Role of SNPs in the biogenesis of mature miRNAs

Ying Wang[1, 2], Jidong Ru[3, *],Xianglian Meng[4]

[1] College of Equipment control, Shenyang Ligong University, No.6, nanping middle road, hunnan new district, Shenyang, Liaoning, 110159, China, [2] College of Computer and Control Engineering, Qiqihar University, No.42, Wenhua Street, Qiqihar, Heilongjiang, 161006, China, [3] College of Light Industry and Textile, Qiqihar University, No.42, Wenhua Street, Qiqihar, Heilongjiang, 161006 China, [4]School of Computer Information & Engineering, Changzhou Institute of Technology, 213032, China

Email addresses of the authors

YW, wangying0129@126.com; RJD, rujidong@126.com; MXL,mengxl@czust.edu.cn.

[#]Correspondence address:

Jidong Ru, PhD

Professor

No.42, Wenhua Street

Qiqihar University

China

E-mail: rujidong@126.com

Keywords

SNP; isomiRs; miRNA; splicing mechanism;

Abstract

Background

SNPs within pre-miRNA regions play a significant role in miRNA generation, processing and function by different molecular mechanisms and are thought to be major contributors to the variations in phenotypes and diseases. Therefore, whole-genome analysis of how SNPs affect mature miRNA biogenesis is important for precision medicine.

Results

In this study, aiming to analyze the role of SNPs in mature miRNA biogenesis genome-wide, we constructed a SNP-pre-miRNA database, named miRSNPBase, consisting of 886 pre-miRNAs and 2640 SNPs based on the latest data. Then, we identified 10574 SNP-pre-miRNAs based on 886 pre-miRNAs and their associated 2640 SNPs, and we performed genome-wide association analyses to identify isoform miRNAs (isomiRs) based on miRFind that are associated with the mechanism of SNPs affecting miRNA maturation. A total of 4235 nor-SNP-pre-miRNAs based on 480 nor-pre-miRNAs and 1250 nor-SNPs were identified. We analyzed the effects of SNP type, SNP location and SNP-mediated free energy change during mature miRNA biogenesis and found that they are closely related to mature miRNA biogenesis. In addition, the MAF distribution of the iso-pre-miRNAs and nor-SNPs was analyzed based on the 1000 Genomes Project. The results demonstrated that individuals who contained the iso-SNPs were in the minority, and those who contained the nor-SNPs were in the majority. Notably, to verify our method and identify important biomarkers, we identified isomiRs and iso-SNPs in 18 GBR individuals of European origin. In the results, 209 iso-pre-miRNA candidates and 71 verified iso-pre-miRNAs of the 18 GBR samples were identified, and 2667 isomiRs of 209 pre-miRNAs were verified by miRNA sequencing data.

Conclusions

Our results clearly indicated that SNPs that altered the mature miRNA splicing mechanism and led to the production of isomiRs, were closely related to and affected normal life processes, and led to epigenetic changes and diseases.

Background

MicroRNAs (miRNAs) are single-stranded small noncoding RNAs 20~24 nucleotides (nt) in length that regulate gene expression at the posttranscriptional level. It is estimated that miRNAs regulate approximately 60% of the transcription process in humans. The abnormal expression of miRNAs is closely related to many diseases, and more than 50% of human miRNAs are located in gene fragment sections and are related to cancers, including breast cancer, lung cancer, colorectal cancer, skin cancer, nasopharyngeal carcinoma, ovarian cancer, and nerve cell carcinoma. In addition, single nucleotide polymorphisms (SNPs) are DNA sequence polymorphisms caused by a single individual diversity variation at the genome level. More than 38 million SNPs, including 140 million small insertion/deletion events and 14,000 structure variations, exist in the human genome and contribute to human phenotypic differences and diseases as molecular markers identified in different research fields [1]. GWAS have reported 1300 variations and 200 phenotypes [2, 3]. Remarkably, approximately 88% of the variations in disease-related SNPs are the most abundant form of noncoding region variations in the human genome, thousands of SNPs exist in the miRNA sequences and their upstream and downstream flanking regions, and more than 40% of pre-miRNAs contain at least one SNP [4]. SNPs within pre-miRNA regions may be responsible for several of the reported associations between SNPs, miRNAs and complex human phenotypes and diseases.

MiRNAs are transcribed from genomic DNA and are spliced in several steps that heavily depend on correct secondary and tertiary structures. The secondary structure is determined by the RNA sequence, which is in turn determined by the genomic sequence. The nucleotide composition of miRNA sequences plays important roles in the biogenesis of mature miRNAs, and variations affect the molecular structure, thermodynamic stability and functional strand selection of miRNA sequences. The sequence features of pre-miRNAs are arbitrary but follow some rules, such as the striking frequency of GGAC in pre-miRNA sequences [5] and the strong preference for 'U' in the first nucleotide position of the 5' arm of mature miRNAs [6].

The SNPs within pre-miRNA regions alter pre-miRNA sequences by adding, replacing, or deleting nucleotides [7], and the structure and the minimum free energy of SNP-related miRNAs are altered [8, 9]. For example, 11 SNPs located in the stem of pre-miRNAs reduce the number of mature miRNA products, and the thermodynamic changes of more than 44% of pre-miRNAs are greater than 2.0 kcal/mole [10].

SNPs within pre-miRNA regions affect the required secondary structure and thermodynamics and alter the miRNA maturation process, including Drosha enzyme processing, Dicer enzyme processing and functional strand choice [6, 7], which are closely related to a variety of phenotypes and diseases [11, 12]. The processing of quite a few pre-miRNAs has been reported to be altered by SNPs; for example, miR-125a, which contains a SNP in the seed region, produces one mature miRNA in one arm, and the miRNA from the other arm is prevented from

undergoing Drosha enzyme processing[9]. Specifically, SNPs change the pre-miRNA processing sites[13]; for example, miR-934, which contains rs73558572, generates five mature miRNAs that are offset (1-2 nt) from the 3’ arm reference mature miRNA[14]. As a result, SNPs lead to imprecise precursor cropping or dicing and affect the expression level of miRNA[15]. For example, when the terminal loop of pre-miRNA is changed, mature miRNA levels are reduced by 60% [16]. Sun et al [14] demonstrated that this process is one of the mechanisms for isomiR generation[17].

Although a multitude of SNPs have been found to play a significant role in miRNA generation, processing and function by different molecular mechanisms and are closely related to various diseases and phenotypes[14], exactly how SNPs affect mature miRNA biogenesis remains unclear; therefore, studies on SNP-affected miRNA maturation mechanisms can provide evidence for causal SNPs and contribute to precision medicine.

In this study, our aim was to perform a genome-wide analysis of the role of SNPs in the biogenesis of mature miRNAs. We present a database, miRSNPBase, that provides comprehensive information about SNPs and SNP-related miRNA loci. Then, all the sequences of pre-miRNAs containing SNPs were surveyed using the position in the human genome as the correlation factor. Four splice sites of each SNP-pre-miRNA were predicted based on miRFind to further identify all the normal miRNAs and isomiRs. Based on the relationship with the produced miRNAs, pre-miRNAs and SNPs were divided into iso-pre-miRNAs and iso-SNPs,

which lead to isomiR production, and nor-SNPs and nor-pre-miRNAs, which contribute to normal miRNA production. Furthermore, the isomiRs and iso-SNP candidates were presented for subsequent studies. Aiming to evaluate how the SNPs affect miRNA maturation mechanisms and penetrance in individuals, the minor allele frequencies of the SNPs included in the normal sequences and isomiRs were statistically analyzed. In addition, we analyzed the effects of SNP type, SNP location and SNP-affected free energy change during the biogenesis of mature miRNAs. Specifically, we verified the predicted isomiRs based on miRNA sequencing data for 18 GBR individuals. The schematic of the overall method is illustrated in Figure 1.

Methods

Data

SNP-related dataset

Information on SNPs was obtained from the 1000 Genomes Project (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/) which was constructed based on 2,504 individuals from 26 populations and included over 84.7 million SNPs and >99% of SNPs with a frequency of >1% for a variety of ancestries.

MiRNASNP dataset

The miRNASNP dataset (http://bioinfo.life.hust.edu.cn/miRNASNP2/download.php) is a database that catalogs a total of 2257 SNPs in 1596 human pre-miRNAs based on miRBase version 19 and the dbSNP database (version 137).

Validation data

The validation data include VCF files, the miRNA sequencing data of GBR (British in England and Scotland) and the Homo sapiens GRCh37 reference sequence.

The 1000 Genomes project consists of a total of 26 subpopulations from five major populations (Americans, Europeans, East Asians, South Asians and Africans). The VCF file contains the final variant call set with phased genotypes for chr1-22, chrX and chrY consisting of 2504 individuals from 26 populations based on the phase3 analysis of the 1000 Genomes sequence data. Eighteen GBR individuals of European origin were selected for our study. We downloaded the VCF files of GBR population-specific SNP data of the 18 individuals from the ftp server of the 1000 Genomes project (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/)[18, 19].

The miRNA sequencing data consist of the set of human lymphoblastic cell line samples from the GBR population. is the dataset was downloaded from https://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-2/samples/.

The *Homo sapiens* GRCh37 reference sequence was from 1000 Genomes project phase 2 and was downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_g1k_v37.fasta.gz.

Identification of SNP-related miRNAs (SNP-miRNAs)

The attribute of miRNA genome position information is described as:

$\mathrm{Attr}_{\mathrm{mir\_pos}}$=(Hogg and Harries);

The attributes of the SNPs are listed as follows:

Attr_{snp}={ Chromosome, Coordinates, SNPID, Reference, Alter, Qual, Filter, Number of allele, Frequency of allele, Amount of allele, Type of variation, Read of the variation };

The position information of miRNAs and SNPs were mapped based on the hg38 and hg19 coordinates, and liftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver) was used to convert the coordinates from hg38 to hg19. Then, based on the coordinates, all SNPs were mapped to the pre-miRNAs, and the SNP extraction conditions are described as $\mathrm{Attr}_{\mathrm{snp}}\{\mathrm{Coordinates}\} \in \mathrm{Attr}_{\mathrm{mir\_pos}}\{\mathrm{End} - \mathrm{Start}\}$. All identified pre-miRNAs and SNPs were used to construct our database miRSNPBase. The SNP-pre-miRNA information in miRSNPBase is defined as $\mathrm{Attr}_{\mathrm{SNP\_pre}}$= {name, plus/trans strand, SNP positions, start of miRNA, end of miRNA, reference nucleotide, alter nucleotide, minor allele frequency}, among them, minor allele frequency, MAF, is the proportion of the SNP variable genes in the species.

Because our miRSNPBase was constructed based on the 1000 Genomes project, we expanded our database with miRNASNP, which was developed on the basis of NCBI dbSNP (version 137).

Identification of four splicing sites of SNP-pre-miRNA

To construct the SNP-pre-miRNA sequences, SNPs were mapped to pre-miRNA sequences based on coordinates. To avoid favoritism, SNP-pre-miRNA sequences of each SNP-associated

pre-miRNA were constructed based on the combination method, and the specific nucleotides of corresponding positions in the pre-miRNAs were substituted by SNPs. The number of SNP-pre-miRNAs is shown as:

$$N_i = C_n^1 + C_n^2 + \cdots + C_n^k + \cdots + C_n^n \qquad (1\text{-}1)$$

where N_i is the number of SNP-pre-miRNAs for the ith pre-miRNA and n is the number of SNPs mapping to one pre-miRNA based on coordinates. An example of constructing a SNP-pre-miRNA sequence is illustrated in Figure 2:

In this process, the SNP position is defined based on the coordinates of the plus or trans strand. For two different situations, we used two different conversion methods. The calculation methods for the variation position of the plus strand and trans strand are illustrated in Figure 3:

To systematically identify mature miRNAs of the SNP-pre-miRNAs, miRFind software developed in our previous work [20] was used to identify the splicing sites. miRFind is software aiming to identify mature miRNAs within pre-miRNA, and it provides five mature miRNA candidates with an accuracy as high as 68%. We defined the start and end sites of the 5' arm mature miRNA as P5_5 and P5_3, similarly, the start and end sites of the 3' arm mature miRNA as P3_5 and P3_3. Based on the identified mature miRNAs, we extracted the normal miRNAs (nor-miRNAs) and isomiRs. After counting the splicing site offsets between the five predicted miRNA candidates and reference mature miRNAs, we defined the SNP-pre-miRNAs, pre-miRNAs and SNPs whose splicing sites were changed as iso-SNP-pre-miRNAs, iso-pre-miRNAs and iso-SNPs, respectively, and not changed as nor-SNP-pre-miRNAs, nor-pre-miRNAs and nor-SNPs, respectively. Specifically, a nor-iso-pre-miRNA is defined as a pre-miRNA with different SNPs that produces normal miRNAs and isomiRs at the same time.

Effects of SNPs on the biogenesis of mature miRNAs

To research the effects of SNP position, SNP type and SNP-affected free energy change on the mature miRNA splicing mechanism, we investigated the enrichment of the iso-pre-miRNAs and nor-pre-miRNAs in the SNP positions and the distribution of SNP types of iso-pre-miRNAs and nor-pre-miRNAs. Moreover, we researched the enrichment of SNP-pre-miRNAs based on free energy change by extracting the free energy changes between the iso-SNP-pre-miRNAs and normal pre-miRNAs and between the nor-SNP-pre-miRNAs and normal pre-miRNAs. The free energy of each normal pre-miRNA and SNP-pre-miRNA was calculated using RNAfold[21].

The MAF of SNPs in the iso-pre-miRNA and nor-pre-miRNA

The iso-SNPs that potentially affect mature miRNA biogenesis are functional SNPs. To research the distribution of iso-SNPs and nor-SNPs in humans, we obtained the MAF of the nor-SNPs and iso-SNPs to analyze the rules of SNP genes in the population based on the 1000 Genomes project. Specifically, to make the investigation results more representative, we only focused on iso-SNPs and nor-SNPs.

Identification and verification of isomiRs based on a GBR population from 1000 Genomes

To verify our method and identify important biomarkers, the VCF files of GBR population-specific SNP data and related miRNA sequencing data were selected in our study. First, we extracted the SNPs of each chromosome based on the VCF files. In this process,

GenomeAnalysisTK.jar was used to compare the samples with the human GRCh37 reference sequence to extract chromosome data and VCF files of the 23 chromosomes. Second, we integrated all the SNPs of the 23 chromosomes. Based on miRSNPBase, we obtained the pre-miRNA-SNP dataset of each sample and then the pre-miRNA-SNP sequences of this sample we constructed. Third, we identified four Dicer sites of the pre-miRNA-SNP sequences, and then the canonical miRNAs and isomiR candidates were extracted. Finally, we aligned all the isomiRs with miRNA sequencing data of the corresponding sample. The miRNAs that can be found in the miRNA sequencing data are the verified isomiRs.

Results

Establishment of the miRSNPBase database

A total of 1881 human pre-miRNAs were extracted based on the miRNA genome position information. On the basis of the coordinates in the genome, the SNPs were mapped to the pre-miRNAs and flanking regions, and we found 2146 SNPs located in the pre-miRNAs. Among them, 995 SNPs were found in the dbSNP, and 1151 SNPs were not in the dbSNP. The results demonstrate that our method affords a great degree of data integrity. As such, the miRNASNP data were integrated to construct our database, named miRSNPBase. In total, miRSNPBase includes 886 pre-miRNAs and 2640 SNPs (Additional file 1: Table 1), among them, 551 pre-miRNAs have mature miRNA in the 5’ arm and 566 pre-miRNAs have mature miRNA in the 3’ arm (Additional file 2: Table 2).

To compare SNP enrichment among the different regions of the pre-miRNAs, we accounted for the number of SNPs located in five regions, the 3’ flanking region, miRNA-5P, terminal loop,

miRNA-3P, and 5’ flanking region of the pre-miRNA. We further characterized the enrichment of SNPs located in the mature miRNAs of each pre-miRNA. The distribution of SNPs in the pre-miRNA sequences is shown in Figure 4:

As shown in Figure 4, the enrichment of SNPs in the miRNA-5P region was higher than that in the other regions. The enrichment of SNPs in the terminal loop was the lowest. In mature miRNAs, SNPs were the most enriched in the 13th nucleotide position. The lowest enrichment was in the 23rd position; since the length of most mature miRNAs is 22 nt, the 23rd position was not considered, and the lowest enrichment of SNPs in pre-miRNAs was found in the 1st, 5th and 17th positions. For most positions, the 13th, 14th and 15th positions had the highest SNP enrichment.

Identification of splicing sites of SNP-pre-miRNA

Each pre-miRNA and related SNPs were used to construct SNP-pre-miRNAs based on nucleotide substitution and the composition method. Taking hsa-mir-1181 as an example, there are 6 SNPs within hsa-mir-1181, based on which we constructed 63 SNP-pre-miRNAs. By the same token, we constructed 10574 SNP-pre-miRNAs using 886 pre-miRNAs and their associated 2640 SNPs. When we identified the mature miRNAs of each SNP-pre-miRNA using miRFind, the five mature miRNA candidates were considered to improve the identified accuracy.

Based on the miRFind results, the mature miRNAs were divided into nor-miRNAs and isomiRs; correspondingly, the SNP-pre-miRNAs associated with nor-miRNAs and isomiRs were named

nor-SNP-pre-miRNAs and iso-SNP-per-miRNAs, respectively. The pre-miRNAs associated with nor-miRNAs and iso-miRNAs were named nor-pre-miRNAs and iso-per-miRNAs, respectively. The result of mature miRNA identification of SNP-pre-miRNAs is illustrated in Table 2:

As shown in Table 2, the enrichment of iso-SNP-pre-miRNAs, iso-pre-miRNAs and iso-SNPs was lower than that of nor-SNP-pre-miRNAs, nor-pre-miRNAs and nor-SNPs in P5_5, and the enrichment of iso-SNP-pre-miRNAs, iso-pre-miRNAs and iso-SNPs was higher than that of nor-SNP-pre-miRNAs, nor-pre-miRNAs and nor-SNPs in other sites. Taking P5_5 as an example, we identified 607 iso-SNP-pre-miRNAs that contained 143 iso-pre-miRNAs and 427 iso-SNPs. Conversely, we identified 4235 nor-SNP-pre-miRNAs that contained 480 nor-pre-miRNAs and 1250 nor-SNPs. All the iso-pre-miRNAs, nor-pre-miRNAs, nor-SNPs and iso-SNPs associated with the four splicing sites are shown in Additional file 3: Table 3.

The distribution of pre-miRNAs and SNPs associated with the normal miRNAs and isomiRs are shown in Figure 5 and Additional file 4: Table 4:

The common nor-pre-miRNAs, nor-SNPs, iso-pre-miRNAs and iso-SNPs of the four splicing sites were extracted, and the results suggested that these nor-pre-miRNAs tended to remain in the normal pre-miRNA splicing process. The nor-SNPs tended to not affect the Drosha and Dicer enzyme splicing of pre-miRNAs. The iso-pre-miRNAs tended to be spliced into isomiRs, and the

nor-SNPs tended to change the splicing sites and produce isomiRs. Remarkably, all these pre-miRNAs and SNPs can be used as candidates for future studies.

Analysis of the MAF of SNPs in the iso-pre-miRNAs and nor-pre-miRNAs

The MAF of iso-SNPs and nor-SNPs is illustrated in Figure 6:

The highest MAF of nor-SNPs is ‘1’, which means that a specific set of SNPs will not affect mature miRNA biogenesis. The highest MAF of the iso-SNPs did not reach ‘1’, which means that there is a specific set of SNPs in the samples that does not affect mature miRNA biogenesis, and a specific set of SNPs that always affects miRNA maturation does not exist.

As shown in Figure 6, we sorted the MAF based on value and compared the value of the bottom 75% and 50%. For all sites, the MAF of iso-SNPs of iso-SNPs was lower than that of nor-SNPs, which means that the population with SNPs that can affect miRNA mature biogenesis is always less than the population with SNPs that do not affect miRNA mature biogenesis.

Based on the MAF of iso-SNPs and nor-SNPs, most molecules with SNPs underwent the traditional mature miRNA splicing process, which is consistent with the existence of heterogeneous miRNAs and mature miRNAs. The results suggested that SNPs altered the

biogenesis of mature miRNAs, which would result in the production of many isomiRs thus changing miRNA expression and affecting normal cellular processes, which can lead to more diseases, but this scenario is obviously not what is actually observed. Therefore, our study illustrates that the MAF of iso-SNPs is lower than that of nor-SNPs, proving that only a few SNPs affect mature miRNA biogenesis.

Effects of SNPs on the biogenesis of mature miRNAs

The iso-pre-miRNAs, nor-pre-miRNA, iso-SNPs and nor-SNPs were analyzed to determine which factors might play an important role in the biogenesis of mature miRNAs. We analyzed the relationships between SNP position, SNP type, SNP-affected free energy change and splice site variation of pre-miRNAs. Because isomiRs can be generated from mutations of all four splicing sites, we analyzed all miRNA sites. The enrichment of pre-miRNAs based on SNP positions is described in Figure 7:

Taking P5_5 as an example, the nor-pre-miRNAs had the highest enrichment and the lowest enrichment when the SNPs were located in section 1 and section 3, respectively. Accordingly, the iso-pre-miRNAs had the highest and lowest enrichment when the SNPs were located in section 4 and section 3, respectively. When the SNPs were located in section 1, the iso-pre-miRNAs had a higher enrichment in that region and the nor-pre-miRNAs had a lower enrichment, and their enrichments were insignificantly different. Additionally, when the SNPs were located

in section 3, the iso-pre-miRNAs had a lower enrichment, the nor-pre-miRNAs had a higher enrichment, and their enrichments were significantly different.

For P5_5-associated pre-miRNAs, when SNPs were located in section 1, the SNP-pre-miRNAs tended to splice and produce isomiRs, and when SNPs were located in sections 2 and 3, the SNP-pre-miRNAs tended to splice and produce normal miRNAs. In a similar way, for P5_3 associated pre-miRNAs, when SNPs were located in sections 2 and 5, the SNP-pre-miRNAs tended to splice and produce normal miRNAs, and when SNPs were located in the other sections, the SNP-pre-miRNAs tended to splice and produce isomiRs. For P3_5- and P3_3-associated pre-miRNAs, when the SNPs were located in sections 1 and 2, the SNP-pre-miRNAs tended to splice and produce normal miRNAs, and when SNPs were located in the other sections, the SNP-pre-miRNAs tended to splice and produce isomiRs.

We analyzed the enrichment of pre-miRNAs in SNPs located in the mature miRNAs, as shown in Figure 8.

We focused on three cases: (1) The enrichment of SNPs in iso-pre-miRNAs/nor-pre-miRNAs is higher than that in nor-pre-miRNAs/iso-pre-miRNAs; (2) the enrichment of SNPs in one class of pre-miRNAs has a large degree of change in specific sites of mature miRNAs; and (3) the

enrichments of SNPs in iso-pre-miRNAs and nor-pre-miRNAs have a slight change in specific sites in mature miRNAs.

For P5_5 splicing-related mature miRNAs, when the SNPs were located in the 7 nt, 9 nt and 15 nt positions, the enrichment in nor-pre-miRNAs was higher and that in iso-pre-miRNAs was lower, and their enrichments were insignificantly different. Additionally, when the SNPs were located in the 6 nt, 8 nt and 13 nt positions, the iso-pre-miRNAs had a lower enrichment, the nor-pre-miRNAs had a higher enrichment, and their enrichments were significantly different. When the SNPs were located in 17-22 nt positions, the enrichment of nor-pre-miRNAs and iso-pre-miRNAs was positive correlated. These findings suggest that pre-miRNAs containing SNPs in the 7 nt, 9 nt and 15 nt positions tend to splice and produce normal miRNAs, pre-miRNAs with SNPs located in the 6 nt, 8 nt and 13 nt positions tend to splice and produce isomiRs, and when the SNPs are located in the 17-22 nt positions, they do not affect the pre-miRNA splicing process.

Similarly, for P5_3 splicing-related mature miRNAs, the results suggest that pre-miRNAs containing SNPs in the 1 nt, 2 nt 4 nt, 6 nt, 11 nt 12 nt and 17 nt positions tend to splice and produce normal miRNAs and tend to splice and produce isomiRs when SNPs are located in the 3 nt, 8 nt 13 nt and 14 nt positions, while the SNPs located in the 18-22 nt positions do not affect the pre-miRNAs splicing process. For P3_5 splicing-related mature miRNAs, the data suggest that pre-miRNAs containing SNPs in the 4 nt, 10 nt, 13 nt, 15 nt and 21 nt positions tend to splice and produce normal miRNAs and tend to splice and produce isomiRs when SNPs are

located in the 2 nt, 9 nt and 22 nt positions, while the SNPs located in the 5 nt, 6 nt and 17 nt positions do not affect the pre-miRNAs splicing process. For P3_5 splicing-related mature miRNAs, the results suggest that pre-miRNAs containing SNPs in the 4 nt, 10 nt, 13 nt, 15 nt and 21 nt positions tend to splice and produce normal miRNAs and tend to splice and produce isomiRs when SNPs are located in the 2 nt, 9 nt and 22 nt positions, while the SNPs located in the 5 nt, 6 nt and 17 nt positions do not affect the pre-miRNAs splicing process. For P3_3-splicing related mature miRNAs, the data suggest that pre-miRNAs containing SNPs in the 2 nt, 4 nt, 15 nt and 22 nt positions tend to splice and produce normal miRNAs and tend to splice and produce isomiRs when SNPs are located in the 3 nt, 4 nt, 16 nt and 21 nt positions, while the SNPs located in the 5 nt, 10 nt, 11 nt, 18 nt, 19 nt and 20 nt positions do not affect the pre-miRNAs splicing process.

In general, the enrichments of iso-pre-miRNAs and nor-pre-miRNAs are negatively correlated in most sites of mature miRNAs, which is consistent with the actual observations; that is, as the SNP-pre-miRNA counts to normal miRNA counts increase, the SNP-pre-miRNA counts to isomiRs counts decrease.

The enrichment of pre-miRNAs and SNPs based on SNP type is described in Figure 9:

For P5_5 splicing-related mature miRNAs, when the SNP types are ‘C’ and ‘G’, the percent of iso-pre-miRNAs is higher; otherwise, when the SNP types are ‘A’ and ‘U’, the percent of nor-pre-miRNAs is higher. Additionally, as the SNP type is ‘A’, the difference in the percent between iso-pre-miRNAs and nor-pre-miRNAs is largest. These observations suggest that when SNP types are ‘C’ and ‘G’, the pre-miRNAs tend to splice and produce isomiRs, while when SNP types are ‘A’ and ‘U’, pre-miRNAs tend to splice and produce normal miRNAs.

Similarly, for P5_3 splicing-related mature miRNAs, the results suggest that when SNP types are ‘C’ and ‘G’, the pre-miRNAs tend to splice and produce isomiRs, while they tend to splice and produce normal miRNAs when SNP types are ‘A’ and ‘U’. For P3_5 splicing-related mature miRNAs, the results suggest that when SNP types are ‘C’, ‘G’ and ‘U’, the pre-miRNAs tend to splice and produce isomiRs, while they tend to splice and produce normal miRNAs when the SNP type is ‘A’. For P3_3 splicing-related mature miRNAs, the data suggest that when SNP types are ‘G’ and ‘U’, the pre-miRNAs tend to splice and produce isomiRs, while they tend to splice and produce normal miRNAs when the SNP type is ‘A’.

The enrichment of SNP-pre-miRNAs based on free energy change is described in Figure 10:

The change in free energy mostly focuses on 0-8 kcal/mol. As free energy increases greater than 4 kcal/mol, SNP-pre-miRNAs tend to shear and produce isomiRs. When the free energy change

decreased by more than 2 kcal/mol, for P5_5 and P3_5 mature miRNA loci, the percentage of iso-SNP-pre-miRNAs is higher than that of nor-SNP-pre-miRNAs, and for P5_3 and P3_3 mature miRNA loci, the percentage of nor-SNP-pre-miRNAs is higher than that of iso-SNP-pre-miRNAs. These findings suggest that the decrease in the free energy tends to alter the splicing sites of the 5' end of mature miRNAs and is less affected by the 3' ends of mature miRNAs.

On the basis of the free energy change, the enrichment distribution of iso-SNP-pre-miRNAs and nor-SNP-pre-miRNAs associated with the 5' end (P5_5 and P5_3) and 3' end (P5_3 and P3_3) sites of mature miRNAs has similar characteristics, indicating that the effects of the free energy change on 5' end and 3' end mature miRNA biogenesis are largely consistent.

For P3_3 mature miRNAs, when the free energy increases, SNP-pre-miRNAs tend to maintain normal mature miRNA biogenesis, and when the free energy decreases, SNP-pre-miRNAs tend to splice and produce normal mature miRNAs.

Identification and verification of isomiRs based on a GBR population from 1000 Genomes

Based on our method, we identified the isomiRs and iso-SNPs of 18 GBR individuals of European origin. Taking HG00097 as an example, because the sample VCF data were provided by karyotype, we extracted the variation information of HG00097 from 23 chromosome files and integrated all the variation information. All SNPs were mapped to pre-miRNAs to construct the SNP-pre-miRNAs for HG00097. Furthermore, we identified the mature miRNA sequences at the four sites. As a result, we predicted 695 isomiRs of 92 pre-miRNAs with 94 SNPs in a different

guide strand with the incorporation of variations in its sequence. The pre-miRNAs, iso-SNPs, and isomiRs of HG00097 are summarized in Additional file 5: Table 5. The results suggest that all SNPs within pre-miRNAs could have a potential impact on miRNA biogenesis and function.

Furthermore, we identified the isomiRs and iso-SNPs of 18 GBR individuals of European origin, and 209 iso-pre-miRNA candidates and 71 verified iso-pre-miRNAs of the 18 GBR samples are shown in Additional file 6: Table 6. In addition, 2667 isomiRs of 209 pre-miRNAs were verified by miRNA sequencing data, and the isomiRs and iso-SNPs of the 18 GBR individuals are shown in Table 3 and are detailed in Additional file 7: Table 7.

We validated them with the miRNA sequencing data. The 158 verified isomiRs of the 18 GBR samples are shown in Table 4, and the details are shown in Additional file 8: Table 8:

Discussion

As an important molecular mechanism by which SNPs significantly contribute to miRNA generation mechanisms and functions, experiments have shown that the molecular structure, thermodynamic stability and functional strand selection are affected by SNPs located in pre-miRNAs. As a result, SNPs influenced the selection of Drosha enzyme processing, Dicer enzyme processing and functional strands in the splicing process of pre-miRNAs and altered the expression levels of miRNAs, which is closely related to various phenotypes and diseases. However, because the expression levels of iso-miRNAs are low, iso-miRNAs are hard to detect,

and the mechanism of SNP-affected mature miRNA biogenesis remains unclear. Therefore, we systematically studied the role of SNPs in the biogenesis of mature miRNAs.

We constructed a SNP-associated pre-mmiRNA database based on the latest data from miRBase and the 1000 Genomes project, and by integrating the miRNASNP database, we obtained our database, named miRSNPBase.

We performed an analysis of the relationships between SNP type, SNP location and SNP-affected free energy change and the biogenesis of mature miRNAs. The results showed that these three factors played important roles in the mature miRNA generation mechanism. On the other hand, we analyzed the penetrance of nor-SNPs and iso-SNPs based on the MAF. The results proved that most SNPs do not alter normal miRNA production, and most individuals contain these SNPs. Moreover, only a few SNPs affect mature miRNA biogenesis, and only a few individuals contain such SNPs.

We identified isomiRs and iso-SNPs in 18 GBR individuals; specifically, these isomiRs were verified by miRNA sequencing data of 18 GBR samples. As a result, we obtained epigenetic-associated isomiRs and SNPs; furthermore, by comparison with the miRNA sequencing data, we found and verified the presence of isomiRs.

Conclusions

Overall, our study suggested that SNPs affect the biological characteristics and lead to changes in the Dicer sites of mature miRNAs undergoing the maturation process, therefore leading to the generation of isomiRs. The distribution of the MAF of SNPs showed that only a few SNPs affect the splicing mechanism of mature miRNAs, which is consistent with reported facts. In addition, some isomiRs were verified based on miRNA sequencing data of 18 GBR individuals. Identification of isomiRs in miRNA sequencing data also indicates that our method is effective. In conclusion, the results suggest that the SNPs play an important role in the biogenesis of mature miRNAs.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The miRNA genome position information used in this paper can be downloaded from ftp://mirbase.org/pub/mirbase/. The information of SNPs described in this manuscript was obtained from the 1000 Genomes Project (ftp://ftp-

trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/). MiRNASNP dataset can be downloaded from http://bioinfo.life.hust.edu.cn/miRNASNP2/download.php. The VCF files used in this paper can be downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/. And the miRNA sequencing data described in this manuscript can be obtained from https://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-2/samples/. The Homo sapiens GRCh37 reference sequence used in this paper can be downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_g1k_v37.fasta.gz.

Competing interests

The authors declare that they do not have any competing commercial interests in relation to the submitted work.

Funding

Publication charges for this article have been funded by the Youth Science Fund of Heilongjiang Province of China (no. QC2017079), the science and technology project of Qiqihar (GYGG-201915, GYGG-201709), the Basic Business of Educational Commission of Heilongjiang Province of China (135109246), Humanities and Social Science Fund of Ministry of Education of China (19YJCZH120), and National Natural Science Foundation (NNSF) of China (61901063).

Authors' contributions

YW designed the work program, drafted the manuscript and wrote the code and implemented the analysis. MXL constructed the database. RJD participated in the writing of the manuscript and the interpretation of the results.

References

1. Shastry BS: **SNPs: impact on gene function and phenotype**. *Methods Mol Biol* 2009, **578**:3-22.
2. Lander ES: **Initial impact of the sequencing of the human genome**. *Nature* 2011, **470**(7333):187-197.
3. Ryan BM, Robles AI, Harris CC: **Genetic variation in microRNA networks: the implications for cancer research**. *Nat Rev Cancer* 2010, **10**(6):389-402.
4. Jin Y, Lee C: **Single Nucleotide Polymorphisms Associated with MicroRNA Regulation**. *Biomolecules* 2013, **3**(2):287-302.
5. Alarcon CR, Lee H, Goodarzi H, Halberg N, Tavazoie SF: **N6-methyladenosine marks primary microRNAs for processing**. *Nature* 2015, **519**(7544):482-485.
6. Hu HY, Yan Z, Xu Y, Hu H, Menzel C, Zhou YH, Chen W, Khaitovich P: **Sequence features associated with microRNA strand selection in humans and flies**. *BMC genomics* 2009, **10**:413.
7. Han J, Lee Y, Yeom KH, Nam JW, Heo I, Rhee JK, Sohn SY, Cho Y, Zhang BT, Kim VN: **Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex**. *Cell* 2006, **125**(5):887-901.
8. Hiard S, Charlier C, Coppieters W, Georges M, Baurain D: **Patrocles: a database of polymorphic miRNA-mediated gene regulation in vertebrates**. *Nucleic Acids Res* 2010, **38**(Database issue):D640-651.
9. Duan R, Pak C, Jin P: **Single nucleotide polymorphism associated with mature miR-125a alters the processing of pri-miRNA**. *Hum Mol Genet* 2007, **16**(9):1124-1131.
10. Gong J, Tong Y, Zhang HM, Wang K, Hu T, Shan G, Sun J, Guo AY: **Genome-wide identification of SNPs in microRNA genes and the SNP effects on microRNA target binding and biogenesis**. *Human mutation* 2012, **33**(1):254-263.
11. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RK, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR *et al*: **RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease**. *Science* 2015, **347**(6218):1254806.
12. Jin Y, Lee CG: **Single Nucleotide Polymorphisms Associated with MicroRNA Regulation**. *Biomolecules* 2013, **3**(2):287-302.
13. Hogg DR, Harries LW: **Human genetic variation and its effect on miRNA biogenesis, activity and function**. *Biochemical Society transactions* 2014, **42**(4):1184-1189.
14. Sun G, Yan J, Noltner K, Feng J, Li H, Sarkis DA, Sommer SS, Rossi JJ: **SNPs in human miRNA genes affect biogenesis and function**. *Rna* 2009, **15**(9):1640-1651.

15. Li MJ, Yan B, Sham PC, Wang J: **Exploring the function of genetic variants in the non-coding genomic regions: approaches for identifying human regulatory variants affecting gene expression**. *Briefings in bioinformatics* 2015, **16**(3):393-412.
16. Zhang X, Zeng Y: **The terminal loop region controls microRNA processing by Drosha and Dicer**. *Nucleic Acids Res* 2010, **38**(21):7689-7697.
17. Guo L, Chen F: **A challenge for miRNA: multiple isomiRs in miRNAomics**. *Gene* 2014, **544**(1):1-7.
18. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA: **A map of human genome variation from population-scale sequencing**. *Nature* 2010, **467**(7319):1061-1073.
19. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA: **An integrated map of genetic variation from 1,092 human genomes**. *Nature* 2012, **491**(7422):56-65.
20. Wang Y, Li X, Tao B: **Improving classification of mature microRNA by solving class imbalance problem**. *Scientific reports* 2016, **6**:25941.
21. Auyeung VC, Ulitsky I, McGeary SE, Bartel DP: **Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing**. *Cell* 2013, **152**(4):844-858.

Figure Legends

Figure 1: The schematics of the overall method

Figure 2: The example of constructing the SNP-pre-miRNA sequence

Figure 3: The calculation method for variation position of plus-strand and trans-strand

Figure 4: The distribution of SNPs in the pre-miRNA sequences

Figure 5: The distribution of pre-miRNAs and SNPs associated with the normal and isomiRs

Figure 6: The MAF of iso-SNPs and nor-SNPs

Figure 7：The enrichment of pre-miRNAs based on SNPs positions

Figure 8：The enrichment of pre-miRNAs as SNPs located in the mature miRNAs

Figure 9：The enrichment of pre-miRNAs and SNPs based on SNPs type

Figure 10：The enrichment of SNP-pre-miRNAs based on free energy change

Table 1. Example of the SNPs information

Table	Parameter	Table	Parameter	Table	Parameter
chrom	1	AC	1	AFR_AF	0
pos	55285	AF	0.000199681	EUR_AF	0
ID	rs532608387	AN	5008	SAS_AF	0.001
REF	T	NS	2504	AA	t
ALT	C	DP	18296	VT	SNP
QUAL	100	EAS_AF	0		
FILTER	PASS	AMR_AF	0		

Table 2. The result of mature miRNA sits identification of SNP-pre-miRNAs

sit	nor-SNP-pre-miRNAs	nor-pre-miRNAs	nor-SNPs	iso-SNP-pre-miRNA	iso-pre-miRNA	iso-SNPs
P5_5	4235	480	1250	607	143	427
P5_3	1825	218	569	3385	405	1065
P3_5	2089	264	707	3435	363	1005
P3_3	2026	267	721	3779	395	1081

Table 3. The isomiRs and iso-snp of 18 GBR populations.

Category	sample	isomiR	SNP	sample	isomiR	SNP
Number	HG00096	697	109	HG00108	719	69
Number	HG00097	601	105	HG00109	718	112
Number	HG00099	426	94	HG00110	663	128

Number	HG00100	710	119	HG00111	711	123
Number	HG00101	646	104	HG00112	463	84
Number	HG00102	657	113	HG00114	665	103
Number	HG00105	683	104	HG00115	682	117
Number	HG00106	732	128	HG00116	714	110
Number	HG00107	315	71	HG00117	639	126

638

639 **Table 4.** The verified isomiRs of 18 GBR.

Category	sample	Veirified-isomiR	sample	Veirified-isomiR
Number	HG00096	96	HG00108	37
Number	HG00097	19	HG00109	19
Number	HG00099	19	HG00110	36
Number	HG00100	67	HG00111	27
Number	HG00101	36	HG00112	27
Number	HG00102	32	HG00114	16
Number	HG00105	23	HG00115	45
Number	HG00106	23	HG00116	33
Number	HG00107	7	HG00117	55

Additional data files

Additional file 1: Table S1: The database miRSNPBase. (xls)

Additional file 1: Table S2: The pre-miRNAs which have mature miRNA in the 5’ arm and 3’arm. (xls)

Additional file 1: Table S3: All the iso-pre-miRNAs, nor-pre-miRNAs, nor-SNPs and iso-SNPs associated with four splicing sits. (xls)

Additional file 1: Table S4: The pre-miRNAs and SNPs associated with the normal and isomiRs. (xls)

Additional file 1: Table S5: The pre-miRNAs, iso-SNPs, and isomiRs of HG00097. (xls)

Additional file 1: Table S6: The isomiRs and iso-snp of 18 GBR populations. (xls)

Additional file 1: Table S7: The verified isomiRs of 18 GBR. (xls)

Additional file 1: Table S8: The iso-pre-miRNA candidates and the verified iso-pre-miRNAs of 18 GBR samples. (xls)