RNA-Seq测序数据分析报告

上海 | 北京 | 南京 | 武汉
电话：400-999-6126
网站：abclonal.com.cn

目录

一. RNA-Seq 服务流程 3
1.1 RNA-Seq 技术的简要介绍 3
1.2 RNA-Seq 数据分析流程 3
二. RNA-Seq 数据分析结果 5
2.1 测序结果原始数据 5
2.1.1 数据格式 5
2.1.2 结果路径 5
2.2 测序数据的质量分析 6
2.2.1 分析目的 6
2.2.2 分析结果 6
2.2.3 结果路径 7
2.3 测序数据比对基因组 8
2.3.1 分析目的 8
2.3.2 分析结果 9
2.3.3 结果路径 10
2.4 转录本的组装，差异表达基因 11
2.4.1 分析目的 11
2.4.2 分析结果 11
2.4.3 结果路径 12
2.5 差异基因功能分析 13
2.5.1 分析结果 13
2.5.2 结果路径 14
2.6 信号通路 15
2.6.1 信号通路图 15
2.6.2 结果路径 15
三、结果分析 16
四、名词解释 17
五、参考文献 18

一. RNA-Seq 服务流程

1.1 RNA-Seq 技术的简要介绍

RNA-Seq技术主要用于细胞或者组织中某一特定时间全基因组的表达水平的研究。真核生物成熟的 mRNA 3’ 端会有一段 poly A的尾巴，通过带有 oligo dT 的磁珠特异性富集细胞或者组织中总mRNA，通过构建mRNA-Seq文库，进行高通量测序，利用生物信息学分析手段，计算出细胞或者组织中不同基因的表达水平。

1.2 RNA-Seq 数据分析流程

RNA-Seq数据分析流程包括：

测序数据质量的分析；

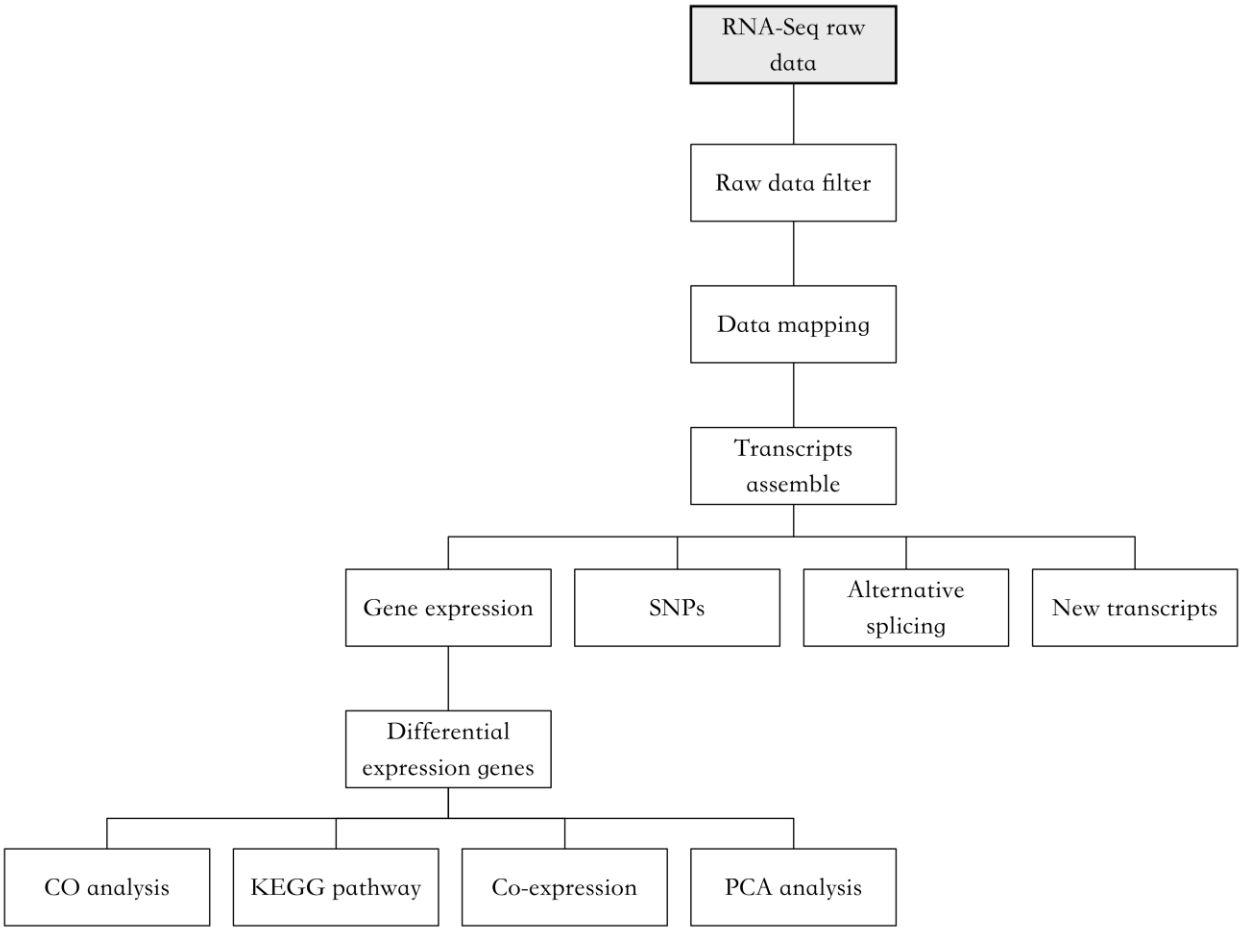
测序数据比对到参考基因组；

转录本的组装和拼接,差异表达基因的查找；

差异表达基因 GO 功能的分析；

差异表达基因 KEGG pathway 通路的分析；

差异表达基因聚类分析等。



图一 RNA-Seq 数据分析流程示意图

二. RNA-Seq 数据分析结果

2.1 测序结果原始数据

2.1.1 数据格式

我们采用Illumina Hiseq-2500测序平台，测序仪将原始的荧光信号转换成对应的碱基序列。测序的结果以 fastq 格式的文件存储，测序得到的原始数据我们称之为 raw data，经过处理之后的数据我们称为 clean data。fastq 格式如下所示:

```
@HWUSI-EAS100R:6:73:941:1973#0/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAG
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCC
```

每一条序列由 4 行信息存储。第一行表示序列的名称，包含了序列在测序仪上的位置信息；第二行是序列的碱基序列；第三行一般是“+”；第四行是序列每个碱基的质量分数。碱基的质量分数通过 ASCII 码编码，对应字符的 ASCII 值减去33即为对应碱基的质量分数。碱基的质量分数越高表示测序错误概率越低。

测序碱基质量分数!	错误发“率!	准确率!
10!	1!in!10!	90%!
20!	1!in!100!	99%!
30!	1!in!1000!	99.9%!
40!	1!in!10000!	99.99%!

图二 碱基质量分数对应测序错误概率

2.1.2 结果路径

1.raw data/Control.fastq.gz 测序原始数据
1.raw data/Treat.fasta.gz 测序原始数据

2.2 测序数据的质量分析

2.2.1 分析目的

测序数据的质量统计涉包括：测序数据质量的评估、低质量数据的过滤以及测序接头序列的去除。只有质量合格的数据才能用于后期的数据分析。

2.2.2 分析结果

测序数据每碱基质量分布图是统计所有测序reads每碱基的质量分布的结果；盒形图是分位数统计的结果；热图表示每个位点质量分数分布的密度，红色表示高密度，蓝色表示低密度。从图下可知，大部分数据质量分数分布都在30以上。

图三 测序数据每碱基质量分布图

测序数据每位点碱基分布结果：由于RNA-Seq建库过程中，mRNA是随机打断的，理论上所有测序的reads中每个位点的碱基是平均分布的。Illumina测序文库构建需要给片段加上接头序列，由于接头序列的存在，测序刚开始的10个碱基分布会有波动。

图四 测序数据每位点碱基分布结果

2.2.3 结果路径

Control.NVC_plot.pdf	Control 组测序数据碱基分布
Control.qual.boxplot.pdf	Control 组测序数据质量分布盒形图
Control.qual.heatmap.pdf	Control 组测序数据质量分布热图
Treat.NVC_plot.pdf	Treat 组测序数据碱基分布
Treat.qual.boxplot.pdf	Treat 组测序数据质量分布盒形图
Treat.qual.heatmap.pdf	Treat 组测序数据质量分布热图

2.3 测序数据比对基因组

2.3.1 分析目的

测序得到的数据一般是 50-100bp 的 DNA 片段，这样的 DNA 片段称为 reads。需将这些 reads 重新比对到基因组上，拼接和组装成转录本后利用 tophat 软件，将测序片段比对到人源基因组hg19。（每条 reads 最多允许2个错误比对碱基）

Sample	reads length	Raw reads number	Clean reads number	Mapped reads number	Mapped ratio
Control	50	12,880,377	12,553,102	11,628,028	92.60%
Treat	50	16,695,704	16,176,209	15,405,335	95.20%

表一 测序结果比对参考基因组统计

Reads length：测序片段的长度
Raw reads number：测序原始数据的 reads 数目
Clean reads number：测序原始数据过滤之后的 reads 数目
Mapped reads number：测序数据比对到参考基因组的数目
Mapped ratio：测序数据比对到参考基因组的比例

2.3.2 分析结果

测序 reads 在基因上分布：用于判断测序文库是否具有偏好性，质量高的文库理论上在基因上的分布没有偏好性。由于建库过程中 RNA 的降解，或者反转录过程中的偏好性，会导致测序的 reads 特异性的 mapping 到基因的 5’端或者 3‘端。

图五 测序 reads 在基因上分布

测序数据饱和度分析：用于判断测序数据量是否能满足分析的要求。随着测序数量的增加找到的可变剪接的 junction 接近饱和，说明测序的数据能满足分析的要求。

图六 测序数据饱和度分析

将基因按照表达值从高到低分成四部分，然后计算随着数据量的增加基因表达值变化的方差。随着数据量的增加，方差降低，说明测序数据量满足分析的要求。

图七基因按照表达值高到低显示图

2.3.3结果路径

3.mapping/Control_mapping/	
- Control_mapping/Control.geneBodyCoverage.curves.pdf	测序 reads 在基因上分布
-Control_mappinh/Control.junctionSaturation_plot.pdf	junction 饱和度分析
-Control_mapping/Control.saturation.pdf	表达值方差分析
-Control_mapping/Control_accepted_hits.bam	比对结果 BAM 文件
3.mapping/Treat_mapping/	
- Treat_mapping/Treat.geneBodyCoverage.curves.pdf	测序 reads 在基因上分布
-Treat_mappinh/Treat.junctionSaturation_plot.pdf	junction 饱和度分析
-Treat_mapping/Treat.saturation.pdf	表达值方差分析
-Treat_mapping/Treat_accepted_hits.bam	比对结果 BAM 文件
3.mapping/Treat_mapping/	
mapping_results_stats.xlsx	比对结果统计

2.4 转录本的组装，差异表达基因

2.4.1 分析目的

根据不同样本中基因的表达水平，找到样本之间差异表达的基因。

2.4.2 分析结果

基因表达的 MA 图：横坐标表示样本表达量的平均值，纵坐标表示样本表达量倍数的差异。红色点表示差异显著的基因，黑色点表达表达差异不显著的基因。

图八 基因表达的 MA 图

基因表达的火山图：图中每个点代表一个基因，横坐标表示基因表达值倍数差异，取 log2 的对数。纵坐标表示基因表达值差异的显著性水平 p value，取 log10 的对数。红色点表示表达差异显著的基因，黑色点表示表达差异不显著的基因。

图九 基因表达的火山图

RNA-Seq 差异表达基因的热图。红色表示基因表达值高，绿色表示基因表达值低。基因的表达值取 log2 的对数。

图十 RNA-Seq 差异表达基因的热图

2.4.3 结果路径

4.differential_expression/	
Control_Treat_MAplot.png	差异表达基因的 MA 图
Control_Treat_Volcano.png	差异表达基因的火山图
Control_Treat_genes_expression.txt	所有表达基因的表达值 txt 格式
Control_Treat_genes_expression.xlsx	所有表达基因的表达值 excel 格式
Control_Treat_significant_differential_expression_genes.txt	显著差异表达基因 txt 格式
Control_Treat_significant_differential_expression_genes.xlsx	显著差异表达基因 excel 格式
Control_Treat_significant_down_regulated_genes.txt	表达下调基因 txt 格式
Control_Treat_significant_down_regulated_genes.xlsx	表达下调基因 excel 格式
Control_Treat_significant_up_regulated_genes.txt	表达上调基因 txt 格式
Control_Treat_significant_up_regulated_genes.xlsx	表达上调基因 excel 格式
Differential expression genes heatmap.png	差异表达基因热图

2.5 差异基因功能分析

2.5.1 分析结果

Gene ontology 简称 GO 分析，是国际基因功能标准分类体系，主要是将分析得到的差异基因与待分析物种中基因组背景比较，通过超几何分布检验，找出差异表达基因显著富集的 GO 条目。筛选得到差异基因，将基因的功能进行注释，找到差异基因参与的生物学过程，细胞内的组分以及分子功能三方面的生物学意义。

图十一 生物学过程

图十二 细胞内组分

图十三 分子功能

GO 分析结果列表

Category	Term	Count	%	PValue	Genes	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR

Category：GO 功能分析分类（BP：生物学过程；CC：细胞内组分；MF：分子功能）
Term：GO 功能的描述，GO 数据库中条目的编号
Count：富集在该 GO 条目的差异基因数目
%：富集在该 GO 条目的基因占所有具有功能注释的差异基因的比例
Pvalue：GO 功能富集的显著性水平
Genes：富集在该条目的基因
List total：具有功能注释的差异基因

Pop hits：基因组背景中与该条目相关的基因数目
Pop total：基因组背景中所有具有注释的基因
Fold enrichment：富集倍数
Bonferroni：Bonferroni 校准之后的 pvalue 值
Benjamini：Benjamini 方法校准之后的 pvalue 值
FDR：错误发生率，也成为 q-value。

2.5.2结果路径

5.gene ontology/differential expression genes GO/
- Control_Treat_BP.png
- Control_Treat_BP.txt
- Control_Treat_BP.xlsx
- Control_Treat_CC.png
- Control_Treat_CC.txt
- Control_Treat_CC.xlsx
- Control_Treat_MF.png
- Control_Treat_MF.txt
- Control_Treat_MF.xlsx

5.gene ontology/down regulated genes GO/
- down_regulated_BP.png
- down_regulated_BP.txt
- down_regulated_BP.xlsx
- down_regulated_CC.png
- down_regulated_CC.txt
- down_regulated_CC.xlsx
- down_regulated_MF.png
- down_regulated_MF.txt
- down_regulated_MF.xlsx

5.gene ontology/up regulated genes GO/
- up_regulated_BP.png
- up_regulated_BP.txt
- up_regulated_BP.xlsx
- up_regulated_CC.png
- up_regulated_CC.txt
- up_regulated_CC.xlsx
- up_regulated_MF.png
- up_regulated_MF.txt
- up_regulated_MF.xlsx

差异基因功能分析结果
- 生物学过程
  - 生物学过程列表 txt 格式
  - 生物学过程列表 excel 格式
- 细胞内组分
  - 细胞内组分列表 txt 格式
  - 细胞内组分列表 excel 格式
- 分子功能
  - 分子功能列表 txt 格式
  - 分子功能列表 excel 格式

下调基因功能分析结果
- 生物学过程
- 生物学过程列表 txt 格式
- 生物学过程列表 excel 格式
- 细胞内组分
- 细胞内组分列表 txt 格式
- 细胞内组分列表 excel 格式
- 分子功能
- 分子功能列表 txt 格式
- 分子功能列表 excel 格式

上调基因功能分析结果
- 生物学过程
  - 生物学过程列表 txt 格式
- 生物学过程列表 excel 格式
- 细胞内组分
  - 细胞内组分列表 txt 格式
- 细胞内组分列表 excel 格式
- 分子功能
  - 分子功能列表 txt 格式
- 分子功能列表 excel 格式

2.6 信号通路

2.6.1 信号通路图

红色表示高表达的差异基因，蓝绿色表示低表达的差异基因

图十四 KEGG 信号通路图

2.6.2结果路径

6.KEGG pathway/	
Control_Treat_pathway.txt	差异基因 KEGG 信号通路列表 txt 格式
Control_Treat_pathway.xlsx	差异基因 KEGG 信号通路列表 excel 格式
KEGG pathway.png	差异基因 KEGG 信号通路
down regulated KEGG pathway.png	下调基因 KEGG 信号通路
down_regulated_pathway.txt	下调基因 KEGG 信号通路 txt 格式
down_regulated_pathway.xlsx	下调基因 KEGG 信号通路 excel 格式
up regulated KEGG pathway.png	上调基因 KEGG 信号通路
up_regulated_pathway.txt	上调基因 KEGG 信号通路 txt 格式
up_regulated_pathway.xlsx	上调基因 KEGG 信号通路 excel 格式
-map*.png	差异基因 KEGG 信号通路图

三、结果分析

我们对两组人源的细胞样品进行了 RNA-Seq 的实验操作和数据分析。测序得到的原始数据，正常组 reads 数目为 12,880,377，实验组 reads 数目为 16,695,704。测序数据质量分数 Q30 比例在 95%以上，测序数据数据量和测序质量都符合预期。测序得到的结果比对到人源的参考基因组 hg19，比对的结果 92.60%和 95.20%的数据都能正确比对到基因组上。比对结果数据的饱和度分析和数据偏好性分析都符合预期，数据的饱和度在 80%以上，测序数据在基因本体上分布均匀，没有偏好性。通过 DESeq 软件进行差异基因的分析，我们找到 681 个差异表达的基因，其中 365 个基因在实验组中表达下调，316 个基因在实验组中表达上调。通过对差异表达基因进行功能分析，结果显示，差异表达基因主要富集在与细胞周期，DNA 复制相关的生物学过程中。细胞内信号通路的分析结果也显示 cell cycle 相关的通路发生了显著的变化。此外，我们分析还发现，细胞周期相关的基因表达主要是降低的。而在实验组中，表达上调的基因主要参与了细胞外基质，TGF-beta 信号通路；表达下调的基因主要参与 cell cycle，DNA replication 等生物学过程。我们建议您对相关信号通路差异表达的基因进行实验验证。

四、名词解释

RPKM 值：每一百万条测序片段中 mapping 到 1KB 碱基长度的基因组上 reads 的数目。RPKM 值主要用于表示基因的表达量。

Mapping：测序片段比对到参考基因组上成为测序数据的 mapping。

Reads：测序得到的片段称为 reads。

测序深度：测序得到的碱基数目相对于待测区域的比例。

五、参考文献

Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments[J]. Bioinformatics, 2012, 28(16): 2184-2185.

Trapnell C, Pachter L, Salzberg S L. TopHat: discovering splice junctions with RNA-Seq[J]. Bioinformatics, 2009, 25(9): 1105-1111.

Anders S, Huber W. Differential expression of RNA-Seq data at the gene level–the DESeq package[J]. Heidelberg, Germany: European Molecular Biology Laboratory (EMBL), 2012.

Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. (2009) Nat Protoc. 4(1):44 -57.