

绝对定量 miRNA 测序

结题报告

客户名字	
客户单位	
服务项目	
项目编号	
报告出具日期	

您值得信赖的基因组学、转录组学与蛋白质组学技术合作伙伴

本报告涉及联川生物技术股份有限公司商业秘密，请确保本报告内容仅供您个人科研使用，并承担保密责任。若擅自向任何第三人或者公司泄露本报告的内容，根据国家有关法律、法规，联川生物技术股份有限公司将有权追究法律责任



目录

一、	前言	5
二、	项目信息	11
2.1	数据库信息	11
2.2	数据分析程序	11
三、	工作流程	12
3.1	实验流程	12
3.2	信息分析流程	12
四、	结果与分析	13
4.1	样品分析名称与分组信息	13
4.2	测序数据预处理	14
4.2.1	UMI 提取及数据去冗余	14
4.2.2	mRNA 比对	15
4.2.3	Rfam 数据库比对分析	15
4.2.4	Repbase 数据库比对分析	16
4.2.5	候选 RNA 长度分布	16
4.3	miRNA 鉴定和预测分析	17
4.3.1	miRNA 鉴定和预测结果	17
4.3.2	miRNA 二级结构预测	19

4.4	miRNA 表达分析.....	20
4.5	miRNA 差异分析.....	20
4.5.1	miRNA 差异分析.....	20
4.5.2	差异 miRNA 上下调统计分析.....	21
4.5.3	差异 miRNA 聚类分析.....	21
4.5.4	差异 miRNA 火山图.....	22
4.5.5	差异 miRNA 维恩图.....	23
4.6	miRNA 表达和差异分析.....	23
4.6.1	差异 miRNA 的靶基因预测.....	23
4.6.2	差异 miRNA 靶基因 GO 富集性分析.....	24
4.6.3	差异 miRNA 靶基因 GO 富集性分析.....	27
4.7	miRNA 成簇分析.....	30
4.8	miRNA 保守性分析.....	30
4.9	miRNA 家族分析.....	31
4.10	miRNA 种子序列分析.....	31
五、	数据挖掘.....	31
5.1	siRNA 分析.....	31
5.2	piRNA 分析.....	31
5.3	miRNA-转录组-降解组联合分析.....	32

5.4	miRNA-转录组-蛋白联合分析.....	33
5.5	ceRNA 网络调控分析.....	35
六、	辅助材料.....	36
6.1	实验方法（英文材料）	36
6.2	生信分析（英文材料）	36
七、	参考文献.....	39
八、	联系我们.....	42

一、前言

MicroRNAs(miRNAs)是在真核生物中发现的一类内源性的具有调控功能的非编码 RNA，其大小长约 18~25 个核苷酸。成熟的 miRNAs 是由较长的初级转录物经过一系列核酸酶的剪切加工而产生的，随后组装进 RNA 诱导的沉默复合体，通过碱基互补配对的方式识别靶 mRNA，并根据互补程度的不同指导沉默复合体降解靶 mRNA 或者阻遏靶 mRNA 的翻译[1]。最近的研究表明 miRNA 参与各种各样的调节途径，包括发育、病毒防御、造血过程、器官形成、细胞增殖和凋亡、脂肪代谢等。

1.1 动植物体内 miRNA 形成过程

在动物体内，miRNA 基因是由细胞核内的 RNA 聚合酶 II (pol II)转录的，首先形成的是 miRNA 的初级体 (pri-miRNA)，pri-miRNA 在核酸酶 Drosha 的作用下，形成了 miRNA 的前体 (pre-miRNA)，pre-miRNA 在特异性蛋白的作用下，从细胞核进入细胞质中，另一个核酸酶 Dicer 将其剪切产生约为 22 个核苷酸长度的 miRNA:miRNA*双链复合体如图 1b 所示[2]。Drosha 和 Dicer 都是 RNaseIII 家族成员，即双链 RNA 特异性核酸内切酶。这种双链复合体很快被引导进入沉默复合体(RISC) 复合体中，其中一条成熟的单链 miRNA 保留在这一复合体中。成熟的 miRNA 结合到与其互补的 mRNA 的位点通过碱基配对调控基因表达。

在植物体内，大多数植物 miRNAs 前体(pri-miRNA)产生于其自身的转录单元，都是由 RNA polymerase II(Pol II)转录而成[3]。并经过 RNA 绑定蛋白 DAWDLE (DDL)绑定 miRNA 进入核加工中心 D-bodies，在 D-bodies 中 DCL1 酶和双链 RNA 结合蛋白 HYPONASTIC LEAVES1 (HYL1)组装成类似于小 RNA 加工复合体的结构，并在 C2H2 锌指结构蛋白 SERRATE (SE) 以及 CBC 的辅助下将 pri-miRNA 切割成 miRNA:miRNA*。接着 miRNA:miRNA*双链体中的 2' 羟基或 3-羟基在 HEN1(Hua Enhancer 1)作用下发生甲基化，这些过程是在细胞核中完成。接着，miRNA: miRNA*在 HST (Hasty)和其他因子的帮助下被运送到细胞质中。其中一个 miRNA*被小 RNA 降解酶(SDN)降解，另外一个成熟的甲基化

miRNA 与 AGO 蛋白(Argonaute)形成 RNA 诱导的沉默复合体(RISC), RISC 中的 miRNA 能切割互补配对的靶 mRNA 或者抑制它的翻译。从而起到转录后负调节的作用, 如图 1a 所示[4]。

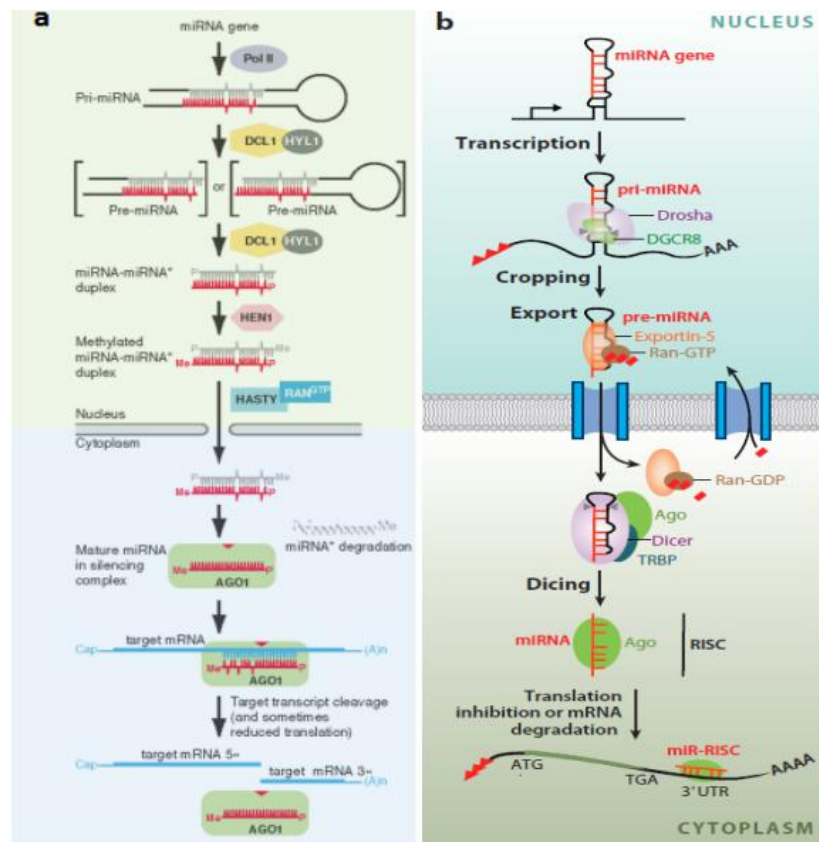


图 1 动物和植物体内 miRNA 的形成原理

1.2 动植物体内 miRNA 作用机制

在动物体内, microRNA 通过与 mRNA 的 3' 端的非翻译区 (Untranslated Regions, UTR) 相结合, 在不减少 mRNA 数量的情况下达到抑制 mRNA 翻译蛋白的功能。一般成熟的 microRNA 的 seed 种子区域与 AGO 蛋白形成复合物后结合在 mRNA 的 3' UTR 区域如图 1, 种子区域指 miRNA 的 5' 端到 3' 端的 2-8 位核苷酸。图中成熟体 microRNA 用天蓝色和红色表示, 2~8 位的 seed 种子区域用灰色表示, mRNA 序列用深绿色和淡绿色表示。当这些成熟体 microRNA 和 AGO 蛋白形成复合物时, 可在 3' 端的 UTR 区域 (图 1 中用黄色表示) 抑制 mRNA 的翻译[5]。

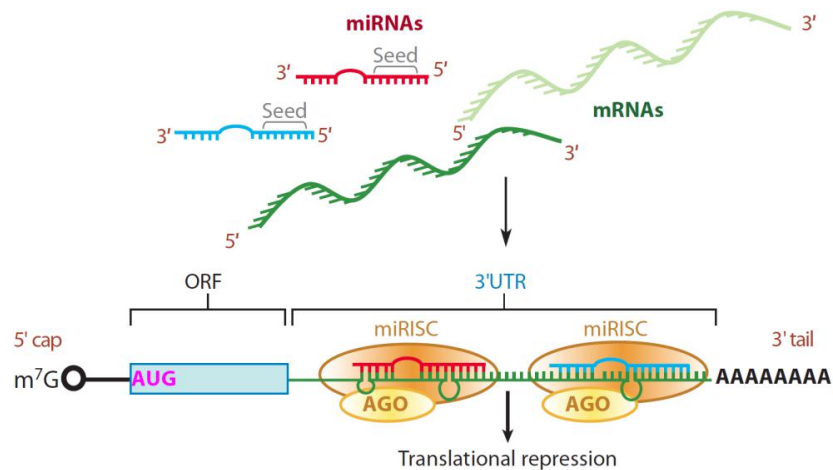


图 2 动物体内 miRNA 的作用机理

在植物体内，microRNA 会对 mRNA 进行剪切，mRNA 被降解后失去编码蛋白的功能。具体来说，microRNA 成熟体与 AGO 蛋白形成复合物，与靶基因相结合，并且在 microRNA 最中间的结合位点对靶基因进行剪切如图 2，该中间位置即 miRNA 的 5' 端到 3' 端的第 10-11 位核苷酸。当 mRNA 被降解成片段后，其编码蛋白的功能亦会消失[6]。

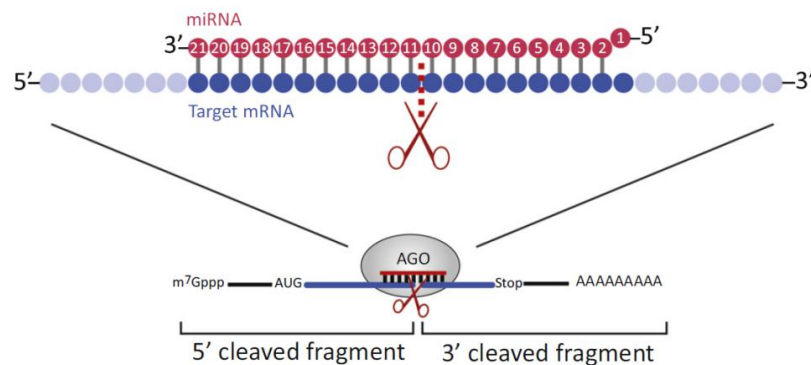


图 3 植物体内 miRNA 的作用机理

miRNA 测序是 miRNA 研究常用的技术手段，常规 miRNA 测序整体流程如下：



图 4：常规 miRNA 测序流程

由于文库构建过程包含 PCR 扩增，会导致出现 PCR Duplications (由于 PCR 扩增的偏好性以及

扩增倍数的不确定性，导致 NGS 测序文库构建时不同片段被扩增的倍数出现偏差，这种因 PCR 扩增产生的重复 reads 即为 PCR Duplications)，PCR Duplications 会对表达定量造成干扰，使得分析结果出现偏差。

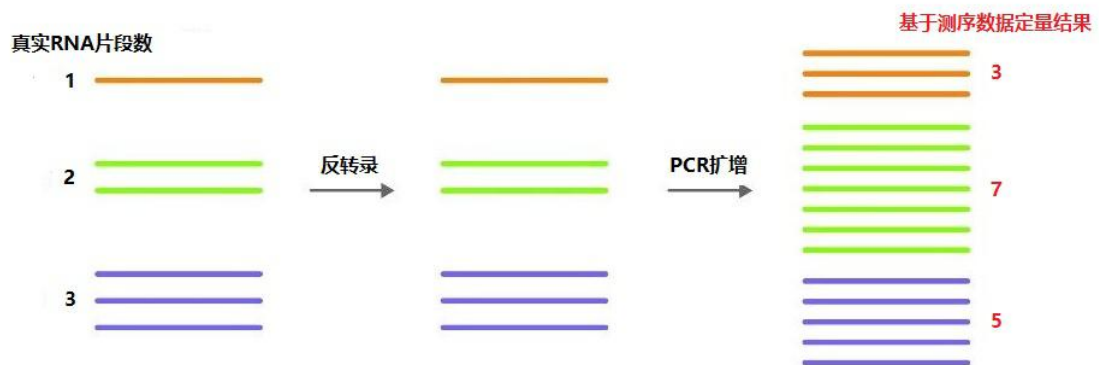


图 5：PCR Duplications 产生过程及影响

为了避免 miRNA 测序文库构建过程中 PCR Duplications 对 miRNA 定量结果造成的干扰，需要去除 PCR Duplication 操作，通常去 PCR Duplications 做法是 mapping coordinates 方法实现，即比对到相同基因组位点的 reads 被认为是 PCR Duplications。**mapping coordinates 法去除 PCR Duplications 的理论基础是长片段 RNA 打断时出现完全相同打断片段的概率很低，序列完全相同的测序数据主要来自于 PCR Duplications。**

由于 miRNA 的几个特征，：

1. miRNA 序列比较短，miRNA 文库构建不需要打断；
2. 基因组上相同和不同位点的基因会转录出完全相同的成熟体 miRNA 序列，miRNA 测序数据中，相同序列即来自同一条 miRNA 的不同拷贝，也来自于 PCR Duplications；
3. 某些细胞中，一种 miRNA 的拷贝数会占到所有 miRNA 拷贝数的 40%[7]，文库构建和测序时 PCR 过程会导致 PCR Duplications 放大。

这些特征使得 miRNA 文库构建时 PCR Duplications 比例比构建长片段 RNA 文库（例如转录组文库、lncRNA 文库等）的 PCR Duplications 更加严重，同时常规 mapping coordinates 法去

除 miRNA 测序 PCR Duplications 时，会过多的去除阳性数据，导致 miRNA 表达水平人为降低。

因此，需要采用绝对定量 miRNA 保证数据真实度，即 UMI 标记技术构建 miRNA 文库（下图）：

序列在反转录后，加入大量的标签（barcode），几乎每个 cDNA 都被唯一的 barcode 标记，然后进行 PCR 扩增获得测序文库（见图 6）。由于序列是被 barcode 唯一标记的，计算序列拷贝数时，不再直接统计同一种 reads 的数量，而是统计每种 reads 有多少个 unique 的 barcode。而具有相同 barcode 的同一种 reads，无论有多少拷贝数，都只计作一个拷贝，即来自 PCR 扩增的假重复（Duplication）被有效去除。

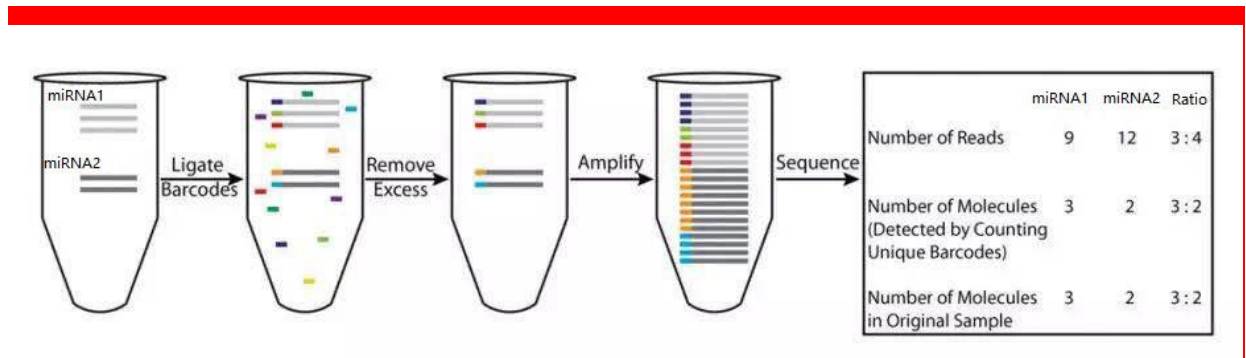


图 6：绝对定量 miRNA 测序去 PCR Duplications

在构建 miRNA UMI 文库时，需要保证每条转录本（miRNA）都能够被唯一的 UMI 所标记，因为当同一转录本被相同的 UMI 标记，表达计算时会错误的被认为是 PCR Duplication，因此 UMI 的数量都要远超过表达丰度最高 miRNA 的拷贝数。上文提到在某些细胞中，丰度最高的 miRNA 表达量（read 数）有时会超过所测数据的 40%，达到几千万 read，因此需要的 UMI 数量是巨大的。为了得到足够的 UMI 种类，UMI 的长度需要更长，例如当长度为 9bp 时，UMI 的数量可达到 $4^9 \times 4^9$ ，数量远超最高表达 miRNA 表达量。如果 UMI 全部由随机碱基构成，连续的 9 个随机碱基在文库扩增时容易和扩展引物随机结合，导致最后的测序数据中出现长度不均一的、长度较短的 read，浪费数据量。同时 miRNA 鉴定的通用标准是分析序列的相似性，而成熟体 miRNA 长度较短，若引入的 UMI 序列出现插入或者缺失的情况，会对 miRNA 鉴定及定量造成干扰。为了避免这些情况出现，可以使用序列固定的 UMI locator 将 UMI 分成不同的区段 [8]，例如 5'-NNN-CGA-NNN--TAC-NNN-3'（图 7）。考虑到测序时需要复杂文库，UMI locator 被设计

成两套 (图 7), 增加文库复杂度, 避免测序数据质量下降。UMI locator 设计和传统 UMI 相比, 一方面通过特定序列将连续的随机碱基分开, 避免出现引物非特异性结合; 另一个方面是可以区分 UMI 是否出现插入缺失的情况, 从而更精准的对 miRNA 鉴定和定量。



图7: 绝对定量miRNA UMI文库结构

二、 项目信息

物种名：人

物种拉丁名：Homo sapiens

2.1 数据库信息

数据库	链接	版本/日期
miR(miRs)数据库	ftp://mirbase.org/pub/mirbase/CURRENT/	22.0
Pre-miRNA(mirs/MIRs)数据库	ftp://mirbase.org/pub/mirbase/CURRENT/	22.0
物种优先级		
RFam	http://rfam.janelia.org	12.0
Repbase	http://www.girinst.org/repbase	22.07
Genome 数据库		
mRNA 数据库		
客户数据库	NA	NA
KEGG pathway 数据库	http://www.genome.jp/kegg	2016.05
Gene Ontology 数据库	ftp://ftp.ncbi.nih.gov/gene/DATA/gene2go.gz	2016.04
KOG 数据库		NA

2.2 数据分析程序

分析项目	名称与软件下载链接	版本/日期
非 miRNA 测序数据&低质量数据去除、长度过滤	FASTP	
UMI 提取和去冗余	umitools	
数据质控	FastQC	0.10.1
miRNA(miRs)基础分析	ACGT101-miR	4.2
miRNA 进阶分析	ACGTUNAFold	3.7
miRNA 靶基因预测	Target Finder(植物)	50
	Targetscan(动物)	3.3a
	miranda(动物)	
靶基因功能分析	Pathway Network	1.6
结果文件绘图	R, cytocape	3.0.1

三、 工作流程

3.1 实验流程

绝对定量 miRNA 实验流程与常规 miRNA 实验流程一致，均包括文库制备和测序。两种文库构建方式一致，流程均包括：1、3' 接头连接；2、5' 接头连接；3、反转录；4、文库扩增。绝对定量 miRNA 文库构建和常规 miRNA 测序文库构建不同点是，绝对定量 miRNA 的 3' 接头和 5' 接头均连接一段 15bp 的 UMI 序列。因此和常规 miRNA 文库相比，绝对定量 miRNA 文库插入片段两侧各多了 15bp，因此测序策略不再是 SE50，而是 PE150 模式。

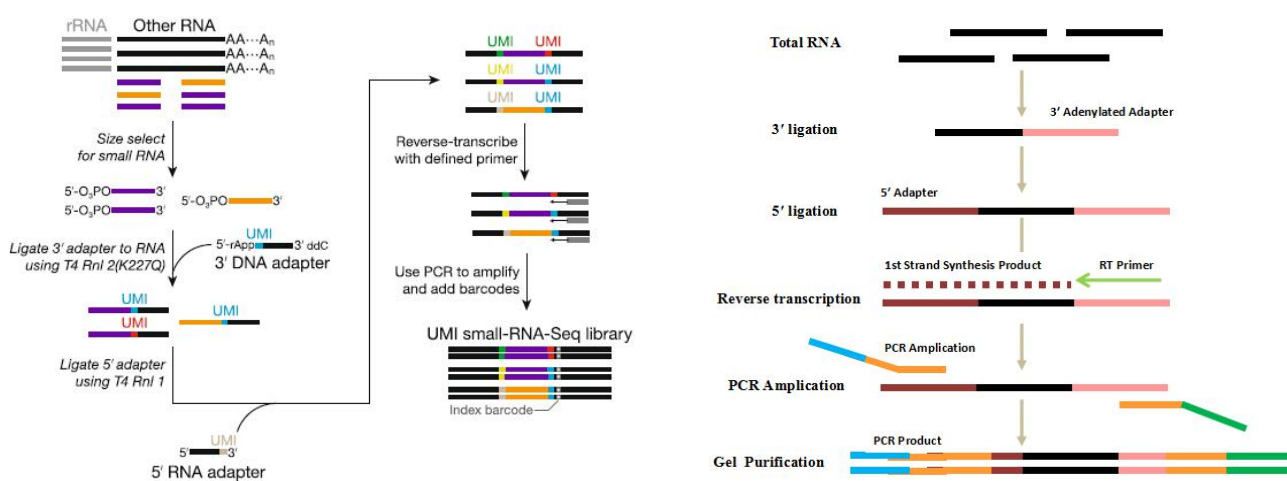


图 4：绝对定量 miRNA（左）和常规 miRNA（右）建库流程

3.2 信息分析流程

150bp 测序数据下机以后，首先使用 FASTP 软件去除 3' 接头以及低质量数据，保留长度 40-70bp 数据，然后再用 umitools 提取保留数据中 UMI 序列并去冗余（去 PCR Duplications：具有相同 5' 和 3' UMI 的数据为 PCR Duplications 数据，选择其中一条序列用于分析，其他去除），最后采用联川生物自主开发的 miRNA 数据分析软件 ACGT101-miR (LC Sciences, Houston, Texas, USA)，该软件的分析流程如下：（1）长度筛选：植物保留碱基长度在 18-25nt，动物保留碱基长度在 18-26nt 的序列；（2）各种 RNA 数据库比对分析：将剩余序列比对（不包含 miRNA）mRNA、RFam 和 Repbase 数据库，并进行过滤；（3）miRNA 鉴定：获取有效数据，并比对前体和基因组进行 miRNA 鉴定；（4）miRNA 差异性分析，

(5) 差异性 miRNA 靶基因预测分析。此外，联川生物提供的数据挖掘包括个性化分析如 siRNA, piRNA, Ta-siRNA 分析，并单独出具报告。高级数据分析主要指组学相关的一些联合分析，如 miRNA-转录组-降解组联合分析，miRNA-转录组-蛋白联合分析，ceRNA 调控网络等。

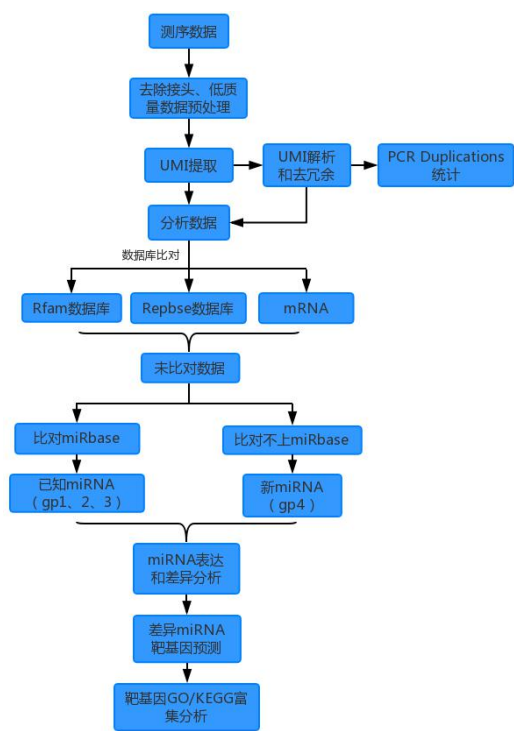


图 5：数据分析流程图

四、 结果与分析

4.1 样品分析名称与分组信息

Sample Name	Sample Species	Biological Replicates
CK_1	Homo sapiens	Control
CK_2	Homo sapiens	Control
CK_3	Homo sapiens	Control
A_1	Homo sapiens	Treat
A_2	Homo sapiens	Treat11
A_3	Homo sapiens	Treat1
B_1	Homo sapiens	Treat2
B_2	Homo sapiens	Treat2
B_3	Homo sapiens	Treat2

文件位置：

summary/1_SequencingData_Overview/Table1_Sample_info.xlsx

4.2 测序数据预处理

原始测序数据首先被去除 3' 接头序列以及低质量数据，保留长度 40-70bp 数据，然后提取 UMI 序列去除 PCR Duplications 并统计 PCR Duplications 比率，PCR Duplications 后剩下的数据（不含有双端 UMI）去除测序碱基长度小于 18nt 的序列，如果序列中有 80% A 或 C 或 G 或 T，3N（不一定是连续的），只有 A、C 没有 G、T，或只有 G、T 没有 A、C，或者是连续的核苷酸二聚体，三聚体，那么这些序列会被过滤掉。然后再与 mRNA（部分物种可能不具备）、RFam（包含 rRNA, tRNA, snRNA, snoRNA 等）以及 Repbase 数据库进行比对过滤，并将过滤后的数据称之为 Valid 数据，并对 Valid 数据进行进一步的 miRNA 比对鉴定与预测分析。

文件位置：

summary/1_SequencingData_Overview/Table2_Overview_of_reads_from_raw_data_to_cleaned_sequence

4.2.1 UMI 提取及数据去冗余

绝对定量 miRNA5' 和 3' UMI 各有一对由 3 个固定碱基构成的 UMI locator，提取双端各 15bp UMI 序列后，结合 UMI locator 对 UMI 序列识别并聚类，依照 UMI 聚类结果去除 PCR Duplications。

图 6：UMI 表及数据比例柱状图（每个样本一个柱子）

样本	测序数据	UMI 标记数据	UMI 标记比率	冗余 UMI	PCR Duplications 率
CK_1					
CK_2					
CK_3					
A_1					
A_2					

4.2.2 mRNA 比对

测序数据通过与本物种 mRNA 比对，去除测得的数据中包含来自 mRNA 降解片段的短序列。

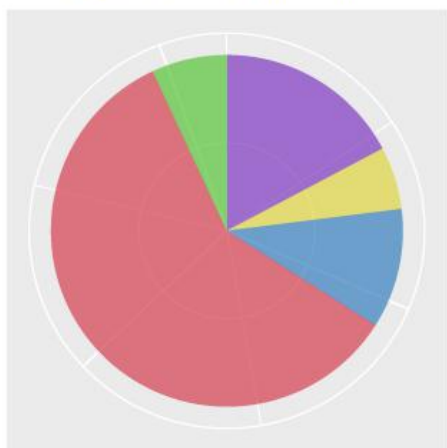
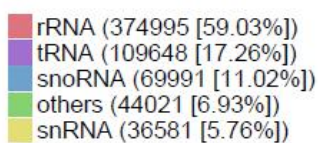
文件位置：

summary/1_SequencingData_Overview/Table2_Overview_of_reads_from_raw_data_to_cleaned_sequence

4.2.3 Rfam 数据库比对分析

Rfam 是非编码 RNA (ncRNA) 家族数据库，包括 rRNA, tRNA, snoRNA, snRNA, miRNA 等其他一些非编码 RNA，我们选取 Rfam 数据库来注释测序得到的小 RNA 序列，尽可能地发现并去除其中可能的 rRNA, snoRNA, snRNA, tRNA 等非 miRNA 序列。

Pie chart for Rfam DSS_1 (Total)



Pie chart for Rfam DSS_1 (Unique)

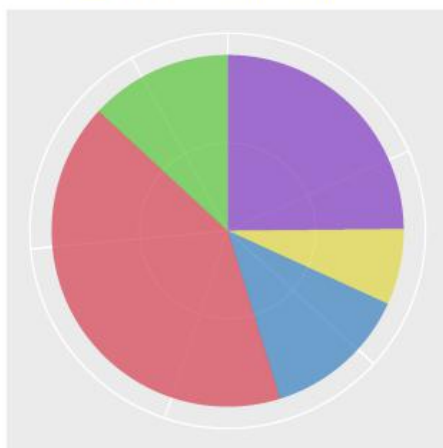
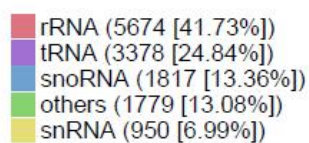


图 7: Rfam 比对饼状图

文件位置：

summary/1_SequencingData_Overview/Table2_Overview_of_reads_from_raw_data_to_cleaned_sequence

summary/1_SequencingData_Overview/ Fig1_Rfam_sequence_category

4.2.4 Repbase 数据库比对分析

Repbase 数据库是真核生物 DNA 中重复序列数据库，包括转座因子，串联重复序列（卫星序列或微型卫星），零星的基因组重复序列和一些多拷贝宿主基因（如 rRNA，tRNA，组蛋白基因等），选取 Repbase 数据库来注释测序得到的小 RNA 序列，尽可能地发现并去除其中可能的 repeat associate sRNA。

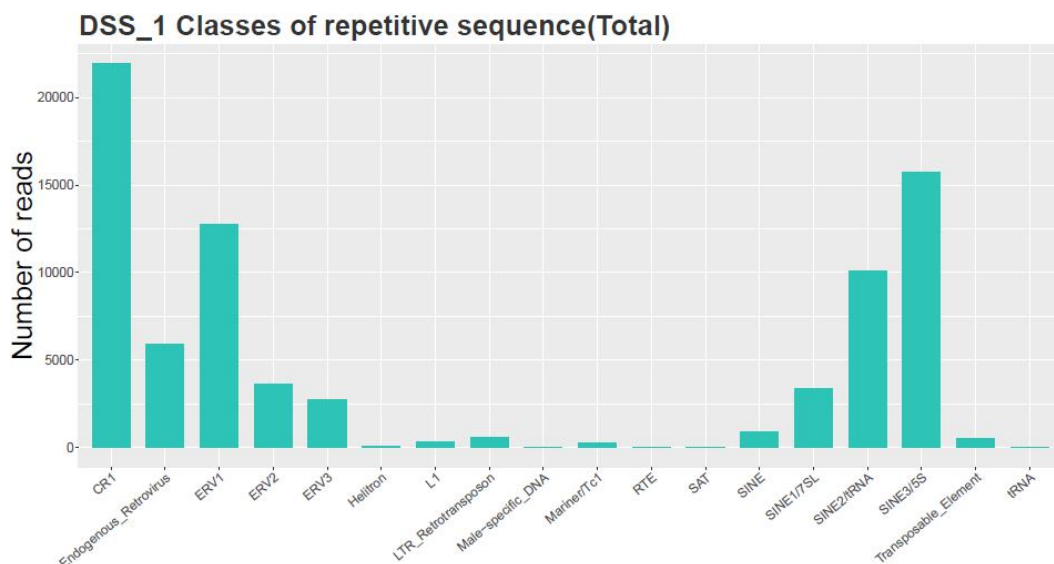


图 8: Repbase 比对饼状图

文件位置:

summary/1_SequencingData_Overview/Table2_Overview_of_reads_from_raw_data_to_cleaned_sequence

summary/1_SequencingData_Overview/Table3

summary/1_SequencingData_Overview/ Fig2_Repeat_sequence_category

4.2.5 候选 RNA 长度分布

在对测序原始数据进行分析与统计的基础上，对过滤后的 Valid 数据进行了长度分布统计。如果大部分数据分布在 20-24nt，则符合 Dicer 酶切割的典型特征。

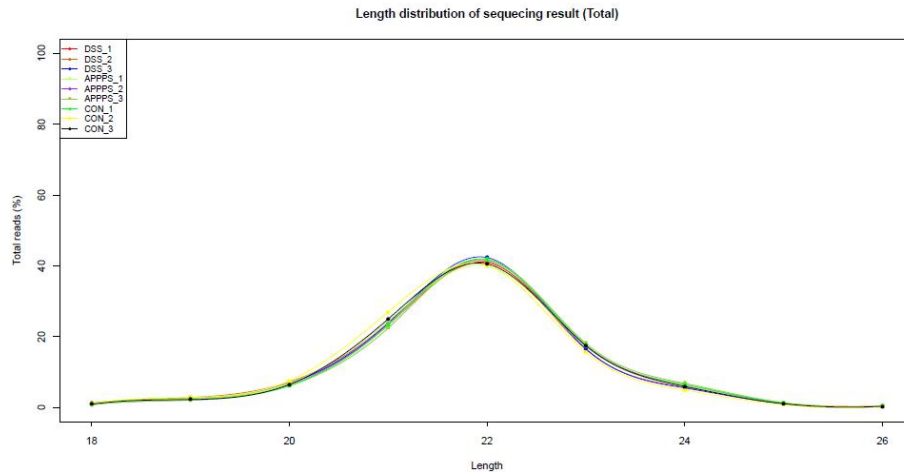


图 9：测序长度分布

文件位置：

summary/1_SequencingData_Overview/Table4

summary/1_SequencingData_Overview/Fig3_Length_distribution_of_counts_of_total_and_unique_sRNAs_in_this_study

4.3 miRNA 鉴定和预测分析

4.3.1 miRNA 鉴定和预测结果

miRNA 序列比较短，在物种之间具有较高的保守性，因此数据分析时，valid data 除了与 miRbase 数据库报道的本物种 miRNA 前体比对，还要和近源物种 miRNA 前体比对，同时 valid data 和能够被测序数据比对上的前体序列，均要再和本物种基因组比对，依照测序数据与前体、基因组，以及前体和基因组比对情况，将鉴定到的 miRNA 分成 gp1、2、3、4 四个组别。各组的定义是有点复杂，但是相对来理解：gp1a 是测序物种本身已报道 的，known miRNAs；gp1b,2a,2b,3 是保守的，conservative miRNAs，gp2 有基因组位置支撑，gp3 没有基因组位置支撑，gp2a 有前体基因支撑，gp2b 没有前体基因支撑；gp4 是新的，novel miRNAs；部分模式物种我们最后的结果中甚至不保留 gp2b 和 gp3 的数据，这个时候只能看到代表 gp2a 的 gp2 的 miRNA。

具体 miRNA 的分组信息如下所示：

分组编号	意义与解释
gp1a	Reads 可以比对到 miRBase 已知的首选物种前体 (pre-miRNAs)，并且前体可以进一步比对到基因组
gp1b	Reads 可以比对到 miRBase 已知的选择物种的前体 (pre-miRNAs)，并且前体可以进一步比对到基因组
gp2a	Reads 可以比对到 miRBase 选择物种的前体 (pre-miRNAs)，前体不能进一步比对到该物种的基因组，但是 reads 可以比对到基因组上。延伸的基因组序列可以形成满足 11 条原则发夹结构
gp2b	Reads 可以比对到 miRBase 选择物种的前体 (pre-miRNAs)，前体不能进一步比对到该物种的基因组，但是 reads 可以比对到基因组上。延伸的基因组序列不可以形成满足 11 条原则发夹结构
gp3a	Reads 可以比对到 miRBase 选择物种的前体 (pre-miRNAs)，前体不能进一步比对到基因组，reads 也不可以比对到基因组上。cluster 中比对上的 Reads 数目拷贝数>1,代表性的 Reads 分值>260,没
gp4a	Reads 不能比对到 miRBase 选择物种的前体 (pre-miRNAs)，reads 可以比对到基因组上。延伸的基因组序列可以形成满足 11 条原则发夹结构
gp1a	Reads 可以比对到 miRBase 已知的首选物种前体 (pre-miRNAs)，并且前体可以进一步比对到基因组
gp1b	Reads 可以比对到 miRBase 已知的选择物种的前体 (pre-miRNAs)，并且前体可以进一步比对到基因组

由于测序实验数据分析检出的 miRNA 主要是以异构体的群体表达形式存在 (这是因为在某些特定阶段的特定时期，很多种类的 miRNA 表达后所呈现的序列存在方式并不一定与 miRBase 数据库中已报道的该种 miRNA 序列完全一致)，因此在鉴定道德各组 miRNA 中，还有 Known (和 miRBase 报道的成熟体 miRNA 一致)、Diff (和 miRBase 报道的成熟体 miRNA 有碱基的差别)、New (新鉴定到的 miRNA)。

字符含义	意义与解释
L+1/R+1	与 miRBase 报道的成熟体 miRNA 相比，测序得到的 miRNA 左侧/右侧多个一个碱基
L-1/R-1	与 miRBase 报道的成熟体 miRNA 相比，测序得到的 miRNA 左侧/右侧少个一个碱基
1SS18TC / 2SS20TG21AT	与 miRBase 报道的成熟体 miRNA 相比，测序得到的 miRNA 的第 18 位碱基由 T 替换成 C/与 miRBase 报道的成熟体 miRNA 相比，测序得到的 miRNA 的第 20 位碱基由 T 替换成 G，第 21 位碱基由 A 替换成 T

各个样本 miRNA 鉴定和预测的结果均整理在 Table5 中，Table5 每列的具体含义请参考下列表格：

Term	意义
Pre-miRNA Index	前体 miRNA 序列号，一个 miRNA 前体可以生成多个 miRNA 成熟体
miRNA_Index	成熟体 miRNA 的序列号
miR_name	miRNA 命名(参考辅助文件：3.6.1 microRNA 的命名规则)
miR_seq	miRNA 成熟体序列
len	miRNA 成熟体长度
genomeID	miRNA 成熟体对应的 miRNA 前体序列所在基因组的染色体名称
Strand	miRNA 成熟体对应的 miRNA 前体序列所在染色体的+链或-链
Start	miRNA 前体在染色体上的起始位置
End	miRNA 前体在染色体上的终止位置
pre-miRNA_seq	miRNA 前体的序列信息，其中大写的为经过加工后成熟体序列

#mir/MIR	匹配上的 miRBase 数据库前体 miRNA 序列的数目
mirIDs	匹配上的 miRBase 数据库前体 miRNA 登陆号，涵盖比对到的各个参考物种
rep mirID	匹配上的 miRBase 数据库上的该物种的前体序列登陆号
rep mirSeq	匹配上的 miRBase 数据库该物种前体 miRNA 序列信息，大写字母代表成熟体序列，其中该序列与 pre-miRNA seq 长度有可能不一致
rep miRID	匹配上的 miRBase 数据库上的该物种的成熟体 miRNA 序列登录号
rep miRSeq	匹配上的 miRBase 数据库上的该物种的成熟体 miRNA 序列信息
type	miRNA 成熟体来源于 miRNA 前体序列的 3' 端或 5' 端
Sequence in miRBase	miRNA 成熟体序列信息是否与 miRBase 是否相关。完全相同 yes，不同 Diff，新序列 New
group	miRNA 报道组别，具体定义见 Table5 表头解释
SS	即 miRNA 前体二级结构表示法。方法一：通过本公司二级结构折叠软件的操作手册 5 Secondary folding software 进行；方法二（推荐）：访问 http://nibiru.tbi.univie.ac.at/forna/ 进行在线绘制
hairpinLen	miRNA 前体二级结构长度
CG%	GC 含量
dG	自由能
MFEI	$MFEI = -dG * 100 / \text{mirLen} / \text{CG}\%$. Reference: Cell. Mol. Life Sci. 63 (2006) 246-254. 最小折叠自由能指标，限定植物专用
raw	下机数据原始 counts 数。这些数据属于未归一化处理，老师一般情况下不用特别关注
norm	归一化表达数据，老师需要关注这部分数据。这些归一化数据可以用于后续分析，可以当做表达量
Expression level	表达水平，高、中、低。对于归一化表达数据从低到高排序，low 为所有样本中拷贝数均小于 10 的 miRNA；middle 是拷贝数只要一个样本中高于 10 且所有样本中拷贝数均小于平均值（拷贝数总和/（样本数目×miRNA 总数））的 miRNA；High 为只要在一个样本中拷贝数大于平均值的 miRNA。

文件位置：src/Summary/2_miRNAExpression/Table5_All_Expressed_miRNA.xlsx

4.3.2 miRNA 二级结构预测

如果测序获得的成熟体 miRNA 存在其前体序列，则用 RNAfold 软件 <http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi> 直接作图，绘制 miRNA 的二级结构。如果测序获得的成熟体 miRNA 不存在其前体序列，则通过 reads 在基因组相应的比对位点的碱基数目延伸（动物 80nt，植物 120nt），预测 miRNA 的前体序列，对预测出的前体序列进行作图。二级结构预测参数和详细内容请参考辅助文件：**3.6.2 microRNA 二级结构分析。**

4.4 miRNA 表达分析

Valid data 通过 Bowtie 比对后，依照比对结果，采用以下归一化方法确定 miRNA 表达谱信息：

1. 找到一个所有样本中的常规序列；
2. 构建一个参考数据系。参考数据系中的每个数据都是所有样本中相关的常规序列的拷贝数的中位数；
3. 对所有的样本和参考数据系对拷贝数做以 2 为底的对数运算转换 $\log_2(\text{copy\#})$ ；
4. 计算各自样本和参考数据系之间 $\log_2(\text{copy\#})$ 的差异 $\Delta\log_2(\text{copy\#})$ ；
5. 选出 $|\Delta\log_2(\text{copy\#})| < 2$ 的序列，也即差异倍数小于 $(2^2 =) 4$ 的参考系；
6. 进行样本之间和子集序列的参考系线性回归分析导出线性公式 $y = a_i x + b_i$ 这里 a_i 和 b_i 是斜率和截距，对于得到的直线， x 就是参考系的 $\log_2(\text{copy\#})$ ， y 就是样品 i 相关序列的期望的 $\log_2(\text{copy\#})$ ；
7. 计算参考系的中值 $x_{\text{mid}} = (\max(x) - \min(x)) / 2$ 。计算相关期望样本 i 的 $\log_2(\text{copy\#})$ ： $y_{i,\text{mid}} = a_i x_{\text{mid}} + b_i$ ，使得 $y_{r,\text{mid}} = x_{\text{mid}}$ ，使得 $\Delta y_i = y_{r,\text{mid}} - y_{i,\text{mid}}$ ，也就是样品 i 的对数校正因数。我们再得到样品 i 的校正因数算法 $f_i = 2^{\Delta y_i}$ ；
8. 通过将原始拷贝数乘以算法校正因数 f_i 来校正各个样本中的拷贝数。

备注 1：相应的英文描述请参考:miRNA 测序表达谱归一化方法.pdf。

备注 2：引用文章请参考：miRNA 测序表达谱数据归一化方法参考文献.pdf。

文件位置：src/Summary/2_miRNAExpression/Table5_All_Expressed_miRNA.xlsx

4.5 miRNA 差异分析

4.5.1 miRNA 差异分析

根据实验设计不同，采用适合的差异检验方法筛选差异显著性基因，例如 Fisher 精确检验、卡方 (2×2)、卡方 ($N \times N$)、T 检验以及 ANOVA 等算法。对于有生物学重复的样本分析，对差异表达基因的筛选以 $P < 0.05$ 为阈值；而对无生物学重复的样本分析，对差异表达基因的筛选以 $|\log_2(\text{Foldchange})| > 1$ （即两倍差异倍数）且 $p_value \leq 0.05$ 为阈值。

各个样本 miRNA 鉴定和预测的结果均整理在 Table8 中，Table8 每列的具体含义请参考下列表格：

Term	意义
Index	序列号
miR_name	miRNA 成熟体的 ID 编号
miR_seq	miRNA 成熟体序列信息
up/down	比较组上下调关系
fold_change	比较组差异倍数，生物学重复样本 norm 值取平均值计算得到
log2(fold_change)	log2 为底的差异倍数
pvalue(t_test)	差异比较组之间通过 t 检验得到的 P 值
Norm	归一化数据
Raw	原始数据

文件位置：src/summary/3_DifferentiallymiRNAs

4.5.2 差异 miRNA 上下调统计分析

差异基因上下调频数统计用于判断不同实验条件下差异表达 miRNA 的个数。其中横坐标表示比较组信息，纵坐标表示上下调 miRNA 的数目，红色代表上调 miRNA，蓝色代表下调 miRNA，其中数字代表上下调 miRNA 的数目。

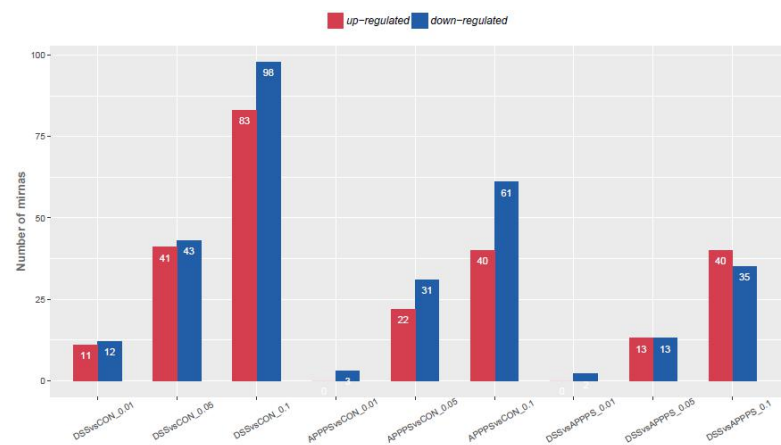


图 10：差异 miRNA 数量柱状图

文件位置：

src/summary/3_DifferentiallymiRNAs/Fig5_barplot_of_differentially_expressed_miRNA.txt..pdf

4.5.3 差异 miRNA 聚类分析

差异 miRNA 聚类分析用于判断 miRNA 在不同实验条件下调控的聚类模式。根据样品 miRNA 表达谱

的相近程度，将 miRNA 进行聚类分析，直观地展示 miRNA 在不同样品（或是不同处理）中的表达情况。

采用 $\log_{10}(\text{norm 值})$ 进行 miRNA 表达展示。同时也可以将差异 miRNA 的 norm 值通过 Z 值方式进行 miRNA 表达展示。其中横坐标为样本，纵坐标为 miRNA，不同的颜色表示不同的 miRNA 表达水平，颜色由蓝色经由白色至红色表示表达量 ($\log_{10}(\text{norm 值})$) 从低到高。红色表示高表达基因，深蓝色表示低表达基因。DIY 热图绘制教程请下载联川生物实用技能手册之 Heatmap 软件 TMEV 使用手册 V2.0.pdf。

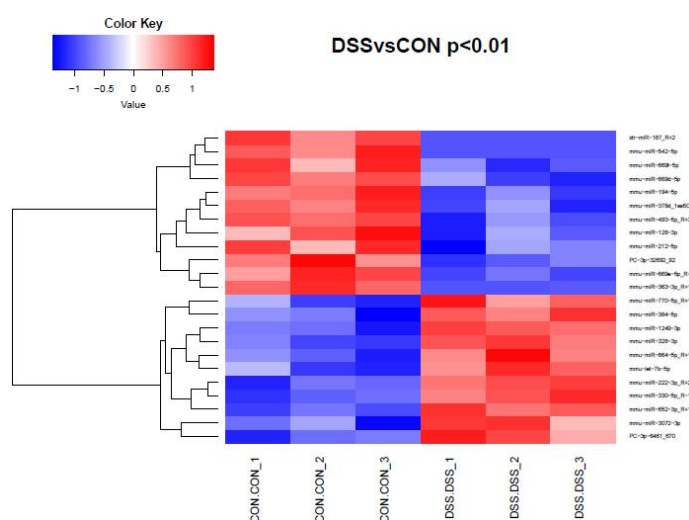


图 11：差异 miRNA 热图

文件位置：

src/summary/3_DifferentiallymiRNAs/Fig6.1_heatmap_of_differentially_expressed_miRNAs.pdf

4.5.4 差异 miRNA 火山图

通过绘制火山图可以了解差异表达 miRNA 的整体分布情况。以 $\log_2(\text{foldchange})$ 为横坐标， $-\log_{10}(\text{pvalue})$ 为纵坐标，对差异表达分析中所有的 miRNA 绘制火山图。其中横坐标代表 miRNA 在不同样本中差异表达倍数变化；纵坐标代表 miRNA 表达量变化差异的统计学显著性；红色代表上调的显著差异表达基因，蓝色代表下调的显著差异表达基因，灰色的点代表非显著性差异表达基因。

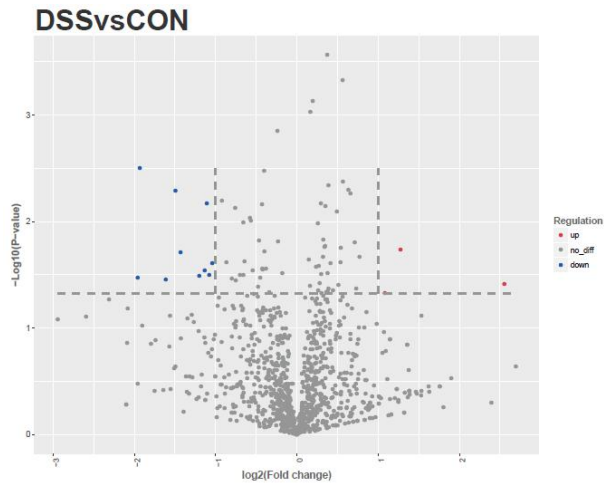


图 12: 差异 miRNA 火山图

文件位置:ummary/3_DifferentiallymiRNAs/Fig6.2_volcano_of_differentially_expressed_miRNAs.pdf

4.5.5 差异 miRNA 维恩图

通过不同比较组之间差异 miRNA 韦恩图, 可以直观地显示出不同比较组之间共同的和特有的差异表达 miRNA 的个数。目前我们的程序最多支持五个组别的基因共表达分析。本报告以 NA 作为范例韦恩图。韦恩图 DIY 教程请下载联川生物实用技能手册之韦恩图绘制.pdf。

图 13: 不同比较组差异 miRNA 韦恩图

文件位置: summary/3_DifferentiallymiRNAs/Fig7_Venn_diagrams_differentially_expressed_miRNA/

4.6 miRNA 表达和差异分析

4.6.1 差异 miRNA 的靶基因预测

针对动物物种，使用 TargetScan[9-11]、miRanda[12-14]两款软件对显著性差异的 miRNA 分别进行靶基因预测。对两款软件预测出的靶基因分别按照每款软件的评分标准进行筛选。TargetScan 算法中去除 context score percentile 小于 50 的靶基因，miRanda 算法中去除最大自由能（Max Energy）大于-10 的靶基因。最后取此两款软件的交集作为差异 miRNA 的最终靶基因。针对植物物种使用 TargetFinder[15] 对显著性差异的 miRNA 进行靶基因预测。此款软件使用基于植物的靶标罚分策略预测靶标。TargetScan, miRanda, TargetFinder 软件进行靶基因预测的详细原理和方法可查看辅助材料：[3.8.1 miRNA 靶基因预测原理。](#)

靶基因预测分析每列的具体含义见下表：

Term	意义
Transcript ID	转录本的 ID 号。转录本指的是一个基因生成的多个 isoform 中的其中一个，也叫 transcript。如果为已知转录本直接在数据库中查询
Species ID	物种的拉丁文或简写等用来区分物种信息
Gene ID	基因对应的 ID 号，一般来源于 NCBI 或 ensembl 等权威数据库，若查询基因具体信息直接搜索 ID 号即可
Symbol	基因名的缩写或简写
Gene Annoation	基因功能的具体描述
miRNA ID	miRNA 成熟体的 ID 号
TargetScan	Targetscan 预测结果，0 为没预测到，1 为预测到
miRanda	miRanda 预测结果，0 为没预测到，1 为预测到
Total	两款软件预测交集，1 表示两款软件任意匹配到后的结果，2 则表示取交集
GO Term	全称 gene ontology，即对基因进行注释，也叫 GO 功能注释。具体的 ID 号可在 GO 官网查阅功能。 GO 具体说明请下载联川生物大讲堂—GO 简介.pdf 查阅
KEGG pathways	小 k，指的是代谢通路的名称，用来表示一个特定的生物路径。 KEGG 具体说明请下载联川生物大讲堂—KEGG 简介.pdf 查阅

4.6.2 差异 miRNA 靶基因 GO 富集性分析

Gene Ontology（简称 GO）是一个国际化的基因功能分类体系，提供了一套动态更新的标准词汇表 (controlled vocabulary) 来全面描述生物体中基因和基因产物的属性。GO 总共有三个 ontology (本体)，分别描述基因的分子功能 (molecular function)、细胞的组成成分 (cellular component)、参与的生物过程 (biological process)。GO 的基本单位是 term (词条、节点)，每个 term 都对应一个属性。

GO 富集分析，首先把所有选定 miRNA 对应的靶基因对应于每个功能或者每个通路注释的 gene 数目统计出来，然后应用超几何检验，找出与所有选定 miRNA 对应的靶基因 mRNA 与注释库中的 GO 对应的 gene 数量（全部具备功能注释的基因，或者全部具备功能注释的 miRNA 靶基因）相比，计算得到 $p\text{-value} \leq 0.05$ 为阈值，满足此条件的功能定义为 miRNA-mRNA 关系对中显著富集的功能。通过功能显著性富集分析能确定 miRNA-mRNA 关系对行使的主要生物学功能。

GO 富集性分析每列的具体含义见下表：

Term	意义
GO Id	Gene ontology ID
Function	GO 的功能分类
GO_term	全称 gene ontology，即对基因进行注释，也叫 GO 功能注释。
S gene number	注释为特定 GO 的具有显著性差异的基因数，即 m
TS gene number	显著性差异的基因数目，即 n
B gene number	注释为特定 GO 的基因数，即 M
TB gene number	基因总数，即 N
pvalue	比较组和对照组比较有显著差异的基因占这个 GO term 所有基因数的比例 是否大于所有 GO term 的总基因数目中的差异基因比例的检验

文件位置：summary/Target_Enrichment_Analysis/Target_Enrichment/3_GO_significant.xlsx

在 GO 富集性分析的基础上，我们分别用柱状图，散点图和网络调控图对富集性分析的结果进行展示。

GO 富集结果

在所有的选定的 miRNA 靶基因中，我们将这些基因对应的 GO 注释按照 Molecular Function、Biological Process 和 Cellar Component 分为三类，并且将每一类中的 GO 功能按照注释到的靶基因个数从高到低排序，并进行作图，横坐标是对 GO 的分类，纵坐标是靶基因所占的百分比，具体的计算公式可参考补充材料。如上图所示，可以直观地看出注释到同一个 GO 的靶基因数目所占的百分比。

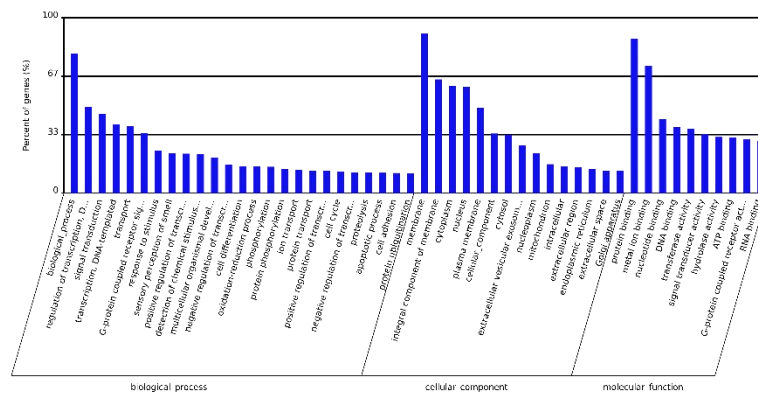


图 14: GO 富集柱状图

文件位置: summary/Target_Enrichment_Analysis/GO_Pictures/GO_bar_graph.png

GO 富集性散点图

采用 ggplot2 对 GO 富集分析结果以散点图展示: Rich factor 表示位于该 GO 的差异基因个数/位于该 GO 的总基因数, P 值越小, GO 富集程度越高。



图 15: GO 富集散点图

文件位置:

summary/Target_Enrichment_Analysis/Target_Enrichment/GO_enrichment_scatterplot.pdf

miRNA-gene-GO 网络图

对靶基因的富集分析结果进行筛选, 可选择以老师感兴趣的 go_term 为节点, 整理与该 GO_term 相

关的 miRNA 和 mRNA，用 cytoscape 软件绘制网络图，如图所示，红色表示 miRNA，绿色表示 mRNA，蓝色表示 GO 的注释。下图仅为展示数据。

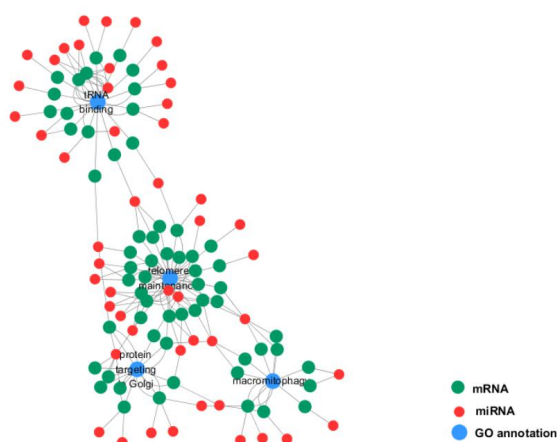


图 16: miRNA-gene-GO 网络图

老师如果想绘制属于自己的 Cytoscape 网络图，请下载 Cytoscape 使用手册。绘图数据在 [summary/Target_Enrichment_Analysis/Target_Enrichment/](#)

4.6.3 差异 miRNA 靶基因 GO 富集性分析

KEGG (Kyoto Encyclopedia of Genes and Genomes) 全称是京都基因和基因组百科全书，是基因组破译方面的公共数据库，基因组信息存储在 GENES 数据库里，包括完整和部分测序的基因组序列；更高级的功能存储在 pathway 数据库里，包括图解的细胞生物化学过程如代谢、膜转运、信号传递、细胞周期；KEGG 另一个数据库是 ligand，包含关于化学物质、酶分子、酶反应等信息。KEGG 提供的整合代谢途径 (pathway) 查询十分出色，包括碳水化合物、核苷、氨基酸等的代谢及有机物的生物降解，不仅提供了所有可能的代谢途径，而且对催化各步反应的酶进行了全面的注解，包含有氨基酸序列、PDB 库等。

KEGG 富集分析，首先把所有选定 miRNA 对应的靶基因对应于每个功能或者每个通路注释的 gene 数目统计出来，然后应用超几何检验，找出与所有选定 miRNA 对应的靶基因 mRNA 与注释库中的 KEGG 对应的 gene 数量（全部具备功能注释的基因，或者全部具备功能注释的 miRNA 靶基因）相比，计算得到 $p\text{-value} \leq 0.05$ 为阈值，满足此条件的功能定义为 miRNA-mRNA 关系对中显著富集的功能。通过功能显著性富集分析能确定 miRNA-mRNA 关系对行使的主要生物学功能。

KEGG 富集性分析每列的具体含义见下表：

Term	意义
Pathway Id	通路的名称，用来表示一个特定的生物路径
pathway	通路的具体描述
S gene number	注释为特定 KEGG 的具有显著性差异的基因数，即 m
TS gene number	显著性差异的基因数目，即 n
B gene number	注释为特定 KEGG 的基因数，即 M
TB gene number	基因总数，即 N
pvalue	比较组和对照组比较有显著差异的基因占这个 KEGG Pathway 代号所有基因数的比例 是否大于所有 KEGG Pathway 的总基因数目中的差异基因比例的检验
Pathway Link	Pathway 地址链接

文件位置：summary/Target_Enrichment_Analysis/Target_Enrichment/4_KEGG_significant.xlsx

在 KEGG 富集性分析的基础上，我们分别用柱状图，散点图和网络调控图对富集性分析的结果进行展示。

KEGG 富集结果

KEGG 通路图

Pathway 通路图展示与说明： 红色代表注释到某个 ko 节点且上调的显著差异表达基因，蓝色或紫色代表注释到某个 ko 节点且下调的显著差异表达基因。方框内的 4 位数字表示各种酶的 EC 编号；空心圆圈表示小分子化合物；实心箭头表示生化反应的方向；虚线箭头连接其他的相关代谢途径。该项目的展示图片为 hsa00010。

图 17：KEGG 通路图

文件位置: summary/Target_Enrichment_Analysis/KEGG_Pictures/

KEGG 散点图

用 ggplot2 对 KEGG 富集分析结果以散点图展示: RichFactor 表示位于该 KEGG 的差异基因数/位于该 KEGG 的总基因数, RichFactor 值越大, KEGG 富集程度越大。

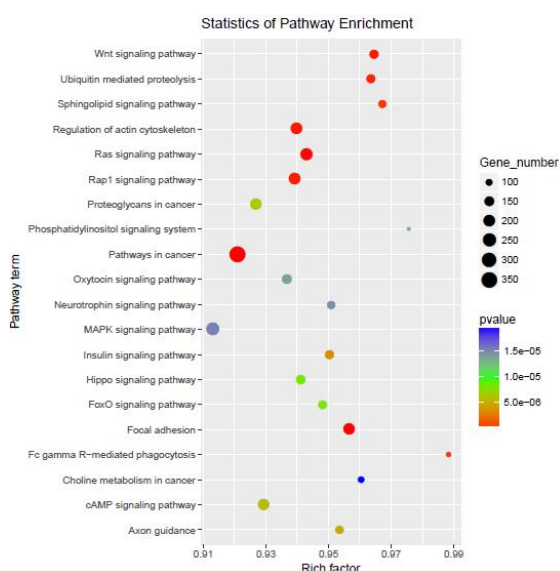


图 18: KEGG 富集散点图

文件位置:

summary/Target_Enrichment_Analysis/Target_Enrichment/KEGG_enrichment_scatterplot.pdf

miRNA-gene-KEGG 网络图

对靶基因的富集分析结果进行筛选, 可选择以老师感兴趣或重点关注的 kegg_pathway 为节点, 整理与该通路相关的 miRNA 和 mRNA, 用 cytoscape 软件绘制网络图, 如图所示, 红色表示 miRNA, 绿色表示 mRNA, 蓝色表示 kegg 通路的注释。下图仅为展示数据。

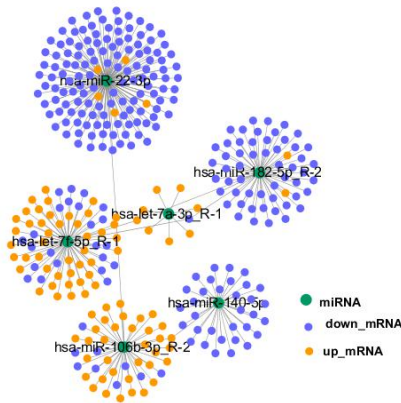


图 19: miRNA-gene-KEGG 网络图

老师如果想绘制属于自己的 Cytoscape 网络图，请下载 Cytoscape 使用手册。绘图数据在

summary/Target_Enrichment_Analysis/Target_Enrichment/

4.7 miRNA 成簇分析

有些 microRNA 在基因组中成簇(cluster)分布,这些 microRNA 来源于同一条初级转录本 Pri-miRNA, 它们转录成一个多顺反子结构,使得它们具有功能上的协同性。这种协同性在一个复杂的细胞信号调节网络中不可或缺,而且比单一的 miRNA 介导的生物学效用更强。由于转录后 miRNA 在成熟加工过程中的调节各异,因此成簇的 microRNA 成熟体的表达量有所不同。通过对成簇 microRNA 的分析,可以在 cluster 水平上观察 microRNA 的差异表达情况。详细内容请参考辅助文件: 3.9 miRNA 成簇分析。

文件位置: summary/

4.8 miRNA 保守性分析

microRNA 在物种间具有高度的保守性, 对于一个物种鉴定出来 microRNA, 可能在进化关系上具有保守性。在分析获得检出的 miRNA 基础上, 我们进一步对选定的物种进行了 miRNA 保守性分析, 对该物种中报道的 miRNA 在其它物种中的出现频率进行了数目统计。详细内容请参考辅助文件: 3.10 miRNA 保守性分析。

文件位置: summary/

4.9 miRNA 家族分析

对检测得到的 miRNA 进行家族分析，探索其所属的 miRNA 家族在其它物种中的存在情况。可以了解该 miRNA 在进化关系上的保守性。详细内容请参考辅助文件：3.11 miRNA 家族分析。

文件位置：summary/

4.10 miRNA 种子序列分析

在动物中，miRNA 与靶基因主要通过不完全互补配对的方式抑制 mRNA 的表达，其中位于 miRNA 5' 端所谓种子序列(第 2-7nt)与靶基因 3' UTR 可形成 Watson-Crick 配对是所有 miRNA 靶基因预测的最重要因素。详细内容请参考辅助文件：3.12 miRNA 种子序列分析。

文件位置：summary/

五、 数据挖掘

针对不同的研究目的，老师会有一些额外的个性化分析需求；因此在上述标准分析结果的基础上，我们为老师展示了一些个性化分析示例结果，具体如下。

5.1 siRNA 分析

siRNAs 是在真核生物中发现的一类内源性的具有调控功能的非编码 RNA，其长主要集中在 21、22 bp。siRNAs 由双链的 RNA 或 RNA 前体经 Dicer 酶加工而成，是 RNAi 的产物，其生成需要 Argonaute 家族蛋白存在。SiRNA 是 siRISC 的主要成员，主要作用是激发与之互补的目标 mRNA 的沉默[16]。

Solexa 深度测序所得的小 RNA(small RNA)几乎涵盖所有 RNA，包括 miRNA、siRNA、piRNA、rRNA、tRNA、snRNA、snoRNA、repeat associate sRNA、mRNA 降解片段等。通过与已知数据库进行比对 small RNA 进行过滤，同时选取没有被注释上的 small RNA，使用联川生物自主开发的软件 ACGT101 通过和基因组比对，鉴定 siRNA。

5.2 piRNA 分析

piRNA (Piwi - interactingRNA) 是从哺乳动物生殖细胞中分离得到的一类长度约为 30nt 的小 RNA，

且 piRNA 需要与 PIWI 蛋白家族成员相结合才能发挥它的调控作用[17]。目前,越来越多的文献表明,piRNA 在生殖细胞的生长发育中的调控是由 Piwi_piRNA 复合物引起的基因沉默导致。

PiRNA 的首位碱基有极强的 T 碱基偏好性,且大部分 piRNA 序列和基因组序列比对时仅有一个比对位点。通过与 Piwi 亚家族蛋白结合形成 piRNA 复合物 (piRC) 来调控基因沉默途径。对 Piwi 亚家族蛋白的遗传分析以及 piRNA 积累的时间特性研究发现, piRC 在配子发生过程中起着十分重要的作用。经过研究人员分析发现,只有 17%-20%的哺乳动物 piRNA 对应于标记的重复序列,包括转座子和逆转录转座子。因此, piRNA 从后生程序化和抑制转录到转录后调控可能会有不同的功能。根据 Piwi 蛋白已知的功能推测 piRNA 的功能可能有 3 个方面。

Solexa 深度测序所得的小 RNA (small RNA)几乎涵盖所有 RNA, 包括 piRNA、siRNA、piRNA, rRNA, tRNA, snRNA, snoRNA, repeat associate sRNA, mRNA 降解片段等。通过与已知基因组数据库和 piRNA 数据库进行比对,同时选取没有比对上 piRNA 数据库的 small RNA, 通过 piRNA 本身特性预测新 piRNA (novel piRNA)。

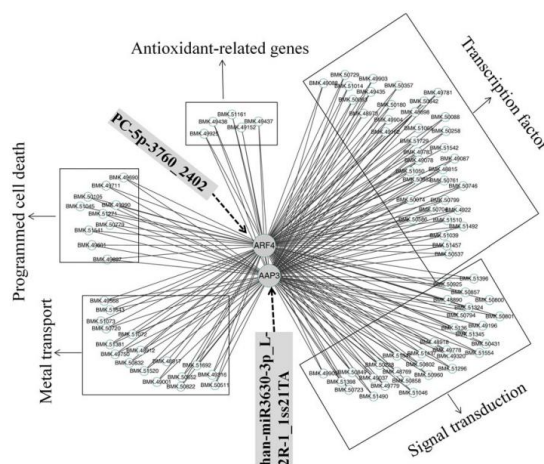
5.3 miRNA-转录组-降解组联合分析

转录组研究能够从整体水平研究基因功能以及基因结构,揭示特定生物学过程以及疾病发生过程中的分子机理。转录组研究是基因功能及结构研究的基础和出发点,了解转录组是解读基因组功能元件和揭示细胞及组织中分子组成所必需的,并且对理解机体发育和疾病具有重要作用。

在植物体内绝大多数的 miRNA 是利用剪切作用调控靶基因的表达,且剪切常发生在 miRNA 与 mRNA 互补区域的第十位或十一位核苷酸上。而降解组测序可以结合高通量测序技术与生物信息学分析各自的优势对植物 miRNA 的靶基因进行筛选,有效的避免假阳性结果的产生。

研究通过 miRNA 测序,转录组测序和降解组测序对东南景天的镉耐受性机制进行研究,超级累型东南景天的根,茎,叶用 400 μ M CdCl₂ 分别处理 0 h, 0.5 h, 6 h, 12 h, 24 h, 48 h, 72 h, 96 h, 小 RNA 测序鉴定出了 356 个 microRNA,其中 79 个差异表达的 microRNA ($P < 0.05$, expression > 10)。降解组测

序鉴定出 180 个保守 miRNA 的 722 个靶基因，14 个新发现的 miRNA 的 32 个靶基因，GO 富集性分析显示，多个靶基因参与镉胁迫应答。通过 WGCNA 软件进行共表达分析，发现 91 个共表达的基因同时受到共调控，并且其中两个 hub 基因 ARF4 和 AAP3 与镉胁迫应答密切相关[18]。

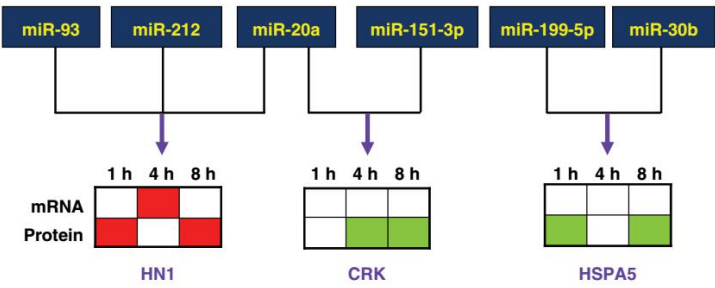


5.4 miRNA-转录组-蛋白联合分析

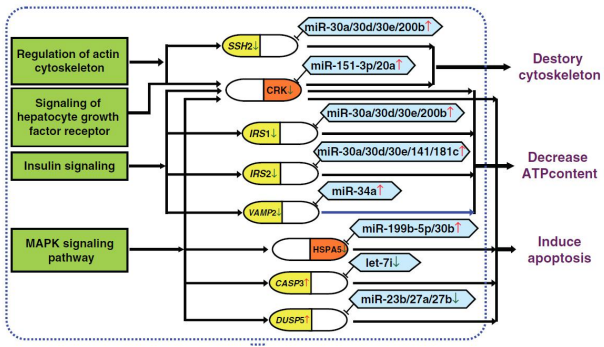
miRNA 和转录组整合分析，是一个系统性双层次联合研究 miRNA 调控基因表达的分子机制的方法。蛋白质是形式细胞功能的主要承担者，蛋白质组学(Proteomics)是从整体水平上研究细胞内蛋白质的组成、活动规律及蛋白质与蛋白质的相互作用，是功能基因组学时代一门新的学科，蛋白质组是细胞功能和状态的直接描述。miRNA-转录组-蛋白组学的联合分析是系统生物学的重要组成部分，也是目前多组学联合分析的基础。

本文的研究目的是探讨 microRNA 在银纳米粒子 (SNP) 对人体成纤维细胞 (HDF) 皮肤毒性作用的影响。本文首先，合成了 20 nm 的 SNP，并用 MTT 法测定了 SNP 对 HDF 增殖的影响，MTT 法显示，200 μ M 的 SNP 处理 1,4 和 8h 不能诱导 HDF 增值。随后又用蛋白组学和 miRNA 测序技术来研究 SNP 对 HDF 的增殖作用,200 μ M 的 SNP 处理的 HDF 细胞中差异蛋白和差异 miRNA 分别有 25 个和 246 个。本文通过整合转录组，蛋白质组和 miRNA 测序，发现 6 个差异表达的 miRNA 通过诱导 mRNA 调节 3 个靶 mRNA -蛋白质对，并且降解和抑制蛋白质翻译，从而导致表达模式的差异。生物信息学分析显示 SNP 诱导的差异表达的 miRNA 靶向 mRNA /蛋白质，共涉及 57 个生物学途径，

其中四个途径同时受到差异表达的 miRNA, 靶 mRNA 和靶蛋白的影响, 这四个生物学途径分别为 “调节肌动蛋白细胞骨架”, “肝细胞生长因子受体信号”, “胰岛素信号” 和 “MAPK 信号通路”。



图中可以看到 HN1, CRK 和 HSPA5 蛋白受到 6 个 miRNA 的调控, 并且 3 个 mRNA-蛋白靶标被 miRNA 共调控, 例如, HN1 的 mRNA-蛋白对受到 miR-93, miR-212 和 miR-20a 的调控, SNPs-1 h 组中, miR-93 并未差异表达, miR-212 上调, miR-20a 下调, HN1-mRNA 没有改变, 而 HN1-蛋白却上调, miR-20a 与 HN1 蛋白之间存在负相关, 表明 miR-20a 抑制 HN1 mRNA 的翻译。在 SNPs-4h 组中, miR-93 和 miR-212 下调, miR-20a 不变, 而 HN1 mRNA 上调, HN1 蛋白质没有变化。在这里, 负相关是在 miR-93, miR-212 和 HN1 mRNA 之间发现, 这表明这两种 miRNA 调控 mRNA 降解。然而, HN1 蛋白水平没有在 mRNA 翻译开始后改变, 其中可能是由于其他调控因素。在 SNPs-8 h 组, 仅 miR-20a 上调, HN1 mRNA 不变, HN1 蛋白上调。并没有观察到 HN1 mRNA /蛋白与 miRNA 之间的负相关性, 因此也许存在其他转录后调控机制。



本文涉及到四条主要的细胞信号通路 “调节肌动蛋白细胞骨架”, “肝细胞生长因子受体信号”, “胰岛素信号” 和 “MAPK 信号通路” 其中, “胰岛素信号通路” 调节细胞能量平衡, 胰岛素受体底物 IRS1 和 IRS2 在此途径中协调骨骼肌生长和通过 Akt 和 AMPK 途径调控代谢物产生。缺乏胰岛素信号转导

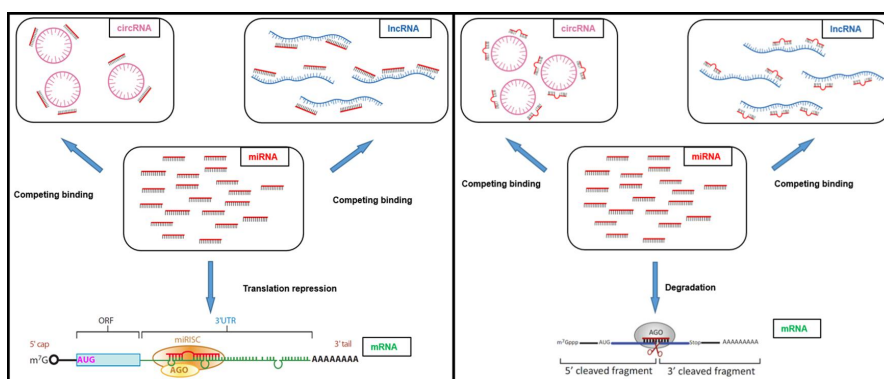
(IRS1, IRS2 等)可能会损害线粒体氧化磷酸化和 ATP 生成, 囊泡相关膜蛋白 2 (VAMP2) 和 CRK 位于 Akt 通路的下游。下调 IRS1, IRS2 和 VAMP2 的 mRNA, 以及 CRK 蛋白, 以上结果表明 SNP 可能抑制 HDF 中 ATP 的产生[19]。

5.5 ceRNA 网络调控分析

得益于高通量测序和生物信息学的发展, 人们通过生物信息学的方法, 在高通量测序的数据中发现了许多传统试验无法挖掘的 RNA。近几年大热的长链非编码 RNA (lncRNA)、环状 RNA (circRNA) 都成为了科研工作者新的研究热点。这类新型的非编码 RNA 除了转录前调控等功能外, 一个重要的功能就是像海绵一样对 microRNA 有吸附作用。这种具有 microRNA 吸附作用的 RNA, 我们称作内源性竞争结合 RNA (Competitive endogenous RNA, ceRNA), 这就是 ceRNA 的由来。

ceRNA 调控网络 (ceRNA networks) 指这些非编码 RNA 如 lncRNA 和 circRNA 等会竞争结合 microRNA, 导致 microRNA 调控的靶基因发生变化, 最终体现在蛋白的表达水平上, 而 microRNA 处于 ceRNA 调控网络中的核心地位。一个 ceRNA 可以结合多个 miRNA, 这些 ceRNA 上与 miRNA 结合的位点被称作 miRNA 识别元件(miRNA recognition elements, MREs)。通常 ceRNA 上有一个甚至多个 MRE, 当 ceRNA 表达量上升时, microRNA 就会被竞争结合, 从而导致 mRNA 转录水平上升, 最后蛋白表达水平上升, 反之亦然。

通常 ceRNA 上有一个甚至多个 MRE, ceRNA 调控网络如下图所示, 左侧为动物右侧为植物:



六、 辅助材料

6.1 实验方法 (英文材料)

和梁洪确认

6.2 生信分析 (英文材料)

动物篇 Raw reads were removed adapter dimers, junk, low complexity in FASTP, and then extract the UMI sequence from the rest read with length in 30~80 bp was used to remove the PCR Duplications. Valid reads were subjected to an in-house program, ACGT101-miR (LC Sciences, Houston, Texas, USA) to common RNA families (rRNA, tRNA, snRNA, snoRNA) and repeats. Subsequently, unique sequences with length in 18~26 nucleotide were mapped to specific species precursors in miRBase 22.0 by BLAST search to identify known miRNAs and novel 3p- and 5p-derived miRNAs. Length variation at both 3' and 5' ends and one mismatch inside of the sequence were allowed in the alignment. The unique sequences mapping to specific species mature miRNAs in hairpin arms were identified as known miRNAs. The unique sequences mapping to the other arm of known specific species precursor hairpin opposite to the annotated mature miRNA-containing arm were considered to be novel 5p- or 3p-derived miRNA candidates. The remaining sequences were mapped to other selected species precursors (with the exclusion of specific species) in miRBase 22.0 by BLAST search, and the mapped pre-miRNAs were further BLASTed against the specific species genomes to determine their genomic locations. The above two we defined as known miRNAs. The unmapped sequences were BLASTed against the specific genomes, and the hairpin RNA structures containing sequences were predicated from the flank 80 nt sequences using RNAfold software (<http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi>). The criteria for secondary structure prediction were: (1) number of nucleotides in one bulge in stem (≤ 12) (2)

number of base pairs in the stem region of the predicted hairpin (≥ 16) (3) cutoff of free energy (kCal/mol ≤ -15) (4) length of hairpin (up and down stems + terminal loop ≥ 50) (5) length of hairpin loop (≤ 20). (6) number of nucleotides in one bulge in mature region (≤ 8) (7) number of biased errors in one bulge in mature region (≤ 4) (8) number of biased bulges in mature region (≤ 2) (9) number of errors in mature region (≤ 7) (10) number of base pairs in the mature region of the predicted hairpin (≥ 12) (11) percent of mature in stem (≥ 80).

植物篇 Raw reads were removed adapter dimers, junk, low complexity in FASTP, and then extract the UMI sequence from the rest read with length in 30~80 bp was used to remove the PCR Duplications. Valid reads were subjected to an in-house program, ACGT101-miR (LC Sciences, Houston, Texas, USA) to remove common RNA families (rRNA, tRNA, snRNA, snoRNA) and repeats. Subsequently, unique sequences with length in 18~25 nucleotide were mapped to specific species precursors in miRBase 21.0 by BLAST search to identify known miRNAs and novel 3p- and 5p-derived miRNAs. Length variation at both 3' and 5' ends and one mismatch inside of the sequence were allowed in the alignment. The unique sequences mapping to specific species mature miRNAs in hairpin arms were identified as known miRNAs. The unique sequences mapping to the other arm of known specific species precursor hairpin opposite to the annotated mature miRNA-containing arm were considered to be novel 5p- or 3p-derived miRNA candidates. The remaining sequences were mapped to other selected species precursors (with the exclusion of specific species) in miRBase 21.0 by BLAST search, and the mapped pre-miRNAs were further BLASTed against the specific species genomes to determine their genomic locations. The above two we defined as known miRNAs. The unmapped sequences were BLASTed against the specific genomes, and the hairpin RNA structures containing sequences were predicated from the flank

120 nt sequences using RNAfold software (<http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi>). The criteria for secondary structure prediction were: (1) number of nucleotides in one bulge in stem (≤ 12) (2) number of base pairs in the stem region of the predicted hairpin (≥ 16) (3) cutoff of free energy (kCal/mol ≤ -15) (4) length of hairpin (up and down stems + terminal loop ≥ 50) (5) length of hairpin loop (≤ 200). (6) number of nucleotides in one bulge in mature region (≤ 4) (7) number of biased errors in one bulge in mature region (≤ 2) (8) number of biased bulges in mature region (≤ 2) (9) number of errors in mature region (≤ 4) (10) number of base pairs in the mature region of the predicted hairpin (≥ 12) (11) percent of mature in stem (≥ 80).

Analysis of Differential expressed miRNAs

Differential expression of miRNAs based on normalized deep-sequencing counts was analyzed by selectively using Fisher exact test, Chi-squared 2X2 test, Chi-squared nXn test, Student t test, or ANOVA based on the experiments design. The significance threshold was set to be 0.01 and 0.05 in each test.

The Prediction of Target Genes of miRNAs

To predict the genes targeted by most abundant miRNAs, two computational target prediction algorithms (TargetScan 5.0 Prepared by LC-Bio [www.lc-bio.com] support@lc-bio.com Hangzhou, Zhejiang, China, 310018 Tel. 86-571-87662413, Fax. 86-571-81951905 and Miranda 3.3a) were used to identify miRNA binding sites. Finally, the data predicted by both algorithms were combined and the overlaps were calculated. The GO terms and KEGG Pathway of these most abundant miRNAs, miRNA targets were also annotated. (动物)

To predict the genes targeted by most abundant miRNAs, computational target prediction algorithms (TargetFinder) were used to identify miRNA binding sites. The GO terms and KEGG

Pathway of most abundant miRNA targets were also annotated. (植物)

七、 参考文献

- [1] Fabian MR, Sonenberg N, Filipowicz W. Regulation of mRNA translation and stability by microRNAs [J]. *Annu. Rev. Biochem.*, 2010, 79: 351-379.
- [2] Garzon R, Calin GA, Croce CM. MicroRNAs in cancer [J]. *Annu. Rev. Med.*, 2009, 60: 167-179.
- [3] Axtell MJ. Classification and comparison of small RNAs from plants [J]. *Annu. Rev. Plant Biol.*, 2013, 64: 137-159.
- [4] Jones-Rhoades MW, Bartel DP, Bartel B. MicroRNAs and their regulatory roles in plants [J]. *Annu. Rev. Plant Biol.*, 2006, 57: 19-53.
- [5] Sun W, Julie Li Y-S, Huang H-D et al. microRNA: a master regulator of cellular processes for bioengineering systems [J]. *Annu. Rev. Biomed. Eng.*, 2010, 12: 1-27.
- [6] Iwakawa H-o, Tomari Y. The functions of microRNAs: mRNA decay and translational repression [J]. *Trends Cell Biol.*, 2015, 25(11): 651-665.
- [7] Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD, Shen Y, Pervouchine DD, Djebali S, Thurman RE, Kaul R, Rynes E, Kirilusha A, Marinov GK, Williams BA, Trout D, Amrhein H, Fisher- Aylor K, Antoshechkin I, DeSalvo G, See LH, Fastuca M, Drenkow J, Zaleski C, Dobin A, Prieto P, Lagarde J, Bussotti G, Tanzer A, Denas O, Li K, Bender MA, Zhang M, Byron R, Groudine MT, McCleary D, Pham L, Ye Z, Kuan S, Edsall L, Wu YC, Rasmussen MD, Bansal MS, Kellis M, Keller CA, Morrissey CS, Mishra T, Jain D, Dogan N, Harris RS, Cayting P, Kawli T, Boyle AP, Euskirchen G, Kundaje A, Lin S, Lin Y, Jansen C, Malladi VS, Cline MS, Erickson DT, Kirkup VM, Learned K, Sloan CA, Rosenbloom KR, Lacerda de Sousa B, Beal K, Pignatelli M, Flicek P, Lian J, Kahveci T, Lee D, Kent WJ, Ramalho Santos M, Herrero J, Notredame C, Johnson A, Vong S, Lee K,

Bates D, Neri F, Diegel M, Canfield T, Sabo PJ, Wilken MS, Reh TA, Giste E, Shafer A, Kutayavin T, Haugen E, Dunn D, Reynolds AP, Neph S, Humbert R, Hansen RS, De Bruijn M, Selleri L, Rudensky A, Josefowicz S, Samstein R, Eichler EE, Orkin SH, Levasseur D, Papayannopoulou T, Chang KH, Skoultschi A, Gosh S, Disteché C, Treuting P, Wang Y, Weiss MJ, Blobel GA, Cao X, Zhong S, Wang T, Good PJ, Lowdon RF, Adams LB, Zhou XQ, Pazin MJ, Feingold EA, Wold B, Taylor J, Mortazavi A, Weissman SM, Stamatoyannopoulos JA, Snyder MP, Guigo R, Gingeras TR, Gilbert DM, Hardison RC, Beer MA, Ren B, Mouse ENCODEC. A comparative encyclopedia of DNA elements in the mouse genome. *Nature*. 2014;515:355–64.

[8] Fu Y , Wu P H , Beane T , et al. Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers[J]. *Bmc Genomics*, 2018, 19.

[9] Agarwal V, Bell GW, Nam J-W et al. Predicting effective microRNA target sites in mammalian mRNAs [J]. *Elife*, 2015, 4: e05005.

[10] Friedman RC, Farh KK-H, Burge CB et al. Most mammalian mRNAs are conserved targets of microRNAs [J]. *Genome Res.*, 2009, 19(1): 92-105.

[11] Nam J-W, Rissland OS, Koppstein D et al. Global analyses of the effect of different cellular contexts on microRNA targeting [J]. *Mol. Cell*, 2014, 53(6): 1031-1043.

[12] Betel D, Koppal A, Agius P et al. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites [J]. *Genome Biol.*, 2010, 11(8): R90.

[13] Betel D, Wilson M, Gabow A et al. The microRNA. org resource: targets and expression [J]. *Nucleic Acids Res.*, 2008, 36(suppl 1): D149-D153.

[14] Enright AJ, John B, Gaul U et al. MicroRNA targets in *Drosophila* [J]. *Genome Biol.*, 2003, 5(1): R1.

[15] Dai X, Zhao PX. psRNATarget: a plant small RNA target analysis server [J]. *Nucleic Acids Res.*, 2011, 39(suppl 2): W155-W159.

[16] Franceschini, A., Meier, R., Casanova, A., Kreibich, S., Daga, N., Andritschke, D., ... & Conde-Álvarez, R. (2014). Specific inhibition of diverse pathogens in human cells by synthetic microRNA-like oligonucleotides inferred from RNAi screens. *PNAS*, 111(12), 4548-4553.

[17] Zhou, L., Canagarajah, B., Zhao, Y., Baibakov, B., Tokuhito, K., Maric, D., & Dean, J. (2017). BTBD18 Regulates a Subset of piRNA-Generating Loci through Transcription Elongation in Mice. *Dev Cell*, 40(5), 453-466.

[18] Han, X., Yin, H., Song, X., Zhang, Y., Liu, M., Sang, J., & Zhuo, R. (2016). Integration of small RNAs, degradome and transcriptome sequencing in hyperaccumulator *Sedum alfredii* uncovers a complex regulatory network and provides insights into cadmium phytoremediation. *Plant Biotechnol J*, 14(6), 1470-1483.

[19] Huang, Y., Lü, X., & Ma, J. (2014). Toxicity of silver nanoparticles to human dermal fibroblasts on microRNA level. *J Biomed Nanotechnol*, 10(11), 3304-3317.

八、 联系我们



地址：杭州经济技术开发区下沙 6 号大街 260 号中自科技园 16 幢 4 层

邮政编码：310018 邮箱：support@lc-bio.com 电话：0571-87662413

如果您有任何问题，欢迎随时与我们联系。

扫描二维码，关注联川生物微信公众号

样品提取怎么办？样品降解怎么办？数据分析不会怎么办？如何发表高质量的论文？更多资讯请随时

关注联川生物公众微信号，回复特定关键词提取相关信息