上海交通大学硕士学位论文

水稻中microRNA介导的基因沉默相关SNP的研究

|  |  |  |
| --- | --- | --- |
| 硕士研究生 | ： | 黄飘飘 |
| 学号 | ： | 1140809051 |
| 导 师 | ： | 张大兵教授、李俊彦博士 |
| 申请学位 | ： | 理学硕士 |
| 学科 | ： | 生物学 |
| 所 在 单 位 | ： | 生命科学技术学院 |
| 答 辩 日 期 | ： | 20 年 月 |
| 授予学位单位 | ： | 上海交通大学 |
|  |  |  |

Dissertation Submitted to Shanghai Jiao Tong University for the Degree of Master

INVESTIGATION OF SNPS INVOLVED IN RICE MIRNA-MEDIATED GENE SILENCING

|  |  |
| --- | --- |
| Candidate： | Piaopiao Huang |
| Student ID: | 1140809051 |
| Supervisor： | Prof. Dabing Zhang & Dr. Junyan Li |
| Academic Degree Applied for： | Master of Science |
| Speciality： | Biology |
| Affiliation： | School of Life Science and Biotechnology |
| Date of Defence： |  |
| Degree-Conferring-Institution： | Shanghai Jiao Tong University |

上海交通大学

学位论文原创性声明

本人郑重声明：所呈交的学位论文《水稻中microRNA介导的基因沉默相关SNP的研究》，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期： 年 月 日

上海交通大学

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权上海交通大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

**保密**□，在 年解密后适用本授权书。

本学位论文属于

不保密□。

（请在以上方框内打“**√**”）

学位论文作者签名： 指导教师签名：

日期： 年 月 日 日期： 年 月 日

水稻中microRNA介导的基因沉默相关SNP的研究

摘 要

miRNA是重要的调节因子，在植物的生长发育和抗逆性中扮演不可替代的角色。和miRNA介导的基因沉默相关的单核苷酸多态性(SNP)可能导致非常严重的植物农艺性状改变。为了进一步了解在miRNA和其靶基因上的进化压力以及SNP如何通过影响miRNA和靶基因对的互补模式从而影响miRNA相关的性状，我们进行了针对miRNA介导的调节的全基因组SNP的研究，并且分析了这些SNP对miRNA和靶基因互补模式以及相关性状的影响。我们发现pre-miRNA上SNP的密度比基因间隔区和外显子区域都要低，这和miRNA是植物中主要调控因子的角色相符。对比成熟的保守miRNA和非保守miRNA发现，两者的SNP分布并不相同，暗示了两者各位点上的进化压力不同；而对比成熟的保守miRNA和其相应的靶基因结合位点，则发现它们之间SNP分布有相关性，这也支持了miRNA和相应靶基因结合位点的共同进化的观点。在本研究中，我们将单倍型分析拓展成联合互补模式分析从而可以应用在miRNA和其结合位点上，并且我们找到了两个靶基因的结合位点上携带有两个可能对miRNA介导的调节产生重要影响的SNP，但是并没有发现这些SNP给水稻带来显著的形状变化。本研究是全基因组分析SNP对miRNA和靶基因相互作用的全新尝试，可能加深甚至改变我们对miRNA介导的基因沉默机理的理解以及SNP给miRNA和靶基因互作带来的影响。

关键词：水稻，microRNA，单核苷酸多态性，互补性，性状改变

INVESTIGATION OF SNPS INVOLVED IN RICE MIRNA-MEDIATED GENE SILENCING

ABSTRACT

MiRNAs are key regulators and play inevitable role in plant growth, development and stress tolerance. Single nucleotide polymorphisms (SNPs) that are involved in miRNA-mediated gene silencing might cause serious changes to plant agronomic traits. To further understand the evolutionary pressure imposed on miRNAs and their targets as well as how SNPs could affect the complementarity of miRNA:target pairs and further bring changes to miRNA-involved phenotypes, we performed a genome-wide scan of SNPs involved in miRNA-mediated regulation, and analyzed their effects on miRNA:target complementarity and related phenotypes. We found that SNP density of pre-miRNAs was lower than both intergenic regions and exons, consistent with their established roles as master regulators in many genetic pathways. And comparison between conserved mature miRNAs and non-conserved mature miRNAs showed the SNP distributions were rather different, implying the differential selection pressure upon them; while comparison between conserved mature miRNAs and their binding sites showed similar SNP distribution, and this supported the co-evolution of miRNAs and their binding sites of cognate targets. In this study, we extended haplotype analysis into combined complementarity pattern analysis to apply on miRNA:target pairs, and found two target genes carrying SNPs which potentially may bring great changes to miRNA-mediated regulation, but we didn’t find obvious phenotypical changes for these SNPs. This study provided a new attempt of analyzing genome-wide SNPs on miRNA:target interactions, and might change and deepen our understand of the mechanism of miRNA-mediated gene silencing and the effects of SNPs on the miRNA:target interaction.

KEY WORDS: Rice, microRNA, SNP, complementarity, phenotypical change

目 录

↑

（黑体3号字，段前0.7厘米，段后0；目录题目与条目之间空两行）

[第一章 正文文字格式 1](#_Toc251590708)

[1.1 论文正文 1](#_Toc251590709)

[1.2 字数要求 1](#_Toc251590710)

[1.2.1 硕士论文字数要求 1](#_Toc251590711)

[1.2.2 博士论文字数要求 1](#_Toc251590712)

[1.3 论文的主要内容与章节安排 1](#_Toc251590713)

[第二章 图表、公式格式 2](#_Toc251590714)

[2.1 图表格式 2](#_Toc251590715)

[2.2 公式格式 4](#_Toc251590716)

[2.3 引用说明 4](#_Toc251590717)

[2.4 本章小结 4](#_Toc251590718)

[第三章 其他格式要求 5](#_Toc251590719)

[3.1 页码 5](#_Toc251590720)

[3.2 页眉 5](#_Toc251590721)

[3.3 目录 5](#_Toc251590722)

[3.4 正文的层次安排 5](#_Toc251590723)

[3.5 打印要求 6](#_Toc251590724)

[3.5.1 页面设置 6](#_Toc251590725)

[3.5.2 字体 6](#_Toc251590726)

[3.5.3 字号 6](#_Toc251590727)

[第四章 结束语 7](#_Toc251590728)

[4.1 主要工作与创新点 7](#_Toc251590729)

[4.2 后续研究工作 7](#_Toc251590730)

[参 考 文 献 8](#_Toc251590731)

[附录**1** 10](#_Toc251590732)

[致 谢 11](#_Toc251590733)

[攻读硕士学位期间已发表或录用的论文 12](#_Toc251590734)

图 录

↑

（黑体3号字，段前0.7厘米，段后0；目录题目与条目之间空两行）

[图2-1论文页面设置图 2](#_Toc251150896)

[图2-2内热源沿径向的分布 3](#_Toc251150897)

表 录

↑

（黑体3号字，段前0.7厘米，段后0；目录题目与条目之间空两行）

[表2-1高频感应加热的基本参数 3](#_Toc251151029)

[表3-1论文的层次代号与说明 6](#_Toc251151065)

# 正文文字格式

# 水稻miRNA信息整理和相应靶基因的预测和整理

## 水稻miRNA整理和相应信息收集

miRBase中收集和整理了目前所有已发表的miRNA的序列以及相关的注释，包括miRNA描述、所在基因家族、“茎环”结构 (stem-loop)、深度测序结果、基因组内容(genome context)等[[1]](#endnote-1)，其中也包含了所有已发表的水稻miRNA的序列和注释，而水稻miRNA是以MSU7.0作为参考基因组。 所以，我们首先从miRBase中将所有的水稻miRNA及其相关注释信息下载下来，主要是包括pre-miRNA序列、成熟产物 (mature miRNA)、基因组内容。在此总共收集得592个pre-miRNA和相应的713个成熟miRNA。

但是，其中有14个pre-miRNA以及相应的24个成熟miRNA的基因组坐标信息并没有提供，其中包括著名的调节MADS-box基因的osa-miR444家族，但由于随后的SNP分析需要miRNA的基因组坐标。所以，我们用本地BLAST，将MSU7.0的水稻基因组序列创建成本地的BLAST数据库，并且用这些pre-miRNA的序列作为信息进行比对。对于BLAST结果的筛选标准是所有的序列完全匹配，其中发现osa-miR444家族的所有pre-miRNA以及osa-miR810a在参照基因组上比对的匹配结果中基因组坐标是间断的，说明它们基因的转录本经过了剪切和拼接。另外在线的BLAST也得出相同的结果，更加有意思的是发现osa-miR444的基因组位置正处在MADS-box基因的反义链上，这正是之前有报道过的天然反义miRNA (nat-miRNA) [[2]](#endnote-2)。通过运用BLAST，最终我们得到了585个pre-miRNA和703个成熟miRNA的所需信息，可以供本实验进一步研究使用。

miRBase中成熟miRNA的基因组位置是其在pre-miRNA上的相对坐标，使用Python脚本将相对坐标结合pre-miRNA的基因组坐标而得到所有相应的成熟miRNA的基因组绝对坐标。

### miRNA根据保守性分类

水稻miRNA可以根据其在不同的植物物种中的保守性分成两类：一类是保守miRNA (conserved miRNA)，也就是在其它的物种中有相应的同源miRNA；另一类则是非保守miRNA (non-conserved miRNA)，通常也称为水稻特异miRNA (rice-specific miRNAs)。两者之间有一些明显的差别，首先就其定义而言，它们的保守性不同；其次，一般说来，多数的保守miRNA的表达量都比非保守miRNA的表达量高；最后，最关键的差别是，保守miRNA几乎都有可辨认的靶基因，而且这些靶基因往往也在植物物种中有保守性，甚至是一些调控因子或其它的调控基因，比如DCL1，ARGONAUTE和生长素受体TIR1，但是非保守miRNA却鲜有可辨认的靶基因[[3]](#endnote-3)。由此可以看出，保守miRNA在植物生长发育等过程中会比非保守miRNA扮演更重要的角色，并且两者之间因其进化过程和靶基因的结合模式不同所以作用机理也可能略有差异，所以在分析研究中将两者分别进行研究和比较非常有必要。

使用miRBase提供的miRNA家族分类标准对pre-miRNA进行保守性分类，分类标准是：该pre-miRNA所在的家族如果还有其它的植物物种的miRNA，就认定其为保守miRNA，否则就认为非保守miRNA。最后的分类结果是，191个pre-miRNA归类为保守miRNA，401个pre-miRNA归类为非保守miRNA；同时，220个成熟miRNA归类为保守成熟miRNA，另外493个归类为非保守成熟miRNA。

## miRNA靶基因预测和整理

鉴于保守miRNA拥有更多重要的靶基因，并且针对这些miRNA的研究更加充分，本实验只针对保守miRNA的靶基因进行研究。而现有的miRNA靶基因的预测方法主要是生物信息方法，比如在线植物小RNA靶基因预测服务器 (psRNATarget, http://plantgrn.noble.org/psRNATarget/)；而实验中鉴定靶基因的方法有miRNA过表达或抗miRNA靶基因 (miRNA-resistant target)，RNA连接酶催化的5’-RACE，降解组测序等[[4]](#endnote-4)。

本实验采用生物信息预测方法结合降解组测试的结果获得miRNA靶基因。使用生物预测在线软件psRNATarget，采用默认的参数进行预测[[5]](#endnote-5)。另外 Li et.al 2010对水稻进行了全转录组的miRNA靶基因鉴定[[6]](#endnote-6)，我们收集了其鉴定的结果。但是这些鉴定出来的靶基因缺乏更加具体的miRNA结合位点的信息，我们使用psRNATarget用这些靶基因反向搜索其对应的miRNA，从而获得相应的结合位点的信息。最终，结合两者的结果，我们得到了823个靶基因，相对应的有2113个miRNA靶基因对。

## 本章小节

本章详细叙述了如何获得所有的miRNA，以及研究所必需其在的水稻基因组的坐标信息，并且使用BLAST得到了一些miRNA所缺失的基因组坐标。进而，根据miRBase的miRNA家族分类，将miRNA分成保守和非保守的miRNA。最后，结合生物信息预测和降解组的数据，得到保守miRNA的靶基因823个，形成miRNA和靶基因的作用对总共2113对。

# 水稻miRNA靶基因的生物相关性筛选

所得到的823个靶基因中，46个靶基因是降解组测序所证实的，另外777个靶基因是基于植物miRNA和其靶基因具有高度互补性(complementarity)的特征而预测出来的。但是用这种生物信息方式得出的预测结果，其生物相关性(biological relevancy)并不能得到保证，也就是这些预测的靶基因并不能确定真的能够被该miRNA调控，并且能够造成植物生理上的影响。因为除了miRNA和靶基因结合位点的序列互补性之外，还有其它因素会影响到miRNA和靶基因的相互结合4：

1. 靶基因的可接近性，这个是由靶基因mRNA的二级结构决定，如果靶基因的结合位点部分需要很高的能量来释放从而可以和成熟miRNA结合的话，miRNA和靶基因的结合能力就相应下降[[7]](#endnote-7)；
2. miRNA和靶基因结合位点序列的整理互补性模式，即使当互补程度相同时，错配出现的地方不同，也会非常影响到miRNA对靶基因的调节[[8]](#endnote-8)；
3. RNA结合蛋白等。

除了上述列举的因素之外，即便是实验的方法也并不能完全保证所鉴定的靶基因的生物相关性，比如miRNA过表达，虽然可以通过过表达的转基因植株表现出的表型来判断潜在的靶基因，但这种人工的过量表达和原本miRNA在植物中的表达水平不同，在原本的植物组织中可能并不具有调节能力的miRNA因此却被鉴定出可以调节某靶基因。

鉴于此，我们需要对所预测出的靶基因生物相关性筛选。由于植物miRNA介导的基因沉默主要是通过靶基因转录子的剪切这种方式，在通常的假设下，也就是，互补性是基因沉默的唯一决定因子和任何给定的装载miRNA的RNA介导的沉默复合体(miRNA-loaded RNA-induced silencing complex, miRISC)都能够独立运作[[9]](#endnote-9),[[10]](#endnote-10)，靶基因mRNA的表达量会被相应的miRNA下调，因此植物miRNA的表达量和其靶基因的表达量就呈现负相关。并且也有实验表明过表达miR319时，五个编码TCP转录因子的基因下调，呈现负相关[[11]](#endnote-11)。因此，我们考虑以miRNA和靶基因的表达量负相关性来筛选具有生物相关性的靶基因。

从RiceFrend[[12]](#endnote-12)的水稻整个生长期间各种组织/器官的基因表达实验中得到表达量的数据，该表达量数据现存于EMBL-EBI数据库，编号为E-GEOD-21396。其中有48个样本以代表生长发育不同阶段的组织／器官，而其中包含了169个miRNA的探针以检测在样本中的miRNA表达量。

## miRNA的表达热图(heatmap)

在所得到的表达量数据中，每个miRNA都有相对应的六个探针的数据，分别是pre-miRNA在3’端、5’端和成熟miRNA序列各有两个探针，所以这些探针中3’端和5’端的探针数据看作是表示pre-miRNA的表达量，而成熟miRNA序列中的探针看作是成熟miRNA的表达量。分别对pre-miRNA的四个探针数据和成熟miRNA的两个探针数据取平均值，就得到相应的pre-miRNA和成熟miRNA的表达量。

利用Genesis软件[[13]](#endnote-13)分别绘制了pre-miRNA和成熟miRNA的表达热图。

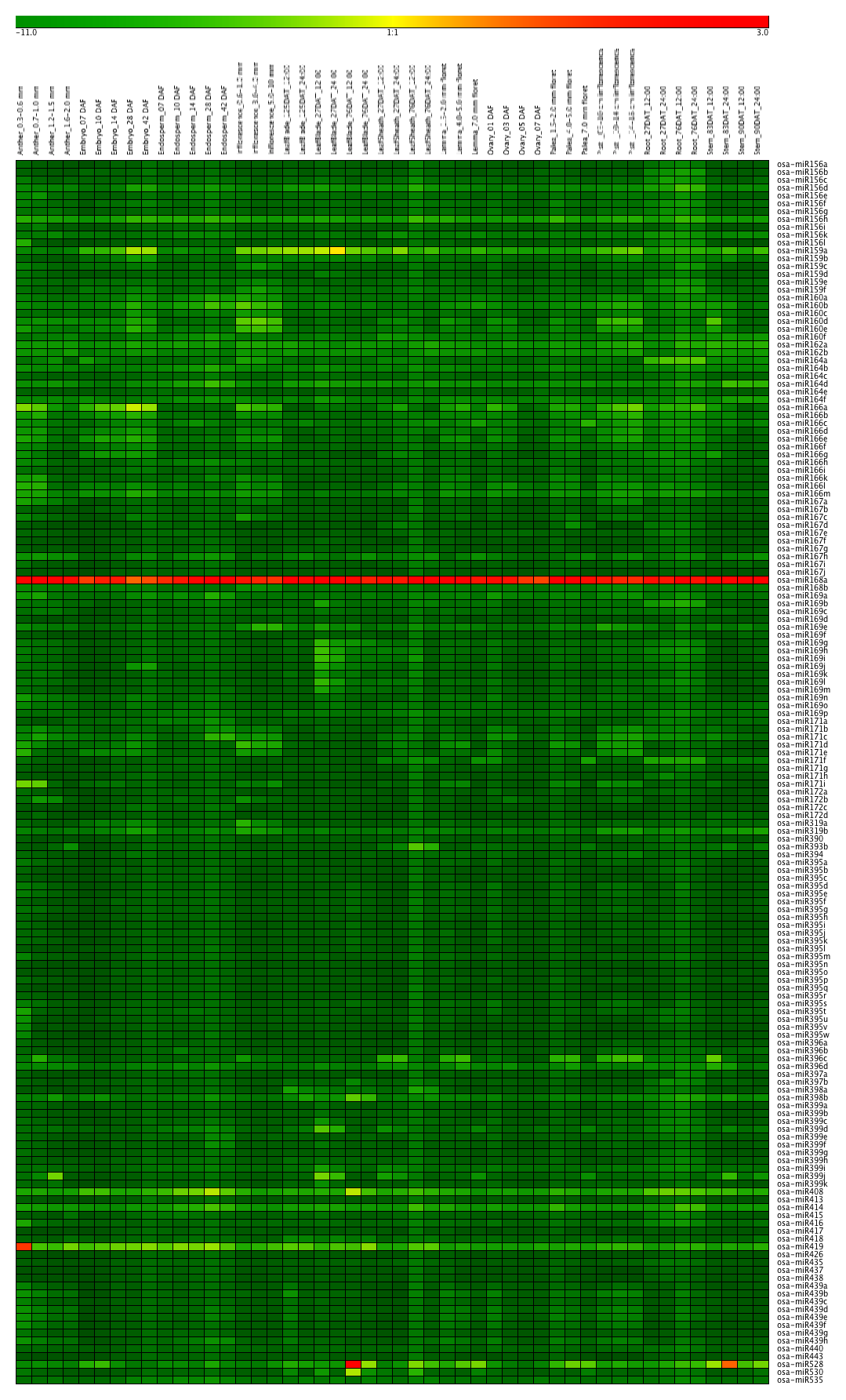


图3-1 pre-miRNA在水稻各个不同时期的组织／器官中的表达热图

Fig.3-1 Heatmap of expression levels of pre-miRNAs in various tissues/organs at various stages

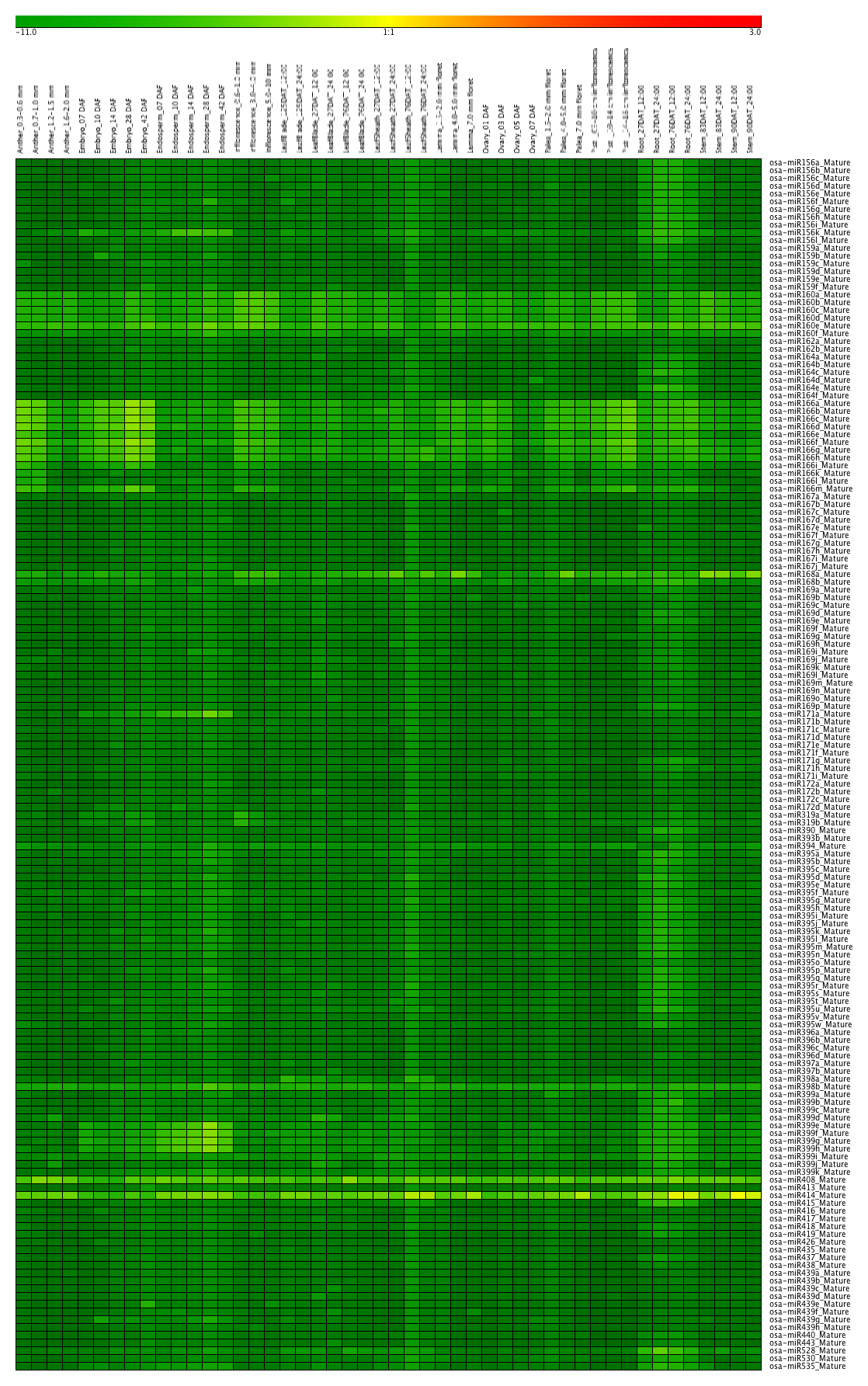


图3-2 成熟miRNA在水稻各个不同时期的组织／器官中的表达热图

Fig.3-2 Heatmap of expression levels of mature miRNAs in various tissues/organs at various stages

以上的图3-1和图3-2中，显示了同一个miRNA在不同阶段的组织和器官中的表达量存在差别，也说明miRNA在不同的组织和器官中的调节强度有差异。另外图3-1中，同一个家族的miRNA表达存在差异，也暗示了miRNA存在的亚功能化 (subfunctionalization)，但是图3-2中，同一个家族的miRNA之间的差异较小，因为同一个家族的成熟miRNA序列非常相似，所以其探针所得到的表达量也很相似。图3-1显示，osa-miR168a是所有检测的pre-miRNA中表达量最高的，图3-2则显示，osa-miR414是所有检测的成熟miRNA中表达量最高的。

## 降解组验证miRNA和靶基因作用对(miRNA:target interaction pair)的表达量相关性检验

收集的靶基因所包含的46个靶基因是用降解组测序验证的，可以作为真正的靶基因。而根据传统的理论，在以上提到的两个前提下，它们和其相应的miRNA之间应该总是有表达量负相关的关联。上节的图3-2结果显示，两者同一个miRNA家族的成熟miRNA表达量几乎相同，源自于探针几乎相同，而真正有区别的是pre-miRNA的表达量，所以此处我们对pre-miRNA和靶基因的表达量进行皮尔森相关性检验(Pearson correlation coefficient test)。

因为降解组测序实验所采用的样本是3周大的水稻幼苗，所以在进行相关性检验时我们选择其中27天大的水稻幼苗样本，选择的组织时叶片和叶鞘。

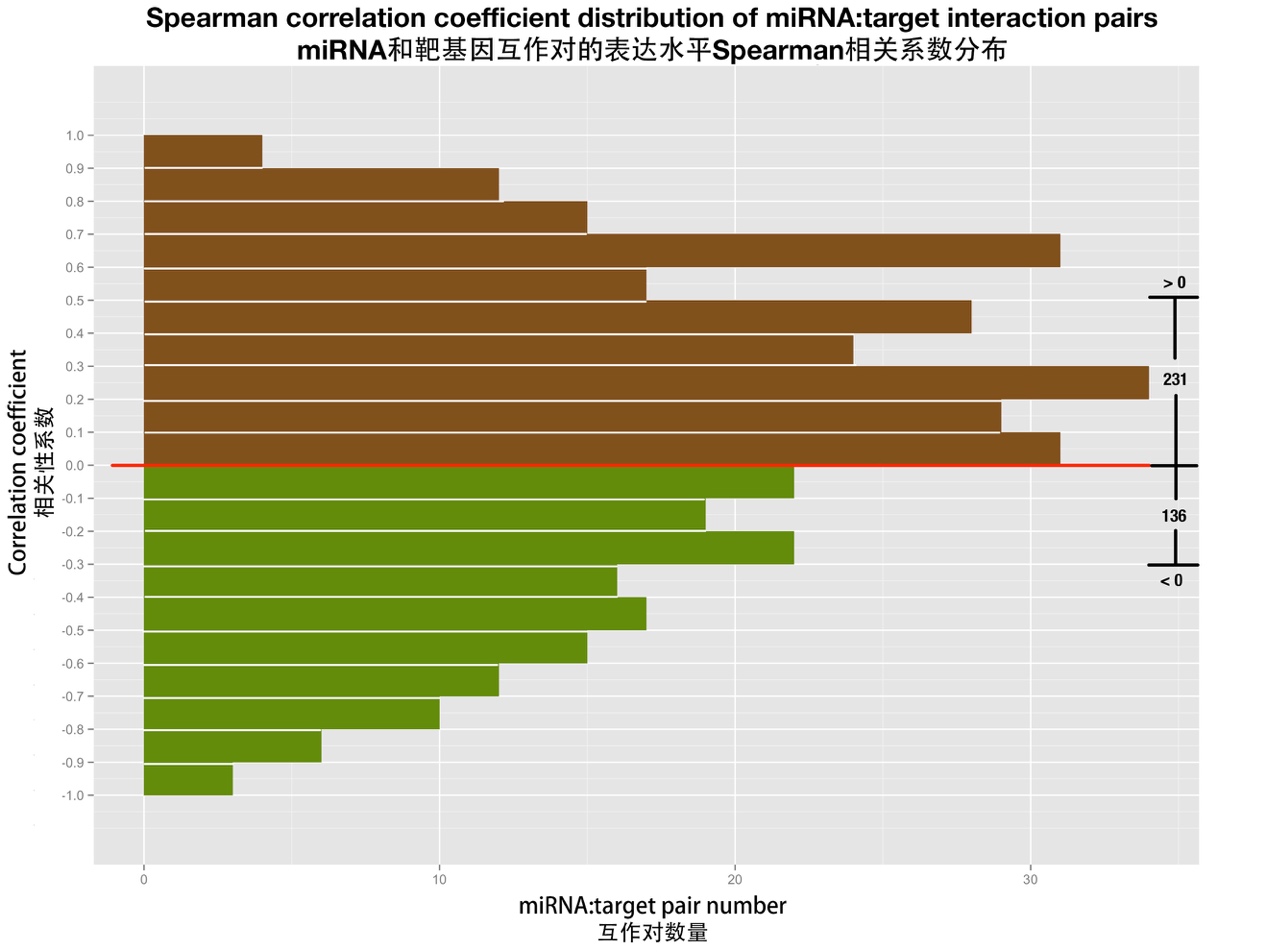


图3-3由降解组验证的miRNA:靶基因作用对之间Spearman相关性系数图

注：X坐标是落在特定相关系数区间的作用对数量，其中下方负相关的作用对的条形图用绿色标出，而上方正相关的作用对则用棕色标出。

Fig. 3- Heatmap of expression levels of mature miRNAs in various tissues/organs at various stages

Note: X-axis is the number of pairs that fall on the specified range of correlation coefficient; so bars with green color in the lower part denote the negatively correlated miRNA:target pairs, by contrast, bars with brown color in the upper part denote the positively correlated miRNA:target pairs.

从图3-3可以看出，367个miRNA和靶基因对中，只有136个呈现负相关，而原定的假设是真正的靶基因和其miRNA在相互作用的组织中总是呈现负相关，所得的结果并不支持原定的假设。不但如此，超过一半(197/367)的作用对之间是弱相关(-0.4~0.4)，这暗示在这些miRNA和靶基因对的表达量之间可能根本就不存在直接的相关性。

Wen et al.,2016也发现了类似的现象，在他们的研究中正相关的miRNA和靶基因作用对也是比负相关的要多。而这种现象可能是由更加复杂的作用机理比如负反馈环路 (negative feedback loops, FBLs) 和不合理正反馈环路 (incoherent feedforward loops, FFLs) 导致[[14]](#endnote-14)。

以上的结果也再次强调了植物miRNA介导的基因沉默还有很多位置的复杂因素未被研究清楚。同时也说明，表达量负相关的方法不能用以筛选具有生物相关性的miRNA和靶基因相互作用对。

## miRNA家族内不同miRNA和靶基因表达量相关性比较

水稻中，很多保守miRNA家族都有不止一个成员，成员之间的成熟miRNA的序列有很高的相似性，同时它们会有重叠度非常高的靶基因集合，而且通常都是来自于同一个基因家族。比如osa-miR156家族有12个成员，它们的靶基因则都是集中在SBP基因家族。对比osa-miR156家族中miRNA和SBP中靶基因的表达量相关性（仍然采用皮尔森相关性系数）。

图3-4中，可以看出对比osa-miR156家族成员和同一个基因的表达量相关性系数，发现它们各自不相同，而且相差很大，这再次暗示了miRNA家族中有亚功能化(subfunctionalization)。；但是值得注意的是，对于每个基因，都有miRNA与之有表达量负相关关系，这暗示着osa-miR156家族中的确有miRNA在调控这个基因。

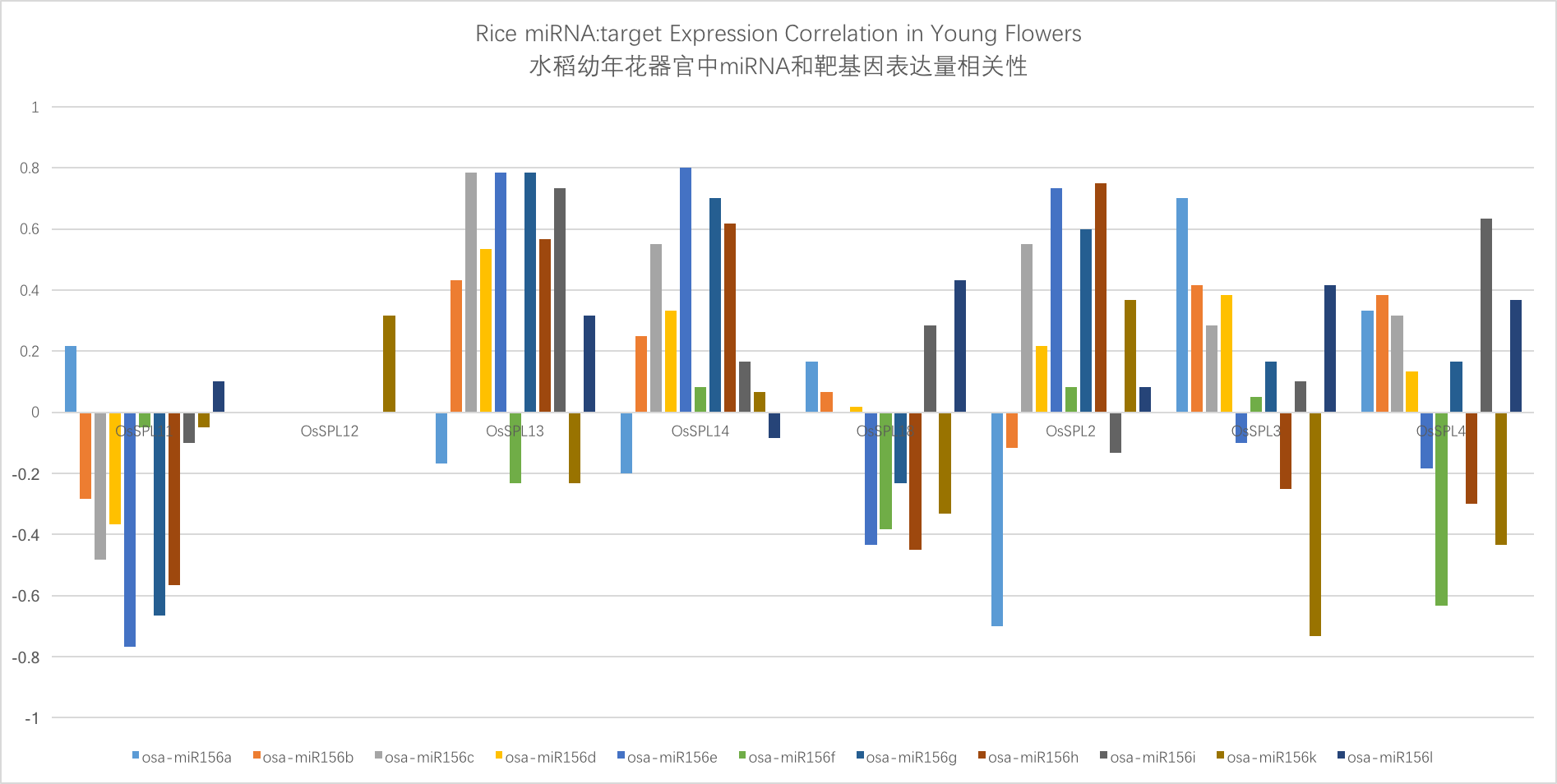


图3-4 osa-miR156家族和SBP基因家族之间表达量的皮尔森相关性系数\*

Fig. 3-4 Pearson correlation coefficient of expression levels of osa-miR156 family and its target gene family SBP

\*幼年花器官包括：0.6-1.0mm花序、3.0-4.0mm花序和5.0-10mm花序

## 本章小节

本章尝试用miRNA和靶基因的表达量水平相关性来筛选出其中具有生物相关性的miRNA和靶基因作用对。但是在对经降解组验证的miRNA和靶基因对进行表达量水平相关性检验时，我们却发现它们之间并没有绝对的负相关性。因此，我们也不能用这种方法对miRNA和靶基因的生物相关性进行筛选，同时也对普遍所认为的miRNA和真正的靶基因之间的表达量具有负相关的观点提出质疑。

另外，对比miRNA家族的表达量和它们与相同靶基因之间的表达量相关性系数，可以发现miRNA家族成员的调节作用并不总是一样，反而呈现亚功能化。

# miRNA及其靶基因上结合位点的SNP获得和分析

## 本地SNP数据库的建立

最近的3K水稻基因组测序项目产生了23M SNP，如此大量的数据可以给水稻中的SNP研究提供很多帮助，而本研究中的所有的SNP都是来源于此(SNP-Seek database, http://snp-seek.irri.org/)。为了方便研究使用，我们将在线的SNP数据下载到本地，并且使用MySQL建立起本地的SNP数据库。

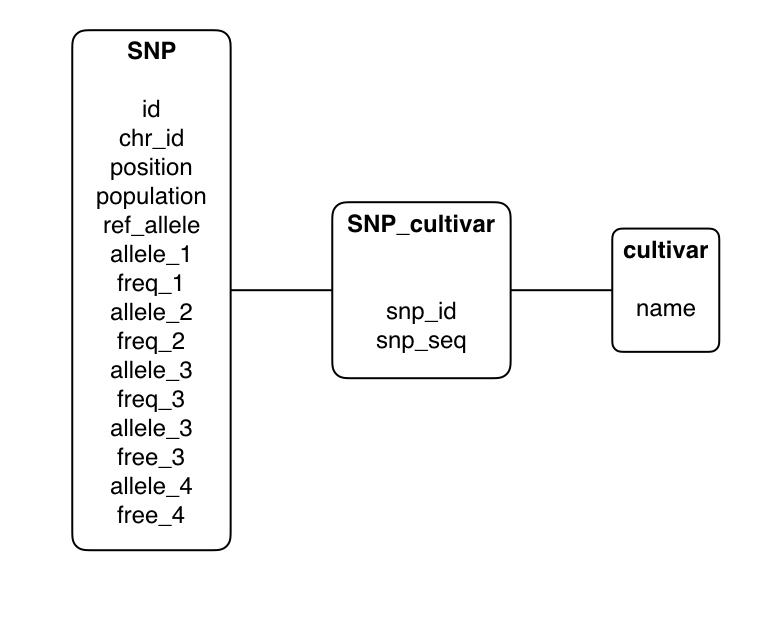


图4-1本地SNP数据库模式图

注：图中三个框内分别代表在本地数据库中的三个表，粗体的是表名，而下面的则是列名称

Fig.4- Basic schema of local SNP database

Note: The three frames are three tables created in local MySQL database, table names are in bold, while the lower ones are column names

使用本地Python脚本，搜索在上述的pre-miRNA、成熟miRNA和靶基因上miRNA结合位点的SNP以及靶基因上结合位点的侧翼区间（上下游各100bp），我们得到大量的SNP数据。从表4-1中，保守miRNA上的平均SNP数量为0.85个，成熟miRNA上的平均SNP数量为1.60个，而在结合位点上的平均SNP数量为1.42个，可以看出保守miRNA的平均SNP数量相较于两者更低；另外在靶基因上miRNA结合位点侧翼的平均SNP数量为9.78个，高于miRNA结合位点的SNP平均数量，但是考虑到其长度是200bp，其上的密度是0.0489 个/ bp，而结合位点上的则是0.0676个/ bp，所以相比之下，结合位点侧翼的SNP密度反倒要比结合位点要低。因为靶基因的可达型(target accessibility)也对于miRNA对靶基因的识别有影响7，而靶基因可达型则是由靶基因mRNA的二级结构决定，特别是结合位点侧翼区段，这可能导致结合位点侧翼SNP密度较小的现象。

表4-1 miRNA和靶基因上的SNP数量统计

Table 4-1 Summary of SNPs on miRNAs and targets

|  |  |  |
| --- | --- | --- |
| Fragment type 基因组片段类型 | Number 总数 | Number of SNPs SNP的数量 |
| pre-miRNAs | 585 | 7193 |
| 成熟miRNA | 793 | 1270 |
| 保守miRNA | 220 | 188 |
| 靶基因上miRNA结合位点 | 823 | 1169 |
| 结合位点侧翼 (~100bp上下游) | 823 | 8048 |

### SNP数据库之间的比较

RiceVarMap (Rice Variation Map, <http://ricevarmap.ncpgr.cn)>[[15]](#endnote-15)是在3K水稻基因测序项目之前的水稻基因组突变数据库，其中包括了从1479个水稻品系测序数据而得的6,551,358个SNP以及1,214,627个INDEL (insertions/deletions)。本实验初期也曾使用RiceVarMap进行分析。由于该数据库所采用的水稻参照基因组是Nipponbare, MSU 6.1，在SNP搜索中就需要进行参照基因组转换 (genome lift)。我们成功转换了560个pre-miRNA和665个成熟miRNA的基因组坐标，并且使用RiceVarMap进行SNP搜索。在pre-miRNA上总共得到1724个SNP，而其中323个SNP落在成熟miRNA上。

对比RiceVarMap和SNP-seek database的SNP结果，发现两者之间1670个SNP是共有的，而只有54个SNP是RiceVarMap所特有的，两者的重合率达到96.9%，所以可以看出SNP-seek database的SNP是具有很高的可信度。

另外，在本实验研究过程中，SNP-seek database经历过版本更新，水稻品系总数量从3,000增加到3,024，同时所得到的SNP总数量也增加了。旧版本中pre-miRNA上SNP数量是4617，成熟miRNA上SNP数量则是798。比较可以发现，新版本SNP-seek database的数据拥有更丰富的SNP数据，而三个数据来源共同包括了1617个SNP，重合率是93.8%，表明了数据库具有很高的可信度。

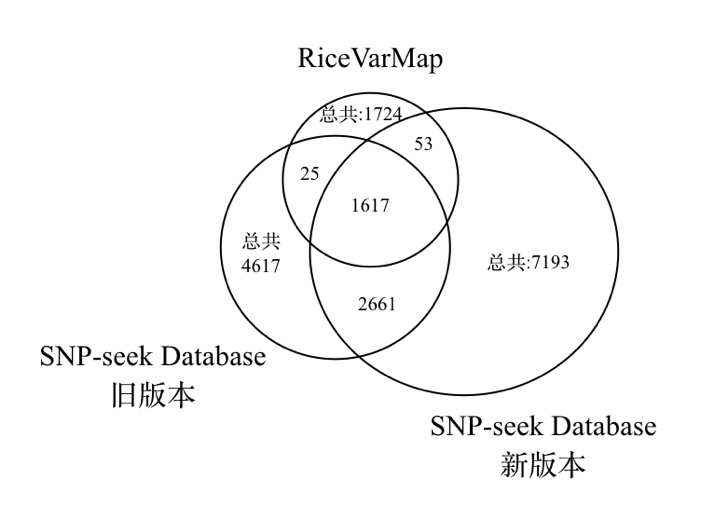


图4-2 RiceVarMap和SNP-seek database新旧版本中pre-miRNA上SNP数量

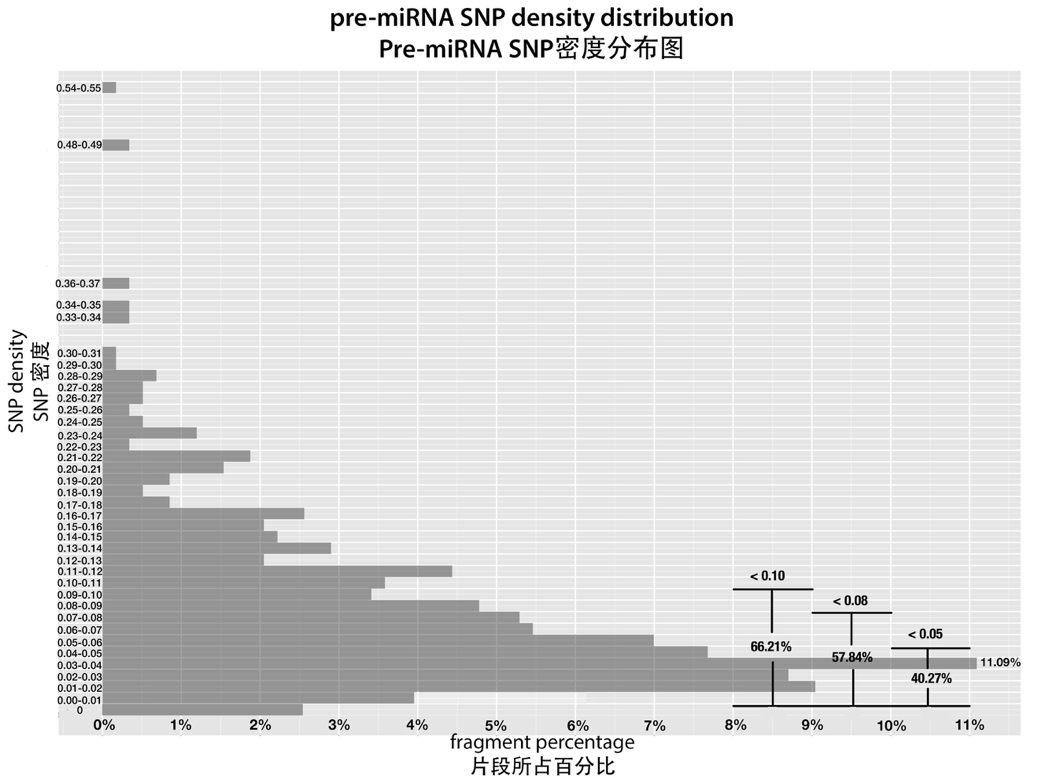
Fig.4- Number of SNPs on pre-miRNAs in RiceVarMap, and in both versions of SNP-seek database

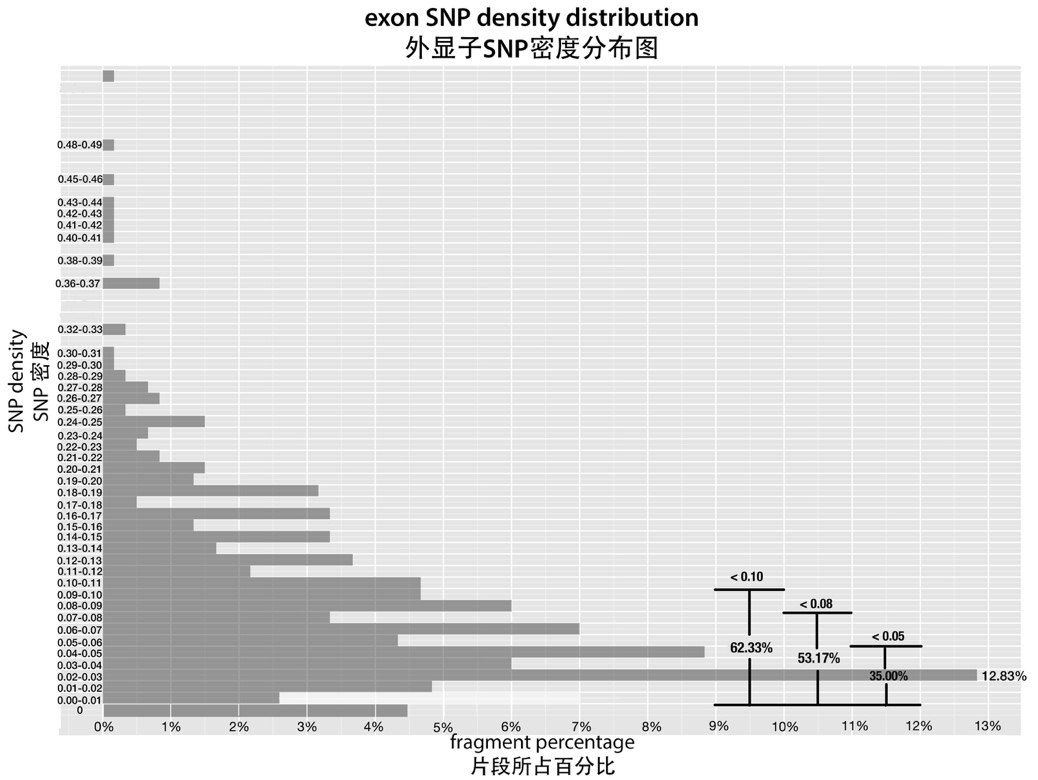
## 水稻全部miRNA的SNP密度分析

SNP密度，也就是SNP数量除以相应基因组区段长度，能够反应该基因组区间的选择压力。其选择压力越大，则其上的SNP密度越低[[16]](#endnote-16)，反之亦然。因为miRNA是功能基因组元件并且是主要调节物，相比于基因间隔区和外显子，它们会经历不同的选择压力[[17]](#endnote-17)。为了研究这个问题，我们在水稻基因组中随机选择了长度为150碱基对长度的基因间隔区 (intergenic region)片段和外显子 (exon)片段作为对照，总共分别选择了600个片段，并且在本地的SNP数据库中对这些片段进行SNP搜索。

正如所预期的，pre-miRNA和外显子上的SNP密度比基因间隔区的更低。对于pre-miRNA，在SNP密度达到0.03-0.04范围之前，基因组片段的百分比随着SNP密度的增加而增加，而在超过0.04之后逐渐下降 (图4-2)。在本研究所取样的外显子也显示了类似的趋势，只是相应的基因组片段百分比最高值是在SNP密度为0.02-0.03区间达到，为12.83% (图4-2)。然而，基因间隔区中则没有展现出类似的趋势 (图4-2)，而是随着SNP密度的增加，不断的“震动”变化，在0.02 – 0.03区间达到最大值，为10.17%。通过分别比较三图中落在0-0.10、 0-0.08 和 0-0.05区间片段的百分比，可以很明显看出落在相应区间的pre-miRNA比例多余外显子，而外显子的比例则多余基因间隔区。这显示了相较于外显子和基因间隔区，pre-miRNA经过了更加严格的进化选择，而这结果也和miRNA在很多调节通路中扮演主要调节因子的角色一致。

其它的研究中也有得到的结果和本实验很类似，在这些研究中，功能性区段比如pre-miRNA特别在种子区域和miRNA结合位点上的SNP非常稀少甚至比3’ 非编码区(UTR)其他保守序列的SNP都少16,17。





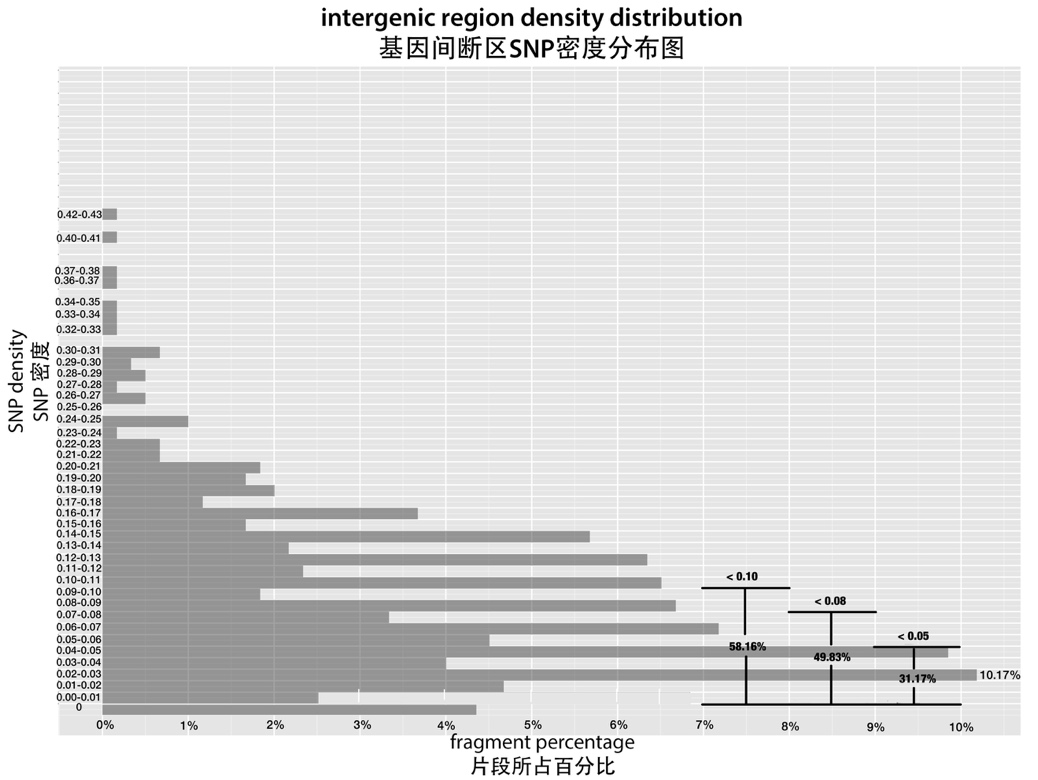


图4-2水稻pre-miRNA、外显子和基因间隔区的SNP密度

注: 其中，SNP密度是SNP的数量除以相应基因组区段的长度。而X坐标则是SNP密度落在相应区间的基因片段百分比，右下方的条形图是SNP密度分别低于0 ~ 0.10, 0 ~ 0.08 和0 ~ 0.05的miRNA比例。

Fig. 4-2 SNP density of pre-miRNAs, exon regions and intergenic regions in rice.

Note: SNP density is the division of SNP number by the length of the genetic region. And x-axis corresponds to the percentage of fragments that have SNP density at the given range. Bar plot on the bottom right shows the percentage of miRNAs whose SNP densities are below 0.10, 0.08 and 0.05, for conserved and non-conserved miRNAs respectively.

由于进化上的保守性不同3,[[18]](#endnote-18),[[19]](#endnote-19)，水稻中保守的miRNA上的SNP密度应该会比非保守的miRNA上的更低。在图4-3中，保守miRNA和非保守miRNA的百分比都是随着SNP密度的增加而上升，两者都在0.03-0.04区间达到最大值，分别是15.1%和8.9%；非保守miRNA最高SNP密度在0.53-0.54，远远高过保守miRNA最高密度。而我们通过分别比较落在0-0.10, 0-0.08 和0-0.05区间的比例，我们可以发现相较于非保守的miRNA，大部分的保守miRNA都聚集在比较低的SNP密度区间中。结果和预期的一样。而保守miRNA和非保守miRNA上SNP密度的差异，甚至比pre-miRNA和基因间断区SNP密度的差异还要大，说明保守miRNA和非保守miRNA上的选择压力相差很大。

之前Liu et al., 2013的研究发现排除一些保守性较差的miRNA，所统计的pre-miRNA上的SNP显著减少，如此也表明在保守miRNA上有更加严格的纯化选择[[20]](#endnote-20)。

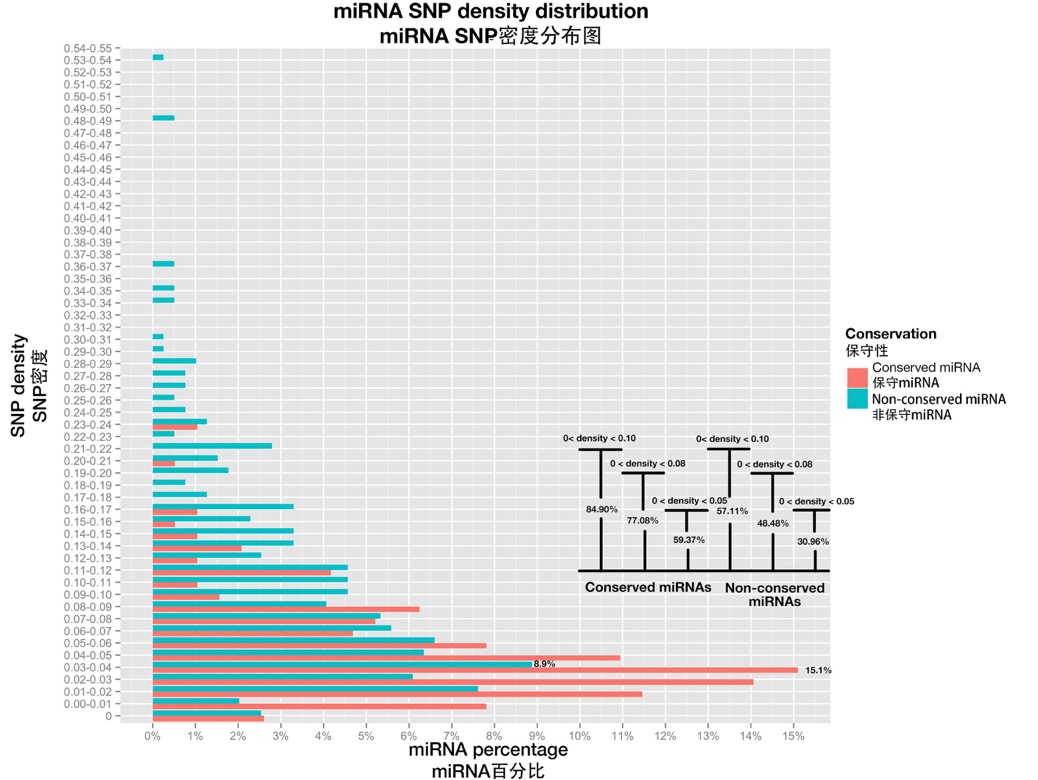


图4-3 pre-miRNA的SNP密度分布图，其中保守miRNA以红色标明，非保守miRNA则以蓝色标明

注: 右下方的条形图是SNP密度分别低于0 ~ 0.10, 0 ~ 0.08 和0 ~ 0.05的miRNA比例。

Fig. 4-3 miRNA SNP density distribution of pre-miRNAs, both conserved miRNAs (in red color) and non-conserved miRNAs (blue color).

Note: Bar plot on the bottom right shows the percentage of miRNAs whose SNP densities are below 0.10, 0.08 and 0.05, for conserved and non-conserved miRNAs respectively.

## 成熟的保守miRNA和非保守miRNA上各位点SNP频率

植物miRNA对不同的靶标基因其调节效率有差异，而这差异归因于它们的互补模式的不同，当然这互补模式是由成熟miRNA和所对应的靶标位点共同决定的。此外，已有证据指明成熟miRNA的21个碱基位点在基因识别和沉默中所起的作用是有差别的8，因为在一些位点上的突变会导致基因沉默调节实效而其它位点上的则不会有明显的影响 17,[[21]](#endnote-21),[[22]](#endnote-22)。这种差异也可以用每个位点的SNP频率的差异性来反应出来（本研究中某位点的SNP频率定义为在该位点出现SNP的miRNA占总的miRNA数量的比例）。更低的SNP频率暗示此位点经受了更大的选择压力。因此，对于成熟miRNA每一个位点进行全体的SNP频率分析可能进一步揭示它们在沉默中所起到的差异性作用。

因为大多数的miRNA都是21 nt的长度，所以本研究主要关注1-21位点上的分析。通过分别计算和比较保守miRNA和非保守miRNA每个位点的SNP频率，我们发现，结果如同预期：保守miRNA上的每一个位点的SNP频率都比非保守的要低。尽管如此，在理论上假如保守miRNA和非保守miRNA都处在相似的进化压力的话，两者在不同位点的SNP频率应该会有相似的秩 (ranking)，并且这相似的秩会导致保守miRNA和非保守miRNA各位点的SNP频率具有线性或者近似线性的关系，这线性关系可以用皮尔森相关性进行检验。然而结果并非和预期相同。位点20是非保守miRNA上SNP频率最高的位点，但在保守miRNA中却是SNP频率第四低的位点，位点12在保守miRNA上是SNP频率最低的位点，但在非保守miRNA上却是频率第二高的位点。同时，皮尔森相关性检验发现两者的各位点SNP频率之间并没有显著的相关性(r=-0.163, p-value=0.4473)。唯一的例外是，两者都在位点一有最低的SNP频率，而这个可能是由位点一在miRNA载入AGO蛋白起到重要作用导致[[23]](#endnote-23),[[24]](#endnote-24)。总结来说，保守miRNA和非保守miRNA上位点SNP频率的不同的秩揭示了两者在各位点经历了不同的选择压力，暗示它们可能使用不同的沉默组份 (silencing component) 来调节靶基因。

通常认为在位点10和11的配对对于植物miRNA的mRNA剪切有非常重要的作用[[25]](#endnote-25),[[26]](#endnote-26),[[27]](#endnote-27)，而这个也对这两个位点在进化上加了一层的限制，并且可能导致它们的SNP频率比其它的位点都要低。然而，我们的结果显示这两个位点在保守miRNA和非保守miRNA中都不是SNP频率最低的位点，这和经验结论并不相符，进而对中心位点（10位和11位）完美配对对植物miRNA介导的沉默必要性的观点提出质疑。

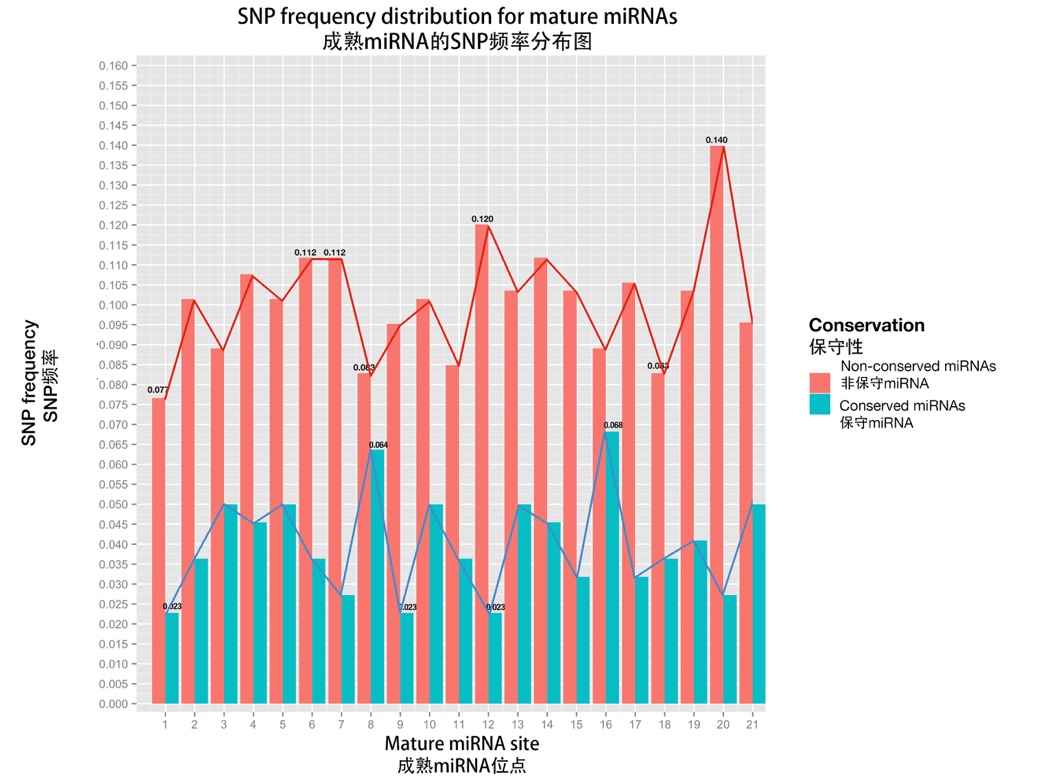


图4-4成熟miRNA每个位点的SNP频率，其中保守miRNA是蓝色而非保守的则是红色

注: X坐标是以成熟miRNA 5’到3’的顺序排列，而Y坐标则是SNP频率。其中标上数字的是SNP频率最高和最低的几个位点

Fig. 4-4 Positional SNP distribution of conserved miRNAs (blue) and non-conserved miRNAs (red).

Note: X-axis is the sites in mature miRNA from 5’end - 3’end, and y-axis is SNP frequency which is calculated as number of SNPs at this site divided by number of miRNAs; the bars marked with digital numbers are those positions that ranks the highest or lowest SNP frequency sites

## 保守miRNA和其靶基因保守位点上SNP频率

下一步，我们尝试分析靶基因上miRNA结合位点上的SNP频率。由于miRNA和靶基因表达量相关性检验发现，它们之间并没有存在必然的负相关，我们就不能使用这种负相关对靶基因进行生物相关性筛选。为此，所有收集而得的823个保守miRNA靶基因都保留进行统计和研究。经过统计和计算在miRNA结合位点上，每个位点的SNP频率，我们得到图4-5。图4-5显示，成熟miRNA的所有位点SNP频率都比miRNA结合位点上的要低，可能由以下因素导致，首先植物保守miRNA通常会调控多个靶基因，从而给成熟miRNA增加更多的约束因而导致SNP频率降低；其次，通过生物信息方法得到的靶基因很可能会有假阳性结果，而这会导致所预测的靶基因不会经历miRNA介导的调节，因而不会有相应的进化压力。

已有研究报导过miRNA和其结合位点存在共同进化23,[[28]](#endnote-28)。两者之间在自然选择的过程中互相影响，不但如此，两者的序列对miRNA的调节作用效率都起到重要的影响。所以，我们推测成熟miRNA和结合位点上各位点的选择压力相似。而这种选择压力的相似性可以表现在两者各位点SNP频率的相关性上。图4-5显示成熟miRNA和结合位点之间SNP频率分布之间的确有相似性，位点1在miRNA结合位点上是SNP频率第二低的位点，同时在成熟miRNA上是频率最低的位点；而位点13在miRNA结合位点上是频率最高的位点，在成熟miRNA上则是频率第三高的位点。通过皮尔森相关性检验，也发现两者之间有显著的正相关 (r=0.5891, p-value=2.455e-3) 。所得的结果也支持了关于成熟miRNA和miRNA结合位点之间有共同进化关系的假设。

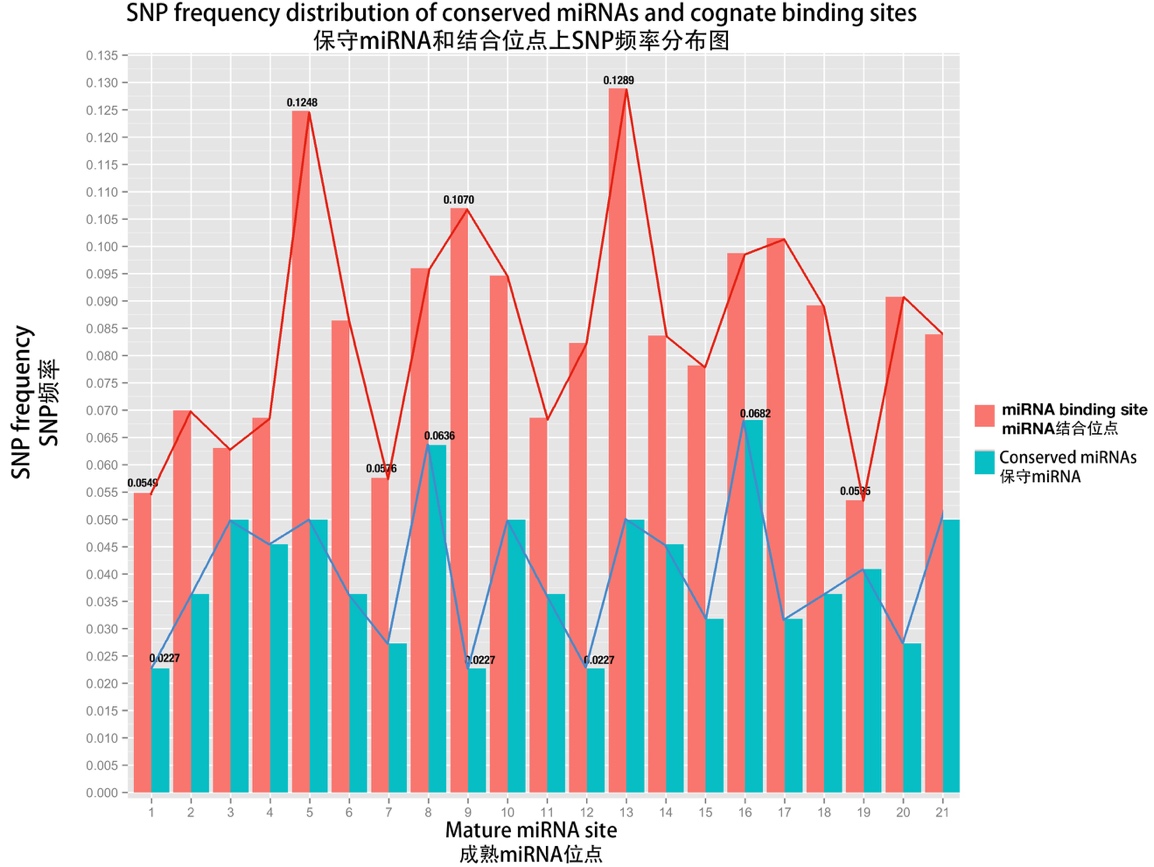


图4-5成熟miRNA和其结合位点上每一个位置的SNP频率分布

注: 结合位点上位置的排列顺序和成熟miRNA的顺序相同，都是按照成熟miRNA 5’到3’的顺序排列，红色表示的是miRNA结合位点，而蓝色表示的则是保守的成熟miRNA

Fig. 4-5 The sites of miRNA binding site are placed in the same order as mature miRNAs Note: The sites of miRNA binding site are placed in the same order as mature miRNAs (from 5’end to 3’end in the miRNA); the blue bars stand for the miRNA binding sites, while red bars stand for conserved mature miRNAs

## 本章小节

本章对实验过程中所得到的SNP进行统计上的分析。首先比较了不同的SNP数据库，并且发现三者之间有很多的SNP是重叠的，这说明这些测序工作所得到的SNP可信度是很高的。对比pre-miRNA、基因间断区和外显子片段上SNP的密度，就会发现pre-miRNA的SNP密度比这两者都要低，再次说明miRNA作为很多调节通路中的主要调节因子受到很大的选择压力。另外，分别对比保守miRNA和非保守miRNA，以及保守miRNA和miRNA结合位点上的SNP频率的相关性，发现保守miRNA和miRNA结合位点上各位点SNP频率有显著的相关性，但是和非保守miRNA却不具有显著的相关性。

# 联合互补模式分析

## 水稻miRNA单倍型分析

单倍型(haplotype, haploid genotype)，也就是一段基因片段上特定突变的集合，经常被用在疾病相关研究中，是其中强大的辨别病例组和对照组的工具。因此，单倍型可以在研究某一段基因片段功能室，帮助将群体根据其基因型不同区分成不同组别。Zhao et al. 就展示了应用单倍型分析的例子15：针对水稻的杂种不育性，粳稻 (*japonica*) 和籼稻 (*indica*) 之间存在杂交不育。而*S5*位点是被鉴定和两者之间杂交不育有很紧密的关系，其中编码了3个连锁基因；他们用其中的一个基因，LOC\_Os06g11010，也就是配子“杀手”，进行单倍型分析。在这个基因中找到14个SNP，其中5个SNP是非同义突变 (non-synonymous mutation)。将这5个SNP放在一起，分析其单倍型，并且根据在这些SNP位点上的碱基类型将水稻品系分成了5个组别。最后发现，几乎所有*aus*品系都集中在一个组别，而*aus*品系是可以克服杂交不育的品系。

而如图5-1所示，通常的利用SNP作分子标记的单倍型分析的过程，是先得到某一片段上所有SNP，每个SNP都有核苷酸多态性，例如在图中SNP1有TC的多态性，SNP2有GA的多态性，SNP3有AT的多态性；然后将群体分成不同组别，每个组别在这些SNP位点的碱基类型相同但是组间是不同的，例如在图中分成三个单倍型，分别是TAA、CGA和CGT，其中TAA是参考基因组模式(reference pattern)，通过归纳不同的水稻品系在这些SNP位点的核苷酸类型会发现Cultivar1, Cultivar2属于参考基因组模式，而Cultivar3属于CGA，Cultivar4属于CGT；最后，比较不同组别，也就是这三个单倍型对应的组别，的待研究特性，从而得出这些特性和相应的SNP是否有关联。具体的定义和操作流程请见附录1。

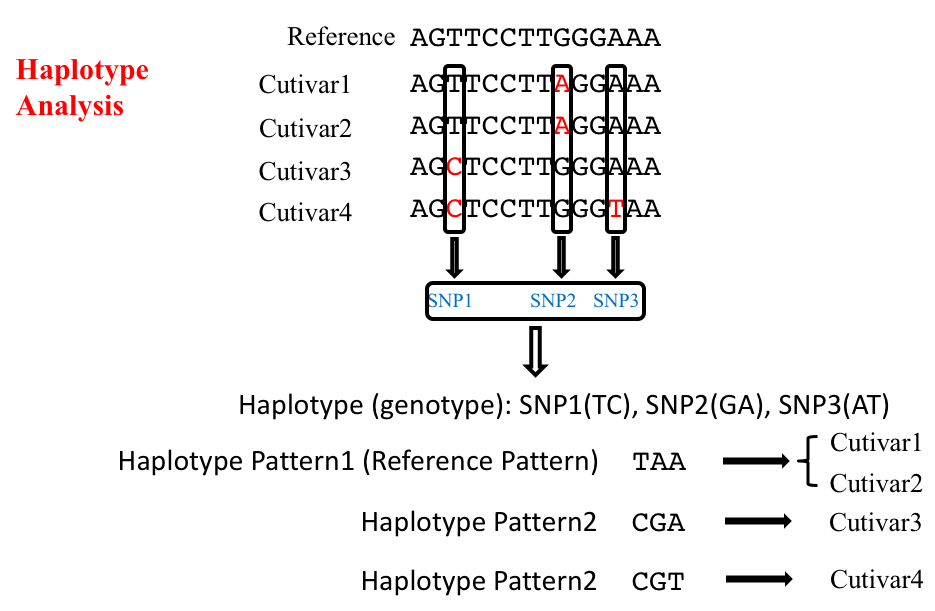


图5-1 单倍型分析图解

Fig. 5-1 Illustration of haplotype analysis

全部585个pre-miRNA中570个上面有SNP，而另外15个没有SNP的pre-miRNA中，osa-MIR156f, osa-MIR156h, osa-MIR396e, osa-MIR397b和osa-MIR528都是属于保守miRNA。图5-1显示的是所有pre-miRNA上SNP数量的情况，其变化趋势是：当SNP数量从0增加到5时，相应pre-miRNA数目不断增加；当SNP数量达到5时，pre-miRNA数目达到最高值，也就是47个；而超过5之后，则呈现下降的趋势。所有pre-miRNA中，SNP数量最多的是osa-miR7695。Pre-miRNA主要集中在0-10个SNP的区域，达到总数的57.95%，另外有部分pre-miRNA上发现了大量的SNP，其中甚至有10个pre-miRNA发现有50个及以上的SNP，而因此pre-miRNA上的SNP平均密度相对很大，平均每个pre-miRNA上有12.3个SNP。以上的数据表明，我们使用SNP-seek database的数据，发现了远远超过Liu Q.P. et al. 研究中的SNP。在他们的研究中，pre-miRNA上平均SNP数量是5.8，并且拥有最多SNP的pre-miRNA是osa-miR2124b，其数量是37。因此表明了，3K水稻基因组测序的确得到了极其丰富的SNP数据。

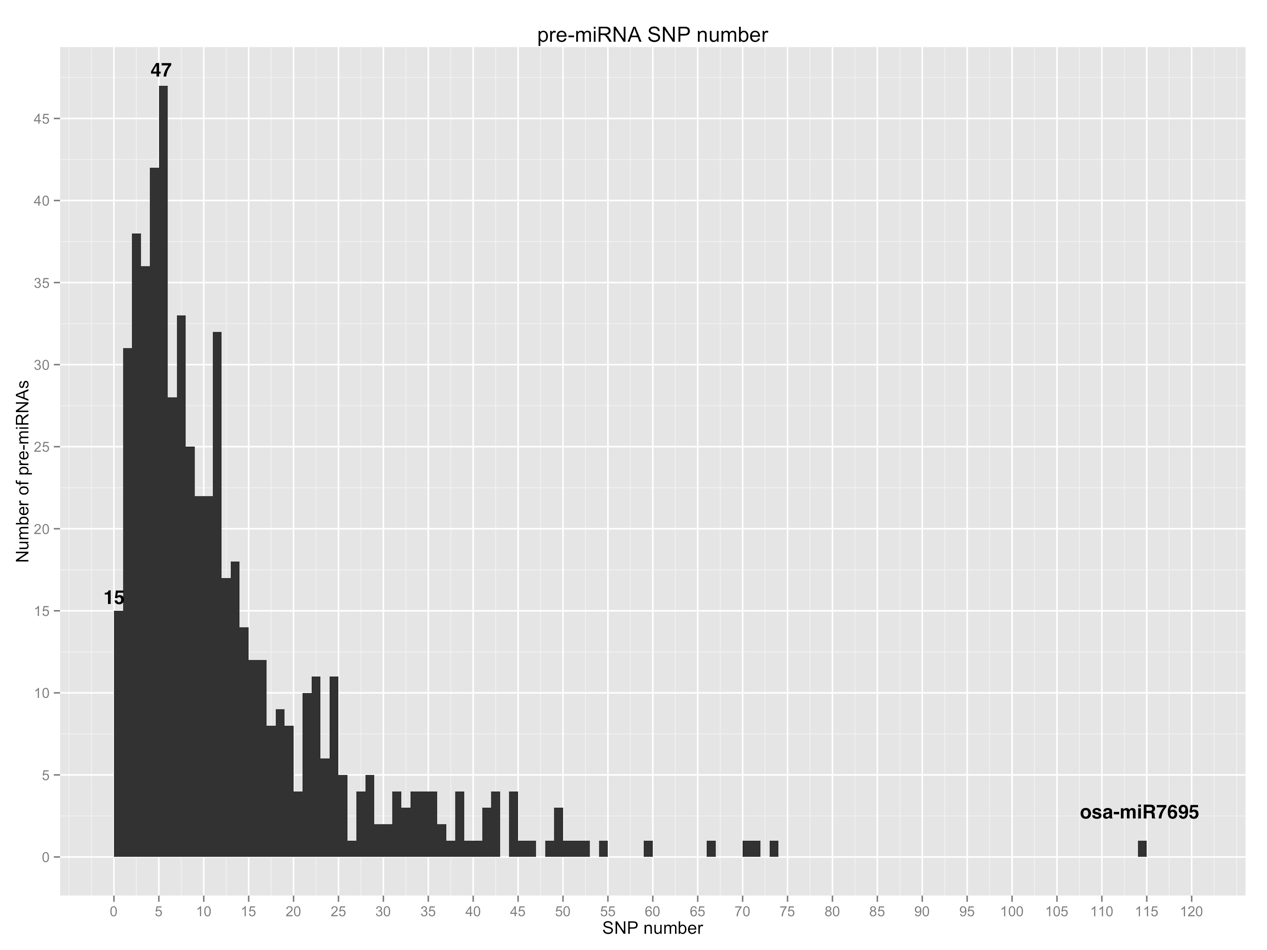


图5-2 所有pre-miRNA上的SNP数量分布图

注：X坐标是pre-miRNA上SNP的数量，而Y坐标是拥有相应SNP数量的pre-miRNA总数

Fig. 5-2 Distribution of SNP numbers of all pre-miRNAs

Note: X axis corresponds to the number of SNPs, while Y axis corresponds to the number of pre-miRNAs that have the given SNP number

将单倍型分析分别应用在这570个含有SNP的pre-miRNA，我们得到了21549个针对不同miRNA的单倍型，平均每个pre-miRNA有37.8个单倍型。导致每个单倍型数量超过SNP数量的原因是，每个pre-miRNA上所有SNP位点不同的等位碱基(alleles)在各种水稻品系中有不同的组合。

## pre-miRNA二级结构分析

植物miRNA在生成和成熟的过程中，会经历两步剪切的阶段：首先是从初级转录本 pri-miRNA (primary miRNA)会被DCL1剪切成pre-miRNA而形成“发夹结构”(hair-pin structure)，然后再次被DCL1剪切从而形成miRNA:miRNA\*二聚体。这两个剪切过程对于成熟miRNA的形成至关重要，不但会影响到成熟miRNA的产量，还可能会影响到最终产物的序列。其中DCL1的精确剪切和pre-miRNA的二级结构，也就是“发夹结构”有密切关系，如果发夹结构比较稳定，则能产生更多的成熟miRNA。有研究报导，当发夹结构的能量变化达到0.3 kcal/mol时，就会影响到成熟miRNA的形成[[29]](#endnote-29)。所以，我们对得到的单倍型进行分析，观察这些单倍型的突变对pre-miRNA的发夹结构能量产生的影响。我们使用RNAfold程序[[30]](#endnote-30)来预测pre-miRNA的发夹结构并且计算了不同的单倍型相对于参考基因组模式（reference pattern, 也就是和参考基因组基因型一样的单倍体模式）的能量变化(∆∆G)。

我们发现，54.7% (11795 / 21549)的单倍型其发夹结构从稳定状态变成不稳定状态，其能量改变范围(∆∆G)是0.3 ~ 39.3kcal/mol，其中能量变化最大的是osa-miR1880 (∆∆G = 39.3kcal/mol)，它上面总共有10个突变，对应1个水稻品系(B227, 闷加高1，源自中国)；另外1215个单倍型并未造成明显的发夹结构能量的变化它们的能量变化是-0.2 ~ 0.2 kcal/mol，其中有583个单倍型，虽然pre-miRNA上面有突变，却没有引起发夹结构能量的变化，其中甚至在osa-miR2872上面发现28个突变，但是能量变化为0 kcal/mol，这个单倍体对应一个水稻品系（IRIS 313-10522, SUKAMANDI 1005，源自印度尼西亚）；最后，有36.8% (7921 / 21549)的单倍型其发夹结构变得更加稳定，能量变化范围是-27.5 ~ -0.3 kcal/mol，其中能量变化最大的是osa-miR5542，其上总共有16个突变，能量变化为∆∆G = -27.5 kcal/mol，这个单倍型对应的水稻品系有三个：CX141, Padi Siam Kuning，源自马来西亚；IRIS 313-10786, KETAN URANG，源自印度尼西亚；IRIS 313-9188，TUMBA，源自印度尼西亚。

图5-2中，(a) osa-miR1880上的10个SNP将该pre-miRNA发夹结构中原本互补性很强的茎部破坏，产生了很多的凸出(bulge)，甚至已经不再是简单的发夹结构（茎环结构, stem），所以很可能会阻碍DCL1的识别和剪切，从而降低成熟miRNA的产量；(b) osa-miR2872上28个SNP， 一部分出现在环部 (loop)，出现在环部的SNP对发夹结构的能量影响不大，而其茎部的SNP则既有形成配对增加其稳定性的，也有产生小的凸出降低稳定性，两者有互相抵消的作用；(c) osa-miR5542上16个SNP，其中很多在茎部促进碱基互补，使pre-miRNA的稳定性大幅增加。

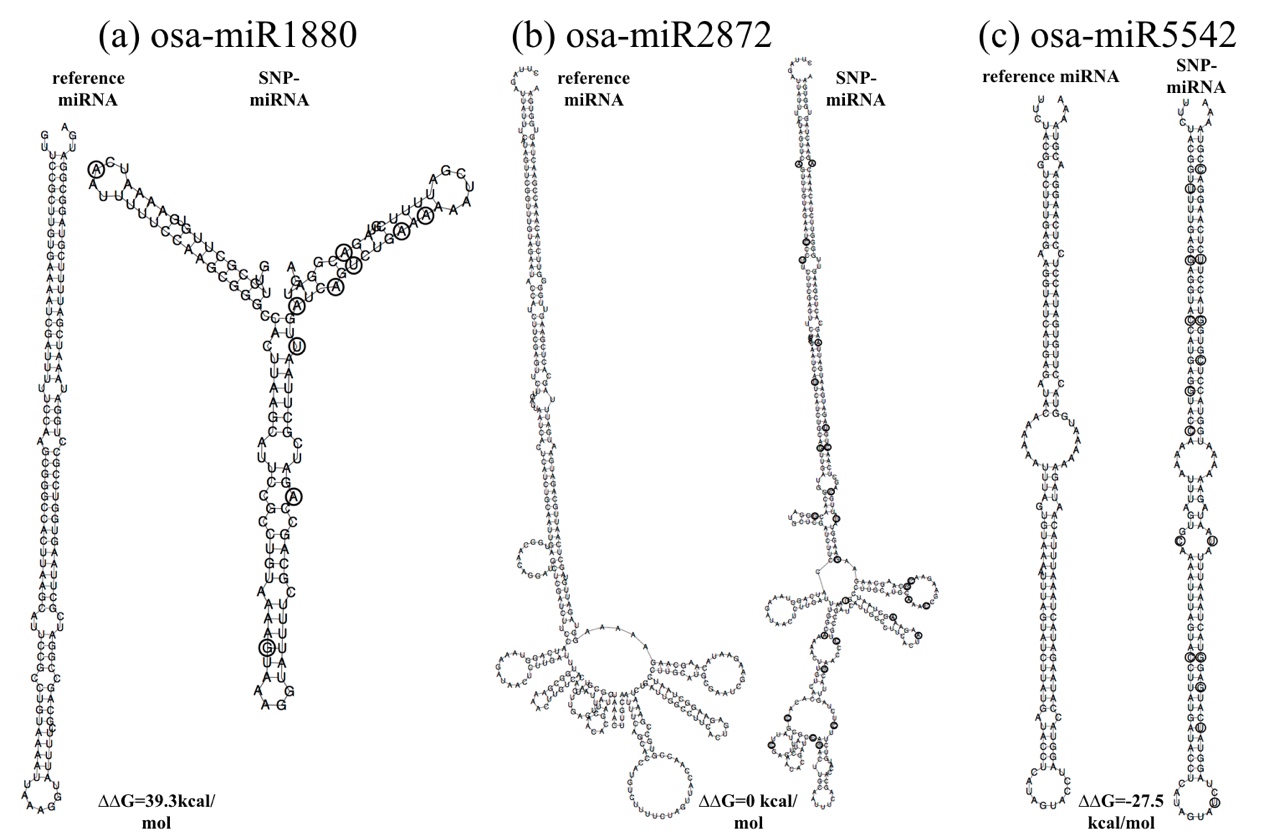


图5-3 osa-miR1880 (a), osa-miR2872 (b), osa-miR5542 (c)的参考基因组模式和单倍体模式的二级结构预测图

注：其中用圆圈圈住的是SNP的位置，这三个miRNA是展示发夹结构因为其上的SNP稳定性下降（∆∆G = 39.3 kcal/mol），稳定性保持不变（∆∆G = 0 kcal/mol）和稳定性增加（∆∆G = -27.5 kcal/mol）的范例。

Fig. 5-3 Illustration of predicted secondary structures for reference and SNP- osa-miR1880 (a), osa-miR2872 (b), and osa-miR5542(c)

Note: The circle marks the SNP position. The three miRNAs are shown as typical examples for displaying the decreased (∆∆G = 39.3 kcal/mol), unchangeable (∆∆G = 0.00 kcal/mol), and increased stability of hairpin structures (∆∆G = -27.5 kcal/mol) by SNP variations.

其中，保守miRNA的二级结构的变化则相对较小，能量变化范围是-8.1 ~ 11.2 kcal/mol，但仍然有73.4% (799 / 1088)的单倍型能量变化绝对值|∆∆G|超过0.3，其中osa-miR399h的一个单倍型的pre-miRNA发夹结构变得不稳定，∆∆G = 11.2 kcal/mol，其相对应的水稻品系是IRIS 313-11415，KAUSHI，源自尼日利亚；而osa-miR156g的一个单倍型的发夹结构变得更加稳定，∆∆G = -8.1 kcal/mol，其对应的水稻品系是IRIS 313-8502，M 102，源自美国。

从图5-3的预测二级结构可以看出SNP对这两个miRNA的影响。(a) osa-miR399h上的三个SNP，都落在茎部，分别是：将茎部配对的UA变成CA造成凸出(bulge)结构变大，将原本配对的GC变成AC导致了一个凸出结构产生以及将GC变成GU使得两者的结合程度降低。以上的SNP带来的影响都是降低发夹结构的稳定性。(b) osa-miR156g上的两个SNP也都是出现在茎部，将原本没有配对的AC变成GC去掉了一个凸出结构，另一个则是GG变成CG将所在的凸出结构减小，因此两者都起到增加茎部互补性的作用，从而增强了发夹结构的稳定性。

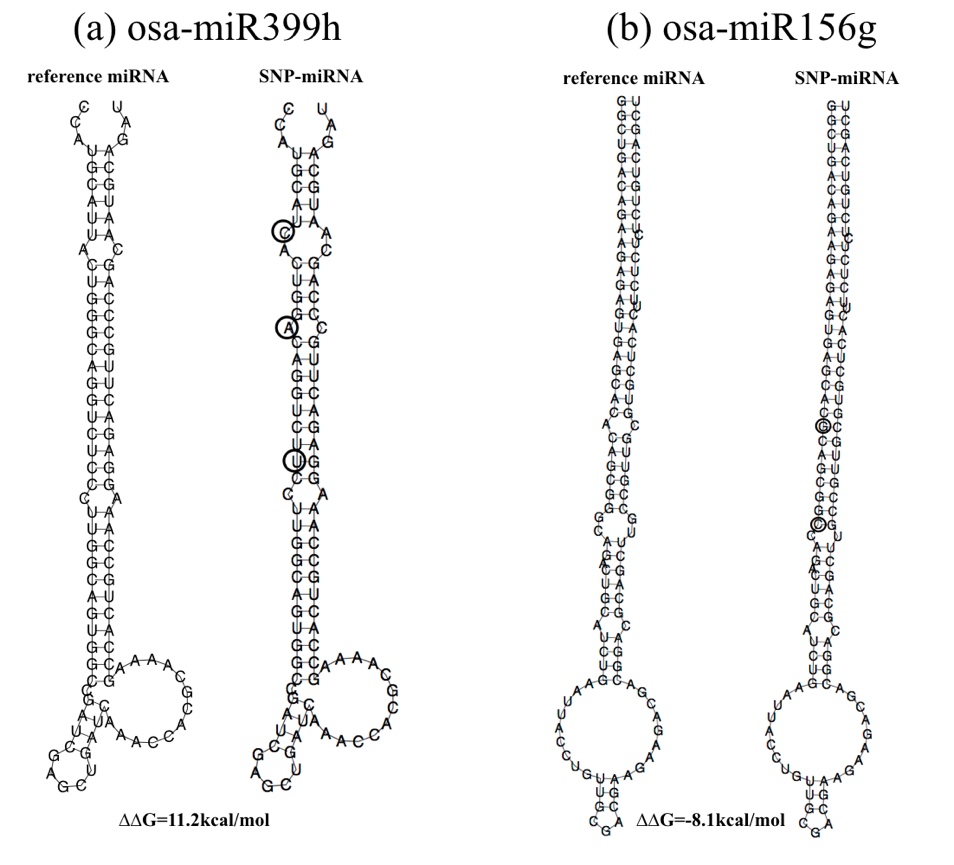


图5-4 保守miRNA中，osa-miR399h (a), osa-miR156 (b) 的参考基因组模式和单倍体模式的二级结构预测图

注：其中用圆圈圈住的是SNP的位置，这两个保守miRNA是展示发夹结构因为其上的SNP稳定性下降（∆∆G = 11.2 kcal/mol）， 和稳定性增加（∆∆G = -8.1 kcal/mol）的范例。

Fig. 5-4 Illustration of predicted secondary structures for reference and SNP- osa-miR399h (a), osa-miR156 (b)

Note: The circle marks the SNP position. The two conserved miRNAs are shown as typical examples for displaying the decreased (∆∆G = 11.2 kcal/mol), and increased stability of hairpin structures (∆∆G = -8.1 kcal/mol) by SNP variations.

对比Liu Q.P. et al.的研究，他们分别计算了单个SNP对于所在pre-miRNA发夹结构稳定性的影响，因而能量变化范围较小 (-10.0 ~ 10.1 kcal/mol)。在本研究中，则是通过单倍体分析，找到在3024种水稻品系中真实发生的突变对pre-miRNA的影响，其中83.0%（17878 / 21549）的单倍型都是两个以上的突变，这也是导致所预测的能量变化范围增大的原因。

## miRNA和靶基因结合位点互补模式(complementarity pattern)研究

植物miRNA对靶基因的识别机制很复杂，不但对互补性(complementarity)有很高的要求，而且其中错配(mismatch)出现的位置对于靶基因的识别和结合也有很大的影响。虽然经验结论告诉我们，植物miRNA和靶基因有很高的互补性，但是实验研究中，也发现植物miRNA也能调控一些错配数量很多的靶基因，比如osa-miR319a能够以*TCP*家族的5个成员作为靶基因，但是它们和osa-miR319a之间却有5个错配，如果将G:U算为0.5个的话，也有4个错配11。为此，在2005年就有研究人员用表达谱在拟南芥中25，对miRNA的识别机制进行研究，发现的经验参数是：

1. 在位点十和位点十一位，不能有错配；
2. 在位点2 ~ 12位之间不能有超过一个的错配；
3. 在位点13及下游，不能有超过两个的连续错配 (consecutive mismatches)；
4. 至少有完全匹配的靶基因自由能的72%；

而2014年，Liu Q.K. et al.则通过研究烟草（Nicotiana benthamiana）的定量过渡相试验，发现在5’区段 (位点2 ~ 12)发生的错配会导致植物miRNA调节效率降低；相反，在3’ (位点13 ~ 21)区段发生的错配甚至可能比完全匹配的调节效率还要高8。

考虑到非保守miRNA并没有很多经验证的靶基因，而保守miRNA则有较多的经验证的靶基因。为了研究在水稻中，miRNA对靶基因的识别机制，我们对所得到的水稻保守miRNA和相应的靶基因的互补模式进行分析。我们收集了193个保守miRNA和对应的616个靶基因，总共有1464个miRNA和靶基因作用对。将其中的经降解组验证(Degradome Validated, DV)的和未经降解(Predicted but Not Degradome Validated, PNDV)组验证的区分，并且分别将其成熟miRNA和结合位点的序列排列在一起，并且统计两组在各位点的匹配情况。其中，DV组有427对，而PNDV组则有1037对，见表5-1。

DV组是经过降解组验证的miRNA和靶基因作用对，所以是真正具有识别和调节的作用对，然而PNDV则未经降解组验证作为参照组。图5-1中，可以看出DV组在第三位是没有错配的，相比之下，在PNDV组有11.62%的错配率；另外DV组在第一，十四和二十位分别有38.06%，49.88%，40.63%的错配率，远远大于PNDV组相应位点的错配率。以上说明了第三位在靶基因识别中对于错配的容忍率很低；而在14位和20位处在13位的下游区段，根据Liu Q.K. et al.的结论，在3’区段的错配甚至会对miRNA的调节效率有积极作用，所以这结果也部分支持了这个结论；至于位点一的高错配率则说明了位点一的匹配对于miRNA的识别并没有起到很大的影响。让人惊讶的是，DV组在10位和11位上存在错配的现象，这可能是miRNA家族中一些并不具有调节能力的成员带来的错配。

表5-1 包含全部成员的DV组和PNDV组在1 ~ 21位点的错配率

注：其中G:U算作0.5个错配，错配率是在该位点错配的miRNA和靶基因作用对占所有作用对的百分比

Table 5-1 Mismatches rate at Position 1 ~ 21 of both DV and PNDV groups, all miRNA family members included

Note: G:U is counted as 0.5 mismatches, and mismatch rate is the percentage of miRNA:target interaction pairs that are mismatched at the particular position

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 位点 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| DV组错配率(%) | 38.06 | 8.43 | 0.00 | 0.94 | 0.70 | 1.05 | 5.74 | 0.47 | 5.04 | 0.70 | 1.06 |
| PNDV组错配率(%) | 8.73 | 9.26 | 11.62 | 7.43 | 13.41 | 11.72 | 6.61 | 6.89 | 12.97 | 8.44 | 7.96 |
| 位点 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |  |
| DV组错配率(%) | 5.16 | 2.46 | 49.88 | 7.84 | 2.34 | 12.88 | 4.22 | 6.80 | 40.63 | 5.39 |  |
| PNDV组错配率(%) | 11.15 | 6.37 | 20.16 | 12.40 | 9.84 | 10.61 | 14.28 | 6.42 | 13.45 | 9.98 |  |

考虑到即使在同一个miRNA家族，也不是所有的成员都能够对靶基因起到调控作用，我们另外对每一个miRNA家族分别取其中的正规成熟miRNA (canonical miRNA)，计算每个位点的错配率，如图5-2，这次DV组总共有54对，而PNDV组则有420对。首先，在这次的实验中，我们发现在位点三、四、十、十一、十六上，错配率都是0，当然同时位点一、十四和二十的错配率仍然维持在较高的水平。所以对比表5-1和表5-2，可以发现，正规成熟miRNA中在位点十和位点十一中并没有出现错配，说明错配很可能是miRNA家族中不能调控该靶基因的成员导致的。

同时在这两次实验中，对比3’区段和5’区段，发现3’区段的错配率比5’区段的错配率要高，说明3’区段对错配的容忍度高于5’区段，这也再次支持了Liu Q.K. et al.的结论，也就是5’区段 (位点2 ~ 12)发生的错配会导致植物miRNA调节效率降低；相反，在3’ (位点13 ~ 21)区段发生的错配甚至可能比完全匹配的调节效率还要高。

表5-2 每个家族只使用正规成熟miRNA (canonical miRNA)作为代表的DV组和PNDV组在1 ~ 21位点的错配率

注：其中G:U算作0.5个错配，错配率是在该位点错配的miRNA和靶基因作用对占所有作用对的百分比

Table 5-2 Percentage of mismatches at Position 1 ~ 21 of both DV and PNDV groups, only canonical mature miRNAs included

Note: G:U is counted as 0.5 mismatches, and mismatch rate is the percentage of miRNA:target interaction pairs that are mismatched at the particular position

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 位点 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| DV组错配率(%) | 30.55 | 13.89 | 0.00 | 0.00 | 1.85 | 1.85 | 7.41 | 0.93 | 3.70 | 0.00 | 0.00 |
| PNDV组错配率(%) | 9.65 | 13.45 | 10.12 | 9.05 | 8.70 | 17.50 | 7.98 | 9.29 | 15.13 | 14.05 | 6.31 |
| 位点 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |  |
| DV组错配率(%) | 3.70 | 3.70 | 42.59 | 9.26 | 0.00 | 9.27 | 6.49 | 11.11 | 28.70 | 12.04 |  |
| PNDV组错配率(%) | 10.96 | 7.15 | 23.93 | 9.05 | 6.19 | 8.33 | 10.36 | 6.91 | 13.81 | 5.83 |  |

## 联合互补模式分析

### 联合互补模式分析方法介绍

单倍型分析适用于分析单个基因组片段上的SNP从而将群体分成具有不同单倍型的类型，因此很适合分析pre-miRNA上的突变。但是当研究SNP对成熟miRNA和靶基因结合位点的相互作用，则需要同时考虑两段序列上的SNP，并且还需要考虑单倍型中的突变对互补模式(complementarity pattern)的影响。所以，我们在本研究中，将单倍型分析拓展成联合互补模式分析(Combined Complementarity Pattern Analysis, CCPA)。

图5-5具体展示了联合互补模式分析的操作流程。首先得到成熟miRNA和靶基因结合位点上的SNP，将两个序列上的SNP按照顺序排列在一起，然后进行上述的单倍型分析，就如同图5-5中miR1和gene1两个序列合成一个序列进行分析。经过单倍型分析后，我们可以根据水稻品系在所有的SNP位点上的核苷酸类型分成不同的单倍型类别，从而得到如图的参考基因组模式和单倍体模式1。在此之后，根据单倍体模式1中和参考基因组模式不同的SNP，也就是所谓的突变，找到SNP在原序列中所在位置并且进行突变，最后将突变后的两条序列（成熟miRNA序列和结合位点序列）比对在一起称为突变后的互补模式。至此，我们可以得到不同的互补模式以及其所对应的水稻品系，然后比较该单倍型的互补模式和参考基因组单倍体的互补模式，我们就可以很明显看出这些SNP对于miRNA和结合位点相互作用的影响。并且比较各中水稻品系待研究特性，从而得出这些特性和相应的SNP是否有关联。

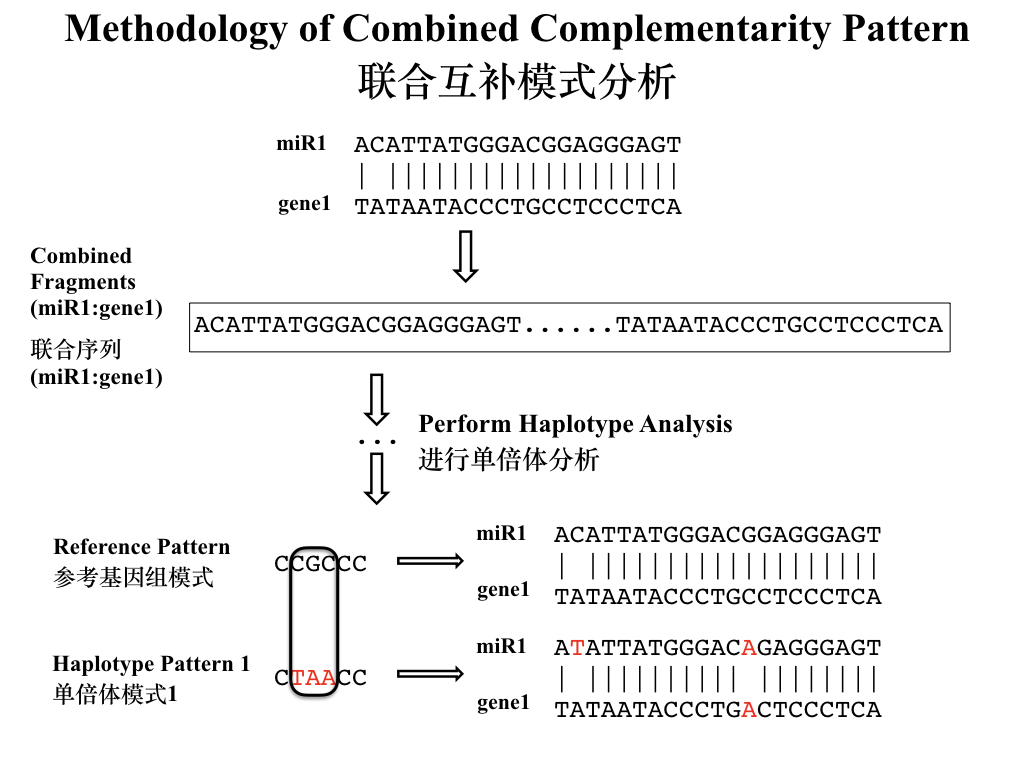


图5-5 联合互补模式分析的流程图解

Fig. 5-5 Illustrations of operation procedures for Combined Complementarity Pattern Analysis (CCPA)

将联合互补模式分析应用在所有220个保守miRNA和823个相应靶基因的结合位点上，也就是2113个miRNA和靶基因作用对，我们总共得到了16867个不同的单倍型，所以平均每一对都有7.98个单倍型。

更深入分析所得到的单倍型产生的突变互补模式，可以发现miRNA和靶基因的互补模式中的位点在SNP的影响下会发生四种改变，分别是：类别1，从配对变成错配；类别2，从错配变成配对；类别3，从配对变成配对(保持配对)和类别4，从错配变成错配（依旧错配）。类别3和类别4并不改变miRNA和结合位点的互补性，所以相对于另外两种类别对miRNA调控的影响相对较小。而类别1会引入新的错配，一般而言会减弱miRNA的调控，类别2则相反，会增强miRNA的调控。将CCPA应用在这823个靶基因和相应的miRNA上，发现74.62%的突变位点是类别1突变，16.67%是类别2突变，0.68%是类别3突变，8.03%是类别4突变。如图5-6所示，左上方是类别4的例子，在osa-miR444b.1和LOC\_Os02g13420作用对中，miRNA序列上第19位发生突变，从G变成A，但是原来在该位点依旧保持错配；右上方则是类别2的例子，在osa-miR396e-5p和LOC\_Os04g24190作用对中，靶基因上结合位点第19位发生突变，从U变成G，结果使原来的错配变成了配对；左下方的是类别1的例子，在osa-miR156c-5p和LOC\_Os05g48800作用对中，miRNA序列中的第13位发生突变，从G变成A，导致该位点原本的配对变成错配；最后右下方的是类别3的例子，在osa-miR1436和LOC\_Os11g31450作用对中，miRNA序列中的第二位从C变成U，同时在靶基因上结合位点的第2位也发生突变从G变成A，两次突变的结果是在该位点的配对状况并没有被破坏，依旧保留了配对。

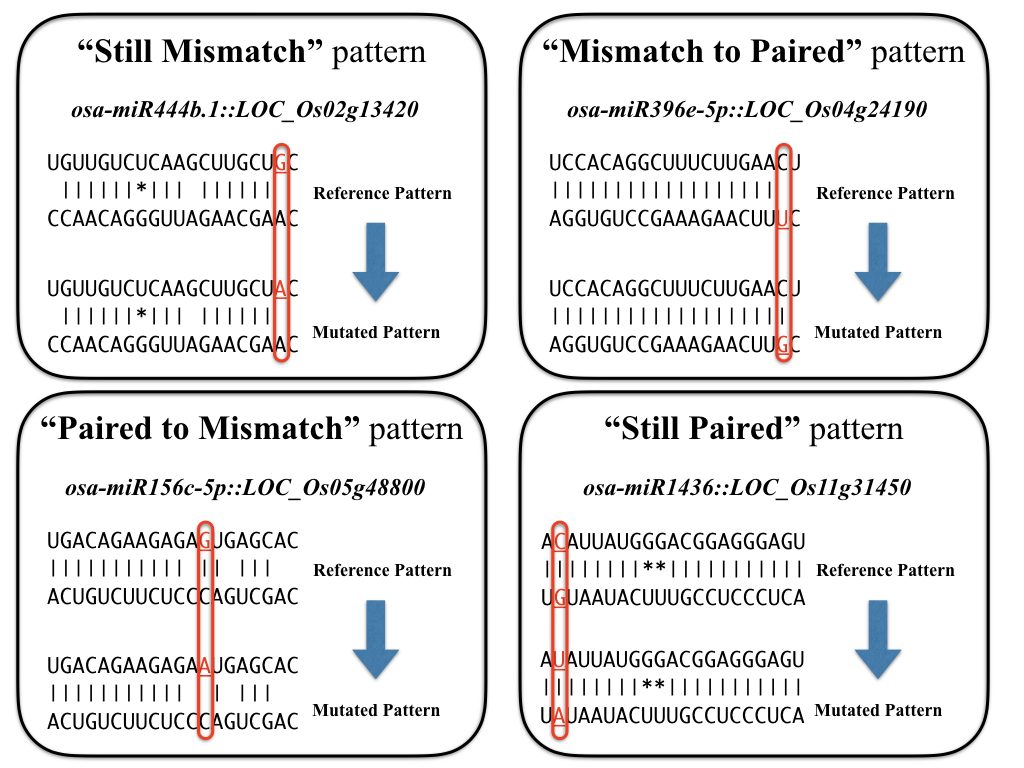


图5-6 互补模式在SNP影响下可能产生的四种互补性变化图例

注：左上方的是类别4（从错配到错配）、右上方的是类别2（从错配到配对）、左下方的是类别1（从配对到错配）、右下方的是类别3（从配对到配对）；每个范例中，上面的互补模式是参考基因组模式，而下面的互补模式则是突变之后的单倍体基因组模式；其中每个互补模式中，上面的序列是miRNA的序列，而下面的序列是靶基因上miRNA结合位点的序列；

Fig. 5-6 Examples of the four classes of changes to the complementarity caused by SNPs

Note: Upper left stands for class 4 (from mismatch to mismatch, still mismatch), upper right stands for class 2 (from mismatch to paired), lower left stands for class 1 (from paired to mismatch) and lower right stands for class 3 (from paired to paired, still paired); in each case, the upper complementarity pattern corresponds to the reference pattern, while the lower one corresponds to mutated haplotype pattern; for each complementarity pattern, the upper sequence corresponds to the sequence of mature miRNA and the lower sequence corresponds to the sequence of miRNA binding site of the cognate target

### 互补性恢复现象和osa-miR818家族

类别3的变化是在引入突变之后，相应位点的互补性仍然保持配对。这种突变发生的概率很低，正如统计的结果看出只有0.68%，其原因是发生这种现象一般需要在成熟miRNA和结合位点两条序列上相对原本是配对的位点分别引入一个SNP，而且突变之后仍旧是配对，这导致发生的概率变的很低。在本研究中，这类突变被称为互补性恢复现象(complementarity recovery phenomenon)。这种现象发生可能经过较复杂的进化过程。在以下的miRNA上我们发现了这种现象：osa-miR1436、osa-miR1439、osa-miR1442、osa-miR1862b、osa-miR2275c、osa-miR444a-3p.1、osa-miR444a-3p.2、osa-miR444b.1、osa-miR444b.2、osa-miR444d.1、osa-miR818a-e。

更深入的研究发现，在osa-miR444家族发现互补性恢复现象，因为这个miRNA家族属于一种特殊的miRNA，也就是天然反义miRNA (natural antisense miRNA, nat-miRNA)。这类miRNA能产生和其靶基因完全配对的成熟miRNA，因为miRNA基因和靶基因在同一段DNA相反的链上2，所以这种互补性恢复现象可以由两者在相反链的特性来解释。其中osa-miR444a-3p.1、osa-miR444a-3p.2在其靶基因*OsMADs23*的相反的链上，而osa-miR444b.1、osa-miR444b.2则是在其靶基因*OsMADs27*的相反的链上，osa-miR444d.1是在起靶基因*OsMADs57*的相反的链上。图5-7的(a)和(c)是其中两个例子，对于(a)中，miRNA和其结合位点的第二位都发生了突变从GC变成UA，而在(c)中，miRNA和其结合位点在第五位和第十一位分别发生了突变，分别从UA变成CG和从AU变成GC。但是实际研究中发现(a)中miRNA和其结合位点两的突变其实只是由SNP 10222300448造成的，(c)中，位点五和位点十二则只是由SNP 10822300429，10822300449造成。

然而，对于其余的osa-miR818a-e、osa-miR1436、osa-miR1439、osa-miR1442和osa-miR1862b上，我们也发现了互补性恢复现象，然而它们发生这种现象则需要在miRNA和其靶基因上就需要引入两个SNP在特定的位置才能造成互补性恢复的现象。并且根据miRBase的miRNA家族分类，这些miRNA由于发夹结构系列相似而归类为同一个miRNA家族，也就是miR818家族。如图5-7中的(b)和(d)，两个例子中互补性恢复现象都发生在位点五，并且都是从CG变成UA，但是它们是分别在miRNA上和miRNA结合位点上的SNP（总共两个SNP）同时突变导致的。

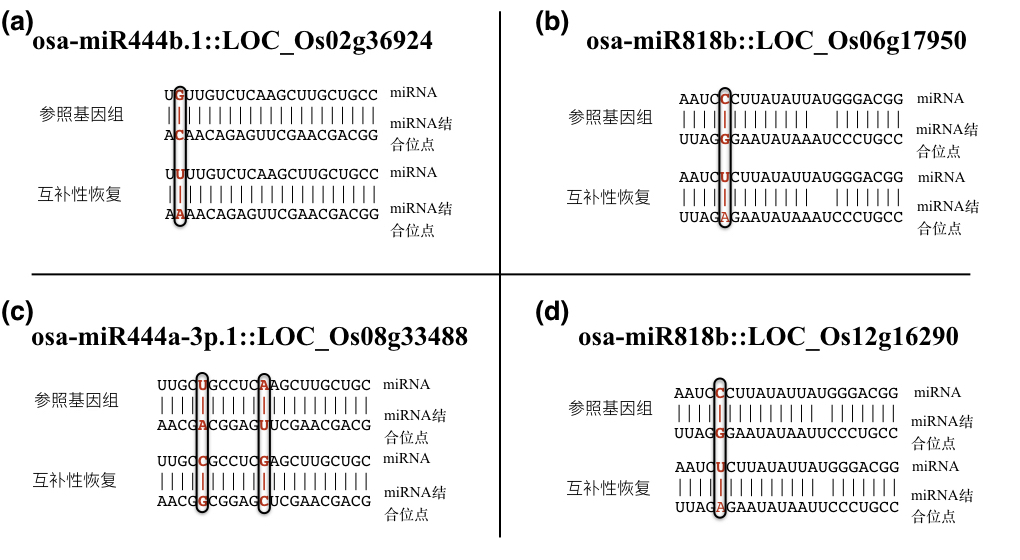


图5-7 互补性恢复模式图

注：图中，竖线代表配对，空格代表错配

Fig. 5-7 Examples of complementarity recovery patterns

Note: In the complementarity pattern, a vertical bar would be placed denoting match, a blank denotes mismatch.

(继续描述osa-miR818家族……)

## miRNA家族和靶基因联合分析

# 单倍体表型分析和对比

# 结束语

## 主要工作与创新点

本文主要……

## 后续研究工作

更深入的研究……

参 考 文 献

[1]

附录1 miRNA单倍型分析

* 名词和定义
  + miRNA单倍型 (miRNA haplotype)：使用SNP作为生物标记，对于每个pre-miRNA，将在其中找到的SNP按照上升的次序排列就得到miRNA单倍型；例如osa-miR156a的miRNA单倍型就是，10122524158, 10122524159, 10122524183, 10122524190, 10122524191, 10122524193, 10122524200, 10122524211, 10122524213, 10122524243；
  + 单倍型模式 (haplotype pattern)：每个SNP的位点都会有频率不同的核苷酸类型，而单倍型模式就是每个SNP位置都用一个具体的核苷酸来代替，因此每个水稻品系都有一个具体的单倍体模式。例如osa-miR156a一个单倍型模式是，GGCAGCTAAC，其中参照基因组的模式被称为参考基因组模式 (reference pattern)，而其它的都称为非参考基因组模式 (non-reference pattern)
* 操作流程
  + 将pre-miRNA上的SNP排列成上升的顺序；
  + 得到每个miRNA的参考基因组模式，以及其它的非参考基因组模式；
  + 根据每个水稻品系的基因型，将水稻品系和单倍型模式对应，也就是将水稻品系根据其单倍型模式分成不同的类型；
  + 最后，将每个单倍型模式转换成pre-miRNA的序列；

致 谢

↑

（黑体3号字居中，段前0.7厘米，段后0厘米，单倍行距，与致谢内容之间空两行）

本文需要感谢……致谢内容：宋体，小四号（“论文正文”样式）。

攻读硕士学位期间已发表或录用的论文

↑

（黑体3号字居中，段前0.7厘米，段后0厘米，单倍行距，与内容之间空两行）

[1] 张三，李四. 已经发表一篇学术论文. XXXXXXX学报 （已录用）

（采用“参考文献内容”样式）

1. Kozomara A, Griffiths-Jones S., miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014, 42:D68-D73 [↑](#endnote-ref-1)
2. Lu, C., et al. Genome-wide analysis for discovery of rice microRNAs reveals natural antisense microRNAs (nat-miRNAs). Proc. Natl. Acad. Sci. USA 2008, 105: 4951–4956. [↑](#endnote-ref-2)
3. Jones-Rhoades M.W., Conservation and divergence in plant microRNAs. Plant Mol Biol. 2011, 80:3–16 [↑](#endnote-ref-3)
4. Li, J.Y. et al. 2014, The functional scope of plant microRNA-mediated silencing. Trends Plant Sci. 2014, 19:785-756. [↑](#endnote-ref-4)
5. Dai X., Zhao P.X., psRNATarget: a plant small RNA target analysis server. Nucleic Acids Res. 2011, 39:W155-W159 [↑](#endnote-ref-5)
6. Li Y.F., Zheng Y., Addo-Quaye C., et al., Transcriptome-wide identification of microRNA targets in rice. Plant J. 2010, 62:742-759 [↑](#endnote-ref-6)
7. Kertesz M., Iovino N., Unnerstall U., et al., The role of site accessibility in microRNA target recognition. Nat. Genet., 2007, 39:1278-1284. [↑](#endnote-ref-7)
8. Liu Q, Wang F, Axtell M.J., Analysis of complementarity requirements for plant microRNA targeting using a Nicotiana benthamiana quantitative transient assay. Plant Cell, 2014, 26: 741-753 [↑](#endnote-ref-8)
9. Iwakawa H., Tomari Y., Molecular insights into microRNA-mediated translational repression in plants. Mol. Cell, 2013, 52:591-601 [↑](#endnote-ref-9)
10. Tang G., Reinhart B.J., Bartel D.P., et al., A biochemical framework for RNA silencing in plants. Genes Dev. 2003, 17:49-63 [↑](#endnote-ref-10)
11. Palatnik J.F., Allen E., Wu X., Schommer C., et al., Control of leaf morphogenesis by microRNAs. Nature, 2003 425:257–63 [↑](#endnote-ref-11)
12. Sato Y., Namiki N., Takehisa H., et al., RiceFREND: a platform for retrieving coexpressed gene networks in rice. Nucleic Acids Research, 2013, 41:D1214-D1221 [↑](#endnote-ref-12)
13. Sturn A, Quackenbush J, Trajanoski Z. Genesis: Cluster analysis of microarray data., Bioinformatics. 2002,18(1):207-8. [↑](#endnote-ref-13)
14. Wen M., Xie M.N., He L., Wang Y.S., Shi S.H., Tang T., 2016, Expression Variations of miRNAs and mRNAs in Rice Oryza sativa. Genome Biology and Evolution 8:3529-3544 [↑](#endnote-ref-14)
15. Zhao H., Yao W., Ouyang Y., et al., RiceVarMap: a comprehensive database of rice genomic variations. Nucleic Acids Research, 2014, 43:D1018–D1022. [↑](#endnote-ref-15)
16. Chen K., Rajewsky N., Natural selection on human miRNA binding sites inferred from SNP data. Nature Genet., 2006, 38:1452–1456 [↑](#endnote-ref-16)
17. Saunders M. A., Liang H., Li, W. H., Human polymorphism at microRNAs and microRNA target sites. Proc. Natl Acad. Sci. USA, 2007 104, 3300–3305 [↑](#endnote-ref-17)
18. Fahlgren N., Jogdeo S., Kasschau K.D., et al., MicroRNA gene evolution in Arabidopsis lyrata and Arabidopsis thaliana. Plant Cell, 2010, 224:1074–1089 [↑](#endnote-ref-18)
19. Rajagopalan R., Vaucheret H., Trejo J., Bartel D.P., A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. Genes Dev., 2006, 2024:3407–3425 [↑](#endnote-ref-19)
20. Liu Q., Wang H., Zhu L., et al., Genome-wide identification and analysis of miRNA-related single nucleotide polymorphisms SNPs in rice., 2013, Rice 6:10 [↑](#endnote-ref-20)
21. Mallory A.C., Reinhart B.J., Jones-Rhoades M.W., et al., MicroRNA control of PHABULOSA in leaf development: Importance of pairing to the microRNA 59 region. EMBO J., 2004, 23: 3356-3364. [↑](#endnote-ref-21)
22. Parizotto E.A., Dunoyer P., Rahm N., et al., In vivo investigation of the transcription, processing, endonucleolytic activity, and functional relevance of the spatial distribution of a plant miRNA. Genes Dev., 2004, 18: 2237-2242 [↑](#endnote-ref-22)
23. Mi S.J., Cai T., Hu Y.G., et al., Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5’ terminal nucleotide. Cell, 2008, 133: 116-127 [↑](#endnote-ref-23)
24. Mallory A., Vaucheret H., Form, function, and regulation of ARGONAUTE proteins. Plant Cell, 2010, 22: 3879–3889 [↑](#endnote-ref-24)
25. Schwab R., Palatnik J.F., Riester M., et al., Specific effects of microRNAs on the plant transcriptome. Dev. Cell, 2005, 8: 517–527 [↑](#endnote-ref-25)
26. Franco-Zorrilla J.M., Valli A., Todesco M., et al., Target mimicry provides a new mechanism for regulation of microRNA activity. Nat. Genet., 2007, 39:1033-1037 [↑](#endnote-ref-26)
27. Todesco M., Rubio-Somoza I., Paz-Ares J., et al., A collection of target mimics for comprehensive analysis of microRNA function in Arabidopsis thaliana. PLoS Genet., 2010, 6: e1001031 [↑](#endnote-ref-27)
28. Arikit S., Zhai J., Meyers B.C., Biogenesis and function of rice small RNAs from non-coding RNA precursors. Curr. Opin. Plant Biol., 2013, 162:170–179. [↑](#endnote-ref-28)
29. Sun G., Yan J., Noltner K., et al., SNPs in human miRNA genes affect biogenesis and function. RNA, 2009, 15:1640–1651 [↑](#endnote-ref-29)
30. Hofacker I.L., Vienna RNA secondary structure server. Nucleic Acids Res., 2003,

    31:3429–3431 [↑](#endnote-ref-30)