# 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 miRNA information from *miRBase.org* including the coordination, strand orientation, sequence of precursors and mature miRNAs; establish 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 miRNAs’ coordinates (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).

After BLASTing the rest 14 pre-miRNAs, I have obtained 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 corresponding 10 mature miRNAs failed in the process.

**Brief conclusion in the collection process:**

While BLASTing with the osa-miR444 family sequences, I have found only the opposite strand of their target gene binding sites perfect match the sequence. This discovery has been documented in a paper published in PNAS 2008, which is the so-called antisense miRNAs.

1. Establish local 3K rice SNP database
   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 doubles the work loading, so currently I have 2 copies of local database including the old version and new version;
2. Database brief info:
   * 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 furnishes our research with abundant data, and 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, they are classified aided by miRNA family file provided by *miRBase.org*, in which file all miRNAs of all species are stored according to their RNA families;
      1. Detail description: the miRNAs are divided into 4 categories, which are rice specific, only monocot conserved, monocot and dicot conserved, dicot but not monocot conserved.
      2. The latter 3 categories are regarded as conserved pre-miRNAs, while the first 1 category is regarded as non-conserved pre-miRNAs;
   2. For mature miRNAs, the canonical mature miRNAs as well as those miRNAs, which have corresponding counterparts in other plant species and relatively big number RNA reads, of the conserved pre-miRNAs are deemed as the conserved miRNAs;

**RESULT:**

* 1. As for pre-miRNAs, 191 are conserved miRNAs, while 401 are non-conserved miRNAs;
  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 the downstream SNP statistics, for there are some meaning in comparing the conserved miRNAs with non-conserved ones;
2. Also, conserved miRNAs are generally well-studied ones, so the afterward function analysis of miRNAs could be very helpful when we focus our eyes on the conserved ones.
3. MiRNA SNP statistics
   1. Pre-miRNA SNP density
      1. For each category of precursor miRNAs, calculate the SNP density of pre-miRNAs by *(Number of SNPs) / (Length of pre-miRNAs)* to see the distribution of each SNP density range;

**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 possibility that a SNP appears in every site along the ~21nt mature miRNAs. Here the possibility is calculated by *(Number of miRNAs that possess SNP at this site) / (Number of miRNAs that possess the specific site)*;

**Conclusions:**

* Overall, site (1, 18 ,11) possess lowest SNP density;
  + - Among conserved miRNAs, site (1,9,12) have 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 classification of miRNAs, we can see there are some differences between the conserved and non-conserved miRNA set;
2. The presence of SNP represents the evolutionary pressure; the higher possibility of SNP means lower level of evolutionary pressure. So the distribution of SNPs along the mature miRNA site, to some extend, implies the evolutionary pressure each site has. 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;
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 revealed by its target gene;

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

***Note:*** Any approach to analyze the function of miRNAs cannot skip the analysis of their target genes. And the negatively expression correlation of miRNA and target gene is thought to be the indicator of biological relevancy of putative target genes.

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 titled "Transcriptome-wide identification of microRNA targets in rice", and here this dataset is taken as the true, biologically relevant target gene set;
   2. Bioinformatics approach: Using *psRNATarget* web server adopting the default parameter to predict targets for all the conserved miRNAs;
   3. To predict the genomic coordination of miRNA binding sites and their flanking regions for target genes using *psRNATarget*.

**Result:**

1. There are 46 experimentally validated target genes collected and 778 miRNA:target interaction pairs;
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 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 are found within the regions flanking 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 are extracted from the experiment data of *RiceFREND (RiceFREND: a platform for retrieving coexpressed gene networks in rice)*.

1. Use experimentally validated target genes as 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;
3. Finally, I just calculate 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

**Result:**

1. The putative conclusion, that miRNAs are negatively expression correlated with their targets, does not stand firm here, for not all the experimentally validated target genes are negatively correlated with miRNAs;
2. In 27-day seedlings, 136 out of 367 pairs are negatively correlated (all these interaction pairs, precursor:target, are experimentally validated, and 27-day seedlings are the closest rice organisms to the ones published in that paper);
3. In 27-day seedling, 82 out of 16 pairs are negatively correlated

**Conclusion:**

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

1. Pre-assumed negative expression correlation of miRNAs and target genes does not hold firm in living cells, which means factors that will affect the silencing of miRNA upon the target mRNA;
2. This expression profile is a collection of all cells in a tissue, so when the expression level are summed, the negative correlation will disappear.

**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 extend, this suggests the sub-fucntionalisation of miRNA family members.

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

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

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

***Analysis procedure:***

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

1. Take mature miRNA genomic region and miRNA binding sites together as one segment and perform the haplotype analysis;
2. Interpret the haplotype pattern into the altered complementary pattern (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 phenotype 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 pattern of the miRNA:target interaction pairs, interpret the combined haplotype pattern into the change of complementary pattern and find out those which are significant changes;
3. Similar to the a), select some important and well-studied miRNAs, to see if there are some combined haplotype patterns have distinct phenotype related with the miRNAs in the corresponding rice accession set; and analyze the alteration of combined complementary pattern;

**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)