# Work Plan and Progress 2016/3/16

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 689 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 BLASTingthe osa-miR444 family sequences, I have found only the 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 (State that why you downloaded the data yourself instead of using the filtered data)

***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 for 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 the old and new Rice SNP-Seek versions respectively;
2. Brief info of the Database:
   * Old version: 2828431 SNPs stored against 3000 rice accessions;
   * New version: 793337 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 statistics and analysis (2015/9-2015/10)**

***Aim:*** To interpret the at-hand SNPs in 2 ways: a) SNP statistics, b) 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 sortedbyfamilies;
      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-conservedmiRNAs;
   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-Seqare 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, for there are some meaning in comparing the conserved miRNAs with non-conserved ones;---- 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 more to focus on them for SNP-relaed function analysis.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. SNP distribution of mature miRNAs
     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)*;

**Conclusions:**

* Overall, position (1, 18 ,11) possess the lowest SNP frequency
  + - 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.

1. 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 ….; the highertheSNP frequency, the lower the evolutionary pressure. So theSNP 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 more things will be revealed;-------- 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 more things including 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 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….. , It is thought that the expression of miRNAs and cognate target genes which are biologically relevant should be in negative correlation .

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 adopting the default parameter to predict targets for all conserved miRNAs;~~
   3. ~~To predict the genomic coordination of miRNA binding sites and their flanking regions for target genes using~~ *~~psRNATarget~~*~~.~~
   4. (More unambiguous way to address b) & c) point) Bioinformatics approach: Using *psRNATarget* web server adopting 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;
   5. To obtain the genomic coordination of miRNA binding sites and their flanking regions for experimentally validated target genes using *psRNATarget*.

**Result:**

1. ~~There are 46 experimentally validated target genes collected and 778 miRNA:target interaction pairs predicted;~~
2. ~~823 target genes are predicted, and 2113 miRNA:target interaction pairs are found in total, out of which 120 pairs are experimentally unique (not overlapped with the bioinformatics prediction)~~
3. ~~The corresponding miRNA binding sites were acquired and stored.~~

**(Unambiguous way)**

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 (training set) 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 provided 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 are negatively correlated with that of their targets, does not stand firm here, as the correlation was not observed forall experimentally validated target genes
2. In 27-day old seedlings, expression of 136 out of 367 miRNA:target pairs are negatively correlated (all these interaction pairs, precursor:target, are experimentally validated in tissues that are at similar developmental stages );
3. In 27-day-old seedlings, ~~82 out of 16 pairs are negatively correlated~~ 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: Combined Complementary Pattern Analysis (2015/11-2016/1)**

***Aim:*** At this stage, we’ve already had 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 procedure:***

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

1. Take genomic region of miRNA precursors and miRNA binding sites together as one segment and perform haplotype analysis;
2. Interpret the haplotype pattern into different complementary patterns (just find out where the SNP appears in the complementary pattern);

Current status:

Now, I’ve acquired the haplotype pattern and corresponding rice accession sets.

Next step is procedure b).

**Plans:**

Link SNP to rice cultivar phenotypes:

1. ~~Select some important and well-studied miRNAs, to see if there are some haplotype patterns have distinct phenotypes related with the miRNAs in the corresponding rice accession set; and analyze the alteration of the 2~~~~nd~~ ~~structure of the pre-miRNAs;~~ Select some important and well-studied miRNAs, to see if there are some haplotype patterns of pre-miRNAs have distinct phenotypes related with the miRNAs in the corresponding rice accession set; and analyze the alteration of the 2nd structure of the pre-miRNAs;
2. Classify the complementary patterns of the miRNA:target interaction pairs, interpret the combined haplotype pattern into changes of complementary patterns and find out those are possibly significant for target recognition you mean??

Similar to a), select some important and well-studied miRNAs, to see if there are some combined haplotype patterns in complementarity patterns have distinct phenotype related with the miRNAs in the corresponding rice accession set;

**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)