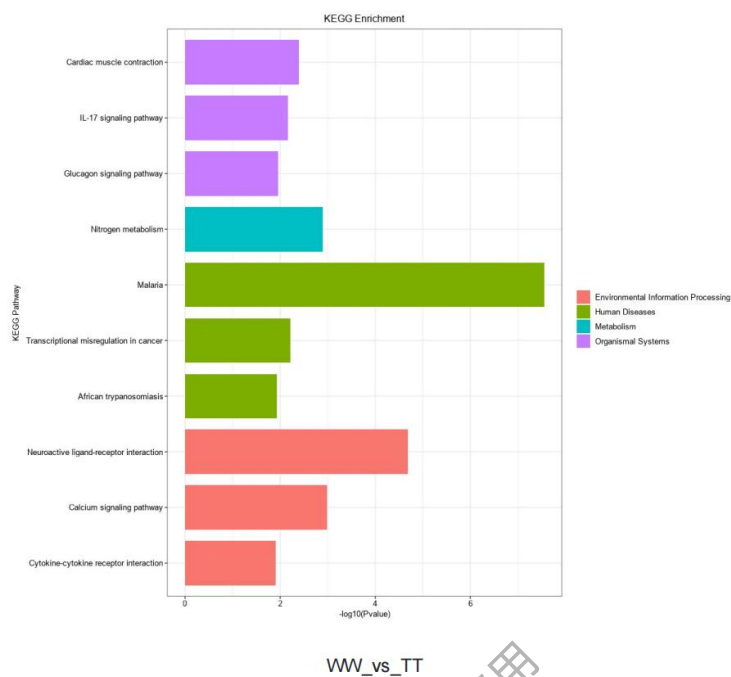


Fig 32: KEGG富集分析柱形图

注：纵坐标为KEGG pathway，横坐标默认为KEGG富集的 $-\log_{10}(\text{p-value})$ ，可选择p-adjust或差异基因（DEG）数目进行展示；

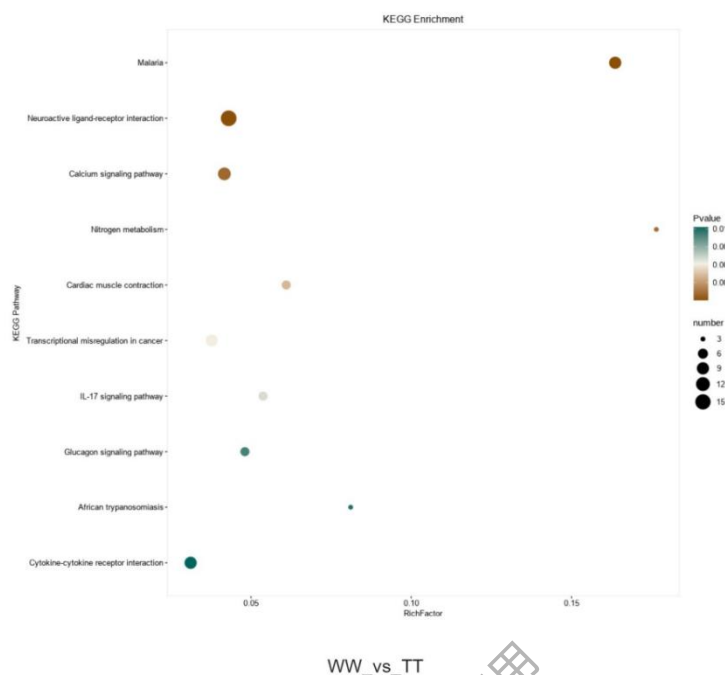


Fig 33: KEGG富集分析因子图

注：横坐标为富集因子（rich factor）（富集到该pathway的差异基因个数/注释到该pathway的总基因数），纵坐标为KEGG pathway，图中点的大小表示相应pathway中富集到的差异（上调或者下调，与分析时选择的基因集有关）基因数目，颜色的深浅表示显著性水平高低；

结果文件

7 GSEA分析

7.1 GO GSEA分析

常规的基于超几何分布的富集分析依赖于显著上调或下调的基因，容易遗漏部分差异表达不显著但有重要生物学意义的基因。基因集富集分析（GSEA）不需要指定明确的差异基因阈值，把所有基因按照在两组样本中的差异表达程度进行排序，然后采用统计学方法检验预先设定的基因集合是否在排序表的顶端或低段富集。GSEA主要包括三个步骤：计算富集得分（Enrichment Score）；估计富集得分的显著性水平；多重假设检验。

Table 19: GO GSEA分析表

ID	Description	SIZE	ES	NES	NOM p val
MYOFIBRIL(GO:0030016)	MYOFIBRIL(GO:0030016)	233	-0.7770025	-2.158914	0.00e+00

GO GSEA分析表

ID	Description	SIZE	ES	NES	NOM p-val

GS: 基因集的名称;
 SIZE: 基因集下包含的基因数目 (经过条件筛选后的值);
 ES: 富集得分 (Enrichment Score) NES: ES的标准化值 (Normalized Enrichment Score), 同时考虑基因集的个数及基因数目, NES的值代表该基因集中的基因在整体基因排序列表中的富集程度, 简单理解NES为正值基因集上调、负值基因集下调;
 NOM p-val: P-value, 针对ES的排列检验, 表示基因集富集的显著性;
 FDR q-val: FDR法校正的p值;
 FWER p-val: 用FWER法 (Bonferonni校正) 校正后的P值;
 RANK AT MAX: 当ES值达到最大时对应的那个基因在排序好的基因列表中所处的位置;
 LEADING EDGE: 核心基因集, 对ES贡献最大的基因成员。该处有3个统计值, tags表示核心基因集占该基因集中基因总数的百分比, list表示核心基因占所有基因的百分比, signal是将前两项统计数据合在一起计算出的富集信号强度;
 一般认为 $|NES| > 1$, $NOM\ p\text{-val} < 0.05$, $FDR\ q\text{-val} < 0.25$ 的通路是显著富集的, 点击detail查看对应基因集富集的详细结果。

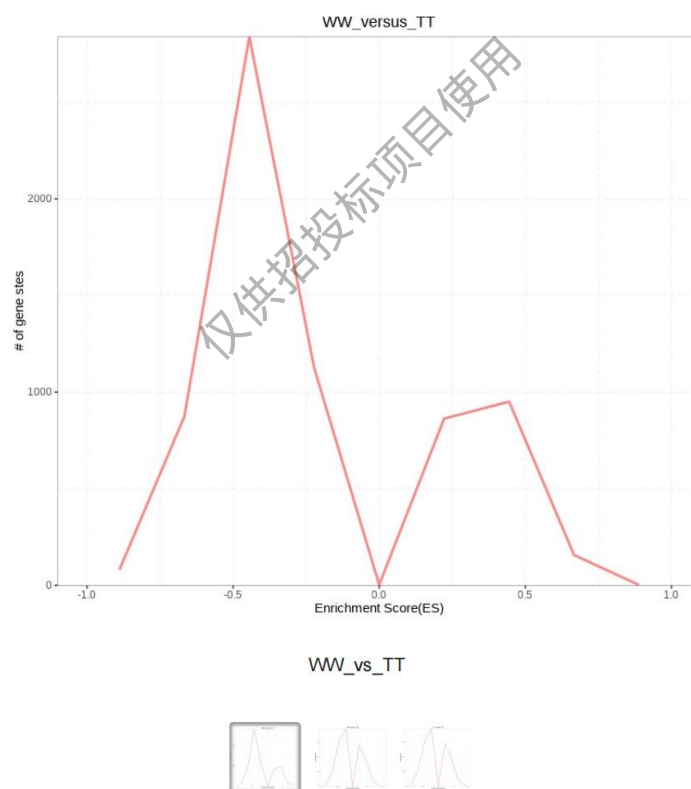
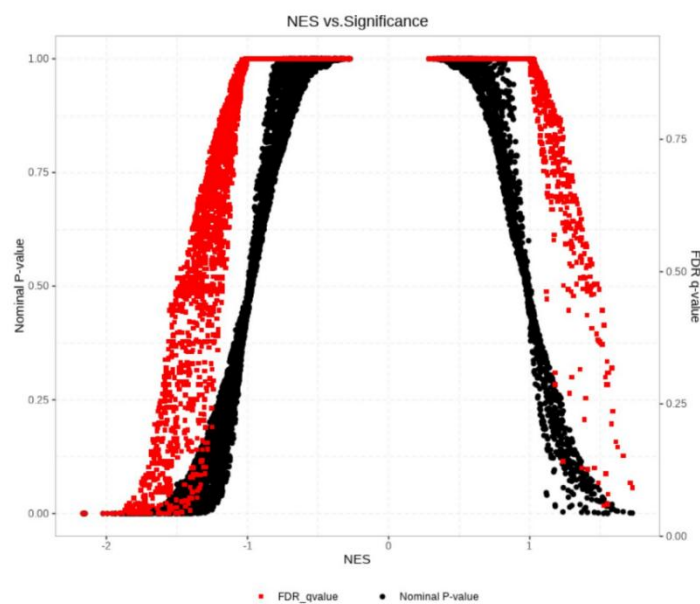


Fig 34: ES图:

注: 基因集的富集分数 (ES) 统计图, 横坐标为ES值, 纵坐标为基因集的数量。



WW_vs_TT

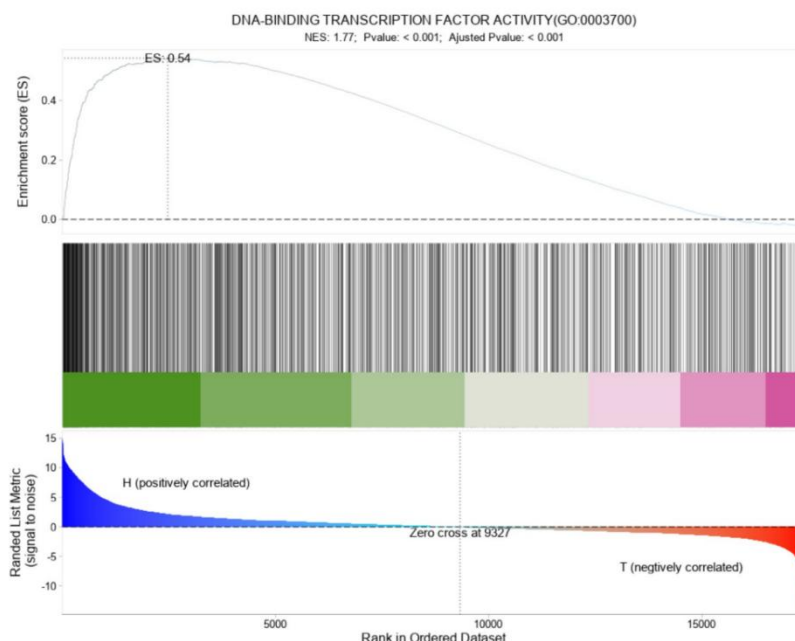


Fig 35: NES图

注：p值与标准化富集分数（NES）的对比图，横坐标为NES值，黑点对应p值结果，红点对应FDR结果；

结果文件

7.1.1 GO-GSEA富集分析图



GO-GSEA富集分析图

注：第一部分：每个位置对应的ES值的分布曲线，最高峰处的得分（垂直距离0.0最远）便是基因集的ES值；
第二部分：用线条标记了对应基因集中基因出现在排序列表中的位置，每条竖线代表一个基因；
第三部分：是所有基因排序后分布情况，其中红色部分对应的基因在处理组中高表达，蓝色部分对应的基因在对照组中高表达。

7.2 KEGG GSEA分析

常规的基于超几何分布的富集分析依赖于显著上调或下调的基因，容易遗漏部分差异表达不显著但有重要生物学意义的基因。基因集富集分析（GSEA）不需要指定明确的差异基因阈值，把所有基因按照在两组样本中的差异表达程度进行排序，然后采用统计学方法检验预先设定的基因集合是否在排序表的顶端或低段富集。GSEA主要包括三个步骤：计算富集得分（Enrichment Score），估计富集得分的显著性水平；多重假设检验。

Table 20: KEGG GSEA分析表

ID	Description	SIZE	ES	NES	NOM p-val
CARDIAC MUSCLE CONTRACTION(mmu04260)	CARDIAC MUSCLE CONTRACTION(mmu04260)	81	-0.7175705	-1.703925	0.0000804

GS: 基因集的名称;
 SIZE: 基因集下包含的基因数目 (经过条件筛选后的值);
 ES: 富集得分 (Enrichment Score) NES: ES的标准化值 (Normalized Enrichment Score), 同时考虑基因集的个数及基因数目, NES的值代表该基因集中的基因在整体基因排序列表中的富集程度, 简单理解NES为正值基因集上调、负值基因集下调;
 NOM p-val: P-value, 针对ES的排列检验, 表示基因集富集的显著性;
 FDR q-val: FDR法校正的p值;
 FWER p-val: 用FWER法 (Bonferonni校正) 校正后的P值;
 RANK AT MAX: 当ES值达到最大时对应的那个基因在排序好的基因列表中所处的位置;
 LEADING EDGE: 核心基因集, 对ES贡献最大的基因成员。该处有3个统计值, tags表示核心基因集占该基因集中基因总数的百分比, list表示核心基因占所有基因的百分比, signal是将前两项统计数据合在一起计算出的富集信号强度;
 一般认为 $|NES| > 1$, $NOM\ p\text{-val} < 0.05$, $FDR\ q\text{-val} < 0.25$ 的通路是显著富集的, 点击detail查看对应基因集富集的结果。

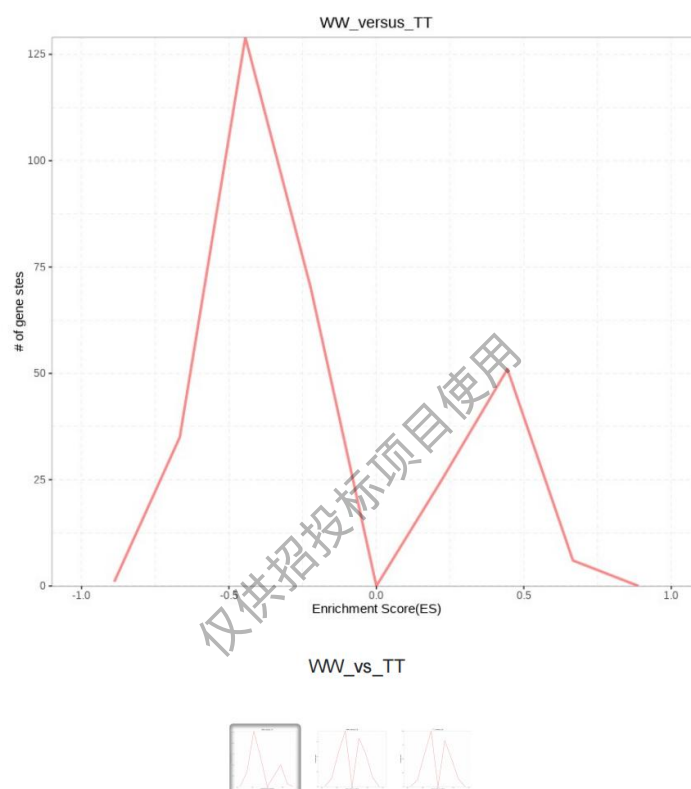
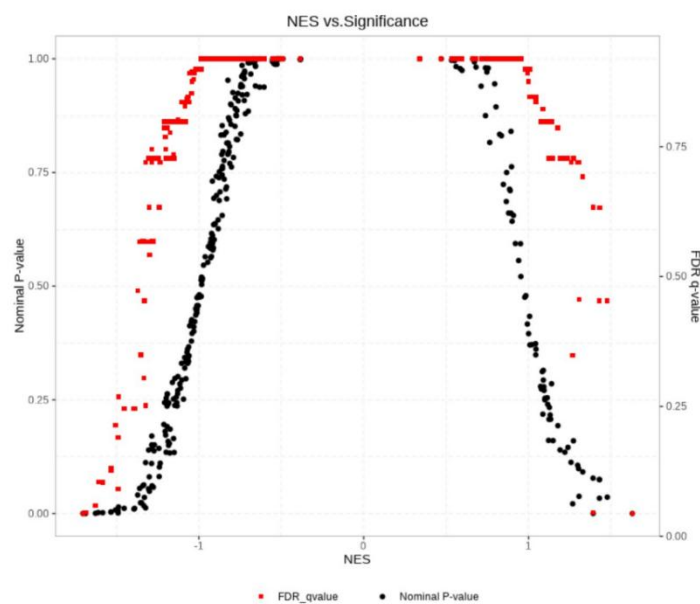


Fig 37: ES图:

注: 基因集的富集分数 (ES) 统计图, 横坐标为ES值, 纵坐标为基因集的数量。



WW_vs_TT

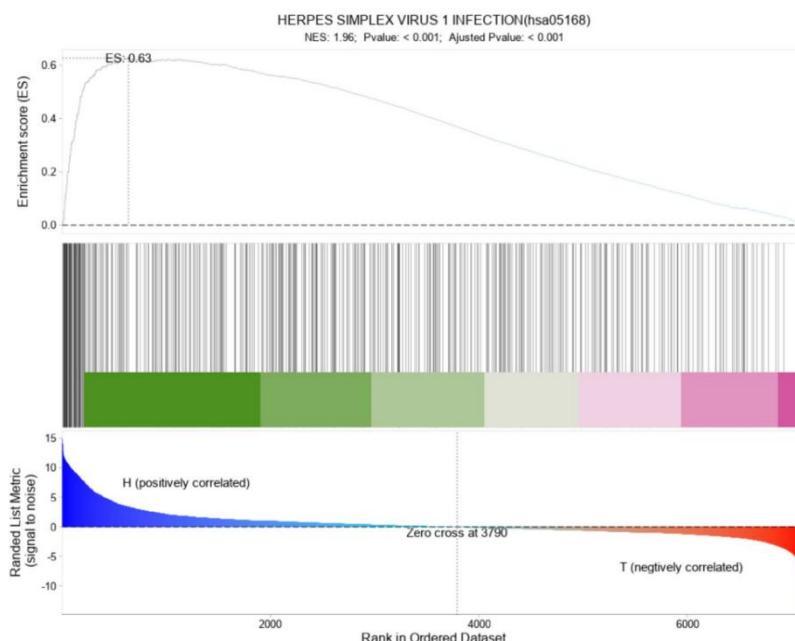


Fig 38: NES图

注：p值与标准化富集分数（NES）的对比图，横坐标为NES值，黑点对应p值结果，红点对应FDR结果；

结果文件

7.2.1 KEGG-GSEA富集分析图



KEGG-GSEA富集分析图

注：第一部分：每个位置对应的ES值的分布曲线，最高峰处的得分（垂直距离0.0最远）便是基因集的ES值；
第二部分：用线条标记了对应基因集中基因出现在排序列表中的位置，每条竖线代表一个基因；
第三部分：是所有基因排序后分布情况，其中红色部分对应的基因在处理组中高表达，蓝色部分对应的基因在对照组中高表达。

8 转录因子分析

8.1 转录因子家族分布

真核生物转录起始过程十分复杂，往往需要多种蛋白因子的协助，转录因子（Transcription Factor, TF）是一类能与基因5'端上游特定序列专一性结合，并与RNA聚合酶II形成转录起始复合体，共同参与转录起始的过程的蛋白质分子。转录因子的预测是分别将植物和动物与PlantTFDB（Plant Transcription Factor Database）和AnimalTFDB（Animal Transcription Factor DataBase）数据库比较，从而预测得到转录因子以及转录因子所属的家族信息，真菌无转录因子家族分析。

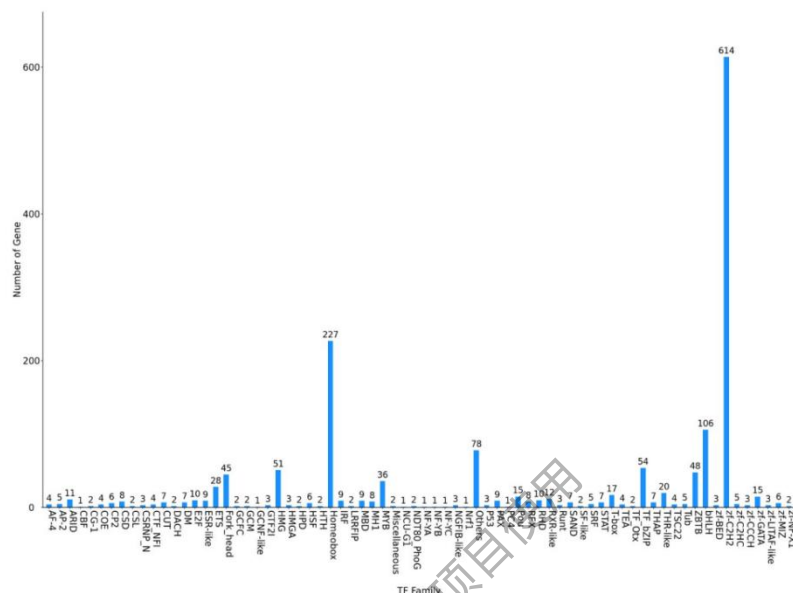
Table 21: 转录因子家族分布表

GeneID	Symbol	Family	Description
ENSMUSG00000000078	Klf6	zf-C2H2	Kruppel-like factor 6 [Source:MGI Symbol;Acc:MGI:1346318]
ENSMUSG00000000093	Tbx2	T-box	T-box 2 [Source:MGI Symbol;Acc:MGI:98494]
ENSMUSG00000000094	Tbx4	T-box	T-box 4 [Source:MGI Symbol;Acc:MGI:102556]
ENSMUSG00000000103	Zfy2	zf-C2H2	zinc finger protein 2, Y-linked [Source:MGI Symbol;Acc:MGI:99213]
ENSMUSG00000000134	Tfe3	bHLH	transcription factor E3 [Source:MGI Symbol;Acc:MGI:98511]
ENSMUSG00000000247	Lhx2	Homeobox	LIM homeobox protein 2 [Source:MGI Symbol;Acc:MGI:96785]

转录因子家族分布表

GeneID	Symbol	Family	Description
ENSMUSG000000000282	Mnt	bHLH	max binding protein [Source:MGI]

注：GeneID：基因ID；
Symbol：基因名称；
Family：基因所属的转录因子家族；
Description：基因描述。



转录因子家族分布图

注：横坐标为不同转录因子家族，纵坐标为该转录因子家族包含的基因数目。

结果文件

8.2 差异表达转录因子分布

对预测为转录因子的差异表达基因进行统计，根据转录因子所属家族信息，对比较组中各转录因子家族包含的差异表达转录因子数目进行柱状图展示。

Table 22: 差异表达转录因子分布表

GeneID	Symbol	Family	Description	Regulation
ENSMUSG000000000938	Hoxa10	Homeobox	homeobox A10 [Source:MGI Symbol;Acc:MGI:96171]	Down Regulation
ENSMUSG000000000942	Hoxa4	Homeobox	homeobox A4 [Source:MGI Symbol;Acc:MGI:96176]	Up Regulation
ENSMUSG000000015619	Gata3	zf-GATA	GATA binding protein 3 [Source:MGI Symbol;Acc:MGI:95663]	Down Regulation
ENSMUSG000000021848	Otx2	TF_Otx	orthodenticle homeobox 2 [Source:MGI Symbol;Acc:MGI:97451]	Down Regulation
ENSMUSG000000022952	Runx1	Runt	runt related transcription factor 1 [Source:MGI Symbol;Acc:MGI:99852]	Down Regulation

差异表达转录因子分布表

GeneID	Symbol	Family	Description	Regulation
--------	--------	--------	-------------	------------

注：GeneID：基因ID；
Symbol：基因名；
Family：基因所属的转录因子家族；
Description：基因描述；
Regulation：上下调信息。

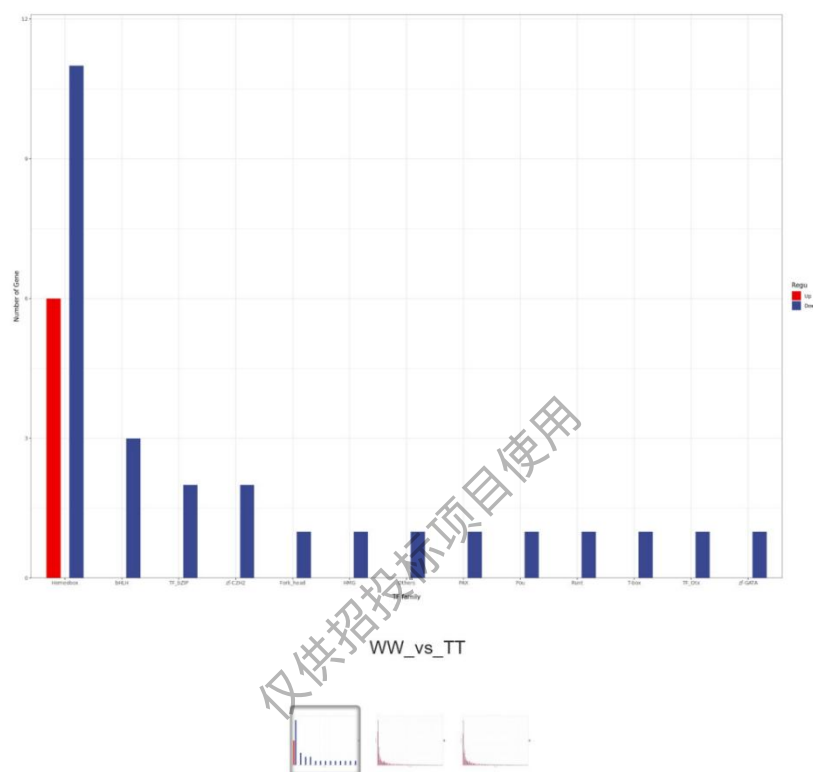


Fig 41: 差异表达转录因子分布图

注：横坐标为不同转录因子家族，纵坐标为该转录因子家族包含的差异基因数目，红色表示上调差异转录因子，蓝色表示下调转录因子。

结果文件

9 结构分析

9.1 转录本拼接

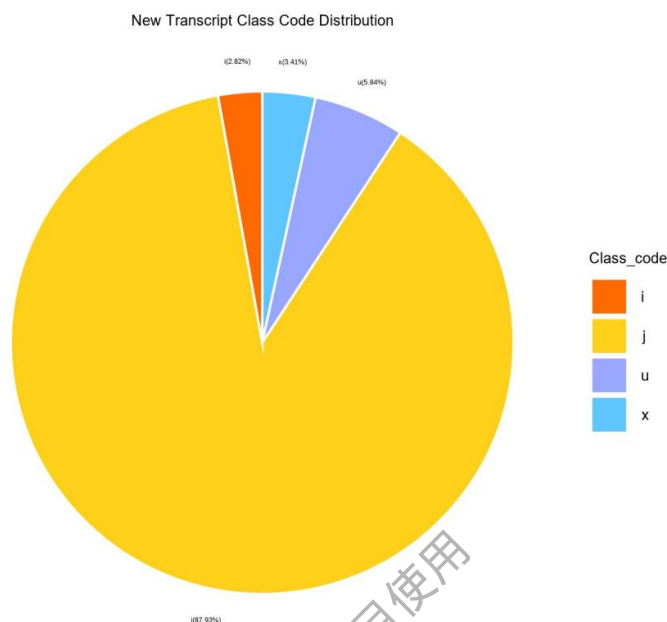
转录本是由基因通过转录形成的一种或多种可编码蛋白质的成熟的mRNA。对于二代转录组测序结果，基于参考基因组的有无情况，转录本的组装主要有两种方式：基于映射的组装和从头组装，在已有参考基因组的基础上，我们选择基于映射的组装方式，使用软件StringTie ([http://ccb.jhu.edu/] (http://ccb.jhu.edu/software/stringtie/) [software/stringtie/] (http://ccb.jhu.edu/software/stringtie/)) 将mapped reads进行组装拼接。

结果文件

9.1.1 新转录本分析

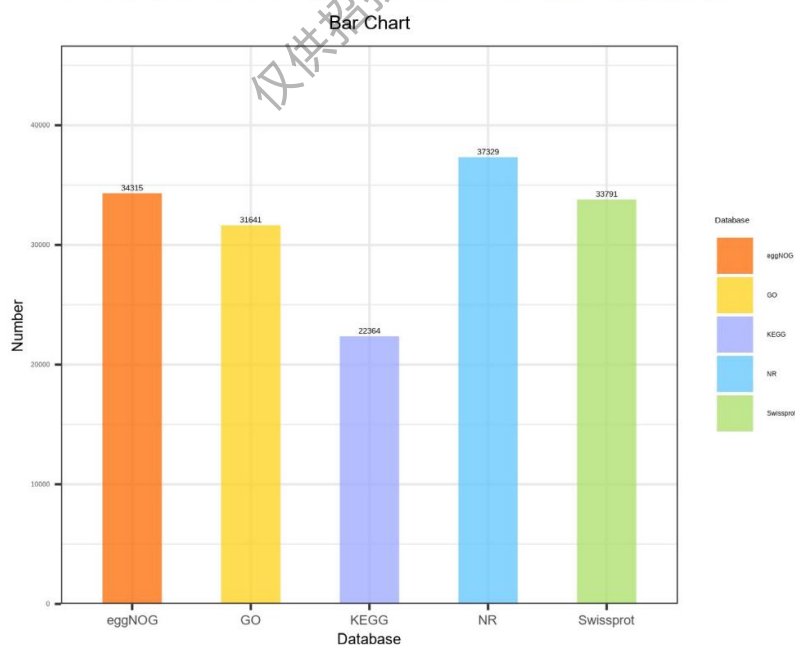
我们将拼接得到的转录本序列与已知转录本进行比较，获得没有注释信息的转录本，将Class Code为“j”，“i”，“u”的转录本作为新转录本，“x”可能为已知转录本的反义转录本，因此也作为新转录本，对所有新转录本进行功能注释。

Class Code是Stringtie给予拼接后的转录本与已知基因和转录本的位置关系的描述，具体见下图：



饼图

注：不同颜色代表不同的Class_code，每块区域为Class_code所占比例。Class_code详细描述见分析介绍；



柱状图：

注：横坐标为各个数据库，纵坐标为对应数据库中注释到的新转录本个数。

[结果文件](#)

9.2 UTR优化分析

UTR (Untranslated Regions) , 即非翻译区, 5'UTR是指成熟mRNA位于编码区 (CDS) 上游不被翻译为蛋白质的区域, 同理3'UTR是CDS下游的非翻译区。UTR包含了基因的调控元件, 主要用于调节基因的表达过程。

我们检查基因CDS上游和下游的Reads覆盖区域作为候选UTR区域。通过与已有的UTR注释信息比较, 列出新发现的可能存在的UTR区域。对于研究较少的物种, 新发现的UTR区域可以优化基因结构, 从而完善它们的基因注释信息。

Table 23: UTR结果

Gene_ID	Gene_Name	Position	5'UTR	3'UTR
ENSMUSG00000100764	Gm29155	1[-]43781121-43783055	43783055-43866960	-
ENSMUSG00000051285	Pcmdt1	1[+]7159144-7243852	-	7243852-7244
ENSMUSG00000057363	Uxs1	1[-]43786126-43866960	-	43781121-4378
ENSMUSG00000033021	Gmppa	1[+]75412574-75419823	75412568-75412574	-
ENSMUSG00000055214	Plid5	1[-]175789872-176102878	176102878-176103094	-
ENSMUSG00000061024	Rrs1	1[+]9615633-9617680	-	9617680-9617
ENSMUSG00000100177	Gm29489	1[+]117958045-117974579	117936499-117958045	-
ENSMUSG00000041763	Tpp2	1[+]43972807-44042160	-	44042160-4404
ENSMUSG00000026504	Sdcccag8	1[+]176642226-176848003	-	176848003-1768
ENSMUSG00000026483	Niban1	1[+]151446937-151597690	-	151597690-1515
ENSMUSG00000063659	Zbtb18	1[+]177269917-177278330	177261620-177269917	-
ENSMUSG00000026049	Tex30	1[-]44125773-44141601	44141601-44157968	-
ENSMUSG00000015962	1700016C15Rik	1[+]177557380-177580890	177555706-177557380	-

Gene_ID: 基因ID;

Gene_Name: 基因名;

Gene_Name: 基因名。5'UTR: 候选的5'端UTR区域;

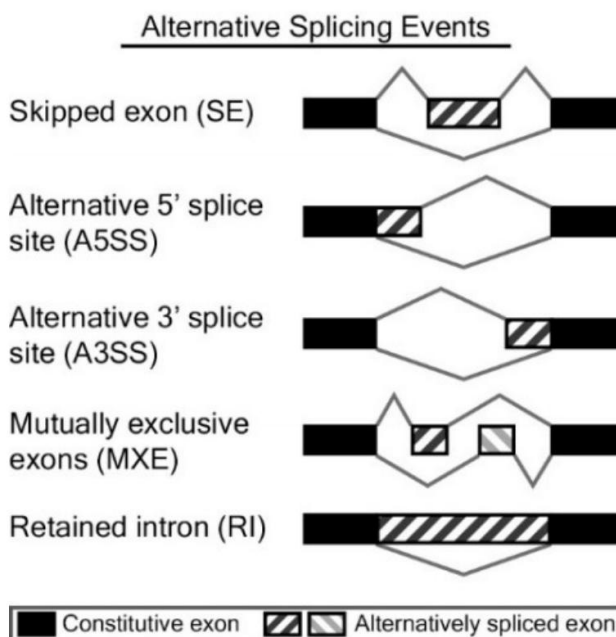
3'UTR: 候选的3'端UTR区域

[结果文件](#)

9.3 差异可变剪切分析

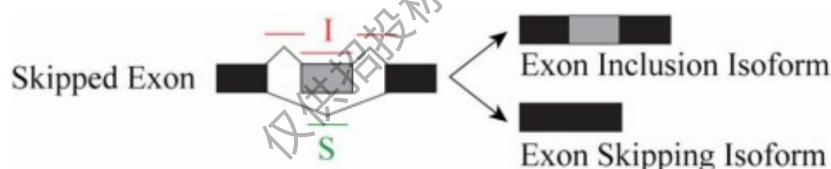
可变剪切 (Alternative Splicing, AS) 指有些基因的一个mRNA前体通过不同的剪接方式 (选择不同的剪接位点) 产生不同的mRNA剪接异构体。它是一种在真核生物中非常普遍的基因表达方式, 是调节基因表达和产生蛋白质多样性的原因。可变剪切研究可分为两个层面: 1、对可变剪切异构体的预测 (定性); 2、对可变剪切异构体相对表达量变化的研究 (定量), 后者往往是实际研究的关注点。差异可变剪切事件, 是指在鉴定到可变剪切事件的基础上, 对可变剪切产生的转录本进行定量, 随后根据比较组方式进行差异分析, 鉴定两组样品在某一剪切位点是否发生差异可变剪切。概况地说, 差异可变剪切分析包含以下4个步骤: 1、转录本组装; 2、可变剪切事件识别; 3、可变剪切转录本定量; 4、差异分析。

我们用rMATS (<http://rnaseq-mats.sourceforge.net/index.html>) 软件进行差异可变剪切分析。该软件可识别的可变剪切事件有5种: 分别是skipped exon (SE) 跳过外显子, alternative 5'splice site (A5SS) 外显子5'剪切位点可变 (即其后的内含子的5'剪切位点可变), alternative 3'splice site (A3SS) 外显子3'可变剪切 (即其前的内含子的3'剪切位点可变), mutually exclusive exons (MXE) 互斥外显子和retained intron (RI) 保留内含子。图示如下:



可变剪切示例图

rMATS的定量方式有两种，一种是Junction counts，只用到了跨越剪接位点的Reads；另外一种方式是Reads On Target And Junction Counts，考虑到了比对到剪接片段的所有Reads。一般情况下，比较两组样品的差异可变剪接只需Junction counts的结果。这里我们给Junction counts的结果，rMATS的统计原理为： $\text{IncLevel} = (I/LI) / (I/LI + S/LS)$ 。



可变剪切统计原理图

IncLevel (exoninclusion level) 用来定量可变剪切，即包含可变剪切事件区域的转录本在包含和跳过可变剪切事件区域的转录本中的百分比。 $\text{IncLevelDiff} = |\text{IncLevel1} - \text{IncLevel2}|$ 和FDR值用来确认两组样品间是否存在差异可变剪切，IncLevel1和IncLevel2分别为两组样品的exon inclusion level，可以通过 $|\text{IncLevelDiff}|$ 与显著性筛选显著差异可变剪切事件。

Table 24: 差异可变剪切分析表

EventType	KnownEventDiff	KnownEventNoDiff	NovelEventDiff	NovelEventNoDiff
A3SS	11	3790	0	0
A5SS	9	2329	0	0
SE	47	11609	53	20244
RI	34	2831	7	537
MXE	4	1247	4	2994
Total	105	21806	64	23775

注：通用表注EventID：差异可变剪切事件编号；

GeneID/GeneName：发生差异可变剪切事件的基因ID/基因名称；

Chr：染色体编号；

Strand: 正负链信息;
 Novel: 该事件是在结构注释文件中已存在的事件 (no) 还是由测序数据预测到的事件 (yes) ;
 UpstreamES/UpstreamEE: 该事件上游的外显子起始位点/终止位点;
 DownstreamES/DownstreamEE: 该事件下游的外显子起始位点/终止位点;
 IJC1/IJC2: 分组1/2中属于包含可变剪切区域的转录本的reads数目;
 SJC1/SJC2: 分组1/2中属于跳过可变剪切区域的转录本的reads数目;
 IncFormLen/SkipFormLen: 表示包含/跳过可变剪切区域的转录本的有效长度, 即该转录本特有的reads的无重复的碱基数目;
 PValue: P值;
 FDR: FDR值;
 IncLevel1/IncLevel2: 处理组/对照组通过计算得到的可变剪切事件表达水平;
 IncLevelDiff: 两组的可变剪切事件表达水平的差异 (IncLevel1-IncLevel2, 处理组-对照组)。

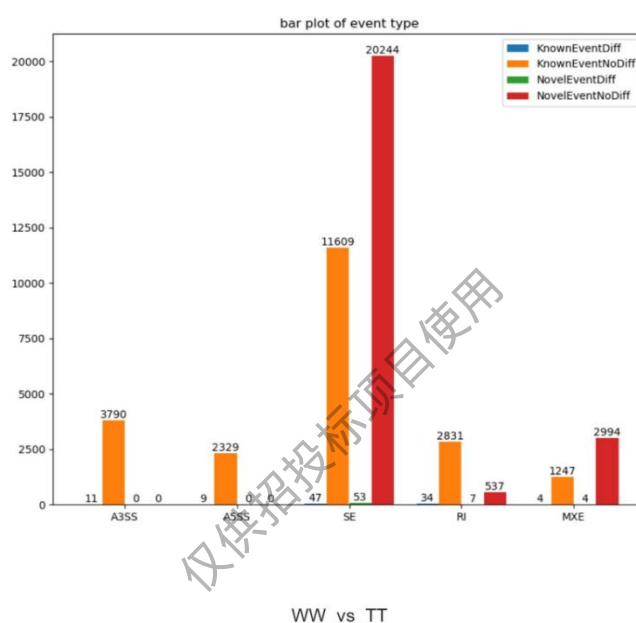


Fig 46: 差异可变剪切分析图

注: 横坐标为rMATS识别的5种可变剪切事件, 纵坐标为对应的可变剪切事件数量。skipped exon (SE): 跳过外显子;
 alternative 5'splice site (A5SS): 外显子3'剪切位点可变 (即其后的内含子的5'剪切位点可变) ;
 alternative 3'splice site (A3SS): 外显子5'可变剪切 (即其前的内含子的3'剪切位点可变) ;
 mutually exclusive exons (MXE): 互斥外显子;
 retained intron (RI): 保留内含子;
 KnownEventDiff: 已知可变剪切事件发生显著差异;
 KnownEventNoDiff: 已知可变剪切事件未发生显著差异;
 NovelEventDiff: 新可变剪切事件发生显著差异;
 NovelEventNoDiff: 新可变剪切事件未发生显著差异。

结果文件

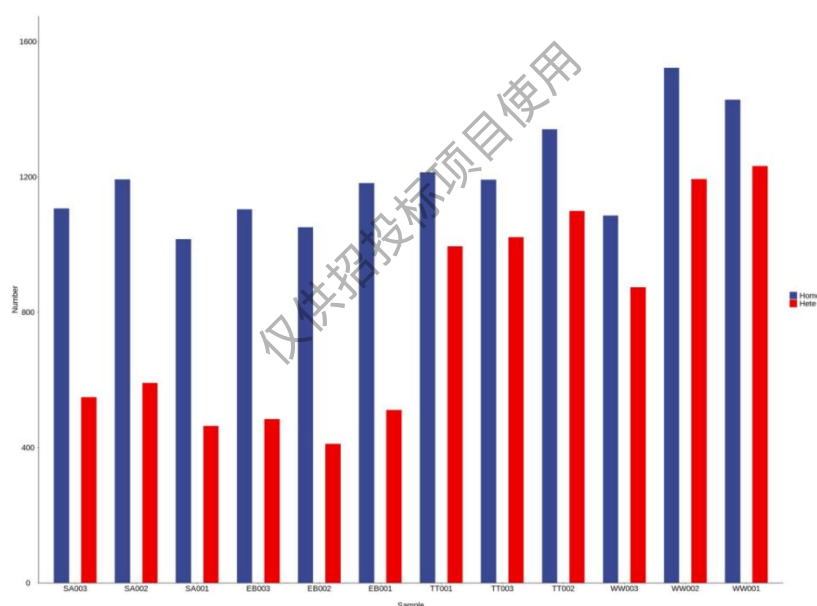
10 变异检测

10.1 SNP

SNP (Single Nucleotide Polymorphisms) 是指在基因组上由单个核苷酸变异形成的遗传标记，其数量很多，多态性丰富。SNP在CG序列上出现的最为频繁，而且多是C转换为T，原因是CG中的C常为甲基化的，自发地脱氨后即成为胸腺嘧啶。SNP出现的原因有很多，有的是遗传背景的单核苷酸多态性，有的是建库技术上造成的突变，有的则可能是测序中读取错误。转录组中的SNP为cSNP，指在编码区出现的SNP。

SNP可对基因的翻译造成影响，对每个样本中各种类型的密码子突变的分布进行了统计，包括：nonsynonymous SNV：非同义单核苷酸突变，指的是一个单核苷酸变异导致对应的氨基酸发生变化；synonymous SNV：同义单核苷酸突变，指的是一个单核苷酸变异虽然改变了碱基，但对应的氨基酸没有发生变化；stopgain：终止密码子增加，突变之后，原本的密码子变成了终止密码子；stoploss：终止密码子减少，突变之后，原本的终止密码子变成了普通密码子；unknown：未知类型。Varscan程序获取SNP位点，过滤标准为：

- 1) SNP位点碱基Q >20;
- 2) 覆盖该位点的Reads数目>8;
- 3) 支持突变位点的Reads数目>2;
- 4) SNP位点的p-value <0.01。



SNP类型分布统图

注：横坐标为样本，纵坐标为对应样本的SNP的数量，Homo (homozygous - variant) 表示纯合子变异体，即这一位点的等位基因都突变了，并且突变相同，Hete (heterozygous - variant) 表示杂合子变异体，即这一位点的等位基因至少有一个突变了，且突变后等位基因不同。

结果文件

10.2 Indel

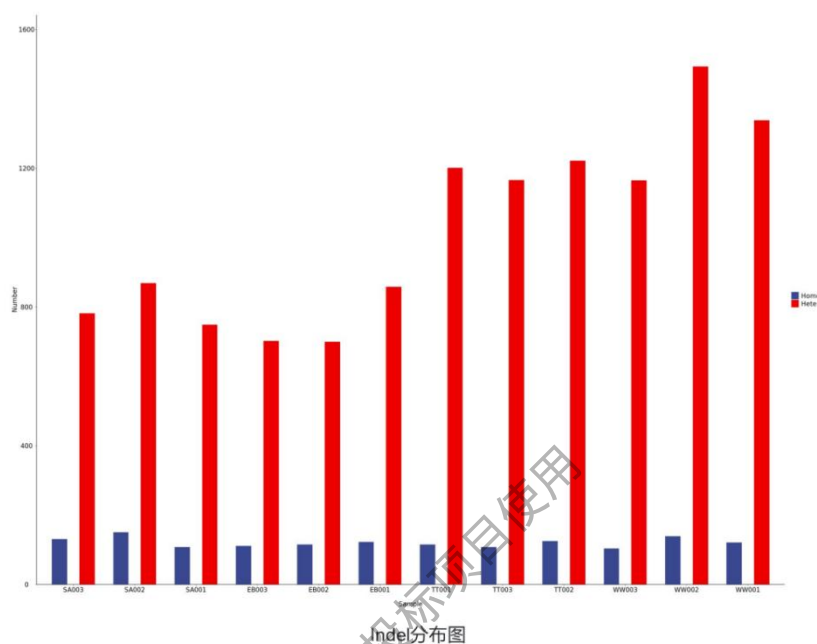
InDel (Insertion-Deletion) 指相对于参考基因组，样本中发生的小片段的插入或者缺失，该插入缺失可能含有一个或多个碱基。InDel可作为一种基因标记用于研究系统进化或物种鉴定。InDel可能造成移码突变，导致mRNA翻译时遇上一个错误的终止密码子。一般InDel不是3的倍数的情况在编码区不常发生，在非编码区相对频繁地发生。除了在高度重复区域附近，InDel发生的频率一般会低于SNP。

Varscan程序获取InDel位点，过滤标准为：1) SNP位点碱基Q >20;

2) 覆盖该位点的Reads数目>8;

3) 支持突变位点的Reads数目>2;

4) SNP位点的p-value <0.01。



Indel分布图

注：横坐标为样本，纵坐标为对应样本的indel的数量。Homo (homozygous - variant) 表示纯合子变异体，即位点的等位基因都突变，并且突变相同，Hete (heterozygous - variant) 表示杂合子变异体，即位点的等位基因至少有一个突变，且突变后等位基因不同。”

结果文件

10.3 Annovar注释

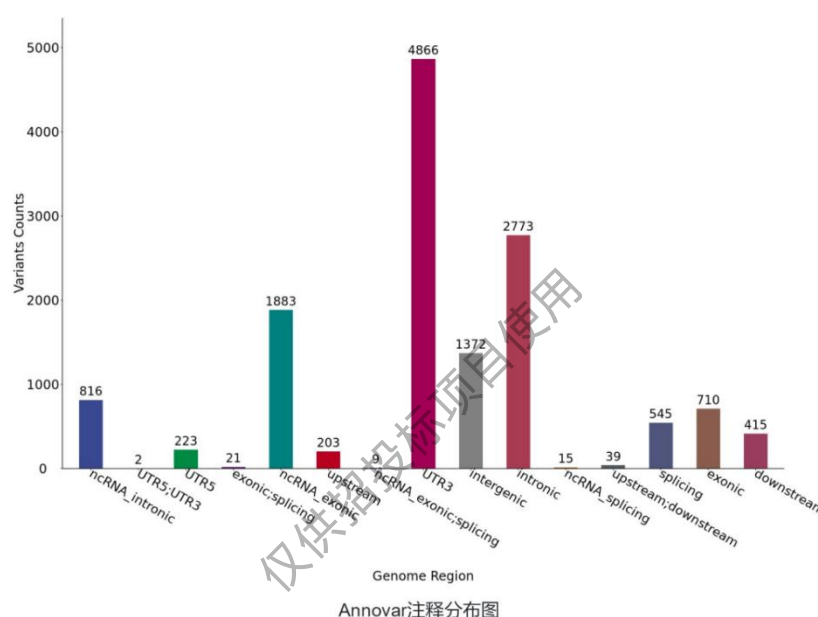
ANNOVAR是一个软件，可以利用最新的数据来分析遗传变异，给出一个列表，包含：染色体、起始位置、终止为止、参考序列上的核苷酸信息和检测到的核苷酸信息。根据ANNOVAR分析结果对SNP/ InDel在几种功能元件上的分布进行统计。

Table 25: Annovar注释表

Type	Count	Percent(%)
ncRNA_intronic	816	5.87
UTR5;UTR3	2	0.01
UTR5	223	1.60
exonic;splicing	21	0.15
ncRNA_exonic	1883	13.55
upstream	203	1.46
ncRNA_exonic;splicing	9	0.06
UTR3	4866	35.02
intergenic	1372	9.87

Type	Count	Percent(%)
intronic	2773	19.96
ncRNA_splicing	15	0.10
upstream;downstream	39	0.28
splicing	545	3.92
exonic	710	5.11
downstream	415	2.98
Sum	13892	100.00

注: Type: SNP/InDel发生的区域。Count: SNP/InDel数目。Percent: 百分比。



Annovar注释分布图

注: 横坐标为SNP/InDel注释的区域, 纵坐标为注释到不同区域的SNP/InDel数目。

结果文件

10.4 突变类型统计

cSNP是指在编码区出现的SNP, 这些SNP直接影响到氨基酸密码子。cSNP包括转换和颠换两种。SNP转换: 嘧啶变成嘧啶或嘌呤变成嘌呤, 即A、G互换, T、C互换。SNP颠换: 嘧啶突变成嘌呤或者相反, 即A、T互换, A、C互换, G、T互换, G、C互换。对各类转换/颠换的数目分别进行统计。

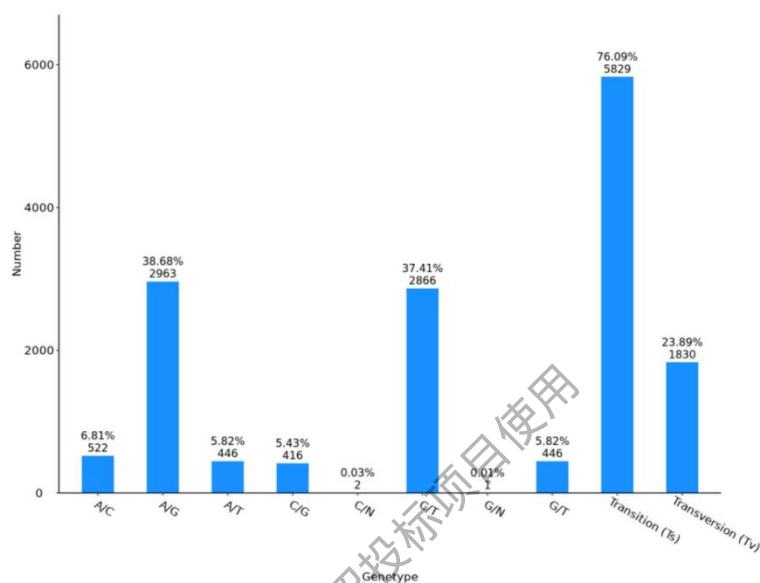
Table 26: 突变类型统计表

Genotype	Number
A/C	522
A/G	2963
A/T	446
C/G	416
C/N	2
C/T	2866

Genotype	Number
G/N	1
G/T	446
Transition (Ts)	5829
Transversion (Tv)	1830

注：Genotype：突变类型；

Number：突变类型数目。



突变类型统计图

注：横坐标为突变类型，纵坐标为该突变类型数目；

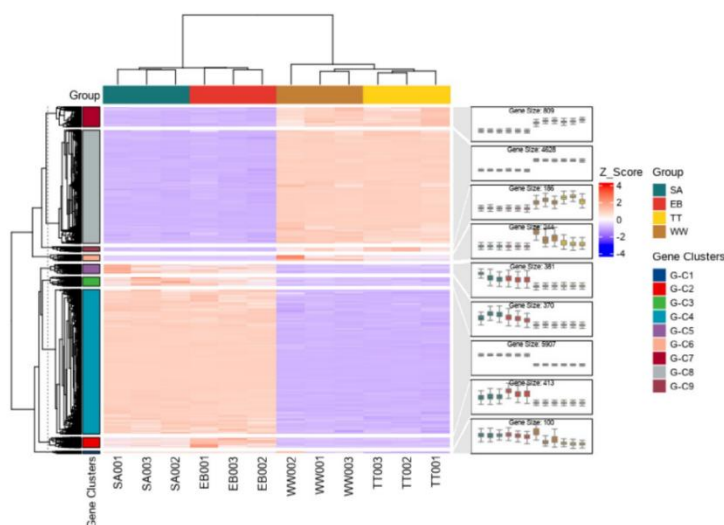
Transition：转换，嘌呤和嘌呤之间的替换，或嘧啶和嘧啶之间的替换；

Transversion：颠换，嘌呤和嘧啶之间的替换。

结果文件

11 高级绘图

11.1 热图+分组箱线图

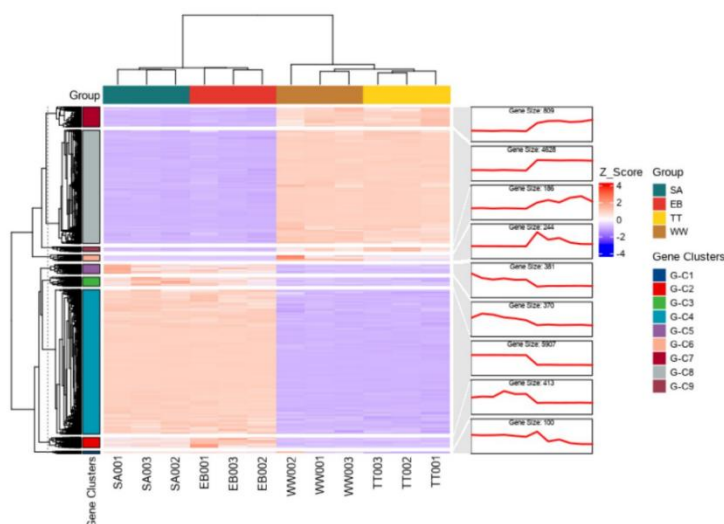


热图+分组箱线图

注：热图主体区域：横向表示基因，每一列为一个样品；
颜色越红，代表基因在该样本中表达量越高，反之颜色越蓝，代表基因在该样本中表达量越低（绘图数据为基因的表达量经过zscore标准化计算后的值）；
左侧样本聚类树：样本聚类情况，表达模式相近的样本聚到一起划分为1个cluster；
右侧色块：聚类到一起的样本使用染色进行划分，不同颜色代表不同的cluster；
箱线图：代表每个cluster下基因在各样本中的表达模式；
箱体表示数据的四分位数，即数据集的中位数和上下四分位数。中位数表示数据的中心趋势，上下四分位数表示数据的分布范围。

结果文件

11.2 热图+趋势图

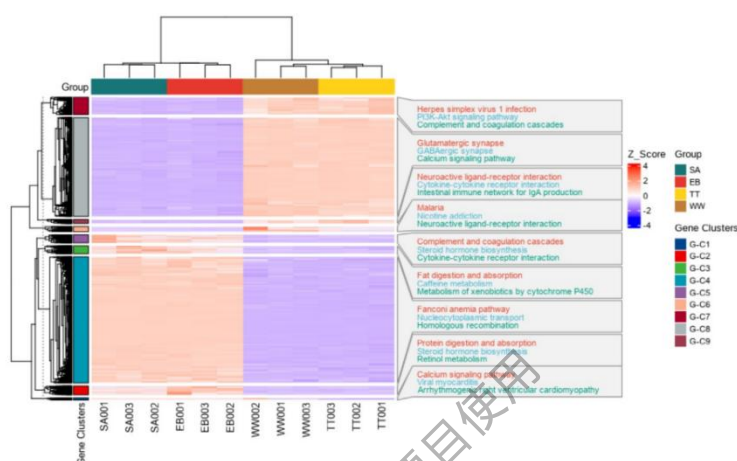


热图+趋势图

注：热图主体区域：横向表示基因，每一列为一个样品；
颜色越红，代表基因在该样本中表达量越高，反之颜色越蓝，代表基因在该样本中表达量越低（绘图数据为基因的表达量经过zscore标准化计算后的值）；
左侧样本聚类树：样本聚类情况，表达模式相近的样本聚到一起划分为1个cluster；
右侧色块：聚类到一起的样本使用染色进行划分，不同颜色代表不同的cluster；
趋势图：用于展示每个cluster下基因在各样本中的表达模式，红线表示Cluster中的所有基因在样品中表达量的平均值。

结果文件

11.3 热图+富集词条图

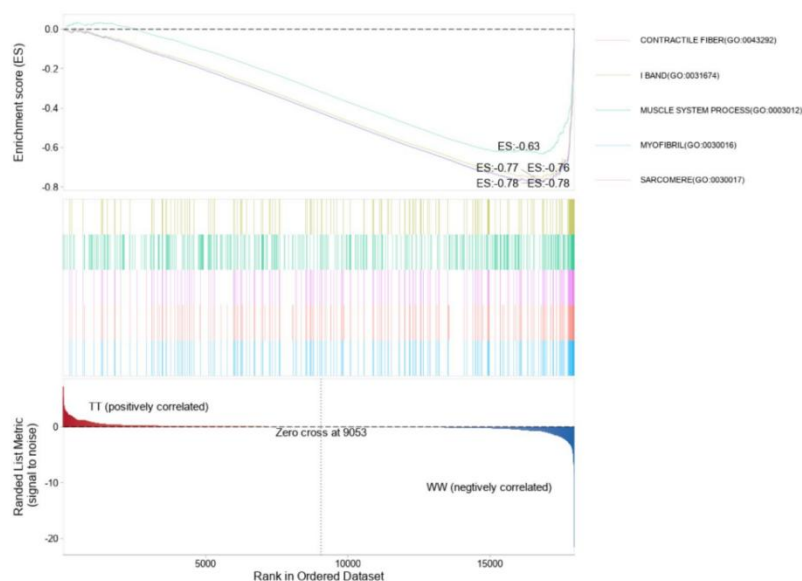


热图+趋势图

注：热图主体区域：横向表示基因，每一列为一个样品；
颜色越红，代表基因在该样本中表达量越高，反之颜色越蓝，代表基因在该样本中表达量越低（绘图数据为基因的表达量经过zscore标准化计算后的值）；
左侧样本聚类树：样本聚类情况，表达模式相近的样本聚到一起划分为1个cluster；
右侧色块：聚类到一起的样本使用染色进行划分，不同颜色代表不同的cluster富集分析结果，并在每一个cluster里面显示前3个通路的具体信息

结果文件

11.4 GSEA-多通路富集分析图



WW_vs_TT



Fig 54: GSEA-多通路富集分析图

注：第一部分：每个位置对应的ES值的分布曲线，最高峰处的得分（垂直距离0.0最远）便是基因集的ES值，不同颜色代表不同的通路/基因集；

第二部分：用线条标记了对应基因集中基因出现在非序列列表中的位置，每条竖线代表一个基因，颜色与第一部分相对应；

第三部分：是所有基因排序后分布情况，其中左侧部分对应的基因在处理组中高表达，右侧部分对应的基因在对照组中高表达。

结果文件

12 材料方法

12.1 中文版材料方法

[中文版材料方法](#)

12.2 英文材料方法

[英文材料方法](#)

12.3 报告解读视频

初次进行转录组测序的老师，可通过报告解读视频详细了解转录组的分析内容及结果；

真核有参转录组报告解读；

链接：<https://pan.baidu.com/s/1PflmeymVAf1uDEVyS0Smbg?pwd=lt49>

提取码：lt49

13 附录

13.1 数据库介绍

13.1.1 GO

基因本体论联合会建立的数据库 (Gene Ontology, <http://geneontology.org/>)。GO的产生主要是为了解决同一基因在不同数据库定义的混乱性以及不同物种的同一基因在功能定义上的混乱性。它是一个国际化的基因功能分类体系, 提供了一套动态更新的标准词汇表 (Controlled Vocabulary) 来全面描述生物体中基因和基因产物的属性。GO涵盖三个方面, 分别描述基因的分子功能 (Molecular Function)、细胞的组件作用 (Cellular Component)、参与的生物学过程 (Biological Process)。基因或蛋白质可以通过ID对应或者序列注释的方法找到与之对应的GO编号, 而GO编号可用于对应到GO Term, 即功能类别或者细胞定位。

GO的基本单元是Term, 每个Term有一个唯一的标示符 (由"GO:"加上7个数字组成, 例如GO:0072669); 每类Ontology的Term通过它们之间的联系 (is_a, part_of, regulate) 构成一个有向无环的拓扑结构。GOSlim是缩减版的GO术语, 它提供了GO注释的概述性结果。

13.1.2 KEGG

京都基因与基因组百科全书 (Kyoto Encyclopedia of Genes and Genomes, <http://www.kegg.jp/>) 是一个整合了基因组、化学和系统功能信息的数据库。把从已经完整测序的基因组中得到的基因目录与更高级别的细胞、物种和生态系统水平的系统功能关联起来是KEGG数据库的特色之一。KEGG注释主要包括: (1) KO (KEGG Ortholog) 注释, 即将分子网络的相关信息跨物种注释; (2) KEGG Pathway注释, 即代谢通路注释, 获得物种内分子间相互作用和反应的网络。

13.1.3 UniProt

UniProt知识库 (UniProt Knowledgebase, <http://www.uniprot.org/help/uniprotkb>) 的子数据库, 是经过有经验的分子生物学家和蛋白质化学家仔细核实的高质量、手工注释的、非冗余的蛋白数据集。SwissProt数据库的每个条目都有详细的注释, 包括结构域、功能位点、跨膜区域、二硫键位置、翻译后修饰、突变体等。该数据库中还包括了与核酸序列数据库EMBL/GenBank/DBJ、蛋白质结构数据库PDB以及Prosite、PRINTS等十多个二次数据库的交叉引用代码。

13.1.4 EC

国际生物化学会酶学委员会 (Enzyme Commission, <http://enzyme.expasy.org/>), 根据酶所催化的反应类型和机理, 把酶分成6大类: 氧化还原酶、转移酶、水解酶、裂合酶、异构酶及合成酶。

13.1.5 eggNOG

真核生物直系同源蛋白质聚类 (Evolutionary Genealogy of Genes: Non-supervised Orthologous Groups, <http://eggno5.embl.de/#/app/home>), 具体信息请参考 (<http://www.ncbi.nlm.nih.gov/COG/>)。我们将列出所有基因的eggNOG ID, 然后把这些eggNOG ID归入适当的eggNOG分类单元 (Category), 由此对基因组的所有基因功能做分类统计, 从宏观上认识该物种的基因功能分布特征。

构成每个eggNOG的蛋白都是被假定为来自于一个祖先蛋白, 并且因此或者是直系同源 (Orthologs) 或者是旁系同源 (Paralogs)。Orthologs是指来自于不同物种的由垂直家系 (物种形成) 进化而来的蛋白, 并且典型的保留与原始蛋白有相同的功能。Paralogs是那些在一定物种中的来源于基因复制的蛋白, 可能会进化出新的与原来有关的功能。

13.2 名词解释

Table 27: 名词解释表

名词	描述
----	----

名词	描述
Raw Data / Raw Reads	测序下机的原始数据
接头 / Adapter	接头是测序时在序列两端分别加上的一段人工序列，接头上含有与测序引物互补结合的序列，通过和测序引物结合来对目的片段进行测序。当加上接头后的序列片段比实际测序读长短时，3'端会测到接头序列，接头序列在分析之前需要去除掉。
模糊碱基 / N	测序中不能确定的碱基，以N表示。一条序列中N越多说明该序列质量越低，一般该种序列需要剔除掉
Clean Data / Clean Reads	过滤后的高质量数据，用于后续分析
Read / Reads	测序中每一条序列称为一个 Read
Read count	比对到一个基因上的 Reads 数目
转录本 / Transcript	是由一条基因通过转录形成的一种或多种可供编码蛋白质的成熟的mRNA
FoldChange	差异表达倍数，同一基因在两个组中的表达量之商，即baseMean_Treat/baseMean_Control
P-value	显著性，统计学根据显著性检验方法所得到的P 值，一般以 $P < 0.05$ 为显著， $P < 0.01$ 为非常显著，其含义是样本间的差异由抽样误差所致的概率小于0.05 或0.01
cSNP	SNP (Single Nucleotide Polymorphisms) 是指在基因组上由单个核苷酸变异形成的遗传标记，其数量很多，多态性丰富。cSNP 是指在编码区出现的 SNP，这些 SNP 直接影响到氨基酸密码子
SNP 转换	嘧啶变成嘧啶或嘌呤变成嘌呤，即 A、G 互换，T、C 互换
SNP 颠换	嘧啶突变成嘌呤或者相反，即 A、T 互换，A、C 互换，G、T 互换，G、C 互换
InDel	Insertion-Deletion，指相对于参考基因组，样本中发生的小片段的插入或者缺失，该插入缺失可能含有一个或多个碱基。InDel 可作为一种基因标记用于研究系统进化或物种鉴定
ϕ (exon inclusion level)	表示包含可变剪切事件区域的转录本在包含和跳过可变剪切事件区域的转录本中的百分比
DAG	有向无环图，在图论中，如果一个有向图无法从某个顶点出发经过若干条边回到该点，则这个图是一个有向无环图。有向无环图是描述一项工程进行过程的有效工具

13.3 常用术语

13.3.1 FASTQ 格式

FASTQ格式 (http://en.wikipedia.org/wiki/FASTQ_format) 是一种文本格式，常用于存储生物学序列及其对应的质量分值。FASTQ格式文件可以采用文本编辑软件（如写字板、UltraEdit、EditPlus等工具）打开，由于文件较大，对电脑的内存要求较高。FASTQ格式中，每个Read由四行信息表示。

第一行为序列名称，以@开头，其后是序列描述；第二行为碱基序列；第三行为"+"号，不代表任何意义；第四行碱基质量，与第二行的碱基序列一一对应。示例如下：

```
@M00200:111:000000000-A6VNV:1:1101:15594:1337 1:N:06
ACGCGGGTATCTAATCCTGTTTGCTCCCCACGCTTTCGCGCCTCAGTGTCAGTTAC
+
ABABADBBDDFFGGGFGGGFGGHBGHBGGHGGGGGGHGGGGGGHGGFBGEGEG
```

13.3.2 质量值

我们使用Sanger质量值来评估下机数据的测序质量。质量值，简称Q值，是碱基读取错误率p的取整映射结果，等于Phred quality score，计算公式为：

$$Q_{phred} = -10 * \log_{10}(p)$$

测序错误率与Q值的简明对应方式如下表所示。

Table 28: 错误率与Q值对应关系表

测序错误率	Q 值
5%	13
1%	20
0.1%	30
0.01%	40

不同的测序平台，采用不同的方案对FASTQ文件中的碱基进行质量编码，Q值与碱基质量的对应关系为：Q值加上一个偏移数值，得到的结果按照ASCII码值对照表（见表2）转换成对应的字符，参考信息如下所示：

```

! " $ % & ' ( ) * + , - . / 0 1 2 3 4 5 6 7 8 9 : ; < = > ? @ A B C D E F G H I J K L M N O P Q R S T U V W X Y Z [ \ ] ^ _ ` a b c d e f g h i j k l m n o p q r s t u v w x y z { | } ~
33 59 64 73 104 126
0.....26...31.....40
-5....0.....9.....40
0.....9.....40
3.....9.....40
0.2.....26...31.....41

S - Sanger Phred+33, raw reads typically (0, 40)
X - Solexa Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64, raw reads typically (-5, 40)
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
(Note: See discussion above).
L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)

```

我们的FASTQ文件采用 1.8+版本编码，将所有字符的ASCII值减去偏移值33，即可得到碱基的Q值。例如，字符I的ASCII值为73，减去33后得到40，那么该字符对应位置的碱基质量为40，测序错误率则为0.01%。

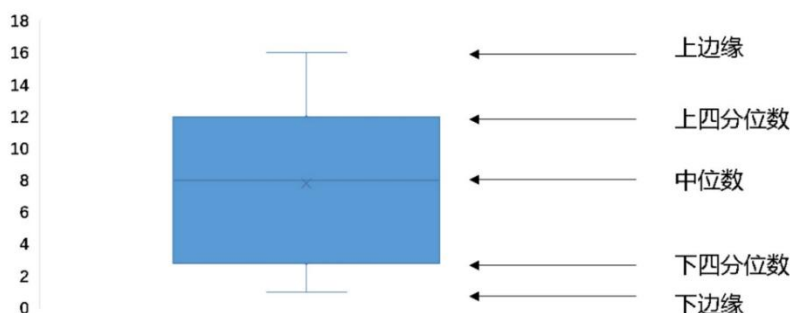
Table 29: ASCII码表

十进制	字符	Q值	十进制	字符	Q值	十进制	字符	Q值	十进制	字符	Q值
32			48	0	15	64	@	31	80	P	47
33	!	0	49	1	16	65	A	32	81	Q	48
34	"	1	50	2	17	66	B	33	82	R	49
35	#	2	51	3	18	67	C	34	83	S	50
36	\$	3	52	4	19	68	D	35	84	T	51
37	%	4	53	5	20	69	E	36	85	U	52
38	&	5	54	6	21	70	F	37	86	V	53
39	'	6	55	7	22	71	G	38	87	W	54
40	(7	56	8	23	72	H	39	88	X	55
41)	8	57	9	24	73	I	40	89	Y	56
42	*	9	58	:	25	74	J	41	90	Z	57
43	+	10	59	;	26	75	K	42	91	[58
44	,	11	60	<	27	76	L	43	92		59
45	-	12	61	=	28	77	M	44	93]	60

十进制	字符	Q值	十进制	字符	Q值	十进制	字符	Q值	十进制	字符	Q值
46	.	13	62	>	29	78	N	45	94	^	61
47	/	14	63	?	30	79	O	46	95	_	62

13.3.3 四分位数

四分位数是指把所有数值由小到大排列并分成四等份，处于第一和第三个分割点位置的数值就是四分位数。



四分位数示例：

13.3.4 Sam / Bam 格式

Sam (sequence alignment/map format) 是一种由Sanger制定的序列比对格式标准，以Tab为分割符的文本格式，可用文本编辑软件打开（如写字板、UltraEdit、EditPlus等工具），主要应用于测序序列比对到基因组上的结果表示，当然也可以表示任意的多重比对结果。当把fastq文件比对到基因组上之后，我们通常会得到一个Sam或者Bam为扩展名的文件。而Bam就是Sam的二进制文件（B取自binary），占用空间更小，不可打开，只能用samtools等软件转换为Sam格式后打开。

Sam分为两部分，注释信息（header section）和比对结果（alignment section）。注释信息可有可无，每一行都是以@开头，用不同的tag表示不同的信息，tag包括@HD（符合标准的版本、对比序列的排列顺序说明）、@SQ（参考序列说明）、@RG（比对上的序列说明）、@PG（使用的程序说明）、@CO（任意的说明信息）。比对结果部分的每一行表示一个片段（segment）的比对信息，包括11个必须的字段（mandatory fields）和一个可选的字段，字段之间用Tab分割。示例及介绍如下：

<pre> @HD VN:1.5 SO:coordinate @SQ SN:ref LN:45 r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG * r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA * r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0; r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC * r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1; r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1 </pre>											Header section
<pre> r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG * r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA * r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0; r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC * r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1; r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1 </pre>											Alignment section
<p>Optional fields in the format of TAG:TYPE:VALUE</p> <p>QUAL: read quality, * meaning such information is not available</p> <p>SEQ: read sequence</p> <p>TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.</p> <p>PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.</p> <p>RNEXT: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.</p> <p>CIGAR: summary of alignment, e.g. insertion, deletion</p> <p>MAPQ: mapping quality</p> <p>POS: 1-based position</p> <p>RNAME: reference sequence name, e.g. chromosome/transcript id</p> <p>FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.</p> <p>QNAME: query template name, aka. read ID</p>											

Sam/bam格式示例图

1. QNAME，比对片段的编号

2. FLAG, 位标识, 1表示该read是pair中的一条 (read表示本条read, mate表示pair中的另一条read), 2表示pair—正—负地比对上参考序列, 4表示这条read没有比对上, 8表示mate没有比对上, 16表示这条read比对上负链, 32表示mate比对上负链, 64表示这条read是read1, 128表示这条read是read2等, FLAG的值是符合情况的数字相加总和, 即83= (64+16+2+1) 表示该read为read1, 比对到负链上, 其mate比对到正链上
3. RNAME, 参考序列的编号
4. POS, 比对上的位置, 注意是从1开始计数, 如果没有比对上, 此处为0
5. MAPQ, 比对的质量, 越高则位点越独特, 计算方法: $Q = -10 \log_{10} p$, p是该序列不来自这个位点的估计值
6. CIGAR (Compact Idiosyncratic Gapped Alignment Report), 使用数字加字母表示比对结果, 如M表示match/mismatch, I表示insertion, D表示deletion等, 数字表示碱基个数, 即42M4I5M为该序列42个碱基匹配, 4个insertion, 5个碱基匹配
7. RNEXT, mate的名称, 如果没有mate, 用*表示
8. PNEXT, mate的位置, 如果没有mate, 用0表示
9. TLEN, paired reads间的距离, 当mate序列位于本序列上游时该值为负值, 如果比对区域仅有一个区段, 或者不可用时, 此处为0
10. SEQ, read序列
11. QUAL, read质量
12. Optional Fields, 可选字段, 格式如: TAG:TYPE:VALUE, 其中TAG由两个大写字母组成, 每个TAG代表一类信息, 如AS表示匹配的得分, XS表示第二好的匹配得分, YS表示mate序列匹配的得分等, TYPE表示TAG对应值的类型, 可以是字符串 (Z)、整数 (i) 等

13.3.5 GFF /GTF 格式

gff格式是一种Sanger研究所定义的, 可以简单方便地描述DNA、RNA以及蛋白质序列的特征的数据格式, 已经成为序列注释的通用格式, 许多软件都支持输入或者输出gff格式。每一行代表一个特征条目 (如基因、转录本、CDS、exon等), 每行有9列, 以Tab为分割符, 每列分别列出该特征条目的一些信息。gff可用文本编辑软件打开 (如写字板、UltraEdit、EditPlus等工具)。

Table 30: GTF/GFF文件示例表

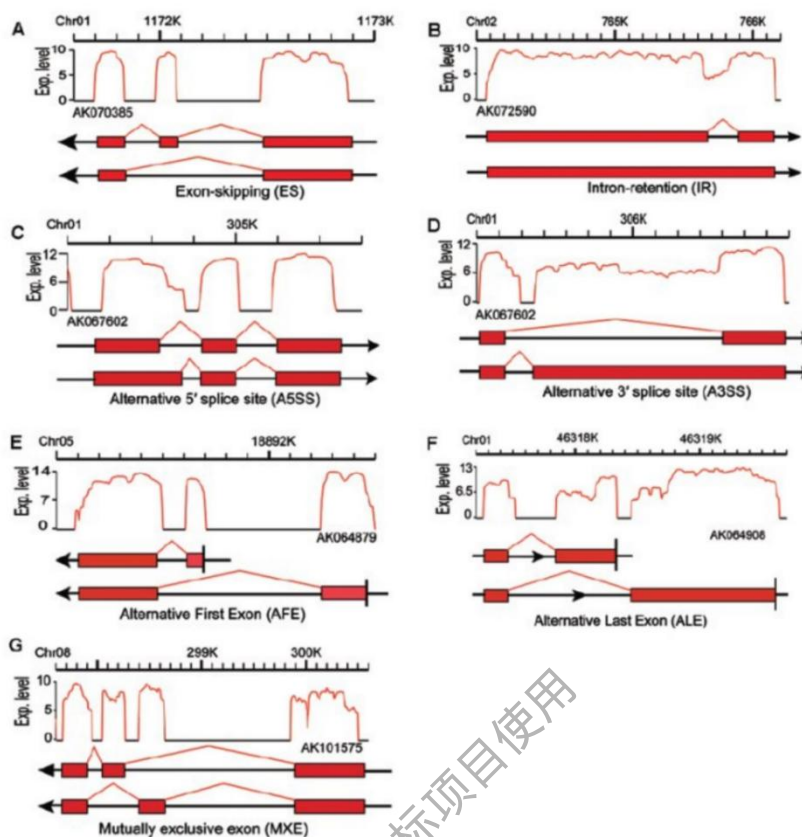
ctg123	PFAM	gene	1000	5000	.	+	.	ID=gene001;Name=EDEN
ctg123	PFAM	TF_binding_site	1000	1012	.	+	.	Parent=gene001
ctg123	PFAM	mRNA	1050	5000	.	+	.	ID=mRNA001;Parent=gene001
ctg123	PFAM	mRNA	1050	5000	.	+	.	ID=mRNA002;Parent=gene001
ctg123	PFAM	exon	1300	1500	.	+	.	Parent=mRNA001
ctg123	PFAM	exon	1050	1500	.	+	.	Parent=mRNA001,mRNA002
ctg123	PFAM	CDS	1201	3902	.	+	0	ID=cds001;Parent=mRNA001
ctg123	PFAM	CDS	3000	4600	.	+	2	ID=cds001;Parent=mRNA001
ctg123	PFAM	CDS	1201	1500	.	+	1	ID=cds002;Parent=mRNA002

1. 序列编号, 可能是染色体或者scaffold的名称
2. 来源, 产生这一特征条目的程序、数据库或者项目

- 3.类型, 如gene, transcript, CDS, mRNA, exon, five/three_prime_utr, start/stop_codon等
 - 4.起始位点, 这一特征条目在序列上的起始位置, 从1开始计数
 - 5.终止位点, 这一特征条目在序列上的终止位置, 不能大于序列的长度
 - 6.得分, 是注释信息可能性的说明, 可以是序列相似性比对时的E-values值或者基因预测是的P-values值。“.”表示为空
 - 7.序列的方向, +表示正义链, -反义链, ?表示未知
 - 8.相位, 仅对类型为“CDS”的条目有效, 有效值为0、1、2, 0表示这一特征条目的第一个碱基是一个密码子的第一个碱基, 1表示这一特征条目的第二个碱基是一个密码子的第一个碱基, 以此类推
 - 9.属性, 以多个键值对组成的注释信息描述, 键与值之间用“=”, 不同的键值对用“;”隔开, 一个键可以有多个值, 不同值用“,”分割。键可以为ID (该特征条目的编号, 在一个gff文件中必须唯一), Name (该特征条目的名称, 可以重复), Parent (该特征条目的父级特征条目, 值为父级特征条目的编号, 比如外显子所属的转录本编号, 转录本所属的基因的编号。值可以为多个) 等
- gtf格式与gff格式前8列基本相同, 不同之处在于第9列, 虽然同样是标签与值配对的情况, 但gtf格式的标签与值之间以空格分开, 且每个属性之后都要有分号; (包括最后一个属性), 而且第9列必须以gene_id以及transcript_id开头。

13.3.6 可变剪切事件

可变剪切 (或选择性剪切) 指有些基因的一个mRNA前体通过不同的剪接方式 (选择不同的剪接位点) 产生不同的mRNA剪接异构体。一般认为, 可变剪接有5种基本形式: ①内含子保留; ②可变的5'端; ③可变的3'端; ④跳过外显子; ⑤互斥外显子 (一组外显子中只能有一个表达)。也有分为7种形式的, 即以上5种可变剪切形式加上可变的第一个或最后一个外显子, 而这两种形式更有可能是可变启动子、可变polyA位点造成的。如下图:



可变剪切事件示例图

13.3.7 FASTA 格式

在生物信息学中，FASTA格式（又称为Pearson格式），是一种基于文本用于表示核苷酸序列或氨基酸序列的格式，可用文本编辑软件打开（如写字板、UltraEdit、EditPlus等工具）。序列文件的第一行是由大于号“>”或分号“;”打头的任意文字说明（习惯常用“>”作为起始），用于序列标记。从第二行开始为序列本身，只允许使用既定的核苷酸或氨基酸编码符号。通常核苷酸符号大小写均可，而氨基酸常用大写字母。文件每行的字母一般不应超过80个字符。示例如下：

Seq1

ADQLTEEQIAEFKEAFSLFDKDGDTITTKELGTVMRSLGQNPTEAELQDMINEVDAD*

13.3.8 Class Code

Class Code是Stringtie给予拼接后的转录本与已知基因和转录本的位置关系的描述。

Table 31: class_code表

优先级	Code	描述
1	=	内含子链完全匹配
2	c	包含
3	j	潜在的新转录本：至少有一个与已知转录本共享的可变剪切位点
4	e	一个外显子片段覆盖了一个已知外显子和至少10bp的已知内含子，可能是 mRNA 前体片段
5	i	转录本片段完全在一个已知内含子中

优先级	Code	描述
6	o	在外显子水平上基本覆盖一个已知转录本
7	p	可能是聚合酶产生的片段（包含在一个已知转录本的2k bases以内）
8	r	重复片段，当参考基因组序列有soft-masked碱基，并且转录本50%的碱基是小写时
9	u	未知片段，处在基因间区的转录本
10	x	外显子覆盖到已知基因的反链上
11	s	转录本的一个内含子覆盖到一个已知内含子的反链上，可能是比对错误导致的
12	.	包含多种情况

13.4 常见问题

问：结果文件中static和images文件夹是干什么的？

答：static文件夹下是网页格式所需的静态文件，不包含结果相关内容，static下的文件请不要改动，否则将会影响网页报告的内容展示。在images文件夹下有所有图片的png格式，其所属分析内容及对应pdf在每一章节下的“结果文件”中均写明了，如果修改或移动images文件夹及里面的内容同样也会影响网页报告的内容展示。

问：当PCA分析、样本相关性检验和聚类分析热图的聚类结果不同时哪个更可靠？

答：PCA分析和样本相关性检验均基于样品所有基因的表达量，而聚类热图是基于差异基因的表达量对样品和基因双向聚类，两者针对的问题不同。而如果PCA和样品相关性检验的聚类结果不一致，则以PCA聚类结果为准，因为PCA分析中会保留对样品贡献大的基因的信息，而样品相关性检验则对所有基因作相同的处理，PCA聚类结果更精确。

问：为什么派森诺在差异分析时使用p value而不使用校正后的p value来筛选？

答：因为DESeq2的p value计算方法已经非常严格了，使用p value足以作为筛选标准，如果使用校正后的p value来筛选可能筛选出较少的差异基因/ LncRNA / CircRNA，甚至没有。如果筛选出的差异基因/ LncRNA / CircRNA数目较多，可以通过调整p value来减少筛选出的差异基因的数目。

问：能否修改筛选条件以获得不同数目的差异基因/ LncRNA / CircRNA？

答：可以，差异表达是一个相对的概念，可通过修改筛选条件获取期望的差异基因/ LncRNA / CircRNA数目，不过一般推荐使用 $pvalue < 0.05$ 和 $|\log_2\text{FoldChange}| > 1$ 。

问：能否只使用部分基因/ LncRNA / CircRNA做差异分析？

答：不能，差异分析是基于所有基因/ LncRNA / CircRNA作为背景做的，如果使用部分基因/ LncRNA / CircRNA做差异分析，则会丢失整体的信息，如测序深度、reads分布特征等，从而产生偏差。

问：为什么有的基因/ LncRNA / CircRNA在两个样本中表达量差别很大，却不属于显著差异的基因/ LncRNA / CircRNA？

答：因为显著差异是一个基于统计学的概念，不能直观地通过表达量的大小来判断基因/ LncRNA / CircRNA是否显著差异，而需要在整体的基础上进行计算后判断。

问：聚类分析热图能否调整基因/ LncRNA / CircRNA或样品的顺序？

答：不能，聚类分析热图是根据表达量自动聚类形成的，相似的基因/ LncRNA / CircRNA或样品会被聚在一起，其结果代表了样品或基因/ LncRNA / CircRNA间的距离。

问：GO功能富集和GO功能分类有何区别？

答：GO功能分类是将基因注释到相应功能的GO分类下，而GO功能富集分析则是将功能相似的基因集通过统计学检验算法富集到一起，从而方便研究具有某一类功能的基因。

问：为什么派森诺在差异基因KEGG富集分析结果中只提供代谢通路图的链接而不直接提供代谢通路图？

答：第一，图片格式所占空间会很大；第二，在链接到的KEGG网页上会提供更丰富的内容，鼠标悬停在节点上会出现该节点的编号等信息，点击节点将会跳转到节点的详细信息，还可通过选择物种查看相应物种特有的代谢通路，如果想保存代谢通路图，则可用鼠标右键点击图片保存。

13.5 参考文献

- [1] Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. The COG database: an updated version includes eukaryotes[J]. BMC Bioinformatics. 2003 Sep 11;4:41.
- [2] Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, Arnold R, Rattei T, Letunic I, Doerks T, Jensen LJ, von Mering C, Bork P. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges[J]. Nucleic Acids Res. Epub 2011 Nov 16; PubMed 22096231.
- [3] The Gene Ontology Consortium, Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology[J]. Nat Genet. 2000 May, 25 (1) : 25–29.
- [4] Minoru Kanehisa,* Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. The KEGG resource for deciphering the genome[J]. Nucleic Acids Res. 2004 January 1; 32 (Database issue) : D277–D280.
- [5] Zhou L., Chen J., Li Z., Li X., Hu X., et al. (2010) . Integrated profiling of microRNAs and mRNAs: microRNAs located on Xq27.3 associate with clear cell renal cell carcinoma. PLoS One 5: e15224.
- [6] Michael I Love, Wolfgang Huber, Simon Anders. (2014) . Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology.
- [7] Wang L., Feng Z., Wang X., Wang X., Zhang X. (2010) . DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics 26, 136-8.
- [8] Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, et al. (2008) . KEGG for linking genomes to life and the environment. Nucleic Acids research 16:D480–484.
- [9] Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L: Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature biotechnology 2013, 31 (1) :46-53.
- [10] Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A: High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic acids research 2008, 36 (10) :3420-3435.
- [11] Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M: KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic acids research 2007, 35 (Web Server issue) :W182-185.
- [12] Rogers MF, Thomas J, Reddy AS, Ben-Hur A: SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. Genome Biol. 2012, 13 (1) :R4.

© 2015-2024 Shanghai Personal Biotechnology Co., Ltd. All rights reserved. 沪ICP备12025704号-1

公司地址：上海市徐汇区银都路218号聚科生物园区2号楼

技术顾问: transsupport@personalbio.cn

联系我们: 021-80118168

E-mail: transsupport@personalbio.cn

[关于我们](#)

仅供招投标项目使用