京房生物 GENEFUND

上海京房生物科技有限公司

RIP-seq 服务

结题报告

合同编号： XX-XXXXXX

2018-00-00

目录

一、背景介绍

RIP-seq RNA Immunoprecipitation 是研究细胞内蛋白与 RNA 相互作用的技术，是了解转录后调控网络动态过程的有力工具，能更有效地发现 miRNA 的调节靶点。这种技术运用针对目标蛋白的抗体把相应的 RNA-蛋白复合物沉淀下来，然后经过分离纯化就可以对结合在复合物上的 RNA 进行测序分析。

RIP 可以看成是普遍使用的染色质免疫沉淀 RIP 技术的类似应用，但由于研究对象是 RNA-蛋白复合物而不是 DNA-蛋白复合物，因此 RIP 实验的优化条件与 ChIP 实验并不相同。RIP 实验下游结合二代测序技术称为 RIP-seq，通过高通量测序和分析，深度解析与目标蛋白相互结合的 RNA 的区域或种类和相互作用强弱。

技术优势：

1. RIP-seq 是目前公认的最有效的确定蛋白质在细胞自然状态下与 RNA 结合的研究手段，可以有效的鉴定一个蛋白是否是 RNA 结合蛋白以及 RNA 结合蛋白与哪些 RNA 直接作用，并确定其结合位点。

2. RIP-seq 可从全转录组范围研究蛋白与 RNA 的相互作用，得知相互作用 RNA 的类型。

3. RIP-seq 分辨率高，可通过分析可得知与蛋白作用的 RNA 序列。

二、流程简介

2.1 实验流程

2.1.1 裂解物准备：

(1) 细胞收集：细胞生长至 90%覆盖度，细胞计数，满足最终获得 2~5mg 蛋白样品的数量，约 5~20 × 10^6 细胞总数；

(2) 交联固定：向细胞悬液中加入适量的甲醛溶液，甲醛终浓度达到 1%，室温放置 10min；向交联体系中加 10 倍体积的 2.66 M 甘氨酸，室温放置 5min，冰浴 10min；

(3) 消化裂解：加入胰酶（溶在 PBS 中，终浓度 0.2%），37℃消化 3min，加入 3ml 培养基（10%FBS）终止消化，加入 1ml RIPA 裂解液，吹打混匀后震荡，冰上放置 15-30min 后震荡，震荡后进行超声；

(4) 裂解物收集：低温离心后，留 100ul input 做 total RNA 抽提，留 100ul input 做 Western，剩下的再平均分到两个 RNase-free 的 2ml 管中。

2.1.2 IP 实验：

(1) 抗体孵育：分别加入 20ul 提前封闭好的 Protein A/G Agarose beads，室温旋转孵育 30min，离心（4℃，1000g，5min），上清转移到新的 2ml 管中；

(2) 孵育：两管上清中分别加入 ER antibody 和 IgG，再加入 yeast total RNA 和 BSA，室温旋转孵育 2h。两管上清中分别加入 100ul 提前封闭好的 Protein A/G Agarose beads，室温旋转孵育 2h；

(3) 清洗收集：RIPA 洗一次， NaCl RIPA 洗两次，RIPA 洗一次。RIPA 中临用前要加入 RiboLock、Proteinase Inhibitor、PMSF、DTT。100ul RIPA 重悬 beads 后，10%留样做 Western 检测，另外 90%加入 Proteinase K 进行消化，95℃失活 Proteianse K 后，加入 DNase I 消化 DNA。

2.1.3 RNA 建库测序：

(1) RNA 抽提建库：Trizol 法抽提 RNA，进行 DNA 片段末端修复、3’端加 A 碱基，连接测序接头。PCR 扩增及 DNA 产物的片段大小选择（一般为 300-400bp，包括接头序列在内）；

(2) RNA 抽提建库：Trizol 法抽提 RNA，进行 DNA 片段末端修复、3’端加 A 碱基，连接测序接头。PCR 扩增及 DNA 产物的片段大小选择（一般为 300-400bp，包括接头序列在内）；

(3) 上机测序：对建好的文库进行文库质检，质检合格后进行上机测序，测序平台为 illumina HiSeq/NextSeq。

2.2 信息分析流程

RIP-seq 测序后获得原始数据（raw reads），经过过滤去接头，去污染、比对参考基因组，使用 Unique mapped reads 进行后续的信息分析，包括：

(1) 使用 cutadapt[1]程序去掉原始下机数据中的接头序列；

(2) 使用 Trimmomatic[2]程序除去低质量的序列得到 clean data；

(3) 使用 Fastqc[3]程序统计 clean data 的数据量，q20 以及 q30 的比例；

(4) 使用 bowtie2[4]程序将 clean data 比对到参考基因组上；

(5) 计算测序实验捕获效率和 RsIP 区域的覆盖度以及平均测序深度；

(6) 使用 MACS2[5]在基因组上进行 peak-calling；

(7) 进行 peak 注释；

(8) 使用 MEME[6]进行结合峰的 motif 检测；

(9) 样品间 peak 的差异及注释。

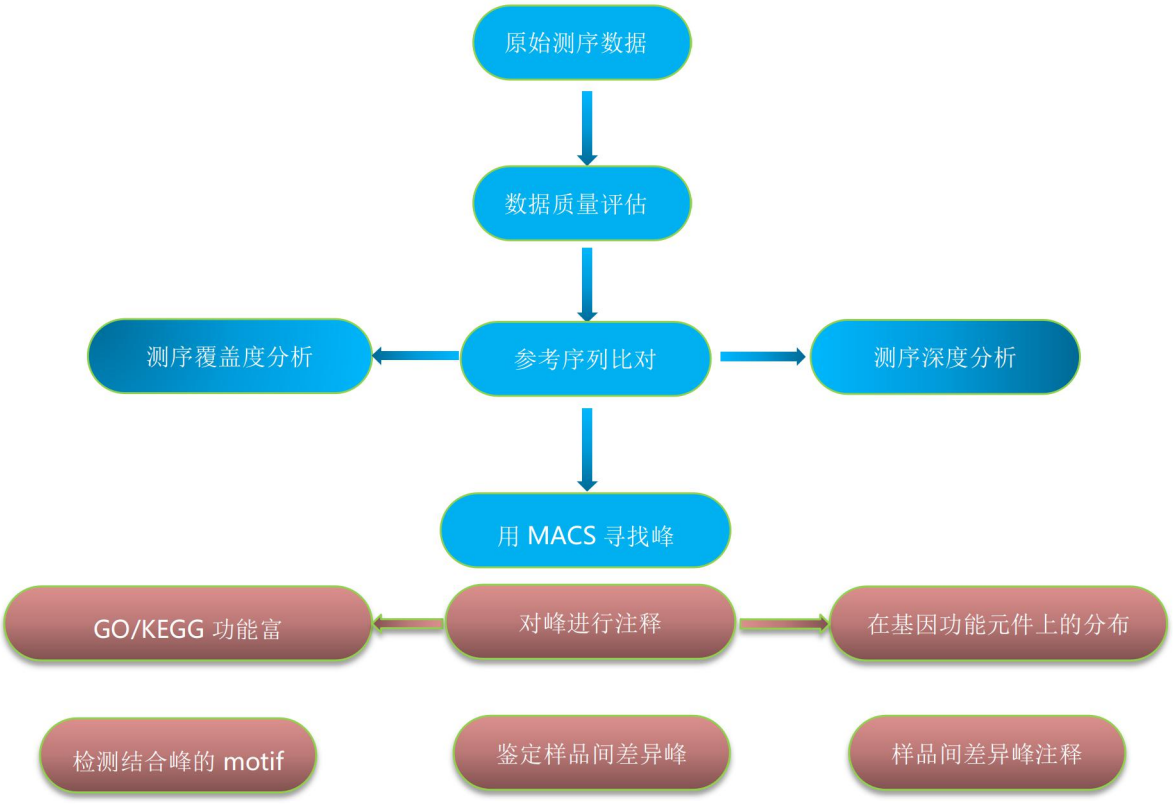


图 2-3 生物信息分析流程

三、RIP-seq 分析结果

3.1 分析结果概述

此次我们共收到了 4 套数据，分别为 A_RIP、A_Inpu1t、B_RIP、B_Input。收到数据后，按 RIP-seq 分析流程对数据进行分析。

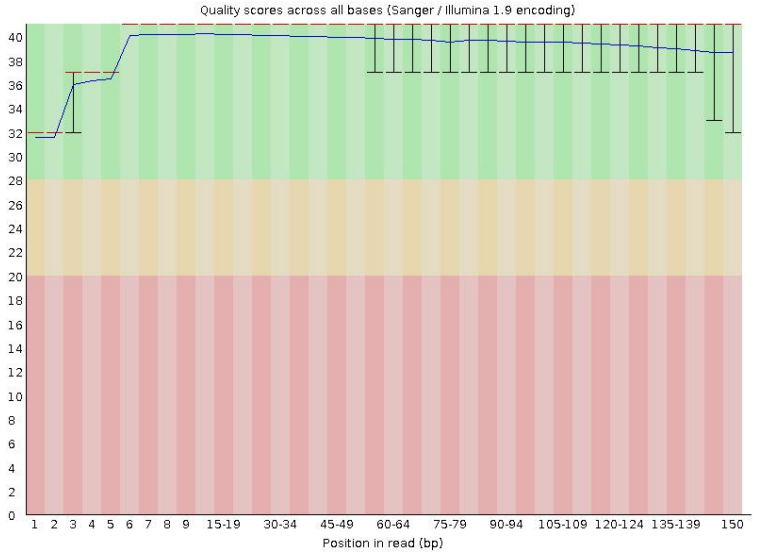
3.1.1 样本信息分组

Sample	RIP	Input
A	A_RIP	A_Input
B	B_RIP	B_Input

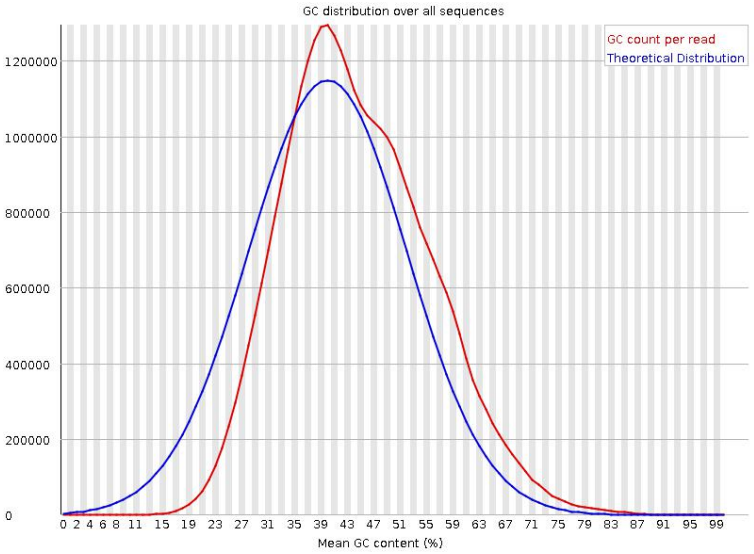
3.2 数据质控分析

通过 FastQC 工具对去除接头和低质量的 Clean Data 进行质控，部分结果如下所示：

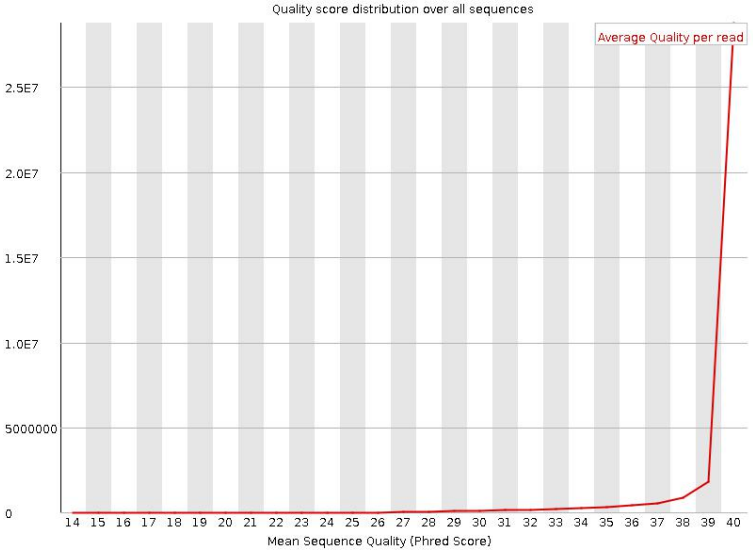
结果存放目录：Result\1.QualityControl\



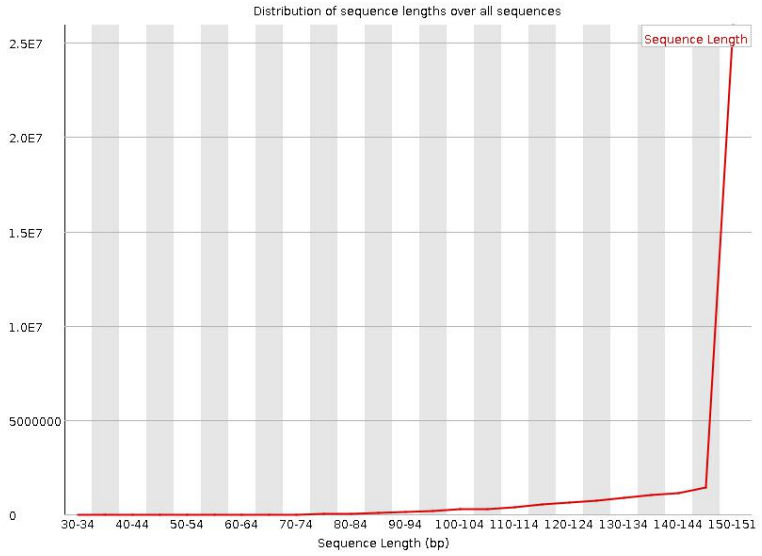
每一碱基质量图：横轴代表位置，纵轴 quality。红色表示中位数，黄色是 25%-75%区间，触须是 10%-90%区间，蓝线是平均数。若任一位置的下四分位数低于 10 或中位数低于 25，报"WARN"；若任一位置的下四分位数低于 5 或中位数低于 20，报"FAIL"。



每一序列的 GC 含量图：横轴为序列的平均 GC 含量，纵轴是 reads 数目，GC 含量表征了 PCR 扩增时的偏差。由图可知，本次样品的 GC 值与理论值相似，不存在明显的 PCR 扩增偏差。



每一序列的质量：碱基质量指征了碱基识别出错的概率。碱基质量值越高表明碱基识别越可靠，测序出错的可能性越小。横轴为碱基质量值，纵轴是 reads 数目，20 或 Q20 是指 100 个碱基中有 1 个会识别出错； Q30 是 1,000 个碱基中有 1 个会识别出错；同样 Q40 表示 10,000 个碱基中才有 1 个会识别出错。



序列长度分布图：横轴为序列的长度，纵轴是 reads 数目，由此图可知本样品测序所获得的大部分序列长度符合预期。

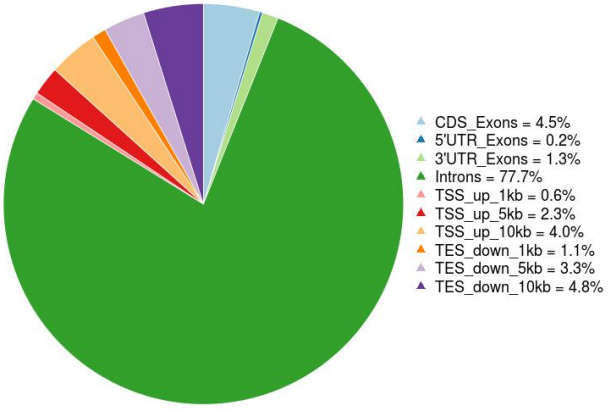
3.3 测序数据量统计

统计所有数据的数据量以及比对情况。

结果存放目录：Result\2.Mapping\

结果展示见下表：

Sample	A_RIP	A_Input	B_RIP	B_Input
Total Raw Reads	30064958	45263312	38577696	37699539
Total Raw Bases	2278190640	3.42E+09	2.91E+09	2.84E+09
Total Clean Reads	29597356	23185580	30436089	7554629
Total Clean Bases	879186950	5.86E+08	8.25E+08	1.97E+08
Alignment Rate	94.58%	95.48%	98.15%	96.46%



Reads 分类饼状图：从图中可以看出此次测序的每一类 Reads 占比。

3.4 Peak-calling 分析

得到比对结果之后，使用 MACS2 工具进行校峰，可以通过比对 RIP 实验相对于 Input 的富集位置（峰）。

Peak-calling 统计结果如下：

Peaks	104515
Q-value cutoff	0.05

结果存放目录：Result\3.CallPeaks\

部分结果展示如下：

chr	start	end	length	abs_summit	pileup	p-value	fold_enrichment	q-value	name
6	81766003	81768027	2025	81766946	256	1.48E-54	39.98377	1.6363E-318	...
1	151987982	151988869	888	151988546	238	5.43E-24	47.13347	6.7644E-318	...
X	41022221	41023317	1097	41022678	220	8.89E-16	56.35078	1.1266E-317	...
1	221393494	221394690	1197	221394228	212	1.43E-11	61.10661	1.814E-317	...
12	12538077	12539992	1916	12539564	231	1.87E-7	50.28303	2.3606E-317	...
13	73481995	73482687	693	73482236	217	3.3102E-5	57.85439	4.2249E-317	...

上表各列含义介绍：

列明	含义
Chr	峰所在染色体
Start	峰的起始位置
End	峰的终止位置
Length	峰区域的宽度
Abs_summit	峰顶的绝对位置
Pileup	峰顶位置的高度
P-value	该峰的 P-value
Fold_enrichment	该峰的富集倍数
Q-value	该峰的 Q-value
name	峰的名称

文件夹下另一个表格的内容与此表格内容相同，是使用其他工具进行后续分析的输入文件。

3.5 峰的注释及相关分析

使用 R 包 ChIPseeker[7]对上一步的校峰结果进行注释以及相关图片的绘制。

结果存放目录：Result\4.Peaks_Annotation\

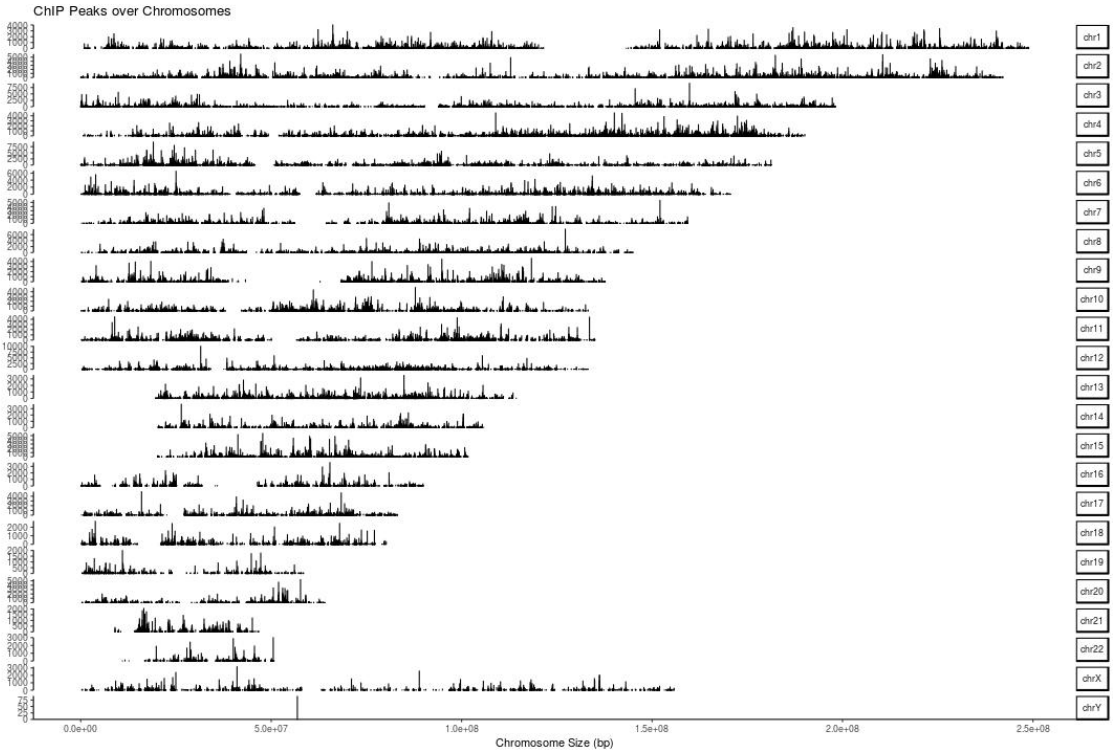
部分结果展示如下：

summit	annotation	geneChr	geneStart	geneEnd	Gene Length	Gene Strand	geneId	transcriptId	distance ToTSS	ENSEMBL	SYMBOL	GENE NAME
81	Promoter (1-2kb)	1	917370	918534	1165	2	100130417	uc057axr.1	1546	ENSG00000223764	LOC100130417	...
393	Promoter (<=1kb)	1	940346	942173	1828	1	148398	uc057axx.1	0	ENSG00000187634	SAMD11	...
410	Promoter (<=1kb)	1	942103	942802	700	1	148398	uc057axz.1	0	ENSG00000187634	SAMD11	...
127	Promoter (<=1kb)	1	1013423	1014540	1118	1	9636	uc001acj.5	0	ENSG00000187608	ISG15	...
96	Distal Intergenic	1	1070966	1074307	3342	2	401934	uc021oen.2	7765	ENSG00000237330	RNF223	...
91	Intron	1	1185036	1197475	12440	1	254173	uc057azy.1	4186	ENSG00000162571	TTLL10	...

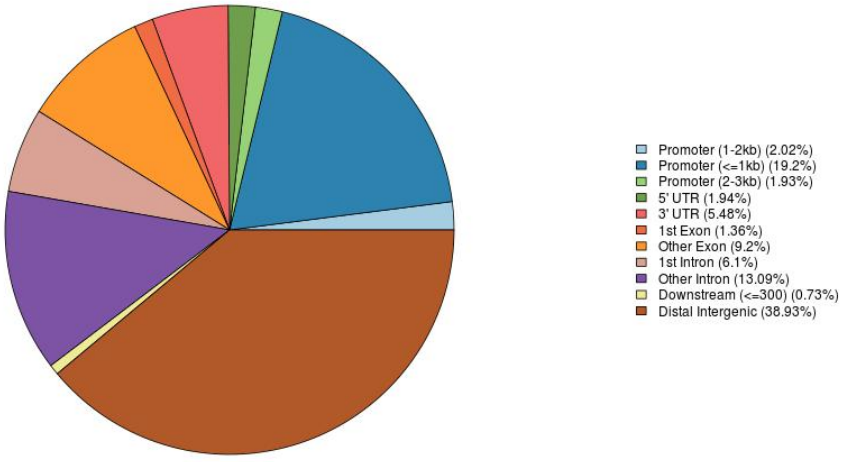
该表格为 3.4 的校峰结果注释，所以未展示的列与 3.4 表格相同，由于篇幅限制在此不做展示，该表格所有列的含义如下：

列名	含义	备注
seqnames	染色体	
start	peak 在基因在染色体上的起始位置	
end	peak 在基因在染色体上的终止位置	
width	peak 的宽度	
name	peak 的名字	
foldchange	富集倍数	
-log10(pvalue)	-log10(pvalue)	
pvalue	p-value	
-log10(qvalue)	-log10(qvalue)	
qvalue	q-value	
summit	相对于 peak start 的 peak 峰顶的位置	
annotation	ChIPseeker 工具对 peak 的注解	
geneChr	peak 对应基因所在染色体	
geneStart	peak 对应基因所在染色体的起始位置	
geneEnd	peak 对应基因所在染色体的终止位置	
geneLength	peak 对应基因长度	
geneStrand	peak 对应基因在染色体的链信息	
geneId	peak 对应基因的标识符	
transcriptId	peak 对应转录本的标识符	
distanceToTSS	peak 到转录起始位点的距离	DistanceToTSS 和 peak 所在的位置（intron，exome，等等），对应的不一定是一个转录本， 正负号只是说明上游或者下游。
ENSEMBL	Peak 对应基因的 ENSEMBL ID	
SYMBOL	peak 对应基因名字	
GENENAME	peak 对应基因的注解	
peaks_overlap_sample	该峰是否与另一个另一个样本中的峰重叠	T 代表是；F 代表否，即该

		峰为当前样本独有
gene_overlap_sample	该峰所属基因拥有的所有峰中，属于两个样本重叠峰的个数	0 代表无
gene_uniq_sample	该峰所属基因拥有的所有峰中，属于该样品独有峰的个数	0 代表无

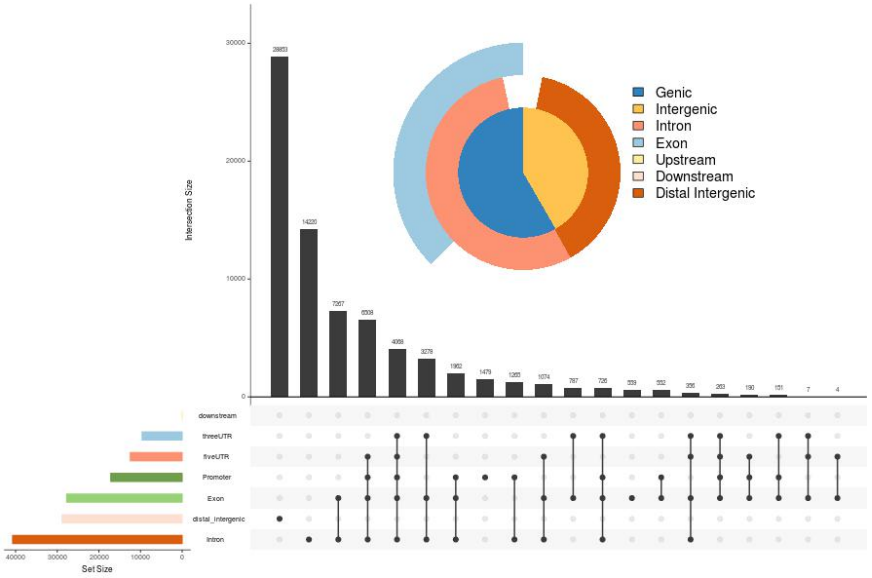


CHIP_Peaks_Coverage.png：此图绘制了所有 peaks 在染色体上的位置分布，横坐标为染色体的坐标，纵坐标为 peaks 的高度。



Peaks_Distribution.png：此图是 ChIPseeker 工具对 peaks 注解类型比例的扇形图。图中 Promoter 代表启动子区域，UTR 分别代表 3’UTR 和 5’UTR，1st Exon 和 1 st Intron 分别代表基因的

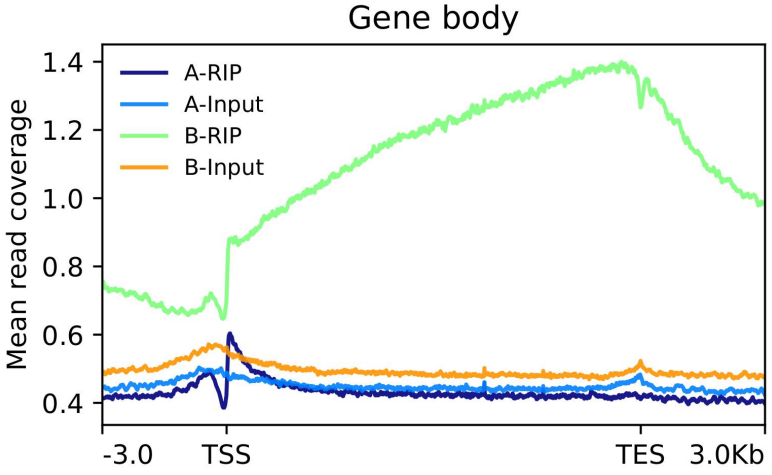
第一个外显子和内含子，Other Exon 和 Other Intron 分别代表基因除第一个以外的外显子和内含子，Downstream 代表基因终止位置下游一定范围，Distal Intergenic 代表除启动子及基因下游之外的基因间区。



Peaks_Distribution2.png：此图右上角为 peaks 基因组注释的维恩饼图，底部为 UpSet 图（图中黑色点表示该位置有数据，灰色的点表示没有，不同点之间的连线表示存在交集。上方条形图为每个交集的具体数据，左侧条形图代表了每种类型的 peaks 数量。图中 Upstream 与 Downstream 分别代表基因的上游与下游一定长度的区域。

3.6 mategene 分析

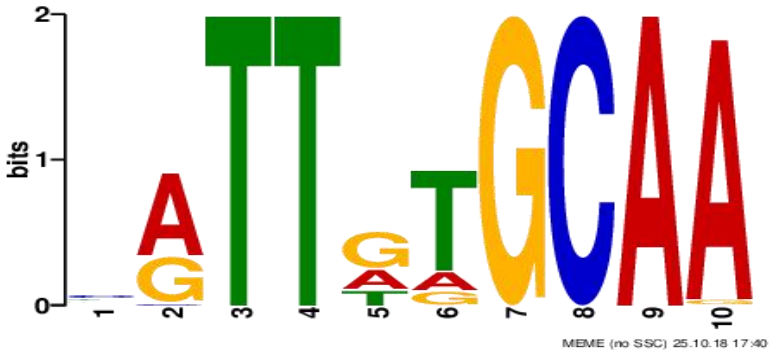
使用 deepTools[8]工具绘制 RIP-seq 在 Genebody 区域的平均信号轮廓，其中横坐标代表基因的起始与终止位点以及上下游 3Kb 范围，纵坐标代表区域的平均覆盖度。



3.7 Motif 分析

结果存放目录：Result\5.Motif\

根据校峰以及注释得到的峰信息，筛选出可靠的峰，取峰顶附近的序列使用 MEME 工具进行 Motif 分析。



Motif 分析结果展示：其中字母的大小和该核苷酸在 Motif 的频率是成比例的。

3.8 差异峰分析

根据两组 RIP 实验差异的校峰结果进行差异分析，结果中的四个表格分别为两组实验的差异峰与重叠峰，内容与 3.5 中的结果一致，在此不做展示。

3.9 差异基因功能富集分析

通过差异峰分析分别得到两个 RIP 实验的差异峰所在的基因作为为差异基因，然后进行功能富集分析，富集结果展示如下：

KEGG 富集结果表格

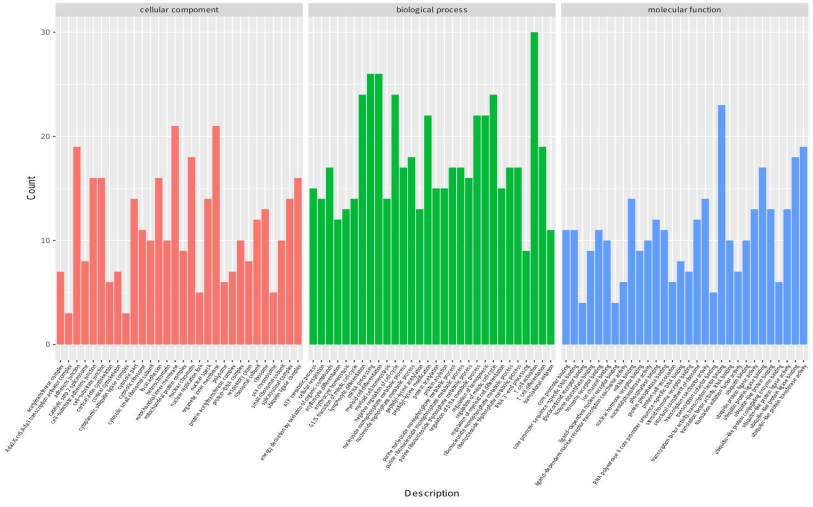
ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count	Hyperlink
hsa03013	...	17/336	171/7469	0.001710145	0.420139708	0.412768836	...	17	...
hsa03060	...	5/336	23/7469	0.003077661	0.420139708	0.412768836	...	5	...
hsa03030	...	6/336	36/7469	0.004916763	0.420139708	0.412768836	...	6	...
hsa00510	...	7/336	49/7469	0.005896698	0.420139708	0.412768836	...	7	...
hsa03410	...	5/336	33/7469	0.015036416	0.720441635	0.707802308	...	5	...

GO 富集结果表格

Ontology	ID	Description	Gene Ratio	BgRatio	pvalue	qvalue	geneID	Count
biological process	GO:0010975	regulation of neuron projection development	199/4139	450/17653	7.03E-23	3.14E-19	PMP22/PTEN/TSC1/	199
biological process	GO:0016358	dendrite development	111/4139	216/17653	3.23E-19	7.21E-16	HPRT1/PTEN/PAFAH1B1/.....	111
cellular compoment	GO:0098794	postsynapse	188/4321	450/18698	4.35E-19	2.26E-16	DMD/PTEN/SNCA/.....	188
biological process	GO:0051962	positive regulation of nervous system development	203/4139	493/17653	6.30E-19	9.37E-16	KIT/PAX2/PTEN/	203
biological process	GO:0050769	positive regulation of neurogenesis	180/4139	430/17653	8.87E-18	9.90E-15	KIT/PTEN/L1CAM/	180
biological process	GO:0031346	positive regulation of cell projection organization	154/4139	353/17653	2.55E-17	2.27E-14	ATP7A/KIT/L1CAM/	154

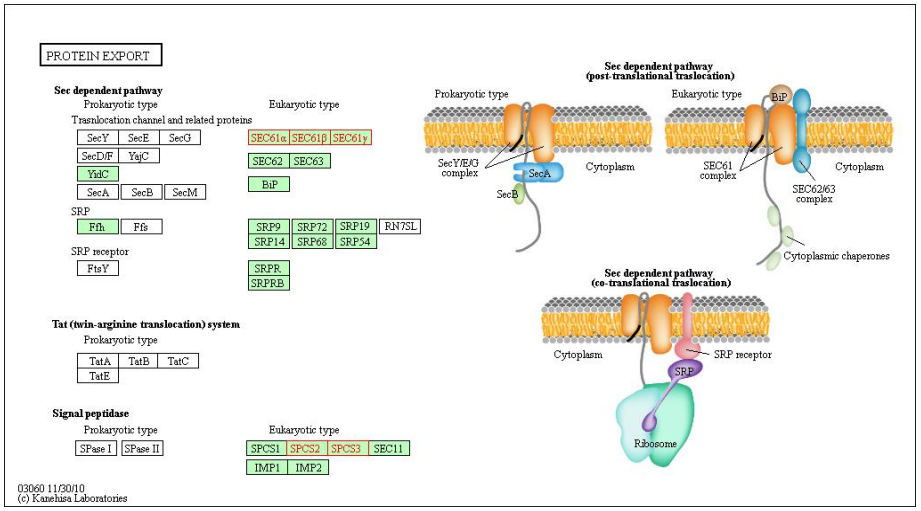
以上两个表格中各列含义介绍如下:

列名	含义
ID	KEGG/Gene Ontology 数据库中唯一的标号信息
Description	对该 KEGG/GO 的描述
GeneRatio	差异基因中与该 Term 相关的基因数与整个差异基因总数的比值
BgRatio	所有（ bg）基因中与该 Term 相关的基因数与所有（ bg）基因的比值
pvalue	富集分析统计学显著水平，一般情况下， P-value < 0.05 该功能为富集项
qvalue	对 p 值进行统计学检验的 q 值
geneID	该 KEGG/GO 涉及的基因
Count	差异基因中与该 Term 相关的基因数



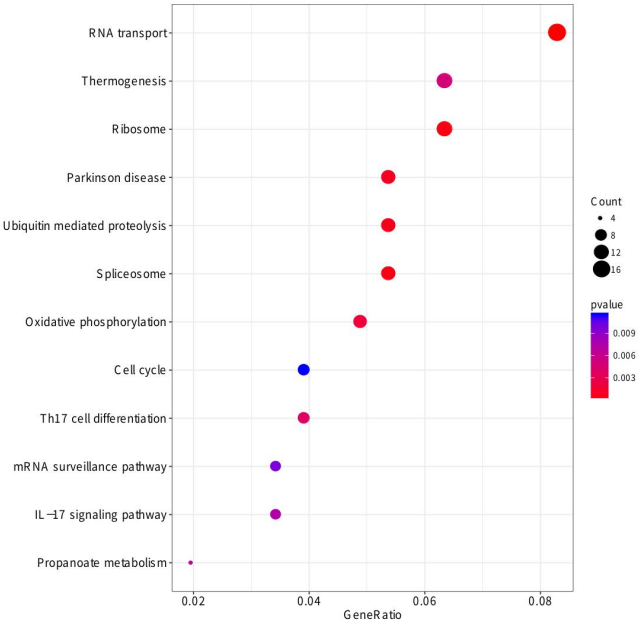
GO 富集的条形图

注:在分析过程中，我们挑选了富集最显著的 30 个 GO term 分析其表达上的差异性，如果不足 30 条，则全部分析。

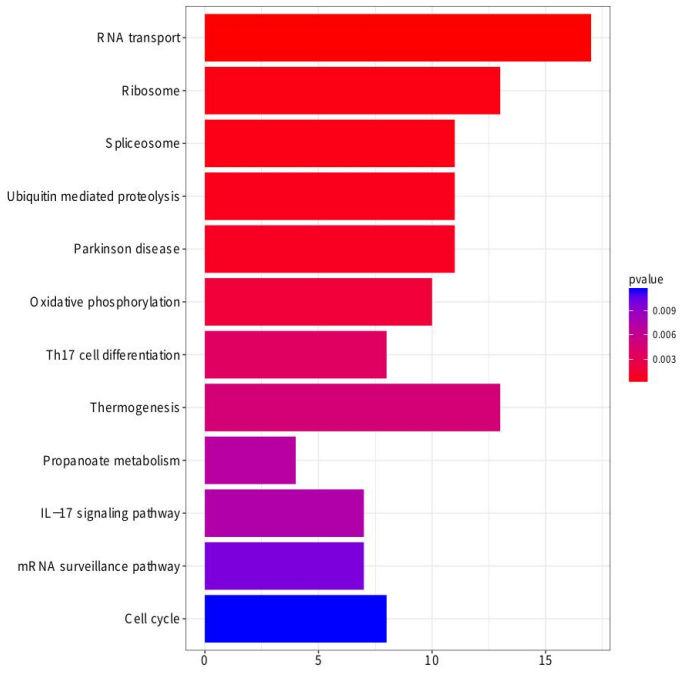


KEGG 通路数据库中 Thyroid hormone signaling pathway 详细信息

注：长方形节点表示基因产物（如酶或一些 RNA 调节子），所有绿色背景的基因产物都属于 KEGG ORTHOLOGY（KO）分类体系（序列高度相似，并在同一条通路上有相似功能的蛋白质被归为一组 KO），而白色背景的基因产物则不在 KO 分类体系之列，红色表示本次测序所研究基因能够注释到这些基因产物上（即认为具有与该节点基因产物相同或相似的功能）；圆形节点表示化合物（即底物或产物）；白色背景圆角长方形表示与本通路相关联的其他通路。箭头说明：酶反应方向或信息传递方向等；实线表示直接作用，虚线表示间接作用。详细说明请参见：http://www.genome.jp/kegg/document/help_pathway.html。



KEGG 富集散点图：图中纵坐标为 KEGG 通路，横坐标为富集在该通路的基因与该通路所有基因的比值，点的大小代表富集在该通路的基因数量，点越大代表富集在该通路的基因越多，点的颜色代表该通路的 P-value，从蓝到红 P-value 越来越小。



KEGG 富集条形图：图中纵坐标代表显著通路，横坐标代表该通路的基因数，条形图颜色代表该通路的 P-value，从蓝到红 P-value 越来越小。

四、分析使用软件及参数设置：

所用软件	参数设置	备注
Trimmomatic-0.35 工具	SLIDINGWINDOW:3:10 LEADING:10 TRAILING:10 MINLEN:50，其他设置选择默认参数	对原始的 reads 进行修剪和过滤，除去低质量的 Reads 和接头序列，接头序列 AGATCGGAAGAG
cutadapt 工具	--max-n 0 --minimum-length 50，其他设置选择默认参数	
FastQC 工具	默认参数	对数据进行质控
Bowtie2 工具	默认参数	将 reads 比对到参考基因组上
samtools 工具	默认参数	将 SAM 格式转变为 BAM 格式文件，用于下一步分析
macs2 callpeak 工具	-g hs -q 0.05，其他设置选择默认参数	找峰
ChIPseeker 工具	-	对峰进行注解，并绘制相关图形

五、Reference

[1] Martin M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.J ., 17, 10–12.. Bioinformatics, 2014,30(15) : 2114–2120.

[2] Anthony M. Bolger, Marc Lohse, Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data[J]

[3] Andrews S.. FastQC: a quality control tool for high throughput sequence data, 2010 Available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[4] Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012, 9:357-359.

[5] Zhang et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol (2008) vol. 9 (9) pp. R137

[6] Timothy L. Bailey, Mikael Bodén, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, William S. Noble, "MEME SUITE: tools for motif discovery and searching", Nucleic Acids Research, 37:W202-W208, 2009.

[7] Guangchuang Yu, Li-Gen Wang, Qing-Yu A. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics 2015, 31(14):2382-2383

[8] Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Research. 2016 Apr 13:gkw257.

[9] Guangchuang Yu, Li-Gen Wang, Yanyan Han and Qing-Yu He.clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology 2012,16(5):284-287