UNIVERSITY OF CALIFORNIA
RIVERSIDE

Exploring Non-Canonical Regulatory Small RNAs in Mammals

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Genetics, Genomics and Bioinformatics

by

Junchao Shi

June 2021

Dissertation Committee:
Dr. Qi Chen, Chairperson
Dr. Xuemei Chen
Dr. Weifeng Gu
Dr. Tong Zhou

The Dissertation of Junchao Shi is approved:

Committee Chairperson

University of California, Riverside

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my mentor Qi Chen. I feel extremely lucky to have him to work with at the start point of my academic journey. He is not only an extraordinarily insightful, supportive, knowledgeable advisor, but also a valuable friend who always inspires me and helps me to lever up. It is my privilege to be his first Ph.D. student and I will take it as my lifetime honor.

I am extremely thankful to my co-mentor Tong Zhou. His prompt inspiration and timely advice is greatly constructive to shape my bioinformatics skills. I also cherish those relaxed and enjoyable moments shared with him in bars after work.

It is my pleasure to acknowledge the valuable input of Xuemei Chen and Weifeng Gu, who have served on my dissertation committee. I very much appreciate their time and support. I would also like to thank Xuemei Chen again together with Shouwei Ding, Tin Nguyen, Lei Yang, Monica Nicolescu and Mihye Ahn for their finest exceptional teaching which well-polished my knowledge and skills to conquer the difficulties encountered.

I want to express the depth of my gratitude to Magdalena Zernicka-Goetz, Ying Zhang, Sihem Cheloufi, Jernej Murn, Changcheng Zhou, Xiudeng Zheng and all other collaborators with their critical, constructive discussions and irreplaceable contributions to my works. I am also eternally grateful to Enkui Duan and Yi Tao for their philosophical view of life and science.

My warmest thanks also go to all past and present lab members. I can achieve nothing without the support of them. To my labmates Yunfang, Dongmei, Shichao, Menghong,

and Ying, from whom I have learned and gained so much. To my roommate Xudong, for the lab and leisure time we spent.

I also want to extend my sincere thanks to Xiaoxia, Rebacca, and Jing for their time, comments and efforts put into proofreading my thesis thoroughly.

My most special thanks go to Songjia, who provided unwavering support, belief and encouragement at every stage of my research project, who contributed fruitful amendments to my dissertation.

Words are powerless to express my gratitude to my family. Without their incredible backing and empathy, it would be unimaginable for me to complete my study.

I am indebted to all the scientists who laid the foundation and pave the way for the research field. All my works are standing on the shoulders of their pioneer accomplishments.

The text of this dissertation, in part, is a reprint of the material as it appears in "SPORTS1. 0: a tool for annotating and profiling non-coding RNAs optimized for rRNA- and tRNA-derived small RNAs" (2018); "Peripheral blood non-canonical small noncoding RNAs as novel biomarkers in lung cancer" (2020), "PANDORA-seq expands the repertoire of regulatory small RNAs by overcoming RNA modifications" and "Origins and evolving functionalities of tRNA-derived small RNAs" (2021). The coauthor Qi Chen listed in those publications directed and supervised the research which form the basis for this dissertation.

Wise lives, civilize ages.

ABSTRACT OF THE DISSERTATION

Exploring Non-Canonical Regulatory Small RNAs in Mammals

by

Junchao Shi

Doctor of Philosophy, Graduate Program in Genetics, Genomics and Bioinformatics
University of California, Riverside, June 2021
Dr. Qi Chen, Chairperson

Small RNAs are short, non-coding RNA molecules that have been identified in a wide range of species across all three domains of life. In mammals, canonical small RNAs such as microRNAs (miRNAs) and Piwi-interacting RNAs (piRNAs) are tissue-specifically distributed and can regulate gene expression at both transcriptional and post-transcriptional levels, associating with various fundamental functions such as gene silencing and retrotransposon control. While miRNAs and piRNAs have been extensively investigated, the existence and mechanism of other non-canonical mammalian small RNAs remain underexplored.

With the extensive use of high-throughput sequencing technologies in the past decade, small RNA diversity is rapidly growing. However, the conventional small RNA library construction method lacks the detection capability for non-canonical small RNAs that carry specific terminal and internal modifications, especially for tRNA-derived small RNAs

(tsRNAs) and rRNA-derived small RNAs (rsRNAs). In addition, existing downstream bioinformatics tools are mostly focused on analyzing canonical small RNAs such as miRNAs and piRNAs, while the annotation of other small RNAs remains rudimentary or ignored.

To reveal a panoramic view of small RNAs, a small RNA annotation pipeline (that is, SPORTS) is developed to simultaneously and comparatively annotate both canonical and non-canonical small RNAs, along with RNA modification prediction capacity. Moreover, a small RNA library preparation procedure (that is, PANDORA-seq) is optimized to comprehensively capture modified small RNAs such as tsRNAs and rsRNAs.

The improved RNA-seq and bioinformatics strategy leads to a new and surprising small RNA landscape that tsRNAs and rsRNAs have a higher abundance than canonical small RNAs in a majority of mouse and human tissues/cells that have been examined. Those mammalian tsRNAs and rsRNAs also exhibit tissue- and cell-specific patterns and the expression level of those small RNAs are dynamically altered during the generation of induced pluripotent stem cells (iPSCs) based on PANDORA-seq. Those newly identified small RNAs also display translational regulation during embryonic stem cell differentiation and have a role in regulating embryonic stem cell lineage fate based on the transcriptomic changes after their transfection.

Table of Contents

Table of Figures

Chapter 1: Introduction

Overview

Small RNAs represent a major family of noncoding RNAs that are universally distributed from bacteria to mammals. However, the concept of 'small RNA' is relatively subjective, only vaguely defined as RNAs that have relatively shorter RNA lengths when compared to transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), and long coding/non-coding RNAs. Short RNAs in bacteria have also been recognized as small RNAs, although they are usually irrelevant to small RNAs in eukaryotes (Kim et al., 2009). The bacterial small RNAs typically range from 50 to 400 nucleotides and are involved in gene expression regulation (Wagner and Romby, 2015). In eukaryotes, especially in mammals, small RNAs are generally referred to some specific types of noncoding RNAs, such as microRNAs (miRNAs, 21-23 nucleotides), and Piwi-interacting RNAs (piRNAs, 21-35 nucleotides) (Bartel, 2018; Ozata et al., 2019). While those canonical small RNAs have already been extensively investigated from various aspects, this thesis focuses on discovering and describing non-canonical small RNAs of 15-50 nucleotides in length. Those small RNAs include tRNA-derived small RNAs (tsRNAs), rRNA-derived small RNAs (rsRNAs), and YRNA-derived small RNAs (ysRNAs) that also show diverse functions, and with great potential as disease biomarkers.

Multiple general small RNA annotation software and pipelines have been developed to analyze miRNAs, piRNAs, *etc*. (Di Bella et al., 2020). However, there is still a lack of specialized tools that can simultaneously and comparatively analyze both canonical and non-canonical small RNAs. In this thesis, a small RNA annotation bioinformatics tool

(small RNA annotation pipeline optimized for rRNA- and tRNA-derived small RNAs, SPORTS) is developed to provide optimized annotation and enhanced visualization, which aims to bridge the gap between raw RNA-seq data and researchers to explore the small RNA world.

Traditional construction of cDNA libraries for small RNA-seq is based on adapter ligation to the 3' and 5' RNA terminals, which is followed by reverse transcription and cDNA amplification. However, some terminal and/or internal RNA modifications in small RNAs can affect the library construction process, thus these small RNAs cannot be efficiently captured during the cDNA amplification step for RNA-seq, generating biased final sequencing results. To overcome the RNA modification causing obstacles during cDNA library construction, a method based on consecutive enzymatic-treatments (panoramic RNA display by overcoming RNA modification aborted sequencing, PANDORA-seq) is developed in this thesis.

Canonical regulatory small RNAs

In general, both miRNAs and piRNAs perform silencing or destruction function through binding Argonaute proteins (for example, Ago subfamily and Piwi subfamily) to target RNA molecules (Hock and Meister, 2008). Such a process is defined as a canonical regulation process in this thesis.

The first miRNA, *lin-4*, was discovered in 1993 in nematode worms (Lee et al., 1993; Wightman et al., 1993). But it did not receive much attention until *let-7* was discovered (Reinhart et al., 2000). Unlike *lin-4*, the miRNA *let-7* is evolutionarily conserved in

metazoan (Pasquinelli et al., 2000). The discovery of *let-7* swiftly boosts large scale miRNA discovery (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001).

Canonical miRNAs are transcribed by RNA Polymerase II (Pol II) as primary miRNAs (pri-miRNAs) with 5' capped, sometimes 3' polyadenylated. Those pri-miRNAs can be hundreds to thousands of nucleotides in length. Cleaved from the pri-miRNA by Drosha, a nuclear RNase III, the precursor miRNA (pre-miRNA, ~70 nucleotides) forms a stem-loop structure. It is further cleaved by another RNase III, Dicer, to generate an ~21nt miRNA duplex (miRNA/miRNA* duplex). Then the single-stranded mature miRNA preferentially assembles into the RNA-Induced Silencing Complex (RISC), which includes an AGO protein, while its passenger strand, miRNA*, degrades. Based on the characteristics of dsRNA cleavage function of RNase III, mature miRNA has 5' phosphate (5'-P) and 3' hydroxy (3'-OH) termini (Bartel, 2018; Du and Zamore, 2005). miRNAs typically recognize and target transcripts by consecutively base-pairing 6-8 nucleotides with mRNA 3' UTR region (Bartel, 2018), which indicates in vivo cross-linking ligation and high-throughput sequencing of hybrids as the best practice to accurately quantify miRNA interactome (Helwak and Tollervey, 2014).

piRNAs that transcribed from the tandem repeat regions were first identified in drosophila testis (Aravin et al., 2001). In 2006, those germline-specific small RNAs were termed piRNAs since they are specifically interacting with PIWI protein (Aravin et al., 2006; Girard et al., 2006; Grivna et al., 2006; Lau et al., 2006). The piRNA sequences are usually distributed in clusters from intergenic loci where are enriched with transposon

fragments, 3′ untranslated regions (3′ UTRs) of messenger RNAs (mRNAs), and long non-coding RNAs (Ozata et al., 2019).

The piRNA clusters can be classified into two forms: uni-strand piRNA clusters, which give rise to piRNA from only one genomic strand; and dual-strand piRNA clusters that generate piRNA precursors from both genomic strands (Czech and Hannon, 2016). While the dual-strand piRNA cluster form has been observed only in arthropods, the uni-strand cluster form has been widely identified in metazoans, including mammals (Ozata et al., 2019). piRNA precursors that have 5’ 7-methylguanosine and 3’ Polyadenylation are generated though Pol II-dependent transcription in most animals. Then the precursors are endonucleolytically cleaved into pre-piRNAs with 5’ monophosphate, which is needed for PIWI protein binding. Pre-piRNAs often begin with a uridine probably due to the intrinsic preference of PIWI protein (Gainetdinov et al., 2018). The 3′ end of the PIWI-bound pre-piRNAs are trimmed by endonuclease and/or 3’ to 5’ exonuclease until they shorten to a certain length (21-35 nucleotides). Concomitantly, the 3’ ends of piRNA intermediates are 2’-O-methylated by an S-adenosylmethionine (SAM)-dependent methyltransferase to form mature piRNAs (Czech and Hannon, 2016; Kirino and Mourelatos, 2007). The piRNAs are generated by the primary pathway, or amplified by ping-pong cycle, which produces secondary piRNAs though pre-existing piRNAs (Beyret et al., 2012). While piRNAs are found in a great variety of species and are of high abundance in mammalian testis, their sequence conservation is modest compared with miRNAs. Thus, the paradigm of piRNA function, especially for pachytene piRNAs, has yet to be determined. In addition to

crosslinking methods, piRNA binding site affinity as well as its molecule number should be taken into account when piRNA targets are being identified (Wu and Zamore, 2021).

Non-canonical regulatory small RNAs

The tRNAs are a type of highly modified and structured RNA that have a well-defined role in mRNA translation. The fragmentation of hundred types of tRNAs at different loci gives birth to a new species of small RNAs: tsRNAs (Schimmel, 2018).

Since the 1960s it has been reported that a tRNA fragment from $tRNA^{fMet}$ can interact with ribosomal subunits in a fashion similar to that of its precursor mature $tRNA^{fMet}$ (Rudland and Dube, 1969); this suggests that tsRNAs may naturally compete with the function of tRNAs under some circumstances. While notable amount of RNAs with the integral tRNA corresponding length (76-90 nucleotides) exist in human urine (Perez-Hernandez et al., 2015), later the RNA sequencing results confirm that tsRNAs are dominant at the small RNA range (El-Mogy et al., 2018; Yeri et al., 2017). In the 2000s, the phenomenon of tRNA cleavage induced by environmental stress was firstly described in protozoa (Lee and Collins, 2005), then in human cell lines (Thompson et al., 2008). Those tsRNAs were formerly regarded as tRNA random degradation intermediates until they were identified under physiological condition in mammalian sperm and serum by RNA-seq (Dhahbi et al., 2013; Peng et al., 2012; Zhang et al., 2014). The mammalian tsRNAs are derived from preferential cleavage sites of tRNAs, and have different length distribution compared with other canonical small RNAs such as miRNAs and piRNAs.

From a structural perspective, the cloverleaf-shaped secondary structure of tRNA is folded into an L-shape in Three-dimensional space (Figure 1.1) (Schimmel, 2018). This L-shaped structure is overall tightly condensed but has two relatively exposed sites: the anticodon at one end of the L and the tRNA elbow at the bending site of the L, where the D-loop and the T-loop meet and interact with each other. The exposed sites of the tRNA structure can be 'points of attack' under a dynamic cellular (and perhaps early proto-cell) environment, being fragmented by one of the following: either nonspecific stress signals such as radiation and oxygen reactive species (ROS), specific recognition by enzymes or ribozymes, or the combination of both. This simple view actually coincides with the prevailing observations that the most abundantly detected tsRNAs are fragmented at the anticodon, and are derived from 5' half of the tRNA (~30 nt), whereas shorter 3' or 5' shorter tsRNAs (~18-22 nt) fragmented at the T-loop or D-loop, respectively, or internal tsRNAs derived from sequences between these loops, are far less abundant (Figure 1.1) (Kumar et al., 2015). Notably, tRNA cleavage can also occur independently of the loop site, such as by targeting a specific tRNA stem position by RNase P that recognizes a specific (that is, GC enriched) sequence (Kikuchi et al., 1990) or, more generally, by enzymes targeting double-stranded (ds) RNA regions. For example, although it is well-known that the Rnase Dicer cleaves dsRNAs to generate siRNA and miRNAs, Dicer is also responsible for the biogenesis of some tsRNAs from tRNAs (Cole et al., 2009; Haussecker et al., 2010; Reinsborough et al., 2019). Other Rnase including RNase T2 (Andersen and Collins, 2012; Thompson and Parker, 2009), RNase L (Donovan et al., 2017), and the vertebrate-specific angiogenin (RNase A family) (Fu et al., 2009; Yamasaki et al., 2009), can cleave tRNAs

at the anticodon loop, resulting in fragmentation into tsRNAs. The cleavage produces RNA fragments with 5'-hydroxyl (5'-OH) and 2',3'-cyclic phosphate (2,3'-CP) termini (Donovan et al., 2017; Luhtala and Parker, 2010; Lyons et al., 2017), providing a unique feature of tsRNAs compared to canonical small RNAs.

In addition to the termini, both tsRNAs and their precursor tRNAs are heavily modified. It has been demonstrated that DNMT2- and NSUN2- dependent addition of a 5-methylcytosine(m^5C) modification to several tRNAs (for example, $tRNA^{Asp}$, $tRNA^{Val}$, $tRNA^{Gly}$ and $tRNA^{Leu}$) increases tRNA stability in flies and mice, whereas deletion of Dnmt2 and/or Nsun2 abolishes m^5C on those tRNAs, making them likely to be cleaved into tsRNAs under stress conditions (Schaefer et al., 2010; Tuorto et al., 2012; Zhang et al., 2018b). The Queuosine (Q) modification catalyzed by QTRT1 occurs at the wobble anticodon position of several tRNAs (for example, $tRNA^{His}$, $tRNA^{Asn}$, $tRNA^{Tyr}$, and $tRNA^{Asp}$) and protects tRNAs against cleavage into tsRNAs in human HEK293T cells (Wang et al., 2018b). Interestingly, recent reports showed that C38 Q-modified tRNA promotes DNMT2-mediated m^5C on C38 of $tRNA^{Asp}$ (Muller et al., 2015; Tuorto et al., 2018); these discoveries resonate with findings that the establishment of one RNA modification can depend on the existence of another (Barraud et al., 2019). Recent evidence also shows that deletion of ALKBH1 (Rashad et al., 2020) or ALKBH3 (Chen et al., 2019) increased the levels of N1-methyladenine (m^1A) in tRNAs, preventing tRNA cleavage and resulting in less tsRNA production. TRMT10A-mediated N1-methylguanine (m^1G) modification also leads to increased $tRNA^{Gln}$ stability and less production of $tsRNA^{Gln}$ (Cosentino et al., 2018). Moreover, 2'-O-methylation of the C34 in human $tRNA^{Met}$ can prevent site-specific

cleavage of $tRNA^{Met}$ by angiogenin and reduce tsRNA production (Vitali and Kiss, 2019). In addition to preventing tRNA cleavage, some RNA modifications can also promote tsRNA biogenesis. For example, PUS7-mediated pseudouridine (Ψ) at the U8 position has been shown to affect tsRNA biogenesis in stem cells, where deletion of PUS7 leads to a decreased level of several types of 5' tsRNAs (~18 nt) with terminal oligo(G), suggesting that Ψ U8 increases the cleavage of these tRNAs to generate tsRNAs (Guzzi et al., 2018). In another example in yeast, 5-methoxycarbonylmethyl-2-thiouridine (mcm^5S^2) at the anticodon wobble position can promote the cleavage of tRNA into tsRNAs (Lu et al., 2008).

In mammals, four cytoplasmic (5S, 5.8S, 18S, and 28S) rRNAs are encoded by the nuclear genome and two mitochondrial (12S and 16S) rRNAs are encoded by the mitochondrial genome. The rRNAs associate riboproteins to form ribosomes that have the fundamental role in synthesizing proteins. The nucleotides that forming the peptidyl-transferase site are extremely conserved across all species in three kingdoms, which resonate with the extensive existence of rRNAs (Lafontaine and Tollervey, 2001). While intact rRNAs make up the majority part of RNAs in somatic cells, it is unveiled that the full length rRNAs are absent in mature human sperm (Ostermeier et al., 2002), indicating rRNA fragmentation may occur during late spermatogenesis or during epididymal transition. Depending on the RNA-seq technology, an appreciable amount of rsRNAs has been systematically recognized and highlighted in mouse sperm (Chu et al., 2017; Zhang et al., 2018b) and other tissues (Wei et al., 2013).

Despite that eukaryotic rRNA maturation has already been detailedly in details as a sequential cleavage process (Aubert et al., 2018; Henras et al., 2015), the understanding

of rsRNA biogenesis remains limited. Previous small RNA-seq results demonstrate that rsRNAs are preferentially generated from 5' and 3' end terminal of rRNAs in both mouse and human cells (Li et al., 2012). It is also found that a prominent rsRNA in size slightly longer than 50 nucleotides derived from the 5' end of mouse 28S rRNA is associated with apoptosis process (King et al., 2000). A 21-nucleotide rsRNA derived from the 5' end of the 18S rRNA is discovered in zebrafish, the sequence of which is identical to it in mammalian cells that can bind to AGO proteins, suggesting that at least some rsRNAs are generated though miRNA biogenesis pathway (Locati et al., 2018). Similar to tsRNA biogenesis, Angiogenin can also cleave 5.8S rRNAs into short fragments with different lengths (Li et al., 2012). Additionally, rsRNAs expression level is also sensitive to environmental exposures such as altered diet and inflammation (Chu et al., 2017; Natt et al., 2019; Zhang et al., 2018b). Interestingly, deletion of a multisubstrate tRNA methyltransferase Dnmt2 in mouse seems to decrease the level of rsRNA while the rRNA m^5C level is not affected (Legrand et al., 2017), suggesting unknown distinguish regulatory mechanisms between rsRNA and rRNA that independent from m^5C (Zhang et al., 2018b).

A complete rRNA modification landscape has been pinpointed with quantitative mass spectrometry, recognizing dozen types of post-transcriptional modifications that are distributed at hundreds of sites in human 5.8S, 18S and 28 rRNAs (Taoka et al., 2018; Wein et al., 2020). More than 90% of the modifications are covered by pseudouridine or 2'-O-methylation, which can also be the majority types of modifications on rsRNAs that are cleaved from mature rRNAs. Impaired 2'-O-methylation can result in ribosome dysfunction that may cause disease (Nachmani et al., 2019). Although reports on rsRNA

modifications remain limited, their potential to contain modifications are high, given their widespread presence and ingenious response to methylation-related enzymes (Zhang et al., 2018b).

Small RNA library generation strategies and problems

High-throughput RNA sequencing (RNA-seq) has substantially facilitated the discovery of small RNAs over the past decade. It does not require a prior knowledge of the RNA sequence compared with polymerase chain reaction (PCR) and microarray detection method. The small RNA-seq workflow includes: RNA samples preparation, complementary DNA (cDNA) libraries construction, high-throughput sequencing, and small RNA-seq data analysis (Figure 1.2).

Small RNA isolation procedure

Two different RNA preparation protocols are commonly chosen when purifying small RNA from cells or tissues. After extracting total RNAs with reagent Trizol (Rio et al., 2010), denaturing urea-Polyacrylamide Gel Electrophoresis (PAGE) based purification is performed to select the specific size of small RNAs. According to the difference in the affinity of RNA size, the glass-fiber filter-based protocol is designed to separate RNAs, and is employed in the MirVana™ miRNA isolation kit. This commercial kit can enrich RNAs less than 200 nucleotides and has better performance than traditional Trizol extraction method in miRNA recovery (Kim et al., 2012). However, the middle-size RNAs such as 5S RNA (~120 nucleotides), 5.8S RNAs (~150 nucleotides), and tRNAs (76-90 nucleotides) are also mixed in the final products of the kit.

It is also a race against time to perform an RNA isolation procedure before the RNase *in vivo* can catalyze the RNA degradation process. The RNA integrity number (RIN) based on a Bayesian learning technique is designed to quantify the integrity of total RNA. The RIN value ranges from 10 to 1, while 10 represents intact RNA and 1 indicates totally degraded RNA (Schroeder et al., 2006). While the algorithm generates a RIN value in an automated and reproducible manner, it also smooths out the species and tissue specificity. For example, the RIN value does not work out when measuring sperm RNA samples and purified small RNAs that lack 18S and 28S rRNA peaks, although the peaks are essential features in the model (Peng et al., 2012).

cDNA library construction

A typical strategy to obtain small RNA library is sequentially composed of 3' and 5' end adapter ligation, reverse transcription, second-strand synthesis, and PCR amplification steps. Generally, RNA ligases involve in adding adapters to canonical small RNAs carrying 5'-P and 3'-OH termini. Both of T4 RNA Ligase 1 (Rnl1) and T4 RNA Ligase 2 (Rnl2) can catalyze the ligation of 5'-P to 3'-OH of DNA or RNA with ATP participation, which may add adapters to either termini of the target RNAs (Nichols et al., 2008). While Rnl1 prefers single-stranded RNA linking, Rnl2 is utilized for ligating nicks in double-stranded RNAs. A truncated form of T4 Rnl2 that contains the first 249 amino acid, is named as Rnl2tr. It is specifically joins the pre-adenylated adapters to the 3' end of RNA because of lacking adenylation domain (Dai and Gu, 2020). Thus, T4 Rnl2tr can reduce the byproducts of RNA self-cyclization and self-ligation process. The 3' end adapter can also be produced by poly (A) polymerase (PAP), which adds an uncertain number of adenine nucleotides to

the small RNA termini bypassing the adapter ligation step (Dard-Dascot et al., 2018). However, adenine nucleotides at the end of original small RNA sequences are identical to the poly (A) tails, thus increasing the difficulty of bioinformatics analysis when adapters are trimmed.

In addition to the intrinsic properties of the enzymes, some modified nucleotides to the terminus of small RNA also hamper adapter ligation. Small RNA with 2'-O methylation at 3' termini (for example, piRNAs) significantly decreases T4 Rnl1 ligation efficiently and PAP tailing efficiency. However, T4 Rnl2tr keeps a certain degree of the enzyme activity when the modification exists (Munafo and Robb, 2010). Small RNAs that contain 5'-OH, 3'-P or 2,3'-CP can be detected in human bioliquid such as serum and plasma (Akat et al., 2019) thus the 5' and 3' adapter ligation will be blocked by the modifications. 3'-P and 2,3'-CP can be removed using T4 Polynucleotide Kinase (T4PNK) while 5'-OH can also be phosphorylated with the same enzyme. Then the ligation step can proceed smoothly (Akat et al., 2019).

Although other terminal modifications may not be abundant in mammalian small RNAs, they may also inhibit ligation process. The 5'-triphosphate (5'-PPP) group is discovered in virus (Abbas et al., 2013) and C. elegans (Gu et al., 2009). It can be dephosphorylated to 5'-P by polyphosphates (Gu et al., 2009). The 5'-cap structure that is commonly found on mRNAs also exists on small nuclear RNAs (Matera et al., 2007). The pyrophosphatases work as a decapping function, which allow RNA ligation to be performed (Dai and Gu, 2020; Kramer and McLennan, 2019). Mature tRNAs can cognate with the 3'-aminoacyl (3'-aa) group to execute translation function. Although the amino acid of the

charged tRNA may obstruct adapter linking (Honda et al., 2015; Raabe et al., 2014), it is able to be released under alkaline condition without enzyme treatment (Evans et al., 2017).

Parts of RNA methylations enriched in tRNA can obstruct the reverse transcription process and generate truncated cDNA products. α-ketoglutarate-dependent hydroxylase (AlkB) enzyme has been described to precisely demethylate such modifications, namely N^1-methyladenosine (m^1A), N^3-methylcytosine (m^3C), N^1-methylguanosine (m^1G), and N^2, N^2-dimethylguanosine ($m^2{}_2G$) (Cozen et al., 2015; Dai et al., 2017; Zheng et al., 2015). Additionally, some other modifications can generate mismatches between RNA template and cDNA sequence during the synthesis. For example, although a few reverse transcriptases can overcome the interference from oxidized guanine, incorrect base pairs can occur (Alenko et al., 2017).

Bioinformatics tools for small RNA deep sequencing data

The amount of sequencing data is estimated to be 1 zetta (10^{21}) bases per year in 2025. It will be doubled every seven months based on historical growth rate (Stephens et al., 2015). However, how to integrate, process and analyze such massive data has become an opening question waiting to be solved.

Quality control

At the present time, small RNA-seq data can be generated by two major sequencing approaches: PCR cluster-based sequencing (Ross et al., 2013) and DNA nanoball-based sequencing (Drmanac et al., 2010). The former uses exponential DNA amplification, and the latter performs linear DNA amplification to obtain enough DNA signals (Fehlmann et al.,

2016). Neither of these sequencing platforms can avoid generating the substitution error when calling bases due to chemical-to-optical signal conversion process. The sequencing errors also accumulate with read length, despite that the error rate of each base in average can be suppressed to 10^{-5} to 10^{-4} (Ma et al., 2019). The sequencing error is quantified as quality scores, indicating the probability that the corresponding base call is wrong. In addition, the length of small RNA varies between 15-50 nucleotides, usually shorter than the sequencer reading length (50bp, 75bp, or 150bp with single-end strategy). Most of the output reads inescapably incorporate 3' end adapter sequences. Therefore, pre-processing reads by trimming the low-quality sequences/bases and adapters is necessary (Figure 1.2). Cutadapt is a Command-Line Interface (CLI) pipeline to find and remove adapter sequences or other unwanted sequence from RNA-seq reads in an error-tolerant way (Martin, 2011). The fastx_clipper tool embedded in FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) is also a common CLI pipeline to remove adapter sequences, but by using different trimming algorithms from Cutadapt (Buschmann et al., 2016). Both of the tools take raw fastq format files as input and generate trimmed fastq files, which is necessary for downstream analysis.

Small RNA database

Most of the small RNA annotation process relies on aligning sequences to the RNA references, which are normally presented by way of a database. The small RNA database is defined as a compilation for one or some kind of small RNA sequences and annotation of eukaryote, prokaryote as well as viruses. miRBase is the conventional online repository for miRNA information. Currently its latest version (release v22.1) contains 38589 pre-

miRNA hairpin and 48885 mature miRNA sequences, most of which are introduced by small RNA-seq data without northern blot validation (Kozomara and Griffiths-Jones, 2014). A recent study indicates that 65% of mature miRNA candidates are likely false-positives (Alles et al., 2019), some of them are even marked as high-confident in miRBase. Therefore, we need to take extra care when annotating miRNAs based on the database.

Several piRNA databases exist for piRNA annotation, including piRBase (Wang et al., 2019), piRNAbank (Sai Lakshmi and Agrawal, 2008), piRNAQuest (Sarkar et al., 2014), IsopiRBank (Zhang et al., 2018a), piRTarBase (Wu et al., 2019), piRNAdb (https://www.pirnadb.org/), and piRNA cluster database (Rosenkranz, 2016). Majority of the databases are the collection of small RNA-seq data from piRNA enriched tissues (for example, testis, ovary, and brain), or the assemblage of RNA-seq results of cross-linking immunoprecipitated (CLIP) with PIWI proteins except for piRNA cluster database. The piRNA cluster database is a web source of predicted piRNA information from existing small RNA-seq data based on piRNA typical length and clustering features by performing the sorting algorithm proTRAC (Rosenkranz and Zischler, 2012). Similar to the problem of miRBase, false-positives also exist in piRNA databases because biological/biochemical verification is absent to the piRNA sequences.

Since tsRNAs and rsRNAs are derived from their precursors, tRNA and rRNA databases can be regarded as the fundamental references for the annotating step. The genomic and mitochondrial tRNAs in both of the databases GtRNAdb (Chan and Lowe, 2016) and tRNAdb (Juhling et al., 2009) are predicted from whole genome sequences by the probabilistic search software tRNAscan-SE (Chan and Lowe, 2019). Additionally, a handful of

tsRNA databases analyzed from small RNA libraries are also available, including tRFdb (Kumar et al., 2015), MINTbase (Pliatsika et al., 2018), PtRFdb (Gupta et al., 2018), OncotRF (Yao et al., 2020), and tsRBase (Zuo et al., 2021). While the rRNA sequences for microbes are well categorized in SILVA (Quast et al., 2013), rRNA sequence repository specialized for animals/plants is still under construction. Alternatively, the rRNA and other kinds of small RNA entries can be collected from comprehensive RNA databases that include both coding and noncoding sequence data. NCBI Nucleotide database (https://www.ncbi.nlm.nih.gov/nuccore/), RNAcentral (https://rnacentral.org/), Ensembl (https://www.ensembl.org/), and Rfam (https://rfam.xfam.org/) are the commonly accessible databases for browsing RNA information, although not perfect for small RNAs.

Small RNA annotation

An efficient and effective computational pipeline is crucial for small RNA annotation procedure. Dozens of frequently used small RNA annotation programs concentrate on interpreting limited types of canonical small RNAs, such as miRDeep2 (Friedlander et al., 2012), piPipes (Han et al., 2015), proTRAC (Rosenkranz and Zischler, 2012), and piRNN (Wang et al., 2018a). These tools focus on discovering the known miRNAs and piRNAs and predicting the novel ones based on their intrinsic sequence features by different statistical algorithms, such as Bayesian Statistics, binomial probability, or convolution neural network.

Due to the accumulating reports on discovering other non-canonical small RNAs in the recent decade, the capability of annotating these small RNA molecules simultaneously becomes increasingly important. One straightforward solution is mapping sequence

reads to respective references in databases mentioned above, since most small RNAs are derived from the annotated RNAs. Frequently used tools for small RNA sequence alignment, are Bowtie (Langmead et al., 2009), Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009), SeqMap (Jiang and Wong, 2008), and PatMaN (Prufer et al., 2008). Bowtie and BWA base on Ferragina-Manziniuse (FM) index searching algorithm, which is designed for short reads mapping (Hatem et al., 2013), while SeqMap bases on the hash table-based algorithm. PatMaN requires no indexing step, but the running time takes 10 times longer on average (Hoffmann et al., 2009). FM index algorithm is preferred when reads are aligned to multiple identical copies in the reference sequences (Shen et al., 2014), which frequently occurs for non-canonical small RNA mapping. Several bioinformatics tools are established for multiple small RNA parallel annotation based on those mapping algorithms. For example, sRNAtoolbox (Aparicio-Puerta et al., 2019) as an integrated tool basing on bowtie focuses on miRNA profiling, while it can also annotate other small RNA sequences. It has both web interface and standalone graphical user interface (GUI), although the offline version depends on virtual machine, which in general causes the performance loss compared with the naïve one (Zhang et al., 2012). Unitas (Gebert et al., 2017) is built upon SeqMap and obtains reference information from Ensembl, piRNA cluster database, SILVA, GtRNAdb, and miRBase. The source code of the latest version of Unitas is provided online while the precompiled executable file to be requested through contact support. The internet connection is required in order to download reference sequences before its usage. The UEA sRNA Workbench (Stocks et al., 2018) is a cross-platform tool calls PatMaN for small RNA annotation. The miRNA, tRNA and rRNA entries are incorporated in the software

although the common species (for example, mouse, rat, and human) they belong to are absent.

The annotation pipeline performance can be quantified by parameters of true positive (TP), true negative (TN), false positive (FP), false negative (FN). They present as Sensitivity (TP/(TP+FN)), Precision (TP/(FP+TP)), and F-measure (2Precision×Sensitivity/(Precision+Sensitivity)) (Di Bella et al., 2020). Additionally, the software customizability and expandability for reference database and input format are usually supportive when non-model organisms are annotated. Another point to consider when choosing annotation software is that the accessible rate of web-based bioinformatics tools descends annually. After 10 years, only ~50% tools published in 2010 are reachable (Kern et al., 2020). Therefore, it is responsible to choose the standalone version of a bioinformatics software that uploads to a common software repositories platform, rather than keep it only on the lab website, during its developing or utilizing.

Objectives

Despite the significant progress achieved, we may still at an early stage to unravel the full complexity of small RNA world. More details of the panoramic small RNA landscape are still waiting to be filled, especially the presence and abundance of non-canonical small RNAs in mammalian tissues and cells. While the small RNA analysis tools are the premise and basis of small RNA exploration, a comprehensive annotation software integrated with well-organized reference database is still some distance away.

Another major challenge in revealing small RNA population is locating small RNAs that commonly carry modified nucleotides. These modified small RNAs usually conceal under regular sequencing method. The recently developed enzymatic treatment procedures that remove specific modifications on either RNA terminus or body holds promising potential to facilitate the detection of non-canonical small RNAs, and the combinatorial usage of multiple enzymes may assist to unveil the small RNA world.

Figures

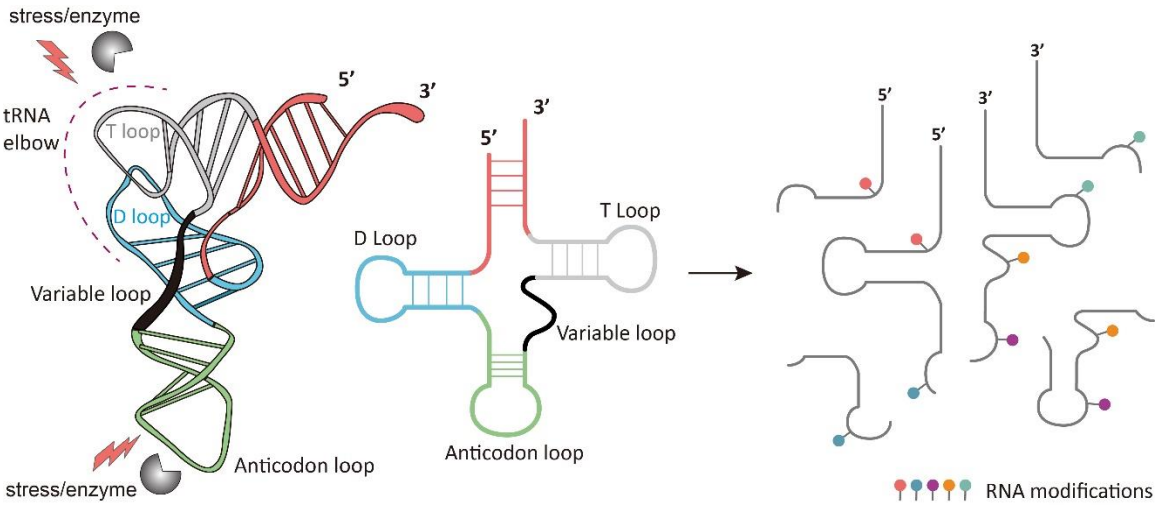

Figure 1.1: tsRNA biogenesis is rooted in tRNA structure and regulated by tRNA modifications

The 2D and 3D structure of a tRNA, showing the loose sites at the anticodon loop and the tRNA elbow (joint of D- and T-loops) that represent the preferred sites of fragmentation to generate various types of tsRNAs.

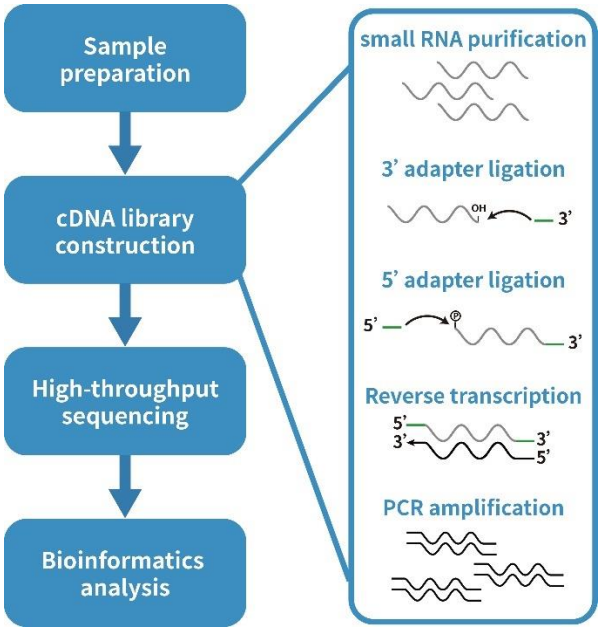

Figure 1.2: Workflow of small RNA sequencing

The main steps of small RNA sequencing and cDNA library construction are outlined in the figure.

References

Abbas, Y.M., Pichlmair, A., Gorna, M.W., Superti-Furga, G., and Nagar, B. (2013). Structural basis for viral 5'-PPP-RNA recognition by human IFIT proteins. Nature *494*, 60-64.

Akat, K.M., Lee, Y.A., Hurley, A., Morozov, P., Max, K.E., Brown, M., Bogardus, K., Sopeyin, A., Hildner, K., Diacovo, T.G.*, et al.* (2019). Detection of circulating extracellular mRNAs by modified small-RNA-sequencing analysis. JCI Insight *5*.

Alenko, A., Fleming, A.M., and Burrows, C.J. (2017). Reverse Transcription Past Products of Guanine Oxidation in RNA Leads to Insertion of A and C opposite 8-Oxo-7,8-dihydroguanine and A and G opposite 5-Guanidinohydantoin and Spiroiminodihydantoin Diastereomers. Biochemistry *56*, 5053-5064.

Alles, J., Fehlmann, T., Fischer, U., Backes, C., Galata, V., Minet, M., Hart, M., Abu-Halima, M., Grasser, F.A., Lenhof, H.P.*, et al.* (2019). An estimate of the total number of true human miRNAs. Nucleic acids research *47*, 3353-3364.

Andersen, K.L., and Collins, K. (2012). Several RNase T2 enzymes function in induced tRNA and rRNA turnover in the ciliate Tetrahymena. Mol Biol Cell *23*, 36-44.

Aparicio-Puerta, E., Lebron, R., Rueda, A., Gomez-Martin, C., Giannoukakos, S., Jaspez, D., Medina, J.M., Zubkovic, A., Jurak, I., Fromm, B.*, et al.* (2019). sRNAbench and sRNAtoolbox 2019: intuitive fast small RNA profiling and differential expression. Nucleic acids research *47*, W530-W535.

Aravin, A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., Morris, P., Brownstein, M.J., Kuramochi-Miyagawa, S., Nakano, T.*, et al.* (2006). A novel class of small RNAs bind to MILI protein in mouse testes. Nature *442*, 203-207.

Aravin, A.A., Naumova, N.M., Tulin, A.V., Vagin, V.V., Rozovsky, Y.M., and Gvozdev, V.A. (2001). Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline. Current biology : CB *11*, 1017-1027.

Aubert, M., O'Donohue, M.F., Lebaron, S., and Gleizes, P.E. (2018). Pre-Ribosomal RNA Processing in Human Cells: From Mechanisms to Congenital Diseases. Biomolecules *8*.

Barraud, P., Gato, A., Heiss, M., Catala, M., Kellner, S., and Tisne, C. (2019). Time-resolved NMR monitoring of tRNA maturation. Nat Commun *10*, 3373.

Bartel, D.P. (2018). Metazoan MicroRNAs. Cell *173*, 20-51.

Beyret, E., Liu, N., and Lin, H. (2012). piRNA biogenesis during adult spermatogenesis in mice is independent of the ping-pong mechanism. Cell research *22*, 1429-1439.

Buschmann, D., Haberberger, A., Kirchner, B., Spornraft, M., Riedmaier, I., Schelling, G., and Pfaffl, M.W. (2016). Toward reliable biomarker signatures in the age of liquid biopsies - how to standardize the small RNA-Seq workflow. Nucleic acids research *44*, 5995-6018.

Chan, P.P., and Lowe, T.M. (2016). GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic acids research *44*, D184-189.

Chan, P.P., and Lowe, T.M. (2019). tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. Methods in molecular biology *1962*, 1-14.

Chen, Z., Qi, M., Shen, B., Luo, G., Wu, Y., Li, J., Lu, Z., Zheng, Z., Dai, Q., and Wang, H. (2019). Transfer RNA demethylase ALKBH3 promotes cancer progression via induction of tRNA-derived small RNAs. Nucleic acids research *47*, 2533-2545.

Chu, C., Yu, L., Wu, B., Ma, L., Gou, L.T., He, M., Guo, Y., Li, Z.T., Gao, W., Shi, H.*, et al.* (2017). A sequence of 28S rRNA-derived small RNAs is enriched in mature sperm and various somatic tissues and possibly associates with inflammation. J Mol Cell Biol *9*, 256-259.

Cole, C., Sobala, A., Lu, C., Thatcher, S.R., Bowman, A., Brown, J.W., Green, P.J., Barton, G.J., and Hutvagner, G. (2009). Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. RNA *15*, 2147-2160.

Cosentino, C., Toivonen, S., Diaz Villamil, E., Atta, M., Ravanat, J.L., Demine, S., Schiavo, A.A., Pachera, N., Deglasse, J.P., Jonas, J.C.*, et al.* (2018). Pancreatic beta-cell tRNA hypomethylation and fragmentation link TRMT10A deficiency with diabetes. Nucleic acids research *46*, 10302-10318.

Cozen, A.E., Quartley, E., Holmes, A.D., Hrabeta-Robinson, E., Phizicky, E.M., and Lowe, T.M. (2015). ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. Nature methods *12*, 879-884.

Czech, B., and Hannon, G.J. (2016). One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing. Trends Biochem Sci *41*, 324-337.

Dai, H., and Gu, W. (2020). Strategies and Best Practice in Cloning Small RNAs. Gene Technol *9*.

Dai, Q., Zheng, G., Schwartz, M.H., Clark, W.C., and Pan, T. (2017). Selective Enzymatic Demethylation of N(2) ,N(2) -Dimethylguanosine in RNA and Its Application in High-Throughput tRNA Sequencing. Angew Chem Int Ed Engl *56*, 5017-5020.

Dard-Dascot, C., Naquin, D., d'Aubenton-Carafa, Y., Alix, K., Thermes, C., and van Dijk, E. (2018). Systematic comparison of small RNA library preparation protocols for next-generation sequencing. BMC genomics *19*, 118.

Dhahbi, J.M., Spindler, S.R., Atamna, H., Yamakawa, A., Boffelli, D., Mote, P., and Martin, D.I. (2013). 5' tRNA halves are present as abundant complexes in serum, concentrated in blood cells, and modulated by aging and calorie restriction. BMC genomics *14*, 298.

Di Bella, S., La Ferlita, A., Carapezza, G., Alaimo, S., Isacchi, A., Ferro, A., Pulvirenti, A., and Bosotti, R. (2020). A benchmarking of pipelines for detecting ncRNAs from RNA-Seq data. Brief Bioinform *21*, 1987-1998.

Donovan, J., Rath, S., Kolet-Mandrikov, D., and Korennykh, A. (2017). Rapid RNase L-driven arrest of protein synthesis in the dsRNA response without degradation of translation machinery. Rna *23*, 1660-1671.

Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G.*, et al.* (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science *327*, 78-81.

Du, T., and Zamore, P.D. (2005). microPrimer: the biogenesis and function of microRNA. Development *132*, 4645-4652.

El-Mogy, M., Lam, B., Haj-Ahmad, T.A., McGowan, S., Yu, D., Nosal, L., Rghei, N., Roberts, P., and Haj-Ahmad, Y. (2018). Diversity and signature of small RNA in different bodily fluids using next generation sequencing. BMC genomics *19*, 408.

Evans, M.E., Clark, W.C., Zheng, G., and Pan, T. (2017). Determination of tRNA aminoacylation levels by high-throughput sequencing. Nucleic acids research *45*, e133.

Fehlmann, T., Reinheimer, S., Geng, C., Su, X., Drmanac, S., Alexeev, A., Zhang, C., Backes, C., Ludwig, N., Hart, M.*, et al.* (2016). cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. Clin Epigenetics *8*, 123.

Friedlander, M.R., Mackowiak, S.D., Li, N., Chen, W., and Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic acids research *40*, 37-52.

Fu, H., Feng, J., Liu, Q., Sun, F., Tie, Y., Zhu, J., Xing, R., Sun, Z., and Zheng, X. (2009). Stress induces tRNA cleavage by angiogenin in mammalian cells. FEBS letters *583*, 437-442.

Gainetdinov, I., Colpan, C., Arif, A., Cecchini, K., and Zamore, P.D. (2018). A Single Mechanism of Biogenesis, Initiated and Directed by PIWI Proteins, Explains piRNA Production in Most Animals. Molecular cell *71*, 775-790 e775.

Gebert, D., Hewel, C., and Rosenkranz, D. (2017). unitas: the universal tool for annotation of small RNAs. BMC genomics *18*, 644.

Girard, A., Sachidanandam, R., Hannon, G.J., and Carmell, M.A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature *442*, 199-202.

Grivna, S.T., Beyret, E., Wang, Z., and Lin, H. (2006). A novel class of small RNAs in mouse spermatogenic cells. Genes & development *20*, 1709-1714.

Gu, W., Shirayama, M., Conte, D., Jr., Vasale, J., Batista, P.J., Claycomb, J.M., Moresco, J.J., Youngman, E.M., Keys, J., Stoltz, M.J.*, et al.* (2009). Distinct argonaute-mediated 22G-RNA pathways direct genome surveillance in the C. elegans germline. Molecular cell *36*, 231-244.

Gupta, N., Singh, A., Zahra, S., and Kumar, S. (2018). PtRFdb: a database for plant transfer RNA-derived fragments. Database (Oxford) *2018*.

Guzzi, N., Ciesla, M., Ngoc, P.C.T., Lang, S., Arora, S., Dimitriou, M., Pimkova, K., Sommarin, M.N.E., Munita, R., Lubas, M.*, et al.* (2018). Pseudouridylation of tRNA-Derived Fragments Steers Translational Control in Stem Cells. Cell *173*, 1204-1216 e1226.

Han, B.W., Wang, W., Zamore, P.D., and Weng, Z. (2015). piPipes: a set of pipelines for piRNA and transposon analysis via small RNA-seq, RNA-seq, degradome- and CAGE-seq, ChIP-seq and genomic DNA sequencing. Bioinformatics *31*, 593-595.

Hatem, A., Bozdag, D., Toland, A.E., and Catalyurek, U.V. (2013). Benchmarking short sequence mapping tools. BMC Bioinformatics *14*, 184.

Haussecker, D., Huang, Y., Lau, A., Parameswaran, P., Fire, A.Z., and Kay, M.A. (2010). Human tRNA-derived small RNAs in the global regulation of RNA silencing. RNA *16*, 673-695.

Helwak, A., and Tollervey, D. (2014). Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH). Nature protocols *9*, 711-728.

Henras, A.K., Plisson-Chastang, C., O'Donohue, M.F., Chakraborty, A., and Gleizes, P.E. (2015). An overview of pre-ribosomal RNA processing in eukaryotes. Wiley interdisciplinary reviews RNA *6*, 225-242.

Hock, J., and Meister, G. (2008). The Argonaute protein family. Genome Biol *9*, 210.

Hoffmann, S., Otto, C., Kurtz, S., Sharma, C.M., Khaitovich, P., Vogel, J., Stadler, P.F., and Hackermuller, J. (2009). Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS computational biology *5*, e1000502.

Honda, S., Loher, P., Shigematsu, M., Palazzo, J.P., Suzuki, R., Imoto, I., Rigoutsos, I., and Kirino, Y. (2015). Sex hormone-dependent tRNA halves enhance cell proliferation in breast and prostate cancers. Proceedings of the National Academy of Sciences of the United States of America *112*, E3816-3825.

Jiang, H., and Wong, W.H. (2008). SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics *24*, 2395-2396.

Juhling, F., Morl, M., Hartmann, R.K., Sprinzl, M., Stadler, P.F., and Putz, J. (2009). tRNAdb 2009: compilation of tRNA sequences and tRNA genes. Nucleic acids research *37*, D159-162.

Kern, F., Fehlmann, T., and Keller, A. (2020). On the lifetime of bioinformatics web services. Nucleic acids research *48*, 12523-12533.

Kikuchi, Y., Sasaki, N., and Ando-Yamagami, Y. (1990). Cleavage of tRNA within the mature tRNA sequence by the catalytic RNA of RNase P: implication for the formation of the primer tRNA fragment for reverse transcription in copia retrovirus-like particles. Proceedings of the National Academy of Sciences of the United States of America *87*, 8105-8109.

Kim, V.N., Han, J., and Siomi, M.C. (2009). Biogenesis of small RNAs in animals. Nat Rev Mol Cell Biol *10*, 126-139.

Kim, Y.K., Yeo, J., Kim, B., Ha, M., and Kim, V.N. (2012). Short structured RNAs with low GC content are selectively lost during extraction from a small number of cells. Molecular cell *46*, 893-895.

King, K.L., Jewell, C.M., Bortner, C.D., and Cidlowski, J.A. (2000). 28S ribosome degradation in lymphoid cell apoptosis: evidence for caspase and Bcl-2-dependent and -independent pathways. Cell death and differentiation *7*, 994-1001.

Kirino, Y., and Mourelatos, Z. (2007). Mouse Piwi-interacting RNAs are 2'-O-methylated at their 3' termini. Nature structural & molecular biology *14*, 347-348.

Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic acids research *42*, D68-73.

Kramer, S., and McLennan, A.G. (2019). The complex enzymology of mRNA decapping: Enzymes of four classes cleave pyrophosphate bonds. Wiley interdisciplinary reviews RNA *10*, e1511.

Kumar, P., Mudunuri, S.B., Anaya, J., and Dutta, A. (2015). tRFdb: a database for transfer RNA fragments. Nucleic acids research *43*, D141-145.

Lafontaine, D.L., and Tollervey, D. (2001). The function and synthesis of ribosomes. Nat Rev Mol Cell Biol *2*, 514-520.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. Science *294*, 853-858.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science *294*, 858-862.

Lau, N.C., Seto, A.G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D.P., and Kingston, R.E. (2006). Characterization of the piRNA complex from rat testes. Science *313*, 363-367.

Lee, R.C., and Ambros, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. Science *294*, 862-864.

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell *75*, 843-854.

Lee, S.R., and Collins, K. (2005). Starvation-induced cleavage of the tRNA anticodon loop in Tetrahymena thermophila. J Biol Chem *280*, 42744-42749.

Legrand, C., Tuorto, F., Hartmann, M., Liebers, R., Jacob, D., Helm, M., and Lyko, F. (2017). Statistically robust methylation calling for whole-transcriptome bisulfite sequencing reveals distinct methylation patterns for mouse RNAs. Genome research *27*, 1589-1596.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, Z., Ender, C., Meister, G., Moore, P.S., Chang, Y., and John, B. (2012). Extensive terminal and asymmetric processing of small RNAs from rRNAs, snoRNAs, snRNAs, and tRNAs. Nucleic acids research *40*, 6787-6799.

Locati, M.D., Pagano, J.F.B., Abdullah, F., Ensink, W.A., van Olst, M., van Leeuwen, S., Nehrdich, U., Spaink, H.P., Rauwerda, H., Jonker, M.J.*, et al.* (2018). Identifying small RNAs derived from maternal- and somatic-type rRNAs in zebrafish development. Genome *61*, 371-378.

Lu, J., Esberg, A., Huang, B., and Bystrom, A.S. (2008). Kluyveromyces lactis gamma-toxin, a ribonuclease that recognizes the anticodon stem loop of tRNA. Nucleic acids research *36*, 1072-1080.

Luhtala, N., and Parker, R. (2010). T2 Family ribonucleases: ancient enzymes with diverse roles. Trends in biochemical sciences *35*, 253-259.

Lyons, S.M., Fay, M.M., Akiyama, Y., Anderson, P.J., and Ivanov, P. (2017). RNA biology of angiogenin: Current state and perspectives. RNA biology *14*, 171-178.

Ma, X., Shao, Y., Tian, L., Flasch, D.A., Mulder, H.L., Edmonson, M.N., Liu, Y., Chen, X., Newman, S., Nakitandwe, J.*, et al.* (2019). Analysis of error profiles in deep next-generation sequencing data. Genome Biol *20*, 50.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011 *17*, 3.

Matera, A.G., Terns, R.M., and Terns, M.P. (2007). Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. Nat Rev Mol Cell Biol *8*, 209-220.

Muller, M., Hartmann, M., Schuster, I., Bender, S., Thuring, K.L., Helm, M., Katze, J.R., Nellen, W., Lyko, F., and Ehrenhofer-Murray, A.E. (2015). Dynamic modulation of Dnmt2-dependent tRNA methylation by the micronutrient queuine. Nucleic acids research *43*, 10952-10962.

Munafo, D.B., and Robb, G.B. (2010). Optimization of enzymatic reaction conditions for generating representative pools of cDNA from small RNA. RNA *16*, 2537-2552.

Nachmani, D., Bothmer, A.H., Grisendi, S., Mele, A., Bothmer, D., Lee, J.D., Monteleone, E., Cheng, K., Zhang, Y., Bester, A.C.*, et al.* (2019). Germline NPM1 mutations lead to altered rRNA 2'-O-methylation and cause dyskeratosis congenita. Nature genetics *51*, 1518-1529.

Natt, D., Kugelberg, U., Casas, E., Nedstrand, E., Zalavary, S., Henriksson, P., Nijm, C., Jaderquist, J., Sandborg, J., Flinke, E.*, et al.* (2019). Human sperm displays rapid responses to diet. PLoS biology *17*, e3000559.

Nichols, N.M., Tabor, S., and McReynolds, L.A. (2008). RNA ligases. Curr Protoc Mol Biol *Chapter 3*, Unit3 15.

Ostermeier, G.C., Dix, D.J., Miller, D., Khatri, P., and Krawetz, S.A. (2002). Spermatozoal RNA profiles of normal fertile men. Lancet *360*, 772-777.

Ozata, D.M., Gainetdinov, I., Zoch, A., O'Carroll, D., and Zamore, P.D. (2019). PIWI-interacting RNAs: small RNAs with big functions. Nature reviews Genetics *20*, 89-108.

Pasquinelli, A.E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M.I., Maller, B., Hayward, D.C., Ball, E.E., Degnan, B., Muller, P.*, et al.* (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature *408*, 86-89.

Peng, H., Shi, J., Zhang, Y., Zhang, H., Liao, S., Li, W., Lei, L., Han, C., Ning, L., Cao, Y.*, et al.* (2012). A novel class of tRNA-derived small RNAs extremely enriched in mature mouse sperm. Cell research *22*, 1609-1612.

Perez-Hernandez, J., Forner, M.J., Pinto, C., Chaves, F.J., Cortes, R., and Redon, J. (2015). Increased Urinary Exosomal MicroRNAs in Patients with Systemic Lupus Erythematosus. PloS one *10*, e0138618.

Pliatsika, V., Loher, P., Magee, R., Telonis, A.G., Londin, E., Shigematsu, M., Kirino, Y., and Rigoutsos, I. (2018). MINTbase v2.0: a comprehensive database for tRNA-derived fragments that includes nuclear and mitochondrial fragments from all The Cancer Genome Atlas projects. Nucleic acids research *46*, D152-D159.

Prufer, K., Stenzel, U., Dannemann, M., Green, R.E., Lachmann, M., and Kelso, J. (2008). PatMaN: rapid alignment of short sequences to large databases. Bioinformatics *24*, 1530-1531.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glockner, F.O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids research *41*, D590-596.

Raabe, C.A., Tang, T.H., Brosius, J., and Rozhdestvensky, T.S. (2014). Biases in small RNA deep sequencing data. Nucleic acids research *42*, 1414-1426.

Rashad, S., Han, X., Sato, K., Mishima, E., Abe, T., Tominaga, T., and Niizuma, K. (2020). The stress specific impact of ALKBH1 on tRNA cleavage and tiRNA generation. RNA biology *17*, 1092-1103.

Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature *403*, 901-906.

Reinsborough, C.W., Ipas, H., Abell, N.S., Nottingham, R.M., Yao, J., Devanathan, S.K., Shelton, S.B., Lambowitz, A.M., and Xhemalce, B. (2019). BCDIN3D regulates tRNAHis 3' fragment processing. PLoS genetics *15*, e1008273.

Rio, D.C., Ares, M., Hannon, G.J., and Nilsen, T.W. (2010). Purification of RNA Using TRIzol (TRI Reagent). Cold Spring Harbor Protocols *2010*, pdb.prot5439.

Rosenkranz, D. (2016). piRNA cluster database: a web resource for piRNA producing loci. Nucleic acids research *44*, D223-230.

Rosenkranz, D., and Zischler, H. (2012). proTRAC--a software for probabilistic piRNA cluster detection, visualization and analysis. BMC Bioinformatics *13*, 5.

Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C., and Jaffe, D.B. (2013). Characterizing and measuring bias in sequence data. Genome Biol *14*, R51.

Rudland, P.S., and Dube, S.K. (1969). Specific interaction of an initiator tRNA fragment with 30 s ribosomal subunits. Journal of molecular biology *43*, 273-280.

Sai Lakshmi, S., and Agrawal, S. (2008). piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. Nucleic acids research *36*, D173-177.

Sarkar, A., Maji, R.K., Saha, S., and Ghosh, Z. (2014). piRNAQuest: searching the piRNAome for silencers. BMC genomics *15*, 555.

Schaefer, M., Pollex, T., Hanna, K., Tuorto, F., Meusburger, M., Helm, M., and Lyko, F. (2010). RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. Genes & development *24*, 1590-1595.

Schimmel, P. (2018). The emerging complexity of the tRNA world: mammalian tRNAs beyond protein synthesis. Nat Rev Mol Cell Biol *19*, 45-58.

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., Lightfoot, S., Menzel, W., Granzow, M., and Ragg, T. (2006). The RIN: an RNA

integrity number for assigning integrity values to RNA measurements. BMC molecular biology *7*, 3.

Shen, B., Teschendorff, A.E., Zhi, D., and Xia, J. (2014). Biomedical data integration, modeling, and simulation in the era of big data and translational medicine. Biomed Res Int *2014*, 731546.

Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., Efron, M.J., Iyer, R., Schatz, M.C., Sinha, S., and Robinson, G.E. (2015). Big Data: Astronomical or Genomical? PLoS biology *13*, e1002195.

Stocks, M.B., Mohorianu, I., Beckers, M., Paicu, C., Moxon, S., Thody, J., Dalmay, T., and Moulton, V. (2018). The UEA sRNA Workbench (version 4.4): a comprehensive suite of tools for analyzing miRNAs and sRNAs. Bioinformatics *34*, 3382-3384.

Taoka, M., Nobe, Y., Yamaki, Y., Sato, K., Ishikawa, H., Izumikawa, K., Yamauchi, Y., Hirota, K., Nakayama, H., Takahashi, N.*, et al.* (2018). Landscape of the complete RNA chemical modifications in the human 80S ribosome. Nucleic acids research *46*, 9289-9298.

Thompson, D.M., Lu, C., Green, P.J., and Parker, R. (2008). tRNA cleavage is a conserved response to oxidative stress in eukaryotes. RNA *14*, 2095-2103.

Thompson, D.M., and Parker, R. (2009). The RNase Rny1p cleaves tRNAs and promotes cell death during oxidative stress in Saccharomyces cerevisiae. The Journal of cell biology *185*, 43-50.

Tuorto, F., Legrand, C., Cirzi, C., Federico, G., Liebers, R., Muller, M., Ehrenhofer-Murray, A.E., Dittmar, G., Grone, H.J., and Lyko, F. (2018). Queuosine-modified tRNAs confer nutritional control of protein translation. The EMBO journal *37*.

Tuorto, F., Liebers, R., Musch, T., Schaefer, M., Hofmann, S., Kellner, S., Frye, M., Helm, M., Stoecklin, G., and Lyko, F. (2012). RNA cytosine methylation by Dnmt2 and NSun2 promotes tRNA stability and protein synthesis. Nature structural & molecular biology *19*, 900-905.

Vitali, P., and Kiss, T. (2019). Cooperative 2'-O-methylation of the wobble cytidine of human elongator tRNA(Met)(CAT) by a nucleolar and a Cajal body-specific box C/D RNP. Genes & development *33*, 741-746.

Wagner, E.G.H., and Romby, P. (2015). Small RNAs in bacteria and archaea: who they are, what they do, and how they do it. Advances in genetics *90*, 133-208.

Wang, J., Zhang, P., Lu, Y., Li, Y., Zheng, Y., Kan, Y., Chen, R., and He, S. (2019). piRBase: a comprehensive database of piRNA sequences. Nucleic acids research *47*, D175-D180.

Wang, K., Hoeksema, J., and Liang, C. (2018a). piRNN: deep learning algorithm for piRNA prediction. PeerJ *6*, e5429.

Wang, X., Matuszek, Z., Huang, Y., Parisien, M., Dai, Q., Clark, W., Schwartz, M.H., and Pan, T. (2018b). Queuosine modification protects cognate tRNAs against ribonuclease cleavage. Rna *24*, 1305-1313.

Wei, H., Zhou, B., Zhang, F., Tu, Y., Hu, Y., Zhang, B., and Zhai, Q. (2013). Profiling and identification of small rDNA-derived RNAs and their potential biological functions. PloS one *8*, e56842.

Wein, S., Andrews, B., Sachsenberg, T., Santos-Rosa, H., Kohlbacher, O., Kouzarides, T., Garcia, B.A., and Weisser, H. (2020). A computational platform for high-throughput analysis of RNA sequences and modifications by mass spectrometry. Nat Commun *11*, 926.

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

Wu, P.H., and Zamore, P.D. (2021). Defining the functions of PIWI-interacting RNAs. Nat Rev Mol Cell Biol *22*, 239-240.

Wu, W.S., Brown, J.S., Chen, T.T., Chu, Y.H., Huang, W.C., Tu, S., and Lee, H.C. (2019). piRTarBase: a database of piRNA targeting sites and their roles in gene regulation. Nucleic acids research *47*, D181-D187.

Yamasaki, S., Ivanov, P., Hu, G.F., and Anderson, P. (2009). Angiogenin cleaves tRNA and promotes stress-induced translational repression. The Journal of cell biology *185*, 35-42.

Yao, D., Sun, X., Zhou, L., Amanullah, M., Pan, X., Liu, Y., Liang, M., Liu, P., and Lu, Y. (2020). OncotRF: an online resource for exploration of tRNA-derived fragments in human cancers. RNA Biol *17*, 1081-1091.

Yeri, A., Courtright, A., Reiman, R., Carlson, E., Beecroft, T., Janss, A., Siniard, A., Richholt, R., Balak, C., Rozowsky, J.*, et al.* (2017). Total Extracellular Small RNA Profiles from Plasma, Saliva, and Urine of Healthy Subjects. Sci Rep *7*, 44061.

Zhang, H., Ali, A., Gao, J., Ban, R., Jiang, X., Zhang, Y., and Shi, Q. (2018a). IsopiRBank: a research resource for tracking piRNA isoforms. Database (Oxford) *2018*.

Zhang, Y., Oertel, R., and Rehm, W. (2012). Performance loss on virtual machines. Studentensymposium Informatik Chemnitz 2012, 52.

Zhang, Y., Zhang, X., Shi, J., Tuorto, F., Li, X., Liu, Y., Liebers, R., Zhang, L., Qu, Y., Qian, J.*, et al.* (2018b). Dnmt2 mediates intergenerational transmission of paternally acquired metabolic disorders through sperm small non-coding RNAs. Nat Cell Biol *20*, 535-540.

Zhang, Y., Zhang, Y., Shi, J., Zhang, H., Cao, Z., Gao, X., Ren, W., Ning, Y., Ning, L., Cao, Y.*, et al.* (2014). Identification and characterization of an ancient class of small RNAs enriched in serum associating with active infection. J Mol Cell Biol *6*, 172-174.

Zheng, G., Qin, Y., Clark, W.C., Dai, Q., Yi, C., He, C., Lambowitz, A.M., and Pan, T. (2015). Efficient and quantitative high-throughput tRNA sequencing. Nature methods *12*, 835-837.

Zuo, Y., Zhu, L., Guo, Z., Liu, W., Zhang, J., Zeng, Z., Wu, Q., Cheng, J., Fu, X., Jin, Y.*, et al.* (2021). tsRBase: a comprehensive database for expression and function of tsRNAs in multiple species. Nucleic acids research *49*, D1038-D1045.

Chapter 2: An annotating tool optimized for non-canonical small RNAs

Abstract

High-throughput RNA-seq has revolutionized the process of small RNA discovery, leading to a rapid expansion of small RNA categories. In addition to the previously well-characterized small RNAs such as miRNAs and piRNAs, recent emerging studies have spotlighted on tsRNAs and rsRNAs as new categories of non-canonical small RNAs that bear versatile functions. Since existing software and pipelines for small RNA annotation mostly focus on analyzing miRNAs or piRNAs, the small RNA annotation pipeline optimized for rRNA- and tRNA-derived small RNAs (SPORTS1.0) is developed. SPORTS1.0 is optimized for analyzing tsRNAs and rsRNAs from small RNA-seq data, in addition to its capacity to annotate canonical small RNAs such as miRNAs and piRNAs. Moreover, SPORTS1.0 can predict potential RNA modification sites based on nucleotide mismatches within small RNAs. SPORTS1.0 is precompiled to annotate small RNAs for a wide range of 68 species across bacteria, yeast, plant, and animal kingdoms, while additional species for analyses could be readily expanded upon end users' input. For demonstration, by analyzing existing small RNA datasets using SPORTS1.0, it is revealing that distinct signatures are present in tsRNAs and rsRNAs from different mouse cell types. Compared to other small RNA species, tsRNAs bear the highest mismatch rate, which is consistent with their highly modified nature. SPORTS1.0 is an open-source software and can be publically accessed at https://github.com/junchaoshi/sports1.0.

Introduction

Expanding classes of small RNAs have emerged as key regulators of gene expression, genome stability, and epigenetic regulation (Cech and Steitz, 2014; Chen et al., 2016b). In addition to the previously well-characterized small RNA classes such as miRNAs and piRNAs, recent analysis of small RNA-seq data has led to the identification of expanding novel small RNA families. These include tsRNA and rsRNAs (Kumar et al., 2016). tsRNAs and rsRNAs have been discovered in a wide range of species with evolutionary conservation, supposedly due, in part, to the highly conservative sequence of their respective precursors, that is, tRNAs and rRNAs (Kumar et al., 2016). Interestingly, tsRNAs and rsRNAs have been abundantly found in unicellular organisms (for example, protozoa), where canonical small RNA pathways such as miRNA and piRNAs are entirely lacking (Garcia-Silva et al., 2014; Lambertz et al., 2015; Liao et al., 2014). The dynamic regulation of tsRNAs and rsRNAs in these unicellular organisms suggests that they are among the most ancient classes of small RNAs for intra- and inter-cellular communications (Szempruch et al., 2016).

Moreover, recent emerging evidence from mammalian species have highlighted the diverse biological functions mediated by tsRNAs, including regulating ribosome biogenesis, translation initiation, retrotransposon control, cancer metastasis, stem cell differentiation, neurological diseases, and epigenetic inheritance (Anderson and Ivanov, 2014; Chen et al., 2016a; Gebetsberger et al., 2017; Ivanov et al., 2011; Kim et al., 2017; Kumar et al., 2016; Martinez et al., 2017; Schimmel, 2018; Schorn et al., 2017). Although tsRNAs are known to be involved in regulating these processes at both post-transcriptional and translational levels (Ivanov et al., 2011; Kim et al., 2017; Luo et al., 2018), the exact molecular

mechanisms of how tsRNAs exert their functions have not been fully understood. Compared to tsRNAs, rsRNAs are more recently discovered and also exhibit tissue-specific distribution. Dynamic expression of rsRNAs is associated with diseases such as metabolic disorders and inflammation (Chu et al., 2017; Wei et al., 2013; Zhang et al., 2018). The diverse biological functions of tsRNAs and rsRNAs and their strong disease associations are now pushing the new frontier of small RNA research.

Currently, there are multiple existing general small RNA annotation software and pipelines (Axtell, 2013; Fasold et al., 2011; Mohorianu et al., 2017; Rueda et al., 2015; Wu et al., 2017), and some have been developed aiming to analyze tsRNAs (Selitsky and Sethupathy, 2015; Thompson et al., 2018; Zheng et al., 2016). However, there still lack the specialized tools that can simultaneously analyze both tsRNAs and rsRNAs in addition to other canonical small RNAs. Here, SPORTS1.0 is provided, which can annotate and profile canonical small RNAs such as miRNAs and piRNAs, and is also optimized to analyze tsRNAs and rsRNAs from small RNA-seq data (Figure 2.1). In addition, SPORTS1.0 can help predict potential RNA modification sites based on nucleotide mismatches within small RNAs.

Results

As an example, SPORTS1.0 was performed to analyze small RNA-seq datasets from mouse sperm (GSM2304822 (Yang et al., 2016)), bone marrow cells (GSM1604100 (Tuorto et al., 2015)), and intestinal epithelial cells (GSM1975854 (Peck et al., 2017)) samples. Graphic output by SPORTS1.0 reveals distinct small RNA profiles in sperm

(Figure 2.2a), bone marrow cells (Figure 2.2b), and intestinal epithelial cells (Figure 2.2c) samples. tsRNAs and rsRNAs are found equally or more abundantly than previously well-known miRNAs or piRNAs (length distribution data for each type of small RNA are exemplified in Table S2.1). In particularly, tsRNAs are dominant in sperm, rsRNAs are highest in bone marrow cells, and intestinal epithelial cells contain an appreciable amount of both tsRNAs and rsRNAs in addition to a miRNA peak.

Importantly, SPORTS1.0 found an appreciable portion of rsRNAs annotated in sperm (48.7%), bone marrow cell (11.1%) and intestinal epithelial cell (61.1%) samples that was previously deemed as 'unmatch genome' (UMG) (Figure 2.2a-c upper pie-chart). This is because these newly annotated rsRNAs are derived from rRNA genes (rDNA), which were not assembled and shown in current mouse genome (mm10) (McStay and Grummt, 2008), and thus were discarded before analysis by previous small RNA analyzing pipelines. SPORTS1.0 can now annotate and analyze these rsRNAs, including providing the subtypes of rRNA precursors (5.8S, 18S, 28S, *etc.*) from which they are derived from (Figure 2.3a-c), as well as the loci mapping information (Figure 2.3d-f). Interestingly, the analyses revealed that the specific loci that generate rsRNAs are completely distinct among sperm, bone marrow cell, and intestinal epithelial cell samples (Figure 2.3d-f), suggesting distinct biogenesis and functions of these rsRNAs. Similarly, SPORTS1.0 also revealed tissue-specific landscape of tsRNAs in terms of their relative abundance (Figure 2.2a-c lower pie chart) and the tRNA loci where they are derived from (5' terminus, 3' terminus, 3'CCA end, *etc.*) (Figure 2.4, and Supplementary Figure S2.1-2.3). Since tsRNAs from different loci bear distinct biological functions (Kumar et al., 2016), the tissue-specific

tsRNA composition may represent features that define the unique functions of respective tissue/cell types.

In addition, SPORTS1.0 also revealed distinct mismatch rates among different types of small RNAs (Figure 2.5 and Table S2.2), with tsRNAs showing the highest. The detected mismatch sites represent the modified nucleotides that might have caused mis-incorporation of nucleotides during the RT process. The relatively higher mismatch rate detected in tsRNA sequences is consistent with their highly modified nature. The mismatch sites detected by SPORTS1.0 could provide a potential source for further analyses of RNA modifications within small RNAs.

Finally, SPORTS1.0 can analyze small RNAs of a wide range of species, depending on the availability of their reference genome and small RNA sequences (Figure 2.6 and Table S2.3). The species to be analyzed and their associated small RNA references are subject to update in future versions, or can be customized by the end users.

Conclusion

SPORTS1.0 is an easy-to-use and flexible pipeline for analyzing small RNA-seq data across a wide-range of species. Using mice as example, SPORTS1.0 provides a far more complicated small RNA landscape than having been previously seen, highlighting a tissue-specific dynamic regulation of tsRNAs and rsRNAs. SPORTS1.0 can also predict potential RNA modification sites based on nucleotide mismatches within small RNAs, and show a distinct pattern between different small RNA types. SPORTS1.0 may set the platform for

many future new discoveries in biomedical and evolution research that is related to small RNAs.

Methods

The source code of SPORTS1.0 is written in *Perl* and *R*. The whole package and installation instructions are available on Github (https://github.com/junchaoshi/sports1.0). SPORTS1.0 can apply to a wide-range of species and the annotation references of 68 species are precompiled for downloading (Table S2.3).

The workflow of SPORTS1.0 consists of four main steps, that is, pre-processing, mapping, annotation output, and annotation summary (Figure 2.1). SRA, FASTQ, and FASTA are the acceptable formats for data input. By calling Cutadapt (Martin, 2011) and *Perl* scripts extracted from miRDeep2 (Friedlander et al., 2008), SPORTS1.0 outputs clean reads by removing sequence adapters and discarding sequences with length beyond the defined range, and those with bases other than ATUCG. The clean reads obtained in pre-processing step are sequentially mapped against reference genome, miRBase (Kozomara and Griffiths-Jones, 2014), rRNA database (collected from NCBI), GtRNAdb (Chan and Lowe, 2016), piRNA database (Sai Lakshmi and Agrawal, 2008; Zhang et al., 2014), Ensembl (Yates et al., 2016) and Rfam (Nawrocki et al., 2015), upon users' setting. Small RNA sequences are first annotated by Bowtie (Langmead et al., 2009). Next, a *Perl* script precompiled in SPORTS1.0 is used to identify the locations of tsRNAs regarding whether they are derived from 5' terminus, 3' terminus, or 3'CCA end of tRNAs. Then an *R* script

precompiled in SPORTS1.0 is applied to obtain rsRNA expression level and positional mapping information regarding their respective rRNA precursors (5.8S, 18S, 28S, *etc.*). SPORTS1.0 can also be used to analyze sequence mismatch information if mismatches are allowed during alignment process. Such information can help predict potential modification sites that have caused nucleotide misincorporation during the reverse transcription (RT) process as previously reported (Ryvkin et al., 2013). In the current version, a mismatch site is designated using criteria as previously described (Ryvkin et al., 2013). Binomial distribution is used to address whether the observed mismatch enrichment is significantly higher than the base-calling error. Here, p_{err} is defined as the base-calling error rate, n_{ref} is defined as the number of nucleotides perfectly fitted to the reference sites, n_{mut} is defined as the number of mismatched nucleotides, and n_{tot} is defined as the sum of n_{ref} and n_{mut}. The probability of observing not larger than k perfectly matched nucleotides out of n_{tot} can be calculated as:

$$P(k \leq n_{ref}) = \sum_{i=0}^{k} pbinom(i; n_{tot}, (1 - p_{err}))$$

SPORTS1.0 provides two methods to evaluate n_{mut} number. The first option is to simply calculate n_{mut} as the read number of sequences containing particular mismatches. Since some sequences may align to multiple reference loci, using this method may result in an increased false-positive rate. A second method is thus included, in which read number of sequences from multiple matching loci is uniformly distributed (based on the assumption that each of these multiple sites will equally express RNAs) and consequently generates an adjusted n_{mut}.

SPORTS1.0 summary output includes annotation details for each sequence and length distribution along with other statistics. (See sample output Figure 2.2 and Figure 2.3**,** Table S2.1 and Table S2.2). User guideline is provided online (https://github.com/juncha-oshi/sports1.0).

Figures

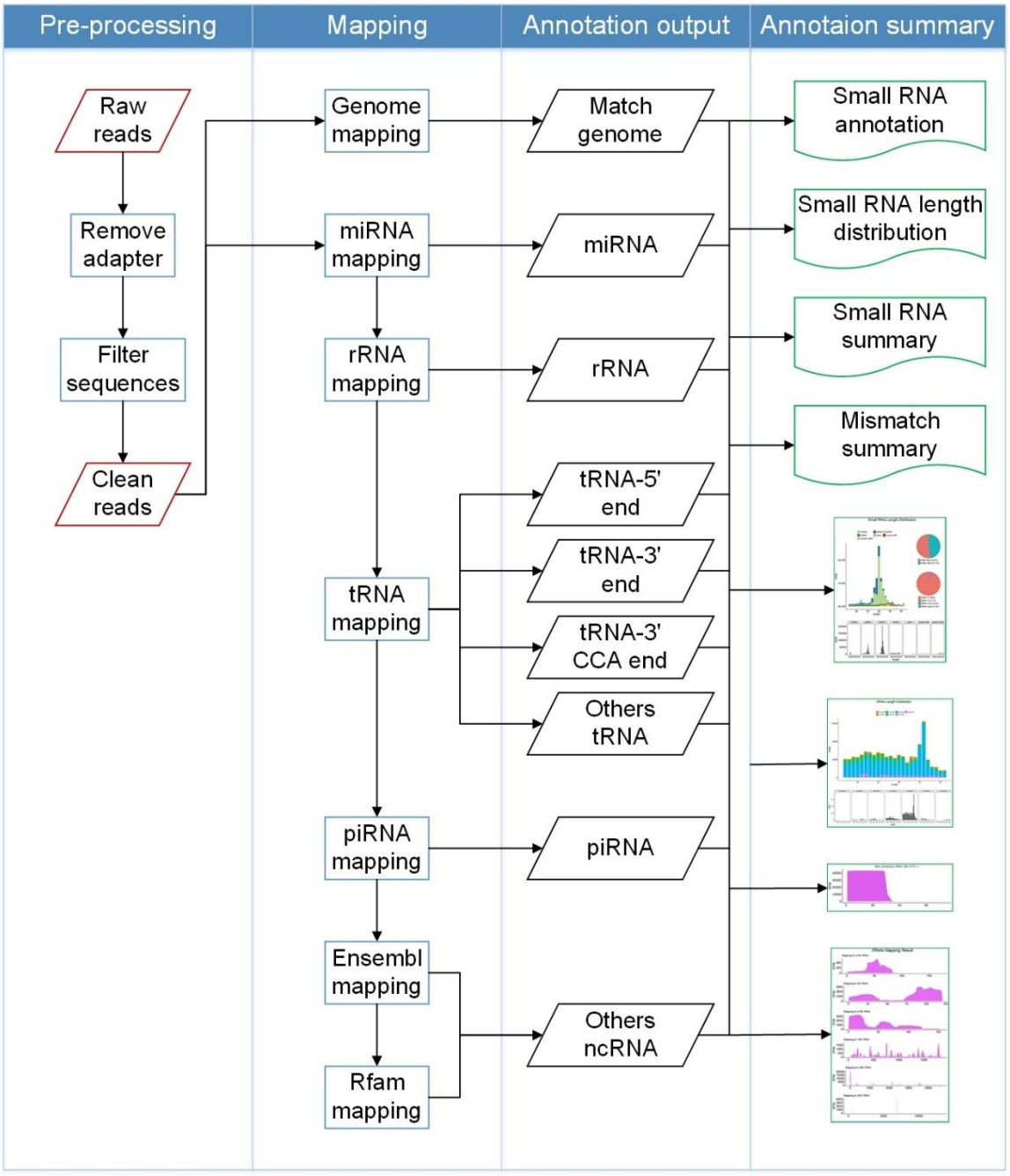

Figure 2.1: Workflow of SPORTS1.0

SPORTS1.0 contains four main steps, that is, pre-processing, mapping, annotation output, and annotation summary, as outlined in the figure.

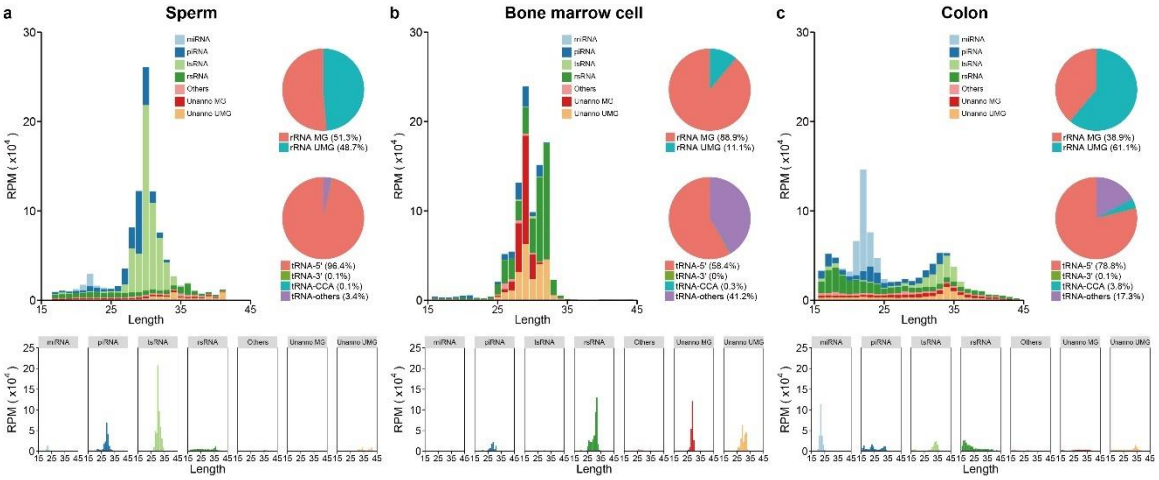

Figure 2.2: Exemplary annotation and profiling of sRNA-seq datasets generated by SPORTS1.0

Categorization and length distribution analysis of different sRNA types in mouse sperm (**a**), bone marrow cell (**b**), and intestinal epithelial cell (**c**) samples. RPM, reads per million clean reads; Unanno: unannotated; MG: match genome; UMG: unmatch genome.

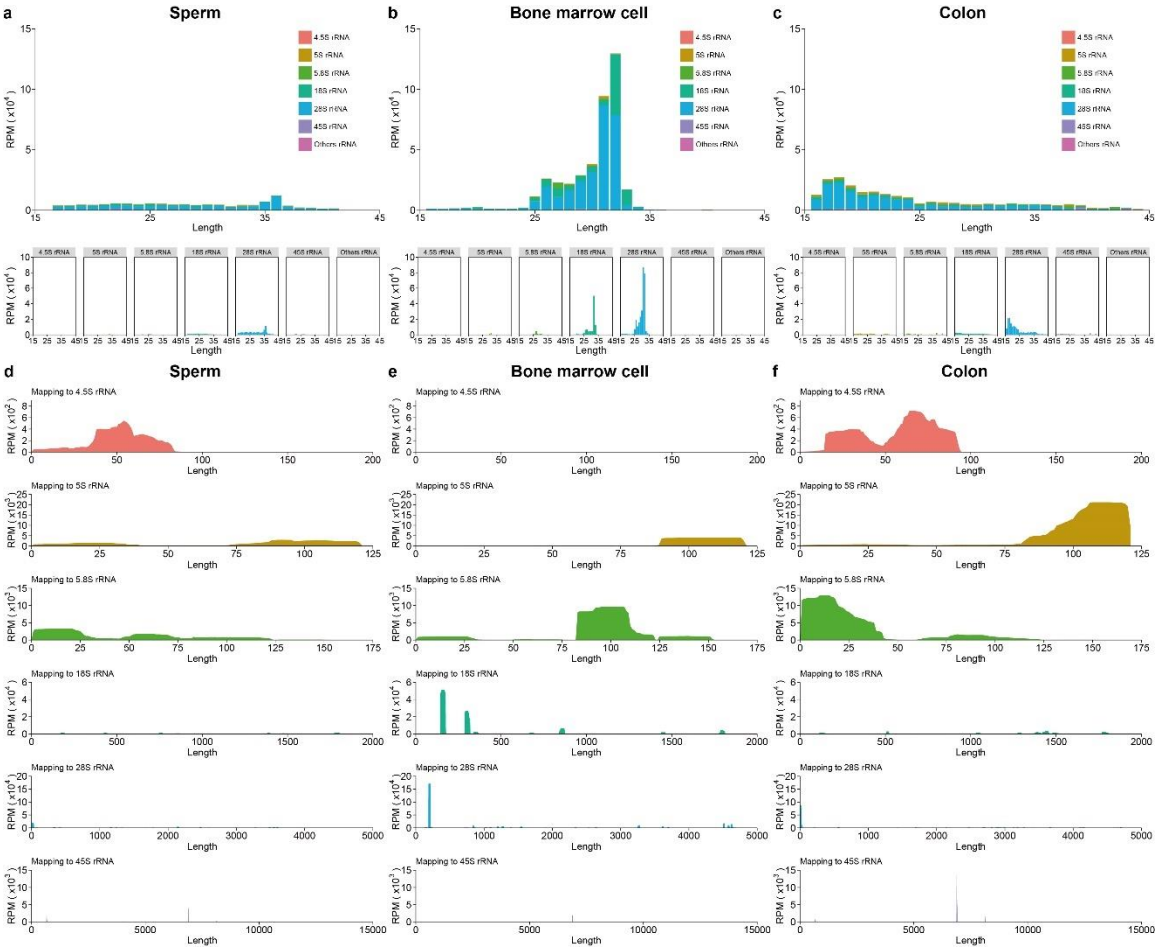

Figure 2.3: Cell-specific rsRNA profiles revealed by SPORTS1.0

Subtypes of rRNA precursors (5.8S, 18S, 28S, *etc.*) for rsRNAs from mouse sperm (**a**), bone marrow cell (**b**), and intestinal epithelial cell (**c**) samples. Comparison of rsRNA-generating loci from different rRNA precursors reveals distinct pattern between sperm (**d**), bone marrow cell (**e**), and intestinal epithelial cell (**f**). RPM, reads per million clean reads.

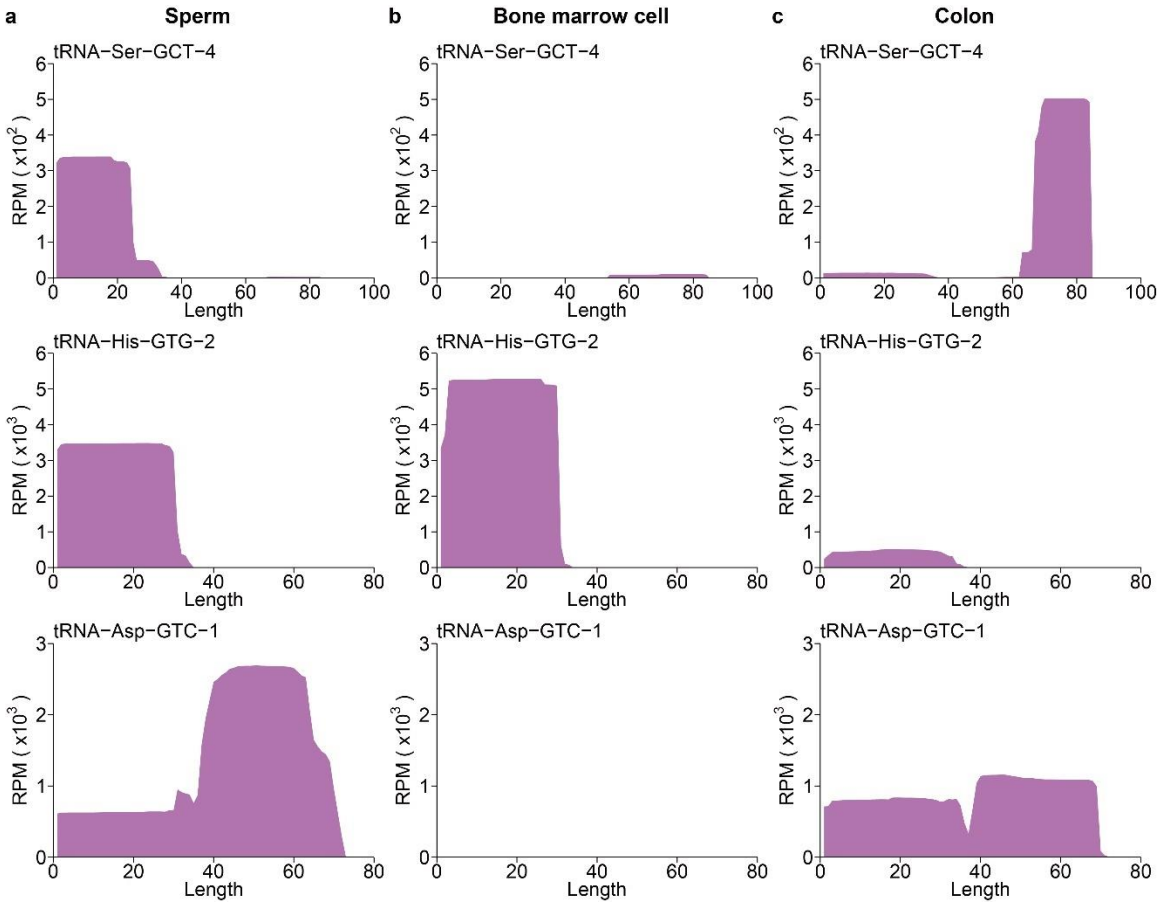

Figure 2.4: Cell-specific tsRNA profiles revealed by SPORTS1.0

Examples of 3 cell-specific tsRNA profiles revealed in mouse sperm (**a**), bone marrow cell (**b**), and intestinal epithelial cell (**c**) samples. Full tsRNA mapping results against tRNA loci are included in Figure S2.1-S2.3 for sperm (Figure S2.1), bone marrow cell (Figure S2.2), and intestinal epithelial cell (Figure S2.3) respectively. RPM, reads per million clean reads.

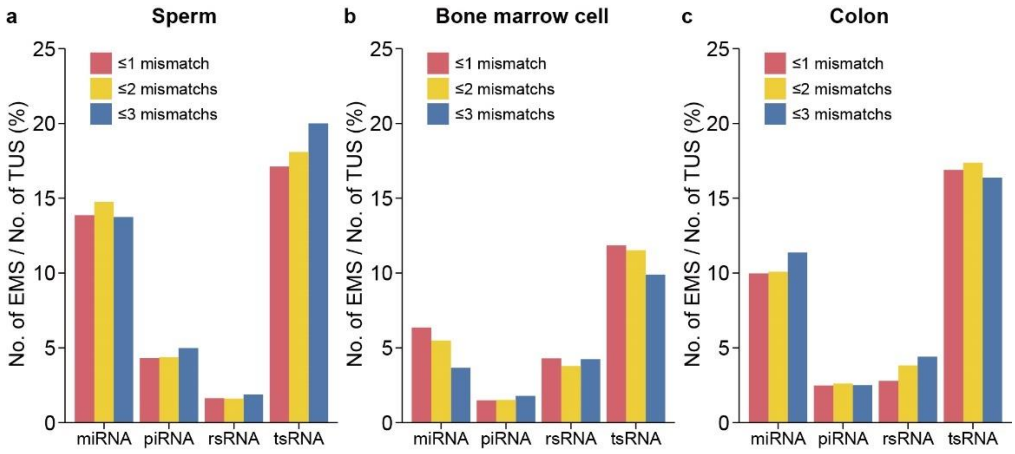

Figure 2.5: sRNA mismatch statistics by SPORTS1.0

The percentage of unique sequences that contain significantly-enriched mismatches out of total number of unique sequences from each subtype of sRNAs (miRNAs, piRNAs, tsRNAs, and rsRNAs) is provided for sperm (**a**), bone marrow cell (**b**), and intestinal epithelial cell (**c**) samples. EMS: enrichment mismatch sequences; TUS: total unique sequences.

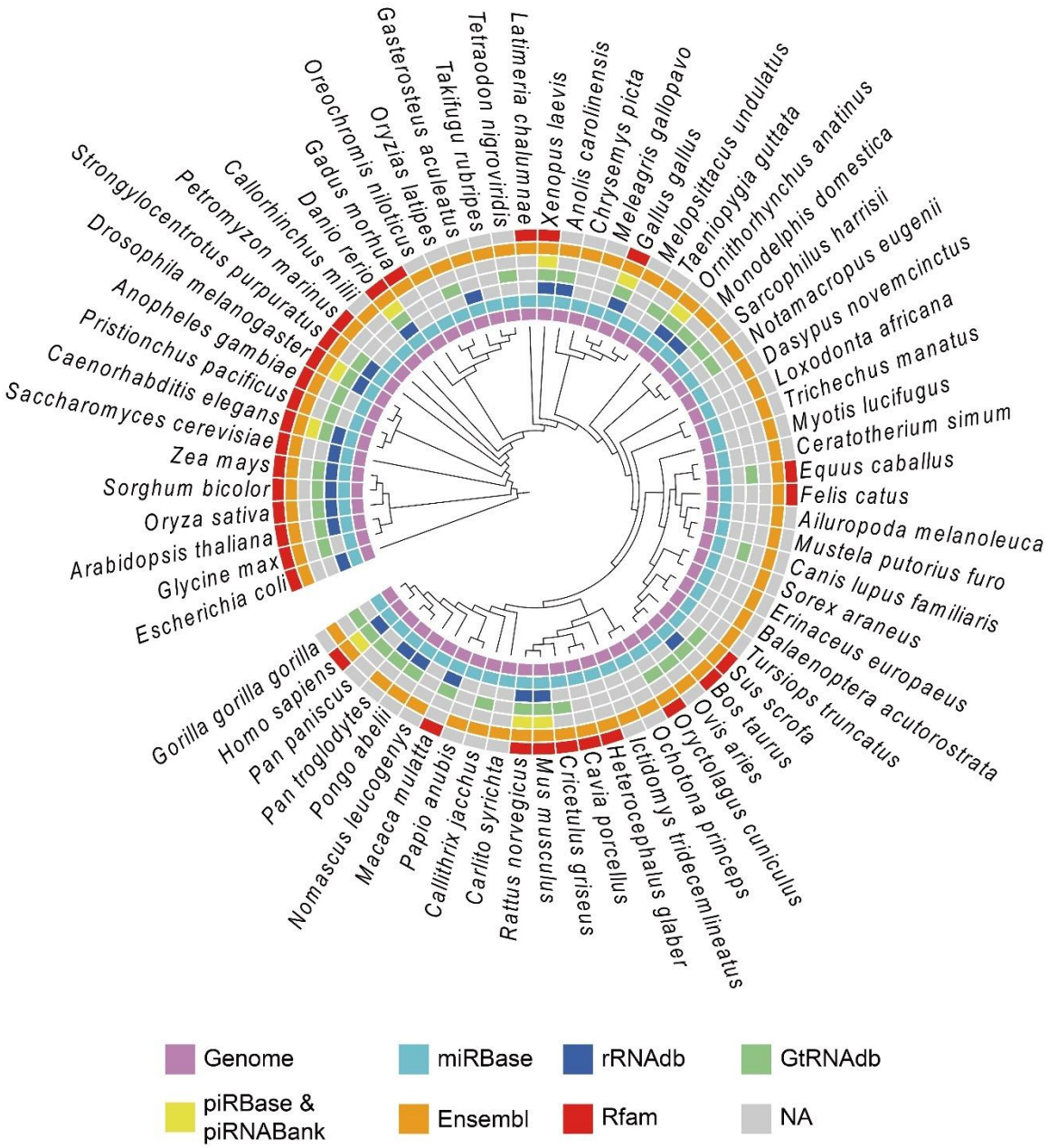

Figure 2.6: Species recompiled for analysis by SPORTS1.0

The 68 species and their respective reference database included in SPORTS1.0 precompiled for analysis.

Supplementary materials

Figure S2.1: The mouse sperm tsRNA mapping results against tRNA loci revealed by SPORTS1.0
Mapping result for each annotated tsRNA was provided.

Figure S2.2: The mouse bone marrow cell tsRNA mapping results against tRNA loci revealed by SPORTS1.0
Mapping result for each annotated tsRNA was provided.

Figure S2.3: The mouse intestinal epithelial cell tsRNA mapping results against tRNA loci revealed by SPORTS1.0
Mapping result for each annotated tsRNA was provided.

Table S2.1: Example output of SPORTS1.0 which includes annotation for each sequence (A), length distribution information (B) and expression level of each annotated category (C) for dataset GSM2304822

Table S2.2: Example output of SPORTS1.0 for sRNA sequence mismatch analysis for dataset GSM2304822 under the alignment criteria of mismatch ≤ 1 (A), ≤ 2 (B), and ≤ 3 (C)

Table S2.3: The list of 68 species and their respective reference database that are precompiled in SPORTS1.0 ready for analyses

References

Anderson, P., and Ivanov, P. (2014). tRNA fragments in human health and disease. FEBS letters *588*, 4297-4304.

Axtell, M.J. (2013). ShortStack: comprehensive annotation and quantification of small RNA genes. RNA *19*, 740-751.

Cech, T.R., and Steitz, J.A. (2014). The noncoding RNA revolution-trashing old rules to forge new ones. Cell *157*, 77-94.

Chan, P.P., and Lowe, T.M. (2016). GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic acids research *44*, D184-189.

Chen, Q., Yan, M., Cao, Z., Li, X., Zhang, Y., Shi, J., Feng, G.H., Peng, H., Zhang, X., Zhang, Y*., et al.* (2016a). Sperm tsRNAs contribute to intergenerational inheritance of an acquired metabolic disorder. Science *351*, 397-400.

Chen, Q., Yan, W., and Duan, E. (2016b). Epigenetic inheritance of acquired traits through sperm RNAs and sperm RNA modifications. Nature reviews Genetics *17*, 733-743.

Chu, C., Yu, L., Wu, B., Ma, L., Gou, L.T., He, M., Guo, Y., Li, Z.T., Gao, W., Shi, H*., et al.* (2017). A sequence of 28S rRNA-derived small RNAs is enriched in mature sperm and various somatic tissues and possibly associates with inflammation. Journal of molecular cell biology *9*, 256-259.

Fasold, M., Langenberger, D., Binder, H., Stadler, P.F., and Hoffmann, S. (2011). DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments. Nucleic acids research *39*, W112-117.

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

Garcia-Silva, M.R., das Neves, R.F., Cabrera-Cabrera, F., Sanguinetti, J., Medeiros, L.C., Robello, C., Naya, H., Fernandez-Calero, T., Souto-Padron, T., de Souza, W*., et al.* (2014). Extracellular vesicles shed by Trypanosoma cruzi are linked to small RNA pathways, life cycle regulation, and susceptibility to infection of mammalian cells. Parasitology research *113*, 285-304.

Gebetsberger, J., Wyss, L., Mleczko, A.M., Reuther, J., and Polacek, N. (2017). A tRNA-derived fragment competes with mRNA for ribosome binding and regulates translation during stress. RNA biology *14*, 1364-1373.

Ivanov, P., Emara, M.M., Villen, J., Gygi, S.P., and Anderson, P. (2011). Angiogenin-induced tRNA fragments inhibit translation initiation. Mol Cell *43*, 613-623.

Kim, H.K., Fuchs, G., Wang, S., Wei, W., Zhang, Y., Park, H., Roy-Chaudhuri, B., Li, P., Xu, J., Chu, K.*, et al.* (2017). A transfer-RNA-derived small RNA regulates ribosome biogenesis. Nature *552*, 57-62.

Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic acids research *42*, D68-73.

Kumar, P., Kuscu, C., and Dutta, A. (2016). Biogenesis and Function of Transfer RNA-Related Fragments (tRFs). Trends in biochemical sciences *41*, 679-689.

Lambertz, U., Oviedo Ovando, M.E., Vasconcelos, E.J., Unrau, P.J., Myler, P.J., and Reiner, N.E. (2015). Small RNAs derived from tRNAs and rRNAs are highly enriched in exosomes from both old and new world Leishmania providing evidence for conserved exosomal RNA Packaging. BMC genomics *16*, 151.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Liao, J.Y., Guo, Y.H., Zheng, L.L., Li, Y., Xu, W.L., Zhang, Y.C., Zhou, H., Lun, Z.R., Ayala, F.J., and Qu, L.H. (2014). Both endo-siRNAs and tRNA-derived small RNAs are involved in the differentiation of primitive eukaryote Giardia lamblia. Proceedings of the National Academy of Sciences of the United States of America *111*, 14159-14164.

Luo, S., He, F., Luo, J., Dou, S., Wang, Y., Guo, A., and Lu, J. (2018). Drosophila tsRNAs preferentially suppress general translation machinery via antisense pairing and participate in cellular starvation response. Nucleic Acids Res.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal *17*, 3.

Martinez, G., Choudury, S.G., and Slotkin, R.K. (2017). tRNA-derived small RNAs target transposable element transcripts. Nucleic acids research *45*, 5142-5152.

McStay, B., and Grummt, I. (2008). The epigenetics of rRNA genes: from molecular to chromosome biology. Annual review of cell and developmental biology *24*, 131-157.

Mohorianu, I., Stocks, M.B., Applegate, C.S., Folkes, L., and Moulton, V. (2017). The UEA Small RNA Workbench: A Suite of Computational Tools for Small RNA Analysis. Methods in molecular biology *1580*, 193-224.

Nawrocki, E.P., Burge, S.W., Bateman, A., Daub, J., Eberhardt, R.Y., Eddy, S.R., Floden, E.W., Gardner, P.P., Jones, T.A., Tate, J.*, et al.* (2015). Rfam 12.0: updates to the RNA families database. Nucleic acids research *43*, D130-137.

Peck, B.C., Mah, A.T., Pitman, W.A., Ding, S., Lund, P.K., and Sethupathy, P. (2017). Functional Transcriptomics in Diverse Intestinal Epithelial Cell Types Reveals Robust MicroRNA Sensitivity in Intestinal Stem Cells to Microbial Status. The Journal of biological chemistry *292*, 2586-2600.

Rueda, A., Barturen, G., Lebron, R., Gomez-Martin, C., Alganza, A., Oliver, J.L., and Hackenberg, M. (2015). sRNAtoolbox: an integrated collection of small RNA research tools. Nucleic acids research *43*, W467-473.

Ryvkin, P., Leung, Y.Y., Silverman, I.M., Childress, M., Valladares, O., Dragomir, I., Gregory, B.D., and Wang, L.S. (2013). HAMR: high-throughput annotation of modified ribonucleotides. RNA *19*, 1684-1692.

Sai Lakshmi, S., and Agrawal, S. (2008). piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. Nucleic acids research *36*, D173-177.

Schimmel, P. (2018). The emerging complexity of the tRNA world: mammalian tRNAs beyond protein synthesis. Nat Rev Mol Cell Biol *19*, 45-58.

Schorn, A.J., Gutbrod, M.J., LeBlanc, C., and Martienssen, R. (2017). LTR-Retrotransposon Control by tRNA-Derived Small RNAs. Cell *170*, 61-71 e11.

Selitsky, S.R., and Sethupathy, P. (2015). tDRmapper: challenges and solutions to mapping, naming, and quantifying tRNA-derived RNAs from human small RNA-sequencing data. BMC Bioinformatics *16*, 354.

Szempruch, A.J., Dennison, L., Kieft, R., Harrington, J.M., and Hajduk, S.L. (2016). Sending a message: extracellular vesicles of pathogenic protozoan parasites. Nature reviews Microbiology *14*, 669-675.

Thompson, A., Zielezinski, A., Plewka, P., Szymanski, M., Nuc, P., Szweykowska-Kulinska, Z., Jarmolowski, A., and Karlowski, W.M. (2018). tRex: A Web Portal for Exploration of tRNA-Derived Fragments in Arabidopsis thaliana. Plant Cell Physiol *59*, e1.

Tuorto, F., Herbst, F., Alerasool, N., Bender, S., Popp, O., Federico, G., Reitter, S., Liebers, R., Stoecklin, G., Grone, H.J.*, et al.* (2015). The tRNA methyltransferase Dnmt2 is required for accurate polypeptide synthesis during haematopoiesis. The EMBO journal *34*, 2350-2362.

Wei, H., Zhou, B., Zhang, F., Tu, Y., Hu, Y., Zhang, B., and Zhai, Q. (2013). Profiling and identification of small rDNA-derived RNAs and their potential biological functions. PloS one *8*, e56842.

Wu, X., Kim, T.K., Baxter, D., Scherler, K., Gordon, A., Fong, O., Etheridge, A., Galas, D.J., and Wang, K. (2017). sRNAnalyzer-a flexible and customizable small RNA sequencing data analysis pipeline. Nucleic acids research *45*, 12140-12151.

Yang, Q., Lin, J., Liu, M., Li, R., Tian, B., Zhang, X., Xu, B., Liu, M., Zhang, X., Li, Y.*, et al.* (2016). Highly sensitive sequencing reveals dynamic modifications and activities of small RNAs in mouse oocytes and early embryos. Science advances *2*, e1501482.

Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L.*, et al.* (2016). Ensembl 2016. Nucleic acids research *44*, D710-716.

Zhang, P., Si, X., Skogerbo, G., Wang, J., Cui, D., Li, Y., Sun, X., Liu, L., Sun, B., Chen, R.*, et al.* (2014). piRBase: a web resource assisting piRNA functional study. Database (Oxford) *2014*, bau110.

Zhang, Y., Zhang, X., Shi, J., Tuorto, F., Li, X., Liu, Y., Liebers, R., Zhang, L., Qu, Y., Qian, J.*, et al.* (2018). Dnmt2 mediates intergenerational transmission of paternally acquired metabolic disorders through sperm small non-coding RNAs. Nature cell biology.

Zheng, L.L., Xu, W.L., Liu, S., Sun, W.J., Li, J.H., Wu, J., Yang, J.H., and Qu, L.H. (2016). tRF2Cancer: A web server to detect tRNA-derived small RNA fragments (tRFs) and their expression in multiple cancers. Nucleic acids research *44*, W185-193.

Chapter 3: Improved small RNA-seq method by overcoming RNA modifications

Abstract

Although high-throughput RNA sequencing (RNA-seq) has greatly advanced small RNA discovery, the currently widely used complementary DNA library construction protocol generates biased sequencing results. This is partially due to RNA modifications that interfere with adapter ligation and reverse transcription processes, which prevent the detection of small RNAs bearing these modifications. PANDORA-seq (Panoramic RNA Display by Overcoming RNA Modification Aborted Sequencing) is presented in this chapter, employing a combinatorial enzymatic treatment to remove key RNA modifications that block adapter ligation and reverse transcription. PANDORA-seq identified abundant modified small RNAs — mostly transfer RNA-derived small RNAs (tsRNAs) and ribosomal RNA-derived small RNAs (rsRNAs) — that were previously undetected, exhibiting tissue-specific expression across mouse brain, liver, spleen and sperm, as well as cell-specific expression across embryonic stem cells (ESCs) and HeLa cells. Using PANDORA-seq, unprecedented landscapes of microRNA, tsRNA and rsRNA dynamics are revealed during the generation of induced pluripotent stem cells. Importantly, tsRNAs and rsRNAs that are downregulated during somatic cell reprogramming impact cellular translation in ESCs, suggest a role in lineage differentiation.

Introduction

High-throughput RNA-seq has substantially facilitated the discovery of functional small RNAs over the last decade. Traditional construction of cDNA libraries for deep sequencing of small RNAs is based on adapter ligation to the 3' and 5' termini, which is followed by reverse transcription. This protocol has been proven efficient for many small RNA species that have a 5'-P and 3'-OH (**Figure 3.1a**), such as miRNAs (Bartel, 2018). However, this protocol has inherent problems when encountering small RNAs bearing specific RNA modifications, including 3' terminal modifications such as 3'-P and 2'3'-CP that block the adapter ligation process (Honda et al., 2015), and RNA methylations such as m^1A, m^3C, m^1G and $m^2{}_2G$ that interfere with reverse transcription (Cozen et al., 2015; Dai et al., 2017; Zheng et al., 2015). small RNAs bearing one or more of these modifications are often inefficiently and incompletely converted into cDNAs, leading to challenges with their detection and quantitation by deep sequencing. This problem is particularly severe for highly modified small RNAs such as tsRNAs and rsRNAs (Chen et al., 2016b; Zhang et al., 2016), because their precursors (tRNAs and rRNAs) are known to harbor a diversity of RNA modifications (Phizicky and Hopper, 2010; Schimmel, 2018; Sergiev et al., 2018) and because 3'-P or 2'3'-CP are commonly implemented during the biogenesis of tsRNAs and rsRNAs (Akiyama et al., 2019; Honda et al., 2015; Shigematsu et al., 2018).

To discover the modified small RNAs that escaped traditional RNA-seq, enzymatic treatment protocols have been developed to address specific RNA modifications. For example, treatment with the dealkylating enzyme AlkB and its mutant forms have been introduced to demethylate RNA modifications (for example, m^1G, m^1A, m^3C, and $m^2{}_2G$) to

enable reverse transcription (**Figure 3.1a**) (Cozen et al., 2015; Dai et al., 2017; Zheng et al., 2015); and T4PNK has been used to convert the 3' terminal 3-'P or 2'3'-CP into 3'-OH and to add a 5'-P, thus facilitating adapter ligation for RNA-seq of small (Akat et al., 2019) and large (Giraldez et al., 2019) RNAs (**Figure 3.1a**). While these methods can reveal the sequence of specific small RNAs bearing targeted modifications, each of these treatments alone cannot capture modified small RNAs beyond their individual enzymatic capacity and therefore are not able to reveal a full small RNA spectrum. In addition, the bioinformatics analyses of small RNAs are currently evolving from previously focusing on miRNAs (Bartel, 2018) to other potentially important small RNA species, including the emerging tsRNAs (Schimmel, 2018; Shi et al., 2019; Su et al., 2020; Zhang et al., 2016) and rsRNAs (Gu et al., 2020; Natt et al., 2019; Zhang et al., 2018b) that can now be systematically analyzed along with miRNAs and piRNAs using the software described in Chapter 2.

To test whether a combinatorial use of enzymatic treatments can overcome both adapter ligation and reverse transcription obstacles and reveal a more in depth composition of small RNAs, PANDORA-seq is developed (**Figure 3.1a,b**). This method, coupled with the improved small RNA bioinformatics pipelines (see Methods), is based on consecutive enzymatic treatments of the small RNA fraction (15- to 50-nucleotide) with T4PNK and AlkB to provide stepwise optimization that improves both adapter ligation and reverse transcription during cDNA library construction, respectively (**Figure 3.1a**). Systematic comparison to existing small RNA-seq methods demonstrated that PANDORA-Seq outperformed both traditional sequencing and individual AlkB or T4PNK treatments by more

extensively and accurately uncovering previously unidentified modified small RNAs in a wide range of mouse and human tissues and cells. PANDORA-seq also revealed unprecedented miRNA, tsRNAs and rsRNAs dynamics during the reprogramming of somatic cells to induced iPSCs, guiding us to probe their function during ESC differentiation. Together, PANDORA-seq and the small RNA repertoire across different lineages open the avenue for future exploration of the hidden layer of functional small RNAs in other biological and disease conditions.

Results

Enzyme validation and protocol optimization

PANDORA-seq is developed by leveraging a combination of two enzymatic treatments that can overcome distinct RNA modifications that either prevent reverse transcription (by AlkB treatment) or adapter ligation (by T4PNK treatment) (**Figure 3.1a**). To this end, AlkB enzyme was generated using a previously reported plasmid with codon optimization (Trewick et al., 2002). Then, its enzymatic efficacy was tested in removing RNA methylations using a high-throughput RNA modification quantitation platform based on liquid chromatography-tandem mass spectrometry (LC-MS/MS) that was developed previously (Chen et al., 2016a; Zhang et al., 2018b). The AlkB efficiency was tested by treating the 15- to 50 -nucleotide RNA fraction extracted from mouse liver, followed by LC-MS/MS examination. As a result, the AlkB treatment efficiently removed m^1A and m^3C and also significantly decreased m^1G and $m^2{}_2G$ to ~20% of their original level (**Figure 3.1c**). The AlkB plasmid (see Methods) has sequence differences at the amino terminus compared to

a previously reported AlkB (Zheng et al., 2015), but generates similar efficacy in removing m^1A, m^3C and m^1G, demonstrating expected enzymatic activity.

The enzymatic efficacy of T4PNK in converting 3'-P and 2'3'-CP into 3'-OH was also tested in regard to its impact in facilitating RNA adapter ligation. As shown in **Figure 3.1d**, synthetic tsRNAs with 3'-P cannot be ligated using T4 ligase, while T4PNK treatment of these 3'-P tsRNAs enabled a high ligation efficiency similar to that of the synthetic 3'-OH tsRNA (**Figure 3.1d**). The effect of T4PNK was further tested on the 25- to 50-nucleotide RNA fraction recovered from mouse tissues, which is expected to contain 5' tsRNAs bearing a 2'3'-CP end such as those generated by angiogenin-mediated cleavage of tRNA (Honda et al., 2015). As an example, using RNAs from the mouse spleen (**Figure 3.1d**), it is found that while T4 ligase alone worked poorly on the untreated samples, T4PNK treatment substantially increased the overall adapter ligation efficiency (**Figure 3.1e**), demonstrating T4PNK's effect in improving adapter ligation for small RNA cDNA library construction.

Notably, although AlkB and T4PNK are not supposed to have ribonuclease activity, and despite the addition of RNase inhibitor during the enzymatic treatment, it was noticed that when treating total RNA from tissues or cells, AlkB can cause detectable RNA degradation, as revealed by increased RNA smear in the small RNA region and increased level of tsRNAs and rsRNAs detected by northern blots (**Figure 3.1f,g**). This phenomenon might be due to the demethylation effect of AlkB on tRNAs and rRNAs, which results in altered RNA structure and increased fragmentation of tRNAs and rRNAs (Pan, 2018). This effect will generate additional tsRNAs/rsRNA in the small RNA library as an artifact, which has

not been addressed in previous publications using AlkB treatment (Cozen et al., 2015; Zheng et al., 2015). To circumvent this problem, the protocol was optimized by applying a pre-size-selection procedure to first obtain the 15- to 50-nucleotide small RNA fraction from the total RNA and then performing enzymatic treatments on this 15- to 50-nucleotide RNA fraction. This procedure pre-eliminated the sources (that is, tRNAs and rRNAs) that generate artificial tsRNAs and rsRNAs from degradation and, importantly, the treatment of AlkB and/or T4PNK in the 15- to 50-nucleotide fraction did not cause further degradation of tsRNAs and rsRNAs (**Figure 3.1h,i**).

The potential impact of treatment order of AlkB and T4PNK was also tested by comparing the RNA-seq results between the treatment order of AlkB first and T4PNK second (AlkB+T4PN) versus T4PNK first and AlkB second (T4PNK+AlkB) in HeLa cells. The results showed a high degree of correlation (ρ=0.995; **Figure 3.1j**) between both treatments orders, indicating that the order of treatment does not result in major differences. With the enzymatic validation and protocol optimization above, PANDORA-seq was established by first size-selecting the 15- to 50-nucleotide RNA fraction, followed by enzymatic treatment in the order T4PNK+AlkB, as applied to all other tissue or cell samples.

A tsRNA- and rsRNA-enriched small RNA landscape

The outcome of PANDORA-seq was assessed in a variety of mouse and human tissue and cell types, including mouse brain, liver, spleen and mature sperm (and sperm head), mouse ESCs (mESCs); human ESCs in primed and naïve (Guo et al., 2017)) states, HeLa cells, and cells during the reprogramming of mouse embryonic fibroblasts (MEFs) into iPSCs (Cheloufi et al., 2015). Three biological repeats were included for most tissues or cell types,

except two biological repeats were included for mouse spleen and naïve hESC sample. The read summaries and differentially expressed small RNAs between individual protocols are presented in **Table S3.1**. small RNA sequence distribution, as exemplified in mouse brain, liver, mature sperm, mESCs, and Hela cells (**Figure 3.2a-e**) (see **Figure 3.3** for other tissue and cell types) reveals that while miRNAs are the dominant small RNAs detected by traditional RNA-seq (except in mature sperm and sperm head as was previously known (Peng et al., 2012)), the treatment with AlkB and T4PNK substantially increased the reads of tsRNAs and rsRNAs in distinct patterns (**Figure 3.2a-e**), and PANDORA-seq showed an overall enhanced effect compared to each treatment alone. Due to the abundantly increased rsRNA reads after T4PNK or PANDORA-seq treatment, which consumed the relative reads of tsRNAs and miRNAs (**Figure 3.2a-e**), the relative tsRNA/miRNA ratio was further separately analyzed under different treatment protocols (**Figure 3.2f, Figure 3.3g-l**), which showed clearer effects of each treatment on tsRNA discovery. Notably, mature sperm heads contained the highest concentration of tsRNAs and showed the highest tsRNA/miRNA ratio across all samples examined under PANDORA-seq (**Figure 3.2c,f**).

The abundant expression of rsRNAs revealed by PANDORA-seq is surprising, yet the results represent the *in vivo* situation. The relative expression levels of representative miRNA, tsRNA and rsRNA were further validated by northern blots in mouse brain, liver and in HeLa cells (**Figure 3.2g-i**). The abundant expression of tsRNAs and rsRNAs has also been previously detected in mouse sperm by northern blots (Chu et al., 2017; Zhang et al., 2018b). Notably, certain miRNAs, such as miR-122, remain highly expressed in the liver compared to tsRNAs and rsRNAs (**Figure 3.2h**), resonating with their crucial role in

liver function (Valdmanis et al., 2016). A further examination of the relative efficiency across different northern blot probes (that is, rsRNA-28S-1, 5' tsRNA^{Glu}, let-7i, mir-122, and mir-21) (**Figure 3.4**) enabled better semi-quantitative analysis of the relative level of the examined small RNAs in the tissues and cells by northern blot signal (**Figure 3.2g-i**), again supporting the abundant existence of rsRNAs and tsRNAs compared with miRNAs, consistent with the result of PANDORA-seq.

Notably, the bioinformatics pipeline discovered appreciable piRNA reads from non-germ cell mouse samples (**Figure 3.2a-e** and **Figure 3.3a-f**). Since the annotation of piRNAs was based on the two existing publicly available piRNA databases (Sai Lakshmi and Agrawal, 2008; Zhang et al., 2014), but not the PIWI pulldown experiments of each tissue, the accuracy of the piRNA annotation largely depends on the quality of the databases. In fact, cautions are exercised in the analyses regarding the true identity of these piRNAs in mice: if one to three mismatches are allowed, the annotation rate of piRNA (but not other types of small RNAs) dramatically decreases and many piRNAs are annotated in other small RNA categories (**Figure 3.5**), which puts the identity of these piRNAs in doubt. Further analyses of piRNAs were avoided in the following context but focused on the other categories of small RNAs that could be reliably annotated (for example, miRNAs, tsRNAs and rsRNAs).

Distinct methylation pattern of miRNAs, tsRNAs and rsRNAs

Next, the response of miRNAs, tsRNAs and rsRNAs was separately analyzed upon T4PNK, AlkB, and PANDORA-seq (T4PNK+AlkB) treatments. Using mESCs (**Figure 3.6a-m**) as an example (see **Figure 3.7** for other tissue and cell types), miRNA profiles were generally

not dramatically changed after the enzymatic treatments, as shown in the correlation for traditional vs AlkB (**Figure 3.6a**), traditional vs T4PNK **(Figure 3.6b**), and traditional vs PANDORA-seq (T4PNK+AlkB) (**Figure 3.6c**). This is consistent with the well-defined biogenesis pathways of miRNAs, which result in 5'-P and 3'-OH termini, and the fact that miRNA populations are less modified than the tsRNA and rsRNA populations (Zhang et al., 2018b).

Compared to miRNAs, tsRNAs are sensitive to both AlkB and T4PNK, as demonstrated by the correlation pattern, with a substantial number of tsRNAs showing upregulation after each treatment alone or after PANDORA-seq treatment (T4PNK+AlkB) in mESCs (**Figure 3.6a-c)** and similarly in other tissue and cell types (**Figure 3.7**). These results resonate with the fact that some reverse transcription-blocking RNA modifications in tsRNAs can be removed by AlkB; and that the 3'-P and 2'3'-CP termini of tsRNAs can be converted to 3'-OH by T4PNK to improve adapter ligation efficiency.

Notably, compared to the effects of AlkB and T4PNK treatment alone, a combinatorial effect of PANDORA-seq is observed when examining the relative expression of tsRNAs of different origins (5' tsRNA, 3' tsRNA, 3' tsRNA with a CCA end, and internal tsRNAs) in mESCs (**Figure 3.6d,** see **Figure 3.8** for other tissue and cell types). The overall mapping of all tsRNAs on a tRNA length scale revealed the preferential loci from which tsRNAs are derived from the full-length tRNA under different protocols (**Figure 3.6e**). In addition to the overall mapping analyses, individual tsRNAs have distinct responses, as exemplified in **f** (data on tsRNA mapping to each kind of tRNA in all tissue and cell types are provided in **Figure S3.1**). In contrast. mitochondrial tRNAs on the other hand showed

an overall different tsRNA production pattern compared with that of genomic tsRNAs (**Figure 3.6d,e**), possibly because mitochondrial tRNAs bear different RNA modifications and structures (Suzuki and Suzuki, 2014) that result in a differential cleavage pattern (see **Figure 3.8** and **3.9** for the tsRNA mapping data in other tissue and cell types).

Compared with tsRNAs, rsRNAs are less sensitive to AlkB treatment but show a dramatic increase after T4PNK treatment (**Figure 3.6g-i**), suggesting that many rsRNAs contain either a 3'-P or 2'3'-CP that can be converted to 3'-OH, or a 5'-OH that can be converted to 5'-P. Detailed mapping data of rsRNAs showed the specific loci of different ribosomal RNAs from which they are derived (as exemplified by 5S, 5.8S 18S and 28S rRNAs in **Figure 3.6j-m**, data for 45S rRNA and mitochondria-encoded 12S, 16S rRNAs are provided in **Figure S3.2**), and the different effects between protocols can be visualized. Notably, PANDORA-seq further increased rsRNA detection compared to T4PNK alone, demonstrating that these small RNAs harbor both adapter ligation-preventing terminal modifications and reverse transcription-blocking internal modifications. The rsRNA mapping data for other tissue and cell types are provided in **Figure S3.2**.

Interestingly, while the majority of miRNAs (annotated in miRBase) are not responsive to AlkB and T4PNK treatment, a small portion of them indeed showed a significant upregulation in their relative expression levels following the PANDORA-seq protocol. Further analyses revealed that most of these distinct miRNA sequences can in fact be annotated to other small RNA categories, with the majority of them annotated to rsRNAs in both mESCs and hESCs (**Figure 3.6n,o**). Similar observations are also shown in other tissue and cell types (**Figure 3.10** and **Table S3.2**), suggesting that these miRNAs are distinct

from canonical miRNAs and await further evaluation in miRBase. Based on this information, the workflow of SPORTS1.0 described in charter 2 was revised (**Figure 3.11**) and the annotation results in this chapter were generated by the upgraded version SPORTS1.1.

Tissue- and cell-specific tsRNA and rsRNA patterns

Using PANDORA-seq, the expression patterns of tsRNAs and rsRNAs were further analyzed across six tissue and cell types in mice (brain, liver, spleen, mESCs, sperm and sperm heads) (**Figure 3.12a-d**) and three cell types in humans (HeLa cell, primed hESCs and naïve hESCs) (**Figure 3.12e-j**). The radar plot of each tissue or cell type shows the relative response of tsRNA subcategory to AlkB, T4PNK and PANDORA-seq treatment compared with the traditional protocol (the levels of tsRNA were normalized to total miRNA reads), revealing tissue- and cell-specific patterns (**Figure 3.12a,e**). Notably, PANDORA-seq increased the relative level of a majority of tsRNA subcategories to a greater extent compared with AlkB or T4PNK treatment alone (**Figure 3.12a,e**). The heatmaps of genomic and mitochondrial tsRNAs further show the relative amount of each tsRNA subcategory (normalized with total miRNA reads) across mouse (**Figure 3.12b)** and human (**Figure 3.12f**) tissue and cell types.

The mapping and overall comparative expression patterns of rsRNAs across different protocols and tissue or cell types are summarized according to their origin from individual ribosomal RNAs (that is, 5S, 5.8S,18S, 28S, 45S and mitochondria-encoded 12S and 16S rRNAs) in **Figure 3.13** and **Figure 3.2**. Overall coverage similarity comparison matrices (**Figure 3.12c,g**) and detailed rsRNA mapping data (**Figure 3.12d,h**) are

presented using rsRNAs from 28S and 18S rRNA as examples, from which the distinct expression patterns of rsRNAs across tissue and cell types can be visualized and compared.

In addition to tsRNAs and rsRNAs, human and mouse samples also contain small RNAs derived from YRNAs, which are defined as YRNA-derived small RNAs (ysRNAs) (**Figure S3.3**). ysRNAs have been reported to be involved in immunological processes (Hizir et al., 2017) and could be harnessed as disease markers along with tsRNAs/rsRNAs (Gu et al., 2020). PANDORA-seq reveals that ysRNAs are differentially expressed between HeLa cells, primed hESCs and naïve hESCs (**Figure 3.12i,j** and **Figure S3.3**) and their biogenesis and functions await further explorations.

Small RNA dynamics during iPSC induction

Finally, PANDORA-seq was used to explore the small RNA dynamics during transcription factor-mediated somatic cell reprogramming to pluripotency. The levels of miRNAs, tsRNAs and rsRNAs showed dynamic changes during the reprogramming process: MEFs (Day0), reprogramming intermediates (Day3) and stably derived iPSCs (**Figure 3.14a**). An overall decrease in the miRNA level during reprogramming was evident by PANDORA-seq (**Figure 3.14b**). The overall tsRNA/rsRNA profiles between different protocols and across different stages were summarized for tsRNAs and rsRNAs in **Figure 3.14c,g** and **Figure 3.13**. Heatmap analyses (**Figure 3.14d**) and exemplary tsRNA loci mapping (**Figure 3.16e,f**) showed a dynamic tsRNA expression pattern during the reprogramming process by PANDORA-seq. The rsRNA comparison matrix (**Figure 3.14g**) showed that PANDORA-seq reveals more dynamic changes in expression patterns across different stages compared with traditional RNA-seq. Representative rsRNAs from 5S, 18S and 28S

rRNAs (**Figure 3.14h-j**) showed statistically significant changes in expression levels during the reprogramming process. Selected individual miRNAs, tsRNAs and rsRNAs between MEFs and iPSCs were validated by northern blots (**Figure 3.14k-r**), with overall consistency with PANDORA-seq results (**Figure 3.14k-r**), but less consistency with the results of traditional RNA-seq (**Table S3.3**).

The results that many miRNAs and tsRNAs are downregulated during iPSC reprogramming are consistent with previous reports that decreased levels of miRNAs (Viswanathan et al., 2008) and tsRNAs (Krishna et al., 2019) are associated with mESC pluripotency (some tsRNAs showing upregulation by PANDORA-seq are actually expressed at low level below the detection limit by northern blots). The changes of rsRNAs during reprogramming are more dynamic, depending on the loci from which they are derived from (**Figure 3.14h-j,q,r**).

tsRNAs and rsRNAs impact mESC differentiation

The tsRNAs (Ala, Arg, Glu, His and Lys) and rsRNA-28S-1 showing downregulation during iPSC reprogramming by PANDORA-seq were further examined by northern blots during mESC differentiation in an embryoid bodies formation assay. The northern blot results showed a trend of upregulation for all these tsRNA and rsRNA candidates during embryoid body differentiation on day 6 and 10 (**Figure 3.15**), suggesting that these tsRNAs and rsRNAs may play a functional role in mESC differentiation. To test this hypothesis, different types of tsRNA and rsRNA (that is, rsRNA-28S-1, individual 5' $tsRNA^{Ala}$, 3' $tsRNA^{Arg}$, 5' $tsRNA^{Glu}$, 5' $tsRNA^{His}$, 3' $tsRNA^{Lys}$, and a pool of the five abovementioned tsRNAs) were transfected into mESCs followed by embryoid body formation. Then transcriptomic

RNA-seq/bioinformatics analyses of embryoid bodies were performed at Days 1, 3 and 6 after transfection (**Figure 3.16a**), during which no significant morphological changes were detected during embryoid body formation after any of the tsRNA or rsRNA transfections.

Gene ontology analyses on the altered mRNAs (**Table S3.4**) suggest that the transfection of rsRNA-28S-1 or the tsRNA pool significantly promoted the lineage differentiation in day 6 embryoid bodies, including the promotion of endoderm (for example, inner ear development), mesoderm (for example, urogenital and muscle/heart development) and ectoderm (for example neurological development) (**Figure 3.16b**). While different effects of individual tsRNA transfections were observed, transfection of tsRNA pool showed an overall combinatory effect (**Figure 3.16b**). It is interesting that the transfection of rsRNA-28S-1 or the tsRNA pool had a similar overall effect in promoting lineage differentiation (**Figure 3.16b**) despite their distinct sequences. This could be due to the fact that both rsRNA-28S-1 and tsRNA pool have a strong effect in downregulating the mitochondria oxidative phosphorylation and translation/ribosome pathways (**Figure 3.16c**), as the alteration of mitochondria oxidative phosphorylation can act as an overarching factor to change cell metabolism and affect cell lineage progression (Zhang et al., 2018a). Moreover, the promotion of embryonic forebrain development has been shown to be associated with downregulation of ribosome/translation pathways (Chau et al., 2018), consistent with the observation. Individual genes involved in the highlighted pathways in **Figure 3.16b,c** were further shown in heatmaps and the overlapping changes shown between each transfection (**Figure 3.16d,e** and **Figure 3.17a-d**), further supporting the discoveries at the pathway level and providing a gene resource for future in-depth investigations.

Next a day 1 to day 3 to day 6 developmental view of the overall trend of selected key pathway was generated shown in **Fig6b,c**, in which an algorithm was applied to compute gene set scores using the rank-weighted gene expression of individual samples (Yang et al., 2012), with a higher level representing an overall upregulation of a specific Gene Ontology biological process (GOBP) term (**Figure 3.16f** and **Table S3.5**). The results recapitulate the conclusion that the main lineage effects appear at day 6 while the effects are minimal at day 1 (**Figure 3.16f**). Indeed, the transcriptomic changes on day1 (from any of the tsRNA or rsRNA transfection groups) were mostly sporadic and the altered genes did not group into clusters into clusters in Gene Ontology analyses under the same criteria that used for the differentially expressed genes on days 3 and 6 (**Figure 3.16b,c**, **Figure 3.17e,f**). This suggests that tsRNA and rsRNA transfection do not directly disrupt mRNAs, but may regulate translational processes (Shi et al., 2019). The embryoid body differentiation effect observed on day 6 would represent the outcome of a cascade reaction during early translational programming (Genuth and Barna, 2018) that results in stem cell differentiation (Li and Wang, 2020). Using a translational assay measuring the nascent protein synthesis, it was indeed found that the transfection of rsRNA-28S-1 or the tsRNA pool in mESCs reduced the translation rate (**Figure 3.16g,h**). Although the exogenous transfection of tsRNAs and rsRNAs may not precisely represent the relative tsRNA and rsRNA quantity and modification status *in vivo*, these proof-of-principle functional data may open future opportunity to investigate how such translational programming may affect cell differentiation.

Conclusion

PANDORA-seq is developed by improving both adapter ligation and reverse transcription during RNA-seq library construction, and it shows major advantages: **(1)** The single and combinational use of T4PNK and AlkB treatments not only enabled the theoretical and practical identification of previously undetected modified small RNAs, but also delineated the small RNAs that respond to different treatments, from which their RNA modification conditions can be partially deduced. **(2)** Importantly, the northern blot-validated PANDORA-seq results in different tissue and cell types (**Figure 3.2)** and during reprogramming (**Figure 3.14**) allowed for discovery of an unprecedented landscape that miRNAs are in fact not the majority small RNA population in many tissue and cell types. **(3)** The pre-size-selection procedure corrected the false positive detection of tsRNAs and rsRNAs that can be induced by AlkB treatment on total RNAs (**Figure 3.1f-i**), which has been previously been overlooked (Cozen et al., 2015). **(4)** The upgraded small RNA analysis pipeline based on SPORTS1.1 (see method) provided direct mapping visualization of tsRNAs and rsRNAs in regard to their sources (tRNAs and rRNAs) and can easily be used for comparison between different protocols and samples, which may provide the benchmark for future small RNA analyses. **(5)** Results from PANDORA-seq also provided a knowledge basis for updating the information in miRBase, including the re-evaluation of miRNA identity according to their sequence origin (for example, sequences that can alternatively be matched to rsRNAs) and modification features judged by their sensitivity to PANDORA-seq (**Figure 3.6n,o**).

Data obtained from PANDORA-seq also provide additional interpretations of previous studies. For example, it has been demonstrated that the injection of the 30- to 40-nucleotide fractions of sperm RNAs from high-fat diet-treated mice can induce metabolic phenotypes in the offspring (Chen et al., 2016a; Sarker et al., 2019; Zhang et al., 2018b), which could be due to the effect of tsRNAs, because tsRNAs were the dominant small RNAs previously detected in 30- to 40-nucleotide fractions by traditional RNA-seq. However, PANDORA-seq revealed that the rsRNAs are, in fact, more abundant in 30- to 40-nucleotide RNA fractions from mature sperm (note that the level of 30- to 40-nucleotide rsRNAs in mature sperm heads are similar to those of tsRNAs) (**Figure 3.2c**); therefore, the phenotypic outcome of injecting the 30- to 40-nucleotide RNA fractions could be a combinatorial effect from both tsRNAs and rsRNAs and may relate to their function in cell fate regulation in the early embryo as exemplified in mESCs (**Figures 3.14, 3.16**).

Methods

Animals

Animal experiments were conducted under the protocol and approval of the institutional animal care and use committees of the University of California, Riverside, the University of Nevada, Reno and the Institute of Zoology, Chinese Academy of Sciences, China. Mice were given access to food and water ad libitum and were maintained on a 12 h light/12 h dark artificial lighting cycle. Mice were housed in cages at a temperature of 22-25 °C, with 40-60% humidity.

Tissue preparation

Male C57BL/6J mice aged 9-10 weeks were sacrificed individually, and brains, livers, and spleens were harvested and frozen in liquid nitrogen. Frozen tissues were pulverized in liquid nitrogen for RNA isolation or were stored at -80 °C.

Sperm isolation

Mature sperm were released from the cauda epididymis of 9-week-old C57BL/6J male mice into 5 ml phosphate-buffered saline (PBS) and incubated at 37 °C for 15 min, after which the sperm were filtered using a 40-μm cell strainer to remove the tissue debris. The sperm were then incubated with somatic cell lysis buffer (0.1% sodium dodecyl sulfate (SDS) and 0.5% Triton X in nuclease-free H_2O) for 40 min on ice to eliminate somatic cell contamination. Sperm were then pelleted by centrifugation at 600g for 5 min. Then, the

sperm pellet was resuspended and washed in 10 ml PBS and centrifuged twice at 600g for 5 min. The precipitation was performed for the RNA isolation procedure.

Sperm head isolation

Sperm head isolation was based on the previous publication (Peng et al., 2012). Mature sperm were released from the cauda epididymis of male mice into 5 ml PBS and incubated at 37 °C for 15 min, after which the sperm were then filtered using a 40-μm cell strainer to remove tissue debris. After centrifugation at 3,000g for 5 min, the sperm were then incubated with lysis buffer (10 mM Tris-HCl (pH 8.0), 10 mM EDTA, 50 mM NaCl, 2% SDS and 7.5% proteinase K) for 15 min at room temperature, followed by centrifugation at 3,000g for 5 min. The pellet (mostly sperm heads) was collected, resuspended, washed in 10 ml PBS and centrifuged at 600g for 5 min, repeated twice. The precipitation was examined under microscopy for sperm head purity (>99%) before being processed for RNA extraction.

Mouse ESCs

E14 mouse ESCs were kindly provided by A. Smith (Stem Cell Institute, Cambridge, United Kingdom). Cells were cultured on gelatin-coated plates in N2B27 supplemented with 2iLIF (1μM MEK inhibitor PD0325901 (Stem Cell Institute), 3μM GSK3 inhibitor CHIR99021 (Stem Cell Institute) and 10 ng ml^{-1} leukaemia inhibitory factor (LIF; Stem Cell Institute)) at 37 °C under 21% O2 and 5% CO2. The N2B27 medium comprised a 1:1 mix of DMEM/F-12 (21331-020; Thermo Fisher Scientific) and Neurobasal A (10888-022;

Thermo Fisher Scientific) supplemented with 1% vol/vol B-27 (10889-038; Thermo Fisher Scientific), 0.5% vol/vol N-2 (homemade), 100μM β-mercaptoethanol (31350-010; Thermo Fisher Scientific), penicillin-streptomycin (15140122; Thermo Fisher Scientific) and GlutaMAX (35050061; Thermo Fisher Scientific). The N-2 supplement contained DMEM/F-12 medium (21331-020; Thermo Fisher Scientific), 2.5 mg ml^{-1} insulin (I9287; Sigma-Aldrich), 10 mg ml^{-1} apo-transferrin (T1147; Sigma-Aldrich), 0.75% Bovine Albumin Fraction V (15260037; Thermo Fisher Scientific), 20μg ml^{-1} progesterone (p8783; Sigma-Aldrich), 1.6 mg ml^{-1} putrescine dihydrochloride (P5780; Sigma-Aldrich) and 6μg ml^{-1} sodium selenite (S5261; Sigma-Aldrich).

Human ESCs

The UK Stem Cell Bank Steering Committee approved all of the hESC experiments. All of the experiments complied with the UK Code of Practice for the Use of Human Stem Cell Lines. The hESC line used was H9, which was kindly provided by L. Vallier (Stem Cell Institute), within an agreement with WiCell. Unless otherwise stated, hESCs were maintained in a humidified incubator set at 37 °C under 21% O_2 and 5% CO_2.

Cells were passaged using Accutase, which was added for 3 min at 37 °C before being diluted in DMEM/F-12 and centrifuged. Cells were then plated in their appropriate medium supplemented with 10μM ROCK inhibitor Y-27632 (72304; STEMCELL Technologies). The ROCK inhibitor was removed after 24 h.

Primed hESCs

Conventional primed hESCs were either cultured on growth factor-reduced Matrigel (Corning)-coated dishes or on irradiated CF-1 MEFs (ASF-1201; AMS Biotechnology). For the Matrigel coating, a 16% Matrigel solution in DMEM/F-12 was incubated for 2 h at room temperature. When cultured on Matrigel, primed hESCs were cultured in mTeSR1 (85850; STEMCELL Technologies), with the medium changed every 24 h. When cultured on MEFs, primed hESCs were cultured in primed medium consisting of DMEM/F-12 (21331-020; Thermo Fisher Scientific) supplemented with 100μM β-mercaptoethanol (31350-010; Thermo Fisher Scientific), penicillin-streptomycin (15140122; Thermo Fisher Scientific), GlutaMAX (35050061; Thermo Fisher Scientific), MEM Non-Essential Amino Acids (11140035; Thermo Fisher Scientific) and 20% vol/vol KnockOut Serum Replacement (10828010; Thermo Fisher Scientific). This was supplemented with 12 ng ml^{-1} bFGF2 (Stem Cell Institute) before use.

Naive hESCs

To convert hESCs into a naive state, the protocol published by A. Smith's laboratory was used (Guo et al., 2017). At 24 h before beginning the resetting protocol, hESCs were plated on MEFs in primed medium. Once reset, cells were maintained in N2B27 supplemented with T2iLGö (1μM CHIR (Stem Cell Institute), 1μM PD03 (Stem Cell Institute), 10 ng ml^{-1} recombinant human LIF (Stem Cell Institute) and 2μM Gö (2285; Tocris) under hypoxic conditions (5% O_2, 5% CO_2 and 37 °C).

Induction of iPSCs

To derive iPSCs, a well-established reprogrammable mouse system that allows reproducible kinetics was used during this process (Cheloufi et al., 2015; Stadtfeld et al., 2010). MEFs were derived from transgenic embryos harbouring two copies of a doxycycline-inducible polycistronic transcription factor cassette (Col1a1::tetOP-OKSM) and a constitutive M2rtTA driver with or without the Oct4-EGFP reporter. Cells were first expanded in DMEM media supplemented with 10% foetal bovine serum (FBS), 100 U ml^{-1} penicillin, 100μg ml^{-1} streptomycin, sodium pyruvate (1 mM), l-glutamine (4 mM), 0.1 mM β-mercaptoethanol and 50μg ml^{-1} sodium ascorbate at 37 °C under normal oxygen levels (21% O_2). MEFs were then trypsinized and plated under reprogramming culture conditions by adding 1,000 U ml^{-1} LIF, 50μg ml^{-1} sodium ascorbate and 2μg ml^{-1} doxycycline to ESC media (knockout DMEM supplemented with 15% FBS, 100 U ml^{-1} penicillin, 100μg ml^{-1} streptomycin, 1 mM sodium pyruvate, 4 mM l-glutamine and 0.1 mM β-mercaptoethanol). Specifically, cells were plated at a density of 2 million, 300,000 and 60,000 cells per 10-cm plate to collect day 0 uninduced MEFs, day 3 reprogramming intermediates and established iPSC cultures, respectively. Doxycycline was replenished every 48 h to sustain expression of the OKSM transcription factors. To establish iPSCs, doxycycline and ascorbic acid were withdrawn at day 5 of reprogramming and cells were cultured for another five days to ensure formation of Col1a1::tetOP-OKSM transgene-independent iPSC colonies. iPSC lines were derived from three independent MEF lines. To reduce epigenetic memory, transgene-independent iPSCs were passaged for an additional five passages and pre-plated for 30 min at 37 °C. Isolated iPSCs were then analysed for Oct4-GFP expression using

flow cytometry and microscopy. Cell pellets for each time point (day 0, day 3 and established iPSCs) were collected and resuspended in TRIzol at a concentration of 10 million cells per ml for subsequent RNA isolation.

Embryoid body assay from ESCs

Mouse ESCs containing an Oct4-GFP reporter were incubated at 37 °C under 5% CO2, passaged every 2 d in gelatin-coated culture dishes and maintained in stem cell media consisting of KO-DMEM (Gibco; 10829) supplemented with 15% FBS (Gibco; 10437; Lot-2190737RP), 2 mM GlutaMAX (Gibco; 35050), 100 U ml^{-1} penicillin (Gibco; 15140), 100µg ml^{-1} streptomycin (Gibco; 15140), non-essential amino acids (100µM each; Gibco; 11140), 55µM β-mercaptoethanol (Gibco; 21985) and 1,000 U ml^{-1} LIF.

Embryoid bodies were formed as previously described (Behringer et al., 2016). ESCs were trypsinized using 0.25% trypsin-EDTA (Gibco; 25200), rinsed twice with Dulbecco's PBS (Gibco; 14190) and resuspended in stem cell media without LIF at 32,000 cells per ml. The cell suspension was then aliquoted into 25-µl drops (800 cells per drop) onto petri dish lids. The lids were then replaced onto a petri dish containing 10 ml Dulbecco's PBS to form hanging drops and incubated for 72 h. Hanging drops were then transferred to suspension culture in ultra-low-attachment 60-mm plates (Corning; 3261) with 6 ml stem cell media, excluding LIF, for up to 3 d. Embryoid bodies were collected from hanging drops at 24 and 72 h and from suspension cultures at day 6 (see below).

tsRNA and rsRNA transfections

ESCs were transfected at the onset of embryoid body formation as hanging drops. The transfection protocol was adapted for hanging drop embryoid bodies from the reverse transfection protocol, as described previously (Schaniel et al., 2006). Briefly, transfection mixtures containing 1.2 μM respective RNA (see below) and 30 μl ml^{-1} Lipofectamine Stem Reagent were incubated for 15 min at room temperature in unmodified DMEM (Gibco-10313). After incubation, ESCs in single-cell suspension with stem cell media (excluding LIF and antibiotics) were added to each transfection mixture to make final concentrations of 32,000 cells per ml, 200 nM total RNA and 5μl ml^{-1} Lipofectamine Stem Reagent. The ESC transfection mixture was then used for the embryoid body differentiation assay. Day 1 and day 3 collections were taken after 24 and 72 h incubation of hanging drops, and day 6 collections were taken after an additional 72 h incubation in suspension culture by low-attachment culture dish (Corning; 3261).

For each transfection, three independent replicates were performed. Vehicle-only transfection was used as a control. The transfection group included one of the following RNA suspensions: rsRNA-28S-1, 5' $tsRNA^{Ala}$, 3' $tsRNA^{Arg}$, 5' $tsRNA^{Glu}$, 5' $tsRNA^{His}$, 3' $tsRNA^{Lys}$ or a tsRNA pool containing the abovementioned five tsRNAs, making a total of 24 samples per time point collection (days 1, 3 and 6).

rsRNA-28S-1 represents a mixture of three sequences of different lengths (27, 30 and 37 nucleotides) mixed together equally. Each transfected small RNA contained two forms, which attached either a hydroxy group or a phosphate group in the 3' terminal of

the synthesized sequence. The total RNA concentration for each transfection group was 200 nM. The transfected tsRNA or rsRNA sequences were as follows:

Name	Sequences
5’ tsRNA^{Ala}	5’P-rGrGrGrGrGrUrGrUrArGrCrUrCrArGrUrGrGrUrArGrAr-GrCrGrCrGrUrGrC-3’OH
	5’P-rGrGrGrGrGrUrGrUrArGrCrUrCrArGrUrGrGrUrArGrAr-GrCrGrCrGrUrGrC-3’P
5’ tsRNA^{His}	5’P-rGrCrCrGrUrGrArUrCrGrUrArUrArGrUrGrGrUrUrArGrU-rArCrUrCrUrGrCrG-3’OH
	5’P-rGrCrCrGrUrGrArUrCrGrUrArUrArGrUrGrGrUrUrArGrU-rArCrUrCrUrGrCrG-3’P
5’ tsRNA^{Glu}	5’P-rUrCrCrCrUrGrGrUrGrGrUrCrUrArGrUrGrGrUrUrArGr-GrArUrUrCrGrGrCrGrCrUrC-3’OH
	5’P-rUrCrCrCrUrGrGrUrGrGrUrCrUrArGrUrGrGrUrUrArGr-GrArUrUrCrGrGrCrGrCrUrC-3’P
3’ tsRNA^{Arg}	5’P-rUrCrGrArCrUrCrCrUrGrGrCrUrGrGrCrUrCrGrCrCrA-3’OH
	5’P-rUrCrGrArCrUrCrCrUrGrGrCrUrGrGrCrUrCrGrCrCrA-3’P
3’ tsRNA^{Lys}	5’P--rArGrGrGrUrUrCrArArGrUrCrCrCrUrGrUrUrCrGrGrGrCrGrCrCrA-3’OH
	5’P--rArGrGrGrUrUrCrArArGrUrCrCrCrUrGrUrUrCrGrGrGrCrGrCrCrA-3’P
rsRNA-28S-1	5’P-rArGrArCrGrUrGrGrCrGrArCrCrCrGrCrUrGrArArUrUrU-rArArGrC-3’OH (27 nucleotides) 5’P-rArGrArCrGrUrGrGrCrGrArCrCrCrGrCrUrGrArArUrUrU-rArArGrC-3’P (27 nucleotides) 5’P-rCrGrCrGrArCrCrUrCrArGrArUrCrArGrArCrGrUrGrGrCrGrAr-CrCrCrGrCrUrGrArArU-3’OH (35 nucleotides) 5’P-rCrGrCrGrArCrCrUrCrArGrArUrCrArGrArCrGrUrGrGrCrGrAr-CrCrCrGrCrUrGrArArU-3’P (35 nucleotides) 5’P-rCrGrCrGrArCrCrUrCrArGrArUrCrArGrArCrGrUrGrGrCrGrAr-CrCrCrGrCrUrGrArArUrUrU-3’OH (37 nucleotides) 5’P-rCrGrCrGrArCrCrUrCrArGrArUrCrArGrArCrGrUrGrGrCrGrAr-CrCrCrGrCrUrGrArArUrUrU-3’P (37 nucleotides)

mESC transfection and Global protein synthesis assay

Before transfection, 3,000 ESCs per well was seeded in 96-well plates coated with 0.1% gelatin and incubated them overnight (~16 h) with mESC medium. The transfection complex was prepared as follows: 0.4µl respective RNA (100µM) with 4µl Lipofectamine Stem Reagent and 20µl Opti-MEM was mixed by vortexing and incubated at room temperature for 15 min. The media was discarded and 180µl new mESC media (excluding antibiotics) was added to the wells. The lipofectamine-RNA transfection complex was added to the wells and incubated for 24 h at 37 °C under 5% CO_2. For each transfection, three independent replicates were used. Vehicle-only transfection was used as a control. The transfection group included one of the following RNA suspensions: scrambled small RNAs, the tsRNA pool or rsRNA-28S-1.

The global protein synthesis assay was performed with the Protein Synthesis Assay Kit (ab235634; Abcam), per the manufacturer's instructions. Briefly, the media was replaced with fresh complete mESC media containing 1×Protein Label. Incubation was performed for 2 h at 37 °C under 5% CO_2. Then, the culture media was removed and the cells were rinsed with PBS. Fixative solution (100µl) was added to each well and the cells were incubated for 15 min at room temperature, protected from light. The cells were washed with wash buffer and incubated with 100µl permeabilization buffer for 10 min at room temperature. The cells were then incubated with 1× reaction cocktail for 30 min, protected from light at room temperature, then washed again. A 1× dilution of DAPI DNA stain was prepared and 100µl was added per well. The cells were incubated for 20 min at room temperature. The DAPI staining solution was aspirated and replaced with PBS. Then, the

samples were analysed by fluorescence microscopy (Lecia DM8 system) with excitation and emission at 440/490 and 540/580 nm, respectively. The intensity of the red signal represented the relative quantity of nascent peptide. The intensity of the sample image was processed and extracted using Fuji (ImageJ) software.

Cell lines

HeLa cells were purchased from the American Type Culture Collection (ATCC; catalogue number CCL-2). HeLa cells were cultured in DMEM medium with 10% FBS and incubated at 37 °C under 5% CO_2. Total RNA was harvested when the confluency reached ~95% in a 100-mm culture dish.

RNA isolation

TRIzol reagent (1 ml; Invitrogen; 15596018) was added to a microtube with pulverized tissues or collected cells and vortexed uniformly. Then, the sample was incubated at room temperature for 5 min. Chloroform (200μl; Alfa Aesar; J67241) was added per ml of sample, vortexed for 15 s, then incubated at room temperature for 2 min and centrifuged for 15 min at 12,000g (4 °C). The aqueous phase was pooled in a microtube and combined with an equal volume of isopropanol (Fisher Scientific; BP2618-212). After gently mixing and incubating at room temperature for 10 min, the tube was centrifuged for 10 min at 12,000g (4 °C). After removing the supernatant, the precipitation was washed with 1 ml 75% ethanol (Koptec; V1001), then centrifuged for 5 min at 7,500g (4 °C). Then, the supernatant

was removed and air-dried for 5 min and the precipitation was resuspended in nuclease-free water, quantified and stored at -80 °C or used for further processing.

Isolation of specified-size RNA from total RNAs

The RNA sample, mixed with an equal volume of 2× RNA loading dye (New England Biolabs; B0363S), was incubated at 75 °C for 5 min. The mixture was loaded into 15% (wt/vol) urea polyacrylamide gel (10 ml mixture containing 7 M urea (Invitrogen; AM9902), 3.75 ml Acrylamide/Bis 19:1, 40% (Ambion; AM9022), 1 ml 10× TBE (Invitrogen; AM9863), 1 g l^{-1} ammonium persulfate (Sigma-Aldrich; A3678-25G) and 1 ml l^{-1} TEMED (Thermo Fisher Scientific; BP150-100)). The gel was run in a 1× TBE running buffer at 200 V until the bromophenol blue reached the bottom of the gel. After staining with SYBR Gold solution (Invitrogen; S11494), gel that contained small RNAs of 15-50 nucleotides was excised based on small RNA ladders (New England Biolabs (N0364S) and Takara (3416)) and eluted in 0.3 M sodium acetate (Invitrogen; AM9740) and 100 U ml^{-1} RNase inhibitor (New England Biolabs; M0314L) overnight at 4 °C. The sample was then centrifuged for 10 min at 12,000g (4 °C). The aqueous phase was mixed with pure ethanol, 3 M sodium acetate and linear acrylamide (Invitrogen; AM9520) at a ratio of 3:9:0.3:0.01. Then, the sample was incubated at -20 °C for 2 h and centrifuged for 25 min at 12,000g (4 °C). After removing the supernatant, the precipitation was resuspended in nuclease-free water, quantified and stored at -80 °C or used for further processing.

Expression and purification of Escherichia coli AlkB

The *E. coli AlkB* gene was cloned into the NdeI/BamHI site of the pET28a (+) plasmid. The constructed plasmid was transformed in the *E. coli* BL21(DE3) strain to express the AlkB protein with a tag of six histidines at the amino terminal. The *E. coli* was cultured in lysogeny broth medium containing 50μg ml^{-1} kanamycin. The medium, with 1 mM isopropyl β-D-1-thiogalactopyranoside added, was incubated at 37 °C for 3 h. The AlkB protein was purified using an Ni-NTA Superflow column and stored in a buffer containing 20 mM Tris-HCl (pH 8.0), 50% glycerol, 0.2 M NaCl and 2 mM dithiothreitol at -80 °C. The purity of the AlkB protein was detected by 12% SDS-polyacrylamide gel electrophoresis (PAGE). The enzyme activity was confirmed by treating RNA with AlkB, followed by LC-MS/MS analysis to quantify the modified nucleosides. The *AlkB* gene sequence used in this study was:

5'--CTGGACCTGTTCGCGGATGCGGAGCCGTGGCAGGAACCGCTGGCGGCGG
GTGCGGTTATCCTGCGTCGTTTCGCGTTTAACGCGGCGGAGCAACTGATCCGT
GACATTAACGATGTGGCGAGCCAGAGCCCGTTTCGTCAAATGGTTACCCCGG
GTGGCTACACCATGAGCGTGGCGATGACCAACTGCGGTCACCTGGGTTGGAC
CACCCACCGTCAGGGTTACCTGTATAGCCCGATCGACCCGCAAACCAACAAG
CCGTGGCCGGCGATGCCGCAGAGCTTCCACAACCTGTGCCAACGTGCGGCGA
CCGCGGCGGGTTACCCGGACTTTCAGCCGGATGCGTGCCTGATTAACCGTTAT
GCGCCGGGTGCGAAGCTGAGCCTGCACCAAGACAAAGATGAGCCGGATCTG
CGTGCGCCGATCGTTAGCGTGAGCCTGGGTCTGCCGGCGATTTTCCAGTTTGG
TGGCCTGAAGCGTAACGACCCGCTGAAACGTCTGCTGCTGGAGCACGGCGAT

GTGGTTGTGTGGGGTGGCGAAAGCCGTCTGTTCTACCACGGTATCCAGCCGCT
GAAAGCGGGCTTTCACCCGCTGACCATTGACTGCCGTTATAACCTGACCTTCC
GTCAAGCGGGTAAGAAAGAA -3'.

Quantification of modified nucleosides in RNA molecules by LC-MS/MS

A total of 1µg 15- to 50-nucleotide RNA from mouse liver was incubated with 0.2 U nuclease P1 (Sigma-Aldrich) and 60µl 50 mM NH4OAc (pH 5.3) in a microtube at 50 °C for 3 h. Then, a sample with 0.04 U phosphodiesterase I (USB) added was incubated at 37 °C for 2 h. After adding 2 U alkaline phosphatase (Sigma-Aldrich), the sample was incubated at 37 °C for 2 h. The mixture was moved into Nanosep centrifugal devices with 3K Omega membrane (PALL; OD003C35) and centrifuged for 20 min at 5,000*g* (4 °C). The liquid phase was lyophilized and stored at -80 °C. Then, the sample was dissolved in 70µl 2 mM ammonium acetate with 175 ng ml^{-1} guanosine (13C, 15N). Afterwards, 65µl of the solution was injected into the LC-MS/MS system. The solution was separated using an Agilent 1200 HPLC system and then detected using an API 4000 QTRAP mass spectrometer (Applied Biosystems) with positive electrospray ionization. The following mass transitions were monitored: *m/z* 244.1 to 112.1 for cytidine (C); *m/z* 268.1 to 136.2 for adenosine (A); *m/z* 284.1 to 152.2 for guanosine (G); *m/z* 245.0 to 113.1 for uridine (U); *m/z* 282.1 to 150.2 for 1-methyladenosine (m^1A); *m/z* 298.1 to 166.1 for 1-methylguanosine (m^1G); *m/z* 258.0 to 126.0 for 3-methylcytidine (m^3C); *m/z* 312.1 to 180.2 for *N*2,*N*2-dimethylguanosine ($m^2{}_2G$); *m/z* 258.1 to 112.1 for 2'-*O*-methylcytidine (Cm); *m/z* 282.1 to 136.2 for 2'-*O*-methyladenosine (Am); *m/z* 259.1 to 113.1 for 2'-*O*-methyluridine (Um); *m/z* 298.1 to

152.1 for 2'-*O*-methylguanosine (Gm); *m/z* 258.1 to 126.1 for 5-methylcytidine (m^5C); *m/z* 298.1 to 166.1 for *N*2-methylguanosine (m^2G); *m/z* 245.2 to 125.1 for pseudouridine (Ψ); and *m/z* 286.1 to 154.1 for *N*4-acetylcytidine (ac^4C). The nucleoside concentration was quantified according to the standard curve running for the same batch of samples. The ratios of m^1A/A, Am/A, m^1G/G, $m^2{}_2G$/G, Gm/G, m^2G/G, m^3C/C, Cm/C, m^5C/C, ac^4C/C, Um/U and Ψ to U were subsequently calculated.

Treatment of RNA with AlkB

The RNA was incubated in 50μl reaction mixture containing 50 mM HEPES (pH 8.0) (Gibco (15630080) and Alfa Aesar (J63578)), 75μM ferrous ammonium sulfate (pH 5.0), 1 mM α-ketoglutaric acid (Sigma-Aldrich; K1128-25G), 2 mM sodium ascorbate, 50 mg l^{-1} bovine serum albumin (Sigma-Aldrich; A7906-500G), 4μg ml^{-1} AlkB, 2,000 U ml^{-1} RNase inhibitor and 200 ng RNA at 37 °C for 30 min. Then, the mixture was added into 500μl TRIzol reagent to perform the RNA isolation procedure.

Treatment of RNA with T4PNK

The RNA was incubated in 50μl reaction mixture containing 5μl 10× PNK buffer (New England Biolabs; B0201S), 1 mM ATP (New England Biolabs; P0756S), 10 U T4PNK (New England Biolabs; M0201L) and 200 ng RNA at 37 °C for 20 min. Then, the mixture was added into 500μl TRIzol reagent to perform the RNA isolation procedure.

RNA adapter ligation capability identification

The synthetic RNA with a 3'-OH end or a 3'-P end, or 25- to 50-nucleotide RNA from mouse spleen were performed in the experiment. Then, 50 ng RNA, dissolved in 5.5μl nuclease-free water mixed with 0.5μl 10μM 3' SR adapter (Takara; sequence: 5'--(rApp)--AGATCGGAAGAGCACACGTCT(NH2)-3') and 2 μl 50% PEG 8000 (New England Biolabs; B1004), was incubated at 70 °C for 2 min. Following this, the sample was immediately incubated on ice for 5 min. Next, 1μl 10× T4 ligase reaction buffer (New England Biolabs; B0216L) and 1μl T4 RNA Ligase 2, truncated KQ (New England Biolabs; M0373L) were added to the sample, which was mixed well. After incubation at 25 °C for 1 h and 75 °C for 5 min, the sample was run on 15% (wt/vol) urea polyacrylamide gel, followed by northern blot using the anti-3' SR adapter probe (Takara; sequence: 5'--(DIG)-AGACGTGTGCTCTTCCGATCT-3') to detect the ligation outcome of the input RNAs.

Northern blot

Total RNA was extracted from mouse tissues and cell lines using TRIzol reagents, per the manufacturer's instructions. RNA was separated by 10% urea-PAGE gel stained with SYBR Gold, and immediately imaged, then transferred to positively charged nylon membranes (Roche; 11417240001) and ultraviolet crosslinked with an energy of 0.12 J. Membranes were pre-hybridized with DIG Easy Hyb solution (Roche; 11603558001) for 1 h at 42 °C. To detect miRNAs, tsRNAs and rsRNAs in the total RNA and 15- to 50-nucleotide small RNAs, membranes were incubated overnight (12-16 h) at 42 °C with DIG-labelled oligonucleotide probes synthesized by Integrated DNA Technologies as follows:

Name	Sequences
rsRNA-28S-1	5’-DIG-ATTCAGCGGGTCGCCACGTCT
rsRNA-28S-2	5’-DIG-GGTCCGCACCAGTTCT
rsRNA-28S-3	5’-DIG-CGCCAGGTTCCACACGAACGT
rsRNA-18S-1	5’-DIG-AGGCACACGCTGAGCCAGTCAGT
5’ tsRNA^{Glu}	5’-DIG-AACCACTAGACCACCAGGGA
5’ tsRNA^{Ala}	5’-DIG-GCACGCGCTCTACCACTG
5’ tsRNA^{His}	5’-DIG-AGTACTAACCACTATACGATCACGG
3’ tsRNA^{Arg}	5’-DIG-TGGCGAGCCAGCCAGGAGTCGA
3’ tsRNA^{Lys}	5’-DIG-TGGCGCCCGAACAGGGACTT
let-7i	5’-DIG-CAGCACAAACTACTACCTCA
let-7f	5’-DIG-AACTATACAATCTACTACCTCA
miR-122	5’-DIG-AAACACCATTGTCACACTCCA
miR-21	5’-DIG-TCAACATCAGTCTGATAAGCTA
3’adapter probe	5’-DIG-AGACGTGTGCTCTTCCGATCT

Then the membranes were washed twice with low stringent buffer (2× SSC with 0.1% (wt/vol) SDS) at 42 °C for 15 min, rinsed twice with high stringent buffer (0.1× SSC with 0.1% (wt/vol) SDS) for 5 min, and then rinsed in washing buffer (1× SSC) for 10 min. Following the washes, the membranes were transferred into 1× blocking buffer (Roche, REF:11096176001) and incubated at room temperature for 3 h, after which the Anti-Di-goxigenin-AP Fab fragments (Roche, REF: 11093274910) were added into the blocking buffer at a ratio of 1:10,000 and incubated for an additional 30 min at room temperature. Then the membranes were washed four times with DIG washing buffer (1× maleic acid buffer, 0.3% Tween-20) for 15 min, sequentially incubated in DIG detection buffer (0.1 M TrisHCl, 0.1 M NaCl, pH 9.5) for 5 min, and coated with CSPD ready-to-use reagent (Roach REF: 11755633001). The membranes were incubated in the dark with the CSPD reagent for 30 min at 37 °C before imaging with ChemiDoc™ MP Imaging System (BIO-RAD).

Small RNA northern blot probe efficiency assay

Synthetic RNA sequences complementary to northern blot probes (that is, rsRNA-28s-1, 5’ tsRNAGlu, let-7i, mir-122 and mir-21) were synthesized by Integrated DNA Technologies as follows:

Name	Sequences
Syn-rsRNA-28S-1	/5Phos/rArGrArCrGrUrGrGrCrGrArCrCrCrGrCrUrGrArA-rUrUrU
Syn-5’ tsRNA-Glu	/5Phos/rUrCrCrCrUrGrGrUrGrGrUrCrUrArGrUrGrGrUrU-rArGrGrArUrUrCrGrGrCrGrCrU
Syn-let-7i	/5Phos/rUrGrArGrGrUrArGrU-rArGrUrUrUrGrUrGrCrUrGrUrU
Syn-miR-122	/5Phos/rUrGrGrArGrUrGrUrGrArCrArArUrGrGrUrGrUrUrU
Syn-miR-21	/5Phos/rUrArGrCrUrUrArUrCrArGrArCrUrGrArUrGrUrUr-GrArC

Small RNA library construction and deep sequencing

The RNA segment was separated by PAGE, then a 15- to 45-nucleotide stripe was selected and recycled. The adapters were obtained from the NEBNext Small RNA Library Prep Set for Illumina (New England Biolabs; E7330S) and ligated sequentially. First, a 3’ adapter system was added under the following reaction conditions: 70 °C for 2 min and 25 °C for 1 h or 16 °C for 18 h (for sperm heads). Second, a reverse transcription primer was added under the following reaction conditions: 75 °C for 5 min, 37 °C for 15 min and 25 °C for 15 min. Third, a 5’ adapter mix system was added under the following reaction conditions: 70 °C for 2 min and 25 °C for 1 h. First-strand cDNA synthesis was performed under the following reaction conditions: 70 °C for 2 min and 50 °C for 1 h. PCR amplification with PCR Primer Cocktail and PCR Master Mix was performed to enrich the cDNA fragments under the following conditions: 94 °C for 30 s; 11-22 cycles of 94 °C for 15 s, 62 °C for

30 s and 70 °C for 15 s; 70 °C for 5 min; and hold at 4 °C. Then, the PCR product was purified from PAGE gel. The qualified libraries were amplified on cBot to generate the cluster on the flow cell. The amplified flow cell was sequenced using the SE50 strategy on the Illumina system by BGI. For sperm heads, the qualified libraries were amplified and sequenced using the SE75 strategy on the Illumina system by the University of California, San Diego IGM Genomics Center.

Quality control of small RNA-seq data

The resulting sequencing reads were processed according to the standard quality control criteria: (1) reads containing N; (2) reads containing more than four bases with a quality score<10; (3) reads containing more than six bases with a quality score<13; (4) reads with 5' primer contaminants or without 3' primer; (5) reads without the insert tag; (6) reads with ploy A; and (7) reads shorter than 15 nucleotides and longer than 44 nucleotides. The sequencing data analyses were performed on the clean reads after data filtration.

Small RNA annotation and analyses for PANDORA-seq data

RNAs of 15-50 nucleotides were subject to the PANDORA-seq protocol. Small RNA sequences were annotated using the software SPORTS1.1 (updated from SPORTS1.0) with one mismatch tolerance (SPORTS1.1 parameter setting: -M 1). Reads were mapped to the following individual non-coding RNA databases sequentially: (1) the miRNA database miRBase 21(Kozomara and Griffiths-Jones, 2014); (2) the genomic tRNA database GtRNAdb (Chan and Lowe, 2016); (3) the mitochondrial tRNA database mitotRNAdb (Juhling

et al., 2009); (4) the rRNA and YRNA databases assembled from National Center for Biotechnology Information nucleotide and gene database, (5) the piRNA databases, including piRBase (Zhang et al., 2014) and piRNABank (Sai Lakshmi and Agrawal, 2008); and (6) the noncoding RNAs defined by Ensembl (Yates et al., 2016) and Rfam 12.3(Nawrocki et al., 2015). The tsRNAs were annotated based on both pre-tRNA and mature tRNA sequences. Mature tRNA sequences were derived from the GtRNAdb and mitotRNAdb sequences using the following procedures: (1) predicted introns were removed; (2) a CCA sequence was added to the 3' ends of all tRNAs; and (3) a G nucleotide was added to the 5' end of histidine tRNAs. The tsRNAs were categorized into four types based on the origin of the tRNA loci: 5' tsRNA (derived from the 5' end of pre-/mature- tRNA); 3' tsRNA (derived from the 3' end of pre-tRNA); 3' tsRNA-CCA end (derived from the 3' end of mature tRNA); and internal tsRNAs (not derived from 3' or 5' loci of tRNA). For the rsRNA annotation, the small RNAs were mapped to the parent rRNAs in an ascending order of rRNA sequence length to ensure a unique annotation of each rsRNA (for example, the rsRNAs mapped to 5.8S rRNA would not be further mapped to the genomic region overlapped by 5.8S and 45S rRNAs).

Differentially expressed small RNA analysis

Pairwise comparison of differentially expressed small RNAs (average RPM > 0.1 in the compared treatments) among different RNA treatment were performed using the R package DEGseq (Wang et al., 2010) with a normalized RPM fold change > 2 and $p < 0.05$.

Atypical miRNA analysis

Here, the miRNAs identified by either traditional RNA-seq or PANDORA-seq (mean RPM>0.1) and can perfectly match to the miRBase (SPORTS1.1 parameter setting: -M 0) were focused on. These miRNAs were re-mapped to the other small RNA databases with one mismatch tolerance (SPORTS1.1 parameter setting: -M 1), which potentially yielded an alternative annotation.

Small RNA secondary structure prediction

The tRNA secondary structure information was obtained from the GtRNAdb, while the YRNA secondary structure was predicted using the RNAfold tool in the ViennaRNA package(Lorenz et al., 2011) with default settings. The RNA secondary structure visualization was performed using the forna tool in the ViennaRNA package.

rsRNA coverage similarity comparison matrix

To calculate the overall rsRNA coverage similarity pairwise comparison among samples, a sensitive method was performed. For one specific rRNA with length n, it was assumed that the rsRNA coverage level of locus i in sample X is x_i and in sample Y is y_i. The rsRNA mapping similarity level between the two samples can be described as:

$$r = \sum_{i=1}^{n} \left| \frac{x_i}{\sum_{i=1}^{n} x_i} - \frac{y_i}{\sum_{i=1}^{n} y_i} \right|$$

The lower r value indicates that samples X and Y are more similar in rsRNA coverage, while higher r value represents the opposite.

Identification of RNA mapping peaks

The peak searching algorithm was modified from the findpeaks function in the R pracma package (version 1.9.9; https://www.rdocumentation.org/packages/pracma/versions/1.9.9/topics/findpeaks). Briefly, a new parameter gradient was added to the original algorithm for RNA peak identification. The expression significance of the RNA mapping region between traditional treatment and PANDORA-seq treatment was analysed by two-way analysis of variance (ANOVA).

mRNA library construction, RNA-seq and quality control

Transcriptome libraries were constructed using the NEBNext Ultra RNA Library Prep Kit for Illumina (New England Biolabs; E7530L) following the manufacturer's recommendations. For each RNA library, six G base pairs (raw data) were generated on the Illumina system. The resulting sequencing reads were processed using standard quality control criteria: (1) reads containing adapters; (2) reads containing N>10% (N represents bases that cannot be determined); and (3) reads containing low-quality (Q score≤5) bases that represent over 50% of the total bases. The data sequencing analyses were performed on the clean reads after data filtration. The mRNA library preparation, quality examination and RNA-seq processes were performed by Novogene.

Transcriptome data annotation

RNA sequences were annotated using kallisto (Bray et al., 2016) with Ensembl mouse cDNA annotation information (GRCm38). The expression level of each gene was normalized to transcripts per kilobase million.

Functional enrichment analysis

The edgeR (Robinson et al., 2010) was employed tool to identify the differentially expressed genes between the control and treated groups during mESC differentiation. The TMM algorithm was used for read count normalization and effective library size estimation (Robinson and Oshlack, 2010). The genes with a false discovery rate < 0.05 and fold change > 1.5 were deemed differentially expressed. The enriched biological process terms of differentially expressed genes were obtained using R package clusterProfiler (Yu et al., 2012), setting a q value threshold of 0.005 for statistical significance. Only the gene sets with ≥2 differentially genes were retained.

GOBP gene set score

The FAIME algorithm (Yang et al., 2012) was applied to assign a gene set score for each GOBP term. The FAIME algorithm calculated gene set scores based on the rank-weighted gene expression of individual samples, which converts each sample's transcriptomic data into pathway-/gene set-based information. A higher gene set score indicates an overall increase in the abundance of the genes within the given GOBP term.

Statistics and Reproducibility

The statistical tests and biological repeats for the RNA-seq samples, LC-MS/MS and northern blot validations are described in the figure captions or Methods. All of the correlation analyses were performed using Spearman's rank correlation test to generate the correlation coefficient (ρ). Multiple t-tests were performed using GraphPad Prism for the statistical analyses of RNA modification dynamics of 15- to 50-nucleotide RNA fractions from mouse liver after AlkB treatment. Fisher's least significant difference (LSD) test was performed for statistical analysis of the different origins of the tsRNAs/miRNA expression ratio under different treatments among mouse and human tissues and cells, miRNA expression during the cell reprogramming using PANDORA-seq, and statistical analysis of representative GOBP terms during days 1, 3 and 6 of embryoid body differentiation under control, rsRNA-28S-1 and pooled tsRNA transfection. Two-way ANOVA was performed for statistical analysis of tsRNA/ rsRNA mapping peaks between MEFs and iPSCs on the corresponding RNA loci. Student's t-test was performed for statistical analysis of the expression level of the northern blot probe targeting small RNAs between MEFs and iPSCs, as well as gene set score comparison for GOBP terms between controls and different RNA transfections. Dunnett's multiple comparisons test was performed using GraphPad Prism for statistical analysis of protein synthesis rates after ESC transfection of scrambled RNA, rsRNA-28S-1 and pooled tsRNA. The radar plots were generated using the radarchart function in the R package fmsb based on a $\log_{10}$-transformed scale. The RNA relative expression heatmaps were generated using the heatmap.2 function in the R package gplots based on a $\log_2$-transformed scale. For each small RNA mapping plot, a shaded band was

included to indicate the standard error of the mean (s.e.m.). The rRNA coverage similarity comparison matrices were generated using the pheatmap function in the R package pheatmap.

Figures

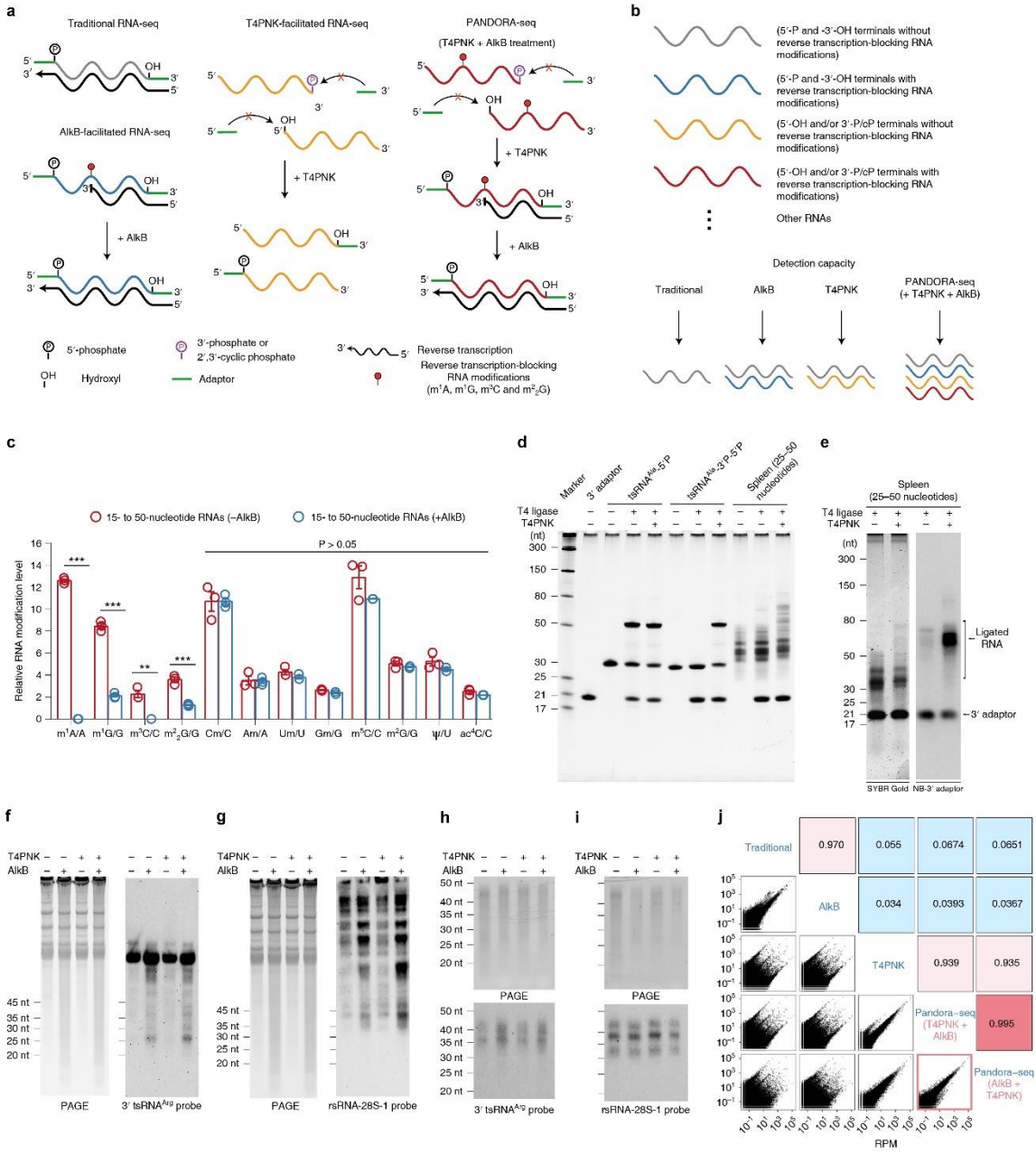
a
Traditional RNA-seq
T4PNK-facilitated RNA-seq
PANDORA-seq
(T4PNK + AlkB treatment)
AlkB-facilitated RNA-seq
+ T4PNK
+ AlkB
5′-phosphate
3′-phosphate or
2′,3′-cyclic phosphate
Hydroxyl
Adaptor
Reverse transcription
Reverse transcription-blocking
RNA modifications
(m1A, m1G, m3C and m22G)
b
(5′-P and -3′-OH terminals without reverse transcription-blocking RNA modifications)
(5′-P and -3′-OH terminals with reverse transcription-blocking RNA modifications)
(5′-OH and/or 3′-P/cP terminals without reverse transcription-blocking RNA modifications)
(5′-OH and/or 3′-P/cP terminals with reverse transcription-blocking RNA modifications)
Other RNAs
Detection capacity
Traditional
AlkB
T4PNK
PANDORA-seq
(+ T4PNK + AlkB)
c
15- to 50-nucleotide RNAs (−AlkB)
15- to 50-nucleotide RNAs (+AlkB)
P > 0.05
Relative RNA modification level
m1A/A
m1G/G
m3C/C
m22G/G
Cm/C
Am/A
Um/U
Gm/G
m5C/C
m2G/G
ψ/U
ac4C/C
d
Marker
3′ adaptor
tsRNAAla-5′P
tsRNAAla-3′P-5′P
Spleen (25–50 nucleotides)
T4 ligase
T4PNK
(nt)
e
Spleen
(25–50 nucleotides)
Ligated RNA
3′ adaptor
SYBR Gold
NB-3′ adaptor
f
T4PNK
AlkB
PAGE
3′ tsRNAArg probe
g
PAGE
rsRNA-28S-1 probe
h
PAGE
3′ tsRNAArg probe
i
PAGE
rsRNA-28S-1 probe
j
Traditional
AlkB
T4PNK
Pandora-seq (T4PNK + AlkB)
Pandora-seq (AlkB + T4PNK)
0.970
0.055
0.0674
0.0651
0.034
0.0393
0.0367
0.939
0.935
0.995
RPM

Figure 3.1: Schematic overview, validation of AlkB and T4PNK enzyme activity, and protocol optimization of PANDORA-seq

a, Schematics of the RNA properties (terminal and internal modifications) and key steps (adapter ligation and reverse transcription) of traditional RNA-seq, AlkB-facilitated RNA-seq, T4PNK-facilitated RNA-seq and PANDORA-seq. **b**, Schematic of the detection capacities of the abovementioned RNA-seq protocols from a small RNA pool. **c**, Demethylation activity of m^1A, m^1G, m^3C and $m^2{}_2G$ with or without AlkB treatment of 15- to 50-nucleotide RNA fractions from mouse tissue (liver), as revealed by LC-MS/MS (n = 3 biologically independent samples). The data represent means ± s.e.m. Statistical significance was determined by two-sided multiple t-test (**$P < 0.01$; ***$P < 0.001$). **d**, Validation of improvements in 3' terminal ligation following T4PNK treatment in synthesized tsRNAs and small RNA fractions extracted from mouse tissue (spleen). nt, nucleotides. **E**, Northern blot analysis of the 3' adapter sequence to show, semi-quantitatively, improvement in the number of adapters being ligated before and after treatment with T4PNK. **f-i**, the improved treatment protocol minimized the potential artificial increase in tsRNAs and rsRNAs due to de novo degradation of tRNAs and rRNAs. In f and g, AlkB treatment on total RNAs (from HeLa cells) resulted in increased tsRNA (**f**) and rsRNA products (**g**), as observed by increased RNA smear (left) and by northern blots (right). In **h** and **i**, northern blot analyses of tsRNAs (**h**) and rsRNAs (**i**) after AlkB and/or T4PNK treatment on pre-size-selected RNA fractions (15- to 50-nucleotide RNA from HeLa cells) did not result in further degradation. For **d-i**, similar results were obtained in three independent experiments. **J**, Comparison of the PANDORA-seq results using treatment with either T4PNK first and AlkB second (T4PNK + AlkB) or AlkB first and T4PNK second (AlkB + T4PNK) in HeLa cells (15- to 50-nucletide RNA) showed highly consistent results (Spearman's correlation; $\rho = 0.995$).

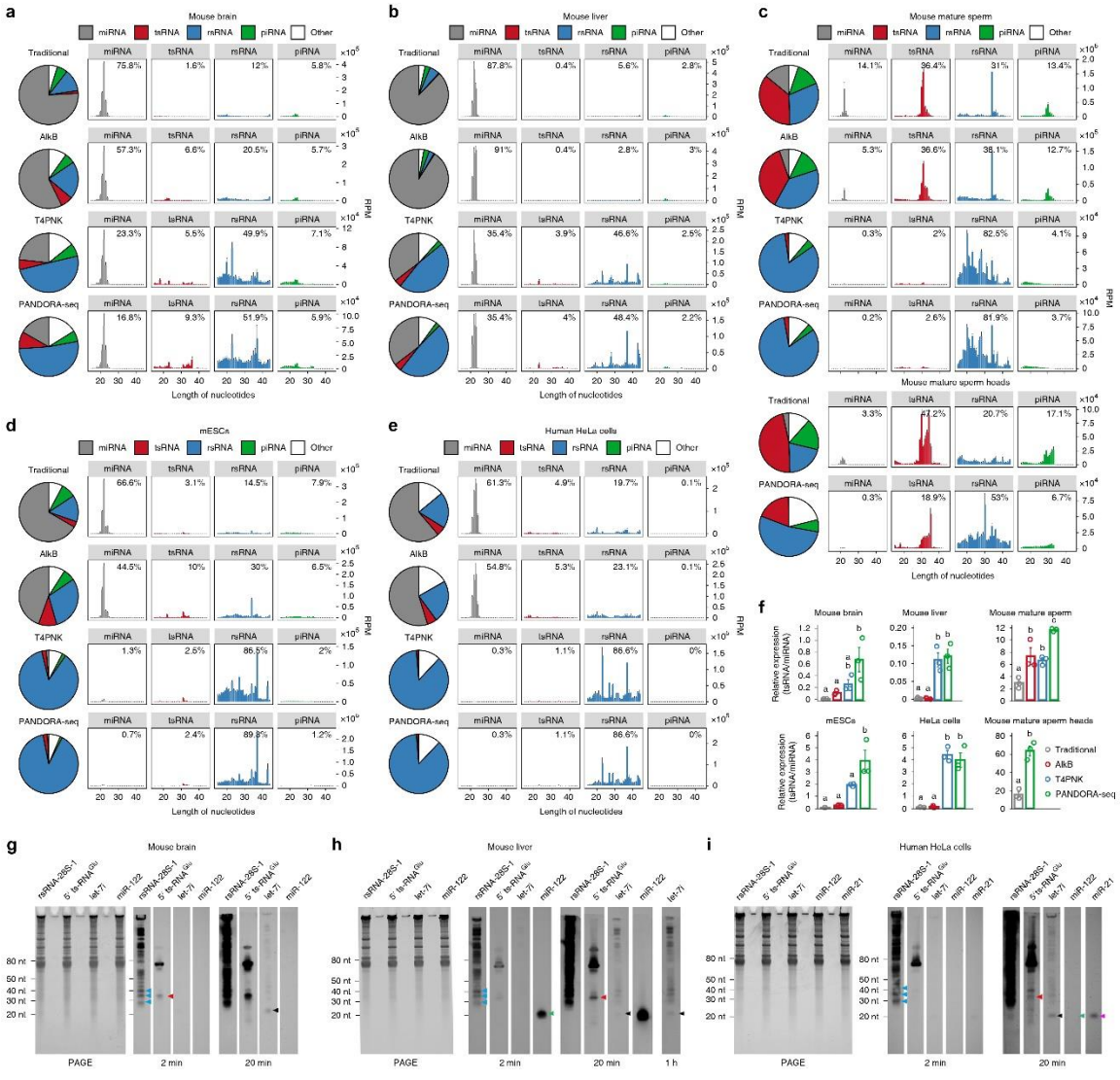
a
Mouse brain
b
Mouse liver
c
Mouse mature sperm
Mouse mature sperm heads
d
mESCs
e
Human HeLa cells
miRNA
tsRNA
rsRNA
piRNA
Other
Traditional
AlkB
T4PNK
PANDORA-seq
RPM
Length of nucleotides
f
Mouse brain
Mouse liver
Mouse mature sperm
mESCs
HeLa cells
Mouse mature sperm heads
Relative expression (tsRNA/miRNA)
Traditional
AlkB
T4PNK
PANDORA-seq
g
Mouse brain
h
Mouse liver
i
Human HeLa cells
80 nt
50 nt
40 nt
30 nt
20 nt
PAGE
2 min
20 min
1 h

Figure 3.2: Read summaries and length distributions of different small RNA **categories under traditional RNA-seq, AlkB-facilitated RNA-seq, T4PNK-facilitated RNA-seq and PANDORA-seq**

a-**e**, Comparison of different protocols in five representative tissue or cell types (from a total of 11; the results for the other tissue and cell types are provided in **Figure 3.3**. 1): mouse brain (**a**), mouse liver (**b**), mouse mature sperm and mature sperm heads (**c**), mESCs (**d**) and HeLa cells (**e**). The results show a dynamic landscape of small RNAs detected by different methods and across different tissue and cell types. The data represent means ± s.e.m. f, Relative tsRNA/miRNA ratios under different protocols (n = 3 biologically independent samples per bar). Different letters above the bars indicate a statistically significant difference ($P < 0.05$). Same letters indicate $P \geq 0.05$. Statistical significance was determined by two-sided one-way ANOVA with uncorrected Fisher's LSD test. All data are plotted as means ± s.e.m. **g**-**i**, The relative expression levels of miRNAs, tsRNAs and rsRNAs, as revealed by PANDORA-seq, were validated by northern blots. The results for mouse brain (**g**), mouse liver (**h**) and HeLa cells (**i**) are shown. For **g**-**i**, similar results were obtained in three independent experiments. Blue arrowheads point to rsRNA-28S-1, red arrowheads point to 5' $tsRNA^{Glu}$, black arrowheads point to let-7i, green arrowheads point to miR-122 and purple arrowheads point to miR-21.

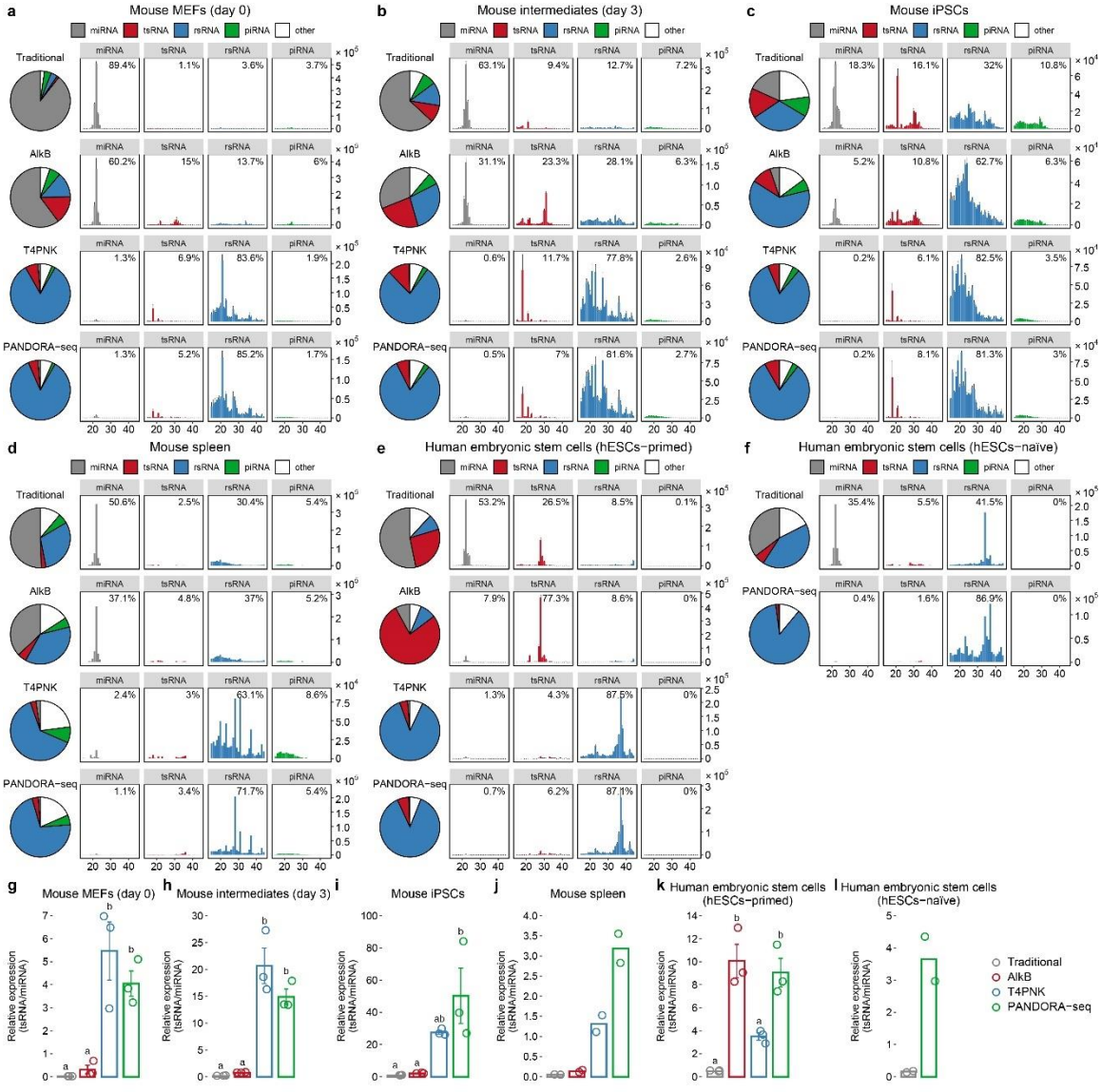
a
Mouse MEFs (day 0)
b
Mouse intermediates (day 3)
c
Mouse iPSCs
d
Mouse spleen
e
Human embryonic stem cells (hESCs−primed)
f
Human embryonic stem cells (hESCs−naïve)
miRNA
tsRNA
rsRNA
piRNA
other
Traditional
AlkB
T4PNK
PANDORA−seq
g
Mouse MEFs (day 0)
h
Mouse intermediates (day 3)
i
Mouse iPSCs
j
Mouse spleen
k
Human embryonic stem cells (hESCs−primed)
l
Human embryonic stem cells (hESCs−naïve)
Relative expression (tsRNA/miRNA)
Traditional
AlkB
T4PNK
PANDORA−seq

Figure 3.3: Reads summary and length distributions of different small RNA **category under Traditional RNA-seq, AlkB-facilitated RNA-seq, T4PNK-facilitated RNA-seq, and PANDORA-seq**

Showing reads summary and length distributions of different small RNA category in six tissue/cell types that are not shown in **Figure 3.2** because of space limitation. (**a**-**c**) Cells during mouse somatic cell reprogramming to iPSC: (**a**) MEFs (day 0), (**b**) intermediates (day 3), (**c**) iPSCs; (**d**) mouse spleen, (**e**) primed human embryonic stem cells (hESCs-primed), and (**f**) naïve human embryonic stem cells (hESCs-naïve) (**g**-**l**) the relative tsRNA/miRNA ratio under different protocols. for **g**,**h**,**i**,**k**, mean ± SEM, n=3 biologically independent samples in each bar; for j,l, n=2 biologically independent samples in each bar; different letters above bars indicate statistical difference, $P < 0.05$; same letters indicate $P \geq 0.05$ (two-sided, one-way ANOVA, uncorrected Fisher's LSD test).

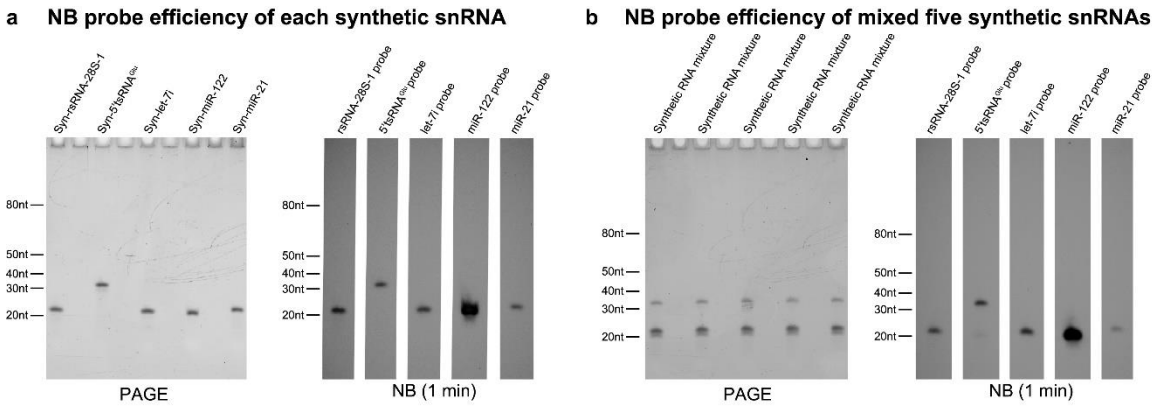

Figure 3.4: Evaluation of Northern blot probe efficiency on synthesized targets

The Northern blot probes used for each target are the same as used in **Figure 3.2 g-i**. **a**, each synthetic small RNAs (that is, rsRNA-28S-1, 5' tsRNA^{Glu}, let-7i, mir-122, mir-21) are individually loaded on PAGE followed by northern blots analyses. **b**, the five synthetic small RNAs were mixed together with the amount tested in (**a**) and then equally separated and loaded on PAGE followed by northern blots analyses. The relative efficiency of each northern blot probe can be shown: the probe efficiency between let-7i, tsRNA^{Glu} and rsRNA-28 are similar; the probe for mir-122 is highest, while the probe for mir-21 has the lowest efficiency. Similar results were obtained in 3 independent experiments.

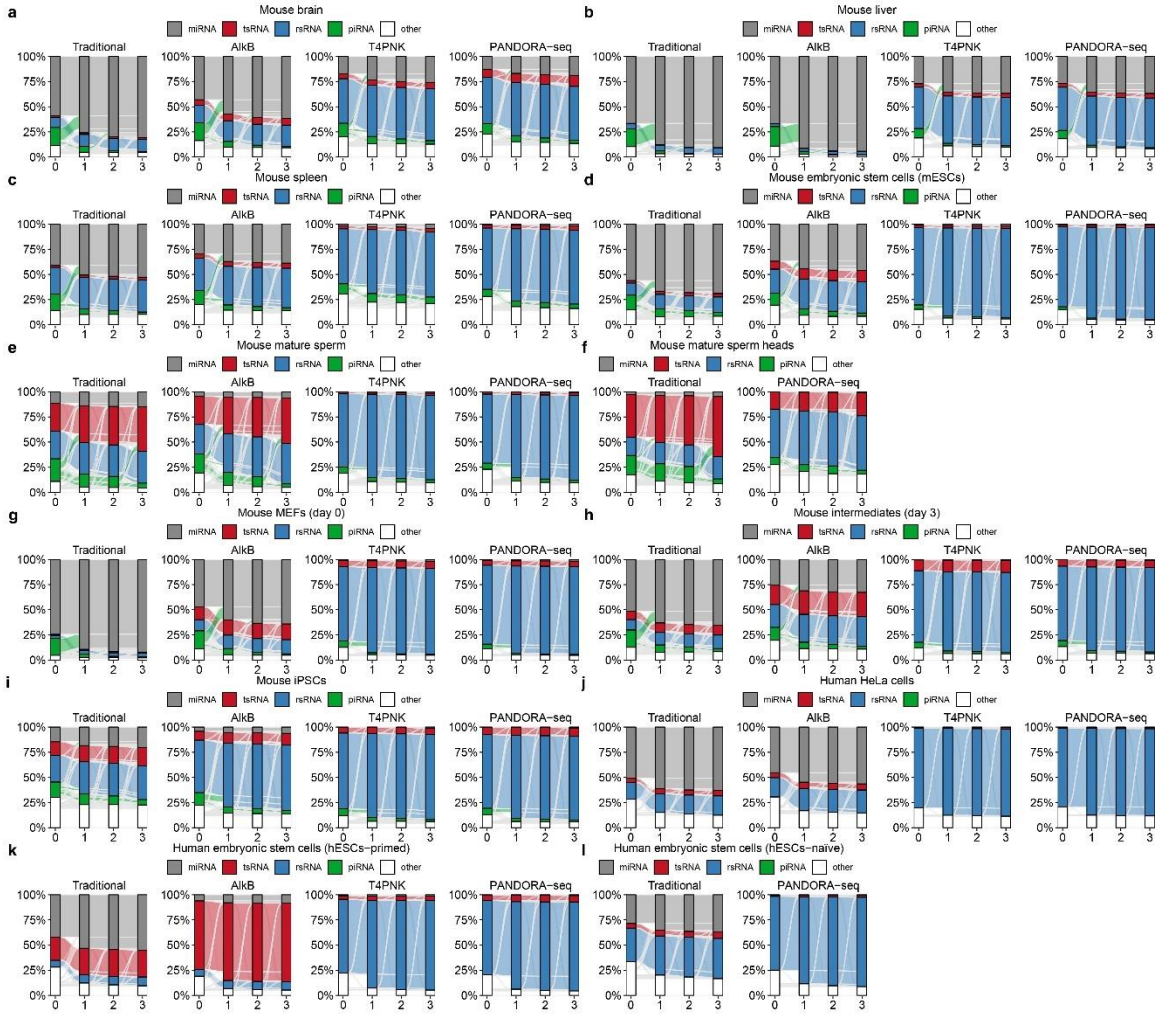

Figure 3.5: Annotation of mouse piRNA in non-germ cell tissue/cell types is not stable when 1-3 mismatches are allowed

When 1-3 mismatches are allowed for small RNAs matching, the piRNA annotation rate (but not other small RNAs types) show significant decrease in mouse tissue/ cell types (**a**) mouse brain, (**b**) mouse liver, (**c**) mouse spleen, (**d**) mouse embryonic stem cells, (**e**) mouse mature sperm, (**f**) mouse mature sperm heads, (**g**) mouse MEFs (day 0), (**h**) mouse intermediate cells (day 3), (**i**) mouse iPSCs. Very few piRNAs were annotated for human cell lines (**j**) human HeLa cells, (**k**) human hESCs-primed, and (**l**) human hESCs-naïve. These data suggest the annotated piRNAs in non-germ cell tissue/cell types could be due to database quality issue and their true identity awaits to be verified.

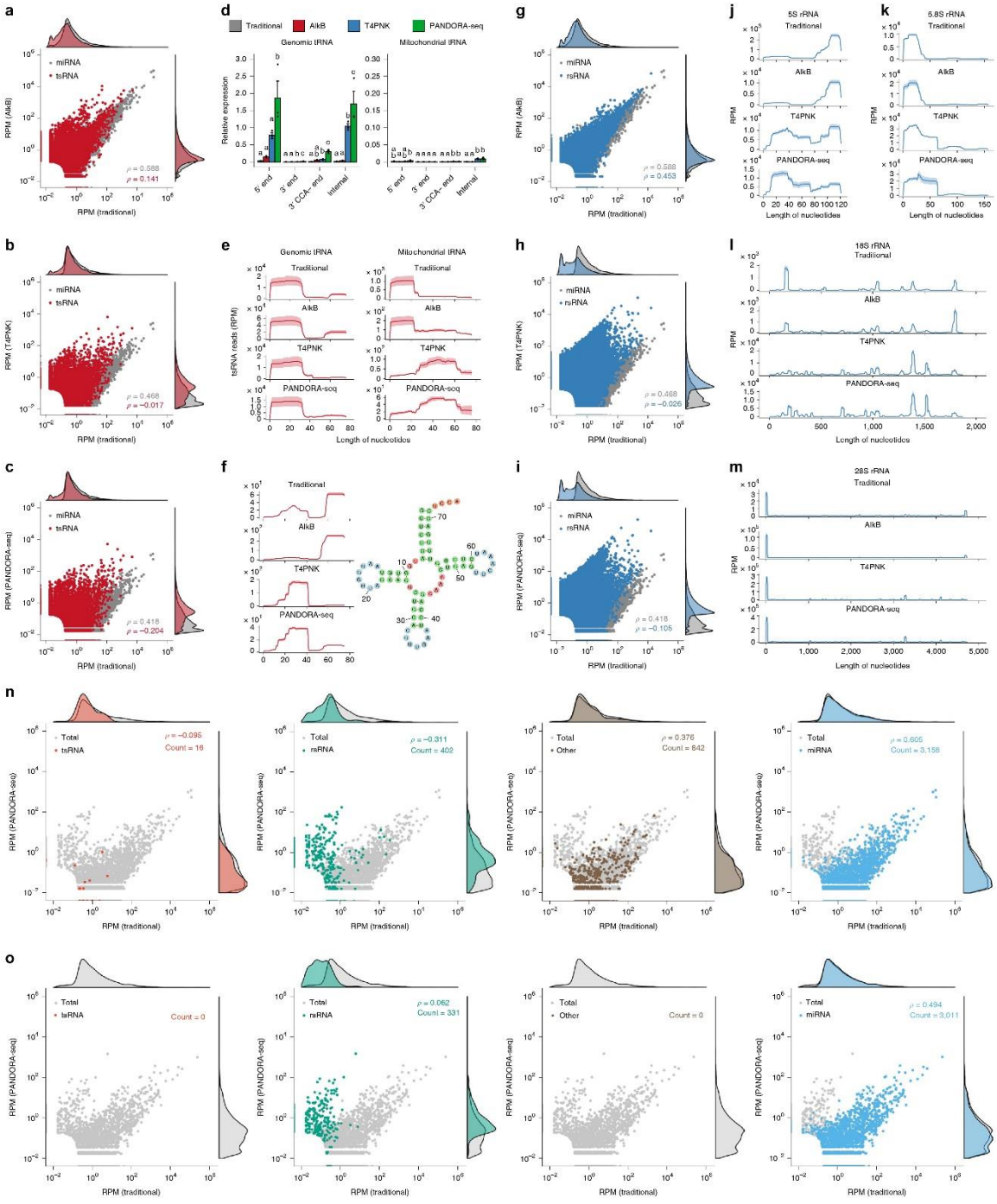
a
d
g
j
k
b
e
h
l
c
f
i
m
n
o
Traditional
AlkB
T4PNK
PANDORA-seq
Genomic tRNA
Mitochondrial tRNA
5S rRNA
5.8S rRNA
18S rRNA
28S rRNA
miRNA
tsRNA
rsRNA
Total
Other
RPM (traditional)
RPM (AlkB)
RPM (T4PNK)
RPM (PANDORA-seq)
Length of nucleotides
tRNA reads (RPM)
5' end
3' end
3' CCA- end
Internal

Figure 3.6: **Dissecting the effects of AlkB, T4PNK and PANDORA-seq on different** small RNA **populations in ESCs**

a-**c**, Scatter plots comparing profile changes in tsRNAs (red dots) and miRNAs (grey dots) detected using AlkB versus traditional (**a**), T4PNK versus traditional (**b**) and PANDORA-seq versus traditional protocols (c). ρ is the Spearman's correlation coefficient. **d**, tsRNA responses to AlkB, T4PNK and PANDORA-seq in regard to different origins (5' tsRNA, 3' tsRNA, 3' tsRNA-CCA end and internal tsRNAs). The *y* axes represent the relative expression level compared with total reads of miRNA ($n = 3$ biologically independent samples per bar). Different letters above the bars indicate statistically significant differences ($P < 0.05$). Same letters indicate $P \geq 0.05$. Statistical significance was determined by two-sided one-way ANOVA with uncorrected Fisher's LSD test. All data are plotted as means ± s.e.m. **e**, Overall length mapping showing the distribution of relative tsRNA reads from mature genomic (left) and mitochondrial (right) tRNA under different RNA-seq protocols. **f**, Dynamic response to different RNA-seq protocols (left) of a representative individual tsRNA (mouse tRNA-Gln-TTG-2; pictured right). **g**-**i**, Scatter plots comparing profile changes in rsRNAs (blue dots) and miRNAs (grey dots) detected using the following protocols: AlkB versus traditional (**g**), T4PNK versus traditional (**h**) and PANDORA-seq versus traditional (**i**). **j**-**m**, Comparison of rsRNA-generating loci by rsRNA mapping data on 5S rRNA (**j**), 5.8S rRNA (**k**), 18S rRNA (**l**) and 28S rRNA (**m**), detected using different RNA-seq protocols. **n**,**o**, Many of the previously annotated miRNAs from miRBase that showed upregulation under PANDORA-seq could also be annotated to other small RNA categories, as exemplified in mESCs (**n**) and primed hESCs (**o**). The mapping plots in **e**, **f** and **j**-**m** are presented as means ± s.e.m.

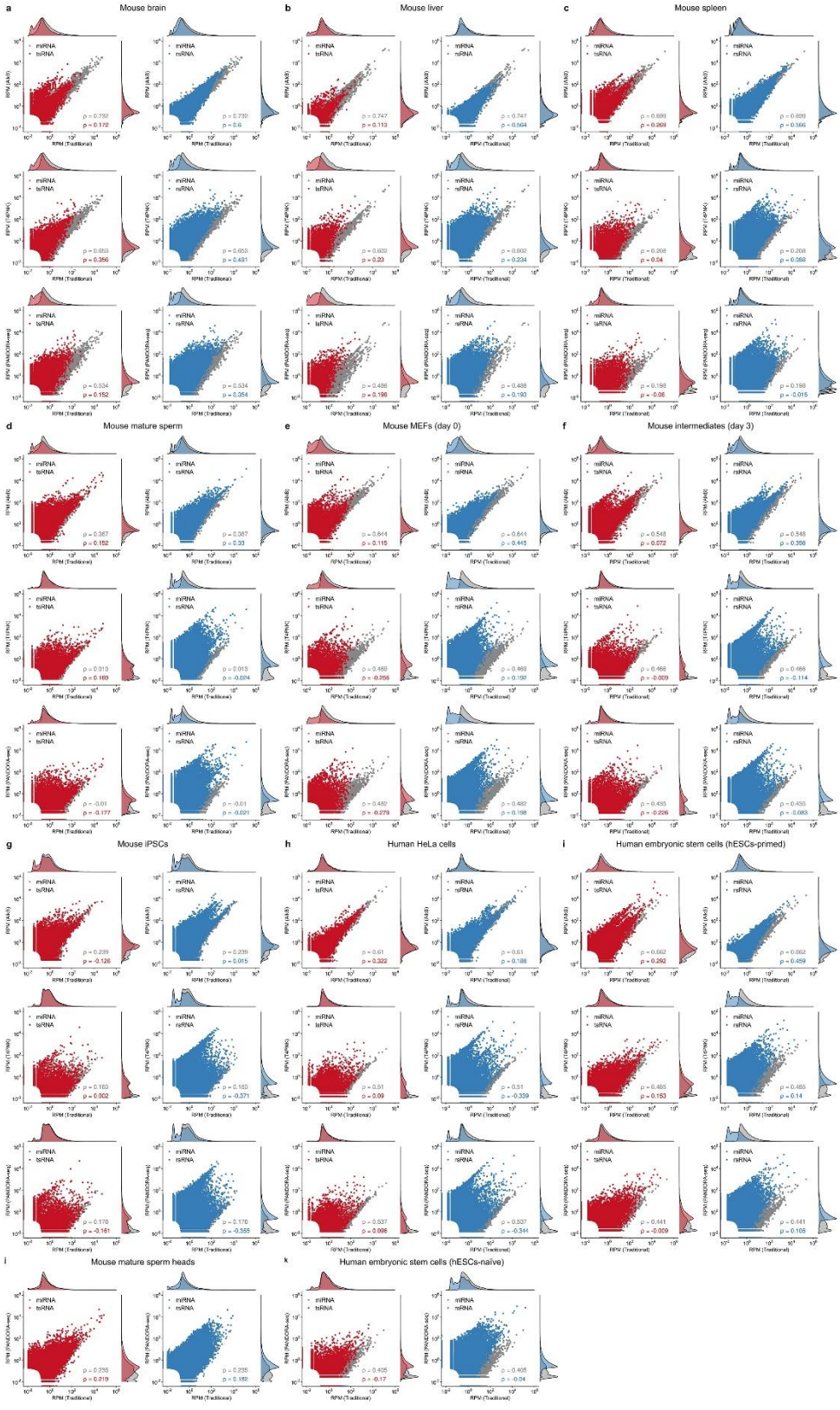
a
Mouse brain
b
Mouse liver
c
Mouse spleen
d
Mouse mature sperm
e
Mouse MEFs (day 0)
f
Mouse intermediates (day 3)
g
Mouse iPSCs
h
Human HeLa cells
i
Human embryonic stem cells (hESCs-primed)
j
Mouse mature sperm heads
k
Human embryonic stem cells (hESCs-naïve)
miRNA
tsRNA
rsRNA
RPM (Traditional)

Figure 3.7: Scattered plot comparison of profile changes in tsRNAs and rsRNAs compared to miRNAs under different treatment protocol

Scattered plot comparison of profile changes in tsRNAs (red dots) and rsRNAs (blue dots) compared to miRNAs (gray dots) under AlkB vs traditional, T4PNK vs traditional and PANDORA-seq vs traditional in (**a**) mouse brain, (**b**) mouse liver, (**c**) mouse spleen, (**d**) mouse mature sperm, (**e**) mouse MEFs (day 0), (**f**) mouse intermediate cells (day 3), (**g**) mouse iPSCs, (**h**) human HeLa cells, (**i**) human hESCs-primed, (**j**) mouse mature sperm heads, and (**k**) human hESCs-naïve.

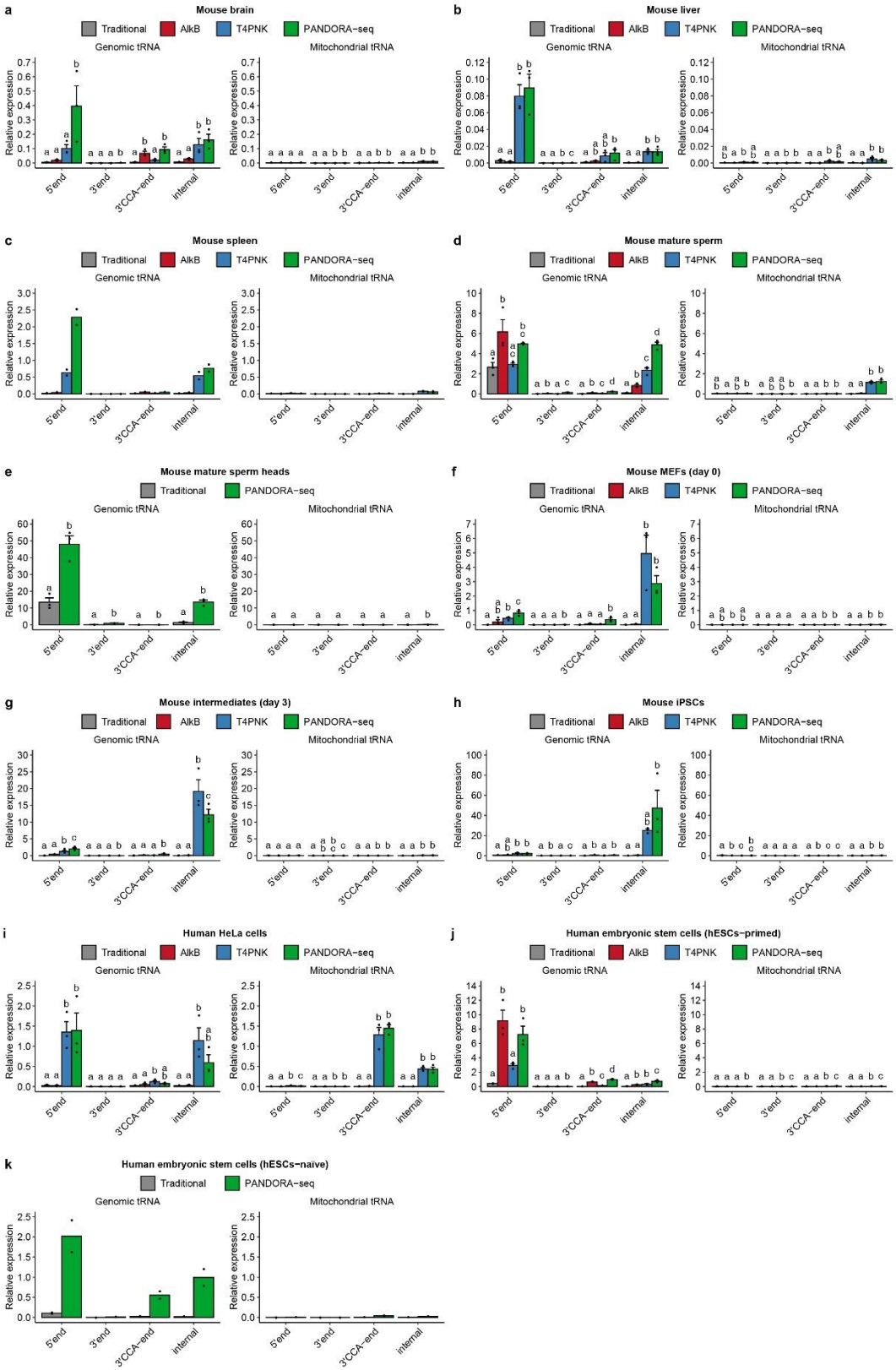
a
Mouse brain
Traditional
AlkB
T4PNK
PANDORA-seq
Genomic tRNA
Mitochondrial tRNA
Relative expression
5'end
3'end
3'CCA-end
internal
b
Mouse liver
c
Mouse spleen
d
Mouse mature sperm
e
Mouse mature sperm heads
f
Mouse MEFs (day 0)
g
Mouse intermediates (day 3)
h
Mouse iPSCs
i
Human HeLa cells
j
Human embryonic stem cells (hESCs-primed)
k
Human embryonic stem cells (hESCs-naïve)

Figure 3.8: The tsRNA responses to AlkB, T4PNK and PANDORA-seq in regard to different tsRNA origin (5' tsRNA, 3' tsRNA, 3' tsRNA with CCA end, and internal tsRNAs)

(**a**), mouse brain, (**b**) mouse liver, (**c**) mouse spleen, (**d**) mouse mature sperm, (**e**) mouse mature sperm heads, (**f**) mouse MEFs (day 0), (**g**) mouse intermediate cells (day 3), (**h**) mouse iPSCs, (**i**) human HeLa cells, (**j**) human hESCs-primed, and (**k**) human hESCs-naïve. For **a**-**b**, **d**-**j**, data are plotted as mean ± SEM (n=3 biologically independent samples in each bar); for c,k, n=2 biologically independent samples in each bar. Different letters above bars indicate statistical difference, $P < 0.05$; same letters indicate $P \geq 0.05$ (two-sided, one-way ANOVA, uncorrected Fisher's LSD test).

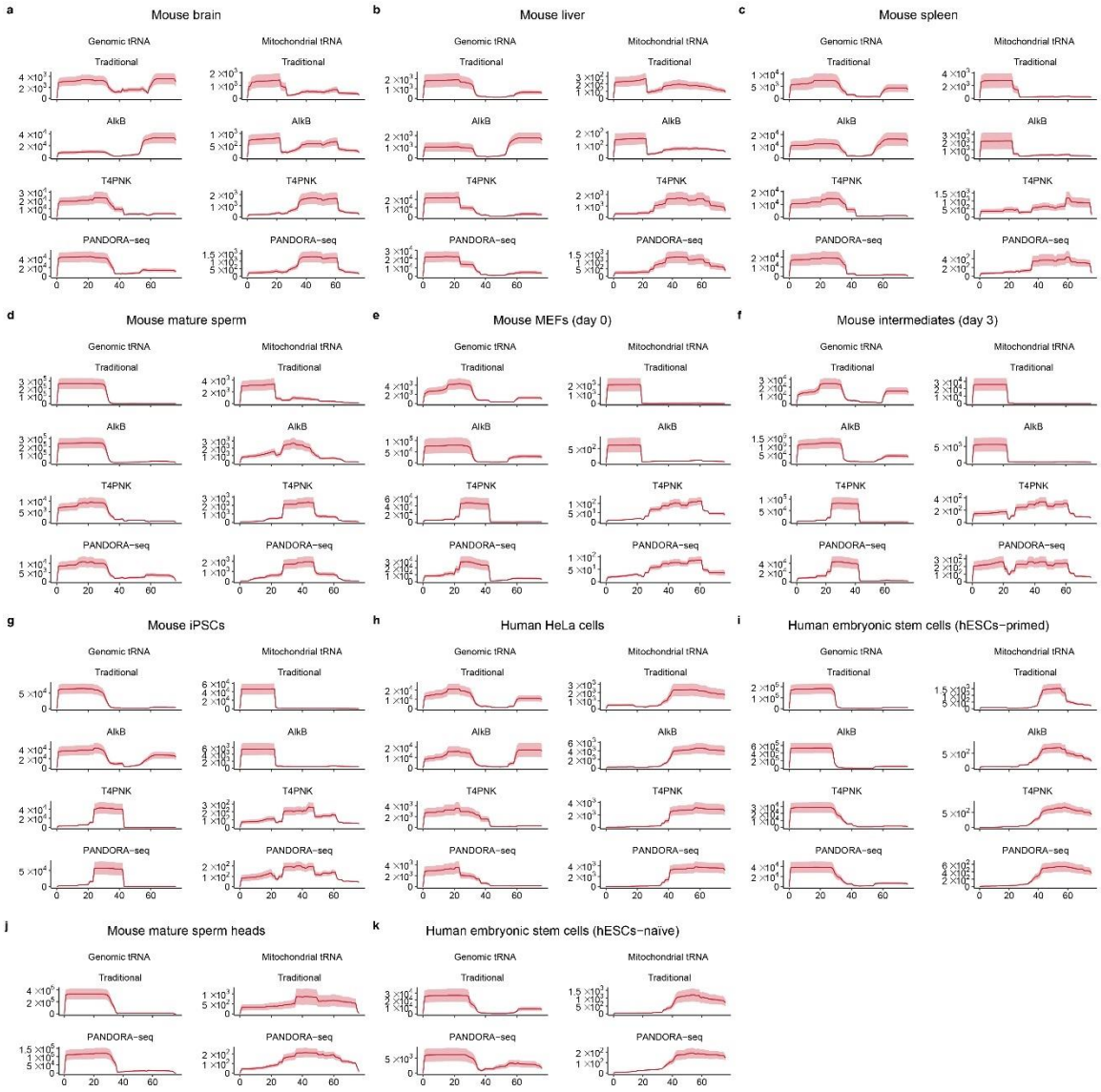

Figure 3.9: Overall length mapping of tsRNA reads in genomic and mitochondrial tRNA under different RNA-seq protocol

Overall mapping of all tsRNAs on a tRNA length scale revealed the preferential loci from which tsRNAs are derived from the mature full tRNA under traditional protocol and different enzymatic treatments. (**a**) mouse brain, (**b**) mouse liver, (**c**) mouse spleen, (**d**) mouse mature sperm, (**e**) mouse MEFs (day 0), (**f**) mouse intermediate cells (day 3), (**g**) mouse iPSCs, (**h**) human HeLa cells, (**i**) human hESCs-primed, (**j**) mouse mature sperm heads, and (**k**) human hESCs-naïve. Mapping plots are presented as mean ± SEM.

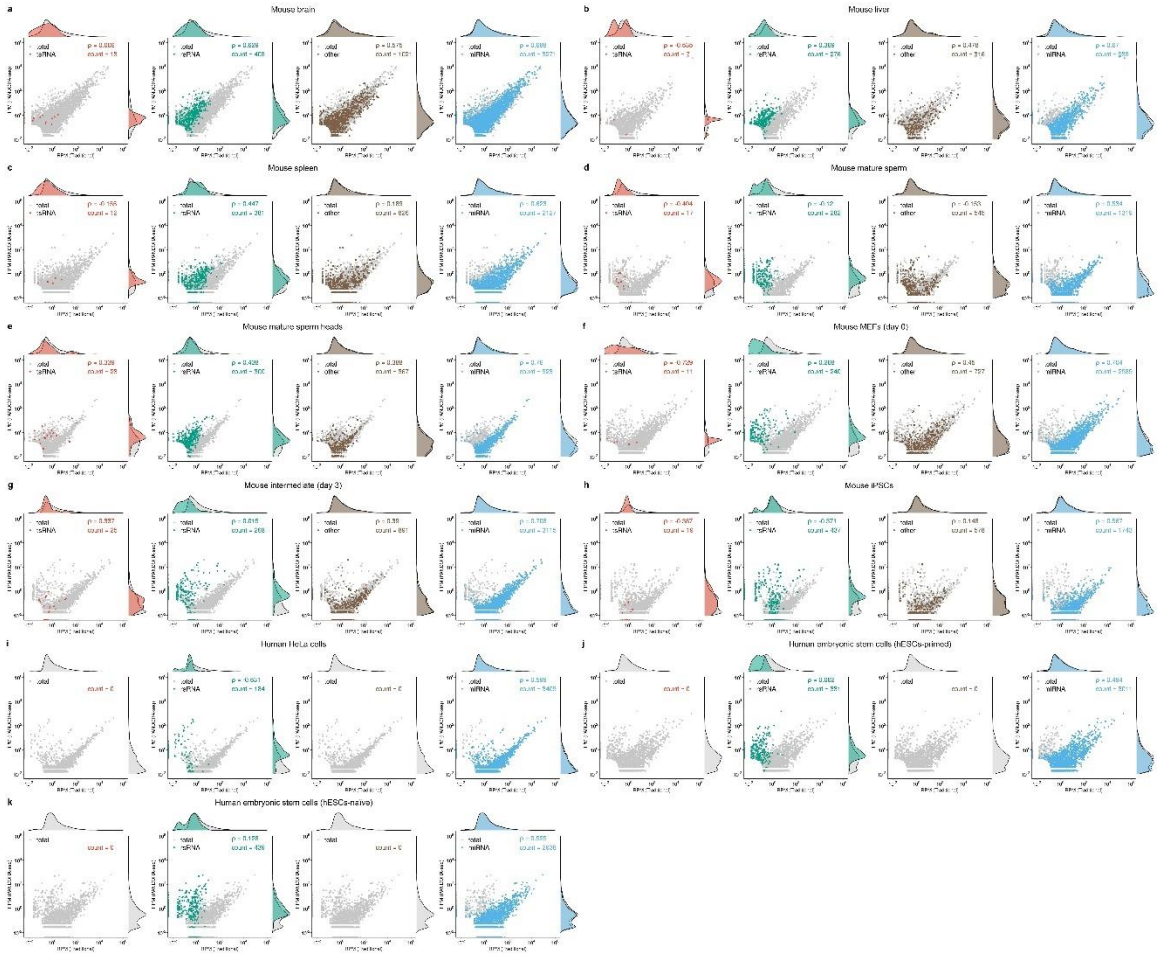

Figure 3.10: The miRNAs that showing sensitive response to PANDORA-seq are in fact rsRNAs

Previously annotated miRNAs from miRBase that showed upregulation under PANDORA-seq could also annotate to rsRNAs (with one mismatch tolerance), as shown in (**a**) mouse brain, (**b**) mouse liver, (**c**) mouse spleen, (**d**) mouse mature sperm, (**e**) mouse mature sperm heads, (**f**) mouse MEFs (day 0), (**g**) mouse intermediate cells (day 3), (**h**) mouse iPSCs, (**i**) human HeLa cells, and (**j**) human hESCs-naïve.

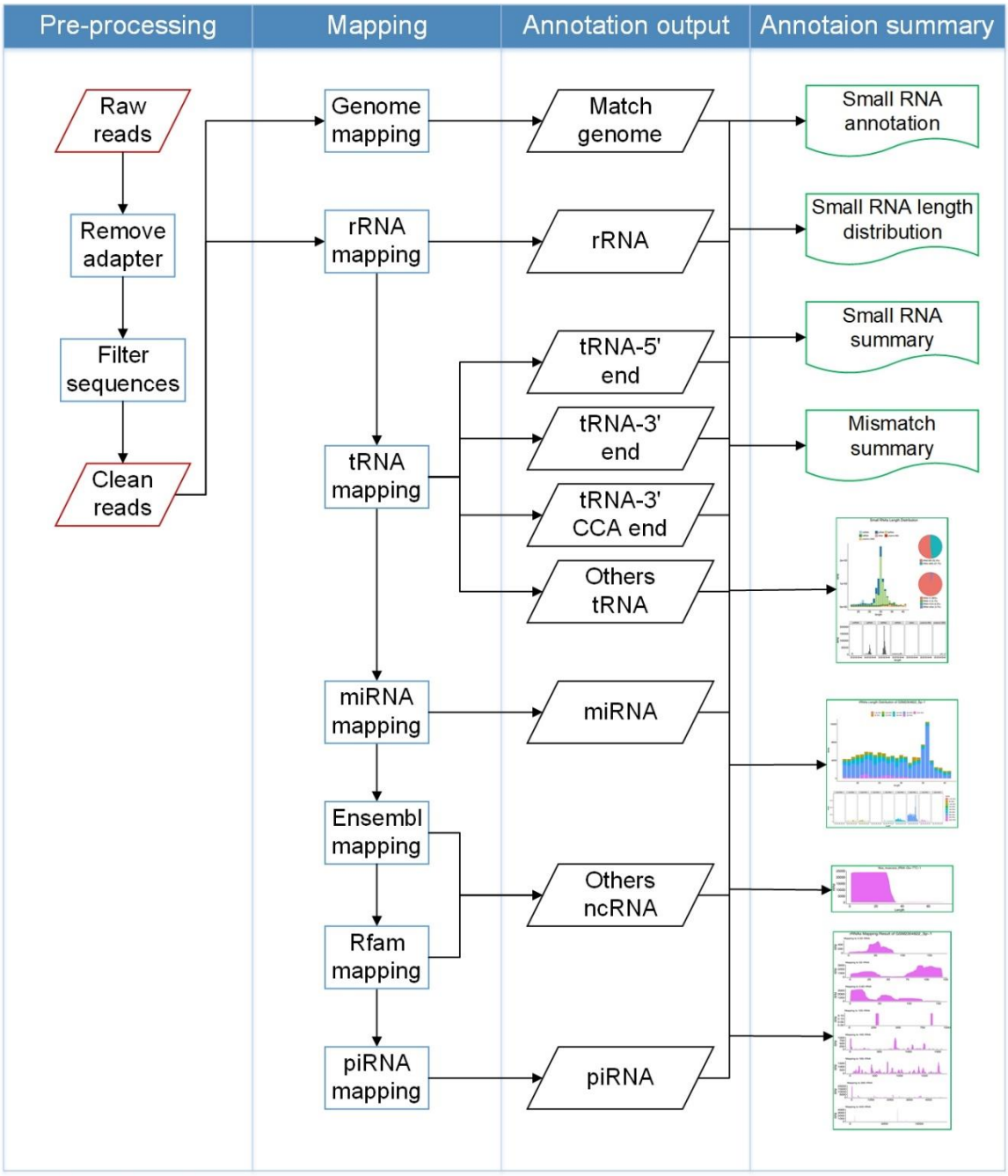

Figure 3.11: Workflow of SPORTS1.1

SPORTS1.1 has changed the mapping order priority of miRNA and piRNA compared with SPORTS1.0.

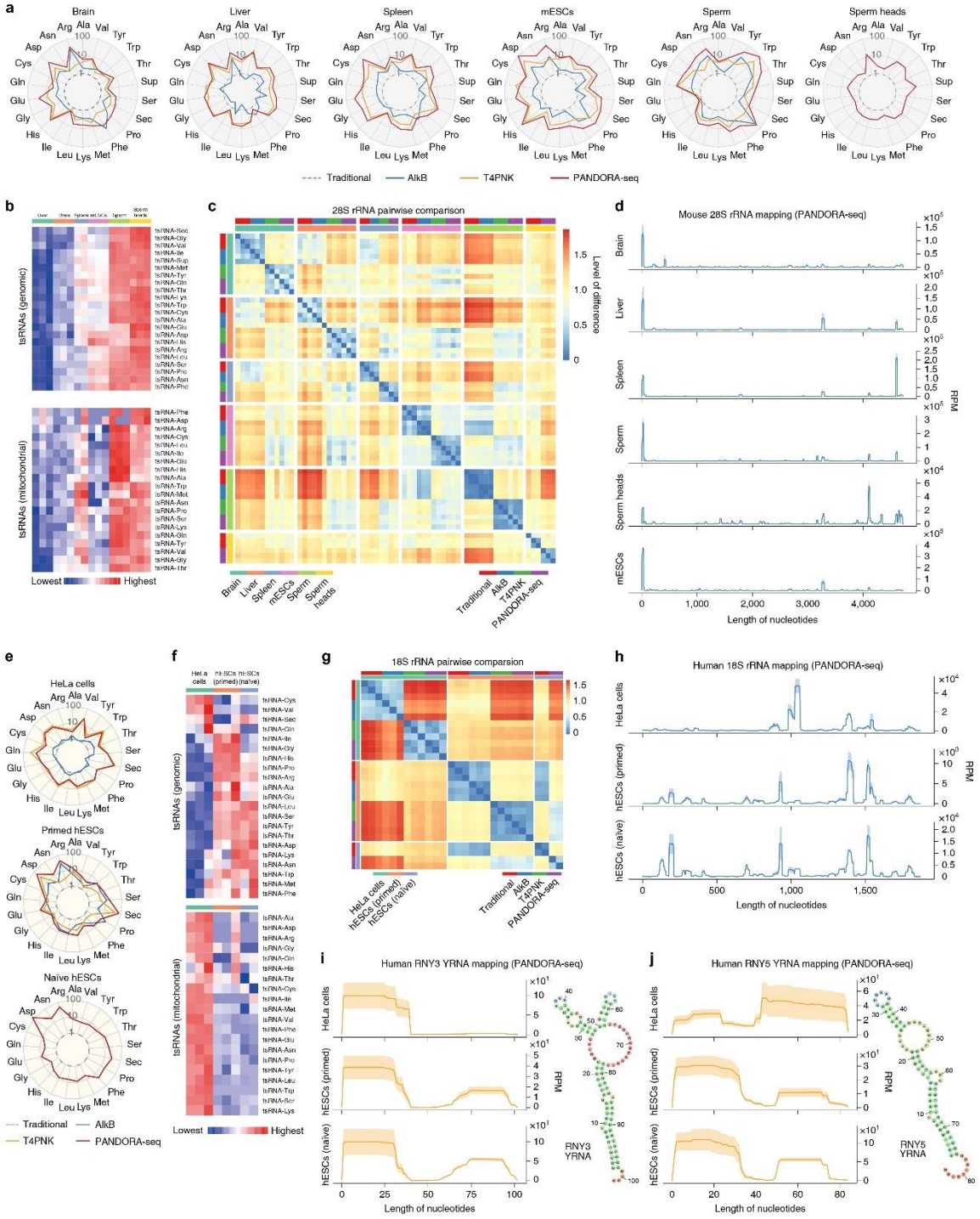
a
Brain
Liver
Spleen
mESCs
Sperm
Sperm heads
Traditional
AlkB
T4PNK
PANDORA-seq
b
tsRNAs (genomic)
tsRNAs (mitochondrial)
Lowest
Highest
c
28S rRNA pairwise comparison
Level of difference
d
Mouse 28S rRNA mapping (PANDORA-seq)
RPM
Length of nucleotides
e
HeLa cells
Primed hESCs
Naïve hESCs
f
tsRNAs (genomic)
tsRNAs (mitochondrial)
g
18S rRNA pairwise comparsion
h
Human 18S rRNA mapping (PANDORA-seq)
i
Human RNY3 YRNA mapping (PANDORA-seq)
RNY3 YRNA
j
Human RNY5 YRNA mapping (PANDORA-seq)
RNY5 YRNA

Figure 3.12: **Tissue- and cell type-specific expression of tsRNAs and rsRNAs in mice and humans**

a, Radar plots showing the different sensitivities of five different mouse tissue or cell types in regard to different RNA-seq protocols. The numbers (1, 10 and 100) on the radius represent log values. **b**, Heatmaps showing the tsRNA (genomic and mitochondrial) relative expression levels (normalized to total miRNA levels and based on a log2-transformed scale in the row direction) of five different mouse tissue or cell types, as detected by PANDORA-seq. **c**, Pairwise comparison matrix showing the overall expression pattern difference of rsRNAs (derived from 28S rRNAs) under different RNA-seq protocols across five mouse tissue or cell types. Blue represents more similarity and red more difference. **d**, Comparison of rsRNA-generating loci from mouse 28S rRNA revealed distinct patterns across tissue and cell types. **e**, Radar plots showing the different sensitivities of three different human cell types in regard to different RNA-seq protocols. The numbers (1, 10 and 100) on the radius represent log values. **f**, Heatmaps showing the tsRNA (genomic and mitochondrial) relative expression levels (normalized to total miRNA levels and based on a log2-transformed scale in the row direction) of three different human cell types, as detected by PANDORA-seq. **g**, Pairwise comparison matrix showing the overall expression pattern difference of rsRNAs (derived from 18S rRNAs) identified using different RNA-seq protocols across three human cell types. Blue represents more similarity and red more difference. **h**, Comparison of rsRNA-generating loci from human 18S rRNA revealed distinct patterns across tissue and cell types. **i**,**j**, Exemplary human ysRNAs (RNY3 (**i**) and RNY5 (**j**)) that are differentially expressed between different cell types, as determined by PANDORA-seq. The mapping plots in **d**, **h**, **i** and **j** are presented as means ± s.e.m.

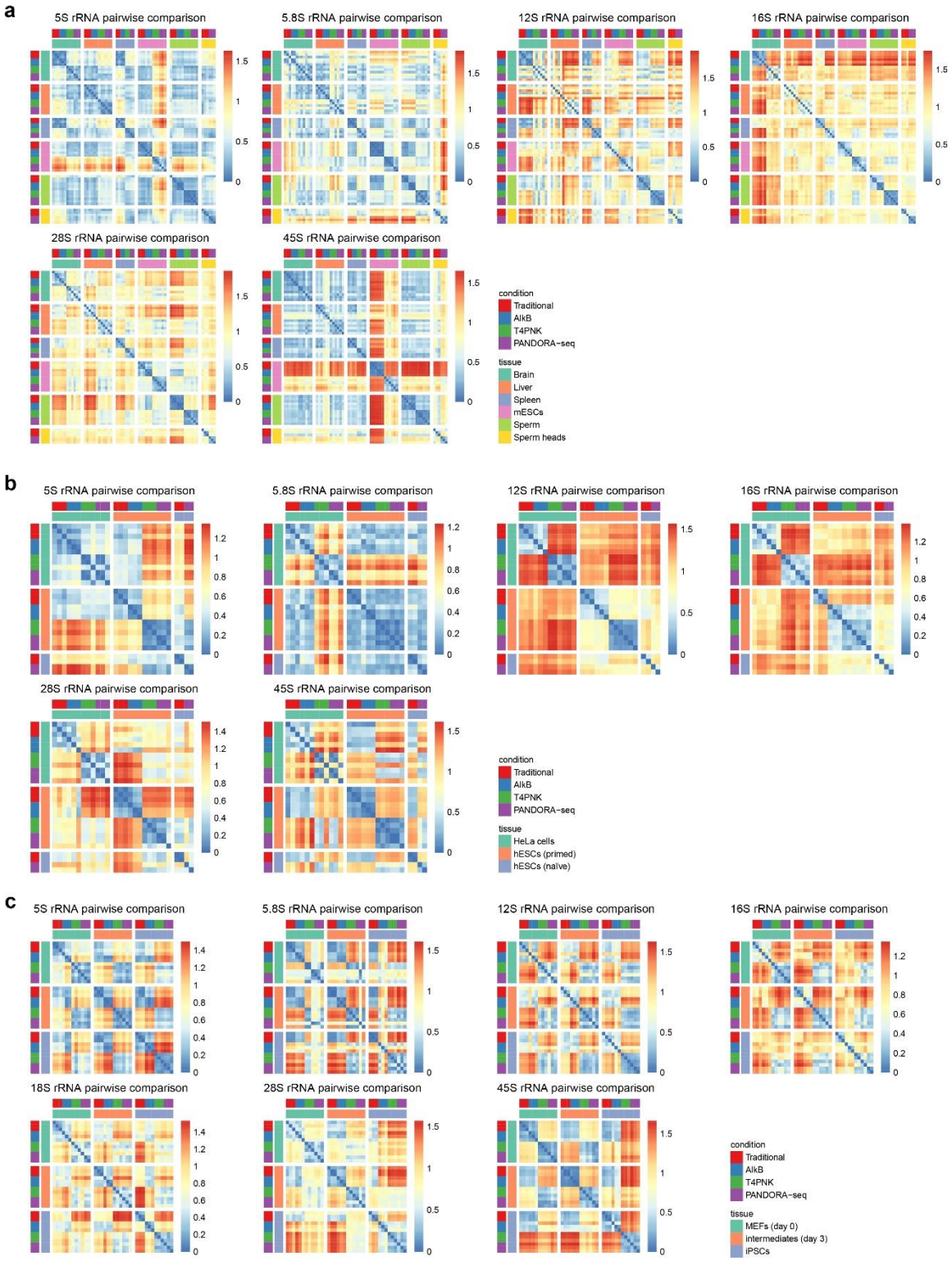
a
5S rRNA pairwise comparison
5.8S rRNA pairwise comparison
12S rRNA pairwise comparison
16S rRNA pairwise comparison
28S rRNA pairwise comparison
45S rRNA pairwise comparison
condition
Traditional
AlkB
T4PNK
PANDORA-seq
tissue
Brain
Liver
Spleen
mESCs
Sperm
Sperm heads
b
5S rRNA pairwise comparison
5.8S rRNA pairwise comparison
12S rRNA pairwise comparison
16S rRNA pairwise comparison
28S rRNA pairwise comparison
45S rRNA pairwise comparison
condition
Traditional
AlkB
T4PNK
PANDORA-seq
tissue
HeLa cells
hESCs (primed)
hESCs (naive)
c
5S rRNA pairwise comparison
5.8S rRNA pairwise comparison
12S rRNA pairwise comparison
16S rRNA pairwise comparison
18S rRNA pairwise comparison
28S rRNA pairwise comparison
45S rRNA pairwise comparison
condition
Traditional
AlkB
T4PNK
PANDORA-seq
tissue
MEFs (day 0)
intermediates (day 3)
iPSCs

Figure 3.13: The pairwise comparison matrices showing the differential expression pattern of rsRNAs under different RNA-seq protocol across tissues and cells

a, Pairwise comparison matrices for six mouse tissue/cell types, including 5S rRNA, 5.8S rRNA, mitochondrial 12S rRNA, mitochondrial 16S rRNA, 28S rRNA and 45S rRNA. Color bar: from blue (more similar) to red (more different). **b**, Pairwise comparison matrices for three human cell types, including 5S rRNA, 5.8S rRNA, mitochondrial 12S rRNA, mitochondrial 16S rRNA, 28S rRNA and 45S rRNA. Color bar: from blue (more similar) to red (more different). **c**, Pairwise comparison matrices for during mouse iPSC reprogramming, including 5S rRNA, 5.8S rRNA, mitochondrial 12S rRNA, mitochondrial 16S rRNA, 18S rRNA, 28S rRNA and 45S rRNA. Color bar: from blue (more similar) to red (more different).

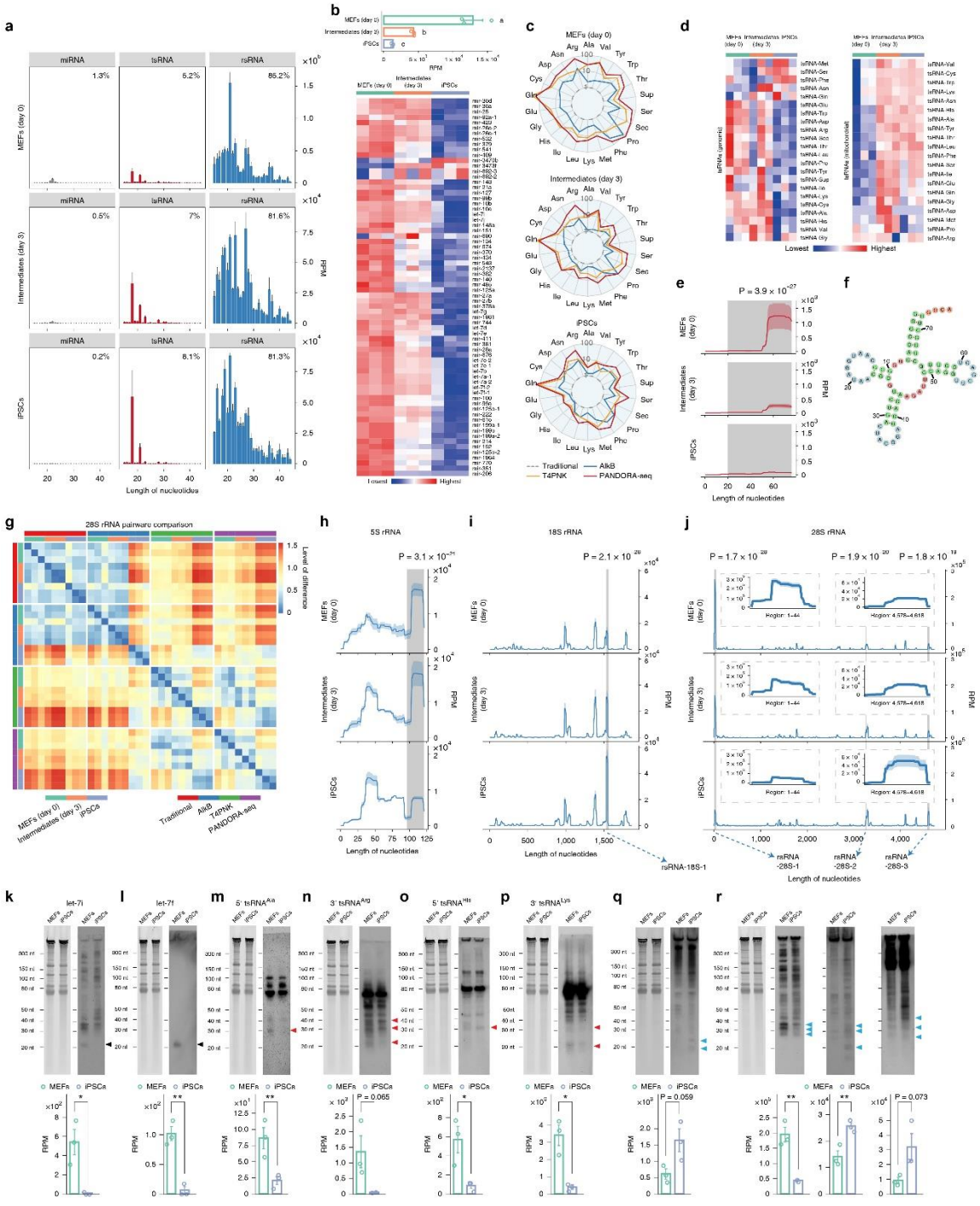
a
miRNA
tsRNA
rsRNA
MEFs (day 0)
Intermediates (day 3)
iPSCs
1.3%
5.2%
86.2%
0.5%
7%
81.6%
0.2%
8.1%
81.3%
RPM
Length of nucleotides
b
MEFs (day 0)
Intermediates (day 3)
iPSCs
RPM
Lowest
Highest
c
MEFs (day 0)
Intermediates (day 3)
iPSCs
Traditional
T4PNK
AlkB
PANDORA-seq
d
tsRNA (genomic)
tsRNA (mitochondrial)
Lowest
Highest
e
P = 3.9 × 10^-27
MEFs (day 0)
Intermediates (day 3)
iPSCs
RPM
Length of nucleotides
f
g
28S rRNA pairwise comparison
Level of difference
MEFs (day 0)
Intermediates (day 3)
iPSCs
Traditional
AlkB
T4PNK
PANDORA-seq
h
5S rRNA
P = 3.1 × 10^-21
MEFs (day 0)
Intermediates (day 3)
iPSCs
RPM
Length of nucleotides
i
18S rRNA
P = 2.1 × 10^-29
MEFs (day 0)
Intermediates (day 3)
iPSCs
RPM
Length of nucleotides
rsRNA-18S-1
j
28S rRNA
P = 1.7 × 10^-29
P = 1.9 × 10^-20
P = 1.8 × 10^-19
Region: 1-44
Region: 4,578-4,618
MEFs (day 0)
Intermediates (day 3)
iPSCs
RPM
Length of nucleotides
rsRNA -28S-1
rsRNA -28S-2
rsRNA -28S-3
k
let-7i
l
let-7f
m
5' tsRNA^Ala
n
3' tsRNA^Arg
o
5' tsRNA^His
p
3' tsRNA^Lys
q
r
MEFs
iPSCs
RPM
P = 0.065
P = 0.059
P = 0.073

Figure 3.14: **PANDORA-seq reveals that tsRNAs and rsRNAs are dynamically regulated during MEF reprogramming to iPSCs (day 0) to intermediate (day 3) and iPSC stages**

a, Dynamic changes in small RNA distribution during iPSC reprogramming from MEFs (day 0) to intermediate (day 3) and iPSC stages (means ± s.e.m.), as determined by PANDORA-seq. **b**, Bar plot (top) and heatmap (bottom) showing miRNA expression changes (based on RPM values) during cell reprogramming using PANDORA-seq. **c**, Radar plots showing the different sensitivities of MEFs, intermediate stages and iPSCs in regard to different RNA-seq protocols. **d**, Heatmaps showing tsRNA (genomic and mitochondrial) expression levels (based on RPM values) during cell reprogramming using PANDORA-seq. **e**,**f**, Dynamic changes (**e**) of a representative tsRNA (tRNA-Arg-ACG-1; pictured in **f**) during the reprogramming process, as determined by PANDORA-seq. **g**, Pairwise comparison matrix showing the correlation of rsRNAs (derived from 28S rRNA) under different RNA-seq protocols during cell reprogramming. Blue signifies more similarity and red more difference. Note that PANDORA-seq revealed a more dynamic change across different stages than traditional RNA-seq. **h**-**j**, Comparison of rsRNA-generating loci by rsRNA mapping data on 5S rRNA (**h**), 18S rRNA (**i**) and 28S rRNA (**j**) under PANDORA-seq, showing dynamic changes during the reprogramming process. In **e** and **h**-**j**, the shaded peaks are marked with the significance value for the comparison between MEFs and iPSCs, as determined by two-way ANOVA. The mapping plots in **e** and **h**-**j** are presented as means ± s.e.m. The highlighted windows in **i** and **j** show the detailed read mappings of rsRNA-18S-1 (**i**) and rsRNA-28S-1, -2 and -3 (**j**), which were used for northern blot validation in **q** and **r** (see arrows). **k**-**r**, Northern blot examination of representative small RNAs (let-7i (**k**), let-7f (**l**), 5' $tsRNA^{Ala}$ (**m**), 3' $tsRNA^{Arg}$ (**n**), 5' $tsRNA^{His}$ (**o**), 3' $tsRNA^{Lys}$ (**p**), rsRNA-18S-1 (**q**) and rsRNA-28S-1, -2 and -3 (**r**)) was performed in MEFs and iPSCs. The northern blot signals (similar results were obtained in three independent experiments) showed overall consistency with their corresponding sequencing reads in MEFs and iPSCs, as revealed by PANDORA-seq ($n = 3$ biologically independent samples per bar). Black arrowheads, miRNAs; red arrowheads, tsRNAs; blue arrowheads, rsRNAs. The data represent means ± s.e.m. Statistical significance was determined by two-sided Student's *t*-test (*$P < 0.05$; **$P < 0.01$).

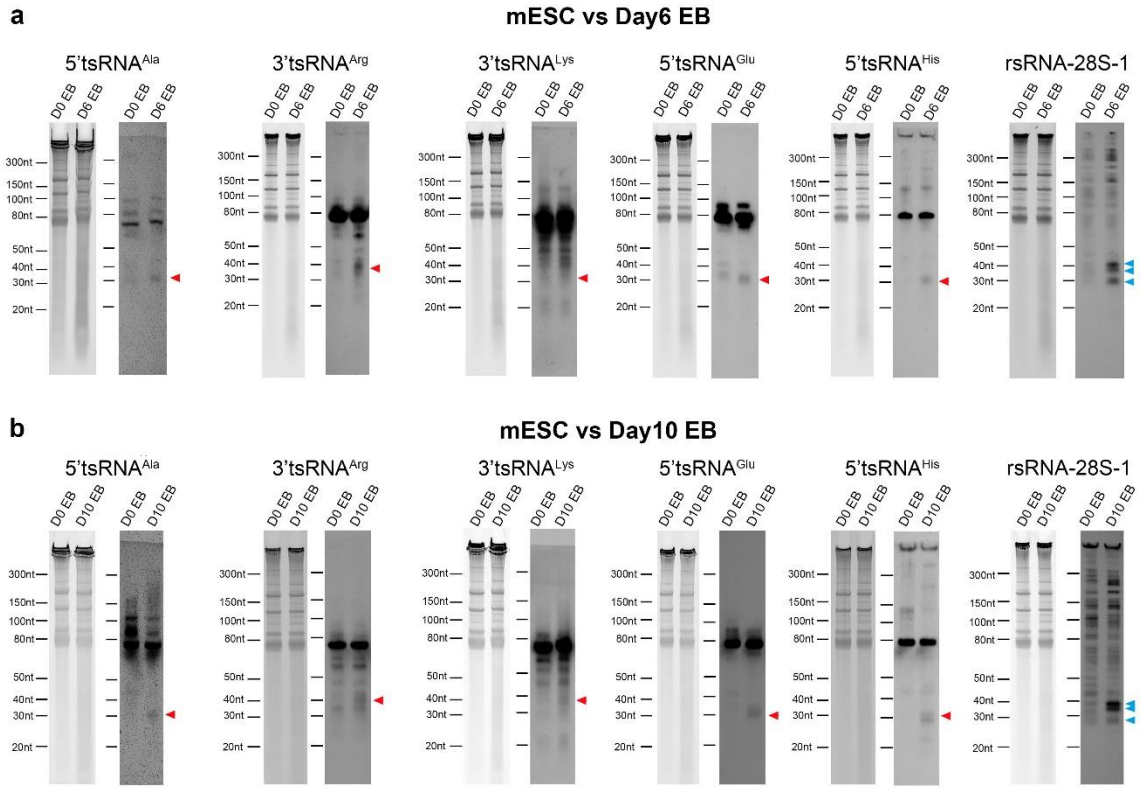

Figure 3.15: Northern blot analyses of tsRNA/rsRNA (that is, tsRNA^{Ala}, tsRNA^{Arg}, tsRNA^{Glu}, tsRNA^{His}, tsRNA^{Lys} and rsRNA-28S-1) changes during mESC to embryoid body differentiation

(**a**) mESC vs Day6 EB; (**b**) mESC vs Day10 EB. Red arrowhead: tsRNAs; Blue arrowhead: rsRNAs. Similar results were obtained in 3 independent experiments for rsRNA-28S-1; and in 2 independent experiments for tsRNA^{Ala}, tsRNA^{Arg}, tsRNA^{Glu}, tsRNA^{His}, and tsRNA^{Lys}. EB: embryoid body.

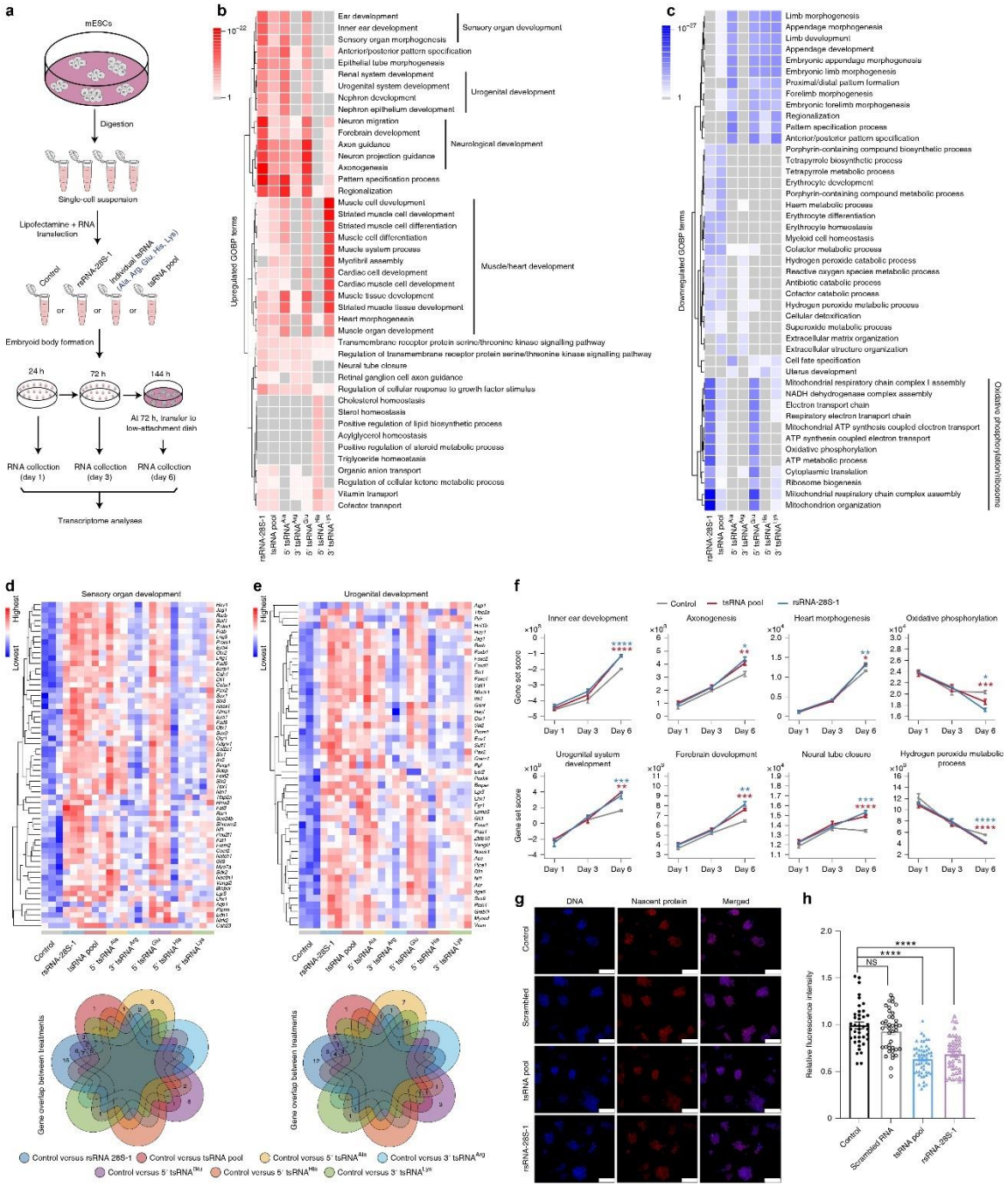
a
mESCs
Digestion
Single-cell suspension
Lipofectamine + RNA translection
Control
rsRNA-28S-1
Individual tsRNA (Ala, Arg, Glu, His, Lys)
tsRNA pool
or
Embryoid body formation
24 h
72 h
144 h
At 72 h, transfer to low-attachment dish
RNA collection (day 1)
RNA collection (day 3)
RNA collection (day 6)
Transcriptome analyses
b
Upregulated GOBP terms
Ear development
Inner ear development
Sensory organ morphogenesis
Anterior/posterior pattern specification
Epithelial tube morphogenesis
Renal system development
Urogenital system development
Nephron development
Nephron epithelium development
Neuron migration
Forebrain development
Axon guidance
Neuron projection guidance
Axonogenesis
Pattern specification process
Regionalization
Muscle cell development
Striated muscle cell development
Striated muscle cell differentiation
Muscle cell differentiation
Muscle system process
Myofibril assembly
Cardiac cell development
Cardiac muscle cell development
Muscle tissue development
Striated muscle tissue development
Heart morphogenesis
Muscle organ development
Transmembrane receptor protein serine/threonine kinase signalling pathway
Regulation of transmembrane receptor protein serine/threonine kinase signalling pathway
Neural tube closure
Retinal ganglion cell axon guidance
Regulation of cellular response to growth factor stimulus
Cholesterol homeostasis
Sterol homeostasis
Positive regulation of lipid biosynthetic process
Acylglycerol homeostasis
Positive regulation of steroid metabolic process
Triglyceride homeostasis
Organic anion transport
Regulation of cellular ketone metabolic process
Vitamin transport
Cofactor transport
Sensory organ development
Urogenital development
Neurological development
Muscle/heart development
rsRNA-28S-1
tsRNA pool
5′ tsRNA^Ala
3′ tsRNA^Arg
5′ tsRNA^Glu
5′ tsRNA^His
3′ tsRNA^Lys
c
Downregulated GOBP terms
Limb morphogenesis
Appendage morphogenesis
Limb development
Appendage development
Embryonic appendage morphogenesis
Embryonic limb morphogenesis
Proximal/distal pattern formation
Forelimb morphogenesis
Embryonic forelimb morphogenesis
Regionalization
Pattern specification process
Anterior/posterior pattern specification
Porphyrin-containing compound biosynthetic process
Tetrapyrrole biosynthetic process
Tetrapyrrole metabolic process
Erythrocyte development
Porphyrin-containing compound metabolic process
Haem metabolic process
Erythrocyte differentiation
Erythrocyte homeostasis
Myeloid cell homeostasis
Cofactor metabolic process
Hydrogen peroxide catabolic process
Reactive oxygen species metabolic process
Antibiotic catabolic process
Cofactor catabolic process
Hydrogen peroxide metabolic process
Cellular detoxification
Superoxide metabolic process
Extracellular matrix organization
Extracellular structure organization
Cell fate specification
Uterus development
Mitochondrial respiratory chain complex I assembly
NADH dehydrogenase complex assembly
Electron transport chain
Respiratory electron transport chain
Mitochondrial ATP synthesis coupled electron transport
ATP synthesis coupled electron transport
Oxidative phosphorylation
ATP metabolic process
Cytoplasmic translation
Ribosome biogenesis
Mitochondrial respiratory chain complex assembly
Mitochondrion organization
Oxidative phosphorylation/ribosome
d
Sensory organ development
Highest
Lowest
Control
rsRNA-28S-1
tsRNA pool
5′ tsRNA^Ala
3′ tsRNA^Arg
5′ tsRNA^Glu
5′ tsRNA^His
3′ tsRNA^Lys
Gene overlap between treatments
e
Urogenital development
f
Control
tsRNA pool
rsRNA-28S-1
Inner ear development
Axonogenesis
Heart morphogenesis
Oxidative phosphorylation
Urogenital system development
Forebrain development
Neural tube closure
Hydrogen peroxide metabolic process
Gene set score
Day 1
Day 3
Day 6
Control versus rsRNA 28S-1
Control versus tsRNA pool
Control versus 5′ tsRNA^Ala
Control versus 3′ tsRNA^Arg
Control versus 5′ tsRNA^Glu
Control versus 5′ tsRNA^His
Control versus 3′ tsRNA^Lys
g
DNA
Nascent protein
Merged
Control
Scrambled
tsRNA pool
rsRNA-28S-1
h
Relative fluorescence intensity
NS
Control
Scrambled RNA
tsRNA pool
rsRNA-28S-1

Figure 3.16: **Transfection of tsRNA or rsRNA impacts mESC lineage differentiation and cell translation**

a, Schematic of the procedure of tsRNA/rsRNA transfection (that is, rsRNA-28S-1, 5' $tsRNA^{Ala}$, 3' $tsRNA^{Arg}$, 5' $tsRNA^{Glu}$, 5' $tsRNA^{His}$, 3' $tsRNA^{Lys}$ and a pool of the five aforementioned tsRNAs (tsRNA pool)), followed by embryoid body formation and transcriptome RNA-seq at days 1, 3 and 6 after transfection. **b**,**c**, Top-ranked upregulated (**b**) and downregulated GOBP terms (**c**) in day 6 embryoid bodies after each tsRNA/rsRNA transfection compared with the control. **d**,**e**, Expression heatmaps of the differentially expressed genes from the representative GOBP terms sensory organ development (**d**) and urogenital development (**e**). Similar analyses for other pathways are shown in **Figure 3.17 a-d**. The Venn diagram beneath each heatmap shows the numbers of overlapped dysregulated genes under different tsRNA/rsRNA transfections. **f**, Gene set score analyses of the representative GOBP terms during days 1, 3 and 6 of embryoid body differentiation under control, rsRNA-28S-1 or pooled tsRNA transfection ($n = 3$ biologically independent samples at each time point). Statistical significance was determined by two-sided one-way ANOVA (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; ****$P < 0.0001$). Data represent means ± s.e.m. **g**,**h**, Global translational assay results. Representative pictures of nascent protein syntheses (**g**) and protein synthesis rates 24 h after transfection of the control (vehicle only; $n = 40$), scrambled RNA ($n = 41$), rsRNA-28S-1 ($n = 44$) and pooled tsRNA ($n = 54$) (**h**) are shown. Scale bars in **g**, 100 μm. The ESC clones were from three independent biological experiments. Statistical significance was determined by two-sided one-way ANOVA (****$P < 0.0001$). NS, not significant. Data represent means ± s.e.m.

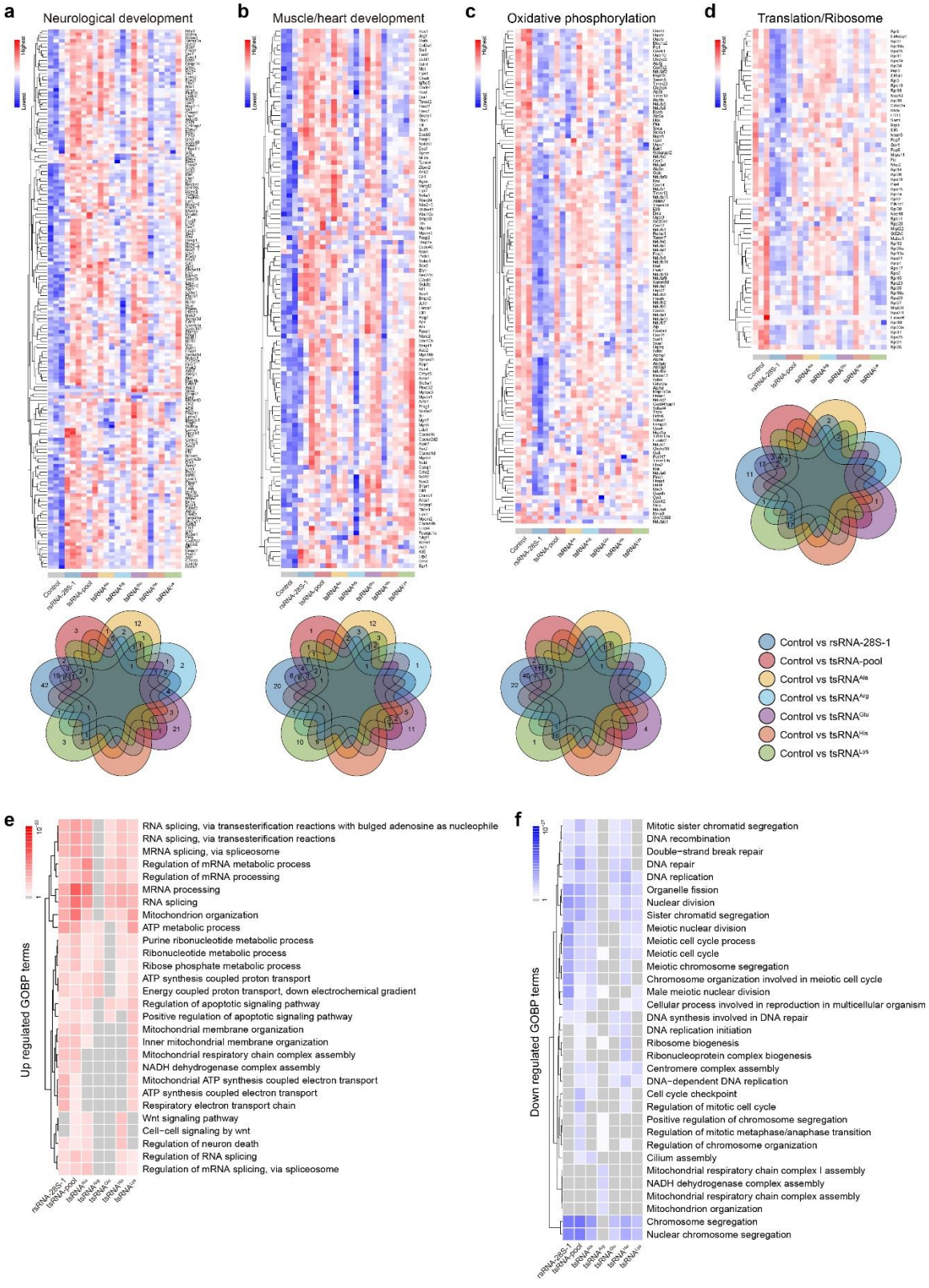
a
Neurological development
b
Muscle/heart development
c
Oxidative phosphorylation
d
Translation/Ribosome
Highest
Lowest
Control
rsRNA-28S-1
tsRNA-pool
tsRNA^Ala
tsRNA^Arg
tsRNA^Glu
tsRNA^His
tsRNA^Lys
Control vs rsRNA-28S-1
Control vs tsRNA-pool
Control vs tsRNA^Ala
Control vs tsRNA^Arg
Control vs tsRNA^Glu
Control vs tsRNA^His
Control vs tsRNA^Lys
e
Up regulated GOBP terms
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
RNA splicing, via transesterification reactions
MRNA splicing, via spliceosome
Regulation of mRNA metabolic process
Regulation of mRNA processing
MRNA processing
RNA splicing
Mitochondrion organization
ATP metabolic process
Purine ribonucleotide metabolic process
Ribonucleotide metabolic process
Ribose phosphate metabolic process
ATP synthesis coupled proton transport
Energy coupled proton transport, down electrochemical gradient
Regulation of apoptotic signaling pathway
Positive regulation of apoptotic signaling pathway
Mitochondrial membrane organization
Inner mitochondrial membrane organization
Mitochondrial respiratory chain complex assembly
NADH dehydrogenase complex assembly
Mitochondrial ATP synthesis coupled electron transport
ATP synthesis coupled electron transport
Respiratory electron transport chain
Wnt signaling pathway
Cell–cell signaling by wnt
Regulation of neuron death
Regulation of RNA splicing
Regulation of mRNA splicing, via spliceosome
f
Down regulated GOBP terms
Mitotic sister chromatid segregation
DNA recombination
Double−strand break repair
DNA repair
DNA replication
Organelle fission
Nuclear division
Sister chromatid segregation
Meiotic nuclear division
Meiotic cell cycle process
Meiotic cell cycle
Meiotic chromosome segregation
Chromosome organization involved in meiotic cell cycle
Male meiotic nuclear division
Cellular process involved in reproduction in multicellular organism
DNA synthesis involved in DNA repair
DNA replication initiation
Ribosome biogenesis
Ribonucleoprotein complex biogenesis
Centromere complex assembly
DNA−dependent DNA replication
Cell cycle checkpoint
Regulation of mitotic cell cycle
Positive regulation of chromosome segregation
Regulation of mitotic metaphase/anaphase transition
Regulation of chromosome organization
Cilium assembly
Mitochondrial respiratory chain complex I assembly
NADH dehydrogenase complex assembly
Mitochondrial respiratory chain complex assembly
Mitochondrion organization
Chromosome segregation
Nuclear chromosome segregation

Figure 3.17: Expression heatmap of the differentially expressed genes from representative GOBP terms in Day6 and Enriched GOBP terms of differential expressed genes in Day3 embryoid bodies after tsRNA/rsRNA transfection

Expression heatmap of the differentially expressed genes from the representative GOBP terms in Day3 embryoid bodies from **Figure 3.16b,c**: (**a**) Neurological development; (**b**) Muscle/heart development; (**c**) Oxidative phosphorylation; (**d**) Translation/ribosome. Venn-diagram beneath each heatmap shows the numbers of overlapped dysregulated genes under different tsRNA/rsRNA transfection. **e**, Top-ranked upregulated GOBP terms in Day3 embryoid bodies after each tsRNA/rsRNA transfection compared to control. **f**, Top-ranked downregulated GOBP terms in Day3 embryoid bodies after each tsRNA/rsRNA transfection compared to control.

Supplementary materials

Figure S3.1: The detailed mapping data of each tsRNA under different treatment protocol
(**a**) mouse brain, (**b**) mouse liver, (**c**) mouse spleen, (**d**) mouse mature sperm, (**e**) mouse mature sperm head, (**f**) mouse MEF (day 0), (**g**) mouse intermediate cell (day 3), (**h**) mouse iPSCs, (**i**) mouse embryonic stem cell (mESC), (**j**) human HeLa cell, (**k**) human embryonic stem cell (hESC-primed), (**l**) human embryonic stem cell (hESC-naïve).

Figure S3.2: The rsRNA mapping data on different rRNAs under different RNA-seq protocol for different tissue/cell types
(**a**) 5S rRNA, (**b**) 5.8S rRNA, (**c**) mitochondrial 12S rRNA, (**d**) mitochondrial 16S rRNA, (**e**) 18S rRNA, (**f**) 28S rRNA, and (**g**) 45S rRNA.

Figure S3.3: The ysRNA mapping data on different YRNAs under different RNA-seq protocol for different tissue/cell types
(**a**) human RNY1 YRNA, (**b**) human RNY3 YRNA, (**c**) human RNY4 YRNA, (**d**) human RNY5 YRNA, (**e**) mouse RNY1 YRNA, (**f**) mouse RNY3 YRNA.

Table S3.1: RNA-seq read summaries and differentially expressed small RNAs by pairwise comparison between individual RNA-seq protocols

Table S3.2: Alternative annotation for miRNA fragments based on miRBase among mouse and human tissues/cells

Table S3.3: Statistics of probes targeting small RNA expression between MEFs and iPSCs under traditional treatment

Table S3.4: List of differentially expressed genes in day 1, 3 and 6 embryoid bodies after tsRNA/rsRNA transfection

Table S3.5: Gene set scores for GOBP terms

References

Akat, K.M., Lee, Y.A., Hurley, A., Morozov, P., Max, K.E., Brown, M., Bogardus, K., Sopeyin, A., Hildner, K., Diacovo, T.G., *et al.* (2019). Detection of circulating extracellular mRNAs by modified small-RNA-sequencing analysis. JCI Insight *5*.

Akiyama, Y., Lyons, S.M., Fay, M.M., Abe, T., Anderson, P., and Ivanov, P. (2019). Multiple ribonuclease A family members cleave transfer RNAs in response to stress. biorxiv.

Bartel, D.P. (2018). Metazoan MicroRNAs. Cell *173*, 20-51.

Behringer, R., Gertsenstein, M., Nagy, K.V., and Nagy, A. (2016). Differentiating Mouse Embryonic Stem Cells into Embryoid Bodies by Hanging-Drop Cultures. Cold Spring Harb Protoc *2016*.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol *34*, 525-527.

Chan, P.P., and Lowe, T.M. (2016). GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic acids research *44*, D184-189.

Chau, K.F., Shannon, M.L., Fame, R.M., Fonseca, E., Mullan, H., Johnson, M.B., Sendamarai, A.K., Springel, M.W., Laurent, B., and Lehtinen, M.K. (2018). Downregulation of ribosome biogenesis during early forebrain development. Elife *7*.

Cheloufi, S., Elling, U., Hopfgartner, B., Jung, Y.L., Murn, J., Ninova, M., Hubmann, M., Badeaux, A.I., Euong Ang, C., Tenen, D., *et al.* (2015). The histone chaperone CAF-1 safeguards somatic cell identity. Nature *528*, 218-224.

Chen, Q., Yan, M., Cao, Z., Li, X., Zhang, Y., Shi, J., Feng, G.H., Peng, H., Zhang, X., Zhang, Y., *et al.* (2016a). Sperm tsRNAs contribute to intergenerational inheritance of an acquired metabolic disorder. Science *351*, 397-400.

Chen, Q., Yan, W., and Duan, E. (2016b). Epigenetic inheritance of acquired traits through sperm RNAs and sperm RNA modifications. Nature reviews Genetics *17*, 733-743.

Chu, C., Yu, L., Wu, B., Ma, L., Gou, L.T., He, M., Guo, Y., Li, Z.T., Gao, W., Shi, H., *et al.* (2017). A sequence of 28S rRNA-derived small RNAs is enriched in mature sperm and various somatic tissues and possibly associates with inflammation. Journal of molecular cell biology *9*, 256-259.

Cozen, A.E., Quartley, E., Holmes, A.D., Hrabeta-Robinson, E., Phizicky, E.M., and Lowe, T.M. (2015). ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. Nature methods *12*, 879-884.

Dai, Q., Zheng, G., Schwartz, M.H., Clark, W.C., and Pan, T. (2017). Selective Enzymatic Demethylation of N(2) ,N(2) -Dimethylguanosine in RNA and Its Application in High-Throughput tRNA Sequencing. Angew Chem Int Ed Engl *56*, 5017-5020.

Genuth, N.R., and Barna, M. (2018). The Discovery of Ribosome Heterogeneity and Its Implications for Gene Regulation and Organismal Life. Molecular cell *71*, 364-374.

Giraldez, M.D., Spengler, R.M., Etheridge, A., Goicochea, A.J., Tuck, M., Choi, S.W., Galas, D.J., and Tewari, M. (2019). Phospho-RNA-seq: a modified small RNA-seq method that reveals circulating mRNA and lncRNA fragments as potential biomarkers in human plasma. The EMBO journal *38*.

Gu, W., Shi, J., Liu, H., Zhang, X., Zhou, J.J., Li, M., Zhou, D., Li, R., Lv, J., Wen, G.*, et al.* (2020). Peripheral blood non-canonical small non-coding RNAs as novel biomarkers in lung cancer. Mol Cancer *19*, 159.

Guo, G., von Meyenn, F., Rostovskaya, M., Clarke, J., Dietmann, S., Baker, D., Sahakyan, A., Myers, S., Bertone, P., Reik, W.*, et al.* (2017). Epigenetic resetting of human pluripotency. Development *144*, 2748-2763.

Hizir, Z., Bottini, S., Grandjean, V., Trabucchi, M., and Repetto, E. (2017). RNY (YRNA)-derived small RNAs regulate cell death and inflammation in monocytes/macrophages. Cell Death Dis *8*, e2530.

Honda, S., Loher, P., Shigematsu, M., Palazzo, J.P., Suzuki, R., Imoto, I., Rigoutsos, I., and Kirino, Y. (2015). Sex hormone-dependent tRNA halves enhance cell proliferation in breast and prostate cancers. Proceedings of the National Academy of Sciences of the United States of America *112*, E3816-3825.

Juhling, F., Morl, M., Hartmann, R.K., Sprinzl, M., Stadler, P.F., and Putz, J. (2009). tRNAdb 2009: compilation of tRNA sequences and tRNA genes. Nucleic acids research *37*, D159-162.

Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic acids research *42*, D68-73.

Krishna, S., Yim, D.G., Lakshmanan, V., Tirumalai, V., Koh, J.L., Park, J.E., Cheong, J.K., Low, J.L., Lim, M.J., Sze, S.K.*, et al.* (2019). Dynamic expression of tRNA-derived small RNAs define cellular states. EMBO reports *20*, e47789.

Li, D., and Wang, J. (2020). Ribosome heterogeneity in stem cells and development. The Journal of cell biology *219*.

Lorenz, R., Bernhart, S.H., Honer Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol Biol *6*, 26.

Natt, D., Kugelberg, U., Casas, E., Nedstrand, E., Zalavary, S., Henriksson, P., Nijm, C., Jaderquist, J., Sandborg, J., Flinke, E.*, et al.* (2019). Human sperm displays rapid responses to diet. PLoS biology *17*, e3000559.

Nawrocki, E.P., Burge, S.W., Bateman, A., Daub, J., Eberhardt, R.Y., Eddy, S.R., Floden, E.W., Gardner, P.P., Jones, T.A., Tate, J.*, et al.* (2015). Rfam 12.0: updates to the RNA families database. Nucleic acids research *43*, D130-137.

Pan, T. (2018). Modifications and functional genomics of human transfer RNA. Cell research *28*, 395-404.

Peng, H., Shi, J., Zhang, Y., Zhang, H., Liao, S., Li, W., Lei, L., Han, C., Ning, L., Cao, Y.*, et al.* (2012). A novel class of tRNA-derived small RNAs extremely enriched in mature mouse sperm. Cell research *22*, 1609-1612.

Phizicky, E.M., and Hopper, A.K. (2010). tRNA biology charges to the front. Genes & development *24*, 1832-1860.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.

Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome biology *11*, R25.

Sai Lakshmi, S., and Agrawal, S. (2008). piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. Nucleic acids research *36*, D173-177.

Sarker, G., Sun, W., Rosenkranz, D., Pelczar, P., Opitz, L., Efthymiou, V., Wolfrum, C., and Peleg-Raibstein, D. (2019). Maternal overnutrition programs hedonic and metabolic phenotypes across generations through sperm tsRNAs. Proceedings of the National Academy of Sciences of the United States of America *116*, 10547-10556.

Schaniel, C., Li, F., Schafer, X.L., Moore, T., Lemischka, I.R., and Paddison, P.J. (2006). Delivery of short hairpin RNAs--triggers of gene silencing--into mouse embryonic stem cells. Nature methods *3*, 397-400.

Schimmel, P. (2018). The emerging complexity of the tRNA world: mammalian tRNAs beyond protein synthesis. Nat Rev Mol Cell Biol *19*, 45-58.

Sergiev, P.V., Aleksashin, N.A., Chugunova, A.A., Polikanov, Y.S., and Dontsova, O.A. (2018). Structural and evolutionary insights into ribosomal RNA methylation. Nature chemical biology *14*, 226-235.

Shi, J., Zhang, Y., Zhou, T., and Chen, Q. (2019). tsRNAs: The Swiss Army Knife for Translational Regulation. Trends Biochem Sci *44*, 185-189.

Shigematsu, M., Kawamura, T., and Kirino, Y. (2018). Generation of 2',3'-Cyclic Phosphate-Containing RNAs as a Hidden Layer of the Transcriptome. Front Genet *9*, 562.

Stadtfeld, M., Maherali, N., Borkent, M., and Hochedlinger, K. (2010). A reprogrammable mouse strain from gene-targeted embryonic stem cells. Nature methods *7*, 53-55.

Su, Z., Wilson, B., Kumar, P., and Dutta, A. (2020). Noncanonical Roles of tRNAs: tRNA Fragments and Beyond. Annual review of genetics *54*, 47-69.

Suzuki, T., and Suzuki, T. (2014). A complete landscape of post-transcriptional modifications in mammalian mitochondrial tRNAs. Nucleic acids research *42*, 7346-7357.

Trewick, S.C., Henshaw, T.F., Hausinger, R.P., Lindahl, T., and Sedgwick, B. (2002). Oxidative demethylation by Escherichia coli AlkB directly reverts DNA base damage. Nature *419*, 174-178.

Valdmanis, P.N., Gu, S., Chu, K., Jin, L., Zhang, F., Munding, E.M., Zhang, Y., Huang, Y., Kutay, H., Ghoshal, K.*, et al.* (2016). RNA interference-induced hepatotoxicity results from loss of the first synthesized isoform of microRNA-122 in mice. Nature medicine *22*, 557-562.

Viswanathan, S.R., Daley, G.Q., and Gregory, R.I. (2008). Selective blockade of microRNA processing by Lin28. Science *320*, 97-100.

Wang, L., Feng, Z., Wang, X., Wang, X., and Zhang, X. (2010). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics *26*, 136-138.

Yang, X., Regan, K., Huang, Y., Zhang, Q., Li, J., Seiwert, T.Y., Cohen, E.E., Xing, H.R., and Lussier, Y.A. (2012). Single sample expression-anchored mechanisms predict survival in head and neck cancer. PLoS Comput Biol *8*, e1002350.

Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L.*, et al.* (2016). Ensembl 2016. Nucleic acids research *44*, D710-716.

Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284-287.

Zhang, J., Zhao, J., Dahan, P., Lu, V., Zhang, C., Li, H., and Teitell, M.A. (2018a). Metabolism in Pluripotent Stem Cells and Early Mammalian Development. Cell metabolism *27*, 332-338.

Zhang, P., Si, X., Skogerbo, G., Wang, J., Cui, D., Li, Y., Sun, X., Liu, L., Sun, B., Chen, R.*, et al.* (2014). piRBase: a web resource assisting piRNA functional study. Database (Oxford) *2014*, bau110.

Zhang, X., Cozen, A.E., Liu, Y., Chen, Q., and Lowe, T.M. (2016). Small RNA Modifications: Integral to Function and Disease. Trends in molecular medicine *22*, 1025-1034.

Zhang, Y., Zhang, X., Shi, J., Tuorto, F., Li, X., Liu, Y., Liebers, R., Zhang, L., Qu, Y., Qian, J.*, et al.* (2018b). Dnmt2 mediates intergenerational transmission of paternally acquired metabolic disorders through sperm small non-coding RNAs. Nat Cell Biol *20*, 535-540.

Zheng, G., Qin, Y., Clark, W.C., Dai, Q., Yi, C., He, C., Lambowitz, A.M., and Pan, T. (2015). Efficient and quantitative high-throughput tRNA sequencing. Nature methods *12*, 835-837.

Chapter 4: Conclusion

Summary

In addition to well-characterized miRNAs and piRNAs (Bartel, 2018; Ozata et al., 2019), the study of other non-canonical small RNAs such as tsRNAs and rsRNAs, is gaining momentum (Lambert et al., 2019; Schimmel, 2018; Su et al., 2020; Wei et al., 2013). The generation of tsRNAs and rsRNAs by cleaving tRNA and rRNA may represent one of the most ancient small RNA biogenesis pathways, as the process exists in all life domains, including archaea, bacteria and eukaryotes (Su et al., 2020). tsRNAs and rsRNAs both exist under physiological conditions and they respond sensitively to various environmental stressors (Andersen and Collins, 2012; Fricker et al., 2019; Garcia-Silva et al., 2014; Lambertz et al., 2015; Lee and Collins, 2005; Liao et al., 2014; Natt et al., 2019; Thompson et al., 2008; Yamasaki et al., 2009; Zhang et al., 2018; Zinskie et al., 2018) that are actively involved in translational regulation (Gebetsberger et al., 2017; Kim et al., 2017; Kuscu et al., 2018; Luo et al., 2018), retrotransposon control (Martinez et al., 2017; Schorn et al., 2017), epigenetic inheritance (Chen et al., 2016; Sarker et al., 2019; Sharma et al., 2016; Zhang et al., 2019; Zhang et al., 2018), and even cross-kingdom regulation between prokaryote and eukaryote (Ren et al., 2019). In particular, RNA modifications in tsRNAs and rsRNAs create additional layers of information regarding secondary structure and binding potential, directing an exciting area of exploration (Frye et al., 2018; Lewis et al., 2017). In contrast, the complicated RNA modification landscapes have caused problems in small RNA high-throughput analyses, because they interfere with RNA-seq library preparation and prevent the detection of tsRNAs and rsRNAs bearing certain modifications.

The small RNA annotation pipeline SPORTS1.0/1.1 that performed in this thesis is convenient for the identification and parallel analyses of both canonical and non-canonical small RNAs in numerous species spreading over three kingdoms with precompiled reference database available. The software can also infer small RNA modification sites in single nucleotide resolution depending on sequence alignment strategy, which shows a distinct pattern among canonical and non-canonical small RNAs.

The improved small RNA-seq method, named PANDORA-seq, is depicted in chapter 3, which expands the repertoire of regulatory small RNAs by resolving RNA modifications that hamper both adapter ligation and reverse transcription process during RNA-seq library construction. Results from PANDORA-seq combined with the northern blot validation confirm that non-canonical small RNAs (for example, tsRNAs and rsRNAs) are the major types of small RNAs in plenty of mammalian tissues and cells, which has previously been underexplored or even excluded in bioinformatics analyses. Transfection of tsRNA and rsRNA sequences discovered by PANDORA-seq to embryoid bodies significantly promoted the lineage differentiation, suggesting that those non-canonical small RNAs play a functional role in cell differentiation.

Future perspectives

PANDORA-seq has limitations and leaves room for future improvement. For example, there are other potential terminal modifications in tsRNAs, or remaining amino acids attached to a tsRNA end that may interfere with adapter ligation (Honda et al., 2015; Raabe et al., 2014), and other tRNA modifications (for example, ms2i6A) that interfere

with reverse transcription (Wei et al., 2015), which can be further addressed through additional enzymatic treatment. PANDORA-seq may also be improved to enable an all-liquid-based protocol (Li et al., 2019) to avoid repeated RNA extraction after enzymatic treatments. Meanwhile, maintaining RNA integrity during every processing is essential, as the degradation of tRNAs/rRNAs may lead to artificial generation of tsRNAs/rsRNAs. Since a considerable RNA amount is requested for enzymatic treatment in PANDORA-seq, it leads to the question that if developing single-cell PANDORA-seq method is realistic with current library construction strategy. Although single-cell small RNA-seq methods based on different approaches are already available (Faridani et al., 2016; Xiao et al., 2018; Yang et al., 2019), the previous protocols do not considered the small RNA modifications that generated biased sequencing results. While the RNA extraction step causes substantial loss of RNA amount that is not acceptable for single cell RNA-seq, a one-pot and fully liquid-based protocol is essential for establishing PANDORA-seq method at single-cell level.

Nonetheless, PANDORA-seq, as well as SPORTS1.1, opens the Pandora's box of small RNAs, especially the hidden world of non-canonical small RNAs that were previously underexplored. The biogenesis and functions of tsRNAs/rsRNAs, as well as the regulatory roles of various RNA modifications, warrant future extensive investigations in different systems. Furthermore, a complete small RNA profile should include not only the expression level of each small RNA sequence, but also the cleavage site/pattern from which they are derived, the overall spectrum of RNA modifications, and the site-specific RNA modifications on each small RNA. The level of complexity buried in the small RNA signature may provide a superior biomarker with better resolution for the diagnosis and

prognosis of complex diseases that were previously difficult to distinguish at the molecular level, facilitating the future development of precision medicine (exemplified in appendix).

References

Andersen, K.L., and Collins, K. (2012). Several RNase T2 enzymes function in induced tRNA and rRNA turnover in the ciliate Tetrahymena. Mol Biol Cell *23*, 36-44.

Bartel, D.P. (2018). Metazoan MicroRNAs. Cell *173*, 20-51.

Chen, Q., Yan, M., Cao, Z., Li, X., Zhang, Y., Shi, J., Feng, G.H., Peng, H., Zhang, X., Zhang, Y*., et al.* (2016). Sperm tsRNAs contribute to intergenerational inheritance of an acquired metabolic disorder. Science *351*, 397-400.

Faridani, O.R., Abdullayev, I., Hagemann-Jensen, M., Schell, J.P., Lanner, F., and Sandberg, R. (2016). Single-cell sequencing of the small-RNA transcriptome. Nat Biotechnol *34*, 1264-1266.

Fricker, R., Brogli, R., Luidalepp, H., Wyss, L., Fasnacht, M., Joss, O., Zywicki, M., Helm, M., Schneider, A., Cristodero, M*., et al.* (2019). A tRNA half modulates translation as stress response in Trypanosoma brucei. Nature communications *10*, 118.

Frye, M., Harada, B.T., Behm, M., and He, C. (2018). RNA modifications modulate gene expression during development. Science *361*, 1346-1349.

Garcia-Silva, M.R., das Neves, R.F., Cabrera-Cabrera, F., Sanguinetti, J., Medeiros, L.C., Robello, C., Naya, H., Fernandez-Calero, T., Souto-Padron, T., de Souza, W*., et al.* (2014). Extracellular vesicles shed by Trypanosoma cruzi are linked to small RNA pathways, life cycle regulation, and susceptibility to infection of mammalian cells. Parasitology research *113*, 285-304.

Gebetsberger, J., Wyss, L., Mleczko, A.M., Reuther, J., and Polacek, N. (2017). A tRNA-derived fragment competes with mRNA for ribosome binding and regulates translation during stress. RNA biology *14*, 1364-1373.

Honda, S., Loher, P., Shigematsu, M., Palazzo, J.P., Suzuki, R., Imoto, I., Rigoutsos, I., and Kirino, Y. (2015). Sex hormone-dependent tRNA halves enhance cell proliferation in breast and prostate cancers. Proceedings of the National Academy of Sciences of the United States of America *112*, E3816-3825.

Kim, H.K., Fuchs, G., Wang, S., Wei, W., Zhang, Y., Park, H., Roy-Chaudhuri, B., Li, P., Xu, J., Chu, K*., et al.* (2017). A transfer-RNA-derived small RNA regulates ribosome biogenesis. Nature *552*, 57-62.

Kuscu, C., Kumar, P., Kiran, M., Su, Z., Malik, A., and Dutta, A. (2018). tRNA fragments (tRFs) guide Ago to regulate gene expression post-transcriptionally in a Dicer-independent manner. Rna *24*, 1093-1105.

Lambert, M., Benmoussa, A., and Provost, P. (2019). Small Non-Coding RNAs Derived From Eukaryotic Ribosomal RNA. Noncoding RNA *5*.

Lambertz, U., Oviedo Ovando, M.E., Vasconcelos, E.J., Unrau, P.J., Myler, P.J., and Reiner, N.E. (2015). Small RNAs derived from tRNAs and rRNAs are highly enriched in exosomes from both old and new world Leishmania providing evidence for conserved exosomal RNA Packaging. BMC genomics *16*, 151.

Lee, S.R., and Collins, K. (2005). Starvation-induced cleavage of the tRNA anticodon loop in Tetrahymena thermophila. J Biol Chem *280*, 42744-42749.

Lewis, C.J., Pan, T., and Kalsotra, A. (2017). RNA modifications and structures cooperate to guide RNA-protein interactions. Nature reviews Molecular cell biology *18*, 202-210.

Li, L., Dai, H., Nguyen, A.P., and Gu, W. (2019). A convenient strategy to clone modified/unmodified small RNA and mRNA for high throughput sequencing. Rna.

Liao, J.Y., Guo, Y.H., Zheng, L.L., Li, Y., Xu, W.L., Zhang, Y.C., Zhou, H., Lun, Z.R., Ayala, F.J., and Qu, L.H. (2014). Both endo-siRNAs and tRNA-derived small RNAs are involved in the differentiation of primitive eukaryote Giardia lamblia. Proceedings of the National Academy of Sciences of the United States of America *111*, 14159-14164.

Luo, S., He, F., Luo, J., Dou, S., Wang, Y., Guo, A., and Lu, J. (2018). Drosophila tsRNAs preferentially suppress general translation machinery via antisense pairing and participate in cellular starvation response. Nucleic acids research *46*, 5250-5268.

Martinez, G., Choudury, S.G., and Slotkin, R.K. (2017). tRNA-derived small RNAs target transposable element transcripts. Nucleic acids research *45*, 5142-5152.

Natt, D., Kugelberg, U., Casas, E., Nedstrand, E., Zalavary, S., Henriksson, P., Nijm, C., Jaderquist, J., Sandborg, J., Flinke, E.*, et al.* (2019). Human sperm displays rapid responses to diet. PLoS biology *17*, e3000559.

Ozata, D.M., Gainetdinov, I., Zoch, A., O'Carroll, D., and Zamore, P.D. (2019). PIWI-interacting RNAs: small RNAs with big functions. Nature reviews Genetics *20*, 89-108.

Raabe, C.A., Tang, T.H., Brosius, J., and Rozhdestvensky, T.S. (2014). Biases in small RNA deep sequencing data. Nucleic acids research *42*, 1414-1426.

Ren, B., Wang, X., Duan, J., and Ma, J. (2019). Rhizobial tRNA-derived small RNAs are signal molecules regulating plant nodulation. Science *365*, 919-922.

Sarker, G., Sun, W., Rosenkranz, D., Pelczar, P., Opitz, L., Efthymiou, V., Wolfrum, C., and Peleg-Raibstein, D. (2019). Maternal overnutrition programs hedonic and metabolic phenotypes across generations through sperm tsRNAs. Proceedings of the National Academy of Sciences of the United States of America *116*, 10547-10556.

Schimmel, P. (2018). The emerging complexity of the tRNA world: mammalian tRNAs beyond protein synthesis. Nat Rev Mol Cell Biol *19*, 45-58.

Schorn, A.J., Gutbrod, M.J., LeBlanc, C., and Martienssen, R. (2017). LTR-Retrotransposon Control by tRNA-Derived Small RNAs. Cell *170*, 61-71 e11.

Sharma, U., Conine, C.C., Shea, J.M., Boskovic, A., Derr, A.G., Bing, X.Y., Belleannee, C., Kucukural, A., Serra, R.W., Sun, F.*, et al.* (2016). Biogenesis and function of tRNA fragments during sperm maturation and fertilization in mammals. Science *351*, 391-396.

Su, Z., Wilson, B., Kumar, P., and Dutta, A. (2020). Noncanonical Roles of tRNAs: tRNA Fragments and Beyond. Annual review of genetics *54*, 47-69.

Thompson, D.M., Lu, C., Green, P.J., and Parker, R. (2008). tRNA cleavage is a conserved response to oxidative stress in eukaryotes. RNA *14*, 2095-2103.

Wei, F.Y., Zhou, B., Suzuki, T., Miyata, K., Ujihara, Y., Horiguchi, H., Takahashi, N., Xie, P., Michiue, H., Fujimura, A.*, et al.* (2015). Cdk5rap1-mediated 2-methylthio modification of mitochondrial tRNAs governs protein translation and contributes to myopathy in mice and humans. Cell metabolism *21*, 428-442.

Wei, H., Zhou, B., Zhang, F., Tu, Y., Hu, Y., Zhang, B., and Zhai, Q. (2013). Profiling and identification of small rDNA-derived RNAs and their potential biological functions. PloS one *8*, e56842.

Xiao, Z., Cheng, G., Jiao, Y., Pan, C., Li, R., Jia, D., Zhu, J., Wu, C., Zheng, M., and Jia, J. (2018). Holo-Seq: single-cell sequencing of holo-transcriptome. Genome Biol *19*, 163.

Yamasaki, S., Ivanov, P., Hu, G.F., and Anderson, P. (2009). Angiogenin cleaves tRNA and promotes stress-induced translational repression. The Journal of cell biology *185*, 35-42.

Yang, Q., Li, R., Lyu, Q., Hou, L., Liu, Z., Sun, Q., Liu, M., Shi, H., Xu, B., Yin, M.*, et al.* (2019). Single-cell CAS-seq reveals a class of short PIWI-interacting RNAs in human oocytes. Nat Commun *10*, 3389.

Zhang, Y., Shi, J., Rassoulzadegan, M., Tuorto, F., and Chen, Q. (2019). Sperm RNA code programmes the metabolic health of offspring. Nat Rev Endocrinol *15*, 489-498.

Zhang, Y., Zhang, X., Shi, J., Tuorto, F., Li, X., Liu, Y., Liebers, R., Zhang, L., Qu, Y., Qian, J.*, et al.* (2018). Dnmt2 mediates intergenerational transmission of paternally acquired metabolic disorders through sperm small non-coding RNAs. Nat Cell Biol *20*, 535-540.

Zinskie, J.A., Ghosh, A., Trainor, B.M., Shedlovskiy, D., Pestov, D.G., and Shcherbik, N. (2018). Iron-dependent cleavage of ribosomal RNA during oxidative stress in the yeast Saccharomyces cerevisiae. The Journal of biological chemistry *293*, 14237-14248.

Appendix: Peripheral blood non-canonical small noncoding RNAs as novel biomarkers in lung cancer

Abstract

One unmet challenge in lung cancer diagnosis is to accurately differentiate lung cancer from other lung diseases with similar clinical symptoms and radiological features, such as pulmonary tuberculosis (TB). To identify reliable biomarkers for lung cancer screening, the recently discovered non-canonical small non-coding RNAs (i.e., tsRNAs, rsRNAs, and ysRNAs) were leveraged in human peripheral blood mononuclear cells and developed a molecular signature composed of distinct ts/rs/ysRNAs (TRY-RNA). The TRY-RNA signature precisely discriminates between control, lung cancer, and pulmonary TB subjects in both the discovery and validation cohorts and outperforms microRNA-based biomarkers, which bears the diagnostic potential for lung cancer screening.

Introduction

One unmet challenge in current lung cancer diagnosis is to accurately differentiate lung cancer from other lung diseases with similar clinical symptoms and radiological features. Imaging-based screening methods, such as low-dose computed topography (LDCT), could sometimes be false positives, as indeterminate pulmonary nodules may also be caused by other lung diseases such as pulmonary tuberculosis (TB) (Lakhani and Sundaram, 2017), which is especially concerning for clinical practice in TB-endemic countries/regions. Therefore, additional noninvasive diagnostic procedures are much needed to avoid a mis-diagnosis in patients with lung cancer mimicking pulmonary TB, or vice versa. Here, a peripheral blood mononuclear cell (PBMC)-based molecular signature is aimed to be developed to differentiate lung cancer patients from healthy controls and pulmonary TB patients by harnessing the novel small RNAs.

Recent small RNA sequencing attempts have ubiquitously detected several non-canonical small RNA types, which are fragments derived from canonically transcribed parent large RNAs, including tRNA-derived small RNAs (tsRNAs), rsRNAs, and ysRNAs. ts/rs/ysRNAs have been discovered in a wide range of species. The biological functions of tsRNAs have become a recent highlight and been linked with various human diseases (Su et al., 2020), including cancers (Balatti et al., 2017; Dhahbi et al., 2014; Farina et al., 2020), while rsRNAs and ysRNAs show sensitive response to pathophysiological conditions (Dhahbi et al., 2014; Farina et al., 2020). In this study, a diagnostic signature is developed composed of distinct ts/rs/ysRNAs (TRY-RNA) in human PBMCs. The TRY-RNA signature accurately discriminates between control, lung cancer, and pulmonary TB subjects in

both the discovery and validation cohorts and outperforms microRNA (miRNA)-based biomarkers. Figure A.1 provides an overview of the experimental design.

Results

Dysregulated non-canonical small RNAs in lung cancer

Small RNA-seq was performed for the PBMC samples collected from 59 human subjects in the discovery cohort, including 13 healthy controls, 10 pulmonary TB patients, and 36 lung cancer patients (Table SA.1). The raw sequencing data were processed by SPORTS1.0. In total, 6673 tsRNA species, 20,172 rsRNA species, 1238 ysRNA species, and 973 miRNA species were identified in human PBMCs (Figure A.2). The co-expression pattern of tsRNAs was investigated across the PBMC samples in the discovery cohort by grouping tsRNA species into subcategories according to their parent tRNA types. It was found that the expression of the tsRNAs derived from the tRNAs of alanine (tsRNA-Ala), asparagine (tsRNA-Asn), leucine (tsRNA-Leu), lysine (tsRNA-Lys), and tyrosine (tsRNA-Tyr) was strongly and positively correlated with that of each other (Spearman's rank correlation test: $\rho > 0.700$ and $P < 10^{-9}$) (Figure A.3a), suggesting shared biogenesis pathways among these tsRNAs. Interestingly, tsRNA-Ala, tsRNA-Asn, tsRNA-Leu, tsRNALys, and tsRNA-Tyr were the only five tsRNA groups that were upregulated in the lung cancer patients relative to the controls (adjusted $P < 0.05$) (Figure A.3b). It was further found that the expression of these five tsRNA groups was also significantly higher in the lung cancer patients than in the pulmonary TB subjects ($P < 0.05$) (Figure A.3b). Next, rsRNA and ysRNA species were grouped into subcategories according to their parent rRNA/YRNA types. Then, it was

found that the rsRNAs derived from rRNA-5S (rsRNA-5S) were significantly upregulated in the lung cancer patients relative to the controls, while the ysRNAs originating from YRNA-RNY1 (ysRNA-RNY1) were downregulated in the lung cancer patients compared with the controls (adjusted $P < 0.05$) (Figure A.3b). More interestingly, the expression of rsRNA-5S and ysRNA-RNY1 showed a completely inverse pattern in the pulmonary TB patients: rsRNA-5S was significantly downregulated in the TB patients relative to the controls, while ysRNA-RNY1 was upregulated in the TB patients compared with the controls ($P < 0.05$) (Figure A.3b). The individual tsRNA-Ala, tsRNA-Asn, tsRNA-Leu, tsRNA-Lys, tsRNA-Tyr, rsRNA-5S, and ysRNA-RNY1 species were further mapped to the corresponding parent RNAs, which followed a nonrandom fragmentation pattern (Figure A.3c-d and Figure A.4), suggesting highly regulated biogenesis of these small RNAs. In addition, the association of these non-canonical small RNA expression was investigated with cancer stage, histological type, lymph node status, metastasis status, and smoking history, but no significant difference was observed (Figure A.5).

The molecular signature composed of noncanonical small RNAs

Next, a molecular signature of small RNAs was developed by harnessing the above prioritized small RNA subcategories (i.e., tsRNA-Ala, tsRNA-Asn, tsRNA-Leu, tsRNA-Lys, tsRNA-Tyr, rsRNA-5S, and ysRNA-RNY1). In total, nine tsRNA species, eight rsRNA species, and eight ysRNA species (Figure A.6a-c) were selected, which consisted of a molecular signature with 25 distinct non-canonical small RNAs (Figure A.6d and Table SA.2), referred to as the TS/RS/YSRNA (TRY-RNA) signature. Both principal component

analysis (Figure A.6e) and hierarchical clustering on RNA expression (Figure A.3e) confirmed the discriminative power of the TRY-RNA signature between the control, lung cancer, and TB groups in the discovery cohort. To systematically evaluate the classification power of the TRY-RNA signature, a TRY-RNA index was assigned to each subject based on the expression of the ts/rs/ysRNAs within the TRY-RNA signature (see methods). The TRY-RNA index was a linear combination of the expression values of the small RNA species within the TRY-RNA signature. A higher TRY-RNA index implies a higher likelihood of lung cancer. It was found that the TRY-RNA index was significantly higher in the lung cancer patients than in the healthy controls, while the TRY-RNA index of the pulmonary TB patients was significantly lower than that of the controls (t-test: $P < 10^{-5}$) (Figure A.7a). The area under the receiver operating characteristic (ROC) curve (AUC) was 1.000 between the cancer and non-cancer subjects and 0.994 between the TB and non-TB subjects (Figure A.7b). In addition, the association of the expression of the individual RNA species was investigated within the TRY-RNA signature with cancer stage, histological type, lymph node status, metastasis status, and smoking history, but significant difference was only observed for ysRNARNY1–28 and ysRNA-RNY1-29a between adenocarcinoma and squamous cell carcinoma patients (Figure A.8-A12).

The performance of the TRY-RNA signature in the validation cohort

The TRY-RNA signature was further assessed in the validation cohort with 35 human PBMC samples collected from 12 healthy controls, 15 lung cancer patients, and 8 pulmonary TB patients (Table SA.3). Unsupervised hierarchical clustering and principal

component analysis demonstrated a totally distinct expression pattern of the TRY-RNA signature between the lung cancer and TB subjects, with the controls largely falling in between in the validation cohort (Figure A.13a-b). The TRY-RNA index in the validation cohort was significantly higher in the lung cancer patients than in the healthy controls, while the TRY-RNA index of the TB patients was significantly lower than that of the controls (t-test: $P < 0.005$) (Figure A.13c). The *AUC* was 0.930 between the cancer and non-cancer subjects and 1.000 between the TB and non-TB subjects (Figure A.13d), which suggests the strong classification power of the TRY-RNA signature for both lung cancer and pulmonary TB screening.

Comparison between the TRY-RNA and miRNAbased signatures

The expression profiles of miRNAs among the control, lung cancer, and pulmonary TB subjects were also compared in the discovery cohort and identified a signature with 43 miRNA species, referred to as the MIR signature (Figure A.14 and Table S4). Similar to the TRY-RNA signature, a MIR index was assigned to each subject based on the expression of the miRNAs within the MIR signature. It was found that in both the discovery and validation cohorts, while the MIR index can differentiate the lung cancer patients from the control and TB subjects (Figure A.7c-d and A.15a-b), a resampling test (See methods) demonstrated a superior classification power of the TRY-RNA signature compared to the MIR signature (Figure A.15c). Whether the MIR signature provided additive classification power to the TRY-RNA signature by combining both signatures was further investigated (referred to as the TRY-RNA $\cup$ MIR signature). Although the performance of the TRY-

RNA ∪ MIR signature was fairly good, the TRY-RNA ∪ MIR signature didn't outperform the TRYRNA signature and barely provided additive information for the classification (Figure A.15d-f).

Conclusions

The TRY-RNA signature derived from the repertoire of PBMC non-canonical small RNAs makes it possible for the early diagnosis of lung cancer and pulmonary TB, which may reflect the host responses to different antigens and would represent an improvement over the previous studies focusing solely on tsRNAs in cancer tissues (Balatti et al., 2017; Farina et al., 2020; Pekarsky et al., 2016). Interestingly, the performance of the TRY-RNA signature shows superiority over the miRNA-based signature, which could be due to the more complex layer of non-canonical small RNAs. For example, tsRNAs and rsRNAs exhibit an unexpected complexity in regards to their RNA modifications as well as their sequence diversities (Shi et al., 2019). Previous study suggests that both tsRNAs and rsRNAs are involved in mammalian epigenetic inheritance, which form a 'RNA code' to convey environmental clue to the offspring (Zhang et al., 2019; Zhang et al., 2018). Also, tsRNAs are thought to regulate translation process and ribosome biogenesis in versatile ways, including the fine-tuning of the ribosome composition that may affect the translational specificity on a selective pool of mRNAs (also referred to as ribosome heterogeneity). In other words, change in tsRNA (and perhaps rsRNA/ysRNA as well) composition may result in altered ribosome heterogeneity that directs the cell to a specific functional state (Shi et al., 2019). The complexity and possible permutations of different tsRNA/ rsRNA/ysRNAs may

endow the superior information capacity and specificity that are needed to distinguish complex diseases, and as being harnessed here, represent a 'disease RNA code' in lung cancer screening.

Methods

Human subjects

This study aimed to develop a non-canonical small RNA-based molecular signature in human PBMCs differentiating lung cancer patients from healthy controls and pulmonary TB subjects. Small RNA-seq was applied to measure the PBMC ts/rs/ysRNA expression for both the discovery (n = 59) and validation (n = 35) cohorts. All the subjects of this study were of Chinese Han descent. Lung cancer patients were recruited from the First Affiliated Hospital of Bengbu Medical College without receiving adjuvant chemotherapy. Both histological and radiological features were collected for the diagnosis of lung cancer. Active pulmonary TB patients were recruited from the Infectious Disease Hospital of Bengbu City before any TB treatment. The diagnosis of TB was based on established international guidelines (Lewinsohn et al., 2017). The healthy controls were recruited from the Physical Examination Center of the First Affiliated Hospital of Bengbu Medical College. Subjects with other concurrent infectious diseases were excluded. All subjects were recruited consecutively over time, with the discovery cohort being recruited first followed by the validation cohort. The detailed information is presented in Table SA.1 and Table SA.2. The Ethics Committee of Bengbu Medical College approved this study, with written informed

consent obtained from all subjects, which conformed to the standard indicated by the Declaration of Helsinki.

PBMC RNA isolation

PBMCs were collected from the subjects in the discovery and validation cohorts. Five milliliters of anticoagulant peripheral blood were drawn from the ulnar vein of each subject. PBMCs were immediately isolated by the Ficoll-Hypaque density gradient centrifugation method. Briefly, the blood samples were diluted with RPMI-1640 basic medium at a ratio of 1:1. The diluted blood was added and spread over the Ficoll-Hypaque separation solution at a ratio of 2:1 and then centrifuged at 2,000 revolutions per minute for 20 minutes at room temperature. After centrifugation, the white misty cell layer was collected into a new centrifuge tube and washed twice with RPMI-1640 basic medium at 2,000 revolutions per minute for 5 minutes at 4 °C. The isolated PBMCs were transferred into 1.5 mL tubes, and 1 mL TRIzol (Ambion, Thermo Fisher Scientific) was added for subsequent total RNA extraction. Total RNA as extracted from PBMCs using TRIzol reagent (Invitrogen) and purified with a mirVana miRNA Isolation Kit (Ambion, Thermo Fisher Scientific) according to the manufacturer's protocol. RNA degradation and contamination were monitored on 1% agarose gels. RNA purity was checked using a NanoPhotometer spectrophotometer (Implen, CA, USA). RNA concentration was measured using a Qubit RNA Assay Kit in a Qubit 2.0 Fluorometer (Life Technologies, CA, USA). RNA integrity was assessed by the RNA Nano 6000 Assay Kit of the Agilent Bioanalyzer 2100 system (Agilent Technologies,

CA, USA). Only the RNA samples with RNA integrity values > 6 were retained for further study.

Small RNA-seq library preparation

A total of ~2 μg total RNA per sample was used as the input for the small RNA-seq libraries. The small RNA-seq libraries were constructed using the NEBNext Multiplex Small RNA Library Prep Set for Illumina® (New England Biolabs, USA). The NEB 3' SR adaptor was ligated to the 3' end of small RNAs, followed by the SR RT Primer being hybridized to the excess 3' SR adaptor, which transformed the single-stranded DNA adaptor into double-stranded DNA. The 5' end adapter was ligated to the 5' ends of small RNAs, followed by the first cDNA strand being synthesized using M-MuLV Reverse Transcriptase. PCR amplification was performed using LongAmp Taq 2X Master Mix for 11-13 cycles, and the products were purified on an 8% polyacrylamide gel (100 V, 80 minutes). DNA fragments were recovered and dissolved in 8 μL elution buffer. The qualified libraries, which were assessed by an Agilent Bioanalyzer 2100, were amplified on the cBot to generate the cluster on the flow cell. The amplified flow cell was sequenced (single-end) on the Illumina System with a read length of 50 nucleotides.

small RNA-seq data processing

SPORTS1.0 was used to parse the raw small RNA-seq data. Clean reads were outputted by removing sequence adapters and discarding sequences with lengths beyond the defined range and those with bases other than ATUCG. The clean reads were sequentially mapped

against miRBase (Kozomara and Griffiths-Jones, 2014), the rRNA/YRNA database (obtained from NCBI), and GtRNAdb (Chan and Lowe, 2016). Because miRNAs were dominant among the sequencing reads, the reads per million (RPM) values were summarized for the small RNA species with lengths ≤ 25 nucleotides and > 25 nucleotides separately. The miRNA-based signature was developed from sequencing reads with lengths ≤ 25 nucleotides, while the signature composed of ts/rs/ysRNAs was derived from sequencing reads with lengths > 25 nucleotides. The non-canonical small RNAs were only retained with at least one read in at least 10 samples. Non-canonical small RNA species were further grouped, i.e., ts/rs/ysRNAs, into individual subcategories according to the parent large RNAs from which they originated. The small RNA subcategories with fold change < 2 between the control and lung cancer groups were excluded from further analyses. A linear model controlling was used for age and sex (McDonough et al., 2019) to compare the expression of each small RNA subcategory between the control and lung cancer groups. The Benjamini-Hochberg procedure was used for *P*-value correction. The same linear model controlling for age and sex was also used to compare the expression of each small RNA subcategory between the pulmonary TB and lung cancer groups. To identify the small RNA species that were differentially expressed between the controls and lung cancer patients and between the controls and TB patients, the edgeR tool (Robinson et al., 2010) was employed controlling for age and sex. The small RNA species (mean RPM > 1) with a false discovery rate < 0.01 were deemed differentially expressed.

Developing the molecular signatures

To develop the TRY-RNA signature, only the small RNA species differentially expressed between the controls and lung cancer patients and between the pulmonary TB and lung cancer patients were retained. To avoid potential biases caused by RNA size fractionation procedures, small RNA species with lengths ≥ 40 nucleotides were further excluded. For each tsRNA subcategory, only the top two tsRNA species with the highest average expression levels across all the PBMC samples if there was more than one tsRNA species within this subcategory were collected. In total, nine tsRNA species were prioritized: tsRNA-Ala-AGC/CGC-30 and tsRNA-Ala-AGC/CGC-31 belonging to tsRNA-Ala, tsRNA-Asn-GTT-26 and tsRNA-Asn-GTT-27 belonging to tsRNA-Asn, tsRNA-Leu-CAG-26 belonging to tsRNA-Leu, tsRNA-Lys-CTT-29 and tsRNA-Lys-CTT-30 belonging to tsRNA-Lys, and tsRNA-Tyr-GTA-31 and tsRNA-Tyr-GTA-32 belonging to tsRNATyr. For rsRNA-5S and ysRNA-RNY1, the RNA species with mean *RPM* > 50 were collected, yielding eight rsRNA species, rsRNA-5S-27, rsRNA-5S-28, rsRNA-5S-30, rsRNA-5S-31, rsRNA-5S-32, rsRNA-5S-37, rsRNA-5S-38, and rsRNA-5S-39, and eight ysRNA species, ysRNA-RNY1-26, ysRNA-RNY1-28, ysRNA-RNY1-29a, ysRNA-RNY1-29b, ysRNA-RNY1- 30, ysRNA-RNY1-31, ysRNA-RNY1-32, and ysRNA-RNY1-36. The expression profiles of miRNAs among the control, lung cancer, and pulmonary TB patients in the discovery cohort were also examined. In total, 43 miRNA species were found to be differentially expressed between the controls and lung cancer patients and between the TB and lung cancer patients. These 43 miRNAs were designated as the MIR signature.

The TRY-RNA, MIR, and TRY-RNA ∪ MIR indices

A scoring scheme was applied which was used in previous studies to assign each human subject a TRY-RNA index (Qian et al., 2018; Qian et al., 2016):

$$I_{TRY-RNA} = \sum_{i=1}^{25} w_i(e_i - \mu_i) / \tau_i$$

Here, $I_{TRY\text{-}RNA}$ was the TRY-RNA index; w_i was the weight of non-canonical small RNA i within the TRY-RNA signature as shown in Table SA.2 (w_i = 1 if small RNA i was upregulated in the lung cancer patients relative to the controls, while w_i = -1 if small RNA i was downregulated in the lung cancer patients); e_i denoted the expression level of small RNA i; and μ_i and τ_i were the mean and standard deviation of the expression of small RNA i across all the samples, respectively. Similarly, the MIR index was defined as:

$$I_{MIR} = \sum_{i=1}^{43} w_i(e_i - \mu_i) / \tau_i$$

Here, I_{MIR} was the MIR index; w_i was the weight of miRNA i within the MIR signature as shown in Table SA.4 (w_i = 1 if miRNA i was upregulated in the lung cancer patients relative to the controls, while w_i = -1 if miRNA i was downregulated in the lung cancer patients); e_i denoted the expression level of miRNA i; and μ_i and τ_i were the mean and standard deviation of the expression of miRNA i across all the samples, respectively. Finally, the TRY-RNA ∪ MIR index was defined as:

$$I_{TRY-RNA \cup MIR} = \sum_{i=1}^{68} w_i(e_i - \mu_i) / \tau_i$$

Here, *ITRY-RNA* ∪ *MIR* was the TRY-RNA ∪ MIR index; *wi* was the weight of RNA *i* within the TRYRNA ∪ MIR signature (including 25 ts/rs/ysRNA and 43 miRNA species); *ei* denoted the expression level of RNA *i*; and μi and τi were the mean and standard deviation of the expression of RNA *i* across all the samples, respectively.

Resampling test

Because the size between the TRY-RNA and MIR signatures was different (25 ts/rs/ysRNA species *vs.* 43 miRNA species), to perform a fair comparison between the two signatures, a resampling test was conducted by randomly selecting 25 miRNA sequences from the MIR signature 1,000 times. For each random 25-miRNA signature, the MIR index for each subject was recalculated, and a multi-class *AUC* was computed among the control, lung cancer, and TB groups according to the generalization model proposed by Hand and Till (Hand and Till, 2001), which represented the classification power of the random signature.

Statistical analysis

All statistical analyses were performed using the R platform. Correlations between continuous variables were measured by *Spearman*'s rank correlation test using the cor.test function. Student's *t*-test was performed for groupwise comparisons of normal distributions, using the t.test function. A linear model controlling for age and sex was applied to prioritize the differentially expressed non-canonical small RNA subcategories using the lm function. If multiple testing should be accounted for, the *Benjamini-Hochberg* procedure was applied

for *P*-value correction using the p.adjust function. Principal component analysis on the expression data of the TRY-RNA signature was performed using the dudi.pca function within the package ade4. The *AUC* and multi-class *AUC* values were computed using the roc and multiclass.roc functions respectively, within the package pROC.

Figures

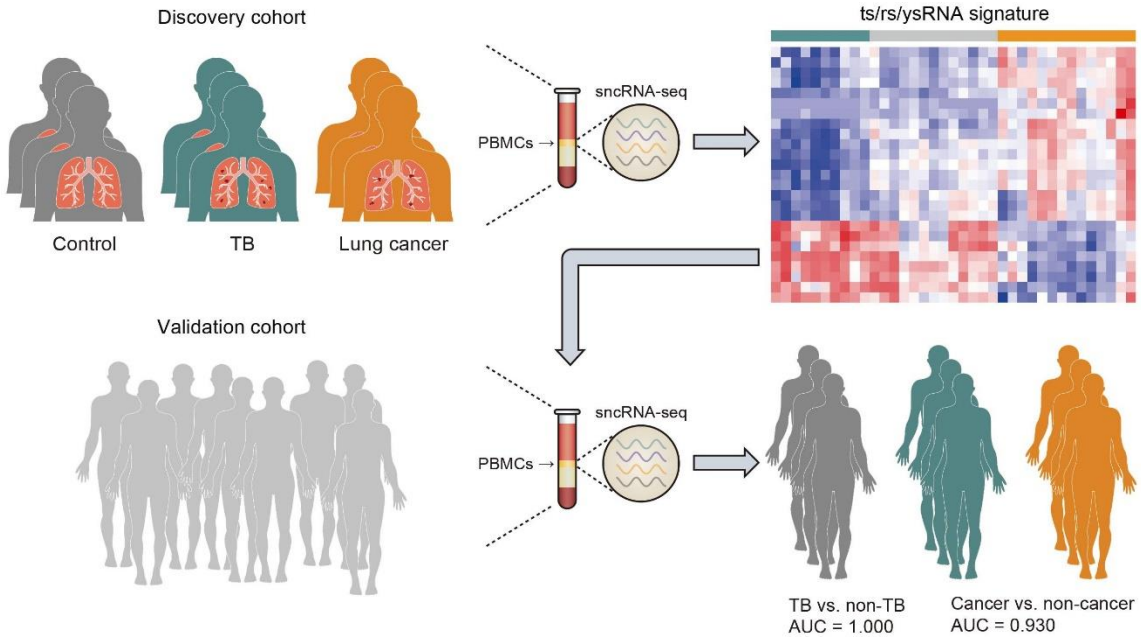

Figure A.1: The workflow of the study. PBMC ts/rs/ysRNA expression of the human subjects in the discovery cohort was profiled by small RNA-seq

A molecular signature composed of ts/rs/ysRNAs was developed to discriminate between healthy controls, lung cancer patients, and pulmonary TB subjects. This signature was validated in the validation cohort with high accuracy. *AUC*: area under the receiver operating characteristic curve.

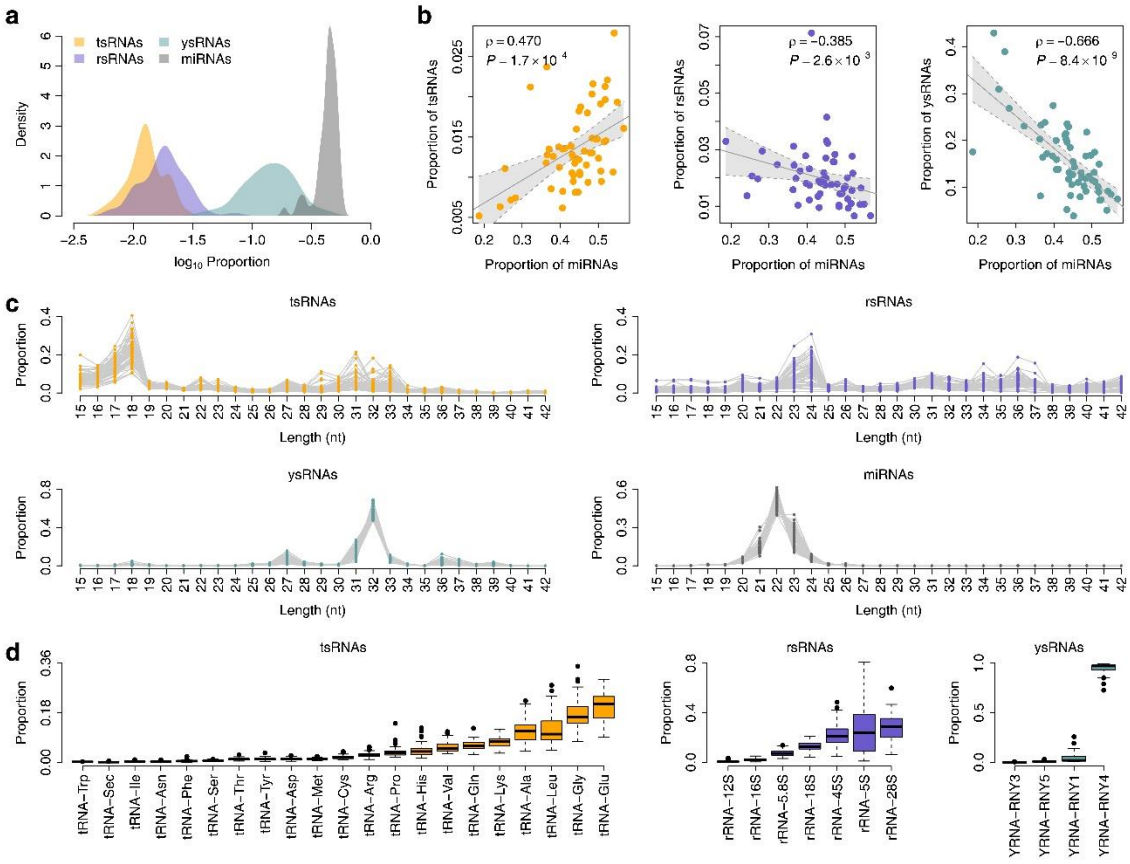

Figure A.2: The landscape of non-canonical small RNAs in human PBMCs

a, The distribution of small RNA read proportions. The X-axis was log10-transformed. For each sample, the read proportions of tsRNAs, rsRNAs, ysRNAs, and miRNAs were computed, respectively. **b**, The correlation in read proportions between miRNAs and non-canonical small RNAs, *i.e.*, tsRNAs, rsRNAs, and ysRNAs. The correlation coeffcients (ρ) and *P*-values were calculated by *Spearman*'s rank correlation test. **c**, The length distribution of small RNAs. Each dot represents one PBMC sample. The Y-axis shows the read proportion within each small RNA category, *i.e.*, tsRNA, rsRNA, ysRNA, and miRNA. **d**, The parent large RNAs from which non-canonical small RNAs originated. The Y-axis shows the read proportion within each non-canonical small RNA category, *i.e.*, tsRNA, rsRNA, and ysRNA.

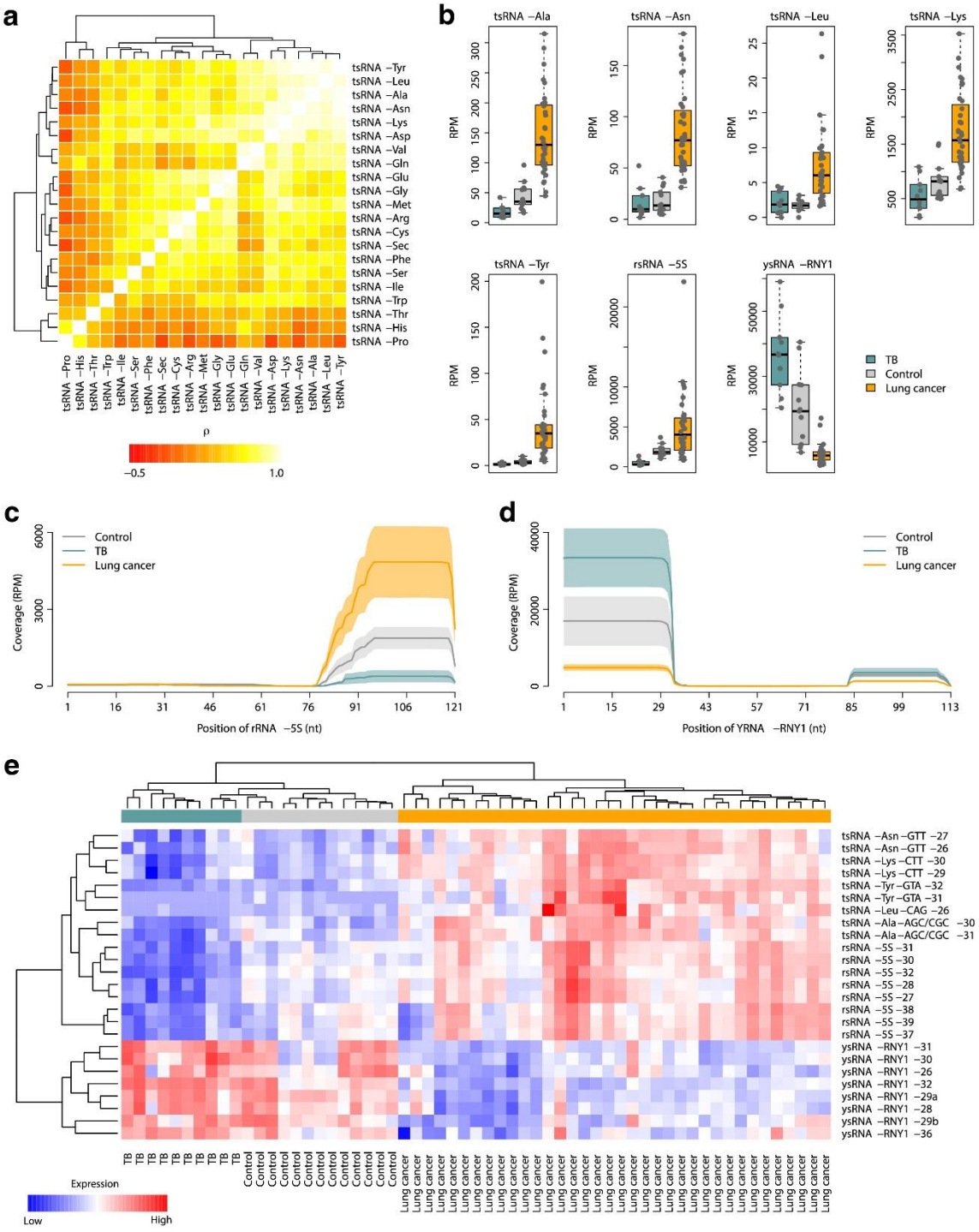
a
tsRNA –Tyr
tsRNA –Leu
tsRNA –Ala
tsRNA –Asn
tsRNA –Lys
tsRNA –Asp
tsRNA –Val
tsRNA –Gln
tsRNA –Glu
tsRNA –Gly
tsRNA –Met
tsRNA –Arg
tsRNA –Cys
tsRNA –Sec
tsRNA –Phe
tsRNA –Ser
tsRNA –Ile
tsRNA –Trp
tsRNA –Thr
tsRNA –His
tsRNA –Pro
ρ
−0.5
1.0
b
tsRNA –Ala
tsRNA –Asn
tsRNA –Leu
tsRNA –Lys
tsRNA –Tyr
rsRNA –5S
ysRNA –RNY1
RPM
TB
Control
Lung cancer
c
Control
TB
Lung cancer
Coverage (RPM)
Position of rRNA –5S (nt)
d
Control
TB
Lung cancer
Coverage (RPM)
Position of YRNA –RNY1 (nt)
e
tsRNA –Asn –GTT –27
tsRNA –Asn –GTT –26
tsRNA –Lys –CTT –30
tsRNA –Lys –CTT –29
tsRNA –Tyr –GTA –32
tsRNA –Tyr –GTA –31
tsRNA –Leu –CAG –26
tsRNA –Ala –AGC/CGC –30
tsRNA –Ala –AGC/CGC –31
rsRNA –5S –31
rsRNA –5S –30
rsRNA –5S –32
rsRNA –5S –28
rsRNA –5S –27
rsRNA –5S –38
rsRNA –5S –39
rsRNA –5S –37
ysRNA –RNY1 –31
ysRNA –RNY1 –30
ysRNA –RNY1 –26
ysRNA –RNY1 –32
ysRNA –RNY1 –29a
ysRNA –RNY1 –28
ysRNA –RNY1 –29b
ysRNA –RNY1 –36
Expression
Low
High

Figure A.3: The dysregulated non-canonical small RNAs in lung cancer

a, The co-expression pattern tsRNA subcategories across the PBMC samples in the discovery cohort. **b**, The expression profile of tsRNA-Ala, tsRNA-Asn, tsRNA-Leu, tsRNA-Lys, tsRNA-Tyr, rsRNA-5S, and ysRNA-RNY1 among the control, lung cancer, and TB subjects. RPM: reads per million. **c** and **d**, The coverage profile of the PBMC rsRNA- 5S and ysRNA-RNY1 sequences along rRNA-5S and YRNA-RNY1, respectively. The solid curves indicate the mean RPM values for the control, lung cancer, and TB groups. The colored bands represent the 95% confidence interval. nt: nucleotide. **e**, Expression heatmap of the small RNA species within the TRY-RNA signature in the discovery cohort.

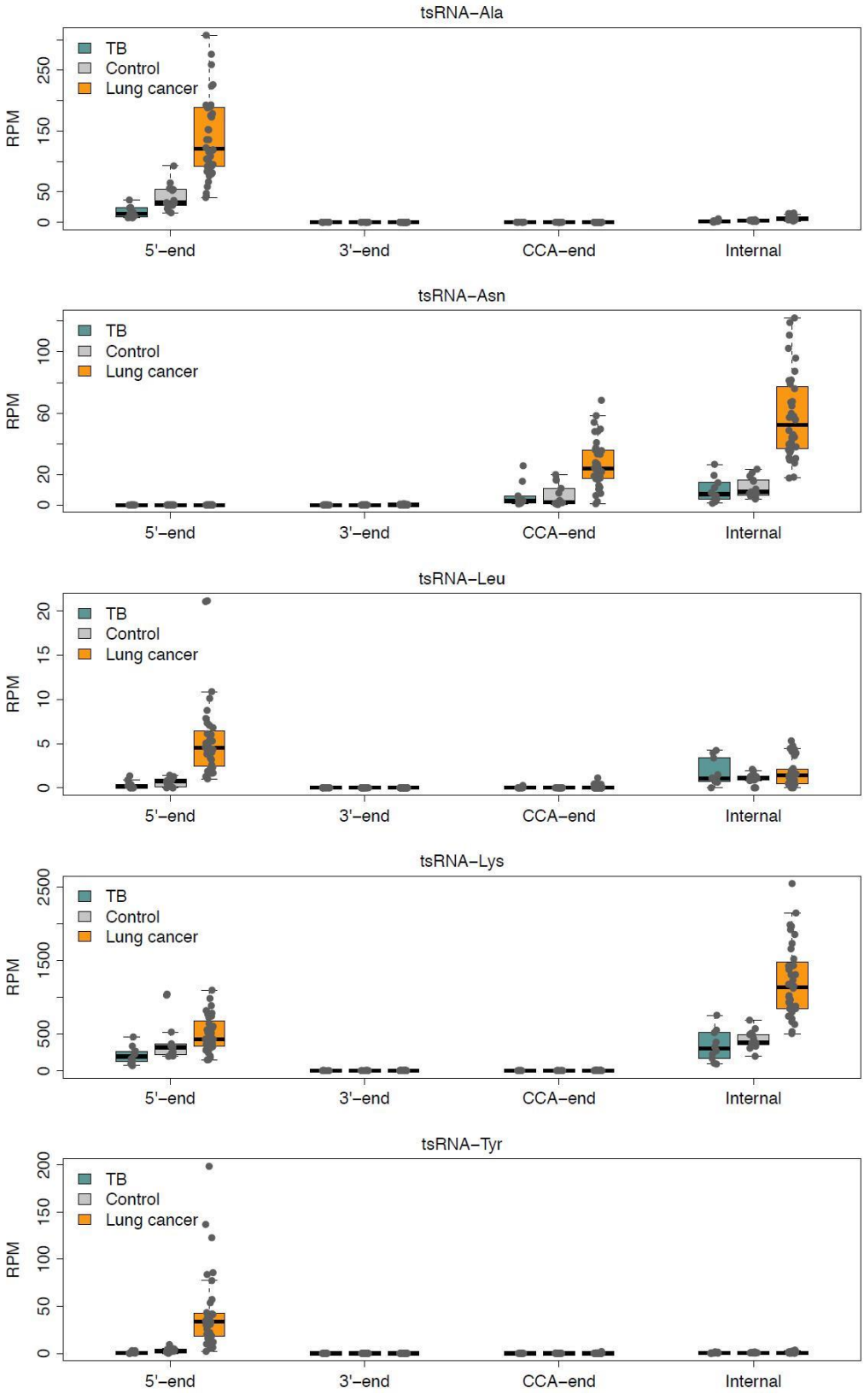
tsRNA–Ala
TB
Control
Lung cancer
RPM
0
50
150
250
5'–end
3'–end
CCA–end
Internal
tsRNA–Asn
TB
Control
Lung cancer
RPM
0
20
60
100
5'–end
3'–end
CCA–end
Internal
tsRNA–Leu
TB
Control
Lung cancer
RPM
0
5
10
15
20
5'–end
3'–end
CCA–end
Internal
tsRNA–Lys
TB
Control
Lung cancer
RPM
0
500
1500
2500
5'–end
3'–end
CCA–end
Internal
tsRNA–Tyr
TB
Control
Lung cancer
RPM
0
50
100
150
200
5'–end
3'–end
CCA–end
Internal

Figure A.4: The mapping profile of tsRNAs

Individual tsRNAs were classified according to the fragment locations on the corresponding parent tRNAs, *i.e.*, the 5 terminus, 3 terminus, 3 CCA-end, or internal region of tRNAs.

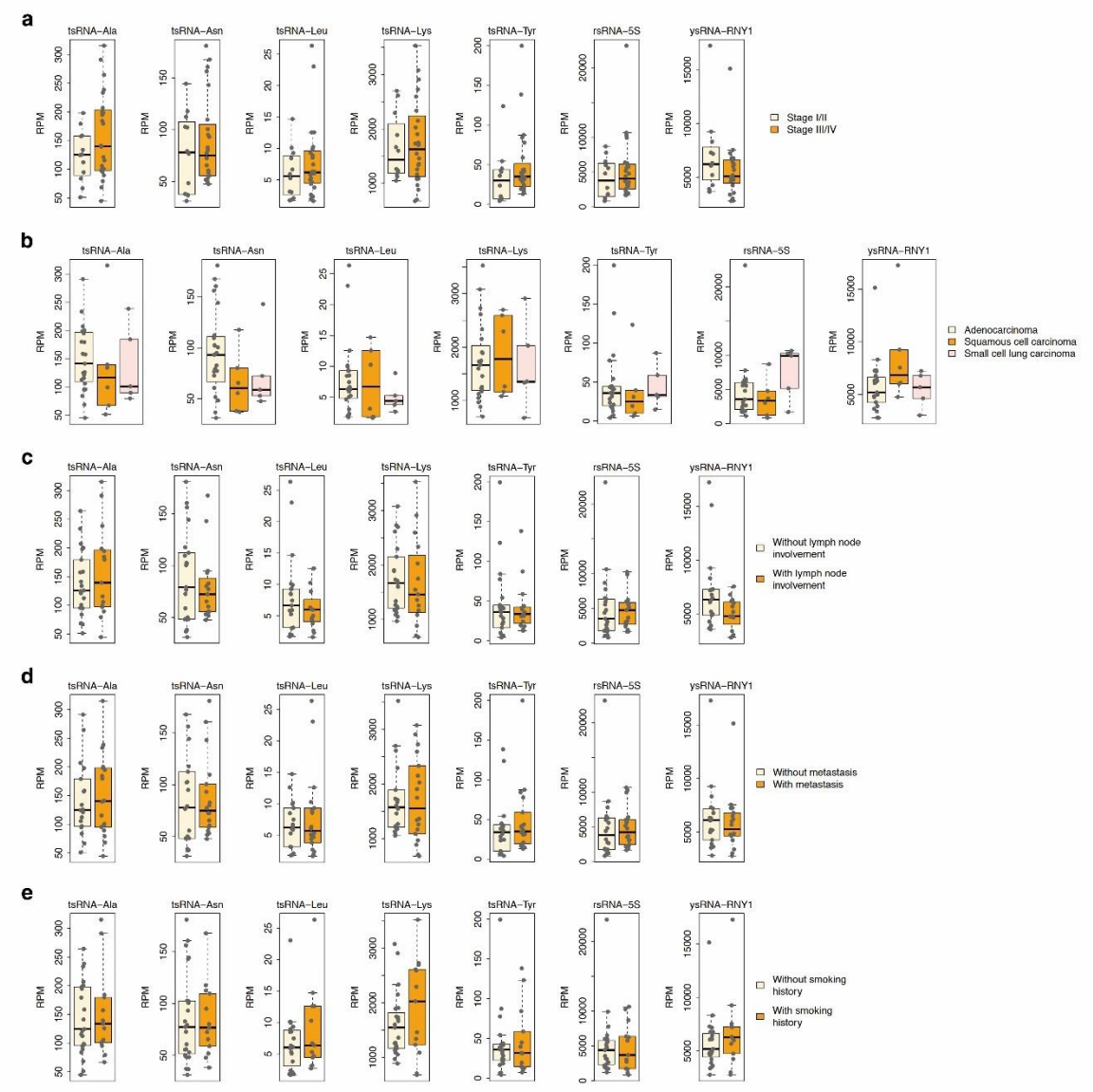

Figure A.5: Comparison of the expression of the prioritized small RNA subcategories between different conditions

Comparison of the expression of the prioritized small RNA subcategories between (**a**) lung cancer stages, (**b**) lung cancer histological types, or the lung cancer patients (**c**) with and without lymph node involvement, (**d**) with and without distant metastasis, or (**e**) with and without smoking history. Groupwise comparisons were performed using a linear model controlling for age and sex. No significant difference was observed (adjusted P > 0.05).

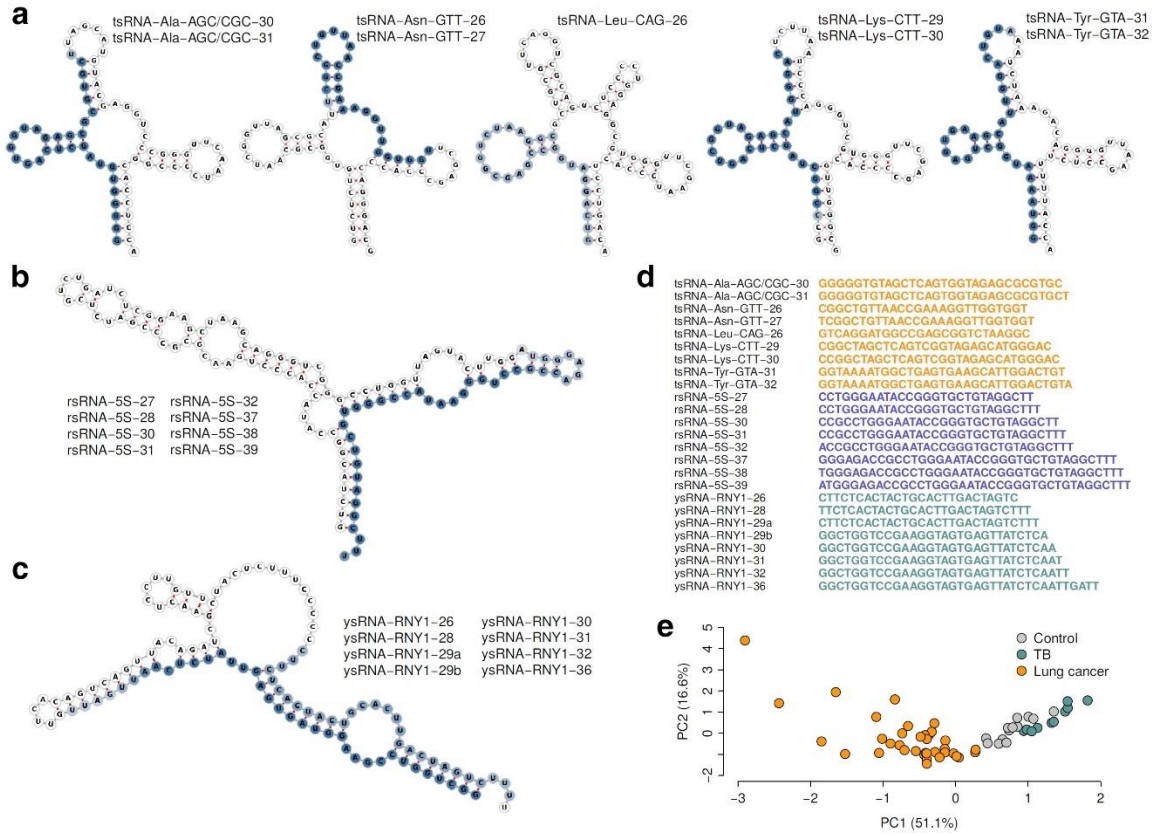

Figure A.6: The TRY-RNA signature

The tsRNA (**a**), rsRNA (**b**), and ysRNA (**c**) species in the signature, respectively. The colored nucleotides indicate the location of the small RNAs on their corresponding parent RNAs. The darkness of the colors (from light blue to steel blue) indicates the overlap level among different small RNA species. Nucleotides with higher overlap levels are darkly colored. **d**, The small RNA sequences of the TRY-RNA signature. **e**, Principal component analysis of the TRY-RNA signature. PC1: the first principal component; PC2: the second principal component.

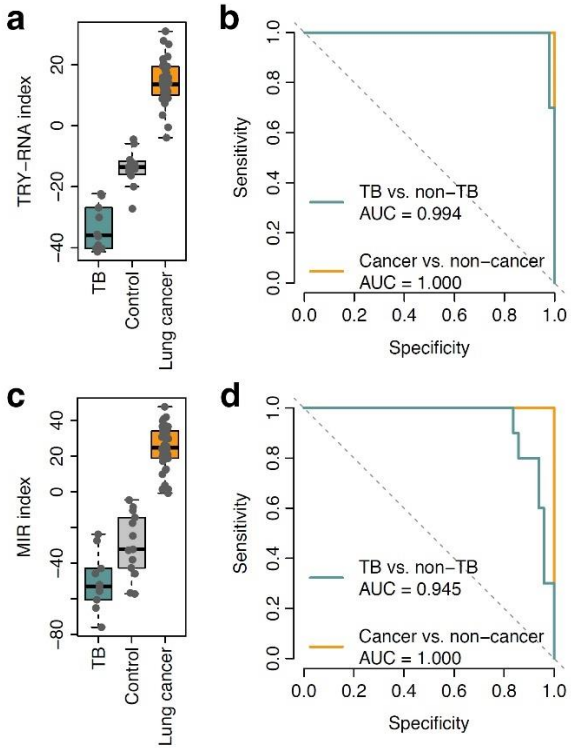

Figure A.7: The TRY-RNA and MIR index in the discovery cohort

Comparison of the TRY-RNA (**a**) and MIR (**c**) index among the control, lung cancer, and TB subjects in the discovery cohort. (B) The ROC curve of the TRY-RNA (**c**) and MIR (**d**) index in distinguishing between lung cancer and non-cancer subjects and between TB and non-TB subjects in the discovery cohort.

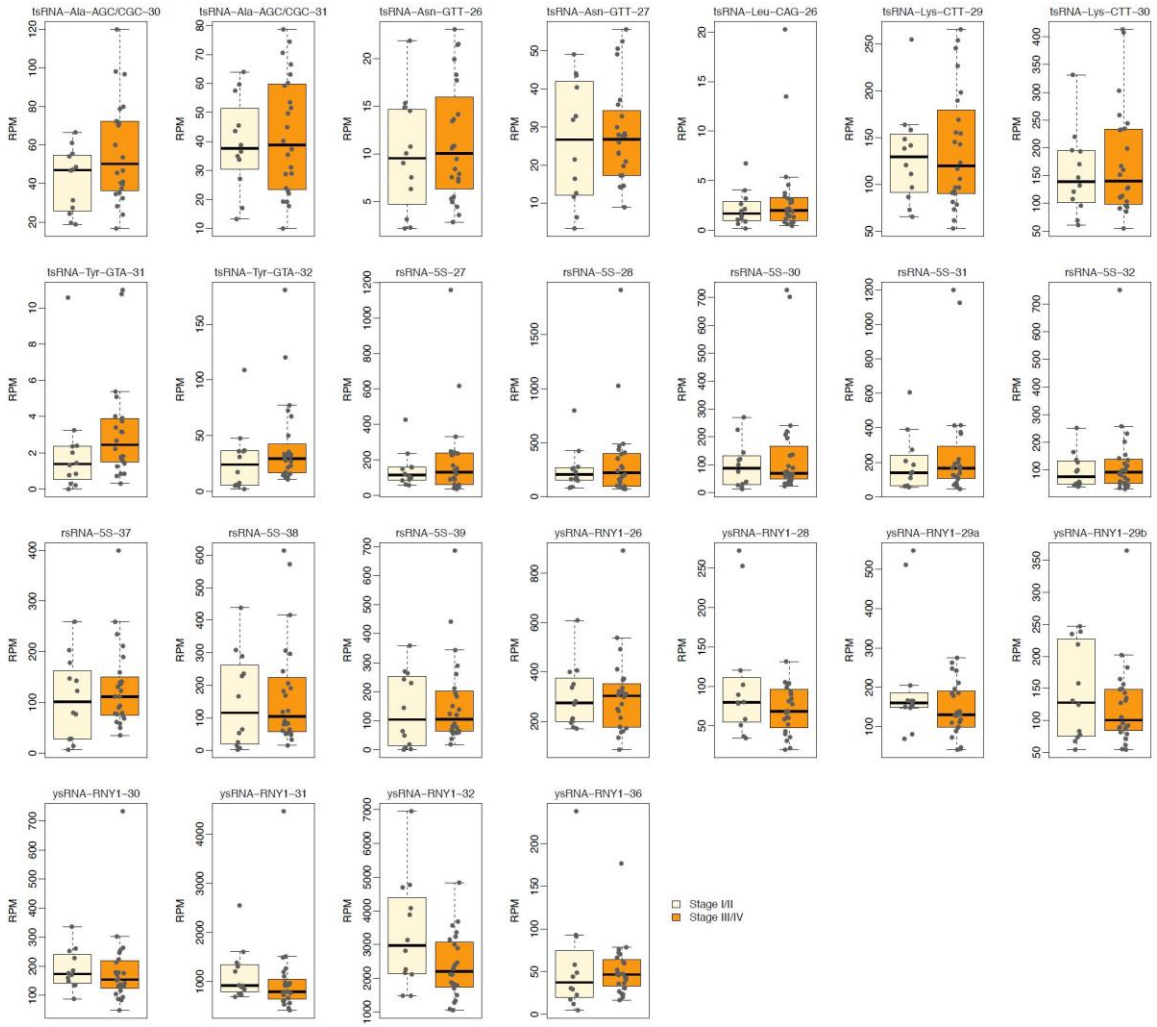

Figure A.8: Comparison of the expression of the small RNA species within the TRY-RNA signature between lung cancer stages

Groupwise comparisons (stage I/II *vs.* III/IV) were performed using a linear model controlling for age and sex. No significant difference was observed (adjusted $P > 0.05$).

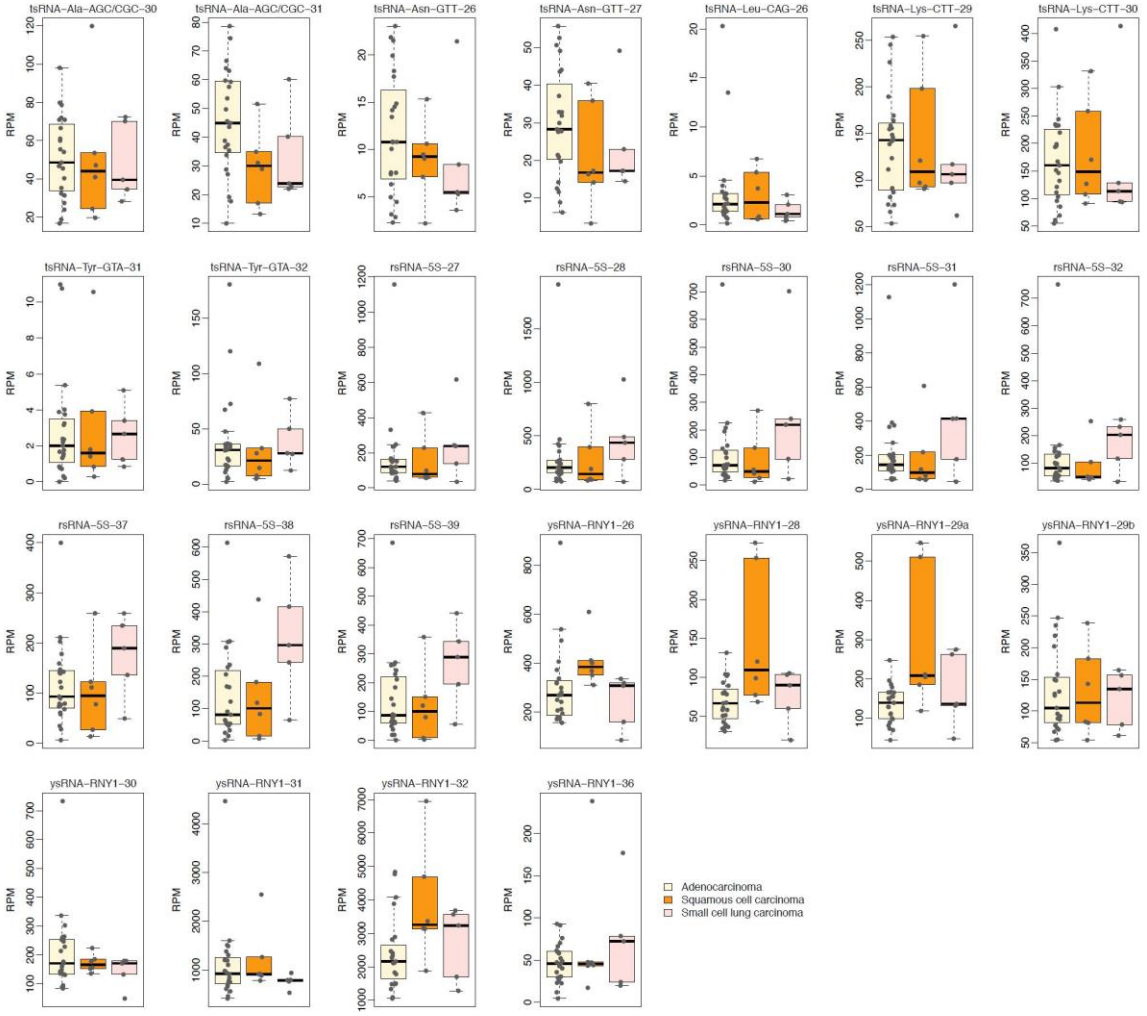

Figure A.9: Comparison of the expression of the small RNA species within the TRY-RNA signature between lung cancer histological types

Groupwise comparisons (*i.e.*, adenocarcinoma *vs.* squamous cell carcinoma, adenocarcinoma *vs.* small cell lung carcinoma, and squamous cell carcinoma *vs.* small cell lung carcinoma) were performed using a linear model controlling for age and sex. Significant difference was only observed for ysRNA-RNY1-28 and ysRNA-RNY1-29a between adenocarcinoma and squamous cell carcinoma (adjusted $P < 0.05$).

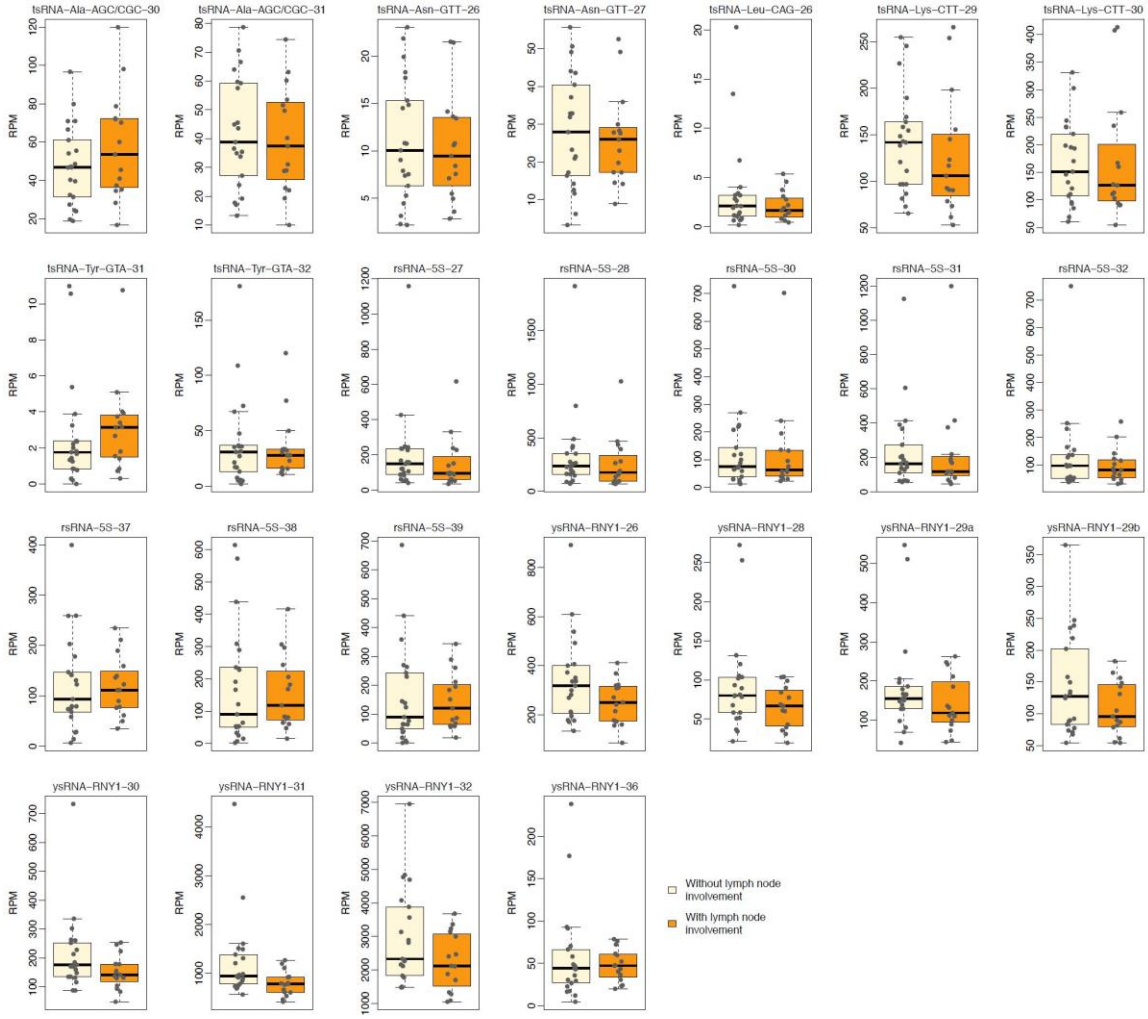

Figure A.10: Comparison of the expression of the small RNA species within the TRY-RNA signature between the lung cancer patients with and without lymph node involvement

Groupwise comparisons were performed using a linear model controlling for age and sex. No significant difference was observed (adjusted $P > 0.05$)

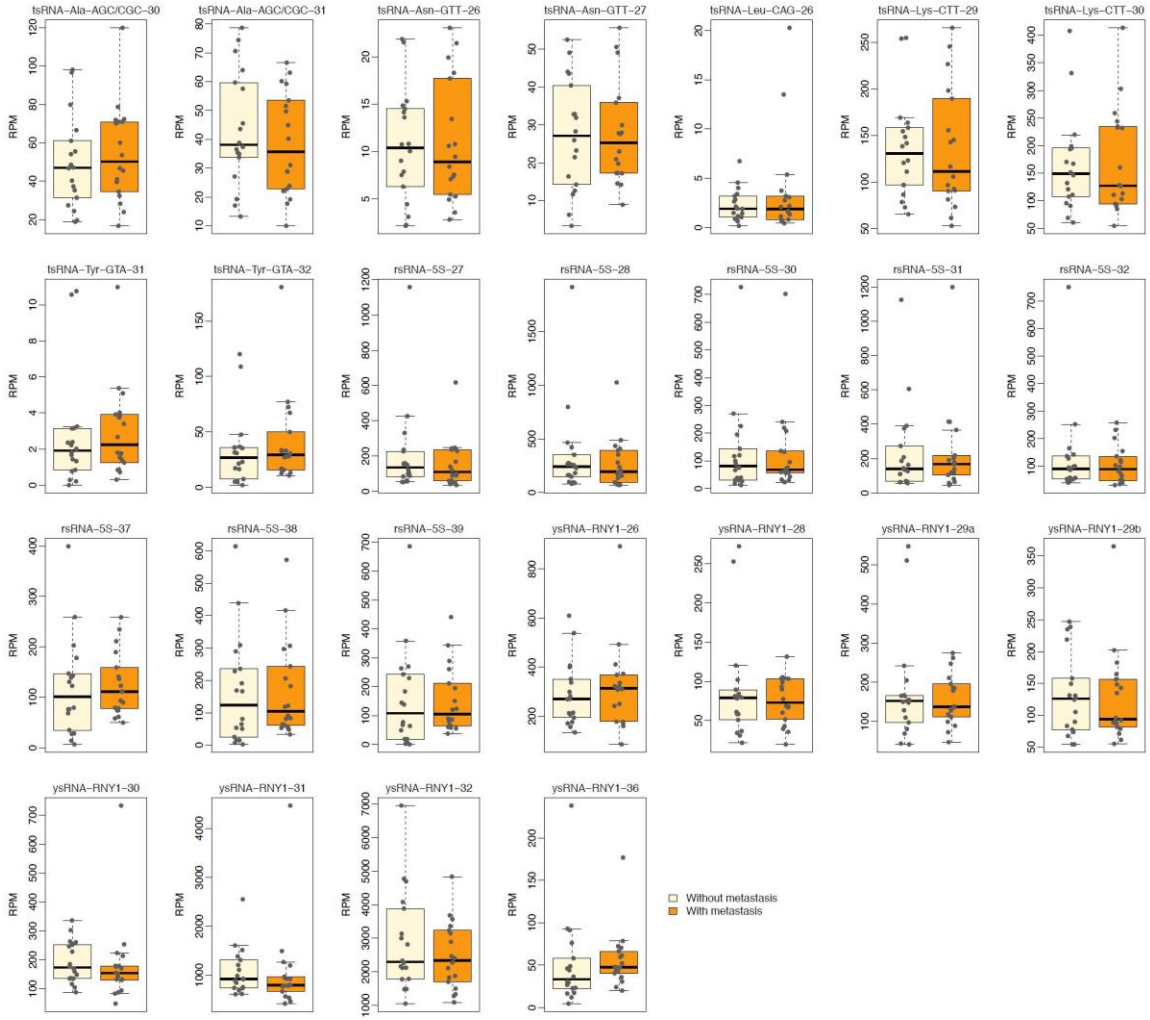

Figure A.11: Comparison of the expression of the small RNA species within the TRY-RNA signature between the lung cancer patients with and without distant metastasis

Groupwise comparisons were performed using a linear model controlling for age and sex. No significant difference was observed (adjusted $P > 0.05$).

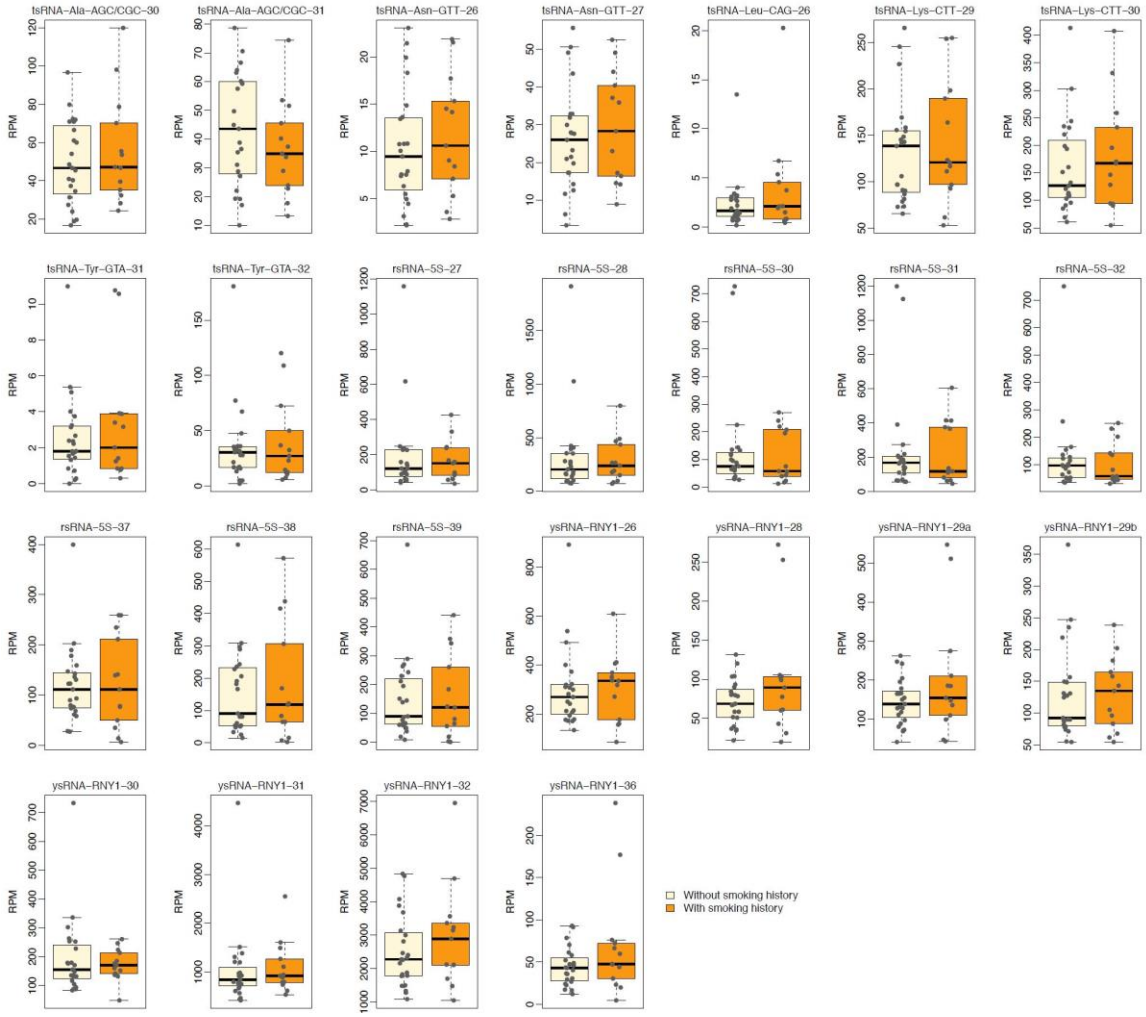

Figure A.12: Comparison of the expression of the small RNA species within the TRY-RNA signature between the lung cancer patients with and without smoking history

Groupwise comparisons were performed using a linear model controlling for age and sex. No significant difference was observed (adjusted $P > 0.05$).

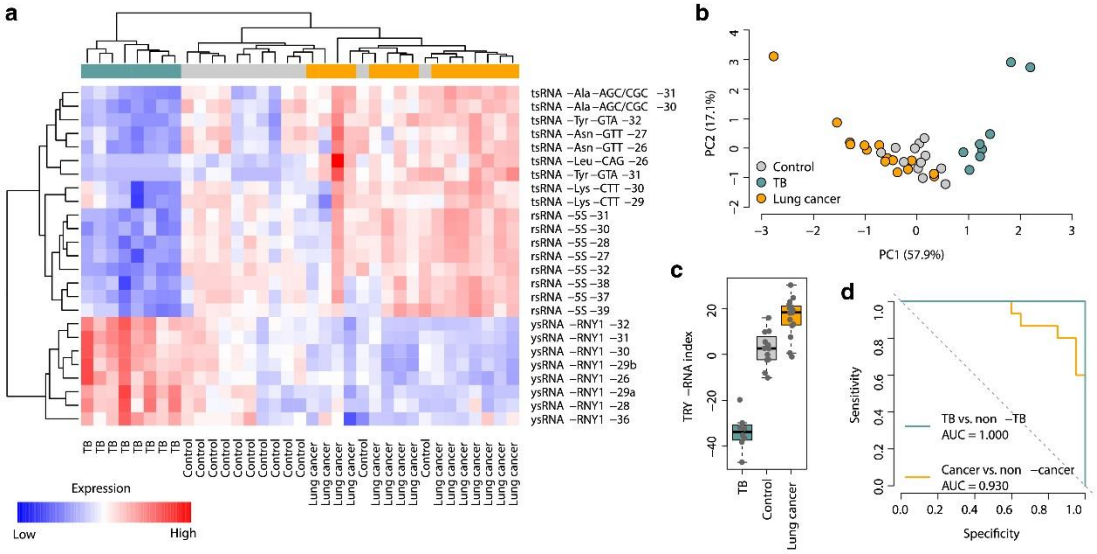

Figure A.13: The performance of the TRY-RNA signature in the validation cohort

a, Expression heatmap of the small RNA species within the TRY-RNA signature in the validation cohort. **b**, Principal component analysis of the TRY-RNA signature. PC1: the first principal component; PC2: the second principal component. PC1 significantly differed between the controls and lung cancer patients (*t*-test: $P = 3.1 \times 10^{-3}$), between the controls and TB patients (*t*-test: $P = 4.7 \times 10^{-6}$), and between the lung cancer and TB patients (*t*-test: $P = 4.1 \times 10^{-8}$). **c**, Comparison of the TRY-RNA index among the control, lung cancer, and TB subjects in the validation cohort. **d**, The ROC curve of the TRY-RNA index in distinguishing between lung cancer and non-cancer subjects and between TB and non-TB subjects in the validation cohort.

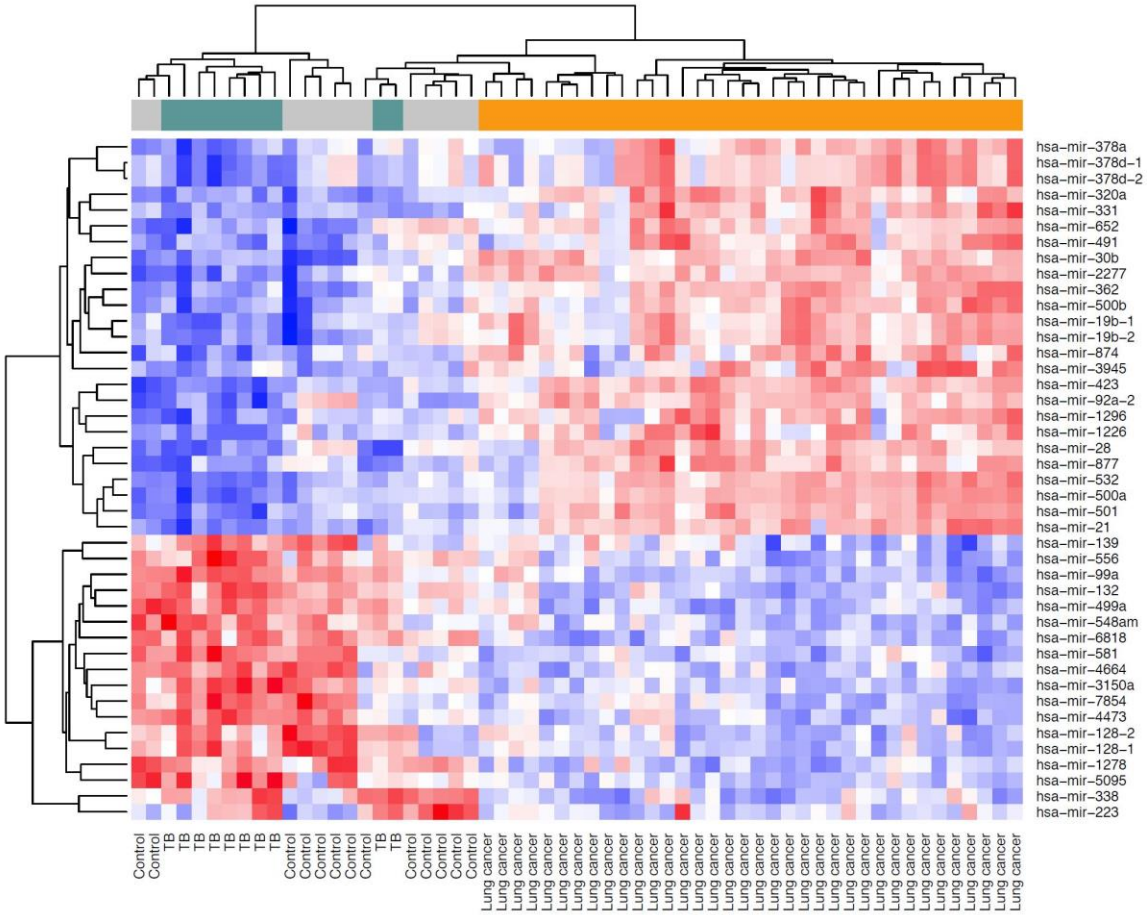

Figure A.14: Expression heatmap of the MIR signature in the discovery cohort

Red represents higher expression while blue stands for lower expression.

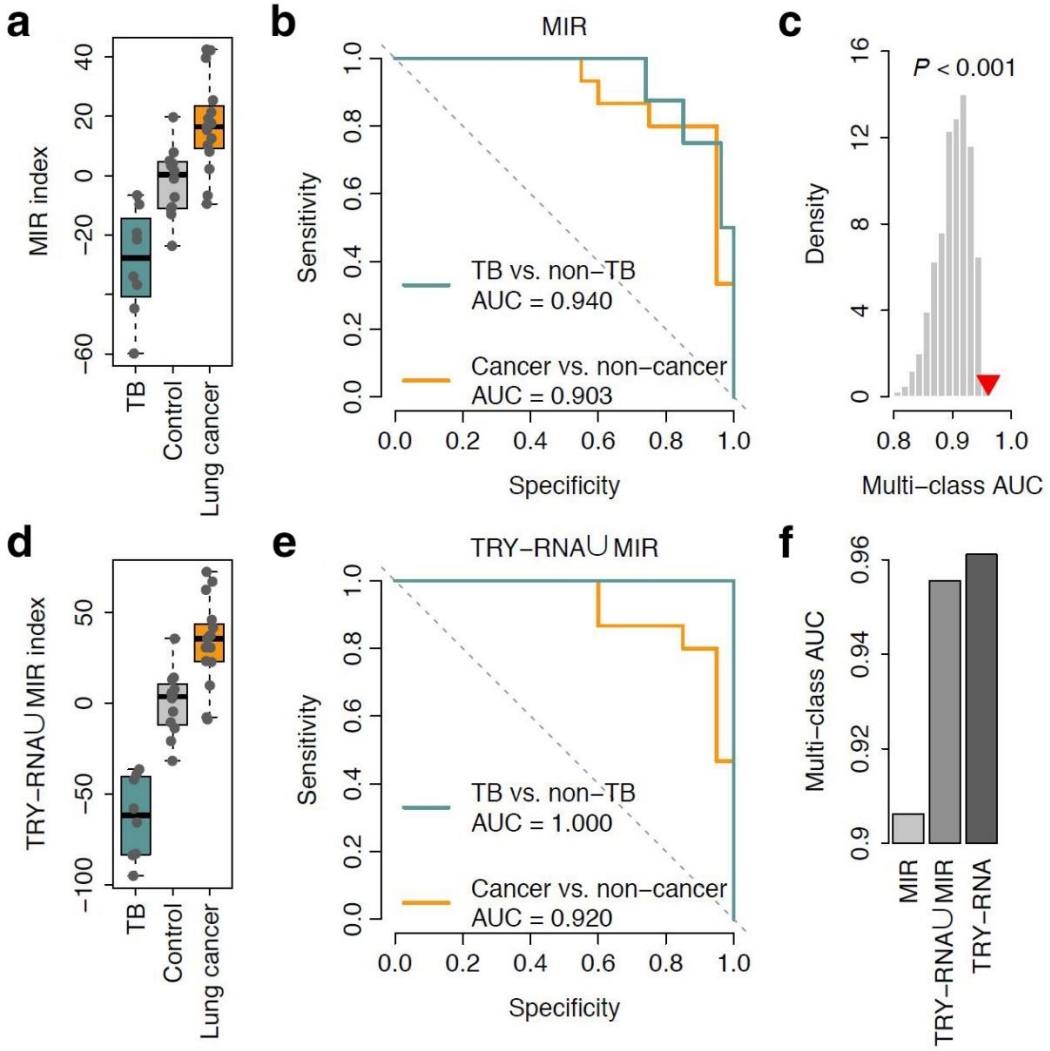

Figure A.15: Comparison between the TRY-RNA and MIR signatures

a, Comparison of the MIR index between the control, lung cancer, and TB subjects in the validation cohort. **b**, The ROC curve of the MIR index in distinguishing between lung cancer and non-cancer subjects and between TB and non-TB subjects in the validation cohort. **c**, The superior classification power of the TRY-RNA signature compared with the MIR signature. The gray histogram shows the distribution of the multi-class *AUC* values of the 1,000 resampled 25-miRNA signatures randomly picked up from the MIR signature. The red triangle represents the multi-class *AUC* of the TRY-RNA signature. The right-tailed *P*-value of the sampling distribution was calculated. **d**, Comparison of the TRY-RNA⁄MIR index between the control, lung cancer, and TB subjects in the validation cohort. **e**, The ROC curve of the TRY-RNA⁄MIR index in distinguishing between lung cancer and non-cancer subjects and between TB and non-TB subjects in the validation cohort. **f**, Comparison of the multi-class *AUC* values between the MIR, TRY-RNA⁄MIR, and TRY-RNA signatures in the validation cohort.

Supplementary materials

Table SA.1: The human subjects of the discovery cohort

Table SA.2: The TRY-RNA signature

Table SA.3: The human subjects of the validation cohort

Table SA.4: The MIR signature

References

Balatti, V., Nigita, G., Veneziano, D., Drusco, A., Stein, G.S., Messier, T.L., Farina, N.H., Lian, J.B., Tomasello, L., Liu, C.G.*, et al.* (2017). tsRNA signatures in cancer. Proceedings of the National Academy of Sciences of the United States of America *114*, 8071-8076.

Chan, P.P., and Lowe, T.M. (2016). GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic acids research *44*, D184-189.

Dhahbi, J.M., Spindler, S.R., Atamna, H., Boffelli, D., and Martin, D.I. (2014). Deep Sequencing of Serum Small RNAs Identifies Patterns of 5' tRNA Half and YRNA Fragment Expression Associated with Breast Cancer. Biomark Cancer *6*, 37-47.

Farina, N.H., Scalia, S., Adams, C.E., Hong, D., Fritz, A.J., Messier, T.L., Balatti, V., Veneziano, D., Lian, J.B., Croce, C.M.*, et al.* (2020). Identification of tRNA-derived small RNA (tsRNA) responsive to the tumor suppressor, RUNX1, in breast cancer. Journal of cellular physiology *235*, 5318-5327.

Hand, D.J., and Till, R.J. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Machine Learning *45*, 171-186.

Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic acids research *42*, D68-73.

Lakhani, P., and Sundaram, B. (2017). Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology *284*, 574-582.

Lewinsohn, D.M., Leonard, M.K., LoBue, P.A., Cohn, D.L., Daley, C.L., Desmond, E., Keane, J., Lewinsohn, D.A., Loeffler, A.M., Mazurek, G.H.*, et al.* (2017). Official American Thoracic Society/Infectious Diseases Society of America/Centers for Disease Control and Prevention Clinical Practice Guidelines: Diagnosis of Tuberculosis in Adults and Children. Clin Infect Dis *64*, 111-115.

McDonough, J.E., Kaminski, N., Thienpont, B., Hogg, J.C., Vanaudenaerde, B.M., and Wuyts, W.A. (2019). Gene correlation network analysis to identify regulatory factors in idiopathic pulmonary fibrosis. Thorax *74*, 132-140.

Pekarsky, Y., Balatti, V., Palamarchuk, A., Rizzotto, L., Veneziano, D., Nigita, G., Rassenti, L.Z., Pass, H.I., Kipps, T.J., Liu, C.G.*, et al.* (2016). Dysregulation of a family

of short noncoding RNAs, tsRNAs, in human cancer. Proceedings of the National Academy of Sciences of the United States of America *113*, 5071-5076.

Qian, Z., Liu, H., Li, M., Shi, J., Li, N., Zhang, Y., Zhang, X., Lv, J., Xie, X., Bai, Y.*, et al.* (2018). Potential Diagnostic Power of Blood Circular RNA Expression in Active Pulmonary Tuberculosis. EBioMedicine *27*, 18-26.

Qian, Z., Lv, J., Kelly, G.T., Wang, H., Zhang, X., Gu, W., Yin, X., Wang, T., and Zhou, T. (2016). Expression of nuclear factor, erythroid 2-like 2-mediated genes differentiates tuberculosis. Tuberculosis (Edinb) *99*, 56-62.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.

Shi, J., Zhang, Y., Zhou, T., and Chen, Q. (2019). tsRNAs: The Swiss Army Knife for Translational Regulation. Trends Biochem Sci *44*, 185-189.

Su, Z., Wilson, B., Kumar, P., and Dutta, A. (2020). Noncanonical Roles of tRNAs: tRNA Fragments and Beyond. Annual review of genetics *54*, 47-69.

Zhang, Y., Shi, J., Rassoulzadegan, M., Tuorto, F., and Chen, Q. (2019). Sperm RNA code programmes the metabolic health of offspring. Nat Rev Endocrinol *15*, 489-498.

Zhang, Y., Zhang, X., Shi, J., Tuorto, F., Li, X., Liu, Y., Liebers, R., Zhang, L., Qu, Y., Qian, J.*, et al.* (2018). Dnmt2 mediates intergenerational transmission of paternally acquired metabolic disorders through sperm small non-coding RNAs. Nat Cell Biol *20*, 535-540.