SNP analysis reveals an evolutionary acceleration of the human-specific microRNAs

Qipeng Zhang[1, 2], Ming Lu[1, 2], and Qinghua Cui[1, 2]

1. Department of Medical Informatics, Peking University Health Science Center, Peking University, 38 Xueyuan Rd, Beijing, China 100083
2. Ministry of Education Key Lab of Molecular Cardiovascular Sciences, Peking University, 38 Xueyuan Rd, Beijing, China 100083

Corresponding author: Cui, Q. (cuiqinghua@bjmu.edu.cn)

MicroRNAs are one class of important gene regulators at the post-transcriptional level by binding to the 3'UTRs of target mRNAs. It has been reported that human microRNAs are evolutionary conserved and show lower single nucleotide polymorphisms (SNPs) than their flanking regions. However, in this study, we report that the human-specific microRNAs show a higher SNP density than both the conserved microRNAs and other control regions, suggesting rapid evolution and positive selection has occurred in these regions. Furthermore, we observe that the human-specific microRNAs show greater SNPs minor allele frequency and the SNPs in the human-specific microRNAs show fewer effects on the stability of the microRNA secondary structure, indicating that the SNPs in the human-specific microRNAs tend to be less deleterious. Finally, two microRNAs hsa-mir-423 (SNP: rs6505162), hsa-mir-608 (SNP: rs4919510) and 288 target genes that have apparently been under recent positive selection are identified. These findings will improve our understanding of the functions, evolution, and population disease susceptibility of human microRNAs.

MicroRNAs (miRNAs) are a class of noncoding small RNAs (~22 nt), which function as regulators of target mRNA expression at the posttranscriptional level by binding to the 3'UTR of target mRNAs through base pairing, resulting in target mRNA cleavage or translation inhibition [1-3]. After being transcribed as primary transcripts (pri-miRNAs), miRNAs are converted into pre-miRNA and then further processed to mature miRNAs (MIRs), which are guided to target mRNAs [4-6]. It is estimated that 1-4% of human genes are miRNAs, and these miRNAs comprise one of the largest classes of gene regulators and a single miRNA can

regulate as many as 200 mRNAs[7]. There is increasing evidence has revealed that miRNAs play critical roles in many key biological processes, such as cell growth, tissue differentiation, cell proliferation, embryonic development, and apoptosis[7]. As such, the dysregulation of miRNAs and their targets may lead to the development of disease, such as cancer[7]. Currently, there are more than 70 diseases have been reported to be associated with miRNAs (see our human miRNA associated disease database: http://cmbi.bjmu.edu.cn/hmdd).

As the most common genetic variants in the human genome[8, 9], single nucleotide polymorphisms (SNPs) occurring in functional regions can affect both phenotypes and diseases susceptibility. Therefore, the SNPs occurring in miRNAs or their target sites may affect the thermodynamics of the RNA-RNA interactions between miRNAs and their target sites, which may result in the dysregulation of target genes and subsequently cause phenotype variation or disease susceptibility. Indeed, it has been reported that polymorphisms in miRNA target sites are associated with Tourette's syndrome[10], muscularity[11], and cancer susceptibility[12]. It has also been reported that polymorphisms in miRNAs are implicated in schizophrenia[13]. Because of their importance to various important biological processes, miRNAs tend to be conserved during evolution[7]. Indeed, recently, upon analyzing the SNPs in human miRNAs or target sites, several studies have reported a lower SNP density in these regions when comparing them with control regions[12-15]. However, all these studies were performed by analyzing the all human miRNAs or target sites together, which may mask some special features of specific miRNAs or target sites. Although most of the miRNAs are conserved during evolution, It is believed that there exists some fast-evolved miRNAs in

genomes and play roles in novel phenotypes and functions[16]. Up to this point, little is known about the SNP distribution in species-specific miRNAs or target sites. In this study, we evaluate whether the human-specific miRNAs still show a significant lower SNP density than control regions, whether there are some miRNAs under recent positive selection, and whether SNPs in the human-specific miRNAs show different features from other miRNAs. These questions may shed light on the functions and evolution of miRNAs.

Results

The human-specific miRNAs show higher SNP density than both the conserved miRNAs and the flanking regions

In order to answer this question, we downloaded all miRNAs from the miRBase[17] (October 2007, see **Methods**). We next classified the human miRNAs into conserved miRNAs and human-specific miRNAs (**Supplementary Table 1**) by the miRNA family annotations in miRBase (see **Methods**), and then identified SNPs in these miRNAs (see **Methods**). As shown in **Supplementary Table 2**, 110 SNPs in 83 pre-miRNAs were identified. For comparison, we also identified the SNPs occurring in the flanking regions (which are defined here as the upstream and downstream regions with specific length as pre-miRNAs) around pre-miRNAs. Generally, these pre-miRNAs exhibite a lower SNP density than their flanking regions (**Supplementary Figure 1**), which corresponds to the result of Saunders et al.[15]. We next calculated the SNP densities for the conserved pre-miRNAs, the human-specific pre-miRNAs, and their flanking regions, respectively. As a result, the human-specific pre-miRNAs show a significant higher SNP density (4.39 SNPs/kb) than the conserved

pre-miRNAs (1.67 SNPs/kb) (**Figure 1**). We tested the significance of the higher SNP density in the human-specific pre-miRNAs by Randomization test ($P<2.0\times10^{-4}$, see **Methods**). In the random case, the SNP density in the human-specific pre-miRNAs follows a normal distribution (P=0.0015, One-sample Kolmogorov-Smirnov test, mean value: 2.35 SNPs/kb, standard deviation: 0.48, maximum value: 4.37 SNPs/kb, minimum value: 0.77 SNPs/kb, **Figure 2**). This result suggests that overall the human specific pre-miRNA are under less negative selection than conserved pre-miRNAs. Furthermore, the conserved pre-miRNAs show a lower SNP density than their flanking regions (**Figure 1a**), suggesting an evolutionary conserved in the conserved pre-miRNAs. In contrast, it is interested that the human-specific pre-miRNAs show a higher SNP density than the flanking regions, suggesting that an evolutionary acceleration has occurred in these regions. This pattern is the inverse of that of the conserved pre-miRNAs (**Figure 1b**). We also performed an analysis to miRNA target sites but didn't find significant patterns (data not shown).

A similar analysis was performed on MIRs (mature miRNA) and a similar pattern was found in these regions. When compared with other regions in pre-miRNAs, the human-specific MIRs show a higher SNP density, whereas conserved MIRs showed a lower SNP density than other regions in pre-miRNAs (**Figure 3**). The significance of the higher SNP density in human-specific MIRs was also tested by Randomization test (**Figure 4,** $P<2.0\times10^{-4}$, Randomization test).

It has been reported that the seed region (base 2-7 from the 5' end of the MIR) in MIR plays a

critical role in its binding with 3'UTR of target mRNA in animals and show low SNP density [15]. In order to investigate this issue for human-specific MIRs and conserved MIRs, we calculated the SNP densities for the seed regions and the other MIR regions excluding seed regions. As a result, we found that the SNP density of seed regions in conserved MIRs is lower that that of the other MIR regions excluding the seed regions, whereas the human-specific MIRs show a converse pattern: the seed regions show a higher SNP density that the other MIR regions excluding seed regions (P=0.02, Randomization test, **Figure 5**). These results suggest that an evolutionary acceleration was occurred predominantly in the human-specific miRNAs, which in turn suggests that these miRNAs may have special functions and might play an important role in human evolution.

The human-specific miRNAs show greater SNP minor allele frequency

Since SNPs in the conserved regions are more likely to result in deleterious effects, negative selection should limit the frequency of deleterious alleles. As a result, these SNPs will show a lower minor allele frequency. Conversely, regions of evolutionary acceleration should show a higher minor allele frequency because of positive selection. We retrieved minor allele frequency data from four populations, YRI (Yoruba in Ibadan, Nigeria), CEU (Utah residents with ancestry from northern and western Europe), CHB (Han Chinese in Beijing), and JPT (Japanese in Tokyo) (**Supplementary Table 3**). We compared the minor allele frequency for SNPs in human-specific miRNAs and conserved miRNAs. As a result, the SNPs in human-specific miRNAs show a trend of greater minor allele frequency (median minor allele

frequency: 0.078 vs. 0.061, P=0.14, Wilcoxon test), although the result is not significant.

SNPs in the human-specific miRNAs affect less on the stability of miRNA secondary structures

A suitable secondary structure is necessary for miRNA to play its role normally. As we observed above, the conserved human miRNAs show a lower SNP density, indicating that SNPs in these miRNAs tend to be more deleterious and then are under stronger negative selections. This observation suggests that SNPs in the conserved miRNAs might have stronger effects on the stability of these miRNAs' secondary structures, which may result from the largely changed free energy of their secondary structures. In order to dissect this issue, we calculated the free energy of secondary structure for each miRNA with different alleles using RNAfold[18] (**Supplementary Table 4**). We next calculated the absolute difference (AD) of free energy for each miRNA with different alleles. We compared the AD of the conserved miRNAs with that of the human-specific miRNAs and found that the conserved miRNAs have greater AD than the human-specific miRNAs as we suspected (median 2.38 vs 1.72, P=0.059, Wilcoxon test). This finding suggests that SNPs occurring in the conserved miRNAs are more likely to affect the stability of miRNA secondary structures. Therefore, the SNPs in the conserved miRNAs are more likely to be functional.

The miRNAs and target sites under recent positive selection

Furthermore, we identified human miRNAs and target sites that have been under recent positive selection based on iHS (integrated haplotype score) test[19] (see **Methods**). As a result, we identified two miRNAs hsa-mir-423 (SNP: rs6505162, |iHS|=2.4376, Asia population) and hsa-mir-608 (SNP: rs4919510, |iHS|=2.8658, Africa population), and 288 target genes (**Supplementary Table 5**) which show high iHS values, indicating that these regions have likely been under recent positive selection. These miRNAs and targets are potential factors that may play important roles in population evolution. Therefore, the functions and roles of these two miRNAs become very important for us to understand human evolution, which, however, remains largely unknown. It was reported that intronic miRNAs (miRNAs that locates within the introns of protein-coding genes) often co-express with host genes, indicating that miRNA often share common functions with their host genes[20]. Therefore, we can predict the functions of an intronic miRNAs through the functions of its host gene. Fortunately, both hsa-mir-423 and hsa-mir-608 are intronic miRNAs. And the host genes of has-mir-423 and has-mir-608 are CCDC55 and SEMA4G, respectively. Both genes are highly expressed in brain and play important roles in nuclear speck and nervous system development, respectively (according to the annotation of GeneCard, http://www.genecards.org/index.shtml). This finding suggests that these two miRNAs may be involved in brain functions and may do some benefits to people survival during population evolution, and therefore are under positive selection. Furthermore, analysis of the function of target genes under recent positive selection revealed that these genes are enriched in metal binding, ion binding, phosphoinositide binding, nucleotide binding, protein transport, and cell cycle control. The locations of these target genes are enriched in membrane region (see

Methods and **Supplementary Table6**).

Discussion

In the previous studies, it has been reported that miRNAs and target sites show a lower SNP density than control regions[12, 14, 15]. However, here, for the first time we revealed a higher SNP density occurred in the human-specific miRNAs than control regions, suggesting that positive selection and evolutionary acceleration has occurred in these regions during human evolution. As miRNAs are involved in gene regulations, these putative regions of evolutionarily accelerated may contribute to various biological processes and play critical roles in disease or population disease susceptibility. We believe that our finding concerns these regions taken with previous reports about other human accelerated regions[21, 22] may improve our understanding of human origin, evolution, and diseases. The human-specific miRNAs also show a trend of greater minor allele frequency. We observed that SNPs in the conserved miRNAs tend to have greater effects on miRNAs' functions than that in the human-specific miRNAs through an analysis of miRNA secondary structure. Finally, we identified two miRNAs hsa-mir-423 and hsa-mir-608, and 288 target genes that have been under recent positive selection. Both the two miRNAs are highly expressed in brain and play important roles in nuclear speck and nervous system development, suggesting that these miRNAs may do some benefits to population survival during evolution. The enrichment of ion binding and membrane genes in these target genes suggests that these genes may play critical roles in response to external stimuli, regulation of neural physiology, regulation of hormone, and

regulation of development. Finally, we would like to suggest that the same analysis employed in this study could be readily applied to other species.

Methods

Human miRNAs data and classification

We download all the miRNAs precursor sequences, mature miRNA sequences, and miRNA genome coordinates data from miRBase[17] (http://microrna.sanger.ac.uk/sequences/index.shtml) on October 2007, in which there are 533 human miRNAs records, but only 528 records have identified genome position (according to NCBI_Human_Genome_BUILD_36.). We classified the human miRNAs into conserved miRNAs and human-specific miRNAs according to the family annotations in miRBase. If a human miRNA has at least one family member in other species, we assigned it to the conserved miRNA group; otherwise we assigned it to the human-specific miRNA group (Supplementary Table 1).

Human SNP data

We downloaded human SNP data from dbSNP by using the University of California, Santa Cruz (UCSC) genome browser[23]. We downloaded the genotype data from International HapMap Project (release 22). We downloaded the SNP allele frequency data in four populations from HapMap by HapMart.

The identification of SNPs in human miRNAs and calculation of SNP density

According to the genome coordinates of SNPs, miRNA precursors, mature miRNAs, and flanking regions, we identified SNPs that locate within these regions by using the Galaxy Intersect tool (http://main.g2.bx.psu.edu/). The SNP densities are calculated as the number of SNPs occurring in one region (for example, miRNA precursors) divided by the total number

of bases in that region.

The identification of miRNAs and target sites that are under recent positive selection

iHS is a statistic developed to detect recent positive selection[19]. Normally, an extreme iHS value (|iHS|>2) of a SNP allele means recent positive selection. We downloaded all iHS scores and ancestral states by population and chromosome for Hapmap phase 2 data from http://hg-wen.uchicago.edu/selection/haplotter.htm. By mapping reference SNP number, we identified two SNPs (rs6505162 and rs4919510) in two miRNAs (hsa-mir-423 and hsa-mir-608) and 320 reference SNP numbers in 288 target genes are under recent positive selection.

Statistical computing

We tested the significance of SNP density of some specific regions, such as human-specific pre-miRNAs, mature miRNAs, and seed regions based on Randomization test. For example, in order to test the significance of high SNP density of human-specific pre-miRNAs (129 human-specific pre-miRNAs), we randomly picked up 129 pre-miRNAs and calculated the SNP density from the 528 miRNAs for 5000 times. We then can compare the random SNP densities with the real SNP density in the human-specific pre-miRNAs and calculate the P value, which indicates that statistical significance of its SNP density. Based on this method, we tested the significance of results for pre-miRNAs, mature-miRNAs, and seed regions.

The One-sample Kolmogorov-Smirnov test and other statistical computations were performed by R, a free statistical software package (http://www.r-project.org/).

Functional enrichment analysis of the 288 target genes that are under recent positive selection

For the 288 recent positively selected target genes identified from by iHS test, we performed the functional enrichment analysis using David Bioinformatics tool (http://david.abcc.ncifcrf.gov/home.jsp). We set the 288 target genes as "Current Gene List" and the human whole genome genes as "Current Background". We then performed function enrichment analysis by Function Annotation Clustering tool (Supplementary Table 6).

Acknowledgements

We thank Dr. Edwin Wang and Dr. Yinghai Deng for advices and corrections of this manuscript. We would also like to thank Prof. Michael A. McNutt for review and comments on this manuscript. This work is supported by the 985 project of Peking University (No. 985-2-108-121).

References

1. Ambros, V. (2004) The functions of animal microRNAs. *Nature* 431, 350-355
2. Bartel, D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281-297
3. Meister, G., and Tuschl, T. (2004) Mechanisms of gene silencing by double-stranded RNA. *Nature* 431, 343-349
4. Carmell, M.A., and Hannon, G.J. (2004) RNase III enzymes and the initiation of gene silencing. *Nat Struct Mol Biol* 11, 214-218
5. Cullen, B.R. (2004) Transcription and processing of human microRNA precursors. *Mol Cell* 16, 861-865
6. Kim, V.N. (2005) MicroRNA biogenesis: coordinated cropping and dicing. *Nat Rev Mol Cell Biol* 6, 376-385
7. Esquela-Kerscher, A., and Slack, F.J. (2006) Oncomirs - microRNAs with a role in cancer. *Nat Rev Cancer* 6, 259-269
8. Kruglyak, L., and Nickerson, D.A. (2001) Variation is the spice of life. *Nat Genet* 27, 234-236
9. Reich, D.E., *et al.* (2003) Quality and completeness of SNP databases. *Nat Genet* 33, 457-458
10. Abelson, J.F., *et al.* (2005) Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science* 310, 317-320
11. Clop, A., *et al.* (2006) A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nat Genet* 38, 813-818
12. Yu, Z., *et al.* (2007) Aberrant allele frequencies of the SNPs located in microRNA target sites are

potentially associated with human cancers. *Nucleic Acids Res* 35, 4535-4541

13. Hansen, T., *et al.* (2007) Brain expressed microRNAs implicated in schizophrenia etiology. *PLoS ONE* 2, e873

14. Chen, K., and Rajewsky, N. (2006) Natural selection on human microRNA binding sites inferred from SNP data. *Nat Genet* 38, 1452-1456

15. Saunders, M.A., *et al.* (2007) Human polymorphism at microRNAs and microRNA target sites. *Proc Natl Acad Sci U S A* 104, 3300-3305

16. Zhang, R., *et al.* (2007) Rapid evolution of an X-linked microRNA cluster in primates. *Genome Res* 17, 612-617

17. Griffiths-Jones, S. (2004) The microRNA Registry. *Nucleic Acids Res* 32, D109-111

18. Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., Schuster, P. (1994) Fast Folding and Comparison of RNA Secondary Structures. *Monatshefte f. Chemie* 125, 167-188

19. Voight, B.F., *et al.* (2006) A map of recent positive selection in the human genome. *PLoS Biol* 4, e72

20. Haygood, R., *et al.* (2007) Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat Genet* 39, 1140-1144

21. Pollard, K.S., *et al.* (2006) An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443, 167-172

22. Sabeti, P.C., *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* 449, 913-918

23. Karolchik, D., *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res* 31, 51-54

Figure Legends

Figure 1 SNP densities in human pre-miRNAs and their flanking regions. (**a**) SNP densities in conserved human pre-miRNAs and their flanking regions (**b**) SNP densities in human-specific pre-miRNAs and their flanking regions.

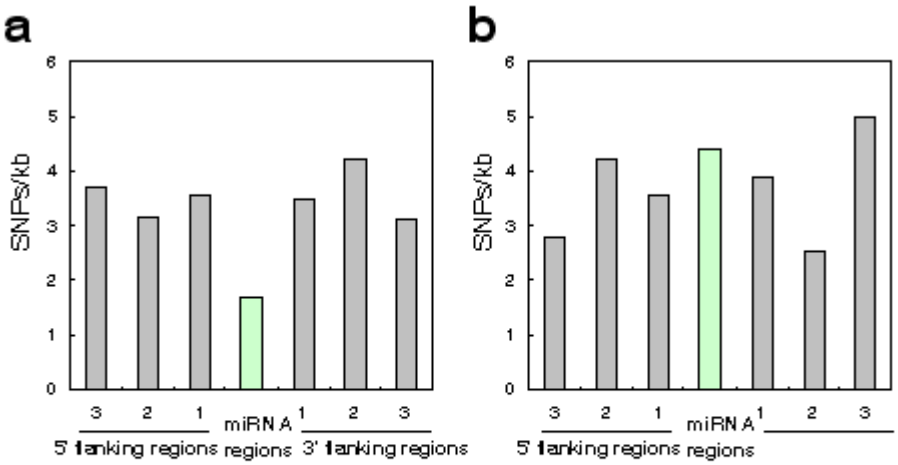


Figure 1

Figure 2 The distribution of SNP density in 5000 random sets of 129 human-specific miRNAs (blue triangle) and the SNP density of real human specific miRNAs (the value indicated by the red arrow).

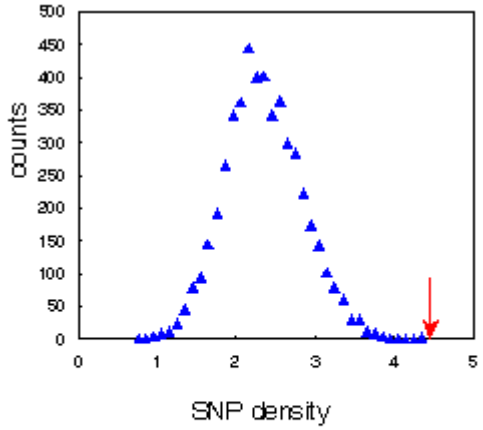


Figure 2

Figure 3 SNP densities in mature miRNAs (MIRs) and pre-miRNAs regions excluding MIR. (**a**) indicates conserved miRNAs and (**b**) indicates the human-specific miRNAs.

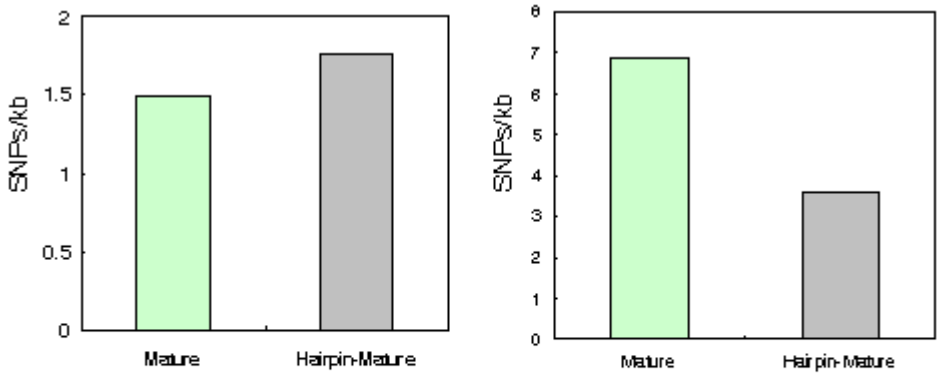


Figure 3

Figure 4 The distribution of SNP density in 5000 random sets of 129 human-specific mature miRNAs (MIRs) (blue triangle) and the SNP density of real human specific MIRs (the value indicated by the red arrow).

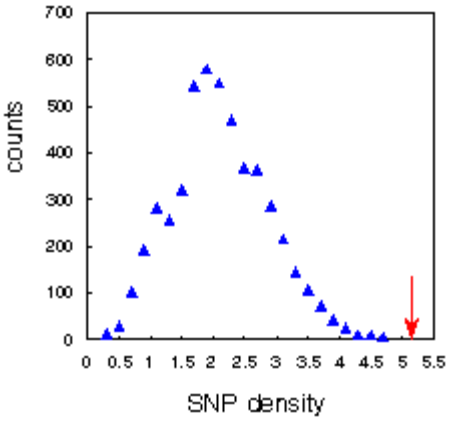


Figure 4

Figure 5 SNP densities in seed regions of mature miRNAs (MIRs) and MIR regions excluding seed regions. (**a**) indicates conserved miRNAs and (**b**) indicates the human-specific miRNAs.

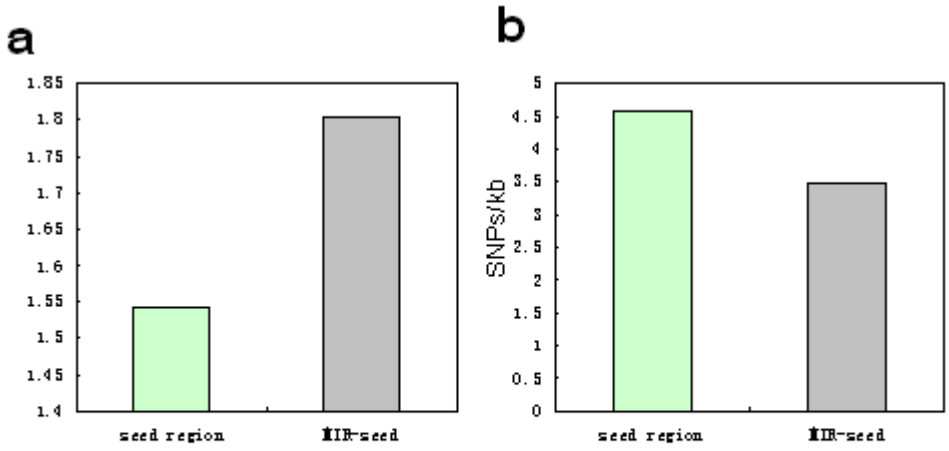
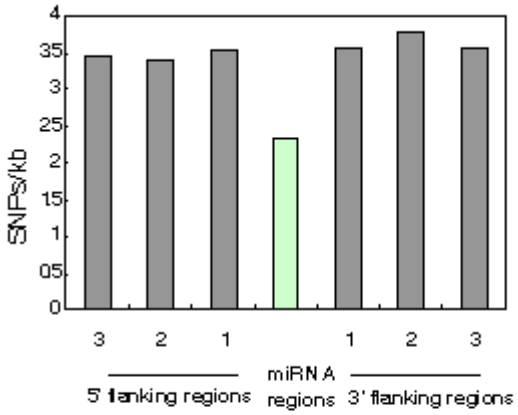


Figure 5

Supplementary Figure 1 SNP densities in total human pre-miRNAs and their flanking regions



Supplementary Figure 1