**Detailed Statements of Experimental Procedures with Figures and Tables**

My Question / Hypothesis: why do I do this part of experiment?

My Methodology: how can I accomplish this part of experiment?

My Expectation: what do I expect as the result of the hypothesis?

My Outcome;

My Discovery: what fits into my expectation or paradox to the known facts? how can this be explained?

**Part I. Overall SNP density of Rice miRNAs**

My Premises:

* SNP density can reflect the selection pressure imposed on given genomic regions. The higher the pressure is; the lower the SNP density will be.
* Conserved miRNAs, and thus are more strictly selected by evolutionary pressure than non-conserved miRNAs.

My Hypothesis:

* Since miRNAs are functional genomic units and master regulators, their SNP density may reflect different selection pressure compared with that of intergenic regions and exons;
* There should be higher selection pressure on conserved miRNAs, therefore SNP density on conserved miRNAs should be lower than non-conserved miRNAs.

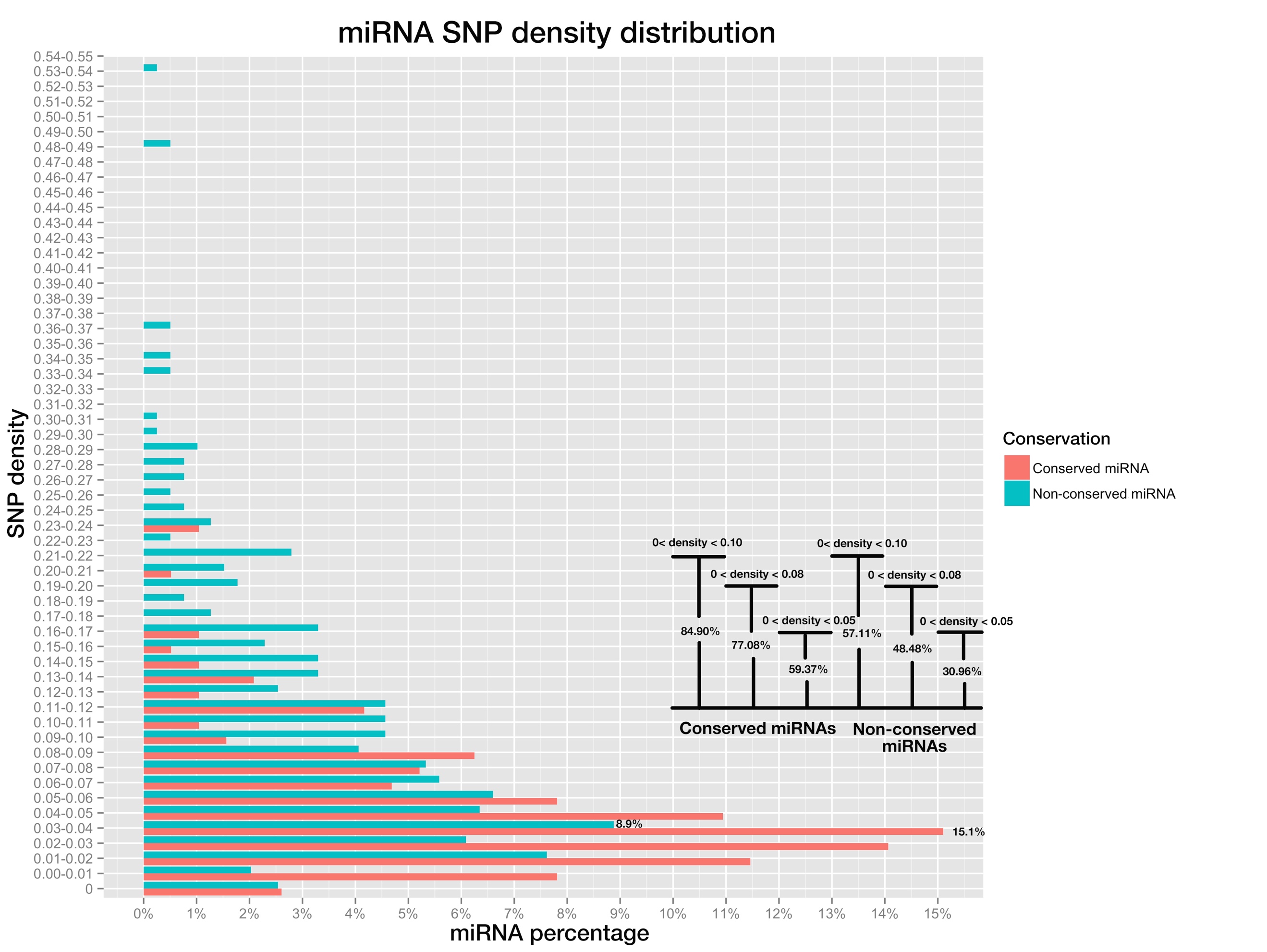
My Methodology:

* First, I obtained all the SNPs database from the 3,000 Rice Genome Project for all pre-miRNAs, and randomly chosen intergenic regions and exons;
* Then the SNP density of conserved and non-conserved miRNAs were compared;
* Lastly, the SNP density of pre-miRNAs, intergenic regions and exons were compared.

My Expectation:

* SNP density of conserved miRNAs should be lower than that of non-conserved miRNAs;
* Compared with intergenic regions, SNP density in pre-miRNAs and exon regions should be lower. But when comparing SNP density in pre-miRNAs with that of exon regions, the results would be unpredictable.

***Results and Explanations***



**Fig 1. miRNA SNP density distribution of pre-miRNAs, both conserved miRNAs (in red color) and non-conserved miRNAs (blue color).**

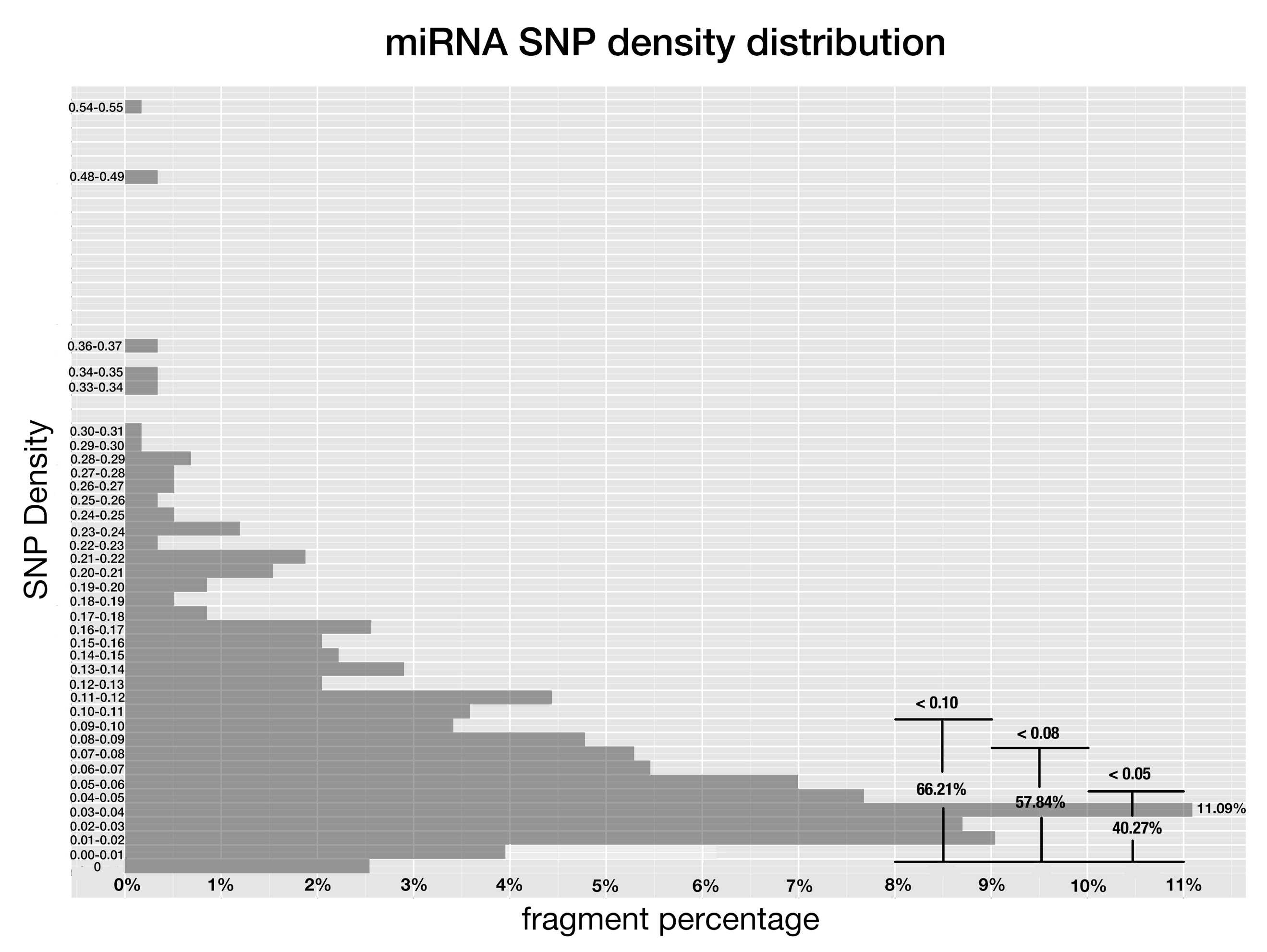
SNP density of a miRNA is the division of SNP number by miRNA length. Bar plot on bottom right shows the percentage of miRNAs whose SNP densities are below 0.10, 0.08 and 0.05, for conserved and non-conserved miRNAs respectively.

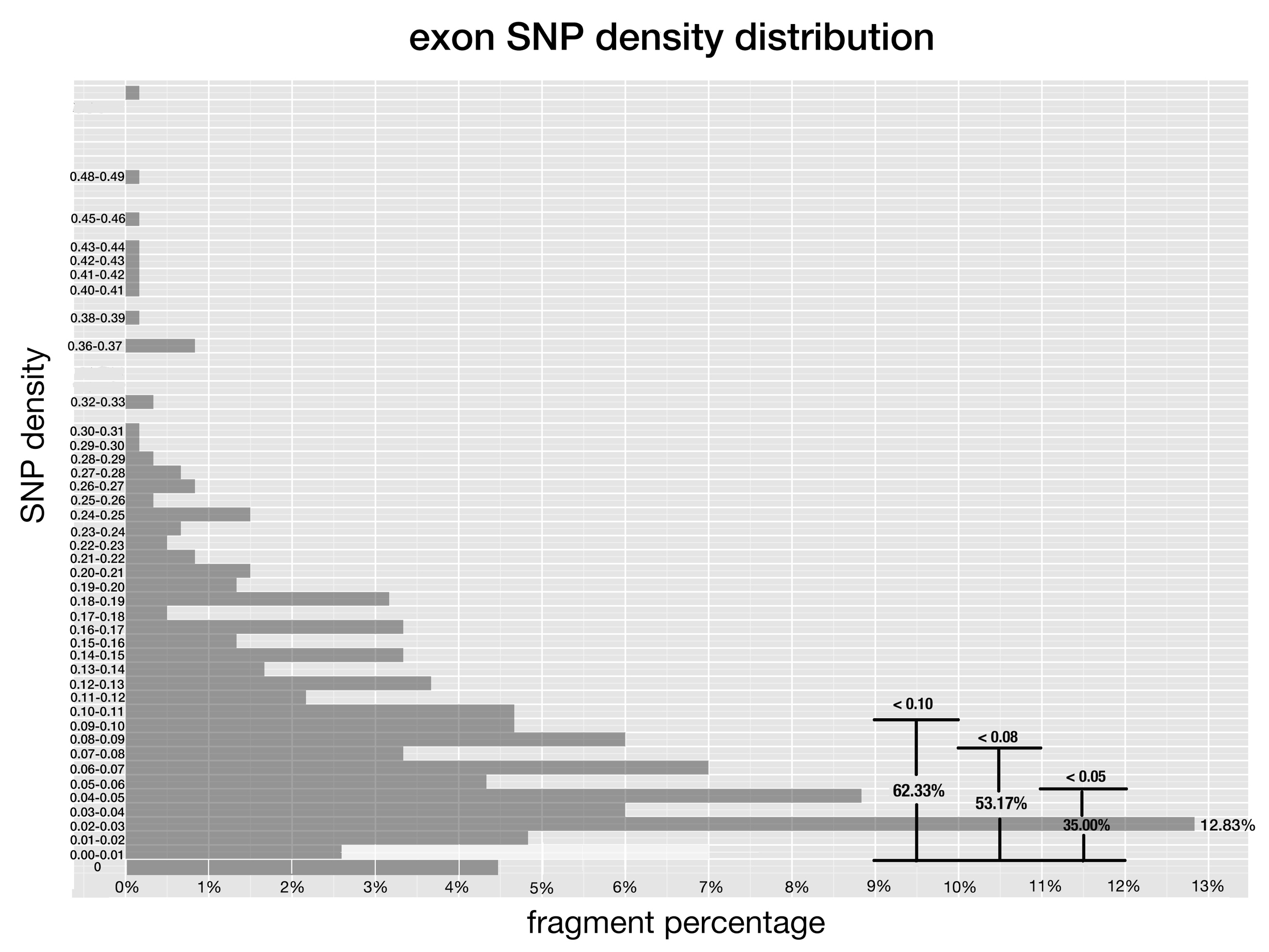
My Outcome:

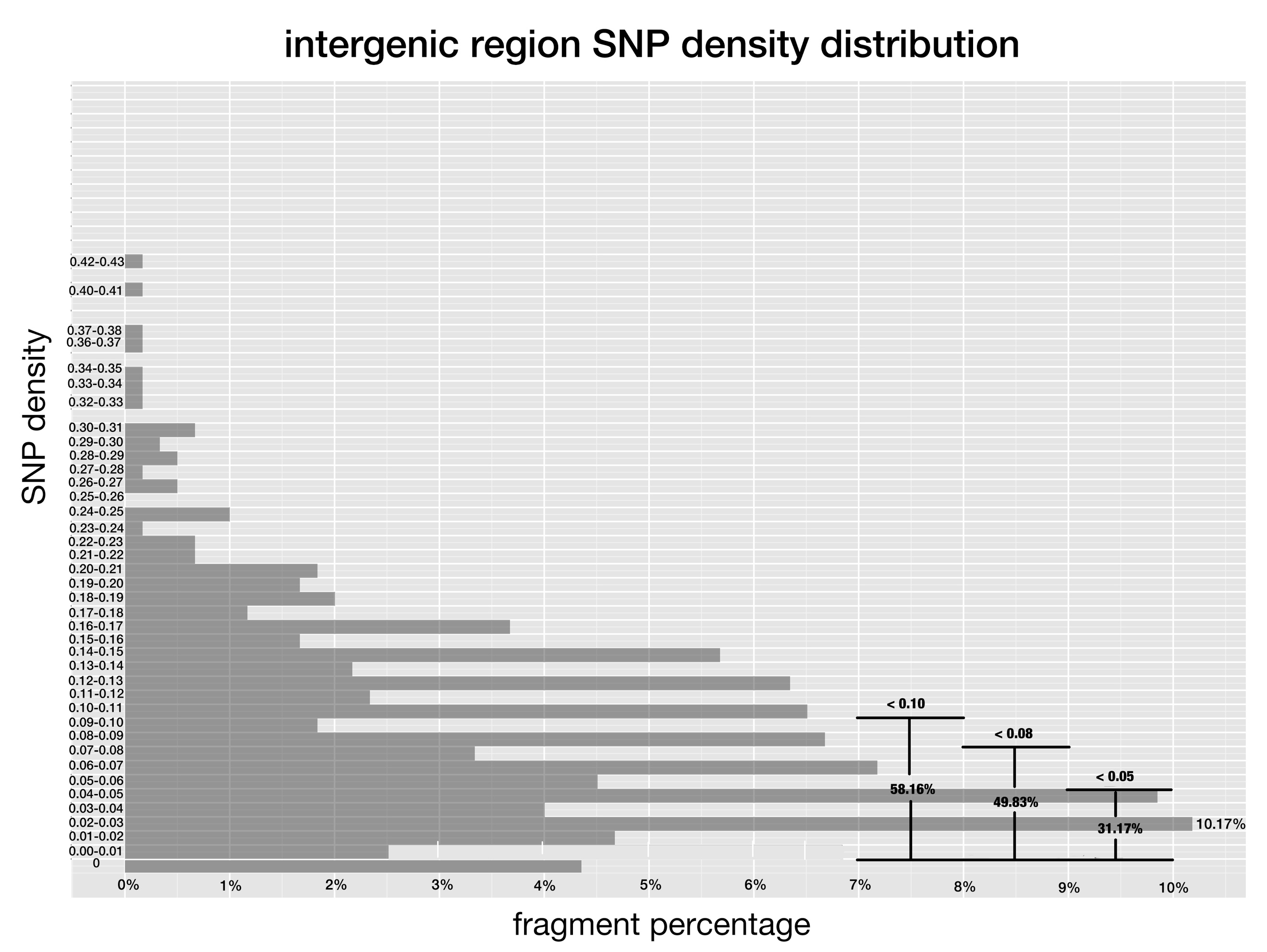
* Most conserved miRNAs and non-conserved miRNAs cluster at 0.03-0.04 range, with a percentage of 16.2% and 9.1% respectively;
* Compared with non-conserved miRNAs, conserved miRNAs cluster at lower SNP density range. As shown in the figure, 87.43% conserved miRNAs had SNP density lower than 0.1, in contrast, 62.82% non-conserved miRNAs fell into this range; while 80.69% conserved miRNAs had SNP density lower than 0.08, in contrast, 52.12% non-conserved miRNAs fell into this range; and 60.60% conserved miRNAs had SNP density lower than 0.05, while 34.66% non-conserved ones fell into this range.

My Discovery:

* Conserved miRNAs cluster at lower SNP density range than non-conserved miRNAs, as expected.







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

SNP density is defined the same as Fig 1, and x-axis corresponds to the percentage of fragments that have SNP density at given range.

My Outcome:

* Fig 2 shows that a large portion of pre-miRNAs cluster at 0.03-0.04 range, as SNP density grows from 0 to 0.04, the fragment percentage increases; while after reaching 0.04, the fragment percentage decreases dramatically as SNP density grows. 69.26%, 61.32% and 43.75% pre-miRNAs have SNP density below 0.10, 0.08 and 0.05, respectively;
* Fig 3 shows most exon regions cluster at 0.02-0.03 range and there are several peaks. 64.83%, 54.17% and 39.50% exons have SNP density below 0.10, 0.08% and 0.05, respectively;
* Fig 4 shows intergenic regions cluster at various SNP density ranges. There are 8 discrete ranges that accumulate fragment percentage to more than 5%. 59%, 50.50% and 35.50% intergenic regions have SNP density below 0.10, 0.08 and 0.05, respectively.
* Compared with intergenic regions, pre-miRNAs and exon regions cluster at lower SNP density range. There is no big distinction between pre-miRNAs and exon regions.

My Discovery:

* The difference between intergenic regions and pre-miRNAs as well as exon regions may due to the reason that pre-miRNAs and exon regions perform biological functions and thus subject to relatively higher selection pressure;

**Part II. SNP distribution along mature miRNAs and miRNA binding sites**

My Hypothesis:

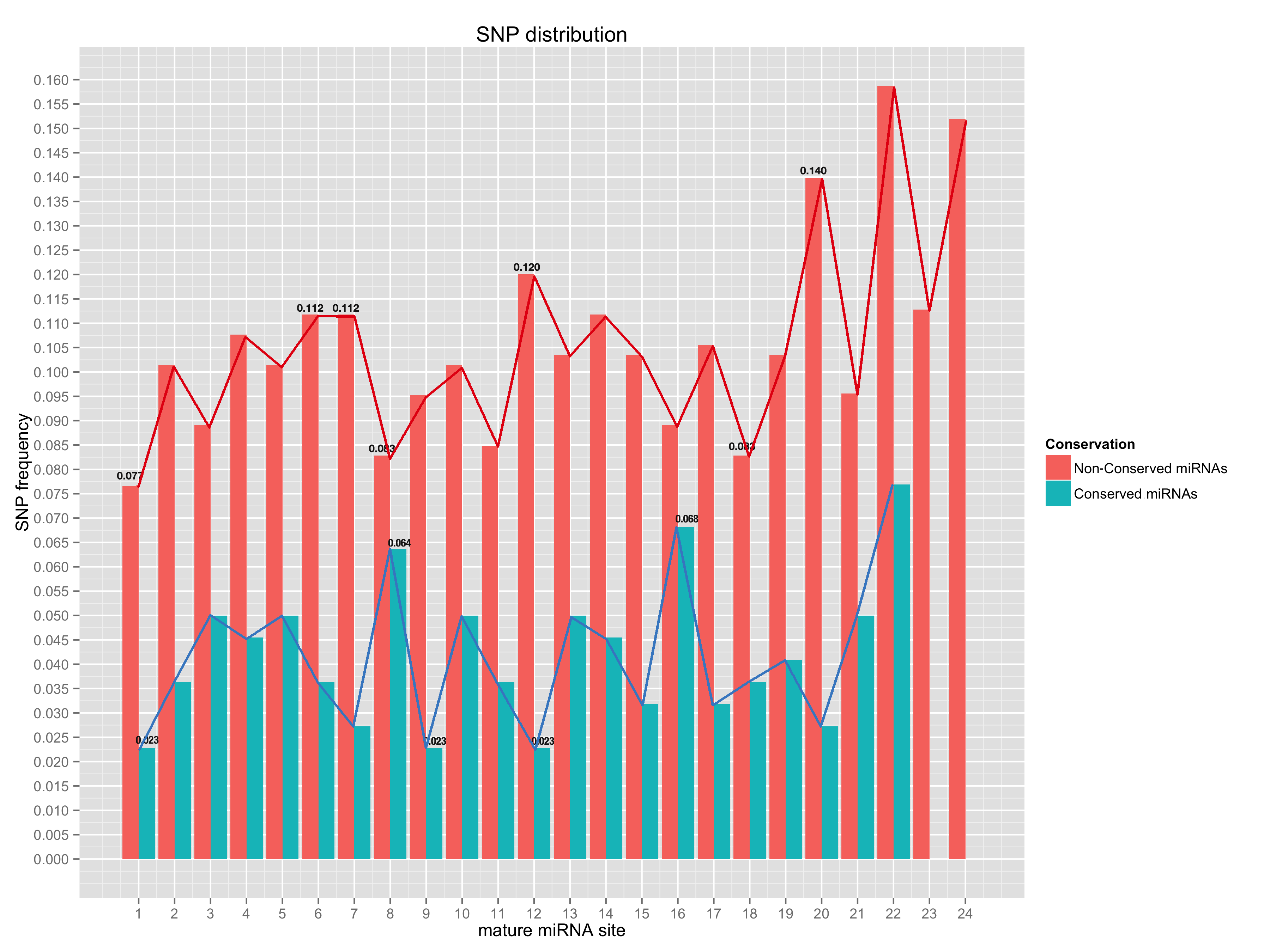
* Different sites in mature miRNAs are diversified in their importance in target recognition and binding; Meanwhile, the sites along miRNA binding sites are equally crucial in the target binding of miRNAs;
* The importance can be reflected by the SNP frequency of the sites and lower SNP frequency implies the given site may be under stricter selection pressure and may be more important for the miRNA:target interaction;
* Conserved miRNAs have more identifiable targets than non-conserved miRNAs, and are more conserved in their regulatory roles in plants through target mRNA silencing. So, we study only binding sites of conserved miRNAs as contrast to mature miRNAs;
* Since miRNA act as mRNA regulator through complementarity to miRNA binding site, the sites in both mature miRNA and miRNA binding site jointly determines the complementarity of miRNA:target, thus the SNP frequency in mature miRNA may have some correlation with that of miRNA binding site;

My Methodology:

* Firstly, I obtained the SNPs that fall in the mature miRNA regions from 3,000 Rice Genome Project;
* Then, with the help of bioinformatic prediction software and published experiment results, targets of conserved miRNAs are collected;
* Finally, SNP distribution along miRNA-binding sites are plotted.

My Expectation:

* SNP frequency of conserved mature miRNAs are lower than non-conserved miRNAs;
* Site 10 and 11 are cleavage site in miRNA silencing and are claimed that must be perfectly complementary to miRNA binding site counterparts, thus SNP frequency of site 10 and 11 should be lower than other sites;



**Fig 5. SNP distribution of all sites along mature miRNA, both conserved miRNAs (blue) and non-conserved miRNAs (red).**

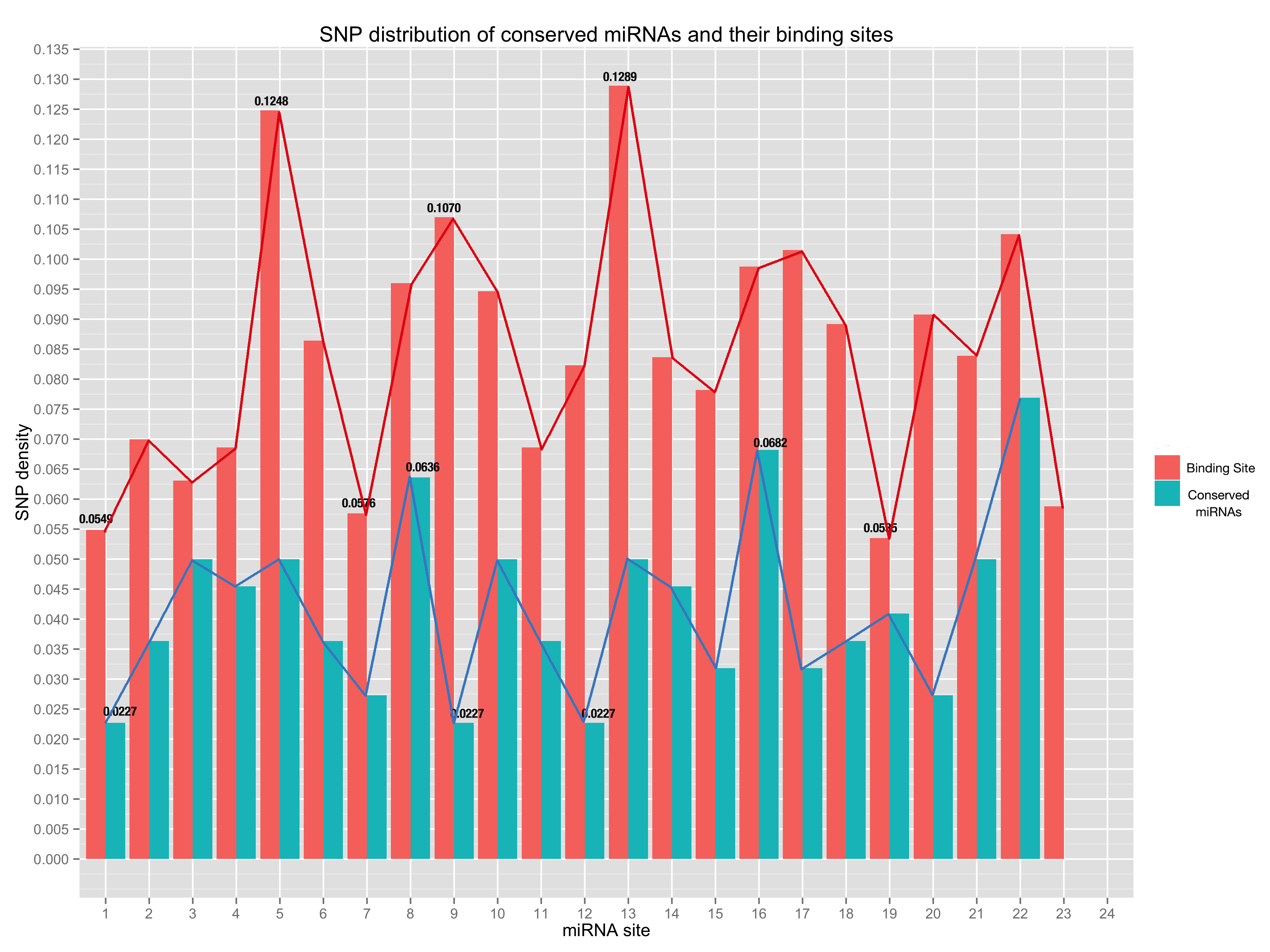
X-axis is the sites in mature miRNA from 5’end – 3’end, and y-axis is SNP frequency which is calculated as number of SNPs at this site divided by number of miRNAs;

My Outcome:

* Overall, SNP frequency of each sites are not the same in both conserved miRNAs and non-conserved miRNAs;
* SNP frequency of non-conserved miRNAs are higher than that of conserved miRNAs at each site;
* For non-conserved miRNAs, sites that have highest 4 SNP frequency are site 20 (SNP frequency, 0.140), site 12 (SNP frequency, 0.120), site 6 and 7 (both have SNP frequency 0.112); while sites that have lowest 3 SNP frequency are site 1 (SNP frequency, 0.077), site 8 and 18 (both have SNP frequency 0.083); (Only site 1 to site 21 are considered, because most miRNAs are 21-nt)
* For conserved miRNAs, sites that have highest 2 SNP frequency are site 16 (SNP frequency, 0.068), and site 8 (SNP frequency, 0.064); while sites that have lowest SNP frequency are site 1, 9 and 12 (all have SNP frequency 0.023); (Only site 1 to site 21 are considered, because most miRNAs are 21-nt);
* The standard deviation of SNP frequency of conserved miRNAs is 0.018. By contrast, standard deviation of SNP frequency of non-conserved miRNAs is 0.020. So, the fluctuation of SNP frequency of non-conserved miRNAs is a little higher than that of conserved miRNAs.

My Discovery:

* SNP frequency of non-conserved miRNAs are higher than that of conserved ones, which implies the selection pressure imposed conserved miRNAs is higher than non-conserved miRNAs;
* Site 10 and 11 are not among the lowest SNP frequency sites, which is not consistent with the empirical claim that the cleavage sites in mature shall be perfectly complementary to miRNA binding sites, and are supposed to have lower SNP frequency than other sites;
* Site 1 in both conserved miRNAs and non-conserved miRNAs have the lowest SNP frequency, and this may be explained by the fact that site 1 determines which Argonaut protein to load for mature miRNAs thus making it under high selection pressure.
* The highest SNP frequency is almost twice as the lowest SNP frequency (site 20 with SNP frequency value 0.140 and site 1 with SNP frequency value 0.77); meanwhile, the highest SNP frequency is nearly thrice as the lowest SNP frequency (site 16 with SNP frequency value 0.068 and site 1 with SNP frequency 0.023). The big difference in SNP frequency of different sites emphasize the difference of selection pressure imposed on the sites and may imply the difference of importance of different sites to the miRNA:target interaction.



**Fig 6. SNP distribution of all sites along conserved mature miRNAs and their binding sites.**

The sites of miRNA binding site are placed in the same order as mature miRNAs (from 5’end to 3’end in the miRNA);

My Outcome:

* Compared with mature miRNAs, SNP frequency of each site in miRNA binding site is higher and the standard deviation of the SNP frequency of miRNA binding site is 0.027 and thus is higher;
* For miRNA binding sites, sites with highest SNP frequency are site 12 (SNP frequency, 0.1289), site 5 (SNP frequency, 0.1248) and site 9 (SNP frequency, 0.1070); while sites with lowest SNP frequency are site 19 (SNP frequency, 0.0535), site 1 (SNP frequency, 0.0549) and site 7 (SNP frequency, 0.0576);
* For miRNA binding site, the highest SNP frequency is more than twice as the lowest SNP frequency (site 13 with SNP frequency value 0.1289 and site 1 with SNP frequency value 0.0549);
* The Pearson Correlation Coefficient between SNP frequencies of all sites in conserved mature miRNAs and that of miRNA binding sites is 0.5891 (a moderate positive linear relationship) and p-value is 0.002455 (< 0.05);

My Discovery:

* Site 1 on miRNA binding site is among the low SNP frequency sites, which is the lowest SNP frequency site in mature miRNA, and this may imply the importance of site 1 in the miRNA:target interaction;
* Concerning the difference between the highest and lowest SNP frequency in miRNA binding sites, the data reinforces the suggestion that the difference of selection pressure imposed on the sites and may imply the difference of importance of different sites to the miRNA:target interaction.
* The correlation test of SNP frequencies in conserved miRNAs and that of miRNA binding sites imply that there’s some relationship between them statistically, and can be explained by the fact that the sites on both mature miRNA and miRNA binding site jointly determines the complementarity of miRNA:target.

**Part III. Expression Correlation Analysis found no obvious negative correlation even between bona fide miRNA:target relationships**

My Premises:

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

My Hypothesis:

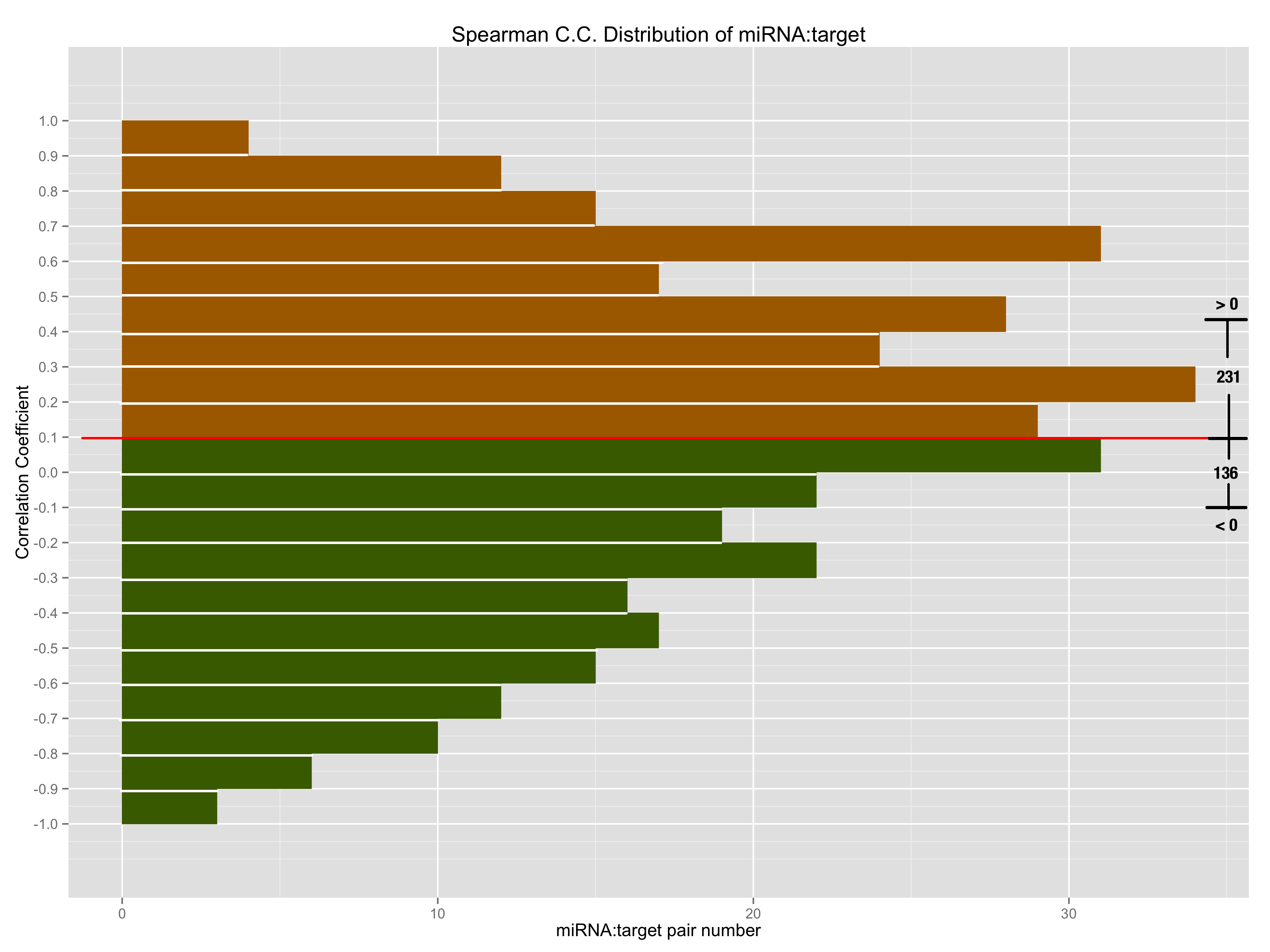
* Since most plant miRNAs regulate their target gene through miRNA-mediated cleavage, under the premises above, we can easily deduce that expression level of miRNAs are negatively correlated with that of their target mRNAs;
* And furthermore, we can use the feature of negative correlation between miRNA and target expression level to sort out bona fide targets for miRNAs;

My Methodology:

* To examine the negative correlation between miRNA and targets, firstly, I have gathered the degradome validated targets for some miRNAs from the paper *“Transcriptome-wide identification of microRNA targets in rice”* as bona fide miRNA targets; (in this paper, they used 3-week-old rice seedlings as their materials)
* Then, I downloaded expression data of miRNAs and rice genes from RiceFREND database (*RiceFREND: a platform for retrieving coexpressed gene networks in rice*);
* Later on, I calculated the spearman correlation coefficient of the identified miRNA and target with the expression data, choosing 27 DAT (day after transplanting) rice seedling as sample set (similar samples as the degradome validation experiment, which is 3-week-old rice seedlings);
* Finally, plot them on a histogram and display the distribution of the correlation coefficient.

My Expectation:

* The correlation coefficients of them are supposed to be negative for all the bona fide miRNA:target relationships;



**Fig 7. The spearman correlation coefficient of degradome validated miRNA:target relationships;**

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.

My Outcome:

* There are 136 miRNA:target pairs negatively correlated, while there are 231 pairs positively correlated. So more than half of all bona fide miRNA:target relationships are positively correlated;
* There are 197 pairs whose correlation coefficient fall in the range of -0.4~0.4 (weak relationship);

My Discovery:

* More than half of all degradome validated miRNA:target relationships are positively correlated, this objects to the hypothesis that bona fide targets are negatively correlated with their miRNAs;
* More than half of all miRNA:target pairs fall in the weak correlated range, which may imply that miRNAs and that of target mRNAs have very weak relationship concerning expression correlation.

**Part IV. Combined Complementarity Pattern Analysis Method**

**Part V. CCPA discovered miR818 family had positions recovered their pairing with two SNPs on both miRNA and binding site**