Novogene 诺禾致源 | 基因科技守护生命健康
Advancing Genomics, Improving life

中国区转录业务线个性化分析手册

Novogene 诺禾致源

中国区产品管理部

2022.06

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

Novogene 诺禾致源

1 差异基因转录因子结合位点预测

1.1 分析背景方法介绍

可用于转录组差异基因转录因子结合位点预测分析，并可以结合 SNP 结果，对预测结果文件进行注释。

转录因子结合位点（TFBS）指转录因子（TF）在 DNA 上的结合位点或区域。通常 TFBS 相当短，长度为 4-30 个碱基对，但通常位于 50-200bp 的较大顺式调控区域中，并且在不同基因中重复或基因内有几次重复；转录因子与 DNA 结合的部分是保守的，但通常相当小（3-5bp）；准确预测 TFBS 对于分析基因的转录模式有着重要意义。

方法介绍：

用 TFBSTools(1.18.0) 进行差异转录本的 TFBS 预测分析;TFBSTools(1.18.0)是分析和处理转录因子结合位点的包，JASPAR 数据库：保存核苷酸谱的 PFM 的集合，最广泛使用的是 JASPAR CORE 集合，它是基于试验证据的多细胞真核生物的 TFBS 谱的非冗余集合。

本项目中的 TFBS 分析过程主要包含以下步骤:

1). 取差异 gene 的前 2Kb 的 DNA 序列作为该 gene 的 promoter 区域序列;

2). 采用 TFBSTools(1.18.0)软件，以及 JASPAR 数据库(2018)，对 gene 的 promoter 区域序列进行 TFBS 预测; (当该物种在 JASPAR 数据库中包含物种本身的 TFBS 位点时，采用本物种的模型进行预测，当该物种在 JASPAR 数据库中无本物种 TFBS 位点信息时，采用“all”即全部 TFBS 进行预测;)

3)基因 promoter 区域的 TFBS 的遗传变异可能会导致转录本表达的差异;会结合 SNP 结果，对 TFBS 文件进行注释。

1.2 结果展示

转录因子靶基因预测结果展示

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

geneID	source	featur	start	end	absScore	relScore		ID	TF	class	siteSeqs	Annotation
ENSMUSG0000002	TFBS	TFBS	1390	1396	10.0238498	0.990960815	+	MA0063	Nkx2-5	Homeo dom	ATAATTG	http://jaspar.
ENSMUSG0000002	TFBS	TFBS	509	515	11.92095771	1	+	MA0087	Sox5	High-mobi	ATTGTTT	http://jaspar.
ENSMUSG0000002	TFBS	TFBS	1669	1675	11.92095771	1	+	MA0087	Sox5	High-mobi	ATTGTTT	http://jaspar.
ENSMUSG0000002	TFBS	TFBS	1359	1366	11.78367816	0.993215651	+	MA0709	Msx3	Homeo dom	GCAATTAA	http://jaspar.
ENSMUSG0000002	TFBS	TFBS	1359	1366	10.32565445	0.992912215	+	MA0880	Dlx3	Homeo dom	GCAATTAA	http://jaspar.
ENSMUSG0000002	TFBS	TFBS	1359	1366	10.50606189	0.990518445	+	MA0881	Dlx4	Homeo dom	GCAATTAA	http://jaspar.
ENSMUSG0000002	TFBS	TFBS	1359	1366	11.20198156	1	+	MA0885	Dlx2	Homeo dom	GCAATTAA	http://jaspar.
ENSMUSG0000006	TFBS	TFBS	297	302	11.35502013	1	+	MA0004	Arnt	Basic hel	CACGTG	http://jaspar.
ENSMUSG0000006	TFBS	TFBS	550	555	10.49384652	1	+	MA0006	Ahr::A	Basic hel	TGCGTG	http://jaspar.

TFBS.xls : seqnames: gene 的 promoter 区域序列 id(命名方式见上述文件)

source: TFBS

feature: TFBS

start: TFBS 位点相对 gene promoter 起点的位点坐标;

end: TFBS 位点相对 gene promoter 终点的位点坐标;

absScore: TFBS 位点得分值;

relScore: TFBS 位点与 gene promoter 区域序列的相关性; 收集 相关性>80%的结果;

strand: TFBS 在正负链上的位置

ID: TFBS id; 查询(http://jaspar.genereg.net/) 数据库获得更多注释;

TF: 相关转录调控因子;

class: 相关转录调控因子所属类别;

siteSeqs: TFBS 位点序列信息;

Annotation ： 可链接到相应网站查看详细信息

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

2 RNA 编辑分析

2.1 分析背景及方法介绍

RNA 编辑是指在 mRNA 水平上改变遗传信息的过程，是多种生命形式的遗传编码变异的重要来源。RNA 编辑的主要机制是前体 mRNA 分子中腺苷的去氨基，脱氨基的事件，即 A-to-I 编辑，将特殊的腺苷（A）转换为肌苷（I）。在翻译中，肌苷被解码为鸟苷（G），从而导致密码子的变化，往往会引起蛋白质产物中的氨基酸替换。

我们使用 REDItools 软件进行 RNA 编辑位点的鉴定，并基于此结果进行多样品编辑位点信息的整合，各样品变异位点编辑水平的比较，RNA 编辑位点类型的分布，RNA 编辑位点在染色体上的分布，RNA 编辑簇的鉴定，A->I 编辑位点的鉴定，RNA 编辑位点有义突变和无义突变的鉴定等分析。

此分析需构建链特异性文库，可以更好的区分正负链的编辑信息，并且最好联合样本的 DNA 重测序数据进行校正，可大幅度降低编辑位点的假阳性，提高鉴定结果的准确性。

2.2 结果展示

2.2.1 RNA 编辑位点鉴定

我们使用主流的 REDItools 软件进行 RNA 编辑位点的鉴定。如果是加入了 DNA 数据，结果会多出几列。鉴定结果如下：

Region	Position	Reference	Strand	Coverage-q25	MeanQ	BaseCount[A,C,G,T]
GL892783.2	4690	A	0	38	38.89	[32, 0, 6, 0]
GL896000.1	7174	C	0	56	40.05	[0, 23, 33, 0]
GL896000.1	16580	G	0	91	40.58	[0, 0, 41, 50]
JH118443.1	104001	T	0	19	38.95	[19, 0, 0, 0]
JH118445.1	47935	G	1	27	39.78	[27, 0, 0, 0]
JH118445.1	47941	G	1	27	40	[11, 0, 16, 0]

因表格太多列，做了表格拆分，横列接上表

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

AllSubs	Frequency	gCoverage-q25	gMeanQ	gBaseCount[A,C,G,T]	gAllSubs	gFrequency
AG	0.16	-	-	-	-	-
CG	0.59	15	41	[0, 7, 8, 0]	CG	0.53
GT	0.55	35	39.86	[0, 0, 18, 17]	GT	0.49
TA	1	15	39	[15, 0, 0, 0]	TA	1
GA	1	16	39.06	[16, 0, 0, 0]	GA	1
GA	0.41	16	39.31	[7, 0, 9, 0]	GA	0.44

表 2.2.1 RNA 编辑位点鉴定

Region：RNA editing 位点所在染色体。

Position：RNA editing 位点坐标。

Reference：参考序列在该位点的坐标。

Strand：参考序列在该位点的链方向，0 表示'+'，1 表示'-'，2 表示未知 。

Coverage-q25：该位点的碱基覆盖度(质量值>=25) 。

MeanQ：该位点上所有碱基的 Qpred 的均值 。

BaseCount[A,C,G,T]：依次为 A,C,G,T 类型的碱基在该位点的 reads 数，用逗号隔开。

AllSubs：RNA editing 类型，'-'表示未发生 RNA editing 。

Frequency：发生 RNA editing 的频率。

gCoverage-q25： DNA 数据，该位点的碱基覆盖度(质量值>=25) 。

gMeanQ：DNA 数据，该位点上所有碱基的 Qpred 的均值 。

gBaseCount[A,C,G,T]：DNA 数据，依次为 A,C,G,T 类型的碱基在该位点的 reads 数，用逗号隔开。

gAllSubs：DNA 数据，突变类型，'-'表示未发生突变 。

gFrequency：DNA 数据，发生突变的频率。

2.2.2 编辑位点在染色体上的分布

我们将参考基因组的每条染色体按照 100000bp 为一个单位进行均分（不足 100000bp 按照实际长度统计），统计每个 bin 内的 RNA 编辑位点数目。

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

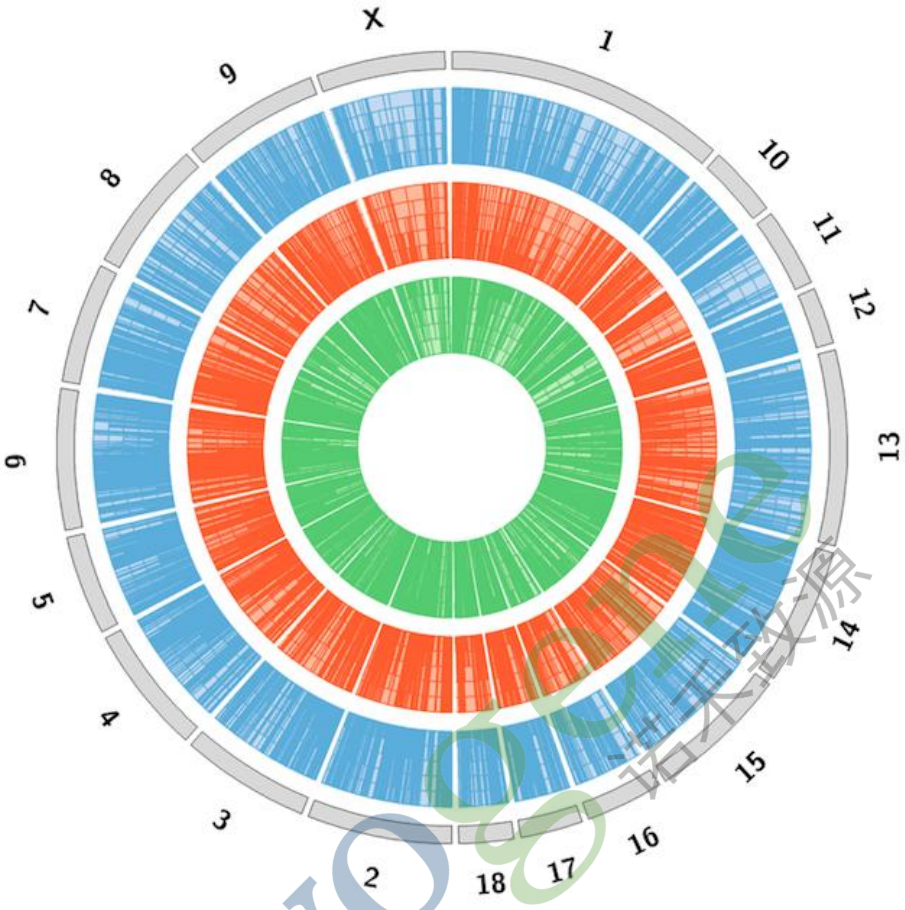


图 2.2.2 编辑位点在染色体上的分布

此图中柱状图的高度代表该 bin 内的 RNA 编辑位点数目，最内圈为负链上鉴定出的 RNA 编辑位点在染色体上的分布信息，中间圈代表的是正链上鉴定出的 RNA 编辑位点在染色体上的分布信息，最外边的一圈代表不区分正负链，总体上鉴定出的 RNA 编辑位点在染色体上的分布信息。

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

2.2.3 RNA 编辑类型分布

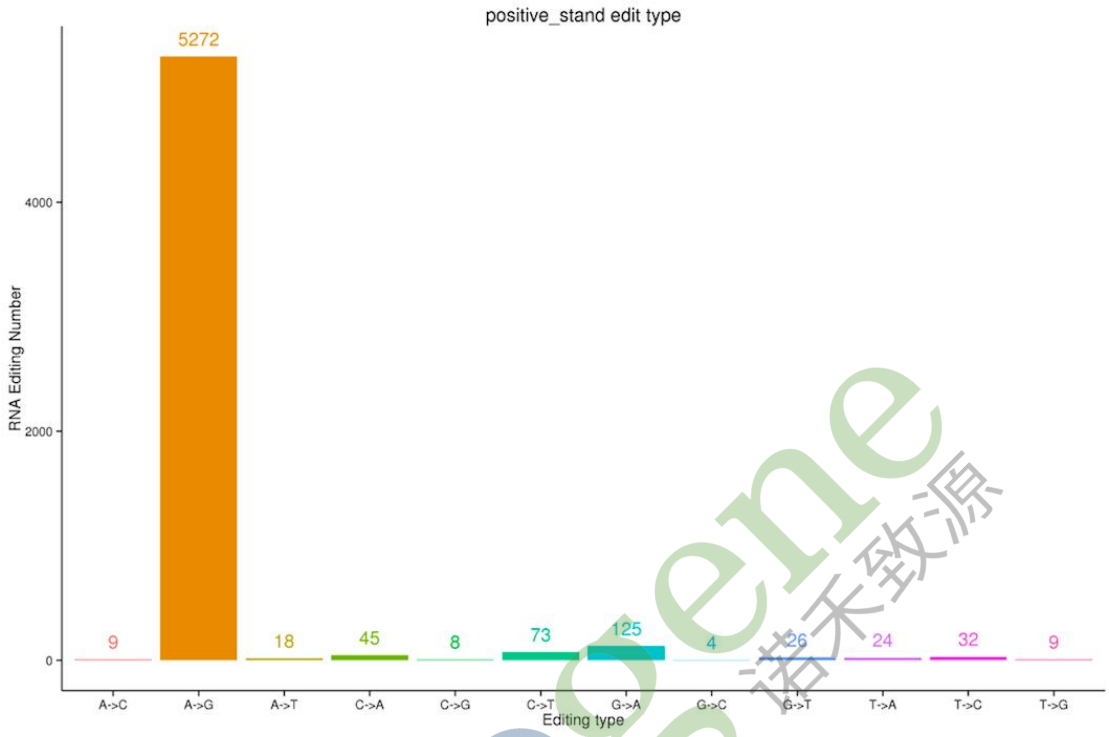


图 2.2.3 RNA 编辑类型分布

此图横坐标代表不同的编辑类型，纵坐标代表发生该编辑类型位点的个数，每个样品会提供三张图，分别是正链，负链，总的编辑类型分布。

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

2.2.4 RNA 编辑类型堆积柱状图

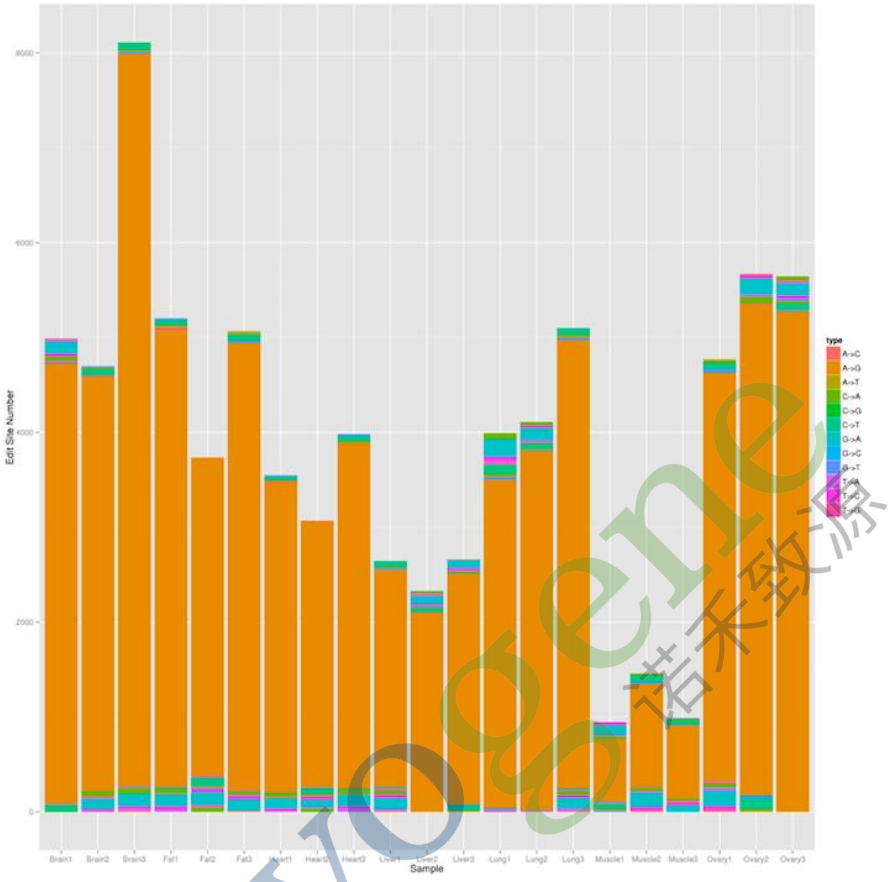


图 2.2.4 RNA 编辑类型堆积柱状图

此图横坐标代表不同样品，纵坐标代表不同编辑类型的个数。

2.2.5 RNA 编辑位点在基因不同功能域的分布

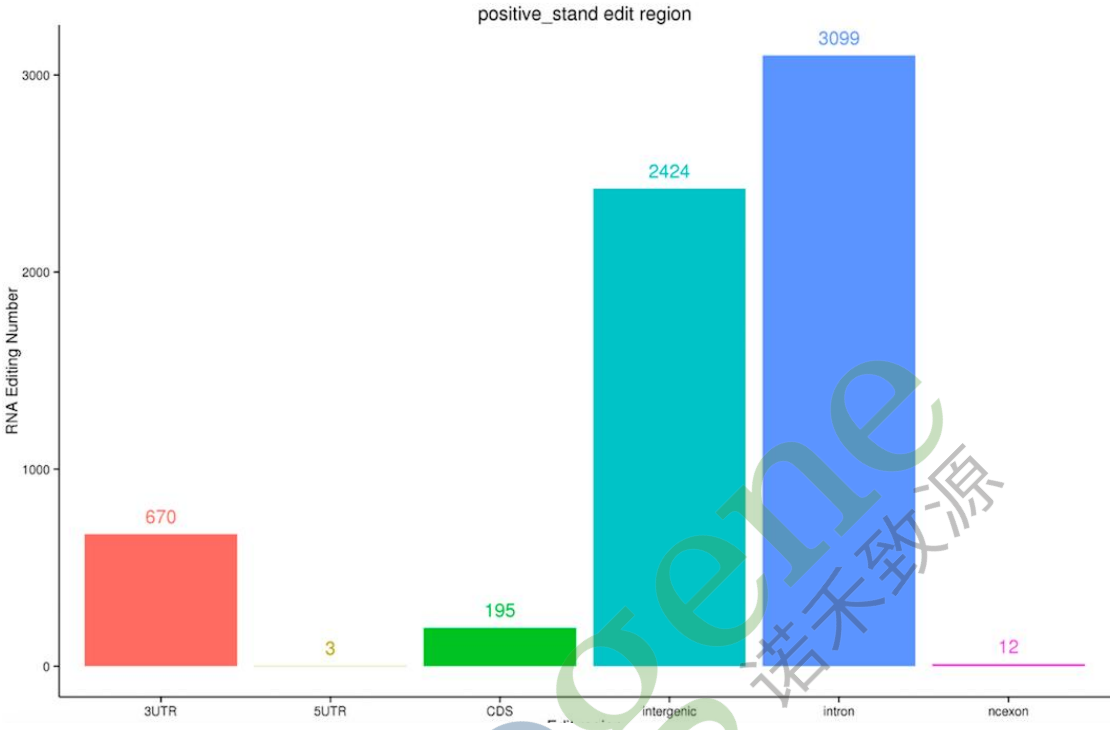


图 2.2.5 编辑位点在基因不同功能域的分布

此图横坐标代表基因不同功能域，纵坐标代表不同功能域的编辑位点个数。

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

2.2.6 编辑位点在不同功能域的堆积柱状图

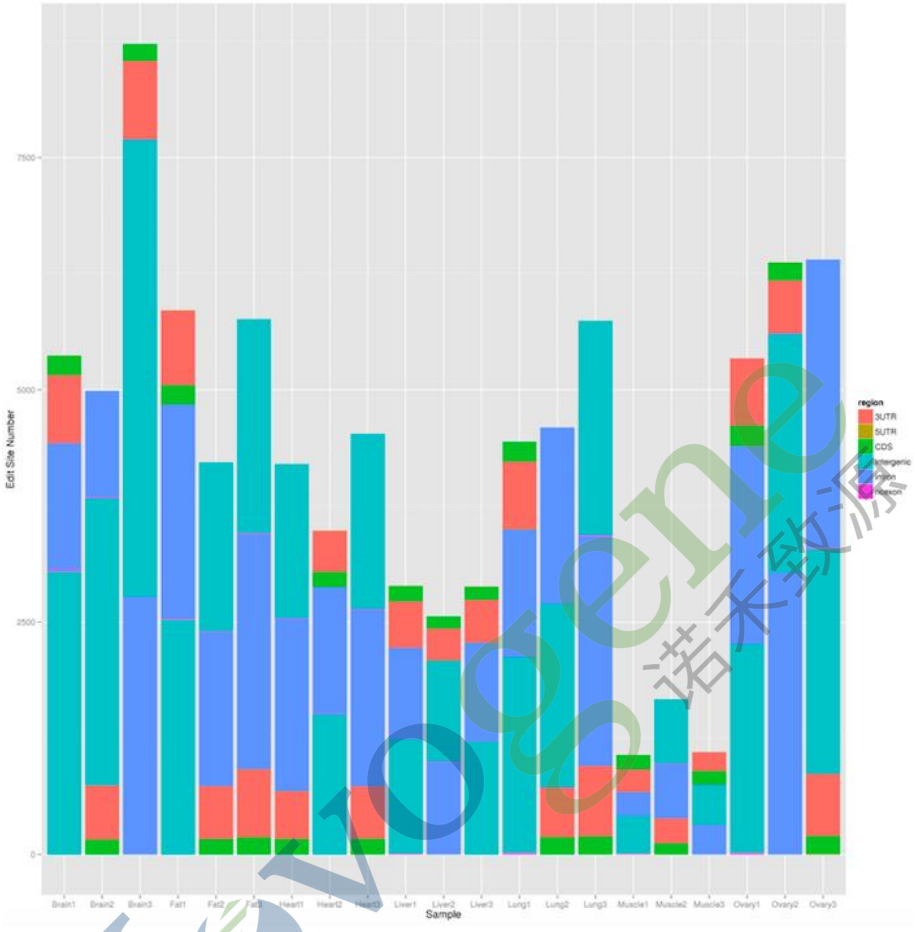


图 2.2.6 编辑位点在不同功能域的堆积柱状图

此图横坐标代表不同样品，纵坐标代表不同的基因功能域中编辑类型的个数。

2.2.7 编辑位点引起的同义突变与非同义突变分析

#chromosome	cordination	ref<->alt	strand	transcript_id	codon_phase
NC_024795.2	4879567	G<->A	-	103856053	1191,2
NC_024795.2	4879624	G<->A	-	103856053	1134,2
NC_024795.2	4879160	C<->T	-	103856053	963,2
NC_024795.2	4879195	G<->A	-	103856053	928,0

因表格太多列，做了表格拆分，横列接上表。

codon_mutate	aa_mutate	synonymous	Nonsynonymous
GAC<->GAT	D<->D	1	0
TGC<->TGT	C<->C	1	0
CTG<->CTA	L<->L	1	0
CCC<->TCC	P<->S	0	1

表 2.2.7 编辑位点引起的同义突变与非同义突变分析

注：该项分析是针对于 exon 上的编辑位点进行分析

#chromosome 染色体

cordination 突变位点

raf>alt 参考基因组 》突变碱基

transcript_id 转录本 ID

codon_phase 密码子相位

codon_mutate 突变前后的密码子

aa_mutate 突变前后的蛋白

synonymous 同义突变

nonsynonymous 非同义突变

2.2.8 编辑簇分析

我们通过编辑簇分析可以查找到染色体编辑位点相对集中的区域。编辑簇鉴定方法：编辑簇内 RNA 编辑位点的最小数目为 5；相邻编辑簇的最低距离为 50bp；编辑簇最小长度为 20bp。

Chrom	**edit box**	**edit site in box**
1	34436936--34437013	34436936,34436943,34436944,34436977,34437013
1	39468190--39468371	39468190,39468197,39468214,39468226,39468238,39468274,39468292,39468310,39468323,39468348,39468371

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

1	75141908--75142038	75141908,75141932,75141939,75141956,75141957,75141980,75141985,75142017,75142038
1	105556331--105556429	105556331,105556342,105556361,105556362,105556407,105556410,105556429
1	106527330--106527443	106527330,106527366,106527379,106527400,106527412,106527423,106527442,106527443
1	132678291--132678363	132678291,132678337,132678350,132678362,132678363

表 2.2.8 编辑簇分析

第一列：染色体。

第二列：编辑簇区间。

第三列：编辑簇内的编辑位点。

2.2.9 A-I 编辑位点的鉴定

一般认为在三个样品以上某个位点均发生了 A->I 的变化，认为该位点有 A->I 的 RNA 编辑事件发生。我们将为客户提供所有 A->I 编辑的位点。

Region	Position	Reference	Strand	Edit type&numbers
NC_024795.2	1006015	A	2	AG,AG,AG,AG
NC_024795.2	1006393	A	2	AG,AG,AG,AG
NC_024795.2	10083330	A	2	AG,AG,AG
NC_024795.2	10083498	A	2	AG,AG,AG,AG
NC_024795.2	10131565	A	2	AG,AG,AG,AG
NC_024795.2	10134439	A	2	AG,AG,AG,AG

表 2.2.9 编辑位点的鉴定

第一列：染色体

第二列：编辑位点

第三列：参考基因组信息

第四列：正负链信息　0 表示'+'，1 表示'-'，2 表示未知 。

第五列：编辑类型，及出现的次数

2.2.10 GO 富集分析

我们对在 exon 区间含有编辑位点的基因进行 GO 富集分析。结果如下表所示：

GO accession	Description	Term type	Over represented p-Value	Corrected p-Value	DEG item	DEG list
GO:1903561	extracellular vesicle	cellular_component	9.0287e-85	1.1892e-80	915	8690
GO:0043230	extracellular organelle	cellular_component	1.2774e-84	1.1892e-80	916	8690
GO:0065010	extracellular membrane-bounded organelle	cellular_component	2.825e-84	1.7532e-80	908	8690
GO:0070062	extracellular exosome	cellular_component	5.4733e-84	2.5475e-80	907	8690

表 2.2.10.1 GO 富集列表

(1) GO accession：Gene Ontology 数据库中唯一的标号信息。

(2) Description：Gene Ontology 功能的描述信息。

(3) Term type：该 GO 的类别(cellular_component：细胞组分；biological_process：生物学过程；molecular_function：分子功能)。

(4) Over represented p-Value：富集分析统计学显著水平。

(5) Corrected p-Value：矫正后的 P-Value，一般情况下，Corrected_pValue < 0.05 该功能为富集项。

(6) DEG item：与该 GO 相关的基因的数目。

(7) DEG list：GO 注释的基因数目。

上述列表后，统计被显著富集的各个 GO term 中的基因数，以柱状图的形式展示。

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

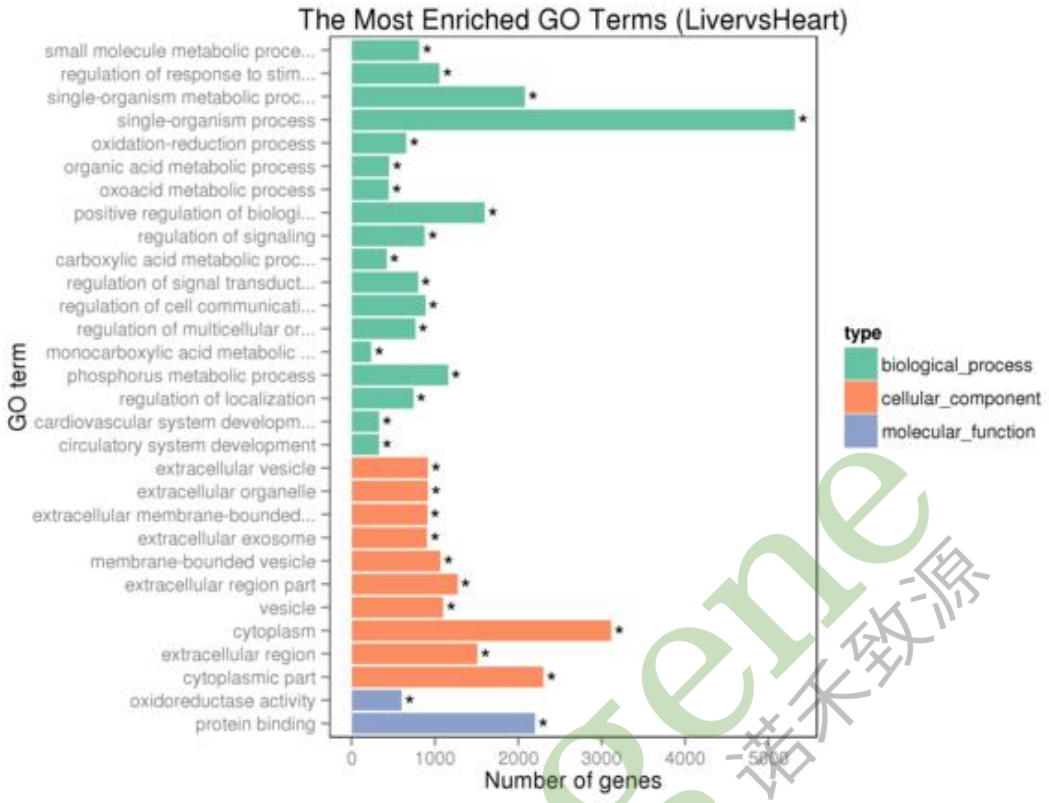


图 2.2.10.2 GO 富集柱状图

此图纵坐标为富集的 GO term，横坐标为该 term 中基因个数。不同颜色用来区分生物过程、细胞组分和分子功能，带“*”为显著富集的 GO term。

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

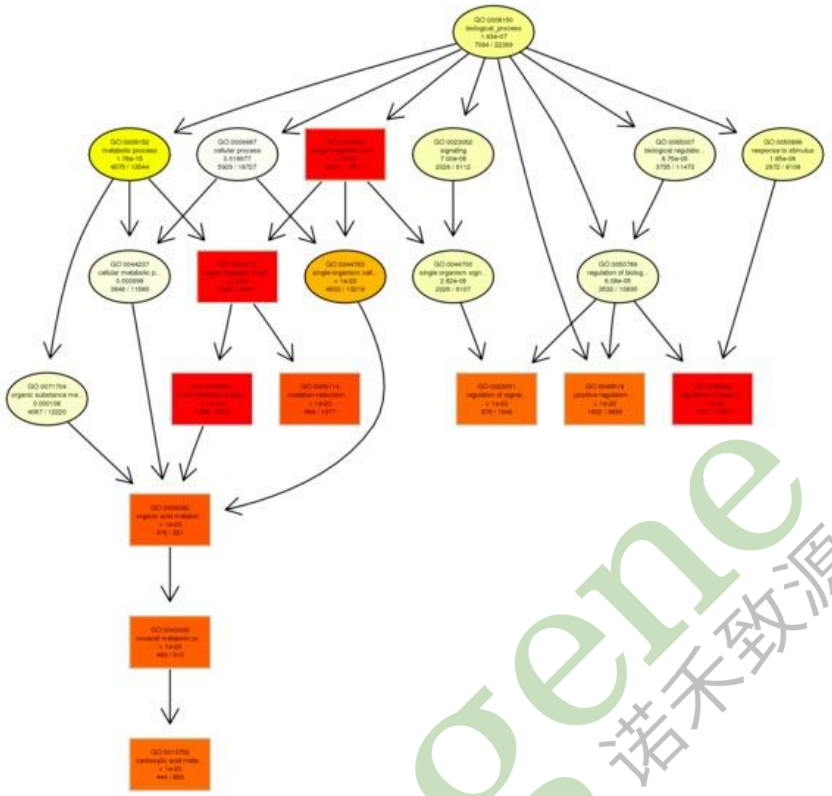


图 2.2.10.3 GO 富集有向无环图

每个节点代表一个 GO 术语，方框代表的是富集程度为 TOP10 的 GO，颜色的深浅代表富集程度，颜色越深就表示富集程度越高，每个节点上展示了该 TERM 的名称及富集分析的 CORRECTED P-VALUE。

2.2.11 KEGG 富集分析

我们对在 exon 区间含有编辑位点的基因进行 KEGG 富集分析。结果如下表所示：

#Term	Database	ID	Sample number	Background number	P-Value	Corrected P-Value
Metabolic pathways	KEGG	ssc01100	770	1166	1.01963364868e-08	2.79379619739e-06
Retinol metabolism	KEGG	ssc00830	54	54	0.000371473789845	0.0366226268339

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

Complement and coagulation cascades	KEGG	ssc04610	67	73	0.000400977666065	0.0366226268339
Peroxisome	KEGG	ssc04146	70	80	0.000737422426016	0.0484083461065

表 2.2.11.1　KEGG 富集列表

(1) #Term：KEGG 通路的描述信息。

(2) Database：KEGG 数据库。

(3) ID：KEGG 数据库中通路唯一的编号信息。

(4) Sample number：该通路下基因的个数。

(5) Background number：该通路下注释基因的个数。

(6) P-value：富集分析统计学显著水平。

(7) Corrected P-value：矫正后的统计学显著水平，Corrected P-value < 0.05 该功能为富集项。

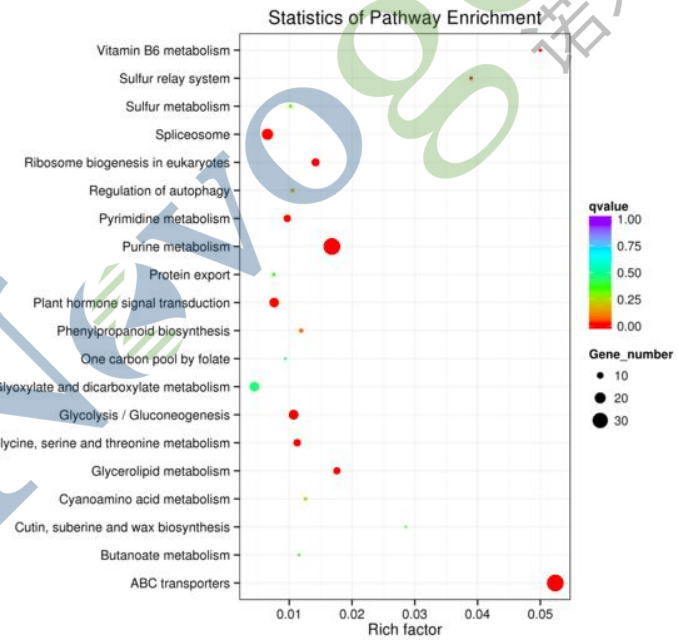


图 2.2.11.2 KEGG 富集散点图

此图纵轴表示 pathway 名称，横轴表示 Rich factor，点的大小表示此 pathway 中候选靶基因个数多少，而点的颜色对应于不同的 qvalue 范围 。

北京诺禾致源科技股份有限公司　网址：www.novogene.com　邮箱：service@novogene.com　电话：010-8283 7801

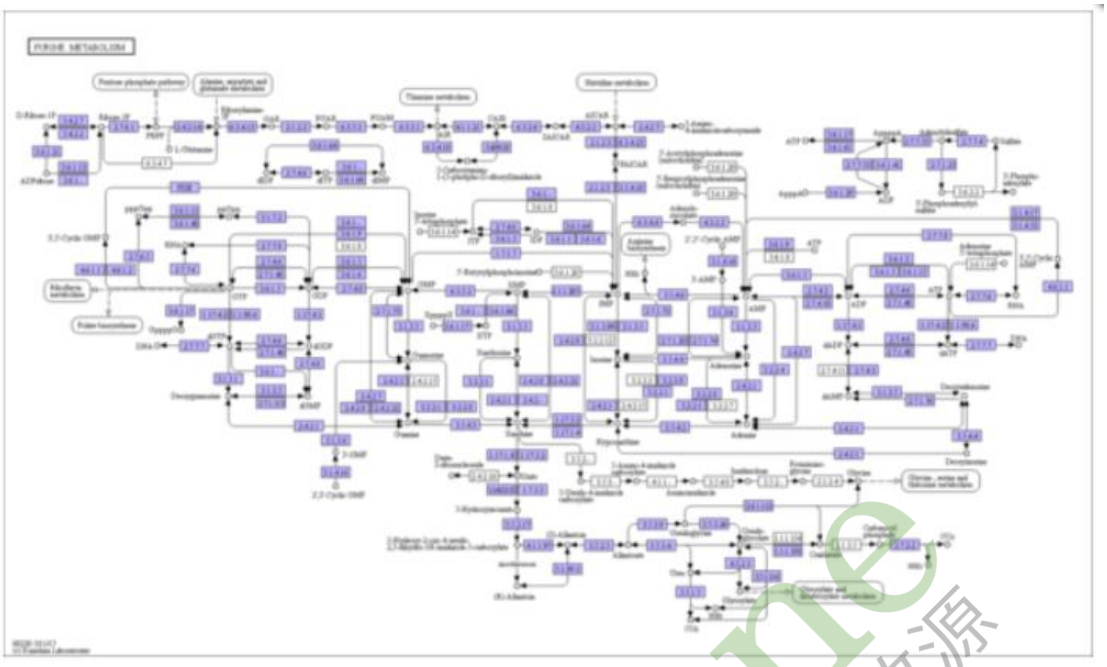


图 2.2.11.3 显著富集的 KEGG pathway 代谢通路图

其中的小方框代表蛋白，红色的小方框代表基因对应的蛋白，鼠标悬停于该节点，会弹出相应的靶基因 id。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。

2.3 参考文献

【1】Chen J Y, Peng Z, Zhang R, et al. RNA Editome in Rhesus Macaque Shaped by Purifying Selection[J]. Plos Genetics, 2014, 10(4):e1004274.

【2】Zhu S, Xiang J F, Tian C, et al. Prediction of constitutive A-to-I editing sites from human transcriptomes in the absence of genomic sequences[J]. Bmc Genomics, 2013, 14(1):206.

3 等位基因差异性表达（ASE）分析

3.1 分析背景及方法介绍

等位基因（allele 又作 allelomorph）一般指位于一对同源染色体的相同位置上控制着相对性状的一对基因。其在生物体内普遍存在差异性表达的现象。我们利用转录组测序数据鉴定的 SNP 结果，通过一系列筛选和标准化处理，最终鉴定出偏向性的结果。

此分析需要测序样本有亲代与子代信息。

3.2 结果展示

3.2.1 等位基因偏向性分析

gene_id	DY_6h_snp.gene.normalized(TW_6h)	DY_6h_snp.gene.normalized(JP_6h)	pvalue	qvalue	signature	bias	Description
evm.model.1	313,588.4	274,417.3	3.57E-20	3.62E-19	TRUE	TW_6h	-//-
evm.model.2	12,28.0	14,42.5	0.01659	0.02706	TRUE	JP_6h	-//-
evm.model.3	16,24.3	13,43.9	0.2682	0.3245	FALSE	None	sp|Q6AY55|DCAKD_RAT
evm.model.4	68,82.8	15,100.2	0.2885	0.3464	FALSE	None	sp|Q91ZN5|S35B2_MOUSE
evm.model.5	57,29.2	25,166.1	0.003353	0.006133	TRUE	TW_6h	sp|Q9H0J9|PAR12_HUMAN

表 3.2.1 等位基因偏向性分析结果

第一列：基因 ID。

第二-三列：子代（父本/母本）标准化后的 reads 数，以逗号分隔。

第四列：二项分布检验计算的 pvalue 值。

第五列：利用 BenjaminiandHochberg 的方法（简称 BH 法）计算的 qvalue 值。

第六列：偏向性指标，FALSE 为不具有偏向性，TRUE 为具有偏向性（阈值 pvalue<0.05）。

第七列：偏向性的亲本信息，none 为非偏向性表达。

第八列：基因注释信息。

3.3 参考文献

【1】刁西洲. 玉米穗部性状遗传和杂种优势分析及雌穗发育杂种优势转录组分析[D]. 中国农业科学院, 2015.

【2】翟荣荣. 超级稻协优 9308 根系杂种优势的转录组分析[D]. 中国农业科学院, 2013.

Novogene 诺禾致源

4 使用 ANNOVAR 注释 snp 位点

4.1 分析背景及介绍

目前转录的分析流程默认使用 snpeff 进行 snp 注释，我们也可以更换软件使用 ANNOVAR 注释 snp 位点（可参考重测的分析流程），流程上是可以实现对应的统一，目前仅有进行人的样本的分析。

4.2 分析结果展示

AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ
GnomAD_	INFO	FORMAT	CON_B_1	Ori_REF	Ori_ALT	shared_ho	shared_he	OMIM	GWAS_Pu	HGMD_II	GO_BP	GO_CC	GO_MF	KEGG_P/	PID_PATI	BIOCART	REACTOME_PATHWAY			
0.3725	AC=1;AF=	GT:AD:DI	0/1:12,6:18	CT	C	0	1	.	NA	.	.	.	.	.	.	.	.			
0.1168	AC=2;AF=	GT:AD:DI	1/1:0,8:8:2	A	AT	1	0	.	NA	.	.	.	.	.	.	.	.			
0.2957	AC=1;AF=	GT:AD:DI	0/1:3,7:10:	G	GGCT	0	1	.	NA	.	.	.	.	.	.	.	.			
0.3511	AC=1;AF=	GT:AD:DI	0/1:14,12:2	CTTAT	C	0	1	.	NA	.	.	.	.	.	.	.	.			
0.1852	AC=2;AF=	GT:AD:DI	1/1:0,4:4:1	AAC	A	1	0	.	NA	.	.	.	.	.	.	.	.			
0.5561	AC=1;AF=	GT:AD:DI	0/1:3,3:6:9	G	GCGC	0	1	.	NA	.	.	.	.	KEGG_U	.	.	REACTOME_IMMUNE_SYSTEM;RE/				
0.8287	AC=2;AF=	GT:AD:DI	1/1:0,10:1(	GTC	G	1	0	.	NA	.	.	.	.	.	.	.	.			
0.8853	AC=2;AF=	GT:AD:DI	1/1:0,27:27	GGA	G	1	0	.	NA	.	.	.	.	.	.	.	.			
0.4682	AC=2;AF=	GT:AD:DI	1/1:0,10:1(	C	CTA	1	0	.	NA	.	.	.	.	.	.	.	.			
0.8839	AC=2;AF=	GT:AD:DI	1/1:0,57:57	GGC	G	1	0	.	NA	.	.	.	.	.	.	.	.			
0.6149	AC=1;AF=	GT:AD:DI	0/1:65,64:1(	C	CT	0	1	.	NA	.	.	.	.	.	.	.	.			
0.1021	AC=2;AF=	GT:AD:DI	1/1:0,420:4	CAGG	C	1	0	.	NA	.	REGULA	CYTOPLA	NUCLEO	.	PID_AR_	IBIOCART	.				
0.161	AC=2;AF=	GT:AD:DI	1/1:0,4:4:1	T	TGGGA	1	0	.	NA	.	.	.	.	.	.	.	.				
0	AC=1;AF=	GT:AD:DI	0/1:7,9:16:	C	CA	0	1	.	NA	.	.	CYTOPLA	PHOSPH(	KEGG_N	.	.	REACTOME_METABOLISM_OF_VIT				
0.8383	AC=2;AF=	GT:AD:DI	1/1:0,17:17	A	AAC	1	0	.	NA	.	.	CYTOPLA	PHOSPH(	KEGG_N	.	.	REACTOME_METABOLISM_OF_VIT				
0.4399	AC=1;AF=	GT:AD:DI	1/2:2,15,1(	C	CT,CTT	0	1	Shprintzen	NA	.	.	.	.	.	PID_SMA	.	REACTOME_SIGNALING_BY_BMP;I				
0.3923	AC=1;AF=	GT:AD:DI	1/2:2,15,1(	C	CT,CTT	0	1	Shprintzen	NA	.	.	.	.	.	PID_SMA	.	REACTOME_SIGNALING_BY_BMP;I				
0.4054	AC=1;AF=	GT:AD:DI	0/1:9,8:17:	T	TTA	0	1	.	NA	.	.	.	.	KEGG_P/	.	.	REACTOME_VITAMIN_B5_PANTOT				
0.3629	AC=1;AF=	GT:AD:DI	0/1:103,79	CCAAA	C	0	1	.	NA	.	.	.	.	.	.	.	.				

AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	
CADD	NovoDb_\	NovoDb_\	GnomAD_	GnomAD_	GnomAD_	GnomAD_	INFO	FORMAT	CON_B_1	Ori_REF	Ori_ALT	shared_ho	shared_he	OMIM	GWAS_Pu	HGMD_II	GO_BP	GO_CC	GO_MF	KEGG_P/	PII
.	.	.	.	.	0.3725	0.3725	AC=1;AF=	GT:AD:DI	0/1:12,6:18	CT	C	0	1	.	NA	.	.	.	.	.	.
-1.119127,(	.	0.0098921	.	.	0.277	0.1168	AC=2;AF=	GT:AD:DI	1/1:0,8:8:2	A	AT	1	0	.	NA	.	.	.	.	.	.
-0.333115,2	0.0061192	0.209507	.	.	0.3187	0.2957	AC=1;AF=	GT:AD:DI	0/1:3,7:10:	G	GGCT	0	1	.	NA	.	.	.	.	.	.
-0.084652,5	0.3403131	0.3626761	.	.	0.6198	0.3511	AC=1;AF=	GT:AD:DI	0/1:14,12:2	CTTAT	C	0	1	.	NA	.	.	.	.	.	.
-0.437370,1	0.2220682	0.1335807	0.1583	0.1561	0.1852	0.1852	AC=2;AF=	GT:AD:DI	1/1:0,4:4:1	AAC	A	1	0	.	NA	.	.	.	.	.	.
-0.334254,2	0.4088063	.	.	.	0.5561	0.5561	AC=1;AF=	GT:AD:DI	0/1:3,3:6:9	G	GCGC	0	1	.	NA	.	.	.	.	KEGG_U	.
-0.649965,1	0.6041261	.	.	.	0.8369	0.8287	AC=2;AF=	GT:AD:DI	1/1:0,10:1(	GTC	G	1	0	.	NA	.	.	.	.	.	.
0.159666,4	0.8688302	0.8573944	0.8848	0.8848	0.8853	0.8853	AC=2;AF=	GT:AD:DI	1/1:0,27:27	GGA	G	1	0	.	NA	.	.	.	.	.	.
-0.103503,3	0.4023865	0.4434629	.	.	0.4682	0.4682	AC=2;AF=	GT:AD:DI	1/1:0,10:1(	C	CTA	1	0	.	NA	.	.	.	.	.	.
1.663711,1	0.857254	0.8280423	.	.	0.8839	0.8839	AC=2;AF=	GT:AD:DI	1/1:0,57:57	GGC	G	1	0	.	NA	.	.	.	.	.	.
0.036462,4	.	0.623227	.	.	0.9491	0.6149	AC=1;AF=	GT:AD:DI	0/1:65,64:1(	C	CT	0	1	.	NA	.	.	.	.	.	.
-0.396888,2	0.0030908	0.090669	.	.	0.1547	0.1021	AC=2;AF=	GT:AD:DI	1/1:0,420:4	CAGG	C	1	0	.	NA	.	REGULA	CYTOPLA	NUCLEO	.	PII
-0.634938,1	0.0014368	0.013465	.	.	0.2232	0.161	AC=2;AF=	GT:AD:DI	1/1:0,4:4:1	T	TGGGA	1	0	.	NA	.	.	.	.	.	.
-0.082131,3	.	.	.	.	0.0001	0	AC=1;AF=	GT:AD:DI	0/1:7,9:16:	C	CA	0	1	.	NA	.	.	CYTOPLA	PHOSPH(	KEGG_N	.
0.280366,5	0.8386594	0.8213028	.	.	0.9197	0.8383	AC=2;AF=	GT:AD:DI	1/1:0,17:17	A	AAC	1	0	.	NA	.	.	CYTOPLA	PHOSPH(	KEGG_N	.
-0.002503,4	0.0003956	0.0617284	.	.	0.5374	0.4399	AC=1;AF=	GT:AD:DI	1/2:2,15,1(	C	CT,CTT	0	1	Shprintzen	NA	.	.	.	.	.	PII
-0.011167,5	0.0377769	0.2636084	.	.	0.4871	0.3923	AC=1;AF=	GT:AD:DI	1/2:2,15,1(	C	CT,CTT	0	1	Shprintzen	NA	.	.	.	.	.	PII
-0.097167,(	0.0156746	.	.	.	0.4054	0.4054	AC=1;AF=	GT:AD:DI	0/1:9,8:17:	T	TTA	0	1	.	NA	.	.	.	.	KEGG_P/	.

P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI
cytoBand	targetScan	tfbsConsSi	genomicSu	Repeat	avsnp150	cosmic82	clinvar_20	gwasCatal	1000g2015	1000g2015	1000g2015	esp6500siv	ExAC_AL	ExAC_EA	SIFT	Polyphen2	Polyphen2	MutationT	gerp++gt2
1p36.33	.	.	Score=0.9	Score=218	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1p36.33	.	.	Score=0.9	Score=151	rs77636979	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1p36.33	.	.	Score=0.9	Score=340	rs14258382	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1p36.33	.	.	.	Score=251	rs14254542	.	.	.	.	0.3512	0.370607	.	.	.	.	.	.	.	.
1p36.33	.	.	.	Score=337	rs11293621	ID=COSN	.	.	.	0.1508	0.121206	.	0.1114	0.1923	.	.	.	.	.
1p36.33	.	.	.	.	rs14236611	.	.	.	.	0.5952	0.309305	.	.	.	.	.	.	.	.
1p36.33	.	.	.	.	rs11153079	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1p36.33	.	.	.	.	rs34091022	.	.	.	.	0.869	0.370807	.	0.3323	0.8939	.	.	.	.	.
1p36.33	.	.	Score=0.9(	.	rs36115387	.	.	.	.	0.4871	0.227037	.	.	.	.	.	.	.	.
1p36.33	.	.	.	.	rs15073703	.	.	.	.	0.869	0.66893	.	.	.	.	.	.	.	.
1p36.33	.	.	Score=0.9	.	rs5772057	ID=COSN	.	.	.	0.5843	0.714657	.	.	.	.	.	.	.	.
1p36.33	.	.	Score=0.9	.	rs14313449	.	.	.	.	0.0853	0.0539137	.	.	.	.	.	.	.	.
1p36.33	.	.	Score=0.9	Score=162	rs3044758	.	.	.	.	0.005	0.009385	.	.	.	.	.	.	.	.
1p36.33	.	.	Score=0.9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1p36.33	.	.	.	.	rs35357728	.	.	.	.	0.8393	0.714457	.	.	.	.	.	.	.	.
1p36.33	.	.	.	.	rs11241321	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1p36.33	.	.	.	.	rs11241321	ID=COSN	.	.	.	.	.	.	.	.	.	.	.	.	.
1p36.32	.	Score=100	.	Score=252	rs3831090	.	.	.	.	0.4385	0.315895	.	.	.	.	.	.	.	.
1p36.32	.	.	.	Score=175	rs13975764	.	.	.	.	0.3313	0.213259	.	.	.	.	.	.	.	.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
CHROM	POS	ID	REF	ALT	QUAL	FILTER	GeneNam	Func	Gene	GeneDeta	ExonicFun	AAChange	Genecode	wgRna	cytoBand
1	128595	.	CT	C	42.73	PASS	.	intergenic	NM_00100	dist=58588	.	.	ENST0000	.	1p36.33
1	673034	rs77636975	A	AT	261.75	PASS	.	intergenic	NR_02832	dist=7303;	.	.	ENST0000	.	1p36.33
1	713786	rs14258382	G	GGCT	247.73	PASS	LOC10028	ncRNA_ex	NR_03390	.	.	.	ENST0000	.	1p36.33
1	900717	rs14254543	CTTAT	C	424.73	PASS	KLHL17	UTR3	NM_19831	NM_19831	.	.	ENST0000	.	1p36.33
1	1158562	rs11293621	AAC	A	137	PASS	SDF4	intronic	NM_01617	.	.	.	ENST0000	.	1p36.33
1	1209183	rs14236611	G	GCGC	78.73	PASS	UBE2J2	UTR5	NM_05816	NM_05816	.	.	ENST0000	.	1p36.33
1	1323945	rs11153079	GTC	G	283.73	PASS	CCNL2	intronic	NM_03093	.	.	.	ENST0000	.	1p36.33
1	1336136	rs34091025	GGA	G	1178.73	PASS	LOC14841	ncRNA_ex	NR_01543	.	.	.	ENST0000	.	1p36.33
1	1424958	rs36115387	C	CTA	412.74	PASS	ATAD3B	intronic	NM_03192	.	.	.	ENST0000	.	1p36.33
1	1510035	rs15073705	GGC	G	2512.73	PASS	SSU72	UTR5	NM_01418	NM_01418	.	.	ENST0000	.	1p36.33
1	1594199	rs5772057	C	CT	1966.73	PASS	SLC35E2B	UTR3	NM_00111	NM_00111	.	.	ENST0000	.	1p36.33
1	1630751	rs14313449	CAGG	C	18813.7	PASS	CDK11B	intronic	NM_03348	.	.	.	.	.	1p36.33
1	1661437	rs3044758	T	TGGGA	142	PASS	SLC35E2	UTR3	NM_18283	NM_18283	.	.	ENST0000	.	1p36.33
1	1683204	.	C	CA	176.73	PASS	NADK	UTR3	NM_00111	NM_00111	.	.	ENST0000	.	1p36.33
1	1690126	rs35357728	A	AAC	714.73	PASS	NADK	intronic	NM_00111	.	.	.	ENST0000	.	1p36.33

Novogene 诺禾致源

5 CARD 数据库抗性基因注释（原核常用）

5.1 分析背景及介绍

随着抗生素药物的发现及使用，越来越多的耐药菌株由此产生。而耐药菌株的发展则会增加疾病治疗的难度和成本，因此耐药微生物的研究则显得尤为重要。目前，通过对耐药基因的鉴定挖掘能够一定程度上帮助我们揭开耐药机制，为疾病的治疗、药物研发提供参考。ARDB 是最先整合了各种微生物中抗药基因的数据库，但它从 2009 年开始就不再更新。而 CARD(the Comprehensive Antibiotic Research Database)数据库包含了 ARDB 数据库中所有抗性信息，并搭建了一个基于志愿者贡献的数据共享平台，做到了实时更新保证了数据的有效性。目前，CARD 数据库收集了超过 1600 个已知的抗生素抗性基因，是一个全面的抗生素耐药性数据库。

CARD 数据库(http://arpcard.mcmaster.ca) 是经过同行评议的耐药性决定因素和相关抗生素的严格收集的，由抗生素耐药性本体论（ARO）和 AMR 基因检测模型进行组织。CARD 数据库下载之后，可以通过 blastp 将基因序列比对到 CARD 数据库，从而进行抗性基因的注释。如果给出的是组装后的 scaffold 序列，则需要先进行基因预测，然后将预测得到的蛋白序列比对到 CARD 数据库，从而获得抗性基因。

5.2 结果展示

Gene_id	hit_id	Name	Description
b0026	ARO:3000521	Staphylococcus mupA conferring resistance to mupirocin	An alternative isoleucyl-tRNA synthetase conferring resistance to mupirocin.
b0048	ARO:3003105	dfrA3	dfrA3 is an integron-encoded dihydrofolate reductase found in Escherichia coli
b0051	ARO:3000600	Erm(34)	ErmD confers MLSb phenotype.
b0092	ARO:3002909	vanG	VanG is a D-Ala-D-Ala ligase homolog that can synthesize D-Ala-D-Ser, an alternative substrate

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

			for peptidoglycan synthesis that reduces vancomycin binding affinity in Enterococcus faecalis
b0127	ARO:3003748	oleC	oleC is an ABC transporter isolated from Streptomyces antibioticus and is involved in oleandomycin secretion.
b0254	ARO:3000620	adeL	AdeL is a regulator of AdeFGH in Acinetobacter baumannii. AdeL mutations are associated with AdeFGH overexpression and multidrug resistance.
b0262	ARO:3003748	oleC	oleC is an ABC transporter isolated from Streptomyces antibioticus and is involved in oleandomycin secretion.

表 5.2　CARD 数据库抗性基因注释表

Gene_id: 比对的基因 ID；

Hit_id：数据库中对应的基因 ID；

Name：数据库中对应的基因名称；

Description：数据库中关于基因的描述信息

5.3 参考文献

[1] McArthur,A.G., Waglechner,N., Nizam,F., Yan,A., Azad,M.A., Baylay,A.J., Bhullar,K., Canova,M.J., de Pascale,G., Ejim,L. et al. (2013) The comprehensive antibiotic resistance database. Antimicrob. Agents Chemother., 57, 3348–3357.

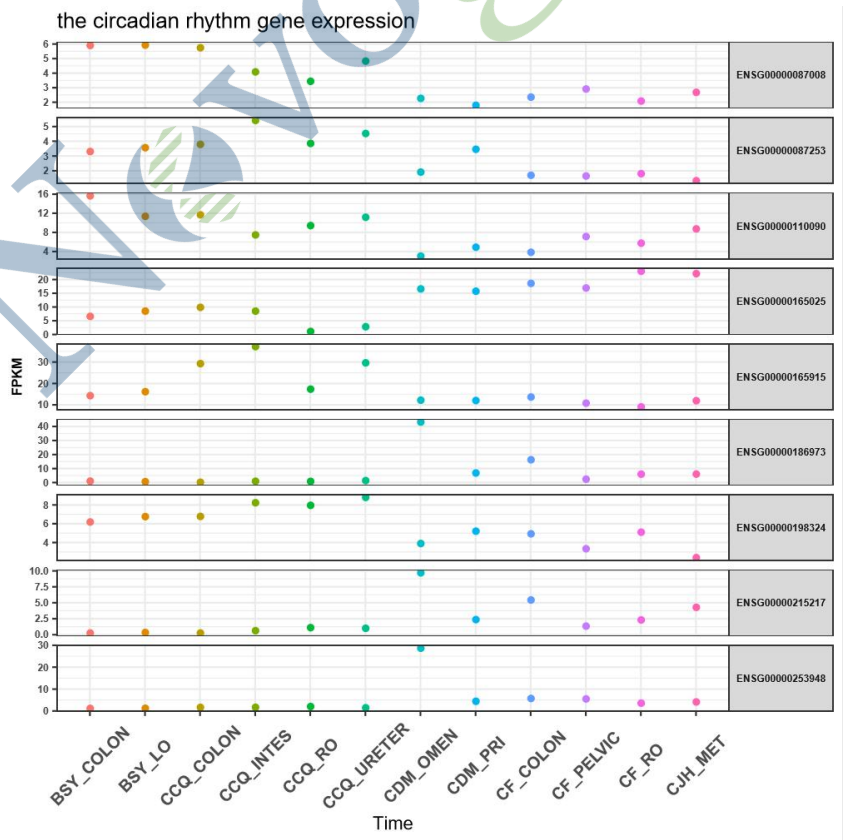
6 利用 Metacycle 进行时序性分析

生物学研究中我们往往处理的数据是具有时序性的，这些数据往往随着时间变化随之发生变化，例如最常见的时序表达谱数据就属于这种类型，这种数据包含着丰富的基因调控信息，当样品为多个时间点时，我们更可能关心基因在多个时间点上的变化规律。通过总体聚类，可以挑选出那些显著周期性波动，符合时间规律的基因，如下图所示，可以看到这些基因都是非常符合周期性变化的，也是潜在的节律相关的调控基因。后续可以结合功能注释，做更深一步的讨论和研究。

6.1 分析方法

使用 Metacycle；输入文件： fpkm 文件

6.2 分析结果



北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

横坐标是时间点（样本名称），右边是基因名称，左侧是基因表达量，每个点表示每个基因在每个时间点中的表达量，通过该图，可以看出基因的变化趋势。

6.3 参考文献

Gang W，Anafi R C，Hughes M E，et al. MetaCycle: an integrated R package to evaluate periodicity in large scale data[J]. Bioinformatics(21):3351-3353.

Novogene 诺禾致源

7 利用 Mfuzz R 包进行时序性分析

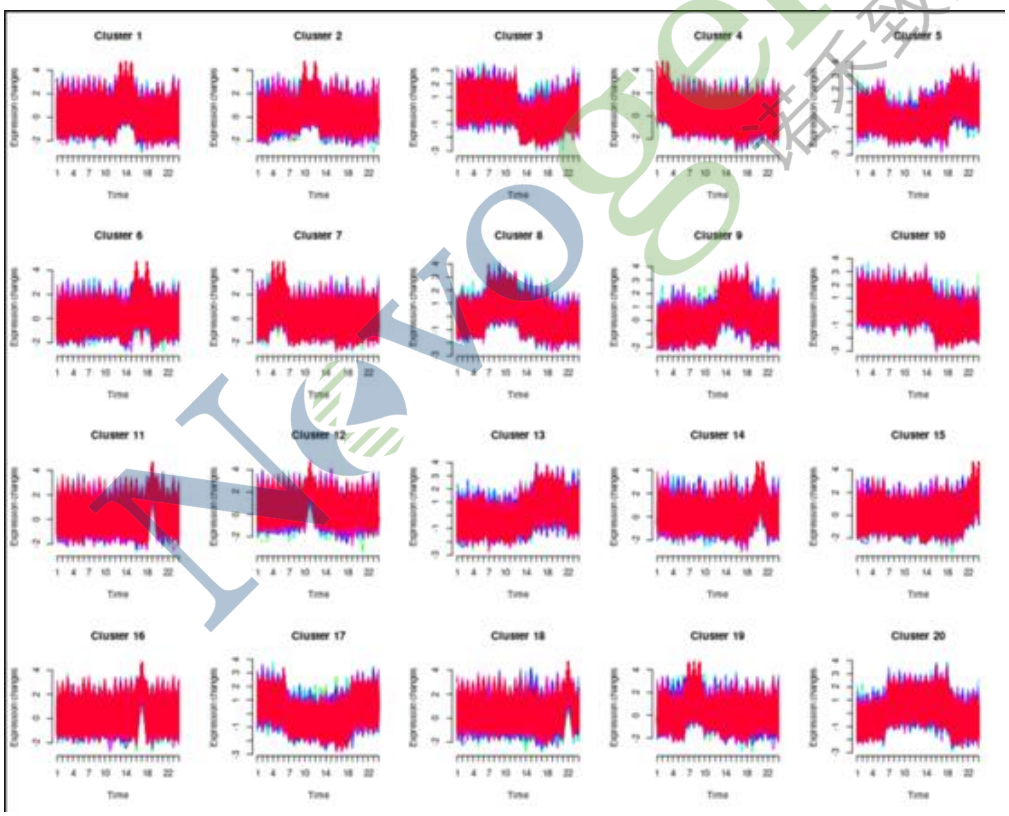
7.1 分析背景及介绍

Mfuzz 能够识别表达谱的潜在时间序列模式，并将相似模式的基因聚类，以帮助我们了解基因的动态模式和它们功能的联系。

7.2 分析使用数据

基因表达量数据

7.3 结果展示



8 利用 maSigPro 包进行时间序列分析

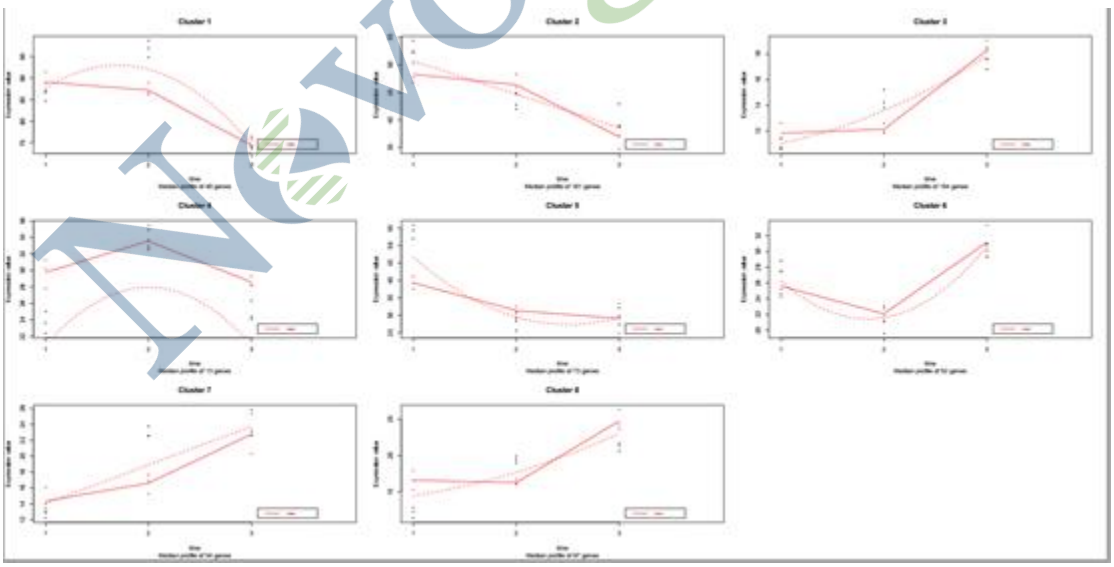
8.1 分析背景及介绍

maSigPro，这是一款可以综合分析时间序列 RNA-seq 的工具。这款工具最初由 Conesa 等人于 2006 年开发，于 2014 年经 Nueda 等人进一步优化而得到的。那么，我们在研究生物学问题的时候，往往是一个动态的过程，即某些基因的表达水平随着时间的推移厄尔发生动态改变。而传统的 DEG 往往只能比 treatment 和 control，却忽略了时间带来的效应。maSigPro 综合考虑了每一时期的 treatment 与 control 的区别，而且也考虑了不同时间点基因表达的动态变化。

8.2 分析使用数据

基因表达量数据

8.3 结果展示



北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

9 利用 TCseq 包进行基因表达趋势分析

9.1 分析背景及介绍

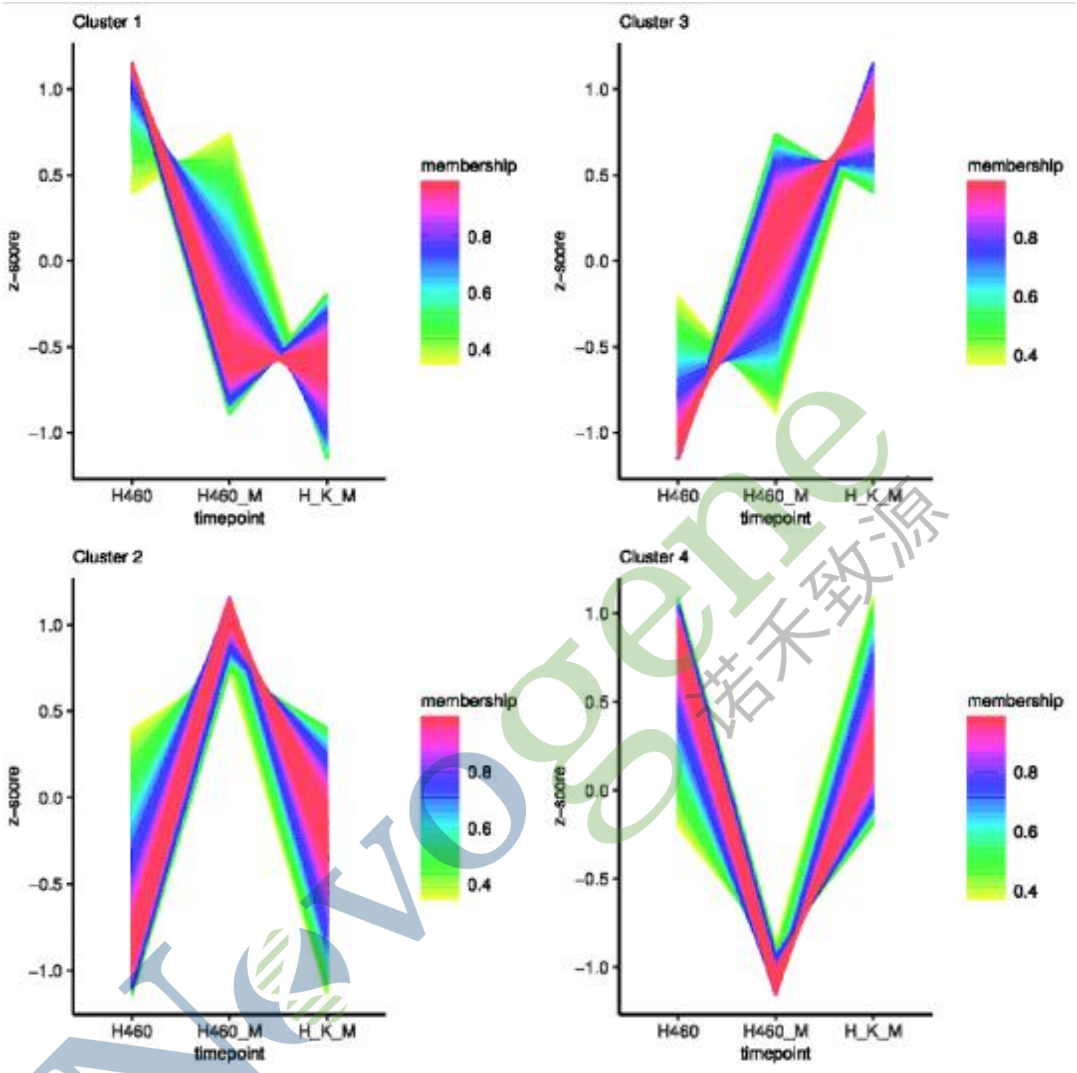
TCseq 包提供了一个统一的套件去处理不同时序类型的数据分析，可以应用于转录组或者像 ATAC-seq、Chip-seq 的表观基因组时序型数据分析。该包主要的集中于不同时间点的差异分析，时间趋势分析及可视化作图。

9.2 分析使用数据及方法

使用基因表达量数据即可进行分析

Novogene 诺禾致源

9.3 结果展示



Novogene 诺禾致源

10 多组合 TCC 差异分析

一般来说，差异基因分析 2 个样品间的差异，而 Evaluation of methods for differential expression analysis on multi-group RNA-seq count data,这篇文献不是比较两组测序数据，而是专注于三组数据之间的差异分析。文献得出结论：TCC 包中基于 DEGES 的流程能有效地对三组数据进行差异分析，其中有（2-6 组生物学重复）的使用 edgeR 的 DEGES 流程(EEE-E)，没有生物学重复的使用 DESeq2 的 DEGES 流程(SSS-S)。

10.1 分析方法

方法：使用 R 语言 TCC 包

输入文件：原始 readcount 文件

10.2 分析结果

结果展示：

gene_id	pvalue	FDR	rank	estimatedDEG
c607819_g1	4.81E-29	3.95E-24	1	1
c574481_g1	4.85E-22	1.99E-17	2	1
c544039_g1	3.98E-20	1.09E-15	3	1
c574481_g2	1.60E-19	3.29E-15	4	1
c150686_g1	7.87E-19	1.29E-14	5	1

gene_id:基因 id

pvalue:统计学差异显著性检验指标

FDR:校正后的 pvalue

rank:排序值

estimatedDEG:是否为差异基因 1:是 0:否

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

10.3 参考文献

Sun.et al. TCC: Differential expression analysis for tag count data with robust normalization strategies

Tanget al. Evaluation of methods for differential expression analysis on multi-group RNA-seq count data

Novogene 诺禾致源

11 差异基因进行双因素方差分析

11.1 分析背景及介绍

涉及自变量和因变量，研究是一个因素起作用还是两个因素都起作用的方差分析如下数据展示，A or O 是自变量，时间是因变量。

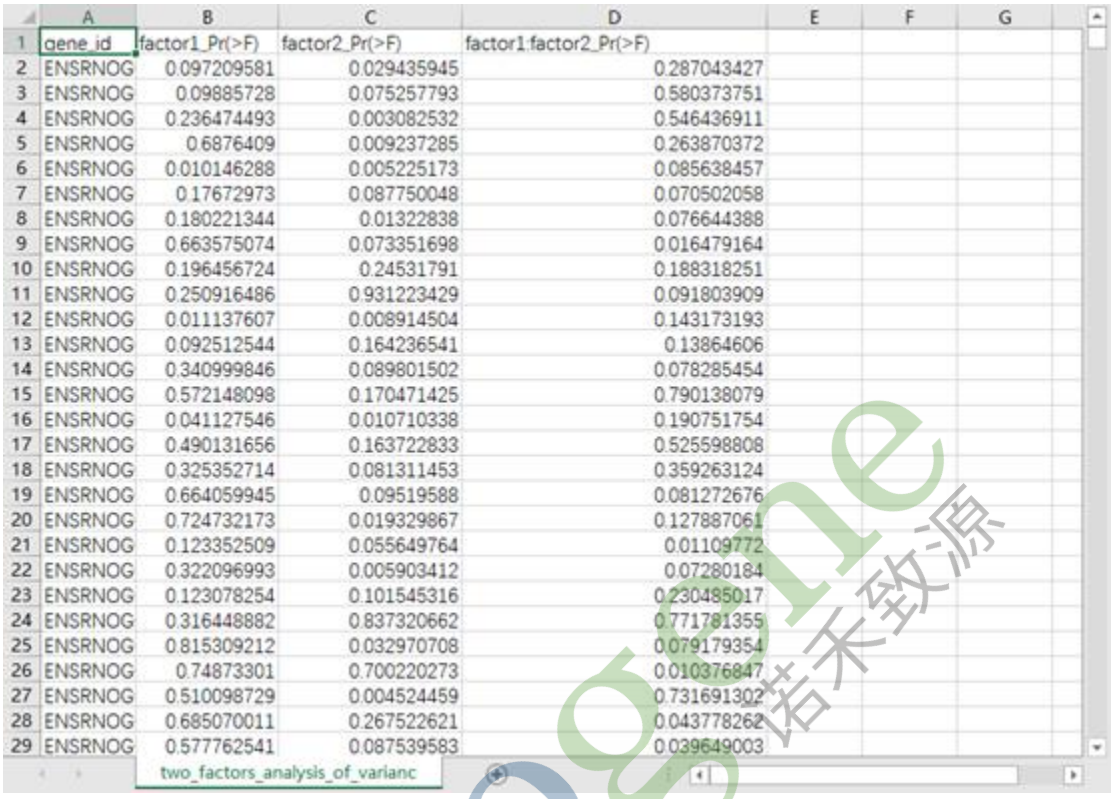
11.2 分析使用数据

数据格式如下：

	A	B	C	D	E
1	gene_ID	A or O	time	FPKM	
2	1	A1	1	1	
3	1	O1	1	1	
4	1	A3	3	1	
5	1	O3	3	1	
6					
7					

Novogene 诺禾致源

11.3 结果展示



	A	B	C	D	E	F	G
1	gene_id	factor1_Pr(>F)	factor2_Pr(>F)	factor1:factor2_Pr(>F)			
2	ENSRNOG	0.097209581	0.029435945	0.287043427			
3	ENSRNOG	0.09885728	0.075257793	0.580373751			
4	ENSRNOG	0.236474493	0.003082532	0.546436911			
5	ENSRNOG	0.6876409	0.009237285	0.263870372			
6	ENSRNOG	0.010146288	0.005225173	0.085638457			
7	ENSRNOG	0.17672973	0.087750048	0.070502058			
8	ENSRNOG	0.180221344	0.01322838	0.076644388			
9	ENSRNOG	0.663575074	0.073351698	0.016479164			
10	ENSRNOG	0.196456724	0.24531791	0.188318251			
11	ENSRNOG	0.250916486	0.931223429	0.091803909			
12	ENSRNOG	0.011137607	0.008914504	0.143173193			
13	ENSRNOG	0.092512544	0.164236541	0.13864606			
14	ENSRNOG	0.340999846	0.089801502	0.078285454			
15	ENSRNOG	0.572148098	0.170471425	0.790138079			
16	ENSRNOG	0.041127546	0.010710338	0.190751754			
17	ENSRNOG	0.490131656	0.163722833	0.525598808			
18	ENSRNOG	0.325352714	0.081311453	0.359263124			
19	ENSRNOG	0.664059945	0.09519588	0.081272676			
20	ENSRNOG	0.724732173	0.019329867	0.127887061			
21	ENSRNOG	0.123352509	0.055649764	0.01109772			
22	ENSRNOG	0.322096993	0.005903412	0.07280184			
23	ENSRNOG	0.123078254	0.101545316	0.230485017			
24	ENSRNOG	0.316448882	0.837320662	0.771781355			
25	ENSRNOG	0.815309212	0.032970708	0.079179354			
26	ENSRNOG	0.74873301	0.700220273	0.010376847			
27	ENSRNOG	0.510098729	0.004524459	0.731691302			
28	ENSRNOG	0.685070011	0.267522621	0.043778262			
29	ENSRNOG	0.577762541	0.087539583	0.039649003			

two_factors_analysis_of_varianc

factor1 代表 A or O

factor2 代表时间

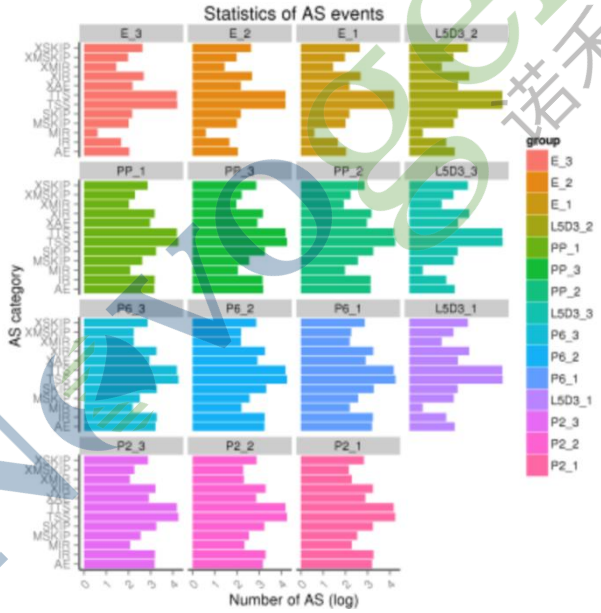
Novogene 诺禾致源

12 ASprofile 可变剪切分析

12.1 分析背景及方法介绍

目前流程中默认的可变剪切分析软件是 rMARTs，该款软件分析的是差异可变剪切。而 ASprofile 可以针对单个样品进行可变剪切分析，软件直接将同一个基因的多个转录本进行比较，进一步鉴定可变剪切事件。ASprofile 软件对 Cufflinks 预测出的基因模型按照单个实验组中所有重复样品的数据和所有实验样品的数据进行分组，并对每组的可变剪切事件分别进行分类和表达量统计。

12.2 结果展示图



软件可预测出 12 类可变剪切事件，分别是：1) TSS: Alternative 5' first exon (transcription start site) 第一个外显子可变剪切；2) TTS: Alternative 3' last exon (transcription terminal site) 最后一个外显子可变剪切；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,

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

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'端剪切；12) XAE: Approximate AE 可变 5'或 3'端剪切（模糊边界）。

12.3 *.fpkm.xls 为样本可变剪切结构及表达量统计表

event_id	event_type	gene_id	chrom	event_start	event_end	event_pattern	strand	fpkm	ref_id
1000001	TSS	CUFF.1	1	11869	12227	12227	+	0	-
1000002	TSS	CUFF.1	1	12010	12057	12057	+	0	-
1000003	TTS	CUFF.1	1	13453	13670	13453	+	0	-
1000004	XSKIP_ON	CUFF.1	1	12975	13052	12697,12975-130	+	0	-
1000005	XSKIP_OFF	CUFF.1	1	12975	13052	1,272,113,221	+	0	-
1000006	XIR_ON	CUFF.1	1	12057	12179	11869-12227	+	0	-
1000007	XIR_OFF	CUFF.1	1	12057	12179	12010-12057,121	+	0	-
1000008	TSS	CUFF.10	1	960587	960800	960800	+	0.70407945	-
1000009	TSS	CUFF.10	1	960639	960699	960699	+	0	-
1000010	TSS	CUFF.10	1	961449	961750	961750	+	0	-
1000011	TSS	CUFF.10	1	963552	964008	964008	+	0	-
1000012	TSS	CUFF.10	1	962694	962917	962917	+	0.67421621	-
1000013	TTS	CUFF.10	1	964963	965715	964963	+	1.37829565	-
1000014	XMSKIP_ON	CUFF.10	1	961293	962047	960800,961293-9	+	0.70407945	-
1000015	XMSKIP_OFF	CUFF.10	1	961293	962047	960,699,962,082	+	0	-
1000016	XIR_ON	CUFF.10	1	961552	961629	961449-961750	+	0	-
1000017	XIR_OFF	CUFF.10	1	961552	961629	961293-961552,9	+	0.70407945	-

注释：

(1) event_id：AS 事件编号

(2) event_type ： AS 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)

(3) gene_id：cufflink 组装结果中的基因编号

(4) chrom：染色体编号

(5) event_start：AS 事件起始位置

(6) event_end：AS 事件结束位置

(7) event_pattern：AS 事件特征 (for TSS, TTS - inside boundary of alternative marginal exon; for *SKIP_ON,the coordinates of the skipped exon(s); for *SKIP_OFF, the coordinates of the enclosing introns; for *IR_ON, the end coordinates of the long, intron-containing exon; for *IR_OFF, the listing of coordinates of all the exons along the path containing the retained intron; for *AE, the coordinates of the exon variant)

(8) strand：基因正负链信息

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

(9) fpkm：此 AS 类型该基因表达量

(10) ref_id：参考基因组 id

12.4 参考文献

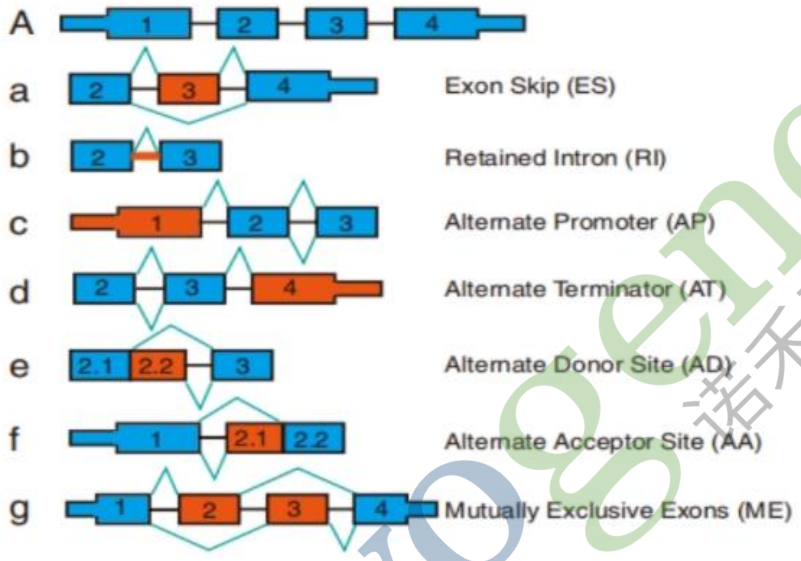
Thousands of exon skipping events differentiate splicing patterns in sixteen human tissues.

Novogene 诺禾致源

13 SpliceSeq 可变剪切分析

13.1 分析背景及介绍

Spliceseq 软件得到 7 种可变剪切类型：在 5 分类的基础上还包括可变性供体位点和可变性受体位点。



ES：外显子跳跃；RI：内含子滞留；AP：可变启动子；AT：可变终止子；AD：可变供体位点；AA：可变受体位点；ME 外显子互斥。

Spliceseq 是 MDAnderson cancer center 开发的，基于 java 探究高通量 RNA-seq 数据可变性剪切模式的软件。测序数据进行比对构建剪切图谱，以量化每个外显子和剪接点的表达水平，然后通过剪切图谱可以预测可变剪切事件翻译成的蛋白质，最后通过 UniProt 数据库的比对注释结果中的蛋白质亚型以识别可变剪接的潜在功能影响。

该工具可用于单个 RNASeq 样本以识别具有多种剪接形式的基因，也可用于一对样本以识别两者之间的差异剪接，或用于样本组以识别剪接模式中具有统计学意义的组间差异。

SpliceSeq 可以从安装界面作为 java web start 应用程序运行分析该数据库中的测序数据，也可以安装在本地分析自己的 mRNA-Seq 数据。

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

13.2 结果展示

13.2.1 单个样本

8.2.1.1 Splice_Event：

	A	B	C	D	E	F	G	H	I	J	K
1	# View type: Sample View										
2	# Sample: T1_work1_1_P1										
3	# Filters: no filter										
4	Gene Symb	Gene RPKM	Splice Type	Exons	Upstream	Downstrea	Included C	Excluded C	PSI	Magnitude	
5	Mbp	1233.34	AT	11			55062	58	0.999586	1	
6	Tpt1	4214.19	ES	4:05	3	6	17536	4	0.99922	0.993514	
7	Ube2i	88.7772	AT	10			996	4	0.998765	1	
8	Atp6v0e	151.425	AT	4			1072	8	0.99858	1	
9	Nhp2l1	116.051	AT	3			4254	5	0.99856	1	
10	Ptma	2787.53	ES	3:04	2	5	12518	7	0.998404	1	
11	Trf	2531.39	AT	17:2			15001	19	0.998391	1	
12	Ywhaq	2364.85	ES	5	4	6	8791	7	0.998375	1	
13	Ttr	33809	ES	3	2	4	163795	109	0.998354	1	
14	Auh	59.3116	AT	13			366	7	0.996849	0.976233	
15	Erp29	43.7182	AT	4			1332	3	0.996693	0.48297	
16	Trf	2531.39	ES	2:2:3:4:5:6:	2:1	10:2	35130	41	0.996145	0.920858	

- ✓ Gene Symbol：对应发生可变剪切的基因名。
- ✓ Gene RPKM：基因的表达量，RPKM 值
- ✓ Splice Type：展示可变剪切的类型，ES、AT、AP 等等
- ✓ Exons：展示发生可变剪切的外显子位点，外显子 1、2、3 等等
- ✓ Upstream Exon：发生可变剪切的外显子的上游外显子位点
- ✓ Downstream Exon：发生可变剪切的外显子的下游外显子位点
- ✓ Included Counts：含有可变剪切外显子的 reads
- ✓ Excluded Counts：不含有可变剪切外显子的 reads
- ✓ PSI：PSI 指数（percent spliced in index)，它是可以量化可变剪切的指标，可以实现某个外显子是否纳入转录本定量，总而用于单个样本或组间多个样本的比较。其计算公式可总结成如下：PSI = splicein / (splicein +splice_out)
- ✓ magnitude：The proportion of isoforms which the splice event affects

8.2.1.2 Gene_Summary

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

View type: Sample View
Sample: T1_work1_1_P1
Filters: no filter

Gene Symb	Total Reac	Gene RPKM	Sum Minor	Splice Even	UniProt Events
0610007P1	667	39.8547	0		
0610009B2	203	17.7362	0		
0610009D(	667	53.1551	0		
0610009O:	578	11.4297	0		
0610010F(	2	0.06629	0		
0610010K1	20	3.68728	0		
0610011F(	632	48.7661	0	RI (5.2)	
0610030E2	135	3.43404	0		
0610031J0	2794	117.041	0	ES (4), AD (2.2)	
0610037L1	186	2.49337	0		
0610037L1	186	2.49337	0		
0610040J0	7	0.315904	0		
1110001A:	77	2.65033	0		
1110001J0	75	8.80292	0		
1110002L(	31	0.134901	0		

- ✓ Gene Symbol：对应发生可变剪切的基因名。
- ✓ Total Reads：基因的 reads 数目。
- ✓ Gene RPKM：基因的表达量，RPKM 值。
- ✓ Sum Minor Isoform % ： Percentage of predicted mRNA expression that is not part of the major isoform. This can be used to find genes where alternative isoforms constitute more than a trivial percentage of gene expression.
- ✓ Splice Events -：检测到的可变剪接事件的分类列表。例如，ES（4）表示外显子 4，外显子跳过
- ✓ UniProt Event ： Categorized list of UniProt annotated regions that are affected by alternate splicing. For example， B(9) indicates a Bind Site on exon 9 that is excluded from some isoforms. See the legend in the Protein Sequence Tab of the Gene Panel for a full list of UniProt annotation categories

13.2.2 组

13.2.2.1 Splice_Event：

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

	A	B	C	D	E	F	G	H	I	J
1	# View type: Group View									
2	# Group: Cancer									
3	# Filters: Avg Gene RPKM>=1									
4	Gene Syml	Avg RPKM	Splice Type	Exons	Upstream	Downstrea	PSI	Magnitude	%Obs.	
5	Kxd1	403.141	AT	10			0.99994	0.982155	1	
6	Nhp2l1	576.48	AT	3			0.999932	1	1	
7	Gadl1	254.265	AT	15			0.999918	0.936112	1	
8	Atp6v0e	216.775	AT	4			0.999903	1	1	
9	Rbm3	418.97	AT	8			0.999824	0.362186	1	
0	Ldha	494.666	AT	10			0.999679	1	1	
1	Ube2i	63.6136	AT	10			0.999627	1	1	
2	Sptbn1	16.7291	AT	37			0.99959	1	1	
3	Actg1	5876.01	ES	3:2:5	3:1	6	0.999563	0.894451	1	
4	Cdk2ap2	198.723	AT	4			0.999508	1	1	
5	Sin3b	60.9983	AT	7			0.999392	0.932786	1	
6	Hnrnpk	76.0538	AT	18:2			0.999381	1	1	
7	Kif1b	15.7205	AT	50			0.999348	0.994932	1	
8	Morn1	53.466	AT	14			0.999339	1	1	
9	4933434E2	315.551	ES	5	4	6	0.999305	1	1	
0	15-Sep	476.468	ES	4	3	5	0.9993	0.970488	1	
1	Tsc22d1	15.2344	AT	8			0.999228	1	1	
2	Apoe	386.931	AT	5			0.999226	0.8335	1	
3	Tpt1	6253.14	ES	5	4	6	0.99905	0.939851	1	

- ✓ Gene Symbol：对应发生可变剪切的基因名。
- ✓ Avg RPKM：基因的平均表达量 RPKM 值。
- ✓ Splice Type：展示可变剪切的类型，ES、AT、AP 等等。
- ✓ Exons：展示发生可变剪切的外显子位点，外显子 1、2、3 等等。
- ✓ Upstream Exon：发生可变剪切的外显子的上游外显子位点。
- ✓ Downstream Exon：发生可变剪切的外显子的下游外显子位点。
- ✓ PSI：PSI 指数（percent spliced in index)：它是可以量化可变剪切的指标，可以实现某个外显子是否纳入转录本定量，总而用于单个样本或组间多个样本的比较。其计算公式可总结成如下：PSI = splicein / (splicein +splice_out)。
- ✓ magnitude：The proportion of isoforms which the splice event affects
- ✓ %Obs.： The percentage of the group expressing the given splice event. If a sample does not express a particular gene，then any event on that gene would not be observed on that sample.（特定剪切事件在组中表达的占比，如果某个基因在样本中不表达，则该样本中无法看到特定基因的任何剪切事件）

13.2.2.2 Gene_Summary

View type: Group View
Group: Cancer
Filters: Avg Gene RPKM>=1

Gene Symb	Avg Reads	Avg RPKM	RPKM Std	Num Samp	Splice Even	UniProt Events
0610007P1	1871	104.415	147.664	1		
0610009D(	1891	146.645	90.3766	2		
0610009O:	315	18.2387	25.7935	1		
0610010K1	306	4.05393	1.85629	2	AP (2.1), AP (1), AA (6.1), ES (5, 6.1), ES (4.2, 6.1), ES (5), ES (5), AD (4	
0610011FC	53	2.69753	3.12716	2	AT (6), AT (5.3), RI (5.2), ES (2, 3, 4), ES (2)	
0610030E2	472	7.34043	10.2527	2		
0610031J0	4403	166.562	19.2545	2	ES (2.2, 3, 4, 5, 6.1), ES (3, 4, 5, 6.1), ES (2.2, 3), ES (3), AD (2.2), ES (2	
0610037L1	2399	57.8033	81.7462	1	RI (7.2), RI (7.2, 7.3, 7.4, 7.5), AA (7.3, 7.4, 7.5), ES (7.2, 7.3, 7.5), ES (7	
1110001A:	211	11.7621	16.6342	1		
1110001J0	636	86.7317	108.556	2		
1110004E(	2091	112.588	152.967	2	ES (3)	
1110004F1	1981	87.2952	7.73812	2	AA (4.1), AD (1.2), ES (1.2, 2, 3)	
1110012L1	816	41.3194	58.4345	1		
1110017D:	793	91.9658	6.71881	2	ES (5)	
1110032A(	2767	98.204	17.4991	2	AP (1), AP (2), ES (3)	
1110037FC	1987	51.8439	73.3184	1	AT (14.2), AT (25)	
1110038F1	1242	60.9952	86.0382	2		
1110051M	4812	174.115	192.461	2	AT (10), AT (9), ES (2)	
1110058L1	511	46.9974	65.2918	2		
1110059G:	39	1.15113	1.62795	1	ES (2)	
1110065P2	238	37.5814	52.9568	2		

- ✓ Gene Symbol：发生可变剪切的基因名。
- ✓ Avg Reads：基因的平均 reads 数目。
- ✓ Avg RPKM：基因的平均表达量 RPKM 值。
- ✓ RPKM Std Dev：基因标准化后的表达量 RPKM 值。
- ✓ Num Sample：该可变剪接事件中组合里有表达的样本数。
- ✓ Splice Events ： Categorized list of the differential splice events detected in the group on average. For example， ES(4) indicates and exon skip on exon number 4 that occurs more frequently in one sample. Use the Splice Event Filter drop down list to see the description and codes for each event type.
- ✓ UniProt Event ： Categorized list of UniProt annotated regions that are affected by alternate splicing events in the group on average. For example， B(9) indicates a Bind Site on exon 9 that is excluded more frequently in one sample's isoforms. See the legend in the Protein Sequence Tab of the Gene Panel for a full list of UniProt annotation categories.

13.2.3 比较组合

13.2.3.1 Splice_Event：

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

1	w									
2	ncer									
3	mary only, |dPSI|>=0.1, p-value<=0.05									
4	Splice Type	Exons	Upstream	Downstrea	Direction	dPSI	p-value	Magnitude	Group1 %C	Group2 %Obs
5	RI	31.2	31.1	31.3	-	-0.99967	2.08E-04	0.792575	0.375	1
6	AT	9			+	-0.79114	4.29E-04	0.855668	1	1
7	ES	4	3	5	-	0.756215	0.002015	0.607716	0.875	1
8	AT	4			-	0.673296	0.004396	0.990179	1	1
9	AT	26.2			-	0.896196	0.004703	1	1	1
10	AT	8.3			+	-0.78712	0.005226	0.666278	0.75	1
11	RI	6.7	6.6	6.8	-	-0.63404	0.006251	0.845429	0.75	1
12	AP	2.1			-	0.445426	0.006428	0.144063	1	1
13	AT	14.2			+	-0.6596	0.0099	0.787603	1	1
14	AT	6			-	0.646811	0.010309	0.847065	1	1
15	AP	2.1			+	-0.68115	0.013951	0.15356	0.75	1
16	AT	30			-	0.597569	0.01594	0.864686	0.875	1
17	ES	4	3	6	-	0.794405	0.016134	0.976764	0.625	1
18	AP	2.1			+	-0.79581	0.016414	0.8	0.625	1
19	AA	18.1	17	18.2	-	0.329409	0.017599	1	0.875	1
20	AT	5			-	0.476388	0.018892	1	1	1
21	AT	3			-	0.662891	0.020813	1	0.75	1
22	AT	5			+	-0.27633	0.02496	1	0.875	1
23	AT	15			-	0.95322	0.031218	1	0.25	1
24	AT	10			+	-0.75947	0.04059	0.97823	0.5	1
25	ES	3.2:5:6:7.1	3.1	7.2	+	-0.21252	0.046842	0.909852	0.375	1
26										

- ✓ Splice Type：展示可变剪切的类型，ES、AT、AP 等等
- ✓ Exons：展示发生可变剪切的外显子位点，外显子 1、2、3 等等
- ✓ Upstream Exon：发生可变剪切的外显子的上游外显子位点
- ✓ Downstream Exon：发生可变剪切的外显子的下游外显子位点
- ✓ Direction：染色体的正负链信息
- ✓ dPSI ：拼接的增量百分比。第二组的平均 PSI 减去第一组的 PSI
- ✓ p-value： 使用各组 PSI 平均值和方差，采用不等方差 t 检验计算
- ✓ magnitude：The proportion of isoforms which the splice event affects

13.2.3.2 Gene_Summary

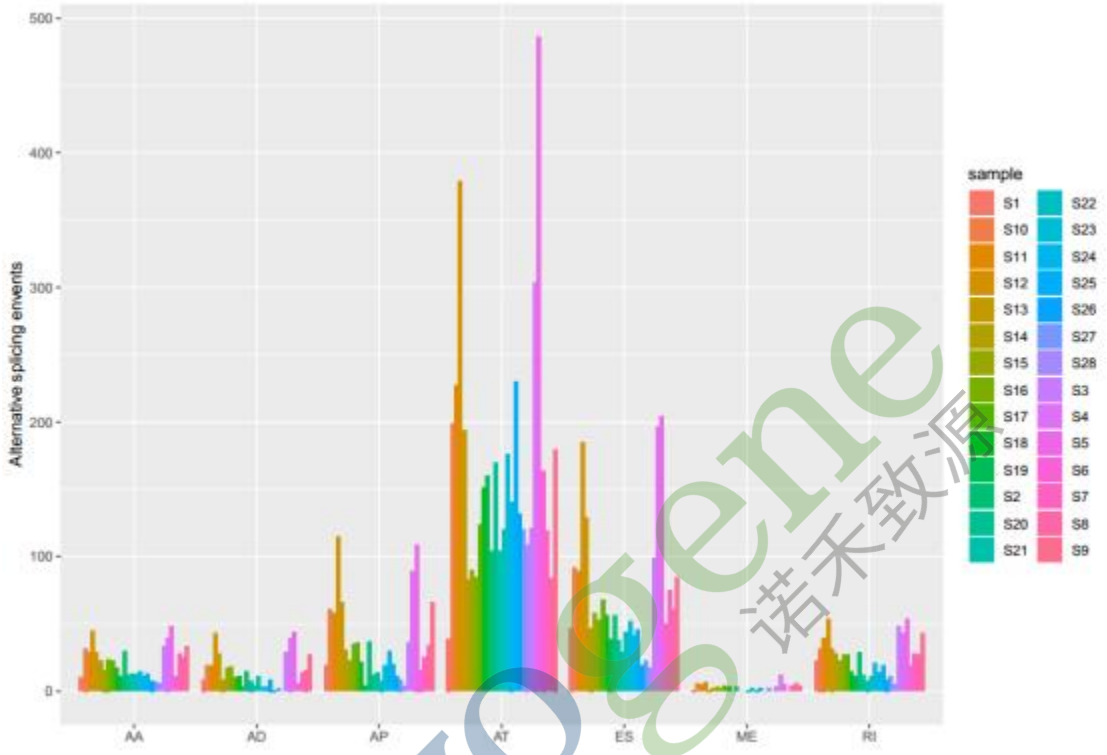
北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

	A	B	C	D	E	F	G	H
1	# View type: Group Compare View							
2	# Group Compare: Normal to Cancer							
3	# Filters: Min Gene RPKM>=1							
4	Gene Symb	Avg RPKM Gro	Avg RPKM Gro	Num Samples Gro	Num Samples	Splice Events	UniProt Events	
5	0610009D(	92.4183	146.645	5	2			
6	0610009O:	50.9756	18.2387	6	1			
7	0610010K1	15.8744	4.05393	4	2	AP- (2.1), AP+ (1), AA- (6.1), ES- (5, 6.1), ES- (4		
8	0610011F(	20.6228	2.69753	5	2	AT- (6), AT- (5.3), RI+ (5.2), ES- (2, 3, 4), ES- (2)		
9	0610031J0	107.562	166.562	8	2	ES- (2.2, 3, 4, 5, 6.1), ES- (3, 4, 5, 6.1), ES- (2.2, 3		
10	0610037L1	98.1791	57.8033	6	1	RI+ (7.2), RI+ (7.2, 7.3, 7.4, 7.5), AA- (7.3, 7.4, 7.5		
11	1110001A:	7.08696	11.7621	3	1			
12	1110004F1	105.869	87.2952	5	2	AA- (4.1), AD- (1.2), ES- (1.2, 2, 3)		
13	1110012L1	79.7825	41.3194	3	1			
14	1110032A(	24.0365	98.204	6	2	AP- (1), AP- (2), ES- (3)		
15	1110037F(	5.99702	51.8439	2	1			
16	1110038F1	40.2532	60.9952	4	2			
17	1110051M	61.7509	174.115	5	2	AT+ (10), AT- (9)		
18	1110058L1	25.4736	46.9974	2	2			
19	1500002C:	9.21945	5.95957	5	2			
20	1500011B(	112.569	2.25426	5	1	AT+ (3), AT- (2)		
21	1500012F(	14.4361	31.0012	2	2			
22	1600012H(	11.415	55.0557	7	2			
23	1600014C:	10.0146	7.11192	2	1	AT- (5), AT- (3.2)		
24	1700021K1	3.30486	1.04518	6	1			
25	1700022I1	35.8374	65.1332	7	2	AT+ (2), AT- (4)		
26	1700037C:	16.6571	1.8164	6	2	AT- (6), AT- (4)		
27	1700037H(	8.00723	26.2132	5	1	ES- (2.2, 3.1), AA- (3.1)		
28	1700047I1	3.09613	61.2759	3	1			

- ✓ Gene Symbol：发生可变剪切的基因名。
- ✓ Avg Reads：基因的平均 reads 数目。
- ✓ Avg RPKM：基因的平均表达量 RPKM 值。
- ✓ RPKM Std Dev：基因标准化后的表达量 RPKM 值。
- ✓ Num Sample：该可变剪接事件中组合里有表达的样本数。
- ✓ Splice Events ： Categorized list of the differential splice events detected in the group on average. For example， ES(4) indicates and exon skip on exon number 4 that occurs more frequently in one sample. Use the Splice Event Filter drop down list to see the description and codes for each event type.
- ✓ UniProt Event ： Categorized list of UniProt annotated regions that are affected by alternate splicing events in the group on average. For example， B(9) indicates a Bind Site on exon 9 that is excluded more frequently in one sample's isoforms. See the legend in the Protein Sequence Tab of the Gene Panel for a full list of UniProt annotation categories.

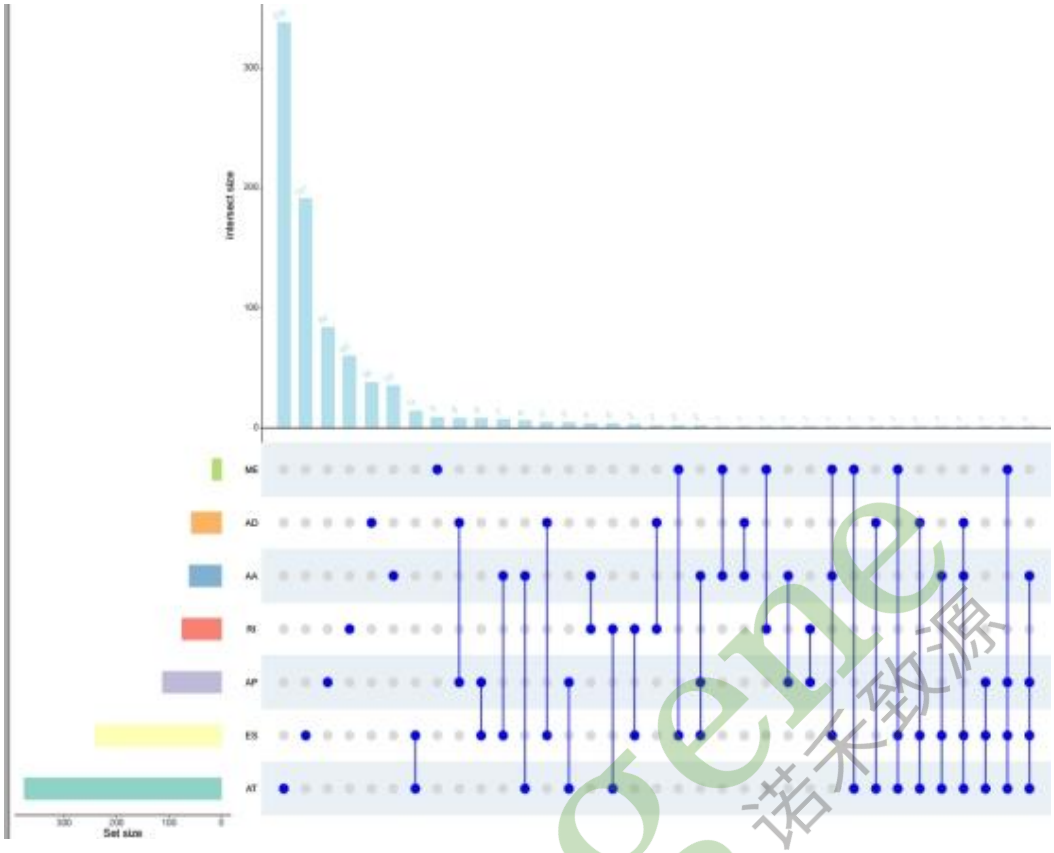
13.2.4 图片展示

13.2.4.1 各样本可变剪切类型柱状图



13.2.4.2 单个样本不同类型可变剪切事件进行 upset 图绘制

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801



13.3 参考资料

软件官网链接：https：//bioinformatics.mdanderson.org/public-software/spliceseq/

相关资料：https：//www.jianshu.com/p/af44e1da6e3c

Novogene 诺禾致源

14 PS1 分析并绘制 SE 图片

14.1 分析背景及介绍

该项分析为基于 AS 可变剪切数据进行的个性化分析，使用的数据为 PSI，对于 ASE 来说，PSI 计算公式为：PSI=splice in/(splice in+splice out)，公式来源：https：//www.jianshu.com/p/183a0c87a4ca，计算得到的 PSI 使用 R 包的单因素方差分析（客户要求使用 anowa 进行方差分析，但 anowa 为 SPSS 商用软件的分析内容，说明中提到该分析为单因素方差分析，即 one-way anova（方差分析的简称为 anova，F 检验），因无法使用 SPSS 软件，我们使用 R 包进行单因素方差分析。

https：//zhidao.baidu.com/question/1899829616755087100.html

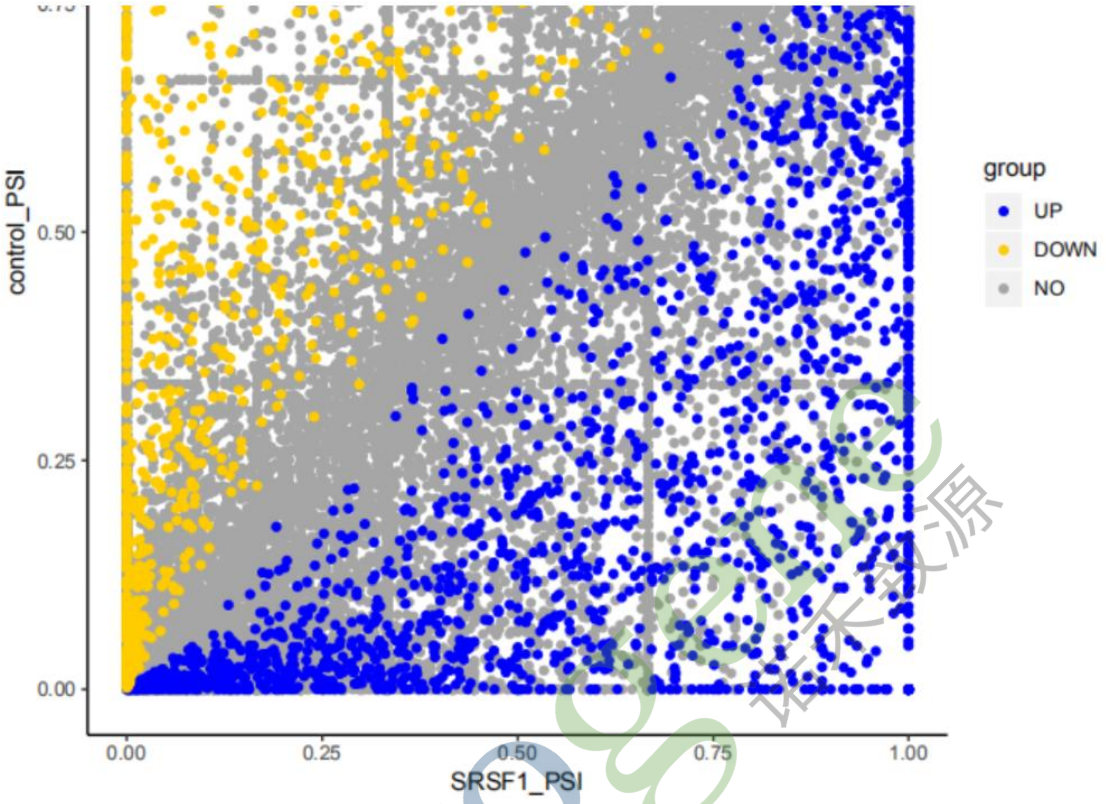
https：//baike.baidu.com/item/%E6%96%B9%E5%B7%AE%E5%88%86%E6%9E%90/1502206?fromtitle=anova&fromid=680591&fr=aladdin）

14.2 分析使用数据及方法

采用上述原理计算 PSI 值，并进行单因素方差分析，使用 PSI 的 FDR 值标注是否是显著的并结合原始结果中的 lncLevelDifference 的正负来标注上下调的信息。

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

14.3 结果展示



点的颜色或个别基因标注等信息可提供具体颜色代码

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

15 iPath 分析

15.1 分析背景及方法介绍

iPath（Interactive Pathways Explorer）（http://pathways.embl.de）是一个网页版的通路图可视化分析软件。iPath3.0 可提供 3 种不同通路的概括图:代谢通路、调控通路以及生物合成和次生代谢通路。其中，代谢通路属于第一层级 Metabolism 大类。代谢通路图中的节点对应各种化合物，边代表一系列的酶类反应。对多个样品来说，通过对它们的代谢途径进行比较与分析，找出它们的共性和特性，特别是对它们特殊代谢途径的分析，能够为进一步深入研究和利用样品的功能提供理论依据。

我们根据 KEGG PATHWAY 数据库的注释结果，提取多个样品功能类一新陈代谢（Metabolism）相关的Kos（KEGG orthologous groups），然后计算每个 KOs 在每个样品中的丰度，根据丰度大小来确定对应代谢途径所表示的边的颜色以及宽度。

15.2 结果展示

15.2.1 提取 KOs 结果

path_id	width	colour/green
map00710	W5.0	#00FF00
map00190	W5.0	#00FF00
map04145	W5.0	#00FF00
map01200	W5.0	#00FF00
map03060	W5.0	#00FF00
map00400	W5.0	#00FF00
map01230	W5.0	#00FF00

表 15.2.1 提取 KOs 结果

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

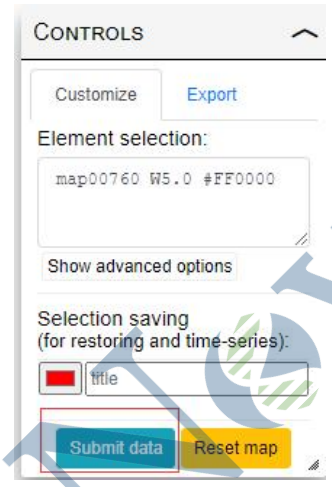
第一列：Path_id　通路 ID

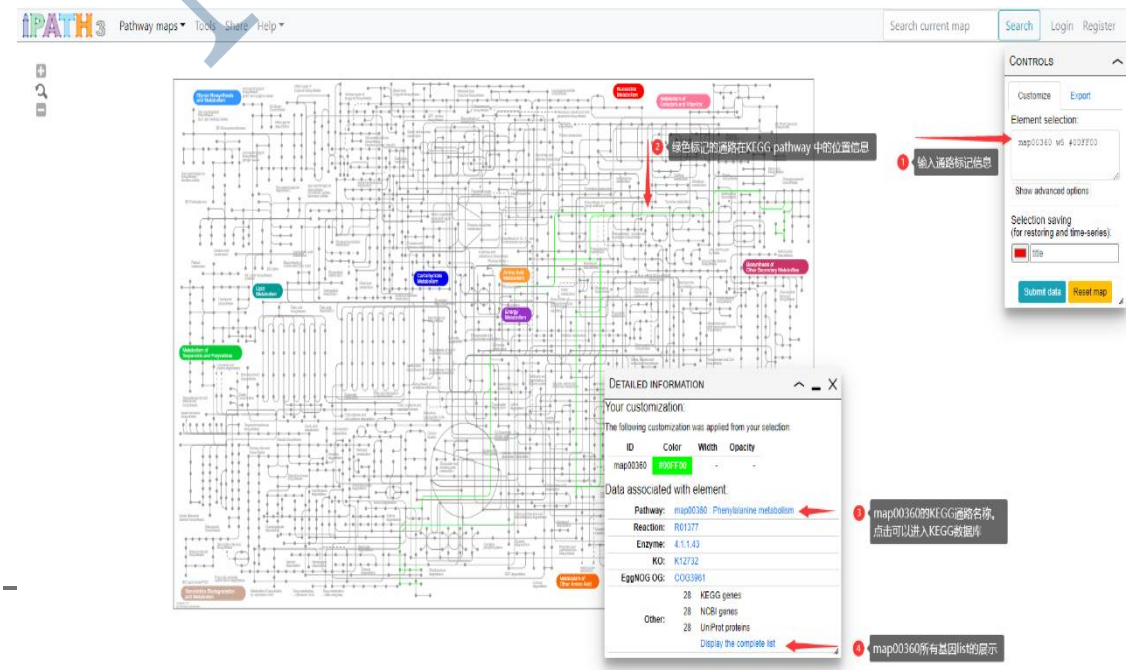
第二列：Width　在 ipath 网页中展示的宽度

第三列：Color　在 ipath 网页中展示的颜色（从左到右绿色黄色红色）

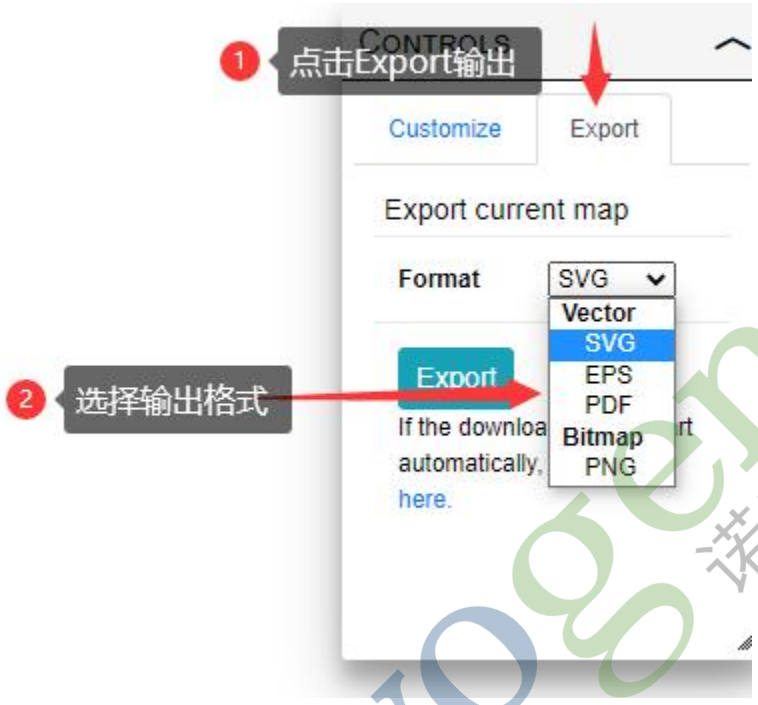
15.2.2 多样品代谢通路比较分析图

登录网址：https://pathways.embl.de/ipath3.cgi?map=metabolic，客户根据结果文件中所展示的数据结果，按照下图中所展示的提交方式即可得到代谢通路图，在 iPath 网站中用户可进行一系列放大，查找等基本操作，更多详细信息可参考官方帮助文档:http://pathways.embl.de/help.html。





结果文件下载：



15.3 参考文献

【1】Letunic I, Yamada T, Kanehisa M, et al. iPath: interactive exploration of biochemical pathways and networks[J]. Trends in Biochemical Sciences, 2008, 33(3):101.

【2】Yamada T, Letunic I, Okuda S, et al. iPath2.0: interactive pathway explorer[J]. Nucleic Acids Research, 2011, 39(Web Server issue):W412.

16 预测分泌蛋白分析

16.1 分析背景及方法介绍

分泌蛋白是指在细胞内合成后，分泌到细胞外起作用的蛋白质。其特点主要有：含有 N 端信号肽，无跨膜结构域，无 GPI 锚定位点，没有将蛋白输送到线粒体或其他胞内细胞器。对于无参项目需先对 Unigene 用 TransDecoderv3.0.1 进行 unigene 的全长 ORF 预测,然后对 complete.pep 预测分泌蛋白,有参考基因组项目直接下载蛋白序列，然后使用 SignalP v4.1 进行信号肽鉴定，获得分泌蛋白序列，再通过 TMHMM v2.0 进行跨膜结构域分析，将不跨膜及单跨膜初步推断为分泌蛋白，最后用 TargetP v1.1 预测真核蛋白的亚细胞定位，筛选 LOC 为 S 的作为最终分泌蛋白。

16.2 结果展示

16.2.1 SignalP v4.1 预测结果

# Measure	Position	Value	Cutoff	signal peptide?
max. C	18	0.334		
max. Y	18	0.560		
max. S	11	0.968		
mean S	1-17	0.935		
D	1-17	0.763	0.450	YES
Name=Gene.39531::Cluster-10020.0::g.39531::m.39531 SP='YES' Cleavage site between pos. 17 and 18: AVA-AV D=0.763 D-cutoff=0.450 Networks=SignalP-noTM				

表 16.2.1 SignalP v4.1 预测结果

C-score：raw cleavage site score

C 分数被训练为在切割位点之后立即处于高位（成熟蛋白中的第一个残基）。

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

S-score：signal peptide score

SP 网络的输出经过训练，可以区分信号肽中的位置与蛋白质成熟部分中的位置以及没有信号肽的蛋白质。

Y-score：combined cleavage site score

C 分数和 S 分数的斜率的组合（几何平均值），比单独的原始 C 分数产生更好的切割位点预测。这是由于：可以在一个序列中找到多个高　　峰 C 分数，其中只有一个是真正的切割位点。Y 分数通过选择 S 分数的斜率陡峭的峰来区分 C 峰。

D-Score: discrimination score

平均值 S 和最大值的加权平均值。

mean S: 可能的信号肽的平均 S 评分（从位置 1 到紧邻最大 Y 评分的位置）。

signal peptide?：是否是信号肽，YES 表示是，NO 表示否。判定标准：D-score 比 D-cutoff 高时即为 YES。

16.2.2 TMHMM v2.0 预测结果

Gene.27111::Cluster-16402.0::g.27111::m.27111	TMHMM2.0	outside	1	312
Gene.143220::Cluster-17511.46455::g.143220::m.143220	TMHMM2.0	inside	105	228
Gene.143220::Cluster-17511.46455::g.143220::m.143220	TMHMM2.0	TMhelix	229	251

表 16.2.2 TMHMM v2.0 预测结果

第一列: GeneID。

第二列：软件名称。

第三列：蛋白位置信息。（例：outside 表示蛋白 1-312 位于细胞膜表面；inside 表示 105-228 位于细胞膜内部；TMhelix 表示 229-251 形成一个典型的跨膜螺旋区）

第四列：序列起始位点。

第五列：序列终止位点。

16.2.3 TargetP v1.1 预测结果

Name	Len	mTP	SP	other	Loc	RC

北京诺禾致源科技股份有限公司　网址：www.novogene.com　邮箱：service@novogene.com　电话：010-8283 7801

Gene.113652::Cluster-16046.39476::g.113652::m.113652	262	0.024	0.962	0.059	S	1
Gene.61144::Cluster-16046.58501::g.61144::m.61144	175	0.019	0.96	0.05	S	1
Gene.143839::Cluster-16046.67724::g.143839::m.143839	287	0.029	0.94	0.062	S	1
Gene.97411::Cluster-16046.81299::g.97411::m.97411	113	0.045	0.881	0.094	S	2
Gene.71563::Cluster-16046.44972::g.71563::m.71563	273	0.052	0.927	0.032	S	1
Gene.8982::Cluster-9386.0::g.8982::m.8982	132	0.05	0.891	0.153	S	2

表 16.2.3 TargetP v1.1 预测结果

第一列 Name:基因 ID。

第二列 Len:长度。

第三，四，五列 mTP,SP,other:Final NN scores on which the final prediction is based (Loc, see below). Note that the scores are not really probabilities, and they do not necessarily add to one. However, the location with the highest score is the most likely according to TargetP, and the relationship between the scores (the reliability class, see below) may be an indication of how certain the prediction is。

第六列 Loc:C Chloroplast, i.e. the sequence contains cTP, a chloroplast transit peptide;

M Mitochondrion, i.e. the sequence contains mTP, a mitochondrial targeting peptide;

S Secretory pathway, i.e. the sequence contains SP, a signal peptide;

_ Any other location;

第七列 RC:Reliability class, from 1 to 5, where 1 indicates the strongest prediction. RC is a measure of the size of the difference ('diff') between the highest (winning) and the second highest output scores. There are 5 reliability classes, defined as follows。

16.3 参考文献

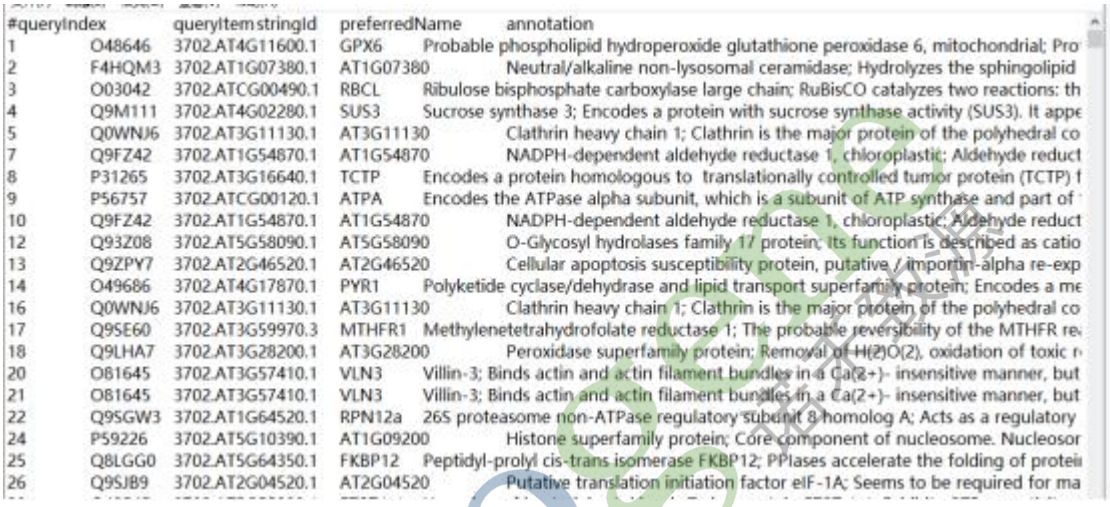
【1】Liu J J, Sturrock R N, Sniezko R A, et al. Transcriptome analysis of the white pine blister rust pathogen Cronartium ribicola : de novo assembly, expression profiling, and identification of candidate effectors[J]. Bmc Genomics, 2015, 16(1):678.

17 PPI 结果中蛋白的注释

17.1 分析背景及介绍

可基于转录组的 PPI 结果中展示的互作蛋白 id，提供 String 数据库中蛋白对应的注释信息。

17.2 结果展示



#queryIndex	queryItem	stringId	preferredName	annotation
1	O48646	3702.AT4G11600.1	GPX6	Probable phospholipid hydroperoxide glutathione peroxidase 6, mitochondrial; Pro
2	F4HQM3	3702.AT1G07380.1	AT1G07380	Neutral/alkaline non-lysosomal ceramidase; Hydrolyzes the sphingolipid
3	O03042	3702.ATCG00490.1	RBCL	Ribulose bisphosphate carboxylase large chain; RuBisCO catalyzes two reactions: th
4	Q9M111	3702.AT4G02280.1	SUS3	Sucrose synthase 3; Encodes a protein with sucrose synthase activity (SUS3). It appe
5	Q0WNJ6	3702.AT3G11130.1	AT3G11130	Clathrin heavy chain 1; Clathrin is the major protein of the polyhedral co
7	Q9FZ42	3702.AT1G54870.1	AT1G54870	NADPH-dependent aldehyde reductase 1, chloroplastic; Aldehyde reduct
8	P31265	3702.AT3G16640.1	TCTP	Encodes a protein homologous to translationally controlled tumor protein (TCTP) f
9	P56757	3702.ATCG00120.1	ATPA	Encodes the ATPase alpha subunit, which is a subunit of ATP synthase and part of
10	Q9FZ42	3702.AT1G54870.1	AT1G54870	NADPH-dependent aldehyde reductase 1, chloroplastic; Aldehyde reduct
12	Q93Z08	3702.AT5G58090.1	AT5G58090	O-Glycosyl hydrolases family 17 protein; Its function is described as catio
13	Q9ZPY7	3702.AT2G46520.1	AT2G46520	Cellular apoptosis susceptibility protein, putative / importin-alpha re-exp
14	O49686	3702.AT4G17870.1	PYR1	Polyketide cyclase/dehydrase and lipid transport superfamily protein; Encodes a me
16	Q0WNJ6	3702.AT3G11130.1	AT3G11130	Clathrin heavy chain 1; Clathrin is the major protein of the polyhedral co
17	Q9SE60	3702.AT3G59970.3	MTHFR1	Methylenetetrahydrofolate reductase 1; The probable reversibility of the MTHFR re
18	Q9LHA7	3702.AT3G28200.1	AT3G28200	Peroxidase superfamily protein; Removal of H(2)O(2), oxidation of toxic r
20	O81645	3702.AT3G57410.1	VLN3	Villin-3; Binds actin and actin filament bundles in a Ca(2+)- insensitive manner, but
21	O81645	3702.AT3G57410.1	VLN3	Villin-3; Binds actin and actin filament bundles in a Ca(2+)- insensitive manner, but
22	Q9SGW3	3702.AT1G64520.1	RPN12a	26S proteasome non-ATPase regulatory subunit 8 homolog A; Acts as a regulatory
24	P59226	3702.AT5G10390.1	AT1G09200	Histone superfamily protein; Core component of nucleosome. Nucleosor
25	Q8LGG0	3702.AT5G64350.1	FKBP12	Peptidyl-prolyl cis-trans isomerase FKBP12; PPIases accelerate the folding of protei
26	Q9SJB9	3702.AT2G04520.1	AT2G04520	Putative translation initiation factor eIF-1A; Seems to be required for ma

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

18 蛋白结构域预测分析

18.1 分析背景及介绍

诺禾可基于蛋白组序列与 pfam 数据库比对以预测蛋白的结构域信息，寻找蛋白结构域。

18.2 结果展示

gene id	Pfam_domain
evm.model.Contig4138.1	PF05786:Condensin complex subunit
evm.model.Contig16180.1	-
evm.model.Contig20053.1	PF01805:Surp module
evm.model.Contig1009.1	PF07714:Protein tyrosine kinase
evm.model.Contig23211.1	-
evm.model.Contig2708.1	-
evm.model.Contig24469.1	-
evm.model.Contig552.8	-
evm.model.Contig552.5	PF07527:Hairy Orange|PF00010:Helix-loop-helix DNA-binding domain
evm.model.Contig552.9	PF00536:SAM domain (Sterile alpha motif)
evm.model.Contig552.7	-
evm.model.Contig552.6	-
evm.model.Contig552.3	PF01431:Peptidase family M13|PF05649:Peptidase family M13
evm.model.Contig5993.1	-
evm.model.Contig27183.1	PF00057:Low-density lipoprotein receptor domain class A
evm.model.Contig34699.1	PF05404:Translocon-associated protein, delta subunit precursor (TRAP-delta)
evm.model.Contig423.1	PF00135:Carboxylesterase family
evm.model.Contig15.36.1	PF00675:Insulinase (Peptidase family M16)|PF16187:Middle or third domain of
evm.model.Contig15.59	PF15313:Hexamethylene bis-acetamide-inducible protein
evm.model.Contig15.30	PF07776:Zinc-finger associated domain (zf-AD)
evm.model.Contig15.47	-
evm.model.Contig15.64	-
evm.model.Contig15.20	PF16399:Intron-binding protein aquarius N-terminus|PF13087:AAA domain|PF1308
evm.model.Contig15.11	PF01900:Rpp14/Pop5 family
evm.model.Contig15.34	PF00071:Ras family
evm.model.Contig15.23	-
evm.model.Contig15.39	PF13639:Ring finger domain
evm.model.Contig15.13	-
evm.model.Contig15.29	PF00009:Elongation factor Tu GTP binding domain
evm.model.Contig15.52	PF01462:Leucine rich repeat N-terminal domain|PF13855:Leucine rich repeat|PF
evm.model.Contig15.18	PF00587:tRNA synthetase class II core domain (G, H, P, S and T)|PF00749:tRNA
evm.model.Contig15.28	PF02373:JmjC domain, hydroxylase

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

19 内参基因分析

19.1 分析背景及方法介绍

将所有基因进行如下处理：保留在所有样品中均表达的基因（有参 fpkm>1,无参 fpkm>0.3，至少 fpkm>0）；与比较组合的差异基因进行比较，过滤掉差异基因；根据表达量大小自大到小排序，选取表达量前 20 的基因利用 NormFinder 软件进行内参基因的筛选。

19.2 结果展示

19.2.1 NormFinder 筛选出的内参基因

Gene name	Stability value
hsa-miR-1307-5p	0.246
hsa-miR-26b-5p	0.269
hsa-miR-29b-3p	0.221
hsa-miR-3157-3p	0.224

19.2.1 NormFinder 筛选出的内参基因

第一列 Gene name：表达量前 20 的基因

第二列 Stability value：基因稳定性值， Stability value 越小，越稳定

19.3 参考文献

Daniel R, Wu Q, Williams V, et al. A Panel of MicroRNAs as Diagnostic Biomarkers for the Identification of Prostate Cancer:[J]. International Journal of Molecular Sciences, 2017, 18(6).

20 密码子偏好性分析

20.1 分析背景及方法介绍

每种氨基酸至少对应一个遗传密码子（一般不超过 6 个），编码同一种氨基酸的密码子称为同义密码子,同义密码子在使用频率中存在一定的差异，这种现象称为密码子使用偏好性（codon usage bias, CUB），并且在某一特定物种或者基因中比较倾向使用的同义密码子称为最优密码子（optimal codons）。

影响密码子偏好性的因素有突变压力（如 GC 含量、基因碱基组成）、选择作用（如基因表达水平、tRNA 丰度、蛋白结构与长度、翻译的起始信号）。此分析主要适用于原核物种和低高等真核生物。我们利用 CodonW 中进行密码子使用偏性的分析。

使用指标：

CAI（密码子适应指数）：基因表达水平测量，0-1

ENc(有效密码子数)：一个基因中所用到的密码子种类的多少。20-61

20：每个氨基酸只使用一个密码子；61：各个密码子均被平均使用

CBI(密码子偏爱指数)：一个具体基因中高表达优越密码子的组分情况

FOP(最优密码子使用频数)：使用最优密码子占总数百分比。

RSCU(相对同义密码子使用度):指某一特定密码子在使用频率与其无偏好性使用时预期频率之间的比值。密码子偏好性程度的有效指标。

RSCU 比值等于 1 说明该密码子无使用偏好性，若 RSCU 比值大于 1 说明该密码子的使用频率较高。

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

20.2 结果展示

20.2.1 密码子偏好性分析结果

	num	RSCU								
Phe	UUU	10722	0.39	Ser	UCU	11258	0.55	Tyr	UAU	7827	0.37	Cys	UGU	11018	0.65
	UUC	43762	1.61		UCC	21608	1.06		UAC	34797	1.63		UGC	23078	1.35
Leu	UUA	3645	0.13		UCA	11062	0.54	TER	UAA	431	0.37	TER	UGA	1902	1.65
	UUG	12720	0.44		UCG	34396	1.68		UAG	1128	0.98	Trp	UGG	20640	1.00
	CUU	11872	0.41	Pro	CCU	10373	0.43	His	CAU	10876	0.43	Arg	CGU	14209	0.62
	CUC	36988	1.29		CCC	22900	0.94		CAC	39732	1.57		CGC	50447	2.22
	CUA	6785	0.24		CCA	15685	0.64	Gln	CAA	12413	0.31		CGA	15548	0.68
	CUG	100022	3.49		CCG	48432	1.99		CAG	67322	1.69		CGG	35712	1.57
Ile	AUU	11417	0.56	Thr	ACU	11655	0.50	Asn	AAU	11347	0.46	Ser	AGU	12072	0.59
	AUC	41057	2.02		ACC	34791	1.50		AAC	38069	1.54		AGC	32295	1.58
	AUA	8573	0.42		ACA	14007	0.60	Lys	AAA	18043	0.51	Arg	AGA	8948	0.39
Met	AUG	38457	1.00		ACG	32417	1.40		AAG	53097	1.49		AGG	11721	0.51
Val	GUU	12087	0.38	Ala	GCU	20790	0.51	Asp	GAU	22588	0.46	Gly	GGU	18692	0.62
	GUC	40273	1.25		GCC	72796	1.78		GAC	76191	1.54		GGC	71447	2.38
	GUA	9710	0.30		GCA	18277	0.45	Glu	GAA	24400	0.48		GGA	16714	0.56
	GUG	66525	2.07		GCG	52052	1.27		GAG	77555	1.52		GGG	13279	0.44

1740652 codons in Average of genes (used Universal Genetic code)

表 20.2.1 密码子偏好性分析结果

RSCU 是指相对同义密码子使用度，RSCU 比值等于 1 说明该密码子无使用偏好性，若 RSCU 比值大于 1 说明该密码子的使用频率较高

20.3 参考文献

Peden J F. Analysis of codon usage.[J]. University of Nottingham, 2011, 90(1):73–74.

21 内含子保留率分析

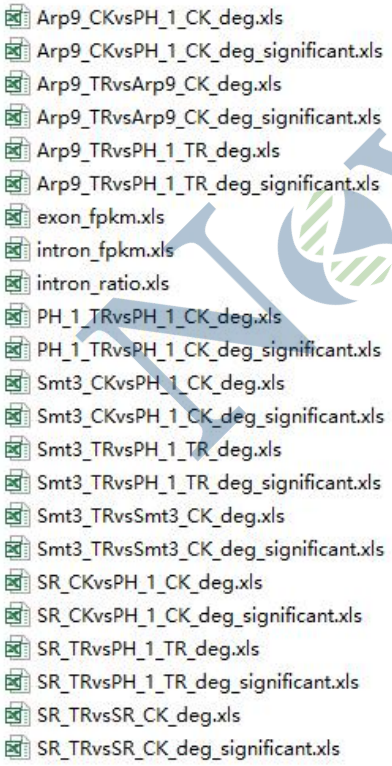
21.1 分析背景及介绍

该分析首先是对 gene 进行定量，去掉 fpkm 小于 1 的 gene，然后对 gene 中 intron 进行定量，计算 fpkm。intron 保留率的计算根据 intron 的 fpkm 比上该 intron 所在 gene 全部 exon 的 reads 计算的 fpkm；在计算 intron 保留率的差异时，选择表达量差异在 2 倍以上，使用 t 检验，阈值一选择矫正后的 p 值小于 0.05。

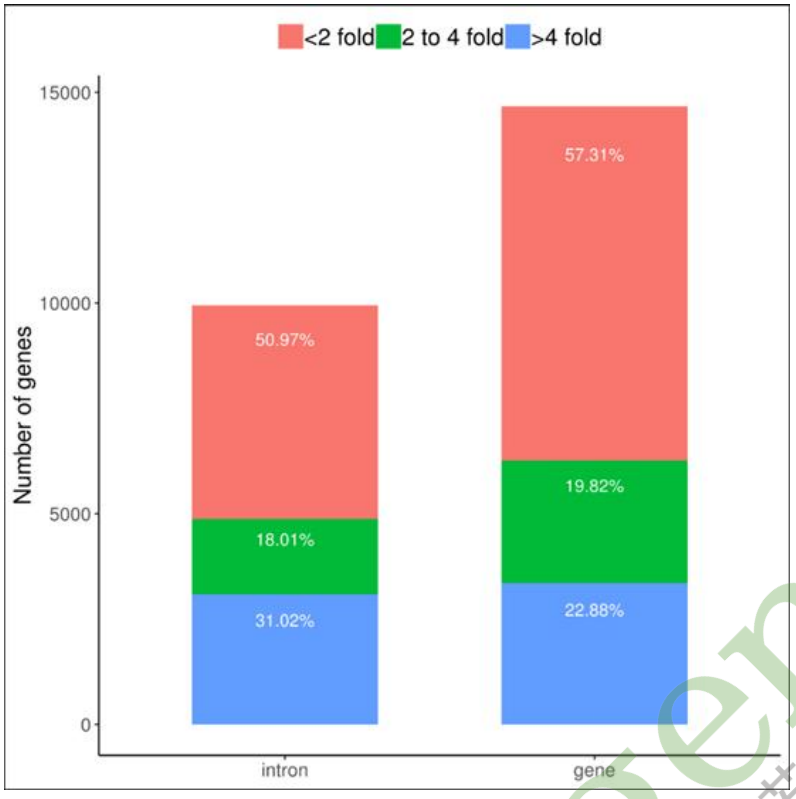
21.2 分析使用数据

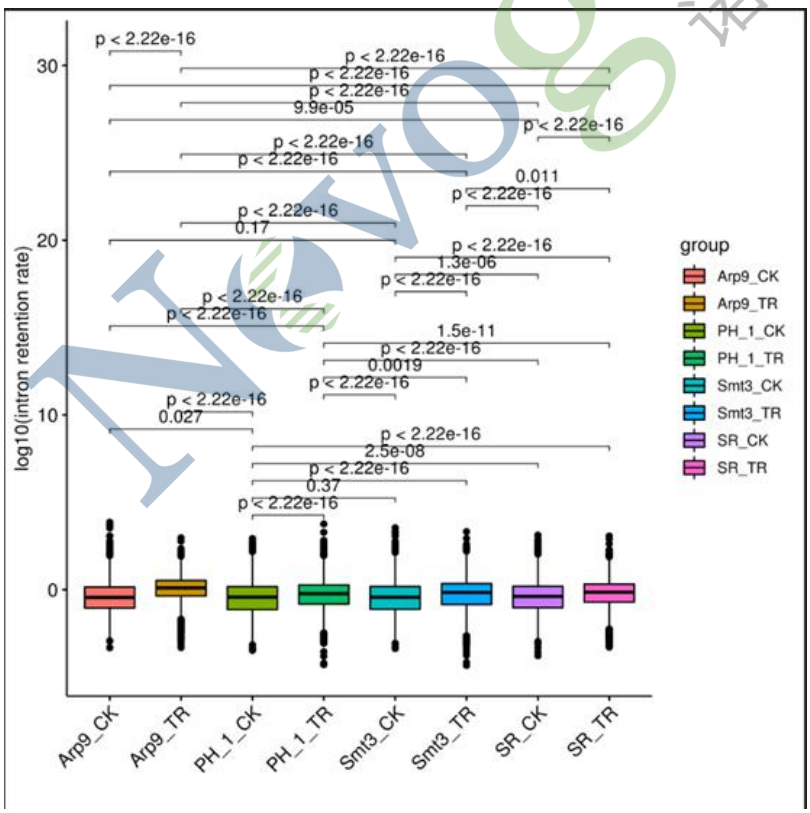
原始数据或 bam 文件

21.3 结果展示：



北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801





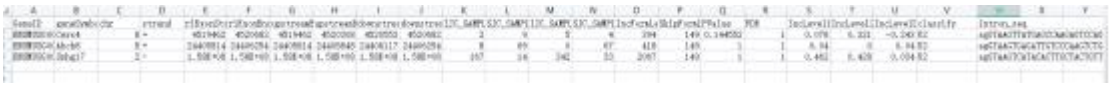
Novogene 诺禾致源

22U2-型内含子、U12-型内含子鉴定

22.1 分析背景及介绍

真核生物基因的一个基本特征是它们被一个或多个内含子所间隔，这些间隔 DNA 在转录后被除去以形成具有完整读码框的 mRNA，这一过程叫做内含子的剪接。根据剪接机制的不同，可将内含子分为 3 类：真核 mRNA 内含子、自我剪接内含子和真核 tRNA 内含子，其中主要是真核 mRNA 内含子。真核 mRNA 内含子有两种类型：U2-型内含子和 U12-型内含子，U2-型内含子存在较为普遍，占总数的 99%，而 U12-型内含子含量较少(<0.4%)。U2-型内含子的 5 剪接位点(5'splicing site，5'ss)具有 AG/GTAAGT 的保守序 列，3′剪接位点具有 TGCAG/G(3′splicing site，3′ss)的保守序列。分支点位于 3′ss 上游大约 20~30nt 处，序列并不保守，一般含有一个腺苷酸，突变或缺失腺苷酸会 降低剪接的效率或导致 pre-mRNA 无法剪接。脊椎动物内含子分支点下游还有一段多聚 嘧啶序列，是其他生物内含子不具有的。植物内含子的一个显著特点就是富含 UA 序列 ， UA 序列均匀的分散在整个内含子当中，对于保证剪接的保真度和精确性起着关键的作 用。虽然 U12-型内含子含量较少，但是在植物、哺乳动物、昆虫的核基因组中均有发现。 第一个发现的 U12-型内含子是以 AT 双核苷酸开始，AC 双核苷酸结束，所以 U12-型内含 子又被称为 AT-AC 型内含子。后来发现 U12-型内含子也有 GT-AG 型的，此外还有一小部 分边界并不规则。U12-型内含子 5′-ss(G/ATATCCTY)和分支点(TCCTTRAY)序列高度保 守，而 3′ss 序列的保守性稍差(YAC/G)，与分支点之间距离大约为 10~20nt。植物 U12- 型内含子也富含 UA 序列，与 U2-型内含子相似。

22.2 结果展示



- ✓ riExonStart_0base：可变剪接事件滞留内含子的起始位置，以 0 开始计数
- ✓ riExonEnd：可变剪接事件滞留内含子的终止位置
- ✓ Classify：U2、U12 剪切类型
- ✓ intron_seq：序列部分大写为内含子，小写为内含子上下游碱基

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

22.3 参考文献

[1]Ko CH,Brendel V,et al. Plant Mol Biol,1998,36:573~583.

[2]Levine A, Durbin R. Nucleic Acids Res,2001,29:4006~4013.

[3]Patel AA,Steitz JA. Nat Rev Mol Cell Biol,2003,4:960~970.

Novogene 诺禾致源

23 tSNE 与 UMAP 聚类分析

23.1 分析背景及介绍

tSNE 与 UMAP 聚类分析是与 PCA 分析类似的两种考量样本间相似度的降维算法。

tSNE （全称为 t-distributed Stochastic Neighbor Embedding，翻译为 t 分布-随机邻近嵌入）是通过将数据点之间的相似度转化为条件概率，原始空间中数据点的相似度由高斯联合分布表示，嵌入空间中数据点的相似度由学生 t 分布 表示 能够将高维空间中的数据映射到低维空间中，并保留数据集的局部特性。

UMAP（全称为 Uniform Manifold Approximation and Projection，翻译为统一流形逼近与投影）是一种新的降维流形学习技术，UMAP 是建立在黎曼几何和代数拓扑理论框架上的，UMAP 是一种非常有效的可视化和可伸缩降维算法。在可视化质量方面，UMAP 算法与 t-SNE 具有竞争优势，但是它保留了更多全局结构、具有优越的运行性能、更好的可扩展性。

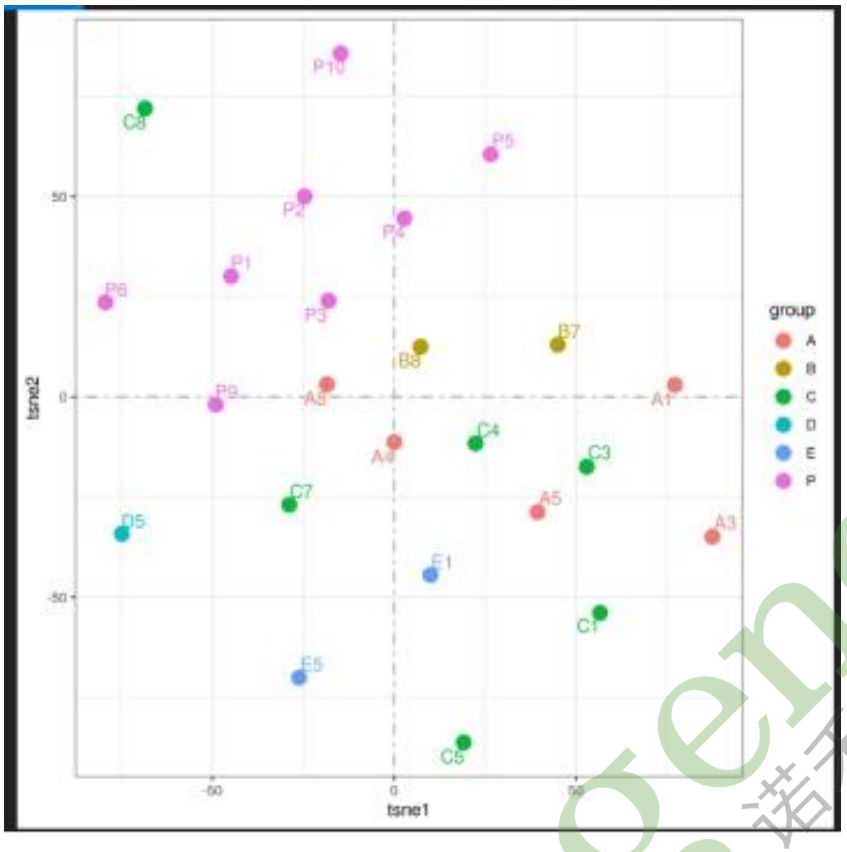
23.2 分析使用数据

表达量数据

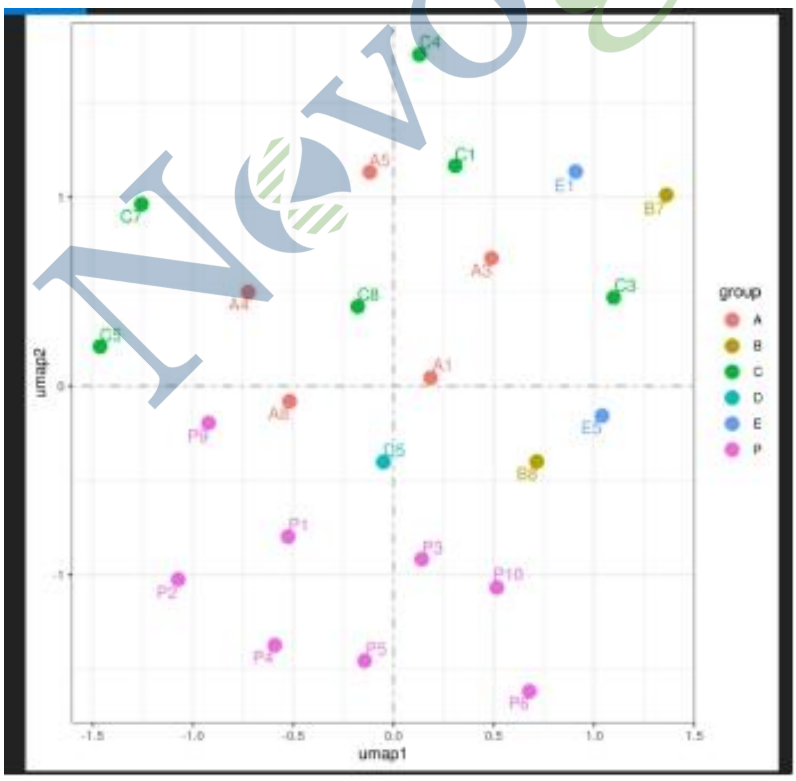
23.3 结果展示

Tsne 结果：

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801



UMAP 结果：



Novogene 诺禾致源

24 样本与表型的聚类分析以及保守性分析

24.1 分析背景及介绍

24.1.1 样本与表型的聚类分析

WGCNA 分析软件中，使用 sample dendrogram and trait heatmap 函数绘制样本与表型的聚类分析（常规分析以外的函数）

参考链接：

http：//www.bio-info-trainee.com/2535.html

https：//blog.csdn.net/weixin_43569478/article/details/83747303

http：//www.360doc.com/content/17/0804/17/46032984_676659287.shtml

https：//blog.csdn.net/wj176623/article/details/53006974

24.1.2 保守性分析

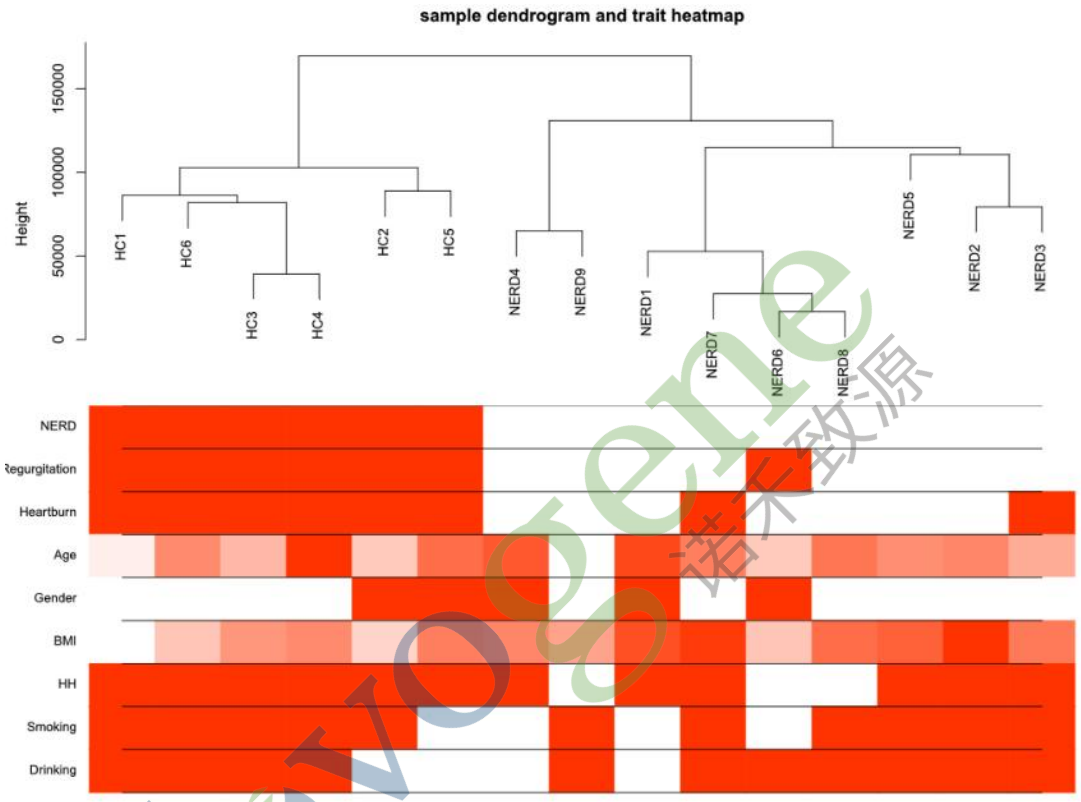
使用 mains = c("Preservation Median rank"，"Preservation Zsummary") 函数绘制保守性图片。

相关的链接为：https：//www.it610.com/article/1226539685594107904.htm

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

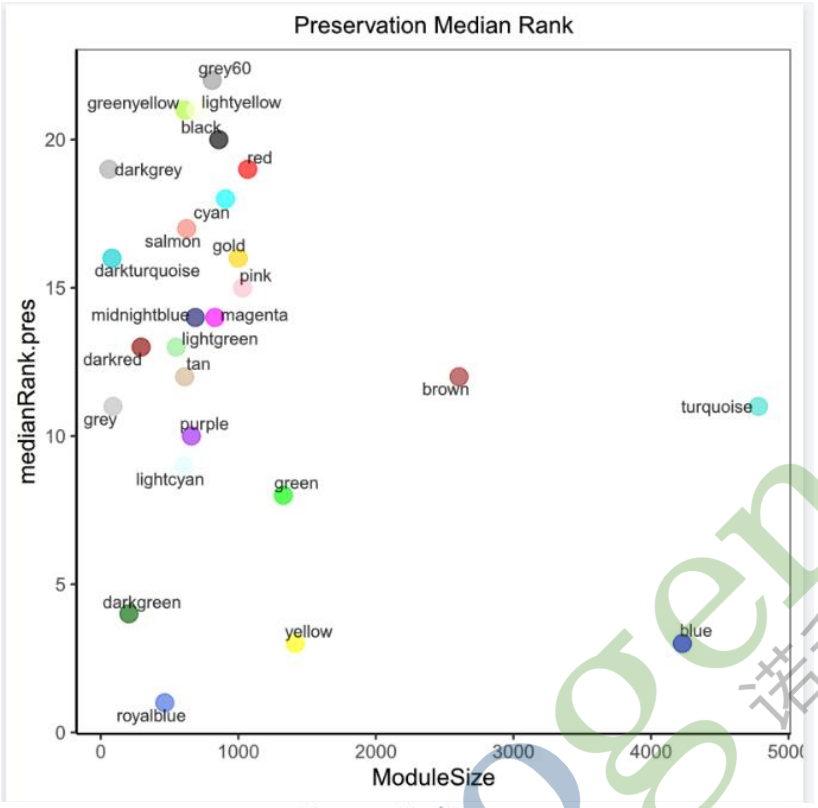
24.2 结果展示

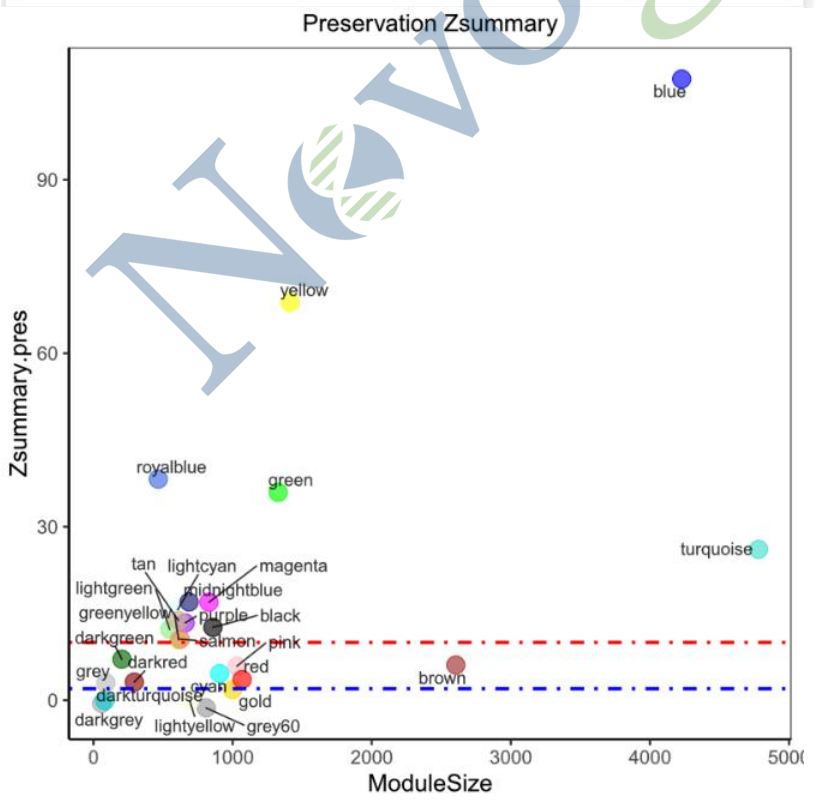
24.2.1 样本与表型的聚类分析



Novogene 诺禾致源

24.2.2 保守性分析





Novogene 诺禾致源

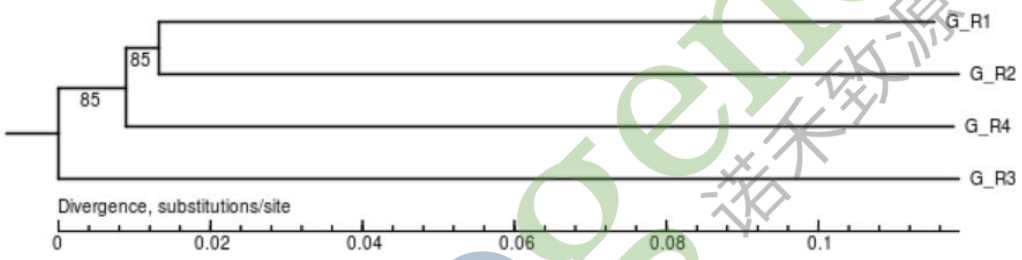
25 SNP 构树

从遗传结构角度分析，从 snp 结果中筛选 reads 支持数>6 的 snp 结果，然后基于不同样品的 SNP 基因型，nj 法构建系统发育进化树。

25.1 分析方法

输入文件：vcf 文件

25.2 结果展示



每一个进化支代表一个样本；每一个节点代表一个分类单元，可以是基因家族或同源物等；进化支的长短表示进化的程度（SNP 的数量），越短代表进化距离越近。最下方为距离标尺，指生物或序列间差异数值的单位长度，相当于进化树的比例尺。

25.3 参考文献

Leaf transcriptome analysis of a subtropical evergreen broadleaf plant, wild oil-tea camellia (Camellia oleifera), revealing candidate genes for cold acclimation

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

26 个性化热图绘制

26.1 多样品差异基因热图分析

在多阶段性或者多时期样品比对分析项目中，客户做了多阶段样品间两两样品间的差异比对分析，此热图可以直观展示多样品两两比较间的上调下调基因个数，可以间接展示各样品间的差异变化。

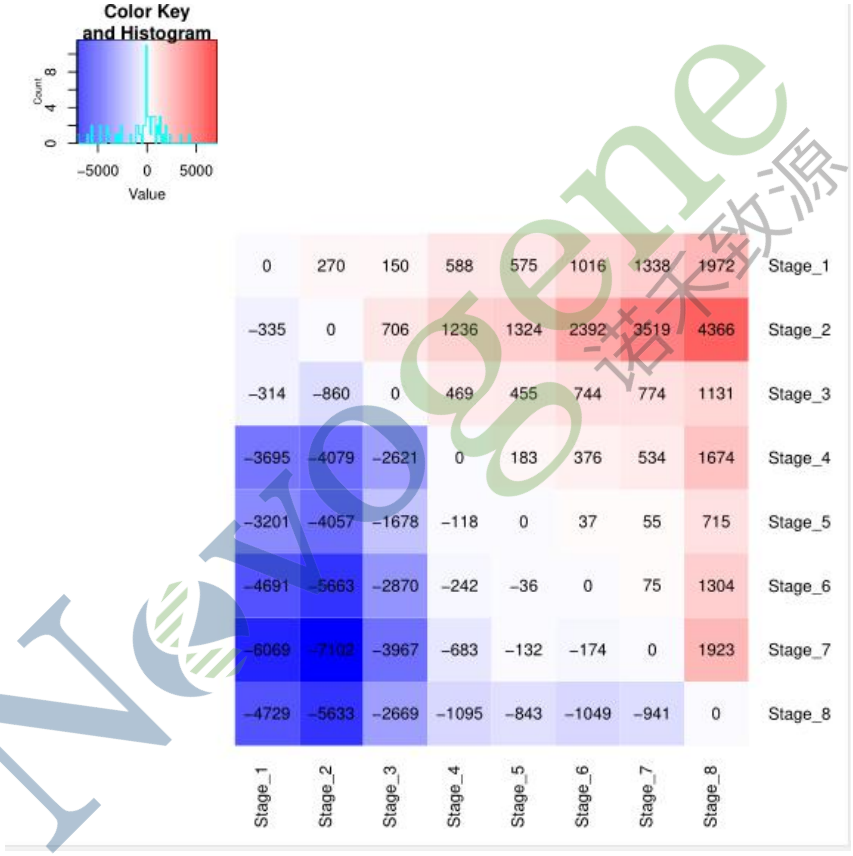


图 5.1 多样品差异基因热图分析

此图颜色越深，代表差异基因越多，蓝色表示下调差异基因个数，红色表示上调差异基因个数。如图所示，stage1vs stage8，可以看出上调基因有 1972 个，下调基因个数 4729 个。

参考文献：

Letunic I, Yamada T, Kanehisa M, et al. iPath: interactive exploration of biochemical pathways and networks[J]. Trends in Biochemical Sciences, 2008, 33(3):101.

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

26.2 展示横纵向标签的热图

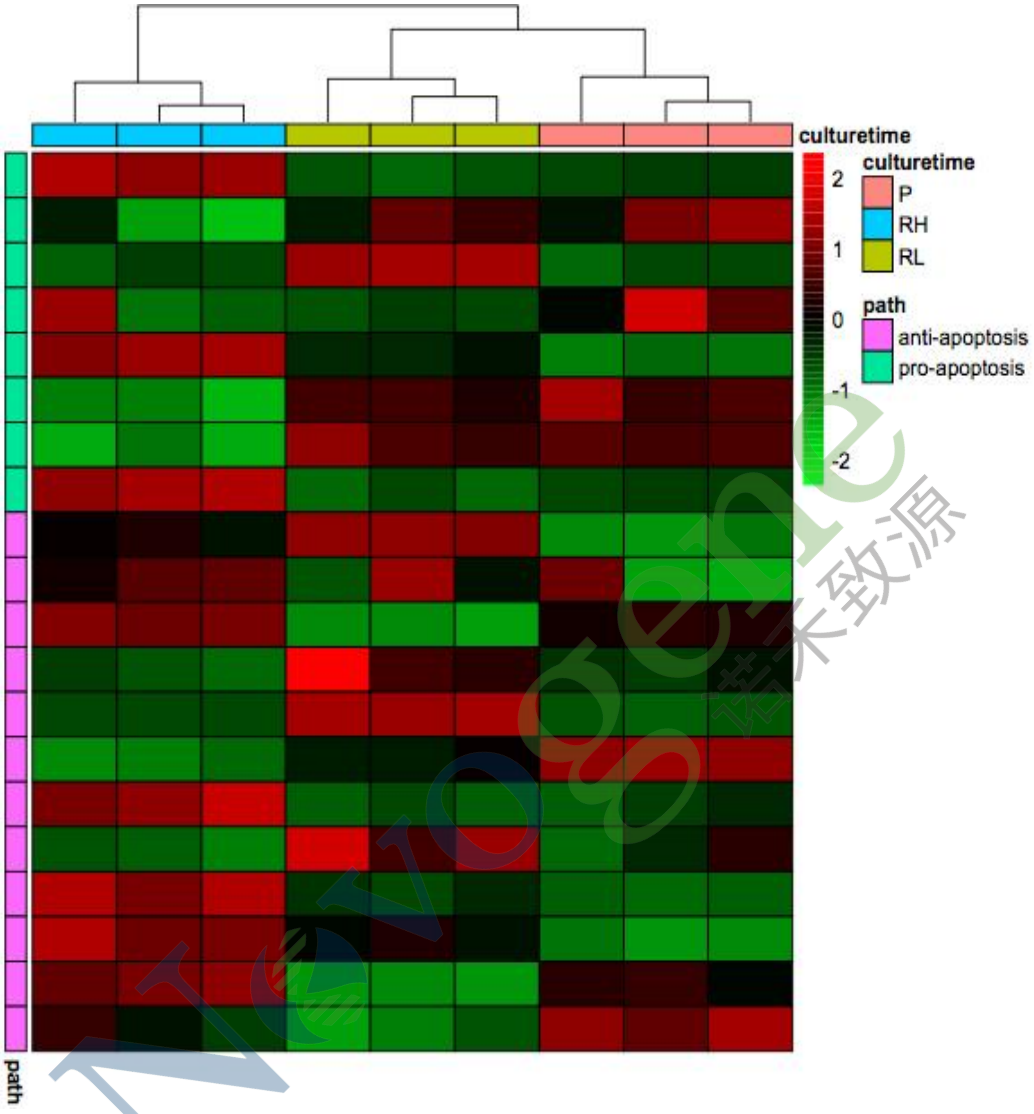


图 5.2 展示横纵向标签的热图

此图根据基因在不同样本中的表达量做聚类，并在横纵轴进行了标签分类标记，横轴表示不同的样本分类，纵轴表示不同的代谢途径。

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

27 个性化 Circos 图绘制

27.1 差异基因在染色体上的分布图

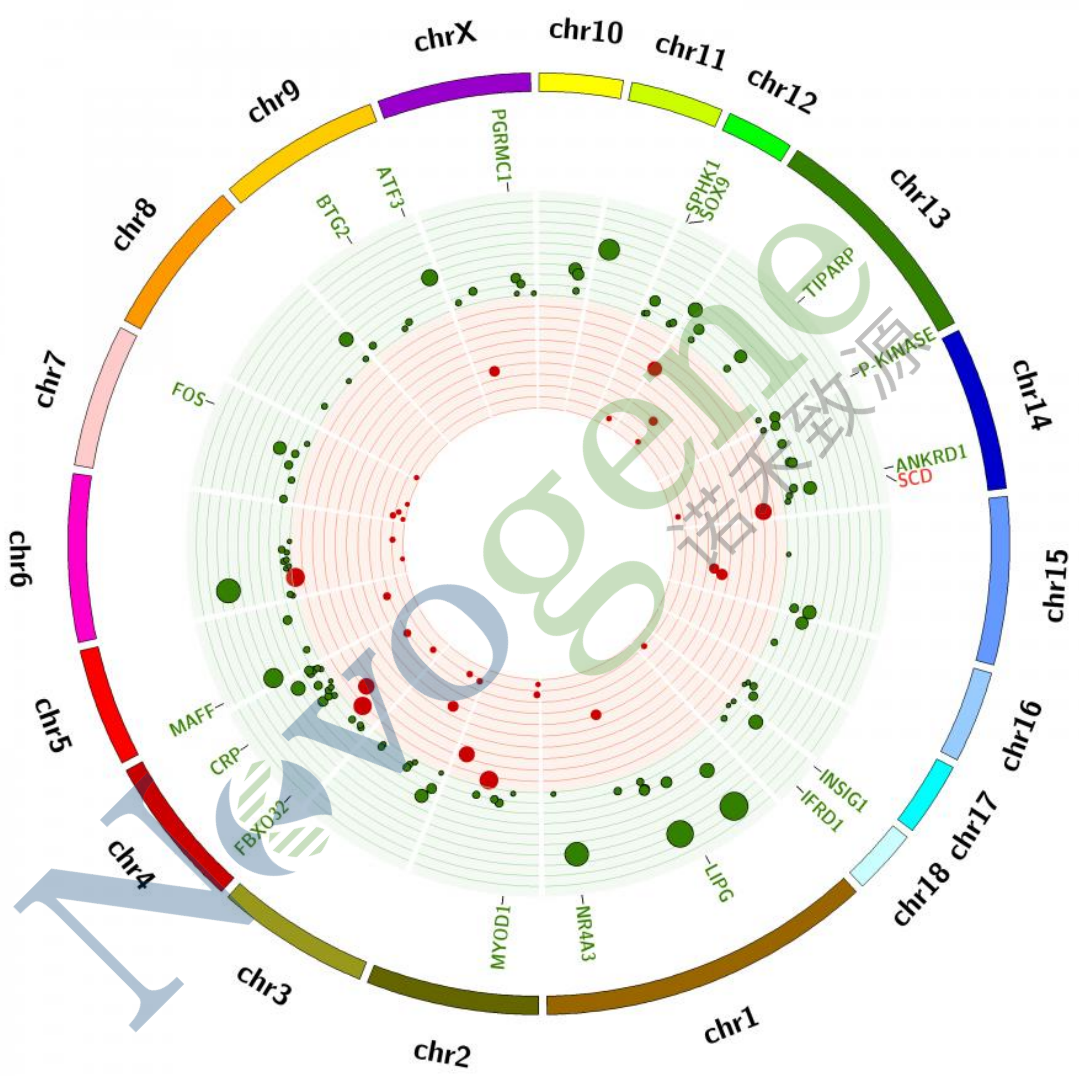


图 6.1 差异基因在染色体上的分布图

此图展示差异基因在染色体分布的 circos 图，绿色为下调差异基因，红色为上调差异基因，并在图中标记了重要基因的名称。

北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

27.2 多信息数据可视化 circos 图

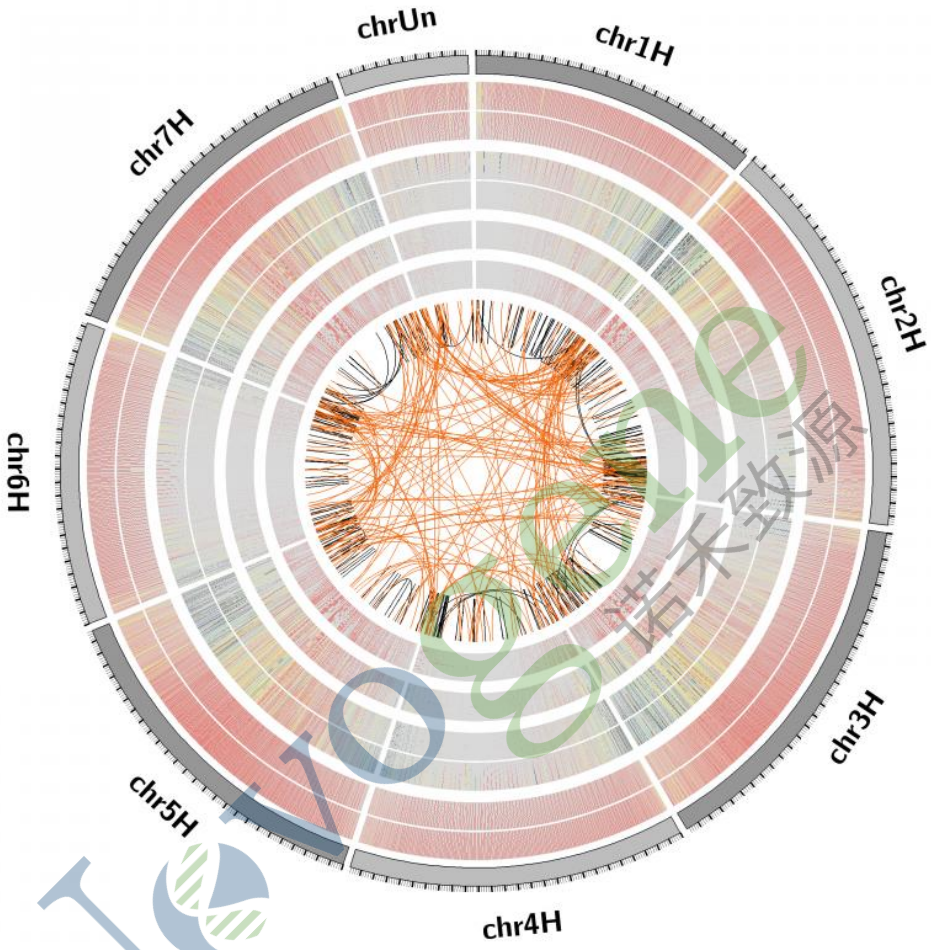


图 6.2 多信息数据可视化 circos 图

此图为多个分析数据可视化的展示 circos 图，从外向内方向，第一圈：染色体位置，第二圈：外层为基于原注释文件的基因分布，内层为基于含新基因注释文件的基因分布，第三圈：外层为基于原注释文件的转录本分布，内层为基于含新基因注释文件的转录本分布，第四圈：AS 分布，第五圈：lncRNA 分布，第六圈：融合基因分布，黑色线代表 intra-chromsome，橘黄色线代表 inter-chromosome。

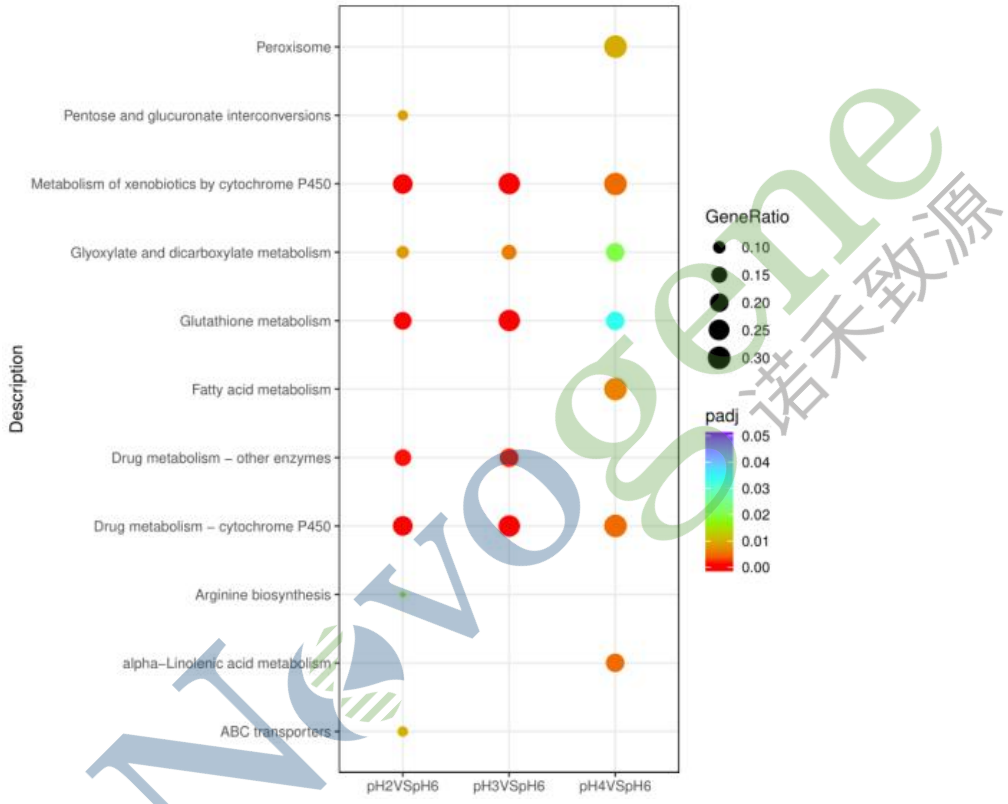
北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

28 多组合气泡图的绘制

28.1 分析背景及介绍

在一张气泡图同时展示多个差异比较组合中目的通路的富集情况。

28.2 结果展示



Novogene 诺禾致源

29 个性化相关系数图

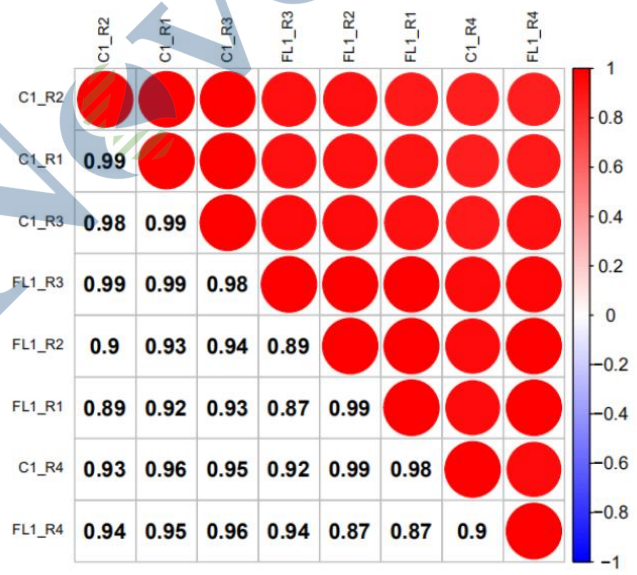
生物学重复是任何生物学实验所必须的，高通量测序技术也不例外(Hansen et al.)。生物学重复主要有两个用途：一个是证明所涉及的生物学实验操作是可以重复的且变异不大，另一个是为了确保后续的差异基因分析得到更可靠的结果。样品间基因表达水平相关性是检验实验可靠性和样本选择是否合理的重要指标。相关系数越接近 1，表明样品之间表达模式的相似度越高。Encode 计划建议皮尔逊相关系数的平方(R2)大于 0.92(理想的取样和实验条件下)。具体的项目操作中，我们要求生物学重复样品间 R2 至少要大于 0.8，否则需要对样品做出合适的解释，或者重新进行实验。为了更美观，可以利用 corrplot 包绘制相关性图。

29.1 分析方法

R 语言 corrplot 包

输入文件：表达量 fpkm 文件

29.2 结果展示



横纵坐标为样本名称，图标为样本间相关性系数，绝对值越大，表示相关性越大。

30 PCA 散点图绘制

30.1 分析背景及介绍

主成分分析（principal component analysis，PCA）是将多个变量通过线性变换以选出较少个数重要变量的一种多元统计分析方法，是常用的数据降维手段。原先的 n 个维度通过线性变换，变成了新的 n 个线性无关的按方差解释度排序的主成分（principal component， PC）。主成分针对的就是方差解释度。

参考链接：

https：//www.jianshu.com/p/8994afcaa757

https：//www.cnblogs.com/leezx/p/6120302.html

30.2 分析使用数据及方法

使用 R 中的 prcomp 做 PCA 图（rotation，维度旋转），使用 PC1 和 PC2 数值，通常会根据 PC1 或 PC2 值挑选部分基因进行标注。

30.3 结果展示

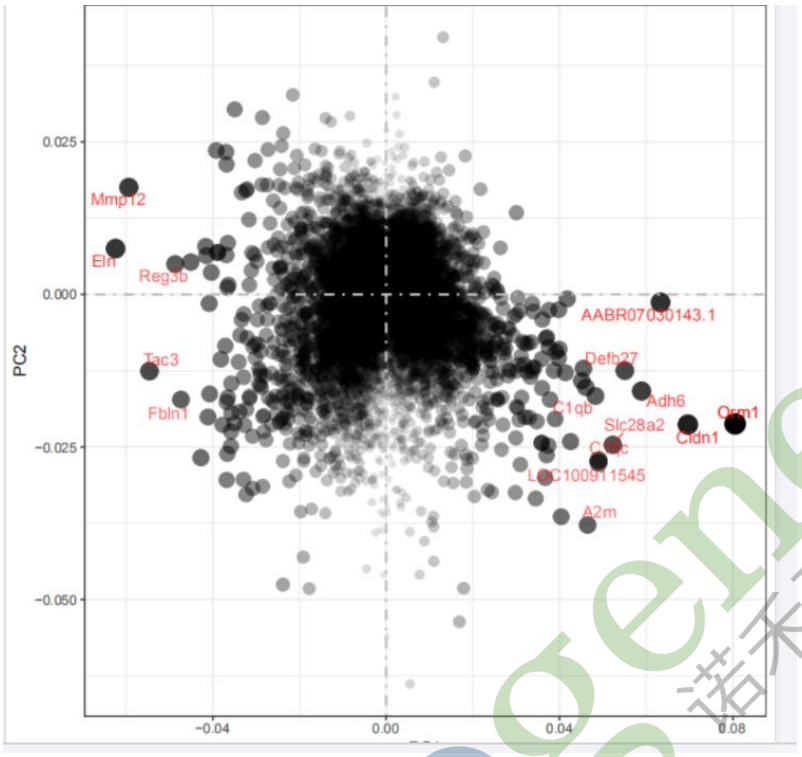
30.3.1 PC 值示例

```
Standard deviations (1, .., p=10):
 [1] 1.610853e+00 1.479822e+00 1.331726e+00 1.282957e+00 8.474849e-01 6.767500e-01 6.082400e-01 4.931341e-01 8.029403e-02 3.945500e-17

Rotation (n x k) = (10 x 10):
              PC1         PC2         PC3         PC4         PC5         PC6         PC7         PC8         PC9         PC10
变量.1  -0.04934193  0.23172642  0.35528789 -0.60701248 -0.038072712 -0.3053328  0.0006953244 -0.04029625  0.49930289 -0.32182186
变量.2   0.42112442  0.32978763 -0.19235190 -0.12045663  0.118650201  0.2018372 -0.5344838809  0.56518371  0.03539948 -0.03479072
变量.3   0.48178172 -0.29027152  0.10326713  0.10245666 -0.369320361  0.3836190  0.0135165199 -0.21789437  0.05283402 -0.57310922
变量.4   0.04654857  0.08031718  0.42419033  0.55536960  0.267812166 -0.1043861  0.3461959304  0.46406197  0.19322793 -0.21874087
变量.5  -0.26887169  0.42051262 -0.07093209 -0.04142334 -0.634521528  0.3531292  0.3625426561  0.27592267  0.02714837  0.09222002
变量.6   0.15949239  0.16929757  0.56222955  0.20947848 -0.438677136 -0.3086376 -0.3799214151 -0.10542960 -0.26438953  0.26993496
变量.7   0.33775547  0.41790066 -0.31597038 -0.02029094  0.004448267 -0.4573559  0.3442102234 -0.10995076 -0.43435004 -0.28741482
变量.8  -0.14465776  0.32942364 -0.35274383  0.50067706 -0.093562088 -0.1265287 -0.2880524712 -0.34418423  0.50570773 -0.11622293
变量.9   0.39607655  0.38020121  0.20781937 -0.03517242  0.307355632  0.3825787  0.2484042221 -0.40493285  0.15670258  0.40383102
变量.10  0.44308931 -0.33708781 -0.23996232 -0.01303115 -0.279383926 -0.3412708  0.2352767697  0.18531279  0.41197113  0.42232049
```


北京诺禾致源科技股份有限公司 网址：www.novogene.com 邮箱：service@novogene.com 电话：010-8283 7801

30.3.2 PCA 散点图添加标注



Novogene 诺禾致源

31 fusioncatcher（医学专用）

31.1 分析背景及方法介绍

基因融合（gene fusion）是指由于某种机制（如基因组变异）使得两个不同基因的部分序列或全部序列融合到一起，形成了一个新的基因。大量研究表明，基因融合与各种疾病，特别是癌症的发生发展密切相关，基因融合也是当前组学大数据分析中的一项重要研究内容。常见的三种发生机制分别是染色体易位（Chromosomal Translocation）、染色体缺失（Interstitial deletion）以及染色体倒位（Chromosomal Inversion）。目前用于融合基因分析的软件有数十种，而 FusionCatcher 该款软件则是通过 Bowtie、BLAT、STAR 和 Bowtie2 这 4 种方式同时检出融合位点，然后将结果与软件内部数据库信息再次进行校正，最终提高融合基因检出结果的真阳性率。

31.2 结果展示图

Gene_1_symbo	Gene_2_sym	Fusion_desc	Counts_of_c	Spanning_p	Spanning_u	Longest_an	Fusion_find	Fusion_poir	Fusion_poir	Gene_1_id(5	Gene_2_id(3	Exon_1_id(5	Exon_2_id(3	Fusion_sequ	Predicted_effect
FGFR3	TACC3	known,adja	0	857	77	42	BOWTIE+BL	4:1806934:+	4:1727977:+	ENSG00000	ENSG00000	013810		AGCAGCTG	in-frame
FGFR3	TACC3	known,adja	0	857	10	42	BOWTIE+BL	4:1807033:+	4:1727832:+	ENSG00000	ENSG00000	013810		CCGTCCCC	intronic/CDS(truncated)
FIP1L1	PDGFRA	known,onc	0	118	48	47	BOWTIE+BL	4:53425965:	4:54274925:	ENSG00000	ENSG00000	134853		TCAGCCGA	in-frame
GOPC	ROS1	known,onc	0	73	20	29	BOWTIE;BO	6:117566854	6:117321394	ENSG00000	ENSG00000	ENSE00003	ENSE00000	GCTAGTTG	in-frame
GOPC	ROS1	known,onc	0	73	2	26	BOWTIE	6:117566854	6:117320030	ENSG00000	ENSG00000	ENSE00003	ENSE00003	GCTAGTTG	out-of-frame
IGH@	CRLF2	known,onc	0	32	10	95	BOWTIE+SP	14:10586325	X:10579929(	ENSG09000	ENSG00000	205755		CGACGGGG	---/intergenic
IGH@	CRLF2	known,onc	0	32	5	32	BOWTIE+ST	14:10586326	X:1228687:-	ENSG09000	ENSG00000	205755		CGTGGTCC	---/intergenic
IGH@	CRLF2	known,onc	0	32	3	96	BOWTIE+SP	14:10586334	X:10581254(	ENSG09000	ENSG00000	205755		TGCTGATGT	---/intergenic
HOOK3	RET	known,onc	0	13	10	29	BOWTIE;BO	8:42968214:	10:43116584	ENSG00000	ENSG00000	ENSE00003	ENSE00001	GGCCAACG	in-frame
AKAP9	BRAF	known,onc	0	12	13	30	BOWTIE;BO	7:92003235:	7:140787584	ENSG00000	ENSG00000	ENSE00003	ENSE00003	TGATAAAC	in-frame
EWSR1	ATF1	known,onc	0	11	10	30	BOWTIE;BO	22:29287134	12:5081428(	ENSG00000	ENSG00000	ENSE00003	ENSE00003	CCAAGTCA	in-frame
TMPRSS2	ETV1	known,onc	0	10	11	37	BOWTIE+ST	21:41494381	7:13935844:	ENSG00000	ENSG00000	006468		GGGTCCTG	in-frame
EWSR1	FLI1	known,onc	0	9	12	29	BOWTIE;BO	22:29287134	11:12880718	ENSG00000	ENSG00000	ENSE00003	ENSE00003	CCAAGTCA	in-frame
ETV6	NTRK3	known,onc	0	7	9	29	BOWTIE	12:11869969	15:87940753	ENSG00000	ENSG00000	ENSE00001	ENSE00001	TCCCCGCC	in-frame
BRD4	NUTM1	known,onc	0	6	9	29	BOWTIE;BO	19:15254152	15:34347965	ENSG00000	ENSG00000	ENSE00001	ENSE00001	TCCAGTGA	in-frame
CD74	ROS1	known,onc	0	6	7	30	BOWTIE;BO	5:150404680	6:117324415	ENSG00000	ENSG00000	ENSE00000	ENSE00000	AGGCACTC	in-frame
CIC	DUX4	known,cosm	0	5	5	34	BOWTIE+ST	19:42295047	4:19017444(	ENSG00000	ENSG00000	260596		CTGCCCCTG	in-frame
DUX4	IGH@	known,canc	0	5	5	84	BOWTIE+SP	4:190174997	14:1902874(	ENSG00000	ENSG09000	000014		CGCAACCT	CDS(truncated)/---

注释：

（1）Gene_1_symbol（5end_fusion_partner）：5'末端融合伴侣的基因符号

（2）Gene_2_symbol（3end_fusion_partner）：3'末端融合伴侣的基因符号

（3）Fusion_description　：融合基因的类型

（4）Counts_of_common_mapping_reads：同时映射在形成融合基因的两个基因上的读数计数。这表明形成融合基因的基因的 DNA / RNA 序列有多相似（即它们的同源性是什么，因为高度同源的基因倾向于显示为候选融合基因）。在涉及形成融合基因的基因的序列完全不同的情况下，则此处期望其值为零

（5）Spanning_pairs：支持融合的读取对数（还包括多重映射读取）

（6）Spanning_unique_reads：融合结上唯一读取（即唯一映射位置）映射的计数

（7）Longest_anchor_found：在融合结的唯一读取映射中找到最长的锚

（8）Fusion_finding_method：用于比对读图和寻找融合基因的比对方法。以下是两种使用的方法：（i）BOWTIE =仅 Bowtie aligner 用于映射基因组上的读数和外显子-外显子融合连接；（ii）BOWTIE + BLAT = Bowtie aligner 用于将基因组上的读数作图和 BLAT 用于作图读段以发现融合结，（iii）BOWTIE + STAR = Bowtie aligner 用于对基因组作图读，而 STAR 用于作图读以发现融合结，（iv）BOWTIE + BOWTIE2 = Bowtie aligner 用于定位基因组上的读

北京诺禾致源科技股份有限公司　网址：www.novogene.com　邮箱：service@novogene.com　电话：010-8283 7801

数，Bowtie2 用于定位读数上的融合结。

（9）Fusion_point_for_gene_1（5end_fusion_partner）：融合连接 5'端的染色体位置

（10）Fusion_point_for_gene_2（3end_fusion_partner）：融合连接 3'端的染色体位置

（11）Gene_1_id(5end_fusion_partner) ：5'末端融合伴侣的 Ensembl 基因 ID

（12）Gene_2_id(3end_fusion_partner) ：3'末端融合伴侣的 Ensembl 基因 ID

（13） Exon_1_id(5end_fusion_partner) ：5'端融合外显子-外显子连接的集合外显子 ID

（14） Exon_2_id(3end_fusion_partner) ：3'端融合外显子-外显子连接的 Ensembl 外显子 ID

（15）Fusion_sequence：推断的融合断点

（16） Predicted_effect：使用 Ensembl 数据库中注释预测候选融合基因的结果。

31.3 参考文献：

[1] FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data.