# miRNA 高通量测序案例

## 上海生物科技有限公司



目录

# 项目基本信息

客户方委托进行“miRNA 测序和生物信息分析”的专项技术服务。

技术服务的内容：

（1）根据客户方提供高质量的 RNA 及准确的背景信息资料，对样品进行检测，样品检测合格后进行样品制备；

（2）利用 illumina 的 Solexa 高通量测序平台构建 miRNA 库进行样本表达谱测序；

（3）技术服务质量：Solexa 测序平台，每个 lane 是 1×50bp 的程序进行测序。

（4）根据客户方要求，完成相关的生物信息分析工作：

a) 测序质量评估；

b) 测序低质量 read 过滤；

c) 已知 miRNA 检测、表达差异分析；

d) novel miRNA 预测、表达差异分析；

e）统计病毒 siRNA；

f）测序结果中各类 RNA 的归类及其比例分析（Rfam 比对）；

项目摘要：

本项目中，我们对测序原始数据进行了质量评估及过滤。根据分析结果，我们认为此次测序得到的数据质量良好(Q20 >98%)，有效数据质量和留存比例均较高(平均为 72.58%)，适用于 miRNA 的后续分析。我们使用 miRDeep2 对过滤后的 read 进行分析，两个样本共检测到已知 miRNA 196 个（其中 144 个为共有）；使用 DEGseq 分析得出差异表达的 miRNA 为 107 个；并依据基因组序列预测得出 novel miRNA 354 个。

# 分析报告

## 1、原始测序 READS 质量基本统计

通过 Solexa RNA 的测序，得到的样本原始数据：测序得到的原始图像数据经 base calling 转化为序列数据，我们称之为原始数据(raw data)。其数据格式为 fastq 格式， 内容包括测序 reads 的名称，序列以及测序质量。在 fastq 格式文件中每个 read 由四行描述：

@HWI-EAS413_4:1:5:1004:56 <= Sequence name - lane:read(1|2):tile:X:Y

GTTTCATTTCTAAACCTGTTTCATTACAAAATGCC <= Sequence

\+ <= repeat of name

ZVZZVZZZZZVZSZZLZFZZZVVZLSZSZZQVLVT <= ASCII quality score. A = low, Z(or other symbol) = high.

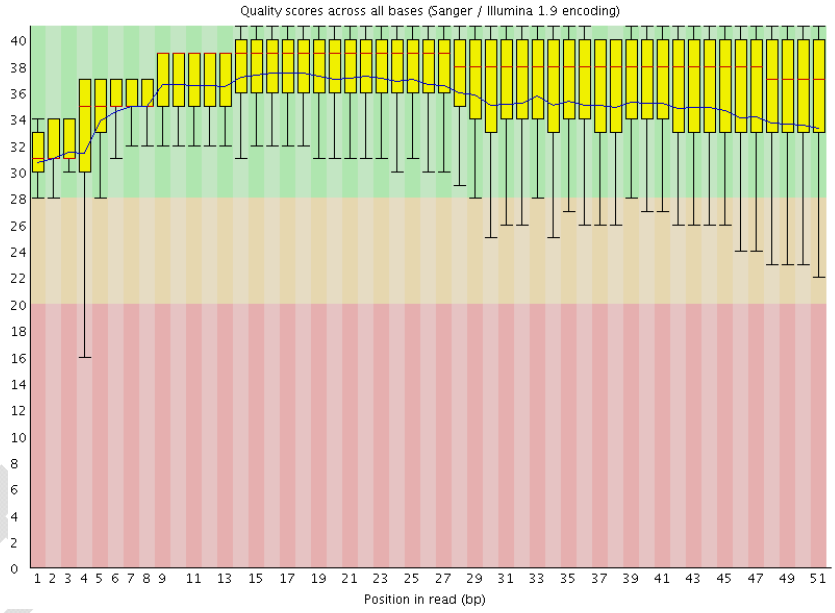
第 1 行和第 3 行是序列名称，由测序仪产生；第 2 行是序列；第 4 行是序列的测序质量 数据文件： /rawdata 目录下*.fq 文件

现对原始测序 READS 进行质量统计：

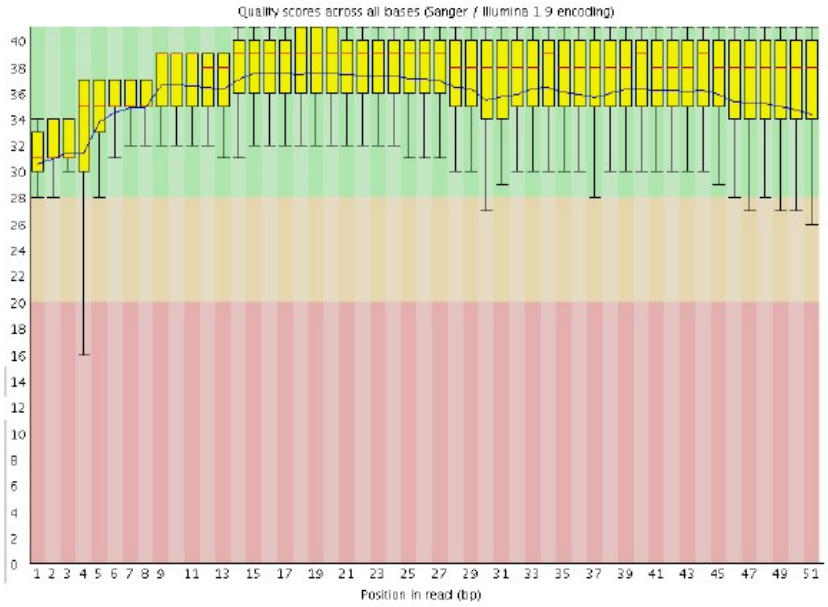
| Sample | 2bunT | Fny2B |
|---|---|---|
| Total read | 10321014 | 9469806 |
| Length | 51 | 51 |
| Q>20 read | 10190281 | 9394619 |
| Q>20 (%) | 98.73% | 99.21% |
| Q>10 read | 10316713 | 9465320 |
| Q>10 (%) | 99.96% | 99.95% |

**注：Q>20(%)=Q>20 read/ Total read*100；Q10 同 Q20。根据 Illumina 的技术文档，当质量值为 10 时，错误率为 10%，20 为 1%。**

S1:




S2:




结论：样本的原始 read 数据质量均很好，各位点碱基质量中位数均大于 30，Q20 read 比例大于 98%，Q10 read 比例近似为 100%，满足后续分析要求。

## 2、READS 质量预处理

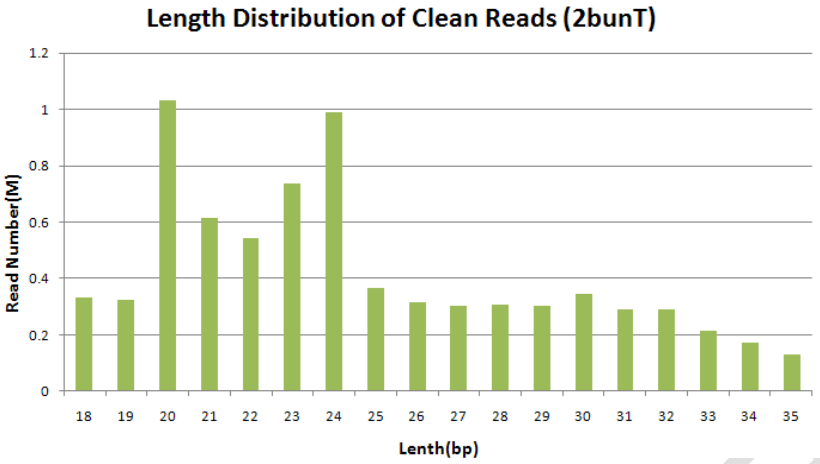
鉴于 Solexa 数据错误率对结果的影响，我们对其进行去接头，去低质量，去污染等处理，得到 Clean 的序列数据，并对 Clean 序列进行长度分布统计。数据质量预处理步骤：

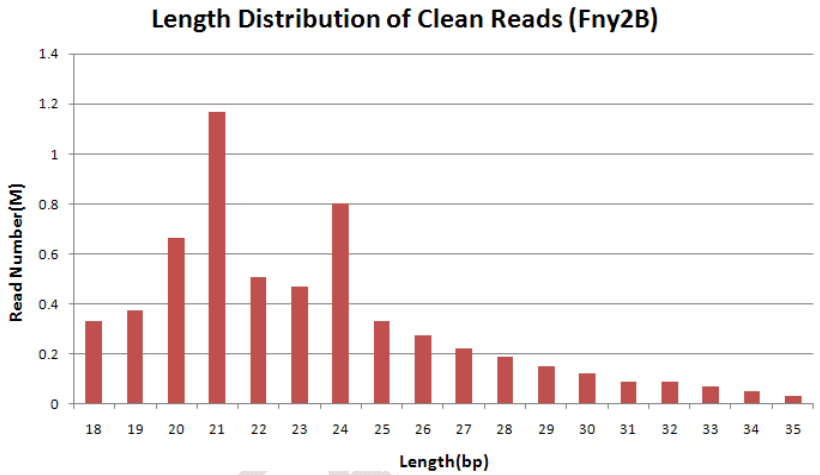
1) 去除低质量 read：质量阈值 20（错误率=1%），比例阈值 40%
2) 去除 reads 中含 N 部分比例较大序列：比例阈值 4%
3) 去除接头序列
4) 截取有效长度区间序列（18-35）
5) 统计 Clean 序列长度分布

质量预处理前后数据结果统计表格如下：

| Sample | Fastq | Clean | Ratio(%) |
|---|---|---|---|
| 2bunT | 10321014 | 8305860 | 80.48% |
| Fny2B | 9469806 | 6125485 | 64.68% |

Clean 序列的长度分布统计图如下：



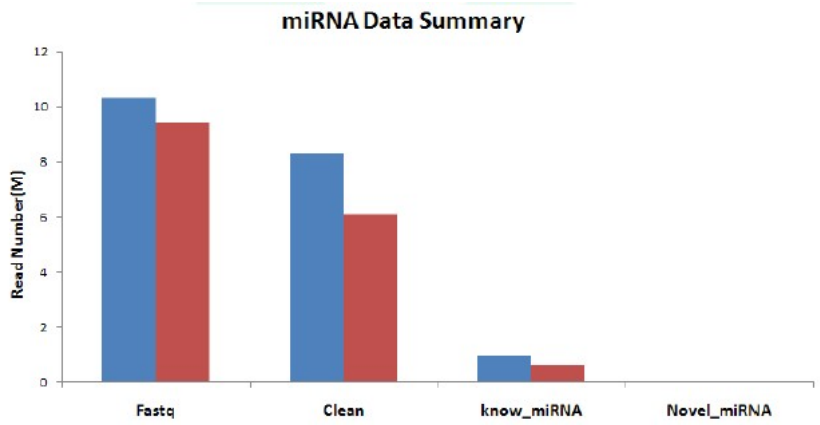

结论：有效数据留存率较高(平均为 72.58%)，有效数据长度主要分布在 20-24bp 左右，符合 miRNA 后续分析要求。

## 3、miRNA 分析摘要

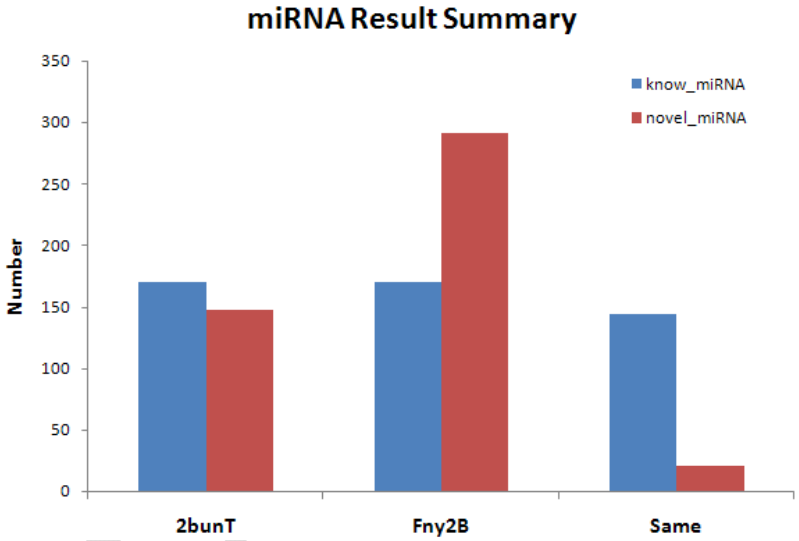
主要使用 miRDeep 软件，对 miRNA RNA-seq 数据进行分析，得出已知 miRNA（来自 miRBase）的种类及表达量，同时预测新的 miRNA，分析结果摘要如下：

数据文件：\data\ summary_data.xlsx

| Sample | Fastq | Clean | know_miRNA | Novel_miRNA |
|---|---|---|---|---|
| 2bunT | 10321014 | 8305860 | 1024292 | 10258 |
| Fny2B | 9469806 | 6125485 | 632821 | 12286 |




| Sample | know_miRNA | novel_miRNA |
|---|---|---|
| 2bunT | 170 | 147 |
| Fny2B | 170 | 291 |
| Same | 144 | 20 |

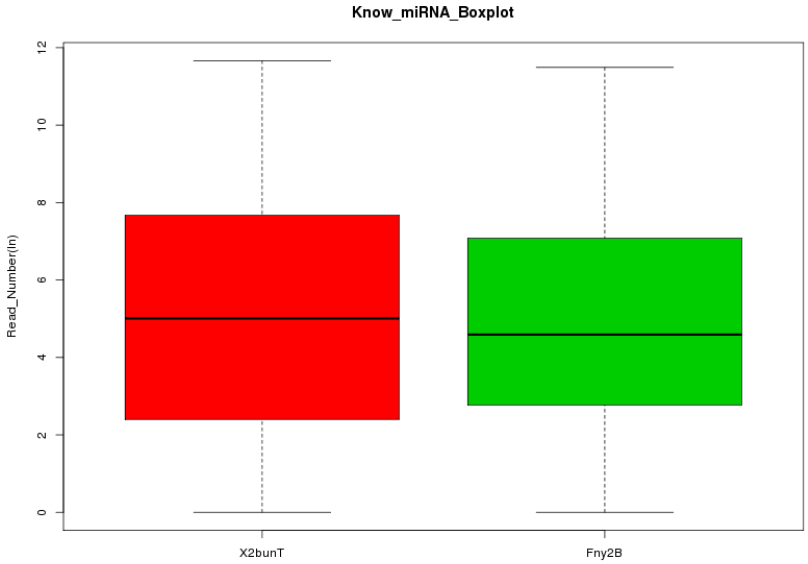



## 4、Known miRNA 分析

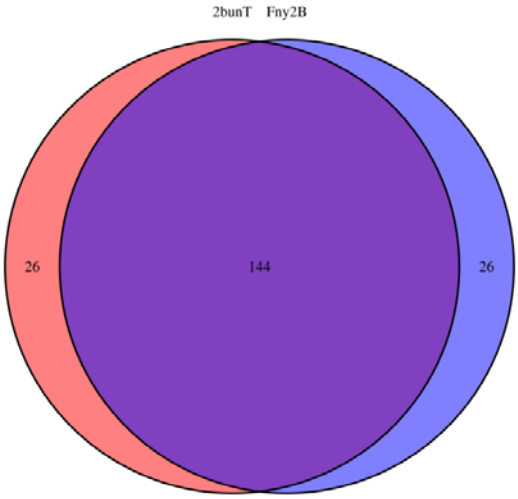
将得到的序列和 miRBase 中已知的 miRNA 进行比对，并对检测到的 miRNA 进行差异表达分析。

数据文件：\data\know miRNA 目录下文件

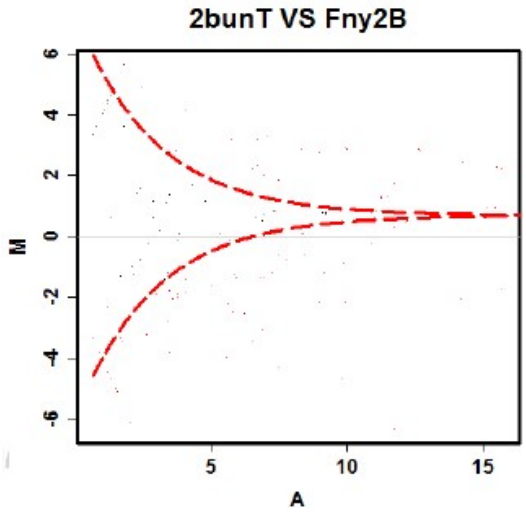
Know miRNA 表达量 boxplot 图：



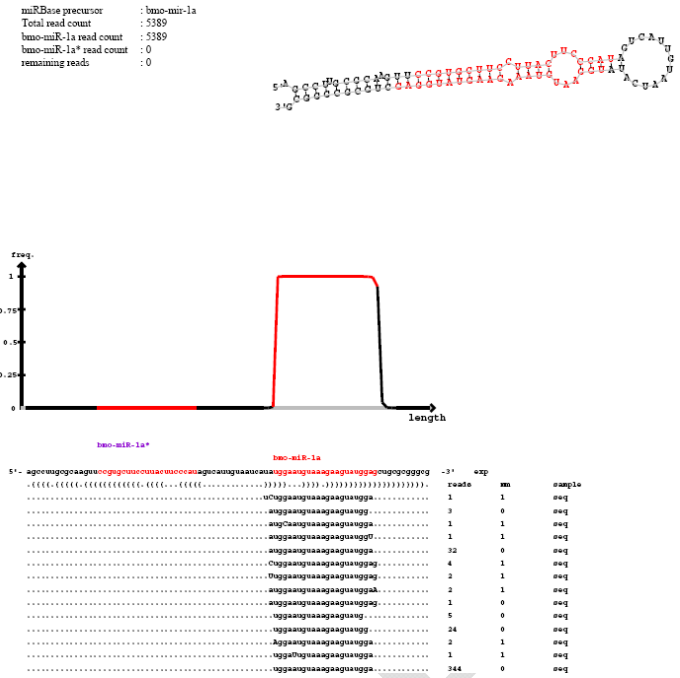

两个样本中检测到的 miRNA 种类 Venn 图：




已知 miRNA 表达量差异分析（DEGseq）：



已知 miRNA 二级结构及检测的 read 与之 mapper 图（pdf）（图例）：




## 5、Novel miRNA 预测

使用 miRDeep2 软件进行 novel miRNA 进行预测，结果文件给出所有的预测结果及最佳预测结果，并对两个样本预测的 novel miRNA 进行了合并处理。

数据文件：\data\novel miRNA 目录下文件

单个样本预测出的 miRNA 概要信息，其中 miRDeep2 score>4 的预测结果为较高可信度阳性结果：

| | Survey of miRDeep2 performance for score cut-offs -10 to 10 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | novel miRNAs | | | known miRBase miRNAs | | | | |
| | miRDeep2 score | predicted by miRDeep2 | estimated false | estimated true positives | in species | in data | detected by miRDeep2 | estimated signal-to-noise | excision gearing |
| Best novel miRNA | 10 | 8 | 10 +/- 3 | 1 +/- 1 (7 +/- 14%) | 338 | 154 | 37 (24%) | 2.7 | 5 |
| | 9 | 8 | 11 +/- 3 | 0 +/- 1 (6 +/- 13%) | 338 | 154 | 38 (25%) | 2.7 | 5 |
| | 8 | 10 | 11 +/- 4 | 1 +/- 2 (9 +/- 15%) | 338 | 154 | 38 (25%) | 2.7 | 5 |
| | 7 | 11 | 12 +/- 4 | 1 +/- 2 (8 +/- 14%) | 338 | 154 | 38 (25%) | 2.6 | 5 |
| | 6 | 12 | 13 +/- 4 | 1 +/- 2 (9 +/- 15%) | 338 | 154 | 38 (25%) | 2.5 | 5 |
| | 5 | 16 | 14 +/- 4 | 3 +/- 3 (17 +/- 18%) | 338 | 154 | 57 (37%) | 3.7 | 5 |
| | 4 | 33 | 17 +/- 5 | 16 +/- 5 (49 +/- 14%) | 338 | 154 | 60 (39%) | 4.2 | 5 |
| | 3 | 48 | 40 +/- 6 | 9 +/- 6 (18 +/- 12%) | 338 | 154 | 60 (39%) | 2.2 | 5 |
| | 2 | 64 | 75 +/- 8 | 0 +/- 1 (0 +/- 2%) | 338 | 154 | 62 (40%) | 1.4 | 5 |
| | 1 | 99 | 122 +/- 11 | 0 +/- 1 (0 +/- 1%) | 338 | 154 | 62 (40%) | 1.1 | 5 |
| | 0 | 147 | 192 +/- 13 | 0 +/- 0 (0 +/- 0%) | 338 | 154 | 62 (40%) | 1 | 5 |
| | -1 | 189 | 282 +/- 16 | 0 +/- 0 (0 +/- 0%) | 338 | 154 | 62 (40%) | 0.8 | 5 |
| | -2 | 241 | 376 +/- 17 | 0 +/- 0 (0 +/- 0%) | 338 | 154 | 62 (40%) | 0.7 | 5 |
| | -3 | 308 | 476 +/- 19 | 0 +/- 0 (0 +/- 0%) | 338 | 154 | 62 (40%) | 0.7 | 5 |
| | -4 | 386 | 575 +/- 21 | 0 +/- 0 (0 +/- 0%) | 338 | 154 | 63 (41%) | 0.7 | 5 |
| | -5 | 464 | 668 +/- 21 | 0 +/- 0 (0 +/- 0%) | 338 | 154 | 64 (42%) | 0.7 | 5 |
| | -6 | 537 | 756 +/- 22 | 0 +/- 0 (0 +/- 0%) | 338 | 154 | 64 (42%) | 0.8 | 5 |
| | -7 | 621 | 830 +/- 24 | 0 +/- 0 (0 +/- 0%) | 338 | 154 | 64 (42%) | 0.8 | 5 |
| | -8 | 695 | 893 +/- 25 | 0 +/- 0 (0 +/- 0%) | 338 | 154 | 64 (42%) | 0.8 | 5 |
| | -9 | 768 | 950 +/- 26 | 0 +/- 0 (0 +/- 0%) | 338 | 154 | 64 (42%) | 0.8 | 5 |
| | -10 | 823 | 998 +/- 26 | 0 +/- 0 (0 +/- 0%) | 338 | 154 | 64 (42%) | 0.9 | 5 |

合并的 novel miRNA 信息，包括 unique_seq_ID、长度、miRNA 前体在 genome 上的位置、前体序列、read 个数、miRDeep2 score、具有相同 seed 序列的 miRNA 名称、对应 PDF 名称

等信息，以及 DEGseq 的差异表达结果。

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | index | unique_seq_ID | Length | precursor coordinate | | consensus precursor sequence | | miRDeep2 score | |
| 2 | | | | 2bunT | Fny2B | 2bunT | Fny2B | 2bunT | Fny2B |
| 3 | N-miRNA_1 | aaaaagagggauggaacacgagga | 24 | chr3:13836 | # | aaaaagagggauggaacacg | # | 0.5 | # |
| 4 | N-miRNA_2 | aaaaggauuuuaaaagcuauuca | 23 | # | chr4:5669972..5670 | # | aaaaggauuuuaaaag | # | 1.6 |
| 5 | N-miRNA_3 | aaaauuagaaguuggagcg | 19 | # | chr3:12898577..128 | # | cucugucuucuuccuu | # | 0.4 |
| 6 | N-miRNA_4 | aaaauucuuaaaaacucuugua | 22 | # | chr5:10531604..105 | # | aaagacuuuauaagag | # | 1.9 |
| 7 | N-miRNA_5 | aaacgacgucguuuuggaagcucc | 24 | chr2:11267 | # | aaacgacgucguuuuggaag | # | 1.5 | # |
| 8 | N-miRNA_6 | aaacgauucugagaaaaacugau | 23 | chr5:46416 | # | aaacgauucugagaaaaacu | # | 0.1 | # |

| | J | K | L | M | N | O | P | Q | R | S | T | U | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Example miRBase miRNA with the same seed | | provisional id(pdf_name) | | read_info | | DEGseq Result | | | | | | |
| | 2bunT | Fny2B | 2bunT | Fny2B | 2bunT | Fny2B | log2(F | log2(F | old z-score | p-value | q-value(E | q-value(S | Signatur |
| | gma-miR1516a-3p | # | chr3_26215 | # | 5 | 0 | 3.32193 | 3.582194 | 2.21237 | 0.026941 | 0.029255 | 0.00026 | FALSE |
| | # | osa-miR5080 | # | chr4_36055 | 0 | 7 | -3.8074 | -3.54709 | -2.383 | 0.017174 | 0.019548 | 0.00017 | FALSE |
| | # | dps-miR-2538-3p | # | chr3_22983 | 0 | 9 | -4.1699 | -3.90966 | -2.7969 | 0.00516 | 0.00738 | 6.34E-05 | FALSE |
| | # | dme-miR-4983-3p | # | chr5_44364 | 0 | 10 | -4.3219 | -4.06166 | -2.9832 | 0.002853 | 0.00459 | 3.93E-05 | FALSE |
| | - | # | chr2_14132 | # | 23 | 0 | 5.52356 | 5.783828 | 5.23176 | 1.68E-07 | 9.36E-07 | 8.22E-09 | TRUE |
| | sme-miR-7b-3p | # | chr5_35198 | # | 16 | 0 | 5 | 5.260266 | 4.34538 | 1.39E-05 | 5.59E-05 | 4.97E-07 | TRUE |

## 6、Rfam 比对

为了分析测序结果中各类 RNA 的归类及其比例，从公共数据库网站下载 XXXX 的 sRNA 数据及注释信息：（ftp://ftp.sanger.ac.uk/pub/databases/Rfam/11.0/，文件名称：Rfam.fasta.gz）

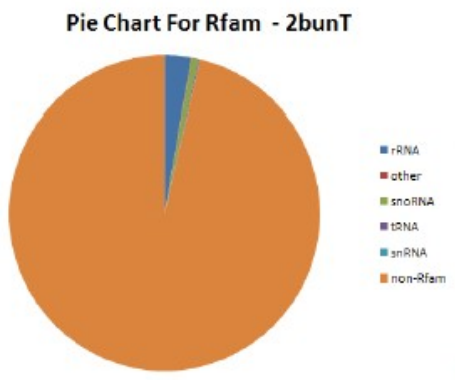
比对方法：使用 bowtie 软件（version 0.12.7）比对 Rfam 11.0 版本中 XXXX 的 sRNA 序列，再统计各 sRNA 的 read 数量。

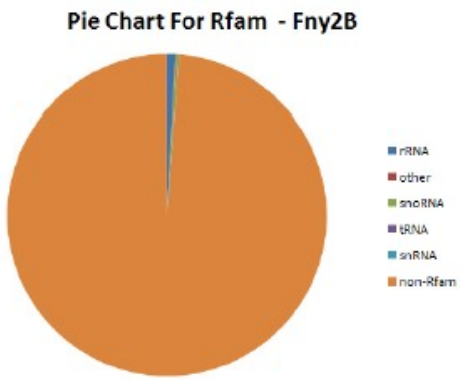
参考文件：\data\reference\Arabidopsis_Rfam_11_0 目录下文件

结果文件：\data\Rfam

各个样本中 Rfam 中比对的 RNA 类型数量如下：

| RNA type | 2bunT | Fny2B |
|---|---|---|
| rRNA | 226584 | 50617 |
| other | 6440 | 4168 |
| snoRNA | 59068 | 17704 |
| tRNA | 10299 | 1030 |
| snRNA | 2565 | 516 |
| non-Rfam | 8000904 | 6051450 |

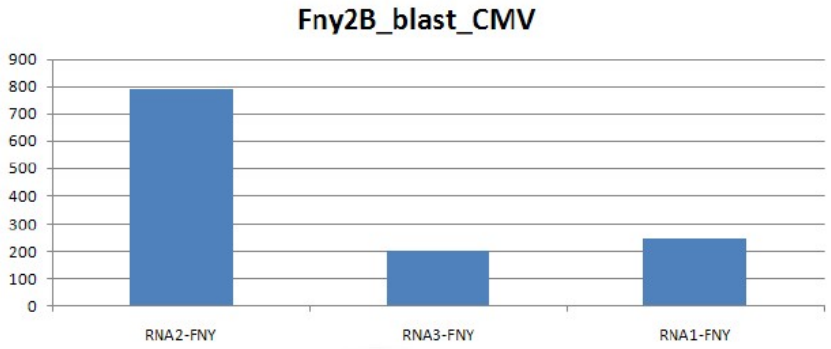


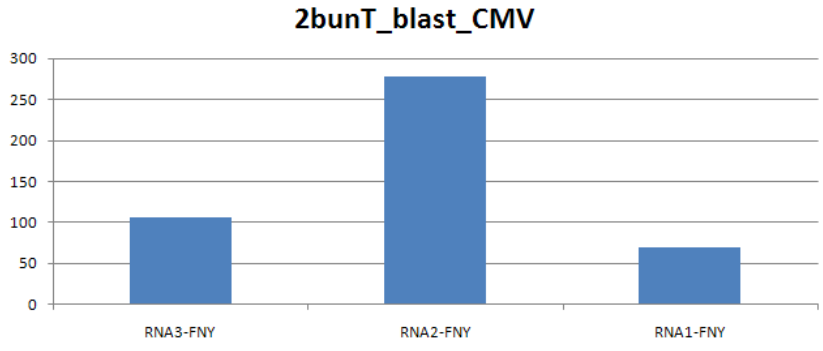



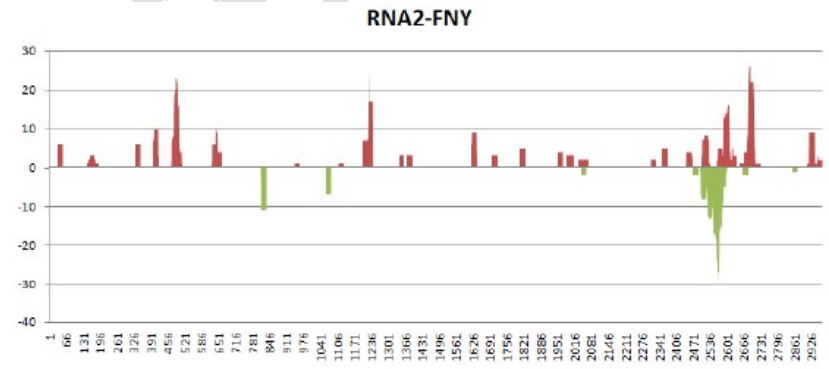
## 7、统计病毒 siRNA 表达信息

先使用 bowtie 软件过滤掉属于 XXXX 的序列，再使用 miRDeep 中的 mapper 子程序和病毒序列进行比对。

结果文件：\data\CMV 目录下文件




## 结果文件

\data\raw_fastq目录下是样本原始数据

\data\reference目录下是下载的公共数据库中的序列

XXXX_genome：XXXXgenome序列

XXXX_Rfam_11_0：Rfam 序列

XXXX：病毒序列（客户提供）
\data\summary_data.xlsx是综合数据及主要分析结果摘要文件
\data\know miRNA目录下是已知miRNA分析结果
know_miRNA_miRBase.xlsx：已知miRNA及差异表达分析结果
pdf：包含检测到的miRNA的PDF的文件夹
know_miRNA_boxplot.jpg：检测到miRNA的表达量boxplot图
know_miRNA_venn.png：检测到miRNA 种类的venn图
DEGseq.png：检测到miRNA DEGseq差异图
\data\novel miRNA目录下是novel miRNA分析结果
novel_miRNA.xlsx：novel miRNA及差异表达分析结果
pdf：包含预测的miRNA的PDF的文件夹
novel_miRNA_boxplot.jpg：检测到miRNA的表达量boxplot图
novel_miRNA_venn.png：检测到miRNA 种类的venn图
DEGseq.png：novel miRNA DEGseq差异图
\data\Rfam目录下是Rfam比对分析结果
\data\CMV目录下是病毒XXXX分析结果

# 参考文件

**Rfam Database**
ftp://ftp.sanger.ac.uk/pub/databases/Rfam/11.0/
**miRBase**
http://www.mirbase.org/
软件信息
**miRDeep2：**
http://www.mdc-berlin.de/en/research/research_teams/systems_biology_of_gene_regulatory_elements/projects/miRDeep/
**DEGseq：**
http://www.bioconductor.org/packages/release/bioc/html/DEGseq.html
**Bowtie：**
http://bowtie-bio.sourceforge.net/index.shtml
**R：**
http://www.r-project.org/