miRNA-seq 数据分析

miRNA-seq Data Analysis

赵婧，刘南*

中国科学院生物与化学交叉研究中心，上海

*通讯作者邮箱: liunan@sioc.ac.cn

引用格式：赵婧，刘南．(2019)．miRNA-seq 数据分析．*Bio-101* e1010249．Doi: 10.21769/BioProtoc.1010249.

How to cite: Zhao, J and Liu, N. (2019). miRNA-seq data analysis. *Bio-101* e1010249. Doi: 10.21769/BioProtoc.1010249. (in Chinese)

摘要：Small RNA 是生物体内一类重要的功能分子，包括 miRNA、siRNA 和 piRNA。它的主要功能是诱导基因沉默，调控细胞生长、发育、基因转录和翻译等生物学过程。基于新一代高通量测序技术的 small RNA 测序，可一次性获得数百万条 small RNA 序列，能够快速鉴定某种组织在特定状态下的所有已知 small RNA，并发现新的 small RNA，为 small RNA 功能研究提供有力工具。本实验教程以 miRNA 为例，详述了 small RNA 测序数据的分析流程。

关键词：miRNA-seq，测序定量，差异表达，靶标预测

仪器设备

1. 服务器 (型号: I620-G20; 操作系统: centos7; CPU: Xeon(R) CPU E5-2620 v4 @ 2.10GHz, 8 核, 32G 内存)

注: 学习者最好是有 mac 或者 linux 系统, 需要 8G+ 的内存, 500G 的存储, 如果是 Windows, 则需要安装 git, notepad++, 虚拟机 (安装 linux 系统) 等, 建议初学者使用 mac。

2. 普通个人电脑 (需要已经安装好 R, 能调用个人终端)

软件

1. FASTX-Toolkit (v0.014)
2. FastQC (v0.11.8)

3. mulitiQC(v1.6)
4. miRDeep2(v2.0.0) (Anders 和 Huber, 2010)
5. DESeq(v1.34.0) (Friedlander 等, 2012)
6. R(v3.5.1)

实验步骤

注:

- 1) 本实验教程需要读者了解基本的 **linux** 命令行, 二代测序原理等。相关知识可以参考如下链接学习 <https://www.plob.org/article/11214.html>。
- 2) 本实验教程是基于已经在服务器和个人电脑上安装好相关软件进行的。如果安装出现问题, 请参考以下链接:

FASTX-Toolkit: http://hannonlab.cshl.edu/fastx_toolkit/download.html

FastQC: <https://www.bioinformatics.babraham.ac.uk/projects/download.html>

mulitiQC: <https://multiqc.info/>

miRDeep2: <https://anaconda.org/bioconda/mirdeep2>
http://blog.sina.com.cn/s/blog_7cffd1400101m3i3.html
https://drmirdeep.github.io/mirdeep2_tutorial.html

DESeq: <https://bioconductor.org/packages/release/bioc/html/DESeq.html>

1. miRNA-seq 数据获得

本实验方案以 2017 年发表在 *Oncotarget* 杂志上的文章: “Systematic analysis reveals tumor-enhancing and -suppressing microRNAs in *Drosophila* epithelial tumors.” (Shu 等, 2017) 中 miRNA-seq 的部分数据作为测试数据。

2. 去掉 adapter

在终端输入命令: `$ fastx_clipper -a adapter -c -i 输入文件名.fastq -z -o`; 输出文件名.fastq.gz, 操作如下图所示:

```
$ fastx_clipper -a AGATCGGAAGAGCACACGTCT -l 18 -c -i 3DH_S1_L001_R1_001.fastq -z -o ../output/3DH.fastq.gz
```

3. 通过 FastQC 对数据进行质量评估

`$ fastqc 所在目录/fastqc -t 4 输入文件`

注：若想指定输出目录，可以增加参数“-o 路径”设置。

操作如下图所示：

```
$ /opt/biosoft/FastQC/fastqc -t 4 3DH.fastq
```

在输出文件夹中每个输入文件都有两个输出结果，一个结果集合压缩包 (.zip 文件) 和一个网页结果 (.html 文件)。一般通过查看网页结果即可知道基本的测序质量。

结果如图 1：

图 1. fastqc 质控结果. A. 表示每个碱基的测序质量，在绿色区域表示测序质量很高；B. 展示了测序测到的长度分布，主要分布在 21~23 nt，这与 miRNA 的大小符合。

当具有多个样品时，查看多个 fastqc 的结果时可以使用 multiQC 软件，将所有的结果综合到一起查看，输入命令：

```
$ multiqc 目标文件夹
```

查看方式与结果与 fastqc 是一致的，在此就不进行赘述。

4. 对测序文件进行比对

目前比对 miRNA 的程序很多，比如 miRDeep2、miRExpress、miRNAkey 以及 sRNAbench 等，同时还有一些在线的分析软件。我们采用综合指标比较好的 miRDeep2 进行比对 (Bisgin 等，2018)。

1) 建立 index 文件，把 fastq 文件转换为 fasta 文件，代码如下：

```
$ bowtie-build 基因组文件输出文件名
```

注：这里的使用的是从 flybase 上下载的果蝇参考基因组文件。

操作如图所示：

```
$ bowtie-build dmel-all-chromosome-r6.24.fasta dmel-all-chromosome-r6.24
```

2) 将测序文件与参考基因组进行比对，比对命令如下：

```
$ mapper.pl 输入文件 -e -h -i -j -m -k adapter -l 18 -p 参考基因组的
index -s 处理过的 reads 输出的文件名 -t 输出文件名.arf
```

操作如图所示：

```
$ mapper.pl 3DH.fastq -e -h -i -j -m -k AGATCGGAAGAGCACAC
GTCT -l 18 -p ../genomo/dme-6.24-all-chromosome -s 3DH_mapper.fa -t 3DH_mapper_v
s_genome.arf -o 16 -n
```

其中所涉及到的参数具体含义如下：

-c 输入文件是 **fasta** 格式

-h 解析为 **fasta** 格式

-i 将 **rna** 转换为 **dna** 字母表 (以映射基因组)

-j 删除所有包含字母序列的条目，除了 **a, c, g, t, u, n, A, C, G, T, U, N** 之外

-m collapse reads

-k 3'适配器序列

-l n 忽视长度低于 n 的序列

-p 将处理过的 **reads** map 到之前建立过索引的基因组上，注意输入的是 **index** 文件的前缀

-s 指出将处理过的 **reads** 输出到某个文件，自己命名

-t 指出将 **mapping** 的结果输出到某个文件，自己命名，必须是 **.arf** 文件

-o 设置线程数

运行结束后，在终端屏幕上会显示出一个比对结果的 **summary**，如图 2 所示：

#desc	total	mapped	unmapped	%mapped	%unmapped
total:	4397972	3899036	498936	0.887	0.113
seq:	4397972	3899036	498936	0.887	0.113

图 2. miRDeep2 比对输出结果

3) 计算比对到已知 **miRNA** 的 **counts** 数

```
$ quantifier.pl -p 前体序列参考文件.fa -m 成熟序列参考文件.fa -r 上
一步处理过的 reads 输出文件 -s 从 miRBase 上下载的 star 序列文件 -t 物种名
称 -y 文件名后缀
```

操作如图所示：

```
$ quantifier.pl -p ../genomo/hairpin.fa -m ../genomo/mature.fa -r 3DH_mapper.fa -s ../genomo/miRNA.str.fa -y now -t dme
```

运行结束后，会生成一个 `miRNAs_expressed_all_samples_now.csv` 文件。所有测到的已知 miRNA 的 counts 等信息都在这个文件中。我们后续的分析都是基于这个文件进行的。

4) 鉴定测序得到的未知 miRNA

一般情况下，鉴定新的基因不常用，但是如果需要也可以进行以下的操作：
miRDeep2.pl 处理过的 reads 输出文件基因组文件处理过的输出文件名.arf 成熟 miRNA 文件其他物种的成熟 miRNA 文件研究物种 miRNA 前体的文件-t 物种名称
2>report.log

如果只有 reads，arf 文件和 genome 文件，需要用 none 表示成熟 miRNA 文件、其他物种的成熟 miRNA 文件和研究物种 miRNA 前体的文件。

操作如图所示：

```
$ miRDeep2.pl 3DH_mapper.fa ../genomo/dmel-all-chromosome-r6.24.fasta 3DH_mapper_vs_genome.arf none none none -t Fruitfly 2>report.log
```

运行结束后，结果会保存在.html 文件中。

5. miRNA 表达矩阵以及差异表达分析

miRNA 可以使用 [DESeq \(DESeq2\)](#), edgeR 以及 limma 等 R 包进行差异表达分析。由于示例数据是没有重复的数据，所以我们采用 DESeq 对数据进行分析 (DESeq2 不支持没有重复组的分析，建议读者使用具有重复的实验数据)。将 R 的工作环境设置到含有输出的 `miRNAs_expression` 的文件夹中，运行如下脚本 (图 3)：

```
library(DESeq)
#Combined the mature microRNA read counts into one excel file
#Read the data to R
miRNA <- read.csv("miRNAs_expressed.csv", sep = ",", header = T,
                  stringsAsFactors = F, row.names = 1)
mirnaCountTable <- miRNA[,c(7,5)]
mirnaDesign = data.frame(
  row.names = colnames(mirnaCountTable),
  condition = c( "HD3", "HD30" ),
  libType = c( "single-end", "single-end" ))

singleSample = mirnaDesign$libType == "single-end"
countTable = mirnaCountTable[ , singleSample ]
condition = mirnaDesign$condition[ singleSample ]
head(countTable)

#Normalisation
cds <- newCountDataSet( countTable, condition )
cds <- estimateSizeFactors( cds )
sizeFactors( cds )
head( counts( cds, normalized = TRUE ) )

#Variance estimation
cds <- estimateDispersions( cds , method="blind", sharingMode="fit-only" )
str ( fitInfo(cds) )

#comparison between two samples
res <- nbinomTest( cds, "HD3", "HD30" )
head(res)

#Data quality assessment by sample clustering and visualization
#Heatmap of 30 high expression microRNA:
cdsBlind <- estimateDispersions( cds, method = "blind" )
vsd <- varianceStabilizingTransformation( cdsBlind )
library("RColorBrewer")
library("gplots")
select <- order(rowMeans(counts(cds)), decreasing=TRUE)[1:30]
hmcol <- colorRampPalette(brewer.pal(9, "GnBu"))(100)
heatmap.2(exprs(vsd)[select,], col = hmcol, trace="none", margin=c(10, 6))
```

图 3. DESeq 分析 R 脚本

运行脚本后我们可以获得 miRNA 差异基因表格以及差异最显著的前 30 个 miRNA 的热图 (如图 4)。

图 4. miRNA 差异基因热图

6. 下游分析

目前进行 miRNA 测序数据分析，大多是要和 mRNA 测序数据关联来看。这里就涉及到了 miRNA 寻找对应的 targets 问题。目前寻找 miRNA 靶点的方法主要分为计算机预测和实验的方法。计算机预测靶点，可以参考该网站的介绍 <https://www.plob.org/article/1157.html>。

而实验方法，目前在果蝇方面，我们实验室首次实现了一种可以检测个体水平上的 miRNA 靶标的实验 Ago1-RIP-seq，通过这种方法我们可以探究已知以及潜在的 miRNA 靶标，为 miRNA 的后续分析提供了非常坚实的实验指导。目前我们正在开发 Ago1-RIP-seq 的数据库，很快可以上线，为广大的科研人员提供服务。

参考文献

1. Anders, S. and Huber, W. (2010). [Differential expression analysis for sequence count data](#). *Genome Biol* 11(10): R106.

2. Bisgin, H., Gong, B., Wang, Y. and Tong, W. (2018). [Evaluation of bioinformatics approaches for next-generation sequencing analysis of microRNAs with a toxicogenomics study design.](#) *Front Genet* 9: 22.
3. Friedlander, M. R., Mackowiak, S. D., Li, N., Chen, W. and Rajewsky, N. (2012). [miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades.](#) *Nucleic Acids Res* 40(1): 37-52.
4. Shu, Z., Huang, Y. C., Palmer, W. H., Tamori, Y., Xie, G., Wang, H., Liu, N. and Deng, W. M. (2017). [Systematic analysis reveals tumor-enhancing and -suppressing microRNAs in *Drosophila* epithelial tumors.](#) *Oncotarget* 8(65): 108825-108839.