Computational and Statistical Analysis of Sequence and Expression Features of MicroRNA and Long Noncoding RNA in Primate Brains

Haiyang Hu

September 2015

Dissertation zur Erlangung des Grades

eines Doktors der Naturwissenschaften (Dr. rer. nat.)

am Fachbereich Mathematik und Informatik

der Freien Universität Berlin

Gutachter:

Prof. Dr. Martin Vingron

Prof. Dr. Philipp Khaitovich

1. Referent: Prof. Dr. Martin Vingron

2. Referent: Prof. Dr. Philipp Khaitovich

Tag der Promotion: 22. January 2016

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe. Ich erkläre weiterhin, dass ich die vorliegende Arbeit oder deren Inhalt nicht in einem früheren Promotionsverfahren eingereicht habe.

Berlin, 14 September 2015 Haiyang Hu

Acknowledgments

I would like to dedicate this work to my family for their unending love and support.

First and foremost, I would like to express my special appreciation and thanks to my supervisors, Prof. Philipp Khaitovich and Prof. Martin Vingron, for their advice and support during my PhD training. I deeply appreciate their concern and patience, especially during the tough times I experienced in pursuing my PhD. I am grateful to Prof. Philipp Khaitovich, who has been a tremendous mentor for me. I would like to thank him for his encouragement and valuable biological insight. His advices on both research and my career have been priceless. I would like to thank Prof. Martin Vingron for allowing me the freedom in choosing my research topics and for his valuable comments to my thesis writing. I would like to express my fullest thanks to all current and former members in Prof. Philipp Khaitovich's group for their friendship, brilliant comments and suggestions. Thank you all! I would like to thank Arthur for helping and encouraging me in my PhD pursuit. I wish to express my warmest thanks to Kirsten, Fabian and Hannes for their kindly support and helping me finish my PhD.

I owe a special thanks to my family for all their unending love, support and encouragement. For my parents, who raised me and supported me in all my pursuits. For my sister, who taught me to have an optimistic attitude toward the future. And most of all for my loving, encouraging wife Amy whose faithful support during the final stages of this PhD is so appreciated. Thank you!

Haiyang Hu

September 2015

Individual Contributions

For the sake of clarity, individual contributions to this thesis will be detailed here. I briefly summarize my contribution and acknowledge the contributions of others.

The experiments of small RNA sequencing and RNA-seq and the Affymetrix exon array have been conducted in the wet lab of Khaitovich's group by the biochemists Zheng Yan and Xi Jiang. Ning Fu conducted the proteomics experiment in Rong Zeng's lab at SIBS. I wish to acknowledge them for generating the excellent data on which I worked. I wish to thank our collaborator, Dr. Wei Chen, by providing sequencing platforms.

I performed all the analysis presented in the thesis in the group of Prof. Philipp Khaitovich at CAS-MPG Partner Institute for Computational Biology (PICB), Shanghai, with supporting analysis of differentially expressed miRNA regulatory effect on target genes by Song Guo, Ning Fu and Guohua Xu (Section 3.6). In addition, I wish to acknowledge all the members of Khaitovich's group for their valuable comments and suggestions.

Summary

Through the study of non-coding RNA (ncRNAs) with known function we have received increasingly insights into the fundamental principles of function and regulation of the transcriptome in recent years. The broadening of the transcriptome coverage by sequencing technologies and the growing multitude of transcriptional data, as well as other types of high-throughput biological measurements, require new computational tools and approaches as well as new analysis pipelines to extract biological meaning from the quickly growing volumes of biological data. In this thesis, I have used current knowledge of ncRNAs features to construct a set of computational and statistical methods and pipelines that can analyze sequence, expression and regulatory properties of two types of ncRNAs: microRNA (miRNA) and long noncoding RNA (lncRNA). To do so, I took advantage of the high-throughput sequencing data collected in primate brains at different ages, allowing me to monitor changes in ncRNA sequence and expression over the evolutionary and the ontogenetic dimensions. In Chapter 3, I described a computational framework I constructed for across-species miRNA comparison on the basis of small RNA sequencing data. The framework includes an efficient small RNA sequencing data preprocessing pipeline, a revised miRNA quantification procedure, a reliable miRNA ortholog prediction method and a pipeline for differentially expressed (DE) miRNA identification. In Chapter 4, I described a systematic study of miRNA 5'-isoforms, including their identification and functionality prediction in the human prefrontal cortex, to deepen our understanding of the complexity of the miRNA repertoire. I developed a comparative approach to predict the functionality of the identified miRNA 5'-isoforms, which resulted in 28 putative functional miRNA 5'-isoforms displaying regulatory features similar to known conserved miRNAs. In Chapter 5, I described a genome-wide lncRNA identification and feature investigation study using strand-specific RNA-seq data covering postnatal ontogenetic stages of human prefrontal cortex. This work integrates de novo transcriptome assembly procedure and downstream lncRNA analysis elements, including a pipeline for lncRNA identification and a detailed lncRNA sequence and expression feature analysis framework. The integrative analysis of lncRNAs expression and genome-wide epigenetic data lead to the identification of a novel class of lncRNA-associated bidirectional promoters that display unique sequence and epigenetic features and preferentially drive the expression of neuronal gene. To conclude, during my thesis work, I developed computational tools that allow researchers to process and integrate different types of large-scale biological data, such as high-throughput transcriptome sequencing, epigenetic data of chromatin modifications and protein abundance data, to identify and characterize two major types of non-coding RNAs: miRNAs and lncRNAs. These results indicate adequacy and appropriateness of the analytical approaches I developed and the statistical tools I used. I hope my work will serve as a useful stepping-stone for both computational and biological studies of the noncoding RNA universe.

Zusammenfassung

Durch die Untersuchung von nicht-kodierender RNA (ncRNAs) mit bekannter Funktion haben wir in den letzten Jahren zunehmend Einblicke in die fundamentalen Prinzipien von Funktion und Regulation des Transkriptoms erhalten. Neue Sequenzierungstechnologien und weitere biologischen Hochdurchsatzanalysen produzieren immense Mengen von Transkriptionsdaten in immer höherer Sequenzierungstiefe. Das exponentielle Wachstum dieser Datenmengen verlangt nach neuen computergestützten Ansätzen und Methoden sowie neuen Analyse-Pipelines. In dieser Dissertation, stelle ich eine Reihe von computergestützten statistischen Methoden vor, die, zur Analyse von Sequenzinformationen, derr Expression und der Eigenschaften regulatorischer Netzwerke von zwei Arten von ncRNA, nämlich microRNA (miRNA) und langer, nicht kodierender RNA (lncRNA), anwendbar sind. Dafür analysiere ich zunächst Hochdurchsatz-Sequenzierungsdaten, die vom Gehirn verschiedener Primatenarten unterschiedlichen Alters her stammen, um Veränderungen in der genomischen Sequenz und Expression während Evolution und Ontogenese zu untersuchen. In Kapitel 3 erörtere ich die Grundlagen der vergleichenden Analyse von miRNA zwischen verschieden Spezies anhand von Sequenzierungsdaten kleiner RNAs. Zu diesen Grundlagen gehört ein Ansatz zur effizienten Vorverarbeitung der Sequenzierungsdaten, ein überarbeitetes Verfahren zur Quantifizierung von miRNAs, eine zuverlässige Methode zur Vorhersage von orthologer miRNA, sowie eine Pipeline zur Identifikation differentiell exprimierter (DE) miRNA. Im vierten Kapitel beschreibe ich eine systematische Studie über verschiedene 5'-miRNA Isoformen, inklusive deren Identifikation und Vorhersage der Funktion im humanen präfrontalen Cortex. Die vorliegende Studie liefert damit einen Ausblick darauf, wie die „versteckte Ebene" des miRNA Repertoires untersuchen werden kann, und stellt zudem eine nützliche Ressource dar, um die Liste der bekannten, funktionellen miRNA durch neue 5'-Isoformen zu erweitern. Kapitel 5 beschäftigt sich mit der genomweiten Identifikation von lncRNA und der Untersuchung derer Eigenschaften basierend auf strangspezifischen RNA-seq Daten vom humanen präfrontalen Cortex, die die komplette postnatale Lebensspanne des Menschen abdecken. Die Ergebnisse dieser Analyse zeigen eine Reihe von bisher unbekannten lncRNAs, die im menschlichen Gehirn exprimiert werden, und demonstrieren eindrücklich, dass weite Teile des menschlichen Transkriptoms noch nicht charakterisiert sind. Darüber hinaus führte die integrative Analyse von lncRNA Expression und genomweiter epigenetischer Informationen zur Identifikation einer neuen Klasse von bidirektionalen Promotoren, die einzigartige Eigenschaften bezüglich Sequenz und epigenetischen Faktoren haben. Zusammenfassend habe ich während meiner Promotion computergestützte Tools und Pipelines entwickelt um verschiedene Arten von biologischen Hochdurchsatzdaten, wie zum Beispiel Hochdurchsatz-Transkriptom-Sequenzierung, epigenetische Chromatin-Modifikationen und quantitative Proteinanalysen zu verarbeiten und zu integrieren. Die Ergebnisse zeigen, dass die hier gewählten analytischen Ansätze und entwickelten Tools angemessen und geeignet für eine solche Analyse sind. Meine Arbeit stellt somit ein nützliches Hilfsmittel für zukünftige computergestützte und biologische Studien im Bereich der nicht-kodierenden RNA dar.

Contents

1. Introduction

The central dogma of molecular biology states that the flow of genetic information moves from DNA to RNA to protein [1]. According to this view, RNA is a bridge in the transfer of genetic information between DNA and proteins. A few exceptions to this paradigm are ribosomal RNA (rRNA), transfer RNA (tRNA), small nucleolar RNA (snoRNA) and small nuclear RNA (snRNA). However, these housekeeping or infrastructural RNAs only work together to turn the genetic information from DNA into protein, thus adhering to the central dogma. Following this paradigm, the DNA sequences except those necessary for synthesizing protein are usually considered "junk" DNA, referring as evolutionary relics. However, in the last two decades, this dogma has been challenged since the completion of the human genome and the genomes of several model species. A big surprise since the completion of the human genome is that there are only ~20,000 protein-coding genes, which neither correlates the organism's complexity nor accounts for the selection pressure during the evolution of the modern organism [2]. The mRNA transcripts of protein-coding gene represent less than ~3% of the human genome. This proportion increases to ~40% if cognate introns are counted (Ensembl gene annotation version 64), which still leaves ~60% of the human genome as junk DNAs.

Recent development of high-throughput methodologies and computational algorithms designed to analyze resulting data indicate that this junk DNA may not be junk after all. Along with the development of high-throughput sequencing techniques and great efforts by large-scale consortia focusing on characterizing functional genomic elements, such as The Encyclopedia of DNA Elements (ENCODE) and The Functional Annotation of the Mammalian Genome (FANTOM) [3-8], our understanding of the architecture, activity and regulation of the eukaryotic genomes has been substantially revolutionized, staring from the observation about transcriptional pervasiveness across 80% of the human genome [7]. Although this initial conclusion based on microarray from ENCODE was criticized due to concerns about high background and cross-hybridization problems, the high solution deep sequencing data from ENCODE clearly demonstrated that more than 70% of the human genome are indeed transcribed into transcripts of various sizes [6, 7]. These findings indicate a general paradigm for functional DNA elements embedded in the non-coding part of mammalian genomes. The discovery of various noncoding RNA species from pervasive transcription further expands our understanding of the extraordinary complexity of the human genome [6, 9-13].

The broadening of the transcriptome coverage by sequencing technologies and the growing multitude of transcriptional data and other types of high-throughput biological measurements require new computational tools and approaches to extract biological meaning from the quickly growing volumes of biological data. This is particularly true for transcripts that encode no proteins (noncoding RNAs) but still may play important roles in the transcriptional, post-transcriptional and translational regulatory networks. In the next section, I will

summarize known noncoding RNA types existing in human cells and focus on two types of noncoding transcripts, long noncoding RNAs and microRNAs, which are particularly relevant to my thesis. This biological background is needed for better design of computational tools aimed at determining the functional role of these transcripts in the regulatory networks.

1.1 The landscape of noncoding RNAs (ncRNAs)

Based on the current estimation from GENCODE (version 22), our human genome encodes more than 60,000 genes [13]. Among them, ~19800 protein-coding genes represent only ~33% of the total gene catalog. The rest encompass a large group of ncRNAs, including >9,000 small RNAs (smRNAs), >15,000 long non-coding RNAs (lncRNAs) and >14,000 pseudogenes. However, the number of annotated ncRNAs is conservative and does not nearly cover the full spectrum of noncoding transcripts present in human cells. For instance, piRNAs are not included in GENCODE annotation. Furthermore, certain classes of ncRNAs, such as lncRNAs, are expressed in a highly spatial- and temporal-specific matter and have not yet been investigated in all tissues at all ontogenetic stages [14, 15]. Thus, the ncRNA catalog is expected to continuously expand based on more transcriptome surveying studies.

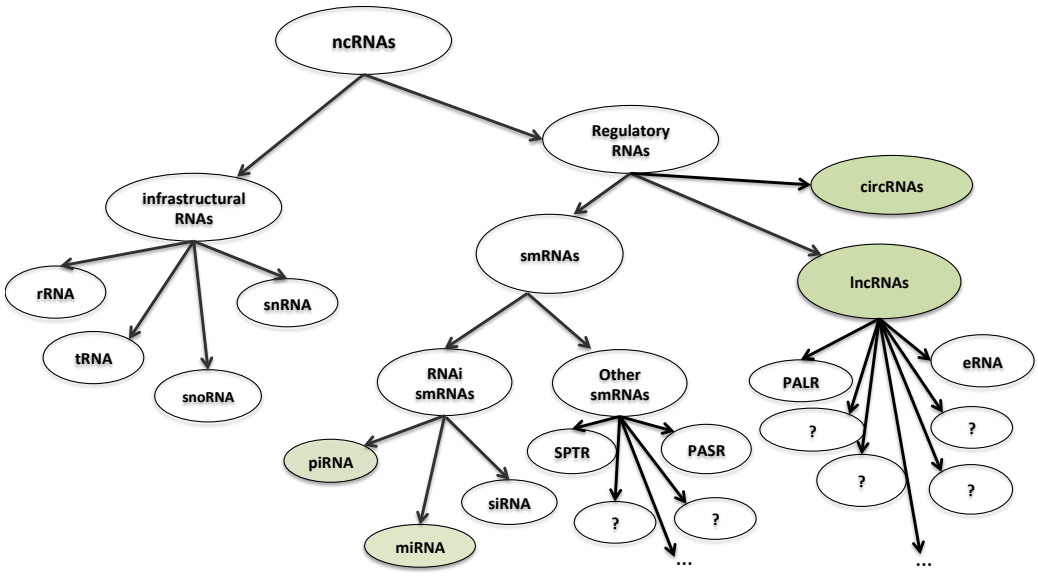


Figure 1.1: The ncRNA landscape and general classification. The ncRNA classification is largely based on the general regulatory potential and transcript length of ncRNAs. The transcript length cutoff of 200nt is arbitrarily chosen based on current RNA purification protocol limitation that takes little consideration of functional meaning. Nevertheless, the size cutoff clearly distinguishes lncRNAs from most smRNAs. The newly discovered circRNAs are listed as a separated class of regulatory RNAs due to their unique circular conformation feature and transcript length range. With regarding to the transcript length, circRNAs fall on both sides of the 200nt cutoff. I studied the ncRNAs in green during my PhD training.

The full ncRNA classification required the knowledge of sequence, structure and functional

features of ncRNAs. However, our current understanding of ncRNAs is just at the very beginning. While remarkable functions have been described for some noncoding transcripts, the importance of most ncRNAs to gene regulation is still unknown. Nevertheless, ncRNAs can be largely classified based on their general regulatory potential and transcript length (Figure 1.1). Accordingly, the ncRNAs can be first divided into infrastructural RNAs and regulatory RNAs. The infrastructural RNAs, which are considered housekeeping, include rRNA, tRNA, snoRNA and snRNA, all of which are mainly involved in or related to the mRNA and protein biogenesis processes. The regulatory RNAs, comprised of a myriad of RNAs of different lengths and with various functions, can be further generally separated into two classes based on transcript length (small and long RNAs): smRNAs (<200nt) and lncRNAs (>200nt), [16].

Since the smRNAs catalog is continuously expanding, smRNAs can be further approximately divided into two groups: smRNAs that involve in RNA interference-related (RNAi) machinery (RNAi-smRNAs) and those that do not (other-smRNAs). As yet, we know little about many newly discovered types of small RNAs in the group of other-smRNAs [17], e.g., promoter associated small RNAs (PASRs) [18], termini associated short RNAs (TASRs) [16], split-site RNAs (SPSRs) [19] and small nuclear-factor 90 associated RNAs [20]. Still, our understanding of the biogenesis and biological functions of RNAi-smRNAs is growing rapidly, thanks to their relatively uniformly functioning specificity. The RNAi-smRNAs include microRNAs (miRNAs) [9, 21], endogenous small interfering RNAs (endo-siRNAs) [22] and PIWI-interacting RNAs (piRNAs), all of which are bound in Argonaute proteins [11]. Briefly, miRNAs are most well studied RNAi-smRNAs, which may be involved in regulation of almost all biological processes [21]. The miRNAs are ~22nt in length, which are associated with Argonaute proteins (AGO1-4) and regulate gene expression post-transcriptionally by partially binding to 3' untranslated regions (3' UTR) of target messenger RNAs (mRNAs) and lead to mRNA degradation or translational inhibition [23]. The endo-siRNAs are 21-26nt in length, which are associated with AGO2 and are involved in post-transcriptional and epigenetic silencing of protein-coding genes and transposons through fully complementary base-pairing [22]. The piRNAs are a class of 24-30nt RNA mainly expressed in the animal germline. piRNAs associated with PIWI subclass of Argonaute proteins and mainly involved in maintaining genome integrity against transposable elements [11].

Being the counterparts of smRNAs, lncRNAs with lengths varying from 200 nt to over 100k nt certainly stand for a heterogeneous group. As would be expected from their general name, no unified functioning mechanisms have yet been identified for them, and only a few have been subjected to extensive experimentation [24-27]. Nevertheless, the verified examples of lncRNA have been found to exert their influence on a diverse range of biological processes from developmental control to disease progression, establishing operable patterns for the functionality of lncRNA [28, 29]. Besides the aforementioned linear ncRNAs, a special class of non-linear RNAs, named circular RNAs (circRNAs), further broadens the ncRNAs repertoire [30]. The circRNAs, possessing the most distinctive feature of being “circular” by forming a covalently closed continuous loop through joining 5’ and 3’ ends, mainly arise from coding exons but exhibit no coding potentials [31-33]. Two previous studies showed that circRNAs could bind and block cognate miRNA as molecular “sponges” to these interfering regulators [31, 32]. However, both validated and predicted candidates with

potential miRNA sponge function are very limited for circRNAs [34], and the existence of other functional possibilities still remains to be determined.

During my PhD training, I mainly studied two kinds of smRNA (miRNA and piRNA) as well as lncRNA and circRNA (shown in green of Figure 1.1), on the basis of high-throughput sequencing data. I got involved in several different analyses, including investigating sequence features associated with miRNA maturation process [35], novel miRNA prediction [36], miRNA promoter prediction, miRNA 5'-isoform identification and functionality predication, identifying and evolutionary studying human specific miRNAs [37], across-species analyzing miRNA profiles and regulations among primates [38], meta-analysis of piRNA-like sequences in a variety of somatic tissues of several species [39], analyzing piRNA profiles during macaque sex maturation, and identification and feature analysis of lncRNA [40] and circRNAs. Although not all my investigations led to publication, the training has undoubtedly broadened and deepened my understanding about the distinct sequences, expressions and evolutionary features of different RNA species. Unfortunately, not all my studies can be summed up in this thesis. The theme of this thesis is computational and statistical analysis of sequence and expression features of two kinds of ncRNAs, miRNA and lncRNA. In the next section, I present the current understanding of miRNA and lncRNA.

1.2 MicroRNA (miRNA)

miRNAs are probably the most well-studied RNAi-smRNAs. In general, miRNAs are ~22 nt in length and regulate gene expression post-transcriptionally by binding to 3' untranslated regions (3' UTR) of target mRNAs, which leads to the mRNA degradation or translational inhibition [21, 41-43]. It is estimated that miRNAs regulate a substantial portion of protein-coding genes in animals and around 60% in human [44]. Controlled genes are involved in a wide range of physiological processes, including development, growth, differentiation and metabolism [21, 41, 45]. Although the extent of the miRNA regulatory universe has only been appreciated within the last two decades, their existence and exemplary power has completely changed our understanding about the fundamental principle of molecular biology. In this section, I summarize current knowledge about miRNA genes, including their discovery, biogenesis, nomenclature and feature characteristics, as well as the approaches for characterizing miRNAs. In addition, potential problems and challenges of miRNA analysis are described.

The discovery of miRNA

The founding member of miRNA, lin-4, was discovered in 1993 through studying developmental timing in worms [9, 46]. Lin-4 was found to encode a small non-protein-coding RNA transcript that contains complementary base pairing to the sequence in the 3'UTR of the lin-14 gene, acts as a negative regulator of lin-14 and represses the accumulation of LIN-14 protein [9, 46]. Seven years later, let-7 was the second miRNA identified, also in worms. It possesses important functions during larval development through complementary to two closely spaced sites in lin-41 3'UTR [47]. Unlike lin-4, the let-7 sequence is deeply conserved across species between worms, flies and humans, a groundbreaking fact that provides initial insight into research for miRNA using genomics. Later, intensified cloning efforts, although laborious, identified numerous additional miRNAs in mammals, fish, worms and flies [48-52]. Nowadays, with ever-more genomes from different species being sequenced and revolutionary advances in high-throughput sequencing technology, sized-fractionated small RNA library construction followed by high-throughput sequencing (small RNA sequencing) coupled with downstream computational prediction and experimental validation has proven to be a reasonable and efficient way for miRNA discovery [53, 54].

Characterization of miRNA expression profile

Common to other functional gene categories, obtaining comprehensive and precise expression level measurements is fundamental to miRNA study. Consequently, several techniques have been developed for characterizing miRNA profiles, mainly including Q-PCR, miRNA microarray and small RNA sequencing [55]. Q-PCR and miRNA microarray are two traditional approaches that measure miRNA abundance using pre-defined primers or probes [56, 57]. Although both approaches have been widely used for miRNA profiling, the advent of small RNA sequencing has revolutionized the manner for characterizing miRNA profiles.

Small RNA sequencing, functioning as a “molecular microscope”, is superior to the other two approaches for both specificity and sensitivity in miRNA quantification [55, 58]. Mainly due

to the short length of miRNAs, both Q-PCR and miRNA microarray have problems distinguishing highly similar sequences like miRNA family members [58]. In addition, the discrimination of mature and unprocessed forms of miRNAs presents another difficulty for Q-PCR and miRNA microarray [58]. In contrast, the high resolution of small RNA sequencing allows easy distinguishing miRNAs that only differ by one nucleotide [55, 58]. Furthermore, small RNA sequencing is more sensitive and displays higher dynamic range gained by the high sequencing depth [59, 60]. Even fragments that occur only a few times in the library will be visible in the data, and the read counts do not show the saturation effects common to microarray derived expression values. Q-PCR and miRNA microarray rely heavily on the availability and accuracy of miRNA sequences for designing primers and probes; thus, both are restricted to measuring known miRNA expression and are strongly affected by erroneously annotated miRNA mature sequences [55, 58, 61]. In contrast, small RNA sequencing is independent of predesigned probes, which makes it suitable not only for the discovery of new miRNAs but also provides the potential to make miRNA quantification more precise by scrutinizing small RNA reads. These unique features open the avenue for unbiased comparative miRNA study across species, provided the genomes are available. The high-throughput and high resolution features of small RNA sequencing also allows uncovering small RNA complexity, which leads to an unexpected finding that miRNAs display heterogeneous ends, suggesting that miRNAs can also have different isoforms similar to protein coding genes despite their short length [54, 61]. This is particularly surprising since it is generally believed that the miRNA processing machinery ensures the generation of a mature miRNA with a fixed sequence. Those miRNA isoforms have been detected using northern blotting in both animals and plants [62, 63]. Notably, their relative abundance also varies among tissues and developmental stages [64]. Whether miRNA isoforms are functional is currently unknown.

There are also limitations for small RNA sequencing. It has been shown that the fragment composition of the sample is significantly altered depending on the methods used for RNA extraction and library preparation [65]. The sequence-specific biases related to enzymatic steps in small RNA cDNA library preparation methods usually favor capturing some miRNAs over others, which makes absolute miRNA quantification difficult. The absolute read counts are therefore not precise representatives of expression levels. As in microarray analysis, the differential expression analysis based on small RNA sequencing is limited to relative comparisons of normalized read counts between samples (fold-changes) to detect miRNA expression differences.

Small RNA sequencing data provide the basis of systematic miRNA profiling. However, the achievement of comprehensive and precise miRNA measurements mainly depends on downstream computational analysis. Huge amounts of data generated through small RNA sequencing provides several computational challenges, including efficient small RNA sequence reads processing procedures, reliable small RNA mapping strategies and reasonable quantification strategies based on mapped reads. The quantification strategy is most important because certain miRNAs with erroneously annotated mature sequences may exist in miRBase. It should be noted that miRBase is a community resource with a somewhat inclusive policy [66]. Although most of the apparently misannotated miRNAs have been excluded from miRBase over the years, some may still remain. As it has been shown based on a large-scale

miRNA cloning study, ~40% of miRNA sequences deposited in miRBase (version 8.2) do not represent the predominantly cloned sequence [67]. Even though the annotation quality is expected to keep improving, the fact that the predominant miRNA sequence may vary across tissues or development stages, in addition to the recent observation of existence a large number of miRNA isoforms, further obscure full reliance upon miRBase annotation for miRNA quantification and downstream miRNA target prediction and function analysis [64, 68]. Consequently, quantification by counting the reads exactly matching annotated miRNAs may not be appropriate. Thus, a better miRNA quantification procedure is required to resolve all the potential problems.

The biogenesis of miRNA

Great efforts in the past two decades enable us to draw a relatively complete miRNA biogenesis pathway. Figure 1.2 depicts the canonical miRNA biogenesis pathway [69]. The majority of miRNAs are transcribed by DNA-dependent RNA polymerase II (RNAPII) to generate a primary miRNA (pri-miRNA) containing a region of imperfect dsRNA, known as the stem loop structure, which harbors the future mature miRNA [70, 71]. The pri-miRNA can be the transcripts of protein-coding genes or independent long noncoding genes in the intergenic region, which have 5' cap structures and polyA tails and may contain introns [72]. The production of canonical miRNAs from these pri-miRNA transcripts proceeds through two site-specific cleavage events by two RNase III enzymes, in the nucleus and cytoplasm, consecutively. In the nucleus, the processing starts with a dsRBD protein, Pasha/DiGeorge syndrome critical region gene 8 (DGCR8), which binds to the pri-miRNA and recruits the RNase III enzyme Drosha to form a multiprotein complex called the Microprocessor [73, 74]. The Microprocessor recognizes the portion with unique hairpin characteristics from the pri-miRNA and cleaves pri-miRNA by Drosha to produce a ~70-nt precursor miRNA (pre-miRNA). The pre-miRNAs displaying 2-nt single-stranded 3′ overhang (the characteristic of RNase III-mediated cleavage) are recognized by the nuclear export protein Exportin 5 and actively transported to the cytoplasm in a Ran-GTP-dependent manner [73, 74]. In the cytoplasm, the pre-miRNA is further cleaved into a ~22-nt miRNA:miRNA* duplex by Dicer with the help of a mammalian Dicer partner, dsRBD protein TAR RNA-binding protein (TRBP) [75]. Similarly, in Drosophila melanogaster, Dicer-1 interacts with a specific isoform of its dsRBD protein partner Loquacious (Loqs) to perform the same function [76]. Small RNA duplexes generated by Dicer and its protein partner also exhibit 2-nt single-stranded 3′ overhangs at both ends, a signature of RNase III cleavage [75]. For small RNAs that are initially produced as ~22-nt duplexes, in most cases, one strand would be chosen to load into Argonaute proteins to form RNA-induced silencing complex (RISC), and the other usually would be discarded—a process called strand selection [21, 77-79]. Strand selection is the final step for miRNA maturation, which is important and must not be random. It is not hard to imagine that loading the wrong strand to functional mature miRNA would cause silencing of the wrong set of genes, which is detrimental to the organism in most cases. Therefore, for most miRNAs, evolutionary pressure has selected one particular strand of the small RNA duplex as a crucial regulator, while in some cases, both strands can be found in RISC [21]. The major determinant of strand selection process resides in the intrinsic structure of the small RNA duplex-thermodynamic property [80, 81]. For miRNAs in both mammals

and flies, the strand with the least stable 5′ end is more often retained. Besides, additional favorable sequence characteristics further enhance the strand selection process, such as 5'-U bias, which I discovered in my previous study [35] and which was further supported and confirmed in vitro and in vivo by several follow-up studies [82-85]. Besides the canonical miRNA biogenesis pathway, several unconventional pathways have also been discovered, including three Drosha-independent and Dicer-dependent pathways for producing non-canonical miRNAs from very short intron [86, 87], 5' capped pre-miRNA [88] and promoter-proximal RNAPII transcription [89], and one Dicer-independent & Ago2-catalytic-activity-dependent pathway for miR-451 maturation [90].

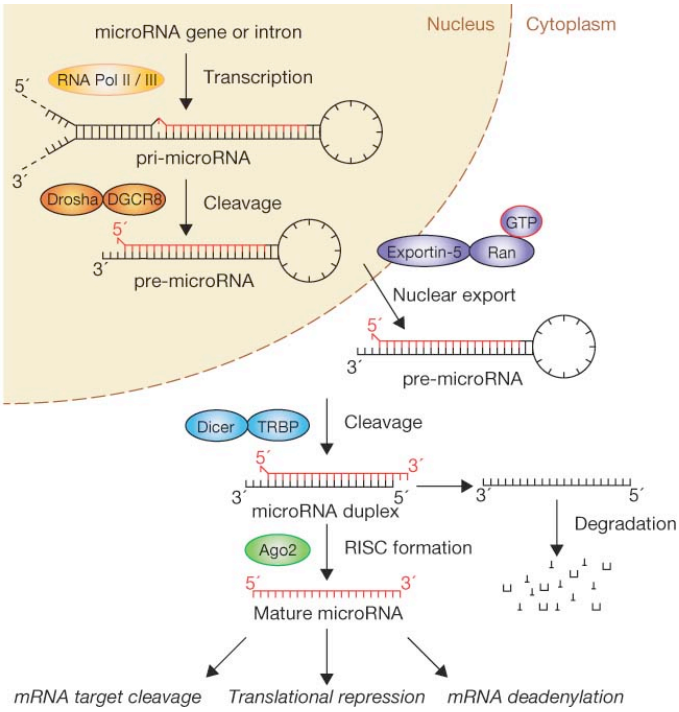


Figure 1.2: The Canonical miRNA biogenesis pathway. The figure was adopted from Winter et al. [69].

Intensive efforts to decipher the miRNA biogenesis process have also brought about several interesting and potentially significant observations. One intriguing observation made during miRNA maturation is that Loqs, the Dicer-1 partner, can tune where Dicer-1 cleaves through using different Loqs isoforms in flies, which results in miRNA mature sequences with a distinct seed sequence and target specificity [91, 92]. The mammalian Dicer-partner TRBP, the Loqs homolog, similarly tunes where Dicer cleaves pre-mir-132 in mice [91, 92]. These observations are particularly interesting because they indicated that one arm of the miRNA precursor may have the potential to produce more than one functional miRNA. In flies, the relative abundance of Loqs isoforms varies widely among tissues and developmental stages [91], which may underlie the previous observation that the relative abundance of miRNA isoforms varies among tissues and developmental stages [61, 64]. Perhaps the relative abundance of Loqs isoforms is regulated across development and differentiation to ensure the correct relative abundance of different miRNA isoforms from various pre-miRNAs. Since the partner proteins of Drosha are also in charge of recognizing the miRNA hairpin structure, it is

not surprising that DGCR8 can also tune the cutting of Drosha, which also creates miRNA sequence variance [93]. These observations suggest that different miRNA isoforms may not merely be miRNA processing noises or sequencing artifacts, but instead potential functional molecules since their generation may occur under specific regulation. Thorough and systematic analysis is needed to estimate the authenticity and functionality of miRNA isoforms.

The miRNA annotation and nomenclature

Numerous miRNA discovery studies including ours have led to identification of thousands of miRNAs in diverse species [36]. When working with such a vast number of miRNAs, proper nomenclature is important to distinguish between gene loci, transcripts and products. Correspondingly, a specific database and nomenclature system were developed for miRNA annotation. Formerly known as the microRNA Registry, the miRBase database represents a widely used primary online repository for miRNA annotation [66, 94]. According to miRBase, miRNAs are named in sequential order of discovery. The miRNA names are taken the form hsa-mir-19a. The first three letters denote the species (*hsa* for *Homo sapiens*). The mature miRNA is designated as miR-19a, while mir-19a refers to pre-miRNA or pri-miRNA. Distinct precursor sequences and genomic loci that produce identical mature sequences are assigned names in the forms hsa-mir-19b-1 and hsa-mir-19b-2. Lettered suffixes signify closely related but not identical mature sequences—for example, hsa-miR-19a and hsa-miR-19b are expressed from precursors hsa-mir-19a and hsa-mir-19b, respectively. miRNA cloning studies sometimes, and small RNA deep sequencing studies more commonly identify two ~22nt miRNAs sequences that originate from two arms of the same predicted precursor [35, 67]. When the relative abundances clearly indicate which is the predominantly expressed miRNA, the mature sequences are assigned names in the forms miR-19a (the predominant product) and miR-19a* (from the opposite arm of the precursor). When the data are not sufficient to determine which sequence is predominant, names like miR-142-5p (from the 5' arm) and miR-142-3p (from the 3' arm) are assigned. It should be noted that although in most cases the miRNA* is degraded from incorporating into RISC, it has been suggested that the miRNA* can act as functional miRNA as well. For instance, through a comparative approach that estimates miRNA functionality based on the conservation status of miRNA target sites, 23 functional miRNA* were identified [95]. Furthermore, the arm that makes the dominant product may change in different tissues, stages and species [96]. Therefore, using -5p/-3p nomenclature to replace miR/miR* nomenclature may be more appropriate for miRNA annotation.

The miRBase contains 8,273 miRNA sequences annotated in primates, rodents, birds, fish, worms, flies, plants and viruses (version 12) [97], providing the potential for cross-species miRNA comparison. However, several caveats should be considered when using miRBase. Although the primary goal of miRBase is to provide miRNA annotation and maintain consistent miRNA nomenclature across all species, miRNA names themselves do not always indicate orthologous relationships across species. miRBase clearly states that miRNA names can convey only limited information and are entirely unsuitable for encoding information about complex sequence relationships [97]. Thus, miRNA orthologous relationship assignment based on miRNA names is inappropriate. Furthermore, largely due to the

difference in the scope of miRNA study in different species, the quantity and quality of annotated miRNAs vary a lot even between closely related species. For example, the best annotated species, human, had 866 mature miRNAs annotated in miRBase (version 12), whereas far fewer miRNAs are annotated in chimpanzee (92) and rhesus macaque (485). Therefore, comparative analysis of microRNA across-species cannot fully rely on the miRNA annotation from miRBase. Instead, a reliable miRNA ortholog prediction and orthologous relationship delineation methods are needed.

The features of miRNA

In the next section, I summarize the features of miRNA genes with respect to genomic arrangement, sequence and structure, seed region and phylogenetic distribution.

The genomic arrangement

With respect to their genome context, miRNAs are ubiquitous in animal genomes [98, 99]. miRNAs are often transcribed as independent units in the intergenic region, many of which consist of polycistronic clusters containing multiple miRNAs [98, 99]. Duplication events contribute to miRNA expansion and cluster formation. Some clusters were formed through miRNA duplication of a single miRNA (homogenous cluster), which leads to local enrichment and amplification of a specific miRNA expression level. The homogenous clusters may help gain large dosage and enlarge the regulatory power. Intriguingly, it is more common for animal miRNA clusters to encode unrelated mature miRNAs (heterogeneous cluster) [100], which means miRNAs of a given cluster can have distinct target pools. In theory, fast diverging speed might be one possible reason for formation of heterogeneous cluster through specific miRNA duplication, but this remains to be fully tested. Nevertheless, considering the co-expression pattern for miRNAs from the same cluster, it is possible that these unrelated mature miRNAs might share certain functional relationships. The most prominent support for this hypothesis comes from the *mir-17-92* cluster, which acts corporately not only in tumor formation but also in development of the lungs and immune system [101-103]. miRNAs are also present in the intragenic region and mostly within introns, which presumably arise from further processing of the excised introns of protein-coding genes [98], whereas some sense-oriented intronic miRNA genes may also have their own promoters located in upstream intronic regions [104, 105]. Intriguingly, intronic miRNAs with their own promoters were more evolutionarily conserved than other intronic miRNAs [104, 105].

The miRNA precursor sequence

Partially due to the miRNA's short mature sequence length, the most notable features of miRNAs are embedded in the secondary structure of their precursor sequences instead of the sequence motif in mature sequences. Metazoan miRNAs are defined by stem-loop secondary structures of their precursor sequences. Inspection of the RNA stem-loops that were predicted to flank miRNAs revealed no sequence similarity. However, these stem-loops were of comparable size and were all predicted to form simple but imperfect hairpins [106]. The canonical miRNA precursors are usually ~70 nt long and have ~16 bp of complementarity between the two arms [107]. Once transcribed, the stem-loop hairpin structure distinguishes

miRNA from other hairpin containing noncoding RNAs for miRNA processing machinery and therefore dominates the miRNA maturation processes. Unique sequence and structure motifs have been revealed through experimental analysis of represented human miRNA precursors, including 48 structure motifs such as internal loops, bulges and mismatches other than G-U wobble pairs [108]. Computational analysis further identified six common features for human precursors, including folding free energy of the longest nonexact stem, the maximum number of consecutive C's in the hairpin, the maximum number of consecutive G's in the hairpin, folding free energy adjusted by the hairpin size, the average folding free energy of the exact stems and the size of bulges [109]. Taken together, the structure analysis and thermodynamic profiling suggest that miRNA precursor hairpin may be considered a mosaic of more and less stable regions that occur at certain sequence intervals and have both a structural and a functional meaning.

The miRNA mature sequence

Although no specific sequence motif has been associated with miRNAs mature sequences [110], several sequence features or propensities do exist. First, the length of mature miRNAs usually ranges from 20 to 24 nt, and a length of 22 nt is most common for miRNA mature sequences [111]. The length preference had been proven to be associated with the general miRNA maturation processing machinery. Structure study indicated that Dicer could act as a "molecular ruler" to cleave double-stranded RNA (dsRNA) substrates at a set distance from one end (~22 nt away from the base of the dsRNA stem) to generate miRNA mature sequence ends [112]. Electron microscopy single-particle reconstruction determined the domain arrangement of human Dicer, which designated the major unannotated region as a "ruler domain" between the "platform"/PAZ tandem and catalytic core (dsRBD and RNasIIIa/b tandem), thus providing an internal 22 nt gauge [113]. Besides the length preference, the mature miRNA sequences are prone to beginning with a uracil (U), which seems to be the only sequence feature of mature miRNAs [114]. 5'-U bias is not only important for miRNA strand selection process during miRNA biogenesis [35, 82], but also is associated with higher efficiency of miRNA-mediated target gene silencing, which might be related to the formation of the ternary miRNA-Argonaute2-mRNA complex and partly associated with the strand selection process [114-116].

The miRNA seed region

With regarding to function and evolution, the most prominent feature of miRNA comes with the seed region. For a given miRNA, the seed region or seed sequence is a heptamer sequence situated at positions 2-8 from miRNA mature sequence 5'-end [117-119]. The first clue of the importance of the miRNA seed region is based on the observation that the lin-14 UTR has core elements of complementary to the 5' region of the lin-14 miRNA [46]. Later on, accumulating observations providing experimental, evolutionary and computational evidence further strongly confirmed that the seed region is the most important.

1) Lai et al. showed that the seed region of miRNAs in *Drosophila melanogaster* is perfectly complementary to the elements in 3'UTR that were previously shown to mediate mRNA decay and translational repression [120].
2) miRNA orthologous sequences comparison demonstrated that the miRNA seed

regions were the most conserved portion among metazoan miRNA homologs [121].

3) Within the miRNA complementary sites of the validated targets of conserved miRNAs in *Drosophila melanogaster*, mRNA residues that paired to the miRNA seed region (seed matches) were perfectly conserved in orthologous regions of other *Drosophila* species [118].
4) For both invertebrate and mammalian miRNA target prediction, requiring the perfect base pairing to the heptamer spanning the miRNA seed region was much more productive and reported significantly fewer false positives than was requiring pairing to any other heptamer of the miRNAs. Conserved pairing to the seed region can be sufficient on its own for predicting conserved targets above the noise of false-positive predictions [117, 118, 122]
5) Small RNA overexpression experiments by overexpression miRNA-like sequences were almost solely affected by the nucleotide substitutions that disrupt seed pairing [123-125].
6) The most significantly enriched heptamer motif from down-regulated genes was the one that was complementary to the miRNA seed region in miRNA mimics transfection experiments [126].
7) miR-141 and miR-200c differ by one nucleotide in their seed region. Deletion miR-141 or miR-200c led to barely overlapped disregulated genes [127].

The importance of miRNA seed region brings about three conclusions.

1) miRNAs sharing the same seed region can be classified into one miRNA family that may regulate the same pool of target genes [122]. miRNAs differ in their seed regions even though one nucleotide may target substantially different targets. Consequently, the complex miRNA isoforms pattern should be treated carefully for miRNA quantification using small RNA sequencing data.
2) The conservation of miRNA seed regions is coincident with the conservation of target sites (seed matches), indicting the existence of co-evolution between miRNAs and regulated target genes [128]. It also suggests that functionality of conserved miRNAs can be inferred from the conservation status of their seed matches on 3'UTR.
3) To predict more confident miRNA target genes, conserved Watson-Crick pairing to the seed region must be considered. In line with this, when evaluated on the basis of proteomic changes after miRNA overexpression or deletion, prediction algorithms that required stringent seed pairing, such as TargetScan, performed better than those that did not [129, 130]. Tools that allowed mismatches or wobbles to miRNA seed region performed much worse [129, 130].

The miRNA phylogenetic distribution

Several features of miRNA phylogenetic distribution have been revealed by a number of studies on the distribution of miRNAs across animals. First, miRNAs are a class of ancient regulators that are present early on in the animal evolution. For instance, miR-100 is conserved between eumetazoans [131]. Second, miRNAs are continuously added to metazoan genomes leading to miRNA expansion. For example, 34 miRNAs that are

conserved between protostomes and deuterostomes indicate a burst of innovation at the base of bilaterian lineage [132]. Additional miRNA expansions have been observed in the lineage leading to placental mammals [133]. Third, miRNAs display low levels of secondary gene loss after their emergences in a particular lineage. Unlike retroposons, miRNAs are not lost at an increasing rate with time, but are instead largely retained in most, and sometimes in all, descendant lineages [134]. The continuous gain of novel miRNAs coupled with rare secondary gene loss lead to the ever-expanding repertoires of miRNAs, which have been shown to be directly correlated with morphological complexity, suggesting that the innovation of miRNA may be important to the emergence of increasingly complex cell types, tissues and organisms [128, 132, 133, 135].

1.3 Long Noncoding RNA (lncRNA)

LncRNAs is a large class of noncoding transcripts representing a heterogeneous group. Functionally, lncRNAs may not be less important than miRNAs. Although understanding of lncRNAs' functional roles is just beginning, the limited verified examples of lncRNA have demonstrated their influences on a diverse range of biological processes from developmental control to disease progression [28]. The best-known example of lncRNA functionality is X-inactive specific transcript (Xist), which is required for transcriptional silencing of one X chromosome during development in females across mammalian species [26].

Determining the function of individual lncRNAs remains a challenge. Nevertheless, rapid progress has been made with regard to approaches for lncRNA identification. Several features associated with lncRNAs were further revealed by globally analyzing the identified lncRNAs. In this section, I summarize the current knowledge about lncRNA genes, especially for the approaches of lncRNA identification and several general sequence and expression features of lncRNAs. The definition of lncRNA is also discussed.

The definition of lncRNA

LncRNAs include all transcripts with lengths varying from 200 nt to over 100,000 nt and represent a heterogeneous group. As would be expected from their general name, no unified definition has been reached for them at the moment. The definition proposed by HUGO Gene Nomenclature Committee (HGNC) describes lncRNAs as spliced, capped and polyadenylated noncoding RNAs [136]. However, because the existence of functional unspliced lncRNAs has been demonstrated [28], this definition is not complete. At present, lncRNAs are operationally defined as a class of RNAs longer than 200 nt and lacking clear protein-coding potential [28]. This definition is also not perfect because the 200nt requirement is arbitrarily chosen by the current RNA purification protocol limitation, which takes little consideration of functional meaning [16]. Nevertheless, the size cutoff clearly distinguishes lncRNAs from small regulatory RNAs such as miRNAs or piRNAs. In a sense, the requirement of a transcript length cutoff may be rather important for distinguishing putative lncRNAs from so-called transcriptional noise generated by the pervasive transcription.

Since the length may not be essential, the core of the lncRNA definition is lack of coding potential. Experimentally, ribosome profiling is useful for determining the translation status of a given transcript [137]. However, it is not realistic to carry out such complex experiments for every lncRNA study due to the requirement of high quality of biological samples, which may leave the application of ribosome profiling to cell lines. Thus, currently, transcript coding potential is mostly accessed through computational analysis. Computationally, one measurement for estimating transcript coding potential is based on the length of open reading frame (ORF). According to the traditional cutoff for protein-coding transcripts, transcripts with a maximal ORF <100 amino acids (aa) were defined as noncoding transcripts [138]. However, functional polypeptides shorter than 100aa have been identified [139]. Furthermore, transcripts containing ORF >100aa are unnecessarily to be translated into proteins. Thus, ORF size alone is not a good indicator for accessing coding potential. Many studies assessed coding potential by translating each lncRNA in all three frames and performing homology queries across large protein family and domain databases (i.e., PFAM) [15]. This analysis is a

good initial indication of protein-coding capacity but may miss a not insignificant portion of protein sequences without clear domains or newly evolved protein sequences. Along with the growing expansion of the lncRNA catalog, more sophisticated methods for estimating transcript coding potential by integrating multiple features were developed. These tools can be largely separated into three categories. The first distinguishes coding from noncoding transcripts based on multiple alignments to calculate the phylogenetic conservation score using codon substitution bias such as that used in the Phylogenetic Codon Substitution Frequencies (PhyloCSF), RNAcode methods [140, 141]. The second category predicts coding potential based on the combination of ORF quality and homologous similarities for protein evidence through pairwise alignments using BLAST such as that used in the Coding-Potential Calculator (CPC) method [142]. The third category estimates coding potential using intrinsic sequence features of coding sequences in an alignment-free way such as that used in the Coding Potential Assessment Tool (CPAT) [143]. Accumulating numbers of lncRNAs have been discovered, resulting in the conclusion that most lncRNAs are less conserved and tend to be lineage-specific [14]. This observation greatly limits the discrimination power of multiple alignments-based methods in the first category, and to a less extent, affect the methods in the second category due to certain species-specific protein-coding genes. By contrast, CPAT can access coding potential independent of transcripts' conservation status [143]. Nevertheless, by taking advantage of the continuous completeness of protein-coding gene collections in diverse species, CPC can be considered one good complementary method for CPAT due to the different features the two methods use.

The identification of lncRNA

Based on their general definition, lncRNAs can be identified by selecting RNA sequences with low coding potential computationally. The advances of coding potential estimation algorithms allow estimating transcript coding potential more precisely; thus, the main challenge of lncRNA identification is detecting novel transcripts or novel transcribed regions. This is a rather straightforward task that can be achieved by the direct detection of the transcribed RNAs. However, conventional gene expression microarrays are almost only designed to detect the expression of protein-coding mRNAs, so unbiased high-throughput RNA detection methods are therefore required that mainly include tiling arrays and high-throughput RNA sequencing (RNA-seq).

Tilling array

Tilling array allows analysis of transcription from specific genomic regions and was initially used for both identification and expression analysis of lncRNAs [16]. In this technique, cDNA is hybridized to microarray slides containing overlapping oligonucleotides that encompass either specific chromosomal regions or a complete genome (whole genome tilling array). Resolution of the hybridized genomic DNA sequence can be adjusted by changing the length of the overlapping sequences between two neighboring probes. For instance, Rinn et al. investigated lncRNAs expressed in the region of the human HOX genes by designing 400,000 probes of 50 bases in length, with each probe overlapping the next one by 45 bases to cover all four human HOX gene clusters. This configuration allowed for the identification of hybridized DNA sequences at 5-base resolution, resulting in the discovery of the lncRNA

HOTAIR that was transcribed from an intergenic region within the HOXC cluster [25]. Although tilling array is useful for detecting transcribed regions at high resolution, several limitations are also obvious. First, being a kind of microarray, tilling array has some limitations inherited from traditional microarrays, such as high background noise owing to cross-hybridization and a limited dynamic range for detecting both very lowly and highly expressed lncRNAs because of both background and saturation of signals. In addition, unless the target region is reasonably designated, a drawback of the tiling array approach is its high cost, especially for using whole genome tilling array to detect transcribed regions genome-wide. Furthermore, because transposable elements (TEs), a major class of repeat, make up a substantial fraction of mature lncRNA transcripts (>30% in human) [144], a notable portion of the genome is difficult to interrogate owing to lack of appropriate probes. Tiling array also generally needs to be custom-made to meet diverse needs, so appropriate probe design raises another potential problem.

RNA-seq

Currently, sequencing of transcriptome using RNA-seq is the most powerful approach for de novo discovery and expression analyses of lncRNAs. In this method, a population of RNA (total or fractionated, such as poly(A)+) is converted to a library of cDNA fragments, with proper amplification. Corresponding molecules are then sequenced in a high-throughput manner to obtain short sequence reads from one end (single-end sequencing) or both ends (paired-end sequencing). There are several types of sequencing technologies, but Illumina platforms are currently the most commonly used for RNA-seq experiments. For instance, as a foundation study, by utilizing RNA-seq reads across 24 tissues or cell types based on Illumina platforms, Cabili et al. identified over 8,000 human lincRNAs, opening the avenue to utilizing RNA-seq for characterizing lncRNA repertoire [15]. RNA-seq possesses several key advantages over tilling arrays. First, the prominent advantage is that RNA-seq can detect transcripts independent of existing genomic sequence and annotation at a genome-wide level, which is most important for lncRNA identification. Compared with tilling array, which struggles to balance throughput and cost, RNA-seq allows sequencing further and deeper due to the ever-deceasing sequencing cost. What $1 used to sequence—1 base—has increased to 10,000 bases. Thus, high sequencing depth enables unbiased and genome-wide lncRNA identification. RNA-seq is further leveraged by strand-specific sequencing protocols, which enable measuring each read with transcription direction. Currently, the strategies that have been developed to generate strand-specific information generally rely on one of three approaches. The first involves the ligation of adaptors in a predetermined orientation to the ends of RNAs or to first-strand cDNA molecules. The known orientations of these adaptors are used as reference points to obtain RNA strand information [145]. The second approach is direct sequencing of the first-strand cDNA products that are generated [146]. The third approach involves selective chemical marking of the second-strand cDNA synthesis products or RNA [147, 148]. Significantly, strand-specific RNA-seq is not only important for assembling and quantifying overlapping transcripts from opposite strands of the genome [148], but also allows for determining the transcription direction for many single exon lncRNAs, such as Neat1, which is necessary for the formation of nuclear paraspeckles [149]. Besides these merits, RNA-Seq can also detect expressed transcripts at a much larger

dynamic range with low background noise. Except for special cases, only sequencing reads that can be mapped back to the genome unambiguously are considered useable; thus, there exists no upper limit for quantification compared to tilling array, which only correlates with the sequencing depth. Therefore, RNA-seq can detect transcripts including lncRNAs that are expressed either at very low or high levels. RNA-Seq also produces better quantification for transcripts expression, and the results of RNA-Seq exhibit high levels of reproducibility for both technical and biological replicates [60].

Although the high resolution and high sequencing depth of RNA-seq data makes it possible to capture all transcriptome elements, including numerous unidentified lncRNAs, billions of short reads pose a significant computational challenge. Thanks to the recent developments in transcriptome assembly approaches, reconstruction of the entire transcriptome by RNA-seq is feasible, even without a reference genome, thus providing the solution to identifying all transcribed RNAs. In general, current transcriptome assembly tools can be separated into two classes, reference-based assembly and de novo assembly [150]. When a reference genome is available, reference-based methods can be used for transcritptome assembly. Generally, reference-based approaches first aligned the RNA-seq reads onto the genome and then clustered the overlapping reads to build a graph representing all possible isoforms. The individual isoforms were finally resolved by traversing through the graph. Cufflinks, one of the most widely used reference-based assembly methods, was developed for efficiently reconstructing transcripts from mammalian-sized data sets [151]. The main advantage of reference-based approaches is high sensitivity, which can assemble transcripts with low expression abundance and generate more full-length transcripts [151]. Reference-based assembly can also be conducted with less memory and more efficiently using parallel computing in that mapped reads are clustered and can be processed independently. However, due to the reference-based algorithm itself, there are also several potential disadvantages. The success of assembly is greatly affected by the reference genome quality [152]. Some transcripts are lost due to incompleteness of the genome assembly. Some transcripts are separated into several parts due to the assembly's gaps of reference genome. The potential erroneously mapped reads are carried over into the assembly, resulting in false positives. The quality of spliced-aligned reads is another problem since the reads alignment tools often only search introns that are smaller than a length cutoff to reduce the required computational resources. Furthermore, multiplied mapped reads were usually discarded from the assembly, which may leave potential gaps in the assembly transcripts where the reads were not uniquely mapped. All these potential drawbacks are partially solved by assembling transcriptome using de novo approaches, which leverage the redundancy of short-reads sequencing to find read overlaps and assemble them into transcripts without any genome reference. Trinity is the most efficient De Bruijn graph-based de novo assembly tool [153]. Trinity implements a unique stepwise strategy by first greedily assembling a set of unique sequences that overlap, creating an independent De Bruijn graph for each group of sequences and assembling isoforms within the group, thus enabling processing in parallel to speed up the assembly process. Since de novo assembly approaches are independent of a reference genome, all problems caused by reference genome quality will not affect assembly quality. De novo assembly approaches are also immune to the potential problems of reads mapping such as erroneously continuous and spliced-aligned reads. Furthermore, multiple mapped reads are also resolved by the de novo

assembly approach, which may potentially make the assembly transcripts more complete. This is important because around half of the human genome is comprised of repeat sequences and repeats make up a substantial fraction of mature lncRNA transcripts (>30% in human). The main drawback of de novo assembly approaches is the requirement of a much higher sequencing depth than reference-based assembly approaches for full-length sequence assembly. The sequencing depth requirement further results in high computational resources needs. In addition, trans-spliced transcripts are not easily discriminated from chimeric reads or assembly artifacts. Nevertheless, with the continuing improvements and advances of de novo assembly algorithms and RNA-seq technologies, along with support from increasing powerful computational resources, de novo assembly approaches will no doubt contribute significantly to the transcriptome studies in the future. The lncRNA identification will therefore benefit from this approach.

The features of lncRNA

Although the lncRNA catalog is far from complete, several sequence and expression features are emerging based on previous lncRNA studies. Being a class of heterogeneous RNA, lncRNAs have been shown to overlap with several functional elements such as enhancer regions (eRNAs) [154], telomeric repeat regions (telomeric repeat-containing RNAs) [155], protein-coding promoter regions (promoter-associated RNAs) [18], 3'UTR regions (3'UTR associated RNAs) [156] and gene antisense regions (antitsense RNAs) [5]. The lncRNAs reside in the intergenic region and are usually called lincRNAs [10]. Previous studies have shown that more than 78% of "dark matter" of transcriptome was in the vicinity of known protein-coding genes [157], suggesting a possible link of the regulation between lncRNA and nearby protein-coding genes. Because many lncRNAs exert regulatory effects on the genomic loci they are derived from or on neighboring loci, it is possible to study the function of lncRNAs based on their genomic context. With regard to transcription processing, many, but not all, lncRNAs are processed as typical mRNAs, including 5' cap structure, splicing and polyadenylation. lncRNAs show a bias for having just one intron and a trend for less efficient cotranscriptional splicing than that of protein-coding genes [158]. Splicing may not occur at all in some cases for many lncRNAs, resulting in a group of single-exon lncRNAs such as Neat1 and Malat1 [149, 159]. The high expression abundance and sequence conservation are believed to be two hallmarks of functional gene categories. However, lncRNAs usually lack both features. The pioneering studies analyzing lncRNAs in mouse and human revealed their low expression nature [10], which may in part account for their invisibility in previous researches. With respect to sequence conservation, lncRNAs are weakly constrained. The average nucleotide substitution ratio was 90%-95% for lncRNAs compared to ~10% for protein-coding genes [160]. Nevertheless, the sequence analysis also showed that lncRNAs are under higher selective pressures than ancestral repeats and random intergenic regions that are considered to be under neutral selection, which suggests lncRNAs are not totally transactional noises [10, 161]. The seminal example of non-conserved but essential lncRNAs comes from the most intensively studied Xist, which displays little sequence conservation throughout the eutherian lineage [162]. The low sequence conservation may also be explained by the high rate of sequence evolution for the lncRNAs that have more plastic structure-function constraints, similar to certain fast-evolving promoters, enhancers or other

regulatory elements [163]. Accumulated results have further suggested that lncRNAs are expressed in a highly tissue-specific or cell type-specific manner compared to protein-coding genes [15], which may underlie the observation of the generally low expression level of lncRNAs in complex tissues encompassing numerous cell types. Some lncRNAs are exported from the nucleus and may perform important functions in the cytoplasm, but the majority are found in the nucleus [14] and are particularly associated with chromatin [164], indicating that lncRNAs may play important roles in transcriptional regulation. Although this summary of lncRNA features might be only partial, it certainly deepens our understanding about lncRNAs and provides clues and guidance to further study of lncRNAs' functions. Besides, all these sequence and expression features are helpful for generally evaluating the quality of the identified lncRNA.

1.4 miRNA and lncRNA expression in the human brain

The human brain, comprised of an extraordinary number of subtypes of glial and neural cells, is the most sophisticated biological organ. It is now clear that these complicated features of the brain are mediated not only by protein-coding genes but also by cell type-, developmental stage- and stimulus-specific profiles of ncRNAs. In fact, because the proportion of non-coding DNAs correlates with the organism's complexity as well as the observation that the number of ncRNAs has increased with the neuronal complexity of metazoans while the number of protein-coding genes has remained relatively stable [165, 166], it is intriguing to speculate that ncRNAs may underlie the unique function repertoires of the brain in higher organisms and mediate the acceleration of human brain evolution. The expression enrichment of ncRNAs, including miRNA and lncRNA, in brains of higher organisms further enhances the link between ncRNAs and evolutionary innovations in brains [167]. Accumulating evidence indicates that both miRNA and lncRNA are essential to various neurobiological functions in the brain.

Multiple studies have shown that miRNAs are key regulators in brains, playing pivotal roles in diverse neurobiological processes ranging from synapse formation to neuronal cell identity establishment. One key experimental strategy for studying the function of miRNAs in the brain globally is to disrupt the miRNA biogenesis pathway. The Dicer knockout animals display several neural developmental defects, including brain size abnormalities and altered dendritic spine morphology in forebrain neurons [168, 169]. In flies, Drosha and Dicer mutations lead to defects in synaptic transmission in photoreceptor neurons [170]. Other complementary experiments that focus on individual miRNAs have further elucidated their propounding roles in modulating diverse neuronal functions [167]. For instance, miR-124 and miR-134 can modulate dendritic growth and arborization. Let-7 and miR-132 play roles in synapse formation. miR-1, miR-132, miR-134 and miR-181a are involved in synaptic function and plasticity. Strikingly, introduction of miR-124 leads to the expression profile of Hela cell line shift toward that of the brain by decreasing expression of dozens of non-neuronal genes [126], revealing the possibility that brain-specific/associated miRNAs can function as master regulators in establishing cell identity. The fact that the expression of a small number of miRNAs can promote reprograming from somatic cells into neuronal cells further highlights the extent to which miRNAs function as pivotal nodes in the regulatory networks that are responsible for establishing cell identity. miR-124, miR-9 and miR-9* are three well-known brain-specific/associated miRNAs. Previous studies have demonstrated that the introduction of miR-124 together with miR-9/miR-9* in human fibroblasts can reprogram them into neuronal cells [171]. The underlying mechanism of the action of these miRNAs may involve regulating the specific composition of subunits of ATP-dependent Brm associated factor (BAF) chromatin remodeling complexes that are essential for neuronal lineage maturation at particular stages. Due to the powerful regulatory potential of miRNAs in brain, it will be intriguing to speculate that miRNAs may also contribute to the unique intelligence and cognitive ability of the human brain, especially in the prefrontal cortex (PFC), which is critical to many cognitive abilities that are considered particularly human [172]. The PFC has remarkably expanded in size throughout human evolution, culminating in modern *Homo sapiens*. While the brain itself has only increased in size about threefold in the past five

million years (the estimated divarication time of human and chimpanzee), the size of the PFC has increased sixfold. The comparative study between human and other closely related primate relatives such as the chimpanzee using miRNA profiles in combination with gene and/or protein expressions may able to elucidate the potential contribution of miRNA to human PFC uniqueness.

Whereas most studies have focused on defining neurobiological roles for miRNAs, recent studies have also begun to characterize the expression and function of lncRNA in brains. The first clue indicating the specific function of lncRNAs in brains is from large scale RNA in situ hybridization analysis in mouse brain based on the resources from Allen Brain Atlas [173]. In situ hybridization showed that many lncRNAs are expressed in specific anatomical regions, cell types or subcellular compartments in the mouse brain, suggesting that many of these lncRNAs may be functional. Examining individual lncRNAs further reveals their diverse roles in the brain. For instance, knocking out lncRNA MEG3 in mice followed by microarray analysis revealed that MEG3 is involved in many processes in brain development, including calcium, Notch and Wnt signaling and long-term potentiation [174]. Malat1 is a ~7kb single exon lncRNA residing in the intergenic region of human genome that interacts with splicing factors and is implicated in nervous system development as it is expressed during later stages of neuronal and oligodendrocyte development [175]. Knockdown of Malat1 in cultured hippocampal neurons results in decreased synaptic density, while overexpression has opposite effects [176]. lncRNAs may also contribute to superior cognitive ability in humans. For example, the lncRNA HAR1F is derived from HAR1A, one genomic region exhibiting prominent signature of positive selection in human lineage [177]. HAR1F is especially expressed in a specific class of neuron cells in human developing neocortex between the 7th and 18th gestational weeks, a crucial period for cortical neuron specification and migration. The correlated expression between HAR1F with cortical patterning protein reelin indicates that HAR1F may participate in coordinating the establishment of regional forebrain organization. Similar to the speculation about miRNAs, the fast-evolving lncRNAs may also play important roles in the formation of human PFC uniqueness. Determining the function of individual lncRNAs remains a pretty tough task. Perhaps what is fundamental and required for the current lncRNA study is the genome-wide identification of lncRNAs in diverse tissues, cell types and developmental stages for fully cataloging lncRNA sequence and expression profiles. The lncRNA study in human PFC is also no exception.

1.5 Thesis outline

The goal of my work was to develop appropriate computational and statistical methods and pipelines for analyzing sequence and expression of miRNA and lncRNA on the basis of high-throughput data generated by small RNA sequencing and RNA-seq in primate brains. The work outlined in this thesis contains three main research tasks: (i) building a framework for across-species miRNA comparison study, (ii) identification of prevalent 5'-isoforms of annotated miRNAs expressed in the primate brain and prediction of their regulatory effects, and (ii) identification of novel lncRNAs expressed in the human brain and investigation of sequence, expression and function features of both novel and annotated lncRNAs throughout human brain development.

More specifically, in Chapter 3 of this thesis, I present a framework for across-species miRNA comparison by analyzing small RNA sequencing data generated from brains of humans and our two close relatives: chimpanzees and macaques. The complexity of the miRNA repertoire is highlighted. Subsequently, a new miRNA quantification method is proposed. A miRNA ortholog prediction procedure, which is fundamental for across-species miRNA comparison study, is developed. By integrating all the developed methods and pipelines described in this chapter, differentially expressed miRNAs between human and other two primate species and their contribution to target gene expression divergence at both mRNA and protein levels in the prefrontal cortex are investigated.

In Chapter 4 of this thesis, I describe a study of miRNA 5'-isoforms including their identification and functionality predictions in the human prefrontal cortex. A comparative approach for predicting functional conserved miRNAs is developed, which allows for identifying the majority of highly expressed functional miRNAs as well as predicting putative functional miRNA 5'-isoforms. The regulatory effects on exclusive target genes of two predicted functional 5'-isoforms were verified by analyzing public microarray data.

In Chapter 5 of this thesis, I describe a genome-wide lncRNA identification and feature investigation study by integrating de novo transcriptome assembly and downstream lncRNA analysis procedures on the basis of age-series RNA-seq data from human prefrontal cortex. Systematic characterization of the identified lncRNAs leads to the identification of a novel class of bidirectional promoters, displaying unique sequence and epigenetic features, which are associated with the expression of neuronal genes.

2. Computational Methods

2.1 Analysis of miRNA sequence and expression quantification

Small RNA sequencing data processing procedure

A pipeline was developed to process small RNA sequencing data from the Illumina sequencing platform. The processing pipeline took raw sequencing reads as input and reported mapped reads after processing using the following three steps.

1) Reads filtering step

First, reads of low quality, reads of low composition complexity and reads derived from adaptor contamination were filtered out. Specifically, reads containing nucleotide bases other than [ATGC] were treated as low quality reads and were removed. Then reads derived from adaptor contamination were identified and filtered out by perfectly matching their 5' first 10nt sequence to the 5' first 10nt sequence of 3' adaptor sequence [5' TCGTATGCCGTCTTCTGCTTGT 3']. Finally, low complexity reads were filtered using mdust algorithm [178]. mdust is a program for identifying and masking out low complexity regions from nucleic acid sequences by searching for regions with poor tri-nucleotide content.

2) 3' adaptor trimming step

The reads passed through the reads filtering step were further processed to remove 3' adaptor sequence. Due to the intrinsic character of Illumina sequencing technology, errors will accumulate at a much higher rate at the reads 3' end [179]. Therefore, two to three mismatches were allowed for 3' adaptor trimming empirically. Specifically, the remaining sequences were trimmed by matching the 3' adapter sequence to the 3'-end, allowing 3 mismatches if the length of the match was greater than 10 nucleotides and allowing 2 mismatches if the length of the match was between 5 and 10 nucleotides. Reads without detectable 3' adaptor sequence or with length no more than 17nt after trimming were discarded.

3) Reads mapping step

The trimmed reads were further mapped to the corresponding genome using Short Oligonucleotide alignment program (SOAP) with the following parameters (*-v* 0, *-g* 0, *-r* 2). SOAP is a program for efficient ungapped and gapped alignment of short oligonucleotides onto the reference sequence. It is specifically designed to map huge amounts of short reads produced by Illumina high-throughput sequencing [180]. Briefly, SOAP first loaded the reference sequence (e.g., the human genome) into memory using 2-bits-per-based encoding (required L/4 bytes for the reference with length size L) and built the seed index with hash tables. Then for each read, SOAP created seeds and searched the corresponding index table for candidate hits based on the number of mismatches between read and reference. In this study, I set parameter *-v* 0 and *-g* 0 to obtain reads that perfectly and ungapped matched to the genome and set parameter *-r* 2 to report all loci of both unique and multiple mapped reads.

The mapped reads were further required to range from 18 to 28 nt in length. Of the mapped reads, less than 1% was expected to be mapped incorrectly, as determined by a mapping of scrambled reads with the same length and mononucleotide composition distribution 100 times. The ungapped match (*-g* 0) was required for two reasons: 1) Biologically, known small RNAs are not under splicing process during their maturation; and 2) Technically, reads have very few insertions and deletions from Illumina sequencing platform. Hence, gapped aligned reads should not be considered for small RNA analysis. The requirement of perfect match (-v 0) is to limit the possible read cross mapping between miRNA members from the same miRNA family. Since many annotated human miRNAs deriving from the same miRNA family only differ by one or two nucleotides (e.g., let-7b and let-7c), allowing mismatches may lead to reads cross mapping between closely related miRNA members. Unlike mRNA RNA-Seq analysis, which is usually based on uniquely mapped reads, multiple mapped reads were allowed (-r 2). This is because the majority of highly expressed miRNAs have more than one loci on the corresponding genome, such as the mature sequence of brain-specific miR-124 that has three loci on the human genome.

miRNA expression quantification procedure

A new quantification procedure was developed to estimate miRNA expression level based on mapped reads from Illumina sequencing data and mature miRNA annotation that was downloaded from miRBase (version 12) [97]. First, all sequences mapping within three nucleotides upstream or downstream of the annotated 5'-position of the mature miRNAs were retained, and then reads from all genomic loci producing the same mature miRNA were united. The union step is indispensable for correct quantification of duplicated miRNAs that have multiple loci on the genome because it avoids double counting the multiple mapped reads deriving from the same miRNA. Next, for each mature miRNA, the sequence with a maximal copy number was designated as the reference sequence that was used to represent the mature sequence for cognate miRNA. Although deemed to be the central global repository for all published miRNAs, miRBase is a community resource with somewhat inclusive policy [97]. Therefore, defining the reference sequence was necessary and useful to further correct false annotated mature miRNA sequences or to define the most frequently used miRNA major isoform sequence in specific tissue or cell type. Finally, the expression level of miRNA *i* denoted as e_i was calculated as a sum of the copy number of the reference sequence r_i and the sequences mapping at the same 5'-end position as the reference sequence s_{ij}. The last step enabled utilizing substantially more reads with 3'end variants to the reference sequence for miRNA quantification.

$$e_i = r_i + \sum_j s_{ij}$$

To evaluate the quantification performance, this new miRNA quantification procedure was compared with another miRNA quantification procedure that measures miRNA expression by counting the number of reads that exactly match the annotated mature miRNA sequences. The performance was estimated according to the following aspects: 1) the total read counts used

for miRNA quantification; 2) the number of quantified miRNA; 3) miRNA expression correlation between human replicate samples; and 4) the number of corrected mature miRNAs that were erroneously annotated in miRBase.

Analysis of miRNA ends heterogeneity

To estimate the 5' and 3' end heterogeneity of miRNA *i*, first, all sequences mapping within six nucleotides upstream or downstream of the annotated 5'-position of the mature miRNA were retained and united. Then the sequence with a maximal copy number was designated as the reference sequence, and the rest of the sequences were designated as shifted sequences. Finally, the heterogeneity of its termini h_i was calculated as a ratio by dividing the sum of the absolute offset distance between the observed 5'- or 3'- ends and the ends of the reference sequence by the copy number of reference sequence r_i. For shifted sequence *j* of miRNA *i*, the absolute offset distance was calculated as the product of the absolute value of the shifted number of nucleotide d_{ij} and its copy number s_{ij} .

$$h_i = \frac{1}{r_i}\sum_j \left| d_{ij} \right| s_{ij}$$

2.2 Across-species miRNA sequence and expression comparison analysis

miRNA orthologous gene prediction

Since the miRNA annotations in chimpanzee and macaque were quite poor in miRBase, I developed a miRNA orthologous gene prediction procedure (MOP) to identify both miRNA precursor and mature sequences in chimpanzee and macaque based on human miRNA annotation. The ortholog finding procedure consists of two consecutive steps: precursor ortholog finding and mature ortholog finding.

To identify human precursor orthologs, human miRNA annotations were downloaded from miRBase (version 12), including both sequence and genomic loci of precursor sequences. Then the best precursor orthologs were extracted by using a combination of reciprocal BLAT, BLAST and liftOver in chimpanzee and rhesus macaque genomes.

BLAST (Basic Local Alignment Search Tool) [181] and BLAT (BLAST-Like Alignment Tool) [182] are the most widely used traditional local alignment tools for sequence searching. One common and important application of BLAST and BLAT is cross-species sequence homolog finding. BLAST searches for high scoring sequence alignments between the query sequence and target sequences in the database using a heuristic approach that approximates the Smith-Waterman algorithm. BLAST can be used for across-species homolog finding for both closely and remotely related species. BLAT is a very fast sequence alignment tool similar to BLAST. For DNA queries, BLAT is designed to quickly find sequences with 95% or greater similarity of lengths of 25 bases or more. In practice, DNA BLAT works quite well in finding human orthologs in other primate species [183]. liftOver is a tool primarily

designed for converting genome coordinates and genome annotation files between genome assemblies [184]. In practice, it can also effectively be used for detecting orthologous coordinates between closely related species such as primates. LiftOver had also been used for finding orthologous coordinates between moderately divergent species such as humans and mice [185]. Sequence coordinates file and chain file are indispensable to executing liftOver. A sequence coordinates file is required in BED format. Creating a chain file is very similar to a whole-genome alignment, in which for each region in the genome, the alignments of the best/longest syntenic regions are used to translate features from one version of a genome to another. The chain files can be downloaded from the UCSC genome browser [184].

Specifically, the ortholog finding procedure mapped all annotated human miRNA precursors to the chimpanzee and rhesus macaque genomes using reciprocal BLAT, BLAST and liftOver and required one precursor ortholog to be supported by at least 2 out of 3 methods. Ortholog finding using reciprocal strategy is important for identifying the authentic orthologous pairs in case lineage specific miRNA duplication and missing happened. It has been efficiently and widely used for protein-coding orthologous gene detection.

For reciprocal BLAT, the ortholog prediction procedure chose the following parameter configuration: [-stepSize = 5 -repMatch = 2253 -minScore = 0 -minIdentity = 0]. This parameter configuration was adopted from the UCSC genome browser; it is an optimal setting for a wide variety of uses and reports all valid matches [184]. The length of each precursor ortholog was further required to be more than 70% and less than 130% of the query sequence. Similarly, for reciprocal BLAST, the ortholog prediction procedure chose the parameter configuration [-F F -b 1 –e 10^{-5}] to search for precursor orthologs and again required the length of the hit sequence to be more than 70% and less than 130% of the query sequence. Parameter –F was set to F to inactivate masking precursor sequence. Parameter –e was set to 10^{-5} to only report hits with E-value of less than 10^{-5}. Parameter –b was set to 1 to report all the best hits for each precursor. For reciprocal liftOver, based on human precursor genomic loci, the ortholog finding procedure chose the website parameter configuration with Perl LWP module [hglft_minMatch = 0.6 hglft_minSizeT = 0 hglft_minSizeQ = 0 boolshad.hglft_multiple = 0] and similarly required the length of the hit sequence to be more than 70% and less than 130% of the query sequence. The Parameter hglft_minMatch was set to 0.6 to find precursor orthologs with lengths longer than 60% of the query precursor sequence length. hglft_minSizeT and hglft_minSizeQ were set to 0 to increase searching sensitivity. boolshad.hglft_multipl was set to 0 to report all best hits for each precursor.

Based on the obtained precursor sequences, the next step is to identify the miRNA mature orthologous sequence in chimpanzee and macaque. For a given human miRNA precursor, the precursor sequences of humans and other two primates (if available) were first aligned using ClustalW2 [186] with default parameters. Then the mature orthologs were extracted based on aligned precursor sequences.

To evaluate performance sensitivity, the predicted miRNA orthologs were compared to the annotated chimpanzee and macaque miRNAs registered in miRBase (version 12). The sensitivity was evaluated as the proportion of known mature miRNAs that can be detected

using the orthologous prediction method. The criteria of the reciprocal ortholog finding strategy and additional hit sequence length filtering as well as the requirement of combining both local and global alignment methods result in the high specificity of the ortholog finding procedure. For the predicted miRNA precursor orthologs, the proportion of miRNA precursors identified by all 3 tools and the orthologous precursor sequence length differences compared to the corresponding human miRNA precursors were analyzed to further estimate the quality of predicted miRNA orthologs. For predicted miRNA mature orthologs, I checked the miRNA expression correlation within species and between species after quantifying miRNA expression in human, chimpanzee and macaque with the new miRNA quantification procedure. The clustering pattern of the species with predicted miRNA mature sequence expression were further plotted and visualized using UPGMA and NJ trees.

miRNA differential expression detection

Characterizing miRNAs that are differentially expressed between conditions or between species is a common but important goal for most miRNA studies. miRNA expression measured by using deep sequencing technology was summarized as a read count matrix in which each column corresponds to a sample and each row is a miRNA. In this study, two approaches were employed to identify differential expression based on miRNA read count data between species.

The first approach is Fisher's Exact Test (FET) based method [54]. Before applying the statistical test, miRNA count data between two species was normalized using the quantile normalization method [187]. Quantile normalization is one of the most robust and effective normalization methods not only for gene microarray data, but also for miRNA count data [188]. miRNA expression matrix was sorted in a descending order in each column (corresponding to each sample), and then the highest expressed value of each column was averaged and used to replace the original highest expressed value in all columns. This process is repeated with what was originally the second highest value in each column, and the third highest, and so on. Finally, quantile normalization makes the entire distribution of data from a different sample the same.

After normalization, for each technical replicate, FET was applied to identify differential expression by using the combination of statistical significance, fold-change and detection level as criteria (FET $p<0.01$, FDR<10%, fold-change>2, at least 10 read counts in at least one of the two species). As a further requirement, the candidate differential expressed miRNA should fulfill these criteria in both technical replicates. The 2 x 2 contingency table below shows the miRNA expression used for testing differential expression of miRNA *i* between human and chimpanzee.

	Human	Chimpanzee	Total
miRNA *i*	n_{11}	n_{12}	n_{1+}
Remaining miRNAs	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

To calculate p-value, a 2 x 2 contingency table of FET is equivalent to the hypergeometric test:

$$p(n_{11}) = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{21}}}{\binom{n}{n_{+1}}}$$

The null hypothesis is rejected at significance level α if $p(x > n_{11}) = \sum_{x>n_{11}} p(x) < \alpha$.

An alternative approach to obtaining differentially expressed miRNA that is more sophisticated and taking reads overdispersion into consideration is using a model-based method. In this study, the edgeR (empirical analysis of digital gene expression data in R) [189] procedure for single factor differential gene expression analysis was used to identify differential expressed miRNA candidates between species. Similar to the aforementioned approach using FET, the first step of edgeR procedure is expression normalization. edgeR considered two technical factors to normalize sample-specific effects on read counts. The first factor is sequencing depth of each sample, which was normalized by varying sequencing depth as represented by library sizes (lib.size). The other factor is miRNA composition between samples. This factor should be considered when a small number of miRNAs took a substantial portion of total library size in sample *a*, but not in another, which will cause the rest of the miRNAs to be undersampled in sample *a*. edgeR normalized miRNA composition effect by using the Trimmed Mean of M-values (TMM) method [190] to find a set of scaling factors (norm.factors) that can minimize the log-fold changes (M-values) for most miRNAs between samples. To calculate these scaling factors, defining r_{ig} and r_{jg} as the number of reads from gene *g* in sample *i* and sample *j* and N_i and N_j as the total number of reads from sample *i* and sample *j*, respectively. Then for the pairs of sample (*i*, *j*), the gene-wise log-fold-changes were calculated as:

$$M_g = \log \frac{r_{ig}}{N_i} - \log \frac{r_{jg}}{N_j}$$

Gene expression levels were represented as the mean of log2 normalized counts:

$$A_g = \frac{1}{2}\left(\log \frac{r_{ig}}{N_i} + \log \frac{r_{jg}}{N_j} \right)$$

Both the M values and the A values were trimmed before taking the weighted average to robustly summarize the observed M values. The product of lib.size and norm.factors representing effective library size was finally used to normalize miRNA read count data.

Next, to address the so-called overdispersion problem, edgeR modeled count data with negative binomial (NB) distribution. The NB model has been shown to be a good fit to RNA-Seq data and is flexible enough to account for biological variability [191, 192]. Denote y_{gj} as the number of reads of gene g of sample j. edgeR assumes that y_{gj} has a mean μ_{gj} and a variance σ^2 as:

$$\sigma^2 = \mu_{gj} + \pi_g \mu_{gj}^2$$

where the dispersion $\pi_g > 0$ represents the overdispersion relative to the Possion distribution. The parameter μ_{gj} is determined by the expression concentration of gene g and sequencing depth of sample j. The dispersion π_g represents the squared coefficient of variation of the true expression levels between biologically independent samples. edgeR used the quantile-adjusted conditional maximum likelihood (qCML) method to estimate the dispersion parameters, including both tagwise dispersion and common dispersion. Tagwise dipersion is a measure of the degree of interlibrary variation for a specific tag. Estimating the common dispersion gives an idea of overall variability across the genome. edgeR first estimates a common dispersion for all the tags and then applies an empirical Bayes strategy for squeezing the tagwise dispersions toward the common dispersion.

Finally, once the NB model was fitted and dispersion estimations were obtained, edgeR conducted an exact negative binomial test to identify differential expressed miRNA [191]. The exact p-value was calculated by summing overall sums of counts that have a probability less than the probability under the null hypothesis of the observed sum of counts. In this study, based on the edgeR result, differential expressed miRNA should fulfill the following criteria: negative binomial test $p<0.001$, FDR <1%, at least 10 read counts in at least one of the two species.

2.3 Across-species miRNA regulatory effect analysis

mRNA expression quantification

The mRNA expression was measured based on Affymetrix Human Exon 1.0 ST Arrays in five human and five chimpanzee prefrontal cortex samples. The Affymetrix Human Exon 1.0 ST Arrays data were processed via several steps. The first step was probe masking. The the human and chimpanzee datasets were processed separately. For the human dataset, the human Exon 1.0 ST probes were mapped to the human genome (hg18) and the probes that matched the genome perfectly and uniquely were retained. For chimpanzee datasets, the same procedure was applied by mapping probes to the chimpanzee genome (panTro2.1). Finally, the probes that match both human and chimpanzee genomes were chosen for human and chimpanzee gene expression quantification. The second step was probe detection. To determine whether the signal intensity of a given probe was above the expected level of background noise, the signal intensity for each probe was compared to a distribution of signal intensities of the anti-genomic probes with the same GC content. Anti-genomic probes are specifically designed by Affymetrix to provide an estimate of the non-specific background hybridization. A probe was classified as detected if its intensity was larger than the 95% percentile of the anti-genomic probes with the same GC content. To further remove any possible systematic experimental bias among arrays, the PM-GCBG correction[1] and quantile normalization were performed [187]. PM-GCBG corrects the probe signal by subtracting the median of the probes' intensity values that have the same GC content as the given probe. Prior to normalization, all intensities were log2 transformed. The last step is gene expression summarization. A transcript was classified as detected if more than 80% of probes and at least ten probes per transcript were classified as detected. The intensities of transcripts were summarized by the median polish method. For the median polish method, given a transcript with j probes, its expression in array i was estimated using the average value of error-corrected j probe signals in array i. The errors were obtained by iteratively repeating the following steps: first subtracting the row median from each point in that particular row and then subtracting the column median from each point in that particular column until medians converge (converge to 0 or a small number, usually the iteration repeats in less than five times). Transcript Cluster Annotations file was used to map the transcript clusters annotated by Affymetrix to Ensembl genes. In cases in which multiple transcript clusters mapped to the same gene, the gene expression was calculated as the median of all corresponding transcript clusters.

Protein expression quantification analysis

The protein identification and quantification were performed in a so-called "bottom-up" manner: Proteins were first converted into peptides, and then the peptide sequences were determined by tandem mass spectrometry (MS/MS), combined with the use of search engines

[1]The original method is described in "White paper: Exon Array Background Correction" on the Affymetrix website.

and protein databases. Specifically, the protein expression abundance was measured using the Label-Free 2D-MS/MS Thermo-LTQ proteomics system with eight humans and eight chimpanzee prefrontal cortex samples. Based on the raw compound mass spectra data, peptide sequence identification was achieved by searching against the combined dataset containing a commonly used human peptide database [193] (IPI human v3.22) and its reversed version representing the mock dataset by using the SEQUEST program [194] in Bioworks 3.2 software suite. The hits from the mock dataset were considered false positives and further used to estimate the false discovery rate (FDR) of the identified peptides. A mass tolerance of 3.0 Da and one missed cleavage site of trypsin were allowed. Cysteine carboxyamidomethlation was set as static modification, and no other modification was checked. At FDR cutoff less than 0.5%, all matches passing a certain Xcorr and delta CN were considered valid. Furthermore, all peptides that assigned unambiguously to one protein were retained, and those assigned to multiple proteins were discarded. The protein expression level of each gene was calculated using the median copy number of all peptides that assigned uniquely to any of the isoforms the corresponding gene. Finally, the genes with more than 5 peptides in human and chimpanzee prefrontal cortex were considered as expressed.

Analysis of miRNA regulatory effect

The regulatory effect of miRNA was estimated on both mRNA and protein levels. The miRNA regulatory effect is determined by whether differently expressed miRNAs can cause significant expression changes on their target genes. For the miRNA differently expressed between humans and chimpanzees, the targets of miRNA highly expressed in humans were expected to be down-regulated in humans. The miRNA target genes were predicted using the TargetScan5 algorithm [195], which has good sensitivity and specificity [129]. In general, the TargetScan5 algorithm predicted the miRNA target gene based on the presence of conserved miRNA binding sites in mRNA 3' UTR regions.

miRNA regulatory effect detection on the mRNA level

Based on predicted targets, to test miRNA regulatory effects on the mRNA level, the gene expression between species was first normalized using quantile normalization and the mean expression difference (d_m) of human and chimpanzee gene expression were further used to represent the mRNA expression difference between species.

$$d_m = m_h - m_c$$

where m_h and m_c represent the mean expression of human and chimpanzee, respectively. The mean expression for human was calculated as:

$$m_h = \frac{1}{n}\sum_{i=1}^{n} X_{h,i}$$

In this study, only the genes with moderate expression differences between species was used to analyze miRNA regulatory effects. The genes with absolute differences between species that were smaller than 0.5 were excluded from the analysis. Then, the Wilcoxon rank sum test was used to compare the expression difference between the targets of miRNA that were highly expressed in humans with targets of miRNA that were highly expressed in chimpanzees. Before applying the Wilcoxon rank sum test, the genes that were targeted by both miRNA highly expressed in humans and miRNA highly expressed in chimpanzees (i.e., targets with inconsistent miRNA effects) were excluded. The Wilcoxon rank sum test, also known as the Mann-Whitney U test, is a nonparametric test of the null hypothesis that two populations are the same against an alternative hypothesis, especially when a particular population tends to have larger values than the other.

miRNA regulatory effect detection on the protein level

Due to greater intraspecies variation in the protein data and small sample size for both human and chimpanzee data (two samples for each species), when testing the miRNA regulatory effects on protein expression, the method was revised to use the effect size, calculated as d_p, to represent the protein expression difference between species. The t-statistics is not suitable to estimate protein expression difference in this study because of the small sample size for human and chimpanzee data.

In general, the simple definition of effect size is the magnitude, or size, of the difference. In this study, d_p was adopted from Cohen's d [196], which was defined as the difference between two means (m_h and m_c) divided by a pooled standard deviation s for the data, which expressed the mean difference of protein expression between human and chimpanzee in standard deviation units:

$$d_p = \frac{m_h - m_c}{s}$$

where m_h and m_c represent the mean value of protein expression in human and chimpanzee, respectively. The pooled standard deviation s was defined as,

$$s = \sqrt{\frac{(n_h - 1)s_h^2 + (n_c - 1)s_c^2}{n_h + n_c - 2}}$$

where the variance for human was defined as

$$s_h^2 = \frac{\sum_{i=1}^{n_h} (X_{h,i} - \bar{X}_h)^2}{n_h - 1}$$

The proteins that with absolute effect size smaller than 0.5 were excluded from the analysis. The Wilcoxon rank sum test was used to compare the expression difference between the targets of miRNA that were highly expressed in humans with targets of miRNA that were highly expressed in chimpanzees.

miRNA regulatory effect detection robustness analysis

To check the robustness of the detected target effect at both mRNA and protein levels, I investiaged whether the significance level of the miRNA regulatory effect was influenced by the expression level of differential expressed miRNA. Specifically, various miRNA expression level cutoffs were used to identify differentially expressed miRNA and the whole miRNA regulatory effect detection procedure was repeated.

To further test the robustness of the detected target effect, I checked whether the significance level of the target effect depended on the choice of the target gene expression divergence cutoff. To test this, the median of the gene expression difference of targets was analyzed at various gene expression divergence cutoffs.

2.4 miRNA 5'-isoform identification, quantification and notation

For a given miRNA *i*, to identify and quantify its 5'-isoforms, first, all sequences mapping within six nucleotides upstream or downstream of the annotated 5'-position of the mature miRNAs were retained. Then sequences from all genomic loci producing the same mature miRNA were united. After applied the new miRNA quantification procedure described in 2.2, all sequences that were not used for canonical miRNA quantification were grouped based on their 5'-position loci identity into *m* sets $S=(S_1, S_2, \ldots, S_m)$. Within set S_j, the highest expressed sequence was designated as the reference sequence to represent 5'-isoform *j*. The expression level of 5'-isoform *j* from miRNA *i* was calculated as a sum of the copy number of all sequences in set S_j.

$$e_j^{(i)} = \sum_{k \in s_j} x_k$$

where $e_j^{(i)}$ denotes the expression level of 5'-isoform *j* from miRNA *i* and x_k denotes the expression of sequence *k* from set S_j .

To label the identified miRNA 5'-isoforms, the combination of the canonical miRNA name annotated in miRBase as well as the 5' end shift direction and offset number were used a notation to represent each miRNA 5'-isoform. Specifically, the negative sign (-) was used to indicate that the 5'end of 5'-isoform was shifted to the left (or upstream) compared to the annotated miRNA, On the other hand, the positive sign (+) was used to represent that the 5'end of 5'-isoform was shifted to the right (or downstream) compared to the annotated miRNA. A number followed by a positive sign or a negative sign indicated the number of nucleotides had shifted away from the annotated 5'end of canonical miRNA. For example, the notation "miR-124|-1" denoted a 5'-isoform with one nucleotide shifted to the left of annotated miR-124 5'end.

2.5 Analysis of miRNA 5'-isoform functionality prediciton

miRNA 5'-isoform functionality prediction methods

Since the miRNA 5'-isoforms identified in human PFC were conserved between human and mouse, I developed a comparative approach to estimate their functionality based on the observation of co-evolution between conserved miRNAs and their target sites (see Introduction: The miRNA seed region). In this study, the term functionality is defined in the context of evolution by measuring whether conserved 5'-isoforms can cause sequence constraint on their target sites, which means the functionality of a conserved miRNA 5'-isoform is estimated based on the its target site conservation status. If the target sites of a conserved miRNA 5'-isoform display significantly excessive conservation between species than expertation, this conserved miRNA 5'-isoform is considered as functional. The terms "seed match", "miRNA target site" and "heptamer" are equivalent in this study.

In general, the procedure of functional heptamer prediction is comprised of three steps. The first is estimating the observed heptamer conservation based on human-mouse 3'UTR alignment by enumerating both conserved and total heptamer occurrence. The second step, which is the most crucial, is to obtain the expected conservation or background conservation for each heptamer based on the control sets that obtained using sequence-shuffling based methods. In the third step, by combining the result of the first two steps, the cutoff 0.05 representing Benjamini-Hochberg (BH) corrected p-value of the binomial test was used to determine whether one heptamer had excessive conservation. In the following section, these three steps are depicted in detail.

Step 1: Estimation of the observed heptamer conservation

The observed heptamer conservation in human 3'UTR region is measured based on the 3'UTR alignment between human and mouse. The calculation is comprised of the following three steps.

1) Obtaining representative 3'UTR sequence dataset in human

   The genomic coordinates of human 3'UTR sequences were based on refseq protein-coding gene annotation and downloaded from the UCSC Genome Brower (hg18) [184]. For the gene with more than one 3' UTR annotation, the longest 3'UTR was retained to represent 3'UTR region of the corresponding gene. The 3'UTR sequences

with lengths shorter than 20nt were removed.

2) Building 3'UTR alignment of human and mouse

The 3'UTR alignment of human and mouse was extracted from human-mouse whole genome alignment based on the representative 3'UTR sequence annotation obtained in step 1. Human-mouse whole genome alignment (hg18&mm9) constructed using blastz was downloaded from the UCSC Genome Brower [184].

3) Enumerating conserved and total occurrence for heptamers.

Based on the extracted 3'UTR alignment, a sliding window of 7nt (heptamer) was scanned along with the alignments with 1nt stepwise. A conserved occurrence for a heptamer was defined as a window that has an identical sequence in both human and mouse. The total occurrence for a heptamer was counted in human 3'UTR. For a given heptamer *j*, the conserved occurrence c_j and total occurrence t_j were enumerated.

Step 2: Estimation of the expected heptamer conservation

A statistically and biologically meaningful result of any sequence motif analysis largely depends on choosing an appropriate control set. In this study, the control set for estimating the expected conservation for a given heptamer was constructed using sequence-shuffling methods. In total, five shuffling procedures were used, including one procedure based on seed match sequence shuffling and four shuffling procedures based on 3'UTR alignment shuffling. Descriptions of the detailed shuffling procedures were described as follows.

Seed Match Shuffling Procedure (SSP)

The procedure for seed match (heptamer) shuffling was adopted from [95]. In general, for a given heptamer, its expected conservation was estimated through the average conservation of a set of heptamer controls sharing the same mononucleotide frequency and similar occurrence in the human 3'UTR.

First, the occurrence of all possible heptamers (16,384) in human 3' UTR was enumerated. Then, for a given heptamer *j*, its heptamer controls with the same mononucleotide frequency were selected. Next, based on their occurrences in the 3'UTR, the heptamer controls were further required to be within ± x% of the occurrence of heptamer *j* where x belongs to one integer from 1 to 15. Whenever possible, the lowest x was chosen with a corresponding set that contained at least ten heptamer controls. If no x met this criterion, the lowest x was chosen that included at least three heptamer controls. Finally, the expected conservation of a given heptamer *j* was calculated as the average conservation level of its heptamer controls.

$$p_j = \frac{1}{n}\sum_{i=1}^{n}\frac{c_{ji}}{t_{ji}}$$

where p_j denotes the expected conservation of a given heptamer j, c_{ji} and t_{ji} denotes the conserved and total occurrences of control i of heptamer j.

Another method for establishing control sets is shuffling 3'UTR alignment. This strategy has not been systemically investigated for estimating miRNA seed match conservation. In general, for a given heptamer, its expected conservation was estimated through the average heptamer conservation calculated based on 1,000 control sets of shuffled 3'UTR alignment. In this study, the alignment is a two-way alignment between human and mouse 3'UTR sequences. For a given position in the 3'UTR alignment, there are four possible alignment patterns between Seq_{human} and Seq_{mouse} (match, mismatch, gap in Seq_{human} and gap in Seq_{mouse}). Considering these four alignment patterns as well as the sequence nucleotide position, five alignment features were derived: global conservation, local conservation, gap pattern, mononucleotide frequency and dinucleotide frequency. According to the combination of these five alignment features, four shuffling procedures were developed to establish the 3'UTR alignment controls used for estimating the expected heptamer conservation.

3'UTR Alignment Shuffling Procedure 1 (USP1)

USP1 shuffled 3'UTR alignment by randomly exchanging the columns from the original 3'UTR alignments. Compared to the original 3'UTR alignment, this shuffling method will produce control 3'UTR alignment with the same global conservation and the same mononucleotide frequency (Figure 2.1B). However, the gap pattern, local conservation pattern and dinucleotide frequency were disrupted.

3'UTR Alignment Shuffling Procedure 2 (USP2)

USP2 is similar to USP1 but with an additional criteria. The control 3'UTR alignment was required to maintain a gap pattern while keeping the same mononucleotide frequency and the same global conservation. Maintaining the gap pattern feature is achieved by separating and memorizing the position of gaps in Seq_{human} and gap in Seq_{mouse} into two groups and further shuffling the columns within each group (Figure 2.1C).

Figure 2.1: Example of shuffled 3'UTR alignment of IMMP1L gene using four 3'UTR shuffling procedures. (A) The original 3'UTR alignment for IMMP1L gene. (B), (C), (D), (E) The shuffled 3'UTR alignment of IMMPlL with USP1, USP2, USP3 and USP4, respectively.

3'UTR Alignment Shuffling Procedure 3 (USP3)

By further incorporating the feature of local conservation, USP3 was established on the basis of USP2. Similar to keeping the gap pattern, the local conservation pattern is maintained by separating and memorizing the positions of matches and mismatches into two groups and further shuffling the columns within each group. Following this procedure, the control 3'UTR alignment retrained the same mononucleotide frequency, the same global conservation, the same gap pattern and local conservation (Figure 2.1D).

3'UTR alignment Shuffling Procedure 4 (USP4)

While the control 3'UTR alignment produced by USP1, 2 and 3 maintained mononucleotide frequency; USP4 was devised to incorporate another feature that can maintain dinucleotide frequency. This is achieved by randomly exchanging the triplets satisfying the following criteria: a) triplets with identical bases at position one and position three; and b) position two of triplets also have identical sequence match pattern (match and mismatch). Figure 2.1E shows an example of shuffled 3'UTR alignment using USP4.

Following USP1, 2, 3 and 4, separately, 1,000 3'UTR alignment control sets were generated. For a given heptamer j, its expected conservation rate p_j was then calculated by averaging the conservation of the corresponding heptamer calculated based on 1,000 3'UTR alignments control sets.

$$p_j = \frac{1}{1000}\sum_{i=1}^{1000}\frac{c_{ji}}{t_{ji}}$$

where c_{ji} and t_{ji} denote the conserved and total occurrences of heptamer j based on shuffled 3'UTR alignment set i.

Step 3: Identification of heptamer with excessive conservation using binomial test

The final step is to determine whether one heptamer has excessive conservation by comparing the observed heptamer conservation obtained in *Step 1* with the expected heptamer conservation obtained in *Step 2*. For a given heptamer j, its conserved occurrence was denoted as c_j, its total occurrence was denoted as t_j and its expected conservation was denoted as p_j. Based on binomial distribution, the probability of observing conserved occurrence occurring more than c_j times can be calculated as the following:

$$p = \sum_{k=c_j}^{t_j} \frac{t_j!}{c_j(t_j - c_j)} p_j (1 - p_j)^{t_j - c_j}$$

Accessing the functionality of a group of seed matches involved the repeated performance of the binomial test, which led to multiple hypothesis testing. The latter caused accumulation of Type I errors of individual tests, which led to an overall higher chance of falsely rejecting at least one tested null hypothesis and therefore increased the chance of false positive discoveries. To reduce the false positives caused by multiple hypotheses testing, the false discovery rate (FDR) was calculated with the Benjamini-Hochberg (BH) procedure [197] to correct for multiple comparisons. FDR controls the expected proportion of incorrectly rejected null hypotheses in a list of rejected hypotheses. It is a less conservative multiple testing correction procedure with greater power than the Bonferroni correction procedure [198], which controls the familywise error rate (FWER). Suppose there are $H_1 \ldots H_m$ null hypotheses and $P_1 \ldots P_m$ corresponding p-values. To control FDRs under the level of q, BH procedure first orders these p-values in increasing order as $P_{(1)} \ldots P_{(m)}$ and further finds the largest index $k \in m$ such that:

$$p_{(k)} \leq \frac{k}{m} q$$

Subsequently, all hypotheses with p-values smaller or equal to $P_{(k)}$ are rejected.

In this study, heptamer with BH corrected p-value (FDR) less than 0.05 were predicted as functional.

miRNA 5'-isoform functionality prediction performance evaluation

To evaluate prediction performance and find the best prediction procedure, conserved miRNA families and nonconserved miRNA families of human were used as a positive set and a negative set to estimate the sensitivity and specificity of the five prediction procedures, respectively.

The reasons for choosing conserved miRNA families and nonconserved miRNA families as training sets are as follows. For functional conserved miRNA families such as those conserved between human and mouse, their target sites (heptamers) were expected to show excessive conservation on the 3'UTR between human and mouse based on the observation of co-evolution (See Introduction: The miRNA seed region). On the contrary, for nonconserved miRNAs families with seed regions that were not shared beyond primates or that only exist in human lineage, one would not expect them to play any role in maintaining the target sites in mouse since no such miRNAs were expressed in mouse. Therefore, the seed matches of nonconserved miRNA families were not expected to show any excessive conservation on the 3'UTR. It should be noted that this framework of performance estimation only focused on predicting the functionality of conservation miRNA families. Using nonconserved miRNA

families as a negative set to estimate specificity does not mean nonconserved miRNA families are not functional.

To obtain conserved and nonconserved miRNA families of human, all annotated miRNAs from five vertebrate species (human, mouse, rat, dog and chicken) were extracted from miRBase (version 12) [97]. miRNAs were further grouped into miRNA families based on the seed sequence identity. The conserved miRNA families were required to be shared between human and mouse and also shared in at least one of the rest of the 3 vertebrate species. The nonconserved miRNA families were defined as human miRNA families that do not exist in the other four vertebrate species. Based on this definition, 162 human conserved and 284 nonconserved miRNA families were obtained, which corresponds to 262 and 326 human miRNAs, respectively. The conserved miRNA and nonconserved miRNA families were represented as positive (*P*) and negative sets (*N*) to evaluate the prediction performance. The true positives (*TP*) are the conserved miRNA families that predicted as functional. The false positives (*FN*) are the conserved miRNA families that were not predicted as functional. The true negatives (*TN*) are the nonconserved miRNA families that predicted as nonfunctional. The false negatives (*FP*) are the nonconserved miRNA families that predicted as functional. Performance was evaluated based on sensitivity (*SN*), specificity (*SP*), accuracy (*ACC*), positive prediction value (*PPV*) and Matthews correlation coefficient (*MCC*) [199]. *MCC* is generally regarded as one balanced measure for evaluating the quality of binary classifications since it takes *TP*, *FN*, *TN* and *FP* into account. It can be used even if the classes are of very different sizes. *MCC* is in essence a correlation coefficient between the observed and predicted binary classifications; it returns a value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and −1 indicates total disagreement between prediction and observation. In this study, *MCC* value was used to determine the prediction performance since the positive and negative sets are unbalanced.

$$SN = \frac{TP}{TP + FN}$$

$$SP = \frac{TN}{TN + FP}$$

$$ACC = \frac{TP + TN}{TP + FN + TN + FP}$$

$$PPV = \frac{TP}{TP + FP}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The method with the best performance was applied to all 16,384 (4^7) heptamers to obtain the proportion of predicted functional heptamer out of 16,384 heptamers.

Since the majority of sequenced reads from known ncRNAs were from degradation fragments that were believed to be nonfunctional, I further compared reads from ncRNA fragments as another set of negative control with known conserved miRNAs and 5'-isoforms with respect to 1) the relationship of expression level and functionality; and 2) heptamer conservation strand bias on 3'UTR. Mapped reads were classified as ncRNA-derived if at least one nucleotide of the sequence fell into a known ncRNA genome annotation region. The genomic annotation of human ncRNA (excluding miRNA and piRNA) was downloaded from UCSC, Ensembl and Refseq. The ncRNA annotations from different sources were further merged if at least one nucleotide was overlapped and further combined into a uniform ncRNA annotation based on the hierarchical order of UCSC, Refseq and Ensembl. The resulting ncRNA annotation mainly included rRNA, tRNA, snoRNA, snRNA, scRNA and misc-RNA classes.

2.6 Analysis of miRNA 5'-isoform functionality verification

The public mouse gene microarray data from miR-124 pri-miRNA overexpression and miR-223 knockout experiments were analzyzed to verify the predicted functional 5'-isoform candidates. The preprocessed miR-124 pri-miRNA overexpression gene microarray data were downloaded from [200]. The preprocessed miR-223 knockout gene microarray data were downloaded from [201]. These two datasets were analyzed separately. For the miR-124 pri-miRNA overexpression dataset, to obtain exclusively conserved targets (ECTs) of miR-124 and its two 5'-isoforms, I first used TargetScan5 to predict conserved target genes (conserved between human and mouse) and further selected those that are exclusively targeted by miR-124 and its two 5'-isoforms (miR-124|+1 and miR-124|-1, respectively). To further remove the potential influence of miR-124 stringently, we excluded the target genes of two 5'-isoforms that were targeted by a weaker 6mer seed match of miR-124 (2-7nt) without considering conservation, although such 6mer seed matches have very weak regulatory effects on targets as was shown before [44]. The nontargets were defined as the genes without conserved target sites of miR-124 as well as two 5'-isoforms, predicted by using TargetScan5. To estimate whether miR-124|+1 and miR-124|-1 are functional, I compared their expression down-regulation magnitude of ECTS to the background nontargets' down-regulation status, before and after miR-124 precursor overexpression using the Kolmogorov-Smirnov test (KS test) and Wilcoxon rank sum test. Significant down-regulation (KS test $p < 0.05$ and Wilcoxon rank sum test $p<0.05$) suggests that 5'-isoforms have a regulatory effect on their target genes. To estimate the specificity of the regulatory effect of miR-124|+1 and miR-124|-1, I applied the same method to further analyze the miR-124 duplex overexpression experiments from human Hela cell line downloaded from [129]. The difference between the

miR-124 precursor overexpression experiment and miR-124 duplex overexpression experiment is that the former experiment can generate both miR-124 and its two 5'-isoforms, whereas the latter only produces the miR-124. The same method was used to analyze the miR-223 knockout dataset. Significant up-regulation (KS test $p < 0.05$ and Wilcoxon rank sum test $p<0.05$) suggests that miR-223|+1 was functional since the miR-223 knockout should in principle eliminate the regulation of miR-223 and its 5'-isoform from their corresponding targets.

2.7 Analysis of transcriptome reconstruction

De novo transcriptome assembly

The quality of raw sequencing reads was first assessed using the FASTX tool kit (http://hannonlab.cshl.edu/fastx_toolkit/index.html). After removing low quality reads (phred score < 20), raw reads from 14 human prefrontal cortex samples were combined, resulting in a total of 284 million 100 nt strand-specific reads. These reads were used as the input data for de novo assembly using Trinity [153].

Trinity is a highly efficient tool for de novo transcriptome assembly that conducts assembly in three successive steps with three modules: Inchworm, Chrysalis and Butterfly. In the first step, Inchworm assembles reads into contigs using a greedy k-mer based approach. First, a k-mer dictionary is built from all sequence reads and error-containing k-mers are eliminated. Then Inchworm ranks k-mers in decreasing order of abundance and selects k-mer based on abundance ranking to seed a contig assembly. Inchworm further assembles contig using a greedy extension based on (k-1)-mer in each direction and concatenating its terminal base to the growing contig sequence. Inchworm usually produces one full-length dominant isoform per locus, generating just the unique portions of other alternatively spliced isoforms. In the second step, Chrysalis clusters related contigs produced by Inchworm into sets of connected components based on shared reads support and further builds de Bruijn graphs for each component and partitions reads among the components, which allows of processing the downstream computations in a massively parallel manner. In the last step, Butterfly reports alternative spliced isoforms and paralogous genes. Using the original RNA-seq reads, Butterfly reconciles individual de Bruijn graphs produced by Chrysalis in parallel and constructs distinct transcripts for splicing isoforms and teasing apart transcripts that correspond to paralogous genes.

Trinity (version r2011-11-26) was downloaded from the Trinity homepage. The assembly parameters were as follows: (--seqType fq --single --CPU 80 --min_contig_length 150 --SS_lib_type F –bflyHeapSpace 260G). The parameters --seqType fq and --single specified that the sequencing data was single end in a fastq format. The parameter --min_contig_length was set to 150 to specify that the minimum length of reported assembled contigs was no shorter than 150nt. The parameter --SS_lib_type was set to F to specify that the sequencing library type was strand-specific and reads were from sense orientation. The final assembly transcripts were further required to have a length no shorter than 300nt.

Assembly transcript contigs mapping

The transcript contigs produced by Trinity were mapped to the human genome (hg19) using GMAP (Genomic Mapping and Alignment Program, version 2011-10-07) [202] with further alignment identity and coverage filtering. The transcript contigs were mapped using GMAP with default parameters except parameter (-A --microexon-spliceprob 0.95 -f 1). The parameter --microexon-spliceprob 0.95 allowed reporting of microexons only if one of the splice site probabilities was greater than 0.95.

GMAP is a standalone program for mapping and aligning cDNA/EST sequences to a genome. Several advantages made it the most suitable for handling the many transcript contigs generated by de novo assembly: 1) It provides fast high-throughput batch processing for large sequence sets using memory mapping and multithreading strategy [202]; 2) it generates accurate gene structures, even in the presence of substantial sequence errors and polymorphisms [202]; and 3) it locates splice sites accurately without the use of probabilistic splice site models, allowing generalized use of cross-species alignment [202].

Unambiguously and uniquely aligned transcript contigs were further required to meet the minimal identity cutoff >0.95 and the coverage cutoff >0.95. Contig clusters were obtained by merging overlapping mapped contigs by at least one overlapping nucleotide. The "known" and "novel" contig clusters classification was based on Ensembl gene annotation (version 64) [203]: Assembled contig clusters that overlapped with at least one annotated transcript by at least one nucleotide were classified as "known," while the remaining contig clusters were classified as "novel"

To identify novel contig clusters that were missing because of the incompleteness of the current human genome (hg19), I first collected contig clusters that could not be mapped to the human genome after mapping with a relaxed mapping cutoff (mapping minimal identity >0.8, coverage >0.5) and further mapped them to four nonhuman genomes (chimpanzee, orangutan, rhesus macaque, mouse and rat) using GMAP with an additional parameter (--cross-species). The parameter (--cross-species) allows mapping contig clusters across species with high sensitivity [202]. Candidate contigs that could be aligned to at least one nonhuman genome were further required to meet the minimal identity cutoff >0.8 and the coverage cutoff >0.8. Putative protein-coding genes and exons were obtained by overlapping the aligned transcript contigs with known annotations from the four nonhuman genomes.

2.8 Identificaiton of novel elements from transcriptome assembly

Identification of novel elements from annotated transcripts

Novel transcribed elements of annotated genes, including novel internal exons, novel splicing donor and acceptor splicing sites and novel 5'UTR and 3'UTR extensions, were identified based on the assembled contig clusters overlapping with at least one transcript, as annotated by the Ensembl database (version 64) [203]. Novel internal exons were defined based on the assembled contig clusters sharing at least one exon of annotated transcripts and were further required to fully reside within the intron region of this annotated transcript. Novel donor and acceptor splice sites were required to share one boundary with an internal exon of an annotated transcript and to contain the canonical donor/acceptor splicing sequence (GT-AG)

at the novel splice boundary. Novel 5'UTR and 3'UTR extensions were required to share at least one exon with annotated transcripts, and each extended region was at least 100nt long.

Identificaiton of novel lncRNAs

Novel lncRNAs were identified as the transcripts without overlapps with known annotations and display low protein-coding potencials. The coding potential of novel transcript contigs was estimated using the two sequence coding potential estimation algorithms, CPC (Coding Potential Calculator) [142] and CPAT (Coding Potential Assessment Tool) [143]. For CPC, novel transcripts with a CPC score of less than 0 were classified as noncoding transcripts. The CPC score cutoff was adopted from the CPC website. For CPAT, the score cutoff 0.364 was used to distinguish coding and noncoding transcripts as recommended by the CPAT website (CPAT score < 0.364 indicates noncoding sequence).

CPC is a protein-homology-based coding potential estimation algorithm that accesses transcript coding potential based on six biologically meaningful features using support vector machine (SVM). The first three features are related to the quality of the open reading frame (ORF) in a transcript, including ORF coverage, ORF log-odds score and the integrity of ORF. The rest three features are derived from the similarity to known protein-coding genes through BLAST search against a common protein dataset, e.g. e.g., UniProtKB/Swiss-Prot [204]. The features include the number of hits to the known protein-coding gene (E-value cutoff 10^{-10}) and the average and variance of E-values of HSPs of three frames.

CPAT is an alignment-free coding potential estimation algorithm that achieves high sensitivity and specificity by using a logistic regression model built with four sequence features: open reading frame (ORF) size, ORF coverage, Fickett TESTCODE statistic and hexamer usage bias. ORF size is the maximum length of all predicted ORFs of a given transcript. ORF coverage is defined as the ratio of ORF size to transcript length. Fickett TESTCODE statistic [205] is a sequence feature combining effects of nucleotide composition and codon usage bias, which is calculated based on four position values and four composition values (nucleotide content) from the DNA sequence. Hexamer score is a log-likelihood ratio score that determines the relative degree of hexamer usage bias in a given transcript. A positive log-likelihood ratio score indicates a coding sequence, whereas a negative score indicates a noncoding sequence.

Since CPC and CPAT utilized different features and strategies to access transcript coding potential, the intersection of the predicted noncoding transcripts of these two methods was used to distinguish novel long noncoding RNAs (lncRNAs) from total novel transcripts from human prefrontal cortex.

Several advantages made CPC and CPAT suitable for predicting the coding potential of assembled novel transcripts. First, both CPC and CPAT have discriminatory power for both conserved and nonconserved transcripts. This feature is crucial since the majority of annotated human lncRNAs and novel transcripts identified in this study are poorly conserved. The multiple-alignment-based coding potential estimation algorithm, such as PhyloCSF [143], is not suitable for analyzing nonconserved transcripts. Second, both algorithms produce highly accurate results. The prediction accuracy of CPC is 95.77% through 10-fold cross-validation in training sets [142]. CPAT achieves sensitivity 0.96 and specificity 0.97 for

a human testing dataset [143].

2.9 Analyis of sequence, expression and genomic context of novel lncRNAs

Sequence and expression property analysis of novel lncRNAs

The sequence and expression properties of novel lncRNAs were analyzed in terms of expression abundance, sequence conservation, transcript splicing site signal, tissue expression specificity, nuclear and cytoplasmic localization preference and temporal expression pattern during human PFC development.

Expression abundance estimation

The expression abundances of lncRNAs were calculated based on uniquely mapped reads and summarized in RPKM (Reads Per Kilobase of exon model per Million mapped reads) [206]. The uniquely mapped reads were obtained through the reads mapping results from TopHat [207] with default parameter and further screened using SAMtools [208]. For each gene, the mean RPKM value of 14 human PFC samples was used to represent its expression abundance. RPKM is the most commonly used gene expression abundance estimation method that normalizes gene expression with gene length and sample sequencing depth, thus facilitating comparison of transcript levels both within and between samples. The expression abundance of gene *i* in sample *j* was estimated in RPKM as:

$$RPKM(i) = \frac{10^9 \cdot C_i}{N_j \cdot L_i}$$

where C_i is the number of uniquely mapped reads that fell into the exons of gene *i*; N_j is the total number of uniquely mapped reads in sample *j*; and L_i is the sum of exons in base pairs of gene *i*.

Exon sequence conservation

Exon conservation was estimated using phastCons score based on 17 vertebrate species' genomes data (phastCons17way) downloaded from UCSC [184].

The PhastCons score is a posterior probability that each nucleotide belongs to a conserved element, which is calculated by the PhastCons program [209]. PhastCons program is a phylogenetic hidden Markov model (phylo-HMM) based method that estimates the probability that each nucleotide belongs to a conserved element, based on the multiple alignment and phylogenetic models for conserved and nonconserved regions. PhastCons scores range from 0 to 1. A high PhastCons score means the site has a higher probability of being conserved.

For each exon, the average of all nucleotides' phastCons scores were used to represent its conservation (require more than 80% of exon's nucleotides to have a valid phastCons score). The same sequence conservation calculation procedure was used for the other five genomic

sequence categories: randomly selected intergenic region, annotated lncRNAs, pseudogenes, and CDS and UTR regions of protein-coding genes.

Transcript splicing site signal

The presence of canonical donor and acceptor site splice signals within novel lncRNAs was identified using the most canonical splicing signal GT-AG motifs. The nucleotide composition at splice sites (positions 11-12) and surrounding region (10nt upstream and downstream) was measured using bits of entropy and illustrated using sequence logo. The same calculation procedure was used for protein-coding genes.

Tissue expression specificity

Tissue specificity was estimated using RNA-seq data from the Human Body map [15]. To increase tissue coverage, two deep sequencing datasets with comparable sequencing coverage (fetal brain and fetal liver) [210] were combined with Human Body map data, resulting in sequencing data from a total of 19 human tissues. The tissue expression specificity of all novel lncRNAs, with a mean expression >0.1 RPKM across tested tissues, was measured using Shannon entropy [211]. For a given novel transcript contig i, its expression in tissue j was denoted as e_{ij}. The Shannon entropy h_i across 19 tissues was calculated as

$$h_i = -\sum_j p_{ij} \log_2(p_{ij})$$

where p_{ij} is measured by

$$p_{ij} = \frac{e_{ij}}{\sum_j e_{ij}}$$

The same tissue expression specificity calculation procedure was used for two other genome sequence categories: annotated protein-coding genes and lncRNAs.

The nuclear and cytoplasmic localization preference

The nuclear and cytoplasmic localization preference of novel lncRNAs was estimated using RNA-Seq data from SK-N-SH cells (GSE30567) from ENCODE/Cold Spring Harbor Labs. SK-N-SH is a human neuroblastoma cell line commonly used as in vitro models of neuronal function and differentiation. The RNA-seq data from nuclear and cytoplasmic fractions of SK-N-SH were mapped to the human genome (hg19) using TopHat [207], with default parameter. The uniquely mapped reads were further screened using Samtools. The expression levels of novel transcript and annotated protein-coding genes as well as annotated lncRNAs were measured in RPKM with uniquely mapped reads. For a given novel transcript, its localization preference was estimated as a ratio between cytoplasmic expression abundance and nuclear expression abundance. The same localization preference calculation procedure

was used for another two genome sequence categories: annotated protein-coding genes and lncRNAs.

Temporal expression patterns

To analyze temporal expression patterns of novel lncRNAs in human PFC development and aging periods, novel lncRNAs were quantified and estimated into RPKM separately in each of 14 human PFC samples with different ages. A polynomial regression-based age test developed in [212] was used to identify novel lncRNAs with age-related expression patterns. For each gene, an age test determined the effect of age on the expression by selecting the best polynomial regression using age (in log2 days scale) as predictor and expression level as response. The best regression model was chosen from all possible linear-to-cubic models by using F-test and the adjusted r^2 criterion [213]. Specifically, age test first fits a third-degree polynomial regression model and all six subregression models with age for gene *i* as:

$$
\begin{aligned}
y_{ij} &= \beta_{0i} + \beta_{1i} a_j + \beta_{2i} a_j^{\,2} + \beta_{3i} a_j^{\,3} + \varepsilon_{ij} \\
y_{ij} &= \beta_{0i} + \beta_{1i} a_j + \varepsilon_{ij} \\
y_{ij} &= \beta_{0i} + \beta_{2i} a_j^{\,2} + \varepsilon_{ij} \\
y_{ij} &= \beta_{0i} + \beta_{3i} a_j^{\,3} + \varepsilon_{ij} \\
y_{ij} &= \beta_{0i} + \beta_{1i} a_j + \beta_{2i} a_j^{\,2} + \varepsilon_{ij} \\
y_{ij} &= \beta_{0i} + \beta_{1i} a_j + \beta_{3i} a_j^{\,3} + \varepsilon_{ij} \\
y_{ij} &= \beta_{0i} + \beta_{2i} a_j^{\,2} + \beta_{3i} a_j^{\,3} + \varepsilon_{ij}
\end{aligned}
$$

where y_{ij} represents the expression level of gene *i* in sample *j*, a_j denotes the age of sample *j* and ε_{ij} denotes the error term.

Next, to find the best regression model, the age test compared all the above seven models to the null model:

$$y_{ij} = \beta_{0i} + \varepsilon_{ij}$$

by using F-test and further selecting the best model that has the highest adjusted r^2. The adjusted r^2 value represents the amount of variance that can be explained by the specific model [213]. Since the number of parameters of the corresponding model penalizes adjusted r^2, an overfitting problem can be largely avoided. One gene is considered age-related if the model with the highest r^2 is significant in F-test at the predetermined FDR cutoff.

In this study, age-related novel lncRNAs were identified using age test at $p<0.01$ under FDR 2%. The p-value cutoff and corresponding FDR were calculated based on 1000 permutations.

Specifically, I randomly made the age assignments across 14 samples 1000 times and repeated the age test for all genes to obtain F-test p-values. At each p-value cutoff, I calculated the number of novel lncRNAs with p-value below the cutoff in 1000 permutations and used the median value as the false positives. The FDR at each p-value was calculated as the ratio between estimated false positives and the original number of age-related genes. The same age test and FDR estimation procedures were applied to the protein-coding gene with mean expression >0.1 RPKM to obtain age-related protein coding genes.

To classify the expression pattern of age-related novel lncRNAs and protein-coding genes, K-means clustering algorithm [214] was used to group age-related novel lncRNAs and known protein coding genes into 12 clusters. K-means is one widely used unsupervised clustering method. For a given set of observations ($x_1, x_2,...,x_n$) and cluster number K, k-means aims to group n observations into k ($k<=n$) sets $S=(S_1, S_2,...,S_k)$. The main idea of K-means is to find the squared error function J that minimizes the total intra-cluster variance:

$$J = \sum_{j\in[k]} \sum_{i\in S_j} || x_i - c_j ||^2$$

where c_j is the mean of points in S_j and $|| x_i - c_j ||^2$ measures Euclidean distance of data point x_i to the cluster center c_j.

Before conducting K-means clustering, the gene expression values were normalized into z-scores for each novel lncRNAs and protein-coding genes. Transforming the expression value into z-scores allows K-means clustering of transcripts based on expression patterns instead of expression abundance. The z-score (z) value of the expression value x of gene i in sample j was calculated by:

$$z = \frac{x - \mu_i}{\sigma_i}$$

where μ_i and σ_i denote the mean and standard deviation of the expression of gene *i.*

Within each cluster, Fisher's exact test was used to calculate the enrichment of novel lncRNAs and protein-coding genes by using all age-related novel lncRNAs and protein coding genes as background. Fisher's exact test $p<0.05$ after Bonferroni correction was considered significant.

Novel lncRNAs classification based on genomic context

Novel lncRNAs located outside of annotated gene regions but within the 4kb region were classified into four categories based on their location with respect to the nearest annotated gene: upstream-sense (UA-lncRNA), downstream-sense (DS-lncRNA), upstream-antisense

(UA-lncRNA) and downstream-antisense (DA-lncRNA). The 4kb distance cutoff used to identify novel lncRNAs-annotated gene pairs was defined using random transcript pairs distance distribution, calculated by 1,000 permutations of novel lncRNAs loci along each chromosome (for each permutation, keeping the same number of novel lncRNAs on each strand of each chromosome). To check whether novel lncRNAs were significantly correlated with nearby protein-coding genes at the expression level, I used a Wilcoxon rank sum test to compare the observed distributions and each of the 200 simulated distributions of the correlation coefficients to determine how many passed the statistical significance cutoff. Specifically, for each permutation, I randomized the relationship between novel lncRNAs and nearby protein-coding genes and estimated the statistical significance of the correlation distribution difference using the Wilcoxon rank sum test.

2.10 Analysis of divergent transcription and function features of NBiPs

Analysis of promoter divergent transcription feature

The divergent transcription from promoters was estimated by deepCAGE (Cap analysis of gene expression) data from brain tissues downloaded from FANTOM4 [215]. deepCAGE measured short (approximately 27 nucleotide) sequence tags originating from the 5' end of full-length mRNAs based on deep sequencing technology and therefore can be used to obtain transcription start sites (TSSs) genome-wide [215]. To define the divergent transcription features specific to the promoters associated with UA-lncRNAs, unidirectional and known bidirectional expressed annotated genes were used as background for comparison. The criteria to select unidirectional, known bidirectional promoters and novel bidirectional promoters were as follows: For known bidirectional promoters (KBiPs) and novel bidirectional promoters (NBiPs), genes were required to form head-to-head gene pairs within the region of 2kb from TSS. For unidirectional promoters (UniPs), genes were required to have no annotated transcripts, or novel transcript contigs identified in this study, within the 5kb region upstream of their TSS. The promoters defined as showing divergent transcription were required to have at least one CAGE tag on each strand. Unidirectional promoters were required to have at least two CAGE tags at the annotated gene's strand and zero tag at the opposite strand. The promoters containing no CAGE tags were excluded. Divergent transcription feature (Di-trans feature) was calculated as the ratio between the number of promoters with bi- and uni-directional expression detected using deepCAGE data. Fisher's exact test was used to calculate the divergent transcription feature enrichment for the promoters associated with UA-lncRNAs, compared to the unidirectional promoters.

Analysis of function feature of genes associated with NBiPs

The protein-coding genes that showed significant positive correlation with the expression of UA-lncRNAs transcribed from NBiPs were selected for function feature analysis (Pearson correlation $p<0.05$ after Benjamini-Hochberg correction). Functional feature analysis includes GO enrichment analysis, gene enrichment analysis with the genes showing mouse brain cell-type specific expression patterns and promoter H3K4me3 modification enrichment analysis between human PFC neurons and non-neuronal cells.

GO functional enrichment was conducted using a hypergeometric test implemented in the

Genetrail package [216]. Functional term with $p<0.05$ after Benjamini-Hochberg correction was considered significant. Protein-coding genes with mean expression >0.1RPKM in human PFC data were used as background. Enriched GO terms were visualized after term redundancy reduction using REVIGO (Reduce Visualize Gene Ontology) [217]. REVIGO processed and summarized a long list of GO terms into a short list of nonredundant ones based on the GO terms' semantic similarity [ref]. The resulting GO terms can be visualized in semantic similarity-based scatterplots. The same functional enrichment analysis procedure was applied to protein-coding genes associated with US-lncRNAs, DS-lncRNAs and DA-lncRNAs.

The list of mouse genes with known cell-type-specific expression patterns was downloaded from [218]. These genes were derived from three brain cell types: neurons, astrocytes and oligodendrocytes. Human orthologs were determined based on 1:1 orthologs between mouse and human using Biomart from Ensembl [203]. Fisher's exact test (FET) was used to test the enrichment significance for overlapping with three cell-type specific genes, and $p<0.05$ after Bonferroni correction was considered significant.

H3K4me3 modification enrichment analysis between neurons and non-neuronal cells from human PFC was conducted using ChIPDiff [219] based on H3K4me3 data from GSE21172 [220]. ChIPDiff is a Hidden Markov model (HMM) based approach for genome-wide identification of differential histone modification sites (DHMSs) from ChIP-Seq data. To identify DHMSs, the most straightforward solution is to partition the genome into bins and calculate the fold-change of the mapped ChIP-Seq fragments in each bin. However, such a fold-change method is sensitive to technical noise caused by randomly sampling ChIP-Seq fragments. ChIPDiff improved the fold-change method by considering the correlation between consecutive bins modeled in a Hidden Markov model (HMM). The HMM transmission probabilities were trained in a unsupervised manner and followed by the inference of the states of histone modification changes using the trained HMM parameters. The uniquely mapped reads were obtained using Bowtie [221], allowing three mismatches. The reads mapped to the same genomic location were counted only once to avoid PCR-amplification artifact. The DHMSs were predicted using ChIPDiff with default parameter. The regions with more than two-fold higher H3K4me3 modification signals in neurons than in non-neuronal cells were considered regions preferentially expressed in neurons (assigned with a "N" flag). The regions with opposite modification signal patterns were considered regions preferentially expressed in non-neural cells (assigned with a "non-N" flag). Significance was assessed by 1,000 permutations of N and non-N flag labels.

Enriched transcription factor binding site identification in NBiP

Transcription factor binding sites (TFBSs) located within NBiP and KBiP regions were predicted using the MATCH algorithm [222] based on TRANSFAC Release 11. MATCH is a TFBS prediction tool based on predefined Position Weight Matrices (PWMs) representing motifs of transcription factors (TFs). For a given TF, MATCH calculated the similarity scores for the whole PWM and five most informative sites of the PWM and used predefined score cutoffs to predict TFBSs. In this study, to minimize false positive matches, the matrix file vertebrate_non_redundant_minFP.prf was used for TFBS prediction. Enriched TFBS in NBiP

regions were identified by Fisher's exact test, using KBiP regions as background. Significantly enriched TFBS were required to fulfill the following criteria: 1) Benjamini-Hochberg adjusted p-value < 0.05; and 2) Fisher's exact test odds ratio >1.3.

To check for potential association between TFs enriched in NBiP and KBiP and neuronal functions, CoCiter [223] was used to estimate the significance of association between enriched TFs with the terms "neuron" and "neural," respectively. CoCiter used a text-mining based approach to infer the association between a gene set and a term set. To determine whether the enriched TFs are significantly co-cited with the term "neuron," CoCiter first obtained the PubMed abstracts with the co-citation for enriched TFs and the term "neuron" and furthermore used the full text search for each term with the PubMed abstracts. Co-citation impact (*CI*), defined as the log-transformed paper count $CI=\log_2(N+1)$, was used to represent the co-citation level, where N is equal to the number of papers that have co-citation for enriched TFs and the terms "neuron." Assessment of the significance of the co-citation, CI_{random}, is calculated by 1000 permutations that select the same gene size as enriched TFs with the terms "neuron." The permutation p-value is defined by the number of times ($CI_{random} >= CI$) divided by 1000.

2.11 Analysis of the DNA sequence and epigenetic features of NBiP

NBiPs, KBiPs and UniPs were defined in Section 2.10. The putative promoter regions were defined as upstream and downstream 2kb regions surrounding the annotated transcription start site (TSS). Three DNA sequence features (GC content, sequence conservation and regulatory potential) and two epigenetic features (H3K4me3 modification profile and DNA methylation status) were explored.

Specifically, GC content was measured as the G+C percentage of the promoter region. Promoter region conservation was estimated using phastCon scores based on 17 vertebrate species' genome data and using the same approach as for estimating novel contig conservation. Regulatory potential was estimated using the Regulatory Potential (RP) Scores [224] downloaded from UCSC [184]. In brief, RP scores are derived from a log-ratio comparison between transition probabilities of two Markov models that are estimated using training data from alignments of experimentally confirmed regulatory elements and aligned ancestral interspersed repeats. RP scores can efficiently distinguish regulatory regions from neutral sites and therefore can be used to identify putative regulatory sites of the human genome. A higher RP score suggests a higher possibility that the corresponding site is functional. For each promoter, the average RP score was used to represent its regulatory potential (require more than 80% of promoter nucleotides to have a valid RP score). The differences with respect to each of the three DNA features among these three promoter types were tested using the Kolmogorov-Smirnov test.

H3K4me3 modification ChIP-Seq data from one adult human PFC was downloaded from [225]. H3K4me3 modification and input control ChIP-Seq data from rhesus macaque PFC was downloaded from [226]. The ChIP-Seq raw reads were mapped to the corresponding genomes (hg19 for human and rheMac2 for macaque) using Bowtie [221] with default parameter except requiring uniquely mapping by using parameter (-m 0). Before mapping, raw reads were collapsed to avoid PCR amplification artifacts. H3K4me3 modification

density differences between different promoter types were tested using the Wilcoxon rank sum test. For DNA methylation data, the DNA methylation status of the human PFC, measured by MeDIP sequencing (Methylated DNA Immunoprecipitation Sequencing), was downloaded from [225]. The same reads mapping procedure for human H3K4me3 modification ChIP-Seq data was applied to MeDIP sequencing data. The DNA methylation level differences between different promoter types were tested using the Wilcoxon rank sum test. Lower MeDIP reads signals indicate a lower DNA methylation magnitude of the corresponding promoter.

2.12 Analysis of general regulator of NBiP

The RNA-Seq data of PABPN1 knockdown and control experiments were downloaded from SRP015926. TopHat [207] was used to map the RNA-Seq reads onto human genome (hg19), allowing at most three edit distances. Only uniquely mapped reads were retained for cufflinks to quantify the expression of known protein-coding genes, known lncRNAs and novel lncRNAs. The regulatory effect of PABPN1 on gene expression was estimated by gene expression difference between PABPN1 knockdown and control conditions (corresponding to the gene expression fold-changes in log2 scale). The positive values indicate a negative regulation relationship between PABPN1 and corresponding genes, and negative values indicate the reverse regulatory relationship.

3. MicroRNA Expression and Regulation in Human, Chimpanzee and Macaque Brains

The across-species miRNA comparison study is an effective way to elucidate the potential function of miRNAs in the evolutionary context. As described in the Introduction (Section 1.4), the human brain, especially the human prefrontal cortex, which is distinguished with regard to both function and evolution, is a target for investigation of the molecular mechanism underlying the human's unique cognitive function. To build the framework for general across-species miRNA comparison and investigate the roles of miRNA in determining gene expression divergence between species in the prefrontal cortex (PFC), I based my analysis on small RNA sequencing (Illumina) data generated from the prefrontal cortex of humans (age: 14-58 years), chimpanzees (age: 12-40 years) and rhesus macaques (age: 6-15 years) using samples containing RNA pooled from multiple individuals. To assess technical variation of the sequencing measurements, small RNA libraries were prepared and sequenced twice. Furthermore, I analyzed small RNA sequenceing data of cerebellum that was obtained from two human samples, one chimpanzee sample and one rhesus macaque sample, all composed from RNA pooled from multiple individuals (Table 3.1).

3.1 Small RNA sequencing data processing and mapping

To compare miRNA expression abundance based on small RNA sequencing data in human, chimpanzee and macaque brains, the first step was to obtain mapped reads on the cognate genomes. Since known miRNAs are shorter than the raw sequencing reads (usually longer than 36nt for the Illumina platform), the 3' sequencing adaptor must be trimmed before reads mapping. Furthermore, low quality reads should be filtered out to increase mapping accuracy and decrease the amount of memory required for mapping. Therefore, a small RNA sequencing data processing pipeline was developed (Section 2.1) that included reads filtering, 3' adaptor trimming and reads mapping (Figure 3.1). By applying this pipeline to raw sequencing data, on average, ~49% of reads can be mapped to the corresponding genome perfectly (Table 3.1).

Based on the mapped reads, the reads length distributions along with the 5' position nucleotide preference were further analyzed. As expected, miRNAs represented the major portion of small RNA transcriptome in brain tissues of 3 species. As shown in Figure 3.2, the majority of mapped reads (>91%) were within the length range between 20 nt to 24 nt with a clear read length peak at 22 nt, which was consistent with the length distribution feature of known miRNAs [111]. In line with another miRNA sequence feature [114], mapped reads with lengths between 20 nt and 24 nt displayed a remarkable 5' position uridine (U) bias. The reads length distribution and 5' position nucleotide bias were highly consistent across all samples of 3 species, which strongly supported the validity of the developed small RNA reads processing pipeline.

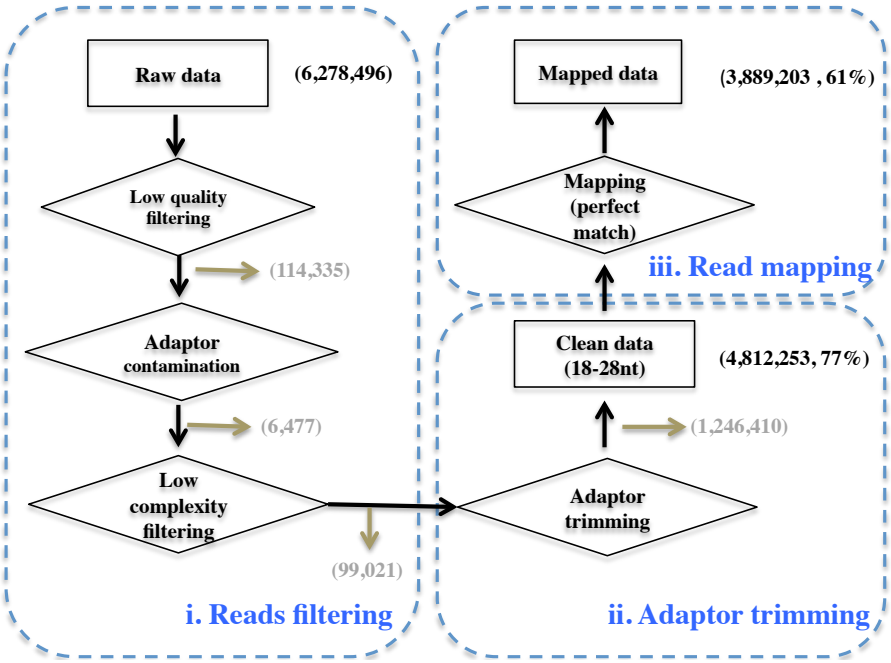


Figure 3.1: Small RNA sequencing data processing pipeline. The workflow illustrates the three consecutive steps of small RNA processing pipeline (reads filtering, adaptor trimming and read mapping) using small RNA data from Hu1 PFC sample as an example. The numbers in black and grey represent the number of reads that were retained and filtered out in each processing step.

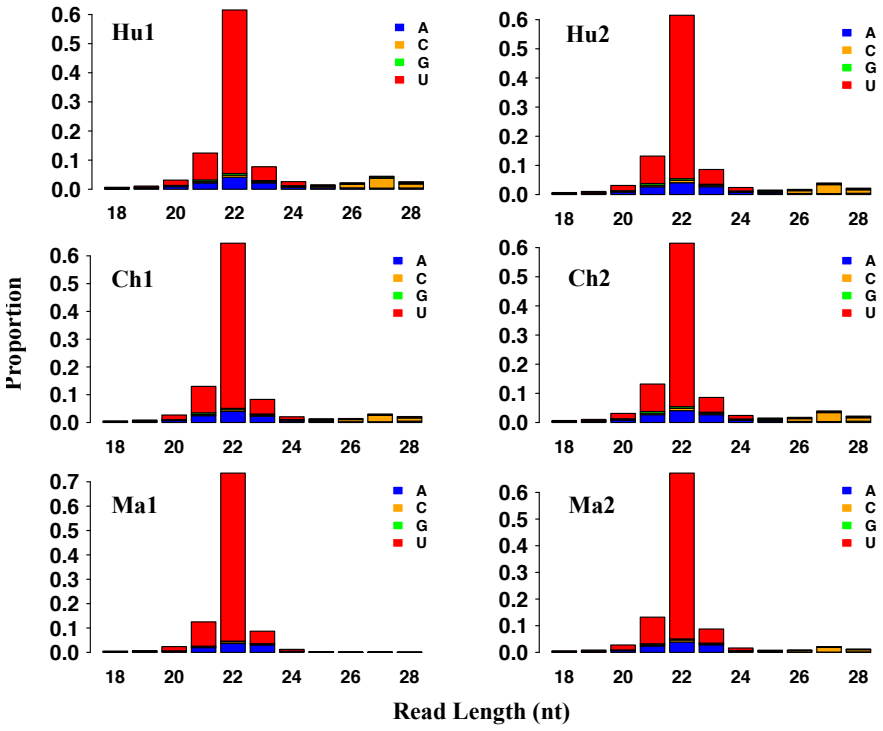


Figure 3.2: The length and 5' position nucleotide distribution of mapped reads in PFC samples. The x-axis represents the length of mapped reads; the y-axis represents the proportion of mapped reads corresponding to each read length (18-28nt). The colors represent reads with 5' position nucleotide: A (blue), C (orange), G (green) and U (red). The labels represent species:

Hu1—Human; Hu2—Human technical replicate; Ch1—Chimpanzee; Ch2—Chimpanzee technical replicate; Ma—Macaque; and Ma2—Macaque technical replicate.

Table 3.1: Sample information and mapping statistics of small RNA sequencing data from human, chimpanzee and macaque brains

Tissue	Sample name	Samples description	Total sequence reads	Total perfectly mapped reads	Mapped percentage
Prefrontal cortex	Hu1	Human	6,486,498	3,889,203	61%
Prefrontal cortex	Ch1	Chimpanzee	7,240,683	3,810,304	53%
Prefrontal cortex	Ma1	Rhesus Macaque	7,241,538	3,868,622	53%
Prefrontal cortex	Hu2	Human technical replicate	7,286,720	2,021,639	28%
Prefrontal cortex	Ch2	Chimpanzee technical replicate	7,828,607	3,124,110	40%
Prefrontal cortex	Ma2	Rhesus Macaque technical replicate	6,517,061	3,560,515	55%
Cerebellum	Hu1	Human	8,241,330	4,112,341	50%
Cerebellum	Hu2	Human biological replicate	9,448,226	4,314,605	46%
Cerebellum	Ch1	Chimpanzee	7,776,308	3,948,506	51%
Cerebellum	Ma1	Rhesus Macaque	8,377,265	4,413,761	53%

3.2 Comparison between mapped reads and annotated miRNAs

To make better miRNA quantification based on mapped reads, as an initial step, the loci of mapped reads were compared with the genomic coordinates of annotated miRNAs in human PFC samples. miRBase is considered the gold-standard miRNA database in which each mature miRNA is annotated as a unique sequence [97]. However, based on the mapped reads

distribution along annotated miRNA precursors, a substantial number of end shifts between mapped reads and annotated miRNA mature sequences were observed. These sequences with end shifts were named “miRNA isoforms.” The sequences with end shifts at the 5’end, 3’end and both ends were called "5’-isoforms”, "3’-isoforms" and "5’&3’-isoforms", respectively. Furthermore, a set of reads from the opposite arm of annotated miRNAs were observed, that in most cases represented reads from novel miRNA* sequences. Take miR-100 as an illustration (Figure 3.3). There are in total 4,086 reads mapped to three nucleotides upstream or downstream of the annotated miR-100 mature sequence. Of these, only 2,178 (53%) reads were exactly matched to the annotated miR-100 mature sequence. The rest of the sequences were mainly derived from 3’ isoforms of miR-100 (1900 reads, 46.5%).

On average, only ~70% of reads from the miRNA precursor region were exactly matched to the annotated mature miRNA sequences in human PFC. Less than 29% of reads were from 3’-isoforms, 5’-isoforms and 5’&3’-isoforms. The remaining ~1% of reads came from novel miRNA* sequences (Figure 3.4A). Most of the reads shifts (~97%) were within 3nt upstream and downstream of annotated mature sequences, and 3’-isoforms showed a broader end shift pattern than 5’- isoforms (Figure 3.4B).

Figure 3.3: The pattern of mapped reads on human miR-100 precursor sequence. The plot shows the pattern of mapped reads on miRNA precursors by taking miR-100 as an example. miR-100 precursor sequence is shown on the top of the plot, with the annotated miR-100 mature sequence labeled in red and the predicted base-pairing secondary structure in dot-bracket notation underneath. The mapped reads can be classified into five types: reads in red represent reads exactly matching the annotated miR-100 mature sequence (annotated); reads in blue, orange and green represent reads that have 3’end shifts (3’isoform), 5’ end shifts (5’isoform) and 5’&3’ ends shifts (5’&3’isoform), respectively; and reads in purple represent reads from novel miRNA* identified in this study (novel miR*). The read count number of mapped reads is shown on the right.

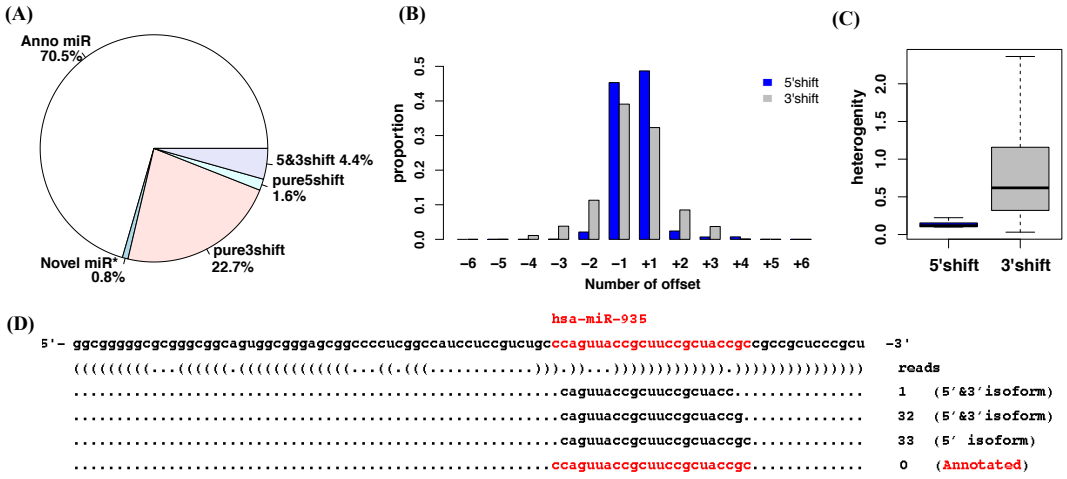


Figure 3.4: The expression abundance and end shift pattern of microRNA isoforms. (A) Proportion of reads from annotated miRNA, miRNA isoforms and novel miRNA* sequences in human PFC. (B) Proportion of reads corresponding to each end offset for reads showing 5' shifts and 3' shifts. (C) miRNA 5'end and 3' end heterogeneity values calculated in Section 2.1. (D) One example of miRNA with potential misannotated mature sequence from miRBase. The annotated miR-935 miRNA mature sequence and corresponding exactly matched reads was labeled in red.

To further quantify the extent of mapped reads end shift magnitude compared with annotated miRNAs, miRNA ends heterogeneity was calculated (Section 2.1). As shown in Figure 3.4C, miRNA 5' end heterogeneity was significantly lower than 3' end heterogeneity (Wilcoxon rank sum test, p<2.2e-16), which was in line with the importance of miRNA seed regions that mostly determined miRNA function [122]. I also found that 160 miRNA mature sequences, representing 28% of expressed miRNAs in human PFC, were probably misannotated since the expression of annotated mature sequences was less abundant than at least one isoform of corresponding miRNAs. One example of potential misannotated miRNA sequence is shown in Figure 3.4D. These results indicated that the quantification method by counting reads exactly matching the annotated miRNA mature sequence should be revised to resolve the substantial reads with end offsets and to solve the miRNA sequence misannotation problem.

3.3 miRNA expression quantification in human brains

Based on these observations, to quantify miRNA expression more precisely, a new miRNA quantification procedure was developed, that can conduct miRNA quantification and mature sequence identification simultaneously (Section 2.1). Take miR-139-3p, for example, to illustrate the quantification process (Figure 3.5A). There were 16 sequences corresponding to 2793 reads mapped within three nucleotides upstream or downstream of the annotated 5'-position of miR-139-3p mature sequence. Of these, the annotated miR-139-3p mature sequence only took 24 reads (0.86%), suggesting the mature sequence of miR-139-3p from miRBase was misannotated. To identify the correct miR-139-3p mature sequence, the new quantification procedure ranked these 16 sequences based on their expression level and furthermore designated the sequence with a maximal copy number as the reference sequence to represent miR-139-3p mature sequence. The newly identified miR-139-3p mature sequence

took 1701 reads, which was 70 times more abundant compared with the previously annotated miR-139-3p mature sequence. To quantify the miR-139-3p expression level, the new quantification procedure used the sum of the copy number of the newly identified miR-139-3p mature sequence and all its 3' isoforms to measure miR-139-3p expression abundance, which covered 98.6% of reads of all 16 sequences.

Based on this new miRNA quantification procedure, the expression of 413 miRNA covered by at least 10 sequence reads were detected in human prefrontal cortex and cerebellum. The new miRNA quantification procedure is superior to that using the annotated mature miRNA sequence as a reference. Take miRNA quantification in human PFC samples as an example. The advantages were as follows. First, the new quantification procedure utilized substantially more reads for miRNA quantification: 1.2 million more reads. The number of reads derived from mature miRNAs increased from 72% to 94% of the total reads mapped to the miRNA precursor region. Second, the new quantification procedure can quantify many more miRNAs. In human PFC, the new quantification procedure measured 86 (29%) more miRNAs with expression of more than five read counts. On average, ~30% more miRNAs were quantified at various miRNA expression level cutoffs (Figure 3.5B). Third, the expression abundance of miRNA measured using this new quantification procedure displayed a comparable and even a slightly better correlation between technical replicates. Fourth, and most importantly, the new miRNA quantification procedure corrected the 5' end annotations for 27 miRNAs and 3'end annotations for 115 miRNAs with expression of more than 10 read counts. Since the functions of miRNAs are predominately determined by the seed region at their 5' ends, identifying the correct 5'end position of mature sequence for these 27 miRNAs was crucial for their target identification and functional studies. Identifying the correct 3'end was also important to designing probe sequences for other miRNA quantification measures such as using Q-PCR and miRNA microarray.

Table 3.2: miRNA quantification performance comparison

Methods	**Total Reads**	**Number of miRNA**		**Expression correlation**		**miRNA annotation correction**	
New quantification procedure	4,742,023	>=1 read	>=10 reads	Pearson correlation	Spearman correlation	5' end	3' end
		532	334	0.994	0.967	27	115
Based on annotation	3,527,780	>=1 read	>=10 reads	Pearson correlation	Spearman correlation		
		432	249	0.989	0.956		

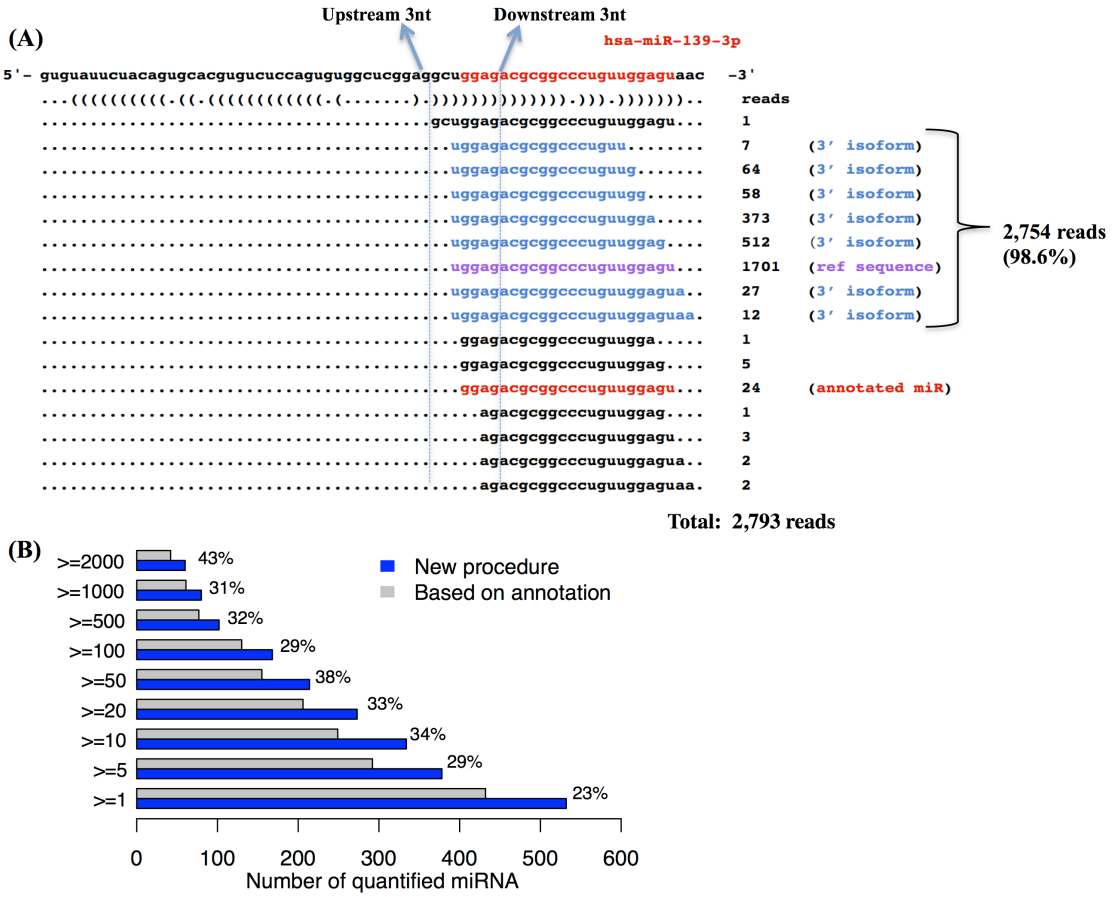


Figure 3.5: The new miRNA quantification procedure. (A) illustrates the new quantification procedure by using miR-139-3p as an example. In total, 16 sequences were retained, corresponding to 2793 reads mapped within three nucleotides upstream or downstream of the annotated 5'-position of miR-139-3p mature sequence. The sequence in red represents the annotated miR-139-3p from miRBase. The most highly expressed sequence was designated as newly identified miR-139-3p mature sequence, indicated in purple. The sequences in blue represent the 3'isoforms of newly identified miR-139-3p. miR-139-3p expression abundance was measured by the sum of the copy number of the newly identified miR-139-3p mature sequence and all its 3' isoforms. (B) The number of miRNAs identified and quantified with the new quantification procedure (in blue) compared with the quantification result based on the miRBase annotation (in grey) at different miRNA expression level cutoffs. The number on the right of each bar represents the proportion of miRNAs quantified exclusively by the new quantification method.

3.4 miRNA expression quantification in chimpanzee and macaque brains

The comparison of miRNA expression across species required miRNA gene annotations in each species as well as miRNA gene orthologous relationships across species. Based on miRNA annotations of miRBase (version 12), humans were well annotated with 692 miRNA precursors corresponding to 866 mature sequences. However, only 92 and 485 mature miRNAs were deposited in miRBase for chimpanzees and macaques (Table 3.3). Therefore, I developed a miRNA ortholog prediction procedure (MOP) to identify both miRNA precursors and mature sequences in chimpanzee and macaque based on human miRNA annotation (Section 2.2). The miRNA ortholog prediction procedure includes two consecutive steps: precursor ortholog identification and mature ortholog identification

(Figure 3.6).

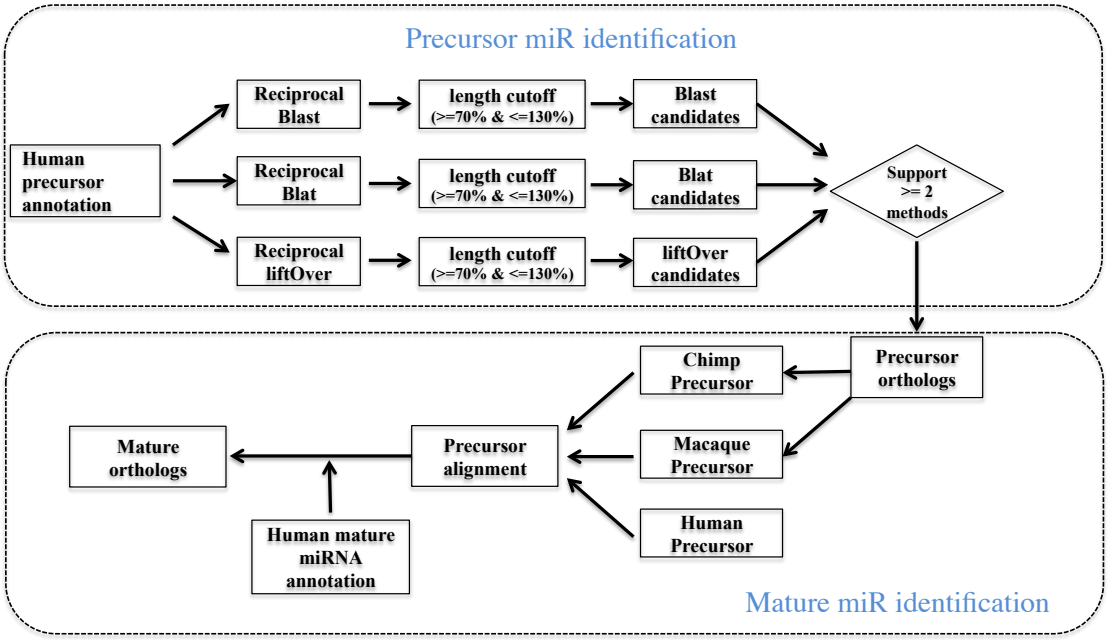


Figure 3.6: The miRNA orthologous gene prediction procedure. The workflow illustrates the two consecutive steps of miRNA orthologous gene prediction procedure: precursor ortholog identification and mature ortholog identification. Briefly, based on human miRNA annotation, the best precursor orthologs were predicted by using a combination of reciprocal BLAT, BLAST and liftOver along with precursor length cutoff in chimpanzee and rhesus macaque genomes. The mature orthologous sequences were further extracted based on ClustalW2 precursor sequence alignments, based on human mature miRNA annotation.

Using the miRNA ortholog prediction procedure, I detected 796 and 752 mature miRNAs in chimpanzee and macaque genomes, which corresponded to 92% and 87% of human annotated mature miRNAs, respectively. The identified miRNA orthologs covered more than 97% of annotated miRNAs of chimpanzee and macaque in miRBase (version 12) and thus greatly expanded the miRNA annotation in these two primate species (Table 3.3). Both precursor and mature orthologs predictions are of high quality. Specifically, 612 out of 639 (95.7%) predicted precursor orthologs were supported by all three methods (reciprocal BLAT, BLAST and liftOver) in chimpanzee (Figure 3.7A). Similarly, 522 out of 604 (86.4%) predicted precursor orthologs were supported by all three methods in macaque (Figure 3.7C). Compared to human miRNA precursors, predicted precursor orthologs have very similar length distribution. The precursor length differences were within 10% for all precursor orthologs in chimpanzee and 97% of precursor orthologs in macaque (Figure 3.7C and Figure 3.7D). For 578 miRNA expressed in human PFCs with at least one read, 543 and 526 mature orthologs could be unambiguously identified in chimpanzee and macaque genomes, respectively. The vast majority of these miRNAs were also expressed in chimpanzee (530) and macaque (504), measured using our new miRNA quantification procedure. In all three species, our miRNA orthologs prediction and miRNA quantification procedures generated highly reproducible miRNA expression measurements, with extremely good positive correlation between technical replicates in all three species (Pearson correlation, r>0.99,

p<10e-15), as shown in Figure 3.8A). Furthermore, miRNA expression divergence among species was evidently greater than the variance within species. The extent of miRNA expression divergence followed the phylogenetic relationship among three species, i.e., human and chimpanzee samples as sister species, with macaque samples forming an outgroup (Figure 3.8B and Figure 3.8C). All these results supported the validity of the developed miRNA ortholog prediction and miRNA quantification procedures.

Table 3.3: Number of miRNA orthologs predicted in chimpanzee and macaque

Species	**Annotated** [a]		**Predicted**		**Mature overlaps** [b]
	Precursor	**Mature**	**Precursor**	**Mature**	
Human	692	866			
Chimpanzee	100	92	639 [c]	796	91
			Blast 649		
			Blat 637		
			LiftOver 631		
Macaque	464	485	604 [d]	752	460
			Blast 626		
			Blat 571		
			LiftOver 577		

[a]miRNA annotation from miRBase (version 12)

[b]Number of overlaps between predicted and annotated mature miRNAs

[c and d]Number of predicted miRNA precursors in chimpanzee and macaque, respectively.

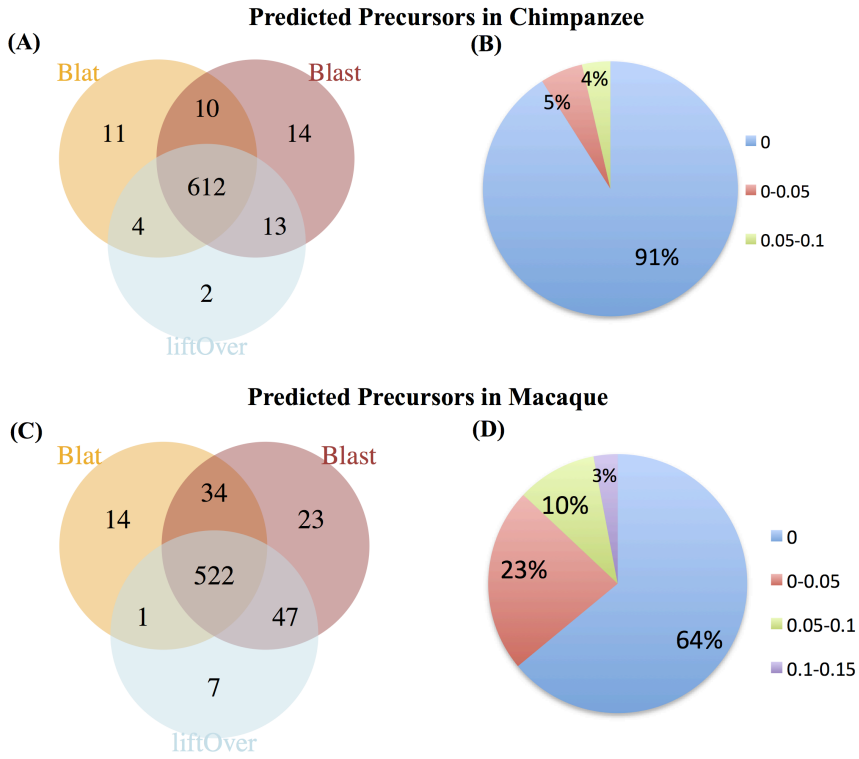


Figure 3.7: Predicted miRNA precursor orthologs in chimpanzee and macaque. Panels A and B describe precursor orthologs in chimpanzee. Panels C and D describe precursor orthologs in macaque. Panels A and C show the overlaps of precursor orthologs using reciprocal BLAT, BLAST and liftOver in chimpanzee and macaque, respectively. Panels B and D depict precursor length differences between human precursors and precursor orthologs in chimpanzee and macaque, respectively. Blue—no length difference; red—length difference between 0 to 5%; green—length difference between 5 to 10%; purple—length difference between 10 to 15%.

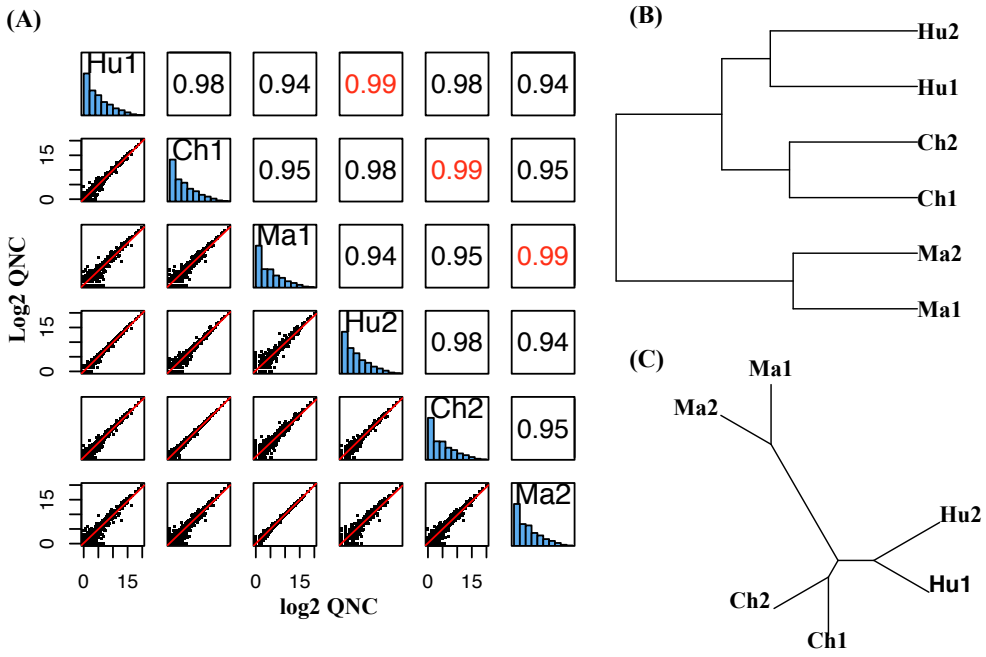


Figure 3.8: Predicted miRNA mature orthologs in chimpanzee and macaque. (A) Pairwise miRNA expression correlation between samples in human, chimpanzee and macaque PFCs. The x-axis and y-axis show quantiled normalized reads counts (QNC) in log2 scale (log2 QNC). The lower panels below the diagonal draw the scatter plot of miRNA expression between samples. The red line in each panel was fitted using loess (local polynomial regression fitting) for miRNA expression comparison. The panels on the diagonal draw the histograms of miRNA expression in each sample. The upper panels above the diagonal report the Pearson correlation coefficient. (B) and (C) show UPGMA (Unweighted Pair Group Method with Arithmetic Mean) and NJ (Neighbor Joining) trees based on miRNA expression of 3 species, respectively. The labels represent species Hu1—Human; Hu2—Human technical replicate; Ch1—Chimpanzee; Ch2—Chimpanzee technical replicate; Ma1—Macaque; Ma2—Macaque technical replicate.

3.5 Differentially expressed miRNA identification

Based on the quantified miRNA expression, I next investigated how many miRNAs were differentially expressed between human and other two primates in the prefrontal cortex. Two approaches, Fisher's exact test (FET) based method and edgeR method, were employed to identify differentially expressed (DE) miRNAs (Section 2.2). Since these two approaches utilized different normalization methods and statistical tests to predict DE miRNAs, the quality of identified DE miRNAs could be evaluated by accessing their prediction agreement. In the human or chimpanzee prefrontal cortex, 325 miRNAs were represented by at least 10 sequence reads in at least one technical replicate of one species. All 325 miRNAs had predicted orthologs in the chimpanzee genome. Based on quantile normalized miRNA expression data, the FET-based method predicted 37 DE miRNAs between human and chimpanzee (FET $p<0.01$, FDR<0.07, fold-change>2, in both technical replicates) (Figure 3.9A), representing ~11% of expressed miRNAs in PFC. Although the combination of statistical significance, fold-change and expression level criteria has been incorporated into the FET-based method, the deep sequencing reads overdispersion feature was not considered. The procedure for single factor differential gene expression analysis in edgeR package was further used to investigate the reads overdispersion effect on DE miRNA identification. Based on normalized miRNA expression data with TMM method, the edgeR method identified 35 DE miRNAs between human and chimpanzee (negative binomial test $p<0.001$, FDR<0.01) (Figure 3.9C). Notably, the vast majority (31, 84% of 37) of DE miRNAs were identified by both methods (binomial test, p<10e-5), which strongly supported the authenticity of identified DE miRNAs. Using the same criteria, 106 out of 338 miRNAs detected in human and macaque prefrontal cortex were differentially expressed between the two species, according to the FET-based method (Figure 3.9B). Similarly, 88 out of these 106 (83%) miRNAs were also classified by edgeR as differentially expressed (Figure 3.9).

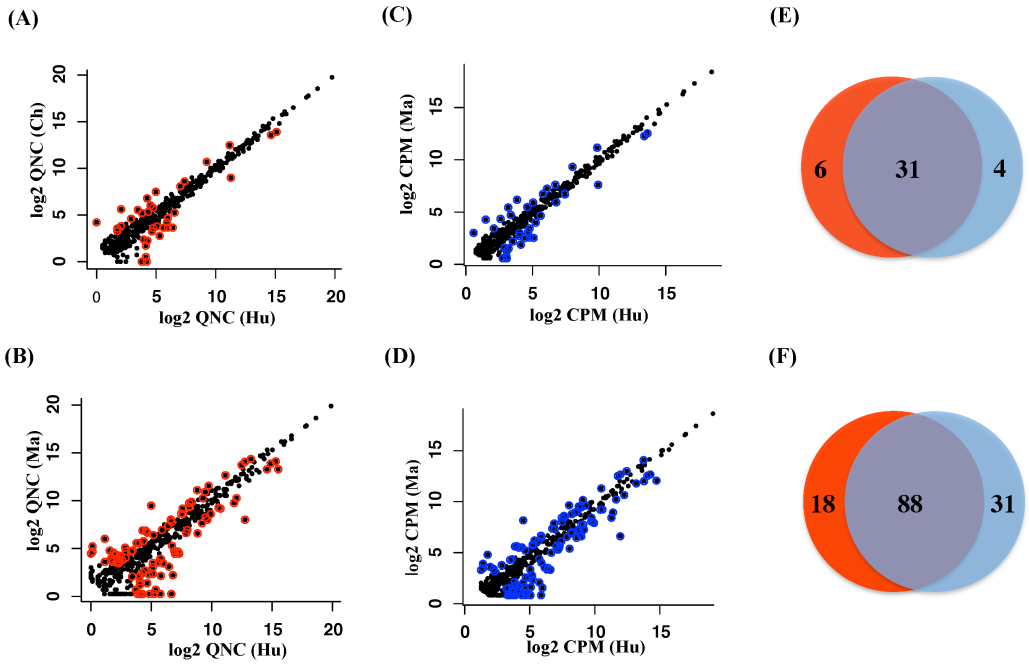


Figure 3.9: Differentially expressed miRNAs between species in PFC. Panels A and B depict DE miRNAs using the FET-based method. (A) for human and chimpanzee comparison; (B) for human and macaque comparison. The x-axis and y-axis represent the mean value of log2-transformed quantile normalized read counts (log2 QNC) of two replicates. DE miRNAs are marked with red outer circles. Panels C and D depict DE miRNAs using the edgeR method. (C) for human and chimpanzee comparison; (D) for human and macaque comparison. The x-axis and y-axis represent mean value of log2-transformed read counts per million (log2 CPM) of two replicates. DE miRNAs are marked with blue outer circles. Venn diagrams showing in Panels E and F depict the overlaps of DE miRNAs between the FET-based method and the edgeR method. (E) for human and chimpanzee DE miRNAs; (F) for human and macaque DE miRNAs. The Venn colors are red for DE miRNA by the FET-based method; blue for DE miRNA by the edgeR method; and purple for the overlaps of DE miRNA by two methods.

Besides the great agreement of DE miRNAs predictions between different methods, the vast majority of DE miRNAs that were found between species in the PFC could be reproduced in the cerebellum. Specifically, out of 37 DE miRNAs between human and chimpanzee in PFC, according to the FET-based method, 31 (84%) displayed consistent expression differences between species in both brain regions (Figure 3.10A and Figure 3.10B). Similarly, out of 106 DE miRNAs between human and macaque in PFC, 82 (77%) showed consistent expression differences between the two species in both brain regions (Figure 3.10C and Figure 3.10D). In both cases, the agreement between the two brain regions was far greater than could be expected by chance (binomial test, $p<10e-5$). Although the PFC and cerebellum are histologically different, previous studies have shown that mRNA expression differences between human and chimpanzee are largely shared between these two brain regions [227]. These results suggest that miRNA divergence is similarly shared between PFC and cerebellum. Furthermore, the good agreement of miRNA divergence estimates between the two brain regions supported the robustness of the miRNA differential expression

identification measurements.

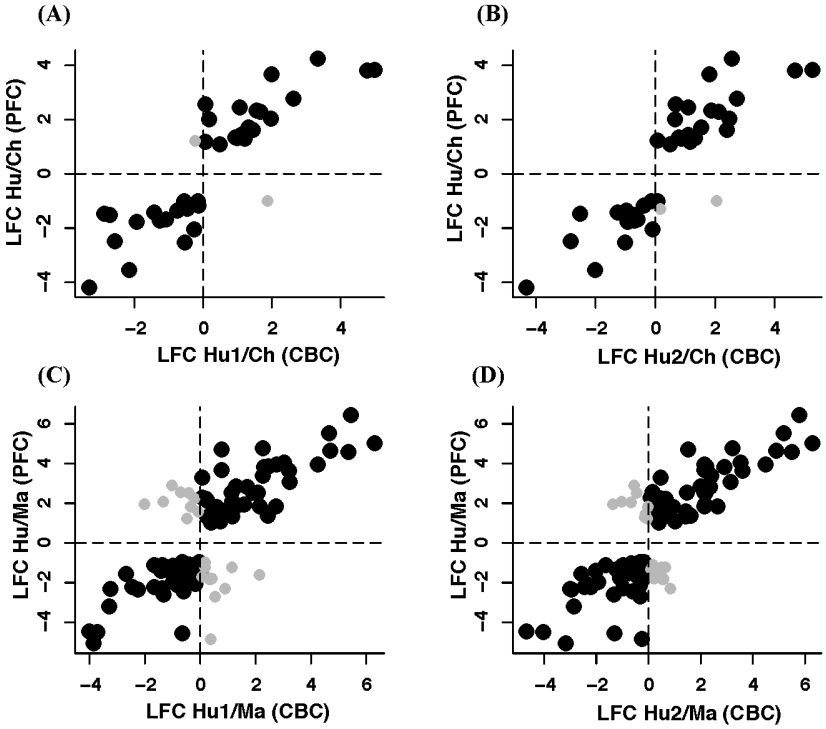


Figure 3.10: miRNA expression divergence between species measured in PFC and cerebellum. miRNA expression divergence was measured by expression log2-transformed fold-changes (LFC) between species. Panels A and B show miRNA divergence between human and chimpanzee measured in PFC and cerebellum. Panels C and D show miRNA divergence between human and macaque measured in PFC and cerebellum. The black dots indicate miRNA showing consistent direction of expression divergence in the two brain regions; grey dots for miRNA show inconsistent directions of expression changes. The labels represent species: Hu—Human; Ch—Chimpanzee; Ma—Macaque; Hu1 and Hu2 are two biological replicates of human cerebellum samples.

3.6 Effect of differentially expressed miRNA on target gene expression

What is the relationship between the expressions of DE miRNA and their targets? Do miRNA expression differences between human and chimpanzee brains contribute to gene expression divergence between these species? Or do miRNA expression differences make the gene expression more similar between human and chimpanzee by balancing transcription fluctuations? To investigate these questions, mRNA and protein expression were measured in human and chimpanzee prefrontal cortex: mRNA expression in five individuals of each species using Affymetrix Exon arrays and protein expression in four individuals of each species with two technical replicates using a label-free 2D-MS/MS Termo-LTQ proteomics methodology (Section 2.3). In total, 13,495 mRNA and 981 proteins were quantified with high confidence in human and chimpanzee prefrontal cortex. After expression normalization using the quantile normalization method, samples from human and chimpanzee could be separated well based on expression divergence for both mRNA and protein data (Figure 3.11),

which supported the validity of the developed cross-species mRNA and protein data quantification procedures.

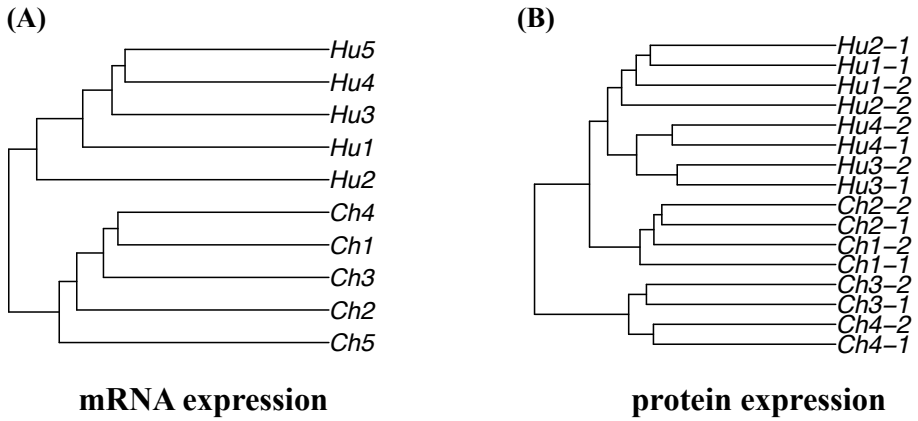


Figure 3.11: The mRNA and protein expression divergence between human and chimpanzee. Panels A and B show UPGMA trees based on the mRNA expression and protein expression, respectively. The mRNA expression was measured in five individuals of each species using Affymetrix Exon arrays. The protein expression was measured in four individuals of each species with two technical replicates using a label-free 2D-MS/MS Termo-LTQ proteomics methodology. The mRNA and protein data processing procedures are described in Section 2.3. Both mRNA and protein data were normalized using the quantile normalization method before clustering analysis.

To analyze the relationship between the expressions of DE miRNA and their targets, I compared the expression divergence between the targets of DE miRNA that were highly expressed in human with targets of DE miRNAs that were highly expressed in chimpanzee (Section 2.3). The target genes were predicted using TargetScan 5 algorithm [195] that predicted targets based on the presence of conserved miRNA binding sites on mRNA 3'UTR regions and are reported to have good sensitivity and specificity [129]. The Wilcoxon rank sum tests demonstrated that DE miRNA expression has a significant negative effect on both mRNA and protein expression in the human and chimpanzee prefrontal cortex, i.e., the targets of highly expressed miRNA were down-regulated in the corresponding species (p<0.05) (Figure 3.12), which was in line with the well-established function of miRNAs playing a role as negative regulators [21]. Notably, the negative regulatory effect did not depend on the choice of miRNA prediction algorithm since similar results were obtained using PITA predictions [228] that predicted targets based on the free energy gained from the formation of the miRNA-target duplex (p<0.05) (Figure 3.13). Furthermore, the negative effect of DE miRNA expression on mRNA and protein expression could be observed at various miRNA expression level cutoffs (Table 3.4). Finally, the negative regulatory effect on mRNA and protein expression divergence could also be observed at various mRNA expression divergence cutoffs (Figure 3.14) and protein expression divergence cutoffs (Figure 3.15). Taken together, the consistent and significant negative relationship between DE miRNA expression and the expression of their target genes, on both mRNA and protein levels,

demonstrated that miRNA expression divergence did contribute to gene expression divergence between human and chimpanzee.

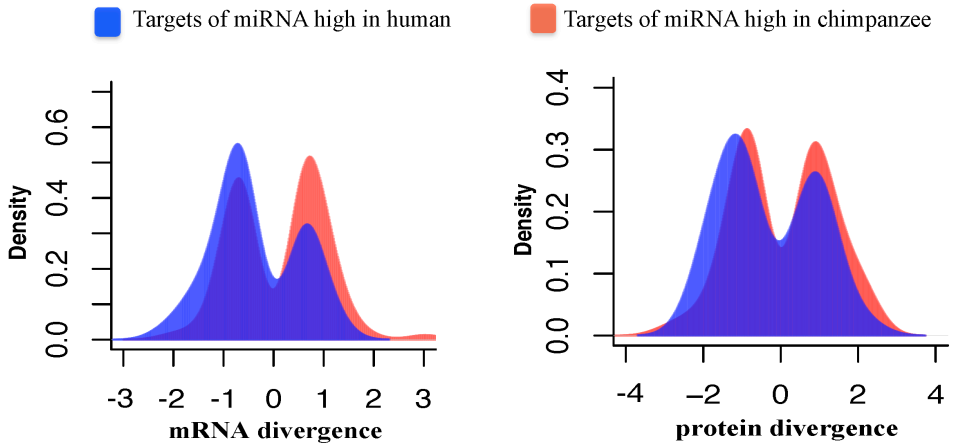


Figure 3.12: Effect of differential expressed miRNA on mRNA and protein expression (TargetScan targets). The left and right panels depict the distribution of mRNA and protein expression divergence for genes targeted by differentially expressed miRNAs between human and chimpanzee prefrontal cortex, respectively. The targets were predicted using TargetScan5 algorithm. The colors indicate genes targeted by miRNA: blue—miRNA highly expressed in human; and red—miRNA highly expressed in chimpanzee. The purple areas show overlap between red and blue distributions. mRNA divergence is represented as log2-transformed fold-change of gene expression between human and chimpanzee prefrontal cortex. Protein divergence is represented as the effect size difference of gene expression between human and chimpanzee prefrontal cortex. For both mRNA and protein divergence, positive values indicate higher expression in human prefrontal cortex.

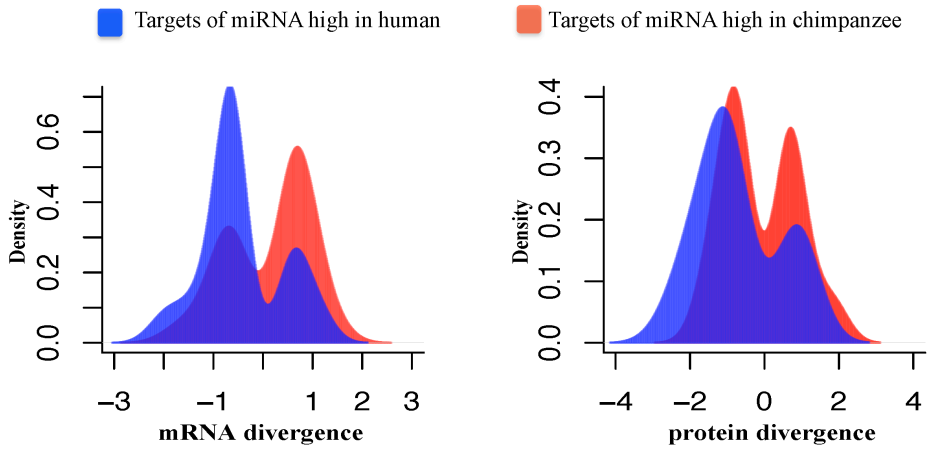


Figure 3.13: Effect of differential expressed miRNA on mRNA and protein expression (PITA targets). The left and right panels depict the distribution of mRNA and protein expression divergence for genes targeted by differentially expressed miRNAs between human and chimpanzee prefrontal cortex, respectively. The targets were predicted using PITA algorithm. The legend descriptions are the same as in Figure 3.11.

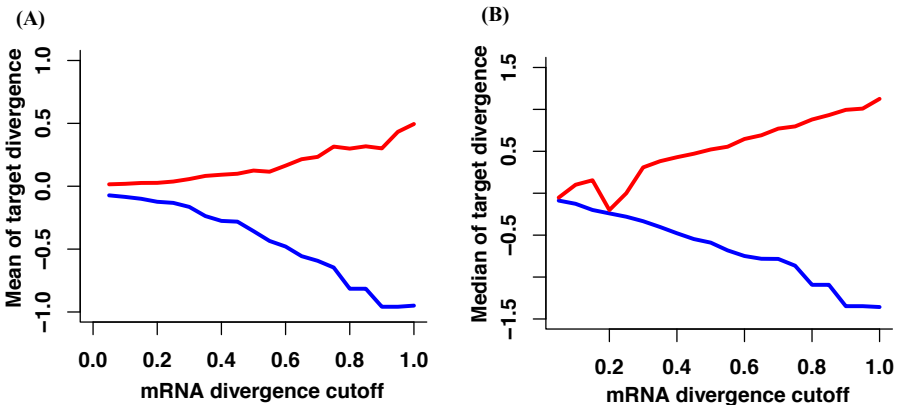


Figure 3.14: Effect of differential expressed miRNA on mRNA expression at different mRNA divergence cutoff. Differences in mean expression divergence (A) and median expression divergence (B) between mRNA targeted by miRNA with high expression in human PFC (blue curves) and mRNA targeted by miRNA with high expression in chimpanzee PFC (red curves) at different mRNA divergence cutoffs. The mRNA divergence cutoff was calculated based on absolute mean difference between human and chimpanzee expression levels.

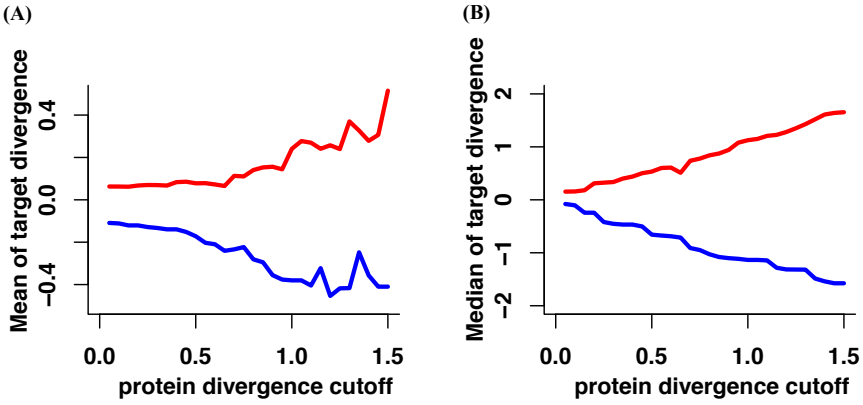


Figure 3.15: Effect of differential expressed miRNA on protein expression at different protein divergence cutoff. Differences in mean expression divergence (A) and median expression divergence (B) between protein targeted by miRNA with high expression in human PFC (blue curves) and protein targeted by miRNA with high expression in chimpanzee PFC (red curves) at different protein divergence cutoffs. The protein divergence cutoff was calculated based on absolute effect size difference between human and chimpanzee expression levels.

Table 3.4: Effect of differential expressed miRNA on mRNA and protein expression at different miRNA expression cutoffs

miRNA regulatory effect on mRNA expression		
miRNA expression cutoff (TPM)	Number of miRNA	Wilcoxon rank sum test P-value
50	26	0.0055
100	19	0.0003
200	11	0.0063
500	7	0.0128

miRNA regulatory on protein expression		
miRNA expression cutoff (TPM)	Number of miRNA	Wilcoxon rank sum test P-value
50	26	0.0716
100	19	0.0355
200	11	0.0279
500	7	0.0124

4. Identification and Functionality Estimation of miRNA 5'-isoforms in the Human Prefrontal Cortex

It is generally believed that the miRNA processing machinery ensures the generation of a single mature miRNA with a fixed sequence. However, as shown in Chapter 3, small RNA sequencing data allows us to scrutinize miRNA repertoire, resulting in an unexpected observation that miRNAs display heterogeneous ends. Although most of miRNA variants are generated due to heterogeneity of the 3' end sequence termination point (3'-isoforms), a notable proportion is also observed at the 5'end (5'-isoforms). The 5'-isoforms are particularly interesting since their seed regions are shifted compared with the annotated sequence, thus directing them to a distinct set of target genes. Currently, little is known about miRNA 5'-isoforms, except for their existence based on the measurements from small RNA sequencing, cloning and northern blotting. To deepen our understanding of miRNA 5'-isoforms abundance and functionality, I studied their authenticity and functionality on the basis of small RNA sequencing data collected in the human prefrontal cortex.

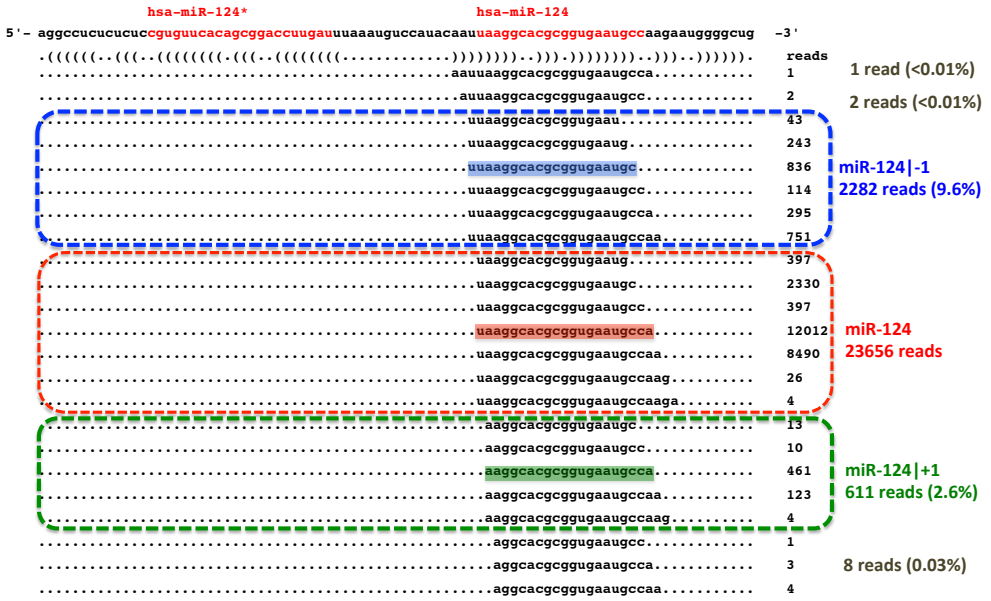


Figure 4.1: miRNA 5'-isoforms from miR-124. The plot depicts sequence and expression abundance of two 5'-isoforms from brain-specific miR-124. The sequences of two 5'-isoforms, miR-124|-1 and miR|124|+1, are marked in blue and green, respectively. The sequence of annotated miR-124 is marked in red. The expression of two 5'-isoforms and annotated miR-124 are measured by the sum of the copy number of the sequences inside blue, green and red boxes, respectively. At the expression level, 5'-isoform of miR-124 was required to take at least 1% of the read count of annotated miR-124. The rest of sequences showing 5' end shifts, with expression less than 1% of the read count of annotated miR-124, were filtered out.

4.1 miRNA 5'-isoform identification in the human prefrontal cortex

By reanalyzing small RNA sequencing data from human prefrontal cortex, in total, 203

5’-isoforms were identified based on the miRNA 5’-isoform identification and quantification procedure (Section 2.4). Figure 4.1 illustrates the sequence and expression abundance of two 5’-isoforms, miR-124|-1 and miR-124|+1, from the brain-associated miR-124. To facilitate the labeling of the identified 5’-isoforms, a notation system was developed (Section 2.4). For instance, miR-124|-1 represented a 5’-isoform with a 5′ terminus that begins one nucleotide to the left (5’ direction) of the annotated miR-124 5’end, while miR-124|+1 represented one 5’-isoform with a 5′ terminus that begins one nucleotide to the right (3’ direction) of the annotated miR-124 5’end (Figure 4.1).

Overall, in the human prefrontal cortex, the expression abundance of the identified 5’-isoforms was much lower than the annotated miRNAs (Wilcoxon rank sum test, p<10e-4) (Figure 4.2A). On the other hand, 5’-isoforms were expressed more abundantly than novel miRNA* sequences (Wilcoxon rank sum test, p<0.05) (Figure 4.2A). By using 10TPM as an expression cutoff that was comparable and higher than the median expression value of annotated miRNAs, 66 out of 203 (32.5%) 5’-isoforms were classified as moderately expressed (Table B.1). Correspondingly, 227 (44.9%) annotated miRNAs and 8 (14.8%) novel miRNA* showed moderate expression level, respectively (Figure 4.2A). Notably, with respect to the number of expressed miRNAs, these 66 5’-isoforms represented more than 20% of total moderately expressed miRNAs (Figure 4.2B).

These 66 5’-isoforms were derived from 50 annotated mature miRNAs. Among these 50 annotated mature miRNAs, 36 produced one 5’-isoform per miRNA and the remaining 14 miRNAs produced 30 5’-isoforms in total (Table B.1). Based on the seed sequence identity, 66 5’-isoforms were grouped into 64 seed families. Importantly, all these 64 seed families are novel compared with known miRNA families in humans, suggesting the identified 5’-isoforms may able to regulate a set of distinct target genes compared to annotated miRNAs.

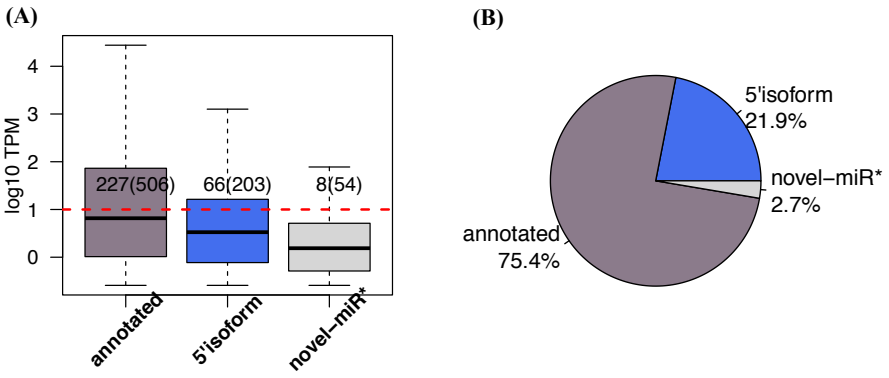


Figure 4.2: The expression abundance of 5’-isoforms in human prefrontal cortex. (A) The expression abundance of identified 5’-isoforms, compared with the expression levels of the other two categories: annotated miRNAs (annotated) and novel miRNA* (novel-miR*). The numbers in parenthesis list the total number of expressed miRNAs in each category. The numbers before parenthesis show the number of moderately expressed miRNAs (>=10TPM) in each category. (B) Proportion of moderately expressed 5’-isoforms out of total moderately expressed miRNAs, in terms of miRNA number.

Are these 66 5'-isoforms bona fide miRNAs? To answer this question, I investigated the expression association between 5'-isoforms and Argonaute 2 (AGO2) protein, which is the key component of RISC complex. Since most functional annotated miRNAs were bound in AGO2 protein, the association between AGO2 and miRNAs is commonly considered the most important criteria for genuine miRNA classification [21]. I tested whether 5'-isoforms were associated with AGO2 protein by quantifying their expression level in human brain based on small RNA sequencing data from AGO2 immunoprecipitation experiment (human brain AGO2-IP data). As shown in Figure 4.3, the vast majority of identified 5'-isoforms (59 out of 66) were readily detected in human brain AGO2-IP data, including five 5'-isoforms that have been studied previously (e.g., miR-101|-1, miR-142-3p|+2 and brain associated miR-9|+1). Besides miR-9|+1, two more brain-associated 5'-isoforms miR-124|-1 and miR9*|+1 were also detected with high expression level in human brain AGO2-IP data (Figure 4.3). Furthermore, a significantly positive expression correlation was observed for 5'-isoform expressions between human PFC and human brain AGO2-IP data (Pearson correlation r= 0.48, p<4e-5), suggesting the majority of identified 5'-isoforms may functional in vivo.

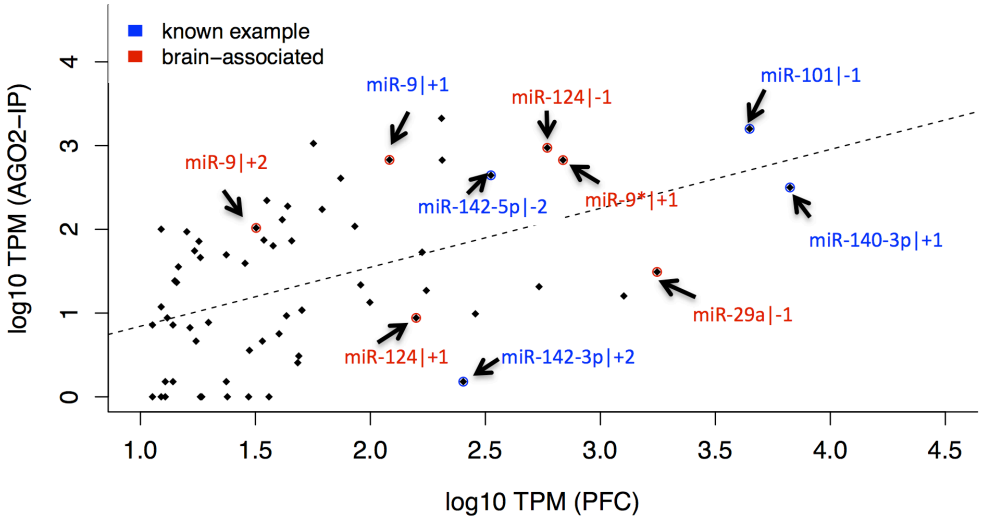


Figure 4.3: 5'-isoform expression in human brain from AGO2-IP experiment. The x-axis displays 5'-isoform expression level in human prefrontal cortex. The y-axis displays 5'-isoform expression level in human brain, based on small RNA sequencing data from AGO2-IP experiment. The miRNA expression was normalized by total mapped reads from corresponding samples into transcripts per million reads (TPM) and showed in a log10-tranformed scale. Examples of known 5'isofroms are marked with blue outer circles. Examples of novel 5'-isoforms from brain-associated miRNAs are marked with red outer circles

At sequence level, all 66 5'-isoform have conserved seed sequences between human chimpanzee, macaque and mouse. In addition to sequence conservation, the expression of 5'-isoforms was significantly positively correlated between human and chimpanzee (Pearson correlation, r=0.93, p<10e-5), and between human and macaque (Pearson correlation, r=0.89,

p<10e-5). In addition, 61 out of 66 5'-isoforms were also expressed in the mouse brain (Figure 4.4). Notably, 5'-isoforms from brain-associated miRNAs (such as miR-9, miR-9* and miR-124) were expressed more abundantly in mouse brain (Figure 4.4). Similarly to the scenario in human brain, a significantly positive expression correlation was observed for 5'-isoform expressions between human PFC and mouse brain (Pearson correlation r=0.33, p<2e-3). Taken together, the expression association with AGO2 protein in human brain and both sequence and expression conservation in the mouse brain strongly indicated the identified 5'-isoforms are bona fide miRNAs. In the following study, I focused on these 66 moderately expressed 5'-isoforms to further investigate their functionality.

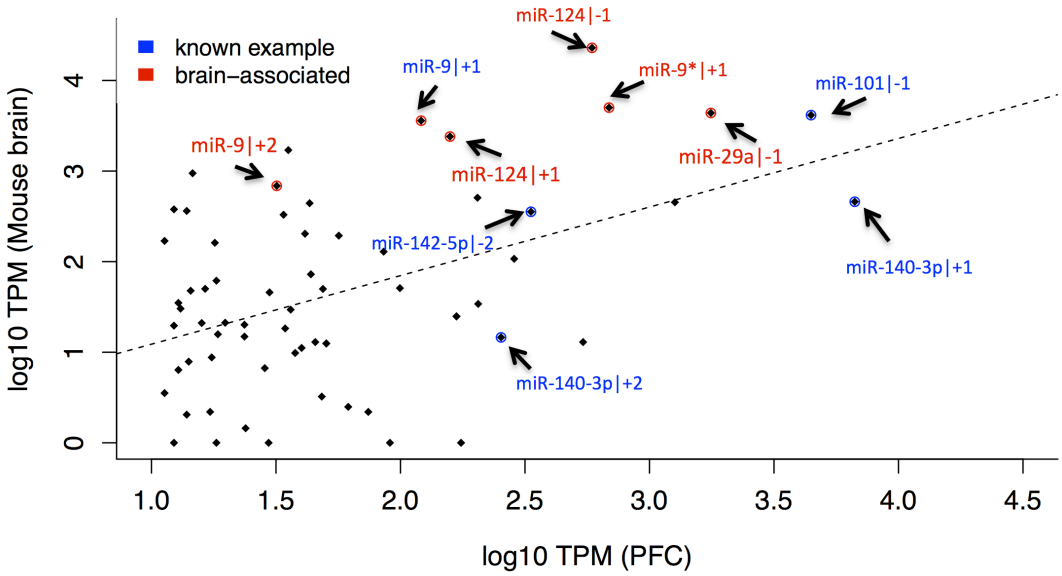


Figure 4.4: 5'-isoform expression in mouse brain. The x-axis displays 5'-isoform expression level in human prefrontal cortex. The y-axis displays 5'-isoform expression level in mouse brain. The miRNA expression was normalized by total mapped reads from the corresponding sample into transcripts per million reads (TPM) and showed in a log10-tranformed scale. Examples of known 5'isofroms are marked with blue outer circles. Examples of novel 5'-isoforms from brain-associated miRNAs are marked with red outer circles

4.2 The procedures for miRNA 5'-isoform functionality prediction

This abovementioned analysis results strongly suggest that the identified 66 5'-isoforms moderately expressed in the human prefrontal cortex were bona fide miRNAs. However, we still lack estimate of their functionality. Since the majority of 5'-isoforms identified in the human PFC were conserved between human and mouse at both sequence and expression levels, I addressed this question using a comparative approach, based on the observation of co-evolution between conserved miRNAs and their target sites (Section 1.2 and Section 2.5). Whether a given 5'-isoform was functional was inferred based on the conservation status of its seed matches in the 3'UTR (heptamer sequence that is complementary to the

corresponding miRNA's seed region; the terms "seed match" "miRNA target site" and "heptamer" are equivalent in this study). The 5'-isoform with seed matches showing excessive conservation was considered to be functional.

In brief, the approach was comprised of three steps. The first step is to estimate the observed heptamer conservation based on human-mouse 3'UTR alignment by enumerating both conserved and total heptamer occurrences. The second step, which is the most crucial, is to obtain the expected conservation or background conservation of heptamer based on the control sets that were generated using sequence-shuffling based methods. In the last step, by combining the result of the first two steps, the cutoff 0.05 representing the Benjamini-Hochberg (BH) corrected p-value of the binomial test was used to determine whether one seed match has excessive conservation. Since a statistically and biologically meaningful result of any sequence motif analysis largely depends on choosing an appropriate control set, I used five sequence-shuffling procedures to construct control sequence sets in the second step. Correspondingly, five prediction procedures were developed to predict functional 5'-isoforms (Figure 4.5), including one based on seed match sequence shuffling (SSP) and four more shuffling procedures based on 3'UTR alignment shuffling (USP1, 2, 3 and 4). Section 2.5 described the five prediction procedures in detail.

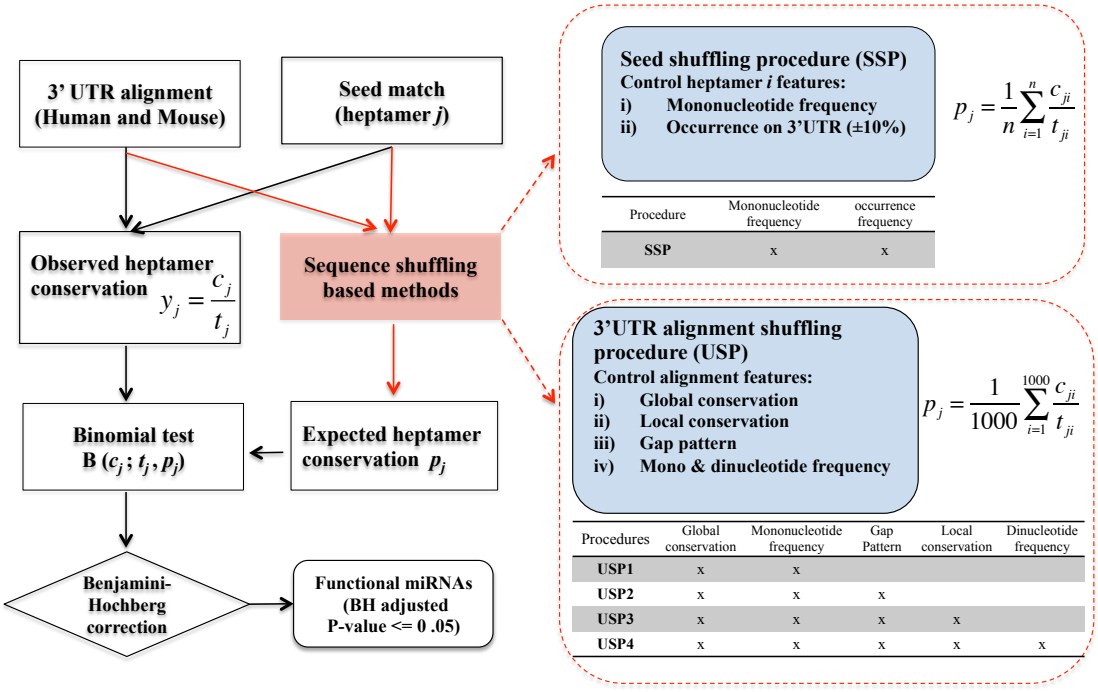


Figure 4.5: Five procedures for functional 5'-isoform prediction. Functional 5'-isoform was inferred based on the conservation status of its seed matches in the human 3'UTR. The significance of excessive conservation of seed matches was determined by comparing observed and expected heptamer conservations with binomial test after Benjamini-Hochberg correction. The criteria for generating shuffled sequences are listed in the red dashed rectangle. Based on five sequence shuffling methods, one seed match shuffling procedure (SSP) and four 3'UTR alignment-shuffling procedures (USP1, 2, 3 and 4) were developed to predict functional 5'-isoforms (Section 2.5). The name of sequence shuffling methods was used to represent the name of the functional 5'-isoform prediction method.

4.3 Performance comparison of 5'-isoform prediction procedures

To identify the best procedure to predict functional 5'-isoforms, the performances of five prediction procedures were evaluated (Section 2.5). The prediction performances were evaluated at two levels, miRNA family level and miRNA level. The conserved and nonconserved miRNA families of human were used as positive and negative sets to estimate prediction performance at the miRNA family level, and the corresponding miRNAs were used to estimate prediction performance at the miRNA level (Section 2.5). Based on miRNA annotations of mouse, rat, dog and chicken in miRBase (version 12), 162 human conserved and 284 nonconserved miRNA families were obtained, which corresponds to 262 conserved and 326 nonconserved miRNAs, respectively. The prediction performance was evaluated based on sensitivity (SN), specificity (SP), accuracy (ACC) and Matthews correlation coefficient (MCC). Table 4.1 and Table 4.2 summarize prediction performances of five prediction procedures on the miRNA family level and miRNA level, respectively.

The performance comparison showed that USP1 and USP2 were much more sensitive than all other methods on both miRNA family and miRNA levels. However, USP1 and USP2 also showed the lowest prediction specificity. On the other hand, SSP, USP3 and USP4 exhibited relatively balanced prediction sensitivity and specificity. USP4 performed better in terms of both prediction sensitivity and specificity compared to SSP and USP3 on both the miRNA family and miRNA levels. On the miRNA family level, SSP performed better than USP3 for both prediction sensitivity and specificity. However, USP3 had higher prediction sensitivity than SSP on the miRNA level. With respect to ACC, USP4 performed best, and SSP was the second best method. In this study, since the positive set and negative set were unbalanced in terms of both miRNA family number and miRNA number, MCC values were finally used to rank prediction performance. A higher MCC value indicates better performance. With respect to MCC, USP4 is superior to all other methods on both the miRNA and miRNA family levels; SSP and USP3 were ranked as the second best method and have similar performance. SSP is better than USP3 on the miRNA family level while the opposite is true on the miRNA level. USP1 and USP2 performed the worst.

In the following section, I further evaluated the prediction performance of the best classifier, USP4, in a more comprehensive way. First, to further evaluate the prediction accuracy of USP4, I compared the predicted functional miRNAs and miRNA families between USP4 and SSP, the second-best classifier. Since USP4 and SSP utilized different approaches to estimate the expected conservation of seed matches, the comparison of their predictions could be considered a good indicator for prediction accuracy. For the 85 functional conserved miRNA families predicted by USP4, 76 (89%) were among the predicted functional conserved miRNA families by SSP (Figure 4.6A). Similarly, at the miRNA level, 151 out of 173 (87%) predicted functional conserved miRNAs were overlapped between USP4 and SSP. The high concordance of predictions between USP4 and SSP indicated the good prediction accuracy of USP4.

Table 4.1: Prediction performance comparison on miRNA family level.

Method	TP	FN	FP	TN	SN	SP	ACC	MCC
SSP	82	80	47	237	0.51	0.83	0.72	0.36
USP1	153	9	245	39	0.94	0.14	0.43	0.13
USP2	137	25	208	76	0.85	0.27	0.48	0.13
USP3	79	83	56	228	0.49	0.8	0.69	0.30
USP4	85	77	40	244	0.52	0.86	0.74	0.41

Table 4.2: Prediction performance comparison on miRNA level.

Method	TP	FN	FP	TN	SN	SP	ACC	MCC
SSP	159	103	58	268	0.61	0.82	0.73	0.44
USP1	251	11	287	39	0.96	0.12	0.49	0.14
USP2	234	28	248	78	0.89	0.24	0.53	0.17
USP3	169	93	61	265	0.65	0.81	0.73	0.47
USP4	173	89	52	274	0.66	0.84	0.76	0.51

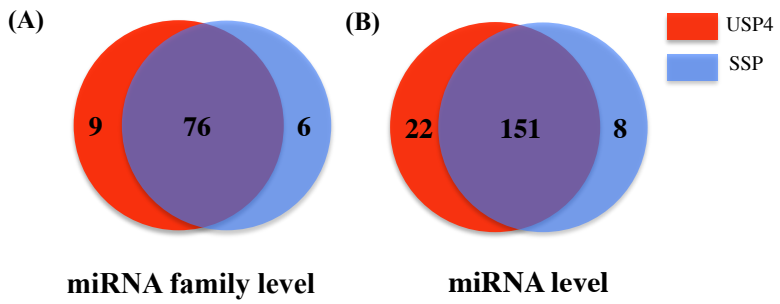


Figure 4.6: The comparison of predicted functional conserved miRNAs and miRNA families between USP4 and SSP. (A) for the overlaps of predicted functional conserved miRNA families between USP4 and SSP. (B) for the overlaps of predicted functional conserved miRNA between USP4 and SSP.

High expression abundance is generally believed to be a hallmark for potential functional regulatory transcripts, including miRNAs. To further evaluate the prediction performance of USP4, I separated conserved miRNA families into functional and nonfunctional categories on the basis of USP4 prediction and further compared the expression abundance of miRNAs from these two categories in the human prefrontal cortex. As shown in Figure 4.7A, the predicted functional conserved miRNA families were expressed significantly higher than nonfunctional conserved miRNA families (Wilcoxon rank sum test, $p<1e-12$). Strikingly, although predicted functional conserved miRNA families only represented 54% of the total expressed miRNA families in terms of family number, they represented 94% of total reads of conserved miRNA family (Figure 4.7B and Figure 4.7C). It should be noted that this result could not be attributed to certain super highly expressed miRNAs, such as miRNAs from let-7 family that represented ~67% of total miRNA read counts in human PFC, because miRNAs from let-7 family were excluded from this analysis. In addition, higher expressed miRNA families tended to have a higher possibility of being predicted as functional (Figure 4.8A). By contrast, no relationship was detected between the expression abundance and functional family predictions from ncRNA fragments that were used as a negative control (Figure 4.7 and Figure 4.8). To further investigate whether the relationship between the status functional miRNA family predictions and expression abundance is only specific in the human prefrontal cortex, the same analysis was conducted using small RNA sequencing data from human placenta and Hela cell line. In both cases, similar results were observed even though expressed miRNAs were poorly correlated between human PFC and other two samples (Figure 4.9 and Figure A. 1). Taken together, the results suggested that USP4 was able to capture the majority of functional conserved miRNA families.

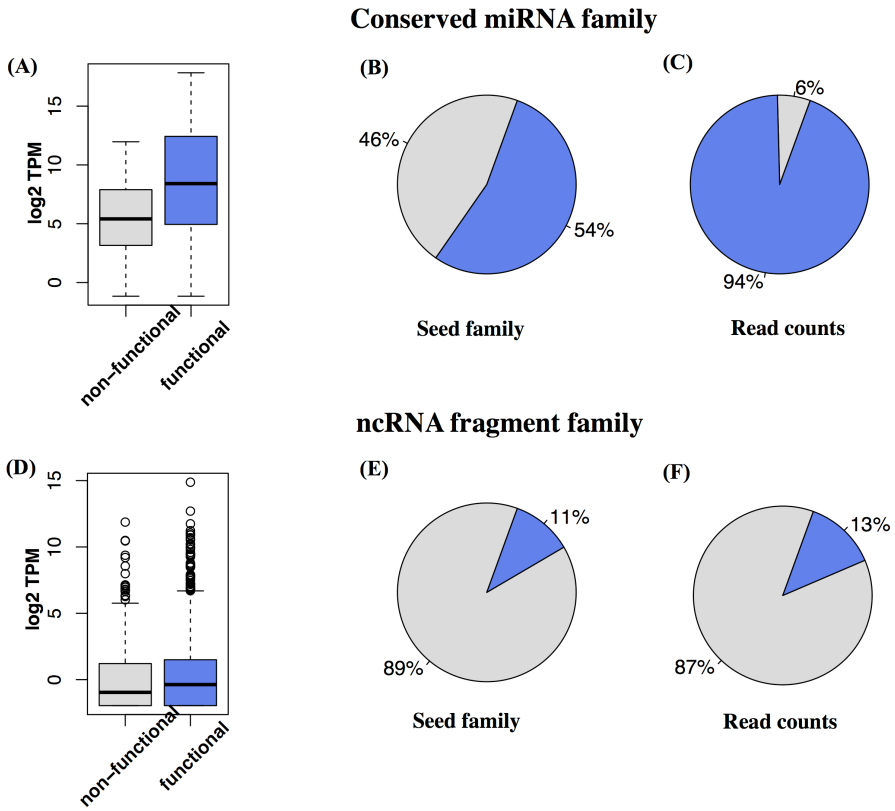


Figure 4.7: The expression abundance comparison between functional and nonfunctional conserved miRNA families based on USP4 prediction in the human prefrontal cortex. Panels A, B and C show the expression of conserved miRNA family. Panels C, D and F show the expression of ncRNA fragment family as a negative control. (A) Expression abundance distributions of predicted functional and nonfunctional conserved miRNA families with boxplot. Panels B and C show the proportion of predicted functional conserved miRNA families in terms of expressed miRNA family number and total read counts. (D) Expression abundance distributions of predicted functional and nonfunctional ncRNA fragment families with boxplot. Panels E and F show the proportion of predicted functional ncRNA fragment families in terms of expressed ncRNA fragment family number and total read counts. The functional conserved miRNA family and ncRNA fragment family were predicted using USP4.

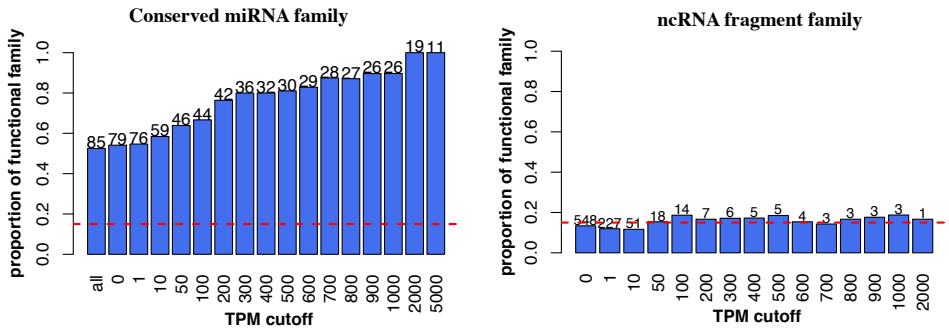


Figure 4.8: Proportion of predicted functional conserved miRNA family on different miRNA families' expression levels. The left panel shows conserved miRNA family. The right panel shows ncRNA fragment family as a negative control. The red dashed line represents the

prediction false positive rate of USP4. The number above the bar represents the number of predicted functional miRNA families at each miRNA family expression cutoff. The miRNA families were binned based on their expression level normalized in TPM. The cutoff of "all" represented all predicted functional miRNA families. The cutoff of 10 TPM represented the predicted functional miRNA families with expression levels larger than 10 TPM.

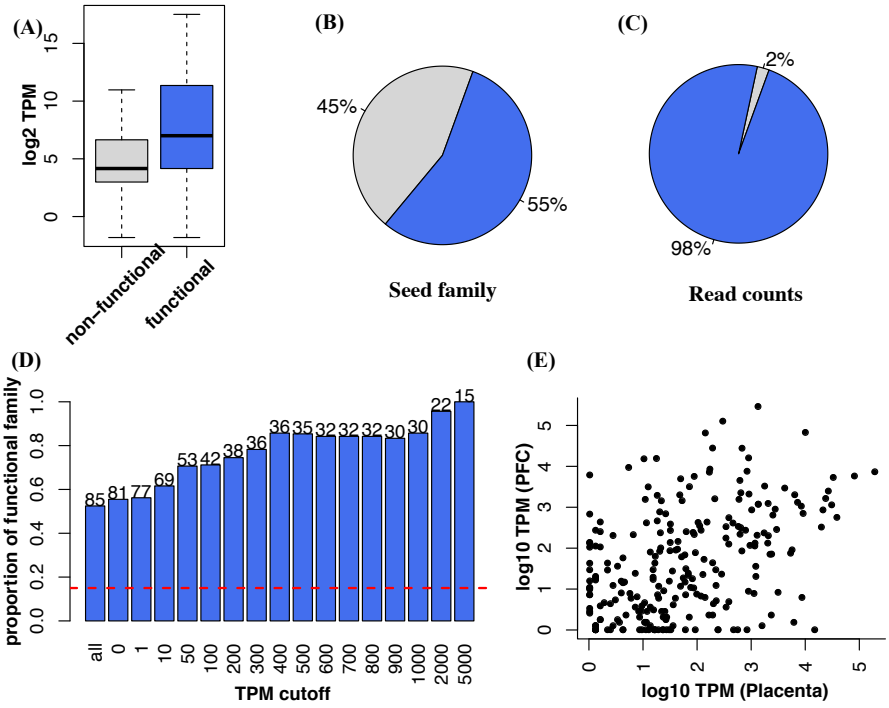


Figure 4.9: The expression abundance comparison between functional and nonfunctional conserved miRNA families based on USP4 prediction in the human placenta tissue. (A) Expression abundance distributions of predicted functional and nonfunctional conserved miRNA families with boxplot. Panels B and C show the proportion of predicted functional conserved miRNA families in terms of expressed miRNA family number and total read counts. The functional conserved miRNA family and ncRNA fragment family were predicted using USP4. (D) Proportion of predicted functional conserved miRNA family on different miRNA families' expression levels. The number above the bar represents the number of predicted functional miRNA families at each miRNA family expression cutoff. (E) Expression comparison of conserved miRNAs between human prefrontal cortex (PFC) and placenta (Pearson correlation r=0.42, p<10e-5). The miRNA expressions were normalized by total mapped reads of corresponding sample into Transcript Per Million reads (TPM).

Finally, I evaluated USP4 prediction performance based on the seed match conservation strand bias of predicted functional conserved miRNA families. The seed match conservation strand bias was measured using the differences of seed match conservation between sense and antisense strand of 3'UTR. Since miRNAs are only bound to the sense strand of 3'UTR, seed match conservation bias on the sense strand was expected if the corresponding conserved miRNA was functional. As shown in Figure 4.10A, a strong seed match conservation strand bias on the sense strand was observed for predicted functional conserved miRNA families. By

contrast, no obvious seed match conservation strand bias was observed for predicted nonfunctional conserved miRNA families (Figure 4.10B). This result again indicated that the majority of functional conserved miRNA families have been identified by USP4.

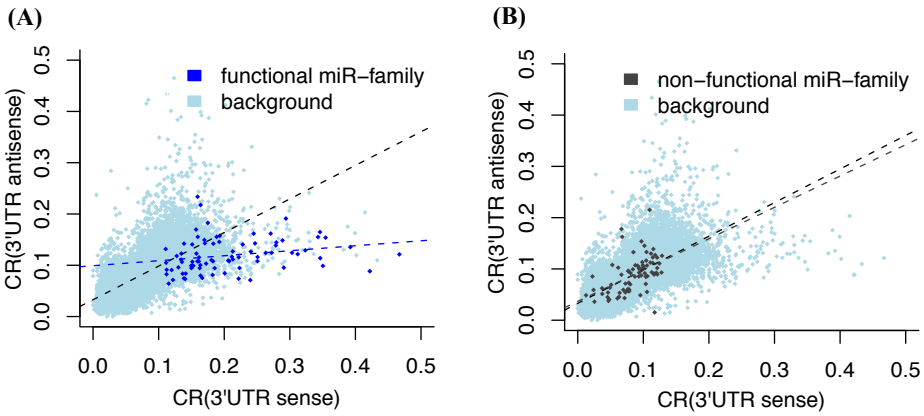


Figure 4.10: Strand bias of seed match conservation of predicted functional conserved miRNA families. Seed match conservation differences between sense and antisense strand of 3'UTR for functional conserved miRNA family (A) and nonfunctional conserved miRNA family (B). CR (seed match conservation rate) was measured as the proportion of conserved seed match occurrences out of the total seed match occurrences. The heptamers with similar seed match occurrences distribution as conserved miRNA families were used as background and are shown in light blue dots.

4.4 Functional miRNA 5'-isoform prediction

Having established the good prediction performance of USP4, I then applied USP4 to obtain functional miRNA 5'-isoform families. Out of 64 5'-isoform families identified from human prefrontal cortex, USP4 predicted 26 (41%) functional 5'-isoform families corresponding to 28 5'-isoforms (Table 4.3). Notably, 23 out of 26 (88%) functional 5'-isoform families were also supported by SSP (Table 4.3). Similarly to functional conserved miRNA families, functional 5'-isoform families were expressed significantly higher than nonfunctional 5'-isoform families, which represented 94% of total reads of 5'-isoforms (Figure 4.11). Furthermore, higher expressed 5'-isoform families tended to have a higher possibility of being predicted as functional (Figure 4.11D). Functional 5'-isoform families also displayed a strong seed match conservation strand bias that was similar to functional conserved miRNA families. The results demonstrated that the identified 26 functional 5'-isoform families might be as functional as functional conserved miRNAs.

Table 4.3: Predicted functional 5’-isoforms based on USP4 method

5'-isoform ID	Seed sequence	Predicted by USP4	Predicted by SSP
hsa-mir-124|-1	TAAGGCA	+	+
hsa-mir-29a|-1	TAGCACC	+	+
hsa-mir-101|-1	TACAGTA	+	+
hsa-mir-27b|-1	TTCACAG	+	+
hsa-mir-9|+1	TTTGGTT	+	+
hsa-mir-30e|+1	TAAACAT	+	+
hsa-mir-330-3p|+1	AAAGCAC	+	+
hsa-mir-181b|-1	AACATTC	+	+
hsa-mir-199a-3p|-1	ACAGTAG	+	+
hsa-mir-199b-3p|-1	ACAGTAG	+	+
hsa-mir-181b|+1	CATTCAT	+	+
hsa-mir-330-3p|2	AAGCACA	+	+
hsa-mir-126|+1	GTACCGT	+	+
hsa-mir-124|+1	AGGCACG	+	+
hsa-mir-199a-3p|+1	AGTAGTC	+	+
hsa-mir-199b-3p|+1	AGTAGTC	+	+
hsa-mir-9|+2	TTGGTTA	+	+
hsa-mir-9*|+1	AAAGCTA	+	+
hsa-mir-137|+1	ATTGCTT	+	+
hsa-mir-487b|+1	TCGTACA	+	-

hsa-mir-323-3p|-1	CACATTA	+	+
hsa-mir-539*|+3	TACAAGG	+	+
hsa-mir-99a|-1	AACCCGT	+	+
hsa-mir-363|+1	TTGCACG	+	+
hsa-mir-142-5p|-2	CCATAAA	+	-
hsa-mir-191|+1	ACGGAAT	+	+
hsa-mir-409-3p|-1	GAATGTT	+	+
hsa-mir-24|+1	GCTCAGT	+	-

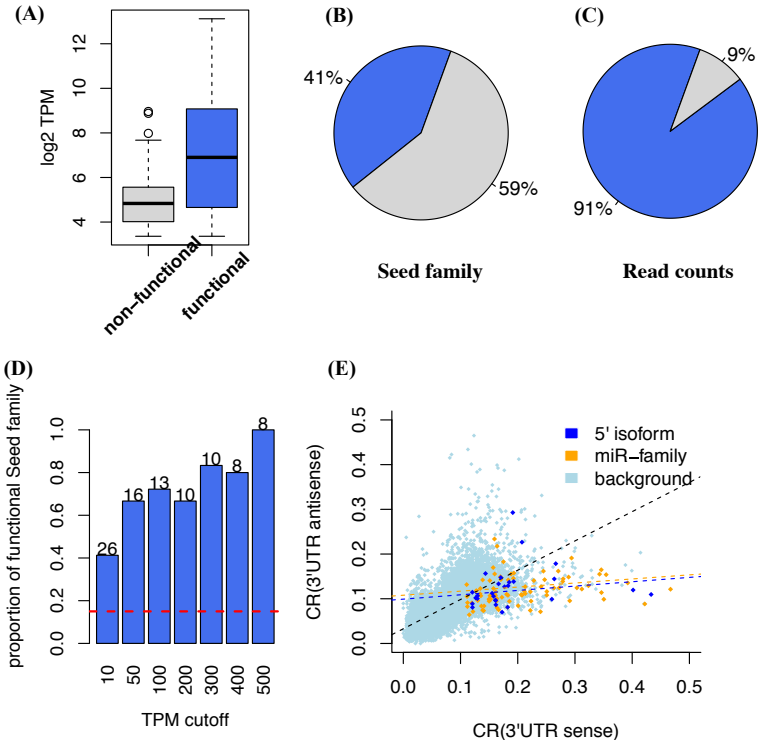


Figure 4.11: Predicted functional 5’-isoform families in the human prefrontal cortex. (A) Expression abundance distributions of predicted functional and nonfunctional 5’-isoform families with boxplot. Panels B and C show the proportion of predicted functional 5’-isoform families in terms of 5’-isoform family number and total read counts, respectively. (C) Proportion of predicted functional 5’-isoform families on different 5’-isoform families' expression levels. (D) Seed match conservation differences between sense and antisense strand of 3’UTR. Colors represent: dark blue—functional 5’-isoform families; orange—functional conserved miRNA families; and light blue—heptamers with similar seed match occurrences

distribution as conserved miRNA families and 5'-isoform families.

4.5 Analysis of regulation of 5'-isoform on the target expression

Finally, to verify the prediction of functional 5'-isoforms of the USP4 method experimentally, I investigated the regulatory effect of 5'-isoforms on the target gene expression based on published miRNA precursor overexpression and miRNA knockout experiments.

I first analyzed mRNA expression profiles before and after miR-124 primary transcript (pri-miRNA) overexpression using RIPmiR-124 plasmids in mouse neuroblastoma CAD cells [200] (Section 2.6). miR-124 produced two conserved 5'-isoforms, miR-124|+1 and miR-124|-1, from both human prefrontal cortex and mouse brain (Table 4.4). At the expression level, miR-124|-1 was more abundant than miR-124|+1 in both human prefrontal cortex and mouse brain, suggesting a potential conserved 5'-isoform biogenesis mechanism across tissues and across species. In this study, I assumed miR-124 pri-miRNA overexpression experiment would produce both miR-124 and two 5'-isoforms in mouse neuroblastoma CAD cells. The northern blotting result from the original paper [200] provided indirect evidence to support this.[2]

Table 4.4: The expression abundance and target gene number for miR124 5'-isoforms

miRNA	Expression (TPM)		USP4 (adjusted p-value)	Exclusive Conserved Targets (ECTs)	Regulatory Effect	
	Human PFC	Mouse brain			Wilcox rank sum test (pvalue)	KS-test (pvalue)
miR-124	6081	78678	2.2e-16	719	2.2e-16	2.2e-16
miR-124|+1	157 (2.6%)	2403 (3.1%)	9.2e-09	86	0.056	0.039
miR-124|-1	586 (9.6%)	22930 (30%)	2.2e-16	112	9.1e-06	2.80e-05

Based on USP4 prediction, both miR-124 and its two 5'-isoforms were predicted to be functional. As observed for functional conserved miRNA family and 5'-isoforms, the expression abundance of miR-124 and its two 5'-isoforms was also positively correlated with the probability to be functional predicted by USP4 (Table 4.4). To estimate the regulatory effect of miR-124|+1 and miR-124|-1 on target gene expression, I first obtained exclusive conserved targets (ECTs) of two 5'-isoforms and further compared mRNA expression

[2]The northern blotting result after miR-124 primary transcript overexpression in mouse CAD cell lines can be found in Figure 1A and Figure S1C of the original paper [200].

differences of the ECTs of two 5'-isoforms and nontarget genes, before and after miR-124 precursor overexpression (Section 2.6). It should be noted that ECTs of miR-124|+1 and miR-124|-1 precluded any targets of miR-124 with both canonical and weaker 6mer seed matches without considering target site conservation. Therefore, the influence of miR-124 on the ECTs of two 5'-isoforms was mostly avoided. As shown in Figure 4.12, compared to nontargets, ECTs of miR-124|-1 were significantly down-regulated after overexpressing miR-124 precursor (KS test, $p<2.8e-05$, Wilcoxon rank sum test, $p<9.1e-06$). miR-124|+1 also showed the tendency to repress its ECTs, although the repression magnitude is only marginally significant (KS test $p<0.039$, Wilcoxon rank sum test $p<0.056$). As expected, ECTs of miR-124 showed the largest magnitude of down-regulation (KS test $p<2.2e-16$, Wilcoxon rank sum test $p<2.2e-16$). Notably, the repression effect of miR-124 and its two 5'-isoforms were positively correlated with their expression abundance, e.g., miR-124|-1 was more abundant than miR-124|+1 and also caused a stronger target repression magnitude than miR-124|+1 (Figure 4.12, Table 4.4), indicating that 5'-isoforms with relatively high expression abundance were sufficient to regulate a distinct set of target genes. The result also suggested that the regulatory effect of 5'-isoforms could not be attributed to miR-124 precursor overexpression artifacts.

To further confirm the regulatory effect of 5'-isoforms, I examined the specificity of detected regulation of miR-124|+1 and miR-124|-1 on their targets more explicitly by conducting the same analysis with published microarray data from miR-124 duplex transfection experiments in human Hela cell line [129]. Theoretically, the miR-124 duplex transfection experiment only delivered the intact miR-124 mature sequence into cells. Therefore, it can be used as a negative control to measure the regulation specificity of miR-124|+1 and miR-124|-1 on their targets. At least for miR-124|-1, miR-124 duplex transfection experiment was an appropriate control since the miR-124|-1 sequence was not nested in the sequences from the miR-124 duplex. As expected, ECTs of miR-124 were significantly repressed after miR-124 duplex overexpression (Wilcoxon rank sum test, $p<2.2e-16$) (Figure 4.13). By contrast, no significant down-regulation was detected from ECTs of miR-124|+1 and miR-124|-1 (KS-test, $p>0.15$; Wilcoxon rank sum test, $p>0.1$) (Figure 4.13). This result demonstrated that the expression repression of ECTs in miR-124 precursor overexpression experiment was mostly attributed to the specific regulation of miR-124|+1 and miR-124|-1.

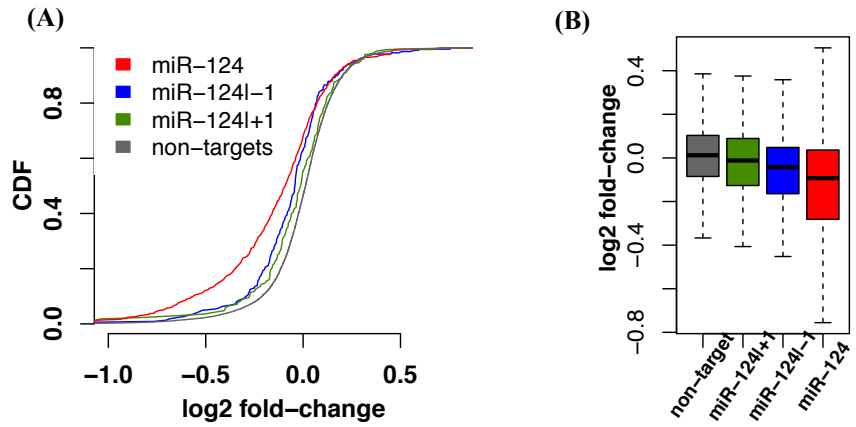


Figure 4.12: Effect of miR-124 and its two 5'-isoforms on target expression based on miR-124 precursor overexpression experiments. Log2 fold-change for exclusive conserved targets of miR-124, miR-124|-1 and miR-124|+ as well as nontargets, after and before miR-124 precursor overexpression in mouse neuroblastoma CAD cells, depicted with cumulative distribution plots in panel A and with boxplot in panel B. The y-axis of panel A shows cumulative distribution function (CDF) of Log2 fold-change distribution.

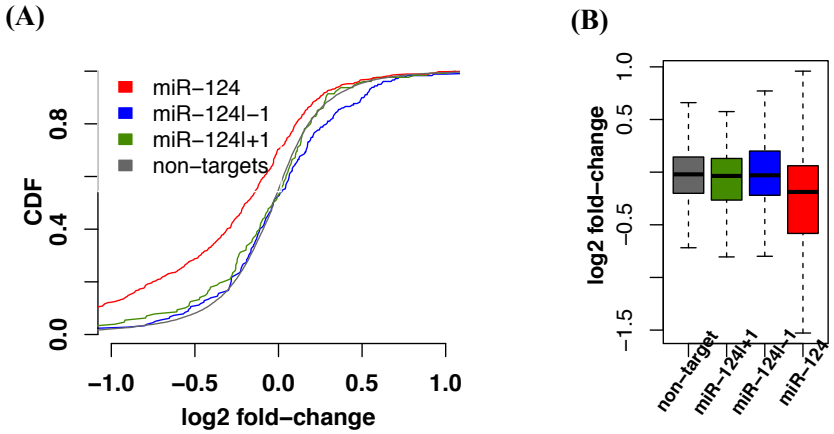


Figure 4.13: Effect of miR-124 and its two 5'-isoforms on target expression based on miR-124 duplex overexpression experiments. Log2 fold-change for exclusive conserved targets of miR-124, miR-124|-1 and miR-124|+ as well as nontargets, after and before miR-124 duplex overexpression in Hela cell line, depicted with cumulative distribution plots in panel A and with boxplot in panel B. The y-axis of panel A shows cumulative distribution function (CDF) of Log2 fold-change distribution.

Finally, I examined the regulatory effect of 5'-isoforms on mRNA expression based on the published miR-223 knockout experiment in mouse neutrophils [201] (Section 2.6). By analyzing small RNA sequencing data from mouse neutrophils, one miR-223 5'-isoform, named "miR-223|+1" was identified, which made up 14.6% of annotated miR-223 expression. Comparison of mRNA expression differences of ECTs of miR-223|+1 and nontarget genes before and after miR-223 knockout showed that loss of miR-223|+1 de-repressed ECTs of

miR-223|+1 significantly (Table 4.5, Figure 4.14). Since miRNA knockout experiments were believed to be the gold-standard approach to unraveling *in vivo* functions of miRNAs, this result suggested that 5'-isoforms were active regulators *in vivo*.

Table 4.5: The expression abundance and target gene number for miR223 5'-isoform

miRNA	Mouse neutrophils (TPM)	USP4 (Adjusted p-value)	Exclusive Conserved Targets (ECTs)	Regulatory effect	
				Wilcox rank sum test (p-value)	KS-test (p-value)
miR-223	25061	3.03e-07	274	2.2e-16	2.2e-16
miR-223|+1	3663 (14.6%)	1.96e-03	380	1.5e-06	1.3e-06

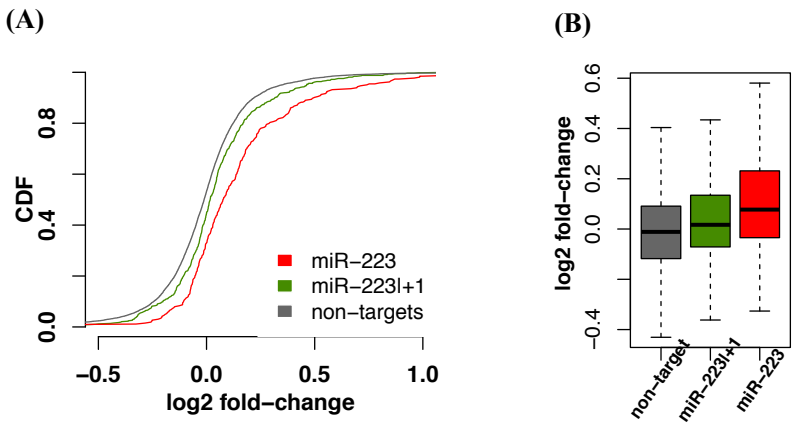


Figure 4.14: Effect of miR-223 and miR-223|+1 on target expression based on miR-223 knockout experiments. Log2 fold-change for exclusive conserved targets of miR-223 and miR-223|+1 as well as nontargets before and after miR-223 knockout in mouse neutrophils, depicted with cumulative distribution plots in panel A and with boxplot in panel B. The y-axis of panel A shows cumulative distribution function (CDF) of Log2 fold-change distribution.

5. Transcriptome Assembly Reveals a Novel Class Bidirectional Promoters Associated with Novel LncRNA and Neuronal Genes

LncRNA remain one of the least characterized types of ncRNA, partially due to its potential heterogeneity. In the study described in this chapter, I attempted to develop and use computational tools to obtain a comprehensive picture of lncRNA types expressed in human prefrontal cortex. As a result, my analysis revealed a specific population of lncRNA expressed from a novel class of bidirectional promoters, displaying unique sequence and epigenetic features that are associated with the expression of neuronal genes.

5.1 Transcriptome assembly in human prefrontal cortex

To comprehensively identify and explore the expression dynamics of novel lncRNAs and novel elements in the known transcripts in the human prefrontal cortex, transcriptome assembly was conducted using strand-specific high-throughput sequencing data collected in the prefrontal cortex of 14 human individuals with an age range from 2 days to 98 years. These data contained an average of 21 million 100nt long reads per sample, with a total of 296 million reads (Table B.2). To avoid the limitations imposed by transcritptome read mapping to the human genome, de novo transcritptome assembly was conducted using the Trinity algorithm [153] (Section 2.7). Of the raw sequence reads, 96% were retained after quality control and subsequently used in the transcript assembly. The assembly resulted in 332,993 transcript contigs with an average length of 1,005 nt and minimum length set to 300 nt. Of these, 307,543 (92.4%) could be unambiguously and uniquely aligned to the human reference genome by using a transcript mapping procedure. Merging transcript contigs that overlapped with each other on the human genome resulted in 92,705 contig clusters. These assembly transcripts covered in total 98,589,683 nt of the human genome. Of them, 61,650,777 nt (64.9%) overlapped with human annotated transcripts based on Ensembl gene annotation [203] (version 64), covering 61% of all annotated exons, while the remaining 36,938,906 nt represented as yet unannotated portions of the human prefrontal cortex transcriptome. Among the unannotated transcripts, 4,123,024 (4.2%) originated from novel elements of annotated genes such as novel exons and novel exon extensions; 3,877,147nt (3.6%) from antisense strand of annotated genes; and 28,937,736 nt (29.7%) from novel intergenic transcripts (Figure 5.1A). Accordingly, of the 92,705 assembly contig clusters, 51,948 (56%) overlapped with at least one annotated transcript, while the remaining 40,758 (44%) originated from gene antisense and intergenic regions (Figure 5.1B). With respect to transcript expression abundance, annotated transcripts accounted for 81% of total transcriptome expression, novel elements of annotated genes and intergenic transcripts for 9% each, and antisense transcripts for the remaining 1% (Figure 5.1C).

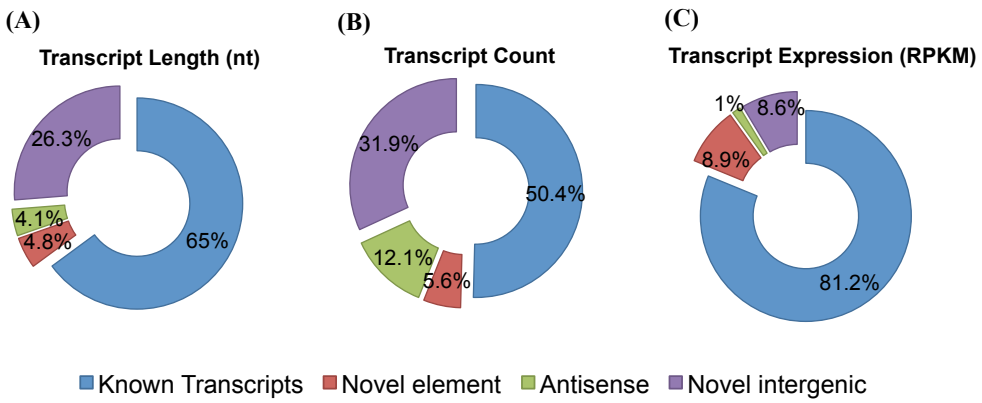


Figure 5.1: Annotated and novel portions of the human prefrontal cortex transcriptome. Panels A, B and C depict the proportion of four transcript types—annotated known transcripts (blue), novel elements of annotated transcripts (red), antisense transcripts (green), and novel intergenic transcripts (purple)—with respect to the total transcript length, transcript count and expression level, respectively.

5.2 Novel elements identification from annotated human transcripts

Among the 51,948 assembly contig clusters that were located within annotated transcripts, 3,699 clusters composing 12,822 transcript contigs contained transcript elements not covered by the existing annotation (Section 2.8). These novel elements included 972 novel internal exons located in 754 protein-coding genes; 926 and 1,211 novel donor and acceptor splice sites containing the most canonical splicing signals (GT-AG) located in 1,687 protein-coding genes; and 1,224 and 4,100 novel 5'UTR and 3'UTR extensions with a length of at least 100nt and located in 1,952 protein-coding genes. Besides protein-coding genes, 267 novel exons, 354 splice boundaries and 1,106 5'/3' terminal exon extensions were found in annotated pseudogenes, lncRNAs and processed transcripts annotated from 1,531 contig clusters (Figure 5.2).

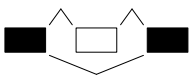
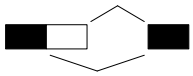
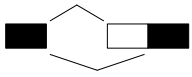
Scheme	Category	Total	Coding gene	Pseudo-gene	lncRNA	Processed transcript
	New exon	1239	972	34	69	164
	New donor site	1100	926	13	87	74
	New acceptor site	1301	1121	18	95	67
	5' extension	1726	1224	25	287	190
	3' extension	4704	4100	50	290	264

Figure 5.2: Categories of novel elements from annotated transcripts in the human prefrontal

cortex transcriptome. Black boxes indicate annotated exons of annotated transcripts; grey boxes indicate UTR of protein-coding genes and terminal exons of pseudogene, lncRNA and processed transcript; and white boxes represent novel transcript elements corresponding to the "Category" column.

5.3 Identification and property analysis of novel lncRNAs

Among the 92,705 contig clusters identified in the data, 40,758 represented novel contig clusters that had no overlap with genome annotation (Ensembl, version 64). To predict novel long noncoding RNAs (lncRNAs) from novel contig clusters, transcript coding potential was estimated based on two coding potential estimation algorithms, CPC [142] and CPAT [143], which employed distinct approaches to estimate transcript coding potential (Section 2.8). The result showed that more than 99% of novel contig clusters had negative coding potential scores calculated by CPC (Figure 5.3A). Consistent with the prediction of CPC, more than 99% of novel contig clusters were predicted to be noncoding transcripts by CPAT (Figure 5.3B and Figure 5.3C).

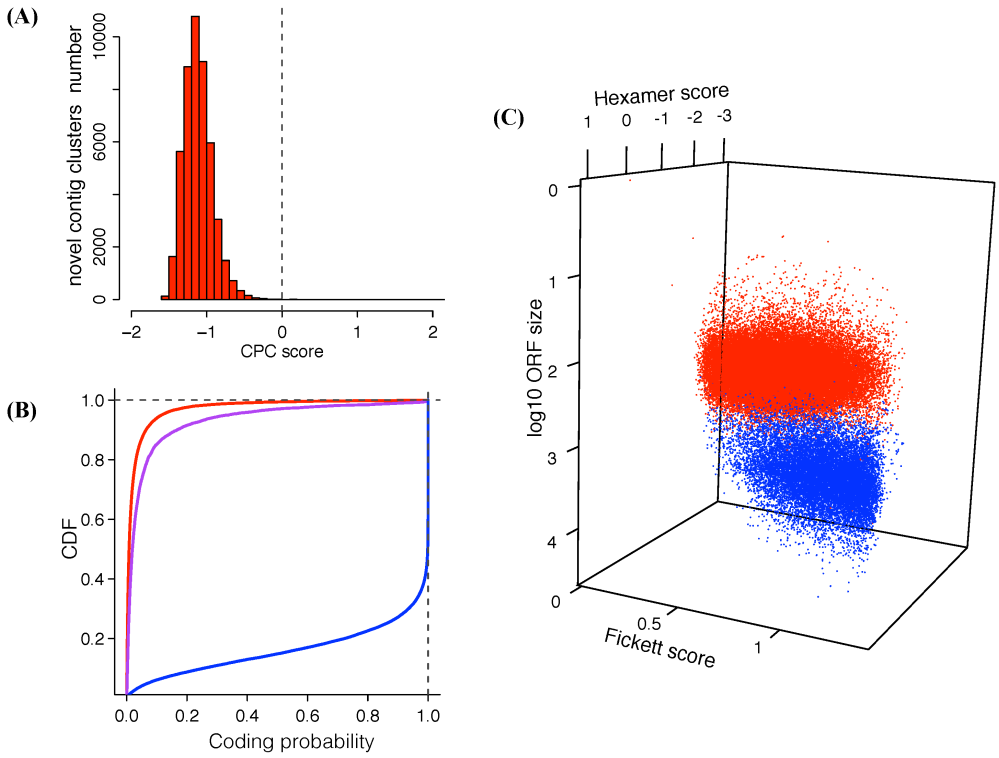


Figure 5.3: Coding potential estimation of novel contig clusters. (A) Distribution of coding potential sores for novel contig clusters using CPC. Negative scores indicate low coding potential. (B) Cumulative distribution function (CDF) of coding probability calculated using CPAT for novel contig clusters (red), annotated lncRNAs (purple) and annotated protein-coding genes (blue). (C) Distributions of Fickett score, Hexamer score and ORF size in log10 scale for novel contig clusters (red points) and annotated coding genes (blue points) in three-dimensional plot.

Notably, novel contig clusters displayed even lower coding probability than annotated lncRNAs (Figure 5.3B), suggesting that novel contig clusters may, in most cases, represent novel lncRNAs or novel lncRNA fragments. Stringently, the intersection of predicted noncoding transcripts from CPC and CPAT were used to define novel lncRNAs, resulting in 38,981 putative novel lncRNAs. The predicted novel lncRNAs displayed all features characteristic of annotated lncRNAs (Section 1.3 and Section 2.9).

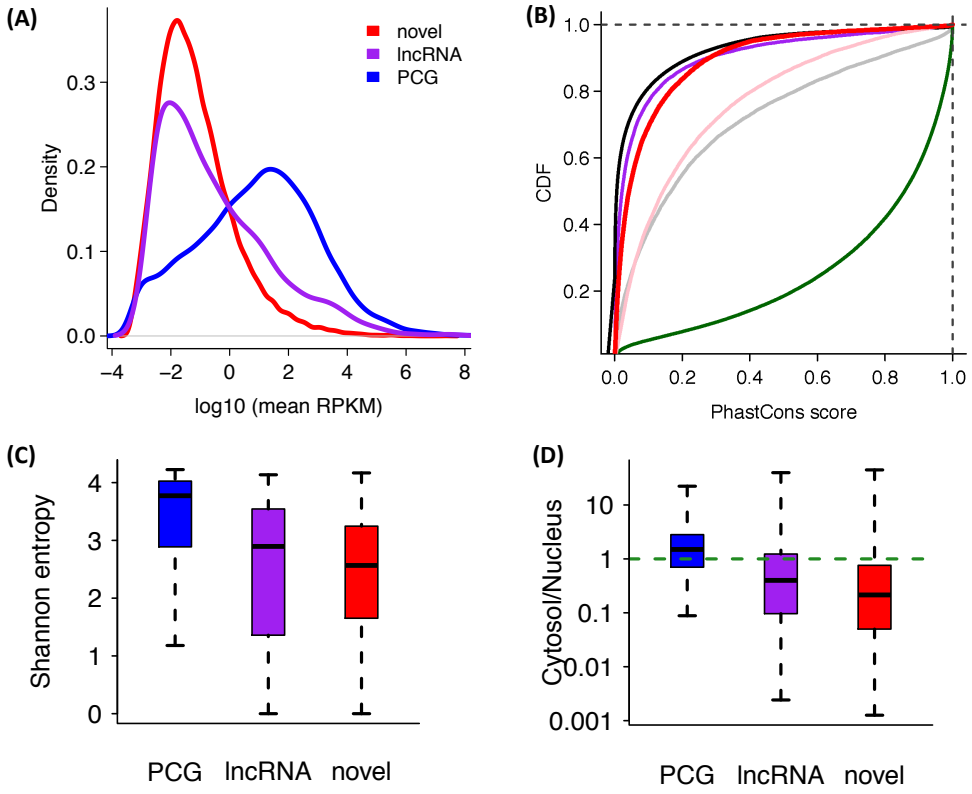


Figure 5.4: Properties of novel lncRNAs. (A) Probability density function of expression abundance across novel lncRNA (novel), annotated lncRNA (lncRNA) and annotated protein-coding genes (PCG). The x-axis displays expression abundance using the mean expression RPKM across 14 human samples in log10 scale. (B) Cumulative distribution of exon sequence conservation estimated using PhastCons scores based on 17 vertebrate species' genome. The colors indicate the random intergenic region (black), annotated lncRNA (purple), novel lncRNA (red), pseudogene (pink), UTR exons (grey) and protein-coding exons (green). (C) Tissue specificity of expression for novel lncRNA (novel), annotated lncRNA (lncRNA) and annotated protein-coding genes (PCG) measured using Shannon entropy based on Human Body Map data. Lower Shannon entropy indicates higher expression tissue specificity. (D) Cellular localization preference (cytosol to nucleus expression level (RPKM) ratio) of novel lncRNA (novel) , annotated lncRNA (lncRNA) and annotated protein-coding genes (PCG).

Specifically, novel lncRNAs were expressed in low expression abundance similar to that of the annotated lncRNAs [15], and both were significantly lower than the expression of annotated protein-coding genes (Wilcoxon rank sum test, p<2.2e-16) (Figure 5.4A). In line

with the low sequence conservation feature of known lncRNAs [15], novel lncRNAs were poorly conserved compared with protein-coding genes at the DNA sequence level among 17 vertebrate species (KS test, p<2.2e-16) (Figure 5.4B). On the other hand, they are significantly more conserved compared to randomly selected intergenic regions (KS test, p<1e-5). In agreement with previous studies reporting the high tissue specificity for lncRNA expression [15], novel lncRNAs and annotated lncRNAs displayed significantly higher expression tissue specificity than annotated protein-coding genes based on Human Body Map data (Wilcoxon rank sum test, p<1e-5) (Figure 5.4C). Furthermore, similarly to known lncRNAs, novel lncRNAs were preferentially localized in the nucleus [15]. Notably, novel lncRNAs showed even higher nucleus localization preference than annotated lncRNAs (Wilcoxon rank sum test, p<2.2e-16) (Figure 5.4D). Novel lncRNAs containing multiple exons also displayed canonical donor and acceptor splice site signals (Figure A.2). Taken together, these features indicate that the predicted novel lncRNAs in this study were mostly authentic lncRNAs or lncRNA fragments.

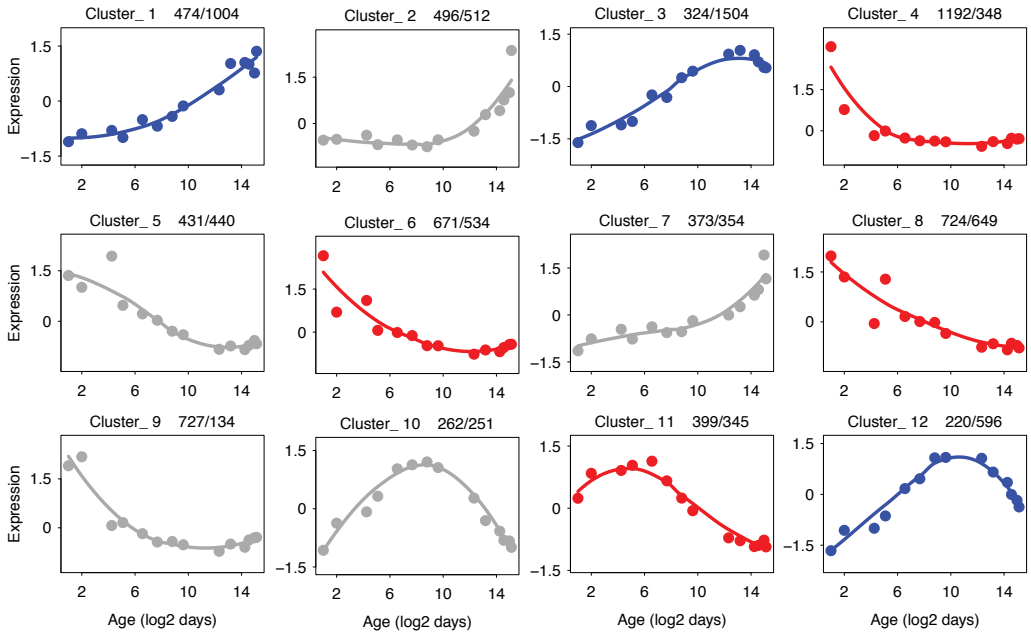


Figure 5.5: The major expression pattern of age-related novel lncRNAs across human postnatal PFC development. The K-means algorithm was used to group expressed age-related novel lncRNAs and protein-coding genes (mean RPKM>=0.1) into 12 clusters. Each panel shows the expression pattern in one of the gene clusters. The x-axis shows the age of individuals on a log2 day scale. The y-axis shows standardized (Z-transformed) expression levels in which each unit indicates one standard deviation difference from the mean. Numbers above each panel show numbers of novel lncRNAs and protein-coding genes in each cluster. Clusters significantly enriched in novel lncRNAs are shown in red, and clusters enriched in protein-coding genes are shown in blue. Clusters showing no significant enrichment are shown in gray.

Since the RNA-Seq data used for transcriptome assembly represents a human prefrontal cortex developmental time series, the temporal expression patterns of novel lncRNAs were

further investigated. Using age-test [212] (Section 2.9), 6,293 novel lncRNAs and 6671 known protein-coding genes displaying significant expression level changes with age were identified (age-test $p<0.01$, $q<0.02$). To explore the temporal expression patterns of these age-related transcripts, we classified their expression profiles into 12 clusters using a k-means algorithm (Figure 5.5) [214]. Interestingly, novel lncRNAs and known protein-coding genes tended to show opposite expression patterns. Specifically, known protein-coding genes were enriched in clusters 1, 3 and 12 (Fisher's exact test, $p<0.0001$ after bonferroni correction, Figure 5.5). All of them showed an expression level increase in development. By contrast, novel lncRNAs were enriched in clusters 4, 6, 8 and 11 and predominantly showed an expression level decrease with advanced age (Fisher's exact test, $p<0.0001$ after bonferroni correction, Figure 5.5). Thus, novel lncRNAs are more highly expressed in young than in aging brains.

5.4 Discovery of a class of novel bidirectional promoter (NBiP)

Previous studies have shown that the majority of the novel transcripts located outside of annotated gene regions, both sense and antisense, were enriched within 10kb from annotated genes, which may represent as yet unannotated extensions of known genes [157]. In this study, there are 26,961 novel lncRNAs located outside annotated gene regions with expression greater than 0.1 RPKM. By analyzing the genomic context of these novel lncRNAs, 14,235 (53%) lncRNAs were located within 4kb from annotated gene boundaries (Simulation test, $p<0.04$, Section 2.9). Based on the DNA strand and relative position with respect to the nearest annotated gene region, these 14,235 novel lncRNAs could be further classified into four categories: upstream-sense (US-lncRNA, 1,323, or 9.3%), downstream-sense (DS-lncRNA, 6,965, or 48.9%), upstream-antisense (UA-lncRNA, 2,964, or 20.7%) and downstream-antisense (DA-lncRNA, 2,983, or 21.1%).

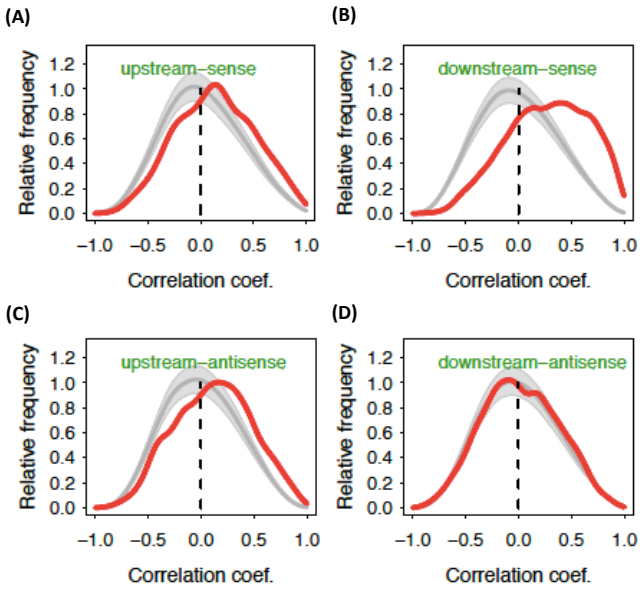


Figure 5.6: Distribution of Pearson correlation coefficients between the expression of protein-coding genes and nearest novel lncRNAs. The red curve shows the Pearson correlation coefficients between the expressions of protein-coding genes and the nearest novel lncRNA:

US-lncRNAs (A), DS-lncRNAs (B), UA-lncRNAs (C) and DA-lncRNAs (D). The grey curves show the average correlation coefficients distribution based on 200 permutations of neighboring novel lncRNAs and protein-coding gene relationships. The grey shaded areas show standard error of the curve estimations.

Significant excess of positive correlations between the expression of novel lncRNAs and the expression of nearby protein-coding genes were detected for the upstream-sense, downstream-sense and upstream-antisense categories (Figure 5.6). No significant correlation signal was found for the downstream-antisense category (Figure 5.6D).

While novel lncRNAs located on the sense strand may still represent potential unidentified 5' and 3' extensions of known genes, transcripts originating from the antisense strand must have an independent origin. Indeed, there is no correlation between the expression of annotated genes and nearby novel DA-lncRNAs. By contrast, a significant excess of positive correlations between annotated genes and novel UA-lncRNAs may indicate shared regulation, presumably at as yet unannotated bidirectional promoters. To investigate this possibility, bidirectional transcription features were analyzed with public human brain tissue deepCAGE data from FAMTOM4 [215] by taking known bidirectional promoters (KBiPs) and unidirectional promoters (UniPs) as positive and negative sets, respectively (Section 2.10). Indeed, a signature of divergent transcription characteristic of bidirectional promoters can be observed for the novel UA-lncRNAs and the corresponding annotated genes (Figure 5.7). The result showed more than five-fold enrichment of the divergent transcription feature from the promoters of novel UA-lncRNA and gene pairs compared to that from UniPs (Fisher's exact test, p<2.2e-16). The divergent transcription characteristic was particularly pronounced for the 273 novel UA-lncRNA and gene pairs that showed a significant positive correlation in the PFC time-series data (Pearson correlation, p<0.05 after Benjamini-Hochberg correction), compared to UniPs (Fisher's exact test, p<0.0001), and was comparable to KBiPs (Figure 5.7). The prominent divergent transcription feature demonstrated novel UA-lncRNA and gene pairs was transcribed from novel class of bidirectional promoters (NBiPs).

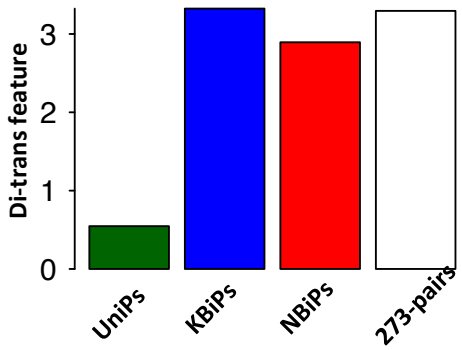


Figure 5.7: Bidirectional transcription feature at different promoter types. The y-axis shows the divergent transcription feature (Di-trans feature) of each promoter type. Di-trans feature was calculated as the ratio between the number of promoters with bi- and uni-directional expression detected using deepCAGE data from brain tissues. Each promoter type is indicated in:

white—novel bidirectional promoters (NBiPs) with significantly positive expression correlation between novel lncRNA and gene expression (273-pairs); red—all NBiPs; blue—known bidirectional promoters (KBiPs); and green—unidirectional promoters (UniPs).

What is the potential function of UA-lncRNA and/or NBiPs? To address this question, Gene Ontology (GO) enrichment analysis was first conducted based on protein-coding genes from 273 novel UA-lncRNA and gene pairs. GO functional analysis revealed a strong and significant enrichment in 21 GO functional terms after redundancy reduction (Section 2.10), including neuronal functions such as "memory," "generation of neurons" and "regulation of synaptic transmission" (hypergeometric test, $p<0.05$ after Benjamini-Hochberg correction, Figure 5.8A). Consistently, the 273 genes are preferentially expressed in neurons, as gauged from H3K4me3 modification data collected in neurons and non-neuronal cells in the human prefrontal cortex (Simulation test, $p<0.00001$, Figure 5.8B, Section 2.10) and neuron-specific gene expression data collected in the mouse brain (Fisher's exact test, $p<0.0001$ after bonferroni correction, Figure 5.8C, Section 2.10). By contrast, protein-coding genes associated with novel lncRNAs from the other three categories did not show any significant functional enrichment. More surprisingly, the protein-coding genes KBiPs, either consisting of two protein-coding genes or a protein-coding gene and known lncRNA pairs, that were expressed in the human prefrontal cortex showed no significant enrichment in neural functions. Instead, these genes were significantly underrepresented in neuronal functions but overrepresented in biological processes related to RNA processing, DNA repair, DNA metabolic process and ribonucleoprotein complex biogenesis (hypergeometric test, $p<0.05$ after Benjamini-Hochberg correction). Similarly, protein-coding genes transcribed from UniPs were not enriched in neuronal functions, but instead in biological processes related to signal transducer activity and receptor activity (hypergeometric test, $p<0.05$ after Benjamini-Hochberg correction). Taken together, these results suggested that NBiPs might represent a separate promoter category that differed from both KBiPs and UniPs and is particular to genes expressed in neurons and/or associated with neuronal functions.

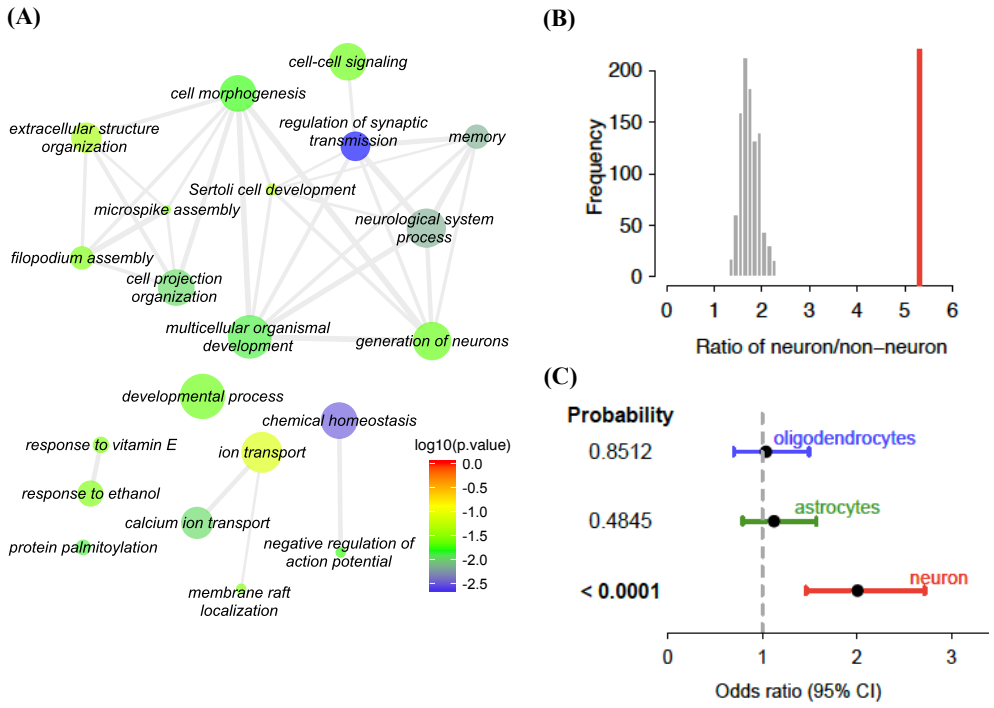


Figure 5.8: Putative function of UA-lncRNA and NBiPs. The potential function of UA-lncRNA and NBiPs were inferred from protein-coding genes from 273 novel UA-lncRNA and gene pairs. (A) GO terms enriched in protein-coding genes from 273 novel UA-lncRNA and gene pairs after redundancy reduction. The node color indicates the GO term's enrichment p-value. The node size is proportional to the GO term's annotated gene number. Dashed rectangle indicates GO terms associated with neuronal functions. (B) Expression specificity of protein-coding genes from 273 novel UA-lncRNA and gene pairs, calculated based on H3K4me3 modification from neurons and non-neural cells of human prefrontal cortex. The red bar represents neuron/non-neuron expression ratio of the 273 genes, and the grey bars represent the ratio distribution by 1,000 permutations of 273 randomly selected expressed genes. (C) Expression specificity of protein-coding genes from 273 novel UA-lncRNA and gene pairs, calculated based on cell type specific expression data from mouse neocortex. The bars show Fisher's exact test odds ratio with 95% confidence interval for enrichment of the 273 genes among mouse orthologs preferentially expressed in oligodendrocytes (blue), astrocytes (green) and neurons (red). The number on the left shows Fisher's exact test p-values.

5.5 Identification of enriched transcription factors in NBiP

The unique functional features of NBiPs prompted me to explore transcription factors that may regulate this promoter type. Comparing transcription factor binding site (TFBSs) density within 2 kb of NBiPs and KBiPs revealed 10 TFBSs that correspond to 11 transcription factors (TFs) enriched in NBiPs and 6 TFBSs corresponding to 8 TFs enriched in KBiPs (Section 2.10, Table 5.1). Notably, with respect to function, TFs enriched in NBiPs were significantly co-cited with the terms "neural" or "neuron" (CoCiter [223], $p<0.01$, Table 5.1). By contrast, TFs enriched in KBiPs showed no such association (CoCiter, $p>0.2$, Table 5.1). Thus, NBiPs may represent an integral part of a regulatory mechanism specific to a set of neuronal genes and involving specific neuron-related TFs. Intriguingly, TFs enriched in

NBiPs included all three methylation-resistant TFs (AP-2 family, EGR family and ZF5) representing three of the top four discriminatory features used to predict the methylation status of CpG islands in the human brain [229], suggesting that NBiPs may display a unique DNA methylation signature.

Table 5.1: Enriched transcription factors in novel and known bidirectional promoters

TFBS ID	TF	p-value	BH adjusted p-value	Odds ratio	Co-cited with term "neural"	Co-cited with term "neuron"
Novel bidirectional promoter (NBiP) enrichment						
V$ETF_Q6	TEAD2	2.39E-18	5.02E-16	1.88		
V$AP2_Q6_01/V$AP2ALPHA_01/V$AP2_Q6	TFAP2A/AP2	3.13E-09	2.38E-07	1.48		
V$HIC1_02	HIC1	3.76E-08	1.57E-06	1.48		
V$KROX_Q6	EGR family (EGR1/2/3/4)	4.48E-08	1.57E-06	1.43	$p = 0.002$	$p = 0.015$
V$ZF5_B	ZFP161/ZF5	8.24E-08	2.47E-06	1.79		
V$LRF_Q2	ZBTB7A	2.98E-07	7.81E-06	1.39		
V$WT1_Q6	WT1	1.62E-06	3.79E-05	1.37		
V$AHRHIF_Q6	AHR	4.26E-05	0.00089506	1.51		
Known bidirectional promoter (KBiP) enrichment						
V$COUP_DR1_Q6	COUP/DR1	0.002709	0.0362	1.37		
V$PPAR_DR1_Q2	PPAR/DR1	0.002754	0.0362	1.40		
V$NFY_Q6_01	NFIC	0.001396	0.0362	1.61		
V$SREBP1_01	SREBF1	0.000268	0.0188	1.35	$p = 0.502$	$p = 0.296$
V$TEL2_Q6	ETV7	0.000191	0.0188	1.30		
V$DR4_Q2	TNFRSF10A	0.000158	0.0188	1.57		

5.6 Analysis of DNA sequence and epigenetic features of NBiP

The unique functional and regulatory features of NBiPs might suggest a specific sequence and epigenetic signature for this promoter type. To investigate this, NBiPs were compared with UniPs and KBiPs with respect to common sequence (GC content, regulatory potential and sequence conservation) and epigenetic features (H3K4me3 modification profile and DNA methylation status) (Section 2.11). The result showed that compared with UniPs and KBiPs, NBiPs show significant differences with respect to all common sequence and epigenetic features. Specifically, NBiPs have a higher GC content and higher regulatory potential, measured as a Regulatory Potential (RP) score, than both UniPs and KBiPs (KS test, $p<0.0001$; Figure 5.9A and Figure 5.9B). Furthermore, NBiPs are more conserved at the DNA sequence level than KBiPs (KS test, $p<0.001$), while both types of bidirectional promoters are more conserved than UniPs (KS test, $p<0.0001$, Figure 5.9C). H3K4me3 modification density, measured in human PFC neurons, is higher at NBiPs and KBiPs than at UniPs, indicating promoter activity (Wilcoxon test, $p<0.00001$, Figure 5.9D). Notably, besides the overall H3K4me3 modification density differences, the shape of H3K4me3 modification profiles differs among the three promoter types (Figure 5.9E). Specifically, UniPs show starkly asymmetric H3K4me3 modification profiles with much of the modification density located downstream of the protein-coding gene transcriptional start site (TSS). By contrast, the shape of H3K4me3 modification profile is more symmetric relative to the TSS for both NBiPs and KBiPs, with the most symmetric signatures observed at KBiPs. This difference in the H3K4me3 modification signature could be reproduced using another H3K4me3 modification dataset obtained from rhesus macaque PFC samples (Figure A.3). By contrast, the input control showed no significant differences in shape and density for H3K4me3 modification profiles among the three promoter types (Figure A.3). Lastly, DNA methylation levels measured in the human PFC also differed among the three promoter types: DNA methylation levels are high at UniPs, intermediate at KBiPs and lowest at NBiPs (Wilcoxon test, $p<0.0001$; Figure 5.9F). The low DNA methylation status was in line with the previous finding that the three methylation-resistant TFs [229] were enriched in NBiPs. Taken together, in the brain, NBiPs formed by UA-lncRNA represent a distinct type of bidirectional promoter with characteristic structural and regulatory properties compared to both KBiPs and UniPs.

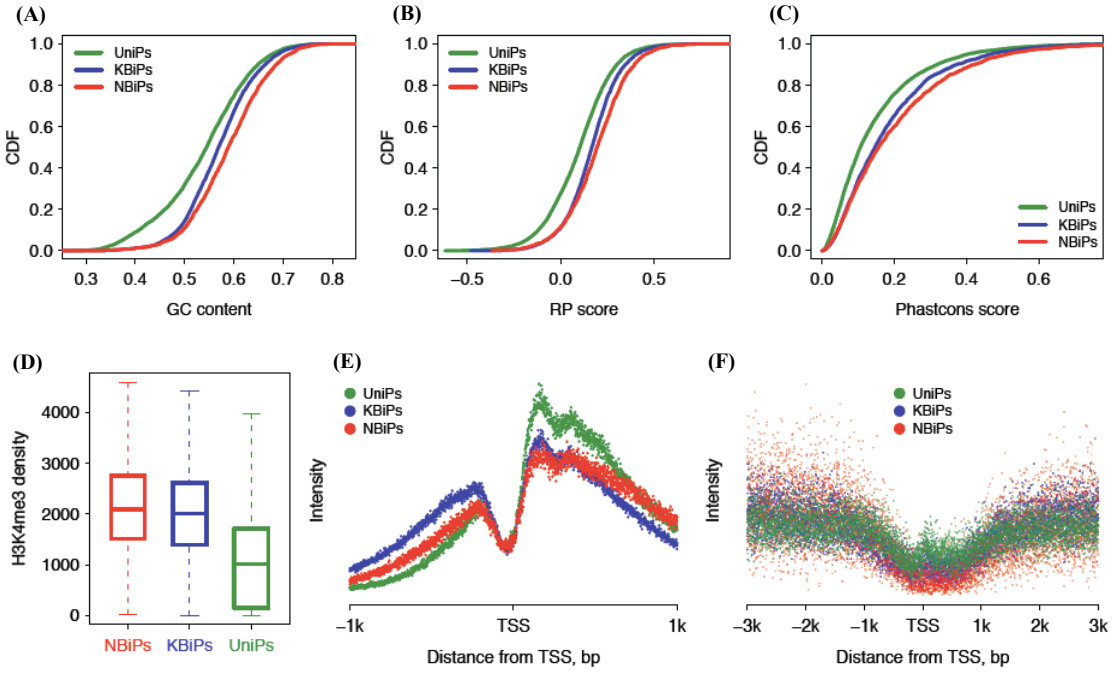


Figure 5.9: Sequence and epigenetic features of different promoter types. (A), (B) and (C) The cumulative distributions of GC content, regulatory potential and sequence conservation for the three promoter types: UniPs (green), KBiPs (blue) and NBiPs (red). All measurements are based on a 2 kb region surrounding the TSS. Promoter sequence conservation was calculated using PhastCons scores based on 17 vertebrate species' genomes. Promoter regulatory potential was calculated using Regulatory Potential (RP) scores (see Methods). (D), (E) The density (panel d) and the shape (panel e) of H3K4me3 modification profiles at each of the three promoter types. (F) DNA methylation profile at each of the three promoter types.

Transcripts expressed in the PFC, and more generally in the brain, are characterized by extended 3'UTR regions [230]. This phenomenon may in part be explained by the low expression of PABPN1, a gene recently shown to play a role in transcript processing in brain tissue [231, 232] (Figure 5.10A). Intriguingly, by reanalyzing data from [233], I found that the expression of novel lncRNAs originating from NBiPs was starkly increased in a PABPN1 knockdown experiment (Section 2.12). Furthermore, this expression increase was significantly greater than for other lncRNA types (Figure 5.10B and Figure 5.10C). This indicates that the strong expression of lncRNAs originating from NBiPs in the human PFC could be related to this general transcript processing mechanism.

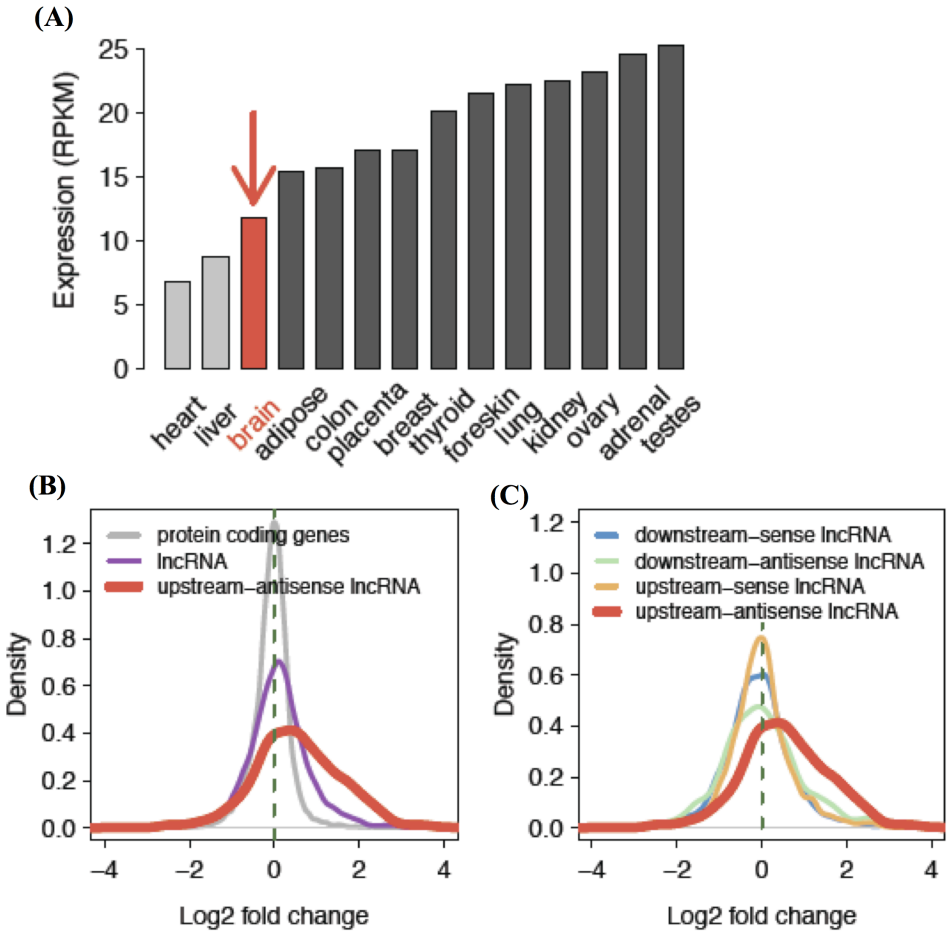


Figure 5.10: PABPN1 knockdown influence on different transcript categories. (A) PABPN1 expression across human tissues calculated using Body Map data. (B), (C) The expression change distribution for different transcript types in a PABPN1 knockdown experiment. The positive values indicate expression upregulation following PABPN1 knockdown.

6. Discussion

6.1 miRNA quantification using deep sequencing data

Appropriate and correct miRNA quantification is fundamental and essential to any miRNA studies. The advent and rapid advancement of high-throughput sequencing technologies has provided the potential to quantify miRNA expression more comprehensively and more accurately, which has also resulted in a growing appreciation of the fact that individual miRNA can be heterogeneous in length [61]. As shown in this study, in the human prefrontal cortex, ~30% of total mapped reads displayed end offsets compared with annotated mature miRNAs annotated in miRBase (version 12). The majority of end offsets occurred at the 3'end, and the end heterogeneity was significantly smaller for the 5'end than the 3'end. Detailed comparison between mapped reads and annotated miRNA mature sequences further indicated that ~28% of expressed miRNAs in human prefrontal cortex were probably misannotated in miRBase (version 12). The complicated nature of the miRNA transcriptome and some erroneously annotated miRNAs in miRBase posed challenges for obtaining accurate estimates of miRNA expression abundance and determining the genuine miRNA mature sequence.

In Chapter 3, I presented an efficient miRNA quantification procedure that can conduct miRNA quantification and mature sequence identification simultaneously (Section 2.1). Compared with a quantification method that quantifies miRNAs using reads with 100% length and sequence match to annotated mature miRNAs [234], the new quantification method exhibited much higher miRNA detection sensitivity, being able to measure ~30% more miRNAs under various miRNA expression detection cutoffs. Furthermore, the expression abundance of miRNAs measured using the new quantification procedure displayed a comparable and even a slightly better correlation between technical replicates. Most importantly, the new miRNA quantification procedure could correct the 5'end annotations for 27 miRNAs and 3'end annotations for 115 miRNAs with expression of more than 10 read counts in the human prefrontal cortex. Obtaining the correct 5'end annotation is crucial since the seed region at the 5'end mostly determines miRNAs' functions. Determining the 3'end is also important since some miRNA quantification assays, such as Q-PCR and miRNA microarray, depend heavily on the integrity of the miRNA sequence, especially at the 3′ end of the sequence, to generate cDNA template or produce labeled probes for miRNA detection. Thus, the new quantification method not only provided better miRNA expression abundance estimation, the correct miRNA mature sequence identified using the new quantification method could also provide important guidance for downstream target prediction analysis and probe design for miRNA microarray and Q-PCR assays.

Some miRNA quantification methods have taken the miRNA ends heterogeneity into consideration for miRNA quantification, including the method that used the sum copy of mapped reads that have minimal 50% overlap [235] or just overlap with annotated miRNAs [236, 237] and the method that used the sum copy of mapped reads falling into a certain interval upstream and downstream of the annotated miRNAs [238, 239]. The potential

drawback of these methods is that the expression abundance of a given miRNA might be estimated using the reads with different seed regions, which would complicate the downstream miRNA target identification that is mainly based on seed sequence identity. By contrast, the new quantification method, taking the knowledge of miRNA targeting mechanism into consideration, first determined the major sequence as the most probable miRNA mature sequence and only used reads that have the same 5'end position as the identified mature sequence for miRNA quantification. Therefore, miRNA target prediction can be conducted using unambiguous and correct miRNA mature sequences. It is also worth noting that the vast majority (>90%) of reads overlapped with annotated mature miRNA have been used for miRNA quantification based on the new miRNA quantification method. Taken together, the new miRNA quantification procedure developed in chapter 3 allows for more comprehensive and accurate miRNA expression abundance estimation.

To resolve substantial reads with 3'end offsets for miRNA quantification, the new miRNA quantification procedure grouped the reads with the same 5' position identity and ignored the 3'end heterogeneity. This quantification strategy is based on the postulation that miRNAs regulate their target genes mainly through miRNA seed region at the 5'end, which has been supported by previous experiments and evolutionary conservation signatures of the miRNA seed region itself and seed matches in the 3'UTRs across species (Section 1.2). Nevertheless, one class of target sites termed 3'-compensatory sites have been reported. The 3'-compensatory sites have a mismatch or wobble in the seed region of the miRNA, but have long stretches of base pairing to the 3' end of the miRNA to make up for the weak binding at the 5' seed. Although 3'-compensatory sites were estimated to represent less than 2% of all preferentially conserved sites detected, it should be noted that the new miRNA quantification procedure is not appropriate for estimating the expression level of miRNAs that exert target base pairing mainly through 3'-compensatory sites. Further advancement of our understanding of miRNA target regulation mode might provide more rules to resolve miRNA end heterogeneity for miRNA quantification.

To quantify miRNA expression precisely, the strategies of 3' adaptor trimming and reads mapping should also be considered carefully. Due to the intrinsic character of Illumina sequencing technology, sequencing errors will accumulate at a much higher rate at the reads 3' end. Thus, a certain number of mismatches should be allowed for 3'adaptor removal to trim the 3'adaptor more completely. Indeed, ~29% of trimmed reads were lost when requiring a perfect match for 3'adaptor removal in the human prefrontal cortex samples. In this study, the number of allowed mismatches was determined empirically by checking distribution of the number of mismatches from the trimmed adaptor sequences. More sophisticated methods might contribute to better 3'adaptor removal. The reads mapping strategy is another consideration for better miRNA quantification. Due to the high sequence similarity of mature sequences within a miRNA family, only perfectly mapped reads were retained in this study to reduce the potential of reads across mapping between miRNA members. Considering that a set of highly expressed miRNAs (e.g., miR-124) has multiple loci on the genome, both unique and multiple mapped were allowed for miRNA quantification. Allowing multiple mapped reads is crucial for correct quantification of duplicated miRNAs. A previous study measured miRNA expression in fly only based on uniquely mapped reads [235], which erroneously quantified miRNA with multiple loci on the genome. Allowing multiple mapped reads also

posed challenges for correct quantification. The new quantification procedure resolved multiple mapped reads by merging reads from all genomic loci producing the same mature miRNA, which could avoid double counting reads for quantification of duplicated miRNAs. Requiring only perfectly mapped reads led to a loss of reads containing sequencing errors for miRNA quantitation. However, this problem can be relieved with the rapid advancement and improvement of sequencing technology.

6.2 Hidden layer of miRNA transcriptome: miRNA 5'-isoforms

Deep sequencing generated millions of small RNA sequencing reads from a given sample profiling and provided the framework for exploring the miRNA transcriptome complexity. It is surprising that miRNAs display heterogeneous ends despite their short length. Little is known about their authenticity and functionality, except for their existence based on the measurements from deep sequencing, cloning and northern blotting. In Chapter 4, a systematic analysis of miRNA 5'isoform was conducted in the human prefrontal cortex. By systematically analyzing small RNA sequencing data from human prefrontal cortex, 66 moderately expressed miRNA 5'-isoforms were identified, representing ~20% of total moderately expressed miRNAs. Importantly, all 66 5'-isoforms were conserved between human and mouse at their mature sequence seed regions, and ~90% of them were associated with AGO2 proteins in human brains and could be readily detected in the mouse brain, indicating that the majority of identified 5'-isoforms are bona fide miRNAs. The results also reflected a well-conserved 5'-isoforms biogenesis mechanism during evolution. In addition, the seed regions of these 66 5'-isoforms were novel compared with known human miRNAs, suggesting that 5'isoforms might be able to regulate a distinct set of target genes. However, authentic miRNAs are not equivalent to functional miRNAs. The essential question of current 5'-isoform study is how many of them are functional. In Chapter 4, I presented one comparative approach to estimating the functionality of individual conserved 5'-isoforms that was based on the co-evolution of conserved miRNAs and corresponding target sites (Section 1.2 and Section 2.5). Detailed performance evaluation showed that the USP4 method outperformed the other proposed methods, highlighting the importance of choosing appropriate control sets for predicting functional conserved miRNAs. The USP4 method performed better than the SSP method that had been used for predicting functional miRNA* sequence in fly. The performance improvement of the USP4 method might be largely contributed by incorporating a dinucleotide-shuffling feature into 3'UTR alignment control set construction because the USP3 method differing from USP4 only on dinucleotide-shuffling feature has similar prediction performance as the SSP method. A dinucleotide-shuffling feature cannot be incorporated into the SSP method because the miRNA seed region is too short to generate enough heptamer controls with the same dinucleotide composition for estimating background seed match conservation. Nevertheless, the great agreement for predicted functional conserved miRNAs between USP4 and SSP demonstrated the high prediction accuracy of USP4. The comparison of the expression levels between functional and nonfunctional conserved miRNAs predicted using USP4 further suggested that the USP4 method could capture the majority of highly expressed miRNAs in different human tissues and cell lines.

The good prediction performance of UPS4 allows for identification of reliable functional

conserved 5'-isoforms. Twenty-six functional 5'-isoforms were predicted in the human prefrontal cortex, including miR-101|-1, miR-142-3p|+1 and mir-9|+1, which have been shown to be functional miRNAs before. Notably, several 5'-isoforms from brain-associated miRNAs were also among the predicted functional 5'-isoforms, including miR-124|+1, miR-124|-1, miR-9*|+1 and miR-9|+2. Importantly, the function of miR-124|-1 can be confirmed based on miR-124 pri-miRNA transfection experiment, suggesting that 5'-isoforms with relatively high expression abundance were sufficient to confer additional mRNA targeting compared with annotated canonical miRNAs. The significant regulatory effect of miR-223|+1 on mRNA expression based on published miR-223 knockout experiment in mouse neutrophils further indicated that 5'-isoforms might be active regulators *in vivo*. Taken together, multiple lines of evidence indicated that these 26 functional 5'-isoforms predicted by the USP4 method might be as important as known miRNAs in regulating their target genes. Previous studies have shown that miR-9/9* and miR-124 can control multiple genes regulating neuronal differentiation and function. Strikingly, the induction of miR-9/9* and miR-124 in human fibroblasts can induce their conversion into neurons. Intriguingly, multiple predicted functional 5'-isoforms have been identified from miR-9/9* and miR-124 in this study. It would be interesting to analyze the potential contribution of 5'-isoforms in neural fate determination.

One potential caveat of USP4 is the quality of 3'UTR alignments. In the current study, the 3'UTR alignments were extracted from human and mouse whole-genome alignments based on human 3'UTR annotation from Refseq. However, no systematic analysis had been conducted to evaluate the quality of human 3'UTR annotation. Furthermore, whether the aligned sequences of mouse were from mouse 3'UTR regions was not determined. It is possible that in some cases, the homologous regions of human 3'UTRs in the alignment were not authentic 3'UTR in mouse, which leads to a wrong estimation of heptamer conservation. In addition, the incompleteness of genome sequence and potential errors in whole-genome alignment would also complicate the miRNA seed match conservation estimation. Nevertheless, with the accumulation of polyA+ RNA-Seq data and new experimental protocols specifying for mRNA 3'end identification, as well as the improvements of genome assembly and alignment algorithm, the 3'UTR alignments are expected to become more complete and accurate in the future. The performance of USP4 might be improved accordingly.

Future improvements of the USP4 method include employing multiple-genome alignment to estimate seed match conservation. In the current study, the USP4 method used human-mouse 3'UTR alignment to measure human miRNA seed match conservation. Target sites have been considered conserved if they were retained at mouse orthologous locations and considered nonconserved if they were missing or have changed in mouse. Such site conservation classification is prone to be affected by the imperfections of mouse genome sequencing and assembly quality, 3'UTR alignment quality and mouse-specific target site gain and loss. Replacing two-way alignment with multiple alignment incorporating 3'UTRs from rat, dog and/or chicken would make the estimation of seed match conservation more accurate. To estimate seed match conservation based on 3'UTR multiple-alignment, one solution is use ad hoc criteria. e.g., conserved target sites were required to be conserved in at least three out of four species in the multiple alignment. Such a solution has proved to be productive for

predicting conserved miRNA target genes [117]. Another possible solution is using Branch Length Score (BLS) [240] to measure target site conservation. The seed match conservation estimation would benefit from BLS measurement since BLS has proven to be more robust against to local alignment inaccuracies, gaps and target site gain and loss. BLS also accounts for different divergence times between species [240]. One difficulty of BLS solution is that BLS calculation using DNAML (DNA Maximum Likelihood program) [241] requires substantial computational resources. Determining BLS cutoff for classifying conserved target site is also much more complicated than using ad hoc criteria.

What is the potential biological significance of 5'-isoforms? Considering the high overlapping nature of the seed region between most 5'-isoforms and their cognate miRNAs, a 5'-isoform might function to reduce the off-target effect for a group of common targets of its cognate canonical miRNA. Supporting this, it has been cogently argued that miR-10|-1 and miR-10 can co-target a large group of common target genes and thus greatly reduced the possible off-target effect of miR-10 [242]. The 5'-isoform can confer additional mRNA targeting to expand previous target repertoire [242], which has been confirmed in this study (such as miR-124|-1) and also in previous studies. Importantly, the existence of 5'-isoforms and corresponding expanded targeting may provide the framework for diversifying the previously established miRNAs. Such function diversification may underlie the previously observed "seed shift" for paralogous miRNAs within a miRNA family, e.g., seed shifting for miRNAs in miR-2 and miR-281 families in *D. melanogaster* [243, 244]. In this case, evolutionary pressure may specifically fix different miRNA isoforms of paralogous miRNAs as major miRNAs to regulate specific target genes. Following the same speculation, "seed shift" may also occur for the orthologous miRNA between species. Indeed, orthologous sequence comparison result suggested that ~13% of conserved miRNAs between *D. melanogaster* and the red flour beetle *Tribolium castaneum* have undergone seed shifting [245]. Perhaps the most prominent seed shift example is the most ancient metazoan miRNA, miR-100, which has a one-base shift to the 5′ end of its mature miRNA in the sea anemone *Nematostella vectensis* compared to all bilaterians [131]. Analyzing the evolution of 5'-isoforms may provide additional insights into their origins and biological functions.

Currently, the molecular mechanism underlying 5'-isoforms biogenesis is still elusive. Although the variation in 5'end processing by Drosha and/or Dicer cutting may produce most 5'-isoforms, a recent study showed that different mRNA isoforms of the Dicer partner protein, loquacious (loqs), was able to determine the mature sequence 5'end identity and length on miR-307a precursor, which finally led to generation of distinct 5'-isoforms from different tissues of *Drosophila*[91]. With the accumulation of public small RNA deep sequencing data, it would be interesting to compare miRNA 5'-isoforms between different tissues and different species comprehensively, which would not only generate more comprehensive 5'-isoforms repertories but also help to shed light on the functions and biogenesis mechanisms of 5'-isoforms. Considering the observations of seed shifting across tissues and species as well as the regulatory potential of 5'-isoforms, current miRNA annotation rules should be revised by incorporating 5'-isoforms with relatively high expression abundance, instead of only selecting one major isoform for annotation as genuine miRNAs.

6.3 miRNA interspecies comparisons

One of the indispensable steps in interspecies miRNA comparison is to find the orthologous relationship of miRNA genes across species. The performance of downstream miRNA expression comparison analysis between species largely depends on the quality of the identified orthologous miRNA pairs, which is a challenging problem due to the short length of miRNA mature sequences. In addition, miRNAs usually have high duplication rates, and many miRNA sequnces only differ one or two nucleotide in their mature sequences within one miRNA family. This further complicates the ortholog finding and orthologous relationship delineation. In Chapter 3, I presented one miRNA orthologous gene prediction procedure (MOP) to identify human miRNA orthologs in chimpanzee and macaque genomes and to construct miRNA orthologous relationship across these three species (Section 2.2). MOP used a two-step strategy to identify reliable miRNA mature orthologous genes in chimpanzee and macaque. In total, MOP successfully predicted 796 and 752 mature miRNAs in chimpanzee and macaque genomes, corresponding to 92% and 87% of human annotated mature miRNAs, respectively. The identified miRNA orthologs covered more than 97% of annotated miRNA of chimpanzee and macaque in miRBase (version 12). Moreover, the extremely high expression correlation between replicates for chimpanzee and macaque samples and gradually increased miRNA expression divergence followed the phylogenetic relationship among three species, further supporting the validity of MOP. The identified miRNAs in chimpanzee and rhesus macaque greatly expanded the miRNA annotation in these two primate species. Rhesus macaque is an important model species for studying human physiology and pathology, so the predicted miRNAs provide valuable resources for further studying the function of individual miRNAs in rhesus macaque.

The identified miRNA orthologs in chimpanzee and rhesus macaque allowed us to estimate miRNA expression divergence between human and other two primates based on small RNA deep sequencing data in prefrontal cortex. Despite high sequence conservation, up to ~11% of the 325 expressed miRNAs diverged significantly between human and chimpanzee, as did up to ~31% between human and macaque. The vast majority of these differences were also found in cerebellum. Notably, the differentially expressed miRNAs with human-specific high expression signature included several brain-specific or brain-associated miRNAs that have been shown to take part in neuronal-related functions and processes. They include miR-184, which is involved in regulation of neural stem cell proliferation and differentiation [246], miR-7, which protects neurons from cell death through downregulation of mTOR signaling [247], and miR-34c-5p, one member of miR-34 family that influences brain aging and neurodegenerative processes [248]. It should be noted that on the DNA sequence level, these miRNAs tend to be conserved: miR-184 mature miRNA sequence is evolutionarily conserved from insects to humans, with only one nucleotide different at 3'end of mature sequence, while miR-7 and miR-34c-5p are classified as broadly conserved. High sequence conservation indicates the functional importance of these miRNAs and shows that expression divergence on the human evolutionary lineage is unlikely to be caused by lack of a selection constraint. Importantly, a significant inverse relationship between human-chimpanzee miRNA expression divergence and expression divergence of the predicted target genes was observed at both mRNA and protein levels. This indicates that miRNA expression divergence plays an

important role in shaping gene expression divergence among species. Further studies are needed to evaluate the functional significance of the miRNA-driven transcriptome changes.

To identify human orthologs in chimpanzee and other primate species, Brameier et al. and Baev et al. used one-way BLAST or one-way BLAT to locate putative miRNA precursor orthologs and aligned human mature miRNAs against predicted precursor sequences with BLAST to obtain mature orthologs [249, 250]. One of the disadvantages of these computational methods is that only local alignment algorithms were considered for precursor ortholog prediction. Although BLAST or BLAT methods were able to find the best local alignment regions and have high sensitivity, ortholog searching based on local alignment algorithms may produce a high number of false positives, especially for miRNA evolved from or overlapped with repeat regions. In addition, the one-way strategy also complicates the orthologous relationship delineation. Compared with previous homolog-based miRNA identification methods, MOP used revised strategies for both miRNA precursor and mature sequence identification. MOP determined the best precursor orthologs by utilizing the combination of reciprocal BLAT, BLAST and liftOver. The latter, which is based on whole genome alignment, can utilize the syntenic homologous blocks to predict miRNA precursors. Since an orthologous finding using liftOver is based on global alignment, some problems due to local alignment algorithm could be resolved, which makes the identification of miRNA precursors more reliable. The requirement of supporting by local alignment (BLAST and BLAT) and global alignment algorithms (liftOver) as well as precursor ortholog length criteria renders the high prediction accuracy of MOP. Incorporation of reciprocal strategy in MOP could not only enhance the ortholog prediction accuracy but also determine the orthologous relationship, which is crucial for downstream miRNA expression comparison across species. To obtain mature orthologs, MOP constructed precursor alignment with human precursors and predicted precursor orthologs of other primates using ClustalW2, and furthermore extracted mature orthologs based on human mature miRNA annotation. Due to the short length of miRNA mature sequence, the BLAST method is less sensitive for locating human mature sequence from the predicted precursor orthologs when substitutions exist between miRNA ortholog pairs. Consequently, Brameier et al. conducted additional manual checks when human mature sequences failed to produce a BLAST hit or the alignment was incomplete [249]. Due to the local alignment algorithm of BLAST, the mature sequence boundaries were also hard to locate precisely when substitutions happened at or near the miRNA sequence ends. By contrast, extracting miRNA mature orthologs based on precursor alignments with a global alignment algorithm such as ClustalW2 allowed for obtaining more accurate miRNA mature orthologs, especially for defining the mature sequence boundaries. It should be noted that although MOP used more stringent criteria for miRNA prediction, prediction sensitivity was not sacrificed. The vast majority (>95%) of annotated known miRNAs in chimpanzee and rhesus macaque could be identified by MOP.

As a general method for miRNA orthologous gene prediction, MOP can be easily applied to other species to obtain putative miRNA orthologs, provided the corresponding genome sequences were available. The homolog-based miRNA finding provided an important strategy for miRNA identification, especially for the miRNAs that were expressed at specific tissue, cell-type or developmental stages or expressed in a low level. The homolog-based miRNA finding coupled with high-throughput sequencing can be considered a good complement to de

novo miRNA prediction methods. Indeed, this strategy has been used to predict miRNAs in the species of both animals and plants [251, 252]. The orthologs predicted by MOP are also useful for species-specific miRNA identification. Recently, by applying MOP with human miRNA annotation in the genomes of 11 species (chimpanzee, gorilla, orangutan, rhesus macaque, marmoset, mouse, rat, dog, cow, opossum and chicken), I successfully identified 22 human-specific miRNAs, including 10 mature human miRNAs with no detectable orthologs in any of the 11 species and 12 mature miRNAs with sequence changes in the seed region that took place in the human lineage after the split with chimpanzee [37]. Further expression pattern and function analysis led to discovery a bona fide human-specific miRNA miR-941 that effected target genes involved in neurotransmitter signaling, hedgehog- and insulin-signaling pathways. miRNA are known for their rapid evolutionary dynamics, with dozens of novel miRNAs emerging in the genomes of individual species of nematode and flies [253, 254]. Novel miRNA emergence could affect expression of hundreds of genes, thus accelerating species-specific gene expression evolution. On the basis of MOP, it would be interesting to obtain species-specific miRNAs in more species, comparing their origins and exploring their putative functions.

6.4 Novel lncRNAs and NBips in the human prefrontal cortex

Detailed classification of lncRNA based on genetic, epigenetic and expression data might be useful in elucidating functional properties of particular lncRNA populations. By taking advantage of strand-specific RNA-seq data collected at different stages of postnatal human brain development, as well as publicly available genetic and epigenetic data, I carried out detailed characterization of annotated lncRNA, as well as novel ones I identified using these data. Several interesting observations have emerged.

First, despite the substantial efforts made toward human brain transcriptome characterization in previous decades, more than 40% of PFC transcripts reconstructed in the current study represent novel transcriptome elements. These elements include novel exons and exon extensions of annotated protein-coding, pseugogene and lncRNA genes as well as plenty of novel lncRNAs. The novel lncRNA identified in this study displayed all the canonical features of known lncRNAs, such as low coding potential, low expression abundance, high tissue expression specificity and high nucleus localization preference. One potentially interesting feature of novel lncRNAs is their temporal expression pattern. Compared with annotated protein-coding genes, age-related novel lncRNAs were preferentially expressed in the young rather than in the aging brain. Previous study has showed that newly evolved protein-coding genes (mostly primate-specific or human-specific) were expressed higher in the human fetal and postnatal young brains compared to other organs or to later life stages [255]. It is appealing to speculate that young genes, independent of their coding potential, may be important in early brain development.

Second, while most of the lncRNAs expressed in the prefrontal cortex (>50%) localize in close proximity (<4 kb) to known protein-coding genes, one fraction of these transcripts, the lncRNAs located upstream of the protein-coding genes on the antisense strand (UA-lncRNA), particularly stands out. Specifically, these transcripts 1) show a significantly positive correlation with the expression of the upstream protein-coding genes; 2) originate from a

specific class of bidirectional promoters showing unique sequence and epigenetic features; 3) are highly enriched upstream of genes that are expressed in neurons and involved in neuronal functions; and 4) are enriched in TFs shown to be linked to neurons.

Bidirectional promoters are a common feature of the human genome and have also been described in the mouse and other species [256, 257]. In humans, 10% of protein-coding genes were annotated to originate from bidirectional promoters [256]. Remarkably, genes preferentially expressed in the brain and involved in neural functions were depleted at these known bidirectional promoters [258]. This result was further confirmed in the current study. By contrast, novel bidirectional promoters showing divergent transcription of novel and potentially brain-specific lncRNAs are highly enriched in neuronal genes. The novel bidirectional promoters identified in the current study are also distinct from both known bidirectional promoters and unidirectional promoters with respect to many aspects of sequence composition and epigenetic features, including H3K4me3 chromatin modifications and DNA methylation. Thus, they may represent a novel promoter type specifically associated with the expression of neuronal genes and regulated by a specific set of TFs. Intriguingly, TFs showing significant association with this promoter type include all three methylation-resistant TFs (AP-2 family, EGR family and ZF5), representing three of the top four discriminatory features used to predict methylation status of CpG islands in the human brain [229]. This fact may explain the unique DNA methylation signature of the NBiPs observed in the current study.

Expression of lncRNAs from bidirectional promoters has been previously shown in many human cell types, including human embryonic stem cells (hESCs) where >60% promoters might be bidirectional and associated with divergent lncRNAs [259]. Notably, even though we find no significant overlap between bidirectional promoters described in hESCs and NBiPs identified in this study, in both cases, expression of protein-coding genes correlated positively with expression of divergent lncRNAs. It is, however, unclear whether this positive correlation represents a regulatory effect of lncRNAs or a passive consequence of the transcriptional activation of the divergent protein-coding genes. Most human promoters bind polymerase complexes in a bidirectional manner and are therefore capable of initiating transcription in both directions [257]. More recently, this knowledge has been challenged. Duttke et al. show that the majority of human promoters are unidirectional and the ones capable of divergent transcription contain their own cognate reverse-directed core promoters [260]. In the current study, I cannot exclude that the presence of lncRNAs at the novel type of bidirectional promoters identified in this study may represent a passive byproduct of neuronal gene transcription from this specific promoter type. Further studies are needed to evaluate the functional significance of UA-lncRNAs.

Bibliography

1. Jacob F, Monod J: **Genetic regulatory mechanisms in the synthesis of proteins.** *J Mol Biol* 1961, **3:**318-356.

2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.

3. Consortium EP, Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, et al: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447:**799-816.

4. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al: **The transcriptional landscape of the mammalian genome.** *Science* 2005, **309:**1559-1563.

5. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, et al: **Antisense transcription in the mammalian transcriptome.** *Science* 2005, **309:**1564-1566.

6. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al: **Landscape of transcription in human cells.** *Nature* 2012, **489:**101-108.

7. Consortium EP: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489:**57-74.

8. Consortium EP: **The ENCODE (ENCyclopedia Of DNA Elements) Project.** *Science* 2004, **306:**636-640.

9. Lee RC, Feinbaum RL, Ambros V: **The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.** *Cell* 1993, **75:**843-854.

10. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, et al: **Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals.** *Nature* 2009, **458:**223-227.

11. Girard A, Sachidanandam R, Hannon GJ, Carmell MA: **A germline-specific class of small RNAs binds mammalian Piwi proteins.** *Nature* 2006, **442:**199-202.

12. Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, Kingston RE: **Characterization of the piRNA complex from rat testes.** *Science* 2006, **313:**363-367.

13. Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, Harte R, Balasubramanian S, Tanzer A, Diekhans M, et al: **The GENCODE pseudogene resource.** *Genome Biol* 2012, **13:**R51.

14. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al: **The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression.** *Genome Res* 2012, **22:**1775-1789.

15. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL: **Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.** *Genes Dev* 2011, **25:**1915-1927.

16. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL, et al: **RNA maps reveal new RNA classes and a possible function for pervasive transcription.** *Science* 2007, **316:**1484-1488.

17. Jacquier A: **The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs.** *Nat Rev Genet* 2009, **10:**833-844.

18. Taft RJ, Kaplan CD, Simons C, Mattick JS: **Evolution, biogenesis and function of promoter-associated RNAs.** *Cell Cycle* 2009, **8:**2332-2338.

19. Taft RJ, Simons C, Nahkuri S, Oey H, Korbie DJ, Mercer TR, Holst J, Ritchie W, Wong JJ, Rasko JE, et al: **Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans.** *Nat Struct Mol Biol* 2010, **17:**1030-1034.

20. Parrott AM, Tsai M, Batchu P, Ryan K, Ozer HL, Tian B, Mathews MB: **The evolution and expression of the snaR family of small non-coding RNAs.** *Nucleic Acids Res* 2011, **39:**1485-1500.

21. Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function.** *Cell* 2004, **116:**281-297.

22. Okamura K, Lai EC: **Endogenous small interfering RNAs in animals.** *Nat Rev Mol Cell Biol* 2008, **9:**673-678.

23. Shukla GC, Singh J, Barik S: **MicroRNAs: Processing, Maturation, Target Recognition and Regulatory Functions.** *Mol Cell Pharmacol* 2011, **3:**83-92.

24. Yoon JH, Abdelmohsen K, Srikantan S, Yang X, Martindale JL, De S, Huarte M, Zhan M, Becker KG, Gorospe M: **LincRNA-p21 suppresses target mRNA translation.** *Mol Cell* 2012, **47:**648-655.

25. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, Chang HY: **Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs.** *Cell* 2007, **129:**1311-1323.

26. Engreitz JM, Pandya-Jones A, McDonel P, Shishkin A, Sirokman K, Surka C, Kadri S, Xing J, Goren A, Lander ES, et al: **The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome.** *Science* 2013, **341:**1237973.

27. Monnier P, Martinet C, Pontis J, Stancheva I, Ait-Si-Ali S, Dandolo L: **H19 lncRNA**

controls gene expression of the Imprinted Gene Network by recruiting MBD1. *Proc Natl Acad Sci U S A* 2013, **110:**20693-20698.

28. Rinn JL, Chang HY: **Genome regulation by long noncoding RNAs.** *Annu Rev Biochem* 2012, **81:**145-166.

29. Mercer TR, Dinger ME, Mattick JS: **Long non-coding RNAs: insights into functions.** *Nat Rev Genet* 2009, **10:**155-159.

30. Cocquerelle C, Mascrez B, Hetuin D, Bailleul B: **Mis-splicing yields circular RNA molecules.** *FASEB J* 1993, **7:**155-160.

31. Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, Maier L, Mackowiak SD, Gregersen LH, Munschauer M, et al: **Circular RNAs are a large class of animal RNAs with regulatory potency.** *Nature* 2013, **495:**333-338.

32. Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, Damgaard CK, Kjems J: **Natural RNA circles function as efficient microRNA sponges.** *Nature* 2013, **495:**384-388.

33. Lasda E, Parker R: **Circular RNAs: diversity of form and function.** *RNA* 2014, **20:**1829-1842.

34. Guo JU, Agarwal V, Guo H, Bartel DP: **Expanded identification and characterization of mammalian circular RNAs.** *Genome Biol* 2014, **15:**409.

35. 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.

36. Shao NY, Hu HY, Yan Z, Xu Y, Hu H, Menzel C, Li N, Chen W, Khaitovich P: **Comprehensive survey of human brain microRNA by deep sequencing.** *BMC Genomics* 2010, **11:**409.

37. Hu HY, He L, Fominykh K, Yan Z, Guo S, Zhang X, Taylor MS, Tang L, Li J, Liu J, et al: **Evolution of the human-specific microRNA miR-941.** *Nat Commun* 2012, **3:**1145.

38. Hu HY, Guo S, Xi J, Yan Z, Fu N, Zhang XY, Menzel C, Liang HY, Yang HY, Zhao M, et al: **MicroRNA Expression and Regulation in Human, Chimpanzee, and Macaque Brains.** *Plos Genetics* 2011, **7**.

39. Yan Z, Hu HY, Jiang X, Maierhofer V, Neb E, He L, Hu YH, Hu H, Li N, Chen W, Khaitovich P: **Widespread expression of piRNA-like molecules in somatic tissues.** *Nucleic Acids Research* 2011, **39:**6596-6607.

40. Hu HY, He L, Khaitovich P: **Deep sequencing reveals a novel class of bidirectional promoters associated with neuronal genes.** *BMC Genomics* 2014, **15:**457.

41. Bartel DP: **MicroRNAs: target recognition and regulatory functions.** *Cell* 2009, **136:**215-233.

42. Treiber T, Treiber N, Meister G: **Regulation of microRNA biogenesis and function.** *Thromb Haemost* 2012, **107:**605-610.

43. Fabian MR, Sonenberg N, Filipowicz W: **Regulation of mRNA translation and stability by microRNAs.** *Annu Rev Biochem* 2010, **79:**351-379.

44. Friedman RC, Farh KK, Burge CB, Bartel DP: **Most mammalian mRNAs are conserved targets of microRNAs.** *Genome Res* 2009, **19:**92-105.

45. Kloosterman WP, Plasterk RH: **The diverse functions of microRNAs in animal development and disease.** *Dev Cell* 2006, **11:**441-450.

46. Wightman B, Ha I, Ruvkun G: **Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans.** *Cell* 1993, **75:**855-862.

47. Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G: **The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans.** *Nature* 2000, **403:**901-906.

48. Lagos-Quintana M, Rauhut R, Meyer J, Borkhardt A, Tuschl T: **New microRNAs from mouse and human.** *RNA* 2003, **9:**175-179.

49. Kim J, Krichevsky A, Grad Y, Hayes GD, Kosik KS, Church GM, Ruvkun G: **Identification of many microRNAs that copurify with polyribosomes in mammalian neurons.** *Proc Natl Acad Sci U S A* 2004, **101:**360-365.

50. Kaufman EJ, Miska EA: **The microRNAs of Caenorhabditis elegans.** *Semin Cell Dev Biol* 2010, **21:**728-737.

51. Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP: **The microRNAs of Caenorhabditis elegans.** *Genes Dev* 2003, **17:**991-1008.

52. Aravin AA, Lagos-Quintana M, Yalcin A, Zavolan M, Marks D, Snyder B, Gaasterland T, Meyer J, Tuschl T: **The small RNA profile during Drosophila melanogaster development.** *Dev Cell* 2003, **5:**337-350.

53. Friedlander MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N: **Discovering microRNAs from deep sequencing data using miRDeep.** *Nat Biotechnol* 2008, **26:**407-415.

54. Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, Prabhu AL, Zhao Y, McDonald H, Zeng T, Hirst M, et al: **Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells.** *Genome Res* 2008, **18:**610-621.

55. Pritchard CC, Cheng HH, Tewari M: **MicroRNA profiling: approaches and considerations.** *Nat Rev Genet* 2012, **13:**358-369.

56. Schmittgen TD, Lee EJ, Jiang J, Sarkar A, Yang L, Elton TS, Chen C: **Real-time PCR quantification of precursor and mature microRNA.** *Methods* 2008, **44:**31-38.

57. Liu CG, Calin GA, Volinia S, Croce CM: **MicroRNA expression profiling using microarrays.** *Nat Protoc* 2008, **3:**563-578.

58. Git A, Dvinge H, Salmon-Divon M, Osborne M, Kutter C, Hadfield J, Bertone P, Caldas C: **Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression.** *RNA* 2010, **16:**991-1006.

59. Zhuang F, Fuchs RT, Robb GB: **Small RNA expression profiling by high-throughput sequencing: implications of enzymatic manipulation.** *J Nucleic Acids* 2012, **2012:**360358.

60. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10:**57-63.

61. Lee LW, Zhang S, Etheridge A, Ma L, Martin D, Galas D, Wang K: **Complexity of the microRNA repertoire revealed by next-generation sequencing.** *RNA* 2010, **16:**2170-2180.

62. Kozlowska E, Krzyzosiak WJ, Koscianska E: **Regulation of huntingtin gene expression by miRNA-137, -214, -148a, and their respective isomiRs.** *Int J Mol Sci* 2013, **14:**16999-17016.

63. Ebhardt HA, Fedynak A, Fahlman RP: **Naturally occurring variations in sequence length creates microRNA isoforms that differ in argonaute effector complex specificity.** *Silence* 2010, **1:**12.

64. Fernandez-Valverde SL, Taft RJ, Mattick JS: **Dynamic isomiR regulation in Drosophila development.** *RNA* 2010, **16:**1881-1888.

65. Linsen SE, de Wit E, Janssens G, Heater S, Chapman L, Parkin RK, Fritz B, Wyman SK, de Bruijn E, Voest EE, et al: **Limitations and possibilities of small RNA digital gene expression profiling.** *Nat Methods* 2009, **6:**474-476.

66. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ: **miRBase: microRNA sequences, targets and gene nomenclature.** *Nucleic Acids Res* 2006, **34:**D140-144.

67. Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, Pfeffer S, Rice A, Kamphorst AO, Landthaler M, et al: **A mammalian microRNA expression atlas based on small RNA library sequencing.** *Cell* 2007, **129:**1401-1414.

68. Marti E, Pantano L, Banez-Coronel M, Llorens F, Minones-Moyano E, Porta S, Sumoy L, Ferrer I, Estivill X: **A myriad of miRNA variants in control and Huntington's disease brain regions detected by massively parallel sequencing.** *Nucleic Acids Res* 2010, **38:**7219-7235.

69. Winter J, Jung S, Keller S, Gregory RI, Diederichs S: **Many roads to maturity: microRNA biogenesis pathways and their regulation.** *Nat Cell Biol* 2009, **11:**228-234.

70. Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, Kim VN: **MicroRNA genes are transcribed by RNA polymerase II.** *EMBO J* 2004, **23:**4051-4060.

71. Cai X, Hagedorn CH, Cullen BR: **Human microRNAs are processed from capped,**

polyadenylated transcripts that can also function as mRNAs. *RNA* 2004, **10:**1957-1966.

72. Morlando M, Ballarino M, Gromak N, Pagano F, Bozzoni I, Proudfoot NJ: **Primary microRNA transcripts are processed co-transcriptionally.** *Nat Struct Mol Biol* 2008, **15:**902-909.

73. Denli AM, Tops BB, Plasterk RH, Ketting RF, Hannon GJ: **Processing of primary microRNAs by the Microprocessor complex.** *Nature* 2004, **432:**231-235.

74. Han J, Lee Y, Yeom KH, Kim YK, Jin H, Kim VN: **The Drosha-DGCR8 complex in primary microRNA processing.** *Genes Dev* 2004, **18:**3016-3027.

75. Bernstein E, Caudy AA, Hammond SM, Hannon GJ: **Role for a bidentate ribonuclease in the initiation step of RNA interference.** *Nature* 2001, **409:**363-366.

76. Jiang F, Ye X, Liu X, Fincher L, McKearin D, Liu Q: **Dicer-1 and R3D1-L catalyze microRNA maturation in Drosophila.** *Genes Dev* 2005, **19:**1674-1679.

77. Ha M, Kim VN: **Regulation of microRNA biogenesis.** *Nat Rev Mol Cell Biol* 2014, **15:**509-524.

78. Hutvagner G: **Small RNA asymmetry in RNAi: function in RISC assembly and gene regulation.** *FEBS Lett* 2005, **579:**5850-5857.

79. Czech B, Hannon GJ: **Small RNA sorting: matchmaking for Argonautes.** *Nat Rev Genet* 2011, **12:**19-31.

80. Khvorova A, Reynolds A, Jayasena SD: **Functional siRNAs and miRNAs exhibit strand bias.** *Cell* 2003, **115:**209-216.

81. Schwarz DS, Hutvagner G, Du T, Xu Z, Aronin N, Zamore PD: **Asymmetry in the assembly of the RNAi enzyme complex.** *Cell* 2003, **115:**199-208.

82. Frank F, Sonenberg N, Nagar B: **Structural basis for 5'-nucleotide base-specific recognition of guide RNA by human AGO2.** *Nature* 2010, **465:**818-822.

83. Czech B, Zhou R, Erlich Y, Brennecke J, Binari R, Villalta C, Gordon A, Perrimon N, Hannon GJ: **Hierarchical rules for Argonaute loading in Drosophila.** *Mol Cell* 2009, **36:**445-456.

84. Okamura K, Liu N, Lai EC: **Distinct mechanisms for microRNA strand selection by Drosophila Argonautes.** *Mol Cell* 2009, **36:**431-444.

85. Ghildiyal M, Xu J, Seitz H, Weng Z, Zamore PD: **Sorting of Drosophila small silencing RNAs partitions microRNA* strands into the RNA interference pathway.** *RNA* 2010, **16:**43-56.

86. Ladewig E, Okamura K, Flynt AS, Westholm JO, Lai EC: **Discovery of hundreds of mirtrons in mouse and human small RNA data.** *Genome Res* 2012, **22:**1634-1645.

87. Okamura K, Hagen JW, Duan H, Tyler DM, Lai EC: **The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila.** *Cell* 2007,

130:89-100.

88. Xie M, Li M, Vilborg A, Lee N, Shu MD, Yartseva V, Sestan N, Steitz JA: **Mammalian 5'-capped microRNA precursors that generate a single microRNA.** *Cell* 2013, **155:**1568-1580.

89. Zamudio JR, Kelly TJ, Sharp PA: **Argonaute-bound small RNAs from promoter-proximal RNA polymerase II.** *Cell* 2014, **156:**920-934.

90. Cheloufi S, Dos Santos CO, Chong MM, Hannon GJ: **A dicer-independent miRNA biogenesis pathway that requires Ago catalysis.** *Nature* 2010, **465:**584-589.

91. Fukunaga R, Han BW, Hung JH, Xu J, Weng Z, Zamore PD: **Dicer partner proteins tune the length of mature miRNAs in flies and mammals.** *Cell* 2012, **151:**533-546.

92. Jaskiewicz L, Zavolan M: **Dicer partners expand the repertoire of miRNA targets.** *Genome Biol* 2012, **13:**179.

93. Wu H, Ye C, Ramirez D, Manjunath N: **Alternative processing of primary microRNA transcripts by Drosha generates 5' end variation of mature microRNA.** *PLoS One* 2009, **4:**e7566.

94. Griffiths-Jones S: **The microRNA Registry.** *Nucleic Acids Res* 2004, **32:**D109-111.

95. Okamura K, Phillips MD, Tyler DM, Duan H, Chou YT, Lai EC: **The regulatory activity of microRNA* species has substantial influence on microRNA and 3' UTR evolution.** *Nat Struct Mol Biol* 2008, **15:**354-363.

96. Griffiths-Jones S, Hui JH, Marco A, Ronshaugen M: **MicroRNA evolution by arm switching.** *EMBO Rep* 2011, **12:**172-177.

97. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ: **miRBase: tools for microRNA genomics.** *Nucleic Acids Res* 2008, **36:**D154-158.

98. Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A: **Identification of mammalian microRNA host genes and transcription units.** *Genome Res* 2004, **14:**1902-1910.

99. Saini HK, Griffiths-Jones S, Enright AJ: **Genomic analysis of human microRNA transcripts.** *Proc Natl Acad Sci U S A* 2007, **104:**17719-17724.

100. Yu J, Wang F, Yang GH, Wang FL, Ma YN, Du ZW, Zhang JW: **Human microRNA clusters: genomic organization and expression profile in leukemia cell lines.** *Biochem Biophys Res Commun* 2006, **349:**59-68.

101. Ventura A, Young AG, Winslow MM, Lintault L, Meissner A, Erkeland SJ, Newman J, Bronson RT, Crowley D, Stone JR, et al: **Targeted deletion reveals essential and overlapping functions of the miR-17 through 92 family of miRNA clusters.** *Cell* 2008, **132:**875-886.

102. Lu Y, Thomson JM, Wong HY, Hammond SM, Hogan BL: **Transgenic over-expression of the microRNA miR-17-92 cluster promotes proliferation and inhibits differentiation of lung epithelial progenitor cells.** *Dev Biol* 2007, **310:**442-453.

103. Xiao C, Srinivasan L, Calado DP, Patterson HC, Zhang B, Wang J, Henderson JM, Kutok JL, Rajewsky K: **Lymphoproliferative disease and autoimmunity in mice with increased miR-17-92 expression in lymphocytes.** *Nat Immunol* 2008, **9:**405-414.

104. Ozsolak F, Poling LL, Wang Z, Liu H, Liu XS, Roeder RG, Zhang X, Song JS, Fisher DE: **Chromatin structure analyses identify miRNA promoters.** *Genes Dev* 2008, **22:**3172-3183.

105. Marsico A, Huska MR, Lasserre J, Hu H, Vucicevic D, Musahl A, Orom U, Vingron M: **PROmiRNA: a new miRNA promoter recognition method uncovers the complex regulation of intronic miRNAs.** *Genome Biol* 2013, **14:**R84.

106. Cullen BR: **Transcription and processing of human microRNA precursors.** *Mol Cell* 2004, **16:**861-865.

107. Tarver JE, Sperling EA, Nailor A, Heimberg AM, Robinson JM, King BL, Pisani D, Donoghue PC, Peterson KJ: **miRNAs: small genes with big potential in metazoan phylogenetics.** *Mol Biol Evol* 2013, **30:**2369-2382.

108. Krol J, Sobczak K, Wilczynska U, Drath M, Jasinska A, Kaczynska D, Krzyzosiak WJ: **Structural features of microRNA (miRNA) precursors and their relevance to miRNA biogenesis and small interfering RNA/short hairpin RNA design.** *J Biol Chem* 2004, **279:**42230-42239.

109. Tran Vdu T, Tempel S, Zerath B, Zehraoui F, Tahi F: **miRBoost: boosting support vector machines for microRNA precursor classification.** *RNA* 2015, **21:**775-785.

110. Berezikov E, Thuemmler F, van Laake LW, Kondova I, Bontrop R, Cuppen E, Plasterk RH: **Diversity of microRNAs in human and chimpanzee brain.** *Nat Genet* 2006, **38:**1375-1377.

111. Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, Dreyfuss G, Eddy SR, Griffiths-Jones S, Marshall M, et al: **A uniform system for microRNA annotation.** *RNA* 2003, **9:**277-279.

112. Bergeron L, Jr., Perreault JP, Abou Elela S: **Short RNA duplexes guide sequence-dependent cleavage by human Dicer.** *RNA* 2010, **16:**2464-2473.

113. Lau PW, Guiley KZ, De N, Potter CS, Carragher B, MacRae IJ: **The molecular architecture of human Dicer.** *Nat Struct Mol Biol* 2012, **19:**436-440.

114. Seitz H, Tushir JS, Zamore PD: **A 5'-uridine amplifies miRNA/miRNA* asymmetry in Drosophila by promoting RNA-induced silencing complex formation.** *Silence* 2011, **2:**4.

115. Felice KM, Salzman DW, Shubert-Coleman J, Jensen KP, Furneaux HM: **The 5' terminal uracil of let-7a is critical for the recruitment of mRNA to Argonaute2.** *Biochem J* 2009, **422:**329-341.

116. Hibio N, Hino K, Shimizu E, Nagata Y, Ui-Tei K: **Stability of miRNA 5'terminal and seed regions is correlated with experimentally observed miRNA-mediated**

silencing efficacy. *Sci Rep* 2012, **2:**996.

117. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB: **Prediction of mammalian microRNA targets.** *Cell* 2003, **115:**787-798.

118. Stark A, Brennecke J, Russell RB, Cohen SM: **Identification of Drosophila MicroRNA targets.** *PLoS Biol* 2003, **1:**E60.

119. Lambert NJ, Gu SG, Zahler AM: **The conformation of microRNA seed regions in native microRNPs is prearranged for presentation to mRNA targets.** *Nucleic Acids Res* 2011, **39:**4827-4835.

120. Lai EC: **Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation.** *Nat Genet* 2002, **30:**363-364.

121. Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP: **Vertebrate microRNA genes.** *Science* 2003, **299:**1540.

122. Lewis BP, Burge CB, Bartel DP: **Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.** *Cell* 2005, **120:**15-20.

123. Doench JG, Sharp PA: **Specificity of microRNA target selection in translational repression.** *Genes Dev* 2004, **18:**504-511.

124. Brennecke J, Stark A, Russell RB, Cohen SM: **Principles of microRNA-target recognition.** *PLoS Biol* 2005, **3:**e85.

125. Lai EC, Tam B, Rubin GM: **Pervasive regulation of Drosophila Notch target genes by GY-box-, Brd-box-, and K-box-class microRNAs.** *Genes Dev* 2005, **19:**1067-1080.

126. Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM: **Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs.** *Nature* 2005, **433:**769-773.

127. Kim YK, Wee G, Park J, Kim J, Baek D, Kim JS, Kim VN: **TALEN-based knockout library for human microRNAs.** *Nat Struct Mol Biol* 2013, **20:**1458-1464.

128. Berezikov E: **Evolution of microRNA diversity and regulation in animals.** *Nat Rev Genet* 2011, **12:**846-860.

129. Baek D, Villen J, Shin C, Camargo FD, Gygi SP, Bartel DP: **The impact of microRNAs on protein output.** *Nature* 2008, **455:**64-71.

130. Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N: **Widespread changes in protein synthesis induced by microRNAs.** *Nature* 2008, **455:**58-63.

131. Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, Degnan BM, Rokhsar DS, Bartel DP: **Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals.** *Nature* 2008, **455:**1193-1197.

132. Christodoulou F, Raible F, Tomer R, Simakov O, Trachana K, Klaus S, Snyman H,

Hannon GJ, Bork P, Arendt D: **Ancient animal microRNAs and the evolution of tissue identity.** *Nature* 2010, **463:**1084-1088.

133. Heimberg AM, Sempere LF, Moy VN, Donoghue PC, Peterson KJ: **MicroRNAs and the advent of vertebrate morphological complexity.** *Proc Natl Acad Sci U S A* 2008, **105:**2946-2950.

134. Wheeler BM, Heimberg AM, Moy VN, Sperling EA, Holstein TW, Heber S, Peterson KJ: **The deep evolution of metazoan microRNAs.** *Evol Dev* 2009, **11:**50-68.

135. Prochnik SE, Rokhsar DS, Aboobaker AA: **Evidence for a microRNA expansion in the bilaterian ancestor.** *Dev Genes Evol* 2007, **217:**73-77.

136. Wright MW, Bruford EA: **Naming 'junk': human non-protein coding RNA (ncRNA) gene nomenclature.** *Hum Genomics* 2011, **5:**90-98.

137. Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES: **Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins.** *Cell* 2013, **154:**240-251.

138. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, et al: **Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs.** *Nature* 2002, **420:**563-573.

139. Frith MC, Forrest AR, Nourbakhsh E, Pang KC, Kai C, Kawai J, Carninci P, Hayashizaki Y, Bailey TL, Grimmond SM: **The abundance of short proteins in the mammalian proteome.** *PLoS Genet* 2006, **2:**e52.

140. Lin MF, Jungreis I, Kellis M: **PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions.** *Bioinformatics* 2011, **27:**i275-282.

141. Washietl S, Findeiss S, Muller SA, Kalkhof S, von Bergen M, Hofacker IL, Stadler PF, Goldman N: **RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data.** *RNA* 2011, **17:**578-594.

142. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G: **CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine.** *Nucleic Acids Res* 2007, **35:**W345-349.

143. Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W: **CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model.** *Nucleic Acids Res* 2013, **41:**e74.

144. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C: **Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs.** *PLoS Genet* 2013, **9:**e1003470.

145. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, et al: **Stem cell transcriptome profiling via**

massive-scale mRNA sequencing. *Nat Methods* 2008, **5:**613-619.

146. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR: **Highly integrated single-base resolution maps of the epigenome in Arabidopsis.** *Cell* 2008, **133:**523-536.

147. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A: **Transcriptome analysis by strand-specific sequencing of complementary DNA.** *Nucleic Acids Res* 2009, **37:**e123.

148. He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW: **The antisense transcriptomes of human cells.** *Science* 2008, **322:**1855-1857.

149. Clemson CM, Hutchinson JN, Sara SA, Ensminger AW, Fox AH, Chess A, Lawrence JB: **An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles.** *Mol Cell* 2009, **33:**717-726.

150. Martin JA, Wang Z: **Next-generation transcriptome assembly.** *Nat Rev Genet* 2011, **12:**671-682.

151. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat Biotechnol* 2010, **28:**511-515.

152. Salzberg SL, Yorke JA: **Beware of mis-assembled genomes.** *Bioinformatics* 2005, **21:**4320-4321.

153. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al: **Full-length transcriptome assembly from RNA-Seq data without a reference genome.** *Nat Biotechnol* 2011, **29:**644-652.

154. Lam MT, Li W, Rosenfeld MG, Glass CK: **Enhancer RNAs and regulated transcriptional programs.** *Trends Biochem Sci* 2014, **39:**170-182.

155. Cusanelli E, Chartrand P: **Telomeric repeat-containing RNA TERRA: a noncoding RNA connecting telomere biology to genome integrity.** *Front Genet* 2015, **6:**143.

156. Mercer TR, Wilhelm D, Dinger ME, Solda G, Korbie DJ, Glazov EA, Truong V, Schwenke M, Simons C, Matthaei KI, et al: **Expression of distinct RNAs from 3' untranslated regions.** *Nucleic Acids Res* 2011, **39:**2393-2403.

157. Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, Ponting CP, Stadler PF, Morris KV, Morillon A, et al: **The reality of pervasive transcription.** *PLoS Biol* 2011, **9:**e1000625; discussion e1001102.

158. Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, Djebali S, Curado J, Snyder M, Gingeras TR, Guigo R: **Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs.** *Genome Res* 2012, **22:**1616-1625.

159. Wilusz JE, Freier SM, Spector DL: **3' end processing of a long nuclear-retained**

noncoding RNA yields a tRNA-like cytoplasmic RNA. *Cell* 2008, **135:**919-932.

160. Ponjavic J, Ponting CP, Lunter G: **Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs.** *Genome Res* 2007, **17:**556-565.

161. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grutzner F, Kaessmann H: **The evolution of lncRNA repertoires and expression patterns in tetrapods.** *Nature* 2014, **505:**635-640.

162. Duret L, Chureau C, Samain S, Weissenbach J, Avner P: **The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene.** *Science* 2006, **312:**1653-1655.

163. Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, Siepel A, Pedersen JS, Bejerano G, Baertsch R, et al: **Forces shaping the fastest evolving regions in the human genome.** *PLoS Genet* 2006, **2:**e168.

164. West JA, Davis CP, Sunwoo H, Simon MD, Sadreyev RI, Wang PI, Tolstorukov MY, Kingston RE: **The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites.** *Mol Cell* 2014, **55:**791-802.

165. Mercer TR, Dinger ME, Mariani J, Kosik KS, Mehler MF, Mattick JS: **Noncoding RNAs in Long-Term Memory Formation.** *Neuroscientist* 2008, **14:**434-445.

166. Taft RJ, Pheasant M, Mattick JS: **The relationship between non-protein-coding DNA and eukaryotic complexity.** *Bioessays* 2007, **29:**288-299.

167. Qureshi IA, Mehler MF: **Emerging roles of non-coding RNAs in brain evolution, development, plasticity and disease.** *Nat Rev Neurosci* 2012, **13:**528-541.

168. Giraldez AJ, Cinalli RM, Glasner ME, Enright AJ, Thomson JM, Baskerville S, Hammond SM, Bartel DP, Schier AF: **MicroRNAs regulate brain morphogenesis in zebrafish.** *Science* 2005, **308:**833-838.

169. Davis TH, Cuellar TL, Koch SM, Barker AJ, Harfe BD, McManus MT, Ullian EM: **Conditional loss of Dicer disrupts cellular and tissue morphogenesis in the cortex and hippocampus.** *J Neurosci* 2008, **28:**4322-4330.

170. Smibert P, Bejarano F, Wang D, Garaulet DL, Yang JS, Martin R, Bortolamiol-Becet D, Robine N, Hiesinger PR, Lai EC: **A Drosophila genetic screen yields allelic series of core microRNA biogenesis factors and reveals post-developmental roles for microRNAs.** *RNA* 2011, **17:**1997-2010.

171. Yoo AS, Sun AX, Li L, Shcheglovitov A, Portmann T, Li Y, Lee-Messer C, Dolmetsch RE, Tsien RW, Crabtree GR: **MicroRNA-mediated conversion of human fibroblasts to neurons.** *Nature* 2011, **476:**228-231.

172. Kuboshima-Amemori S, Sawaguchi T: **Plasticity of the primate prefrontal cortex.** *Neuroscientist* 2007, **13:**229-240.

173. Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS: **Specific expression of long noncoding RNAs in the mouse brain.** *Proc Natl Acad Sci U S A* 2008,

105:716-721.

174. Gordon FE, Nutt CL, Cheunsuchon P, Nakayama Y, Provencher KA, Rice KA, Zhou Y, Zhang X, Klibanski A: **Increased expression of angiogenic genes in the brains of mouse meg3-null embryos.** *Endocrinology* 2010, **151:**2443-2452.

175. Mercer TR, Qureshi IA, Gokhan S, Dinger ME, Li G, Mattick JS, Mehler MF: **Long noncoding RNAs in neuronal-glial fate specification and oligodendrocyte lineage maturation.** *BMC Neurosci* 2010, **11:**14.

176. Bernard D, Prasanth KV, Tripathi V, Colasse S, Nakamura T, Xuan Z, Zhang MQ, Sedel F, Jourdren L, Coulpier F, et al: **A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression.** *EMBO J* 2010, **29:**3082-3093.

177. Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A, et al: **An RNA gene expressed during cortical development evolved rapidly in humans.** *Nature* 2006, **443:**167-172.

178. Morgulis A, Gertz EM, Schaffer AA, Agarwala R: **A fast and symmetric DUST implementation to mask low-complexity DNA sequences.** *J Comput Biol* 2006, **13:**1028-1040.

179. Minoche AE, Dohm JC, Himmelbauer H: **Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems.** *Genome Biol* 2011, **12:**R112.

180. Li R, Li Y, Kristiansen K, Wang J: **SOAP: short oligonucleotide alignment program.** *Bioinformatics* 2008, **24:**713-714.

181. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.

182. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Res* 2002, **12:**656-664.

183. Bhagwat M, Young L, Robison RR: **Using BLAT to find sequence similarity in closely related genomes.** *Curr Protoc Bioinformatics* 2012, **Chapter 10:**Unit10 18.

184. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, et al: **The UCSC Genome Browser Database: update 2006.** *Nucleic Acids Res* 2006, **34:**D590-598.

185. Marques AC, Ponting CP: **Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness.** *Genome Biol* 2009, **10:**R124.

186. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al: **Clustal W and Clustal X version 2.0.** *Bioinformatics* 2007, **23:**2947-2948.

187. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19:**185-193.

188. Garmire LX, Subramaniam S: **Evaluation of normalization methods in**

mammalian microRNA-Seq data. *RNA* 2012, **18:**1279-1288.

189. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26:**139-140.

190. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data.** *Genome Biol* 2010, **11:**R25.

191. Robinson MD, Smyth GK: **Small-sample estimation of negative binomial dispersion, with applications to SAGE data.** *Biostatistics* 2008, **9:**321-332.

192. Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, Robinson MD: **Count-based differential expression analysis of RNA sequencing data using R and Bioconductor.** *Nat Protoc* 2013, **8:**1765-1786.

193. Griss J, Martin M, O'Donovan C, Apweiler R, Hermjakob H, Vizcaino JA: **Consequences of the discontinuation of the International Protein Index (IPI) database and its substitution by the UniProtKB "complete proteome" sets.** *Proteomics* 2011, **11:**4434-4438.

194. Eng JK, McCormack AL, Yates JR: **An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database.** *J Am Soc Mass Spectrom* 1994, **5:**976-989.

195. Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP: **MicroRNA targeting specificity in mammals: determinants beyond seed pairing.** *Mol Cell* 2007, **27:**91-105.

196. Cohen J: **Statistical power analysis for the behavioral sciences (2nd ed.).** *NJ: Lawrence Earlbaum Associates* 1988.

197. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.** *Journal of the Royal Statistical Society Series B-Methodological* 1995, **57:**289-300.

198. Dunn OJ: **Estimation of the Medians for Dependent-Variables.** *Annals of Mathematical Statistics* 1959, **30:**192-197.

199. Matthews BW: **Comparison of the predicted and observed secondary structure of T4 phage lysozyme.** *Biochim Biophys Acta* 1975, **405:**442-451.

200. Makeyev EV, Zhang J, Carrasco MA, Maniatis T: **The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing.** *Mol Cell* 2007, **27:**435-448.

201. Guo H, Ingolia NT, Weissman JS, Bartel DP: **Mammalian microRNAs predominantly act to decrease target mRNA levels.** *Nature* 2010, **466:**835-840.

202. Wu TD, Watanabe CK: **GMAP: a genomic mapping and alignment program for mRNA and EST sequences.** *Bioinformatics* 2005, **21:**1859-1875.

203. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, et al: **Ensembl 2011.** *Nucleic Acids Res* 2011, **39:**D800-806.

204. UniProt C: **UniProt: a hub for protein information.** *Nucleic Acids Res* 2015, **43:**D204-212.

205. Fickett JW: **Recognition of protein coding regions in DNA sequences.** *Nucleic Acids Res* 1982, **10:**5303-5318.

206. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5:**621-628.

207. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics* 2009, **25:**1105-1111.

208. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.

209. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.** *Genome Res* 2005, **15:**1034-1050.

210. Ameur A, Zaghlool A, Halvardson J, Wetterbom A, Gyllensten U, Cavelier L, Feuk L: **Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain.** *Nat Struct Mol Biol* 2011, **18:**1435-1440.

211. Shannon CE: **A Mathematical Theory of Communication.** *Bell System Technical Journal* 1948, **27:**623-656.

212. Somel M, Franz H, Yan Z, Lorenc A, Guo S, Giger T, Kelso J, Nickel B, Dannemann M, Bahn S, et al: **Transcriptional neoteny in the human brain.** *Proc Natl Acad Sci U S A* 2009, **106:**5743-5748.

213. Xie Y, Pan W, Khodursky AB: **A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data.** *Bioinformatics* 2005, **21:**4280-4288.

214. MacQueen JB: **Some Methods for classification and Analysis of Multivariate Observations.** *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* 1967**:**281–297.

215. Balwierz PJ, Carninci P, Daub CO, Kawai J, Hayashizaki Y, Van Belle W, Beisel C, van Nimwegen E: **Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data.** *Genome Biology* 2009, **10**.

216. Keller A, Backes C, Al-Awadhi M, Gerasch A, Kuntzer J, Kohlbacher O, Kaufmann M, Lenhof HP: **GeneTrailExpress: a web-based pipeline for the statistical evaluation of microarray experiments.** *Bmc Bioinformatics* 2008, **9**.

217. Supek F, Bosnjak M, Skunca N, Smuc T: **REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms.** *Plos One* 2011, **6**.

218. Cahoy JD, Emery B, Kaushal A, Foo LC, Zamanian JL, Christopherson KS, Xing Y, Lubischer JL, Krieg PA, Krupenko SA, et al: **A transcriptome database for astrocytes, neurons, and oligodendrocytes: A new resource for understanding brain development and function.** *Journal of Neuroscience* 2008, **28:**264-278.

219. Xu H, Wei CL, Lin F, Sung WK: **An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data.** *Bioinformatics* 2008, **24:**2344-2349.

220. Cheung I, Shulha HP, Jiang Y, Matevossian A, Wang J, Weng Z, Akbarian S: **Developmental regulation and individual differences of neuronal H3K4me3 epigenomes in the prefrontal cortex.** *Proc Natl Acad Sci U S A* 2010, **107:**8824-8829.

221. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biology* 2009, **10**.

222. Quandt K, Frech K, Karas H, Wingender E, Werner T: **MatInd and MatInspector: New fast and versatile tools for detection of consensus matches in nucleotide sequence data.** *Nucleic Acids Research* 1995, **23:**4878-4884.

223. Qiao N, Huang Y, Naveed H, Green CD, Han JDJ: **CoCiter: An Efficient Tool to Infer Gene Function by Assessing the Significance of Literature Co-Citation.** *Plos One* 2013, **8**.

224. Elnitski L, Hardison RC, Li J, Yang S, Kolbe D, Eswara P, O'Connor MJ, Schwartz S, Miller W, Chiaromonte F: **Distinguishing regulatory DNA from neutral sites.** *Genome Research* 2003, **13:**64-72.

225. Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong CB, Nielsen C, Zhao YJ, et al: **Conserved role of intragenic DNA methylation in regulating alternative promoters.** *Nature* 2010, **466:**253-U131.

226. Liu Y, Han DL, Han YX, Yan Z, Xie B, Li J, Qiao N, Hu HY, Khaitovich P, Gao YA, Han JDJ: **Ab initio identification of transcription start sites in the Rhesus macaque genome by histone modification and RNA-Seq.** *Nucleic Acids Research* 2011, **39:**1408-1418.

227. Khaitovich P, Muetzel B, She XW, Lachmann M, Hellmann I, Dietzsch J, Steigele S, Do HH, Weiss G, Enard W, et al: **Regional patterns of gene expression in human and chimpanzee brains.** *Genome Research* 2004, **14:**1462-1473.

228. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E: **The role of site accessibility in microRNA target recognition.** *Nature Genetics* 2007, **39:**1278-1284.

229. Fang F, Fan SC, Zhang XG, Zhang MQ: **Predicting methylation status of CpG islands in the human brain.** *Bioinformatics* 2006, **22:**2204-2209.

230. Ramskold D, Wang ET, Burge CB, Sandberg R: **An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data.** *PLoS Comput*

Biol 2009, **5:**e1000598.

231. Jenal M, Elkon R, Loayza-Puch F, van Haaften G, Kuhn U, Menzies FM, Oude Vrielink JA, Bos AJ, Drost J, Rooijers K, et al: **The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites.** *Cell* 2012, **149:**538-553.

232. Simonelig M: **PABPN1 shuts down alternative poly(A) sites.** *Cell Res* 2012, **22:**1419-1421.

233. Beaulieu YB, Kleinman CL, Landry-Voyer AM, Majewski J, Bachand F: **Polyadenylation-dependent control of long noncoding RNA expression by the poly(A)-binding protein nuclear 1.** *PLoS Genet* 2012, **8:**e1003078.

234. Wang WC, Lin FM, Chang WC, Lin KY, Huang HD, Lin NS: **miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression.** *BMC Bioinformatics* 2009, **10:**328.

235. Buermans HP, Ariyurek Y, van Ommen G, den Dunnen JT, t Hoen PA: **New methods for next generation sequencing based microRNA expression profiling.** *BMC Genomics* 2010, **11:**716.

236. Zhang Y, Xu B, Yang Y, Ban R, Zhang H, Jiang X, Cooke HJ, Xue Y, Shi Q: **CPSS: a computational platform for the analysis of small RNA deep sequencing data.** *Bioinformatics* 2012, **28:**1925-1927.

237. Muller S, Rycak L, Winter P, Kahl G, Koch I, Rotter B: **omiRas: a Web server for differential expression analysis of miRNAs derived from small RNA-Seq data.** *Bioinformatics* 2013, **29:**2651-2652.

238. Friedlander MR, Mackowiak SD, Li N, Chen W, Rajewsky N: **miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades.** *Nucleic Acids Res* 2012, **40:**37-52.

239. Pantano L, Estivill X, Marti E: **SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells.** *Nucleic Acids Res* 2010, **38:**e34.

240. Kheradpour P, Stark A, Roy S, Kellis M: **Reliable prediction of regulator targets using 12 Drosophila genomes.** *Genome Res* 2007, **17:**1919-1931.

241. Felsenstein J: **Evolutionary trees from DNA sequences: a maximum likelihood approach.** *J Mol Evol* 1981, **17:**368-376.

242. Cloonan N, Wani S, Xu Q, Gu J, Lea K, Heater S, Barbacioru C, Steptoe AL, Martin HC, Nourbakhsh E, et al: **MicroRNAs and their isomiRs function cooperatively to target common biological pathways.** *Genome Biol* 2011, **12:**R126.

243. Ruby JG, Stark A, Johnston WK, Kellis M, Bartel DP, Lai EC: **Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs.** *Genome Res* 2007, **17:**1850-1864.

244. Marco A, Hooks K, Griffiths-Jones S: **Evolution and function of the extended**

miR-2 microRNA family. *RNA Biol* 2012, **9:**242-248.

245. Marco A, Hui JH, Ronshaugen M, Griffiths-Jones S: **Functional shifts in insect microRNA evolution.** *Genome Biol Evol* 2010, **2:**686-696.

246. Liu C, Teng ZQ, Santistevan NJ, Szulwach KE, Guo W, Jin P, Zhao X: **Epigenetic regulation of miR-184 by MBD1 governs neural stem cell proliferation and differentiation.** *Cell Stem Cell* 2010, **6:**433-444.

247. Fragkouli A, Doxakis E: **miR-7 and miR-153 protect neurons against MPP(+)-induced cell death via upregulation of mTOR pathway.** *Front Cell Neurosci* 2014, **8:**182.

248. Liu N, Landreh M, Cao K, Abe M, Hendriks GJ, Kennerdell JR, Zhu Y, Wang LS, Bonini NM: **The microRNA miR-34 modulates ageing and neurodegeneration in Drosophila.** *Nature* 2012, **482:**519-523.

249. Brameier M: **Genome-wide comparative analysis of microRNAs in three non-human primates.** *BMC Res Notes* 2010, **3:**64.

250. Baev V, Daskalova E, Minkov I: **Computational identification of novel microRNA homologs in the chimpanzee genome.** *Comput Biol Chem* 2009, **33:**62-70.

251. Sunkar R, Jagadeeswaran G: **In silico identification of conserved microRNAs in large number of diverse plant species.** *BMC Plant Biol* 2008, **8:**37.

252. Gerlach D, Kriventseva EV, Rahman N, Vejnar CE, Zdobnov EM: **miROrtho: computational survey of microRNA genes.** *Nucleic Acids Res* 2009, **37:**D111-117.

253. Lu J, Shen Y, Wu Q, Kumar S, He B, Shi S, Carthew RW, Wang SM, Wu CI: **The birth and death of microRNA genes in Drosophila.** *Nat Genet* 2008, **40:**351-355.

254. de Wit E, Linsen SE, Cuppen E, Berezikov E: **Repertoire and evolution of miRNA genes in four divergent nematode species.** *Genome Res* 2009, **19:**2064-2074.

255. Zhang YE, Landback P, Vibranovski MD, Long M: **Accelerated recruitment of new brain development genes into the human genome.** *PLoS Biol* 2011, **9:**e1001179.

256. Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM: **An abundance of bidirectional promoters in the human genome.** *Genome Res* 2004, **14:**62-66.

257. Wei W, Pelechano V, Jarvelin AI, Steinmetz LM: **Functional consequences of bidirectional promoters.** *Trends Genet* 2011, **27:**267-276.

258. Yang MQ, Koehly LM, Elnitski LL: **Comprehensive annotation of bidirectional promoters identifies co-regulation among breast and ovarian cancer genes.** *PLoS Comput Biol* 2007, **3:**e72.

259. Sigova AA, Mullen AC, Molinie B, Gupta S, Orlando DA, Guenther MG, Almada AE, Lin C, Sharp PA, Giallourakis CC, Young RA: **Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells.** *Proc Natl Acad Sci U S A* 2013, **110:**2876-2881.

260. Duttke SH, Lacadie SA, Ibrahim MM, Glass CK, Corcoran DL, Benner C, Heinz S, Kadonaga JT, Ohler U: **Human promoters are intrinsically directional.** *Mol Cell* 2015, **57:**674-684.

Appendix A: Supplementary Figures

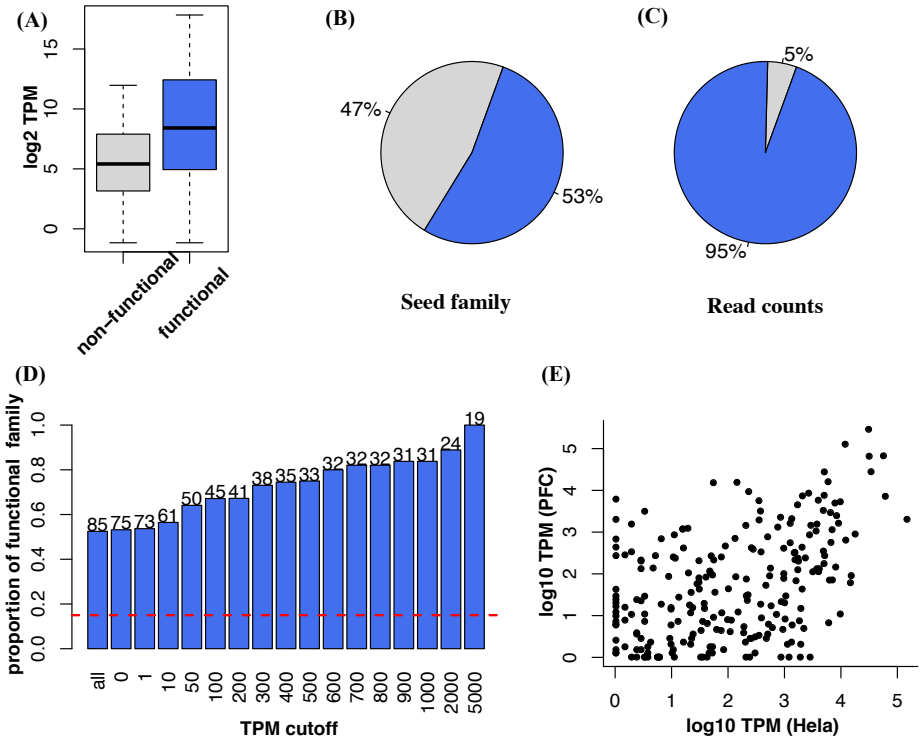


Figure A.1: The expression abundance comparison between functional and nonfunctional conserved miRNA families based on USP4 prediction in human Hela cell line. (A) Expression abundance distributions of predicted functional and nonfunctional conserved miRNA families with boxplot. Panels B and C shows the proportion of predicted functional conserved miRNA families in terms of expressed miRNA family number and total read counts. The functional conserved miRNA family and ncRNA fragment family were predicted using USP4. (D) Proportion of predicted functional conserved miRNA family on different miRNA families' expression levels. The number above the bar represents the number of predicted functional miRNA families at each miRNA family expression cutoff. (E) Expression comparison of conserved miRNAs between human prefrontal cortex (PFC) and Hela cell line (Pearson correlation r=0.39, p<10e-5). The miRNA expressions were normalized by total mapped reads of corresponding sample into transcript per million reads (TPM).

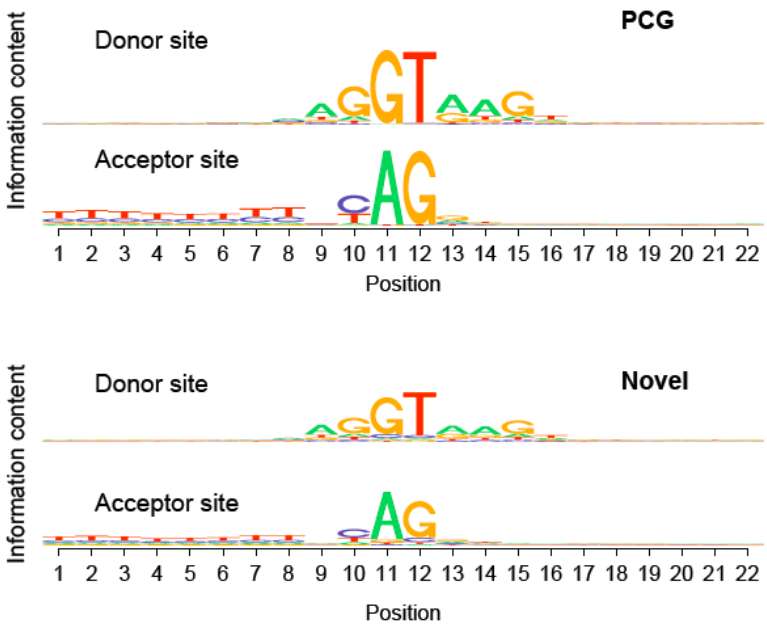


Figure A.2: Splicing site signals at donor and acceptor sites for novel lncRNAs and known protein-coding genes. Nucleotide composition at and around the splice sites (positions 11-12) of annotated protein-coding genes (PCG, upper panel) and novel transcripts (bottom panel).

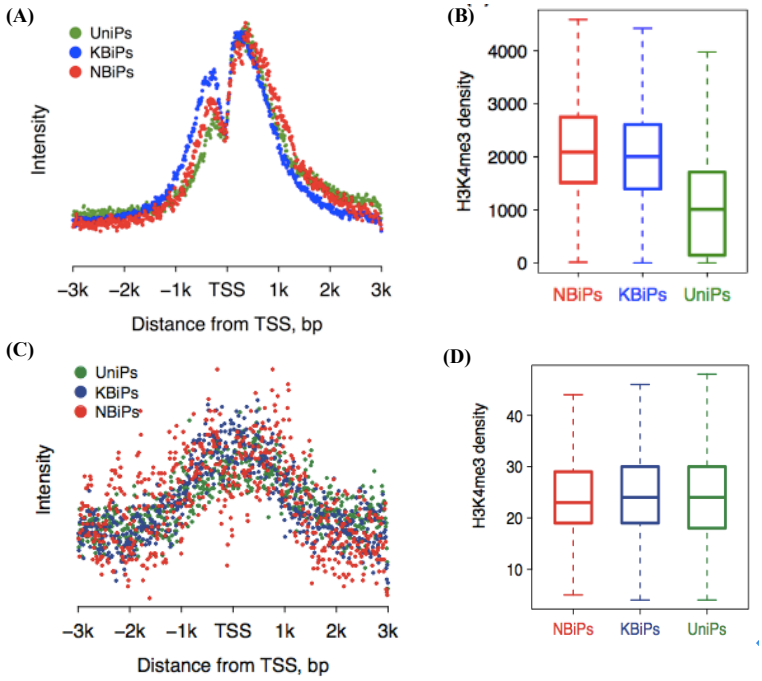


Figure A.3: H3K4me3 modification and input control profiles at three promoter types in rhesus macaque prefrontal cortex. (A) and (B) are based on H3K4me3 modifications; and (C) and (D) are based on H3K4me3 input controls. The left panels show the H3K4me3 modification profiles and the right panels show the H3K4me3 modification density at promoter regions. Green—UniPs, blue—KBiPs and red—NBiPs.

Appendix B: Supplementary Tables

Table B.1: The 5'-isoforms identified from human prefrontal cortex

5'-isoform ID	Seed sequence	Human PFC	Chimp PFC	Macaque PFC	Human brain	Mouse brain
hsa-mir-101|-1	TACAGTA	4454.64 [a]	5129.25	5129.25	1584.79	4145.75
hsa-mir-124|-1	TAAGGCA	586.75	541.43	541.43	940.11	22930.30
hsa-mir-124|+1	AGGCACG	157.10	170.85	170.85	7.76	2403.12
hsa-mir-125b|+2	CTGAGAC	40.37	43.04	43.04	129.35	202.30
hsa-mir-126|+1	GTACCGT	34.45	34.12	34.12	219.89	1702.46
hsa-mir-127-3p|+2	GATCCGT	11.83	14.17	14.17	0.52	34.04
hsa-mir-127-5p|-1	CTGAAGC	11.31	13.91	13.91	0.00	18.66
hsa-mir-127-5p|+2	AAGCTCA	42.17	41.47	41.47	8.28	440.73
hsa-mir-128|+2	CAGTGAA	285.15	291.05	291.05	8.80	106.45
hsa-mir-132*|-1	ACCGTGG	16.97	0.00	0.00	70.88	160.50
hsa-mir-135a*|-1	TGTAGGG	11.83	20.73	20.73	0.00	5.37
hsa-mir-137|+1	ATTGCTT	13.63	24.15	24.15	34.67	944.77
hsa-mir-137|+2	TTGCTTA	10.28	12.60	12.60	6.21	168.41
hsa-mir-140-3p|+1	CCACAGG	6674.63	5958.84	5958.84	315.09	457.60
hsa-mir-140-3p|+2	CACAGGG	252.49	201.82	201.82	0.52	13.59
hsa-mir-142-5p|-2	CCATAAA	333.23	368.47	368.47	440.82	354.44
hsa-mir-151-3p|-1	CTAGACT	13.11	14.17	14.17	23.28	6.87
hsa-mir-151-3p|-2	ACTAGAC	60.68	59.31	59.31	172.29	1.49
hsa-mir-181a*|-1	ACCATCG	12.86	13.38	13.38	6.21	1.05

hsa-mir-181b|-1	AACATTC	73.28	95.01	95.01	406.16	1.19
hsa-mir-181b|+1	CATTCAT	22.63	16.01	16.01	48.64	19.11
hsa-mir-191|+1	ACGGAAT	55.54	97.89	97.89	1059.11	192.75
hsa-mir-192|+1	GACCTAT	204.15	238.56	238.56	670.03	33.14
hsa-mir-199a-3p|-1	ACAGTAG	166.87	151.69	151.69	52.26	23.89
hsa-mir-199a-3p|+1	AGTAGTC	49.37	37.00	37.00	9.83	11.50
hsa-mir-199b-3p|-1	ACAGTAG	166.87	151.69	151.69	52.26	23.89
hsa-mir-199b-3p|+1	AGTAGTC	49.37	37.00	37.00	9.83	11.50
hsa-mir-221*|-1	ACCTGGC	47.31	59.84	59.84	1.55	2.24
hsa-mir-23b*|-1	TGGGTTC	17.23	13.91	13.91	0.00	0.00
hsa-mir-23b*|+1	GGTTCCT	22.88	27.56	27.56	0.00	0.45
hsa-mir-24|+1	GCTCAGT	12.86	5.25	5.25	0.52	361.90
hsa-mir-26b|+3	AGTAATT	35.23	25.46	25.46	0.00	28.52
hsa-mir-27b|-1	TTCACAG	17.23	14.17	14.17	45.01	60.62
hsa-mir-29a|-1	TAGCACC	1763.60	93.43	93.43	30.01	4380.89
hsa-mir-30a*|+1	TTCAGTC	13.37	18.90	18.90	22.25	46.73
hsa-mir-30e|+1	TAAACAT	11.31	12.07	12.07	99.34	376.98
hsa-mir-320a|-1	AAAAGCT	39.08	48.03	48.03	4.66	10.15
hsa-mir-320a|+1	AAGCTGG	98.48	160.62	160.62	12.42	50.02
hsa-mir-320a|+2	AGCTGGG	22.63	39.63	39.63	0.52	13.88
hsa-mir-323-3p|-1	CACATTA	84.59	108.39	108.39	107.62	127.95
hsa-mir-330-3p|+1	AAAGCAC	1263.76	1502.77	1502.77	15.00	451.18
hsa-mir-330-3p|+2	AAGCACA	540.21	598.12	598.12	19.66	11.94
hsa-mir-342-3p|+2	CACACAG	42.68	32.54	32.54	187.82	71.51
hsa-mir-363|+1	TTGCACG	16.20	0.00	0.00	54.33	1.19

hsa-mir-378|+1	TGGACTT	28.80	48.55	48.55	2.59	44.79
hsa-mir-382|+1	AGTTGTT	47.82	71.12	71.12	2.07	48.97
hsa-mir-383|-1	AGATCAG	32.91	56.43	56.43	3.62	328.31
hsa-mir-409-3p|-1	GAATGTT	44.48	39.63	39.63	71.92	11.94
hsa-mir-411|-1	TAGTAGA	202.87	205.23	205.23	2113.57	506.42
hsa-mir-423-3p|-1	AGCTCGG	36.77	27.56	27.56	62.61	8.81
hsa-mir-433|+1	CATGATG	17.48	19.42	19.42	0.00	14.78
hsa-mir-485-3p|-1	GTCATAC	18.77	22.57	22.57	6.73	20.16
hsa-mir-485-3p|+3	TACACGG	10.28	7.87	7.87	0.00	2.54
hsa-mir-487b|+1	TCGTACA	15.43	10.76	10.76	5.69	49.12
hsa-mir-495|+1	ACAAACA	12.08	15.75	15.75	7.76	29.26
hsa-mir-504|+1	ACCCTGG	27.51	24.15	24.15	38.29	5.67
hsa-mir-539*|-1	ATCATAC	11.31	3.15	3.15	10.87	0.00
hsa-mir-539*|+2	ATACAAG	89.74	27.56	27.56	20.70	0.00
hsa-mir-539*|+3	TACAAGG	174.07	54.59	54.59	17.59	0.00
hsa-mir-664*|+1	TGGCTAG	16.46	29.92	29.92	3.62	7.76
hsa-mir-886-5p|+1	GGTCGGA	28.54	29.66	29.66	0.00	0.00
hsa-mir-9*|+1	AAAGCTA	687.03	837.47	837.47	670.55	5015.42
hsa-mir-99a|-1	AACCCGT	33.43	28.61	28.61	73.47	17.32
hsa-mir-99b|+1	CCCGTAG	14.91	13.12	13.12	92.61	20.01
hsa-mir-9|+1	TTTGGTT	120.08	156.42	156.42	674.17	3596.62
hsa-mir-9|+2	TTGGTTA	30.85	47.50	47.50	102.96	686.33

[a] expression normalized by total mapped reads in Transcript Per Million (TPM)

Table B.2: Sample information of 14 strand-specific RNA-seq data

Sample index	Age (days)	Tissue	Accession number	Total reads
S1	2	Prefrontal cortex	SRR107727	21,277,649
S2	4	Prefrontal cortex	SRR111895	21,284,713
S3	19	Prefrontal cortex	SRR111896	20,754,409
S4	34	Prefrontal cortex	SRR111897	23,722,421
S5	94	Prefrontal cortex	SRR111898	23,416,250
S6	204	Prefrontal cortex	SRR111899	22,698,303
S7	443	Prefrontal cortex	SRR111900	23,934,412
S8	787	Prefrontal cortex	SRR111901	17,759,057
S9	5,105	Prefrontal cortex	SRR111902	19,901,399
S10	9,277	Prefrontal cortex	SRR111903	23,201,284
S11	19,457	Prefrontal cortex	SRR111904	16,019,209
S12	24,090	Prefrontal cortex	SRR111905	20,948,595
S13	32,120	Prefrontal cortex	SRR111906	21,032,459
S14	35,770	Prefrontal cortex	SRR111907	20,255,260
Total	--	--	--	296,205,420

Appendix C: Curriculum Vitae

For reasons of data protection, the curriculum vitae is not included in the online version.

Der Lebenslauf ist in der Online-Version aus Gründen des Datenschutzes nicht enthalten.

Appendix D: Publications (during PhD training)

1. Wei, Y.N., **Hu, H.Y**., Xie, G.C., Fu, N., Ning, Z.B., Zeng, R. and Khaitovich, P. (2015) Transcript and protein expression decoupling reveals RNA binding proteins and miRNAs as potential modulators of human aging. *Genome biology*, **16**, 41.

2. **Hu, H.Y**.**, He, L. and Khaitovich, P.** (2014) Deep sequencing reveals a novel class of bidirectional promoters associated with neuronal genes. *BMC genomics*, **15**, 457.

3. Meunier, J., Lemoine, F., Soumillon, M., Liechti, A., Weier, M., Guschanski, K., **Hu, H.Y.**, Khaitovich, P. and Kaessmann, H. (2013) Birth and expression evolution of mammalian microRNA genes. *Genome research*, **23**, 34-45.

4. Marsico, A., Huska, M.R., Lasserre, J., **Hu, H.Y.**, Vucicevic, D., Musahl, A., Orom, U. and Vingron, M. (2013) PROmiRNA: a new miRNA promoter recognition method uncovers the complex regulation of intronic miRNAs. *Genome biology*, **14**, R84.

5. Weng, K., **Hu, H.Y.**, Xu, A.G., Khaitovich, P. and Somel, M. (2012) Mechanisms of dietary response in mice and primates: a role for EGR1 in regulating the reaction to human-specific nutritional content. *PloS one*, **7**, e43915.

6. **Hu, H.Y.***, He, L.*, Fominykh, K., Yan, Z., Guo, S., Zhang, X., Taylor, M.S., Tang, L., Li, J., Liu, J., Wang, W., Yu, H. and Khaitovich, P. (2012) Evolution of the human-specific microRNA miR-941. *Nature communications*, **3**, 1145.

7. Yan, Z.*, **Hu, H.Y.***, Jiang, X., Maierhofer, V., Neb, E., He, L., Hu, Y., Hu, H., Li, N., Chen, W. and Khaitovich, P. (2011) Widespread expression of piRNA-like molecules in somatic tissues. *Nucleic acids research*, **39**, 6596-6607.

8. Somel, M., Liu, X., Tang, L., Yan, Z., **Hu, H.Y.**, Guo, S., Jiang, X., Zhang, X., Xu, G., Xie, G., Li, N., Hu, Y., Chen, W., Paabo, S. and Khaitovich, P. (2011) MicroRNA-driven developmental remodeling in the brain distinguishes humans from other primates. *PLoS biology*, **9**, e1001214.

9. Liu, Y., Han, D., Han, Y., Yan, Z., Xie, B., Li, J., Qiao, N., **Hu, H.Y.**, Khaitovich, P., Gao, Y. and Han, J.D. (2011) Ab initio identification of transcription start sites in the Rhesus macaque genome by histone modification and RNA-Seq. *Nucleic acids research*, **39**, 1408-1418.

10. **Hu, H.Y.***, Guo, S.*, Xi, J., Yan, Z., Fu, N., Zhang, X., Menzel, C., Liang, H., Yang, H., Zhao, M., Zeng, R., Chen, W., Paabo, S. and Khaitovich, P. (2011) MicroRNA expression and regulation in human, chimpanzee, and macaque brains. *PLoS genetics*, **7**, e1002327.

11. Somel, M., Guo, S., Fu, N., Yan, Z., **Hu, H.Y.**, Xu, Y., Yuan, Y., Ning, Z., Hu, Y., Menzel, C., Hu, H., Lachmann, M., Zeng, R., Chen, W. and Khaitovich, P. (2010)

MicroRNA, mRNA, and protein expression link development and aging in human and macaque brain. *Genome research*, **20**, 1207-1218.

12. Shao, N.Y.*, **Hu, H.Y.***, Yan, Z., Xu, Y., Hu, H., Menzel, C., Li, N., Chen, W. and Khaitovich, P. (2010) Comprehensive survey of human brain microRNA by deep sequencing. *BMC genomics*, **11**, 409.

13. **Hu, H.Y.***, Yan, Z., Xu, Y., Hu, H., Menzel, C., Zhou, Y.H., Chen, W. and Khaitovich, P. (2009) Sequence features associated with microRNA strand selection in humans and flies. *BMC genomics*, **10**, 413.

* First/Co-first author

** Correponding author