# Work Plan and Progress 2015/9/12

Rice miRNA SNP analysis:

**Part I: RiceVarMap SNP analysis (2015/6-2015/8)**

(By then, the 3K Rice SNP-seek Database was not at service. So we decided to work with RiceVarMap as data resource and take it as a practice to know better about the down-stream analysis)

Aim: Obtain the SNPs that fall in miRNA precursor genome regions and their targets. It helps to build the down-stream SNP analysis pipeline and in the future the result can be compared with the 3K SNP data.

Analysis Process:

1. Get the genome coordination of all miRNA precursors and mature miRNAs
   1. Majority of them can be found in miRBase, but due to the genome version adopted by miRBase(MSU7.0) is different from the that of RiceVarMap (MSU6.1). In addition to that, a small portion of miRNAs’ coordinates are not provided;
   2. Use genome lift-over program and MSU6.0 genome coordination (which is found the same as MSU6.1) to get the majority of the genome coordination;
   3. Use BLAST to get the rest miRNAs’ coordinates (which are not provided by the website).

RESULT: There are 585 miRNAs’ coordinates are obtained. But still there are 6 miRNAs whose coordinates are not obtained. They are osa-MIR812k, osa-MIR812I, osa-MIR1862f, osa-MIR3980a, osa-MIR3980b, osa-MIR812o.

1. MiRNA classification (into conserved and non-conserved families)
   1. As for miRNA precursors, they are classified by the rules of miRBase (for they have classified all the miRNAs of all species into miRNA familes);
   2. As for mature miRNAs, they are divided as canonical and non-canonical miRNAs. Canonical mature miRNAs are classified by the same rule as the corresponding miRNA precursors and non-canonical miRNAs are defined as conserved miRNAs.

RESULT: miRNA precursors are already classified into 4 categories

(Group1: Conserved in monocots and dicots;

Group2: Conserved only in monocots;

Group3: Rice-specific;

Group4: Conserved only in dicots;)

While mature miRNAs are not yet classified.

1. Search SNPs against RiceVarMap

RESULT: There are 1724 SNPs found falling in the region of precursors and 322 SNPs within within mature miRNA regions.

1. Haplotype analysis

Note: Further analysis would involve the analysis of SNPs’ impact on the miRNAs. Connecting the SNP with actually sequenced cultivars is very necessary. So we decided to do haplotype analysis. More information and illustrations about haplotype analysis is attached to the end of the report (Appendix I).

RESULT: The haplotype analysis is finished, and we have got the corresponding sequence to the haplotype patterns.

**Overall comments:**

1. This part is suspended right now until the 3K SNP analysis is finished. Then the result can offer as a comparison to the 3K SNP anlaysis

**Part II: 3K SNP-seek Database analysis (2015/8-NOW)**

(From 2015/8, the 3K SNP-seek Database has returned back to normal. So immediately we put aside the analysis on RiceVarMap and refocus on 3K SNP-seek Database)

Aim: Obtain the SNPs with the 3K SNP-seek Database and build the complete pipeline of miRNA’s SNP analysis;

Analysis Process:

1. Download raw data from the 3K SNP-seek Database

Note: The database really is hard to use, even for downloading the SNP data. I plan to download all genome-wide SNPs from the database, but on the database, I can only query at most 50kbp genome region at each time. So there would be more than 6700 queries for me to download all the SNPs. Each query would take ~1.5mins. So estimated time of downloaded would be 167hours (nearly a month’s work loading without doing other work)

RESULT: Currently, I’ve downloaded about 1100 files (a file for each query), covering all the miRNA precursor genome regions.

1. Inserting the SNP data into local database

RESULT: There are 2502167 SNPs, and it takes 819 mins (~13 hours) to insert them into the database. For ~20M SNPs, it would take 109hours (~a week’s time)

1. Haplotype analysis
   1. We’ve just discussed and modified the haplotype analysis criteria, which is illustrated in Appendix II

**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 resourses\*\*

**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)