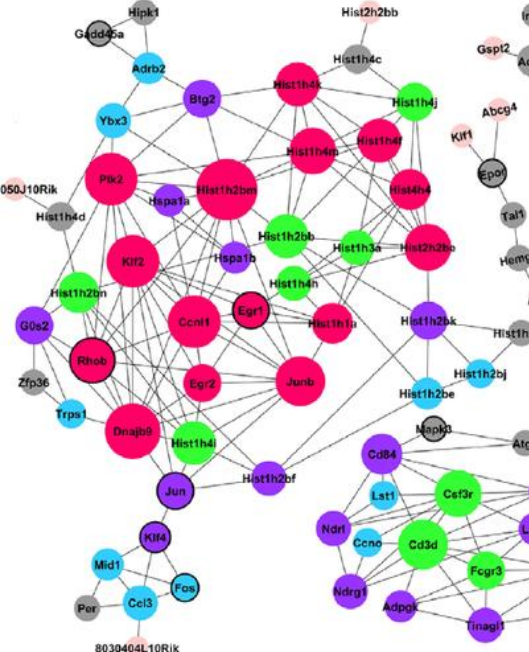


转录组 RNAseq 生物信息分析结题报告

-- 数据分析与结果展示



RNA-seq 是一种定量分析转录组的高通量测序技术。这种技术可以用来测量基因的表达水平，检测可变剪接，发现新基因和转录本，检测发生在外显子区域的基因突变，检测基因融合等等。相比于传统的芯片杂交技术(microarray)，RNA-seq 有着精度更高，应用更广泛的优点。自从 2008 年首篇文章(Marioni J. C., et al)发表以来，RNA-seq 的实验技术已经非常完善，成为一种常规的高通量实验技术，并且已经成为转录组分析的主要工具

上海嘉因生物有限科技公司

地址：上海市杨浦区赤峰路 65 号 611 室

技术服务热线: 021-61539657

網址: www.rainbow-genome.com

邮箱: marketing@rainbow-genome.com



目录

一、 建库测序流程.....	4
1. Total RNA 样品检测.....	4
2. 文库构建和质检.....	5
3. 上机测序.....	6
二、 生物信息分析流程.....	7
三、 结果展示及说明.....	8
1. 原始序列数据.....	8
2. 测序数据质量评估.....	8
2.1 测序错误率分布检查.....	8
2.2 碱基含量分布检查.....	10
2.3 Adapter Content 分析.....	10
2.4 测序数据过滤.....	11
3. 参考序列比对分析.....	12
3.1 Reads 与参考基因组比对情况统计.....	13
3.2 Reads 在参考基因组不同区域的分布情况.....	14
3.3 Reads 在染色体上的密度分布情况.....	15
4. 可变剪接分析.....	15
4.1 可变剪接注释及差异分析统计.....	16
4.2 差异可变剪接可视化.....	18
5. 基因表达水平分析.....	18
5.1 基因表达水平分析.....	18
5.2 基因表达相关性分析&样本组成成分分析.....	19
5.3 差异表达基因筛选.....	20
5.4 差异基因的 GO 富集分析.....	23
5.5 差异基因的 KEGG 富集分析.....	24
5.6 差异基因的 GSEA 富集分析.....	27
5.7 差异基因中核心基因 Signal-Net 分析.....	28
6. lncRNA 表达水平分析.....	29
6.1 lncRNA cis-regulation 预测.....	29
6.2 lncRNA trans-regulation 预测.....	30

6.3	lncRNA_mRNA 联合表达网络	
	分析	32

四.	Methods 英文版	33
五.	参考文献	36
	关于公司	37

一. 建库测序流程

从 RNA 样品到最终数据获得，样品检测、建库、测序每一个环节都会对数据质量和数量产生影响，而数据质量又会直接影响后续信息分析的结果。为了从源头上保证测序数据的准确性、可靠性，嘉因生物对样品检测、建库、测序每一个生产步骤都严格把控，从根本上确保了高质量数据的产出。流程图如图 1 所示：

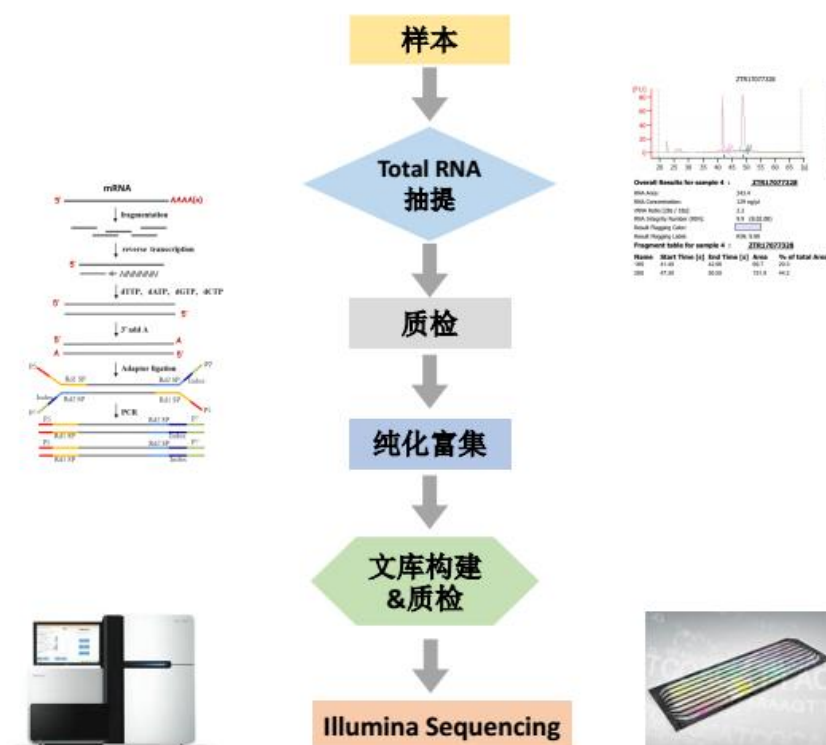


图 1. 建库测序流程示意图

1. Total RNA 样品检测

Total RNA 样品的检测方法主要包含以下 4 种：

- (1) 琼脂糖凝胶电泳分析 RNA 降解程度以及是否有污染；
- (2) Nanodrop 检测 RNA 的纯度（OD260/280 比值）；
- (3) Qubit 对 RNA 浓度进行精确定量；
- (4) Agilent 2100 精确检测 RNA 的完整性；

2. 文库构建和质检

mRNA 的获取主要有两种方式: 一是利用真核生物大部分 mRNA 都带有 polyA 尾的结构特征, 通过 Oligo(dT)磁珠富集带有 polyA 尾的 mRNA。二是从总 RNA 中去除核糖体 RNA, 从而得到 mRNA。随后在 NEB Fragmentation Buffer 中用二价阳离子将得到的 mRNA 随机打断, 按照 NEB 普通建库方式或链特异性建库方式进行建库。

NEB 普通建库: 以片段化的 mRNA 为模版, 随机寡核苷酸为引物, 在 M-MuLV 逆转录酶体系中合成 cDNA 第一条链, 随后用 RNaseH 降解 RNA 链, 并在 DNAPolymerase I 体系下, 以 dNTPs 为原料合成 cDNA 第二条链。纯化后的双链 cDNA 经过末端修复、加 A 尾并连接测序接头(1), 用 AMPure XP beads 筛选 200bp 左右的 cDNA, 进行 PCR 扩增并再次使用 AMPure XP beads 纯化 PCR 产物, 最终获得文库。建库原理如下图左所示。

链特异性建库: 逆转录合成 cDNA 第一条链方法与 NEB 普通建库方法相同, 不同之处在于合成第二条链时, dNTPs 中的 dTTP 由 dUTP 取代, 之后同样进行 cDNA 末端修复、加 A 尾、连接测序接头和长度筛选, 然后先使用 USER 酶降解含 U 的 cDNA 第二链再进行 PCR 扩增并获得文库。链特异性文库具有诸多优势, 如相同数据量下可获取更多有效信息; 能获得更精准的基因定量、定位与注释信息; 能提供反义转录本及每一 isoform 中单一 exon 的表达水平。建库原理如下图右所示。

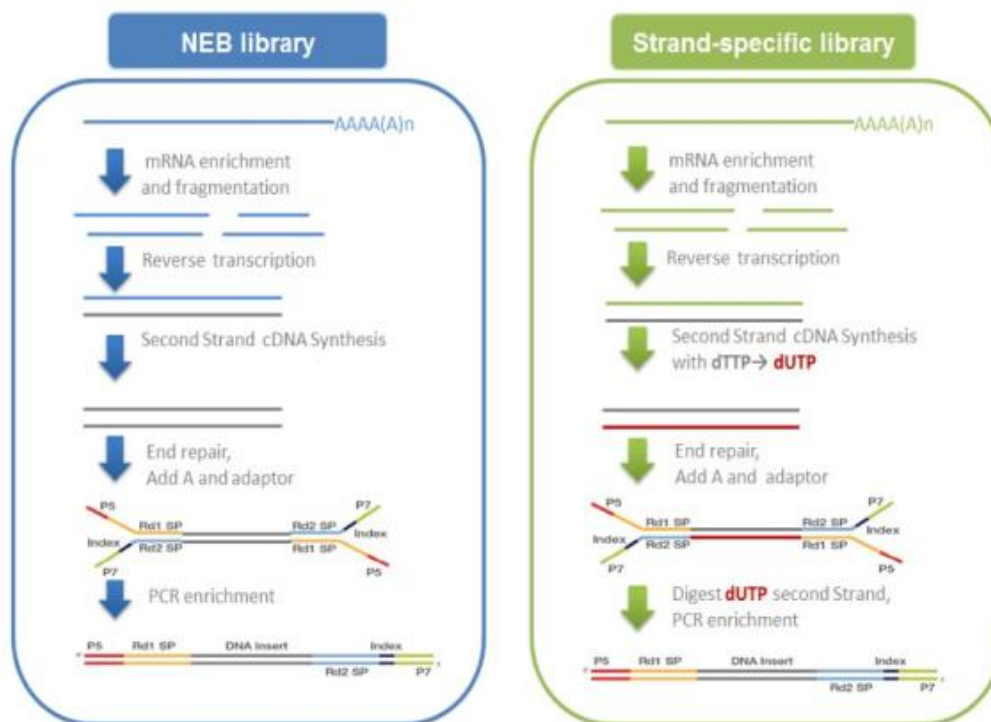


图 1.2 文库构建流程示意图

文库构建完成后，先使用 Qubit2.0 进行初步定量，稀释文库至 1ng/ul，随后使用 Agilent 2100 对文库的插入片段长度（insert size）进行检测，insert size 符合预期后，使用 Q-PCR 方法对文库的有效浓度进行准确定量（文库有效浓度 > 2nM），以保证文库质量。

3. 上机测序

库检合格后，把不同文库按照有效浓度及目标下机数据量的需求 pooling 后进行 Illumina HiSeq 测序。测序基于边合成边测序（Sequencing by Synthesis)的原理，在序的 flow cell 中加入四种荧光标记的 dNTP、DNA 聚合酶以及接头引物进行扩增，在每一个测序簇延伸互补链时，每加入一个被荧光标记的 dNTP 就能释放出相对应荧光，测序仪通过捕获荧光信号，并通过计算机软件将光信号转化为测序峰，从而获得待测片段的序列信息。测序过程如下图所示：

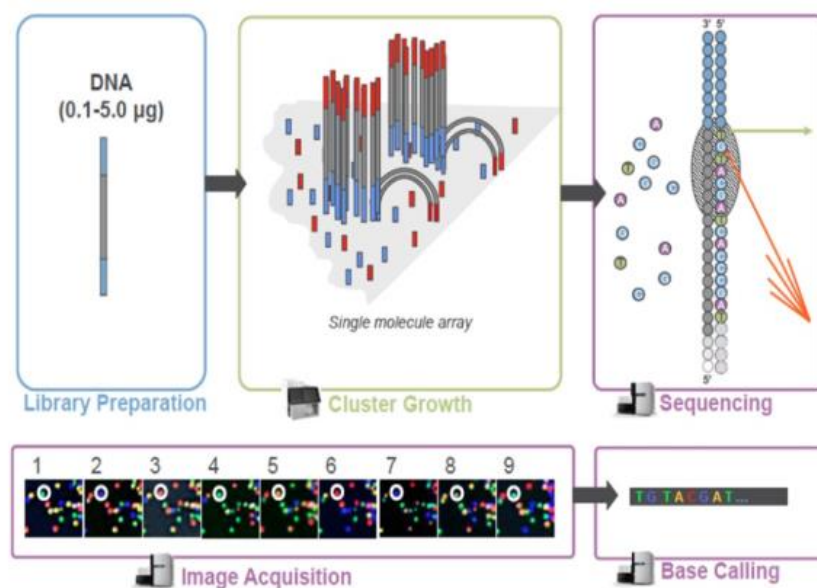


图 1.3 Illumina HiSeq 测序流程示意图

二. 生物信息分析流程

测序数据下机之后，我们就可以获得原始测序序列(Sequenced Reads)，在有相关物种参考序列或参考基因组的情况下，我们通过如下(图 2)标准流程进行生物信息分析：

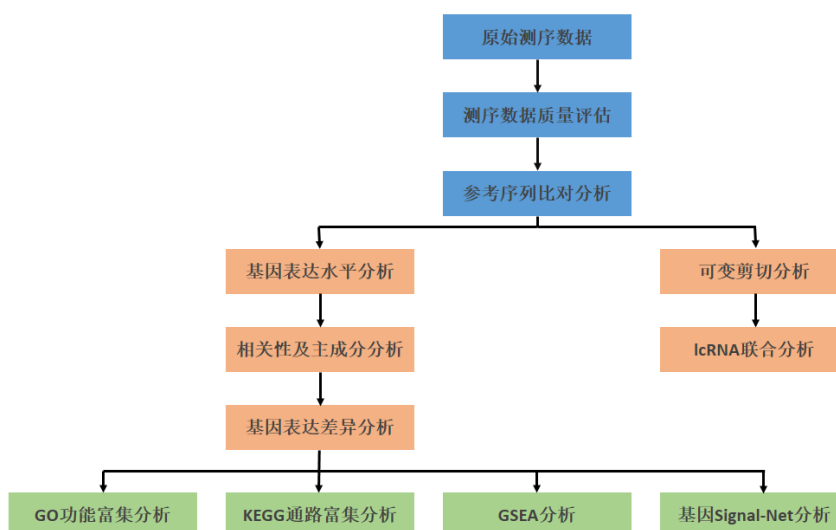


图 2 生物信息分析流程

对于上图分析内容，若其存在于合同信息分析内容中，则进行此项分析；若不存在，则不进行。

三. 结果展示及说明

1. 原始序列数据

NGS (Next Generation Sequencing, 二代高通量测序) 技术应用高通量测序仪 (如: illumina Hiseq 2000/2500、Miseq) 对 cDNA 进行测序。首先得到的原始图像文件, 然后经过碱基识别及误差过滤, 最终得到可以用于分析的原始测序片段, 我们称之为 Reads, 结果以 fastq 格式储存, 它包括序列的碱基组成信息以及其对应的序列质量信息, 双端测序(pair-end)会分为两个 Reads 文件: _1, _2 两个文件。

```
@ST-E00493:238:H5CVYCCXY:4:1101:1377:1801 1:N:0:CCTCCT↓
GCCTGGCGTTCCTCCGCGGTTCCTGTACAAAGGCGACGACAAGTCCCGGGTCCGCGAGCCGCTCCGCGCCATCCACGAGTCGCCCTCCGTTATCCTGGGCC
+↓
AAAFFAAAF<JFA<FJJJJ<FJJJJJFJFFAJJ7F-7A---<FJFJJJJJ7J<A7A7-F---7A7F7AA-7AA-AAJJFJ-A7-7A-7F-AA-FAJJ7FJAAFF7<-JF-AA7F<7-7)FFJF77---->-7
@ST-E00493:238:H5CVYCCXY:4:1101:3630:1801 1:N:0:CCTCCT↓
GGATTGAGCCGCCGATTTTTTAACCTAGATCTCGAAATGCATCGTATTCTGTCCATTGGACTGTAAGGTTTATGTAGGCAATCTTGGAACAATGGCAACAA
+↓
AAAFFJFJJFJ<AFJJ<JJJJJJFJ7FFJFJFJJFFFAFFJAJAJFJJJJ7AAAAA-FAAAJAFJFJF<FAJJAAFJ77FJAAFF7<FJJJJ77F<AFF-7FJA<F-F--AAJ<F--><AA7-))
```

- 第一行: reads 名称, 通常以@开头, 随后为 Illumina 测序标识符和描述信息;
- 第二行: 碱基序列;
- 第三行: 以+开头, 存储与第一行相同的信息或为缺省值;
- 第四行: 碱基的测序质量值, 该行字符为第二行对应碱基的质量值加上 33 后转换为 ASCII 码, 逆向转化即可直观得到每个碱基的质量信息。

2. 测序数据质量评估

2.1 测序错误率分布检查

每个碱基测序错误率是通过测序 Phred 数值(Phred score, Q_{phred})通过公式 (公式 1 : $Q_{phred} = -10\log_{10}(e)$) 转化得到, 而 Phred 数值是在碱基识别(Base Calling)过程中通过一种概率模型计算得到, 这种模型可以准确地预测碱基判别的错误率。Phred 分值, 不正确的碱基识别率, 碱基正确识别率以及 Q-score 的对应关系如下表所显示:

表 1 质量与错误率对照

测序错误率(E)	碱基正确识别率	测序质量值(Q)
5%	90%	13
1%	99%	20
0.1%	99.9%	30
0.01%	99.99%	40(max Q)

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

(1) 测序错误率随着测序序列(Sequenced Reads)长度的增加而升高。这是由测序过程中化学试剂的消耗导致的, 为 Illumina 高通量测序平台所具有的特征。

(2) 前 6 个碱基具有较高的测序错误率, 此长度恰好为 RNA-seq 建库过程中反转录所需的随机引物长度。前 6 个碱基测序错误率较高是因为随机引物和 RNA 模版的不完全结合(Jiang et al.)。

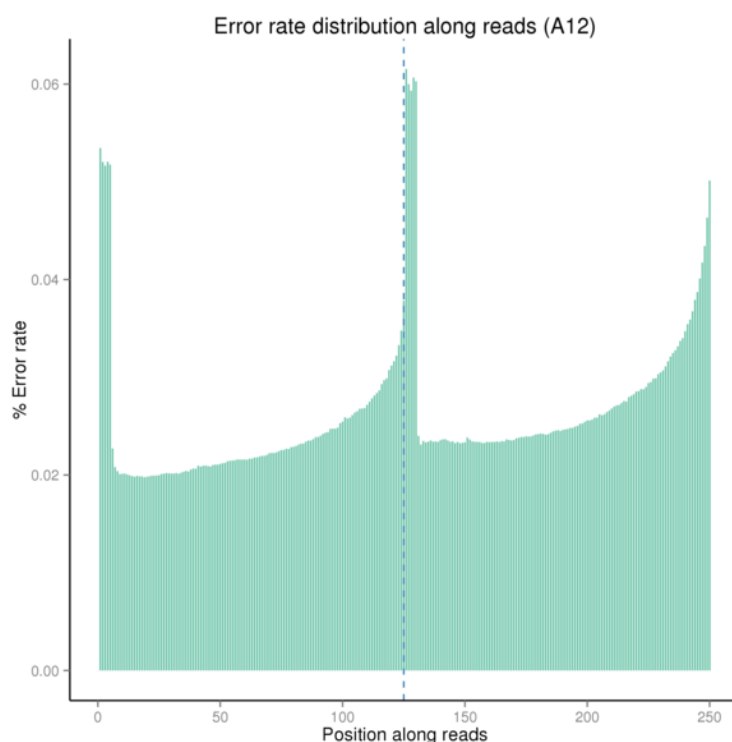


图 2.2.1 测序错误率分布图

2.2 碱基含量分布检查

通过碱基分布示意图，图 3.2.2，我们可以观测测序过程碱基均衡性，一般来说，因采用随机引物扩增但随机引物种类有限而导致 reads 的前 10 个碱基的比例不均衡，会出现波动是正常现象。在此波动之后的碱基互补配对原理 GC 及 AT 碱基对会分别均衡分布。

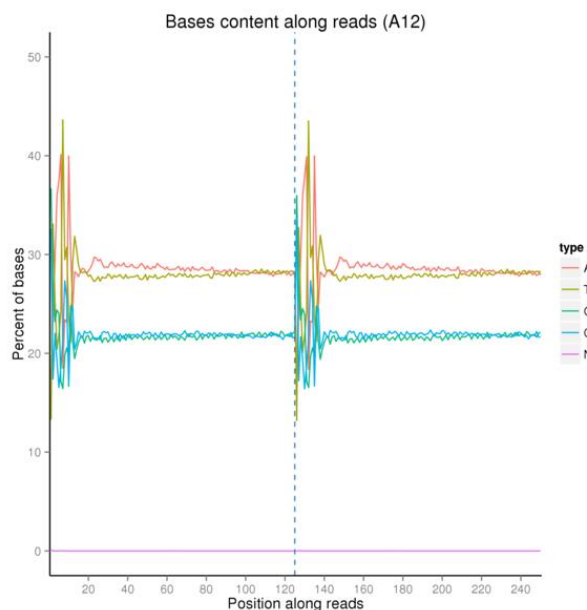


图 3.2.2 碱基分布示意图

2.3 Adapter Content 分析

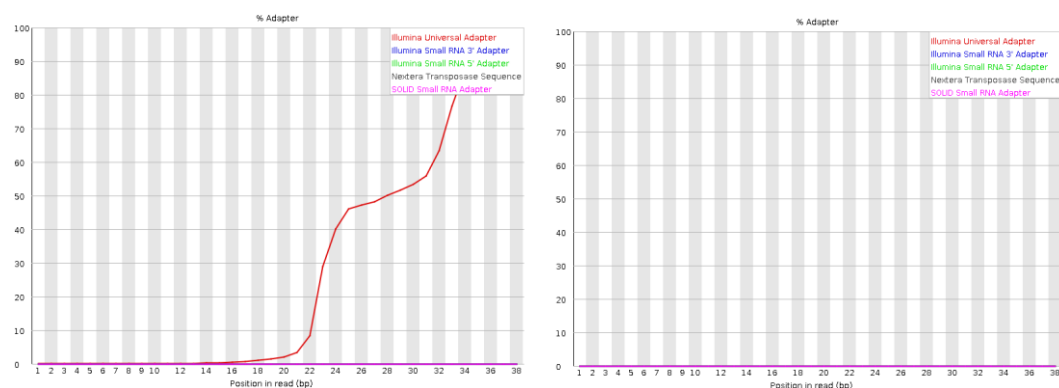


图 3.2.3 接头去除前(左图)后(右图)Adapter Content 统计图

注：

- 1、Adapter Content 反映序列 reads 在每个接头序列文库的立即百分比率；
- 2、正常情况，在完全去除接头情况下，序列 reads 在接头序列文库中占比是趋于

零的，即 Adapter Content 图形中曲线贴近基线。

3、接头信息：3' adapter, 5' -AGATCGGAAGAGCACACGTCT-3'；

5' adapter, 5' -GTTTCAGAGTTCTACAGTCCGACGATC-3

4、这里只展示 treat1 样本去除接头前后的 Adapter Content 统计结果

2.4 测序数据过滤

测序下机得到的原始序列，包含有带接头及低质量的 reads，为了保证信息分析质量，必须对 raw reads 进行过滤，得到 clean reads，后续分析都基于 clean reads。数据的处理大致如下：

- 1) 去除带接头(adapter)的 reads；
- 2) 去除 N(N 表示无法确定碱基信息)的比例大于 10%的 reads；
- 3) 去除低质量 reads($Q_{\text{phred}} \leq 20$ 的碱基数占整个 read 长度的 50% 以上的 reads)。

Classification of Raw Reads (A11)

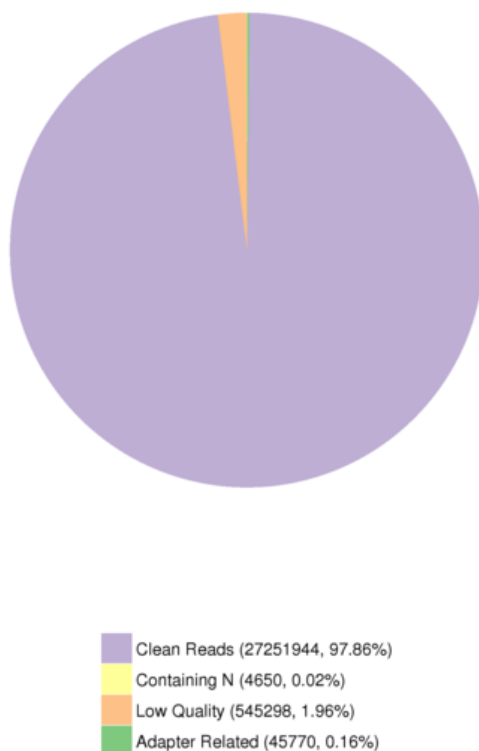


图 3.2.4 原始数据组成

注：不同颜色的比例分别代表不同成分比例

- (1)Adapter related: 因有接头, 过滤掉的 reads 数及其占总 raw reads 数的比例。
- (2)Containing N: 因 N 含量超过 10%, 过滤掉的 reads 数及其占总 raw reads 数的比例。
- (3)Low quality: 因低质量, 过滤掉的 reads 数及其占总 raw reads 数的比例。
- (4)Clean reads: 最终得到的 clean reads 数及其占总 raw reads 数的比例。

3. 参考序列比对分析

参考序列比对(Reads Mapping)是指将经过下机处理的原始数据(Sequenced Reads)比对到参考基因组上。嘉因生物采用主流分析软件 STAR 对 RNA-seq 测序数据进行比对分析。STAR 采用 Maximal Mappable Prefix (MMP) 搜索方法, 可以对 junction reads 进行精确定位, 如下图所示, 其综合性能在同类比对软件中表现较为突出。

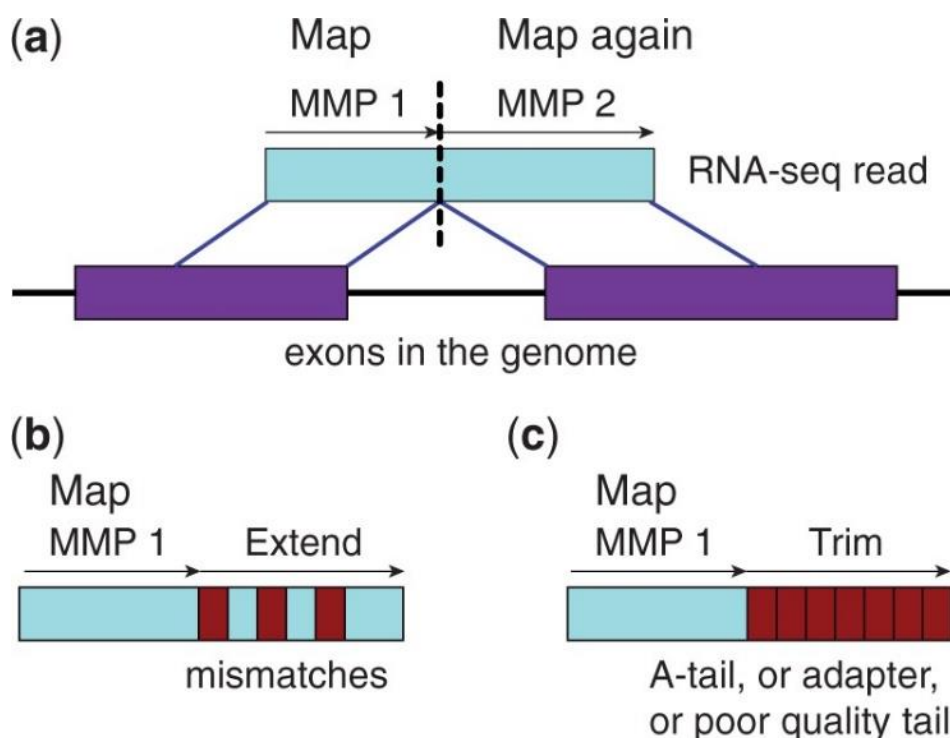


图 3.3 数据回帖分析原理

3.1 Reads 与参考基因组比对情况统计

表 2 参考序列比对结果统计表

	Control1	Control2	Control3	treat1	treat2	treat3
Number of input reads	23624094	23967327	24004478	21566180	21623553	21555629
Average input read length	300	300	300	300	300	300
UNIQUE READS:						
Uniquely mapped reads number	21894148	21858030	21986062	19906147	20067278	19758528
Uniquely mapped reads %	92.68%	91.20%	91.59%	92.30%	92.80%	91.66%
Average mapped length	296.92	296.65	296.56	296.81	297.05	296.63
Number of splices: Total	23702756	23606514	23938109	21700625	21958587	21420931
Number of splices: Annotated (sjdb)	23161183	23113486	23444298	21214359	21509837	20962507
Number of splices: GT/AG	23439822	23352787	23678758	21464862	21727207	21193348
Number of splices: GC/AG	207449	201223	205535	188090	184709	180643
Number of splices: AT/AC	18286	17244	17639	15743	16268	15773
Number of splices: Non-canonical	37199	35260	36177	31930	30403	31167
Mismatch rate per base, %	0.31%	0.36%	0.38%	0.33%	0.35%	0.36%
Deletion rate per base	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%
Deletion average length	2.16	2.19	2.14	2.16	2.11	2.19
Insertion rate per base	0.02%	0.01%	0.01%	0.01%	0.01%	0.01%
Insertion average length	1.51	1.53	1.53	1.51	1.53	1.53
MULTI-MAPPING READS:						
Number of reads mapped to multiple loci	401512	421307	413327	353764	372265	382722
% of reads mapped to multiple loci	1.70%	1.76%	1.72%	1.64%	1.72%	1.78%
Number of reads mapped to too many loci	3546	3865	3874	3315	3129	3567
% of reads mapped to too many loci	0.02%	0.02%	0.02%	0.02%	0.01%	0.02%
UNMAPPED READS:						
% of reads unmapped: too many mismatches	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
% of reads unmapped: too short	5.57%	6.99%	6.63%	6.00%	5.42%	6.50%
% of reads unmapped: other	0.04%	0.04%	0.04%	0.04%	0.04%	0.04%

注：

- 1) Number of input reads: RNA-seq 原始数据下机处理后，参考序列比对分析中输入序列数目统计；
- 2) Average input read length: 序列平均长度统计；
- 3) Uniquely mapped reads number: 特异性回帖至参考基因组上的序列数目统计，即序列只回帖至参考基因组某一特定位置；
- 4) Uniquely mapped reads %: 特异性回帖至参考基因组上的序列百分比统计；
- 5) Average mapped length: 回帖至参考基因组上的序列平均长度统计；
- 6) Number of splices: Annotated(sjdb): 已经注释的可变剪接数目统计（即这部分可变剪接出现在参考文档.gtf 中）；
- 7) Number of non-canonical splices: 未经注释的可变剪接数目统计；
- 8) Mismatch rate per base: 错误匹配百分比统计（一般 $\leq 0.8\%$ ）；
- 9) Deletion rate per base: 缺失百分比统计；

- 10) Deletion average length: 缺失长度平均值统计;
- 11) Insertion rate per base: 插入百分比统计;
- 12) Insertion average length: 插入长度平均值统计;
- 13) Number of reads mapped to multiple loci: 回帖至参考基因组多个位点的序列数目统计;
- 14) % of reads mapped to multiple loci: 回帖至参考基因组多个位点的序列百分比统计;
Number of reads mapped to too many loci: 如果序列回帖至参考基因组的位置个数大于 10, 则被认为是改序列为 “too many loci”;
- 15) % of reads mapped to too many loci: “too many loci” 百分比统计;
- 16) % of reads unmapped: too many mismatches: 序列中包含太多的错配而造成的回帖失败;
默认值为序列长度的 30%(0.3*read length)或 10 个以上错配;
- 17) % of reads unmapped due to too short: 序列太短造成的回帖失败; 默认值为 2/3 序列长度;
- 18) % of reads unmapped due to other reasons: 其他原因造成的回帖失败。

3.2 Reads 在参考基因组不同区域的分布情况

将比对到基因组上的 reads 分布情况进行统计, 定位区域分为 Exon(外显子)、Intron(内含子)、Intergenic(基因间区)和。在基因组注释较为完全的物种中, 比对到 Exon (外显子) 的 reads 含量最高, 比对到 Intron (内含子) 区域的 reads 来源于 pre-mRNA 的残留及可变剪接过程中发生的内含子滞留事件导致的, 而比对到 Intergenic (基因间区) 的 reads 是因为基因组注释不完全。

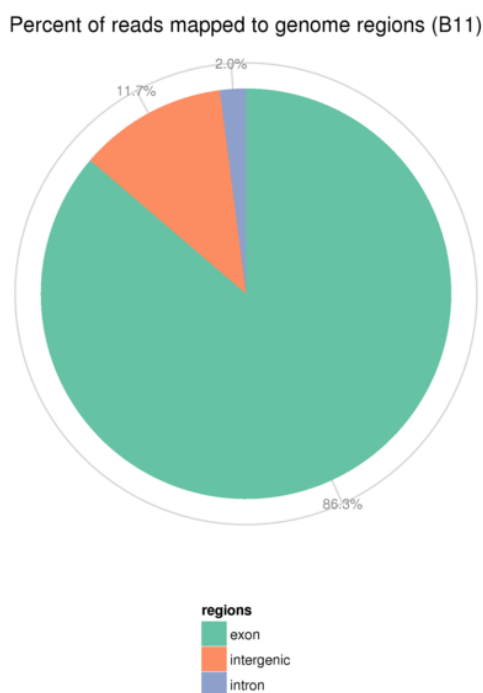


图 3.3.2 Reads 在参考基因组不同区域的分布情况

3.3 Reads 在染色体上的密度分布情况

对 Total mapped reads 比对到基因组上各个染色体（分正负链）的密度进行统计，如下图所示，正常情况下，整个染色体长度越长，该染色体内部定位的 reads 总数会越多(Marquez et al. 2012)。从定位到染色体上的 reads 数与染色体长度的关系图中，可以更加直观看出染色体长度和 reads 总数的关系。（图中最多只展示其中 15 条染色体）

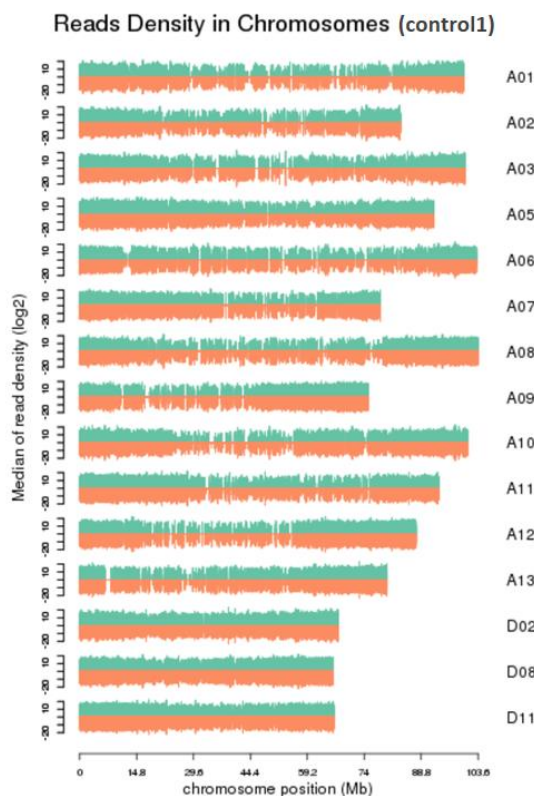


图 3.3.3 Reads 在染色体上的密度分布图

4. 可变剪接分析

针对可变剪接分析(Alternative splicing, AS)，我们采用 rMATS 作分析，它是一种从重复的 RNA-seq 数据中检测不同 AS 的强大而灵活的统计分析方法，它不仅可以对可变剪接事件进行分类，还可以进行不同样本间可变剪接事件的差异分析。除了进行非配对重复样品的分析，我们的方法还包括了专为配对重复样品分析的模式，比如针对对照模式匹配的临床 RNA-seq 数据集的案例。

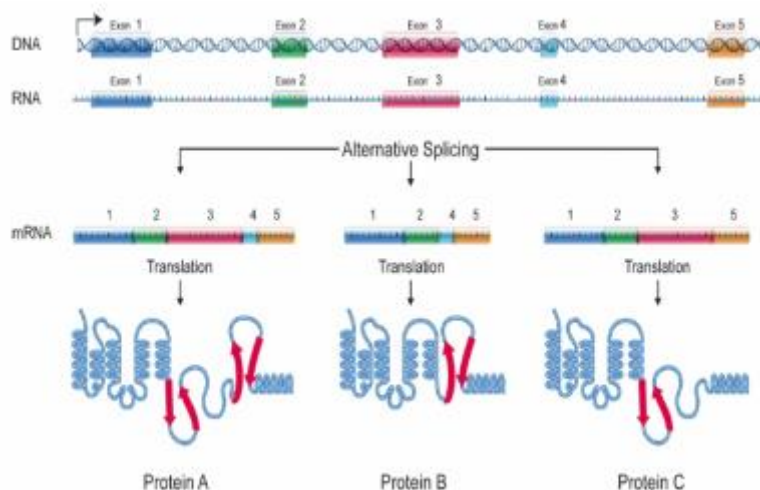


图 3.4.a 可变剪接形成示意图

rMATS 软件对可变剪接事件分类如下图所示：

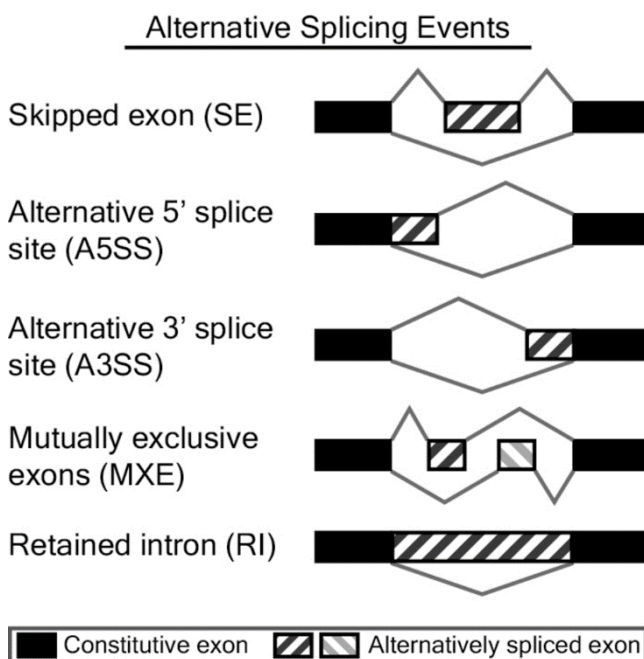


图 3.4.b 可变剪接分类示意图

4.1 可变剪接注释及差异分析统计

可变剪接差异分析包括可变剪接事件定量及表达差异显著性分析。每个可变剪接事件对应两个 Isoform，分别为 Exon Inclusion Isoform 和 Exon Skipping Isoform，如下图所示。分别对两个 Isoform 进行表达量统计，并除以其有效长度，得到校正后表达量，然后计算 Exon Inclusion Isoform 在两个 Isoform 总表达量的比值，

最后进行差

异显著性分析，结果如下表所示。

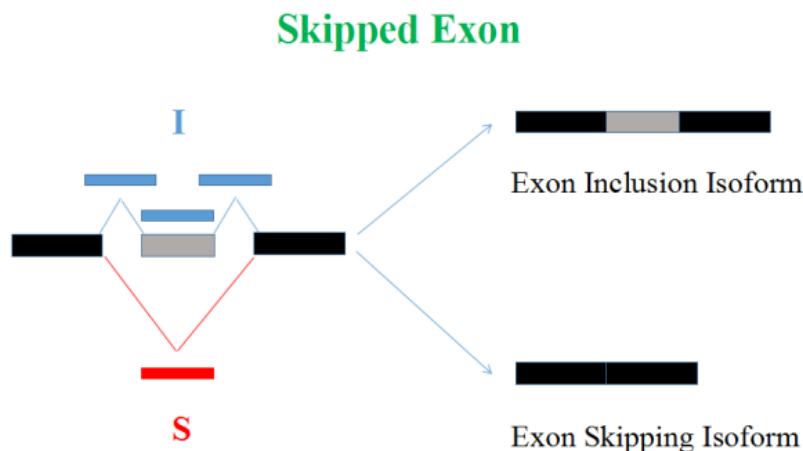


图 3.4.1 可变剪接结果统计示意图

表 3 可变剪接差异分析结果

ID	GeneID	geneSymbol	chr	strand	exonStart	exonEnd	upstreamES	upstreamEE	downstreamES
18017	ENSG00000140416	TPM1	chr15	+	63353911	63353987	63353396	63353472	63354413
18035	ENSG00000140416	TPM1	chr15	+	63353396	63353472	63353067	63353138	63353911
18038	ENSG00000140416	TPM1	chr15	+	63340746	63340906	63336255	63336351	63349183
19449	ENSG00000039560	RAI14	chr5	+	34813678	34813765	34812284	34812313	34814687
21227	ENSG00000104852	SNRNP70	chr19	+	49605370	49606844	49604646	49604728	49607890
31373	ENSG00000112081	SRSF3	chr6	+	36567597	36568053	36566625	36566760	36568928
33032	ENSG00000138326	RPS24	chr10	+	79799961	79799983	79796951	79797062	79800372
33035	ENSG00000138326	RPS24	chr10	+	79799958	79799983	79796951	79797062	79800372
34795	ENSG00000198467	TPM2	chr9	-	35684728	35684743	35684484	35684547	35685060
35439	ENSG00000163359	COL6A3	chr2	-	238303229	238303268	238296224	238296827	2.38E+08

注：

(1)ID：可变剪接事件编号

(2)GeneID：可变剪接事件所在基因编号

(3)geneSymbol：可变剪接事件所在基因名称

(4)chr：可变剪接事件所在染色体

(5)strand：可变剪接事件所在链的方向

(6)exonStart_0base：发生该可变剪切事件的外显子起始位置

(7)exonEnd：发生该可变剪切事件的外显子终止位置

(8)upstreamES：发生该可变剪切事件上游 exon 起始位置

(9)upstreamEE：发生该可变剪切事件上游 exon 终止位置

(10)downstreamES: 发生该可变剪切事件

下游 exon 起始位置

(11)downstreamEE: 发生该可变剪切事件下游 exon 终止位置

4.2 差异可变剪接可视化

对于表达差异显著性的可变剪接事件，进行可视化展示，如图 3.4.2 所示。图中跨外显子比对的 reads 使用连接外显子 junction 边界的弧线表示。弧线的粗细和比对到 junction 上的 reads 数成正比，同时弧线上的数字指出了 junction reads 的数目。

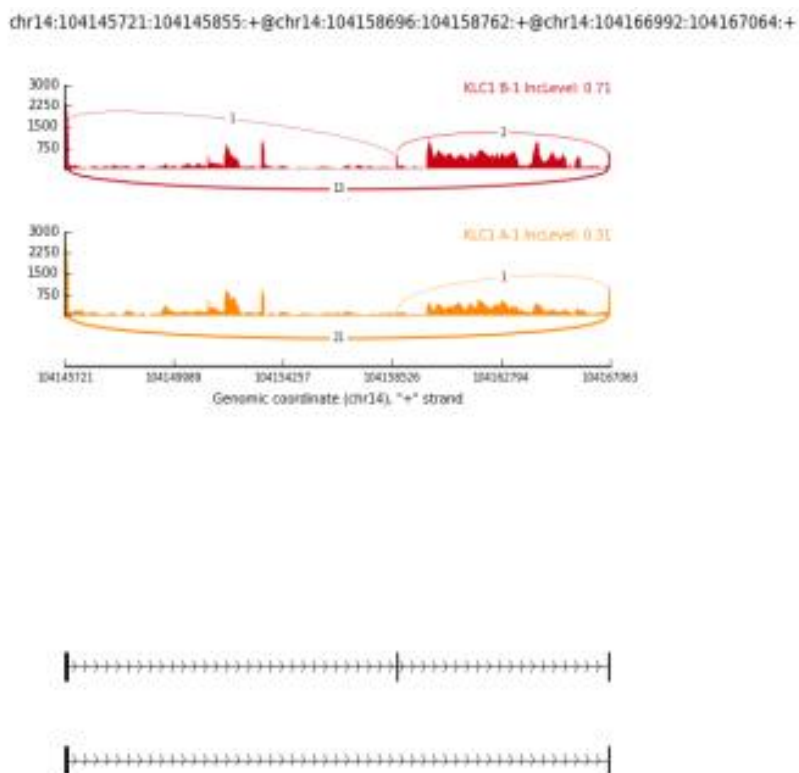


图 3.4.2 差异可变剪接事件可视化展示图

5. 基因表达水平分析

5.1 基因表达水平分析

通过回贴定位到基因区域的 Reads 数目来估计相应基因的表达水平。Reads 数目的多少直接反应了基因的表达量，因此直接统计基因上回贴到的 Reads 数是目前最为常用的评估基因表达水平的方法。

表 4 基因表达水平统计样表

Gene Symbol	Treat1	Treat2	Treat3	Control1	Control2	Control3
-------------	--------	--------	--------	----------	----------	----------

A1BG	3	1	8	4	8	9
A1BG-AS1	0	0	0	0	0	0
A1CF	589	538	460	625	474	509
A2M	1	0	2	3	0	6
A2M-AS1	5	4	7	2	6	2
A2ML1	1	0	0	1	0	0
A2MP1	0	0	0	0	0	0
A3GALT2	0	0	0	0	0	0
A4GALT	527	560	527	513	594	649
A4GNT	0	3	0	2	0	2
...

注：

- 1) 每一列为样本，每一行为基因，表格中数字为基因在对应样本中的表达值，即 read count;
- 2) 完整列表见附件“ / 结果文件/phase1-AllExpGenes/expression_level_all.xls”中。

5.2 基因表达相关性分析&样本组成成分分析

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。根据客户提供的样本我们进行样本基因表达的相关性分析，具体结果如图 3.5.2.a。

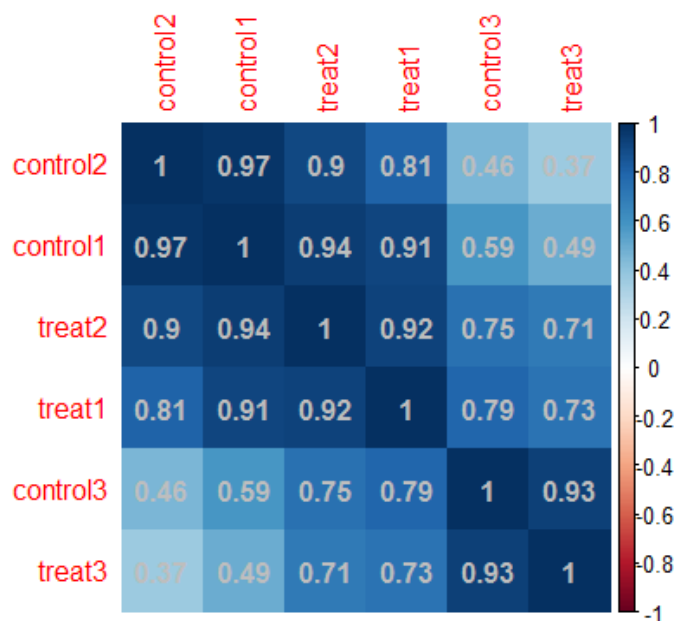


图 3.5.2.a 样本基因先关性分析

当样本数目较多时, 根据各样本所有基因的表达水平(read counts), 我们可以对样品进行主成分分析(Principal Component Analysis, PCA)。主成分分析是一种将多个变量通过线性变换以选出较少个数重要变量的多元统计分析方法, 它通过降维的手段将大量影响表达模式的基因变换成少数几个主成分, 尽可能多地保留原始变量的信息, 且彼此间互不相关, 将这几个主成分作线性组合, 作为新的综合指标来对样品进行聚类分析, 具体结果如图 3.5.2.b5.2.2 所示

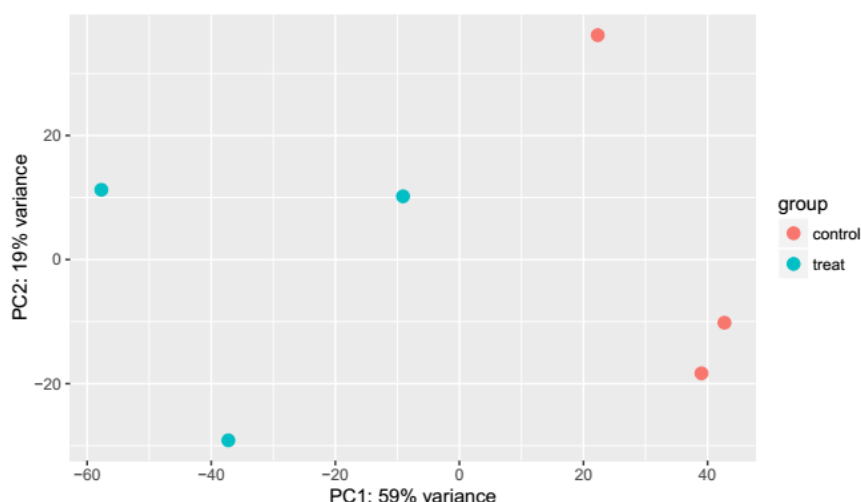


图 3.5.2.b 样本表达水平的 PCA 聚类图

5.3 差异表达基因筛选

在得到基因表达值以后, 差异基因筛选是整个 RNA-seq 分析的关键一步。现在科研领域中针对 RNA-seq 数据差异基因筛选的软件非常多, 如 edgeR, limma, 或者 DEseq 等等。选择什么样的软件和参数都会对结果产生相当大的影响, 因此这一步骤中软件 and 参数的选择显得至关重要。我们根据客户的数据特点, 选择合适软件进行差异基因筛选。


 差异表达基因结果展示:

表 5. 差异基因结果示意图

gene_id	baseMean	log2FC	lfcSE	stat	pval	padj
PER1	423.3092719	-0.572766157	0.168512867	-3.398946125	0.00067646	0.999509376
UNC5B	42.94226372	-1.093134279	0.34716899	-3.148709452	0.001639932	0.999509376
CHRN1	65.68120407	0.753102655	0.254082558	2.964007692	0.003036607	0.999509376

NCOA5	346.6340181	0.497894443	0.171045329	2.910891781	0.003603988	0.999509376
CLDN7	151.9938813	-0.551411491	0.192098016	-2.870469468	0.004098628	0.999509376
MRPL32	657.9713521	0.305511511	0.106912607	2.857581726	0.004268827	0.999509376
gene id	treat1	treat2	treat3	Treat4	Control1	Control2
PER1	392.646376	293.5164911	334.9691862	561.9314193	544.8259897	411.966169
UNC5B	25.7641979	21.55191018	34.98095932	65.27486183	65.80491528	44.27673779
CHRNA1	81.41486538	80.04995211	85.86235469	45.40859954	53.22456383	48.1268889
NCOA5	350.3930915	449.5112696	417.6514537	255.4233724	259.3487837	347.4761379
CLDN7	127.7904216	113.9172395	128.2635175	210.9607853	164.5122882	166.5190356
MRPL32	714.1835659	747.1328864	720.8197678	601.6639438	599.0182729	565.0096757

注：

- 1) gene_id: 基因名称，默认格式为 gene symbol ID;
- 2) baseMean: 基因在样本中的平均表达数值统计;
- 3) log2FC: log2(Fold Change), 实验组与对照组比值取 LOG;
- 4) lfcSE: log2FC 的标准方差，用以算 p-value 值;
- 5) stat: logFC 与 lfcSE 的比值，原则上服从 t 分布，用来计算 p-value; 该数值如果为正数，代表基因上调；该数值如果为负数，代表基因下调;
- 6) pval(p-value): 基因差异表达显著性统计参数，数值越小代表显著性越高；一般认为 $p\text{-value} < 0.05$ 即为显著性具有统计意义;
- 7) padj(adjusted p-value): 是经过多重假设检验调整过的 p-value;
- 8) 本次课题中差异表达基因分析选用软件 Deseq2;
- 9) 所有基因表达值完整列表见附件“/结果文件/phase2-DiffExpGenes/Treat_vs_control_diff.xls”。

差异表达基因结果统计：

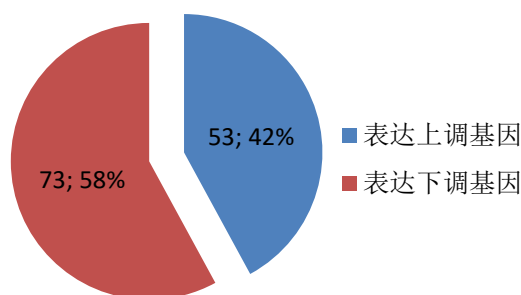


图 3.5.3.a 差异表达基因结果统计饼图

注：

- 1) 差异表达基因一上调基因：2105 个；完整列表见附件“/结果文件/phase2-DiffExpGenes/genes_up.xls”;
- 2) 差异表达基因一下调基因：2613 个；完整列表见附件“/结果文件/phase2-DiffExpGenes/genes_down.xls”;
- 3) 差异表达基因筛选条件：cutoff: $p\text{-val} \leq 0.05$ 。

差异表达基因热图结果展示：

热图分析可以更直观的看到差异基因在各个处理组中的表达值，并且 hierarchical 聚类也会使得差异表达基因聚类，便于后续分析。

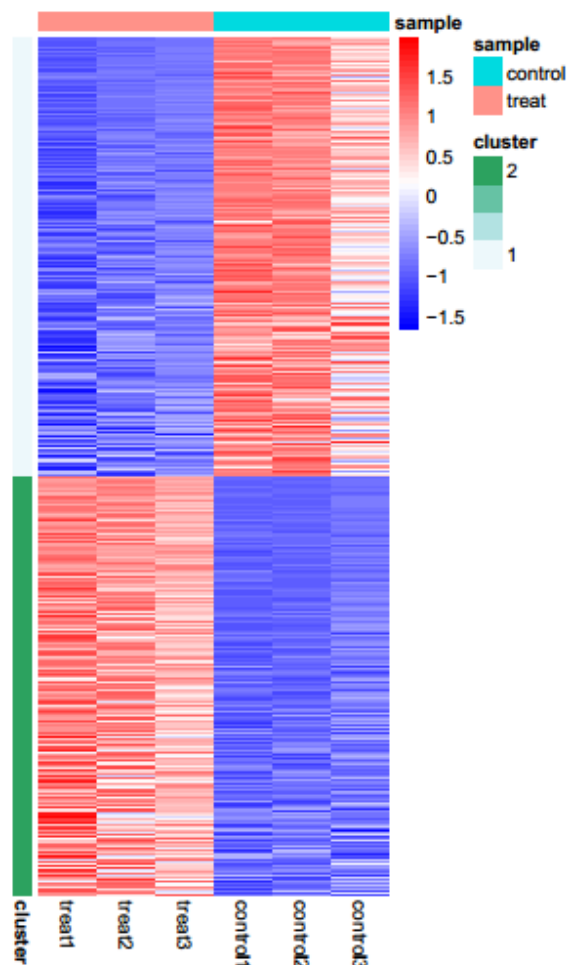


图 3.5.3.b 差异表达基因结果统计饼图

注：

- 1) 每一行代表一个差异表达基因，红色表示上调，蓝色表示下调；
- 2) 每一列代表一个样本，从左往后分别为 3 个实验组，3 个对照组样本；
- 3) treat1, treat2, treat3 分别对应 3 个实验组；
- 4) control1, control2, control3 对应 3 个对照组；
- 5) 根据基因表达量变化将基因分为两类：上调基因、下调基因；

原始文档见附件“/结果文件/phase2-DiffExpGenes/Treat_vs_control_diff.pdf”。

差异表达基因火山图展示：

火山图可直观显示表达差异显著性基因的整体分布情况，横坐标表示基因在不同样本中的表达倍数变化($\log_2\text{FoldChange}$)，纵坐标表示表达差异的显著性水平($-\log_{10}\text{padj}$)。若比较组合无表达差异显著性基因，默认调整筛选表达差异显著性的阈值进行火山图的绘制。上调基因用红色点表示，下调基因用绿色点表示，如

图 3.5.3.c 所示。

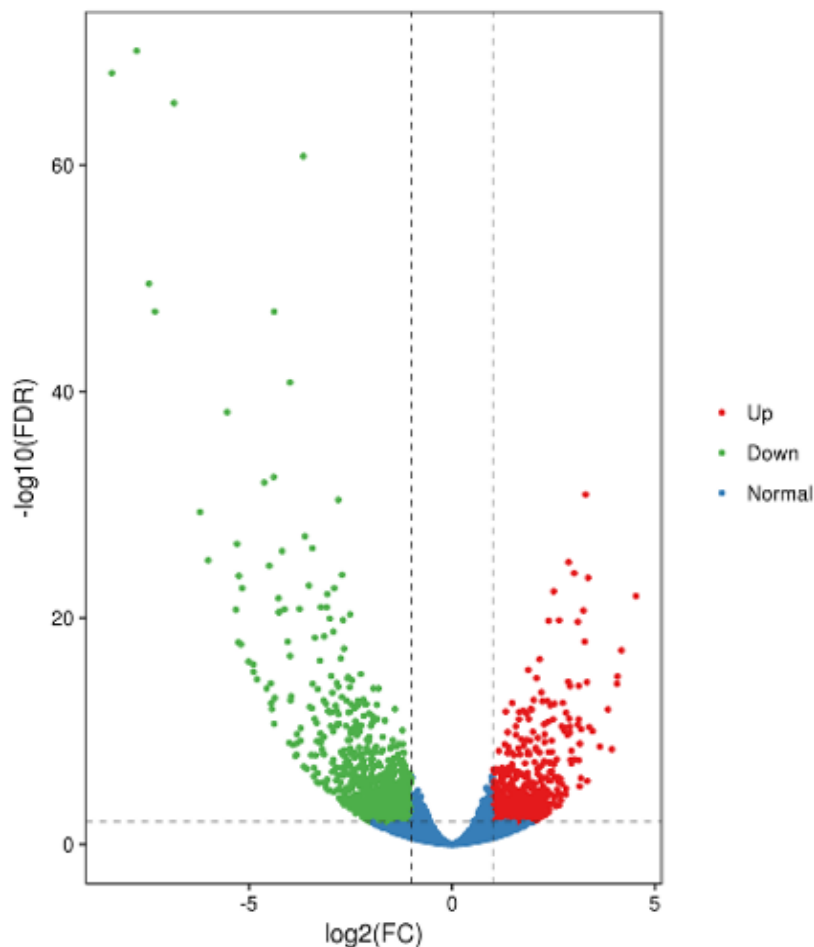


图 3.5.3.c 差异表达基因结果统计饼图

原始文档见附件“/结果文件/phase2-DiffExpGenes/Treat_vs_control_diff.pdf”。

5.4 差异基因的 GO 富集分析

GO (Gene ontology) 是国际通用的基因功能分类体系，按照基因的细胞成分 (cellular component)，分子功能 (molecule function) 以及生物过程 (biological process) 分为 3 类。在我们拿到了差异基因的集合以后，进行 GO 富集能够看到不同样本在这三大类中的基因分布情况，也可以用于对目标功能的基因的聚集。通过计算数据集的超几何分布，得出基因富集对应 GO terms 的排序。

🌈 差异表达基因 GO 富集分析结果展示（上调 / 下调基因）：

表 6 差异基因的 GO 富集示意图表

Up-regulated Genes Gene Ontology Analysis						
GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0006270	2.74E-06	49.56483516	0.098693503	4	39	DNA replication initiation

GO:0034421	1.87E-05	826.7317073	0.007591808	2	3	post-translational protein acetylation
GO:0006259	2.23E-05	5.693782383	2.469868173	11	976	DNA metabolic process
GO:0006261	2.47E-05	16.63632766	0.349223164	5	138	DNA-dependent DNA replication
GO:0006260	8.57E-05	9.481650071	0.736405367	6	291	DNA replication

...

...

Down-regulated Genes Gene Ontology Analysis

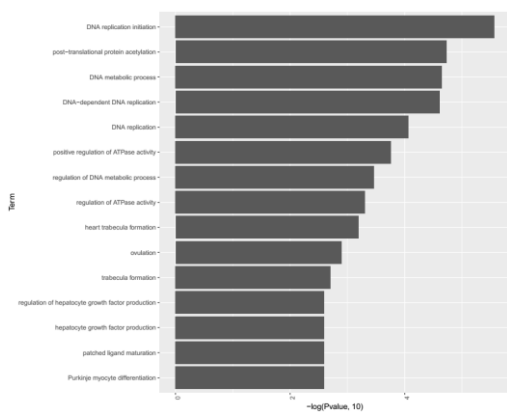
GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0001568	4.76E-06	6.491166078	2.037429379	11	577	blood vessel development
GO:0001944	6.90E-06	6.228924847	2.118644068	11	600	vasculature development
GO:0072358	7.94E-06	6.131799877	2.150423729	11	609	cardiovascular system development
GO:0072359	1.59E-05	4.864259523	3.262711864	13	924	circulatory system development
GO:0019220	2.92E-05	3.783207038	5.71680791	17	1619	regulation of phosphate metabolic process

...

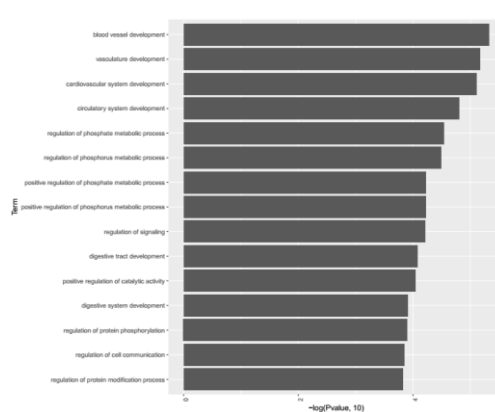
...

注：

- 1) GOBPID: gene ontology biological process ID; GO 数据库中生物学过程分类编号信息;
- 2) Pvalue: 富集分析显著性统计参数;
- 3) Count: 差异表达基因中隶属于该生物学过程的基因个数统计;
- 4) Size: 该生物学过程所包含的所有基因个数统计;
- 5) Term: 生物学过程名称;
- 6) Gene Ontology 分析选用数据库 GO 生物过程 (biological process);
- 7) GO 富集分析按照显著性统计参数 pvalue 排序;
- 8) GO 富集分析显著性筛选阈值: cutoff: pvalue < 0.05;
- 9) 上、下调差异基因 GO 富集分析分别选择 top20 biological process 做结果展示, 完整列表见附件 “/结果文件/phase3-GO_KEGG/up/genes_up_go.xls; /结果文件/ phase3-GO_KEGG/down/ genes_down_go.xls ”。



a. 上调基因 GO 富集分析



b. 下调基因 GO 富集分析

图 3.5.4. 差异基因的 GO 富集示意图

注：选取结果中 top15 富集进行绘图展示。

5.5 差异基因的 KEGG 富集分析

随着科学研究的发展, 科学家们发现不同的基因间存在着相互作用关系。这些基因相互协调, 发挥其生物学功能, 展现出生物学现象。针对这一现象, 科学家们分门别类的制定出各种 pathway。较为著名的有 KEGG (Kyoto Encyclopedia of Genes and Genomes) 数据库。利用 pathway 显著性富集, 确定差异基因所属的 pathway, 进而找出可能相互作用的基因, 挖掘现象背后的机制。


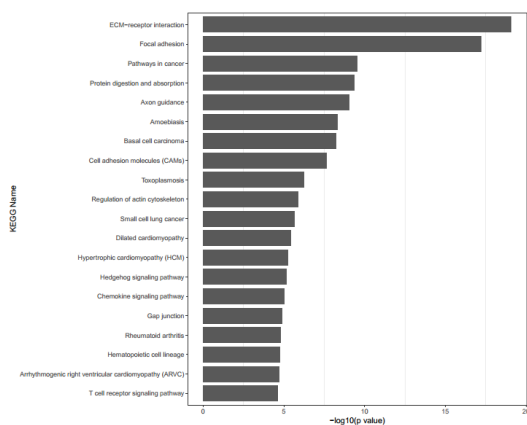
 差异表达基因 KEGG pathway 富集分析结果展示 (上调 / 下调基因):

表 7 差异基因的 KEGG 富集列表

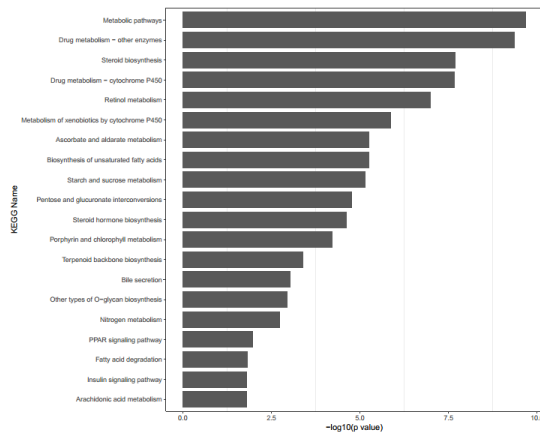
Up-regulated Genes KEGG pathway Analysis						
KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
4512	7.87E-20	10.5046208	5.659773585	35	86	ECM-receptor interaction
4510	5.70E-18	5.369824806	13.16226415	51	200	Focal adhesion
5200	2.81E-10	3.021918626	21.32286792	53	324	Pathways in cancer
4974	4.38E-10	6.210961919	5.133283019	23	78	Protein digestion and absorption
4360	9.12E-10	4.453982344	8.621283019	30	131	Axon guidance
...						...
Down-regulated Genes KEGG pathway Analysis						
KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
1100	2.10E-10	4.88481203	12.60316981	36	1176	Metabolic pathways
983	4.30E-10	21.76313148	0.632301887	10	59	Drug metabolism - other enzymes
100	2.05E-08	50.32307692	0.19290566	6	18	Steroid biosynthesis
982	2.16E-08	13.78965297	0.932377358	10	87	Drug metabolism - cytochrome P450
830	1.00E-07	13.84582543	0.825207547	9	77	Retinol metabolism
...						...

注:

- 1) KEGGID: KEGG pathway ID; KEGG 数据库中生物学通路编号信息;
- 2) Pvalue: 富集分析显著性统计参数;
- 3) Count: 差异表达基因中隶属于该生物学通路的基因个数统计;
- 4) Size: 该生物学通路所包含的所有基因个数统计;
- 5) Term: 生物学通路名称;
- 6) KEGG pathway 富集分析按照显著性统计参数 pvalue 排序;
- 7) KEGG pathway 富集分析显著性筛选阈值: cutoff: pvalue < 0.05;
- 8) 上、下调差异基因 KEGG pathway 富集分析分别选择 top20 biological process 做结果展示, 完整列表见附件 “/结果文件/phase3-GO_KEGG/up/genes_up_kegg.xls; /结果文件/phase3-GO_KEGG/down/genes_down_kegg.xls”。



a. 上调基因 KEGG 富集分析



b. 下调基因 KEGG 富集分析

图 3.5.5.1 差异基因的 KEGG 富集示意图

注：选取结果中 top20 富集进行绘图展示。

差异表达基因在 KEGG pathway 中具体通路信息展示：

以表达上调基因 KEGG pathway 富集分析结果为例，其中表达下调基因 AMPK, ACC, GK, G6PC, PP1 等蛋白因子显著富集在生物学通路“INSULIN SIGNALING PATHWAY”中，并以绿色字体标出。所有显著富集的通路信息见附件“/ 结果文件 / phase3-GO_KEGG”。

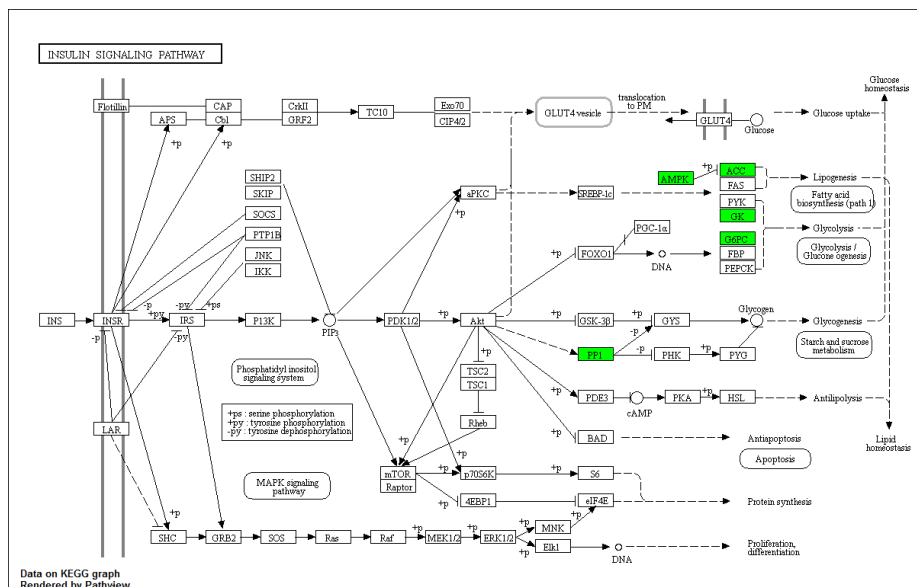


图 3.5.5.2. KEGG pathway 富集分析结果展示示意图

注：

- 1) 绿色标记的基因为本次课题中差异表达基因；
- 2) 由于每一个 pathway 中基因很多，很多方格代表了一类基因而并不是一个基因名；

- 3) 以上图为例，可以在对应的 excel 文件中找到此 pathway，在 Symbol 栏显示了基因。有时，在图中绿色显示了的基因或数字与 Symbol 栏中显示的基因不相同。在这种情况下可以百度或者 google 搜出 KEGG pathway 中绿色基因或数字在 KEGG pathway 的含义，可以发现 Symbol 栏中的一个或多个基因包含在此绿色基因或者数字中，此时此绿色基因或数字代表了一类基因。

5.6 差异基因的 GSEA 富集分析

GSEA 软件是由 broad institute 开发的用于分析差异表达基因所富集的通路和 GO term 的软件。用该软件分析差异表达基因的特征被国际上许多顶尖实验室所采用。GSEA 软件的分析是基于广泛的数据库的，其中较为著名的有 Gene Ontology 数据库和 KEGG 数据库。

该分析的好处是，可以一次性的扫描所有 GO term 以及 pathway，并直观的了解每一个 pathway 或者 GO term 中所有基因的差异表达情况，且该方法不依赖于差异表达的 p-value cutoff。如上三点好处，让 GSEA 备受推崇。

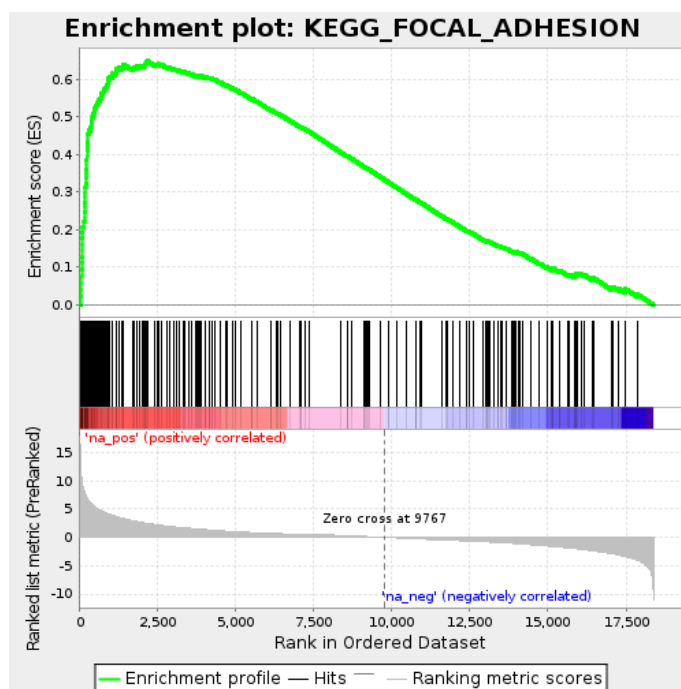


图 3.5.6 GSEA 分析核心图(举例)

注：

- 1) 文件在数据汇总的 GESA 文件夹中 index.html 文件，以上图为例，说明 GSEA 的核心分析图；
- 2) 我们根据基因的差异表达情况，对基因组中的所有基因进行排序，其中上调的基因依次排在最前面（显示为图中中间颜色符红色的部分），下调的基因依次排在最后面（显示为图中中间颜色符蓝色的部分），差异表达不明显的基因排在中间（显示为图中中间颜色符白色的部分）。

红蓝交界的部分)；

- 3) 颜色符上面的黑色竖线 (bar) 代表此图中的基因在排序队列中的位置。若图中大部分黑色竖线对应 (hit) 于红色的颜色符区域, 表明在此图示状态下的大部分基因是上调的。若图中大部分黑色竖线对应 (hit) 于蓝色的颜色符区域, 表明在此图示状态下的大部分基因是下调的；
- 4) 图中最上面的绿色曲线, 与黑色竖线的富集区域相对应。其峰值出现的区域 即为黑色竖线 (bar) 集中出现的区域；

完整信息见附件 “/结果文件/phase4-GSEA/”

5.7 差异基因中核心基因 Signal-Net 分析

Signal-Net 整合所有 KEGG 通路中基因和基因之间的调控关系, 构建基因与基因间的调控网络图。通过 Singal-Net, 可以获得网络中具有核心调控地位的基因, 重要的桥梁基因 (eg. 连接主网络图的基因) 以及核心基因的上下游基因和它们之间的相互关系。

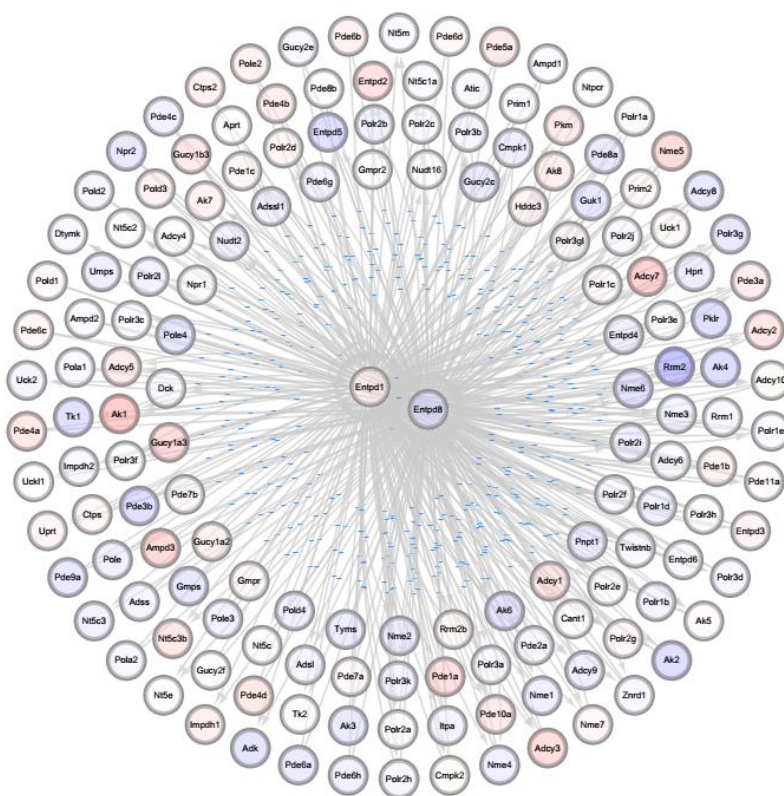


图 3.5.7 Signal-Net 分析核心基因图(部分截图)

注:

- 1) 文件在数据汇总的 Signal-Net 文件夹中, 以上图为例, 图中的节点为基因, 红色代表上调, 蓝色代表下调。红色越深, 代表上调程度越高; 蓝色越深, 代表下调程度越高;
- 2) 各个基因间的连线代表实验验证的来自 KEGG 数据库的调控关系, 可以将图放大后即可

清楚看到：

- 3) 差异表达显著且连接的边众多的节点即为核心基因；核心基因互相连接的即为重要的桥梁基因；
- 4) 原始文档见附件“/结果文件 / phase5-SignalNet/”。

6. lncRNA 表达水平分析

6.1 lncRNA cis-regulation 预测

这里，我们基于 cis_regulation 的原理，定位关键基因。我们首先对每一个 lncRNA 搜索距离其 100kb 以内的 mRNA，组成“lncRNA—mRNA 匹配对”。对这些匹配对，再利用 mRNA 和 lncRNA 的基因表达量进行关联分析，从而定位表达协同的“lncRNA—mRNA 匹配对”。在得到匹配对以后，我们可以通过 mRNA 的功能推测与其匹配的 lncRNA 功能。

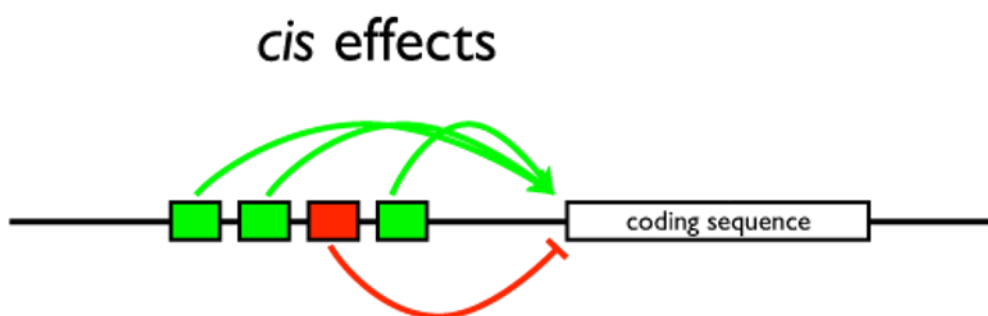


图 3.6.1 cis_regulation 示意图

老师可以通过自己倾向的几个原则之一，利用 excel 对我们的输出结果自行进行简单操作，从而定位关键基因：

1) 老师可以通过 mRNA 或者 lncRNA 的差异表达程度，进一步过滤筛选更为严格的“lncRNA-mRNA 匹配对”。（我们的输出文件会有每一个 lncRNA 以及 mRNA 的差异表达程度）；

2) 与老师感兴趣的 mRNA 有关联的 lncRNA 是关键 lncRNA。如果老师有感兴趣的 mRNA 的话，通过 excel 搜索命令，就可以找到与之关联的关键 lncRNA。

3) 案例：Genome-wide Mapping and Characterization of Notch-Regulated Long Noncoding RNAs in Acute Leukemia 的作者就是通过该方法成为寻找到关键 lncRNA。

表 8 lncRNA cis_regulation 预测结果

lncRNA_name	lncRNA_logFC	lncRNA_p_value	mRNA_name	mRNA_logFC	mRNA_p_value	distance
RP11-862L9.3	-1.337504919	0.018773892	ZNF236	-0.126316469	0.596851459	78023
RP11-862L9.3	-1.337504919	0.018773892	MBP	-0.026029997	0.960765809	81565
RP11-527H14.2	-1.519015199	0.023924117	ANKRD30B	-0.301324528	0.603648636	155762
RP1-290I10.7	1.135807811	0.04445406	TFAP2A	-0.048655661	0.930942806	27894
RP1-290I10.7	1.135807811	0.04445406	GCNT2	0.087585436	0.855855377	126479
...						...

注：

- 1) 第一列为lncRNA，第二列和第三列是这个lncRNA的差异表达log（fold-change）和显著性p_value；第四列为mRNA，第五列和第六列是这个mRNA 的差异表达log（fold-change）和显著性p-value。每一行是距离在100kb以内的一对lncRNA和mRNA，其中距离是在最后一列；
- 2) 文件结果汇总在lncRNA/cis-regulation 中；
- 3) 如果是windows系统，请用excel文件打开所有的文件。

6.2 lncRNA trans-regulation 预测

基于由 microRNA 介导的 sponge 原理，定位关键基因。

在 lncRNA 的众多功能中，lncRNA 会通过结合 microRNA 来抑制 microRNA 对 mRNA 的抑制或降解的作用（这里，lncRNA 被形象的比喻成海绵，被吸收的 microRNA 就是水滴）。我们在这里通过预测的方法预测潜在的 lncRNA-microRNA-mRNA 相互作用,从而定位关键 lncRNA。

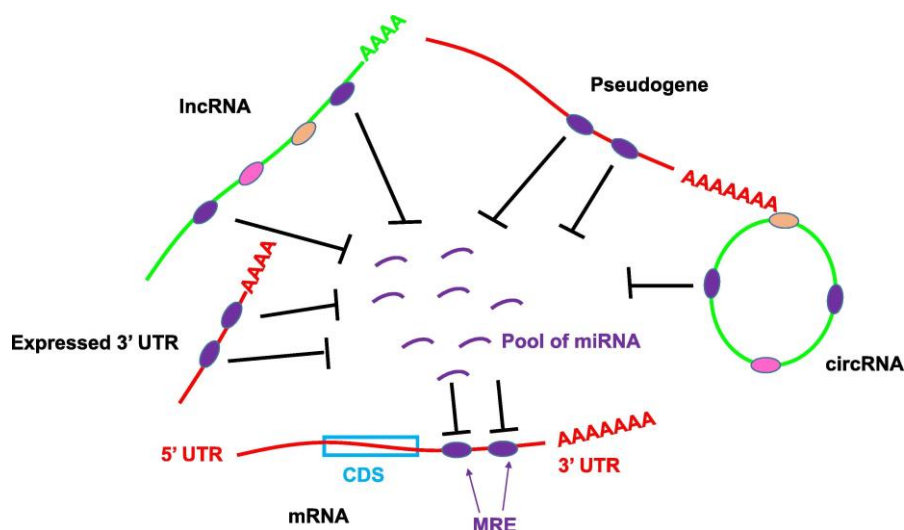


图 3.6.1 cis_regulation 示意图

我们首先对每一个 lncRNA 通过计算机软件，计算潜在的组成“lncRNA-microRNA-mRNA 匹配对”。进一步，考虑到基于这一海绵原理结合的 lncRNA 与

mRNA 会呈现正向表达的关系,所以我

们再检测 mRNA 和 lncRNA 的基因表达量是否协调一致,从而定位高度正向共表达的“lncRNA—microRNA—mRNA 匹配对”。在得到匹配对以后,我们可以通过 mRNA 的功能推测与其匹配的 lncRNA 功能。

老师可以通过自己倾向的几个原则之一,利用 excel 对我们的输出结果自行进行简单操作,从而定位关键基因:

1) 老师可以通过 mRNA 或者 lncRNA 的差异表达程度,进一步过滤筛选更为严格的“lncRNA—microRNA—mRNA 匹配对”。(我们的输出文件会有每一个 lncRNA 以及 mRNA 的差异表达程度);

2) 与老师感兴趣的 mRNA 有关联的 lncRNA 是关键 lncRNA。如果老师有感兴趣的 mRNA 的话,通过 excel 搜索命令,就可以找到与之关联的关键 lncRNA。

3) 案例:发表于 oncogene 的 A long noncoding RNA critically regulates Bcr-Abl-mediated cellular transformation by acting as a competitive endogenous RNA 的作者就是通过该分析,定位了通过海绵作用介导的 lncRNA—mRNA 搭配对,并进一步定位到了与 PTEN (tumor suppressor) 相关的关键 lncRNA。

表 9 lncRNA cis_regulation 预测结果

lncRNA_name	lncRNA_ logFC	lncRNA_ p_value	mRNA_ name	mRNA_ logFC	mRNA_ p_value	microRNA	TargetScan_ Score_mRNA	TargetScan_ Score_lncRNA
HOXB-AS1	0.69	0.03	TKT	-0.78	0.04	hsa-miR-485-5P,hsa...	-0.10,-0.33,...	-0.12,-0.17,...
RP11-443A13.5	-0.58	0.03	SLC27A1	-0.80	0.03	hsa-miR-124a,hsa...	-0.09,-0.08,...	-0.07,-0.07,...
HOXB-AS1	0.69	0.03	SLC27A1	-0.80	0.03	hsa-miR-136,hsa...	-0.07,-0.06,...	-0.06,-0.12,...
RP11-498C9.13	0.53	0.04	TKT	-0.78	0.04	hsa-miR-520a,hsa...	-0.03,-0.03,...	-0.10,-0.10,...
RP11-498C9.13	0.53	0.04	CDC42SE1	0.63	0.02	has-378,hsa...	-0.20,-0.30,...	-0.13,-0.09,...
...								...

注:

1) 第一列为lncRNA, 第二列和第三列是这个lncRNA的差异表达log (fold-change) 和显著性p—value; 第四列为mRNA, 第五列和第六列是这个mRNA的差异表达log (fold-change) 和显著性p—value。第七列是介导sponge作用的microRNA列表(因为是用软件targetscan预测的, 所以数量往往很多)。第八列和第九列分别是mRNA 和lncRNA 与每一个第七列中的microRNA 的结合强度;

3) 文件结果汇总在lncRNA/trans-regulation中的inquiry_trans_regulation.xls文件中;

3) 对应文件夹中的lncRNA_targetScan_results_hg19和mRNA_targetScan_results_hg19是涉及到的microRNA与lncRNA, mRNA结合强度预测的相关参数。具体说明请参见targetscan官网 <http://www.targetscan.org>;

4) 如果是windows系统, 请用excel文件打开所有的文件。

6.3 lncRNA_mRNA 联合表达网络分析

lncRNA 和 mRNA 的联合表达分析是生物调控机制中非常有意义的研究，我们根据 mRNAseq 及 lncRNAseq 获得的差异基因分别构建两者的关系网络。如图 3.6.1 所示，蓝色圆形表示 mRNA，绿色圈代表 lncRNA。节点的大小表示节点之间的相互关系的力量，两个节点之间的边代表了基因之间的相互作用。在基因上有更多的边缘，连接到它的基因越多，这个基因在网络中的作用就越重要。

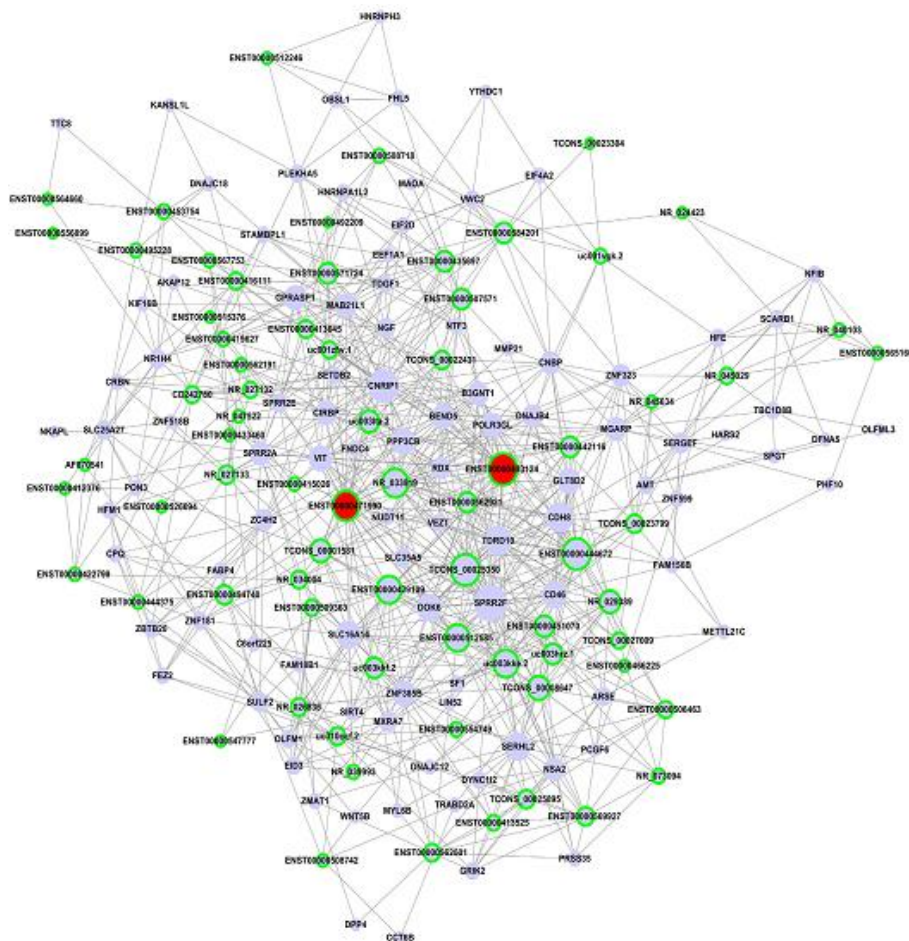


图 3.6.1

四. Methods 英文版

Sample collection and preparation:

RNA quantification and qualification

- 1) RNA degradation and contamination was monitored on 1% agarose gels.
- 2) RNA purity was checked using the NanoPhotometer[®] spectrophotometer (IMPLEN, CA,USA) .
- 3) RNA concentration was measured using Qubit[®] RNA Assay Kit in Qubit[®] 2.0 Fluorometer (Life Technologies, CA, USA).
- 4) RNA integrity was assessed using the RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, CA, USA).

Library preparation for Transcriptome sequencing

A total amount of 3 µg RNA per sample was used as input material for the RNA sample preparations. Sequencing libraries were generated using NEBNext[®] Ultra[™] RNA Library Prep Kit for Illumina[®] (NEB, USA) following manufacturer's recommendations and index codes were added to attribute sequences to each sample. Briefly, mRNA was purified from total RNA using poly-T oligo-attached magnetic beads. Fragmentation was carried out using divalent cations under elevated temperature in NEBNext First Strand Synthesis Reaction Buffer (5X) . First strand cDNA was synthesized using random hexamer primer and M-MuLV Reverse Transcriptase (RNase H⁻) . Second strand cDNA synthesis was subsequently performed using DNA Polymerase I and RNase H. Remaining overhangs were converted into blunt ends via exonuclease/polymerase activities. After adenylation of 3' ends of DNA fragments, NEBNext Adaptor with hairpin loop structure were ligated to prepare for hybridization. In order to select cDNA fragments of preferentially 150~200 bp in length, the library fragments were purified with AMPure XP system (Beckman Coulter, Beverly, USA). Then 3 µl USER Enzyme (NEB, USA) was used with size-selected, adaptor-ligated cDNA at 37°C for 15 min followed by 5 min at 95 °C before PCR. Then PCR was performed with Phusion High-Fidelity DNA polymerase, Universal PCR primers and Index (X)

Primer. At last, PCR products were purified (AMPure XP system) and library quality was assessed on the Agilent Bioanalyzer 2100 system.

Clustering and sequencing

The clustering of the index-coded samples was performed on a cBot Cluster Generation System using TruSeq PE Cluster Kit v3-cBot-HS (Illumina) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina HiSeq platform and 125 bp/150 bp paired-end reads were generated.

Sample collection and preparation:

Quality control

Raw data (raw reads) of fastq format were filtered by CUTADAPT^[1]. File were then processed by FASTQC^[2] for quality control. reads containing ploy-N and low quality reads from raw data. At the same time, Q20, Q30 and GC content the clean data were calculated. All the downstream analyses were based on the clean data with high quality. Quantification of gene expression level.

Reads mapping to the reference genome

Reference genome and gene model annotation files were downloaded from genome website directly. Index of the reference genome was built using STAR^[3] and paired-end clean reads were aligned to the reference genome using STAR .

Alternative splicing analysis

rMATS^[4] (v3.2.1 beta) was used for identification and differential expression analysis of alternative splicing events. We used hg19 genome as our reference genome ,and transcripts annotation from RefSeq for alternative splicing events annotation. Differential alternative splicing events were accepted if they could achieve an FDR lower than 5%. Functional interpretation of differential expression for alternative splicing events were performed by Database for Annotation, Visualization and Integrated Discovery (DAVID) program. Fisher's exact test $p < 0.05$ was

chosen as the threshold of significant change functions and KEGG pathways.

Quantification of gene expression level

HTSeq^[5] v0.6.1 was used to count the reads numbers mapped to each gene.

Differential expression analysis

Differential expression analysis of two conditions/groups (two biological replicates per condition) was performed using the DESeq^[6] R package (1.18.0). DESeq provide statistical routines for determining differential expression in digital gene expression data using a model based on the negative binomial distribution. The resulting P-values were adjusted using the Benjamini and Hochberg's approach for controlling the false discovery rate . Genes with an adjusted P-value <0.05 found by DESeq were assigned as differentially expressed.

五. 参考文献

- [1] Andrews S. *FastQC: A quality control tool for high throughput sequence data*[J]. Reference Source, 2010.
- [2] Martin M. *Cutadapt removes adapter sequences from high-throughput sequencing reads*[J]. *EMBnet. journal*, 2011, 17(1): pp. 10-12.
- [3] Dobin A, Davis C A, Schlesinger F, et al. *STAR: ultrafast universal RNA-seq aligner*[J]. *Bioinformatics*, 2013, 29(1): 15-21.
- [4] Shen S, Park J W, Lu Z, et al. *rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data*[J]. *Proceedings of the National Academy of Sciences*, 2014, 111(51): E5593-E5601.
- [5] Anders S, Pyl P T, Huber W. *HTSeq—a Python framework to work with high-throughput sequencing data*[J]. *Bioinformatics*, 2014: btu638.
- [6] Love M I, Huber W, Anders S. *Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2*. *bioRxiv*, 2014[J]. URL: <http://biorxiv.org/content/early/2014/02/19/002832>,arXiv:<http://biorxiv.org/content/early/2014/02/19/002832.full.pdf>, doi, 2014, 10: 002832.

关于公司

上海嘉因生物科技有限公司, 是致力于高通量实验设计, 高通量实验实施及生物信息学分析的医学科研服务企业。我们的核心成员来自于哈佛大学、加州大学洛杉矶分校、清华大学和同济大学等国内外知名学府。截至目前, 团队成员已在 Nature, Cancer Cell, Genome Research, Nature Structural & Molecular Biology, Leukemia 和 Nucleic Acids Research 等国际顶尖学术期刊发表 20 余篇学术论文, 累计影响因子超过 220。

- 1) *The modENCODE Consortium. modENCODE and ENCODE resources for analysis of metazoan chromatin organization. Nature, 2014. (IF:42.4)*
- 2) *The JARID1B/KDM5B histone demethylase is a luminal lineage-driving oncogene in breast cancer. Cancer Cell, 2014. (IF:23.9)*
- 3) *Canonical nucleosome organization at promoters forms during genome activation. Genome Research, 2014. (IF: 13.9)*
- 4) *Integrative genomic analyses reveal clinically relevant long non-coding RNA in human cancer. Nature Structural & Molecular Biology, 2013. (IF:11.6)*
- 5) *Integrating Gene and Mir Expression Profiles and Regulatory Network Structures to Define Aberrant Feed Forward Loops with Functional and Clinical Implications in Myeloma. Blood, 2012. (IF:9.8)*
- 6) *Transcription factor-pathway coexpression analysis reveals cooperation between SP1 and ESR1 on dysregulating cell cycle arrest in non-hyperdiploid multiple myeloma. Leukemia, 2013. (IF:9.7)*
- 7) *CR Cistrome: a ChIP-Seq database for chromatin regulators and histone modification linkages in human and mouse. Nucleic Acids Research, 2013. (IF: 8.8)*
- 8) *PlantGSEA: a gene set enrichment analysis toolkit for plant community. Nucleic Acids Research, 2013. (IF: 8.8)*
- 9) *agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Research, 2010. (IF: 8.8)*
- 10) *Target analysis by integration of transcriptome and ChIP-seq data with BETA.*

Nature Protocol, 2013. (IF: 7.8)

11) *Identifying ChIP-seq enrichment using MACS. Nature Protocol*, 2012. (IF: 7.8)

12) *Genome-wide analysis of histone modifications: H3K4me2, H3K4me3, H3K9ac, and H3K27ac in Oryza sativa L. Japonica. Molecular Plant*, 2013. (IF:6.6)

13) *BSeQC: Quality Control of Bisulfite Sequencing Experiments. Bioinformatics*, 2013. (IF: 5.3)

14) *CistromeFinder for ChIP-seq and DNase-seq data reuse. Bioinformatics*. 2013. (IF: 5.3)

15) *DiNuP: a systematic approach to identify regions of differential nucleosome positioning. Bioinformatics*, 2012. (IF: 5.3)

16) *GFOLD: a generalized fold change for ranking differentially ex-pressed genes from RNA-seq data. Bioinformatics*, 2012. (IF: 5.3)

17) *CistromeMap: A knowledgebase and web server for ChIP-Seq and DNase-Seq studies in mouse and human. Bioinformatics*, 2012. (IF: 5.3)

18) *PHF8 and REST/NRSF co-occupy gene promoters to regulate proximal gene expression. Scientific report*, 2014. (IF:5.1)

19) *Local chromatin dynamics of transcription factors imply cell lineage-specific functions during cellular differentiation. Epigenetics*, 2012. (IF: 5.1)

20) *GeSICA: genome segmentation from intra-chromosomal associations. BMC Genomics*, 2012. (IF: 4.0)

21) *ProFITS of maize: a database of protein families involved in the transduction of signaling in the maize genome. BMC genomics*, 2010. (IF: 4.0)

22) *plantsUPS: a database of plants' Ubiquitin Proteasome System. BMC genomics*, 2009. (IF: 4.0)

23) *Classify hyperdiploidy status of multiple myeloma patients using gene expression profiles. PLoS ONE*, 2013. (IF:3.5)