# Work Summary 2016/10/11

Rice miRNA and target SNP analysis:

**Part I: MiRNA SNP searching (2015/8-2015/11)**

***Aim:*** Collect all rice miRNA information from *miRBase.org* including their genome coordination, strand orientation and sequences of precursors and mature products; establish a local 3K rice genome SNP database and search SNPs that fall into the genomic regions of miRNAs against the local database.

***Analysis Procedure:***

1. Collect all related information of miRNAs, especially genomic coordination
   1. Majority of them (both pre-miRNAs and mature miRNAs) can be found in *miRBase.org*. While a small number of miRNAs’ genomic coordination are not provided, some of which are well-known miRNAs e.g. osa-miR444 families;
   2. Use BLAST to get the rest of miRNAs’ coordination (which are not provided by the website).

**RESULT:** There are 592 pre-miRNAs and 713 mature miRNAs reside in *miRBase.org*, out of which 14 pre-miRNAs’ genomic coordination are not provided (in other words, 578 pre-miRNAs and 703 mature miRNAs are described in full detail).

By BLASTing the rest 14 pre-miRNAs, I have obtained the genomic coordination of 7 pre-miRNAs (along with 13 corresponding mature miRNAs).

**Short summary:**

Finally, 585 pre-miRNAs and 703 mature miRNAs are available for downstream analysis; while 7 pre-miRNAs and the 10 corresponding mature miRNAs failed in the process.

**Brief conclusion in the collection process:**

While BLASTing the osa-miR444 family sequences, I have found only their binding sites locate on the non-coding strand of their target gene. This has been documented in a paper published in PNAS 2008, which is the so-called antisense miRNAs.

1. Establish a local 3K rice SNP database

***Note:*** There are 2 kinds of 3K rice SNP data available on-line, one is filtered SNP data set, another original SNP data set. The former set of data is very convenient to get as it is provided for batch download, while the latter one requires manually download for each fragment of chromosomes.

But the original SNP data set contains tri-/tetra-allelic SNPs and much more abundant SNPs most of which are thrown away in the filtering process, but are still relatively credible, and these extra SNPs are very precious wealth for us to explore the detailed genotypes of the sequenced ~3000 rice accessions; so, we chose the latter data set.

* 1. Download SNP files from Rice SNP-Seek Database (*oryzasnp.org/iric-portal/*) manually;
  2. Design and establish local database with MySQL (a database management software);
  3. Write scripts to parse the downloaded SNP files and load them into the local databases.

**RESULT:**

1. During the process, the Rice SNP-Seek Database has updated their SNP data, which directly doubled the workload, so currently I have 2 different local databases based on the old and new Rice SNP-Seek versions respectively;
2. Brief info of the Database:
   * Old version: 2,828,431 SNPs stored against 3000 rice accessions;
   * New version: 793,337 SNPs stored against 3024 rice accessions;
   * Downloading files only cover the genomic regions of miRNAs and their target genes, not all the SNP files from the Rice SNP-Seek Database;
3. Search SNPs against local SNP database using self-written scripts

**RESULT:**

* 1. SNPs for 585 pre-miRNAs:
     1. Old version: 4617 SNPs;
     2. New version: 7193 SNPs;
     3. 4278 SNPs are consistent in both copies of databases;
  2. SNPs for 703 mature miRNAs:
     1. Old version: 793 SNPs;
     2. New version: 1270 SNPs;

**Short conclusion:**

SNP population has increased a lot in the new version SNP data partly due to the increasing of rice cultivar accessions from 3000 to 3024.

**Overall comments for Part I:**

1. This is the major and fundamental part of the whole research, also it costs a large amount of time;
2. 3K rice genome SNP database provided our research with abundant data, and a large number of SNPs have been found.

**Part II: MiRNA SNP analysis (2015/9-2015/10)**

***Aim:*** To interpret the SNPs at hand in with miRNA haplotype analysis.

***Analysis Procedure:***

1. MiRNA classification (by conservation)
   1. For pre-miRNAs, their classification was aided by the miRNA family file provided by *miRBase.org*, in which all miRNAs from many species are sorted by families;
      1. Detailed description: all miRNAs are divided into 4 categories, which are rice specific, only conserved in monocot, conserved in both monocot and dicot, conserved in dicot but not monocot.
      2. The latter 3 categories are regarded as conserved miRNAs, while the first category is regarded as non-conserved miRNAs;
   2. Canonical mature miRNAs as well as those non-canonical miRNAs generated from conserved precursors, which have corresponding counterparts in other plant species and relatively high reads in sRNA-Seq are deemed as the conserved miRNAs;

**RESULT:**

* 1. For pre-miRNAs, 191 are classified as conserved, while 401 are non-conserved;
  2. As for mature miRNAs, 220 are conserved mature miRNAs, while 493 are non-conserved mature miRNAs.

**Short comment:**

1. MiRNA classification is the basis for downstream SNP statistics, because generally conserved miRNAs are well studied and we’ll focus on conserved ones in our downstream analysis, in addition, non-conserved miRNAs are relatively new in the process of miRNA evolution which means comparing of conserved miRNAs and non-conserved ones helps us understand the evolutionary difference between the 2 groups;
2. Also, most well-studied miRNAs are generally conserved, so it makes us more to focus on them for SNP-related function analysis.
3. MiRNA haplotype analysis (Appendix I & Appendix II in detail)

**Overall Comment for Part II:**

1. After the classification of miRNAs and consequent SNP analysis, we can see that there are some differences between the conserved and non-conserved miRNA sets;
2. The frequency of SNP represents the evolutionary pressure on mature miRNA sites; the higher the SNP frequency, the lower the evolutionary pressure. So the SNP frequencies along the mature miRNA site, to some extent, implies the significance of each site in terms of target recognition. But this shall further be combined with the SNP distribution over the corresponding miRNA binding site of the target genes, then the co-evolution of miRNAs and their target genes will be revealed;
3. The haplotype serves as a connection between genotype and actual phenotype of rice accession, but in this stage, we cannot rashly come to any conclusion because the function of any miRNA is determined by its target gene;

**Part III: MiRNA target prediction and examination of biological relevancy (2015/11-2016/1)**

***Note:*** To analyze the function of miRNAs, the analysis of their target genes cannot be excluded. But it is impossible to differentiate biologically relevant targets from the rest by simply comparing complementarities. A possible helpful way to screen targets with possible biological relevancy is to calculate and filter the negative expression correlation of miRNA and target interaction pairs. It is thought that the expression of miRNAs and cognate target genes which are biologically relevant should be in negative correlation under the premises:

* 1. Complementarity is the sole determinant of silencing;
  2. Plant miRNA-loaded RNA-induced silencing complex (miRISC) is able to act independently.

In this part, we mainly focus on the conserved miRNAs.

***Aim:*** Collect miRNA target through bioinformatics prediction and experimentally validation, then filter the target gene dataset with their expression correlation.

***Analysis Procedure:***

1. Collection of miRNA targets:
   1. Experimentally validated targets are collected from the paper "*Transcriptome-wide identification of microRNA targets in rice*", and the corresponding genes with clear degradome signal are taken as true, biologically relevant targets.
   2. Bioinformatics approach: Using *psRNATarget* web server with the default parameter (maximum number of mismatches as 3.0, length of complementarity scoring as 20 and range of central mismatch as 9-11) to predict targets for all conserved miRNAs;
   3. To obtain the genomic coordination of miRNA binding sites and their flanking regions for experimentally validated target genes using *psRNATarget*.

**Result:**

1. For experimentally validated targets: there are 46 collected and 778 miRNA:target interaction pairs found;
2. Overall collection: there are 823 target genes collected (include the literally recorded ones and bioinformatics prediction) and corresponding 2113 miRNA:target interaction pairs, out of which 120 pairs are experimentally unique (not overlapped with the bioinformatics prediction)
3. The corresponding miRNA binding sites were acquired and stored.

**Notice:**

1. The alternative splicing forms of target genes are reserved in our data lest alternative splicing modifies the miRNA binding sites on the target genes.
2. SNP searching within the miRNA binding sites and flanking regions against the local 3K rice genome SNP database

**Result:**

1. 9217 SNPs were found within the regions flanking (~100bp upstream and downstream) miRNA binding sites (include) of the collected 823 target genes; while 1169 of them fall within the miRNA binding sites;
2. Expression correlation of miRNA:target

Note: Expression profiles of miRNAs and target genes were extracted from the experiment data of *RiceFREND (RiceFREND: a platform for retrieving coexpressed gene networks in rice)*.

1. Use experimentally validated target genes as authentic target genes to train the correlation method, and filter the predicted target genes to get the biologically relevant ones;
2. I’ve tried a method called TaLasso (described in “Quantification of miRNA-mRNA Interactions”), and found this method is not applicable in our research;

***Note:*** The Talasso method aims to decipher the miRNA-mRNA relationship with the aid of expression data. It adopts the Lasso regression and has several assumptions:

a) miRNAs are the only regulators of mRNA expression, considering other possible effects as part of the noise;

b) the miRNAs down-regulate their corresponding mRNA targets;

c) TaLasso will only quantify the down-regulation effect on those miRNA-mRNA interactions from an initial set of putative miRNA-mRNA pairs.

In the paper, the Talasso method was tested with both the experimentally validated targets as well as predicted ones, and ranked them with the parameters it generated.

There are several points which make this method not applicable to our research:

a) It assumes that miRNAs down-regulate true targets——In fact, in plants, this assumption remains to be validated, and when we use the Pearson Correlation Coefficient to test the expression correlation of degradome validated interaction pairs, not all interactions are negative correlated; which indicates that this assumption does not hold the truth;

b) Although the top-ranked predictions of miRNA-mRNA interactions in the resulted Talasso method are enriched in experimentally validated interaction set. There was no cut-off parameter provided to judge whether a predicted interaction pair is an authentic one or not.

1. Therefore, I just calculated the Pearson Correlation Coefficient and Spearman Correlation Coefficient of miRNA and corresponding target expression profile. I’ve performed the correlation of pre-miRNAs and target genes and that of mature miRNAs and target genes.

***Note:***

* + - 1. In the calculation of correlation coefficient, choosing of sample sets is very crucial, for different samples sets will result in very different coefficients. In my research I’ve performed the correlation test on 4 samples sets: a) all sample, b) all seedlings, c) 27-day seedling, d) all-flowers. Because the paper *"Transcriptome-wide identification of microRNA targets in rice"* from which I collected the experimentally validated targets, adopts the 4-week (28 days) seedling as samples, then the 27-day seedling samples used by *RiceFREND* would be most similar.
      2. Correlation of mature miRNAs with target genes serves as a validation test for the assumption that there’s the negative correlation between authentic miRNA:target interactions, while the correlation of precursors and target genes would provide insight into the individual relationship between each miRNA family member and the putative target genes.

**Result:**

1. The putative conclusion, that the expression of miRNAs is negatively correlated with that of their targets, does not stand firm here, as the correlation was not observed for all experimentally validated target genes
2. In 27-day-old seedlings, expression of 136 out of 367 precursor:target pairs are negatively correlated (all these interaction pairs are experimentally validated in tissues that are at similar developmental stages );
3. In 27-day-old seedlings, 116 out of 363 mature:target pairs are negatively correlated.

**Conclusion:**

There are 2 possible explanations to the unexpected results of the expression correlation:

1. Pre-assumed negative expression correlation between miRNAs and cognate target genes does not hold firm in the dataset we examined, which implies that there are other factors other than complementarity = affect the silencing of target mRNA;
2. This expression profiling was performed in a collection of tissues containing different organs and cell types, and the calculation was performed between the total expression of miRNA and targets. So it cannot be ruled out the possibility that there might be instances that the targets are only expressed in certain organs/cells，or expressed exclusively with cognate miRNAs, therefore the summed values used will screw the result. .

**Short comments:**

Though the result does not seem promising, it actually tells us the correlation between miRNAs and their targets is not as simple as we thought previously. And to some extent, this suggests the possible sub-functionalization of miRNA family members.

Further experiments may be needed to validate the correlation between them.

**Part IV: SNP Statistics of miRNAs and miRNA binding sites (2016/9-2016/10)**

***Aim:*** To do statistical analysis of SNPs found within miRNAs and miRNA binding sites helps to explore the evolutionary relationship of conserved and non-conserved miRNAs as well as the that of mature miRNAs and corresponding binding sites;

***Analysis Procedure:***

1. MiRNA SNP statistics
   1. Pre-miRNA SNP density
      1. For precursor miRNAs, calculate the SNP density of by *(Total Number of SNPs) / (Length of pre-miRNAs)*

**Conclusion:**

* SNP density of non-conserved miRNAs is larger than that of conserved miRNAs, indicating larger evolutionary pressure on conserved miRNAs;
  1. Compare pre-miRNA SNP density with controls of randomly chosen exon fragments and intergenic regions

***Note:***

Averagely, the length of pre-miRNAs is 152 and there are592 pre-miRNAs in total. 2 papers addressing the topic of how to select the controls for statistical experiments, suggest the ratio of controls to cases is set to 1;

* + 1. Acquire the exon regions and intergenic regions, randomly choose 600 fragments with the length of 150bp from both kinds of regions respectively, out of which from every chromosome, will 50 fragments be randomly selected;
    2. Retrieve the SNP number of each fragment chosen by searching the online database, and do the plot of SNP density of exon regions and intergenic regions separately.

**Conclusion:**

* Compared with the density plot of intergenic region, higher percentage of pre-miRNAs fall into the lower SNP density category than that of intergenic fragments, which may imply pre-miRNAs are relatively more evolutionary conserved than intergenic regions;
  1. SNP distribution of mature miRNAs and miRNA binding sites
     1. For each category of mature miRNAs, calculate the frequency that a SNP appears in every site along the ~21nt mature miRNAs. The frequency is calculated by *(Number of miRNAs that possess SNP at this site) / (Number of miRNAs)*;
     2. I have specially chosen the binding sites of degradome validated targets, as well as the that of genes targeted by conserved miRNAs, and have done the plot of SNP distribution along the binding sites;

**Result:**

***For SNP distribution over mature miRNAs***

1. Overall, positions (1, 18 ,11) possess the lowest SNP frequency;
2. Among conserved miRNAs, position (1,9,12) have the lowest SNP density, which is contrary to empirical parameters that cleavage site (10,11) must be complementary to target genes, indicating high evolutionary pressure.

***As for SNP distribution over miRNA-binding site of degradome validated miRNA:target***

Notice:

     The total number of degradome validated miRNA::target interaction pairs are relatively small, so the SNP distribution may not be that reliable statistically;

1. Positions (11, 13) are free from SNPs; apart from them, Positions (1,2,10,12) possess the lowest SNP density.
2. Positions(14, 15, 17) possess the highest SNP density;

**Conclusions:**

* Position 1 has the lowest SNP density both from the miRNA aspect and the binding site aspect, which may imply that Position 1 is under high evolutionary pressure, and this may due to the fact that position determine the type of AGO to load the mature miRNAs;
* The controversy to the empirical parameters remains, that Positions (10, 11) on miRNA-binding site do not possess the lowest SNP density;

**Part V: Combined Complementary Pattern Analysis (2015/11-2016/1)**

***Aim:*** At this stage, we’ve already gotten the SNPs of mature miRNAs and miRNA binding sites in target genes, so we can combine them together to see how SNPs will affect the complementarity between miRNAs and targets.

***Analysis procedures:***

\*\* The process is similar to the previously mentioned miRNA haplotype analysis, apart from the following steps:

1. Classify the complementarity pattern of some conserved miRNAs & their targets and compare the results with current empirical parameters in complementarity requirements published by other researches (Liu, Q., et al. (2014). Plant Cell **26**(2): 741-753. & Schwab, R., et al. (2005). Dev Cell **8**(4): 517-527.)
   1. Complementarity patterns of miRNA:target interaction pairs are divided into 2 groups, namely, degradome validated (DV) group and predicted but not degradome validated
   2. (PNDV, their cleavage products of target genes are not detected, 2 possibilities: a. not true target genes; b. the target gene does not express at the detected organ at the chosen stage) group (all member comparison);
   3. Considering of making the contribution of each miRNA family even, I’ve summarized the pattern by choosing only one family member of each miRNA family as representative pattern (representative comparison); this may be more accurate for the result would not be skewed by the different number of miRNA family members;
   4. Note: we choose "miRXXXa” member as the representative, for generally the miRXXXa is the most common and highest expressed ones of the whole family.

**Basic Statistics:**

1. Classification is performed on a data set composed of 193 conserved miRNAs and 616 targets, and in total 1464 complementarity patterns;
   1. For all member comparison:
      * 1037 patterns in the PNDV set;
      * 427 patterns in the DV set;
   2. For representative comparison:
      * 420 patterns for PNDV set;
      * 54 patterns for DV set;

**Results:**

1. **Comparison between DV set for representative mode and DV set for all member mode**

***Note:*** First, we actually cannot distinguish the true regulating members of a miRNA family to a corresponding target gene from those members that can’t regulate. Or maybe other members just regulate with different silencing efficacies because their complementarity patterns have been altered.

     So we suppose that “miRXXXa” member is the true regulator, and all other members’ complementarity patterns are the isoforms to the representative one.

***Conclusion:***

In the representative comparison, DV set shows that position 4, 10, 11, 16 are free from mismatches;

While in the all member comparison, DV set possesses mismatches at these 4 positions at very low frequencies, so this can be explained by the existence of some weird family members of miRNA families. Also the low frequencies show that these positions are very important to the silencing efficacy, and that the miRNA members that possess mismatches at these 4 positions may not regulate the target genes well or even cannot silence the target genes;

1. **Representative comparison of DV set and PNDV set** (Compare the 2 sets only with chosen representative miRNAs)

***Conclusions:***

1. All positions of PNDV set possess mismatches at some frequencies, while position 3, 4, 10, 11, 16 in DV set are free from mismatches, suggesting these 5 positions require base pairing to function well;
2. Mismatch frequencies of position 14 and 20 of the DV set is almost twice that of the PNDV set, suggesting the existence of mismatches at 3’ region would increase the possibility of turning to be a true regulating miRNA:target interaction especially at position 14 and 20;
3. Concatenate genomic regions of miRNA precursors and miRNA binding sites together as one fragment and perform haplotype analysis;

**Results:**

1. We have analyzed 180 miRNAs together with their 538 target genes, and finally we obtained 28732 unique combined haplotype patterns; and also, for each combined pattern, we have generated the mutated RNA sequence of their mature miRNA and miRNA binding site in preparation for the next step analysis, and collected the lists of the corresponding cultivar index numbers searching through the local database;

|  |  |
| --- | --- |
| Length of Combined pattern | Number of Patterns |
| 1 | 845 |
| 2 | 1443 |
| 3 | 1541 |
| 4 | 1933 |
| 5 | 2471 |
| 6 | 2697 |
| 7 | 3011 |
| 8 | 2672 |
| 9 | 2494 |
| 10 | 1900 |
| 11 | 2311 |
| 12 | 2367 |
| 13 | 1095 |
| 14 | 748 |
| 15 | 862 |
| 16 | 342 |

1. Interpret the haplotype pattern into different complementary patterns (just find out the position of the SNPs in the complementary patterns);

***Note:*** Transform the combined haplotype pattern into visualized complementarity pattern of mature miRNA and its binding site with genomic coordination of SNPs, mature miRNAs and target genes;

1. Analysis into the results of combined complementarity analysis
   1. Classify SNP effects on the complementarity pattern into 4 categories, namely: i. still mismatch (after the point mutation, the state of complementarity of the site remains mismatched); ii. mismatch to paired (the state of complementarity has turned from mismatch to pairing); iii. paired to mismatch (the site turned mismatch after the point mutation); iv. still paired (the state of complementarity remains paired);

***Note***: The first category (still mismatch) is valueless in our study, so we’ll just skip this category;

**Result:**

* Still paired patterns: 179 patterns possess 181 positions,
* Paired to mismatch: 13840 patterns with 19845 positions;
* Mismatch to paired: 3979 patterns with 4431 positions;
  1. Deeper analysis into “Still paired patterns”

***Note***: This is a very intriguing phenomenon because both nucleotides of the same site in a complementarity pattern have mutated, and they still remain Watson-Crick pairing.

**Result:**

* + - 1. These 179 “still paired” patterns fall into 5 miRNA families: osa-miR1436, osa-miR1439, osa-miR1442, osa-miR1862b, osa-miR2275c, osa-miR818, osa-miR444;

**Conclusion:**

1. The reason why this kind of pattern appears in osa-miR444 can be explained by mechanism of the anti-sense miRNAs in this family, which means the mature miRNA gene is in the anti-sense strand of its targeting gene and any SNP appear in this genomic region would surely change the both nucleotides and they would surely remain paired;
2. The other 4 miRNA families (including osa-miR1436, osa-miR1439, osa-miR1442, osa-miR1862b) are in fact grouped into the same miRNA family by *miRBase.org*, which is called MIR818 family. This piece of information added to the mystery of the phenomenon;
3. The former phenomenon can be partly explained by the large number of SNPs found within the mature miRNAs and their miRNA binding sites. (ps. They possess 6-7 SNPs averagely). The top 5 frequent sites that possess the “still paired” mutations are site2, site5, site20, site18 and site16.
   1. Analyze 5 well-studied miRNA families

***Note***: The 5 well-studied miRNA families are osa-miR156, osa-miR444, osa-miR172, osa-miR159, osa-miR397;

Except osa-miR444, the miRNA family members are highly similar among each another in the other 4 miRNA families.

* + 1. Multifold Combined Complementarity Analysis (MCCA) into 5 miRNA families
       1. Detailed description about MCCA can be found at Appendix IV;

**Result:**

1. In osa-miR156 family, 7 haplotype patterns have been discovered and each pattern consists of only one point mutation, out of which 5 patterns fall onto the miRNAs and 2 patterns fall onto the target genes. A SNP in osa-miR156c-5p caused a mismatch at site 11, which may potential have big impact on the regulation and should be noticed;
2. In osa-miR172 family, 3 haplotype patterns have been discovered and each pattern consists of only one point mutation, and they all fall onto the miRNAs;
3. But for osa-miR159 family and osa-miR397 family, no haplotype patterns have been found;
   * 1. SNP effects on protein-coding target genes of the miRNAs under study

**Result:**

**osa-miR444 family target genes**

* OsMADs57 (LOC\_Os02g49840)
  1. 10230461208——Gln97Gln, synonymous variant
  2. 10230461213——Leu99Gln, missense variant
  3. 10230461236——Lys107Gln, missense variant
* OsMADs27 (LOC\_Os02g36924)
  1. 10222300431——Ala93Val, missense variant
  2. 10222300448——Gln99Lys, missense variant
* OsMADs23 LOC\_Os08g33488
  1. 10820899277——Gln98Arg, missense variant
  2. 10820899283——Leu96Ser, missense variant
  3. 10820899289——Ala94Gly, missense variant

**osa-miR156 family target genes**

* OsSPL13 (LOC\_Os07g32170)
  + 10719100460, 10719100463——3\_prime\_UTR\_variant
* OsSPL18 (LOC\_Os09g32944)
  + 10919647853——Leu347Phe, missense variant
  + 10919647862——Leu350Leu, synonymous

**osa-miR397 family target gene**

* LOC\_Os12g15680 laccase precursor protein
  + 11208962185——Ala249Ala, synonymous variant
  1. Analyze the phenotypes of rice cultivars to find out the effects of haplotype patterns on phenotypes
     1. Download all the descriptive phenotypes from the Rice SNP-Seek Database ([*http://oryzasnp.org/iric-portal/\_variety.zul*](http://oryzasnp.org/iric-portal/_variety.zul)) of all the 3024 rice cultivars;
     2. As for osa-miR156, the concerning phenotypes in the phenotype list provided by IRRI are as follow:
        1. 100-grain weight (gm);
        2. Grain length (mm);
        3. Grain Width (mm);
        4. Grain size (derived from grain length/width);
        5. Secondary branching at reproductive;
     3. As osa-miR172, they are:
        1. Heading days to 80% fully headed;
        2. Spikelet fertility;

**Result:**

1. For the above 7 phenotypes of the 2 miRNAs, generally cultivars that belong to reference haplotypes cover wide ranges of quantified phenotypes; and all haplotypes in question which are not reference, are without the ranges of reference haplotypes;

**Appendix I**

**miRNA haplotype analysis:**

***miRNA haplotype***

\* Adopt SNP as biological marker, for each miRNA precursor, SNPs distributed within its genome region form the miRNA haplotype (in ascending order of genome coordination)

\* e.g. osa-MIR443's miRNA haplotype: sf0330014542, sf0330014549, sf033001458, sf0330014600

\*\*haplotype pattern\*\*

\* For each miRNA precursor, every locus of SNP is occupied with a nucleotide acid, so haplotype pattern means a specific sequence of nucleotide; and because every SNP possess 2 alleles (commonly, but not always), theoretically there are ```2^len(miRNA haplotype)```haplotype patterns for each haplotype

\* e.g. one haplotype pattern of osa-MIR443: CGGA

\* Special haplotype patterns:

\* Reference pattern: all loci are possessed by allele in reference genome

\* Non-reference pattern: all loci are possessed by allele different from the on in reference genome

\*\*trinary pattern\*\*

\* This is a newly coined term, in which reference allele is replaced by 0, non-reference allele is replaced by 1 and 'N' is replaced by 2(Note that because the sequencing of rice genome got a miss-calling at the specific SNP position, an 'N' will occur)

\* e.g. reference pattern of osa-MIR443: CGGA <===> 0000; while AATT <===> 1111

***Steps of analysis***

\* step1: Classify SNPs into their corresponding precursor intervals in ascending order (This is the so-called miRNA haplotype)

\* step2: Obtain reference pattern and non-reference pattern of each miRNAs (as reference)

\* step3: For each precursor along with its haplotype, grasp the haplotype pattern and the corresponding cultivars

\* step4: Transform the haplotype pattern into trinary pattern\* (To compare each haplotype pattern visually with 0-1-2 digits)

\* step5: For each haplotype pattern, mutate the original RNA sequence with specific SNPs

**Appendix II**

**Haplotype Pattern** (proposal, concerning heterozygotes)

the haplotype pattern will be described in this format:

- ATCG for the alleles

- N for miss-calling allele

- lower-capped letter for heterozygotes

     \*\*for example, a heterozygote pair A/T, where freq(A) > freq(T), then it would be represented as "t"\*\*

     \*\*Because, a minor allele in the heterozygote would be more precious and may be more potential in exploring the gene resources\*\*

**Pentanary Pattern**

0: Reference allele

1-3: Non-reference allele in descending order of their frequency

4: N (miss-calling)

**Processing criteria**:

1) Threshold of #cultivars corresponding to each haplotype pattern is 10 (include 10, which means >= 10, in the RiceVarMap, they also use 10 as threshold)

2) Concerning heterozygote such as aTT, take it as a different haplotype pattern from ATT, but when it was converted to pentanary pattern, they would be converted into the same pattern (In this way, we can trace back the heterozygotes)

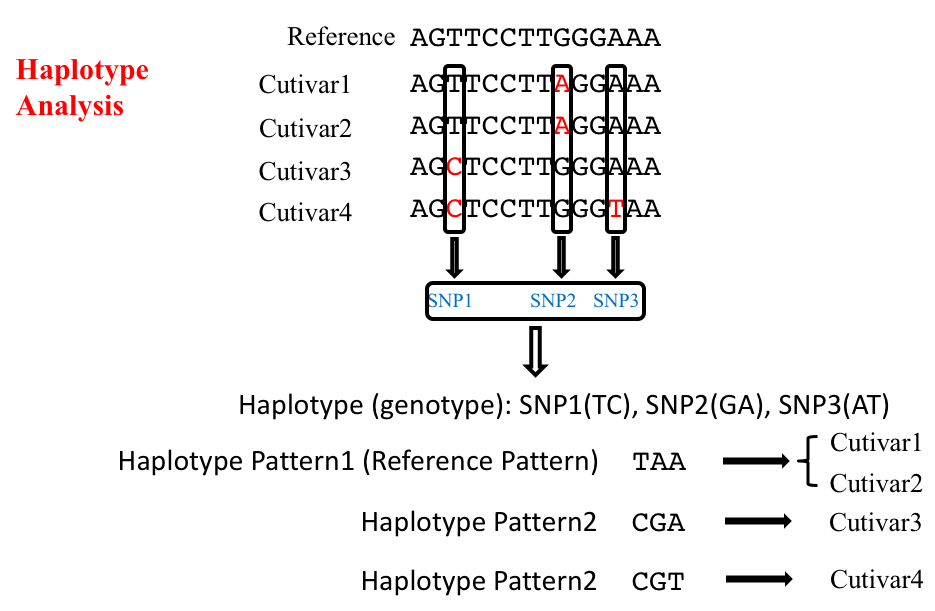
**Appendix III**

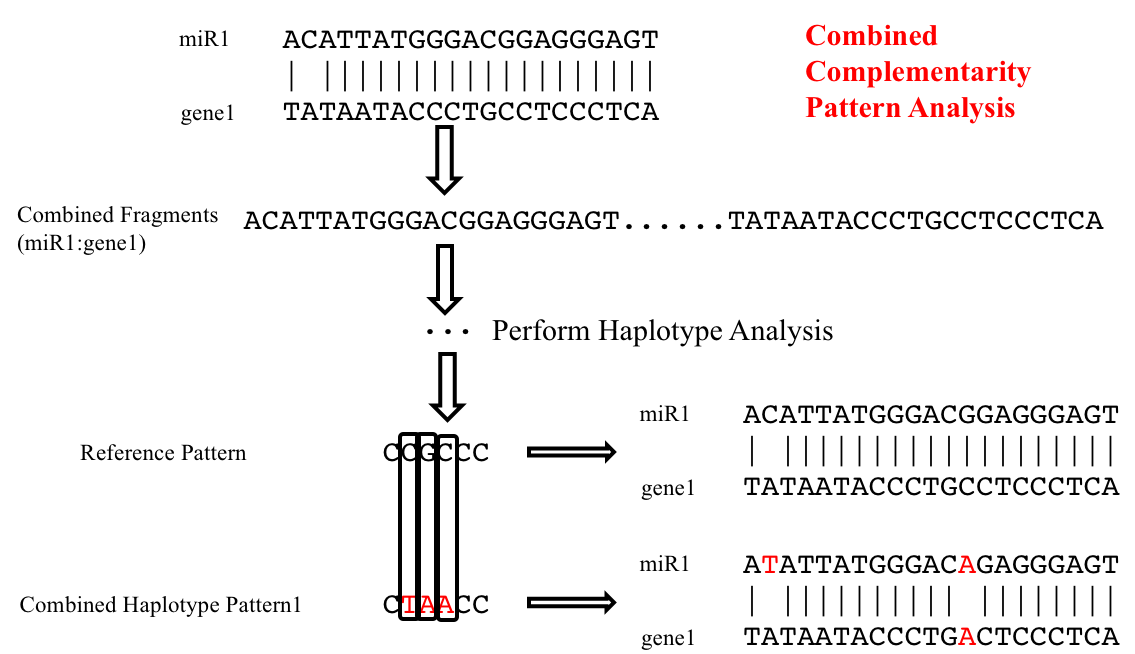
**Combined Complementarity Pattern Analysis**

***[Analysis procedures]***

1. Combine together the genomic region of mature miRNA and that of miRNA binding site into one fragment and perform the haplotype analysis;
2. Obtain the combined haplotype pattern, in which the SNPs come from both fragments and are listed in ascending order;
3. Interpret the combined haplotype pattern into complementarity pattern, where mature miRNA sequence is paired with miRNA binding site (Watson-Crick base pair), to see the effects of point mutations on miRNA pairing in real cultivars.

***[Figures to illustrate the procedures****]*





**Appendix IV**

**Multifold Combined Complementarity Analysis**

***Note:*** *In some well-studied miRNA families, the miRNA members are similar to each other by their sequence and they are often predicted to have the same targets, so in order to see the overall effects of mutations have on the regulation of a specific target gene, we decided to combine together all the miRNA members as well as the target gene binding site under study to perform the multifold combined complementarity analysis.*

***Description:*** *Combine together the genomic regions of all members belonging to the same miRNA family as well as that of the target gene under study;*

***[Figure to illustrate the Multifold combined complementarity analysis, osa-miR156 as an example]***

