

RIC-seq结题报告

2023-09-06



目录

1 产品简介	4
1.1 RIC-seq测序原理	4
1.2 生物信息学分析流程图	4
2 质控	5
2.1 测序数据格式	5
2.2 测序数据过滤	5
3 比对	6
3.1 rRNA序列比对	6
3.2 基因组序列比对	7
4 全局互作分析	7
4.1 每个样本的互作鉴定分析	7
4.2 全局互作统计	8
5 RNA-RNA互作差异分析	9
5.1 差异分析	9
5.2 差异基因热图	10
5.3 火山图	11
6 差异基因富集分析	12
6.1 差异基因GO富集分析	12
6.2 差异基因KEGG分析	13

目录	3
7 HubRNA鉴定分析	14
7.1 联合RNA-seq进行hubRNA鉴定	14
7.2 hubRNA表达量分布	15
8 hubRNA差异分析	16
8.1 差异分析	16
8.2 差异基因热图	17
8.3 火山图	18
9 hubRNA富集分析	19
9.1 hubRNA的GO富集分析	19
9.2 hubRNA的KEGG分析	19
10 基序分析	20
11 eRNA-uaRNA互作分析	21
11.1 eRNA-uaRNA互作简介	21
11.2 eRNA-uaRNA互作靶基因GO富集分析	23
11.3 eRNA-uaRNA互作靶基因KEGG分析	23
12 特定基因互作分析	24
13 参考文章	26

1 产品简介

1.1 RIC-seq测序原理

近年来，在RNA互作研究中，高度结构化的RNA分子通常彼此相互作用，并与各种RNA结合蛋白结合，以调节关键的生物过程。近期开发的RIC-seq（RNA In situ conformation sequencing）[1]是一种能在细胞原位水平上捕获RNA高级结构及分子之间相互作用位点的新技术。利用RIC-seq技术还可系统分析重大疾病相关突变对RNA高级结构和作用靶标的影响，这将有望揭示非编码区突变的致病机理，并为临床诊断和治疗奠定基础。

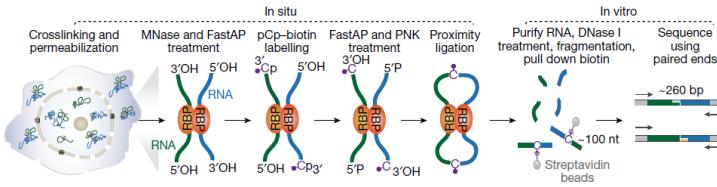


图 1: RIC-seq 测序原理

1.2 生物信息学分析流程图

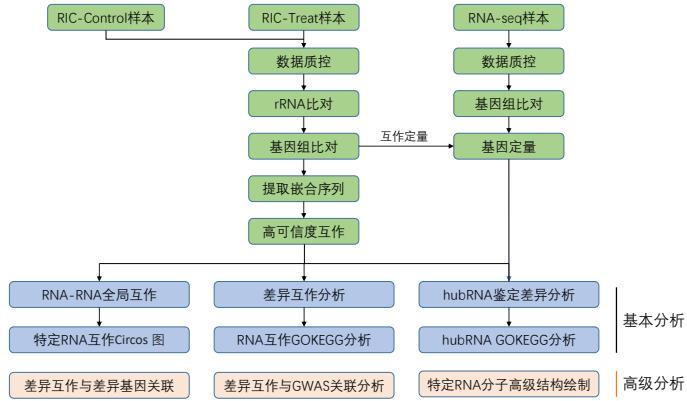


图 2: RIC-seq 分析流程

2 质控

2.1 测序数据格式

高通量测序得到的原始图像文件须经过碱基识别及误差过滤，得到序列数据，我们称之为Raw data或Raw reads，结果以压缩后的fastq格式储存，该文件包括reads的序列信息及其对应的质量信息。

FASTQ格式文件中每个read由四行描述，如下：

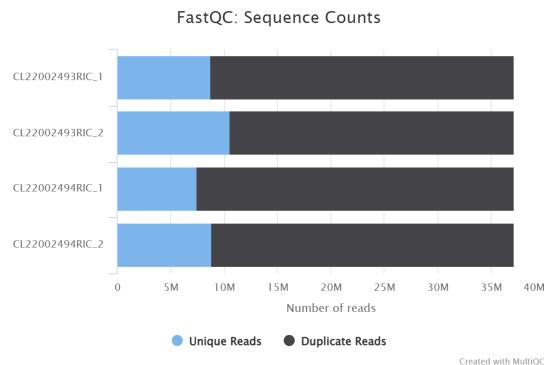
```
@A00679:14:HF7F3DSXX:3:1101:3314:1000 1:N:0:TAAGGCAG+GTAAGGA
TGTGATGAGCAAGATCGGCTCGCGAACGCCCTGGTTCCAAGTGATCATCCCGCTGTCGCTACTGGGCTGCACGTGGTGC
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFF:FFFFFF:FFFFFFFFFFFFFFFFFFFFFFFF,FFFFFF
@A00679:14:HF7F3DSXX:3:1101:9010:1000 1:N:0:TAAGGCAG+GTAAGGA
TCCTCATTGGAAAGGAACCTGTGCTGTGCTCTCTGCTGCTTCCATGTAGGTGGGCAACTGGACCAGGGGATGACTCTAAGTCTAGATTGAAACCTCTTGA
+
FFFFFF
-
```

每个序列共有4行，第1行以'@'开头，为reads的ID；第2行是序列；第三行是测序方向；第4行是每个碱基的质量值。

2.2 测序数据过滤

原始数据中会存在一部分接头序列和低质量的序列，为避免影响后续数据分析，首先对原始数据进行去接头和质量控制处理，获取用于后续数据分析的Clean data。利用FastQC分析测序数据质量，获得测序质量分布，碱基含量分布，重复测序片段比例等信息：

```
## 结果文件:1.qc_map/*_report.html
```



3 比对

3.1 rRNA序列比对

我们采用STAR program (v020201)对过滤后的测序序列进行rRNA序列比对的比对分析:

```
## 结果文件:1.qc_map/*_rRNA_map_stat.xls
```

表 1: 数据比对统计表

sample	Reads number	length	Uniquely mapped reads	Uniquely mapped rate	unmapped too short rate	unmapped other rate	chimeric reads
C2C12-1_read1_torRNA	169294380	128	14671407	8.67%	49.47%	41.49%	0.38%
C2C12-1_read2_torRNA	161845055	135	12867149	7.95%	47.73%	44.13%	0.40%

- (1) sample:样品名称。
- (2) Reads number:统计原始序列数据。
- (3) length:reads平均长度。
- (4) Uniquely mapped reads:比对上的reads。
- (5) Uniquely mapped rate:比对上的reads的比对率。
- (6) unmapped too short rate:由于序列太短, 未比对上序列的比率。
- (7) unmapped other rate:其他为比对上序列的比率。
- (8) chimeric reads:嵌合序列比对率。

3.2 基因组序列比对

我们采用STAR program (v020201)对过滤后的测序序列进行参考基因组的比对分析：

```
## 结果文件:1.qc_map/*_Genome_map_stat.xls
```

表 2: 数据比对统计表

Sample	Number of input reads	Average input read length	Uniquely mapped reads	of reads unmapped too short	of reads unmapped other	of chimeric reads
C2C12_1_read1	15399807	128	45.22%	43.18%	0.75%	4.21%
C2C12_1_read2	148665388	136	41.66%	47.98%	0.73%	2.61%

- (1) sample:样品名称。
- (2) Reads number:统计原始序列数据。
- (3) length:reads平均长度。
- (4) Uniquely mapped reads:比对上的reads。
- (5) Uniquely mapped rate:比对上的reads的比对率。
- (6) unmapped too short rate:由于序列太短, 未比对上序列的比率。
- (7) unmapped other rate:其他为比对上序列的比率。
- (8) chimeric reads:嵌合序列比对率。

4 全局互作分析

4.1 每个样本的互作鉴定分析

由比对完成的reads进行RNA-RNA互作互作分析鉴定, 可获得如下全局互作图:

```
## 结果文件:2.global_stat/*_num_of_interactions_from_part.list
```

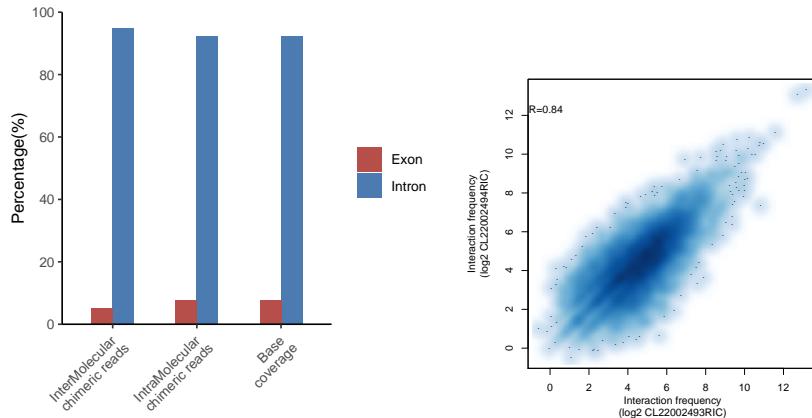


图 3: 样本互作统计图

图中3的左图统计了分之间，分子内互作和嵌合序列在外显子区域和内含子区域的比例统计，右图是两个重复样本的相关性分析。

4.2 全局互作统计

对于每个样本鉴定的互作reads，进行统计，得到全局互作reads比对统计柱状图、配对reads统计柱状图和在重复样本中都鉴定出来的高可信度互作reads分布图

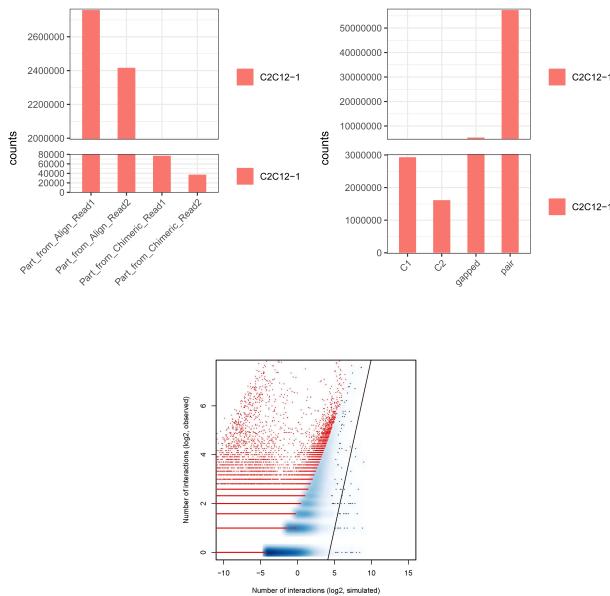


图 4: 全局互作统计图

图中4依次统计了互作reads在R1和R2的分布、配对reads分布和在重复样本中都鉴定出来的高可信度互作reads分布

5 RNA-RNA互作差异分析

5.1 差异分析

采用 DESeq2[3] R 包分析组间差异，默认筛选条件为：2 倍差异， $FDR < 0.05$ 的差异基因。

```
## 结果文件:4.DE_RNA_RNA_interaction/All_sample_interaction_TVSC.All.txt
```

表 3: 数据比对统计表

AccID	log2FoldChange	pvalue	padj
EIF3K_TMEM182	1.3268008	0.2825128	0.9991606
INSIG2_STAG3L1	0.2239053	0.8181824	0.9991606
C10orf35_ELF2	0.3475381	0.7849327	0.9991606
TUFM_TUFMP1	-0.2307588	0.8127408	0.9991606

- (1) geneid: 基因 ID
- (2) baseMean: 所有样本标准化后的平均值
- (3) Log2FoldChange: 差异倍数的 Log2 对数转换值 lfcSE: 差异倍数对数值的标准误
- (4) Stat: 统计量 pvalue: p 值
- (5) padj: 校正后 p 值, 即 FDR gene_name: 基因名称 gene_type: 基因类型

5.2 差异基因热图

为了全面的直观的展示样品之间的关系及差异情况, 将差异表达基因做聚类分析。用挑选的差异基因归一化的表达量表格作为输入文件。一般来说, 同一类样品能通过聚类出现在同一个簇 (cluster) 中, 聚在同一个簇的基因可能具有类似的生物学功能。热图的作用为:

- (1) 直观呈现多样本多个基因的全局表达量变化;
- (2) 呈现多样本或多基因表达量的聚类关系。本分析提供了 3 种配色方案的热图供选择。

```
## 结果文件:2.global_stat/*TVSC.All.txt
```

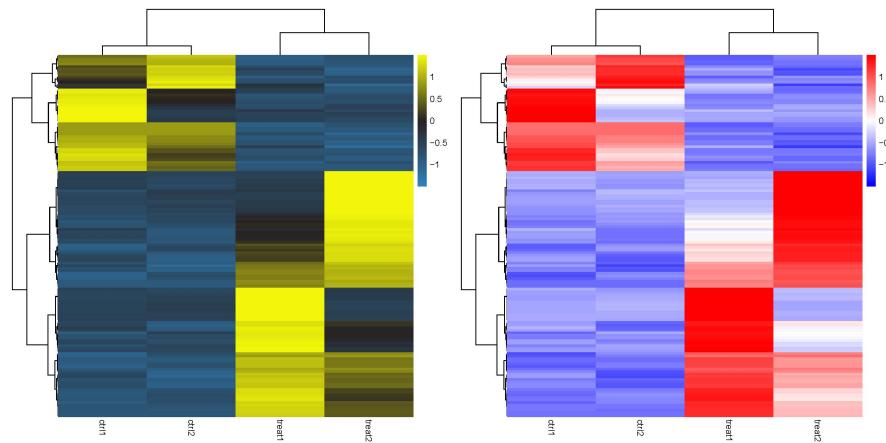


图 5: 差异基因热图

5.3 火山图

火山图（Volcano Plot）用于显示两组样品数据的差异基因分布情况，火山图在一张图中显示了两个重要的指标（Fold change 和FDR），可以非常直观且合理地筛选出在两样本间发生差异表达的基因。显著性差异基因数量越多，图的形状越接近“火山”，反之则稀疏。

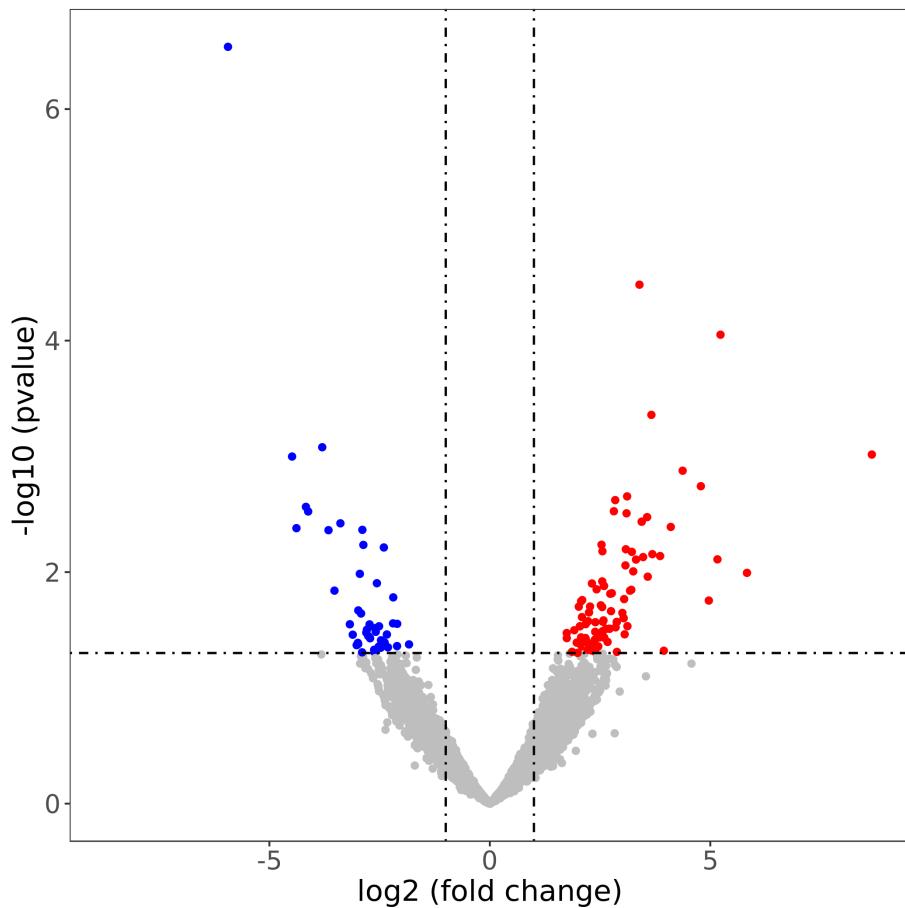


图 6: 火山图

6 差异基因富集分析

6.1 差异基因GO富集分析

基因功能分析[4,5] (GeneOntology) 将显著性差异基因分配到不同的功能分类中，从各方面描述基因的功能，可以分为三个主要的类群，生物学进程 (Biological Process, BP)，分子功能 (Molecular Function, MF) 和细胞组分 (Cellular Component, CC)。

```
## 结果文件:4.DE_RNA_RNA_interaction
```

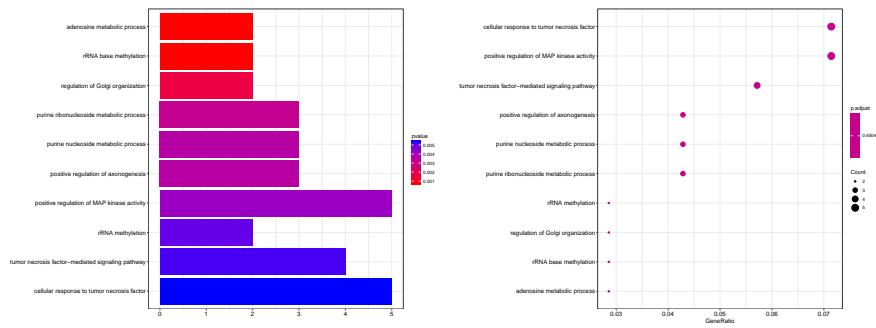


图 7: GO富集分析

6.2 差异基因KEGG分析

信号通路分析[6]的目的是基于 KEGG 数据库去寻找显著性差异基因显著性富集的信号通路。对差异 mRNA 进行 KEGG 通路分析，校正后的 p 值小于 0.05 的 term 展示

```
## 结果文件:4.DE_RNA_RNA_interaction
```

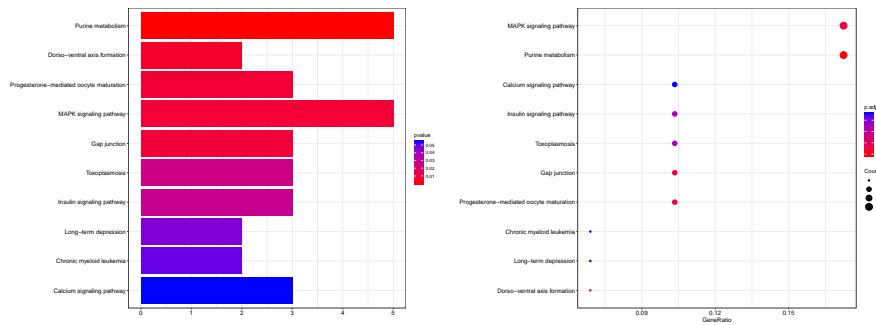


图 8: KEGG富集分析

7 HubRNA鉴定分析

hubRNA是基因座转录的 RNA 类似于与全局不同 RNA 片段相互作用的中枢，2020年，薛愿超研究团队利用RIC-seq技术绘制了HeLa细胞的转录组RNA 3D交互图，验证了2307个RNA拓扑结构域。通过对目标基因数量和相互作用强度进行划分，确定了642个RNA相互作用中心，命名为hub RNA。RNA 70%来源于含有大内含子(>50 kb)的编码基因，只有5%来源于lncRNA和伪基因，而hub lncRNA表现出比 hub pre-mRNA更强的反式作用。Hub RNA与非hub RNA相比具有更强的反式作用并且其结构更为保守。(1)它们通常与来自同一染色体和亚细胞结构的目标RNA相互作用；(2)根据187种已知的RBP结合模体，hub RNA及其作用靶标可分为三类，来自相同类的hub RNA通常相互作用；(3)在相同的GO term 中的hub RNA不表现优先的相互作用；(4)13%的hub RNA与超级增强子(SE)重叠，这些hub RNA作为SE富集到CCUUCCCC模体，并被RBP所占据。

7.1 联合RNA-seq进行hubRNA鉴定

hubRNA在RIC-seq和RNA-seq中都表现出具有高表达量，从而在细胞内行使广泛的互作。下图展示在重复样本中鉴定的hubRNA和hubRNA种类分布图

```
## 结果文件:5.hubRNA
```

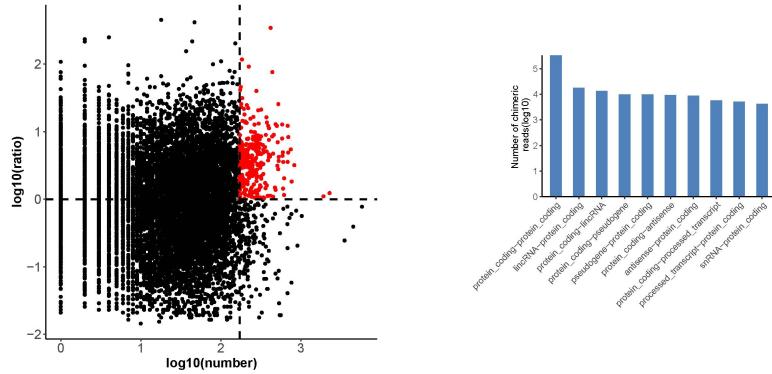


图 9: hubRNA分布

图9显示hubRNA是在RNA-seq和RIC-seq中具有高表达的RNA分子，同时可以看出各种互作类型的分布。

7.2 hubRNA表达量分布

hubRNA的在编码区和非编码区的分布不同，hubRNA和nohubRNA之间的也有显著的差异，下图展示了hubRNA在编码区和非编码区以及hubRNA-nohubRNA之间的差别。

```
## 结果文件:5.hubRNA
```

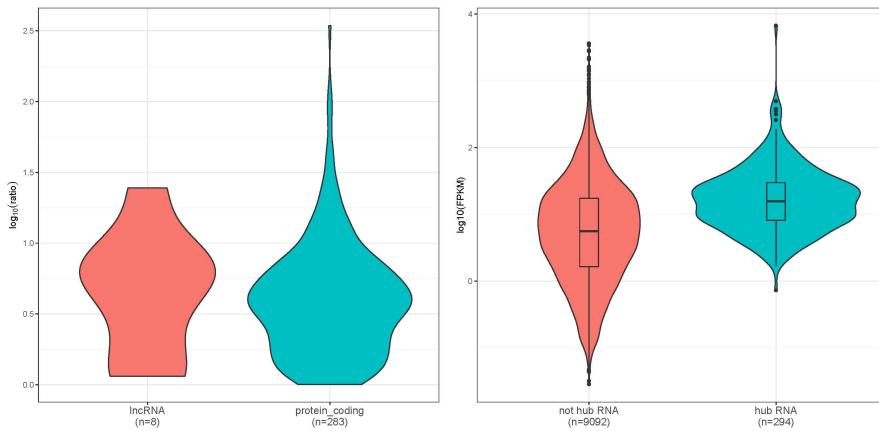


图 10: hubRNA 表达分布

图10显示非编码区的表达量分布以及hubRNA和no-hubRNA之间的表达量上的不同.

8 hubRNA差异分析

8.1 差异分析

采用 DESeq2[3] R 包分析组间差异， 默认筛选条件为： 2 倍差异， FDR<0.05 的差异基因。

```
## 结果文件:6.DE_hubRNA/All_sample_interaction_TVSC.All.txt
```

表 4: 数据比对统计表

AccID	log2FoldChange	pvalue	padj
EIF3K_TMEM182	1.3268008	0.2825128	0.9991606
INSIG2_STAG3L1	0.2239053	0.8181824	0.9991606
C10orf35_ELF2	0.3475381	0.7849327	0.9991606
TUFM_TUFMP1	-0.2307588	0.8127408	0.9991606

- (1) geneid: 基因 ID
- (2) baseMean: 所有样本标准化后的平均值
- (3) Log2FoldChange: 差异倍数的 Log2 对数转换值 lfcSE: 差异倍数对数值的标准误
- (4) Stat: 统计量 pvalue: p 值
- (5) padj: 校正后 p 值, 即 FDR gene_name: 基因名称 gene_type: 基因类型

8.2 差异基因热图

为了全面的直观的展示样品之间的关系及差异情况, 将差异表达基因做聚类分析。用挑选的差异基因归一化的表达量表格作为输入文件。一般来说, 同一类样品能通过聚类出现在同一个簇 (cluster) 中, 聚在同一个簇的基因可能具有类似的生物学功能。热图的作用为:

- (1) 直观呈现多样本多个基因的全局表达量变化;
- (2) 呈现多样本或多基因表达量的聚类关系。本分析提供了 3 种配色方案的热图供选择。

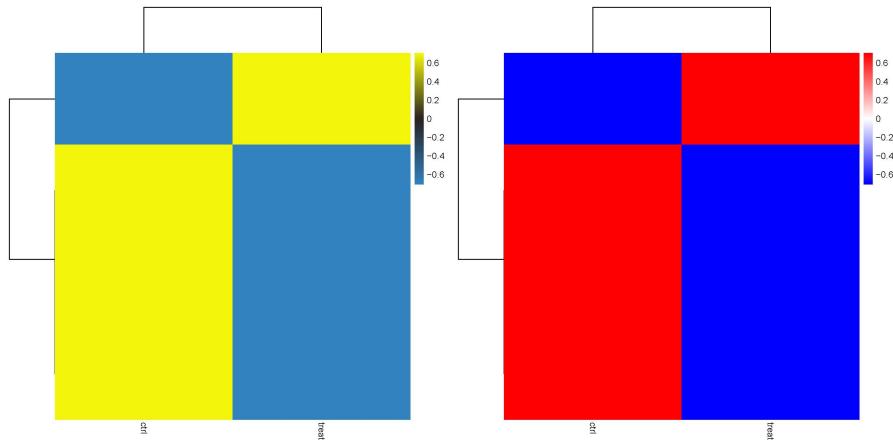


图 11: 差异基因热图

8.3 火山图

火山图（Volcano Plot）用于显示两组样品数据的差异基因分布情况，火山图在一张图中显示了两个重要的指标（Fold change 和FDR），可以非常直观且合理地筛选出在两样本间发生差异表达的基因。显著性差异基因数量越多，图的形状越接近“火山”，反之则稀疏。

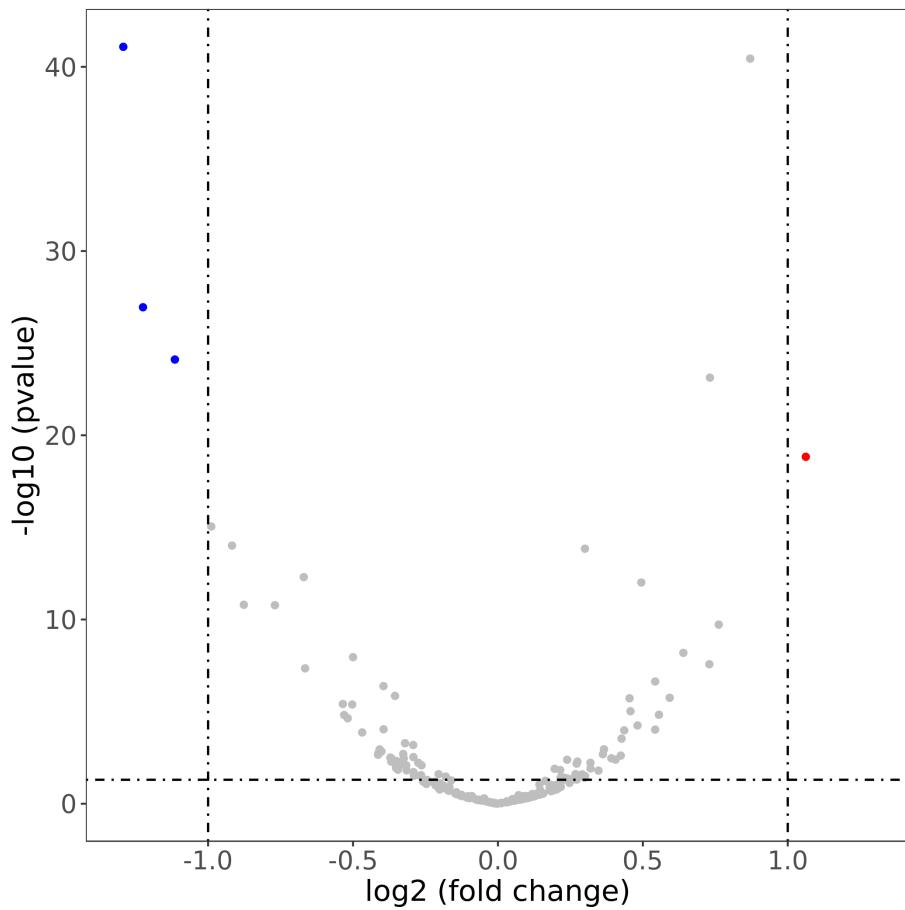


图 12: 火山图

9 hubRNA富集分析

9.1 hubRNA的GO富集分析

基因功能分析[4,5]（GeneOntology）将显著性差异基因分配到不同的功能分类中，从各方面描述基因的功能，可以分为三个主要的类群，生物学进程（Biological Process, BP），分子功能（Molecular Function, MF）和细胞组分（Cellular Component, CC）。

```
## 结果文件:5.hubRNA
```

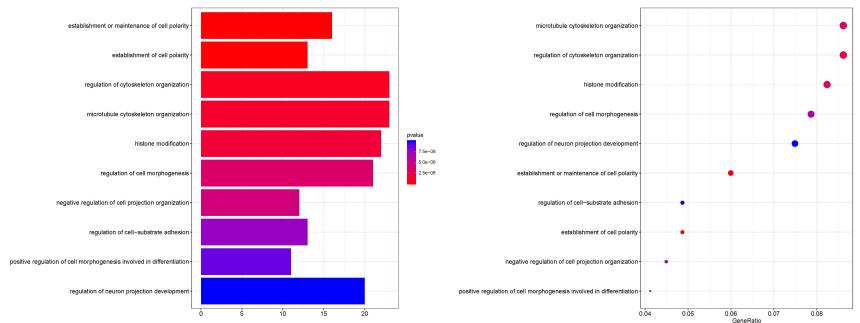


图 13: GO富集分析

9.2 hubRNA的KEGG分析

信号通路分析[6]的目的是基于 KEGG 数据库去寻找显著性差异基因显著性富集的信号通路。对差异 mRNA 进行 KEGG 通路分析，校正后的 p 值小于 0.05 的 term 展示

```
## 结果文件:5.hubRNA
```

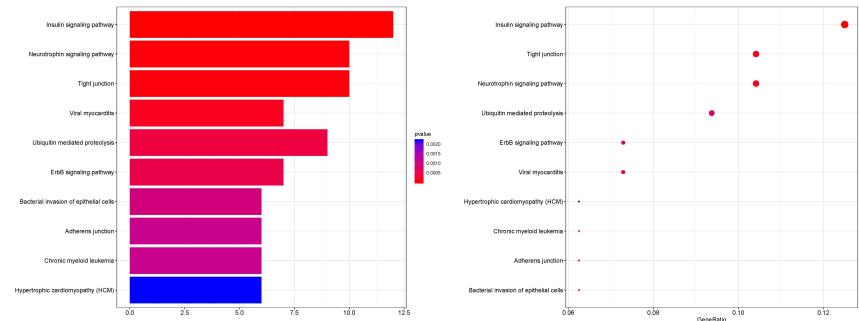


图 14: KEGG富集分析

10 基序分析

我们利用HOMER软件对互作序列进行motif分析，可以通过motif找到互作序列共同的序列，及转录因子调节区域：

结果文件:8.motif

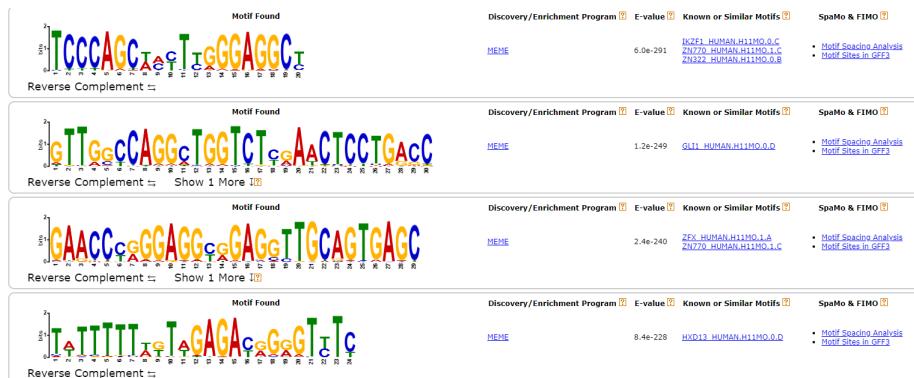


图 15: motif图

图15所示，互作reads的共同序列，可发现新的motif。

11 eRNA-uaRNA互作分析

11.1 eRNA-uaRNA互作简介

增强子（enhancer），是DNA上一小段可与蛋白质（反式作用因子；trans-acting factor）结合的区域，与蛋白质结合之后，基因的转录作用将会加强。增强子可能位于基因上游，也可能位于下游。且不一定接近所要作用的基因，甚至不一定与基因位于同一染色体。这是因为染色质的缠绕结构，使序列上相隔很远的位置也有机会相互接触。增强子具有以下特征)：

(1)、增强子DNA 序列处于染色体疏松的区域，与核小体中组蛋白的修饰，转录因子的结合有关； (2)、增强子活性与其DNA 序列结合的组蛋白H3 的第4 位赖氨酸单甲基化(H3K4me1)和第27 位赖氨酸乙酰化(H3K27ac)修饰程度成正相关； (3)、增强子发挥功能需要增强子区域和启动子的区域的直接相互作用，形成三维环状结构(3D-loop)。增强子和启动子的相互作用由多种蛋白介导，如Mediator 复合体、Cohesin 等。

在真核生物细胞里，DNA的染色质复合体结构像原核生物的超螺旋一样折叠。所以虽然增强子与基因相距很多核苷酸，但在几何上两者距离很近。使增强子与总转录因子及RNA多聚酶II的相互作用成为可能。增强子可以在被它调控基因的上游或下游。而且增强子不一定靠近转录起始位点才能调控基因转录，已发现有些距离达几十万碱基对。增强子通过与激活蛋白的结合对启动子（不直接对启动子本身）起作用。这些激活蛋白与中介复合物（辅激活物）相互作用，后者通过使用多聚酶II和总转录因子开始基因转录。曾经在内含子内发现增强子。有时增强子的方向颠倒后并不影响它的功能。而且增强子可被切除并插入到染色体的其他位置，仍然影响基因转录。这就是虽然内含子并不转录但其多态性会起作用。

超级增强子（Super enhancer, SE）是一类具有超强转录激活特性的顺式调控元件，2013年由美国白头生物医学研究所（Whitehead Institute for Biomedical Research）学者Richard A. Young首次提出。与普通增强子（Typical enhancer, TE）相比，超级增强子区域跨度范围通常可达 8-20 Kb，远高于普通增强子的200-300 bp跨度范围。更重要的是，超级增强子比普通增强子具有更高密度的转录激活相关组蛋白修饰（H3K27ac、H3Kme1等）、Mediator复合体和Bromodomain

containing 4 蛋白(BRD4, 与组蛋白乙酰化修饰位点结合)的结合; 辅因子(Mediator等)及转录因子富集密度。以上特点决定了超级增强子具有强大的调控功能。

(1)、超级增强子具有高密度的H3K27ac 和H3K4me1 修饰, 以及Mediator复合体和Bromodomain containing 4 蛋白(BRD4, 与组蛋白乙酰化修饰位点结合)的结合; (2)、超级增强子结合的转录因子以及与转录活性相关的染色体的标记比普通增强子高很多; (3)、超级增强子调控的基因比普通增强子调控的基因表达水平高很多; (4)、组成超级增强子的单个增强子也可以像普通增强子一样激活基因转录; (5)、超级增强子可以结合组织中特异的转录因子; (6)、与普通增强子相比, 超级增强子活性对于转录因子的阻断更敏感。在细胞内, 启动子和增强子区都可转录产生RNA, 且增强子和启动子在空间上邻近配对后才能激活转录。eRNA(Enhancer和Super enhancer区域转录RNA)与uaRNA(promoter区域转录RNA)的互作是细胞内非常重要的调控机制, RIC-seq (RNA in situ conformation sequencing)技术可一次性捕获细胞内所有直接的RNA-RNA配对或者由蛋白质介导的间接RNA-RNA近距离相互作用, 捕获eRNA-uaRNA互作用于探索Enhancer和Super enhancer与promoter的互作机制:

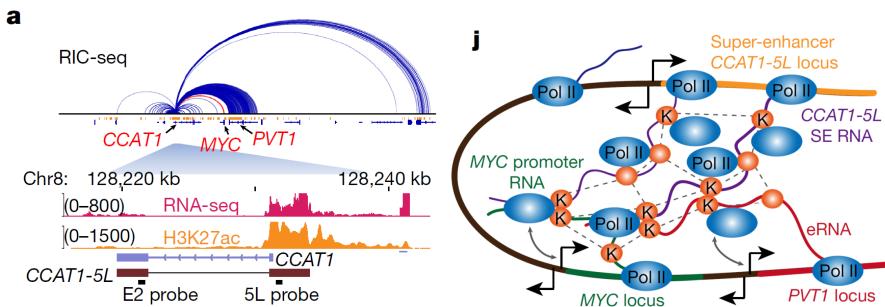


图 16: eRNA互作机制

图16所示, 通过RIC-seq鉴定得CCAT1是一个hub RNA, 同时CCAT1与超级增强子638重叠, 并且已被证明在多种癌症中上调; RIC-seq数据还显示它的异构体CCAT1-5L与MYC启动子RNA及PVT1 eRNA相互作用。于是研究者选取了super enhancer CCALT1-5L作为下游继续研究的基因。

11.2 eRNA-uaRNA互作靶基因GO富集分析

基因功能分析[4,5]（GeneOntology）将显著性差异基因分配到不同的功能分类中，从各方面描述基因的功能，可以分为三个主要的类群，生物学进程（Biological Process, BP），分子功能（Molecular Function, MF）和细胞组分（Cellular Component, CC）。

```
## 结果文件:9.eRNA/
```

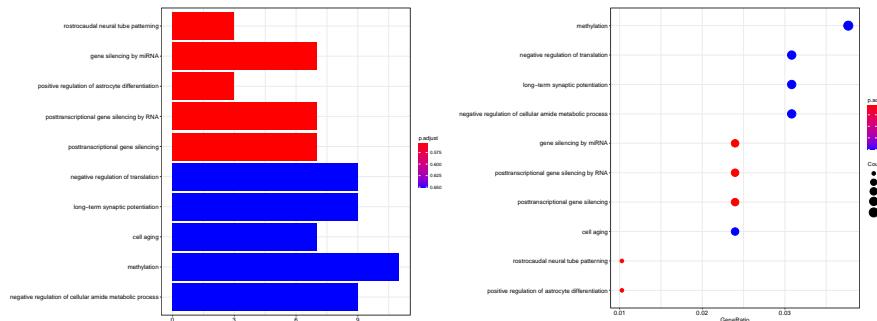


图 17: GO富集分析

11.3 eRNA-uaRNA互作靶基因KEGG分析

信号通路分析[6]的目的是基于 KEGG 数据库去寻找显著性差异基因显著性富集的信号通路。对差异 mRNA 进行 KEGG 通路分析，校正后的 p 值小于 0.05 的 term 展示

```
## 结果文件:9.eRNA/
```

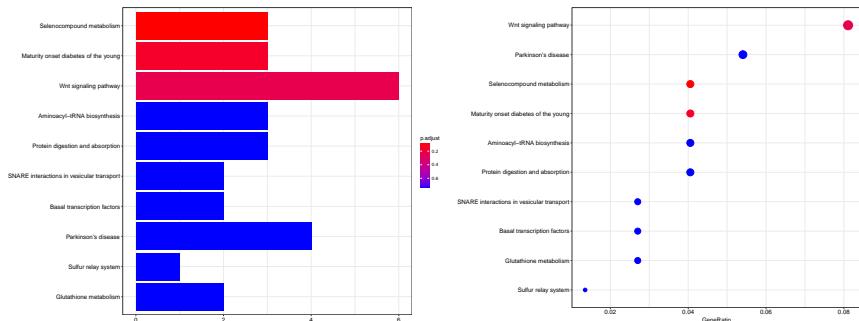


图 18: KEGG富集分析

12 特定基因互作分析

circos图能够展示特定基因在全基因组中的互作情况，如下图所示：

```
## 结果文件:2.global_stat/*_num_of_interactions_from_part.list
```

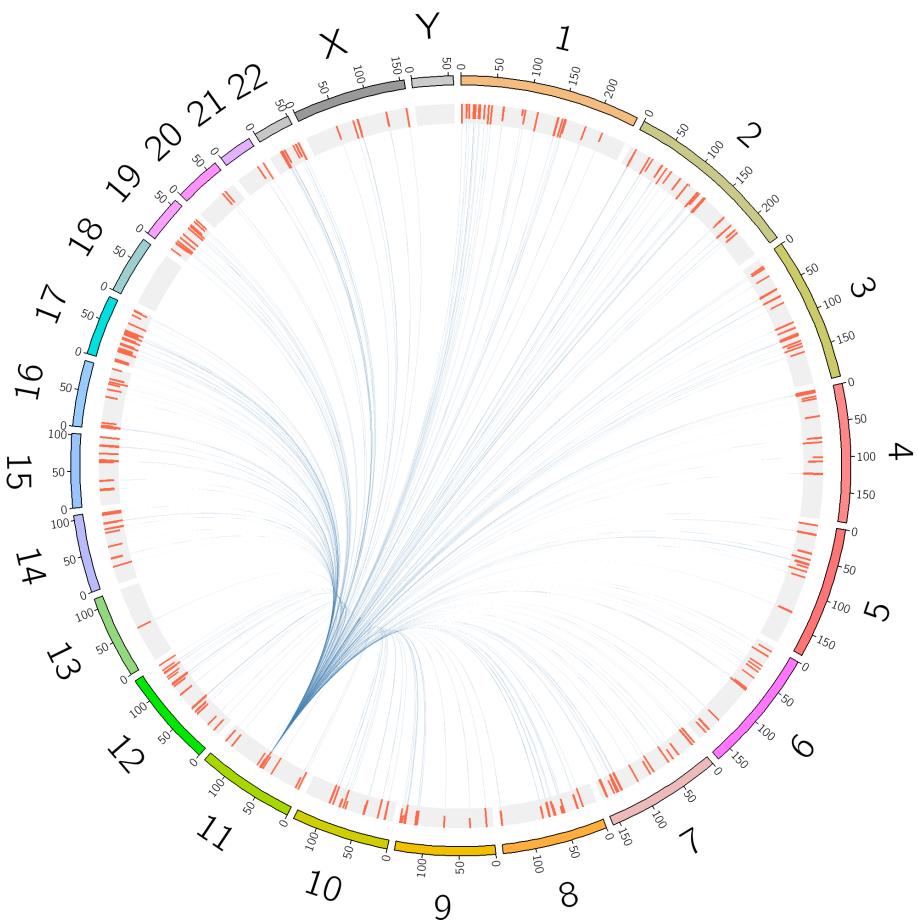


图 19: GO富集分析

图19所示，NEAT1基因在全基因组上的互作强度。

13 参考文章

- [1] Cai Z, Cao CC, Ji L. et al. RIC-seq for global in situ profiling of RNA-RNA spatial interactions. *Nature*, 2020, 582, 432-437.
- [2] Kim D, Langmead B, et al. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, 2015, 12(4):357-360.
- [3] Meng J, Cui X, et al (2013). Exome-based analysis for RNA epigenome sequencing data. *Bioinformatics*, 29, 1565-1567.
- [4] Botstein D, Cherry J M, et al. Gene Ontology: tool for the unification of biology. *Nat genet*, 2000, 25(1): 25-9.
- [5] Draghici S, Khatri P, et al. A systems biology approach for pathway level analysis. *Genome research*, 2007, 17(10): 000-000.