Tailored Analysis in Studying Transcriptome Landscape

Dissertation zur Erlangung des Grades

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

vorgelegt von

Xintian Arthur YOU

am Fachbereich Mathematik und Informatik

der Freien Universität Berlin

Berlin 2015

Supervisor: Prof. Dr. Knut Reinert

Second supervisor: Prof. Dr. Wei Chen

Date of the viva voce/defense: 2015-12-14

Selbststandigkeitserklärung

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.

I hereby declare that this thesis is my own original research work and has not been submitted in any form for another degree of diploma at any university or other institute of education. Contributions from others have been clearly acknowledged in the text and references to literatures are given.

Xintian Arthur YOU

2015-08-30, Berlin

Preface

All the results of studies presented here originated from discussions and collaborations with other researchers. I summarize my contributions and acknowledge the contributions of my collaborators in below.

Chapter 2 describes a novel approach for *de novo* transcriptome assembly and is published on Genome Research [1]. The idea for building up a better transcriptome reference by combining the advantages of two different sequencing technologies was proposed by Prof. Nikolaus Rajewsky and Prof. Wei Chen. My contribution was developing a hybrid assembly pipeline and assessing the quality of assembled transcriptome. I would like to acknowledge contributions of Catherine Adamini, Yongbo Wang, Guido Mastrobuoni, Pinar Oenal, Agnieszka Rybak-Wolf, Dominic Grün, Dominic Tolle, Matthias Dodt, Sebastian D. Mackowiak, Andreas Gogol-Döring, Eric Ross, Alejandro Sanchez Alvarado, Stefan Kempa and Christoph Dieterich.

Chapter 3 describes an approach for direct identification of gene transcripts in full-length. Parts of the results have been published in EMBO J [2] and PLOS One [3], and the projects are still on going. Prof. Wei Chen initiated the idea to sequence full-length transcripts with Pacific Biosciences (PacBio) sequencing-based strategy. My contribution was to analyze PacBio sequencing data, develop computational pipeline (iPEC) to correct sequencing errors and characterize full-length transcripts. I would like to acknowledge contributions of Wei Sun, Claudia Quedenau, Sophie A. O. Armitage, Andreas Gogol-Döring, Haihuai He, Yoshiaki Kise, Madlen Sohn, Tao Chen, Prof. Ansgar Klebes, and Prof. Dietmar Schmucher.

Chapter 4 describes a pipeline for identification and functional validation of microRNAs without reference genomic sequences. Parts of the results have been published in Nucleic Acids Research [4]. The idea stemmed from a project to characterize pre-miRNAs in MCF-7 cell lines with Na Li and

supervised by Prof. Wei Chen. My contribution was to develop the computational pipeline (miRGrep) for *de novo* miRNA prediction. Na Li and Tao Chen carried out most of the wet-lab experiments. I would like to acknowledge Zisong Chang, Martina Weigt, and Hang Du for their experimental contributions. Besides, I would like to thank Sebastian D. Mackowiak, Marc R. Friedländer, Andreas Gogol-Döring, Yuhui Hu and Christoph Dieterich for fruitful discussions.

Chapter 5 describes the identification and functional analysis of circular RNAs (circRNAs). Parts of the results have been published in Nature Neuroscience [5]. Prof. Erin M. Schuman and Prof. Wei Chen promoted a genome-wide profiling of non-coding RNAs in neurons. My contribution was to develop a computational pipeline (acfs) for circRNA identification, annotation and quantification. Ana Babic made synaptosomes preparations that led to our discovery that circRNAs are enriched in brain synapses. Together with Claudia Quedenau, we demonstrated the full-length sequence identity of several circRNAs using PacBio technology. I would also like to acknowledge the contributions of Irena Vlatkovic, Tristan Will, Irina Epstein, Georgi Tushev, Güney Akbalik, Mantian Wang, Caspar Glock, Xi Wang, Jingyi Hou, Hongyu Liu, Wei Sun, Sivakumar Sambandan and Tao Chen.

Acknowledgments

I would like to thank my supervisors Prof. Wei Chen, Prof. Knut Reinert and Prof. Martin Vingron for their guidance and support during my PhD. I am grateful to Prof. Wei Chen for letting me work with so many cutting-edge sequencing technologies, encouraging me to participate and lead many research topics and providing such stimulating research environment at the MDC Berlin. I shall benefit from it all my life. I am grateful to Prof. Knut Reinert for his advice and support on both scientific research and career planning. I am grateful to Prof. Martin Vingron for his valuable mathematical insight and culturing a warm environment at the MPIMG in Berlin. I am grateful to the following people for proofreading my thesis: Prof. Wei Chen, Prof. Knut Reinert, Prof. Martin Vingron, Kun Song. I am very grateful to Verena Heinrich, Robert Schöpflin, Anna Ramisch and Edgar Steiger for helping me with the Zusammenfassung.

I would like to thank Jennifer Stewart, Sabrina Deter for helping my get settled in Berlin, and Kirsten Kelleher and Hannes Luz for very kind assistance along all steps of my PhD. I thank the entire group of Prof. Wei Chen in MDC, the Algorithmic Bioinformatics group at Freie Universität Berlin and the Computational Molecular Biology group at MPIMP for fruitful collaborations, inspiring conversations and always-helpful feedback. I would like to thank Yongbo Wang for his excellent experiments that lead to my first publication. I would like to thank Na Li for her excellent experiments, lots of discussions with me and even debugging of my scripts. I would like to thank Wei Sun for his brilliant and innovative approaches towards difficult biological questions; one of them is still unsurpassable since we published it. I would like to thank Ana Babic for her devotion in synaptosomes that opens a door for me to neuroscience. I would like to thank Wei Sun, Tao Chen, Hang Du, Meisheng Xiao, Hongyu Liang, Zhong Wang for augmenting my biological knowledge through experiments and discussions. I would like to thank Mirjam Feldkamp,

Claudia Langnick, Madlen Sohn, Claudia Quedenau and Anna-Maria Ströhl for excellent sequencing management.

Finally, I would like to thank my parents and in-laws for their support. And especially, I want to thank my wife Dr. Xinyi Yang for her support throughout my thesis, who had to listen to my nonsense from cell to entropy. Their long lasting support and love is the foundation of all my achievements. I love you.

Xintian Arthur YOU
2015-08-30, Berlin

Table of Contents

Chapter 1

Introduction

1.1 Transcriptome: center of the central dogma

1.1.1 Genome

Genome, as encoded in deoxyribonucleic acid (DNA), is the complete instruction set for the development and function of all living organisms. Being able to be faithfully replicated and pass on to future generations, it is also referred to as the “blueprint of life”. The building blocks of DNA consist of four kinds of nucleotides, and therefore genetic information can be encoded on an array of nucleotides. The DNA sequence can be roughly partitioned into two groups in terms of their functions: genes and regulatory elements. The DNA sequences of genes instruct enzymes to synthesize RNA (ribonucleic acid) molecules in the nucleus. Regulatory sequences are generally not transcribed into RNA but are recognized by proteins such as transcriptional factors (TFs) to facilitate the transcription regulation of the genes either located proximal or distal to a gene. According to their regulatory roles, they can be further classified into promoters, enhancers, insulators and silencers. The sequence integrity of the genome is so important that even a single nucleotide alteration could lead to diseases [6].

Genetic information can be extracted from DNA sequences, for example, the number of genes, the possible function of genes and the genetic variants that might explain the etiology of many diseases, etc. Therefore, great efforts have been made to identify the genome of living organisms, such as the Human Genome Project [7] that finished in 2004 and the 1000 Genome Project that launched in 2008. In comparison to the sheer number of sequences genomes, it has been still far from adequacy of functional interpretation of these

genomes in depth, and the completeness of which should shed light on better understanding of biology and medicine.

1.1.2 Transcriptome

Transcriptome constitute of RNA, which is another kind of macromolecule that encodes genetic information. RNA is also the genome for many viruses that take RNA instead of DNA as their genetic materials [8]. Although RNA can store as much genetic information as DNA does, it is a “mortal surrogate” of DNA in most cases due to the following two reasons. On one hand, RNA does not have the information fail-safe mechanism as DNA does. Due to the fact that RNA is usually single-stranded, there is no backup from which the genetic information could be restored. One the other hand, RNA is much less stable than DNA as it is more prone to hydrolysis.

The biogenesis of RNA, a process known as transcription, is catalyzed by RNA polymerases in the fashion of scanning one strand of the DNA, namely the “template” strand, to produce a complementary copy. Most RNAs need to be modified in order to exert their biological functions. Some modifications are simple, requiring only one specific enzyme, such as the 2’-O-methylation of ribosomal RNAs (rRNAs) and the termini processing of transfer RNAs (tRNAs); some are carried out by an orchestrated ensemble of complexes named ribonucleoproteins (RNPs), such as the splicing of messenger RNAs (mRNAs). The mRNA splicing is a regulated process that contributes to the production of mature mRNA, in which the intron sequences are removed and the exon sequences are concatenated in cell nuclei mostly. Then the mature mRNAs are exported to the cytoplasm and translated into proteins. The removed introns, in the form of lariat, can either be degraded completely by exonuclease, or be processed to give rise to functional small RNAs such as microRNA (miRNAs) [9] or circular RNAs (circRNAs) [10].

The splicing of mRNA is a highly dynamic and versatile procedure. Alternative processing of RNAs could yield many different transcript isoforms from the very same gene locus, which is a procedure named alternative splicing. The

proteins consequently translated from the alternative spliced mRNAs could have similar or antagonizing (dominant negative effect, such as the Homer1b/c~Homer1a pair [11]) or even unrelated (multi-ORF transcripts, such as Cers1/Gdf1 gene [12]) functions. Since alternative splicing could render multiplexed biological functions of the same gene, it greatly increases the diversity of proteins encoded by the genome by expanding their functions. Therefore, identification of a comprehensive catalog of RNA at the transcript level, instead of only at gene level, and measuring their cellular abundance in different tissues/developmental stages is of utmost importance to understand the molecular mechanism of development and diseases. With the drastic improvement of massive parallel sequencing techniques, including both next-generation sequencing (NGS) and third-generation sequencing (Single-cell/-Molecule Real-Time sequencing, or SMRT), we are able to profile the transcriptome landscape in an unbiased and cost-effective manner. The increasing knowledge of transcriptome would greatly facilitate the advance of biology and medicine.

1.1.3 Proteome

Proteins are macromolecules consist of amino acid residues in chains. They are synthesized by ribosomes in a way that up to 20 different kinds of amino acids are sequentially linked by peptide bonds, the order of which is coded in the mRNAs. Proteins are the final functioning units in all forms of life, as they contribute to the production, modification, transport and turnover of DNA, RNA and proteins themselves.

1.1.4 The center of molecular biology

“DNA makes RNA and RNA makes protein” is one simplified version of the central dogma of molecular biology. As DNA does not catalyze biochemical reactions and proteins do not encode genetic information, RNA stands in the middle and serves as the bridge between the “blueprint” and the “workforce”. The fact that the diversity of transcriptome expands and evolves much faster

than that of genome further indicates that transcriptome stands in the center of molecular biology [13].

1.2 Complexity of transcriptome

1.2.1 Protein-coding or non-coding

Transcripts can be categorized in different ways owning to their great diversity on different aspects. Out of many ways of looking at transcription products, the ability to instruct protein synthesis can be one straightforward classification. In the sequence of protein coding genes (mRNAs), the region between the start codon (AUG in eukaryotes) and the stop codon (UAG, UAA or UGA) is called the coding region (CDS) as its sequence is decoded by ribosomes and translated into proteins. Due to the importance that they make proteins, mRNAs were once considered to be the only informative part of the genome, and the rest of the genome was viewed as “junk DNA” [14], [15] . Since the discovery of tRNAs in 1970s, more and more functional non-coding genes (ncRNAs) have been identified, such as signal recognition particle RNA in 1982 [16], antisense RNA in 1984 [17], microRNAs in 2001 [18], piwi-interacting RNA (piRNA) in 2006 [19], long intergenic non-coding RNA (lincRNA) [20] and most recently circRNAs in 2012 [21] . There are roughly 20,000 coding and 10,000 non-coding genes in the human genome. Although the cellular abundance of coding genes are in general higher than that of non-coding genes, increasing volume of studies shows that the functional importance of non-coding genes is no less than that of coding genes.

1.2.2 Linear or non-linear

Another way of categorizing transcripts is to see whether the RNA molecule has termini. Most of the RNA molecules are linear, with distinct 5’ and 3’ termini. They can fold into secondary and ternary structure. Although folding could render some resistance to RNA turnover, linear RNAs can be degraded from the two ends by exoribonucleases. Circular transcripts (circRNAs), however, are largely immune to exoribonucleases such as RNase R thanks to

the closed structure. They are transcribed linear, but the 5' and 3' termini are then covalently linked together post-transcriptionally. Recent studies have shown that splicing aided by inverted repeat elements in the upstream and downstream region contributes to the biogenesis of circRNAs [22], yet the exact mechanism remains to be determined.

1.2.3 Transcriptional regulation

Gene expression regulation at transcriptional level contains many key steps. First, an open chromatin structure must be present to make DNA accessible to RNA polymerase. The DNA accessibility can be regulated by modulating the extent of CpG methylation on DNA and the combination of various chromatin modifications. Second, transcription factors bind to promoter region of the DNA (such as TATA box) to recruit RNA polymerase for transcription, or to enhancer region to augment the transcription activity, or to silencer region to block RNA transcription. Different promoter binding sites could give rise to several 5' sequences of transcripts, especially in the case of first exon, and the variable sequences harboring regulatory elements might further contribute to post-transcriptional processing and translational efficiency [23]. Third, topoisomerases modulate the overwinding or underwinding of the DNA that is crucial for transcriptional elongation [24]. Fourth, alternative splicing could not only generate transcript isoforms that encode different proteins, but also influence the fate of transcripts via inclusion or exclusion of regulatory elements [25]. Fifth, alternative transcription termination generates transcripts with different regulatory elements in their 3' untranslated region (3' UTR). Those elements, upon being bound by various trans-factors such as miRNAs or RBPs, can modulate the stability of the transcripts in post-transcriptional regulation [26], [27].

1.2.4 Post-transcriptional regulation

Post-transcriptional regulation is the modulation of gene expression on the level of RNA transcripts, including mechanisms involving ncRNAs and RNA

binding proteins (RBPs). MicroRNAs can guide the RNA-induced silencing complex (RISC) to the target transcripts in order to repress the translation efficiency and RNA stability [28]. Small interfering RNAs (siRNAs) can guide RISC to the target RNA to degrade it [29]. PiRNAs can lead PIWI proteins to transposable elements and modulate their expression [30]. RNA editing enzymes (such as cytidine and adenosine deaminase) could alter the function and stability of RNAs by changing specific nucleic acid residues [31]. Various RNA binding proteins can alter the stability or localization of their targets in an allosteric manner. For example, cytoplasmic poly(A)-binding protein (PABPC) can bind to the poly(A) tail thus facilitates the protein translation and increases the transcript stability [32]. Zip code binding protein (ZBP) binds to zip-code sequence in the 3' UTR of transcripts and directs the transport of the target RNAs to distal dendrites [33]. Staufen proteins bind to the double-stranded structure of RNAs (either intra-RNA or inter-RNA) and mediate the decay of the non-translating RNAs [34]. NcRNAs can regulate the stability of mRNAs by competing for the cofactors via various mechanisms. For example, Pten-p1, one pseudogene of mRNA Pten, can decoy the miRNAs targeting Pten due to their sequence similarity. Augmentation of Pten-p1 abundance leads to increased Pten mRNA stability and its protein expression [35]. CircRNAs can sponge and/or transport miRNAs and RBPs due to the circular structure and innate stability. For example, Cdr1_as can sequester about 74 miR-7 molecules [36], [37]. Moreover, the interaction among coding genes, non-coding genes and RBPs can form a multi-layer regulatory network.

1.3 Opportunities and challenges

The advance of massive parallel sequencing technology in the past decade not only enables the genomic identification of thousands of organisms, it also allows us to profile the more complex transcriptome of any given cell population in a cost-effective yet unbiased manner. With much-anticipated prospects, many fundamental questions remain to be answered: (1) How many genes are there in the genome of human (or any other organism of interest) and what are they? (2) What are the full-length sequences of each

and every of transcripts in the genome? (3) How many functional isoforms are there per gene locus? (4) What are the differences among those transcript isoforms in terms of coding and regulatory potential? (5) Are there more types of ncRNAs? (6) What are the regulatory mechanisms of the ncRNAs and what are their cofactors? (7) How does the transcriptome, including coding and non-coding RNAs, change dynamically during development or exposure to stimuli? (8) To what extent the expression variation allows robustness other than pure noise? (9) What is the architecture of the gene regulatory network containing layers of DNA, RNA and protein? (10) What is the capacity of the regulatory network and how will it react when challenged? (11) How can we translate the knowledge learnt from the regulatory network to medical applications?

To address these questions, we first need to have a comprehensive understanding of the transcriptome landscape, as in my belief that all phenotypes should have distinct manifestations in transcriptome. More specifically, we need to scrutinize every aspect in the RNA biology, with customized tools for specific questions. Only after that, a unified framework could be established for data integration, information extraction and hypothesis testing, and is scalable and fit to biological questions.

1.4 Thesis objective and structure

Objective

With the advance of sequencing technology and the general research interest on gene expression regulation, the number of genome-wide studies keeps increasing. Due to the different nature of each individual study, customized analysis has been in much demand to draw proper conclusions, which requires a constant bidirectional communication between experimental and computational parts throughout the study. The general goal of this thesis is to present examples in which tailor-made analysis reveals biological insights either by developing new tools or by scaffolding a pipeline that makes better

use of the existing tools. Furthermore, the approaches developed in those case-by-case studies can be used in conjunction in order to shed light on a more comprehensive understanding of biological phenomena.

Structure

Chapter 2 describes a hybrid approach to assemble the transcriptome. This pipeline bridges two different types of datasets and two assembly techniques.

Chapter 3 describes an approach to directly characterize gene transcripts in full-length. The highlight of this approach lies in the fact that the detailed information of the transcript, including the exact sequences of 5' UTR, 3' UTR and the combination of alternative exons, can be experimentally determined instead of being estimated as in methods described in Chapter 2.

Chapter 4 describes a method for miRNA identification and validation independent of reference genome sequences. Rich information regarding to the biogenesis and post-transcriptional processing of miRNAs can be extracted from the parallel sequencing approach.

Chapter 5 describes a framework for the identification and functional analysis of circular RNAs. We further demonstrate that circRNAs are enriched in brain and their functional relevance is examined.

Chapter 2

De novo transcriptome assembly

2.1 Introduction

A catalogue of RNA transcripts of organisms is the most important resource in molecular biology nowadays. Only a handful of eukaryotic organisms, including human and *C. elegans*, have high-quality transcriptome reference, whilst the genomic sequences of the vast majority of organisms are still in a rather immature state, leaving their often-predicted transcriptome models even more error-prone. A flatworm named *Schmidtea mediterranea* (hereafter refereed to as Smed) is one of the organisms with many spectacular features and biomedical implications (such as having a strong regeneration potential) and yet the study of which has been impeded for years due to the lack of high-quality transcriptome reference. The current gene annotation in the Smed genome is far from complete, and is largely based on computational predictions complemented with partial supporting evidence from EST libraries [38]. In the Smed genome database (SmedGD, http://smedgd.neuro.utah.edu/), 30,930 "MAKER" transcripts from 30,333 genome loci were predicted [39]. Many of these gene transcripts await further validation, and the number of missing transcripts in SmedGD is unknown.

In order to provide a much-needed high-quality resource of the Smed transcriptome annotation, we developed a general strategy for sequencing and assembling of a complex transcriptome without relying on the availability of the genome sequence. The strategy described here is of great practical importance for decoding the transcriptome of complex organisms since genomes are known to be extremely difficult to assemble. The major reasons of these difficulties are polyploidy and low complexity. Recently, first attempts have been made to sequence and assemble the transcriptomes of animals

such as the butterfly [40], coral [41], and whitefly [42]. However, in all cases the mean length of assembled transcripts (197 nt for butterfly, 266 nt for whitefly, and 440 nt for coral) was substantially shorter than the estimated average mRNA length (>1000 nt). We demonstrated that the assembly performance can be improved with the following two strategies: (1) the combination of complementary sequencing technologies that provide either long and relatively few sequencing reads (454 Life Sciences [Roche] technology), or short and magnitudes more sequencing reads (for example, Illumina technology); (2) the efficient normalization of cDNA libraries prior to sequencing. Transcripts of high abundance occupy the majority of sequencing capacity, limiting the detection sensitivity for rare but functional important transcripts. This bias can be solved to a large extent by normalizing the cDNA libraries before sequencing. The combination of both strategies leads to the success of the mRNA transcriptome characterization of Smed for the first time in a genomic information-independent fashion.

2.2 Methods

2.2.1 Library construction and normalization

The construction of normalized full-length-enriched cDNA libraries works in three steps: (1) the synthesis of double strand cDNA via RACE (Rapid Amplification of cDNA Ends) technique, (2) the removal of poly(A/T) tails followed by the ligation of a DNA adapter, and (3) the normalization of the resulting cDNA library using duplex-specific nuclease (DSN). The DSN normalization method is based on the denaturation–reassociation of double-stranded (ds) DNA coupled with the preferential degradation of the ds-DNA fraction formed by abundant transcripts [43].

2.2.2 Sequencing protocol

The 454 sequencing library was prepared from 5 μg normalized cDNA library, and then sequenced by 200 cycles on a 454 GS FLX sequencer per manufacturer's instructions. Five micrograms of normalized cDNA library was

used to construct the single-end Illumina sequencing library, which was then sequenced by 36 cycles on the GAIIX per manufacturer's protocols. In parallel, 300 ng of poly(A) RNA was used to construct paired-end Illumina sequencing library, which was then sequences by 2x76 cycles on the GAIIX per manufacturer's protocols.

After sequencing, the PCR primer sequences have to be trimmed off for both 454 and Illumina reads, otherwise they would be treated as part of the RNA transcripts and would lead to false assembly. Since the 5' and 3' PCR primers for the 454 full-length cDNA library are different, the position and the sequence of the recognized primers can be used to determine the strand information of the read and assembled transcript. A small portion (0.8%) of the 454 sequencing reads were discarded as the PCR primer sequences were detected in the middle of reads and therefore likely to be the PCR artifacts.

2.2.3 De novo transcriptome assembly

There are two general methods for sequence assembly: (1) Overlap Layout Consensus (OLC) method and (2) De Bruijn Graph (DBG) method. As the name suggests, the OLC assemblers work in three steps. First, short overlapping sub-strings are located in an all-against-all comparison, with a hashing indexing method to increase the efficiency. Using the index information, reads can be grouped into different bundles, where a bundle consists of reads that overlap and only overlap within the bundle consistently. A layout is then generated for each bundle by weighted voting. Second, layouts are optimized through all-against-all comparison, breaking into contiguous parts (contigs) at repetitive or common regions, re-joining of contigs using contigs-overlapping reads (guaranteeing read coherence) and a quality control step (with minimal length and number of supporting reads). Finally, a consensus sequence (isotig) is generated for each of the refined layouts, representing one RNA transcript.

The OLC method works well for sequencing datasets of high quality, low-to-

medium throughput and long read length. Due to the all-against-all comparison, it would become computational unfeasible as the number of input sequences increases [44]. On the contrary, the DBG based assembly method suits better for the high-throughput sequencing datasets.

The DBG assembly works in the following steps. First, all sequencing reads were split into substrings of a given length (k-mers) and added to a de bruijn graph, with each edge represents a k-mer that connects the two nodes representing k-1-mer prefix and suffix. For example, an edge of “ABC” connects node “AB” and “BC”. Then, a Eulerian trail of the graph can be found in linear time, which might represent one RNA transcript. Last, all possible Eulerian trails of the same DBG are compared and some trails satisfying certain criteria, such as parsimoniousness, are reported as transcript sets for a gene locus. While DBG can easily scale up with large number of input sequences, it has a drawback of loss of the read coherence, which might lead to false assembly for complex gene loci.

We established a hybrid strategy that integrated two assembly methods designed for different types of sequencing data. In practice, the Illumina paired-end and single-end reads were first assembled into contigs using SOAPdenovo [45] (a DBG method, with parameters tuned best for Smed: max_rd_len=100, agv_ins=200, reverse_seq=0, asm_flags=3, re_len_cutoff=3, pair_num_cutoff=3, map_len=25, K=25). Then, the contigs longer than 100 nt (with an expectation of median length of 1000 nt in mind) from the Illumina assembly were combined together with 454 reads for the final assembly using Newbler (a OLC method, version 2.3, Roche; with parameters tuned best for Smed: -cdna -it 100 -ig 500 -icc 100 –icl 3).

2.2.4 Redundancy filtering

In order to remove redundant transcripts and retain a set of potentially unique isoforms, the following procedure was applied: The mutual overlap of candidate transcripts was determined by running BLAT [46] with default

parameters on all possible pairs. From each pair, the shorter of the two transcripts was discarded if the number of nonaligned nucleotides fell below a threshold of 35 nt (corresponding to shorter than 95% of all exons) and if the longer one was not discarded previously.

2.3 Results

2.3.1 Sequencing and transcriptome assembly

We established a hybrid strategy that finalized a polished transcriptome reference based on the integration of two independent assembly methods designed for different sequencing technologies (Figure 2.1). At first, poly(A) RNA extracted from Smed worms was used to constructed a full-length cDNA library, which was then normalized using a duplex-specific nuclease. We sequenced the normalized library on 454 GS FLX platform and obtained 1,370,473 reads with a median length of 340 nt that passed the 454 quality filter (Table 2.1).

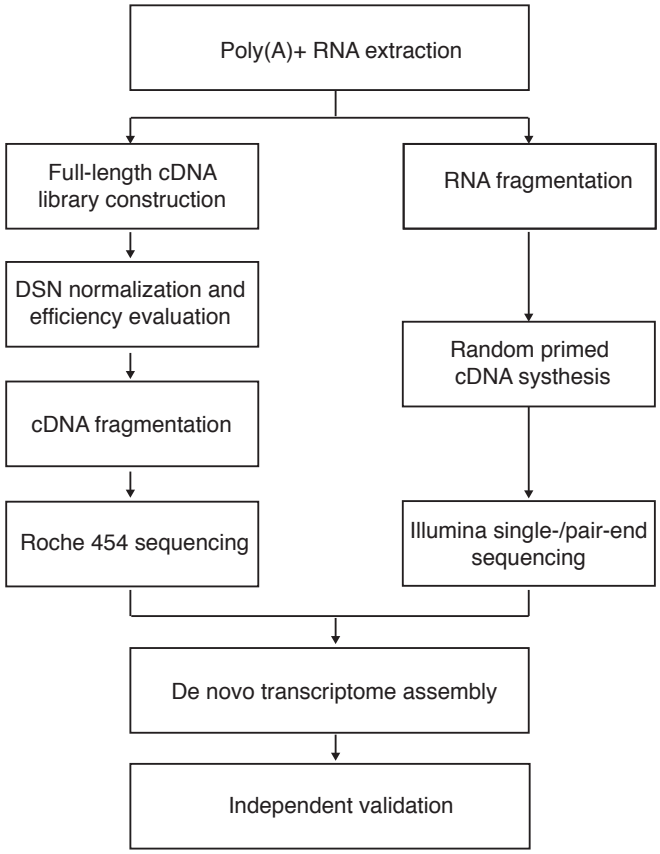

Figure 2. 1 Scheme of hybrid sequencing and assembly

	454	Illumina Paired-End (PE)		Illumina Single-End (SE)	
	454	Rep1	Rep2	Norm	Non-norm
Number of reads	1,370,473	29,009,277	27,560,500	9,043,682	11,204,306
Read length	Median 340 nt	76 nt Insert size of 200 nt	76 nt Insert size of 200 nt	36 nt	36 nt

Table 2. 1 Summary of sequencing results

After trimming off the 454 sequencing adapters and the cDNA PCR primers, strand information of all 454 reads was determined and 454 reads were used as input for the Newbler assembler, a 454 proprietary software distributed together with the 454 sequencing machine. The majority of reads (83.99%) was successfully used for assembly and resulted 24,630 contigs with a median length of 953 nt and a maximum length of 7,009 nt (Table 2.2). Due to the low efficiency of the reverse transcription (RT) efficiency for long transcripts, the assembled transcripts longer than 6 kb were under-represented in our full-length cDNA library. To compensate for this, we constructed another cDNA library, in which random priming was used to reverse transcribe the fragmented poly(A) RNA (Figure 2.1), at the cost of losing the strand information. To increase the effective length of the sequencing reads, we sequenced from both ends of this library using the pair-end sequencing method. We obtained 28 million read pairs, each consisting of two reads of length of 76 nt (Table 2.1). Using the SOAPdenovo software, these read pairs together with 20 million single-end sequencing reads from the full-length cDNA library were corrected for sequencing errors using a k-mer frequency method prior to assembly. In total, we obtained 41,501 contigs that are longer than 100 nt, with the median and maximum length of 252 nt and 14,628 nt respectively. To expand the 454-only assembly, we used 454 reads together with processed Illumina contigs as input for a final assembly using Newbler (Figure 2.1). This hybrid assembly transcript set (hereafter referred to as “HA transcripts”) consisted of 26,669 contigs (from 14,793 loci) with a drastically enhanced median length of 1,107 nt. Although the maximum length of 14,740 nt (partial Titin transcript, note that the complete Titin

transcript in human is of length of 65535 nt) indicates that the some of the longest transcripts are still missing in the assembly, it matches with the expected length of the full-length cDNA library. Identifying the full-length sequences of the extremely long transcripts would be more suited to carry out in a gene/transcript-specific fashion to ensure the quality.

	Number of transcripts	Mean length (nt)	Median length (nt)	Max. length (nt)
454	26,904	1,048	953	7,009
Illumina	41,501	451	252	14,628
454 + Illumina	26,669	1,300	1,107	14,740
HA transcripts (Final assembly)	18,618	1,228	1,078	14,740

Table 2. 2 Summary of *de novo* assembly results

2.3.2 Enhanced sensitivity from cDNA normalization

To demonstrate that library normalization dramatically increases the probability of identifying lowly expressed transcripts, we did simulations to examine the percentage of the HA transcripts that could be recovered from the non-normalized and normalized cDNA libraries. The abundance of the HA transcripts represented in the non-normalized and normalized library was estimated form the single-end Illumina sequencing of the two libraries. Out of 18,618 HA transcripts, 16,734 with measurable expression (RPKM > 1) in both the non-normalized and normalized library were used for simulation. A simulated cDNA pool was first generated containing a total of 10 million molecules for each library, where the number of molecule per transcript is proportional to its RPKM value. Simulated 454 read pools then were generated by sampling cDNA fragments from each cDNA pool.

We then randomly sampled 0.5, 1, 1.5 and 1.95 M (million) reads from the read pool for assembly. The assemblies were then aligned to the HA transcripts to see how well they could recover HA transcripts, at minimum

percentages of transcript length of 50%, 75% and 90% (Figure 2.2). The assembly generated from the normalized datasets recovered much more HA transcripts than those from the non-normalized datasets (Figure 2.2). For example, at the depth of 1.5 M simulated 454 reads, an additional 60% of HA transcripts can be assembled at almost full length (>=90%) with the normalized library compared to the non-normalized library. It suggests that transcriptome assembly benefits greatly from normalization strategy.

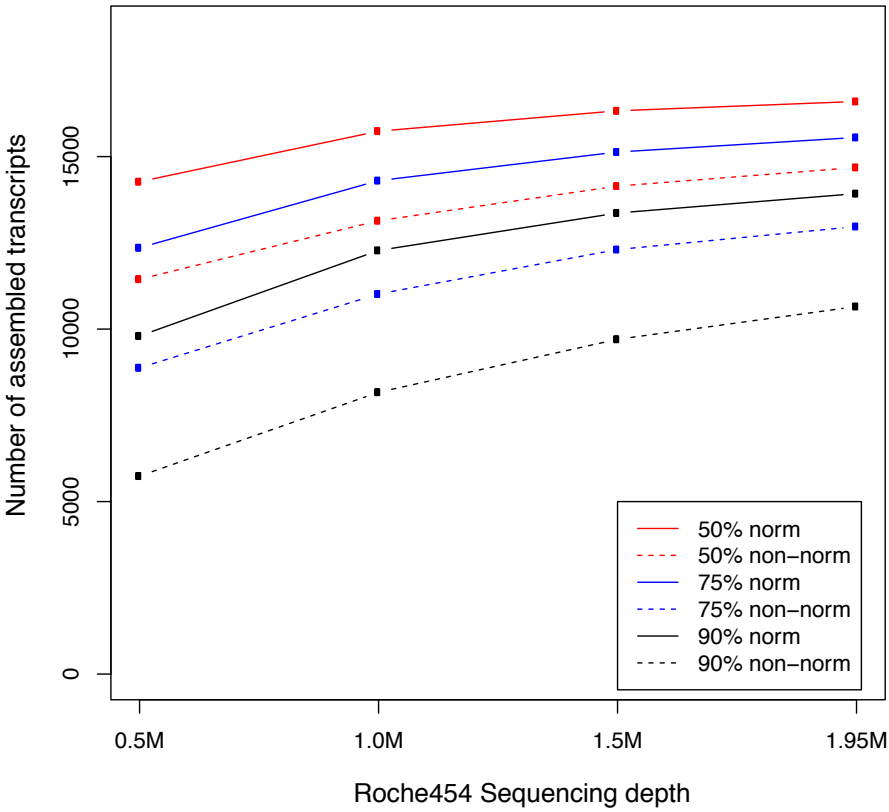

Figure 2. 2 Library normalization enhances transcript recovery. X-axis represents the number of simulated 454 reads used in the assembly simulation. Y-axis marks the number of HA transcripts that can be recovered at certain minimum percentage (red for 50%, blue for 75% and black for 90%) from the normalized (solid line) or non-normalized (dashed line) cDNA libraries.

2.3.3 Cost-efficient hybrid assembly

It is an important and yet difficult question to find the optimal mixture of the two sequencing technologies, considering the length and the number of sequencing reads and the available budget. Simulation assembly using subsets of our 454 and Illumina sequencing data demonstrated that long read length is crucial to the completeness of assembled transcripts (Table 2.3). However, the cost of one 454 run is similar to that of two Illumina lanes. Of note, the most efficient strategy depends on the landscape of the expressed transcripts in different organisms, therefore there might not exist a single best-balanced combination.

Data	No. of isotigs	Median Length	Max Length	Recover 50%	Recover 75%	Recover 90%	Recover 95%
454 1-lane	13460	797	4810	53.53%	37.06%	24.09%	16.03%
454 2-lanes	18987	917	5410	74.43%	60.91%	45.96%	36.13%
454 3-lanes	23585	982	6352	87.67%	80.11%	71.03%	63.44%
Illumina 2-lanes	52906	206	6774	67.55%	41.95%	17.68%	8.68%
454 1-lane + Illumina 2-lanes	18494	850	6770	76.17%	61.41%	41.55%	27.81%
454 2-lanes + Illumina 2-lanes	22155	951	6770	88.45%	80.40%	65.55%	53.02%

Table 2. 3 Influence of sequencing depth on transcript recovery

2.3.4 Assembly quality evaluation

We aligned the HA transcripts to the draft reference sequences of the Smed genome using Blastn [47]. The transcripts covered 7.37% of the genome references and 154 (0.6%) of them cannot be aligned to the draft genome. We then aligned all Illumina reads to the HA transcripts (as the transcriptome) and the genome references in parallel. Combining the two biological replicates,

there were 33.4 and 27.9 million PE reads aligned to the assembled transcripts and genome reference, respectively (Table 2.4). There were 3.0 million PE reads that can be aligned to the draft genome but not aligned to the transcriptome. We speculated that this small proportion (10.72%) of reads were likely to originated from lowly expressed genes as well as the intronic and intergenic regions, as a similar percentage is also found in RNA-Seq studies in human [48] or mouse [5]. On the other hand, 8.5 million of the PE reads were transcriptome specific, which can contribute to the transcribed regions that are absent in the current draft genome reference.

	PE rep1	PE rep2	SE rep1	SE rep2
Number of Reads passed filter	22,081,765	21,134,818	7,629,531	10,502,913
Number of Mapped to genome	14,325,352	13,631,467	4,301,018	5,833,220
Percentage of Mapped to genome	64.87%	64.50%	56.37%	55.54%
Number of Mapped to transcripts	17,127,363	16,290,632	4,681,661	6,235,694
Percentage of Mapped to transcripts	77.56%	77.08%	61.36%	59.37%

Table 2. 4 Comparison of HA transcripts with the draft genome reference

Moreover, 1% of the transcriptome-specific PE reads readily revealed un-characterized genomic sequences since one end of the pair were aligned to both HA transcripts and the current genome reference while the other end could only be aligned to our HA transcripts (Figure 2.3). As an example, we visualized the relationship between one HA transcript that encodes PIWI2 protein and the draft genome reference contigs (Figure 2.4). There were not only gaps in the genomic contigs, but also duplications likely owing to the incomplete genome assembly.

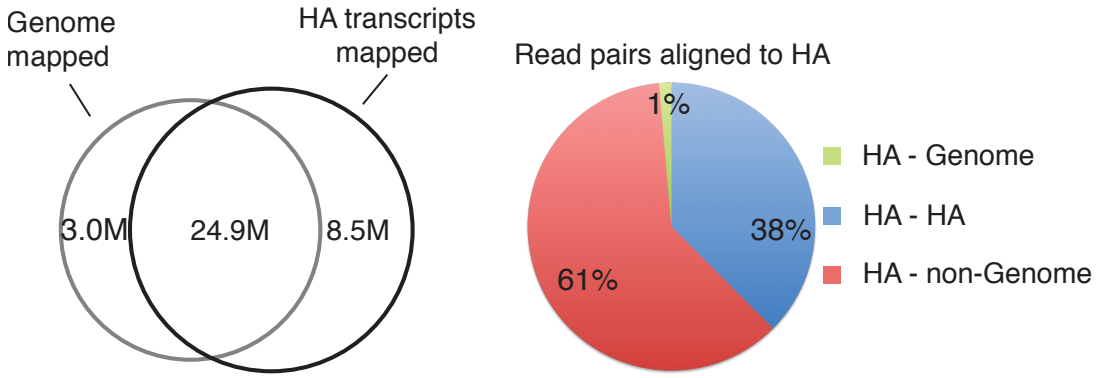

Figure 2. 3 HA transcripts expand the genome reference. Venn diagram to the left marks the number of pair-end reads aligned only to genome reference (3.0 million), or only to HA transcripts (8.5 million) or both (24.9 million). Pie chart to the right shows the classification of the 8.5 million pair-end reads according to the alignments of both ends: 1% (green) have one end mapped to HA transcripts and the other to genomic reference; 38% (blue) have both end mapped only to HA transcripts; the rest (red) have only one end mapped to HA transcripts and the other end unmapped.

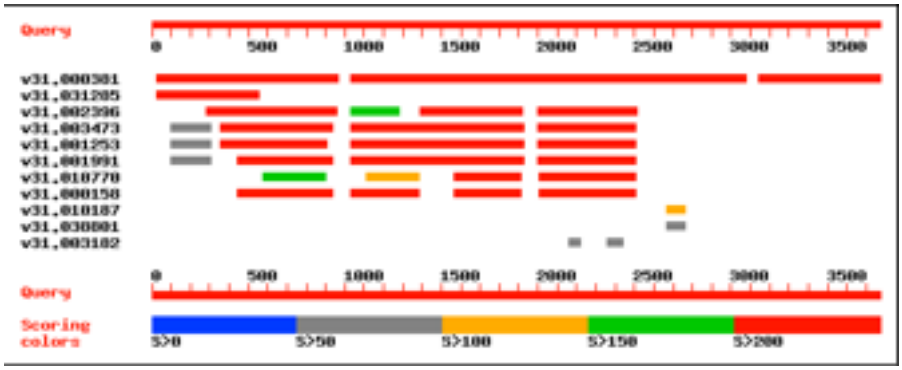

Figure 2. 4 An example of HA transcripts that connect genome scaffolds. The query track on top marks the HA transcript for Piwi2, which could be aligned to 11 genomic locations of Smed.

2.3.5 Transcript annotation

We annotated the HA transcripts according to their sequence homology to proteins identified in other organisms. Using Blastx at a threshold of 1E-10, 11,542 assembled transcripts could be aligned to NCBI non-redundant protein database (nr), with 2,101 additional ones could be further annotated when lowering the threshold to 1E-3. To assess the quality of HA transcripts

regarding to whether they could code full-length proteins, we calculated a C value (for Completeness) for each assembled transcript as the ratio between the length of predicted coding region and the length of the homologous protein (Figure 2.5). The majority of the HA transcripts are of relative high C value, indicating near complete assembly at least in terms of the coding sequences. Notably, not all these annotated assembled transcripts represent protein-coding genes, since they can be pseudogenes and processed transcripts that are highly similar to the homologous coding genes in sequence.

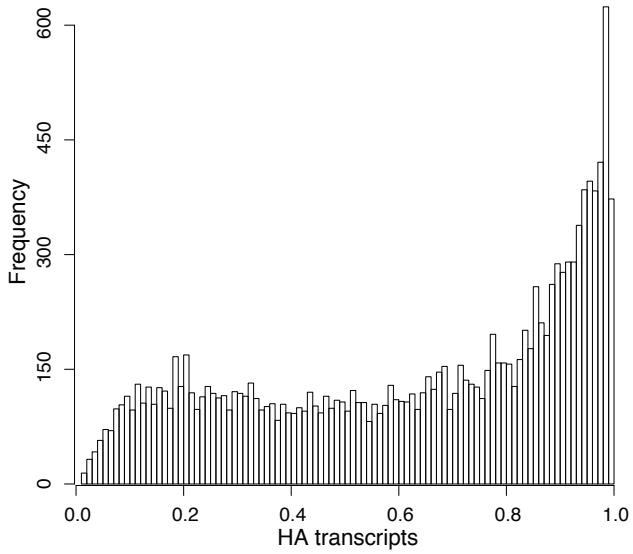

Figure 2. 5 The completeness of HA transcripts. The histogram of the C value (on x-axis) of HA transcripts indicates the majority of the HA transcript can code full-length annotated proteins.

To estimate the contribution of non-coding transcripts in our assembly, we computationally translated the transcripts into peptides. The coding potential of each transcript could be estimated by the ratio (denoted as P) between the length of the longest ORF and the length of the transcript (Figure 2.6). The higher this ratio is, the more potential a certain transcript could code a protein. There are 1828 transcripts with P value < 0.3, and therefore they are likely non-coding transcripts. Interestingly, there are about 1000 transcripts that are of > 0.9 P value but no homologue can be found in the NCBI nr database, which could be a set of Smed-specific protein coding genes. Interestingly,

transcripts with P value very close to one might suggest for either short UTRs or incomplete assembly.

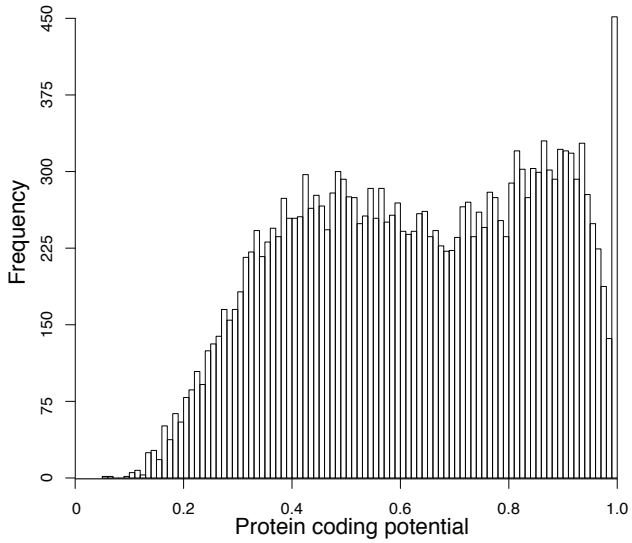

Figure 2. 6 The coding potential of HA transcripts. The histogram of the P value of HA transcripts indicates that about 10% of the HA transcripts are non-coding transcripts while the majority code for proteins.

2.3.6 Comparison to genome-guided transcriptome assembly

To further evaluate the accuracy of the *de novo* transcriptome assembly, we compared HA transcripts to a conventional genome-guided transcript assembly. First, all sequencing reads were aligned to the genome reference sequences using TopHat [49] (version 1.2.0). Then, Cufflinks [49] (version 0.9.3) was used to assemble transcripts using default parameters. A total of 32,235 transcripts (hereafter referred to as “Cufflinks transcripts”) were obtained, with a mean length of 903 nt and a maximum length of 18,510 nt (Table 2.5). We then aligned the HA transcripts with to Cufflinks transcripts in a mutually manner using BLAT and regarded those as similar if the two are of >90% sequence similarity over >80% of the shorter transcript length. With these criteria, 17,000 HA transcripts have counterparts with 24,443 Cufflinks transcripts. 1,619 and 7,792 transcripts remain to be HA and Cufflinks specific. Compared with HA transcripts, Cufflinks set contained much more shorter contigs.

	Number of transcripts	1stQ Length (nt)	Median Length (nt)	Mean Length (nt)	3rdQ Length (nt)	Max Length (nt)
HA	18,619	643	1,078	1,228	1,625	14,740
Cufflinks	32,235	370	654	903	1,163	18,510
Common-in-HA	17,000	683	1,123	1,335	1,670	14,740
Common-in-Cufflinks	24,443	422	766	1,018	1,333	18,510
HA-specific	1,619	483	631	792.7	993	4,077
Cufflinks-specific	7,792	288	450	543.7	679	6,271

Table 2. 5 Comparison of HA transcripts with Cufflinks transcripts

2.3.7 Validation using RACE

To validate the quality of our hybrid assembly with regard to the complete 5' and 3' ends of the transcripts, we performed 5' and 3' RACE on 24 randomly picked HA transcripts of high, moderate and low abundance, respectively. The majority (22 out of 24) of RACE products were successfully sequenced by conventional Sanger method. We confirmed that these 22 HA transcripts have complete or nearly complete 5' and 3' ends (at most 50 nt shorter than RACE products). For the remaining two, both 3' ends were validated by RACE but the 5' ends failed, which could be explained by the mild 5' bias in the cDNA cloning protocol [48]. In addition to the completeness of 5' and 3' ends, we also estimated the sequence accuracy of HA transcripts. Out of a total of 16,520 nt that can be aligned with our transcripts, we identified substitutions for 116 nt, and deletions/insertions affecting 9/24 nt and therefore the overall error rate is around 0.90%. With independent experimental validation, we demonstrated that the HA transcripts we assembled is of high quality.

2.4 Discussion

Genome annotation is traditionally based on sequencing cDNA libraries. The procedures of the traditional and yet still gold standard Sanger sequencing is both laborious and cost prohibitive for large sets of targets. Nowadays RNA-

Seq, as the second-generation technology, has mostly been applied to quantify the expression level of already annotated loci and to identify differentially expressed genes. RNA-Seq has also been used to refine annotated gene structures such as alternative splicing, alternative 5' and 3' ends, or to even build up gene models in a *de novo* manner [50]. However, all of these tasks rely on knowledge of genomic sequences, which, in many situations, is not available or very difficult to obtain. Here, we present an elegant method to obtain a high-quality characterization of a complex animal transcriptome without using genomic sequences. Foreseeable development and application of this approach could be: 1) use cDNA derived from specific cell types and development stages; 2) polish both 5' and 3' ends of the transcripts with special datasets designed for UTR analysis, since the UTRs harbor many regulatory elements that are important for post-transcriptional regulations and 3) augment the current transcript catalogue by identifying more comprehensive isoform sets, instead of only the dominant ones as achieved in the present study.

With the rapid improvement of Illumina sequencing technology, transcriptome assembly in DBG method is more popular. Many tools have bee developed recently, such as Oases [51], Trinity [52] and Scripture [53]. At the time of this study, Velvet [54] and AllPath [55] were tested for possible substitutes for SOAPdenovo, but they performed poorly (resulting in too many short and unconnected fragments) since they were designed for small genome assembly. PCR bias, highly similar domain and repetitive regions would lead to over-assembly. Moreover, it is difficult to *de novo* assemble complex gene loci, such as the Dscam gene that has theoretically 38,016 isoforms. Sequencing technologies that render long read length, such as Roche 454, PacBio and Nanopore, are therefore naturally better means in resolving transcripts that are long and complex. Roche 454 is the state-of-the-art and the only high-throughput long read-length sequencing technology available at the time of the project. Now, as Pacbio sequencing can sequence DNA as long as 30kb, it should be wise to replace 454 sequencing technique with it. In

fact, Pacbio sequencing can fully assembly the genome of a microbial [56]. It could be expected that in the near future long read-length sequencing methods would be widely used for transcriptome structure profiling, especially in the fields of cancer biology and neuroscience, where the exact sequence identity sometimes matters more than the quantity.

Chapter 3

Full-length transcriptome identification

3.1 Introduction

As transcriptome-wide studies have become pivotal in understanding biological processes, stark importance emerges on the task of cataloging a complete set of full-length transcripts in different cell types and conditions. Variants in both coding and non-coding regions affect the stability and functionality of the mRNA via post-transcriptional regulatory pathways. Cloning followed by Sanger sequencing works well for short single-isoform transcripts, as done traditionally a decade ago. For transcripts longer than 3 kb, it is still technically challenging to capture both 5' and 3' ends simultaneously. With special protocols such as SMARTer (Clontech), the power to capture full-length transcripts is greatly enhanced, with high-quality RNA as starting materials and applying 5' RACE and 3' RACE in conjunction. Yet, quantification of multiple isoforms from the same gene locus remains elusive using microarrays or shotgun sequencing due to the combinatorial problem. Many computational tools have been developed to tackle this combinatory problem, applying various heuristics. Cufflinks reports a minimum number of isoforms per gene locus that can best explain all the reads that are aligned to a reference genome [57]. Trinity reports a set of all possible isoforms supported by reads independent of reference genome sequences using de bruijn graph [52]. However, the performance drop significantly in the situation that there are more than one isoform transcripts [58]. Hybrid sequencing strategy (as described in Chapter 2) could render more reliable results when given additional information, comparing to current strategies based on single sequencing method. Given the noise introduced during the RNA extraction, library prep, shotgun sequencing and/or assembly procedures, the aforementioned methods could at the best estimate the

dominant transcripts. Single molecule sequencing is the only method, to date, that establishes proof instead of estimation of the full-length sequence of RNA transcripts. Single Molecule Real-Time (SMRT) sequencing technique developed by PacBio provides the sequencing capacity of over 3kb on average thus allowing direct characterization of full-length transcripts.

Nevertheless, PacBio technology also bears some innate pitfalls aside from its advanced design for sequencing long reads. First, although the long sequencing length allows the identification of full-length transcripts for many genes theoretically, error rate as high as 15% forbids its direct application to transcriptome profiling in practice. Second, the cloning efficiency is biased among transcripts of different lengths during reverse transcription, PCR amplification and sequencing step. Furthermore, due to the limited throughput (150,000 reads per sequencing run), transcripts of lower abundance are less likely to be detected, in sharp contrast to Illumina RNA-Seq (which provides 100-fold more coverage at the same cost). Therefore, optimizations in both experimental and computational analysis should be made to fully exploit the advantages of long read-length sequencing technologies.

3.2 Methods

3.2.1 PacBio SMRT technology

The PacBio sequencing is similar with Illumina RNA-Seq in the sense of sequencing-by-synthesis, but the implementation is vastly different as PacBio operates on a single DNA molecule whilst Illumina operates on a cluster of thousands identical DNA molecules [Table 3.1].

Platform	Illumina HiSeq2000	PacBio RS
Sequence depth	~150,000,000	150,000
Sequence length /nt	<= 150	<= 30,000
Insert size /nt	<= 800	<= 10,000
Cost per run/lane/Euro	~1000	~100
Error rate	< 0.5%	~ 13%
Starting material /ng	50	1000

Table 3. 1 Comparison of the performance between PacBio and Illumina sequencing technology

In brief, PacBio sequencing works in four steps (Figure 3.1). First, DNA template is prepared. The DNA template could be fragmented genomic DNA, or targeted DNA PCR products, or RT-PCR products of RNA samples. Second, the DNA library is generated by ligation of hairpin-shaped SMRTbell adapters to both termini of a double-stranded DNA (the resulting single-strand DNA circular molecule is named as SMRTbell). Third, sequencing primers are added to the cDNA library together with DNA polymerase. At last, once the polymerase binds to the sequencing adapter on a SMRTbell, it starts to extend the primer by synthesis in a rolling cycle fashion (images modified from the company website), emitting light that is recorded and analyzed to determine the nucleotide sequence. Here, nucleotides labeled with different fluorophores were incorporated into the newly synthesized DNA strand and the fluorescence is monitored by a two-photon CCD camera. The signal intensity, duration of the fluorescent pulse and the interval between pulses are measured to determine the DNA base identity. As the recording of the fluorescence is at real time, currently at 100 frames per second, the length of a stretch of the same nucleotides can be determined with high confidence.

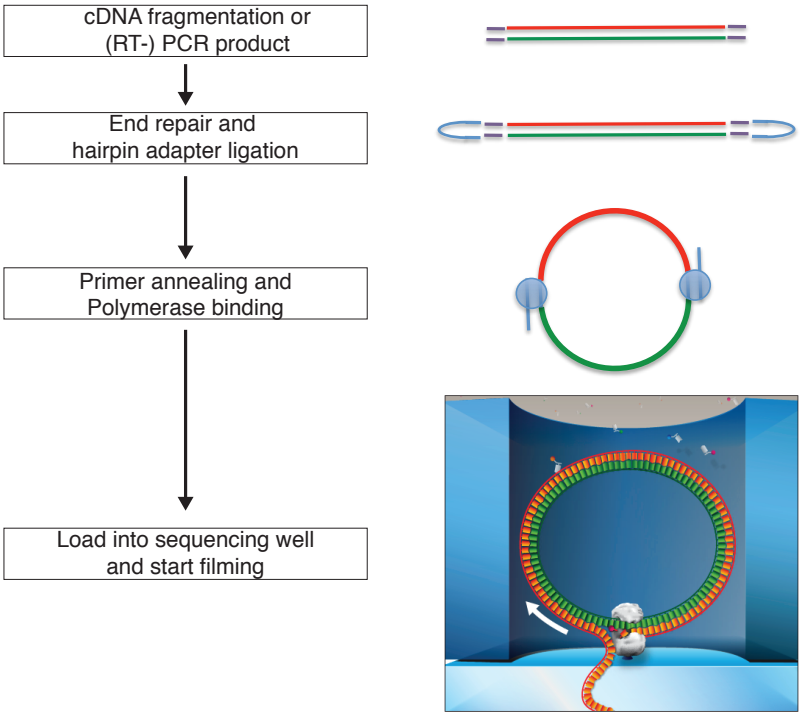

Figure 3. 1 Schematic workflow of PacBio sequencing

3.2.2 Experimental improvement

Due to the fact that RNA transcripts are of various abundance spanning over seven orders of magnitudes, a few highly abundant transcripts will saturate the sequencing capacity. Despite the potential influence of transcript length, the 100 most abundant genes constitute 15%~20% of the transcriptome. In order to increase the representation of the lowly abundant transcripts, we employed the DSN (Duplex-Specific Nuclease, see Chapter 2) normalization method to our full-length cDNA library.

Moreover, another modification in our experiment is made associated with different length of the transcripts, since it also influences the sequencing outcome. The longer the transcript is, the more difficult it is to be amplified in full-length during both RT and subsequent PCR due to a higher possibility that reaction enzymes drop off the templates. Furthermore, shorter cDNA are easier to be loaded into the sequencing well than the longer ones in practice. Therefore, unlike Illumina RNA-Seq, it is challenging for PacBio to sequence all transcripts at the same efficiency. In order to identify as many different transcripts as possible, we separate cDNAs into several groups according to their length so that the bias of length is reduced. We size-selected cDNAs on electrophoresis gel in the following four ranges: 0.3~1k, 1~2k, 2~3k and >=3k bps and sequenced them in parallel.

3.2.3 Error correction

In order to make direct usage of PacBio long reads, random errors introduced during the sequencing should be minimized to the extent below the thresholds as required in specific applications. For example, in the gap closure step in genome assembly, the error rate should be less than 5% where the PacBio long reads are used as scaffolds to bridge the assembled contigs [59].

There are currently two major computational approaches for error correction. The PacBio-alone method exploits the rolling-cycle amplification nature of the PacBio sequencing to reduce the randomly distributed errors. By generating

short DNA libraries (~400nt as recommended by the manufacturer), PacBio reads the DNA molecule back and forth several times during sequencing. Computing multiple sequence alignments between passes (the sequence between two adjacent SMRTbell adapters) followed by consensus calling could drastically reduce the error rate. This method is implemented in HGAP, and has been proved useful for bacterial genome assembly [56]. However, the performance relies on the number of passes. In this sense, the advantage on exceptional long reads sequencing is traded for higher accuracy.

The other approach that keeps the long reads advantage reduces the sequencing errors by using additional information as implemented in PacBioToCA [59], LSC [60] and LoRDEC [61]. These methods take advantage of the much higher quality and depth of the RNA-Seq reads. By aligning the short reads (from RNA-Seq, such as Illumina) to long reads (from PacBio), high quality local alignments can be made and then piled up to vote for a consensus. PacBioToCA is both slow and resource demanding due to the fact that it uses a modified version of blast for alignment and generates all-to-all alignments. LSC, as one of the tools dedicated to PacBio error correction, compresses the homopolymers before the alignment of short to long reads, based on the observation that most of the errors in PacBio is insertions and deletions. If an indel takes place within a homopolymers, LSC can easily spot it and correct it. However, if it happens on a nucleotide outside of homopolymer, or the indel is of several different nucleotides (heteropolymers), the region might not be corrected due to the difficulty in the alignment. LoRDEC builds a de bruijn graph using short reads first, then finds all possible paths that best support the long reads, finally reports the best path corresponding to the correct sequence. LoRDEC is much faster than PacBioToCA, but is still slow when dealing with complex transcriptomes. Therefore, none of these present tools could meet the demand of efficient error correction for PacBio sequencing on higher eukaryotes.

Here, we present a prototype of error correction pipeline named iPEC

(acronym “Illumina Pacbio Error Correction”). IPEC follows one very straightforward strategy: erroneous bases can only be corrected if they are covered by high quality alignments. The length of alignment is the most important determinant of the alignment quality. The long PacBio reads can be viewed as concatenations of good regions and bad regions, where the good ones have low error rates (<= 2% that can be tolerated by most aligners) and the bad ones bear higher error rates. The good regions can be locally aligned with short reads, supported by tools such as Bowtie and BWA. Because of the ambiguity of the borders between good and bad regions, alignments to the good region often cover part of the adjacent bad regions. Consolidations of the local alignments can correct not only the errors in the good regions, but also those belonged to the bad ones, thereby effectively shrinks the size of bad regions. Iteratively doing so could expand the good region to the whole long read. The rational behind this iterative correction is that, through prior rounds of correction, the error-free regions are expanded so that they can serve as longer seeds for subsequent alignment, in which the neighboring erroneous regions can be covered and thus be corrected. IPEC aligns all short reads (from Illumina) directly to long reads (from PacBio) and then retain only the local alignments of upper-quartile scores. The alignments are piled up to vote for consensus sequences. These consensus sequences, whose error-free regions are extended, are subjected to further rounds of correction. This iterative correction halts when either every base of the long read is covered or no more base covering could be made.

3.2.4 Transcript clustering

With long reads of improved quality, one can directly characterize transcripts. Instead of depending on likelihood inference, as done in the past decade on short-read sequencing data, now we can identify full-length transcripts by simply sorting and clustering the processed PacBio long reads. There are in general two approaches: genome-guided clustering and *de novo* clustering. *De novo* transcriptome clustering is the preferred option when there are no available genome reference sequences or the existing reference sequences

are of impoverished quality. The first step of *de novo* clustering is the assignment of the long reads into groups, where each group should corresponds to one gene locus or one gene family. This can be achieved by either performing an all-to-all alignment of long reads or using the short reads alignment information learnt from the error correction. It is advisable to do all-to-all mapping when the number of long reads is relatively small, otherwise, the memory consumption and the run time would be too much to be practically feasible. Pairwise alignments exceeding minimum thresholds of length and score are selected, and a greedy algorithm can be used to separate reads into groups. On the other hand, existing alignment between short reads and long reads (intermediate products of the error correction step) can be used to infer the grouping of the long reads. One long read can be represented by a set of non-redundant aligned short reads. Two long reads are classified to the same group if the overlap between their representative short reads is large enough and larger than all overlaps with other groups using a greedy algorithm. Noticeably, it is crucial to use the alignment of non-redundant short reads here, otherwise the results could be wrong due to the over-counting of the same sequences.

The second step of *de novo* clustering is to extract representative reads from the multiple sequence alignments within each group. To prepare for multiple sequence alignments, all reads from the same group should be checked for strandedness using the poly(A) tail or the stranded RNA-Seq information. Reads should be reverse-complemented if they are not of the sense strand. Highly similar reads were not reported if there were no difference in the internal sequence and the tandem difference at the 5' or 3' end was less than an arbitrary threshold such as 50 nt. External evidence such as the 3P-Seq for the exact 3' end or CAGE data for the 5' exact end can be added into this step in order to get a comprehensive and reasonable transcript sets.

Alternatively, genome-guided transcript clustering would be a better choice when the available genome reference sequences are of decent quality,

because of the simple fact that pairwise alignment is much easier to compute than multiple sequence alignments. Corrected long reads are aligned to the genome using splice-aware aligners, such as GMAP [62], and both the genomic coordinates and the exon structures are determined. Long reads are simultaneously separated into stranded, non-overlapping groups corresponding to gene loci. Identification of representative reads within each group can be done using either of the two following notions: intron notion or exon notion. In the intron notion, two reads represent two different isoforms if there is a different intron (marked by a pair of splicing junction sites). All alternative splicing events, including different 5' or 3' usage of internal exons, exon skipping and intron retention, are easily reflected on the set of splicing junctions one read contains. However, additional information must be taken into account to distinguish isoforms where tandem first exon (from alternative promoters) or last exon (alternative PAS) is used. On the contrary, exon notion does not have this problem, as both 5' and 3' of every single exon of each read are recorded. Reads with different sets of exon boundaries naturally represents different isoforms, although downstream analysis of alternative splicing would be less intuitive than using the intron notion. In either of these two notions, a read can be viewed as an ordered binary string of all possible coordinate pairs (junctions in intron notion or exon boundaries in exon notion) within the gene locus. After assigning the read with the highest number of non-zero elements in its string representation as the first isoform, a greedy algorithm can assign the new coming read either to a new isoform, or to one of the existing isoforms where all the elements are continuously matched. Special care is necessary for the terminal exons, where a minimum difference threshold should be set to report reasonable results. According to the study of alternative polyadenylation using 3P-Seq [63], the minimal distance between two adjacent PASs is about 50nt. It is reasonable to argue that a difference larger than 50nt in either 5' site of the first exon or 3' site of the last exon originated from different transcription events, and therefore they should be treated as two different isoforms.

3.3 Application on drosophila Dscam gene

3.3.1 Fly Dscam gene has 38016 isoforms

The Dscam gene (Down syndrome cell adhesion molecule) encodes a family of transmembrane proteins that play profound roles in both neuronal patterning and pathogen recognition. In human, the gene locates in the Down Syndrome critical region of chromosome 21 [64] and its over-expression contributes to the physiological defects in Down Syndrome [65]. Whilst the human copy of Dscam encodes only three transcript isoforms, its homolog in Drosophila melanogaster (fruit fly) could theoretically encode a maximum of 38,016 isoforms. The isoform diversity of the Dscam gene locus is huge, bearing in mind that there are merely 17,321 genes in the entire fly genome (Flybase version 6.02). As shown in Figure 3.2, out of 115 exons contained in the Dscam gene locus, 20 are constitutive, and the rest 95 are separated into the following four clusters. Twelve variable exons make up exon cluster-4, 48 for cluster-6, 33 for cluster-9 and 2 for cluster-17. Exon cluster 4, 6, and 9 encode three ectodomains (immunoglobulin domains with specific binding capacity) and cluster 17 encodes a transmembrane domain. Assuming a mutually exclusive splicing pattern within the ectodomains, every isoform contains only one of the variable exons from each exon clusters, leading to a combination of 19,008 different ectodomains. Furthermore, exon skipping can contribute to at least four additional isoforms [66]. Chemoaffinity experiment demonstrated that the identities of the ectodomains are crucial for neuronal self-recognition, as each isoform binds to itself but only rarely to other isoforms [67]. Genetic experiments have shown the importance of maintaining a sufficient number of isoforms for synaptogenesis and neuronal guidance [68].

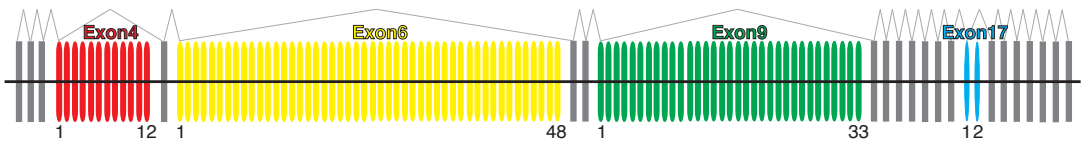

Figure 3. 2 Structure of Dscam gene. Each vertical bar represents one of the 115 exons, with color grey represent constitutive exons; and color red, yellow, green and blue for exon-4, exon-6, exon-9 and exon-17 respectively.

The full configuration of Dscam splicing pattern has been under scrutiny since 2001. Single-strand conformation polymorphism followed by Sanger sequencing [69] and customized microarray [70] were employed to study the single exon usage. However, the major limitation is that the usage of exon4, 6 and 9 cannot be measured simultaneously. Instead, the abundance of each isoform was inferred based on the assumption that the alternative splicing occurs independently at each exon cluster. As a result, the conclusion of "stochastic yet biased expression of Dscam splice variants" [70] is merely a reflection of this assumption. Microarray can be substituted by standard short-read NGS, but again, the sequencing length of NGS is not long enough to identify the combination of exon4, exon6 and exon9. To address the issue in an unbiased manner, we developed two methods characteristic of full-length, high-throughput profiling of massive Dscam splicing variants.

3.3.2 Direct sequencing of Dscam ectodomains using PacBio

The long read length feature of PacBio technology enables direct sequencing of the variable ectodomain part of Dscam cDNA, as the sequence from exon4 to exon9 is about 2 kb in length. The cDNA product, hereafter referred to as 2kb-cDNA, can be generated as illustrated in Figure 3.3 and directly sequenced using PacBio RS system.

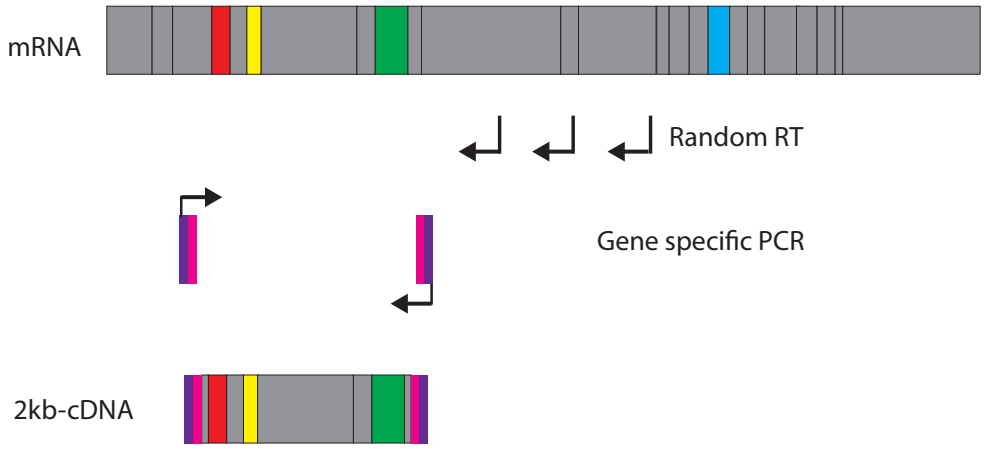

Figure 3. 3 PacBio sequencing of Dscam ectodomain. The scaled Dscam mRNA is depicted at the top, omitting the poly(A) tail to the right. The 2kb-cDNA is generated

by random primed reverse transcription followed by gene specific PCR, the primers of which targets exon3 and exon10.

The PacBio reads can be aligned to reference sequences to identify the combinations of the alternative exons in cluster 4, 6, and 9. As the minimum sequence difference among the exons from the same exon clusters is 18%, a read could only be confidently assigned to its true origin if the error rate is less than 9%. For example, assume Read-1 is actually originated from Isoform-1 with an error rate of 15% and there exists an Isoform-2 with an error rate of 20% compared to Isoform-1. After aligning Read-1 to all 19,008 possible isoforms, we found the error rate is 15% between Read-1 and Isoform-1 but only 10% between Read-1 and Isoform-2. Therefore Read-1 could be erroneously assigned to Isoform-2. CCS (circular consensus) reads satisfy the requirement on error rate (median of 2.7%) whilst the raw reads do not (median error rate of 11%). To compensate for the loss of non-CCS reads, we performed 10 PacBio sequencing runs for the S2 cell culture. The CCS reads were aligned to Dscam exon references using BLAT with a very sensitive parameter setting (tileSize = 8; stepSize = 5; oneOff = 1; minScore = 20; minIdentity = 70). After the alignment, high quality reads were selected using the following criteria: 1. Exon4, exon6 and exon9 can all be unambiguously identified; 2. Each exon (from exon3 to exon10) appears at most once and their order in the read is the same as that in the Refseq mRNA.

A total of 137,676 high-quality PacBio reads support 4,666 isoforms in S2 cells. The dynamic range of the abundance of the isoforms spans over three orders of magnitudes, where top 100 isoforms accounts for 25% of all reads. Some alternative exons were used more frequently than others. For example, more than half of the isoforms contain exon9.6, which is consistent with early single-exon studies [68]. Now, the hypothesis of the independent splicing choice among exon clusters can be tested by comparing the observed frequency to the expected frequency. Noticeably, a great advantage of the PacBio technology applied here over both array-based and standard NGS

technologies is that PacBio technology strictly renders uniform coverage along transcripts, which is an important yet often violated assumption for the other technologies. Assuming independent splicing choice among exon clusters, the expected frequency of one particular isoform can be calculated as the product of the expected frequencies of its exon4, 6 and 9. As shown in Figure 3.4, the independent hypothesis is strongly supported by the fact that the variance in expected frequency can explain 83% of the variance in the observed frequency (p-value < 2.2E-16 in a paired Pearson correlation test).

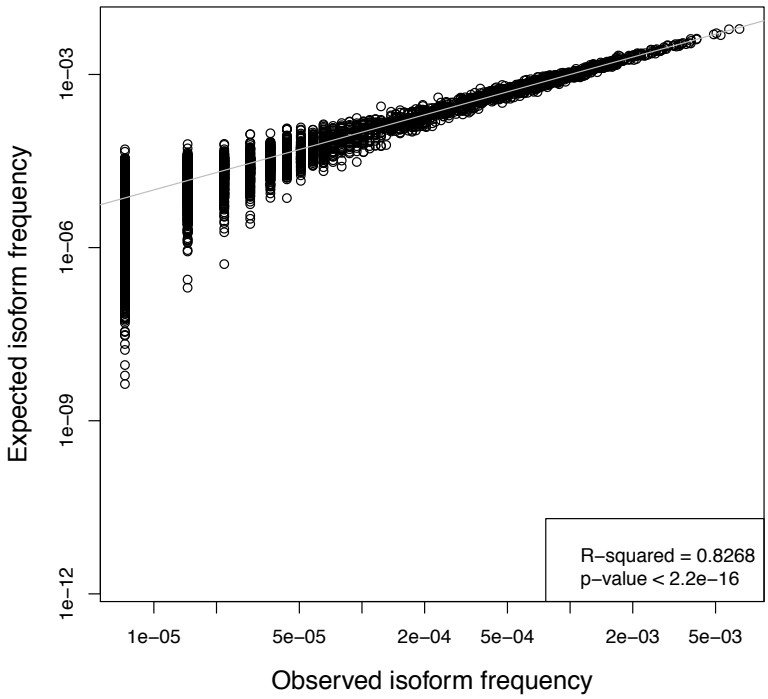

Figure 3. 4 PacBio sequencing supports independent splicing model. Each circle represents one Dscam ectodomain isoform. The value on X-axis marks the observed isoform frequency from PacBio sequencing data, and the value on Y-axis marks the expected frequency assuming independent splicing model. The grey line marks the diagonal, and the linear regression results are shown in the bottom-right corner.

3.3.3 Fly Dscam isoforms do not respond to immune challenge

As one of the proposed functions of Dscam being pattern recognition receptor in the immune system, Dscam was shown to be upregulated and its isoforms

were differentially expressed upon exposure to parasites [71]. Loss of Dscam resulted in the deficiency of uptake of pathogenic bacteria [72]. However, there is little evidence that the specificity of Dscam isoforms contribute to the immune response. We therefore repeated the study in order to measure the possible differential isoform regulation upon immune challenge. Here, adult flies were exposed to *E. coli* for 18 hrs and showed immune activation, since the gene expression of antibacterial peptide (AMP) was elevated, serving as a molecular marker for immune activation. However the expression of Dscam remained largely unchanged (Figure 3.5 A). Similar phenomena were also observed on S2 cells treated with E. coli for 12 or 18 hrs (Figure 3.5 B). Then what about the relative representation of each individual isoforms? With more than 30,000 sequencing reads per sample, a total of 3,885 different isoforms can be detected. As shown in Fig 3.5 C, there was no significant difference between the control and the treated samples. Admittedly, the sequencing depth is too low to capture all isoforms, however, it is still sufficient to test whether the isoform distribution changed between the control and treated groups. Therefore, we concluded that the stimulation of *E. coli* in flies did not lead to elevated Dscam gene expression or a detectable change in alternatively spliced exons or isoforms.

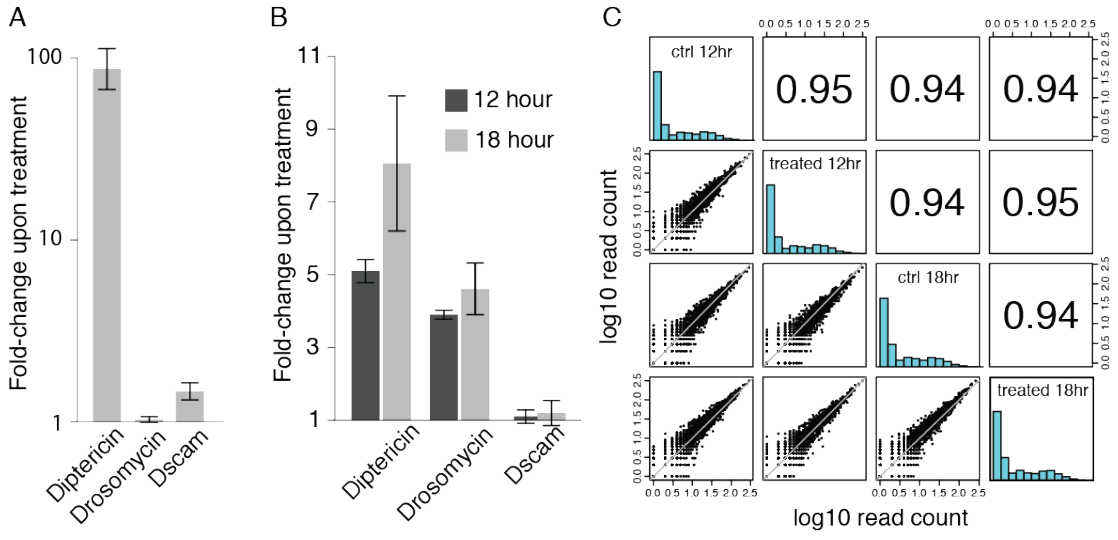

Figure 3. 5 Dscam isoforms do not respond to immune challenges. (A) At 18 hours after challenging with *E. coli* to living flies, Diptericin increased drastically, whilst Drosomycin and Dscam remained largely unchanged. Biological replicates n=2, each with 20 flies. (B) At 12 and 18 hours after challenging with *E. coli* to S2 cells, both

Diptericin and Drosomycin increased substantially, whilst Dscam remained unchanged. Biological replicates n=3. Error bar denotes standard error. (C) Pairwise comparison of Dscam isoform abundances between control (12hr), treated (12hr), control (18hr) and treated (18hr) in the left-lower parts. Pearson correlation coefficients are shown on the right-upper parts.

3.3.4 Discussion

Direct identification of the Dscam ectodomains by PacBio sequencing demonstrated that, as proof of principle, the splicing choice between alternative exon clusters is largely independent. However, the sequencing depth was not enough to profile all isoforms, detecting less than 25% of 19,008 possibilities. The limitation mostly came from the relative high cost of current PacBio technology. To quantify Dscam isoforms more precisely, we then developed a highly cost-effective customized method, namely CAMSeq (Circularization Assisted Multi-Segment Sequencing), which sequences exon4, 6 and 9 simultaneously using Illumina GAII. Unlike direct sequencing using PacBio, CAMSeq is indirect in the sense that it requires manipulation of cDNA (circularization) before sequencing, which could introduce noise such as intermolecular chimeras (at an average rate of 0.97% estimated by the embedded barcode system). However, CAMSeq is tailor-designed for Dscam gene locus that focusing only on the combination of the three mutually exclusive exon clusters and therefore not suitable for genome-wide transcript profiling. On the contrary, with the capability of rendering much longer read length, PacBio is the promising technology that allows a genome-wide full-length transcriptome profiling.

3.4 Application on rat transcriptome

3.4.1 Introduction

As a pilot task in a still on-going project on gene regulation in brain using rat as a model organism, we would like to first improve the transcriptome annotation using sequencing data. The current rat annotation is in such an

impoverished status that often the exon sequences registered in the RefSeq annotation do not agree with the genome reference. It is shown that there are over 2500 mRNA species in the dendrites [73], serving as a potent RNA transcript pool poised to maintain philological synaptic functions and respond to stimuli. Although this rich dataset is already a great step forward since the discovery of dendritic local translation [74], detailed functional analysis of the majority of those 2500 mRNAs is hindered by the lack of knowledge of the exact sequences and isoform variants. It has been reported that different isoforms originated from the same gene locus can exert diverse or even counteractive functions [11]. Alternative splicing could lead to inclusion or exclusion of functional domain(s) of a transcript and therefore alter its functions [75]; alternative use of poly-adenylation sites could lead to inclusion or exclusion or regulatory elements in the 3' UTR and thus alter the sub-cellular localization, translational efficiency and turnover efficiency of the transcript [76].

Since PacBio technology allows capturing RNA transcripts in their full-length, assembly of transcriptome is no longer needed (see discussion of Chapter 2). However, there are two challenges that come with the PacBio sequencing technology: low throughput and high error rate. Therefore we developed a customized analysis pipeline that reports high quality transcripts in full- or near full-length (Figure 3.6). First, we addressed the low throughput limitation by optimization of transcript representation in the sequencing library. By efficient normalization of cDNA library and separate processing of cDNA with different size ranges, we are able to detect as many different transcripts as possible at a given sequencing budget. Second, to reduce the sequence errors, we resort to the Illumina sequencing featured with high throughput and high accuracy. Using the Illumina reads derived from the same RNA samples, we can correct most of the random errors introduced in the PacBio sequencing. Together, we developed a computational pipeline named iPEC (Illumina to PacBio Error Correction), the workflow of which is shown in Figure 3.6.

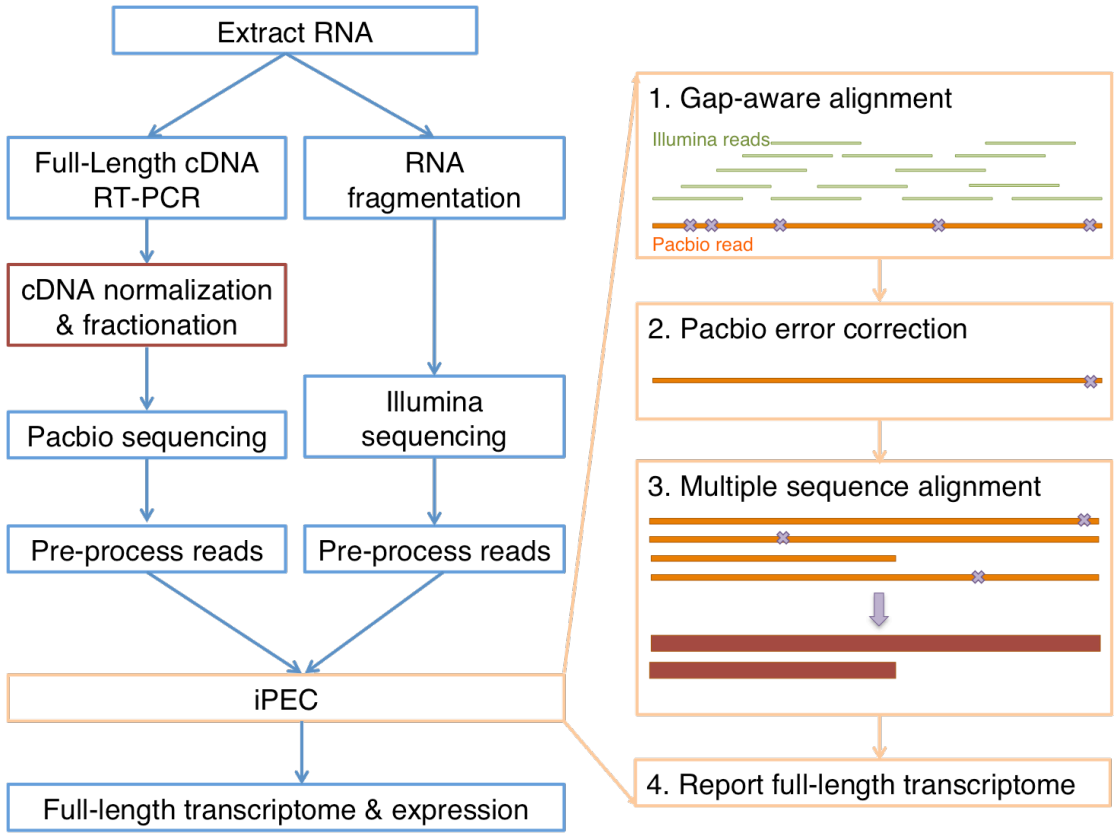

Figure 3. 6 Workflow of hybrid sequencing and iPEC

3.4.2 Sequencing results

Using PacBio technology, we sequenced full-length cDNA libraries of four different size ranges obtained from the hippocampus CA1 region of rat brains. We obtained in total 4 million long reads from 95 sequencing runs (Table 3.2). The sequencing length of cDNA libraries match with the intended size ranges (Figure 3.7). The distribution of length of polymerase extension is similar for all sequencing libraries (Figure 3.7), indicating that the sequencing procedure is not biased to libraries at any particular length.

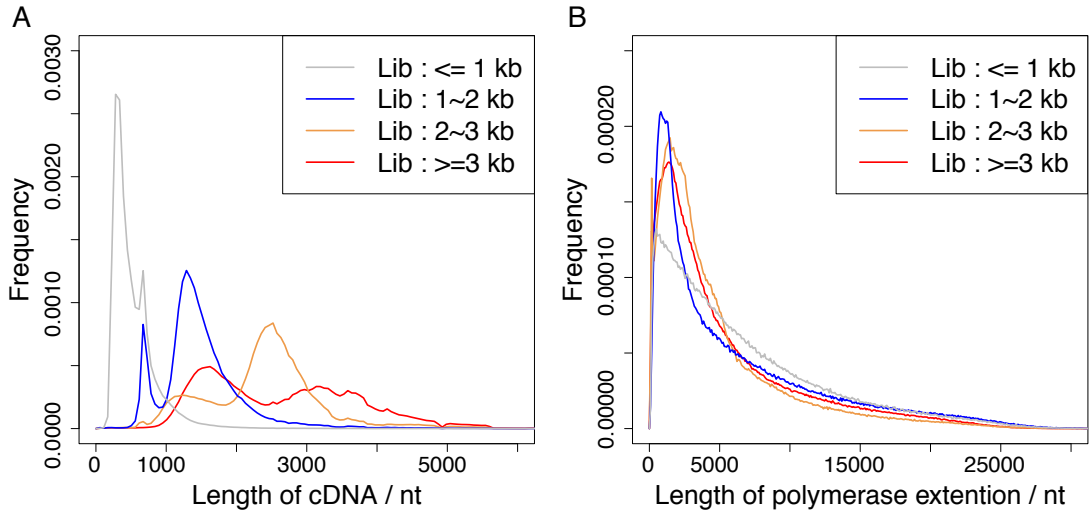

Figure 3. 7 Summary of PacBio sequencing length. (A) The distribution of the size of cDNA of four libraries, with the major peak centered in the range corresponding to the selection size. The left-shoulder on the profiles of the longer libraries, especially for the longest library (red curve), can be explained by the leftover of the shorter cDNA resulted from an imperfect size selection. (B) The robustness and the capacity of PacBio technology, with over half of the polymerase extension beyond 5 kb.

After aligning the PacBio reads to rat RefSeq gene annotation, 71% (13,090 out of 18,436) of the RefSeq genes could be detected, which is similar to most current RNA-Seq studies [5]. Based on the independent Illumina sequencing results, the abundance of those gene loci spans seven orders of magnitudes (Figure 3.8). In contrast, their representation in PacBio sequencing results spans only four magnitudes (Figure 3.8). The three-fold reduction of global dynamic range demonstrated that our approach of cDNA normalization and size fractionation could efficiently address the low-throughput issue.

Libraries	<= 1kb	1~2 kb	2~3 kb	>= 3 kb	Total
No. of sequencing run	23	21	20	31	95
No. of reads	568,636	777,440	1,034,557	1,694,533	4,075,166
No. of Refseq genes (partial)	9,961	10,923	11,453	11,706	13,090
No. of Refseq genes (3' complete)	8,684	9,616	10,450	10,876	12,298
No. of Refseq genes (FL)	1,091	3,712	5,558	6,253	7,283

Table 3. 2 PacBio sequencing statistics of rat hippocampus

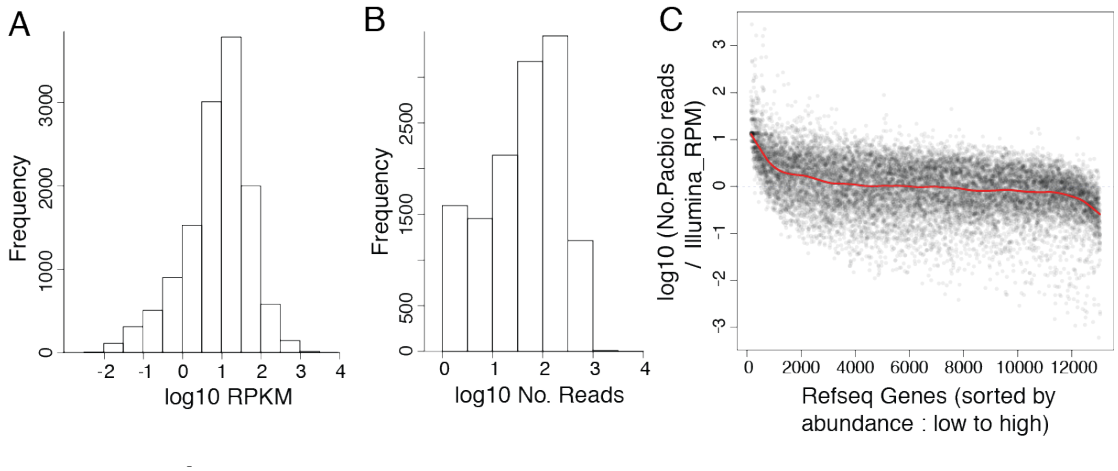

Figure 3. 8 Normalization of PacBio library. (A) The histogram of the dynamic range measured by RNA-Seq. (B) The histogram of the dynamic range of the normalized cDNA library measured by PacBio. (C) The normalization effect for each gene (grey dot), with the average effect shown in red (LOWESS curve)

Then we went on to estimate the extent to which the Refseq genes can be recovered in terms of 5' and 3' completeness. From the 13,090 Refseq gene loci that were at least partially covered by the PacBio reads, 10,876 genes could be recovered when requiring their 3' termini are covered. This number appeared to be counter-intuitively low since the cDNA was generated from the 3' ends (poly(A) tail) of the transcripts and therefore should, in theory, cover all the 3' ends. However, the incomplete generation of cDNA and/or the sequencing limit could lead to sequencing reads corresponding to partial transcripts, especially for long transcripts. Moreover, this could also be explained by the impoverished RefSeq annotation in rat. RefSeq tends to annotate the most abundant or longest isoform of a gene locus, neglecting isoforms with shortened 3' UTR that would otherwise appear to be 3' incomplete.

More stringently, 6,253 genes could be recovered in their full length, from 5' end to 3' end. Following the same argument as above, this number is likely also underestimated. Two factors could influence the transcript recovery. The first one is the length of the transcripts, since longer transcripts are obviously more difficult to be converted to full-length cDNA and subsequently be

sequenced in full-length. Refseq transcripts longer than 5 kb are rarely represented as “full-length” in our PacBio sequencing result, whilst the majority of the transcripts that is shorter than 1kb could be identified end-to-end (Figure 3.9 A). The second factor affecting the transcript recovery is the abundance of transcripts. After considering all transcripts that are shorter than 5 kb in length, it appears that it is less likely for lowly abundant transcripts to be sequenced in an end-to-end manner compared to the highly abundant ones (Figure 3.9 B). For the most highly expressed genes, more than 95% of them were at least partially covered, 90% were sequenced with complete 3’ ends and 65% were sequenced in full-length.

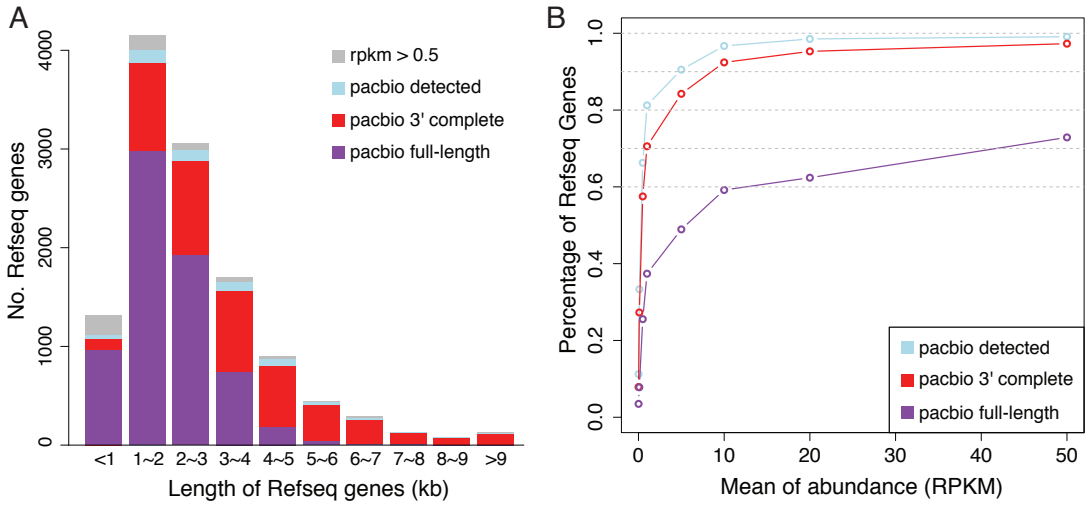

Figure 3. 9 PacBio sequencing covers most of RefSeq genes. (A) The effect of transcript length on gene recovery. Refseq genes are grouped by length (on X-axis), and the height of each colored bar (drawn from 0 on Y-axis) denotes the number of RefSeq transcripts identified according to different criteria, as shown in the inset. (B) The effect of transcript abundance on gene recovery. RefSeq genes are grouped by abundance (RPKM, on X-axis), and it shows the percentage of the genes (on Y-axis) that can be identified according to different criteria, as shown in the inset.

3.4.3 Error correction removed 95% of the sequencing errors

After applying IPEC to our PacBio reads, the error rate is drastically reduced from 11.5% down to 0.25% (Figure 3.10 A). Moreover, there are over 10 thousand additional reads that could be aligned to the rat genome references only after error correction. The major issue affecting the iPEC performance is

the coverage of the corresponding transcripts in the Illumina sequencing dataset. This is evident when we compare the sequence accuracy for transcripts of different abundance before and after error correction. The accuracy of the raw PacBio reads is quite equal (around 88%) for transcripts with different abundance, however, the accuracy of the iPEC corrected reads for highly abundant transcripts increased significantly more than that for the lowly abundant ones (Figure 3.10 B). For rare transcripts (RPKM < 1, corresponding to roughly one molecule per cell), iPEC cannot improve the sequence accuracy due to the lack of high-quality Illumina reads.

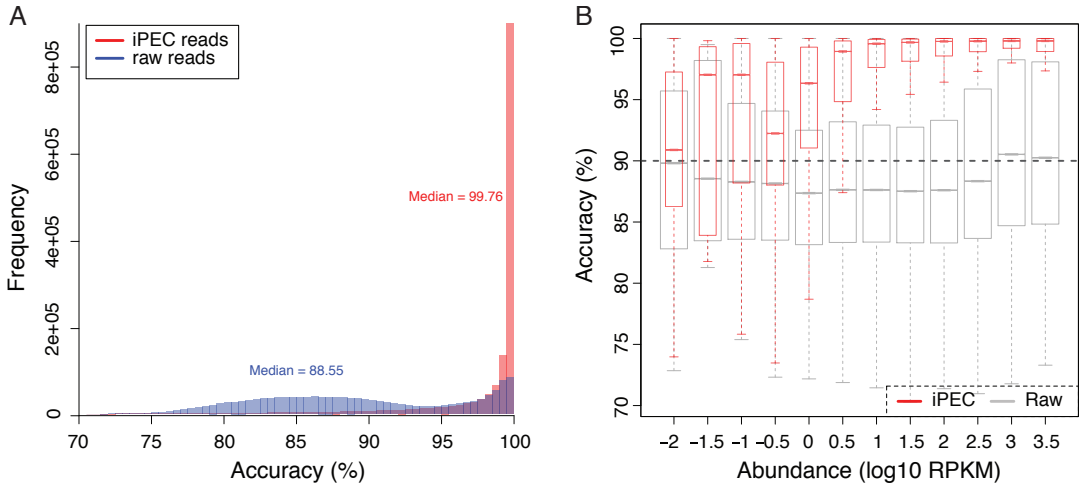

Figure 3. 10 Performance of iPEC. (A) The histogram of sequence accuracy that can be aligned to genome before (blue) and after (red) error correction. (B) For transcripts belonging to different expression bin (X-axis), the accuracy (Y-axis) is shown for raw reads (grey) and corrected reads (red). Whiskers show extreme data points no more than 1.5 times the interquartile range.

3.4.4 Transcriptome landscape of rat CA1 hippocampus

After evaluating the performance of our pipeline on transcript recovery and error correction, we sought to characterize the landscape of transcript isoforms in the CA1 region of rat hippocampus. To be sure to report full-length transcripts, we used only full-pass reads which are characterized by having Clontech SMART PCR primer sequences at both ends and containing a stretch of poly(A) or poly(T) on one end according to PacBio sequencing strategy. There are about 1.1 million full-pass reads combining all SMRT

libraries, and as expected, the percentage of the full-pass reads dropped with increasing size range (Table 3.3). More than 60% of the sequences could be recognized as full-length transcripts (as full-pass) from cDNA libraries shorter than 1 kb, the percentage dropped below 15% for reads from cDNA libraries longer than 3 kb. The low percentage of full-pass read is the bottom-neck of our non-assembly transcriptome profiling and should be alleviated with the improvement of both cDNA generation protocol and sequencing technology. As depicted in Figure 3.11, it is unlike to identify transcripts longer than 5kb solely on the current dataset and without probability based end extension.

Libraries	<= 1kb	1~2 kb	2~3 kb	>= 3 kb	Total
No. of sequencing run	23	21	20	31	95
No. of reads	568636	777440	1034557	1694533	4075166
No. of full-pass reads	358016	230111	279336	296698	1164161
Percentage of full-pass reads (%)	62.96	29.60	27.00	17.51	28.57

Table 3. 3 Summary of PacBio full-pass reads

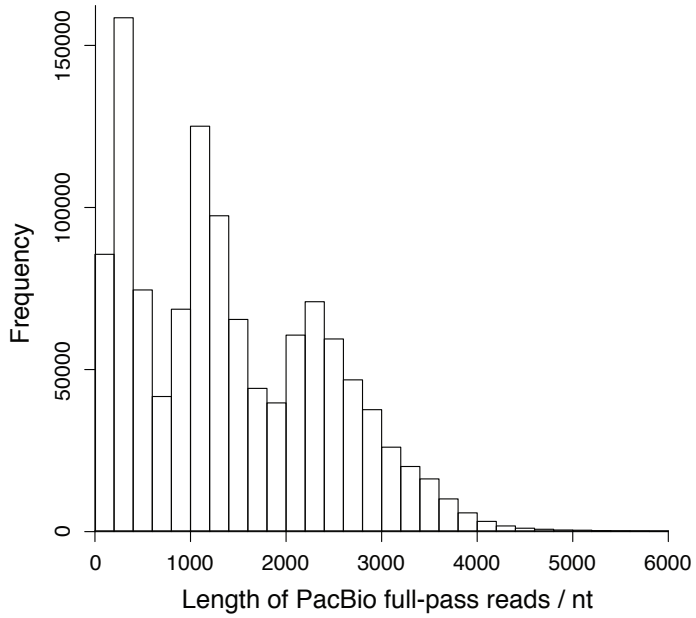

Figure 3. 11 Length distribution of PacBio full-pass reads. Few full-pass reads are of length longer than 5000 nt. The three peaks at around 300, 1000 and 2000 nt are likely artifacts resulting from the size selection step.

Since several full-pass reads could potentially correspond to the same RNA

transcript, sequence similarity can be used to deduce a representative set of references via transcript clustering. Transcript clustering, in brief, selects a minimal set of the distinct sequences that could represent all full-pass reads. The longer sequence is selected from the two if the shorter one is a substring of the longer one, and the only difference between them is at the 5' end, considering the potential falloff of reverse transcriptase towards 5' end of transcripts. After transcript clustering, we obtained 74 thousand transcripts derived from 12 thousand gene loci. The full-pass transcript set is of length distribution similar to that of the annotated gene in mouse, although transcripts longer than 5000 nt were under-represented (Figure 3.12).

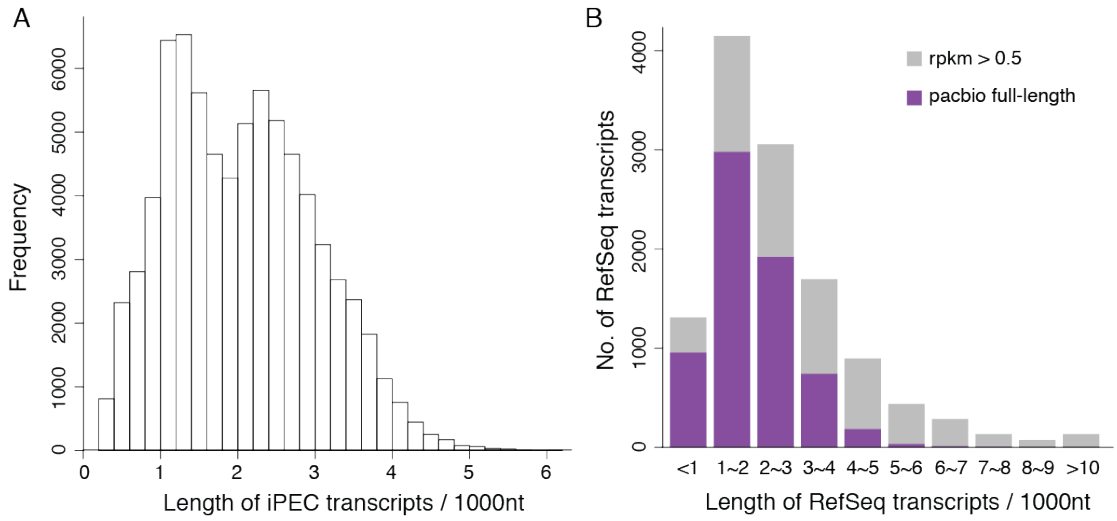

Figure 3. 12 Length distribution of iPEC transcripts. (A) The histogram of the length distribution of iPEC transcripts. (B) The histogram of the length distribution of expressed RefSeq transcripts (grey) and those with supporting full-length iPEC transcripts (purple). For RefSeq transcripts longer than 5k nt, few full-length iPEC transcripts were identified.

Out of 74,011 iPEC transcripts, only 10,774 were annotated "as is" in the RefSeq database and the rest are different transcript isoforms of the known gene loci or novel gene loci (Figure.3.13 A). The lack of transcript isoform diversity is not restricted to RefSeq annotation in rat. Most of the rat gene loci are annotated in both RefSeq and Ensembl with only one transcript isoform, whilst the majority of iPEC transcripts and mouse Ensembl gene loci have

multiple isoforms (Figure 3.13 B). The iPEC transcripts greatly enrich the complexity of existing gene annotation in rat.

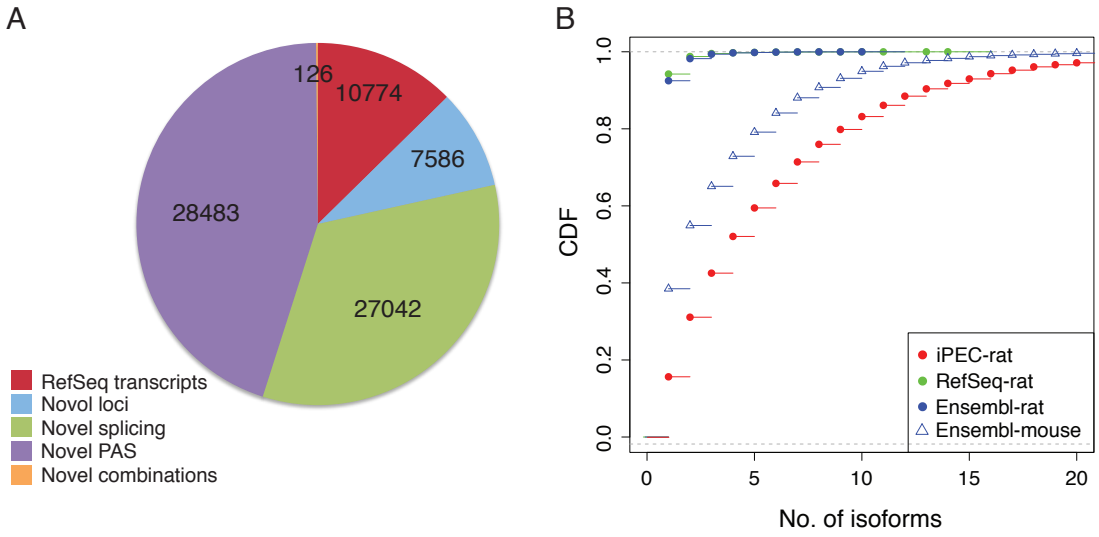

Figure 3. 13 Annotation of iPEC transcripts. (A) The annotation pie chart of iPEC transcripts. The proportion of iPEC transcripts identical to RefSeq transcripts is shown in red, transcript from novel loci in blue, transcript with novel splicing events in green, transcripts with novel alternative poly-adenylation sites (PAS) in purple, and transcript with combinations of splicing events different from RefSeq transcripts in orange. (B) Cumulative distribution of the number of isoforms per gene locus is shown for iPEC transcript (red dots), RefSeq annotation in rat (green dots), Ensembl annotation in rat (blue dots) and Ensembl annotation in mouse (blue triangles). The complexity of iPEC transcripts identified here from one type of tissue (brain) is even higher than that of mouse Ensembl annotation that contains several of tissue types and developmental stages.

3.5 Discussion

The past decade witnesses the increasing understanding towards the complex transcriptome landscape, which has revolutionized many views of molecular biology, including the “one gene-one enzyme” hypothesis. Now we know that over 95% of human multiexon genes have multiple transcript isoforms that are of potentially diverse functions [25]. As a proxy of human biology, we would expect no less in transcriptome complexity in mouse [77] or

rat or many other potential organisms. There are many studies focusing on several areas of transcriptome, from alternative splicing (AS) [25], alternative transcription start site (aTSS) [78] and alternative poly-adenylation site (aPAS) [76] to RNA secondary structure [79] and RNA-editing [31]. However, not much is known on the full-length transcriptome profile, where the aforementioned processing events are linked together and with resolution at single molecule level. In fact, many concurrent studies start out with over-simplified situations. For example, in the studies of genome-wide post-transcriptional regulation mediated by miRNAs, the miRNA binding sites analyzed are almost always restricted to the 3' UTR of RefSeq genes [80]. RefSeq gene contains the least number of transcript isoforms comparing to all other gene annotations, especially when considering the diversity of 3' UTR. Therefore studying genome-wide miRNA-mRNA interaction based on a more comprehensive set of full-length would render better results.

Here we demonstrate that PacBio technology can be used to characterize full-length transcripts. The experimentally determined full-length transcripts can not only greatly improve the transcriptome annotation, but also directly shed light on the interactions of the mechanisms of various RNA processing events. As an example, we confirmed the independent splicing choice of the ectodomains of Dscam gene, which underwent extensive debates in the past ten years. More importantly, we developed a computational framework that reports high-quality transcript references of the sample of interest with or without the reference genome. It would be useful to identify molecular markers for clinical diagnosis, especially in the cases of cancer, where the genomic sequences could be very different from the reference. Unbiased identification the global profile of transcriptome landscape, including the aberrant ones [81] will help us to learn more about the underlying mechanisms of various physiological or pathological processes from embryonic development to cancer, which will in turn lead us towards better diagnosis and potential treatment as true precision medicine.

Chapter 4

De novo pre-microRNA identification

4.1 Introduction

MicroRNAs constitute an important class of small non-coding RNAs that regulate gene expression at the post-transcriptional level through sequence-specific base pairing. Most miRNA genes are transcribed by the RNA polymerase II to generate primary miRNA (pri-miRNA) transcripts. Alternatively, pre-miRNAs can be generated from debranched short introns with hairpin-forming potential (mirtron) by the spliceosome complex, or can be derived from other non-coding RNAs such as snoRNAs. After being transported into the cytoplasm by the exportin-5 complex, pre-miRNAs are further processed by an enzyme named Dicer into double stranded mature miRNA duplexes (miRNA-5p:miRNA-3p, or historically miRNA:miRNA*), one strand of which is preferentially incorporated into the RNA-induced silencing complex (RISC) and bind to the target mRNAs. In mammals, at least 1/3 of protein-coding genes are thought to be under miRNA regulation [82].

Sanger sequencing was first used to identify miRNA genes [18]. It does not allow a global profiling of miRNAs since it is both biased (towards the highly abundant ones) and resource prohibitive. With the introduction of RNA-Seq technology, the detection sensitivity has been dramatically improved. Using sequencing data, currently a total of 28,645 miRNAs from 223 organisms have been registered and annotated in a database named miRBase [83].

In contrast to the analysis of mature miRNAs, attempts to systematically characterize pre-miRNAs are limited, despite more functional and regulatory information can be learned from the latter. The precise sequences of most, if not all, pre-miRNA sequences are not determined by experiments. Instead, they are often inferred from the sequences of the corresponding mature

miRNA pair, therefore ambiguity could arise if one of the pair was not identified. To date, the expression patterns of known pre-miRNAs are analyzed by using northern blot, *in situ* hybridization and qPCR. Again due to the laborious procedure, such experiments are seldom done at the global level.

In order to gain a deeper understanding of mammalian miRNAs, we sequenced in parallel miRNAs and pre-miRNAs derived from ten different tissues of adult mice. We developed a computational pipeline, miRGrep (miRNA Genome Reference free Prediction, available at https://github.com/arthuryxt/miRGrep), to search for genuine miRNA genes solely based on sequencing datasets, without using genomic sequences. Using miRGrep, 239 known mouse pre-miRNAs could be recovered and 41 novel ones were predicted with high confidence. Similar to the well-studied miRNAs, the mature miRNAs derived from most of these novel loci showed reduced abundance following Dicer knockdown. Moreover, Argonaute2 immunoprecipitation (Ago2 IP) experiment confirmed that novel miRNAs could bind to Ago2/RISC complex, strongly indicating their functional roles as miRNAs. Evaluation on another dataset obtained from *C. elegans* demonstrated that miRGrep could be widely used for miRNA discovery in metazoans, especially in the absence of a reference genome. Moreover, we observed several new aspects of processing and modification of mouse miRNAs, including Ago2 cleaved pre-miRNAs, new editing events and exclusively 5' tailed mirtrons. These new insights are not only valuable to a better understanding of miRNA biology but also might serve as diagnostic biomarker of various diseases.

4.2 Methods

4.2.1 Small RNA sequencing protocols

Small RNA sequencing libraries (10-40nt fraction and 50-100nt fraction) were prepared using Illumina small RNA library preparation kits. Note, for 50-100nt

fraction RNAs, the denaturation temperature is elevated to 98°C in order to disrupt the hairpin structure of the pre-miRNAs. Small RNA (10-40nt fraction) libraries were sequenced for 36 cycles using Illumina GAIIx. Ago2 IP RNA library was sequenced for 50 cycles using Illumina HiSeq2000. Normalized small RNA (50-100nt) libraries were sequenced for 100 cycles using Illumina HiSeq2000.

4.2.2 Normalization of sequencing library

The small RNA (50-100nt fraction) sequencing library was normalized by using Duplex specific Nuclease (DSN, Evrogen) per manufacturer's instructions (as in Chapter 2).

4.2.3 Small RNA sequencing reads mapping

First, 3' adapter sequences were removed from the sequencing reads using a custom Perl script. The reads of length between 17 and 30 nt from small RNA 10-40 nt fraction were retained. The reads from mouse and *C. elegans* samples were mapped to genome reference sequences (UCSC genome browser mm9 and ce6) and known pre-miRNA sequences deposited in miRBase (mouse and *C. elegans*, version 16.0) (http://www.mirbase.org/) [83] without allowing any mismatch using soap1 and soap.short [84], respectively. To be considered as a known miRNA, the 5' and 3' ends of a sequencing read should be within 1nt and 3nt from the 5' and 3' ends of the miRNA annotated in miRBase, respectively. For the small RNA 50-100nt fraction, the sequencing reads of length between 40 and 94 nt were aligned to genome reference sequences (UCSC genome browser mm9 and ce6) allowing 2 mismatches using soap2 [85]. To determine the mouse reads derived from full-length pre-miRNAs, we mapped the first 40nt to the mouse pre-miRNA sequences deposited in miRBase allowing 2 mismatches using soap2 and then further extended the alignment to the 3' end. The 5' and 3' ends of mouse pre-miRNAs in miRBase were manually annotated based on the secondary structure if one of the miRNA pair is not identified. Reads regarded

as full-length pre-miRNAs should satisfy the following criteria:

1) The 5' and 3' end of the alignment were within 2 and 5 nt from 5' and 3' end of the pre-miRNA, respectively; considering alternative processing events on pre-miRNAs.

2) No more than five mismatches were found in the alignment, considering higher error rates towards the end of long sequencing reads.

4.2.4 Identification of Ago2-cleaved pre-miRNAs

After aligning the sequencing reads of length between 40 and 94 nt on known mouse pre-miRNA sequences as above, we applied the following filters to extract the reads derived from potential ac-pre-miRNAs according to [86]:

1) When aligning a read to pre-miRNA annotations, if one end of the read matches to one end of pre-miRNA, the other end of the read should be 9-12 nt shorter than the annotation;

2) The truncated part pair extensively to the other arm (at least 8 nt);

3) No bulge is allowed within 4 nt from the potential cleavage site;

4) Ac-pre-miRNA candidates should be supported by at least two reads.

4.2.5 Identification of miRNA editing events

To examine the mouse miRNA editing events, we mapped the non-genome-mapping reads from mouse small RNA (10-40 nt) libraries to mouse reference miRNA sequences, allowing one mismatch. The uniquely mapped reads with one mismatch at least 1nt away from the 3' or 5' end of known miRNAs were retained. For each of the mismatches identified in these reads, we calculated the fraction of certain mismatch at one position as the number of reads bearing that mismatch divided by the number of all reads containing mismatches at the same position. We obtained a set of highly confident A-I (conversion from Adenosine to Inosine) editing sites by searching for A-G changes that could pass the following filters (modified from [31]):

1) The fraction was higher than 90%;

2) The change was found in at least 10 reads;

3) The same change was found in at least one pre-miRNA read, and the

sequencing quality score of that base was higher than 30;

4) The same change was not annotated as a SNP in dbSNP (build 128).

Editing frequency was calculated as the number of reads containing the edited A-G change divided by the total number of reads mapped to the same miRNA.

4.2.6 De novo prediction of pre-miRNAs

In order to predict miRNAs based on the sequencing reads obtained from the two small RNA fractions corresponding to potential miRNAs and pre-miRNAs, we mapped the sequencing reads of length between 17 and 30 nt on the sequencing reads of length between 40 and 94 nt using soap.short without allowing any mismatch. We selected the 40-94nt (long) reads as potential pre-miRNAs on which the mapping pattern of 17-30nt (short) reads was compatible with Dicer processing in the following four steps.

Step-1. Define read clusters.

On one long read, a cluster of mapped short reads is defined as a complete set of overlapping short reads while the maximal distance between the start position of any two reads within the cluster does not exceed 14nt, a length learned from known miRNAs. If the long and short reads were originated from genuine precursor and mature miRNAs, the short reads should form at most three clusters at the 5' end, 3' end and the middle of the long read, corresponding to the miRNA/miRNA* and the loop, respectively. Also, it is reasonable to expect that the 5' and 3' end clusters contain more short reads than the middle cluster, since the middle cluster correspond to the loop region of the hairpin structure and is fast degraded. Furthermore, given the length distribution of canonical mature miRNAs, the majority of short reads from 5' and 3' end clusters should be of length between 17 and 25nt, according to well-accepted size range of mature miRNAs. Therefore, based on these rules, long reads are discarded if they exhibit any of the following patterns:

1) Number of clusters exceeded 3;

2) The minimal distance between any two reads in different clusters is shorter than 5 nt;

3) The number of reads in the middle cluster exceeded that in the 5' end and 3' end cluster;

4) Less than 66% of distinct/non-redundant short reads or less than 90% of all short reads from 5' and 3' end cluster were of length between 17 and 25 nt.

Step-2. Require Dicer compatibility.

After filtering out the obvious non-Dicer compatible reads, we further selected the potential pre-miRNA reads. For each remaining long read, we first identified the most abundant distinct/non-redundant short reads from the 5' and 3' end clusters. The long reads were retained only if the most abundant short reads start or end less than 5nt away from the 5' or 3' end of the long read respectively. We then counted the number of short reads that start at most 1nt away from the 5' end of the most abundant reads in the 5' and 3' end clusters. The term "Sharpness" denoted the percentage of these reads out of all short reads mapped on the same long read. Because most short reads that mapped to a genuine pre-miRNA should origin from the miRNA duplex, we selected long reads with a sharpness value above the threshold of 0.75 (corresponding to 95% of known miRNAs). The selected reads were then clustered if:

1) The most abundant distinct/non-redundant short reads mapped on their 5' and 3' clusters were identical;

2) Their length differed less than 5 nt in length;

3) Their sequence similarities were above 90%. One representative read with the highest abundance from each cluster was selected.

Step-3. Predict 2nd structure

We predicted the secondary structures of the selected long reads using RNAfold (parameters: -p –d 2 –noLP) [87] and randfold (parameter: -d 199) [88], respectively. Only the long reads which could fold into unbifurcated hairpin structures were retained.

Step-4. Estimate 2nd structure stability

The remaining long reads satisfying the following criteria were selected as potential pre-miRNA candidates. The values of the threshold correspond to 95% of known miRNAs. The rest were used as "background" in the probabilistic scoring of potential pre-miRNA candidates.

1) The randfold p-value was smaller than 0.2;

2) More than 60% of the nucleotides in the "mature" part (the most abundant distinct/non-redundant short reads from 5' or 3' end clusters) were base paired.

4.2.7 Probabilistic scoring of pre-miRNAs

We scored the potential pre-miRNA candidates using a Naïve Bayesian classifier with six features:

f_1: Minimal folding free energy calculated by RNAfold divided by the sequence length;

f_2: Randfold p-value

f_3: Number of unpaired nucleotides at 5' end

f_4: Length of 3' overhang (number of unpaired nucleotides at 3' end minus that at 5' end)

f_5: Average length of the most abundant distinct/non-redundant short reads from the 5' and 3' end cluster that corresponded to potential miRNA/miRNA*

f_6: Length of candidate pre-miRNA

The "positive training dataset" was pre-miRNAs from miRBase. We calculated the probability of a given potential pre-miRNA candidate to be a genuine pre-miRNA using the following formula:

$$\Pr(pre \mid data) = \frac{P(data \mid pre) * P(pre)}{P(data \mid pre) * P(pre) + P(data / non) * P(non)}$$

where

$$P(data \mid pre) = \prod_{i=1}^{6} P(f_i \mid pre)$$

and

$$P(data \mid non) = \prod_{i=1}^{6} P(f_i \mid non)$$

$P(pre)$ is the prior probability that a long read was a genuine miRNA precursor. $P(non)$ is the prior probability that a long read was non-miRNA background stem-loop and was equal to $1 - P(pre)$. Both $P(pre)$ and $P(non)$ are set to 0.5 by default, but could be changed based on the expected pre-miRNA sequences in the deep sequencing samples.

$P(f_i \mid pre)$ denotes the probability that a miRBase pre-miRNA has the value of f_i, where $i \in \{1...6\}$

$P(f_i \mid non)$ denotes the probability that a non-miRNA background stem-loop-like sequence has the value of f_i, where $i \in \{1...6\}$

4.3 Results

4.3.1 miRNA and pre-miRNA sequencing

We sequenced small RNA (10-40nt) libraries from 10 different mouse tissues and obtained 167 million reads between 17 and 30 nt in length (hereafter referred to as “short reads”). Of these, 75.2% aligned to the mouse genome without mismatch and 52.8% derived from known mouse miRNA loci (Figure 4.1). In parallel, we sequenced pre-miRNAs. To characterize as many pre-miRNAs as possible, we pooled the total RNA from the 10 mouse tissues equally and extracted small RNA between 50 and 100 nt in length. After library normalization, we obtained over 57 million reads between 40 and 94 nt in length (hereafter referred to as “long reads”), of which 86.7% could be mapped to the mouse genome. In stark contrast to that of short reads, only 0.80% of the long reads were originated from known pre-miRNA loci (Figure 4.1).

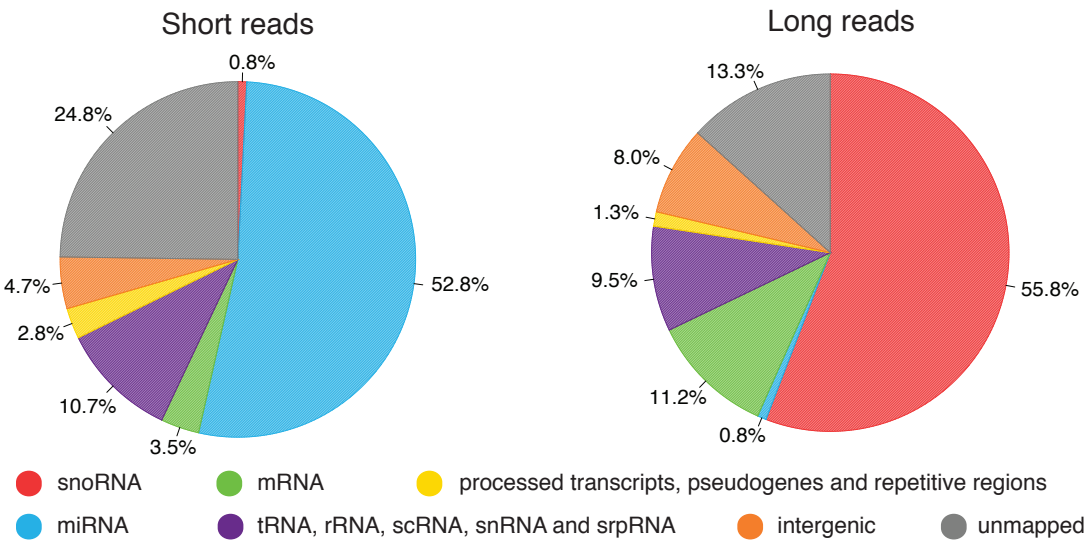

Figure 4. 1 Parallel sequencing of miRNAs and pre-miRNAs. The pie charts represent the genomic origins of the short reads and the long reads, respectively.

We identified 887 known mouse miRNAs originated from 568 pre-miRNAs from the short reads. 687 miRNA from 481 pre-miRNAs were expressed (with RPM more than 1; RPM: Reads Per Million total miRNA reads) in at least one tissue (Figure 4.2). In comparison, only 281 known pre-miRNAs are identified using the long reads. The distribution of length and RNA secondary structure of these 281 pre-miRNAs are similar to that of all mouse pre-miRNAs deposited in miRBase, indicating that our detection of pre-miRNAs was not biased towards any particular subset of pre-miRNAs. 278 out of 281 detected pre-miRNAs had the corresponding miRNA present in at least one tissue. As shown in Figure 4.3, miRNAs with their precursors detected were expressed at a significant higher level than those without (two-sided Wilcox rank-sum test, $P < 2.2e\text{-}16$), whereas the correlation between the abundance of a certain miRNA and that of its precursor was low ($R^2=0.1501$). Such low correlation might be explained by the fact that pre-miRNAs are RNA intermediates that undergo fast processing.

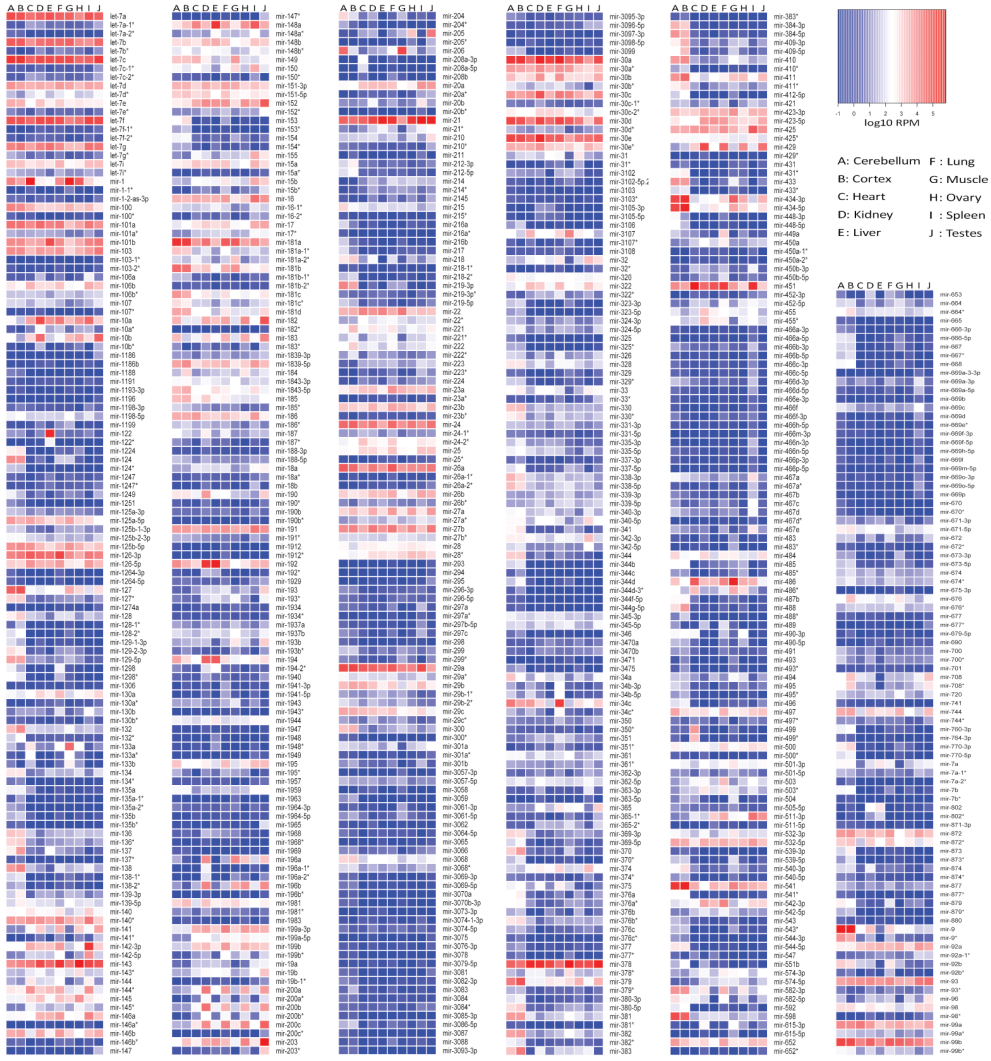

Figure 4. 2 Relative expression of 687 expressed miRNAs in ten mouse tissues

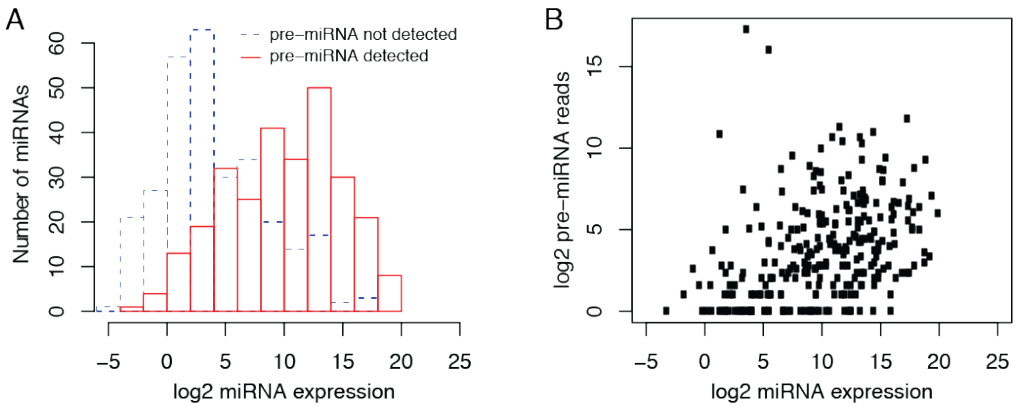

Figure 4. 3 Abundance of miRNAs and pre-miRNAs. (A) The histogram depicts the abundance of miRNAs whose pre-miRNAs were detected (red) are in general much higher than that of those whose pre-miRNAs were not detected (blue). (B) The

scatter plot shows the poor correlation between the abundance of miRNAs and the corresponding pre-miRNAs.

4.3.2 De novo prediction of mouse pre-miRNAs

Since the long reads datasets (corresponding to pre-miRNAs) greatly reduce the search space for miRNAs, and obtaining them is much more cost-efficient than establishing high quality genome reference sequences for most organisms, we developed a computational pipeline for miRNA Genome reference free prediction (miRGrep) solely relying on short and long reads. In brief, potential pre-miRNA sequence candidates are extracted by selecting the long reads that could form stable hairpins and exhibit a mapping pattern of short reads that is compatible with Dicer processing. These pre-miRNA candidates are then scored with a Naive Bayesian classifier and finally a short list of highly confident candidates are reported.

We applied miRGrep to predict mouse miRNAs on our sequencing data set. In total, 155,760,811 short reads were perfectly mapped to 5,789,406 distinct long reads. The vast majority (5,524,656) of the long reads were discarded because the mapping position of short reads did not fit with the model of Dicer processing. The remaining 264,750 long reads were then merged into 131,207 clusters and one representative read from each cluster was selected to predict the secondary structure. Out of these, 1,277 long reads that could form stable hairpin structures were selected for the probabilistic scoring. As a result, 538 long reads were scored higher than 0.95, of which 324 with at least five supporting short reads were retained as pre-miRNA candidates. One example is shown in Figure 4.4.

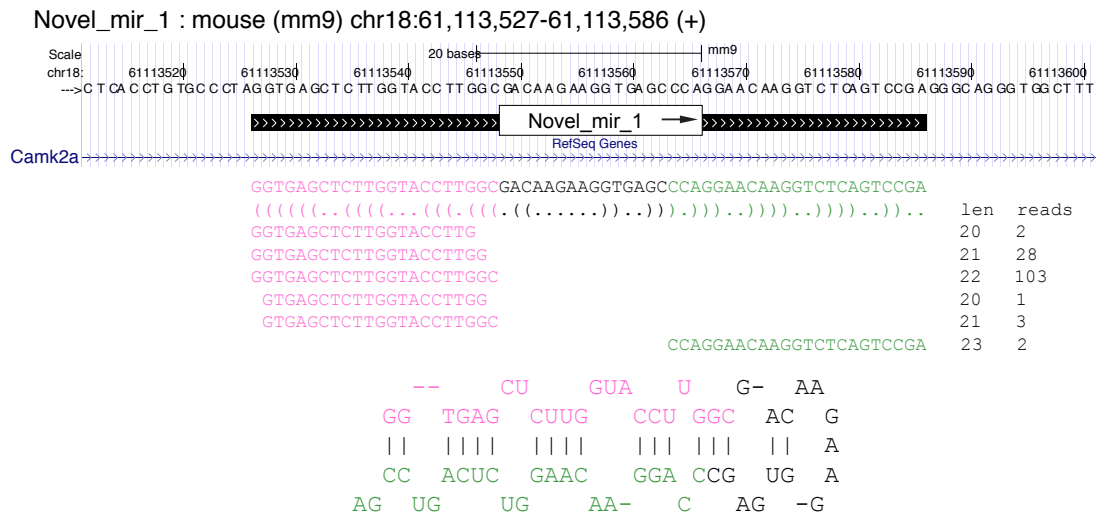

Figure 4. 4 An example of novel pre-miRNAs in mouse. This novel pre-miRNA locates within an intron of the gene Camk2a, the sequences of 5p- and 3p- mature miRNAs are shown in red and green, respectively.

Among these 324 pre-miRNA candidates, 245 corresponded to 239 known mouse pre-miRNAs. Although the nominal false negative rate is nearly 15%, some pre-miRNAs discarded by miRGrep could as well be true negatives owing to false annotation in miRBase. For instance, miR-1944 and miR-715, which were not reported by miRGrep, have been recently removed in miRBase v18.0. Moreover, three out of the 41 novel pre-miRNA candidates have been recently registered in miRBase v18.0. From these novel pre-miRNAs, 66 mature miRNAs were supported by short reads. Compared to known miRNAs, these novel miRNAs were expressed at a much lower level (Figure 4.5), suggesting the importance of deep sequencing on charting a comprehensive miRNA landscape.

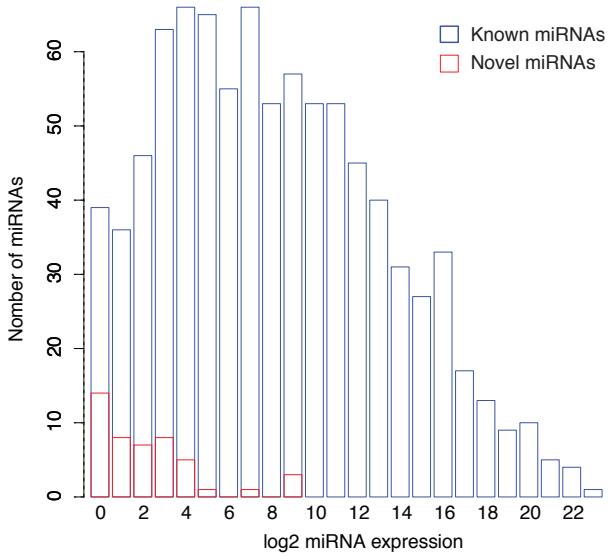

Figure 4. 5 Abundance of novel and known miRNAs. The abundance of novel miRNAs (red) is in general much lower than that of known miRNAs (blue).

Regarding to the genomic locations, 34 novel pre-miRNAs identified in this study located to introns of protein coding genes (Table 4.1). Among them, 13 are distant to splicing sites, i.e. both 5' and 3' ends are at least 10nt away from the ends of the hosting introns. Of the remaining 21 intron-containing pre-miRNAs, whereas four had the 'nearly' exact boundary as the hosting introns and thus resembled canonical mirtrons, 17 had only one end generated by spliceosome while the other end likely matured through Drosha independent trimming [89]. Interestingly, all the pre-miRNAs from the latter category were 5' tailed mirtrons, which shared only their 3' ends with the hosting introns. Indeed, we also found the long reads possibly derived from the intermediate tailing products for several tailed mirtrons. To check whether the tailed mirtrons in mouse were exclusively 5' tailed, we analyzed the boundary of known mouse pre-miRNAs located in introns and found that all 21 known tailed mirtrons are tailed from 5' end. In contrast to our findings in mouse, the tailed mirtrons identified so far in drosophila are all from 3' end [90]. It awaits further investigation whether the inconsistence between the two organisms is due to the difference in underlying processing mechanisms such as more efficient usage of 5'-3' (mouse) versus 3'-5' (fly) exoribonuclease

after splicing.

Genomic location	Number of novel pre-miRNAs
LINE, SINE	2
Intergenic region	3
Exons	2
Introns : mirtron	4
Introns : tailed-mirtron	17
Introns : other	13

Table 4. 1 Genomic annotation of novel miRNAs

4.3.3 Validation of novel mouse miRNAs

We validate the authenticity of the miRGrep novel miRNAs using three independent approaches.

First, to investigate whether the novel miRNAs were indeed dependent on Dicer for expression, we used RNA interference to knockdown Dicer in a mouse N2a cell line. RT-qPCR showed that the level of Dicer mRNA transcripts in cells treated with siRNA was significantly decreased by 85% (Figure 4.6 A). After sequencing the small RNAs from unperturbed and siRNA treated cells, we compared the abundance of different non-coding RNA derived transcripts between the two samples. Comparing to the controls, both rRNAs and tRNAs showed a median increase of 21% and 19% after silencing Dicer, whereas both known and novel miRNAs decreased in abundance with a median reduction of 32% and 55%, respectively (Figure 4.6 B-E). Using TaqMan assay we confirmed that the expression level of one novel miRNA (miR-Novel-2) decreased after Dicer knockdown, similar to that of one known miRNA (Figure 4.6 F). These results demonstrated that the novel miRNAs identified in this study were enriched in Dicer dependent small RNAs.

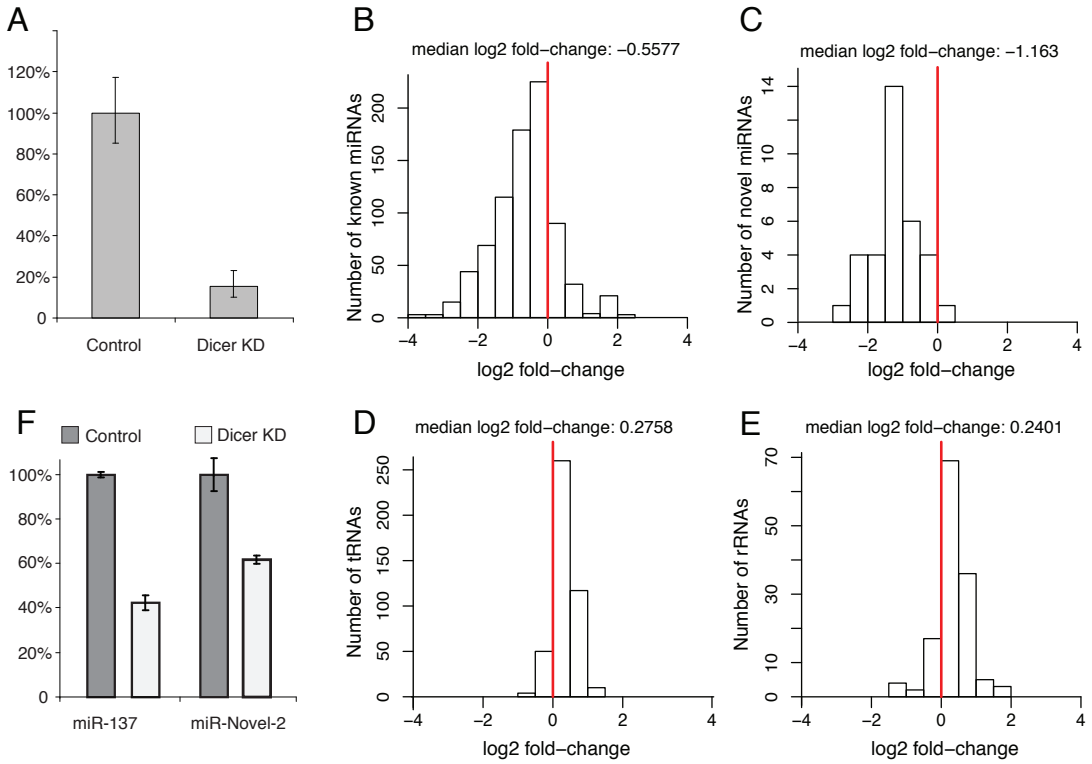

Figure 4. 6 Novel miRNAs depend on Dicer for expression. (A) The abundance of Dicer mRNA decreased to 15% in Dicer knockdown (KD) compared to scramble knockdown (Control). (B-E). The log2 fold-change of small RNAs abundance after Dicer knockdown for known miRNAs (B), novel miRNAs (C), tRNAs (D), and rRNAs (E). (F) The abundance of novel miRNA miR-Novel-2 decreased by 40% after Dicer knockdown, similar to that of miR-137. Error bars represent standard deviation.

Second, miRNAs mediate target mRNA silencing by directly bind to Ago2 proteins. To further confirm the functionality of our miRNA candidates, we isolated and sequenced Ago2 associated RNA in N2a cells by Ago2 IP. A total of 49 novel miRNAs derived from 37 candidate pre-miRNAs could be detected in IP sample. The novel miRNAs showed a similar Ago2 binding profile as that of known miRNAs, whereas both tRNAs and rRNAs showed significant depletion (Figure 4.7 A). The enrichment of miR-137 and one novel miRNA was further validated by TaqMan assay (Figure 4.7 B). These results showed that our novel miRNAs were indeed incorporated into Ago2/RISC complex, indicating their potential functionality.

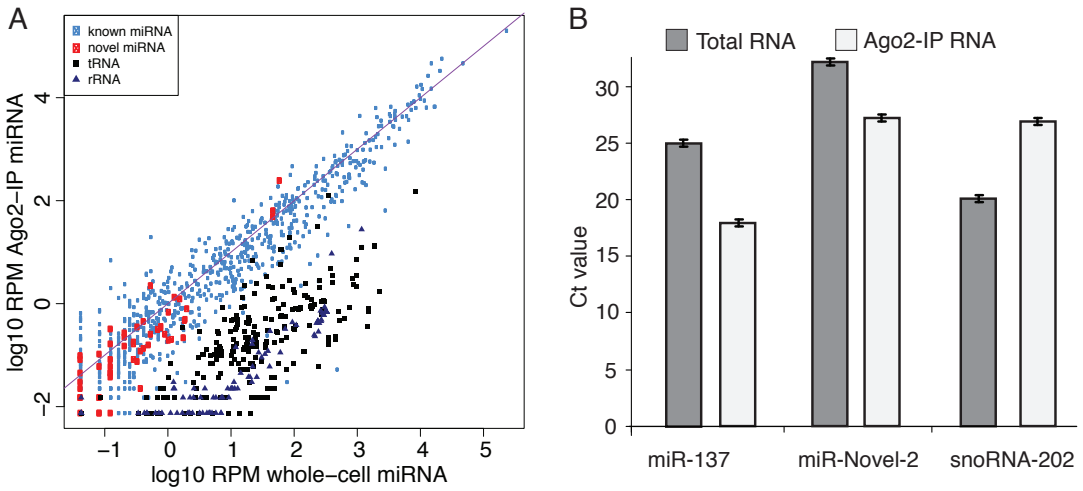

Figure 4. 7 Novel miRNAs bind to Ago2. (A) Each point in the scatter plot represents one gene: known miRNAs (light blue), novel miRNAs (red), tRNAs (black) and rRNAs (dark blue). Value on X- and Y- axis denote the log10 RPM of Ago2-IP RNA and total RNA, respectively. (B) TaqMan assay validation of Ago-2 association, the values on Y-axis mark the cycle value. The novel miRNA miR-Novel-2 are enriched in Ago2-IP RNA, similar to that of miR-137; while as a negative control, snoRNA-202 is depleted in the Ago2-IP RNA.

Third, two most abundant novel pre-miRNAs loci were chosen for further experimental investigation. First, we demonstrated the presence of novel miRNAs using northern blotting (Figure 4.8 A-B). Then we further investigated whether the processing of the two pre-miRNA candidates was dependent on Dicer. When we incubated the in vitro transcribed and 32P-labeled pre-miRNAs with recombinant Dicer, both pre-miRNA transcripts were efficiently processed into mature miRNAs (Figure 4.8 C-D). Most importantly, we showed that both novel miRNAs could significantly repress the target mRNAs as other canonical miRNAs using *in vitro* luciferase assay (Figure 4.8 E).

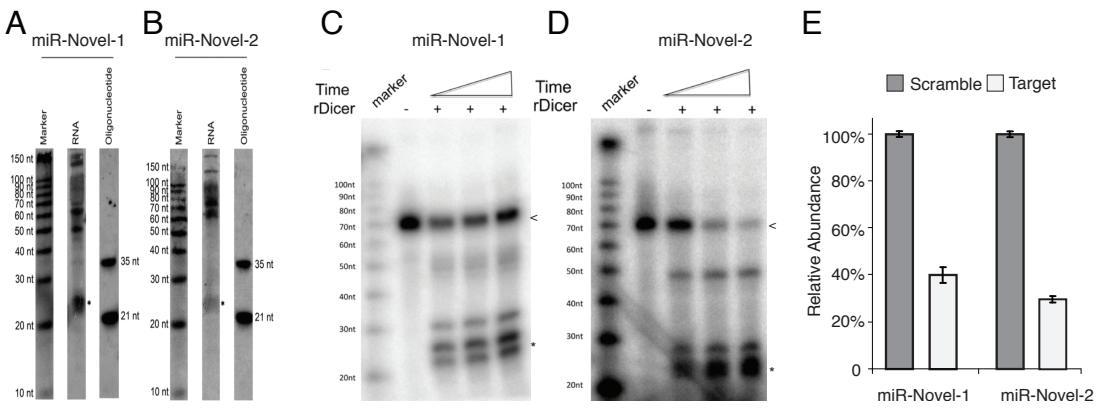

Figure 4. 8 Novel miRNAs have miRNA-like functions. (A-B) Northern blot of two novel miRNAs reveals specific bands (marked by asterisk sign) corresponding to mature miRNAs. (C-D) Efficient *in vitro* rDicer processing of pre-miRNAs (marked by "<" sign) in to mature miRNAs (marked by asterisk sign). Lane marked by "+" or "-" sign denote RNAs are incubated with or without rDicer. (E) The luciferase assay of the two novel miRNAs. Sequences complementary to novel miRNAs (Target) or scrambled sequences (Scramble) were inserted into the 3'UTR of Renilla gene. Comparing with scramble controls, inclusion of Target sequences resulted in substantial decrease of protein production. Experiments were done in triplicates, and error bars represent standard deviation.

Taken together the results of three independent approaches, we successfully demonstrated that the novel miRNAs identified by miRGrep are *bona fide* miRNAs.

4.3.4 Evaluation of miRGrep

Most miRNA discovery tools rely on the alignment of sequencing reads to reference genome sequences [91]–[93]. Obviously, these tools are of limited applicability in the study of organisms whose genome has not been sequenced. Other tools that do not depend on genome sequences take advantage of evolutionary conservation while neglecting lineage-specific ones [94]. In contrast to these tools, miRGrep takes advantage of parallel sequencing of potential mature and precursor miRNAs. Of 438 known mature miRNAs recovered by miRGrep, 11% (48) are not conserved in other species and could not be identified only by homology search. It demonstrates that

miRGrep can discover not only the conserved miRNAs, but also lineage specific ones.

In probabilistic scoring of pre-miRNA candidates, the known mouse miRNAs are used to estimate the model parameters. To investigate the potential bias arguably introduced by over-trained using known mouse miRNAs, we trained our model again using known miRNAs in human, fruit fly and *C. elegans*, respectively. As illustrated in Table 4.2, the predictions based on known miRNAs from different organisms were nearly identical, indicating that the features included in our model represent the miRNA characteristics common to all metazoans.

	Specific to mouse training set	Common	Specific to other training set
Human	8	316	13
C. elegans	21	303	11
Fruit fly	14	310	9

Table 4. 2 miRGrep models conserved features of miRNAs. The values denote the number of miRNAs predicted only using mouse training set (1st col), common to two trains sets (2nd col), only using human or C. elegans or fly training sets (3rd col).

Furthermore, to assess whether miRGrep could be applied to other metazoans, we chose another well-studies organism, *C. elegans*, for validation. We applied our parallel sequencing approach to RNAs extracted from adult *C. elegans*, and miRGrep identified 108 pre-miRNA candidates. Ninety-eight of them corresponded to 88 known *C. elegans* pre-miRNAs, which covered 50% of all miRNAs identified along all developmental stages in the past decade. One of the 10 newly identified pre-miRNAs is shown in Figure 4.9 as an example.

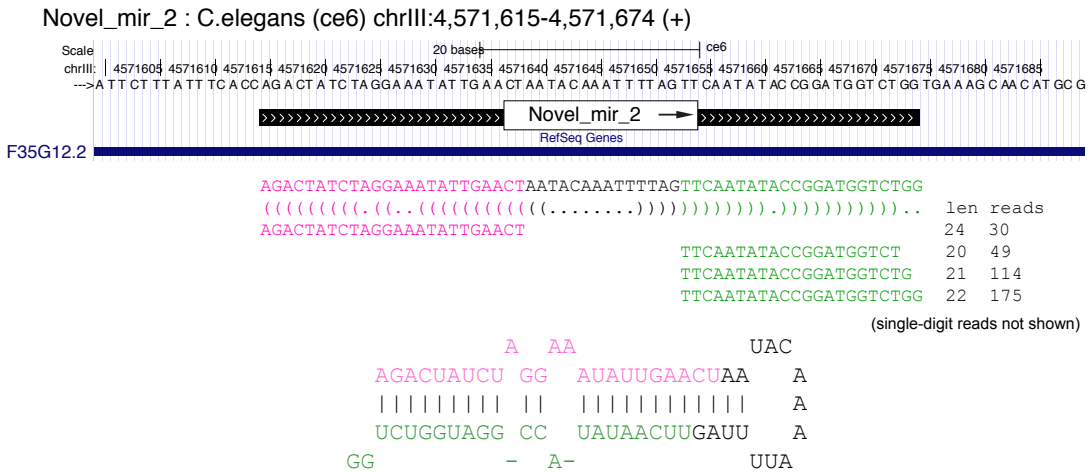

Figure 4. 9 An example of novel pre-miRNAs in *C. elegans*. This novel pre-miRNA locates within the 3'UTR of the protein-coding gene F35G12.2. The 5p- and 3p-miRNAs are colored in red and green, respectively.

4.3.5 Identification of pre-miRNA processing intermediates

Half of the long reads originated from known miRNA loci are not full-length pre-miRNA transcripts, likely representing processing intermediates or degradation products of pre-/pri-miRNAs. For example, long reads with truncation in one arm of the hairpin structure resemble endogenous processing intermediates resulted from Ago2-mediated endonucleolytic cleavage, termed as Ago2-cleaved pre-miRNAs (ac-pre-miRNA) [86]. In our data, we identified eight potential ac-pre-miRNAs in mouse (Table 4.3), out of which seven originated from let-7 family. Importantly, 3' end uridylation events were observed for all the ac-pre-miRNAs, which further indicates that ac-pre-miRNAs are *bona fide* pre-miRNA processing intermediates.

Ac-pre-miRNA	Distance to 3'-end (nt)
mmu-let-7a-1	11
mmu-let-7b	10
mmu-let-7c-2	12
mmu-let-7d	10
mmu-let-7f-1	13
mmu-let-7i	11
mmu-mir-98	11
mmu-mir-30b	9

Table 4. 3 List of pre-miRNA processing intermediates

4.3.6 Identification of miRNA editing events

RNA editing is a process that alters nucleotide(s) of RNA molecule post-transcriptionally, most commonly observed as ‘A-to-I’ or ‘C-to-U’ base substitutions. The ‘A-to-I’ editing events, catalyzed by double-strand RNA binding enzymes named ADARs, have been reported in mammalian pre-miRNAs, and such events can affect miRNA biogenesis as well as targeting [95]. Since ADAR enzymes are enriched in brain, we went on to identify A-to-I editing events on miRNAs using our brain data of both mature and precursor miRNAs. The rationale is that since ADARs bind to double-strand RNA, true miRNAs editing events should also be observed on precursor miRNAs. As shown in Figure 4.10, nine editing sites were found both in cortex and cerebellum, but their editing frequencies were different in the two different regions, suggesting potential differential regulation and functions. 11 of 15 sites located in the seed regions of miRNAs, which would affect selection of mRNA targets, as previously described for the miR-376 cluster [95]. For the remaining 4 sites outside of seed regions, the A-to-I editing at pre-mir-497 could affect its secondary structure by forming 'I-C' wobble base pair, thereby impacting the processing by Dicer. Depending on the miRNA expression level and their sub-cellular localization pattern, even of low editing frequency, the copy number of edited miRNAs can still be high enough to be biological relevant with altered functions. Therefore the new editing events identified here contribute to a better understanding of miRNA regulation and function.

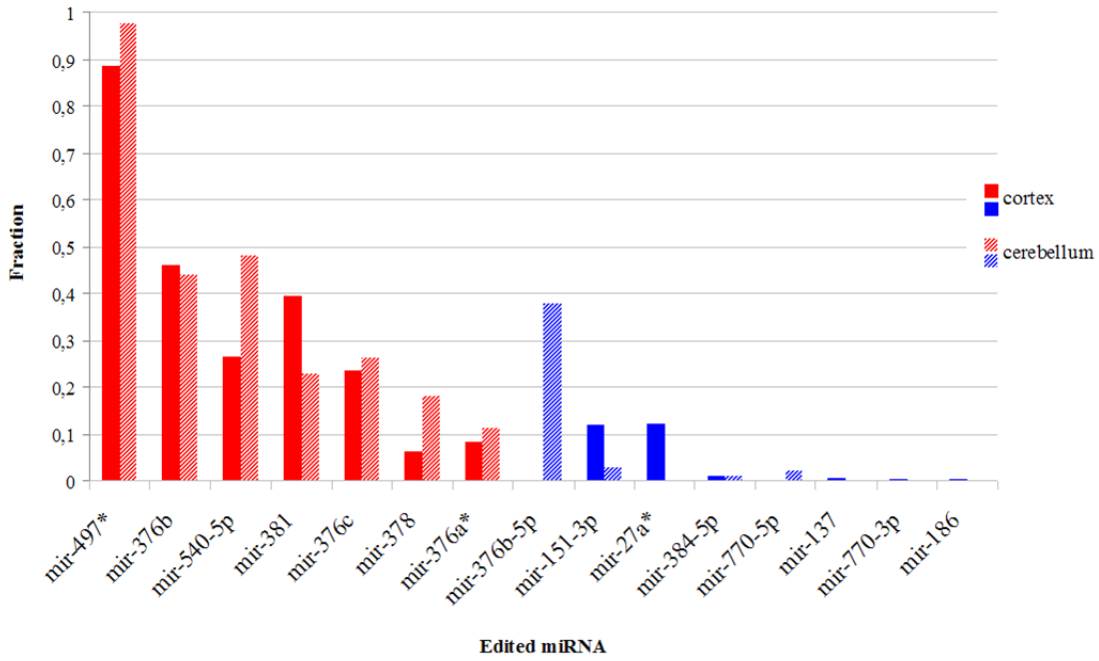

Figure 4. 10 Relative frequency of miRNA editing events. Editing frequency in mouse cortex and cerebellum (in solid color and shaded color, respectively) of previously identified sites (in red) are in general higher that that of the newly identified ones (in blue).

4.4 Discussion

Comparing to mature miRNA sequencing, our pre-miRNA sequencing has rather limited efficiency, and further experimental improvements should make our pipeline more powerful. Using locked nucleic acids (LNA) to remove about 300 species of abundant non-pre-miRNA transcripts prior to sequencing resulted to an even lower percentage (0.2% compared to 0.8% in this study) of pre-miRNA sequencing reads at a much higher cost (mainly due to the synthesis of LNA oligos) [96]. In comparison, a PCR based approach yields higher efficiency of over 50% in sequencing pre-miRNAs [97]. However, this method only works on known miRNAs in a gene-specific manner. To our knowledge, we provided the first unbiased genome-wide profiling of miRNA and pre-miRNAs in a cost-effective manner.

Since miRNAs serve as a potent gene expression regulator, the knowledge of the identification, biogenesis, and processing of them are of great importance to the understanding of miRNA-related functions. Here, the tailor-designed computational pipeline, namely miRGrep, allows *de novo* identification as well as expression profiling of miRNAs without relying on the availability of genome reference sequences. This pipeline could be widely used in the miRNA-related studies where the genome reference is of low quality or even absent. In fact, out of 223 organisms whose miRNAs are registered in miRBase (version 21), only half (111) have genome references sequenced, albeit of various quality. For the other half, the miRNA annotation is merely inferred from the mature miRNA sequencing experiments via sequence homology to the ones annotated with genome references, and is therefore likely of high false positive (RNA fragments or degradation products that are not generated by the miRNA biogenesis pathway) and false negatives (organism-specific miRNAs with little homology). MiRGrep could greatly improve the miRNA annotation for metazoans independent of genomic sequences, even for well-studied organisms such as mouse in this study. Moreover, miRGrep might prove to be useful not only to organisms without available genome references, but also samples where the genome sequences differ significantly from the reference, such as cancer. Furthermore, with novel insight gained from this approach, we can improve the current understanding of miRNA processing and modifications that are likely of potential medical implications.

Chapter 5

Circular RNAs identification

5.1 Introduction

5.1.1 Circular RNAs, old acquaintance and new roles

Cellular RNAs can be classified in terms of their structure: linear or circular form. In contrast to a linear RNA that possesses distinct 5' and 3' termini, reflecting the start and end sites of transcription, a circular RNA (circRNA) has no terminal. The closed structure is formed by covalently joining the 3' end of the precursor to the 5' end (Figure 5.1).

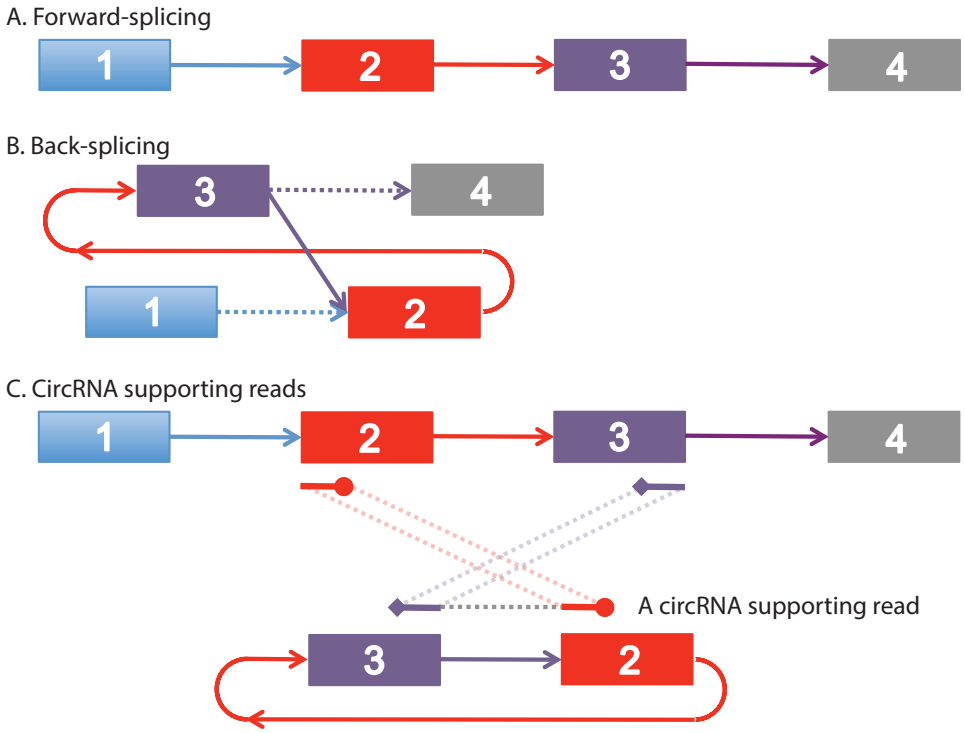

Figure 5. 1 The structure of circRNAs. (A) The Scheme of normal (forward) splicing where the upstream exon2 connect to downstream exon3 in a 5'-to 3' manner. (B) The scheme of back splicing where the downstream exon3 connect to upstream exon2 in a 5'-to-3' manner. (C) A sequencing read that span the back-splice junction of a circRNA cannot be directly aligned to genomic sequences, but can be partially aligned in a reversed manner.

Circular RNAs comprising exonic sequences from a few gene loci were first identified in eukaryotic cells more than 20 years ago [98], [99]. At that time, the observation of circRNAs was regarded as transcriptional artifacts or splicing noise [98], although some circRNAs are extremely abundant (e.g. the one derived from the SRY gene locus in adult testis) [99] and can represent the dominant form of RNA transcript derived from one specific genomic locus [21]. Only recently, the advance of ribosome-depleted RNA sequencing techniques combined with tailored computational tools enables identification of thousands of new circRNAs in organisms ranging from archaea to human [21], [37], [100]. In most cases, circRNAs comprises exonic sequences of protein-coding transcripts (see the discussion section of this chapter for other types). For many years, no clear function was attributed to any of the circRNAs, but it was recently demonstrated that two previously annotated circRNAs could serve as miRNA sponges by sequestering miRNAs and preventing their interactions with target mRNAs [36], [37]. Although this observation offers one function model for circRNAs, circRNAs represent a heterogeneous group of transcripts that likely also exert diverse cellular functions via as yet undiscovered mechanisms. In addition to miRNA regulation, it has been proposed that cytoplasmic circRNAs could sequester RNA-binding proteins (RBPs) and thereby could also regulate the intracellular transport of associated miRNAs, RBPs, or mRNAs [101], [102]. Nuclear circRNAs, on the other hand, can interact with transcription-splicing complex and therefore regulate the abundance of hosting transcripts [10], [103].

RNA-mediated regulation of cellular function and protein translation is crucial for polarized cells to functionalize cellular compartments. This is particularly true for neurons, where the complex morphology and distal location of synapses mandate a high degree of local regulation [104]. Localized protein synthesis has been observed in both dendrites and axons, contributing by the localization of translational machinery and over 2000 protein coding genes in each compartment [73], [105]. And in other cell types, RBPs are important for

both RNA transport and translational regulation [106]. In recent years, other classes of RNA species and RNA-based regulation have been identified in neurons including miRNAs and lncRNAs [107], [108].

In order to study the potential role of circRNAs as a novel RNA-mediated regulatory mechanism, we set out to characterize and profile the expression pattern of circRNAs in several tissues/samples from mice and rat. CircRNAs are found to be enriched in the brain compared to other tissues. Moreover, a large fraction of circRNAs is derived from genes that code synaptic proteins. Using PacBio Sequencing for a subset of circRNA candidates, we identified the rolling circle cDNA products that, for the first time, elucidate the true circular structure of circRNAs. Furthermore, based on the separate profiling of the RNA localized in neuronal cell bodies and synaptic processes (axons and dendrites), we found that, on average, circRNAs are more enriched in the distal part of neurons than their linear isoforms. Using high resolution *in situ* hybridization, circRNAs were visualized directly in the dendrites of neurons. Finally, the abundance of several circRNAs changes at developmental stages that correspond to synapse formation and also following homeostatic plasticity.

5.1.2 Challenges

The property of circularity and low abundance has contributed to the relative anonymity of circRNAs. Although the circRNA population in a cell is about 3% to that of mRNAs as estimated in this study, their representation in high-throughput sequencing results is less than 0.1%. This ostensible discrepancy is due to the fact that most of the reads originated from circRNAs are exactly the same as those derived from linear transcripts, whereas only those reads that span the back-splice junction sites can be unambiguously assigned to circRNAs. Since the circRNA-specific reads cannot be directly mapped to genome or transcriptome, they have to be identified using partial alignment. Therefore to characterize circRNAs in an unbiased manner, computational pipelines should be designed to identify the back-spliced junction sites.

Furthermore, given circRNAs also undergo alternative splicing, the exact sequences of circRNAs should be carefully examined. We present acfs (acronym for Arthurian CircRNA Finder Suite), a user-friendly analysis pipeline that satisfies the challenges (available at https://github.com/arthuryxt/acfs).

5.2 Methods

5.2.1 Sequencing protocol

Ribosomal RNA (including mitochondrial rRNA) is depleted from total RNA using the RiboZero Gold kit (Epicentre Bio-technologies). RNA-seq library is then generated from rRNA-depleted RNA using Illumina stranded RNA Sample Prep kit per manufacturer's instruction, and is subsequently sequenced for 150 nt on Single-End (SE) mode on an Illumina HiSeq 2500. Note that at RNA fragmentation step, one should make sure that the majority of the resulting single-strand RNA fragments are of length larger than the sequencing length, otherwise it will be a waste to sequence that long and the possibility to detect circRNAs will decrease.

5.2.2 Sequencing data preparation

After removing the Illumina sequencing adapter at 3' end, the reads are aligned to the corresponding genome reference (bdgp5 for fly, mm9 for mouse, rn5 for rat and hg19 for human) and annotated transcriptome sequences using Tophat2 [49], allowing up to six mismatches. Any other tools that align reads to both genome and transcriptome can also be used. Cufflinks [49] (v2.21) is then used to estimate the total transcriptional output based on Ensembl gene annotation (v5.72 for fly, v67 for mouse, v72 for rat, and v71 for human). Genes annotated as "protein coding" or "lincRNA" are retained for further analysis. To compare gene expression between two samples, we convert the FPKM (Fragments Per Kilobase per Million) to TPM (Transcripts Per Million) using the following formula: TPM = FPKM * 1000000 / (sum_of_FPKM) [109]. Due to the back-spliced structure of circRNAs, the RNA-Seq reads derived from the junction sites cannot be directly mapped to

either genome or transcriptome references. Therefore, those unmapped reads are collected for the identification of circRNAs.

5.2.3 Fusion reads identification

For each sequencing data set, the unmapped reads are further aligned to the respective genome reference sequences by BWA [110] using local mode (with parameter: -mem –k 16). Maximum exact matches (MEM) are located using Borrows-Wheeler transformation (BWT) and the FM-index. Longer alignments with tolerable mismatches and/or gapes are finalized on the extended MEMs using Smith-Waterman (SW) algorithm. There are seven categories of alignment results:

1) unmapped;

2) only one segment is aligned but not end-to-end match;

3) two segments aligned to the same chromosome and on the same strand;

4) two segments aligned to the same chromosome but on the opposite strands;

5) more than two segments aligned to the same chromosome and all on the same strand;

6) more than two segments aligned to the same chromosome but on opposite strands;

7) more one segments aligned to different chromosomes.

Alignments of category 3 and 5 that satisfying the following requirements are retained as candidates supporting back-spliced junctions:

1) regions on the same chromosome and no more than 1Mb away from each other;

2) on the same strand;

3) in reverse order;

4) alignment score (AS, the Phred-sacled probability of the alignment being incorrect) higher than 30 for all segments to ensure 99.9% accuracy

There still can be circRNA supporting reads in category 1 and 2 due to the requirement of alignment length and score, thus they should be re-examined

after circRNA identification. Alignments of category 4 and 6, where alignments are on opposite strand of the same chromosome, could partly be explained by the RNA-dependent RNA Polymerase (RDRP) activity in the library preparation [111]. Alignments of category 4, 6 and 7 could contribute to trans-splicing, if the break points match the canonical splicing sites of host genes and the partial alignments are in the same orientation of the host genes.

5.2.4 Back-splice site identification

Due to the sequence similarity of the termini of exons, the partial alignments within a single back-spliced read tend to show an overlap most frequently of 2nt, and the exact splicing site is usually located in the vicinity of the overlap region. To identify the splicing sites in an unbiased manner, the most probable sites are determined by their splicing strength, instead of just looking for one of the canonical configurations (GU-AG pair). The strength of potential splicing sites supported by these candidate back-spliced reads is estimated using MaxEntScan [112]. The exact junction site is then determined by selecting the donor and acceptor site pair with the highest splicing strength score that is allowed by the read. Let n be the length of a back-spliced read originated from a gene locus on the +strand, and for simplicity, there are only two partial alignments. Let $[1,i]$ mark the first alignment and $[j,n]$ mark the second, where $j <= i$. Let $S_5(k)$ mark the strength of 5' splicing site at the genomic position given by the k^{th} position of the read and $S_3(k)$ mark the strength of 3' splicing site at the genomic position given by the k^{th} position of the read. The objective is to maximize the function:

$$SSum(x) = S_5(i - x) + S_3(j + x)$$

where $0 <= x <= (i - j + 1)$

For cases with more than two partial alignments, the two adjacent partial alignments $[a,i]$ and $[j,b]$ are considered, where $(0 <= a) \& (b <= n) \& (j <= i)$ and the genomic coordinates $J < B < A < I$. An authentic back-splice site should be of SSum value higher than 10, corresponding to about 95% of all pairs of splice site of mouse Ensembl genes (Figure 5.2).

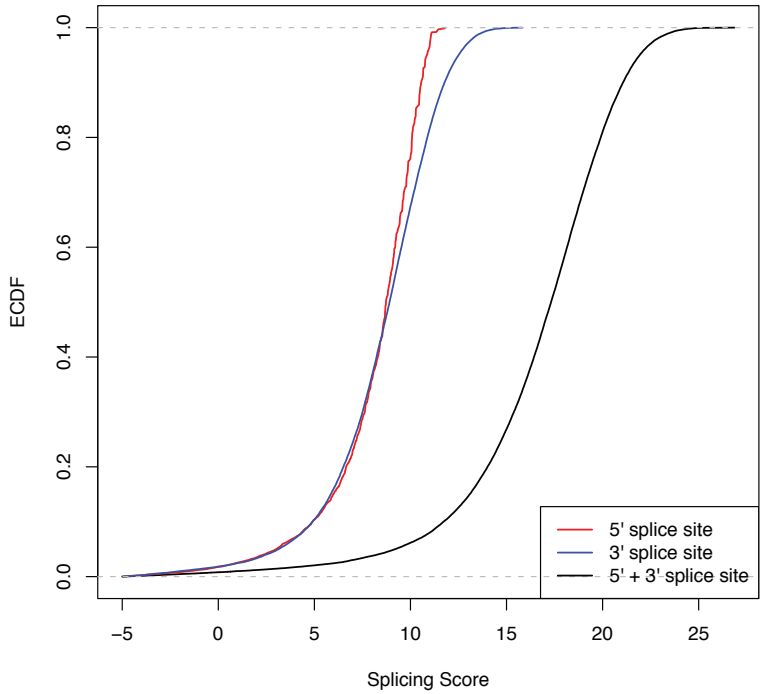

Figure 5. 2 Strength of canonical splicing sites. The empirical cumulative distribution of splicing strength on 5' splice site (red), 3' splice site (blue) or a pair of splice site (black) of Ensembl genes.

5.2.5 Filtering

Since the genome reference sequences to which back-spliced reads aligned are of various qualities, i.e. the genome reference for human is of high quality whereas that for rat is still in a rather impoverished status, the authenticity of the predicted back-splice sites should be carefully examined. If two exons of a certain gene locus exchange the order in the annotation by mistake (possibly owing to misassembly), then the reads originated from that locus would appear to match with back splicing. To filter out possible false-positives introduced by faulty genome assembly, the sequences of the identified back-splice sites (upstream 50nt and downstream 50nt or shorter if the adjacent exon is shorter than 50nt) are extracted and aligned to the transcript database (Refseq RNA sequences or Nucleotide collection) using Blast. Back-splice sites with successful alignments of over 95% similarity are excluded, since they are highly likely to originate from linear transcripts.

5.2.6 Abundance estimation

To estimate the abundance of circRNAs, all unmapped reads are re-aligned to circRNA sequences in an end-to-end fashion. As for most of the circRNAs, there is no direct evidence for their exact sequences; the exonic sequences are filled in using existing annotation. For those circRNAs containing short exons, the internal sequences could be fully or partially determined if the sequencing length is long enough. For circRNAs whose back-splice sites do not overlap with the exon borders of existing annotations, especially when one of the splicing sites locates within the intronic or intergenic region, the whole sequence between the back-splice sites is taken with all splice-able introns removed.

In order to mimic the circular structure, the sub-sequence from the 5' end of each circRNA candidates of length equivalent to sequencing length is appended to the 3' end. RNA-Seq reads that mapped to the junction (with an overhang of at least 6nt as default to ensure unambiguous alignment) are counted for each circRNA candidate. TPM (Transcript Per Million) is calculated for each candidate, where the effective length was set to: (sequencing length - 2 * 6). Note that the sum of TPM of all circRNAs gives more accurate estimation of the relative abundance of circRNAs as a group with regard to total RNAs.

5.2.7 Conservation analysis

The positions of rat circRNAs were converted to mouse (mm9) genome coordinates using the UCSC liftOver tool, then were intersected with mouse circRNAs using BEDTools. To examine he evolutionary conservation of the para-junctional sequences of mouse circRNAs, PhastCons scores for alignment of 29 vertebrate genomes with mouse (mm9) was downloaded from (http://hgdownload.soe.ucsc.edu/goldenPath/mm9/phastCons30way/vertebrate/). To rule out possible biases, the sequences around the splicing sites involved in the back-spliced junctions were compared to those not involved on the same gene locus.

5.2.8 PacBio sequencing of RT-PCR products

The RT-PCR products obtained from the mouse brain and rat brain samples were directly sequenced using PacBio RS system as in Chapter 3. The circular consensus reads (CCS reads) obtained from the PacBio sequencing were aligned to custom database (consisting of sequences fro both linear mRNAs and circRNAs) using Blast (parameters: -evalue 1E-10 –word_size 9). Alignments with identity higher than 95% were reported, and can be subsequently inspect for rolling-cycle products.

5.2.9 MiRNA binding potential

To quantify the density of miRNA binding sites on circRNAs, the number of predicted miRNA binding sites (nearly fully complementary, 7mer-1A, 7mer-8m and 8mer sites) [113] was counted for all miRNAs (deposited in miRBase version19). As a control, the same procedure was preformed on CDS and 3'UTR of the protein-coding genes.

5.2.10 RBP binding potential

The RBP binding sites on circRNAs were predicted based on their sequence motifs deposited in RBPDB [114]. As a control, the predicted RBP binding sites on the circRNAs were compared to those on CDS and 3'UTR of protein-coding genes.

5.2.11 Peptide translation potential

The translational capacity of circRNAs could be estimated from their association with ribosome complex. Polysome profiling was done on mouse brain samples and ribosome footprinting [115] was done on rat brain samples. Sequencing reads from five fractions of mouse brain (non-ribosome, 40S sub-unit of ribosome, 60S sub-unit of ribosome, mono-ribosome, and poly-ribosome) as well as Ribosome Protected Fragments of rat brain were aligned

to circRNAs using BWA, and the reads spanning the circular junctions were counted and converted to TPM as described above.

To directly test the potential peptides predicted from circRNAs, a liquid chromatography mass spectrometry sequencing was done on total lysate from 21-day-old primary neurons without any pharmacological or electrophysiological treatment. The circRNA sequences were translated in three potential frames, and the position of the circRNA junction was recorded. This custom database was then merged together with the rat protein RefSeq database, and subsequently used as a template for peptide matching with Mascot. Peptides spanning the circular junction sites were recorded.

5.3 Results

5.3.1 Enrichment in brain

To systematically examine the possibility of tissue-specific expression pattern of mammalian circRNAs, ribosomal RNA (rRNA) depleted total RNA samples from different mouse tissues, including brain, liver, lung, heart and testes, were subject to RNA-Seq. Both sequencing depth and mappability were similar in all biological replicates (Table 5.1). Reads that map directly to reference genome sequences or canonical exon-exon junctions can be derived from either linear mRNAs or circRNAs and therefore were used to estimate the expression of the total transcriptional output (hereafter referred to as TTO) of the corresponding gene loci. To specifically identify circRNAs, we used the remaining reads that spanned the 5' and 3' splicing sites of exon(s) of individual genes; but in reverse order (Figure 5.1). From the five tissues, we detected a total of 10641 unique circRNAs.

Samples	No. total reads	No. mapped reads	No. circRNAs
Brain rep1	19794174	18765595	6186
Brain rep2	19164999	18283420	5664
Heart rep1	16507635	14636897	989
Heart rep2	19876852	19049052	1315
Liver rep1	21001677	20195781	912
Liver rep2	19056514	18322339	816
Lung rep1	20390058	19595737	1556
Lung rep2	18406517	17282775	1320
Testes rep1	19940919	18802286	2943
Testes rep2	20222389	19198630	3093

Table 5. 1 Summary of circRNA sequencing results

Although circRNAs were identified in all tissues we examined, their abundance was clearly highest in brain (Figure 5.3 A), where 20% of the protein-coding genes produced circRNA (Figure 5.3 B). Two factors contributed to the higher abundance of circRNAs in brain. First, many circRNA-hosting genes were expressed exclusively in brain (Figure 5.3 C). Second, on average, when a host gene was expressed in brain as well as other tissue(s), the proportion of transcription output that is directed to biogenesis of circRNA is significantly higher in brain than in other tissue(s). To examine the second factor, we compared the relative contribution of circRNAs (defined as the ratio of TPM values between a circRNA and the TTO of its hosting gene) between samples (Figure 5.3 D).

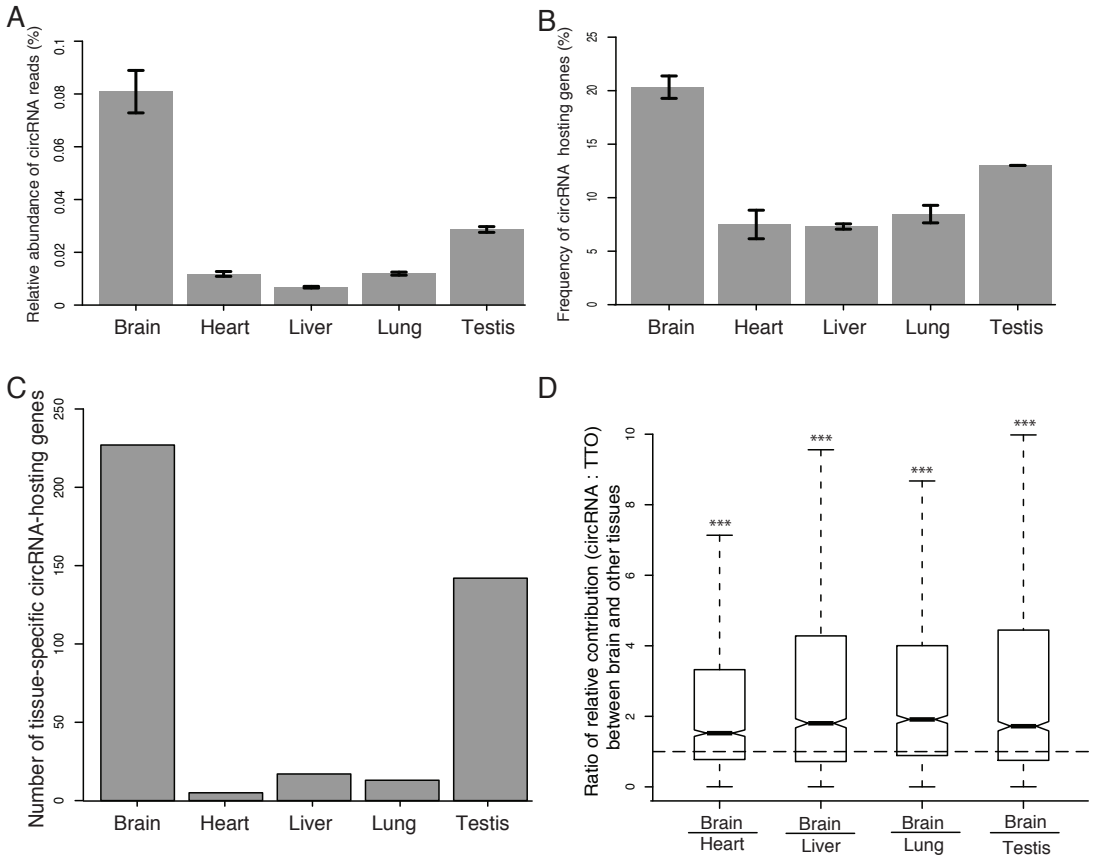

Figure 5. 3 Mouse circRNAs are enriched in brain. (A) The percentage of circular junction reads from all mapped reads (both genomic and transcriptomic) is shown for the five tissues, with the highest in brain. (B) The percentage of circRNA-hosting genes from all genes is shown for the five tissues, with the highest in brain. (C) The number of tissue-specific genes that host circRNAs is shown for the five tissues, with the highest in brain. (D) The relative contribution of circRNA to TTO of the same gene locus is significantly higher in brain compared to all other tissues. Error bars represent standard deviation.

The observation that there are three, on average, sequencing reads per circular junction would lead to questioning the robustness of the estimation of such enrichment. Indeed, there are many factors at play, including sequencing depth, sequencing read length and the threshold for calling an expressed circRNA. As sequencing read length clearly has a positive contribution to the detection of circRNAs, we did simulations with various thresholds and sub-sampled datasets. As shown in Figure 5.4, the relative contribution of circRNAs in brain is consistently and significantly larger than

those in other tissues.

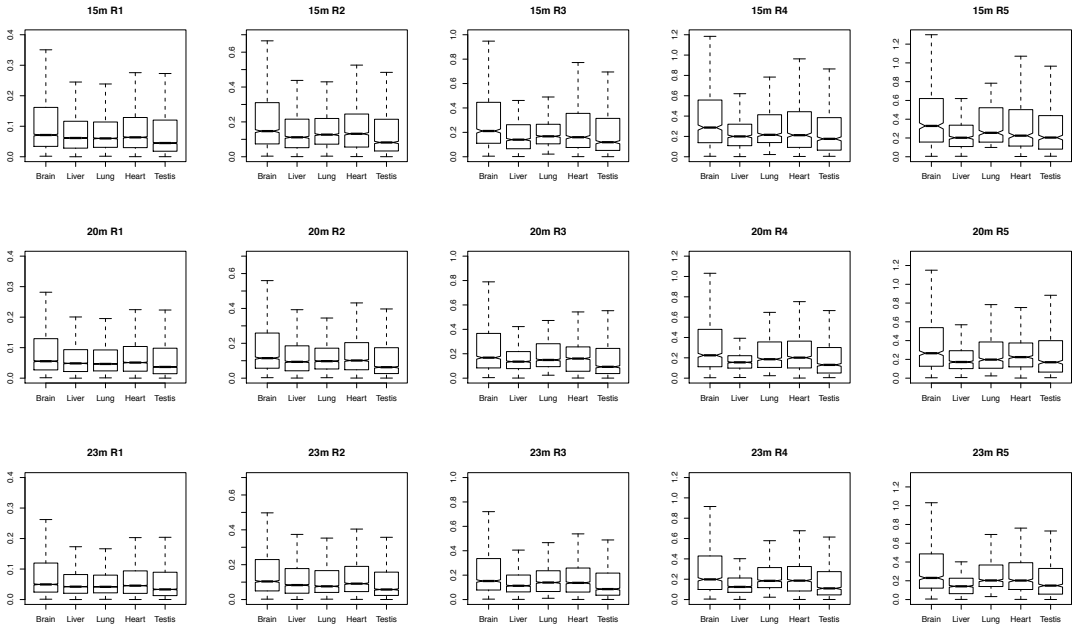

Figure 5. 4 Simulation of sequencing depth and thresholds. The caption “15m R1” stands for “sub-sampling down to 15 million reads” and “requiring at least 1 circular junction read”, and alike. The relative contribution of circRNAs is shown as box plot for the five tissues.

We re-analyzed published RNA-Seq datasets from various tissues of rats. Although the sequencing length (50nt) is much shorter in this datasets, which drastically reduce the chance to detect circRNAs, we observed a similar enrichment of circRNAs in brain (Figure 5.5).

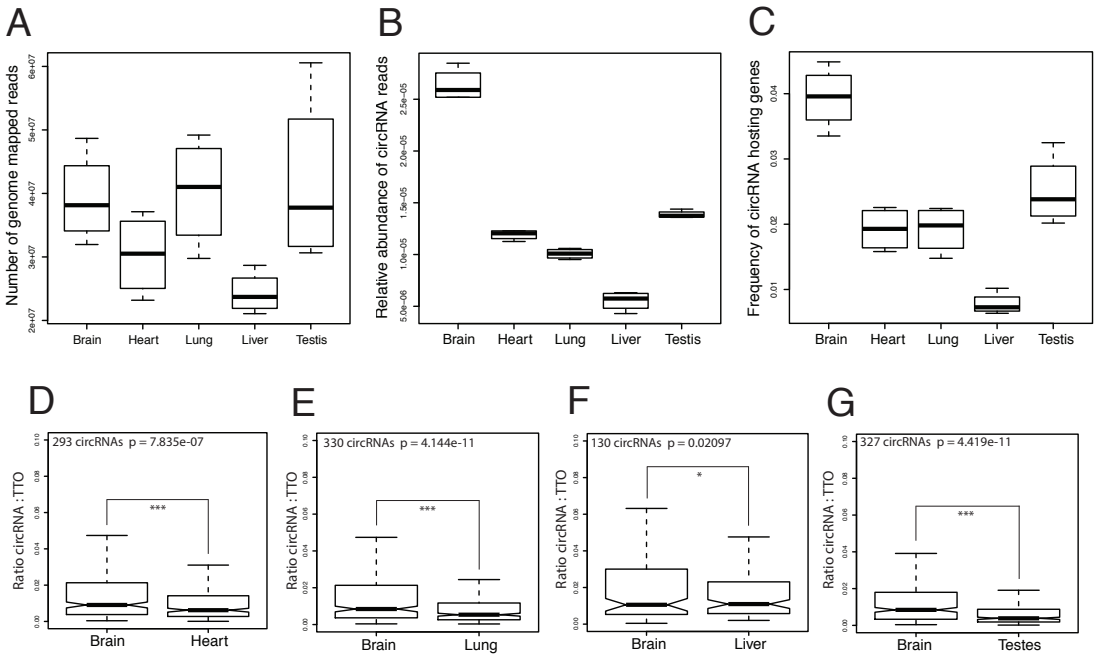

Figure 5. 5 Rat circRNAs are enriched in brain. (A) The number of mapped reads is shown for the five tissues. (B) Relative abundance of circRNA reads is shown for the five tissues. (C) The frequency of circRNA hosting genes is shown for the five tissues. (D-G) The relative contribution of circRNAs is significantly higher than that of heart (D), hung (E), liver (F) and testes (G).

In the study of fly samples, we also found that circRNAs were more abundant in brain compared to the body (Figure 5.6), which is consistent with a recent report [116]. We further make suggestions for future sequencing studies of circRNAs: sequencing in single-end mode with at least 100 nt and 20 million reads per sample.

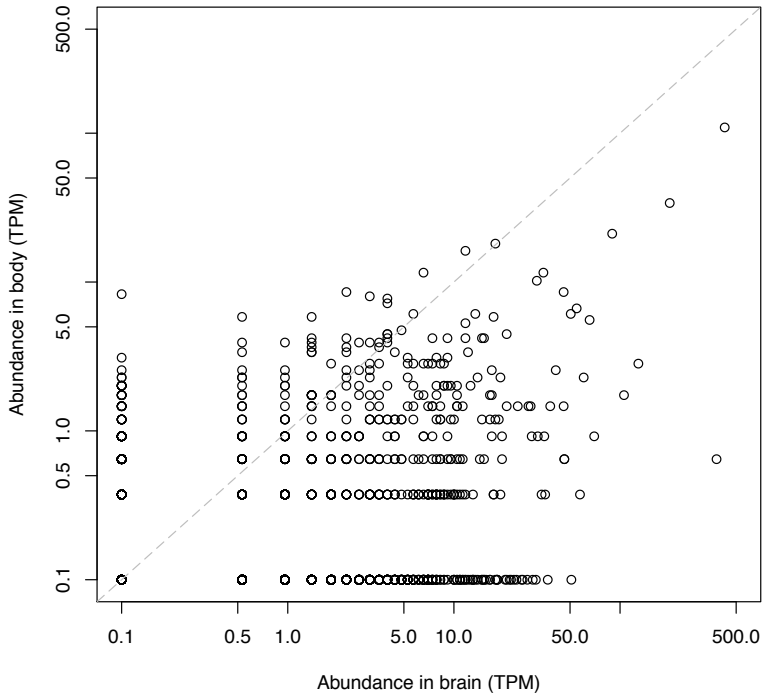

Figure 5. 6 Fly circRNAs are enriched in brain. Each circle represents one fly circRNA, with the values on X- and Y-axis represent the abundance (TPM) in brain and body of the fly, respectively.

5.3.2 Independent validations

We validated the authenticity of circRNAs predicted in our study by three independent methods. First, as circRNAs do not possess a poly(A) tail, their representation should be depleted in a poly(A)-enriched sequencing library. Compared with rRNA-depleted total RNA sequencing library, poly(A)-enriched RNA sequencing library from the same sample produced a much lower number of reads originated from circRNA population (Figure 5.7). Of note, this protocol cannot completely erase the representation of circRNAs, since the oligo(dT) probes used for poly(A)-enrichment can still capture or even enrich some of the circRNAs whose sequences contain a stretch of As.

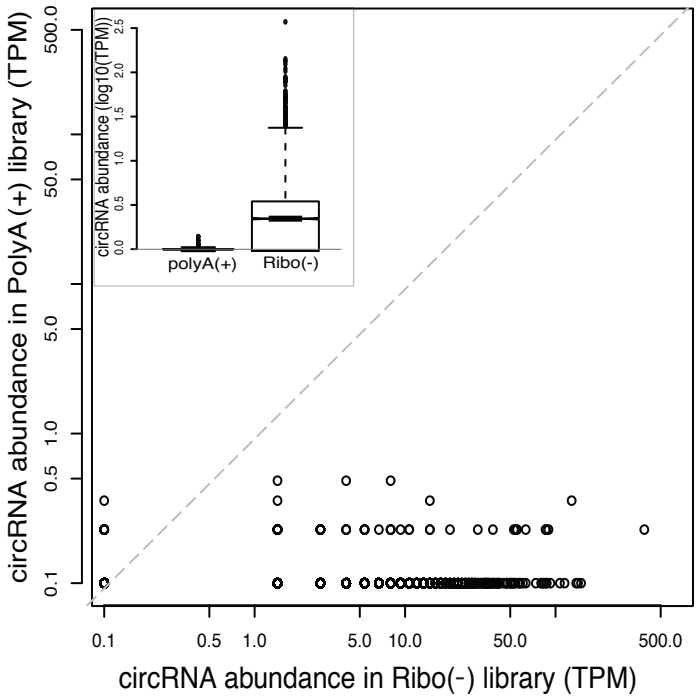

Figure 5. 7 CircRNAs are depleted in a poly(A) library. Each circle represents one mouse circRNA. Values on X- and Y-axis denote the abundance (TPM) of circRNAs in Ribo(-) and PolyA(+) library, respectively. Inset shows that circRNAs are significantly depleted in the PolyA(+) library.

Second, since circRNAs benefit from the closed structure that endows strong resistance to the exonucleases such as RNase R, they should be more stable than the linear transcripts upon such treatment. We therefore quantified the RNase R resistance of 20 circRNA candidates, and all of them exhibited stability at least five-fold higher than their linear counterparts (Figure 5.8). The increase of abundance for some circRNAs after RNase R treatment suggests that they are more resistant to RNase R than the well accepted control 5S rRNA.

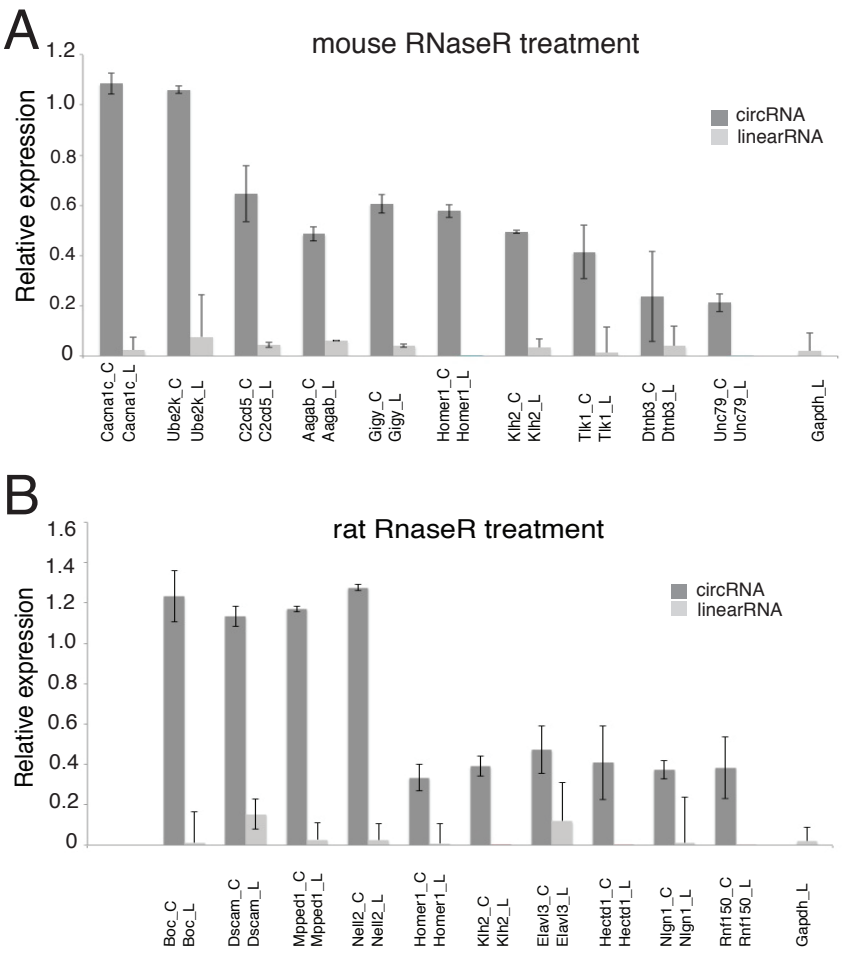

Figure 5. 8 CircRNAs are resistant to RNase R. CircRNAs from mouse (A) and rat (B) are at least 5-fold more resistant to RNase R compared to their hosting linear transcripts.

Third, we sequenced the RT-PCR products derived from 12 circRNA candidates using the PCR primers that anchor on the circular junctions. Expected PCR products were detected for all 12 candidates. Moreover, for 11 of them, we observed the PacBio sequencing reads corresponding to the rolling circle RT products (Figure 5.9).

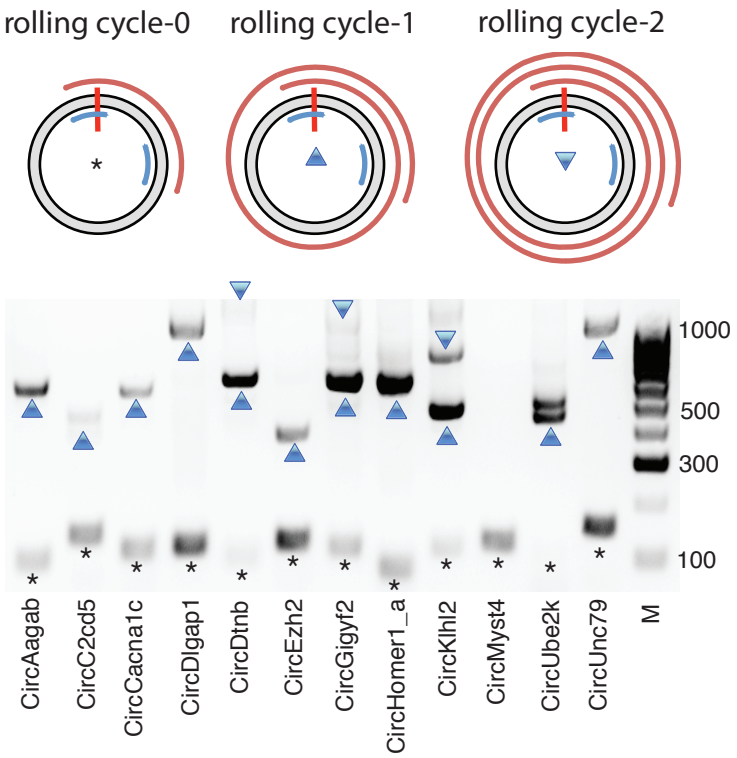

Figure 5. 9 Rolling cycle products of circRNAs. Each grey ring stands for one circRNA, the red vertical bar for the back-spliced junction, two blue arcs for the PCR primers, and the red spirals outside of circRNAs for the RT-PCR products that were deep sequenced using PacBio technology. The asterisk, upward and downward triangle symbols on the gel image denote the 0-, 1- and 2-cycle RT-PCR products, respectively.

For the only one circRNA candidate without detected rolling circle RT-product (circMyst4), we speculated that the RT process would be difficult to complete even one rolling circle, since the estimated length of the circRNA is about 3 kb. As a control, we also did PacBio sequencing for the PCR products of the linear transcripts of the same 12 gene loci using poly(A)-enriched RNA. From the control sequencing results, we found only expected PCR products from the linear transcripts, without a single read that can be unambiguously attributed to circRNAs. This observation serves as direct evidence for the circular nature of the circRNA structure and, to our knowledge, is the first time that the full-length sequences of circRNAs have been identified. Notably, for two circRNAs (circDtnb and circEzh2), in addition to the “canonical” forms that

encompass all of the annotated exons between the two back-splicing sites, we also observed circular isoforms that consisted of the same junction sequences, but with one internal exon skipped (circDtnb) or one unannotated exon inserted (circEzh2). As observed in our study and previous reports, multiple circRNAs with different back-splicing junctions could be produced from the same gene loci. The identification of circRNA isoforms with the same back-splicing junction, but different internal sequences, adds another layer to circRNA diversity and possibly regulatory functions. The fact that the internal exon composition cannot be simply predicted using back-splicing junctions necessitates the experimental determination of full-length sequence of a circRNA before further functional investigations.

5.3.3 Synaptic gene origin and dendritic localization

The observation of brain circRNAs enrichment prompts the question on the potential selected production of circRNAs: Is there a positive correlation between the relative contribution of a circRNA and the function of its host gene? To address it, we conducted a Gene Ontology analysis for the genes that give rise to circRNAs in brain. Interestingly, several functional groups related to synaptic functions such as synapse, presynaptic active zone and postsynaptic density were significantly enriched categories in the neuronal circRNA population (Figure 5.10).

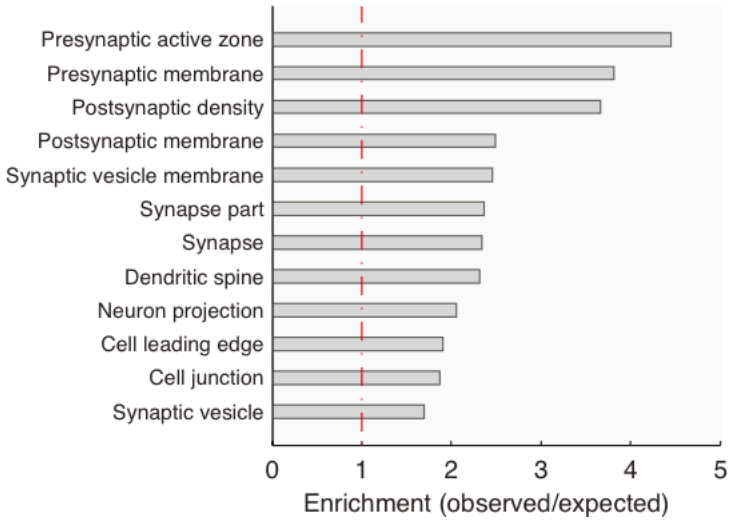

Figure 5. 10 GO analysis of brain circRNAs. Functional groups related to synaptic functions are overrepresented in the genes that host brain circRNAs.

Given the enrichment of host genes with synapse relate functions, we next examined whether the circRNAs are enriched in synaptic tissue. To address it, we prepared synaptosomes, a biochemically purified preparation that is enriched in synapses [117] or microdissected the synaptic neuropil from the hippocampus, a brain structure that exhibits robust synaptic plasticity and is important for learning and memory [118]. We then compared the abundance of circRNAs in these compartments (synaptosomes or neuropil) to that in a whole hippocampal homogenate or a microdissected layer comprising primarily hippocampal neuronal somata. We found that most circRNAs are indeed enriched in the synaptic fractions examined (Figure 5.11 A, B) and the overlap between the two synaptic fractions was statistically significant (p-value < 2.2E-16, Fisher Exact Test). The same pattern of results was obtained when the tissue was obtained from rat (Figure 5.11 C) and there was substantial overlap between the circRNAs identified in mouse and rat (Figure 5.11 D-F).

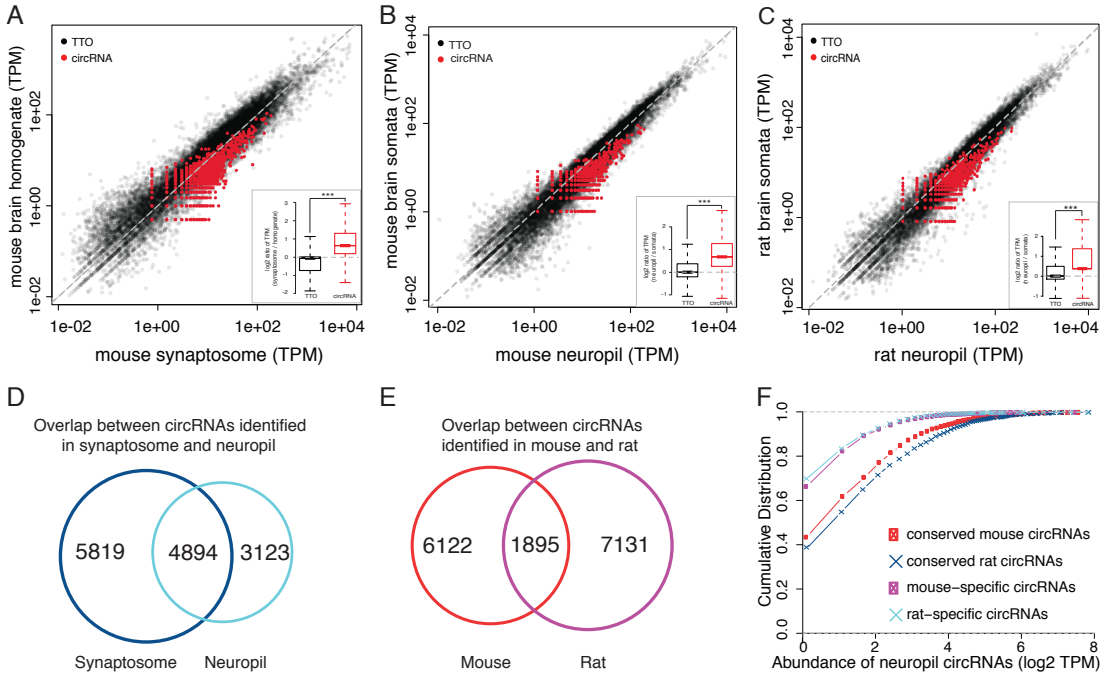

Figure 5. 11 Brain circRNAs are enriched in synapse. (A-C) The abundance of circRNAs and TTO of protein-coding gene loci (TPM) were compared between the synaptosomes (X-axis) and whole brain (Y-axis) in mouse (A); neuropil (X-axis) and somatic layer (Y-axis) of the hippocampus in mouse (B); or neuropil (X-axis) and

somatic layer (Y-axis) of the hippocampus in rat (C). Inset shows that the abundance of circRNAs, but not TTO, was significantly higher in the synaptic fractions. (D) The overlap of circRNAs identified in mouse synaptosomes and neuropil layer. (E) The overlap of circRNAs identified in mouse neuropil and rat neuropil. (F). Cumulative distribution of circRNA abundance shown in (E), in which circRNAs identified in both mouse and rat neuropil layer are of higher abundance than the others.

Moreover, we directly visualized the circRNAs using high-resolution in situ hybridization [73]. In cultured hippocampal neurons, we detected circRNA particles distributed in the cell body as well as in the dendrites, visualized using an antibody against a dendritic marker (anti-MAP2) (Figure 5.12). Similarly, we demonstrated via in situ hybridization that there are substantial expression of circRNAs in both somata and neuropil layers of CA1 hippocampal region (Figure 5.13).

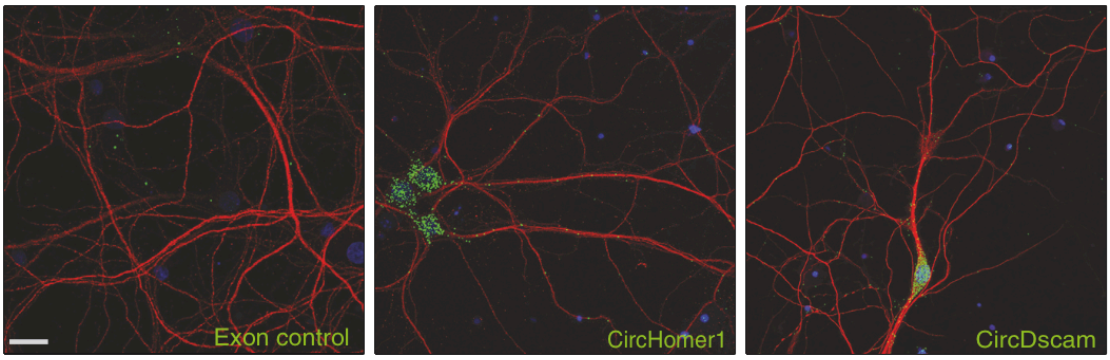

Figure 5. 12 Visualization of circRNAs in cultured neurons. CircRNA-positive particles (green) are apparent in the cell bodies (blue, nuclei stained with DAPI) and in the dendritic processes, which is illustrated via an antibody to MAP2 (red). As a negative control, a control exon probe designed to detect non-contiguous region of two exons that do not form a back splicing junction yielded few background particles (left panel). Scale bars represent 20 microns.

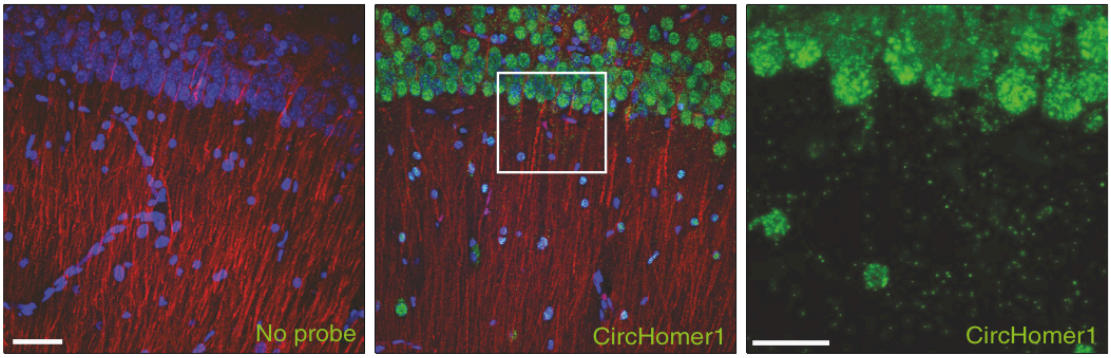

Figure 5. 13 Visualization of circRNAs in hippocampal slices. CircRNA-positive particles (green) are apparent in the cell bodies (blue, nuclei stained with DAPI) and in the dendritic processes, which is illustrated via an antibody to MAP2 (red). As a negative control, no signal was detected when no probe was used for hybridization (left panel). Scale bars represent 50 microns.

We validated the specificity of our circRNA in situ hybridization by comparing the signal intensity of circRmst, circKlhl2 and circGigyf2 in brain, liver and lung (Figure 5.14 A). Consistent with the RNA-Seq data, the in situ hybridization data revealed only background levels of expression of circRmst and circKlhl2 in liver and lung, in comparison to their evident enrichment in hippocampal neurons. In contrast, circGigyf2 was expressed in all examined tissues as expected from RNA-Seq data (Figure 5.14 B). To test whether the circRNA localization can mimic that of its host transcript, we performed in situ hybridization of circRims1 and its host mRNA Rims1 in cultured hippocampal neurons. Although signals for the circRNA and mRNA were apparent in both cell body and dendrites, they clearly did not co-localize (Figure 5.14 C). Given the anticipated diversity of circRNA populations, however, one must be open to counterexamples of co-localization of circRNA and mRNA when more cases are examined.

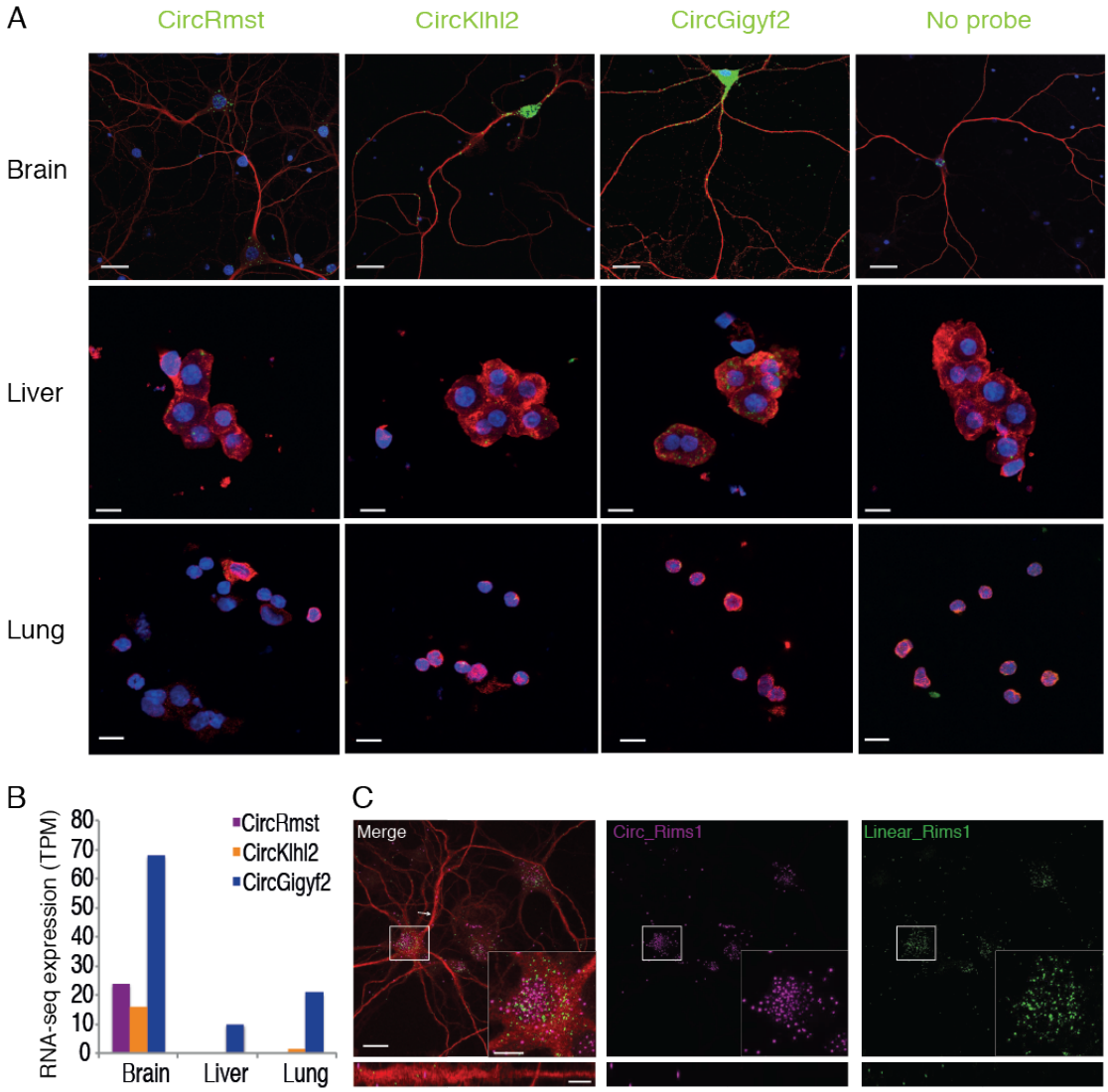

Figure 5. 14 Validation of circRNA localization. (A) Validation of abundance of circRNAs in brain, liver and lung using in situ hybridization. No probe control indicates background level of fluorescence. Nuclei were stained by DAPI (blue), circRNAs in situ hybridization punctae were shown in green, antibody staining (MAP2 for dendrites in brain, Albumin for liver and Hemo-Oxygenase 1 for lung) is shown in red. Scale bar = 20 microns for brain and 10 microns for liver and lung. (B) Abundance (TPM) of three circRNAs in brain, liver and lung based on RNA-Seq measurement. (C) In situ hybridization of circRims1 (purple) and linear Rims1 transcript (green) in the same neuron shows lack of co-localization. Whole image and zoom-in of soma and dendrites are shown. Scale bar = 20 microns for the whole image and 10 microns for zoom-in.

5.3.4 MiRNA binding potential

Recent studies of two individual circRNAs suggested that they function as miRNA “sponges”, sequestering miRNAs [36], [37]. By searching the potential miRNA binding sites, we estimated the potential of the brain circRNA population to serve as miRNA sponges. Although there are many cases that circRNAs possess several miRNA binding sites, the brain circRNAs as a group do not exhibit a greater capacity to serve as miRNA sponges than linear mRNAs (Figure 5.15), consistent with recent analysis from other groups [119]. However, it should be noted that in order to function as miRNA sponge, circRNAs do not have to outcompete the linear transcripts with regard to binding density, their extraordinary stability and localization pattern could render strong regulatory effect at a specific niche and/or upon specific stimulations.

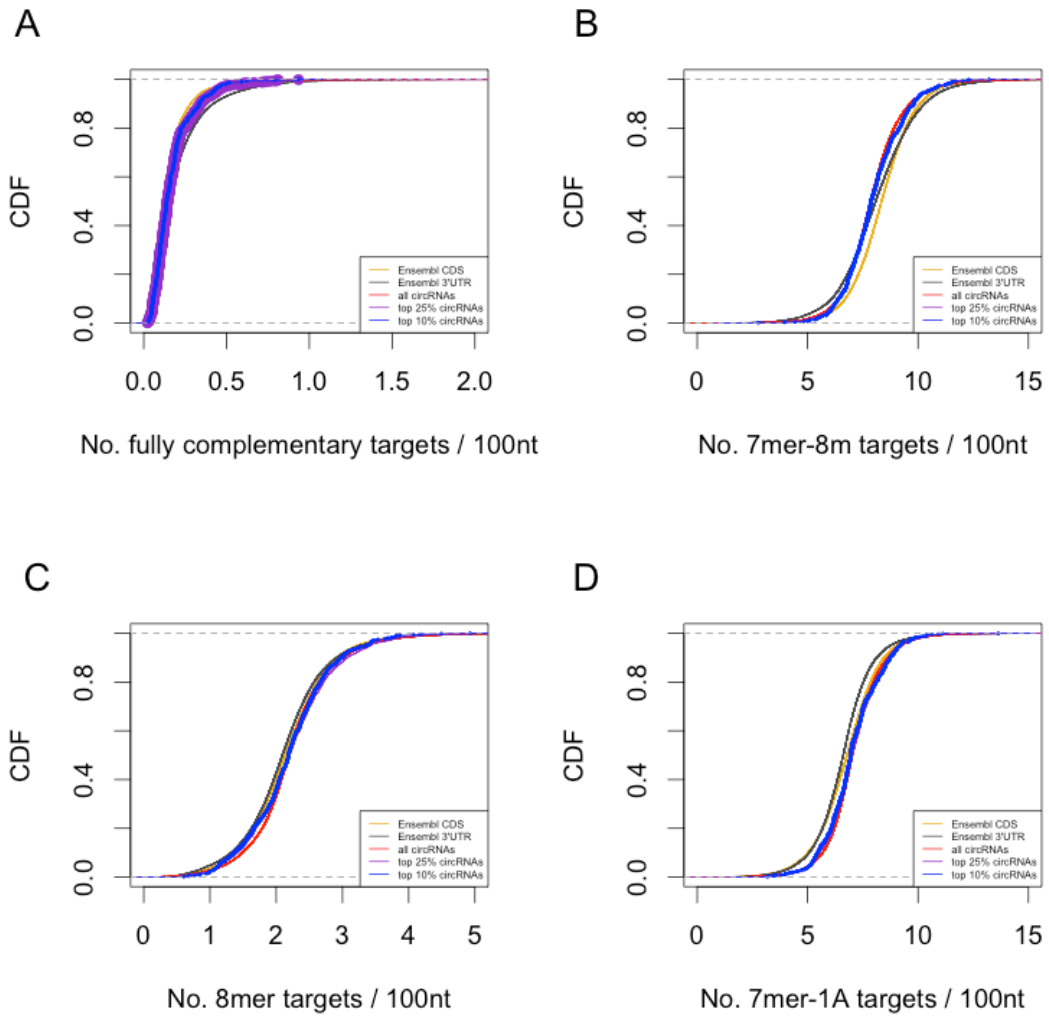

Figure 5. 15 MiRNA binding potential of circRNAs. The density of miRNA binding site of fully complementary targets (A), 7mer-8m targets (B), 8mer targets (C) and 7mer-1A (D). Based on nucleotide sequences, circRNAs (red) as a population do not possess a higher density of miRNA binding sites than that of either 3' UTR (black) or CDS (coding sequence, yellow) of the mRNAs. This trend remains when circRNAs of different abundances are examined.

5.3.5 RBP binding potential

We also examined the possibility that circRNAs might function to bind or sequester RBPs. Here, we predicted the binding sites of 38 RBPs based on the binding sequence motifs deposited in the RBPDB. CircRNAs possess a lower RBP binding density, when compared to either the coding sequences or the 3’ UTR of protein-coding genes (Figure 5.16). This trend is consistently

observed for circRNAs with different abundances.

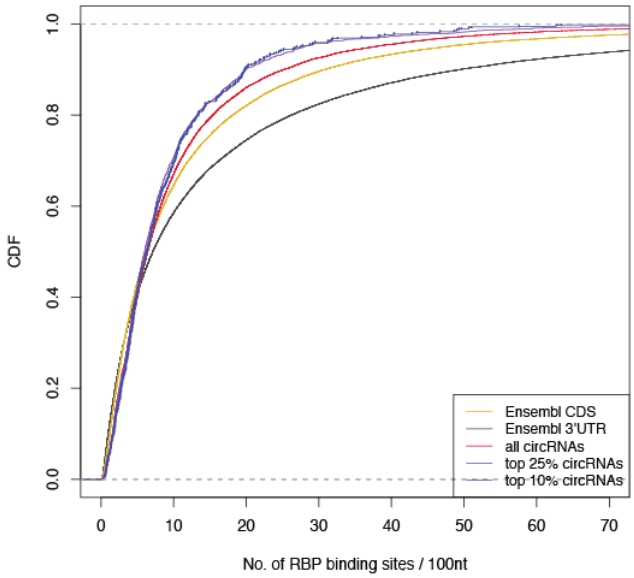

Figure 5. 16 RBP binding potential of circRNAs. CircRNAs (red) have a lower density of RBP binding sites than that of either 3' UTR (black) or CDS (coding sequence, yellow) of the protein coding genes. This trend remains when circRNAs of different abundances are examined.

5.3.6 Peptide translation potential

Given the fact that neuronal circRNAs were mostly composed of protein-coding exons, we investigated their potential to be translated into peptides with three independent approaches. Using a large mass spectrometry (MS) dataset obtained from hippocampal neurons, we searched for peptides predicted by circular junctions but without any success. The inability to detect a circRNA-derived peptide, however, could be a result of the well-known low detection sensitivity of MS-based shotgun proteomics approaches. Thus, we further studied the association of circRNAs with ribosomes. We performed ribosome profiling on rat brain samples in order to find the footprints of ribosome on circRNAs. Similar to what was recently reported for circRNAs from a human cell line [119], we did not detect a single ribosome protected fragments (RPF) that overlap with circular junctions. This negative observation could be resulted from the short read length of RPFs and the possibility that ribosome might bind to sequences outside of circular junctions.

To circumvent this limitation, we performed polysome profiling on mouse brain samples. In contrast to mRNAs, circRNAs were enriched in the non-ribosomal RNA fraction and strongly depleted in the ribosome/polysome-bound RNA fractions (Figure 5.18). Together, these results indicate that circRNAs as a group are unlikely to be translated into peptides.

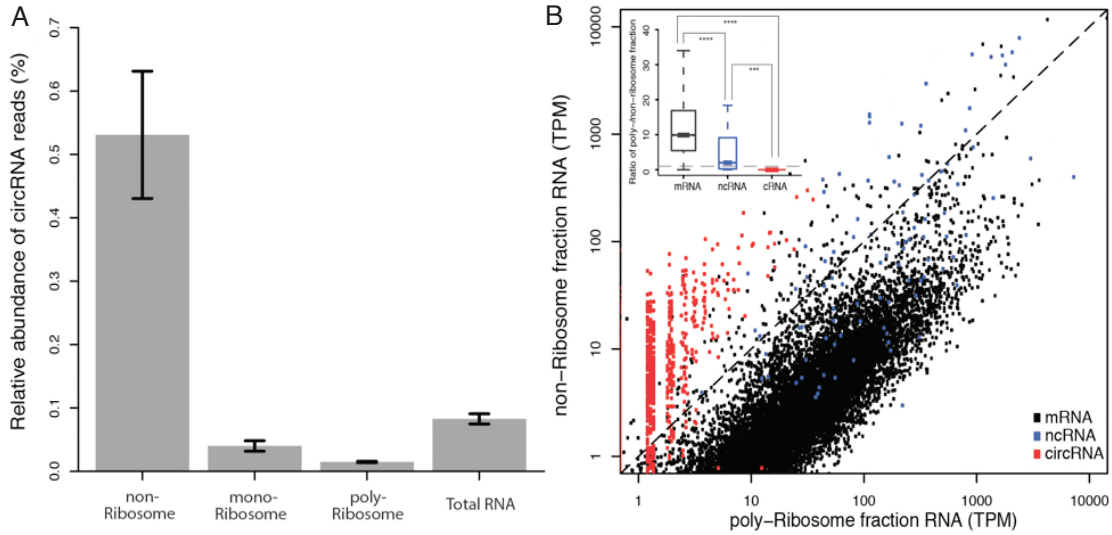

Figure 5. 17 CircRNAs associate less with ribosomes. (A) CircRNAs are enriched in the non-ribosome fraction and are drastically depleted from the mono-/poly-ribosome fraction. Values on the Y-axis denote the percentage of circRNAs in the RNA fractions marked on the X-axis. (B) CircRNAs (red) are enriched in the non-ribosome fraction, whilst protein-coding genes (black) are enriched in the poly-ribosome fractions. Classical non-coding RNAs such as snRNAs and snoRNAs are shown in blue. The inset shows that circRNAs are of significantly less associations with ribosomes that that of classical non-coding RNAs.

5.3.7 Conservation

As functionally important elements are often evolutionarily conserved, we examined the exonic sequence conservation around the mouse circRNA junctions. Compared with splicing sites of the same host gene those are not involved in forming back-splicing junctions, the exonic sequences around the back-splicing junctions are more conserved (Figure 5.19). Moreover, for circRNAs that are conserved in both mouse and rat, their para-junctional sequences are on average 10% more conserved than that of the non-conserved circRNAs, almost reaching the maximum PhastCons scores.

Moreover, we analyzed the overlap of circRNAs detected in rat and mouse, and found 23.6% of the circRNAs identified in mouse neuropil were also expressed in rat neuropil (Figure 5.19). This observation is consistent with a recent study in which 20% of mouse circRNAs were detected in human cell lines [119], but it was higher than the estimation in another study in which only 4% of the mouse circRNAs were identified in human samples [37]. The difference might be explained by different sampling depths, as most identified circRNAs were expressed at low levels and might therefore 'stochastically' escape detection. Indeed, the circRNAs detected in both mouse and rat samples were clearly of much higher abundance than those detected in only one sample (Figure 5.19). This further suggests that circRNAs are conserved, and our observation of a 23% overlap between mouse and rat circRNAs may prove to be an underestimate. Together with the para-junctional sequence conservation, the conservation of circRNAs expression strongly suggests functional relevance.

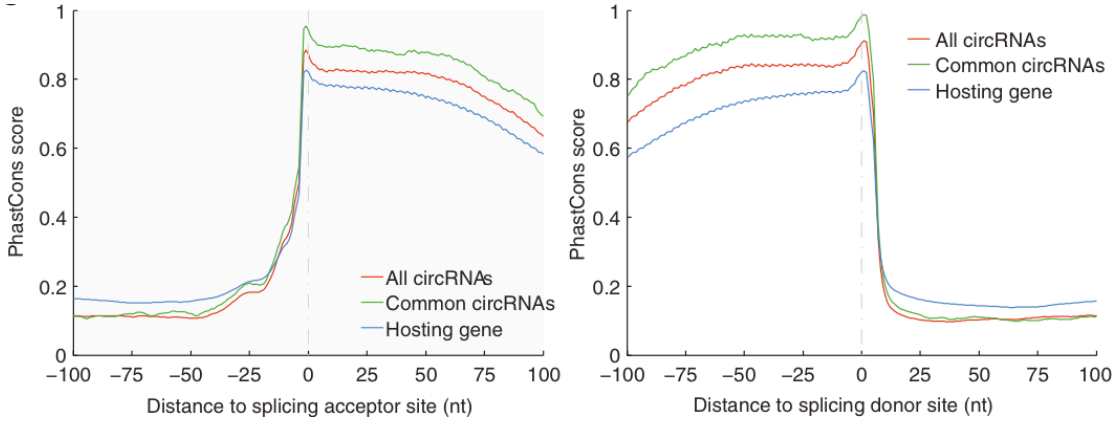

Figure 5. 18 CircRNAs are evolutionarily conserved. The exonic sequences around the splicing sites (left, splicing acceptor; right, splicing donor) involved in the formation of mouse circRNA back-spliced junctions (red) are more conserved than those from the same gene locus but not involved (blue). Values on Y-axis denote the average PhastCons score, values on X-axis denote the distance to the splicing site (negative and positive values means upstream and downstream, respectively). Importantly, the para-junctional sequences common in mouse and rat (green) are even more conserved, almost reaching the maximum PhastCons score.

5.3.8 CircRNAs are regulated in brain during development

The development of the CNS and brain involves neuronal maturation, neurite outgrowth and synaptogenesis. Non-coding RNAs such as miRNAs and lncRNAs have emerged as important components for regulating these developmental processes [107], [120]. To determine whether the expression of circRNAs is developmentally regulated in brain, we profiled the circRNA population in the hippocampus over several stages: embryonic (E18), early postnatal (P1), postnatal at the beginning of synapse formation (P10) and late postnatal hippocampus following the establishment of mature neural circuits (P30). There was a clear shift in the circRNA expression pattern associated with the onset of synaptogenesis at P30 (Figure 5.20 A). Notably, the circRNAs that were consistently upregulated during hippocampal development were produced from the gene loci that also code for proteins enriched with synapse-related functions (Figure 5.20 B). In contrast, no enrichment of any functional categories could be observed for the gene loci showing the opposite (downregulated) circRNA dynamic expression pattern.

We next examined the relationship between the expression of a circRNA and its linear host comparing the earliest (E18) and latest (P30) developmental stages. We found that many circRNAs change their expression independent of their host transcripts during synaptogenesis (Figure 5.20 C). We validated 13 circRNA and mRNA pairs with different expression patterns using quantitative PCR (Figure 5.20 D). Dlgap1, whose protein product is a core component of postsynaptic density (PSD), showed an >20-fold increase in circRNA expression at P30 when compared with E18, whilst the mRNA abundance increased by less than four fold. Genes such as Myst4 (aka Kat6b, associated with Ohdo and Genitopatellar syndrome), Klhl2, and Aagab (Alpha-and-gamma-adaptin binding protein, involved in clathrin-coated vesicle trafficking) drastically increased their circRNA expression over development whereas their mRNA expression decreased markedly. In contrast, Cacna1c showed substantial decreases in circRNA expression along developmental stages, while the mRNA abundance remain unchanged. Transcripts from

lncRNA RMST locus were recently identified as a factor that is important for neuronal differentiation as well as a co-regulator for SOX2, a mediator of neural stem cell fate [121]. Two circRNAs as well as the linear transcript from the RMST locus were downregulated during the development, thereby supporting a potential function of circRNAs in brain development. Using high-resolution in situ hybridization in cultured hippocampal neurons, we further validated the developmental regulation of circKlhl2 that exhibited strong upregulation during development (Figure 5.20 E). Analysis of the average fluorescence intensity at an early and late developmental stage (neurons cultured beginning at P1, days in vitro = 4 or 21) revealed a significant enhancement of the circKlhl2 expression levels ($P < 0.0001$; Figure 5.20 F). Thus, taken together the data from high-throughput sequencing, quantitative PCR and in situ hybridization, indicate that the expression of circRNAs is developmentally regulated in neurons and that many circRNAs change their expression independent of their host linear transcripts, suggesting a circRNA-specific regulation of biogenesis and/or turnover (Figure 5.20 C,D).

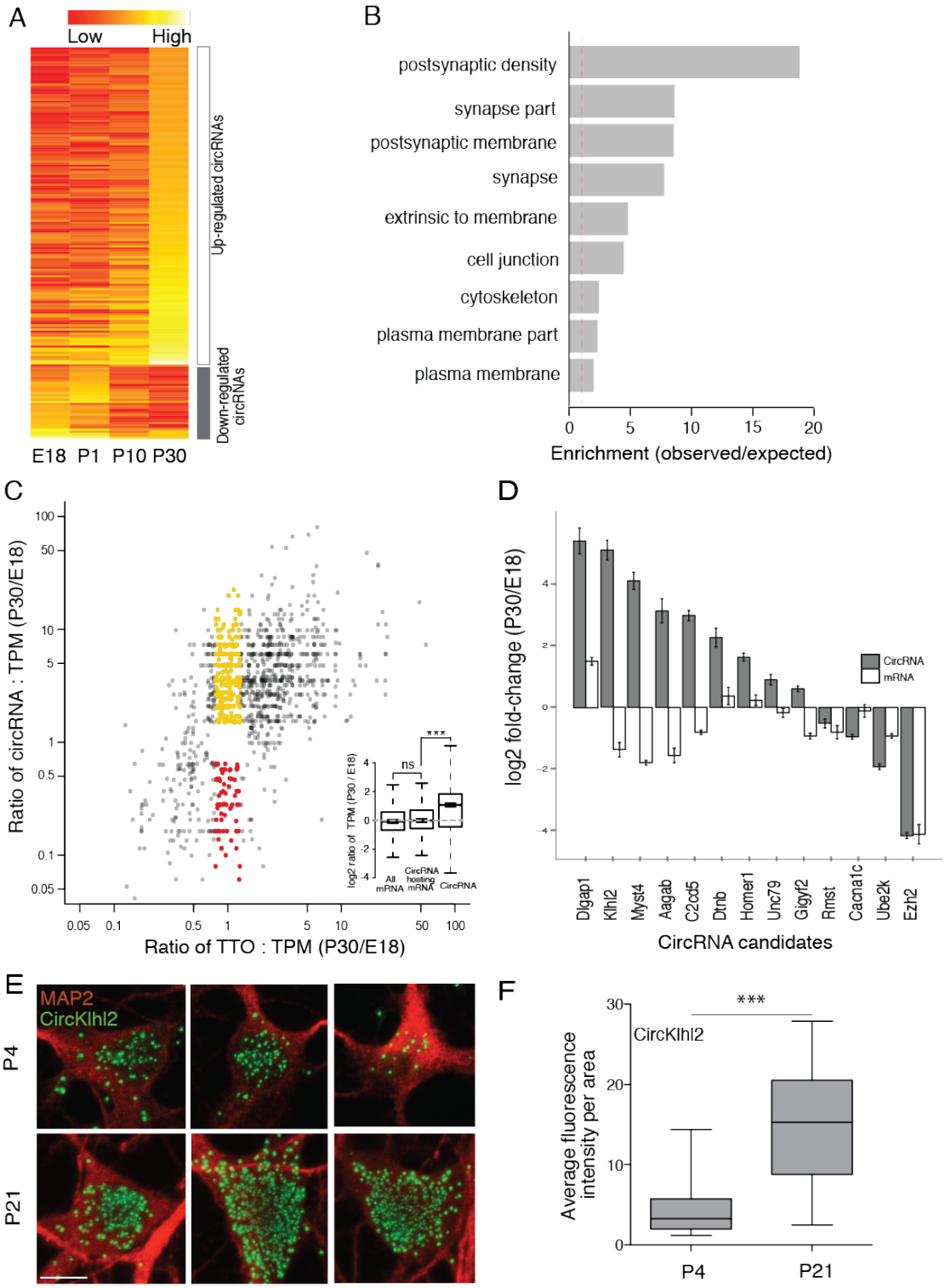

Figure 5. 19 Regulation of circRNAs during brain development. (A) Heat map of circRNA expression along developmental stages showing the regulation of several clusters between P1 and P10, when synapses typically form. The abundance of circRNAs is shown on a scale from red (low) to yellow (high). A developmentally downregulated cluster reaches a peak in expression at E18 or P1, and the declined in the later stages. In contrast, a much bigger cluster of circRNAs is low in

abundance in the earlier stages and then peaks at P10 or P30 in expression. (B) The significantly enriched GO terms. The hosting genes of circRNAs that are upregulated along development are enriched for synaptic functions, whereas the downregulated group is not enriched for any GO terms. (C) Fold change of abundance of both circRNAs (Y-axis) and the TTO of the host gene loci (X-axis) between the two extreme stages E18 and P30. Each dot represents one pair of circRNA/gene locus. Dots in red and yellow highlight circRNAs whose abundance changed significantly while the TTO of the corresponding host gene did not. Inset shows that although most circRNA hosting gene loci did not change much in abundance, circRNAs were significantly upregulated. (D) The change of abundance for both circRNAs and hosting protein-coding genes was validated using quantitative PCR. Error bar stands for standard deviation. (E,F) Validation of change of circRNA abundance between developmental stages using high-resolution *in situ* hybridization. Quantification of the florescence intensity is shown in (F). The outline of neuronal soma was illustrated using an antibody against MAP2 (red). Scale bar represents 10 microns.

5.3.9 CircRNAs change their expression as a result of neuronal plasticity

As our data suggests that circRNAs might regulate synaptic functions, we went on to study the possibility that the abundance of circRNAs could be modulated by alterations in neuronal activity and plasticity. We induced homeostatic synaptic plasticity in cultured hippocampal neurons by manipulating neuronal activity using bicuculline, an antagonist to the GABAA receptor. Treatment with bicuculline enhanced excitatory neuronal network activity, leading to a homeostatic decrease in the mini-excitatory postsynaptic current (mEPSC) amplitude, without a change in mEPSC frequency (Figure 5.21 A) [122]. Following the induction of homeostatic plasticity, the circRNA population exhibited dynamic behavior: the expression of 37 circRNAs was enhanced (Figure 5.21 B), whereas that of 5 circRNAs was reduced. In contrast, most of their linear host transcripts showed no substantial change in expression level (Figure 5.21 B). We validated the plasticity-induced changes in four circRNA candidates using quantitative PCR (Figure 5.21 C). We also

visualized directly the circRNA expression changes after homeostatic plasticity for additional candidates using in situ hybridization. Notably, a circRNA (circHomer1_a) derived from the Homer1 linear transcript was significantly upregulated circRNA after plasticity induction in primary hippocampal neurons ($P < 0.0005$ for somata and $P < 0.0208$ for dendrites) (Figure 5.21 D,E) and hippocampal slices (Figure 5.21 F). Taken together, these data indicate that circRNA expression levels are regulated by neural plasticity, suggesting that they are important for regulating synaptic transmission and/or local translation.

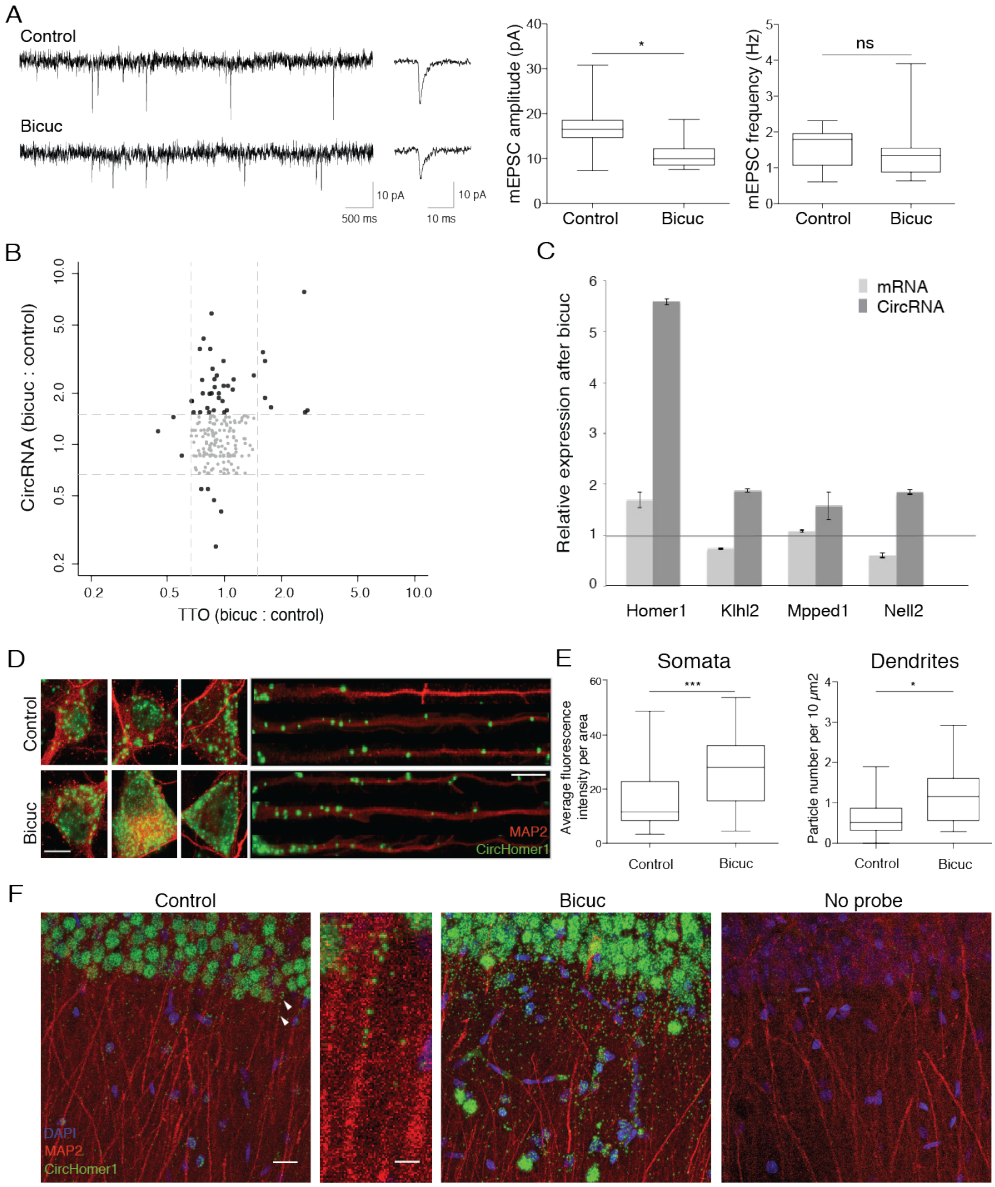

Figure 5. 20 Regulation of circRNAs during homeostatic plasticity. (A) Electrophysiology traces of mEPSCs from control neuronal cultures and bicuculline treated (12hr) neuronal cultures. Representative recordings are shown to the left, and

the average mEPSC waveform, amplitudes and frequency are shown to the right. (B) Scatter plot shows the change of abundance of circRNAs (Y-axis) and the TTO of corresponding gene loci (X-axis) after included homeostatic plasticity. Each dot represents one pair of circRNA/gene locus. Grey dots represents pairs whose abundance remained largely unchanged (less than 30% change) for both circRNA and TTO. (C) Quantitative PCR validation of abundance change after homeostatic plasticity using high-resolution in situ hybridization in control or bicuculline-treated neurons. Dendrites are illustrated using an antibody against MAP2. Scale bar represents 10 microns. (E) The abundance of circHomer1_a upregulated significantly in both cell body and dendrite after homeostatic plasticity. (F) The abundance of circHomer1_a upregulated significantly in hippocampus slices after homeostatic plasticity. Control slice, zoom-in of the arrowhead indicated regions, bicuculline-treated slice and no-probe control are shown from left to right. Scale bar represents 20 microns with the exception of 5 microns in zoom-in. Bicuculline-treatment, as a well-established method to induce homeostatic plasticity, resulted in a significant upregulation of circHomer1_a in both stratum pyramidale (somatic layer) and stratum radiatum (neuropil layer).

5.4 Discussion

5.4.1 Mechanisms of functions

Eukaryotic circRNAs are a class of low-abundance, but biochemically stable, cellular RNAs that possess neither a 5' nor a 3' end. The property of circularity has contributed to their relative anonymity (until recently), as most of the transcriptome-wide studies begin with the purification of a poly(A) RNA fraction. Similar to other recent studies, we sequenced and analyzed rRNA-depleted samples that allow one to analyze both circRNAs and their linear hosting transcripts in an unbiased and quantitative manner. Although circRNA identification relied on available genome annotation in previous studies, we established a computational pipeline that does not rely on gene annotations or assume canonical splice sites, and can therefore identify circRNAs derived from previously unannotated genomic regions. This allowed us to identify the circRNAs in rat, which, to date, has a relatively impoverished genome

reference and annotation.

We found that circRNAs are most abundant in brain compared to other tissues and the brain-expressed circRNAs are derived from genes that code for proteins with synapse-related functions. Moreover, we, for the first time, visualized the sub-cellular localization of individual circRNA species both in vitro and in vivo. While many schemes of possible functions have been proposed for circRNAs, we found that they can bind to miRNAs and RBPs but in general not exceed the extent of that of the host linear transcripts. However, circRNAs exhibit strong evolutionary conservation in both sequence level and expression level, which serves as a strong indicator of their functional relevance.

While it has been reported circRNAs with developmental-stage-specific expression in *C. elegans* (oocyte, sperm and embryonic stages) [37], we present the evidence for developmental regulation of circRNAs in neurons in this study. The development of the CNS/brain involves neuronal maturation, neurite outgrowth and synaptogenesis. Non-coding RNAs such as miRNAs and lncRNAs have emerged as key players in regulating these developmental processes [123], [124]. However, for most of them the molecular mechanisms by which they function are still unknown. Recently, the lncRNA RMST was identified as a factor important for neuronal differentiation as well as functioning as a co-regulator of SOX2, a mediator of neural stem cell fate [121]. We identified a set of circRNAs to be differentially expressed in the mouse hippocampus at different developmental stages (E18 to P30). Intriguingly, a circRNA that was significantly downregulated at later stages arises from the linear transcript coding for Rmst, thus supporting a potential function of circRNAs in brain development. In contrast, the expression of circKlhl2 was increased at P30 (P21) compared to E18 (P4) indicating a putative role of this circRNA during synaptogenesis or when mature synapses have formed. In summary, we showed a shift in the expression pattern for a large set of circRNAs associated with the onset of synaptogenesis, indicating

role of circRNAs in hippocampal development.

The brain is the most plastic organ and its circuits underlay tight regulation and modification throughout the entire lifespan of animals. Both the stability and flexibility of neuronal networks is key to all perception, behavior as well as learning and memory. Experience-dependent alterations in the connectivity of neural networks can result in plasticity of intrinsic excitability and synaptic connections. We induced homeostatic plasticity by treating cultured hippocampal neurons with bicuculline and observed a dynamic change in circRNA expression. Interestingly, a circRNA (circHomer1_a) derived from the Homer1 linear transcript was the most significantly upregulated circRNA after plasticity. The Homer1 protein plays a major role in the organization of the postsynaptic densities. It regulates mGluR function [11] and is implicated in neurological disorders such as Parkinson Disease and Schizophrenia [125]. It is known that neuronal activity causes an increase in expression of immediate early gene variants of Homer1 [126]. The increase in circHomer1_a expression levels we observed coincided with an up-regulation of Homer1a mRNA suggesting a potential co-regulation of the circRNA and its linear host. However, this co-regulation may be an exception since only few circRNAs showed co-regulation with their host genes; more common was the observation that circRNAs exhibited changes opposite to those shown by the linear host mRNA following plasticity. Thus, our findings indicate a potential for the existence of diverse mechanisms of action for different sets of circRNAs in synaptic plasticity.

Finally, as a heterogeneous group of transcripts, it is very likely that circRNAs affect cellular and neuronal function via a diverse set of mechanisms. The different datasets accumulated in this study should serve as a rich resource for future functional research, where genetic perturbation of specific circRNAs followed by careful phenotypic examination in different in vitro and/or in vivo neuronal systems will be needed to shed more light on circRNA function in the nervous system and specifically to address their role in learning and memory.

5.4.2 Other types of circular RNAs

Recently, two other types of circular RNAs have been identified: circular intronic long non-coding RNAs (ciRNAs) and exon-intron circular RNAs (EIciRNAs). Among many shared or specific features of the circular RNAs (Table 5.2), ciRNAs are not generated from the splicing process and therefore is the only type (so far) of circular RNAs that cannot by identified using our computational pipeline acfs. In fact, EIciRNAs can be viewed as a variant of circRNAs, where the introns between the circular junctions are not completely spliced out. Together with the fact that there are multiple isoforms of circRNAs with the same circular junctions, the exact full-length sequence of circular RNAs must be identified before any further functional analysis.

	circRNA	ciRNA	EIciRNA
Biogenesis	5'-3' ligation mediated by spliceosome	Un-debranched intron lariat	5'-3' ligation mediated by spliceosome
Sub-cellular localization	Mostly in cytoplasm	Enriched in nucleus	Exclusively in nucleus
MiRNA binding sites	Many	Few	Few
Functions	Post-transcriptional regulation	Transcriptional regulation (mostly on host genes)	Transcriptional regulation (mostly on host genes)
Can be identified using acfs	Yes	No	Yes

Table 5. 2 Types of circular RNAs

5.4.3 Fusion transcripts

Similar to circRNAs, fusion transcripts originate from coupled splicing and joining of two different primary RNA transcripts (termed as trans-splicing, in contrast to back-splicing for circRNAs). Many trans-splicing event have been observed in cancer samples [127]–[130], and their importance lies in a simple fact that the generation of fusion transcripts with novel (often pathological) functions does not rely on genomic rearrangement and therefore is much

easier to take place. Given the similarity of biogenesis between fusion transcripts and circRNAs, acfs can also be used in fusion transcript identification and expression profiling. The existence and more importantly the change of their abundance may be indicative of the cellular status or response to certain stimuli, either endogenous or exogenous, in a physiological or pathological context.

5.4.4 Improvements

High throughput sequencing data, especially those with long read length and/or paired-end mode, can be used to infer the internal sequences of circRNAs. Moreover, the absolute and relative coverage between the exons and introns could be used to estimate the exon and intron composition of the circRNA sequences. Our data suggest that circRNAs are more abundant in vivo (from cells with physiological context) than in vitro (cell cultures), which explains the low detection rate in most current profiling studies using cell lines, and suggests broader in vivo profiling studies of circRNAs.

Summary and discussion

The knowledge of the transcriptome landscape is crucial for understanding various mechanisms of molecular biology, and more importantly for disease diagnosis and precise treatments. Broadly speaking, three layers contribute to the importance of the transcriptome landscape. First, the profile of all isoforms of protein-coding genes determines, by and large, the development path of cells and organisms. As over 90% of the protein-coding genes in human undergo alternative splicing, abnormal alternative splicing is observed in all eight hallmarks of cancer and neurological diseases. The abnormal isoforms play critical roles in promoting oncogenesis. In cancer, the aberrant usage of 5' and/or 3' UTR, via differential choice of the alternative promoters and/or polyadenylation sites, alters not only the coding sequences but also the regulatory elements in the UTRs, thereby affects the fate of the RNA transcripts and influences their functions. Second, the profile of various regulatory elements modulates the activity of protein-coding genes. Small non-coding RNAs participate in development and various diseases. RBPs can alter the fate of RNA by either bind to RNA directly or indirectly. Proteins modulating DNA methylation and a variety of chromatin modifications have direct influence on RNA landscape. Third, the interplay of RNA transcripts and regulatory elements shapes the dynamic property of transcriptome landscape. Expressional changes in RNA lead to difference in protein abundances, which in turn modulates the expression profile of the trans-regulatory elements such as ncRNAs and RBPs, thus forms a dynamic co-regulatory network consisting of feed-forward-loops and feed-back-loops. Mechanistic information could be extracted from data profile in the form of time-series, or differential compartmentalization, or even heterogeneous population. In fact, such information is not only valuable to reverse engineering molecular biology, it can also be used to guide clinical diagnosis and make treatment plans. Identifying the players in the regulatory network is the first step of decoding

molecular biology. In this thesis, I present tailored analysis on four specific projects belonging to the above two layers.

First, a hybrid assembly pipeline is developed for identification of transcriptome independent of genomic sequences. By combining two complementary sequencing technologies in conjunction with efficient cDNA normalization, a high quality transcriptome can be characterized. It out-performs other assembly tools that focus on one type of data input using one algorithm, and the results are experimentally validated.

Second, an analysis framework is developed for characterization of full-length transcripts. By tailoring tools for long read-length sequencing technology, transcriptome landscape could be examined with greater detail. Moreover, the association of different RNA processing events could be experimentally measured. The application on fly Dscam gene transcripts resolved the independent splicing hypothesis and called for re-examination of previous experiments. The application on rat brain greatly enhanced the transcriptome annotation, which is crucial for the neuroscience community that use rat as a model organism.

Third, a *de novo* microRNA prediction tools is presented. By designing sequencing experiments that capture snapshots of miRNA biogenesis process, not only mature miRNAs and precursor miRNAs could be identified, but also the information on miRNA processing and modification could be learnt. Proof-of-principle experiments on well-studies organism like mouse and *C. elegans* demonstrate the efficacy and application potential of this method.

Finally, a customized pipeline is developed for characterizing circRNAs as a novel group of regulatory RNAs. By examining potential splicing junctions based on local alignments, circRNAs can be identified from the “junk” RNA-Seq data. Tens of thousands of circRNAs are identified and quantified when

applied to mouse, rat and fly. Further experiments demonstrate that circRNAs are enriched in brain synapses and participate in brain development and neuronal homeostatic plasticity.

In summary, this thesis presents tailored analysis on four different aspects of transcriptome landscape. With big data to come, the methods can be used in conjunction towards an integrated understanding of molecular biology and medicine.

Bibliography

[1] C. Adamidi, Y. Wang, D. Gruen, G. Mastrobuoni, X. You, D. Tolle, M. Dodt, S. D. Mackowiak, A. Gogol-Doering, P. Oenal, A. Rybak, E. Ross, A. Sánchez Alvarado, S. Kempa, C. Dieterich, N. Rajewsky, and W. Chen, "De novo assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics.," *Genome Res.*, vol. 21, no. 7, pp. 1193–200, Jul. 2011.

[2] W. Sun, X. You, A. Gogol-Döring, H. He, Y. Kise, M. Sohn, T. Chen, A. Klebes, D. Schmucker, and W. Chen, "Ultra-deep profiling of alternatively spliced Drosophila Dscam isoforms by circularization-assisted multi-segment sequencing.," *EMBO J.*, vol. 32, no. 14, pp. 2029–38, Jul. 2013.

[3] S. a O. Armitage, W. Sun, X. You, J. Kurtz, D. Schmucker, and W. Chen, "Quantitative profiling of Drosophila melanogaster Dscam1 isoforms reveals no changes in splicing after bacterial exposure.," *PLoS One*, vol. 9, no. 10, p. e108660, Jan. 2014.

[4] N. Li, X. You, T. Chen, S. D. Mackowiak, M. R. Friedländer, M. Weigt, H. Du, A. Gogol-Döring, Z. Chang, C. Dieterich, Y. Hu, and W. Chen, "Global profiling of miRNAs and the hairpin precursors: insights into miRNA processing and novel miRNA discovery.," *Nucleic Acids Res.*, vol. 41, no. 6, pp. 3619–34, Apr. 2013.

[5] X. You, I. Vlatkovic, A. Babic, T. Will, I. Epstein, G. Tushev, G. Akbalik, M. Wang, C. Glock, C. Quedenau, X. Wang, J. Hou, H. Liu, W. Sun, S. Sambandan, T. Chen, E. M. Schuman, and W. Chen, "Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity.," *Nat. Neurosci.*, vol. 18, no. 4, pp. 603–10, Apr. 2015.

[6] P. I. W. de Bakker, G. McVean, P. C. Sabeti, M. M. Miretti, T. Green, J. Marchini, X. Ke, A. J. Monsuur, P. Whittaker, M. Delgado, J. Morrison, A. Richardson, E. C. Walsh, X. Gao, L. Galver, J. Hart, D. a Hafler, M. Pericak-Vance, J. a Todd, M. J. Daly, J. Trowsdale, C. Wijmenga, T. J. Vyse, S. Beck, S. S. Murray, M. Carrington, S. Gregory, P. Deloukas, and J. D. Rioux, "A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC.," *Nat. Genet.*, vol. 38, no. 10, pp. 1166–72, Oct. 2006.

[7] E. S. Lander, "Initial impact of the sequencing of the human genome.," *Nature*, vol. 470, no. 7333, pp. 187–97, Feb. 2011.

[8] W. Fiers, R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, F. Molemans, A. Raeymaekers, A. Van den Berghe, G. Volckaert, and M. Ysebaert, "Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene.," *Nature*, vol. 260, no. 5551, pp. 500–7, Apr. 1976.

[9] Y.-K. Kim and V. N. Kim, "Processing of intronic microRNAs.," *EMBO J.*, vol. 26, no. 3, pp. 775–83, Mar. 2007.

[10] Y. Zhang, X.-O. Zhang, T. Chen, J.-F. Xiang, Q.-F. Yin, Y.-H. Xing, S. Zhu, L. Yang, and L.-L. Chen, "Circular intronic long noncoding RNAs.," *Mol. Cell*, vol. 51, no. 6, pp. 792–806, Sep. 2013.

[11] P. R. Brakeman, A. A. Lanahan, R. O'Brien, K. Roche, C. A. Barnes, R. L. Huganir, and P. F. Worley, "Homer: a protein that selectively binds metabotropic glutamate receptors.," *Nature*, vol. 386, no. 6622, pp. 284–8, Mar. 1997.

[12] S. J. Lee, "Expression of growth/differentiation factor 1 in the nervous system: conservation of a bicistronic structure.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 88, no. 10, pp. 4250–4, May 1991.

[13] P. Khaitovich, G. Weiss, M. Lachmann, I. Hellmann, W. Enard, B. Muetzel, U. Wirkner, W. Ansorge, and S. Pääbo, "A neutral model of transcriptome evolution.," *PLoS Biol.*, vol. 2, no. 5, p. E132, May 2004.

[14] S. Ohno, "So much 'junk' DNA in our genome.," *Brookhaven Symp. Biol.*, vol. 23, pp. 366–70, Jan. 1972.

[15] L. E. Orgel and F. H. Crick, "Selfish DNA: the ultimate parasite.," *Nature*, vol. 284, no. 5757, pp. 604–7, Apr. 1980.

[16] P. Walter and G. Blobel, "Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum.," *Nature*, vol. 299, no. 5885, pp. 691–8, Oct. 1982.

[17] T. Mizuno, M. Y. Chou, and M. Inouye, "A unique mechanism regulating gene expression: translational inhibition by a complementary RNA transcript (micRNA).," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 81, no. 7, pp. 1966–70, Apr. 1984.

[18] R. C. Lee and V. Ambros, "An extensive class of small RNAs in Caenorhabditis elegans.," *Science*, vol. 294, no. 5543, pp. 862–4, Oct. 2001.

[19] N. C. Lau, A. G. Seto, J. Kim, S. Kuramochi-Miyagawa, T. Nakano, D. P. Bartel, and R. E. Kingston, "Characterization of the piRNA complex from rat testes.," *Science*, vol. 313, no. 5785, pp. 363–7, Jul. 2006.

[20] M. J. Hangauer, I. W. Vaughn, and M. T. McManus, "Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs.," *PLoS Genet.*, vol. 9, no. 6, p. e1003569, Jun. 2013.

[21] J. Salzman, C. Gawad, P. L. Wang, N. Lacayo, and P. O. Brown, "Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types.," *PLoS One*, vol. 7, no. 2, p. e30733, Jan. 2012.

[22] W. R. Jeck, J. a. Sorrentino, K. Wang, M. K. Slevin, C. E. Burd, J. Liu, W. F. Marzluff, and N. E. Sharpless, "Circular RNAs are abundant, conserved, and associated with ALU repeats.," *RNA*, vol. 19, no. 2, pp. 141–57, Feb. 2013.

[23] T. Aid, A. Kazantseva, M. Piirsoo, K. Palm, and T. Timmusk, "Mouse and rat BDNF gene structure and expression revisited.," *J. Neurosci. Res.*, vol. 85, no. 3, pp. 525–35, Feb. 2007.

[24] J. J. Champoux, "DNA topoisomerases: structure, function, and mechanism.," *Annu. Rev. Biochem.*, vol. 70, no. 1, pp. 369–413, Jan. 2001.

[25] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.," *Nat. Genet.*, vol. 40, no. 12, pp. 1413–5, Dec. 2008.

[26] P. Ivanov and P. Anderson, "Post-transcriptional regulatory networks in immunity.," *Immunol. Rev.*, vol. 253, no. 1, pp. 253–72, May 2013.

[27] H. Siomi and M. C. Siomi, "Posttranscriptional regulation of microRNA biogenesis in animals.," *Mol. Cell*, vol. 38, no. 3, pp. 323–32, May 2010.

[28] D. P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function.," *Cell*, vol. 116, no. 2, pp. 281–97, Jan. 2004.

[29] S. M. Elbashir, J. Harborth, W. Lendeckel, A. Yalcin, K. Weber, and T. Tuschl, "Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells.," *Nature*, vol. 411, no. 6836, pp. 494–8, May 2001.

[30] J. Brennecke, C. D. Malone, A. a Aravin, R. Sachidanandam, A. Stark, and G. J. Hannon, "An epigenetic role for maternally inherited piRNAs in transposon silencing.," *Science*, vol. 322, no. 5906, pp. 1387–92, Nov. 2008.

[31] Z. Peng, Y. Cheng, B. C.-M. Tan, L. Kang, Z. Tian, Y. Zhu, W. Zhang, Y. Liang, X. Hu, X. Tan, J. Guo, Z. Dong, Y. Liang, L. Bao, and J. Wang, "Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome.," *Nat. Biotechnol.*, vol. 30, no. 3, pp. 253–60, Mar. 2012.

[32] A. W. Craig, A. Haghighat, A. T. Yu, and N. Sonenberg, "Interaction of polyadenylate-binding protein with the eIF4G homologue PAIP enhances translation.," *Nature*, vol. 392, no. 6675, pp. 520–3, Apr. 1998.

[33] E. A. Shestakova, R. H. Singer, and J. Condeelis, "The physiological significance of beta -actin mRNA localization in determining cell polarity and directional motility.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 13, pp. 7045–50, Jun. 2001.

[34] C. Gong and L. E. Maquat, "lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements.," *Nature*, vol. 470, no. 7333, pp. 284–8, Feb. 2011.

[35] L. Poliseno, L. Salmena, J. Zhang, B. Carver, W. J. Haveman, and P. P. Pandolfi, "A coding-independent function of gene and pseudogene mRNAs regulates tumour biology.," *Nature*, vol. 465, no. 7301, pp. 1033–8, Jun. 2010.

[36] T. B. Hansen, T. I. Jensen, B. H. Clausen, J. B. Bramsen, B. Finsen, C. K. Damgaard, and J. Kjems, "Natural RNA circles function as efficient microRNA sponges.," *Nature*, vol. 495, no. 7441, pp. 384–8, Mar. 2013.

[37] S. Memczak, M. Jens, A. Elefsinioti, F. Torti, J. Krueger, A. Rybak, L. Maier, S. D. Mackowiak, L. H. Gregersen, M. Munschauer, A. Loewer, U. Ziebold, M. Landthaler, C. Kocks, F. le Noble, and N. Rajewsky, "Circular RNAs are a large class of animal RNAs with regulatory potency.," *Nature*, vol. 495, no. 7441, pp. 333–8, Mar. 2013.

[38] R. M. Zayas, A. Hernández, B. Habermann, Y. Wang, J. M. Stary, and P. A. Newmark, "The planarian Schmidtea mediterranea as a model for epigenetic germ cell specification: analysis of

ESTs from the hermaphroditic strain.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 51, pp. 18491–6, Dec. 2005.

[39] S. M. C. Robb, E. Ross, and A. Sánchez Alvarado, "SmedGD: the Schmidtea mediterranea genome database.," *Nucleic Acids Res.*, vol. 36, no. Database issue, pp. D599–606, Jan. 2008.

[40] J. C. Vera, C. W. Wheat, H. W. Fescemyer, M. J. Frilander, D. L. Crawford, I. Hanski, and J. H. Marden, "Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing.," *Mol. Ecol.*, vol. 17, no. 7, pp. 1636–47, Apr. 2008.

[41] E. Meyer, G. V Aglyamova, S. Wang, J. Buchanan-Carter, D. Abrego, J. K. Colbourne, B. L. Willis, and M. V Matz, "Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx.," *BMC Genomics*, vol. 10, p. 219, Jan. 2009.

[42] X.-W. Wang, J.-B. Luan, J.-M. Li, Y.-Y. Bao, C.-X. Zhang, and S.-S. Liu, "De novo characterization of a whitefly transcriptome and analysis of its gene expression during development.," *BMC Genomics*, vol. 11, no. 1, p. 400, Jan. 2010.

[43] P. A. Zhulidov, E. A. Bogdanova, A. S. Shcheglov, L. L. Vagner, G. L. Khaspekov, V. B. Kozhemyako, M. V Matz, E. Meleshkevitch, L. L. Moroz, S. A. Lukyanov, and D. A. Shagin, "Simple cDNA normalization using kamchatka crab duplex-specific nuclease.," *Nucleic Acids Res.*, vol. 32, no. 3, p. e37, Jan. 2004.

[44] Z. Li, Y. Chen, D. Mu, J. Yuan, Y. Shi, H. Zhang, J. Gan, N. Li, X. Hu, B. Liu, B. Yang, and W. Fan, "Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph.," *Brief. Funct. Genomics*, vol. 11, no. 1, pp. 25–37, Jan. 2012.

[45] R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, S. Li, H. Yang, J. Wang, and J. Wang, "De novo assembly of human genomes with massively parallel short read sequencing.," *Genome Res.*, vol. 20, no. 2, pp. 265–72, Feb. 2010.

[46] W. J. Kent, "BLAT--the BLAST-like alignment tool.," *Genome Res.*, vol. 12, no. 4, pp. 656–64, Apr. 2002.

[47] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool.," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–10, Oct. 1990.

[48] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq.," *Nat. Methods*, vol. 5, no. 7, pp. 621–8, Jul. 2008.

[49] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter, "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.," *Nat. Protoc.*, vol. 7, no. 3, pp. 562–78, Mar. 2012.

[50] I. Birol, S. D. Jackman, C. B. Nielsen, J. Q. Qian, R. Varhol, G. Stazyk, R. D. Morin, Y. Zhao, M. Hirst, J. E. Schein, D. E. Horsman, J. M. Connors, R. D. Gascoyne, M. a Marra, and S. J. M. Jones, "De novo transcriptome assembly with ABySS.," *Bioinformatics*, vol. 25, no. 21, pp. 2872–7, Nov. 2009.

[51] M. H. Schulz, D. R. Zerbino, M. Vingron, and E. Birney, "Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels.," *Bioinformatics*, vol. 28, no. 8, pp. 1086–92, Apr. 2012.

[52] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. a Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev, "Full-length transcriptome assembly from RNA-Seq data without a reference genome.," *Nat. Biotechnol.*, vol. 29, no. 7, pp. 644–52, Jul. 2011.

[53] M. Guttman, M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. J. Koziol, A. Gnirke, C. Nusbaum, J. L. Rinn, E. S. Lander, and A. Regev, "Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs.," *Nat. Biotechnol.*, vol. 28, no. 5, pp. 503–10, May 2010.

[54] D. R. Zerbino and E. Birney, "Velvet: algorithms for de novo short read assembly using de Bruijn graphs.," *Genome Res.*, vol. 18, no. 5, pp. 821–9, May 2008.

[55] I. Maccallum, D. Przybylski, S. Gnerre, J. Burton, I. Shlyakhter, A. Gnirke, J. Malek, K. McKernan, S. Ranade, T. P. Shea, L. Williams, S. Young, C. Nusbaum, and D. B. Jaffe, "ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads.," *Genome Biol.*, vol. 10, no. 10, p. R103, Jan. 2009.

[56] C.-S. Chin, D. H. Alexander, P. Marks, A. a Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E. E. Eichler, S. W. Turner, and J. Korlach, "Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data.," *Nat. Methods*, vol. 10, no. 6, pp. 563–9, Jun. 2013.

[57] C. Trapnell, B. a Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.," *Nat. Biotechnol.*, vol. 28, no. 5, pp. 511–5, May 2010.

[58] T. Steijger, J. F. Abril, P. G. Engström, F. Kokocinski, J. F. Abril, M. Akerman, T. Alioto, G. Ambrosini, S. E. Antonarakis, J. Behr, and P. Bertone, "Assessment of transcript reconstruction methods for RNA-seq.," *Nat. Methods*, vol. 10, no. 12, pp. 1177–84, Dec. 2013.

[59] S. Koren, M. C. Schatz, B. P. Walenz, J. Martin, J. T. Howard, G. Ganapathy, Z. Wang, D. a Rasko, W. R. McCombie, E. D. Jarvis, and Adam M Phillippy, "Hybrid error correction and de novo assembly of single-molecule sequencing reads.," *Nat. Biotechnol.*, vol. 30, no. 7, pp. 693–700, Jul. 2012.

[60] K. F. Au, J. G. Underwood, L. Lee, and W. H. Wong, "Improving PacBio long read accuracy by short read alignment.," *PLoS One*, vol. 7, no. 10, p. e46679, Jan. 2012.

[61] L. Salmela and E. Rivals, "LoRDEC: accurate and efficient long read error correction.," *Bioinformatics*, vol. 30, no. 24, pp. 3506–14, Dec. 2014.

[62] T. D. Wu and C. K. Watanabe, "GMAP: a genomic mapping and alignment program for mRNA and EST sequences.," *Bioinformatics*, vol. 21, no. 9, pp. 1859–75, May 2005.

[63] C. H. Jan, R. C. Friedman, J. G. Ruby, and D. P. Bartel, "Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs.," *Nature*, vol. 469, no. 7328, pp. 97–101, Jan. 2011.

[64] K. Yamakawa, Y. K. Huot, M. A. Haendelt, R. Hubert, X. N. Chen, G. E. Lyons, and J. R. Korenberg, "DSCAM: a novel member of the immunoglobulin superfamily maps in a Down

syndrome region and is involved in the development of the nervous system.," *Hum. Mol. Genet.*, vol. 7, no. 2, pp. 227–37, Feb. 1998.

[65] V. Cvetkovska, A. D. Hibbert, F. Emran, and B. E. Chen, "Overexpression of Down syndrome cell adhesion molecule impairs precise synaptic targeting.," *Nat. Neurosci.*, vol. 16, no. 6, pp. 677–82, Jun. 2013.

[66] H.-H. Yu, J. S. Yang, J. Wang, Y. Huang, and T. Lee, "Endodomain diversity in the Drosophila Dscam and its roles in neuronal morphogenesis.," *J. Neurosci.*, vol. 29, no. 6, pp. 1904–14, Feb. 2009.

[67] S. L. Zipursky and J. R. Sanes, "Chemoaffinity revisited: dscams, protocadherins, and neural circuit assembly.," *Cell*, vol. 143, no. 3, pp. 343–53, Oct. 2010.

[68] D. Schmucker, J. C. Clemens, H. Shu, C. A. Worby, J. Xiao, M. Muda, J. E. Dixon, and S. L. Zipursky, "Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity.," *Cell*, vol. 101, no. 6, pp. 671–84, Jun. 2000.

[69] A. M. Celotto and B. R. Graveley, "Alternative splicing of the Drosophila Dscam pre-mRNA is both temporally and spatially regulated.," *Genetics*, vol. 159, no. 2, pp. 599–608, Oct. 2001.

[70] G. Neves, J. Zucker, M. Daly, and A. Chess, "Stochastic yet biased expression of multiple Dscam splice variants by individual cells.," *Nat. Genet.*, vol. 36, no. 3, pp. 240–6, Mar. 2004.

[71] Y. Dong, H. E. Taylor, and G. Dimopoulos, "AgDscam, a hypervariable immunoglobulin domain-containing receptor of the Anopheles gambiae innate immune system.," *PLoS Biol.*, vol. 4, no. 7, p. e229, Jul. 2006.

[72] F. L. Watson, R. Püttmann-Holgado, F. Thomas, D. L. Lamar, M. Hughes, M. Kondo, V. I. Rebel, and D. Schmucker, "Extensive diversity of Ig-superfamily proteins in the immune system of insects.," *Science*, vol. 309, no. 5742, pp. 1874–8, Sep. 2005.

[73] I. J. Cajigas, G. Tushev, T. J. Will, S. tom Dieck, N. Fuerst, and E. M. Schuman, "The local transcriptome in the synaptic neuropil revealed by deep sequencing and high-resolution imaging.," *Neuron*, vol. 74, no. 3, pp. 453–66, May 2012.

[74] H. Kang and E. M. Schuman, "A requirement for local protein synthesis in neurotrophin-induced hippocampal synaptic plasticity.," *Science*, vol. 273, no. 5280, pp. 1402–6, Sep. 1996.

[75] G. Rudenko, T. Nguyen, Y. Chelliah, T. C. Südhof, and J. Deisenhofer, "The structure of the ligand-binding domain of neurexin Ibeta: regulation of LNS domain function by alternative splicing.," *Cell*, vol. 99, no. 1, pp. 93–101, Oct. 1999.

[76] C. Mayr and D. P. Bartel, "Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells.," *Cell*, vol. 138, no. 4, pp. 673–84, Aug. 2009.

[77] A. I. Su, M. P. Cooke, K. A. Ching, Y. Hakak, J. R. Walker, T. Wiltshire, A. P. Orth, R. G. Vega, L. M. Sapinoso, A. Moqrich, A. Patapoutian, G. M. Hampton, P. G. Schultz, and J. B. Hogenesch, "Large-scale analysis of the human and mouse transcriptomes.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 7, pp. 4465–70, Apr. 2002.

[78] P. Carninci, A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C. a M. Semple, M. S. Taylor, P. G. Engström, M. C. Frith, A. R. R. Forrest, W. B. Alkema, S. L. Tan,

C. Plessy, R. Kodzius, T. Ravasi, T. Kasukawa, S. Fukuda, M. Kanamori-Katayama, Y. Kitazume, H. Kawaji, C. Kai, M. Nakamura, H. Konno, K. Nakano, S. Mottagui-Tabar, P. Arner, A. Chesi, S. Gustincich, F. Persichetti, H. Suzuki, S. M. Grimmond, C. a Wells, V. Orlando, C. Wahlestedt, E. T. Liu, M. Harbers, J. Kawai, V. B. Bajic, D. a Hume, and Y. Hayashizaki, "Genome-wide analysis of mammalian promoter architecture and evolution.," *Nat. Genet.*, vol. 38, no. 6, pp. 626–35, Jun. 2006.

[79] Y. Wan, K. Qu, Q. C. Zhang, R. A. Flynn, O. Manor, Z. Ouyang, J. Zhang, R. C. Spitale, M. P. Snyder, E. Segal, and H. Y. Chang, "Landscape and variation of RNA secondary structure across the human transcriptome.," *Nature*, vol. 505, no. 7485, pp. 706–9, Jan. 2014.

[80] Z. Fang and N. Rajewsky, "The impact of miRNA target sites in coding sequences and in 3'UTRs.," *PLoS One*, vol. 6, no. 3, p. e18067, Jan. 2011.

[81] D. Feng and J. Xie, "Aberrant splicing in neurological diseases.," *Wiley Interdiscip. Rev. RNA*, vol. 4, no. 6, pp. 631–49, 2013.

[82] B. P. Lewis, C. B. Burge, and D. P. Bartel, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.," *Cell*, vol. 120, no. 1, pp. 15–20, Jan. 2005.

[83] A. Kozomara and S. Griffiths-Jones, "miRBase: integrating microRNA annotation and deep-sequencing data.," *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D152–7, Jan. 2011.

[84] R. Li, Y. Li, K. Kristiansen, and J. Wang, "SOAP: short oligonucleotide alignment program.," *Bioinformatics*, vol. 24, no. 5, pp. 713–4, Mar. 2008.

[85] R. Li, C. Yu, Y. Li, T.-W. Lam, S.-M. Yiu, K. Kristiansen, and J. Wang, "SOAP2: an improved ultrafast tool for short read alignment.," *Bioinformatics*, vol. 25, no. 15, pp. 1966–7, Aug. 2009.

[86] J. Krol, I. Loedige, and W. Filipowicz, "The widespread regulation of microRNA biogenesis, function and decay.," *Nat. Rev. Genet.*, vol. 11, no. 9, pp. 597–610, Sep. 2010.

[87] I. L. Hofacker and P. F. Stadler, "Memory efficient folding algorithms for circular RNA secondary structures.," *Bioinformatics*, vol. 22, no. 10, pp. 1172–6, May 2006.

[88] E. Bonnet, J. Wuyts, P. Rouzé, and Y. Van de Peer, "Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences.," *Bioinformatics*, vol. 20, no. 17, pp. 2911–7, Nov. 2004.

[89] V. N. Kim, J. Han, and M. C. Siomi, "Biogenesis of small RNAs in animals.," *Nat. Rev. Mol. Cell Biol.*, vol. 10, no. 2, pp. 126–39, Feb. 2009.

[90] A. S. Flynt, J. C. Greimann, W.-J. Chung, C. D. Lima, and E. C. Lai, "MicroRNA biogenesis via splicing and exosome-mediated trimming in Drosophila.," *Mol. Cell*, vol. 38, no. 6, pp. 900–7, Jun. 2010.

[91] M. R. Friedländer, W. Chen, C. Adamidi, J. Maaskola, R. Einspanier, S. Knespel, and N. Rajewsky, "Discovering microRNAs from deep sequencing data using miRDeep.," *Nat. Biotechnol.*, vol. 26, no. 4, pp. 407–15, Apr. 2008.

[92] M. Hackenberg, N. Rodríguez-Ezpeleta, and A. M. Aransay, "miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments.," *Nucleic Acids Res.*, vol. 39, no. Web Server issue, pp. W132–8, Jul. 2011.

[93] D. Hendrix, M. Levine, and W. Shi, "miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data.," *Genome Biol.*, vol. 11, no. 4, p. R39, Jan. 2010.

[94] D. Gerlach, E. V Kriventseva, N. Rahman, C. E. Vejnar, and E. M. Zdobnov, "miROrtho: computational survey of microRNA genes.," *Nucleic Acids Res.*, vol. 37, no. Database issue, pp. D111–7, Jan. 2009.

[95] Y. Kawahara, B. Zinshteyn, P. Sethupathy, H. Iizasa, A. G. Hatzigeorgiou, and K. Nishikura, "Redirection of silencing targets by adenosine-to-inosine editing of miRNAs.," *Science*, vol. 315, no. 5815, pp. 1137–40, Feb. 2007.

[96] a M. Burroughs, M. Kawano, Y. Ando, C. O. Daub, and Y. Hayashizaki, "pre-miRNA profiles obtained through application of locked nucleic acids and deep sequencing reveals complex 5'/3' arm variation including concomitant cleavage and polyuridylation patterns.," *Nucleic Acids Res.*, vol. 40, no. 4, pp. 1424–37, Feb. 2012.

[97] M. a Newman, V. Mani, and S. M. Hammond, "Deep sequencing of microRNA precursors reveals extensive 3' end modification.," *RNA*, vol. 17, no. 10, pp. 1795–803, Oct. 2011.

[98] C. Cocquerelle, B. Mascrez, D. Hétuin, and B. Bailleul, "Mis-splicing yields circular RNA molecules.," *FASEB J.*, vol. 7, no. 1, pp. 155–60, Jan. 1993.

[99] B. Capel, A. Swain, S. Nicolis, A. Hacker, M. Walter, P. Koopman, P. Goodfellow, and R. Lovell-Badge, "Circular transcripts of the testis-determining gene Sry in adult mouse testis.," *Cell*, vol. 73, no. 5, pp. 1019–30, Jun. 1993.

[100] P. L. Wang, Y. Bao, M.-C. Yee, S. P. Barrett, G. J. Hogan, M. N. Olsen, J. R. Dinneny, P. O. Brown, and J. Salzman, "Circular RNA is expressed across the eukaryotic tree of life.," *PLoS One*, vol. 9, no. 6, p. e90859, Jan. 2014.

[101] M. W. Hentze and T. Preiss, "Circular RNAs: splicing's enigma variations.," *EMBO J.*, vol. 32, no. 7, pp. 923–5, Apr. 2013.

[102] W. R. Jeck and N. E. Sharpless, "Detecting and characterizing circular RNAs.," *Nat. Biotechnol.*, vol. 32, no. 5, pp. 453–61, May 2014.

[103] Z. Li, C. Huang, C. Bao, L. Chen, M. Lin, X. Wang, G. Zhong, B. Yu, W. Hu, L. Dai, P. Zhu, Z. Chang, Q. Wu, Y. Zhao, Y. Jia, P. Xu, H. Liu, and G. Shan, "Exon-intron circular RNAs regulate transcription in the nucleus.," *Nat. Struct. Mol. Biol.*, vol. 22, no. 3, pp. 256–64, Mar. 2015.

[104] C. Hanus and E. M. Schuman, "Proteostasis in complex dendrites.," *Nat. Rev. Neurosci.*, vol. 14, no. 9, pp. 638–48, Sep. 2013.

[105] K. H. Zivraj, Y. C. L. Tung, M. Piper, L. Gumy, J. W. Fawcett, G. S. H. Yeo, and C. E. Holt, "Subcellular profiling reveals distinct and developmentally regulated repertoire of growth cone mRNAs.," *J. Neurosci.*, vol. 30, no. 46, pp. 15464–78, Nov. 2010.

[106] R. B. Darnell, "RNA protein interaction in neurons.," *Annu. Rev. Neurosci.*, vol. 36, pp. 243–70, Jul. 2013.

[107] E. Huntzinger and E. Izaurralde, "Gene silencing by microRNAs: contributions of translational repression and mRNA decay.," *Nat. Rev. Genet.*, vol. 12, no. 2, pp. 99–110, Feb. 2011.

[108] J. L. Rinn and H. Y. Chang, "Genome regulation by long noncoding RNAs.," *Annu. Rev. Biochem.*, vol. 81, pp. 145–66, Jan. 2012.

[109] B. Li and C. N. Dewey, "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.," *BMC Bioinformatics*, vol. 12, p. 323, Jan. 2011.

[110] H. Li and R. Durbin, "Fast and accurate long-read alignment with Burrows-Wheeler transform.," *Bioinformatics*, vol. 26, no. 5, pp. 589–95, Mar. 2010.

[111] S. D. Wagner, P. Yakovchuk, B. Gilman, S. L. Ponicsan, L. F. Drullinger, J. F. Kugel, and J. a Goodrich, "RNA polymerase II acts as an RNA-dependent RNA polymerase to extend and destabilize a non-coding RNA.," *EMBO J.*, vol. 32, no. 6, pp. 781–90, Mar. 2013.

[112] G. Yeo and C. B. Burge, "Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals.," *J. Comput. Biol.*, vol. 11, no. 2–3, pp. 377–94, Jan. 2004.

[113] R. C. Friedman, K. K. Farh, C. B. Burge, and D. P. Bartel, "Most mammalian mRNAs are conserved targets of microRNAs.," *Genome Res.*, vol. 19, no. 1, pp. 92–105, Jan. 2009.

[114] K. B. Cook, H. Kazan, K. Zuberi, Q. Morris, and T. R. Hughes, "RBPDB: a database of RNA-binding specificities.," *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D301–8, Jan. 2011.

[115] N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, and J. S. Weissman, "Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling.," *Science*, vol. 324, no. 5924, pp. 218–23, Apr. 2009.

[116] J. O. Westholm, P. Miura, S. Olson, S. Shenker, B. Joseph, P. Sanfilippo, S. E. Celniker, B. R. Graveley, and E. C. Lai, "Genome-wide analysis of drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation.," *Cell Rep.*, vol. 9, no. 5, pp. 1966–80, Dec. 2014.

[117] P. R. Dunkley, P. E. Jarvie, and P. J. Robinson, "A rapid Percoll gradient procedure for preparation of synaptosomes.," *Nat. Protoc.*, vol. 3, no. 11, pp. 1718–28, Jan. 2008.

[118] L. R. Squire, "Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans.," *Psychol. Rev.*, vol. 99, no. 2, pp. 195–231, Apr. 1992.

[119] J. U. Guo, V. Agarwal, H. Guo, and D. P. Bartel, "Expanded identification and characterization of mammalian circular RNAs.," *Genome Biol.*, vol. 15, no. 7, p. 409, Jan. 2014.

[120] M. Guttman and J. L. Rinn, "Modular regulatory principles of large non-coding RNAs.," *Nature*, vol. 482, no. 7385, pp. 339–46, Feb. 2012.

[121] S.-Y. Ng, G. K. Bogu, B. S. Soh, and L. W. Stanton, "The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis.," *Mol. Cell*, vol. 51, no. 3, pp. 349–59, Aug. 2013.

[122] G. G. Turrigiano, K. R. Leslie, N. S. Desai, L. C. Rutherford, and S. B. Nelson, "Activity-dependent scaling of quantal amplitude in neocortical neurons.," *Nature*, vol. 391, no. 6670, pp. 892–6, Feb. 1998.

[123] T. R. Mercer, I. A. Qureshi, S. Gokhan, M. E. Dinger, G. Li, J. S. Mattick, and M. F. Mehler, "Long noncoding RNAs in neuronal-glial fate specification and oligodendrocyte lineage maturation.," *BMC Neurosci.*, vol. 11, p. 14, Jan. 2010.

[124] M. Guttman, J. Donaghey, B. W. Carey, M. Garber, J. K. Grenier, G. Munson, G. Young, A. B. Lucas, R. Ach, L. Bruhn, X. Yang, I. Amit, A. Meissner, A. Regev, J. L. Rinn, D. E. Root, and E. S. Lander, "lincRNAs act in the circuitry controlling pluripotency and differentiation.," *Nature*, vol. 477, no. 7364, pp. 295–300, Sep. 2011.

[125] K. A. Newell and N. Matosin, "Rethinking metabotropic glutamate receptor 5 pathological findings in psychiatric disorders: implications for the future of novel therapeutics.," *BMC Psychiatry*, vol. 14, p. 23, Jan. 2014.

[126] D. Bottai, J. F. Guzowski, M. K. Schwarz, S. H. Kang, B. Xiao, A. Lanahan, P. F. Worley, and P. H. Seeburg, "Synaptic activity-induced conversion of intronic to exonic sequence in Homer 1 immediate early gene expression.," *J. Neurosci.*, vol. 22, no. 1, pp. 167–75, Jan. 2002.

[127] S. A. Tomlins, D. R. Rhodes, S. Perner, S. M. Dhanasekaran, R. Mehra, X.-W. Sun, S. Varambally, X. Cao, J. Tchinda, R. Kuefer, C. Lee, J. E. Montie, R. B. Shah, K. J. Pienta, M. A. Rubin, and A. M. Chinnaiyan, "Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer.," *Science*, vol. 310, no. 5748, pp. 644–8, Oct. 2005.

[128] M. Soda, Y. L. Choi, M. Enomoto, S. Takada, Y. Yamashita, S. Ishikawa, S. Fujiwara, H. Watanabe, K. Kurashina, H. Hatanaka, M. Bando, S. Ohno, Y. Ishikawa, H. Aburatani, T. Niki, Y. Sohara, Y. Sugiyama, and H. Mano, "Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer.," *Nature*, vol. 448, no. 7153, pp. 561–6, Aug. 2007.

[129] Y. W. Asmann, B. M. Necela, K. R. Kalari, A. Hossain, T. R. Baker, J. M. Carr, C. Davis, J. E. Getz, G. Hostetter, X. Li, S. A. McLaughlin, D. C. Radisky, G. P. Schroth, H. E. Cunliffe, E. A. Perez, and E. A. Thompson, "Detection of redundant fusion transcripts as biomarkers or disease-specific therapeutic targets in breast cancer.," *Cancer Res.*, vol. 72, no. 8, pp. 1921–8, Apr. 2012.

[130] F. Mertens, B. Johansson, T. Fioretos, and F. Mitelman, "The emerging complexity of gene fusions in cancer.," *Nat. Rev. Cancer*, vol. 15, no. 6, pp. 371–81, Jun. 2015.

List of Figures

List of Tables

Appendix A: Curriculum Vitae

For reasons of data protection, the curriculum vitae is not published in the electronic version.

For reasons of data protection, the curriculum vitae is not published in the electronic version.

Appendix B: Zusammenfassung

Eine genaue Kenntnis des Transkriptoms ist von entscheidender Bedeutung im Bereich der Molekularbiologie und gewinnt Bedeutung bei der Diagnose von Krankheiten und deren Behandlung. Drei entscheidende Aspekte des Transkriptoms tragen zu dessen vielschichtiger Bedeutung bei. Zunächst definiert das Profil aller Isoformen der Protein-kodierenden Gene den Entwicklungspfad der Zellen und Organismen. Zweitens moduliert das Profil der regulatorischen Elemente die Aktivität der Protein-kodierenden Gene. Drittens prägt das Zusammenspiel der Protein-kodierenden Gene und regulatorischen Elemente die Dynamik des Transkriptoms. Die Identifizierung der einzelnen Bestandteile des regulatorischen Netzwerks ist der erste Schritt im Bereich des Reverse Engineering in der Molekularbiologie. In der vorliegenden Arbeit beschreibe ich vier Analysemethoden für Anwendungen, die sich mit den ersten beiden Aspekten beschäftigen.

Als Erstes wurde eine Software-Pipeline entwickelt, die ohne Referenzgenom ein Assembly zur Identifizierung des Transkriptoms durchführt. Ein qualitativ hochwertiges Transkriptom konnte erstellt werden, indem zwei sich ergänzende Sequenziertechnologien und zusätzlich eine effiziente cDNA-Normalisierung kombiniert wurden. Die vorgestellte Pipeline übertrifft bestehende Programme, die nur auf eine einzige Art von Eingabedaten setzen. Darüber hinaus wurden die Ergebnisse experimentell bestätigt.

Als Zweites wurden Analysemethoden erarbeitet, um vollständige Transkripte zu charakterisieren. Es wurden Werkzeuge für die Auswertung von Daten aus Sequenziertechnologien, die lange Reads liefern, entwickelt, mit denen das Transkriptom genauer untersucht werden kann. Des weiteren konnte damit das Zusammenspiel verschiedener Schritte der RNA-Prozessierung experimentell untersucht werden. Eine Untersuchung des Transkripts des Gens Dscam in der Fruchtfliege bestätigte die Hypothese des unabhängigen Spleißens, wodurch eine Neuauswertung früherer Experimente notwendig wird. Die Anwendung der Methode auf Sequenzierdaten des Rattengehirns verbesserte deutlich die Annotation des Transkriptoms. Dies ist von großer Bedeutung für die Neurobiologie, in der die Ratte als Modellorganismus eingesetzt wird.

Als drittes wurde ein de-novo-miRNA-Vorhersagewerkzeug implementiert. Durch die Entwicklung von Sequenzierexperimenten, welche eine Momentaufnahme der miRNA-Entstehung liefern, können nicht nur prozessierte und Vorläufer-miRNAs identifiziert werden, sondern auch Details der miRNA-Prozessierung und -Modifikation beobachtet werden. Erste Experimente in Modellorganismen wie der Maus und *C.elegans* zeigten die Effizienz und das Anwendungspotential der Methode.

Schließlich ist eine Pipeline zur Charakterisierung von zirkulärer RNA entwickelt worden. Durch die Untersuchung von potentiellen Spleißstellen basierend auf lokalen Alignments können zirkuläre RNAs aus ansonsten nicht berücksichtigten RNA-Sequenzdaten identifiziert werden. Zehntausende zirkuläre RNAs in Maus, Ratte und Fruchtfliege konnten identifiziert und quantifiziert werden. Weitere Experimente zeigen, dass zirkuläre RNAs in Gehirnsynapsen angereichert sind und bei der Entwicklung des Gehirns und neuronalen homöostatischen Plastizität beteiligt sind.

Zusammenfassend beschreibt diese Arbeit vier Analysemethoden für verschiedene Aspekte des Transkriptoms. Die vorgestellten Methoden tragen gemeinsam zu einem ganzheitlichen Verständnis der Molekularbiologie und Medizin bei.