1-1-2013

Methods in and Applications of the Sequencing of Short Non-Coding RNAs

Paul Ryvkin
University of Pennsylvania, paulnik@gmail.com

Methods in and Applications of the Sequencing of Short Non-Coding RNAs

Abstract
Short non-coding RNAs are important for all domains of life. With the advent of modern molecular biology their applicability to medicine has become apparent in settings ranging from diagonistic biomarkers to therapeutics and fields ranging from oncology to neurology. In addition, a critical, recent technological development is high-throughput sequencing of nucleic acids. The convergence of modern biotechnology with developments in RNA biology presents opportunities in both basic research and medical settings. Here I present two novel methods for leveraging high-throughput sequencing in the study of short non-coding RNAs, as well as a study in which they are applied to Alzheimer's Disease (AD). The computational methods presented here include High-throughput Annotation of Modified Ribonucleotides (HAMR), which enables researchers to detect post-transcriptional covalent modifications to RNAs in a high-throughput manner. In addition, I describe Classification of RNAs by Analysis of Length (CoRAL), a computational method that allows researchers to characterize the pathways responsible for short non-coding RNA biogenesis. Lastly, I present an application of the study of non-coding RNAs to Alzheimer's disease. When applied to the study of AD, it is apparent that several classes of non-coding RNAs, particularly tRNAs and tRNA fragments, show striking changes in the dorsolateral prefrontal cortex of affected human brains. Interestingly, the nature of these changes differs between mitochondrial and nuclear tRNAs, implicating an association between Alzheimer's disease and perturbation of mitochondrial function. In addition, by combining known genetic factors of AD with genes that are differentially expressed and targets of regulatory RNAs that are differentially expressed, I construct a network of genes that are potentially relevant to the pathogenesis of the disease. By combining genetics data with novel results from the study of non-coding RNAs, we can further elucidate the molecular mechanisms that underly Alzheimer's disease pathogenesis.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Genomics & Computational Biology

First Advisor
Li-San Wang

Keywords
Alzheimer's disease, machine learning, non-coding RNA, RNA, RNA modification, sequencing

Subject Categories
Bioinformatics | Genetics | Molecular Biology

METHODS IN AND APPLICATIONS OF THE SEQUENCING OF SHORT NON-CODING RNAS

Paul Ryvkin

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2013

Supervisor of Dissertation

_____

Li-San Wang, Ph.D.
Assistant Professor of Pathology and Laboratory Medicine

Graduate Group Chairperson

_____

Maja Bucan, Ph.D.
Professor of Genetics

Dissertation Committee:
James Eberwine, Ph.D. (Chair)
Professor of Pharmacology
Brian Gregory, Ph.D.
Assistant Professor of Biology
F. Bradley Johnson, M.D. Ph.D.
Associate Professor of Pathology and Laboratory Medicine
Tandy Warnow, Ph.D.
Professor of Computer Sciences at University of Texas at Austin

METHODS IN AND APPLICATIONS OF THE SEQUENCING OF SHORT NON-CODING RNAS

Paul Ryvkin

Dedication

This work is dedicated to my loving parents, Mark and Yelena Ryvkin, to whom I owe everything I've achieved and am yet capable of achieving.

Acknowledgements

First I must thank my thesis advisor, Li-San Wang, whose careful guidance and perseverance show through in this work. Thanks also go to my thesis committee who graciously took the time to provide useful input throughout the process.

This section would not be complete without thanking all of my former and current labmates, particularly: Fan Li for transforming my ugly hacks into useful apps, Kajia Cao for keeping my head out of the clouds, Fanny Leung for helping with the benchwork and the machine learning algorithms, Otto Valladares for keeping the servers humming, Micah Childress for putting a public face on my software, and everyone else in the Wang lab.

My appreciation goes to the staff at the Institute for Biomedical Informatics, particularly Hannah Chervitz and Tiffany Barlow for their organizational prowess. I'd also like to thank all the GCB students I've known; the elder for their sage advice, the co-matriculating for their commiseration, and the younger for excellent times had.

In the course of my time at Penn I've worked with many, many other researchers, without whom this work would not have been possible. I'd like to thank Brad Johnson, who taught me so many important molecular biology techniques ranging from nucleic acids extraction to keeping the supernatant. Also to thank is Brian Gregory and everyone in his lab, particularly Isabelle Dragomir for her help with the sequencing library preparation and Lee Vandivier and Ian Silverman for their help with validating experiments. Thanks also go to Alice-Chen Plotkin for her help with tissue processing, Vivianna Van Deerlin for her help with large-scale RNA extraction, Theresa Schuck for her help with tissue dissection, and Virginia Lee for her ever insightful input. I particularly appreciate everyone at the Center for Neurodegenerative Disease Research and everyone in Gerard Schellenberg's lab for being gracious hosts for much of this work.

A special thank-you goes to John Trojanowski and the Institute on Aging for providing the funding for the Alzheimer's study, which yielded almost all of the data necessary for this work. Additional funding came from the National Institutes of Health, National Institute of General

Medical Sciences, the National Human Genome Research Institute, the National Institute on Aging, Penn Alzheimer's Disease Center, and the National Science Foundation.

Finally I'd like to thank my office in Blockley Hall for sheltering my computer from the elements, my bicycle for faithfully transporting me from point A to point B, my two cats for being endlessly fascinating felids, the never-boring city of Philadelphia, the food and company at Grace Tavern, the scenery of Rittenhouse Park, and the game of Bridge (special thanks to Kathleen Sprouffske, Miler Lee, Rumen Kostadinov, and Aaron Goodman). I also thank my wonderful girlfriend Chrystelle Browman for supporting me despite the unique challenges of dating a PhD student.

ABSTRACT

METHODS IN AND APPLICATIONS OF THE SEQUENCING OF SHORT NON-CODING RNAS

Paul Ryvkin

Li-San Wang

Short non-coding RNAs are important for all domains of life. With the advent of modern molecular biology their applicability to medicine has become apparent in settings ranging from diagonistic biomarkers to therapeutics and fields ranging from oncology to neurology. In addition, a critical, recent technological development is high-throughput sequencing of nucleic acids. The convergence of modern biotechnology with developments in RNA biology presents opportunities in both basic research and medical settings. Here I present two novel methods for leveraging high-throughput sequencing in the study of short non-coding RNAs, as well as a study in which they are applied to Alzheimer's Disease (AD). The computational methods presented here include High-throughput Annotation of Modified Ribonucleotides (HAMR), which enables researchers to detect post-transcriptional covalent modifications to RNAs in a high-throughput manner. In addition, I describe Classification of RNAs by Analysis of Length (CoRAL), a computational method that allows researchers to characterize the pathways responsible for short non-coding RNA biogenesis. Lastly, I present an application of the study of non-coding RNAs to Alzheimer's disease. When applied to the study of AD, it is apparent that several classes of non-coding RNAs, particularly tRNAs and tRNA fragments, show striking changes in the dorsolateral prefrontal cortex of affected human brains. Interestingly, the nature of these changes differs between mitochondrial and nuclear tRNAs, implicating an association between Alzheimer's disease and perturbation of mitochondrial function. In addition, by combining known genetic factors of AD with genes that are differentially expressed and targets of regulatory RNAs that are differentially expressed, I construct a network of genes that are potentially relevant to the pathogenesis of the disease. By combining genetics data with novel results from the study of non-

coding RNAs, we can further elucidate the molecular mechanisms that underly Alzheimer's disease pathogenesis.

CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

1. Introduction

1.1. RNA Biology

1.1.1. The Central "Dogma"

The central hypothesis (or as Francis Crick infamously and erroneously coined it, the "central dogma") [35,36] of molecular biology outlines the relationship between three important types of organic molecules: DNA (deoxyribonucleic acids), RNA (ribonucleic acids), and proteins (**Figure 1.1**). The totality of each type of molecule in the cell is referred to as the *genome*, the *transcriptome*, and the *proteome*, respectively. Under this framework, information flows from DNA to RNA and then to proteins; DNA serves as a template for *transcription* of RNA, which in turn serves as a template for *translation* into protein. Proteins form *enzymes* which carry out a range of functions throughout the cell and are generally responsible for *phenotype*, or the appearance and behavior of the organism. While we now know that there are many exceptions to this view [12,91,112,140,150] (**Figure 1.2**), it is a useful start for describing molecular biology.

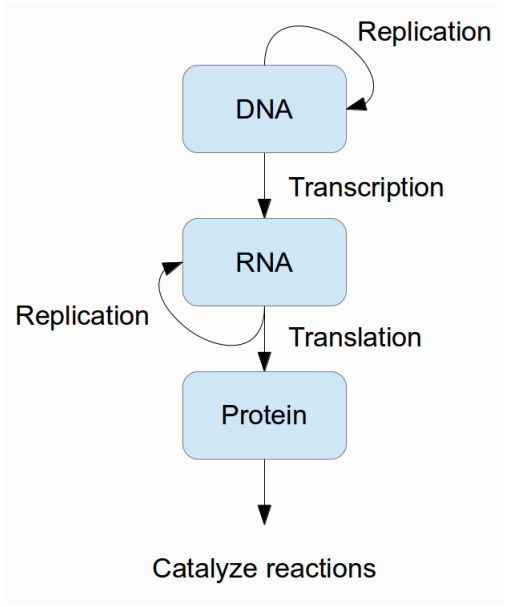

Figure 1.1 – The central hypothesis of molecular biology.

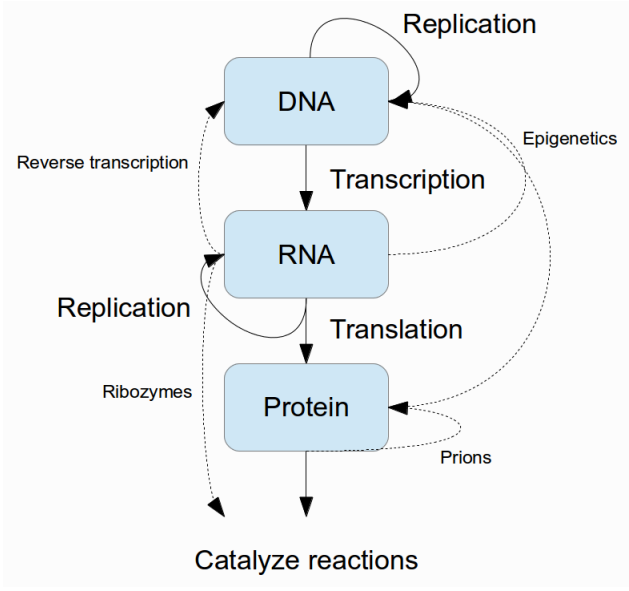

Figure 1.2 – The central hypothesis revised.

DNA can be considered a fixed information storage medium for the cell. Exceptions to this picture of DNA include the entire field of *epigenetics* which seeks to describe dynamic modifications to DNA, as well as the study of the processes of DNA replication and repair. In general, however, DNA serves only as a template and is not responsible for catalyzing other types of reactions.

RNA, in contrast, exists in a constant state of flux via creation (transcription from DNA) and destruction (finely controlled turnover by enzymes). Similarly, proteins, which comprise enzymes, exist in a constant state of flux. For many years, proteins alone were considered to be the workhorse of the cell – after all, they catalyze nearly all of the reactions necessary to support life while DNA and RNA "merely" store information. However, with the discovery of catalytic RNAs (*ribozymes*) [27,69,94], these molecules are now appreciated as more than simple "messengers" between DNA and proteins. It is especially difficult to write off RNAs since the machinery that translates RNA into protein (the *ribosome*) is itself made up of RNA; indeed, it has been shown that the RNA (not the protein) component of this machinery is responsible for its activity [120]. RNA is therefore a key component of the cellular machinery and not simply a transitory messenger.

Like the other ubiquitous organic polymers central to life (DNA and proteins), RNA primarily stores information by way of its *sequence*. While DNA is a polymer of the deoxyribonucleotides deoxyadenosine (dA), deoxycytidine (dC), deoxyguanosine (dG), and deoxythymidine (dT), RNA is a polymer of the ribonucleotides adenosine (A), cytidine (C), guanosine (G), and uridine (U) [6]. The key differences are RNA's inclusion of a hydroxyl group where DNA is missing one, the substitution of uridine for thymidine, and RNA's propensity to exist in a greater variety of structural forms. Analogous to DNA, it is the sequential order of the ribonucleotides that form the primary information content of RNA. Another form of information stored by RNA is its structure; RNAs are prone to fold into particular geometries which can be important for their catalytic functions [123,127] (**Figure 1.3**). The *primary structure* of an RNA is

its sequence. Its *secondary structure* is a graph whose nodes are nucleotides and whose edges represent Watson-Crick and wobble-pairing interactions between pairs of these nucleotides. Its *tertiary structure* describes long-range interactions between its base-paired and/or unpaired sections. Finally, the *quaternary structure* of an RNA models its interactions with other molecules. In addition to the folding geometry of the RNA, a third form of information is the presence of non-canonical nucleotides formed by covalent modification of the standard four [3,4,37,166,169] – in Chapter 3 I present a method for detecting these non-canonical nucleotides.

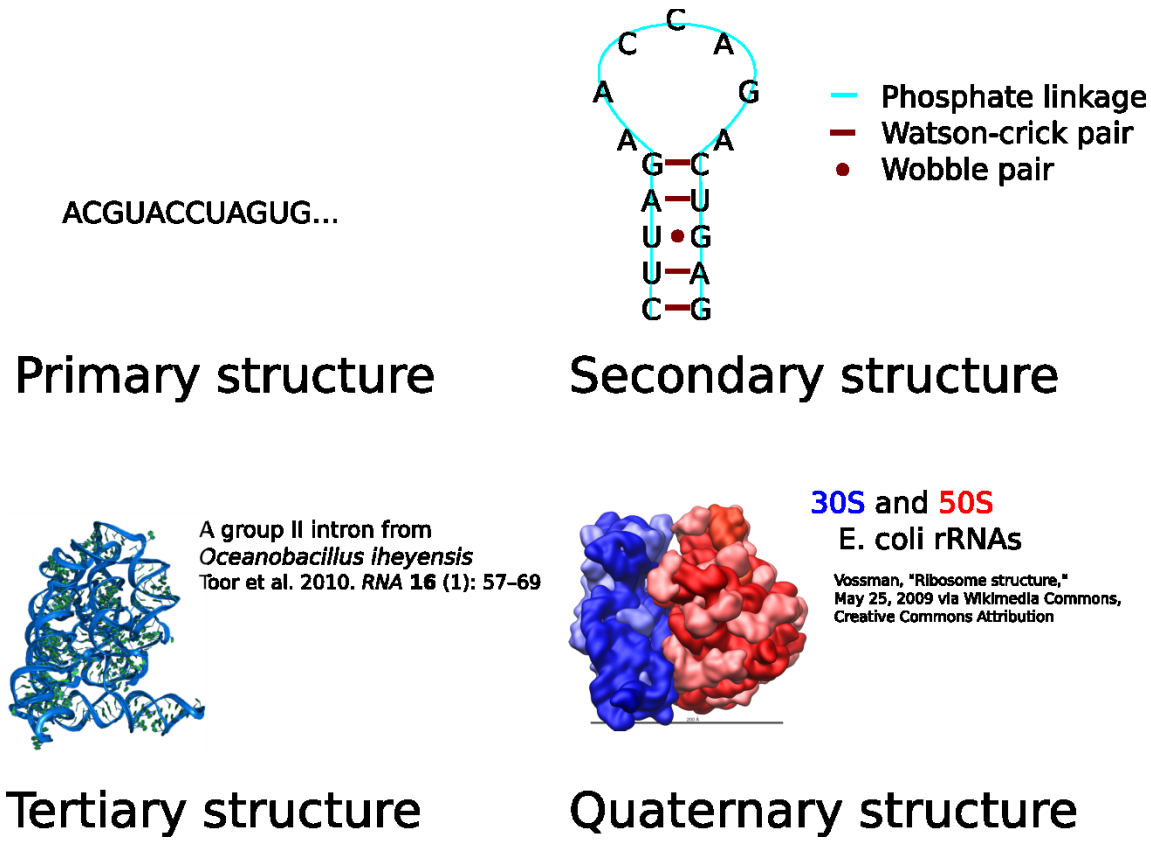

Figure 1.3 – The structure of RNAs.

1.1.2. Protein-coding RNAs

RNAs can be broadly categorized into two groups: those that code for proteins (coding RNAs) and those that do not (non-coding, or ncRNAs). The only extant class of coding RNAs is messenger RNAs (mRNAs) – however, not all mRNAs code for proteins. In higher eukaryotes, mRNAs are transcribed by the enzyme RNA polymerase II and undergo a sophisticated maturation process from the original mRNA transcript [117]: they can be *spliced* into various *isoforms* [65], they are capped by a special chemical structure on the *5' end* [133], they are *polyadenylated* on the *3' end*, and their sequence can be dynamically changed [139] (*RNA editing*) and chemically modified (*RNA modification)*. The terms 5' and 3' correspond to the exposed atom of the ribose sugar in the ribonucleotide – generally a 5' triphosphate on one end and a 3' hydroxyl group on the other end. Since mRNAs are translated into proteins from 5' to 3', these are conventionally depicted as the left and right ends of the molecule, respectively. In eukaryotic splicing, multiple alternative forms of an RNA transcript are generated when the cellular splicing machinery removes sections called *introns* and concatenates together sections designated as *exons*, which usually contain the coding portion of the transcript (i.e., the sequence that will determine the translated protein). Thus one gene may produce many distinct mRNAs with varying sequences which are then translated into proteins with a variety of functions. Capping, in eukaryotic organisms, refers to the addition of the ribonucleotide N7-methylguanosine (m^7G) to the 5' carbon of the mRNA via an unconventional 5'-5' triphosphate linkage. This cap serves to stabilize the mRNA and promote its export from the nucleus. Polyadenylation is a process whereby a homopoylmer of adenosines is sequentially added to the 3' end of the mRNA. Among other functions, this poly(A) "tail" regulates enzymatic degradation of the mRNA from the 3' end. Nearly all eukaryotic mRNAs are polyadenylated with the notable exception of the histone genes, where the 3' terminus is designated by a small stem-loop RNA structure. The process of RNA editing generally consists of post-transcriptional changes in the sequence of an RNA. In the case of eukaryotic mRNAs, this is usually a deamination of adenosine to inosine or deamination of cytidine to uridine [85]. Inosine has similar base-pairing properties to guanosine, but overall it is

far less specific in its base-pairing specificity. These changes to an mRNA's sequence can affect its alternative splicing, stability, and even the eventual protein sequence that is coded. Other types of changes to an RNA's sequence, which always produce non-canonical nucleotides, are termed RNA modifications. Examples of RNA modifications are the methylation of guanosine at the carbon 2 amine (producing N2-methylguanosine or m^2G) and the isomerization of uridine into its C-glycoside pseudouridine (Ψ). These types of modifications are believed to be rare in protein-coding mRNAs, but the search for them is an active field of research. So far, it seems that the non-canonical nucleotides 5-methylcytidine (m^5C) and N6-methyladenosine (m^6A) can be found in mRNAs transcriptome-wide [115]. Furthermore, the recent discovery that a gene whose variants are found to be associated with obesity in humans, FTO, is an adenosine N6-methyltransferase suggests that these modifications may play a very important role in human disease [57,84].

1.1.3. Non-coding RNAs

While protein-coding mRNAs are important for deciding the sequences of proteins, the most abundant RNAs in the cell by far are non-coding RNAs; ribosomal RNA (rRNA) can make up over 80% of all the RNA in mammalian cells. The next most abundant class of non-coding RNAs, transfer RNAs (tRNAs), can make up another 10%. Not only are non-coding RNAs the most abundant RNAs in the cell, they are also the most evolutionarily conserved: all cellular life on earth relies on ribosomes, and thus ribosomal RNA, and the similarity of its sequence among disparate organisms is great enough for it to act a universal phylogenetic character [128]. The universality of ribosomal RNA, combined with its sufficiency for ribosomal function is a central piece of evidence supporting the hotly-debated "RNA world" hypothesis which claims that the use of RNA as an information storage medium preceded DNA's on Earth [26].

Aside from their lack of protein-coding capacity, there are many fundamental differences between coding and non-coding RNAs, ranging from how they encode information to how they

are processed. For example, while the non-canonical nucleotide modifications described in Section 1.1.2 are thought to be rare in mRNAs, they are ubiquitous in non-coding RNAs. The most abundant non-canonical nucleotide in the cell, pseudouridine, is commonly found in rRNA and tRNA [73].

Unlike protein-coding mRNAs, there is great diversity in the non-coding RNA population [83,114] (**Table 1.1**). Unfortunately, producing a consistent nomenclature of non-coding RNAs is a difficult task, and currently it proceeds in an *ad hoc* manner publication by publication. For example, while some classes of RNA are defined by their location in the cell, others are defined by the genomic neighborhood of their DNA template. An initial useful subdivision of non-coding RNAs is by their size: generally ncRNAs shorter than around 50 nucleotides (nt) are considered short non-coding RNAs, or "small RNAs," while longer ones are referred to as long non-coding RNAs (lncRNAs). Section 1.1.4 describes the many types of short non-coding RNAs, while this section focuses on the longer ones.

Table 1.1 – A compendium of non-coding RNAs found in animals.

Abbreviation	Name	Biological role	Example(s)
rRNA	Ribosomal RNA	Translation	5S rRNA
tRNA	Transfer RNA	Translation	$tRNA^{Met}_{CAU}$
snoRNA	Small nucleolar RNA	RNA modification	SNORD115
snRNA	Small nuclear RNA	mRNA splicing	U1
scRNA	Small cytoplasmic RNA	Various	hY1
srpRNA	Signal recognition particle RNA	Protein localization	
lincRNA	Long intergenic non-coding RNA	Various	XIST, TSIX, MALAT1
miRNA	Micro RNA	mRNA silencing	let-7
piRNA	*Piwi*-interacting RNA	Transposon silencing	piR-53941
tRF	tRNA fragment	Unknown	$tRNA^{Met}_{CAU}$ 5' half
paRNA	Promoter-associated RNA	Unknown	EF1a promoter
vtRNA	Vault RNA	Unknown; drug resistance	VTRNA1-1
aRNA	Antisense RNA	mRNA regulation	BACE1-AS
natRNA	Natural antisense transcript RNA	Unknown	HAS2-AS1
-	Transposable elements	Self replication	SINEs and LINEs
Hammerhead	Hammerhead ribozyme	mRNA regulation	C10orf118
TERC	Telomerase RNA component	Telomere extension	TERC
RNase P	Ribonuclease P RNA component	Cleavage of pre-tRNAs	RPPH1

Ribosomal RNA is largely transcribed by RNA polymerase I and is central to an organelle within the cell called the ribosome [122]. Ribosomes are responsible for translating mRNAs into proteins. In eukaryotes the ribosome is made up of a small subunit (SSU) and large subunit (LSU). In the human genome ribosomal RNA exists in many copies (as rDNA), and often in long tandem arrays, which have long presented an obstacle to assembly of the human genome due to their repetitive nature. Ribosomal RNA maturation takes place in the nucleolus, a small substructure of the nucleus, where it is spliced and modified in myriad ways by other RNAs and ribonucleoprotein (RNP) complexes. Importantly, it is the structure of the rRNA that is responsible for its function, not necessarily its sequence; structure-over-sequence is a common theme among ncRNAs.

The next most abundant class of ncRNA is transfer RNA [127]. Transfer RNAs are transcribed by RNA polymerase III and tend to be around 70 nt in length; they fold into a distinctive “cloverleaf” secondary structure with an L-shaped tertiary structure. Like rRNA, the DNA genes from which they are transcribed (tDNA) exist with high copy number in mammalian genomes [13]. The function of tRNA is to act as an intermediary between mRNA and the ribosome. The *acceptor arm* of a tRNA is covalently bonded to a specific amino acid by a highly conserved family of proteins called tRNA aminoacyl synthases. The *anticodon loop* of a tRNA contains a three-nucleotide sequence called the “anticodon.” When an mRNA is being translated by a ribosome, the appropriate tRNA associates with the mRNA’s current *codon* (three-letter code associated with an amino acid) by way of sequence complementarity. Thus a tRNA provides a link between particular codon sequences and particular amino acids, giving rise to the genetic code (**Table 1.2**). In tRNAs, both the structure and sequence are of critical importance – their structure allows for the appropriate interaction with the ribosome while their sequence provides specificity for particular codons. Notably, there are fewer tRNA anticodons encoded in the genome than there are complementary codons in the genetic code. This is because one tRNA can bind to multiple codons by way of covalent RNA modifications in the anticodon loop, yielding nucleotides with degenerate base pairing properties (e.g., inosine). Structural perturbations thus

induced adjacent to the anticodon can also alter the specificity of the codon-binding. Like mRNAs and rRNAs, tRNAs can also have introns that are spliced out [1].

Table 1.2 – The eukaryotic nuclear genetic code.

RNA Codon	Amino acid
UUA	Serine (Ser)
UUG	
UCU	
UCC	
AGU	
AGC	
UUA	Leucine (Leu)
UUG	
CUU	
CUC	
CUA	
CUG	
GUU	Valine (Val)
GUC	
GUA	
GUG	
CCU	Proline (Pro)
CCC	
CCA	
CCG	
ACU	Threonine (Thr)
ACC	
ACA	
ACG	
GCU	Alanine (Ala)
GCC	
GCA	
GCG	
UGU	Cysteine (Cys)
UGC	
UAU	Tyrosine (Tyr)
UAC	

RNA Codon	Amino acid
CGU	Arginine (Arg)
CGC	
CGA	
CGG	
AGA	
AGG	
GGU	Glycine (Gly)
GGC	
GGA	
GGG	
AUU	Isoleucine (Ile)
AUC	
AUA	
UUU	Phenylalanine (Phe)
UUC	
CAU	Histidine (His)
CAC	
CAA	Glutamine (Gln)
CAG	
AAU	Asparagine (Asn)
AAC	
AAA	Lysine (Lys)
AAG	
GAU	Aspartic acid (Asp)
GAC	
GAA	Glutamic acid (Glu)
GAG	
UAA	Stop codon
UAG	
UGA	
UGG	Tryptophan (Trp)
AUG	Methionine (Met)

Small nucleolar RNAs (snoRNAs) are, as their name suggests, non-coding RNAs that are generally localized to the nucleolus (but also Cajal bodies) [49]. There are three main subclasses of small nucleolar RNAs, each having a different set of structural and sequence motifs: C/D box, H/ACA box, and small Cajal body-specific (scaRNA). The main function of snoRNAs is to guide covalent modification of other RNAs, ranging from rRNA to small nuclear RNAs (snRNAs), via small nucleolar ribonucleoprotein (snoRNP) complexes. In general, C/D box snoRNAs guide methylation of RNAs while H/ACA box snoRNAs guide pseudouridylation of RNAs. One notable exception is the C/D box snoRNA SNORD115, which has complementarity to the serotonin 2 C receptor mRNA and alters its splicing [88].

Small nuclear RNAs (snRNAs) largely comprise the RNA component of the spliceosome; that is, they make up the machinery responsible for splicing of RNAs [51,118,154]. As their name suggests, they are largely localized to the nucleus. There are several families of snRNAs with names such as U1, U2, and so on. In conjunction with proteins they form small nuclear ribonucleoprotein (snRNP) complexes, which form the spliceosome. As with the other ncRNA types described, their genes exist in high copy number scattered throughout mammalian genomes [107].

A somewhat mysterious and only recently described class of non-coding RNAs is that of long intergenic non-coding RNAs (lincRNAs) [24,87]. lincRNAs look very similar to mRNAs – they are transcribed by RNA polymerase II and tend to be polyadenylated and spliced – but they do not code for proteins and often localize to the nucleus rather than the cytoplasm. While some notable examples of lincRNAs, such as Xist [32] and MALAT1 [82] have been well known for quite some time, the recent application of high-throughput RNA-sequencing has illuminated many more lincRNAs with varying levels of abundance and tissue specificity. Their function and biological relevance are largely unknown.

1.1.4. Short non-coding RNAs (small RNAs)

Short non-coding RNAs, or small RNAs (smRNAs), play an important role in higher eukaryotic transcriptomes. RNAs that are considered small RNAs tend to be less than 45 nt in length, although there is no standard cutoff for the definition. They are almost always the product of processing a longer transcript rather than being independently transcribed directly from the genome. The pathways responsible for generation of smRNAs generally consist of a number of proteins and ribonucleoprotein complexes that process the precursor transcript in tandem and in parallel. These pathways tend not to be as conserved across evolutionary distances as some highly conserved proteins. Plants and animals, for example, have rather distinct smRNA pathways that behave in quite different ways as a whole.

The best characterized class of smRNA to date is the microRNA [145]. MicroRNAs are a particular subtype of small interfering RNA (siRNA) [55]. Small interfering RNAs were first described by Craig C. Mello, Andrew Fire, and others in their 1998 *Nature* article, for which Mello and Fire won a Nobel Prize in 2006. They are short (~21 nt) double-stranded RNAs which promote gene silencing through a variety of methods – usually by either catalyzing degradation of an mRNA transcript or inhibition of translation of an mRNA into its concomitant protein. They target specific mRNAs by nature of having sequence complementarity (full or partial) to a particular site on the mRNA, usually in its 3' untranslated region (3' UTR). The distinguishing features of microRNAs are that they tend to be processed either from larger transcripts called primary miRNAs (pri-miRNAs) or from introns that have been spliced out of pre-mRNAs (so called mirtrons). In animals, the processing of pri-miRNAs into pre-miRNAs is accomplished in the nucleus by the *microprocessor complex*, a protein complex that includes the Drosha and Pasha/DGCR8 proteins; this complex recognizes hairpins on pri-miRNAs and cleaves them out, creating pre-miRNAs. Mirtrons bypass this processing as they originate from introns and not pri-miRNA transcripts. The resulting pre-miRNA, which generally consists of a *stem* and a *loop* structure, is then exported to the cytoplasm by the protein Exportin-5. In the cytoplasm, an endonuclease called Dicer further processes the stem-loop pre-miRNA into a mature

miRNA:miRNA* duplex by cleaving out the loop and a part of the stem. Each strand of the duplex forms a distinct single-stranded mature miRNA with full or near-complementarity between the two. The convention for which one is dubbed the “star” miRNA is usually set by their order of discovery, the method by which the miRNA was discovered, and the relative expression levels of each miRNA strand in the tissue in which it was discovered. The resulting mature miRNAs tend to be about 22nt long in animals. It is these mature miRNAs, in conjunction with the RNA-induced silencing complex (RISC), which form a regulatory ribonucleoprotein complex that carries out silencing activity on mRNAs. The class of protein that is central to RISC’s silencing activity is the *Argonaute* family. They are responsible for guiding the miRNA to its target mRNA. The miRNA-RISC (miRISC) then silences the mRNA transcript either by inhibiting translation into protein by the ribosome or degradation of the mRNA via cleavage.

Another type of small RNA found in animals is the *Piwi*-interacting RNA (piRNA), named after the *Piwi* class of proteins, a subclass of the *Argonaute* family [64,81,132,137]. Unlike miRNAs, piRNAs tend to be significantly longer (26-32 nt versus 22nt) and also tend to have a uridine on their 5’ end. The process by which they are generated is not yet fully clear. However, their functional role has been partially elucidated: they are involved in the silencing of “selfish” genetic elements known as transposons as well as in the placement of epigenetic marks on chromatin. They are also highly active in mammalian testes and are required for mammalian spermatogenesis.

There are a variety of other types of small non-coding RNAs, and in the literature they are generally labeled by their precursor RNA. Small RNAs can be produced from any type of precursor, ranging from protein-coding mRNA to non-coding RNA types such as rRNA, tRNA, snRNA, and snoRNA. There is evidence that some of these small RNAs are processed like and behave like microRNAs: they are produced by cleavage of stem-loop structures by the Dicer protein and go on to have regulatory effects on mRNAs [9,18,103,126]. The fact that they originate from precursor transcripts other than pri-miRNAs does not preclude them from behaving

like miRNAs. However, the vast majority of non-miRNA small RNAs that are commonly found in small RNA-seq datasets, for example, are entirely uncharacterized other than their annotated precursor transcript. For example, tRNA-derived smRNAs (known as tRFs, or tRNA fragments) [99], are thought to be the result of a combination of endolytic cleavage under stress response conditions and non-specific cleavage by Dicer – but whether they are a simply non-specific byproduct of smRNA processing pathways or go on to have functional regulatory roles has yet to be determined. In Chapter 3 I present a quantitative method that can help researchers characterize these largely unstudied populations of small RNAs.

1.2. Measuring the transcriptome

The advent of high-throughput sequencing (HTS), the most common subtype of which is shotgun sequencing, has heralded in a new age of computational biology. In current-generation shotgun sequencing, DNA (or RNA) is fragmented into smaller pieces and then a machine produces "reads" by reading the sequence of these fragments from either one end (single-end sequencing) or both ends (paired-end sequencing). While the sequencing of genomes (DNA-seq) has gained recent attention, researchers are starting to see the value in applying these technologies to the sequencing of RNA (RNA-seq) [124] (**Figure 1.4**). In DNA-seq, researchers seek genetic variants that are uncommon, as well as types of variants that are difficult to detect with genotyping methods. This same level of sensitivity can be applied to RNA, where the goal is to not only determine the sequences of RNA transcripts, but also to infer changes in their abundance and alternative splicing between experimental conditions or in disease states. In this dissertation I present alternative facets of the transcriptome that can be measured using this data, but have not yet been fully explored (Chapters 2 and 3). While the clinical applications of RNA-seq have yet to be fully realized, it can already be used for biomarker discovery and in the generation of target hypotheses for, e.g., drug discovery. Traditional RNA-sequencing focuses solely on polyadenylated messenger RNAs, perturbations of which are more amenable to interpretation

when the function of the coded protein is known. However, alternative forms of RNA-sequencing, such as those that I present in this dissertation, are just as important in assaying the impact of the full (coding and non-coding) transcriptome (Chapter 4). Examples of alternate forms of RNA-sequencing are: Ribosomal RNA-depleted RNA-seq (rRNA(-) RNA-seq) [30], small RNA-seq (smRNA-seq) [95], cross-linking immunoprecipitation-high-throughput sequencing (CLIP-seq) [105], Bisulfite RNA-seq [143], methylated RNA immuniprecipitation RNA-seq (MeRIP-seq) [115], double-stranded RNA-seq (dsRNA-seq) [172], single-stranded RNA-seq (ssRNA-seq), and degradome-seq (PARE, GMUCT) [2,63,66,159].

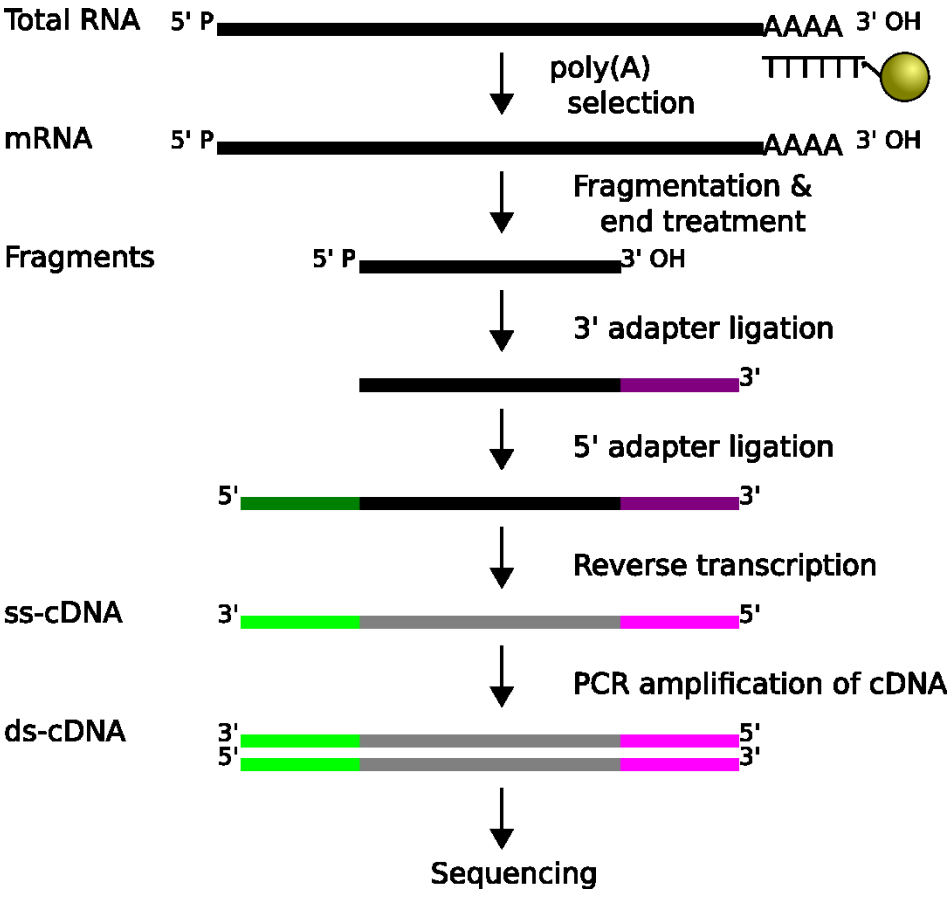

Figure 1.4 – Strand-specific polyA(+) RNA-sequencing.

In rRNA(-)-seq, the goal is similar to that in regular polyA(+) seq – measure abundance of and detect alternative splicing of transcripts. However, instead of limiting the experiment to only those RNAs with poly(A) tails, depleting ribosomal RNA allows one to assay a wider range of transcripts. One downside, however, is that the presence of highly abundant non-coding, polyA(-) RNAs can reduce the dynamic range of the estimated read counts. Although with recent increases in sequencing depth capabilities, this disadvantage has grown considerably less important. In Chapter 4 I describe an application of rRNA(-)-seq to a study of the differences in non-coding RNAs in the Alzheimer's disease brain.

Small RNA-seq is similar to rRNA(-) seq in that its intended purpose is to infer the abundance of non-coding RNAs. However, the method focuses on a subgroup of non-coding RNAs that are shorter than a particular length; the desired range for sequenced RNAs is usually 15-45nt. In small RNA-seq, usually the rRNA depletion is forgone and instead an additional size-fractionation step is added: the shorter fraction of RNAs is selected by polyacrylamide gel electrophoresis (PAGE) and subsequent gel extraction. Again, this type of RNA-sequencing is applied to Alzheimer's disease in Chapter 4.

While the previously described methods are used to assay the abundance and splicing changes in RNAs, there are other aspects of the transcriptome that can be measured. For example, in CLIP-seq, the goal is to elucidate the binding-specificity of an RNA-binding protein. In short, the RNA and all its bound proteins are cross-linked, and an antibody specific to one protein is used to pull-down a fraction of RNA that is enriched for the protein of interest. Then after fragmentation and removal of the proteins, RNA-sequencing is performed on the enriched fraction. After mapping these reads back to the genome, one can infer all of the sites in the transcriptome where the protein has some binding affinity. This can be used to determine general rules for the specificity of this particular protein by performing *de novo* sequence- or structural-motif searches within the enriched sequences. The procedure is analogous to chromatin-immunoprecipitation sequencing (ChIP-seq), where the goal is to find binding sites of chromatin-

binding proteins. Among other studies, CLIP-seq has been applied to the study of an RNA-binding protein called TDP-43 which is implicated in the pathogenesis of neurodegenerative disorders such as amyelotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD). CLIP-seq is also useful for finding *in vivo* sites of microRNA-mRNA binding by using an antibody specific to proteins in the miRNA silencing machinery; such studies are extremely important for elucidating regulatory targets of short RNAs such as microRNAs.

Another example of a DNA-sequencing protocol that has been adapted to RNA-seq is that of bisulfite sequencing. The goal in bisulfite RNA-seq is to detect sites in the transcriptome where a cytidine has been replaced by a 5-methylcytidine (m^5C) – that is, one is searching for a particular RNA modification in all RNAs. The protocol consists of treating the RNA with bisulfite before sequencing. Treatment with bisulfite deaminates cytosine to uracil, but 5-methylcytidine resists this conversion. After sequencing, one can detect conversion at cytidines and infer that the cytidines that are not converted into uridine must be methylated at the N5 position. Computationally, this presents issues as the conversions induce mismatches between the RNA sequences and the genomic sequence, which complicates the process of mapping these sequences back to the genome. However, specific alignment methods have been developed to mitigate this particular issue.

An alternative to bisulfite sequencing is another method called MeRIP-seq. Here, instead of using bisulfite to produce a signal at unmodified cytidines, one instead uses an m^5C-specific antibody to immunoprecipitate m^5C-enriched RNA. By sequencing this fraction one can infer that enriched sequences relative to a non-specific immunoprecipitation are likely to have m^5C sites. In addition, this method can be applied to any modification rather than just m^5C. An advantage over bisulfite RNA-seq is that it does not induce mismatches in the RNA sequences; a disadvantage is that it may lack nucleotide-by-nucleotide resolution of the specific sites that are methylated.

Another facet of the transcriptome that is a very active research area is RNA structure prediction. Historically, structural prediction of biomolecules such as RNA has been a laborious

and expensive low-throughput process. Additionally, *in silico* predictions based on annotated sequences alone have limited accuracy. Now high-throughput RNA-sequencing, in conjunction with biochemical methods, has allowed researchers to predict RNA structures transcriptome-wide. There are several methods for accomplishing this, but they largely rely on similar biochemical treatments: the differences lie in the algorithms used to infer structure from sequencing data. Briefly, RNA is digested by a structure-specific RNAse enzyme and the remaining undigested RNA is sequenced. When the desired fraction is that of double-stranded RNA (dsRNA), the RNA is treated with an ssRNAse (single-stranded RNAse). Similarly, when single-stranded RNA (ssRNA) is desired, the RNA is treated with a dsRNAse. By sequencing each of these types of libraries in parallel and using computational methods to infer base-pairing probabilities, one can begin to infer RNA secondary structure transcriptome-wide.

Another variant on RNA-seq that is a high-throughput extension of existing low-throughput methods is degradeome sequencing. Degradeome sequencing is a high-throughput version of 5' RACE (rapid amplification of cDNA ends). The degradome is the fraction of RNA resulting from regulatory cleavage of transcripts – these transcripts are silenced by particular types of cleavage. These cleavage events leave particular biochemical marks on the 5' ends of the resulting fragments – in particular, the lack of the 5' cap of the original transcript. By selecting for these types of fragments with biochemical methods, one can sequence such fragments and infer sites where cleavages like this have occurred. This method is largely used in plant transcriptomes, where short regulatory RNAs such as microRNAs carry out their silencing activity largely by catalyzing endolytic cleavage of the target transcript. In doing so, one can find *in vivo* target sites of these regulatory RNAs.

1.3. Alzheimer's Disease

Alzheimer's disease (AD), the most common form of dementia, was discovered by the German neuropathologist Alois Alzheimer in 1906 [20]. As of 2013, it is the most expensive disease in the

US; its immense societal burden is estimated at $157-$215 billion per year [173]. The FDA currently approves of four drugs for its treatment, all of which are cholinesterase inhibitors, and none of which are particularly effective at treating the disease. While its prevalence increases drastically with age (the risk doubles every 5 years after age 65), we still do not know what fundamentally causes it. Also, while it is estimated to be around 70% heritable, the genetics of AD have yet to be fully elucidated [8].

Alzheimer's disease is clinically characterized by progressive memory loss, cognitive impairment, and behavioral changes. The hallmarks of its neuropathology are structures known as *senile plaques* and *neurofibrillary tangles*. Senile plaques are extracellular protein aggregates consisting mainly of the pepide amyloid beta (Aβ), whose precursor protein is encoded by the gene APP (amyloid precursor protein), and whose function is yet unclear. The neurofibrillary tangles are composed of the hyperphosphorylated protein tau (gene: MAPT), which normally associates with microtubules, structures that maintain the internal structure and morphology of cells.

Broadly, AD cases can be broadly classified into two categories based on their genetic underpinning (familial or sporadic) and also by the age of onset (early or late). The familial form of the disease is almost always caused by autosomal dominant mutations in a small number of genes related to production of the Aβ peptide: presenilins 1 and 2, which help process APP, and APP itself. The onset of the disease when it is familial (before 65) tends to be much earlier than when it is LOAD. Familial cases, however, are extremely rare: they only account for 0.5% to 2.5% of all AD cases. The vast remainder of AD cases are of unknown genetic etiology, although several risk factors have been identified by recent genome-wide association studies (GWAS) [14,79,119]. What these studies have shown is that the largest genetic risk factor for sporadic AD by far is apolipoprotein E (ApoE) on chromosome 19, and alleles in a small number of other genes confer additional risk (**Table 1.3**). It is not yet fully understood what roles are played by these genes in the pathogenesis of AD, and functional studies of them are a very active area of

research. It is hoped that these studies will lead to earlier and more accurate diagnosis of AD and ultimately to treatments for the disease. In Chapter 4 I integrate these known genetic factors with RNA-sequencing data in order to increase the impact of correlative functional data by connecting them to causative genetics data.

Table 1.3 – Genes implicated in LOAD by genome-wide association in Caucasian populations.

Gene symbol	Chromosome	Gene name
ApoE	19	Apolipoprotein E
TREM2	6	Triggering receptor expressed on myeloid cells 2
TOMM40	19	Translocase of outer mitochondrial membrane 40 homolog (yeast)
BIN1	2	Briding integrator 1
CLU	8	Clusterin
ABCA7	19	ATP-binding cassette sub-family A member 7
CR1	1	Erythrocyte complement receptor 1
PICALM	11	Phosphatidylinositol binding clathrin assembly protein
MS4A6A	11	Membrane-spanning 4-domains, subfamily A
CD33	19	Myeloid cell surface antigen CD33
CD2AP	6	CD2-associated protein
EPHA1	7	Ephrin type-A receptor 1

1.4. Outline of dissertation

In Chapter 2 I present a computational method that, in conjunction with one of many types of RNA-sequencing methods, can be used to detect modified ribunocleotides transcriptome-wide. The method can be considered a high-throughput generalization of already-existing low-throughput methods that capitalizes on the availability of modern RNA-sequencing technology. In addition to detecting modified nucleotides, it can also differentiate between different types of RNA modifications.

In Chapter 3 I describe a method for characterizing and classifying many different kinds of non-coding RNAs using small RNA-sequencing data. The key innovation of this method is that

it digests RNA-sequencing into biologically relevant features, rather than *black box*-style features that can hinder the interpretability of the results, particularly by domain experts. Using these more interpretable features, which were selected based on their known relevance to RNA processing pathways, the software can predict with a high degree of accuracy the class of small non-coding RNA. In addition, this method has been validated by comparing across independent datasets where different tissue types were used for sequencing.

In Chapter 4 I present an integrative analysis of the rRNA-depleted and small RNA transcriptomes of the Alzheimer's disease prefrontal cortex. I describe the genes that are differentially expressed and classify them by their coding potential, their known precursor RNAs, and their predicted and experimentally verified regulatory targets. By integrating rRNA(-) and small RNA transcriptome data with loci known to be genetically associated with AD, we can begin to build a network that connects AD risk-associated variants with functional genomics data from the human brain.

2. High-throughput Annotation of Modified Ribonucleotides (HAMR)

Appeared in: Ryvkin P*, Leung YY*, Silverman IM*, Childress M, Valladares O, Dragomir I, Gregory BD, Wang L-S. HAMR: high-throughput annotation of modified ribonucleotides. RNA. 2013. (*Joint first authors)

2.1. Introduction

Covalent post-transcriptional modifications of specific nucleotide bases in RNA molecules are known to be highly prevalent and physiologically important. However, their overall abundance and biological function are not well understood. This gap is even more surprising given that RNA modifications play a role in maintaining structure, catalytic activity, and cellular abundance of RNAs, and that all known classes of RNA molecules harbor various levels of diverse modifications. Additionally, the recent discovery that an RNA methyl-6 adenosine demethylase (FTO) is a risk gene in obesity highlights the significance of RNA modifications to human biology [57,62,84].

Methods for detecting such modifications are well established [23,34,68,70,76,77,115,130,170]. One such method is primer extension, which relies on the differential ability of reverse transcriptase to produce cDNAs with base-pair substitutions at positions occupied by modified nucleotides [161]. Interestingly, all high-throughput RNA sequencing library preparation protocols require RNA to cDNA conversion by reverse transcription (RT), thus we reasoned it is possible to identify sites of modified nucleotides in all RNAs transcriptome-wide by uncovering nucleotides with significant sequence error rates. Using this idea, we developed HAMR, and demonstrate that this software allows fast and reliable

identification of modified nucleotides at single-nucleotide resolution in all RNA classes transcriptome-wide through the analysis of nucleotide substitutions found in various RNA-seq datasets. This software will provide an important tool for future work on RNA modifications, which are emerging as important regulators of human biology and physiology [43,84].

2.2. Methods

2.2.1. RNA extraction and sequencing

Frozen human brain tissue from four female patients without neurological pathology was obtained from the Center for Neurodegenerative Disease Research. Trizol extraction was performed to obtain total RNA. cDNA libraries for sequencing were generated following the Illumina small RNA library preparation procedure. The libraries were sequenced on an Illumina GAIIx machine to 50bp and were submitted to NCBI GEO database (GSE43335). The reads were 3' adapter-trimmed, requiring at least 6 bp of adapter sequence with at most a 6% mismatch rate. All untrimmed reads and trimmed reads shorter than 14bp were discarded. The remaining reads were mapped to the human genome (hg19) [59] using Bowtie [97] under "-v 2" mode with a maximum 6% mismatch rate and allowing up to 100 mappings per read. Any unmapped reads were re-aligned to the set of tRNA transcripts with -CCA tails appended, and these were merged into the final alignment. For the whole transcriptome libraries, the same extractions were performed on brain samples from the same four patients, plus an additional male patient (GSE46523). Instead of initial size-fractionation, RNAs were depleted by one round of Ribominus (Invitrogen). Additionally, sequences mapping to known rRNA sequences were masked out of the dataset, and both adapter-trimmed and untrimmed reads were used.

The alignments were also performed using a different alignment program, BWA [102]. The results obtained using BWA were nearly identical to those given by Bowtie's alignments (195

modified sites versus Bowtie's 202). Reads aligning to repeat regions or annotated RNAs other than tRNAs were discarded. Nuclear tRNA annotations were taken from the "tRNAs" table in the UCSC genome browser (hg19). Annotations for mitochondrial tRNAs were generated by running tRNAscan-SE (v1.23) set to organelle mode on the mitochondrial genome ("chrM" in hg19). Multi-mapping reads were partially resolved by taking those alignments whose mismatches aligned to SNPs (dbSNP 135) as the true hits, prioritizing them over alignments whose mismatches had no apparent explanation. The yeast data, consisting of 20.8 million reads sequenced on an Illumina Genome Analyzer I, was obtained from from the NCBI Sequencing Read Archive (GSM775340).

2.2.2. tRNA locus clustering

tRNA loci were taken from the tRNAscan annotation at UCSC and were required to have a tRNAscan score of 60.0. The loci were merged into families based on an empirical measure of sequence similarity computed from the number of reads mapping across them simultaneously, resulting in a clustering of tRNA loci that minimizes the number of cross-mapping reads. Each ordered pair of loci *(i,j)* is assigned a similarity value

$$s(i,j) = \frac{N_{ij}}{\max_k N_{ik}}$$

where N_{ij} is the number of reads mapping to both loci and the denominator is taken over all loci *k*. Then the symmetric similarity is

$$S(i,j) = S(j,i) = \max\{s(i,j), s(j,i)\}$$

and the distance is set to be

$$D(i,j) = 1 - S(i,j).$$

Hierarchical clustering with *k*=84 clusters yielded the fewest cross-mapping reads with the fewest rogue clusters (those whose tRNAs decode to more than one amino acid). The two rogue clusters were Gly(SMC)1 containing 1 $tRNA^{Val}_{CAC}$ and Cys(NVM)1 containing 6 $tRNA^{Ala}_{AGC}$, 1 $tRNA^{Ala}_{CGC}$, 3 $tRNA^{Ala}_{UGC}$, 1 $tRNA^{Ser}_{AGA}$, and 1 $tRNA^{Val}_{AAC}$.

2.2.3. Detecting candidate RT misincorporation sites

The read alignment was converted to a pileup format and bases with quality score below 30 were discarded. Candidate RT misincorporation sites were taken to be those covered by at least 10 reads and significantly enriched (FDR<5%) for mismatches by the binomial test, assuming a base call error rate of 1%. We tested two null hypotheses. The first, H_0^1, consists of the hypothesis that the genotype is homozygous reference. Therefore, the probability of seeing fewer than k out of n_{tot} reads matching the reference nucleotide at a given site is

$$\Pr(k_{ref} < k \mid n_{tot} \text{ reads, site genotype is homozygous reference nucleotide}) = \sum_{i=1}^{k} Binom(i; n_{tot}, p_e)$$

where p_e is the base calling error rate. A more conservative null hypothesis, H_0^2 assumes only that the genotype is biallelic. It is a composite hypothesis consisting of sub hypotheses for each of the 10 possible genotypes. HAMR tests each possible biallelic genotype and takes the maximal p-value among all the tested genotypes. The advantage of using H_0^2 is that it will not falsely call significant any site that looks like a heterozygous or homozygous SNP. The main disadvantage is that it will cause HAMR to miss simple RNA edits as well as modifications that produce one- or two-nucleotide patterns in the cDNA. H_0^2 is more appropriate when one wishes to avoid false positives due to polymorphisms, but H_0^1 can be used if corroborating DNA evidence or other means are available to rule out such false hits. During the scan of the entire small RNA

transcriptome, the single nucleotides corresponding to the 5' and 3' ends of reads were discarded to reduce false positives resulting from elevated base calling error and ligation errors on read-ends.

2.2.4. tRNA modification identification

RNA modification data was taken from the RNA modification database [129]. Specific locations of tRNA modifications were taken from the eukaryotic entries in tRNAdb 2009 and from the curated S. cerevisiae data at MODOMICS [38]. The tRNAdb data were given precedence over MODOMICS in all cases. Within the tRNAdb data, if multiple modifications were annotated for the same site, precedence was given to the organism closest in evolutionary distance from the target organism (either human or *S. cerevisiae*), using divergence time estimates as the means reported at timetree.org [75]. For each candidate modification site, an evidence level was assigned based on its overlap with the known modification data. The highest confidence overlap is one where a candidate modification occurs at a particular site in a particular tRNA for both the prediction and in the annotation. The next lowest confidence overlap is one where a known modification occurs at that site in any isoacceptor tRNA. Finally, the lowest level of evidence is the presence of a known modification in any eukaryotic tRNA at that site. Higher evidence data always takes priority over lower evidence data. If multiple possible modifications of the same evidence level are annotated at the same site, the modification data is marked as ambiguous. Modified sites were plotted on the RFAM consensus tRNA structures using SAVoR [100]. The classifier for identifying specific modifications by mismatch pattern is a 3-nearest-neighbor classifier in three dimensions, with the features being the sequenced proportions of the three non-reference nucleotides, after Laplace smoothing. For training data we only used the highest level of evidence (same site, same tRNA) and only modifications supported by at least 3 instances in the RNA-seq data were used.

2.2.5. Software

The HAMR program takes as input a sequence-read alignment in BAM format (consisting of uniquely-mapped reads) and produces a table of genome coordinates and nucleotide frequencies at those coordinates. Given an assumed sequencing error-rate, it then performs a statistical analysis to select those sites whose mismatch rates are higher than expected by chance. The result is a set of sites that consist of both potential SNPs and candidate RNA modifications. These sites may optionally be classified as particular modifications based on the models built from no-chemical-treatment tRNA data.

The web interface allows specification of a remote, indexed BAM file and BED file with targeted intervals for querying. The user may specify parameters for the preprocessing steps, such as minimum base call quality score, minimum coverage at a site, assumed sequencing error rate, and significance level. Additionally, the user may use the software to predict the modification type based on mismatch patterns in tRNA data.

2.3. Results

Our method, HAMR, is able to detect the presence of multiple types of modifications present in RNA sequenced only once, without chemical treatment. In addition, the signals produced by these modifications via modulation of RT activity are present in all types of RNA sequencing datasets, which means that HAMR could be invaluable in gleaning more data from previous studies or publicly available data. We demonstrate that the method is able to detect modifications in two newly generated human RNA datasets as well as a publicly available yeast dataset and there is significant overlap in the signal detected.

2.3.1. Small RNA-sequencing of tRNA families

tRNAs are the most highly modified cellular RNAs. Since they are highly represented in small RNA sequencing libraries as tRNA fragments [22] we developed our approach on this type of data, although in principle our method can be applied to any type of RNA-seq dataset. We analyzed small RNA-seq data obtained using the dorsolateral prefrontal cortex of four deceased human patients who showed no signs of neuropathology. We found that the majority of reads (57%) mapped to known microRNAs, 23% to tRNAs, and the rest to other types of known RNAs and intergenic regions.

Since tRNA loci exist in multiple copies across the human genome, their associated short RNA-seq reads will often map to multiple loci. Simply eliminating the ambiguously mapped reads would greatly reduce our data. We reasoned that the exact identity of the tRNA locus was not as important as the family producing each read with regards to RNA modification specificity. Given that isoacceptor tRNAs (those accepting the same amino acid) tend to have similar sequences and isodecoders (those with the same anticodon) even more, we were able to combine similar tRNA loci into families and refer to them by their predicted amino acid and anticodon. The 386 high-scoring tRNA loci annotated by tRNAscan-SE [106] fell into 84 tRNA families that were distinct enough to greatly reduce read mapping ambiguity. The post-clustering cross-mapping rate (proportion of reads that map to one or more tRNA families) ranged from 9% for shorter reads (18 – 20 nucleotides (nt)) down to 2% for longer reads (>31nt). Furthermore, only two families included so-called rogue tRNAs, or tRNAs that share sequence identity with their siblings but code for a different amino acid.

2.3.2. Detecting modified sites by mismatch rates

In order to detect true post-transcriptional RNA sequence differences, we needed to exclude other sources of mismatches such as base calling error and DNA polymorphisms. It is noteworthy

that we observed an elevated mismatch rate for tRNA-derived smRNA reads, as would be expected when a large number of modified bases are present. In fact, when comparing the mismatch rates of reads mapping to tRNAs, microRNAs, and other types of RNAs, we found that tRNAs showed an overall elevated level of mismatches, microRNAs showed a spike corresponding to the ends of mature miRNAs, and other RNAs showed a gradual increase in mismatches towards the 3' ends of reads (**Figure 2.1**). These data were consistent with high numbers of modified bases spread across tRNA reads, with edits/additions at the ends of mature microRNAs [21,162], and with simple base calling error, which is expected to increase at the 3' ends of longer reads, respectively. The elevated-mismatch sites throughout the length of tRNA-derived small RNA reads, not just their 3' ends, suggested that data from smRNA-seq allowed us to identify true base pair modifications and not merely sequencing errors. Additionally, the distribution of PHRED quality scores at mismatch-containing sites 38.33 (std dev. 2.28) was nearly identical to that at non-mismatching sites 38.37 (std dev. 2.28).

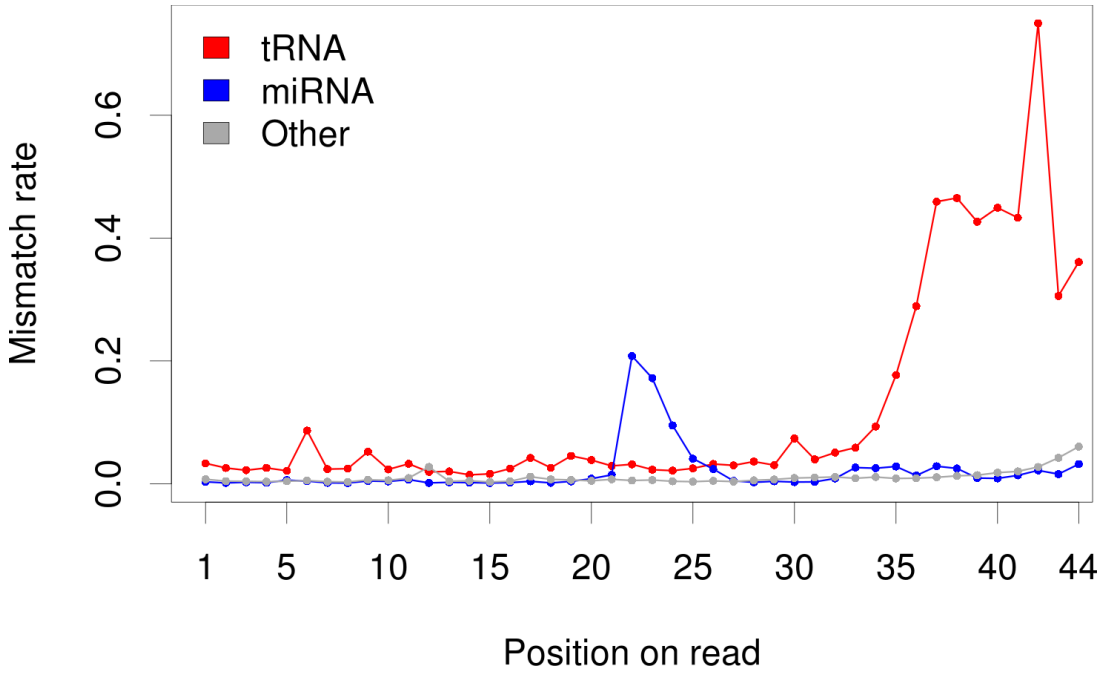

Figure 2.1 – Mismatch rates in small RNA reads mapping to three types of RNA

Taking advantage of this observation to identify base modifications transcriptome-wide, we developed a model for allowing statistically significant identification of RNA modification sites based on nucleotide misincorporation by RT, while ignoring sequencing errors and single nucleotide polymorphisms (SNPs) due to genotype. The model assumes a fixed base calling error rate, and makes a set of assumptions about the underlying genotype to model the mismatch rate due to chromosomal polymorphism. The simplest null hypothesis, H_0^1, assumes that the site is homozygous with the reference allele. Taking this as the null hypothesis results in any non-reference nucleotide above the base calling error rate being called as a candidate modification. A more conservative null hypothesis, H_0^2, assumes only that the genotype is biallelic. Under this assumption, we call candidate modifications where three or more nucleotides are sequenced at a rate higher than base call errors. Such patterns will arise at sites of RT misincorporation due to modifications and not at biallelic polymorphic sites.

We estimated library-wide base calling error to be around 1% based on the observed library-wide mismatch rate and on previous reports of error rates in Illumina sequencing [108]. We also required coverage of at least 10 reads per nucleotide, including reads with the same start and end positions. Under H_0^2, HAMR called 228 candidate modifications out of 5,487 sequenced tRNA sites. Of these, 201 (88%) did not overlap with any known SNP in dbSNP release 135 [134]. Among these 201 sites, 123 (61%) coincided perfectly with a known modification as listed in tRNAdb 2009 [142] or MODOMICS [38] and 187 (93%) coincided with sites known to be modified on any tRNA (**Figure 2.**2 and **Figure 2.3**).

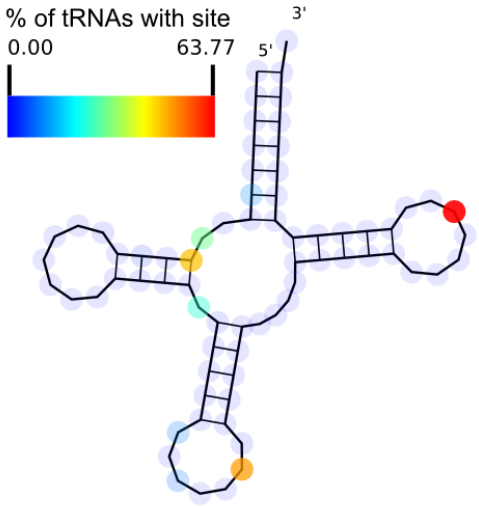

Figure 2.2 - Locations of known tRNA modifications predicted to affect RT incorporation

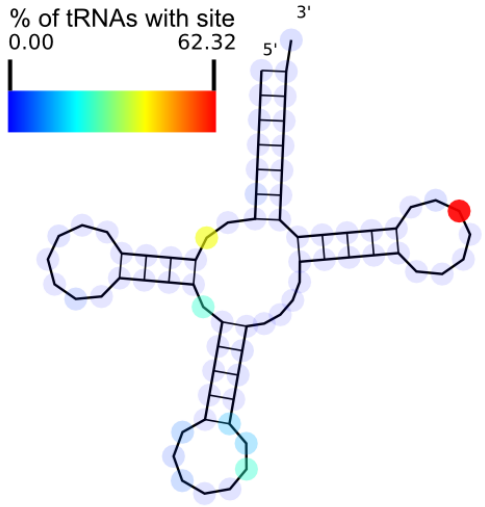

Figure 2.3 - Modification sites predicted by HAMR

In order to test for possible violations of the biallelicity assumption under H_0^2, we ascertained the overlap between our called sites and known CNVs. Of the 233 genomic sites where we called a modification under H_0^2, 36 (15%) of the candidate sites fall within gain-of-copy CNVs listed in the Toronto CNV database [171]. Of the 36 sites in CNVs, 20 fall within rare CNVs (only 1 observation) and 16 fall within recurrent CNVs (observed more than once). This suggests

that, if the results are false positives due to undiscovered SNPs compounded by copy number variation, such instances are only a small fraction of the sites called by HAMR.

Since no chemical treatment that allows the identification of a specific post-transcriptional modification is used, our approach is limited to detecting modifications that modulate RT incorporation during normal sequencing library preparation. We predicted the RT effect of the remaining modifications based on their presence along the Watson-Crick edge (on the Watson-Crick bonds) of the nucleoside (**Table 2.1**). We found that HAMR exhibits higher sensitivity where these types of modifications are predicted to occur (Fig. 3). While inosine (I) is known to produce an A>G substitution in cDNA [11] this nucleotide pattern is indistinguishable from an A/G SNP and so is discarded under the conservative null hypothesis H_0^2. When we used the less conservative null hypothesis, H_0^1, 60% of known inosine edit sites were called (**Figure 2.5**).

Table 2.1 – Selected RNA modifications and their known and predicted effects on RT

Modification	**Symbol**	**RT effect**	**W-C edge**
Inosine	I	Mistranscription [67]	Y
N^1-methylinosine	m^1I		Y
N^1-methyladenosine	m^1A	Can't pair [67]	Y
N^2-methyladenosine	m^2A		N
N^6-threonylcarbamoyladenosine	t^6A	Stop [67]	Y
N^6-isopentenyladenosine	i^6A		Y
5-methoxycarbonylmethyl-2-thiouridine	mcm^5s^2U		N
2-methylthio-N^6-isopentenyladenosine	ms^2i^6A		Y
N^6-methyl-N^6-threonylcarbamoyladenosine	m^6t^6A		Y
2-methyladenosine	m^2A		N
N^1-methylguanosine	m^1G		Y
N^2-methylguanosine	m^2G	Pause [67]	Y
N^2,N^2-dimethylguanosine	$m^2{}_2G$		Y
7-methylguanosine	m^7G		N
Wybutosine, peroxywybutosine	yW,o_2yW	Stop [67]	Y
Queuosine, mannosyl-queuosine	Q, manQ		N
3-methylcytidine	m^3C		Y
5-methylcytidine	m^5C	Pairs [67]	N
N^4-acetylcytidine	ac^4C		Y
5-methyluridine (ribothymidine)	m^5U / T	Pairs [60]	N
5-carbamoylmethyluridine	ncm^5U		N
Dihydrouridine	D	Can't pair [67], Pairs [60]	N
Pseudouridine	Ψ / Y	Pairs [67], Pairs [60]	N
2'-O-methyl nucleosides	Am, Cm, Gm, Um	Pause w/ low dNTP [67]	N

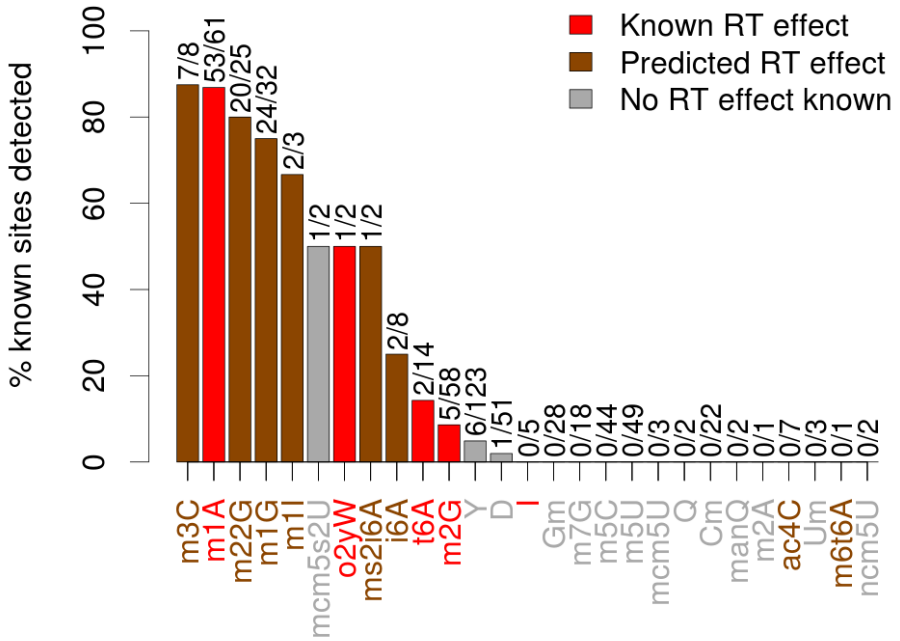

Figure 2.4 – HAMR's sensitivity for detecting different types of RNA modification

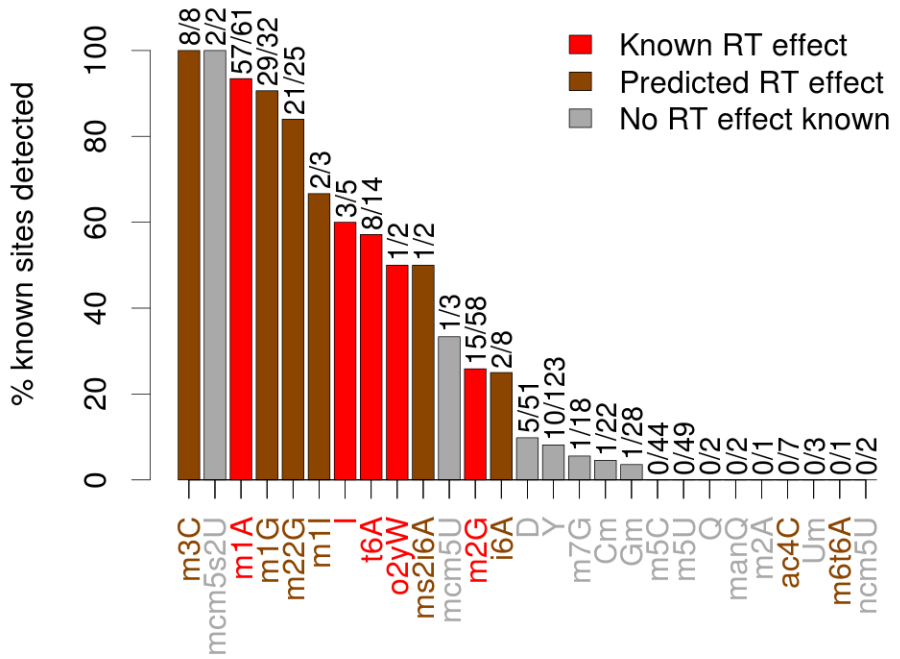

Figure 2.5 – HAMR's sensitivity under the loose model H_0^1

2.3.3. Calling modification types by incorporation patterns in RT

We hypothesized that different types of modifications affecting RT incorporation would have distinct incorporation patterns due to the differential base-pairing properties of the modified ribonucleotides. In order to visualize the incorporation patterns we mapped each potentially modified site (excluding known SNPs and using the conservative null hypothesis H_0^2) onto a ternary plot with the three dimensions corresponding to observed fractions of the three non-reference nucleotides. This can be done for each precursor nucleotide separately (A, C, G, and U). The ternary plots clearly show clustering by modification type for modified adenosines and guanosines (**Figure 2.6**a,b). Using this approach, we observed thirteen sites for cytidine (m^3C) (**Figure 2.7**), while predicting two RT-effecting sites for uridine (**Figure 2.8**). Interestingly, despite U>D (dihydrouridine) and U>Y (pseudouridine) not being predicted to affect RT incorporation, we were able to detect these sites and they tended to cluster together. We also found that the m^3C sites were sequenced with a very similar nucleotide pattern in all four human brain samples and so those observations cluster together.

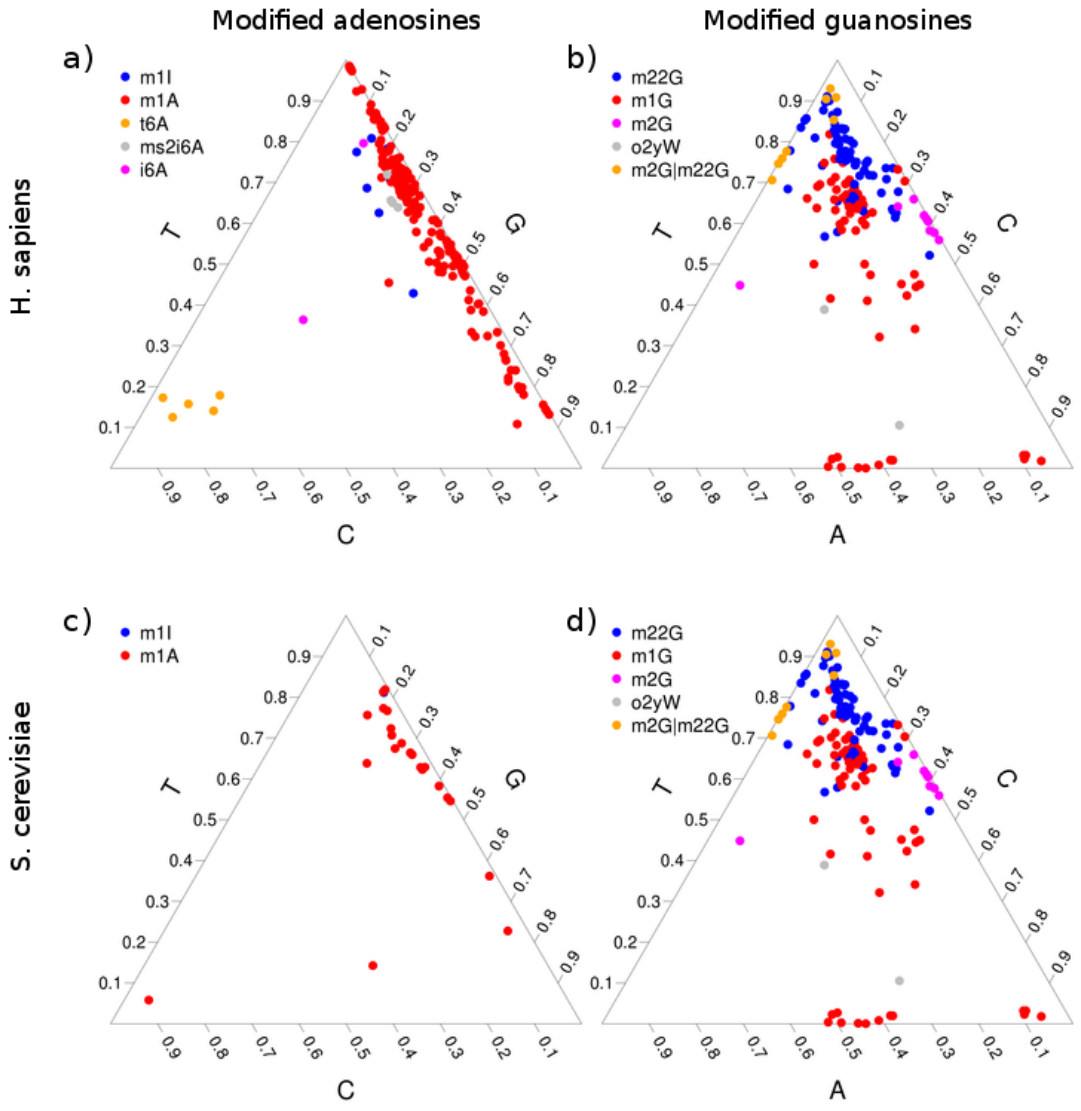

Figure 2.6 – Observed nucleotide frequencies in cDNA for different modification types and in different organisms

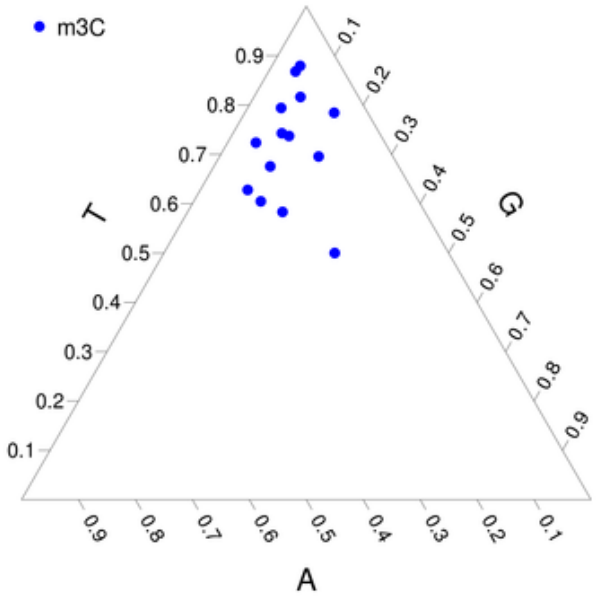

Figure 2.7 – Sequenced nucleotide frequencies at known tRNA m^3C sites in the human brain

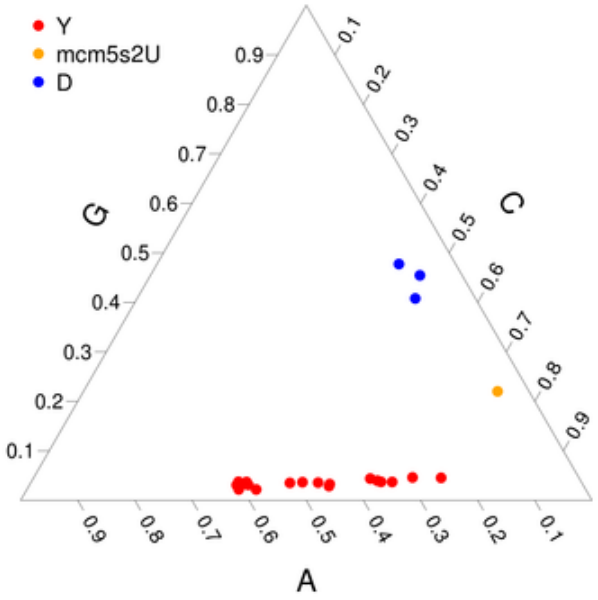

Figure 2.8 – Sequenced nucleotide frequencies at known modified tRNA uridines in the human brain

Amongst modified adenosines, m^1A shows a bias towards sequencing of T with varying amounts of G, and m^1I shows a very similar pattern. In contrast, t^6A shows a strong bias towards sequencing of C in the cDNA. Under the less conservative H_0^1, 60% of the known inosine sites were detected and found to be very strongly associated with a G in the cDNA, as is expected (**Figure 2.9**). At guanosines, both $m^2{}_2G$ and m^1G heavily favor sequencing of T with varying amounts of C and A, while peroxywybutosine (o_2yW) shows more variation. Observations for peroxywybutosine were insufficient for us to draw strong conclusions about its RT incorporation patterns.

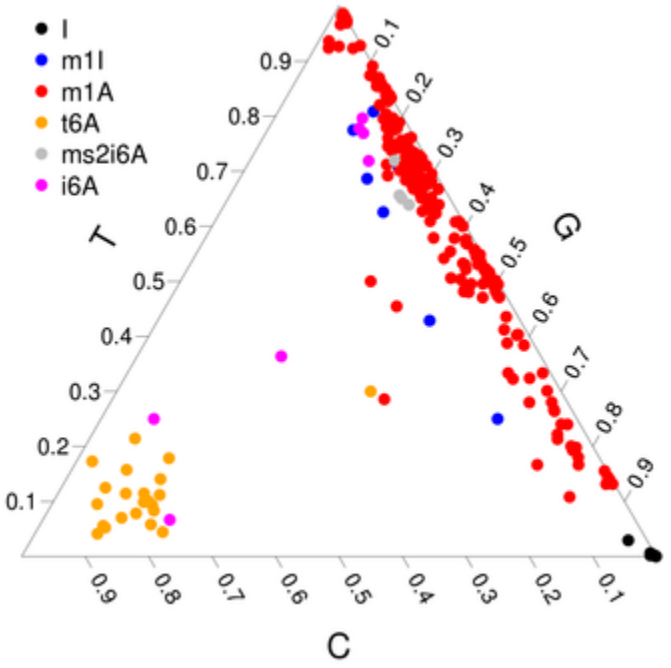

Figure 2.9 – Sequenced nucleotide frequencies at guanosines when using the loose model H_0^1

We set out to design a classifier that could take these patterns as input and predict the most likely modification at a site using these ternary plots. Given that m^1A, m^1I, and ms^2i^6A and i^6A and t^6A co-cluster, we decided to merge these two sets of modifications into the combined classes $m^1A|m^1I|ms^2i^6A$ and $i^6A|t^6A$. Similarly, we merged m^2G and $m^2{}_2G$ into a single class,

$m^2G|m^2{}_2G$. These two may be especially difficult to resolve because m^2G is a chemical precursor of $m^2{}_2G$. Using a 3-nearest-neighbor classifier and leave-one-out cross-validation (LOOCV) we were able to differentiate between the two groups of adenosine modifications with 98% accuracy. For the guanosine modification types m1G and $m^2G|m^2{}_2G$ we were able to achieve 78% accuracy. For the 18 observations of significant uridine sites, we were able to distinguish between D and Y modifications with 86% accuracy. As there was only one type of cytidine modification that was detected, m^3C, a classifier was not necessary. It is informative, however, that without chemical treatment the only cytidine modification we detected was m^3C.

2.3.4. Expanding the tRNA modification annotation

Given the incomplete nature of the annotation we used, we set out to see if our classifier could expand the annotation by predicting modifications across all human tRNAs. We expected that the universally conserved modifications, e.g., m^1A, would appear in all sequenced tRNAs despite those sites sometimes being absent from known annotations. Most of the undetected modifications were m^2G sites, and our low sensitivity for m^2G is likely due to its mild effect on RT incorporation [168].

In total, we predicted 78 modification sites that were absent from the annotation (Supplementary Table 2). In many cases the modifications were absent because the specific tRNA was not listed. First, we looked at isoacceptor tRNAs and matched 25 sites to m^1A9, m^1A58, m^1G9, $m^2{}_2G26$, m^1G37, m^3C32, and Y39. For the other 53 sites not previously uncovered, we then searched across all tRNAs; this led to an additional 39 matched sites that were known to be modified in at least one type of tRNA. The remaining 14 sites were considered completely novel.

Table 2.2 – All tRNA sites predicted to be modified by HAMR

tRNA	Site	Predicted mod type	Matches DB	Type
Glu(UUC)2	C3	m^3C	*N/A*	Novel
Leu(CAA)1	A26	$m^1A\|m^1I\|ms^2i^6A$	*N/A*	Novel
Met(CAU)1	C20	m^3C	*N/A*	Novel
Thr(HGU)1	A39	t^6A	*N/A*	Novel
Thr(UGU)2	A39	t^6A	*N/A*	Novel
Val(UAC)1	A26	$m^1A\|m^1I\|ms^2i^6A$	*N/A*	Novel
mtAsn(GUU)1	A72	$m^1A\|m^1I\|ms^2i^6A$	*N/A*	Novel
mtAsn(GUU)1	G73	m^1G	*N/A*	Novel
mtCys(GCA)1	G45	m^1G	*N/A*	Novel
mtCys(GCA)1	G49	m^1G	*N/A*	Novel
mtCys(GCA)1	U73	Y	*N/A*	Novel
mtGln(UUG)1	G73	m^1G	*N/A*	Novel
mtLys(UUU)1	A59	$m^1A\|m^1I\|ms^2i^6A$	*N/A*	Novel
mtPhe(GAA)1	A59	$m^1A\|m^1I\|ms^2i^6A$	*N/A*	Novel
Arg(CCK)1	G39	m^1G	N	Known site on other tRNA
Arg(UCU)3	A38	t^6A	N	Known site on other tRNA
Arg(YCG)1	A38	t^6A	N	Known site on other tRNA
Asp(GUC)1	A9	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on other tRNA
Asp(GUC)2	A9	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on other tRNA
Glu(YUC)1	U33	Y	Y	Known site on other tRNA
Lys(UUU)1	A38	t^6A	N	Known site on other tRNA
Met(CAU)1	A38	t^6A	N	Known site on other tRNA
Met(CAU)2	A38	t^6A	N	Known site on other tRNA
Met(CAU)3	A38	t^6A	N	Known site on other tRNA
Pro(HGG)1	G6	*No consensus*	*N/A*	Known site on other tRNA
Thr(CGU)1	A38	t^6A	N	Known site on other tRNA
Thr(CGU)2	A38	t^6A	N	Known site on other tRNA
Thr(HGU)1	A38	t^6A	N	Known site on other tRNA
Thr(UGU)1	A38	t^6A	N	Known site on other tRNA
Thr(UGU)2	A38	t^6A	N	Known site on other tRNA
Trp(CCA)2	A38	t^6A	N	Known site on other tRNA
Val(CAC)1	U33	Y	Y	Known site on other tRNA
Val(UAC)2	G39	m^1G	N	Known site on other tRNA
mtAla(UGC)1	A9	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on other tRNA
mtAla(UGC)1	G37	*No consensus*	*N/A*	Known site on other tRNA
mtArg(UCG)1	A16	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on other tRNA
mtAsn(GUU)1	U1	Y	Y	Known site on other tRNA
mtAsn(GUU)1	G26	m^1G	N	Known site on other tRNA
mtCys(GCA)1	G9	m^1G	Y	Known site on other tRNA
mtCys(GCA)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on other tRNA
mtGln(UUG)1	G9	m^1G	Y	Known site on other tRNA
mtGln(UUG)1	U34	Y	Y	Known site on other tRNA
mtGln(UUG)1	G37	$m^2G\|m^2{}_2G$	N	Known site on other tRNA
mtGlu(UUC)1	U33	D	N	Known site on other tRNA

tRNA	Site	Predicted mod type	Matches DB	Type
mtGlu(UUC)1	U34	D	N	Known site on other tRNA
mtLys(UUU)1	U34	D	N	Known site on other tRNA
mtMet(CAU)1	C32	m^3C	Y	Known site on other tRNA
mtPhe(GAA)1	U33	Y	Y	Known site on other tRNA
mtPro(UGG)1	A9	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on other tRNA
mtPro(UGG)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on other tRNA
mtVal(UAC)1	U33	Y	Y	Known site on other tRNA
mtVal(UAC)1	U34	D	N	Known site on other tRNA
mtVal(UAC)1	U40	D	N	Known site on other tRNA
Arg(ACG)1	U39	Y	Y	Known site on isoacceptor
Arg(UCG)1	G9	$m^2G\|m^2_2G$	N	Known site on isoacceptor
Arg(UCG)1	G26	$m^2G\|m^2_2G$	Y	Known site on isoacceptor
Arg(UCG)1	G37	m^1G	Y	Known site on isoacceptor
Arg(UCG)1	U39	Y	Y	Known site on isoacceptor
Arg(UCG)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on isoacceptor
Arg(YCG)1	U39	Y	Y	Known site on isoacceptor
Ile(UAU)1	G9	m^1G	Y	Known site on isoacceptor
Ile(UAU)1	G26	*No consensus*	*N/A*	Known site on isoacceptor
Thr(CGU)1	G9	*No consensus*	*N/A*	Known site on isoacceptor
Thr(CGU)1	G26	$m^2G\|m^2_2G$	Y	Known site on isoacceptor
Thr(CGU)1	C32	m3C	Y	Known site on isoacceptor
Thr(CGU)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on isoacceptor
Thr(CGU)2	G9	*No consensus*	*N/A*	Known site on isoacceptor
Thr(CGU)2	G26	$m^2G\|m^2_2G$	Y	Known site on isoacceptor
Thr(CGU)2	C32	m^3C	Y	Known site on isoacceptor
Thr(CGU)2	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on isoacceptor
Thr(UGU)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on isoacceptor
Thr(UGU)2	C32	m^3C	Y	Known site on isoacceptor
Thr(UGU)2	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on isoacceptor
Thr(UGU)3	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on isoacceptor
mtGlu(UUC)1	A9	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on isoacceptor
mtHis(GUG)1	A9	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on isoacceptor
mtLys(UUU)1	A9	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on isoacceptor
mtTrp(UCA)1	A9	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on isoacceptor
Ala(AGC)1	G26	$m^2G\|m^2_2G$	Y	Known site on this tRNA
Ala(AGC)1	A37	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Ala(AGC)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Ala(AGC)3	G26	$m^2G\|m^2_2G$	Y	Known site on this tRNA
Ala(HGC)1	A37	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Ala(HGC)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Arg(ACG)1	G9	m^1G	Y	Known site on this tRNA
Arg(ACG)1	G26	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Arg(ACG)1	G37	m^1G	Y	Known site on this tRNA
Arg(ACG)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Arg(CCK)1	G9	m^1G	Y	Known site on this tRNA
Arg(UCU)1	G9	m^1G	Y	Known site on this tRNA

tRNA	Site	Predicted mod type	Matches DB	Type
Arg(UCU)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Arg(UCU)2	G9	m^1G	Y	Known site on this tRNA
Arg(UCU)2	G26	$m^2G\|m^2_2G$	Y	Known site on this tRNA
Arg(UCU)2	C32	m^3C	Y	Known site on this tRNA
Arg(UCU)2	U39	Y	Y	Known site on this tRNA
Arg(UCU)2	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Arg(UCU)3	G9	m^1G	Y	Known site on this tRNA
Arg(UCU)3	C32	m^3C	Y	Known site on this tRNA
Arg(UCU)3	U39	Y	Y	Known site on this tRNA
Arg(UCU)3	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Arg(YCG)1	G9	m^1G	Y	Known site on this tRNA
Arg(YCG)1	G26	$m^2G\|m^2_2G$	Y	Known site on this tRNA
Arg(YCG)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Asn(GUU)1	G9	m^1G	Y	Known site on this tRNA
Asp(GUC)1	G6	$m^2G\|m^2_2G$	Y	Known site on this tRNA
Asp(GUC)2	G6	$m^2G\|m^2_2G$	Y	Known site on this tRNA
Cys(NVM)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Gln(CUG)1	G9	m^1G	Y	Known site on this tRNA
Gln(CUG)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Gln(UUG)1	G9	m^1G	Y	Known site on this tRNA
Gln(UUG)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Glu(UUC)1	G9	m^1G	Y	Known site on this tRNA
Glu(UUC)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Glu(UUC)2	G9	m^1G	Y	Known site on this tRNA
Glu(UUC)2	U34	Y	N	Known site on this tRNA
Glu(UUC)2	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Gly(CCC)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
His(GUG)1	G37	m^1G	Y	Known site on this tRNA
His(GUG)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Ile(RAU)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Ile(UAU)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Int(CAU)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Leu(CAA)2	G26	$m^2G\|m^2_2G$	Y	Known site on this tRNA
Leu(CAA)2	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Leu(UAA)1	G26	$m^2G\|m^2_2G$	Y	Known site on this tRNA
Leu(UAA)1	G37	m^1G	Y	Known site on this tRNA
Leu(UAA)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Leu(UAG)1	G26	$m^2G\|m^2_2G$	Y	Known site on this tRNA
Leu(UAG)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Leu(UAG)2	G26	$m^2G\|m^2_2G$	Y	Known site on this tRNA
Leu(UAG)2	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Leu(WAG)1	G26	$m^2G\|m^2_2G$	Y	Known site on this tRNA
Leu(WAG)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Lys(CUU)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Lys(UUU)1	A37	t^6A	Y	Known site on this tRNA
Met(CAU)1	G26	$m^2G\|m^2_2G$	Y	Known site on this tRNA

tRNA	Site	Predicted mod type	Matches DB	Type
Met(CAU)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Met(CAU)2	G6	$m^2G\|m^2_2G$	Y	Known site on this tRNA
Met(CAU)2	G26	$m^2G\|m^2_2G$	Y	Known site on this tRNA
Met(CAU)2	U39	Y	Y	Known site on this tRNA
Met(CAU)2	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Met(CAU)3	G26	$m^2G\|m^2_2G$	Y	Known site on this tRNA
Met(CAU)3	U39	Y	Y	Known site on this tRNA
Met(CAU)3	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Phe(GAA)1	G37	m^1G	N	Known site on this tRNA
Phe(GAA)1	U39	Y	Y	Known site on this tRNA
Phe(GAA)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Phe(GAA)2	A14	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Phe(GAA)2	G26	$m^2G\|m^2_2G$	Y	Known site on this tRNA
Phe(GAA)2	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Pro(HGG)1	U38	Y	Y	Known site on this tRNA
Pro(HGG)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Ser(CGA)1	G26	$m^2G\|m^2_2G$	Y	Known site on this tRNA
Ser(CGA)1	C32	m^3C	Y	Known site on this tRNA
Ser(CGA)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Ser(GCU)1	G26	$m^2G\|m^2_2G$	Y	Known site on this tRNA
Ser(GCU)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Ser(WGA)1	G26	$m^2G\|m^2_2G$	Y	Known site on this tRNA
Ser(WGA)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Ser(YGA)1	G26	$m^2G\|m^2_2G$	Y	Known site on this tRNA
Ser(YGA)1	C32	m^3C	Y	Known site on this tRNA
Ser(YGA)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Thr(HGU)1	C32	m^3C	Y	Known site on this tRNA
Thr(HGU)1	A37	t^6A	Y	Known site on this tRNA
Thr(HGU)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Trp(CCA)1	G9	m^1G	Y	Known site on this tRNA
Trp(CCA)1	G26	$m^2G\|m^2_2G$	Y	Known site on this tRNA
Trp(CCA)1	G37	m^1G	Y	Known site on this tRNA
Trp(CCA)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Trp(CCA)2	G9	m^1G	Y	Known site on this tRNA
Trp(CCA)2	G37	m^1G	Y	Known site on this tRNA
Trp(CCA)2	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Trp(CCA)3	G9	m^1G	Y	Known site on this tRNA
Trp(CCA)3	G26	$m^2G\|m^2_2G$	Y	Known site on this tRNA
Trp(CCA)3	G37	m^1G	Y	Known site on this tRNA
Trp(CCA)3	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Tyr(GUA)1	U20	D	Y	Known site on this tRNA
Tyr(GUA)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Val(CAC)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Val(HAC)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Val(UAC)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
Val(UAC)2	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA

tRNA	Site	Predicted mod type	Matches DB	Type
mtArg(UCG)1	A9	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
mtAsn(GUU)1	A9	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
mtGly(UCC)1	A9	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
mtLeu(UAA)1	G9	m^1G	Y	Known site on this tRNA
mtLeu(UAA)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
mtLeu(UAG)1	A9	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
mtLeu(UAG)1	G37	m^1G	Y	Known site on this tRNA
mtLeu(UAG)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
mtPhe(GAA)1	A9	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
mtPhe(GAA)1	A37	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
mtPro(UGG)1	G37	m^1G	Y	Known site on this tRNA
mtSer(UGA)1	C32	m^3C	Y	Known site on this tRNA
mtSer(UGA)1	A58	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
mtThr(UGU)1	A9	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA
mtThr(UGU)1	C32	m^3C	Y	Known site on this tRNA
mtTrp(UCA)1	A37	$m^1A\|m^1I\|ms^2i^6A$	N	Known site on this tRNA
mtTyr(GUA)1	G9	m^1G	Y	Known site on this tRNA
mtTyr(GUA)1	A37	$m^1A\|m^1I\|ms^2i^6A$	N	Known site on this tRNA
mtVal(UAC)1	A9	$m^1A\|m^1I\|ms^2i^6A$	Y	Known site on this tRNA

2.3.5. Validation in *S. cerevisiae* small RNA dataset

In order to validate HAMR and demonstrate its utility in other organisms, we tested the software using a previously published yeast small RNA dataset [45]. We remapped the reads to the latest *Saccharomyces cerevisiae* genome release (sacCer3, UCSC) and applied the same procedure as with the human data to collapse the yeast tRNA loci into families. Of the 3,783 sequenced yeast tRNA sites with coverage greater than 10, 67 were called as potentially modified sites. Of these, 56 (84%) corresponded exactly to known modifications in tRNAdb or MODOMICS. Six more sites corresponded to positions that were not annotated as being modified on their particular tRNAs, but were known to be modified in an isoacceptor tRNA. The final five sites were known to be modified in other tRNAs. The sensitivity for RT-affecting modification was higher than those not predicted to affect RT incorporation (**Figure 2.10**). Similar to the human data, when we used the less conservative null hypothesis H_0^1, we were able to detect 100% of the inosine sites, as well as a t^6A, an m^3C, and an ac^4C site (**Figure 2.11**).

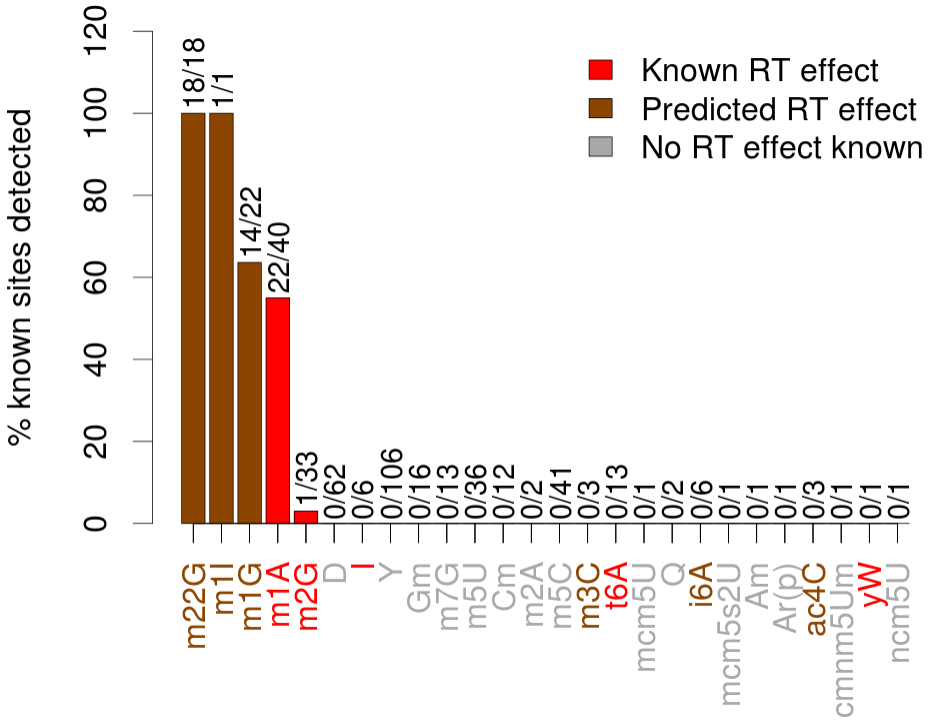

Figure 2.10 – HAMR's sensitivity in an independent *S. cerevisiae* dataset using the strict model H_0^2

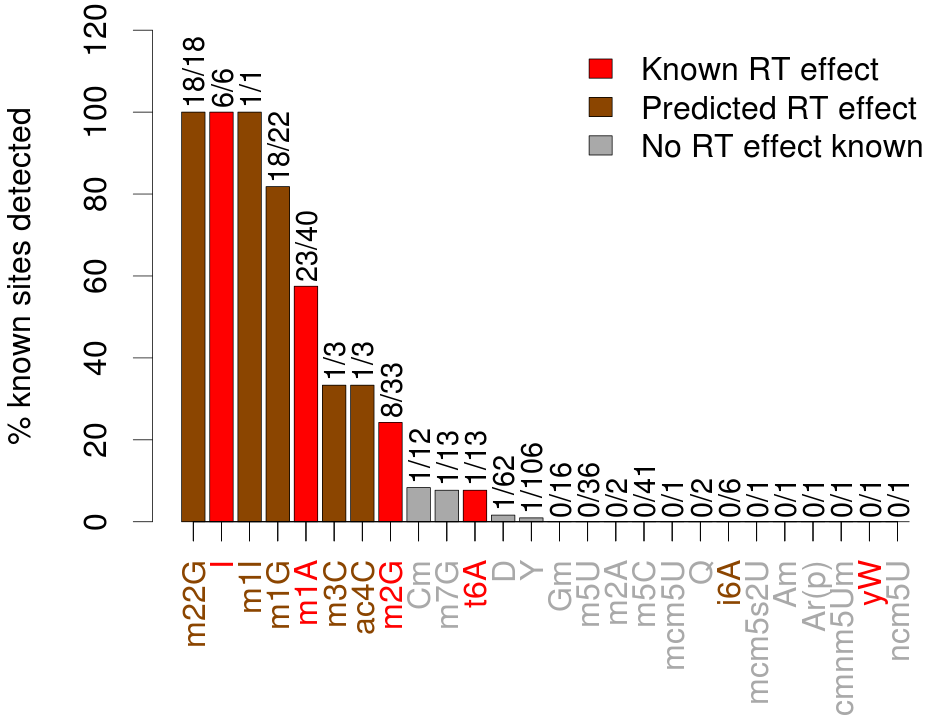

Figure 2.11 - HAMR's sensitivity in an independent *S. cerevisiae* dataset using the loose model H_0^1

The sequenced nucleotide patterns in yeast were similar to those in the human brain data. (**Figure 2.6**c,d). The fact that the two datasets were generated using different library preparations, sequenced by different versions of Illumina sequencers attests to the robustness of the statistical model we have developed. In fact, the classifier trained on human tRNAs was able to achieve 90% accuracy for modified adenosines and 65% accuracy for modified guanosines in yeast tRNAs.

2.3.6. Validation in human rRNA(-)-seq dataset

In order to ascertain the reproducibility of the tRNA modifications that were not directly present in the databases, we generated additional RNA-seq data from whole transcriptome (rRNA-depleted) libraries, which include entire tRNAs as opposed to only tRNA fragments. We compared both the "seminovel" and "novel" tRNA sites in the small RNA libraries to the whole transcriptome libraries (**Table 2.3**). Seminovel here means the site is not annotated as modified on that particular tRNA, but is annotated on some other tRNA accepting a different amino acid. Of the 23 seminovel sites that were called in more than half of the smRNA libraries, 10 (43%) are also called in at least one whole transcriptome library. Two had drastically lower coverage in the whole transcriptome libraries. The remaining 13 (mostly ms^2i^6A38) sites could not be detected in the whole transcriptome libraries, possibly due to a real difference in ms^2i^6A modification rates between tRNA fragments and whole tRNAs. Of the 6 novel sites detected in more than half of the smRNA libraries, 4 were detected in the whole transcriptome libraries. The remaining two had drastically lower read coverage in the whole transcriptome libraries.

Table 2.3 – Comparison of novel sites in smRNA data to same loci in an rRNA(-) libraries.

tRNA	Site	Predicted mod	smRNA (/4)	rRNA(-) (/5)	Low coverage
mtLys(UUU)1	A59	$m^1A\|m^1I\|ms^2i^6A$	4	5	N
Val(UAC)1	A26	$m^1A\|m^1I\|ms^2i^6A$	4	2	Y
mtPhe(GAA)1	A59	$m^1A\|m^1I\|ms^2i^6A$	4	1	Y
Thr(HGU)1	A39	t^6A	4	0	Y
Met(CAU)1	C20	m^3C	3	5	
mtAsn(GUU)1	A72	$m^1A\|m^1I\|ms^2i^6A$	3	0	Y
Glu(UUC)2	C3	m^3C	2	5	N
Leu(CAA)1	A26	$m^1A\|m^1I\|ms^2i^6A$	2	0	N

Table 2.4 – Comparison of seminovel sites to rRNA(-) libraries.

tRNA	Site	Known mod	smRNA (/4)	rRNA(-) (/5)	Low coverage
Asp(GUC)1	A9	m^1A	4	5	N
Asp(GUC)2	A9	m^1A	4	5	N
Pro(HGG)1	G6	m^2G	4	5	N
mtAla(UGC)1	A9	m^1A	4	5	N
mtAla(UGC)1	G37	$m^1G\|o_2yW$	4	5	N
mtGln(UUG)1	G37	$m^1G\|o_2yW$	4	5	N
mtPro(UGG)1	A9	m^1A	4	5	N
mtCys(GCA)1	G9	$m^1G\|m^2G\|xG$	4	4	N
mtGlu(UUC)1	U34	xU	4	2	N
Arg(YCG)1	A38	ms^2i^6A	4	0	N
Lys(UUU)1	A38	ms^2i^6A	4	0	Y
Met(CAU)3	A38	ms^2i^6A	4	0	N
Thr(HGU)1	A38	ms^2i^6A	4	0	N
mtPhe(GAA)1	U33	Y	4	0	N
mtVal(UAC)1	U34	xU	4	0	N
mtLys(UUU)1	U34	xU	3	3	N
Arg(CCK)1	G39	Gm	3	0	N
Arg(UCU)3	A38	ms^2i^6A	3	0	Y
Met(CAU)1	A38	ms^2i^6A	3	0	N
Met(CAU)2	A38	ms^2i^6A	3	0	N
Thr(CGU)1	A38	ms^2i^6A	3	0	N
Thr(UGU)2	A38	ms^2i^6A	3	0	N
mtAsn(GUU)1	U1	Y	3	0	N
Glu(YUC)1	U33	Y	2	0	N
Thr(CGU)2	A38	ms^2i^6A	2	0	N
Thr(UGU)1	A38	ms^2i^6A	2	0	N
Val(UAC)2	G39	Gm	2	0	N
mtCys(GCA)1	A58	m^1A	2	0	N

2.3.7. Detecting modifications in other RNAs

Scanning the entire human small RNA transcriptome and excluding tRNAs revealed 73 sites with mismatch patterns potentially corresponding to RNA modifications (**Table 2.5**). Nearly half (36) of these sites fell within known pre-microRNAs. Since the microRNA sites nearly always fell within 2 nucleotides of the 3' ends of mature microRNAs as annotated by mirBase [93], they most likely correspond to untemplated nucleotide additions, a phenomenon that has previously been observed in small RNA-seq datasets [31].

Table 2.5 – Candidate sites of modification across the entire small RNAome

RNA type	No. sites
tRNA	166
miRNA	36
mt-tRNA	13
intergenic	11
mRNA_intron	5
rRNA	5
transposon	4
ncRNA_exon	3
Antisense mRNA exon	2
Antisense transposon	2
snRNA	2
Antisense mRNA intron	1
Antisense ncRNA exon	1
scRNA	1

2.3.8. Software

Users may submit a link to a remote indexed BAM (read alignment) file to the online version of HAMR. HAMR detects candidate modification sites either transcriptome-wide or at selected loci specified by transcript ID or genomic coordinates. Users may also opt to filter out known dbSNP sites for human data and select various options affecting the stringency of the analysis, including p-value or FDR thresholds, minimum coverage, and which null hypothesis to use. The web version of HAMR is available at http://wanglab.pcbi.upenn.edu/hamr.

2.4. Discussion

Here we present HAMR, a high-throughput method to map RNA modifications within all classes of RNAs by identifying misincorporation of nucleotides by reverse transcriptase during production of cDNA products. While traditional methods use chemical treatment of the RNAs prior to RT, many modifications are still detectable even without treatment due to their effect on RT incorporation. This is advantageous because it allows for retrospective assays of potential RNA modifications in existing RNA-seq datasets, and also because it allows for the detection of RNA modifications with only one sequencing run. However, it is worth noting that the use of different chemical treatments in addition to different types of RT enzymes should expand the range of modifications that are detectable by HAMR. Since many modifications also cause complete halts in RT, a future research direction is to develop a method that allows the utilization fragment endpoint locations for modification mapping.

We have also found that the number of allowed mismatches in read alignment places a limit on the detection of nearby modifications. Improvement of methods, like the one presented here, will thus necessitate development of an alignment method that allows mismatches at arbitrary sites. This would be similar to the mapping methods used for bisulfite sequencing data [163], which are designed to map reads accurately in the face of cytosine deamination.

2.5. Acknowledgements

Brain samples were obtained from the Center for Neurodegenerative Disease Research at the University of Pennsylvania. RNA-seq experiments were carried out with help from Vivianna M. Van Deerlin, Virginia Lee, John Q. Trojanowski, Alice Chen-Plotkin, Gerard D. Schellenberg, and

Steven E. Arnold, and their lab members. We thank Mingyao Li, Zissimos Mourelatos, and the members of the Wang and Gregory labs for their comments.

3. Classification of RNAs by Analysis of Length (CoRAL)

Appeared in: Leung YY*, Ryvkin P*, Ungar LH, Gregory BD, Wang L-S. CoRAL: predicting non-coding RNAs from small RNA-sequencing data. Nucleic Acids Res. 2013. (*Joint first authors)

3.1. Introduction

One of the most significant biological discoveries of the last decade includes the discovery of new types of RNAs and their specific functions in eukaryotic cells [48,153]. For instance, non-coding RNAs (ncRNAs) are transcripts that are not translated into proteins but serve other important biological functions. ncRNAs have highly diverse functions including protein translation (transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs)), regulation of gene expression (microRNAs (miRNAs) and long intergenic non-coding RNAs (lincRNAs)) [71,87], pre-mRNA splicing (small nuclear RNAs (snRNAs)) [16], RNA modification (small nucleolar RNAs (snoRNAs)) [111], and the list is still expanding. Advances in high-throughput sequencing technologies have led to the unexpected discovery that up to 93% of the human genome is transcribed in some tissues [25]. Thus, it is not surprising that the non-coding RNA database [19] includes 135 different ncRNA classes. Unfortunately, the classification of most RNAs in this database is more representative of the historical process by which the ncRNAs were discovered such as sedimentation coefficient (e.g. 4.5S RNA) or cellular location (e.g. snoRNA), than of their true cellular function. This gap highlights the fact that most transcribed regions are still of unknown molecular function and biological significance.

Given that little is known about most ncRNAs, a potential approach is to gather an enormous amount of experimental data efficiently and systematically using RNA-seq, and analyze these

data using sophisticated computational approaches. Unlike microarrays, RNA-seq does not rely on target probe hybridization, and thus one does not need to know in advance which regions are being transcribed. These properties make RNA-seq a promising tool to study ncRNA biology. Additionally, RNA-seq is highly versatile in that it can be modified to study specific properties, e.g. small RNA sequencing (smRNA-seq) [95] where gel-based size selection is used to enrich for RNAs with particular sequence lengths.

While traditional methods predict RNA function using primary sequence or alignment information, new approaches using RNA-seq data have been proposed. For example, the miRDeep2 algorithm [58] searches for genomic regions that fold into hairpin structures and are enriched for sequenced reads next to the hairpin loop region (the expected location of mature miRNAs) to identify potential miRNA loci. Additionally, Langenberger *et al.* [96] pioneered the use of smRNA-seq features such as abundance and block length distribution to classify ncRNAs. Their method DARIO [53] uses random forest (RF) classifiers to differentiate between tRNA, miRNA, and snoRNA loci with reasonable performance. However, features generated from DARIO are not normalized by transcript-wide abundance; as a result, the most informative feature for miRNA identification is their overall abundance. This does not generalize well to other ncRNAs and is simply a result of the fact that miRNAs are highly abundant in human smRNA-seq datasets.

Erhard and Zimmer [50] used similarities between RNA transcripts to classify ncRNAs. Their similarity measure was created based on the relative positions and lengths obtained from sequencing experiments. However, relative positions of reads require good knowledge on the start- and end-points of transcripts within a genome sequence, which is a challenge for newly discovered classes of ncRNA. Evaluation of their method on two classes of RNA (miRNAs and tRNAs) yielded performance with recall values of 98% and precision of 60% for miRNAs and ~80% for tRNAs, which leaves room for improvement.

To address the limitations of these previous RNA function classifiers, we have developed a framework for classifying RNA transcripts by functional categories using smRNA-seq data (Fig.

1), which can then be applied to identify unannotated RNAs with similar functions in other organisms in the future. To do this, we first designed algorithms to generate several types of features from smRNA-seq data based on read length distribution, strand specificity, and the secondary structure of the transcript for transcribed genomic regions. We then applied a multi-class classification algorithm with feature selection and cross validation schemes included to train classifiers among a collection of known RNA functional classes including lincRNAs, miRNAs, scRNAs, C/D box snoRNAs, snRNAs, and transposon-derived RNAs. For each RNA class, we identified the most informative features that might be associated with the molecular mechanisms and metabolic processes of the functional classes. Trained models, informative features, and annotation results have been validated using: 1) external datasets, 2) SAVoR [100], a visualisation tool for RNA structures [101], and 3) curation of the primary literature.

3.2. Methods

3.2.1. Processing of small RNA-seq data

The smRNA-seq data used for our analysis came from four sources: human brain data generated as part of this study (GSE43335), a previously published dataset from human skin (GSE31037) [86], and published datasets from human liver (SRR040571) and muscle (SRR040572) [52]. The human brain data was obtained by sequencing small RNAs (smRNAs) extracted from the dorsolateral prefrontal cortex of four deceased human patients with no apparent pathology. All reads were trimmed to remove the Illumina 3' adapter sequence using cutadapt [110], and only those reads containing the adapter were taken as true smRNA reads. Reads were mapped to the reference genome GRCh37/hg19 using Bowtie [97] and those mapping to multiple loci were discarded. In order to merge reads into transcribed loci, we used the RSEQTools' [72] bgrSegmenter tool. (**Table 3.1**)

Table 3.1 – Number of reads and loci at each stage of smRNA-seq processing

	Raw reads (millions)	3' adapter trimmed reads (millions)	Uniquely mapped reads (millions)	Small RNA loci, ≥ 1 read	Small RNA loci, ≥ 15 reads
Brain	104.1	51.9 (50%)	15.4 (30%)	6,246	4,525 (72%)
Skin	307.0	188.4 (61%)	85.4 (28%)	11,423	8,638 (76%)
Liver	3.37	1.48 (44%)	1.15 (78%)	269	216 (80%)
Muscle	3.79	3.42 (90%)	0.368 (11%)	218	178 (82%)

3.2.2. Labelling training data

Functional categories were assigned to loci by overlapping their coordinates with RNA annotations from the UCSC Genome Browser [59]. While there are many different types of ncRNA described, we focused on a subset of functional classes where sufficient numbers of confirmed loci were available to train predictive models.

For quality control purposes, loci covered by fewer than 15 reads were discarded. This value was chosen as a compromise between selecting high quality sufficiently transcribed regions and identifying significant levels of loci for each class (**Figure 3.1**). Based on these criteria the following six RNA classes were selected: lincRNAs, miRNAs, scRNAs, C/D box snoRNAs, snRNAs and transposon-derived RNAs (**Figure 3.2** and **Figure 3.3**). We excluded rRNAs and tRNAs because they are easily identifiable by sequence homology alone.

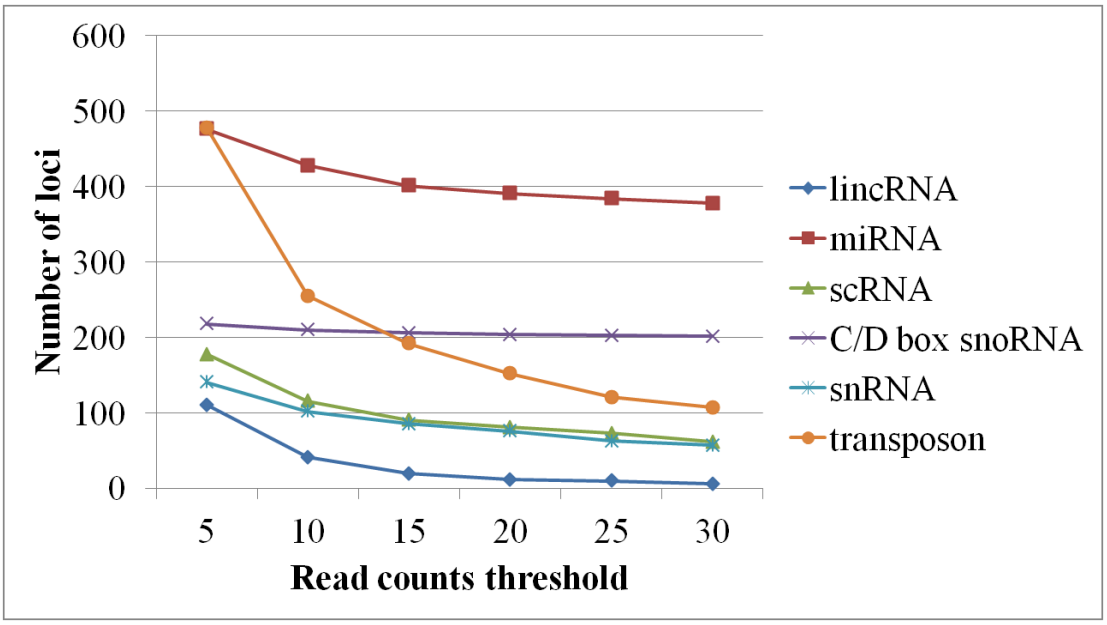

Figure 3.1 – The effect of read count thresholds on the ability to detect smRNA loci

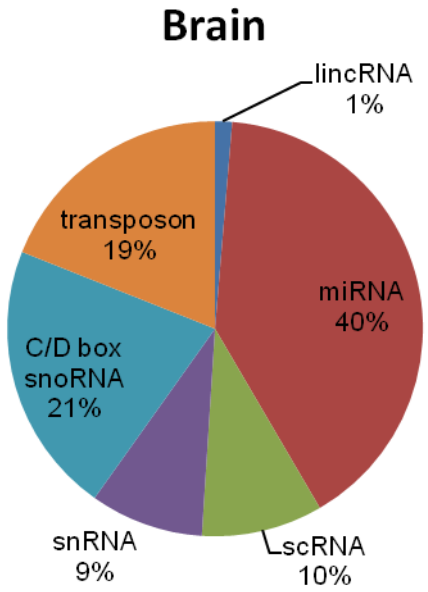

Figure 3.2 – Summary of RNA classes in the brain smRNA-seq

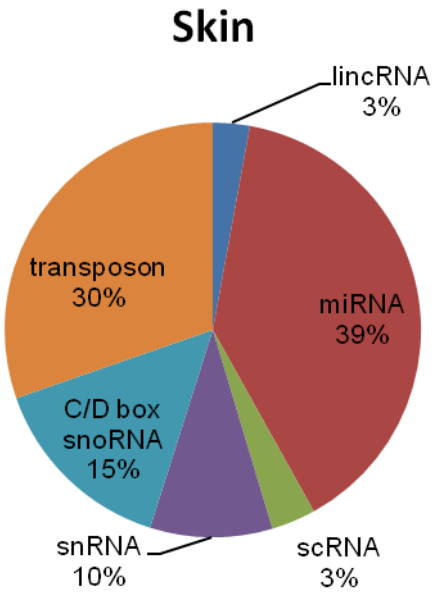

Figure 3.3 – Summary of RNA classes in the skin smRNA-seq

3.2.3. Feature generation

We noted that features used for classification purposes should be flexible, comprehensive, efficient, and scalable. Therefore, we developed features that would be most likely to reflect the underlying biological properties of small ncRNAs. For example, microRNAs are consistently processed into their mature form of 22 nucleotide (nt) fragments as a consequence of Dicer's activity on the stem-loop structure of pre-microRNAs [10]. It is reasonable to assume, then, that the lengths of smRNAs are consistent with some aspects of their biogenesis, which should also be consistent within classes sharing the same molecular function. Thus, for a transcribed locus *i* that starts at genomic position *a* and ends at position *b,* we define the length features as:

$$s_{iL} = \sum_{k=a}^{b} \frac{N_{Lk}}{\text{Length}\quad (i)}$$

for read lengths $14 \le L \le 30$, where N_{Lk} is the number of reads of length *L* mapping to base *k* and Length (*i*) is the length of locus *i*. The values of these 17 features are then transformed into log-odds-ratios via the following normalization procedure:

$$p_{iL} = \frac{1+s_{iL}}{\sum_{14 \le L \le 30} s_{iL}}, \quad x_{iL} = \log\left[\frac{p_{iL}}{1/17}\right]$$

In addition to the read lengths, we introduced a feature based on the abundance of antisense transcription. The numerical value of this feature reflects the number of reads mapped to the antisense strand of the transcribed locus. This feature is generated based on the assumption that the presence of antisense transcription at a locus is relevant to the biogenesis of smRNAs from this region. Another important feature that is likely to be specific to smRNA biogenesis is the specificity of cleavage positions. We encode this as two features: 5' and 3' positional entropy. The entropy is computed based on the distributions of the 5' and 3' end positions of all smRNA reads mapped to a given locus, respectively. This entropy feature is designed to capture the specificity (or degeneracy) of RNA cleaving-enzymes specific to the production of different types of smRNAs. For example, the processing of mature microRNAs from pre-microRNAs tends to produce fragments with a more stable 5' cleavage position (low entropy) and more variable 3' end

(higher entropy). We also generate features corresponding to the base composition of the reads, weighted by their expression: these are the four nucleotide frequencies transformed into a log-odds ratio relative to equal base frequencies. Additionally, we compute the predicted minimum free energy of the genomic region surrounding the transcribed locus (40 bp on either side) using RNAfold with the default parameters [78].

3.2.4. Feature selection and classification framework

In order to identify features that are most representative of the six ncRNA classes, we used the R package varSelRF (version 0.7-3) [41], which finds a small, optimal set of non-redundant features for each class. When computing the feature importance we used varSelRF with parameters (mtryFactor=4, vars.drop.fac = 0.35, ntree = 1e3). For the number of variables mtryFactor setting we tried various values and saw no difference in performance, so we used a value corresponding to the square root of the number of features as recommended in the literature [144]. Similarly, the number of trees did not greatly affect accuracy but had a large impact on running time. The selected variable drop factor yielded classifiers with the highest training accuracy. Random forest was used as a classifier to distinguish between multiple RNA classes. The feature selection portion uses both backwards variable elimination and selection based on the variable importance index outputted by the RF model. When training the models, 100 RF models comprised of 1000 trees were built to determine the stability of results.

3.2.5. Evaluation of performance

Typically the performance of a binary-class classifier is evaluated by comparing values from the confusion matrix, including rates of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Other commonly used measures for binary classification are accuracy, recall/sensitivity, and positive predictive value. Measures for multi-class classification are generalized from measures used in binary classification. ACC_k is the overall accuracy, which is

the proportion of predictions that are correct: $ACC_k = (TP_k+TN_k)/(TP_k+TN_k+FP_k+FN_k)$. For every class C_k, the class-specific evaluation measures are defined by recall (REC_k) and positive predictive value (PPV_k), derived from counts of C_k from the confusion matrix. REC_k is defined as the proportion of positive labelled samples that are predicted as positive: $REC_k = TP_k/(TP_k+FN_k)$, whereas PPV_k is defined as the proportion of positive samples that are correctly identified: $PPV_k = TP_k/(TP_k+FP_k)$.

3.3. Results

3.3.1. Visualization of the length features

We hypothesized that the lengths of some small ncRNAs are specific to particular classes of precursor ncRNAs. Therefore, we tested the distribution of the read length feature for three of the ncRNA classes in the human brain and skin datasets. miRNAs demonstrated a strong peak at 22 nt in length (**Figure 3.4**, **Figure 3.3**, and **Figure 3.10**), which is consistent with what is known about the length of mature miRNAs in animals. Products coming from C/D box snoRNAs tend to be depleted of shorter RNAs and enriched for longer RNAs (**Figure 3.6**, **Figure 3.7**, and **Figure 3.11**). Transposon-derived smRNAs appear to show slightly different distributions depending on the tissue type. For example, they show a weak broad peak around 19 – 23 nt in the brain data and a flatter, weaker bias towards 16 – 22 nt in the skin data (**Figure 3.8**, **Figure 3.9**, and **Figure 3.12**).

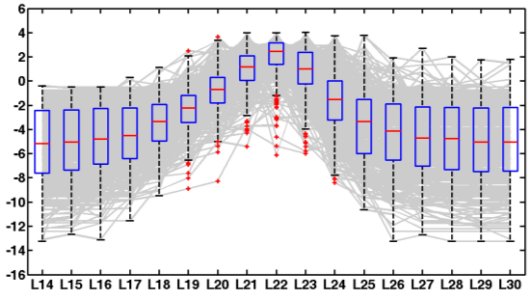

Figure 3.4 – Read length spectrum for brain miRNAs

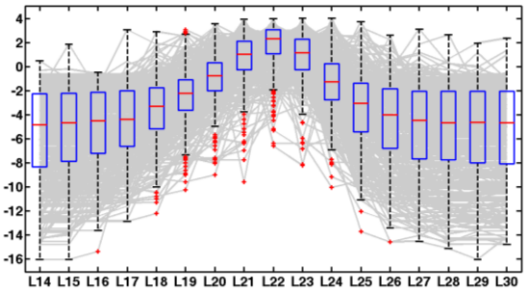

Figure 3.5 – Read length spectrum for skin miRNAs

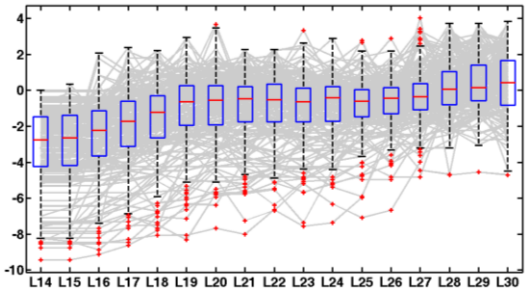

Figure 3.6 – Read length spectrum for brain C/D box snoRNAs

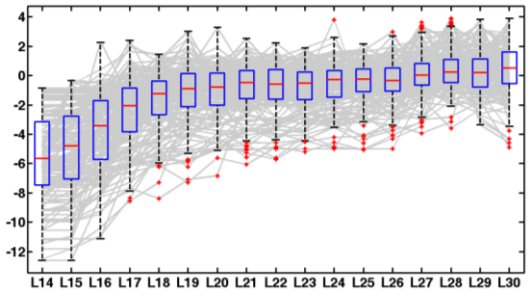

Figure 3.7 – Read length spectrum for skin C/D box snoRNAs

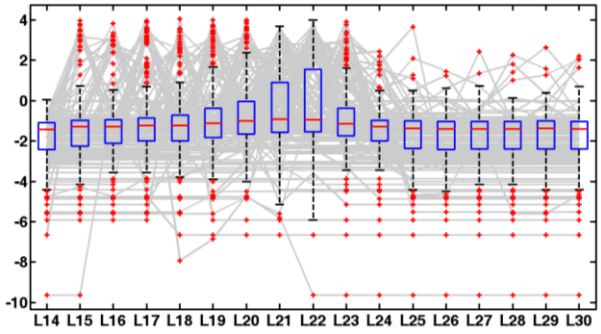

Figure 3.8 – Read length spectrum for brain transposon-derived smRNAs

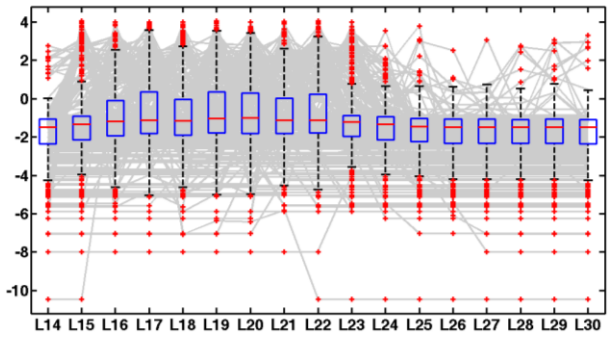

Figure 3.9 – Read length spectrum for skin transposon-derived smRNAs

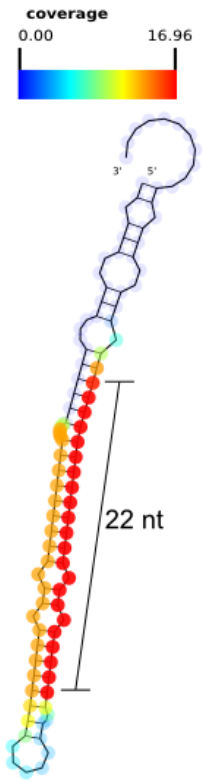

Figure 3.10 – SAVoR plot for a brain microRNA

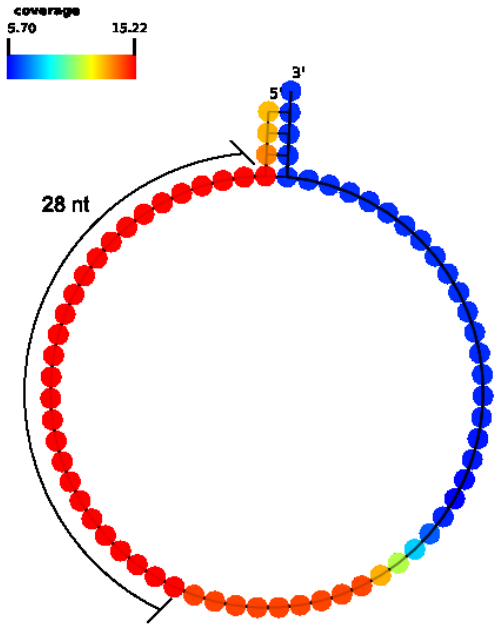

Figure 3.11 – SAVoR plot for a brain C/D box snoRNA

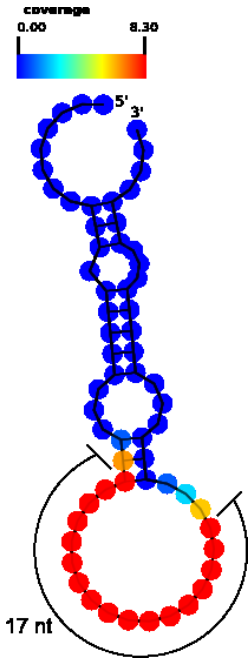

Figure 3.12 – SAVoR plot for a brain transposon-derived smRNA locus

In addition, we examined the correlations among the features in the brain dataset (**Figure 3.13**). Unsurprisingly, features corresponding to adjacent lengths correlate very strongly. Interestingly, there appear to be four clusters of lengths: 14 – 18 nt, 19 – 20 nt, 21 – 23 nt, and 24 – 30 nt. These results suggest that specific classes of smRNAs tend to have coherent lengths. We also found that positional entropy at both ends of human brain small RNAs strongly correlate. This suggests that small RNAs with high 5' cleavage specificity tend to also have high 3' cleavage specificity.

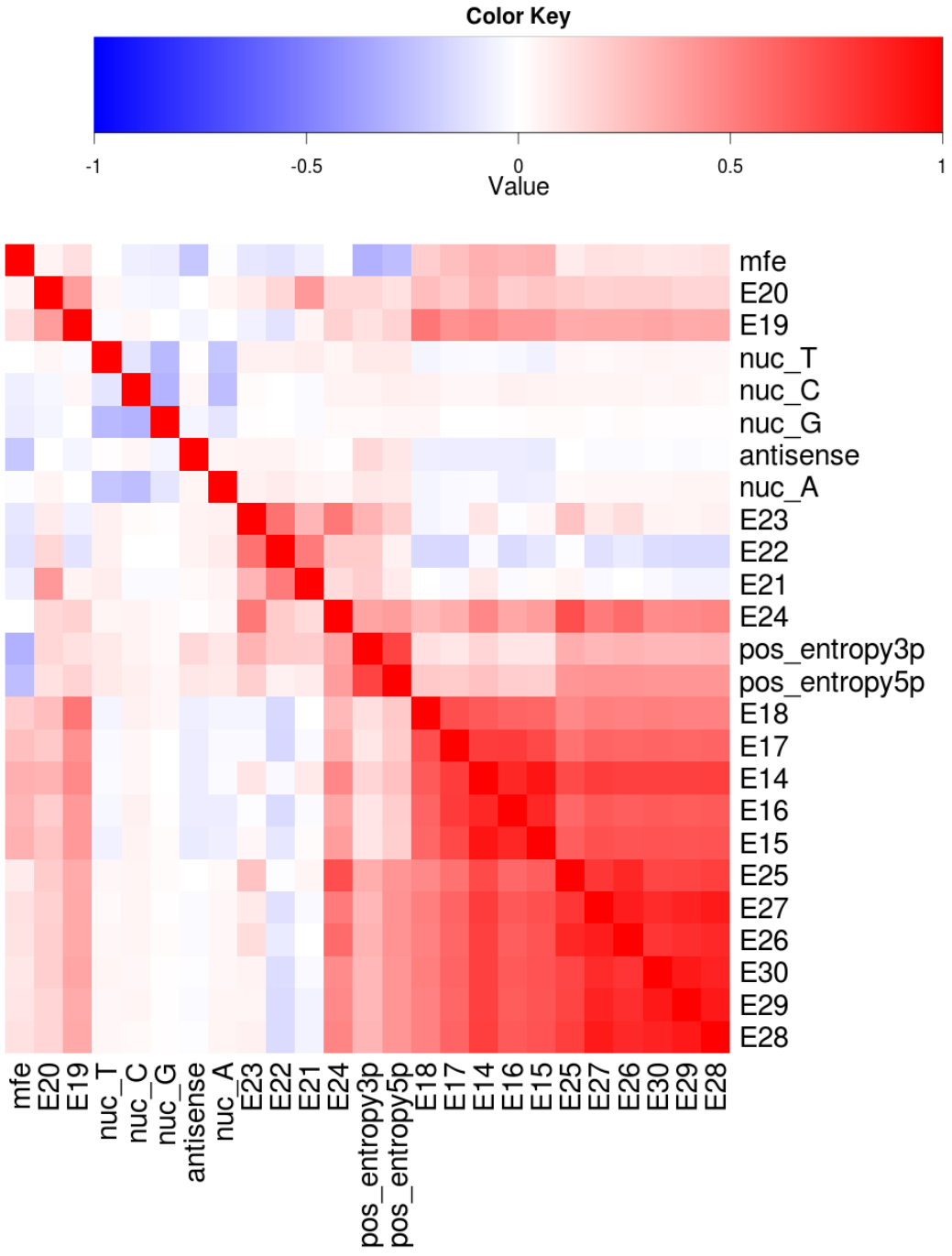

Figure 3.13 – Correlation heatmap of all the features in the brain data

3.3.2. Discriminative power of features

Due to the varying number of loci within each ncRNA class, it can be challenging to visualize all loci in a dataset. In order to determine how well the length features were able to separate the loci, we built RF trees by classifying one ncRNA class versus all other classes. We then applied multidimensional scaling (MDS) to the proximity matrix obtained from the RF trees. miRNA, C/D box snoRNAs, and transposon-derived RNAs were the most visually distinguishable classes of smRNAs using our features (**Figure 3.14** and **Figure 3.15**), and this pattern was found to be consistent between the two (brain and skin) datasets.

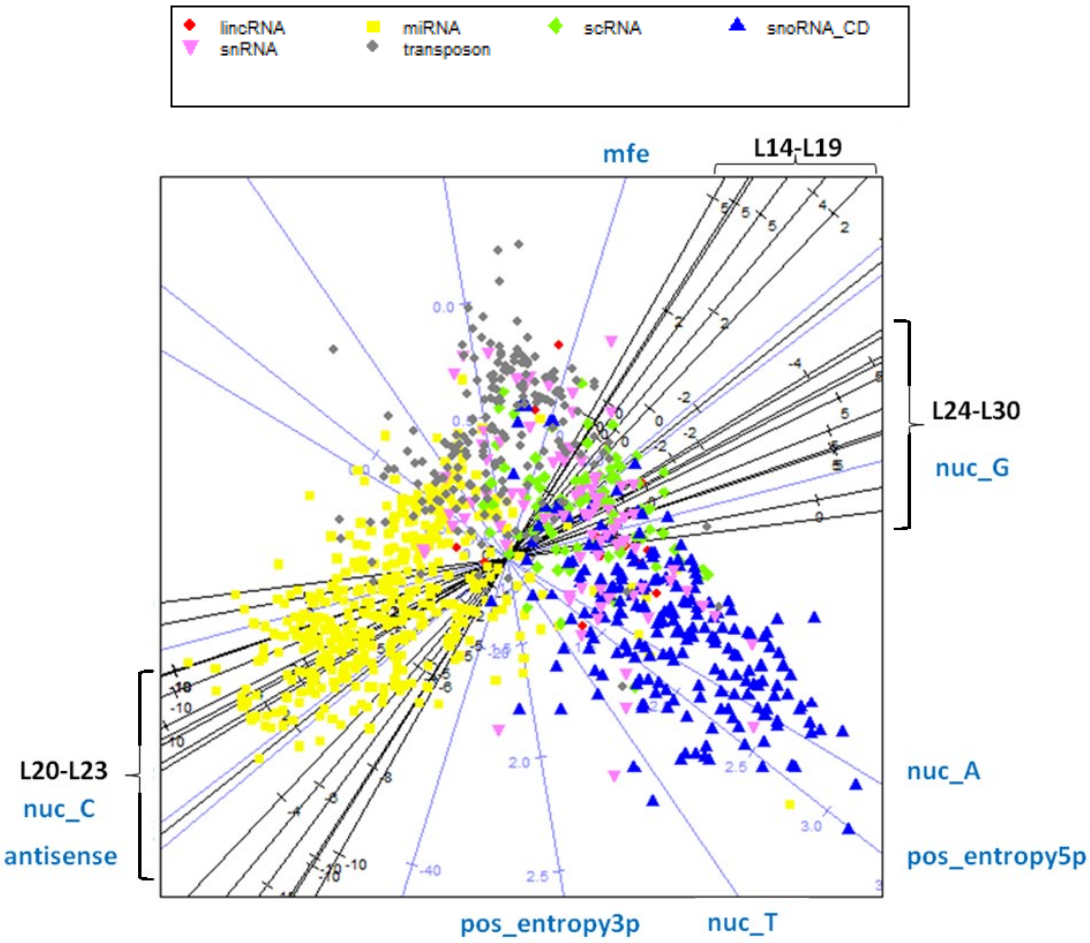

Figure 3.14 – Multidimensional-scaling projection of the features in the brain data

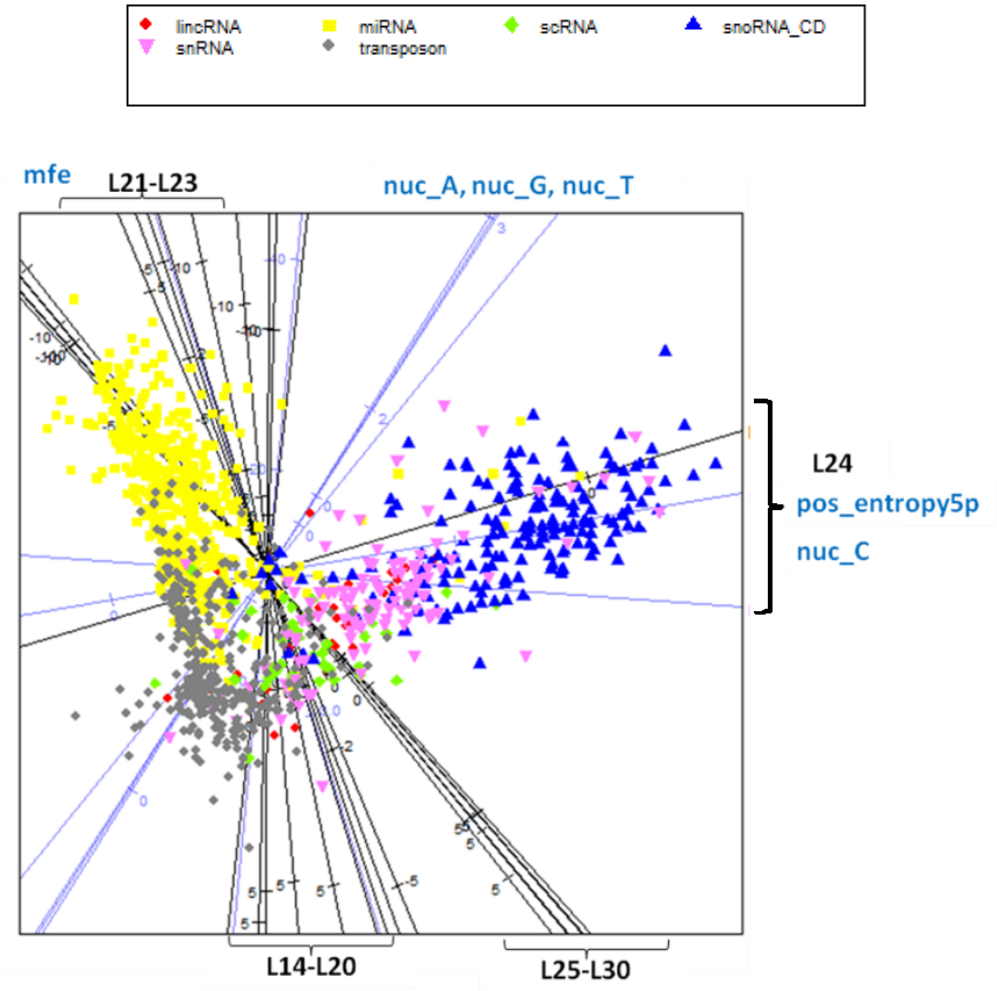

Figure 3.15 - Multidimensional-scaling projection of the features in the skin data

3.3.3. Comparison with existing classification approaches – DARIO and miRDeep

We compared our method with a published method (DARIO), which was designed for classifying smRNAs by their precursor ncRNA loci. Since DARIO only uses three classes of ncRNAs (miRNAs, C/D box snoRNAs, and tRNAs) for building its classification model, we ran CoRAL while limiting the data to those three classes only (**Table 3.2**).

Table 3.2 – Comparison of a 3-class CoRAL model to DARIO

		DARIO	*CoRAL*
miRNA	**REC (%)**	90	94
	PPV (%)	92	95
C/D box snoRNA	**REC (%)**	N/A	88
	PPV (%)	N/A	91
tRNA	**REC (%)**	84	90
	PPV (%)	81	87
Overall accuracy (%)		87	91

CoRAL gives the best results for all three classes, with an improvement of ~ 3 – 4% for miRNAs and tRNAs. DARIO reported none of the loci as being annotated as snoRNAs and so that class was unable to be compared, but demonstrates that CoRAL is able to identify these RNAs that cannot be distinguished by DARIO. When restricting the comparison to miRNAs and tRNAs, CoRAL's predictive performance is 91%, which is a 4% improvement over the same analysis performed by DARIO.

Additionally, we compared our results to those produced by miRDeep2 on the brain data (ran with default parameters). miRDeep2 had a recall of 81% and PPV of 98%, whereas CoRAL had a recall of 88% and PPV of 91% for miRNAs, while also predicting 5 other RNA classes. Thus, CoRAL has increased functional classification capabilities as well as improved overall performance compared the to currently available classifier options.

3.3.4. Building a classification model using 6 classes of ncRNAs

There are currently more than 135 classes of ncRNAs in the NONCODE database. Here, we focused on a subset of functional classes where sufficient numbers of confirmed loci were available for us to build our predictive models. A total of six classes were included: lincRNAs, miRNAs, scRNAs, C/D box snoRNAs, snRNAs, and transposon-derived smRNAs. Performance measures were averaged over 1000 different seeds of RF classifiers (**Table 3.3**).

Table 3.3 – Cross-tissue comparison of a 6-class CoRAL classifier

		Brain		***Skin***	
		CoRAL	***Baseline***	***CoRAL***	***Baseline***
lincRNA	**Count**	13		34	
	Recall (%)	16	0	1	1
	PPV (%)	62	0	38	2
miRNA	**Count**	397		465	
	Recall (%)	91	78	89	71
	PPV (%)	88	43	86	42
scRNA	**Count**	93		41	
	Recall (%)	78	1	29	0
	PPV (%)	81	7	49	0
C/D box snoRNA	**Count**	209		176	
	Recall (%)	94	14	88	5
	PPV (%)	79	22	81	15
snRNA	**Count**	87		113	
	Recall (%)	28	1	57	1
	PPV (%)	67	7	67	9
transposon	**Count**	187		361	
	Recall (%)	77	5	80	24
	PPV (%)	74	15	77	28
Overall	**Count**	986		1190	
	Accuracy (%)	81	33	79	33

For both datasets, the overall accuracy is approximately 80%, which is a significant improvement over the baseline of 33%. The best performing classes are miRNA, C/D box snoRNA, and transposon-derived RNAs. The performance of these three classes is also consistent between the two tissue types. In contrast, the lincRNA, scRNA, and snRNA classes

performed more poorly. The lower performance of these classes can possibly be attributed to their smaller representation among loci, since there were fewer smRNA loci present from these regions for both tissue types. Another potential reason for the lower performance is that these classes are less cohesive than the other classes. lincRNAs generally do not share any structural properties and are known to have diverse functional roles [24]. scRNAs are in fact an umbrella group for two distinct types of RNAs: human Y (HY) RNAs and the BC200 small cytoplasmic RNA [152], which have different secondary structures and likely different functions in the cell. Finally, the snRNA class is a highly incoherent grouping due to the structural diversity among its members. For example, while the U1 and U2 RNAs are both small, localized to the nucleus, and involved in pre-mRNA splicing, they perform very different functions and have very different secondary structures [16]. Therefore, it is reasonable to expect more diversity in the properties of smRNAs being produced by cleavage of snRNAs as opposed to the three better performing RNA classes.

3.3.5. Features that can discriminate between classes of small RNAs

While we were interested in comparing the reproducibility of the smRNA features for various ncRNA classes, an important biological question to ask is which features are specific to which ncRNA classes. To determine this, we counted the number of times a feature is selected out of the 1000 RF models (**Figure 3.16**). In order to provide potentially biologically informative insights, we also marked features as being lower- or higher-valued in one class than in the others. We found that smRNAs from C/D box snoRNAs often have a higher positional entropy at their 5' end and are very short (< 16 nt) or long (> 25nt). Interestingly, the length bias for these smRNAs is more marked in the brain data than in the skin data, but the entropy bias is consistent between tissues. snRNAs do not have many discriminative features in the skin dataset but in the brain they seem to preferentially produce shorter RNAs. Transposon-derived RNAs show very low positional entropy – suggesting that their cleavage positions tend to be very consistent. They also seem to

be depleted of miRNA-length products (22 – 24 nt), while being enriched for shorter products (< 19nt) and having high minimum free energy (MFE) values for their secondary structure.

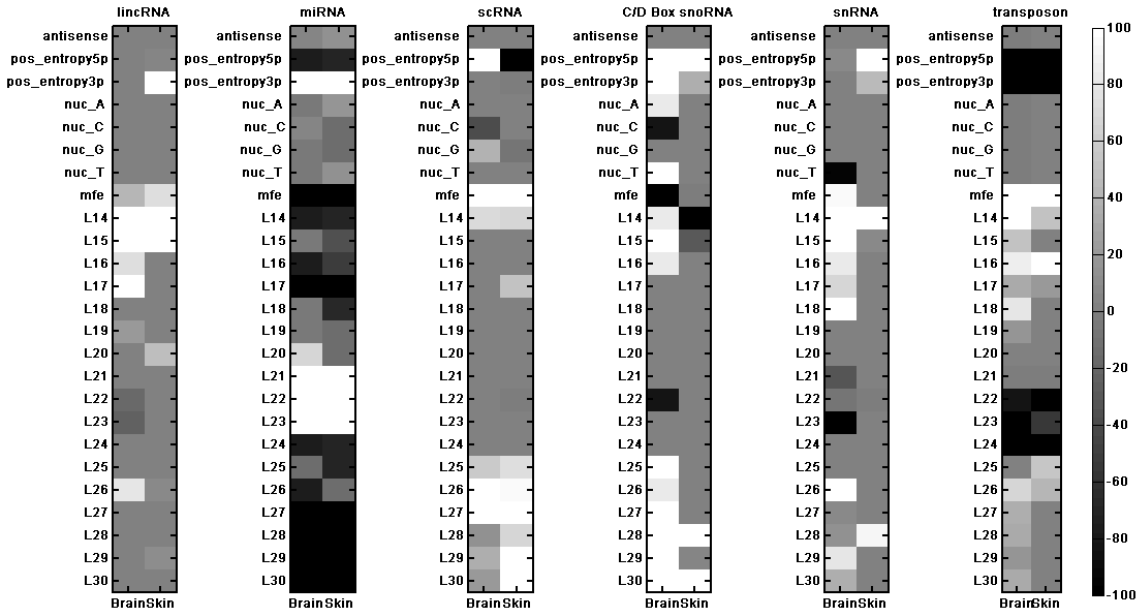

Figure 3.16 - Feature importance map of the 6-class classifier for each tissue

We found the class-specific features were largely consistent across the two tissues, but vary widely for the ncRNA classes under study. For instance, lincRNAs show a propensity to produce shorter RNAs (14 – 17 nt), with slightly longer RNAs being produced in the skin data. Additionally, miRNAs were broadly distinguished by the production of fragments between 20 and 23 nt long, and this was very consistent between the tissue types. They also display a strong bias for low 5' positional entropy and high 3' entropy. This mirrors what is already known about lower variability of miRNA cleavage at the 5' end and higher variability at the 3' end [46].

Small cytoplasmic RNA (scRNA)-derived smRNAs demonstrated a broad peak of discrimination at 27 nt for both tissue types, with skin RNAs showing longer lengths. It has previously been shown that Y RNA (a type of scRNA) fragments do produce miRNA-like smRNAs but their potential function is still unclear [158]. scRNA-derived RNAs are moderately consistent between the two tissue types, but consistently show a preference for longer products with high MFE values.

Similar to scRNAs, C/D box snoRNAs were found to produce longer fragments. In both tissues, the positional entropy at both ends of the resulting smRNAs tended to be high, indicating a great degree of variability in cleavage positions. The pattern for snRNAs was less clear because their processing was highly inconsistent between the tissue types, with the exception of the production of 14 nt fragments, which was seen in both the brain and skin datasets. This may be due to the heterogeneity in the properties (especially structural) of RNAs that are collectively referred to as snRNAs. In contrast, we found that the features distinguishing transposable element-derived smRNAs were almost entirely consistent between the two tissues. With the most discriminative features being high cleavage specificity, high MFE, smaller products, and the absence of miRNA-sized products. Thus, determining the mechanism of transposon-derived smRNA processing and their functions will likely be an interesting future research direction.

In order to determine whether a subset of features was the most useful for overall classification we selected the first five dimensions from the MDS analysis. This resulted in a drop in overall accuracy of 8% (data not shown). This suggests that while a small number of features capture most of the differences between the classes, many other features are still highly informative. More importantly, results obtained from the original features are more conducive to interpretation than a model that is only generated based on a projection of the original features.

3.3.6. Validation of the classification models between datasets

In order to evaluate the robustness of our classification models, we performed validation using independent datasets. In order to do this, we trained RF models on the brain data and applied them to the skin data and vice versa. Overall, the models were found to work fairly well, showing an accuracy of approximately 80% in both cases (**Table 3.3**). This suggests that patterns of smRNAs produced from ncRNAs are generally consistent and mostly non-tissue specific. However, we found that the degree of consistency varies among the classes of smRNAs. miRNAs, C/D box snoRNAs, and transposon-derived RNAs show the most consistent results both within and between tissue types. However, the lincRNA and snRNA classes display very

tissue-specific patterns of smRNA processing (**Table 3.3**). This is expected for lincRNAs given their tissue-specific patterns of expression. Besides tissue specificity, one other potential reason why certain classes perform much better across tissue types may be the number of loci present within the tissues being used for analysis. Since we are using a fixed minimum of 20 reads mapping to each locus, differences in overall expression between the tissue types will result in a different number of loci in each class (**Figure 3.17**, **Figure 3.18**, **Figure 3.19**, **Figure 3.20**, **Figure 3.21**, and **Figure 3.22**). Therefore, while the cross-tissue classifier performs well overall, it is limited by not only the number of loci in each class but also the consistency in these numbers across the tissue types being studied.

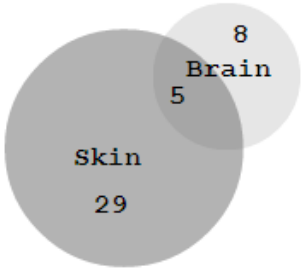

Figure 3.17 – lincRNA-derived smRNA locus overlap between brain and skin

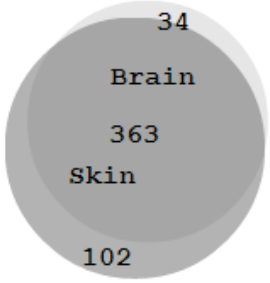

Figure 3.18 - miRNA locus overlap between brain and skin

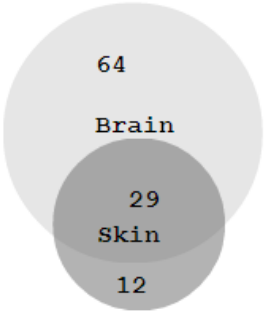

Figure 3.19 - scRNA-derived smRNA locus overlap between brain and skin

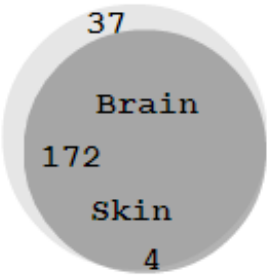

Figure 3.20 – C/D box snoRNA-derived smRNA locus overlap between brain and skin

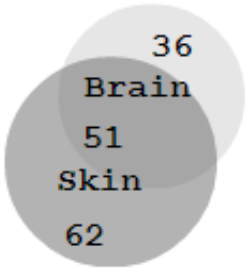

Figure 3.21 - snRNA-derived smRNA locus overlap between brain and skin

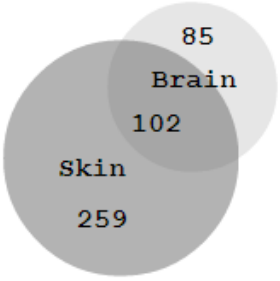

Figure 3.22 - Transposon-derived smRNA locus overlap between brain and skin

In order to further vaildate the robustness of the classifier when applied to different datasets, we tested additional publically-available smRNA-seq datasets for human liver and muscle (**Table 3.4**). We restricted the classes to those represented by at least 10 loci in all four datasets (miRNA, C/D box snoRNA, and tRNA). For each pair of datasets we trained the model on one and tested on the other. Overall the accuracies (65-93%) suggest that the model can classify across tissue types fairly well, conditional on the training dataset having high enough sequencing depth to fully characterize the lower-abundance small RNAs. For example, the liver dataset has far fewer reads than the others and thus performed poorest (<70%) when used as the training dataset. Despite this, the model was able to classify liver smRNAs fairly well (77-93%) when trained on the other tissue types. Overall, our results suggest that CoRAL is a comprehensive and robust method for classifying RNAs using smRNA-seq datasets.

Table 3.4 – Four-way independent cross-validation of the 3-class classifier

	Test			
Train	**Brain**	**Skin**	**Liver**	**Muscle**
Brain	*91%*	87%	93%	91%
Skin	81%	*89%*	81%	90%
Liver	71%	67%	*93%*	92%
Muscle	63%	67%	93%	*100%*

3.4. Conclusions

Patterns of cleavage in human ncRNAs appear to be non-random and reflect specificity in the processes that produce smRNAs from the corresponding precursors. This is despite the fact that the classes of ncRNAs studied here are defined based on differing criteria (sequence homology, secondary structure homology, biological function, cellular localization, and transcript length). While it is unknown whether these fragments or the cleavage of the precursors have some biological function, the non-random nature of the cleavage events hints at some role.

We also found that the classification features that distinguished each class of ncRNA are generally consistent across tissue types in humans, suggesting there are as-yet unknown biological pathways regulating their biogenesis. We also demonstrated that some types of ncRNAs show more tissue specific properties (lincRNAs, scRNAs, and snRNAs). However, the other three RNA classes (miRNAs, C/D box snoRNAs, and transposon-derived RNAs) are highly reproducible and consistent across two of the tissue types tested in our study.

As compared to previous work like DARIO, one of the significant contributions of CoRAL is the development of biologically interpretable features such as fragment length, cleavage specificity, and antisense transcription. These features are able to capture the essence of ncRNAs, i.e., how they are processed into smaller fragments. It seems likely that the features revealed by CoRAL can serve as a basis for further exploration and validation.

The ability of CoRAL to consistently annotate loci between tissue types suggests that it may be useful in annotating ncRNAs in other organisms and even more tissue types using only smRNA sequencing data. Thus, it will be a powerful tool for the annotation of future non-coding transcriptomes in this era of genomic progress, which complements other currently available comparative genomics methodologies. It is worth noting that our approach may even outperform homology-based methods, given the lower homology due to compensatory evolution in many classes of RNAs [113].

3.4.1. Software Availability

The CoRAL source code required genome annotation files, and prediction results are available at: http://wanglab.pcbi.upenn.edu/coral .

4. Characterizing the non-coding transcriptome of Alzheimer's Disease

4.1. Introduction

Studies of the transcriptome in Alzheimer's disease are not new [15,17,33,104,116,155]. However, all of these studies tend to focus on the protein-coding portion of the transcriptome while ignoring the non-coding portion. While there is a preponderance of functional data for proteins (often of a non-directional nature such as simple binding or association), an advantage of studying non-coding RNAs is that there are also many known directional relationships amongst them. For example, snRNAs direct splicing of pre-mRNAs; another example is that of ribosomal RNAs, differential expression of which might indicate a global downregulation of translation in the face of environmental stress. Another example is that of tRNAs as, again, markers of cellular stress – in particular, small RNAs deriving from their cleavage. Here we present a study of the non-coding transcriptome (both long and short) transcripts in the dorsolateral prefrontal cortex of the AD-affected brain.

4.2. Methods

4.2.1. RNA-sequencing

The small RNA and rRNA(-) sequencing were performed as described in section 2.2.1.

4.2.2. Calling small RNA loci and building smRNA locus families

By taking uniquely-mapping reads only, we can lose up to 60% of the reads in a small RNA library. We can mitigate this issue by keeping some cross-mapping reads and losing the ability to distinguish between different loci that are copies of the same or very similar gene (this is acceptable as they will never be resolvable at a given read-length.) In order to accomplish this, the method for clustering tRNAs by empirical cross-mapping rates in Chapter 2 was generalized and adapted to all small RNA loci in a transcriptome. Here the clustering quality criterion was

defined so as to minimize inclusion of multiple RNA classes into one cluster, and the value of K (# clusters) was chosen to minimize this value. Clusters were heavily penalized if they contained loci coming from *a priori* pairs defined as incompatible (**Table 4.1**).

Table 4.1 – RNA classes defined as incompatible when clustering loci.

RNA class	**Incompatible classes**
miRNA	tRNA, mt-tRNA, snoRNA, snRNA, rRNA
tRNA	miRNA, snoRNA, snRNA, rRNA
mt-tRNA	miRNA, snoRNA, snRNA, rRNA
snoRNA	miRNA, tRNA, mt-tRNA, snRNA, rRNA
snRNA	miRNA, tRNA, mt-tRNA, snoRNA, rRNA
rRNA	miRNA, tRNA, mt-tRNA, snoRNA, snRNA

4.2.3. Predicting the impact of tRNA activity changes on protein translation

Given a list of tRNAs that whose activities are predicted to increase, and the set of codons in the coding sequence of every human gene (obtained from Ensembl), we can assign to each gene the importance of that particular subset of tRNAs in its translation. First we take the reverse complement of each anticodon to determine approximately the set of codons it recognizes (ignoring post-transcriptional modifications of the anticodon sequence and non-canonical base-pairing). Then we can compute, for each protein-coding sequence in the genome, a score indicating the importance of this set of anticodons to that gene's translation (assuming no positional biases along the transcript in translational efficiency):

$$S_i = \frac{\sum_j I_{ij}}{N_i}$$

where S_i is the score for sequence *i*, I_{ij} is an indicator whose value is 1 when codon *j* of sequence *i* is targeted by the list of given anticodons, and N_i is the length of sequence *i* in codons.

Sequences were required to be at least 50 codons long and the maximally scoring transcript within a gene's set of transcripts was used as the score for that gene.

4.2.4. Building a network of AD-related genes

MicroRNA targets were taken from two databases of experimentally validated microRNA-mRNA interactions: mirTarBase and miRecords. Targets of the minor spliceosome were considered to be those genes that have U12 introns according to the database U12DB [5]. Protein-protein interactions were taken from the STRING tool [56].

4.3. Results

4.3.1. Sample characteristics and RNA-seq processing statistics

In order to characterize the prefrontal cortex non-coding transcriptome in AD, we performed two types of RNA sequencing on brain tissue from seven patients (**Table 4.2**). The patients were all non-Hispanic Caucasian females matched for age-of-death (mean 79 years vs. 79.25 years). The first set of libraries was generated by depleting ribosomal RNA (rRNA) from total RNA, allowing us to assay the expression of long non-coding and all coding transcripts independent of the presence of polyadenylated tails. The second set of libraries was generated by selecting for small RNAs (smRNA) by size fractionation, and these libraries were expected to elucidate the expression of, among other transcripts, short regulatory RNAs known as microRNAs.

Table 4.2 – Summary of samples and RNA-seq data processing

			rRNA(-) reads (millions)				**smRNA reads (millions)**		
Dx	ApoE	Age	Raw	Filtered	Mapped	Unique	Raw	Trimmed	Mapped
AD	ε3/ε4	81	83.6	38.3	24.1	14.1	32.5	21.3	17.3
AD	ε3/ε3	83	72.1	33.6	18.3	7.9	31.6	17.8	14.2
AD	ε3/ε4	73	73.1	32.5	19.1	8.2	25.3	15.6	11.0
N	ε3/ε3	92	59.2	27.2	15.2	9.4	26.0	12.3	8.8
N	ε2/ε3	72	81.2	32.6	18.7	11.6	30.0	13.9	11.6

N	ε3/ε3	68	39.3	21.0	12.1	7.0	29.4	18.6	15.2
N	ε3/ε3	85	66.5	27.7	14.1	8.8	18.7	7.1	6.1

4.3.2. Global changes in non-rRNA transcription in the AD brain

The rRNA(-) libraries, as expected, were dominated by mRNAs – first exonic reads at 30% of the total, followed by intronic reads at 20% (**Figure 4.1**). Despite introns being far longer than exons, they were underrepresented in total RNA due to splicing; the abundance of mature mRNA versus pre-mRNA transcripts resulted in higher coverage of exons. The next most sequenced class of RNAs was mitochondrial tRNAs, containing 10-20% of the reads. Strikingly, there was a great degree of variability in total mt-tRNA abundance between the samples – this variability is not seen for any of the other types of transcripts, including nuclear tRNAs. Broadly, the data suggest that mt-tRNAs are slightly less abundant in the AD samples, although this difference is not statistically significant. This depletion of mt-tRNA transcripts in the AD samples is potentially an indicator of cell death in the AD-affected cells.

Following mt-tRNA transcripts in abundance were sequences corresponding to transposable elements. The data suggest that transposable elements are widely expressed in this region of the brain, but there did not seem to be a difference between the AD and normal samples. The next most-sequenced class of reads consisted of those with no known annotation, which we call “intergenic.” While 10% of these regions are pseudogenes, it is likely that a large fraction of these reads are spurious mappings due to SNPs, sequencing errors, splicing, RNA editing, and cryptic exons or introns confounding the alignment. After intergenic reads the next most abundant classes were sno- and sn- (small nucleolar and small nuclear) RNAs. Approximately 10% of reads fell on the strand opposite to the annotated RNA and were labeled as “antisense.” The class of RNAs showing the most antisense transcription was transposable elements (**Figure 4.2**). This antisense transcription could be indicative of antisense RNA-mediated silencing of transposable elements in the brain transcriptome.

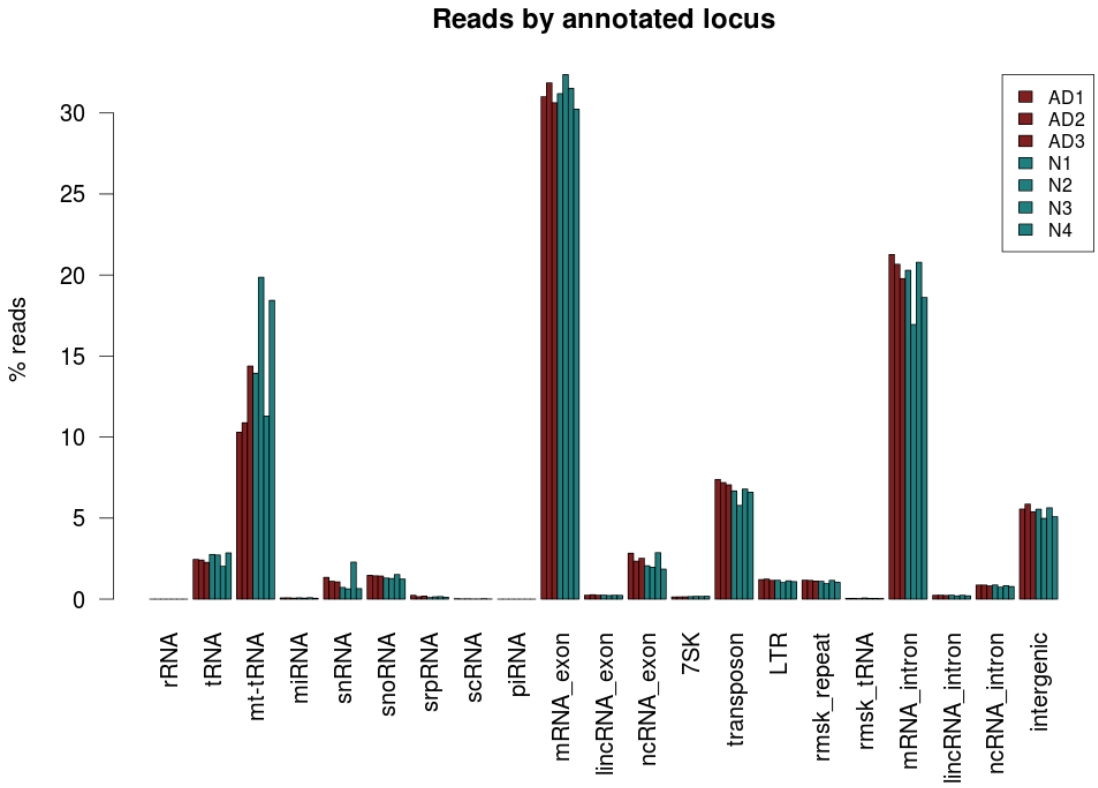

Figure 4.1- Summary of sequenced RNAs in the rRNA(-) libraries

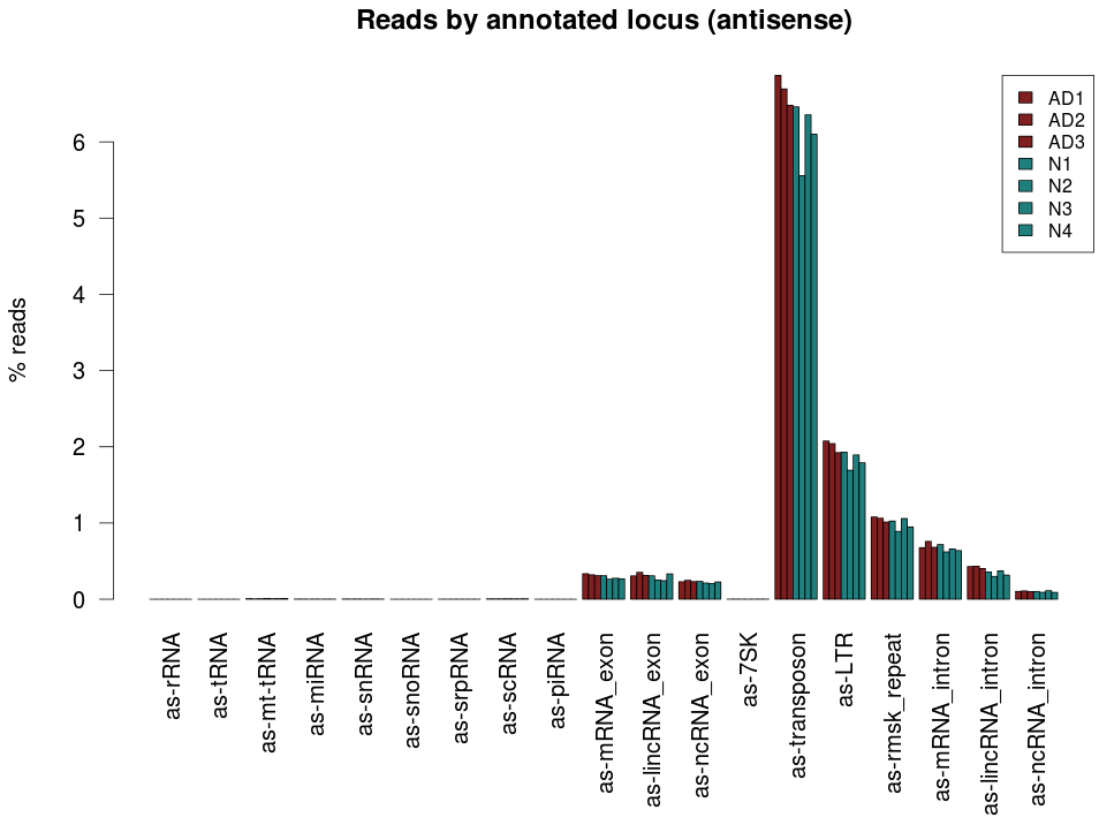

Figure 4.2 – Summary of antisense transcription in the rRNA(-) libraries

4.3.3. Global changes in small RNA biogenesis in the AD brain

In contrast to the rRNA(-) libraries, the small RNA (smRNA) libraries were largely composed of microRNAs (**Figure 4.3**), which accounted for 60-70% of the reads. The next most abundant class of smRNAs were tRNA-derived fragments (tRFs), containing 10-20% of the reads. There was a very striking pattern in global read counts: miRNAs were overall depleted in the AD brain while tRFs were overall enriched. Since the next largest class, snoRNA-derived smRNAs, did not differ between the conditions, the data do not suggest that the apparent drop in overall miRNA abundance was a simple consequence of an increase in tRF abundance or that the apparent increase in tRF abundance was an artifact resulting from a drop in miRNA abundance. Interestingly, despite the difference in tRF abundance, the rRNA(-) libraries showed

no change in full-length tRNA transcript levels. This suggests a difference in the activity of the tRF biogenesis processes (e.g., tRNA cleavage) in the AD state. Increased tRNA cleavage is known to be an indicator of cellular stress response in eukaryotes [28,61,151]. By contrast, the inverse relationship was seen for mt-tRNAs: the full-length transcripts were depleted in AD, but there was no apparent difference in fragments derived from mt-tRNAs. This suggests that nuclear and mitochondrial tRNAs are perturbed by distinct mechanisms in the AD state.

In contrast to the rRNA(-) libraries, only a very small proportion of the smRNA reads derived from the opposite strand to the annotated one (i.e., are antisense) (**Figure 4.4**). In general, however, the trend of the antisense smRNA reads is skewed towards repetitive elements similar to the rRNA(-) reads.

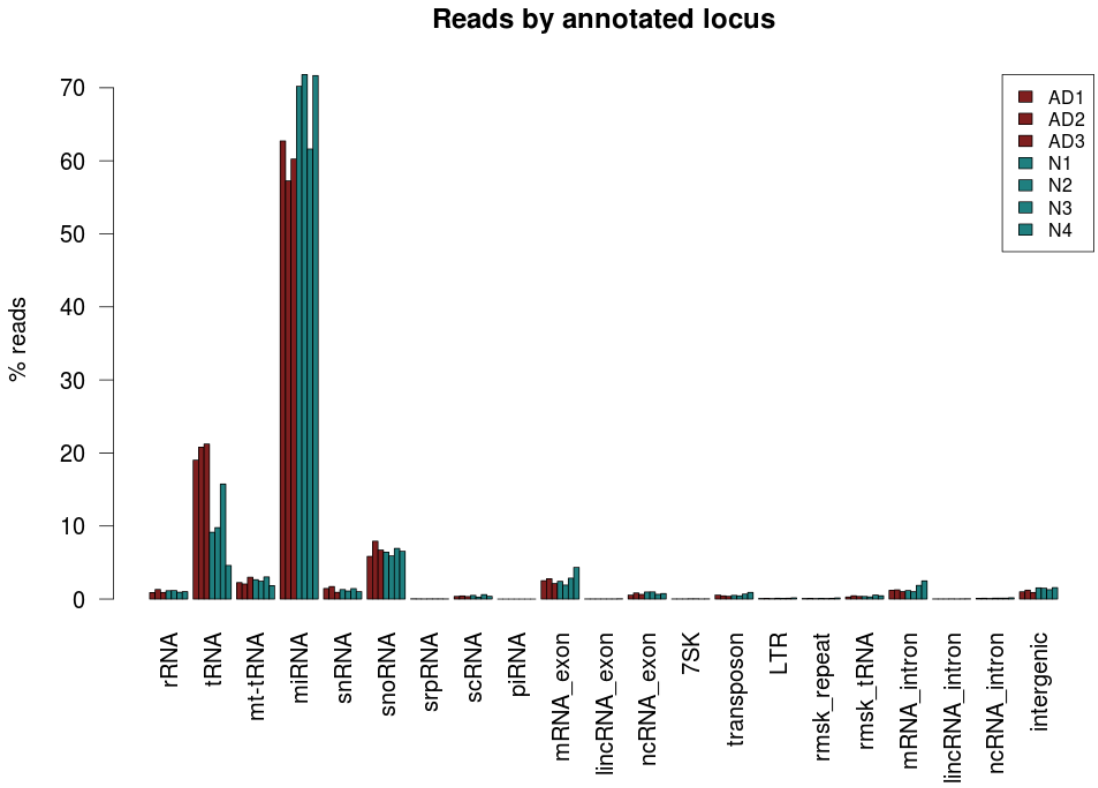

Figure 4.3 – Summary of sequenced RNAs in the smRNA-seq libraries

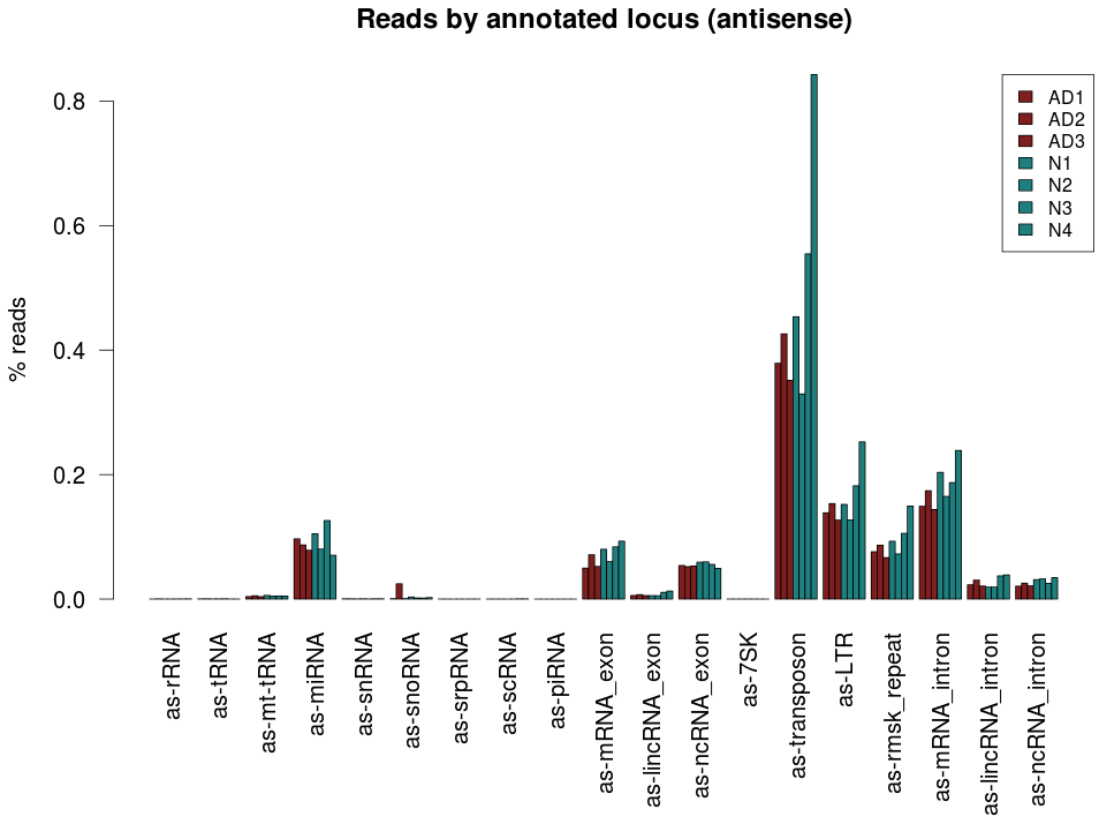

Figure 4.4 – Summary of antisense transcription in the smRNA libraries

Differentially expressed transcripts in the AD brain

Within the rRNA(-) libraries we found 215 differentially expressed RNAs, 62 of which were non-coding. The most commonly differentially expressed non-coding RNAs were tRNAs, followed by lincRNAs and snoRNAs (**Figure 4.5**). Interestingly, while the total number of reads mapping to tRNAs wasn't significantly different between the AD and normal groups, there were still many tRNA transcripts that were differentially expressed. We also detected 6 downregulated transcripts that were transcribed in a head-to-head fashion adjacent to protein-coding mRNAs (suggesting a bidirectional promoter), which in the literature have been termed promoter-associated RNAs (paRNAs) [40,74,149].

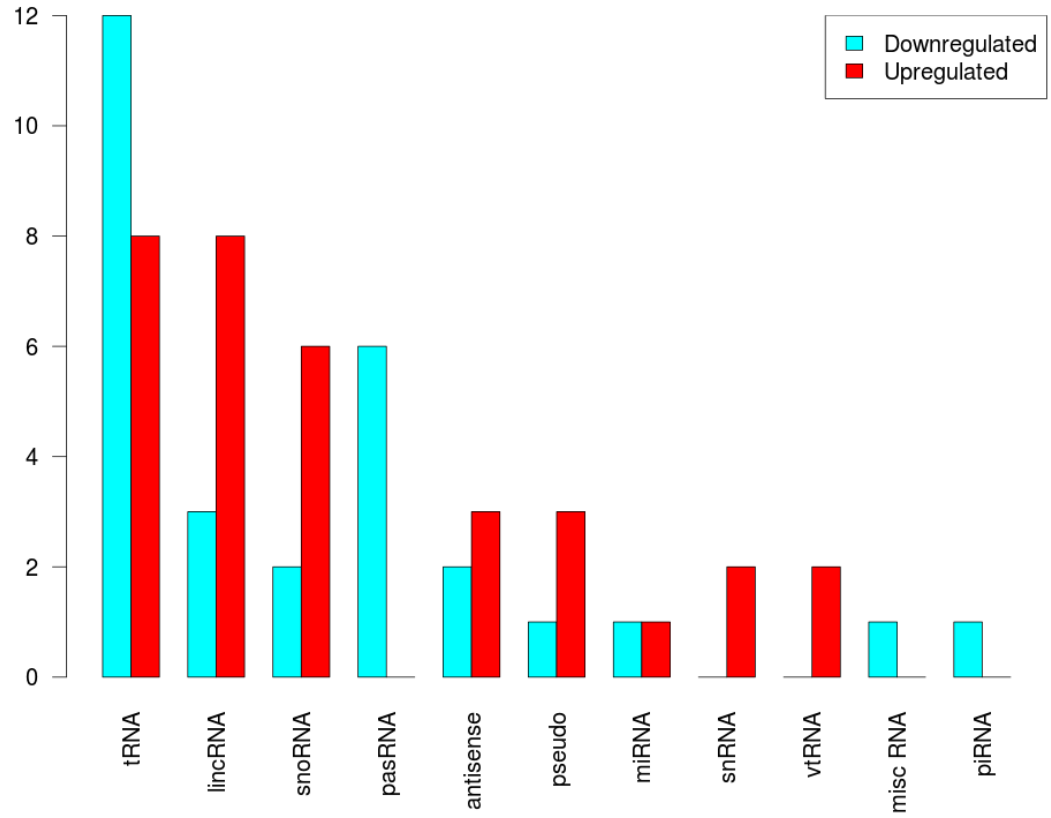

Figure 4.5 – Number of differentially expressed ncRNA transcripts by RNA class

Of the top 10 downregulated genes by significance, there were 5 mRNAs, 2 snoRNAs, 2 tRNAs, and a ncRNA that is antisense to the gene STXBP5 (**Table 4.3**). Included among the top 10 downregulated genes was ADCYAP1/PACAP; this gene has been shown to be neuroprotective in mouse models of AD via its upregulation of the alpha-secretase pathway [89]. Another top 10 gene, FAM190A, is a top GWAS hit for attention deficit disorder [98]. Also on the list was PUM2, a gene whose deficiency causes nesting behavior abnormalities in mice, a feature shared with mouse models of APP [54,136].The two tRNAs both code for cysteine: $tRNA^{Cys}_{GCA}$. One resides in a tandem cluster of $tRNA^{Cys}$ on chromosome 7 and the other lies in an intron of the gene CPNE4 on chromosome 3, near a tandem copy of itself.

The top downregulated snoRNA is the C/D box snoRNA SNORD79 – it lies in an intron of the GAS5 gene, an ncRNA thought to function as a host for snoRNAs [138]. SNORD79 is predicted to guide 2'O-ribose methylation of 28S rRNA at residue A3809; therefore, its downregulation could negatively impact rRNA function. Similarly, the other snoRNA on the top 10 downregulated list is the H/ACA box snoRNA SNORA36A, found on chromosome X in the intron of DKC1, a gene whose mutations can cause X-linked dyskeratosis congenita. Interestingly, DKC1 is itself a member of H/ACA box snoRNPs, and is also a member of telomerase. SNORA36A is predicted to guide pseudouridylation of 18S rRNA at residues U105 and U1244. Therefore its downregulation, like that of SNORD79, could negatively impact rRNA maturation. The ncRNA transcript STXBP5-AS is antisense to the gene STXBP5 which encodes for a protein thought to be involved in neurotransmitter release – mutations in the gene are associated with venous thrombosis. It is unclear what the relationship between antisense transcripts and the regulation of their sense-strand counterparts is.

Table 4.3 – Top 10 AD-downregulated transcripts in the rRNA(-) libraries.

Symbol	UCSC id	log2(fold change)	P-value	Description
ADCYAP1	uc010dkh.3	-0.70	1.5E-06	Homo sapiens adenylate cyclase activating polypeptide 1 (pituitary)
TRNA_Cys	uc021xee.1	-0.69	1.5E-06	tRNA Cys (anticodon GCA)
SNORD79	uc009wwk.1	-0.67	1.2E-07	small nucleolar RNA, C/D box 79
TRNA_Cys	uc022aox.1	-0.66	7.4E-06	tRNA Cys (anticodon GCA)
PDCD10	uc003fez.3	-0.60	2.5E-05	Programmed cell death 10
SNORA36A	uc004fmn.3	-0.58	4.4E-05	Small nucleolar RNA, H/ACA box 36A
STXBP5-AS	uc003qls.2	-0.56	1.7E-05	ncRNA antisense to STXBP5
LINC00086	uc004eyv.4	-0.55	7.7E-05	lincRNA
ANKRD30BL	uc002tti.3	-0.55	1.8E-04	Ankyrin repeat domain 30B-like ncRNA
JA040723	uc022bqt.1	-0.53	1.2E-04	piRNA piR-31490

In addition to the downregulated transcripts, there were also many upregulated transcripts (**Table 4.4**). Among the top 10 upregulated RNAs were 3 tRNAs, 6 mRNAs, and 1 uncharacterized lincRNA. The tRNAs consisted of two serine-charged tRNAs ($tRNA^{Ser}_{CGA}$,

$tRNA^{Ser}_{GCU}$) and one $tRNA^{Glu}_{UUC}$. One of the $tRNA^{Ser}$ loci lies in the promoter of B3GNT1 and the other lies just downstream of the circadian-rhythm-related gene PER1. Included among the mRNAs is that of MLXIPL, a candidate gene for genetic association with plasma triglyceride levels [90]. This is especially intriguing since defects in ApoE are associated with plasma triglyceride levels. Another top upregulated gene is CNST (cosortin) which interacts with connexins, a class of proteins that could be involved in AD pathogenesis pathways [92]. Yet another top upregulated gene, ZMYM5, is known to repress the transcription of PSEN1, one of the genes whose familial variants can cause AD [125]. Interestingly, two components of the minor spliceosome (U6atac and U12 snRNAs) are both significantly upregulated in the AD brain. The minor spliceosome responds positively to stress [167], and so the cell stress-related component of AD may play a role in its upregulation. One of the genes whose expression is modulated by the minor splicesome is PTEN, which can be found in NFTs in AD [141].

Table 4.4 – Top 10 AD-upregulated transcripts in the rRNA(-) libraries

Symbol	UCSC id	log2(fold change)	P-value	Description
TRNA_Ser	uc021tps.1	0.85	2.4E-12	tRNA Ser (anticodon CGA)
TRNA_Ser	uc021qlw.1	0.78	4.5E-09	tRNA Ser (anticodon GCT)
TRNA_Glu	uc021vol.1	0.63	1.7E-06	tRNA Glu (anticodon TTC)
TPD52L1	uc003pzu.1	0.55	8.2E-06	Tumor protein D52-like 1
AK054921	uc004fbf.1	0.63	7.7E-06	Highly similar to 40S RIBOSOMAL PROTEIN S15A.
MLXIPL	uc003tyn.1	0.66	7.1E-06	MLX interacting protein-like
SPAG5	uc002hbq.3	0.55	2.4E-05	Sperm associated antigen 5
TCONS_00016137	N/A	0.59	5.7E-05	N/A
CNST	uc001ibp.3	0.41	6.8E-05	Homo sapiens consortin, connexin sorting protein
ZMYM5	uc010tcn.1	0.55	9.3E-05	Homo sapiens zinc finger, MYM-type 5 (ZMYM5)

4.3.4. Differentially expressed small RNAs in the AD brain

There were a total of 456 small RNA loci differentially expressed in the human prefrontal cortex (Figure 4.6). The largest class of D.E. smRNAs was the intergenic class, which consists of those small RNA loci that don't overlap any annotated RNA whatsoever. It is likely that a large fraction of these are the result of cross-mapping errors whether from SNP-induced mismatches or extremely high-abundance RNAs combined with the normal rate of base-calling error.

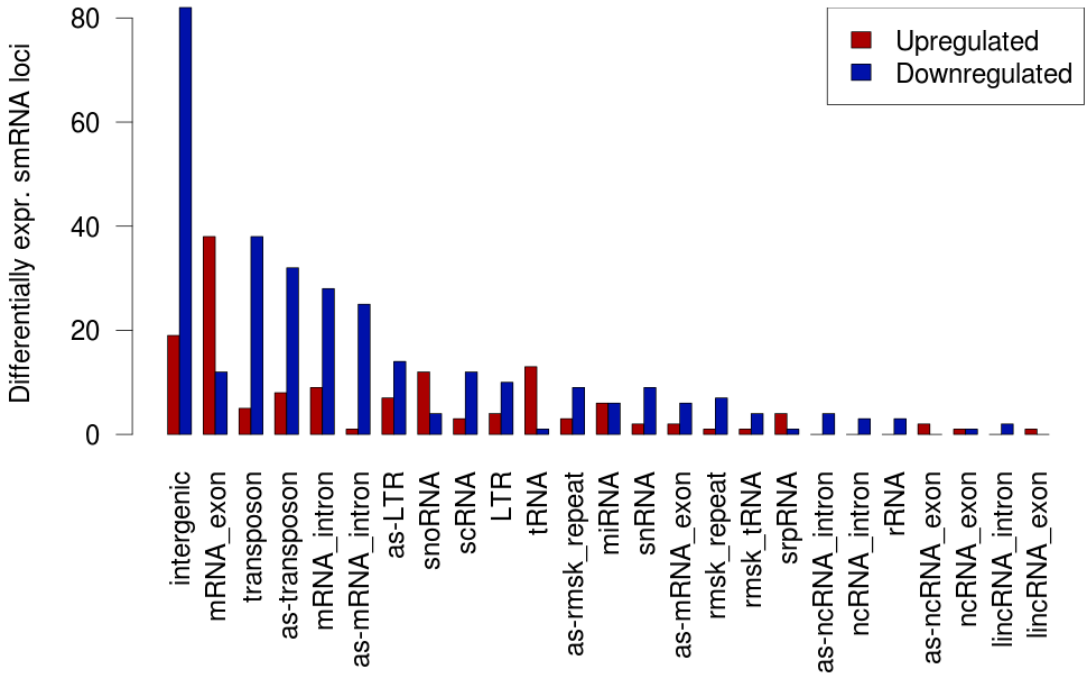

Figure 4.6 - Number of differentially expressed smRNA loci by ncRNA class.

The next largest class was small RNAs deriving from mRNAs. It is unknown whether these are specifically processed from double-stranded structures on mRNAs or are simple mRNA fragments resulting from exonuclease activity on degrading transcripts. Since 60% of these mRNA-derived smRNA loci had lengths shorter than 50nt, it is likely that the class is a mix of the two: small, well defined loci and larger loci that are home to a larger number of more varied cleavage events. This proportion (60%) was not different between upregulated and downregulated mRNA-derived smRNA loci. Interesting members of the list include APOD, genetic variants in which are associated with AD in Chinese and Japanese populations [29,135]. GLRX2, glutaredoxin 2, is a mitochondrial protein involved in protection against oxidative stress in mitochondria. SCD5, involved in fatty acid metabolism [7], has been shown to be upregulated in

the AD brain. SPP1, or osteopontin, is a gene whose protein product is a biomarker for mild cognitive impairment [147] and AD. SQSTM1 has been shown to be overexpressed in the AD brain, and it also associates with the neurofibrillary tangles (NFTs) characteristic of AD [121]. CRYAB (Alpha-b crystallin) associates with NFTs as well [109]. CLSTN1 (calsyntenin I) has been shown to regulate amyloid beta production [157].

Table 4.5 – Differentially expressed small RNAs derived from mRNAs or antisense transcripts

Type	Gene	Locus	log2(FC)	FDR
mRNA exon	CLDN12	chr7:90044707-90044724(+)	-1.91	0.02
Antisense mRNA	ENTPD4	chr8:23315089-23315113(+)	1.89	0.03
Antisense mRNA	ADAD1	chr4:123332506-123332525(+)	-2.08	0.04
mRNA exon	KIF5C	chr2:149829885-149829908(+)	1.76	0.05
mRNA exon	UHMK1	chr1:162470009-162470034(+)	1.95	0.06
mRNA exon	TOB1	chr17:48943684-48943732(-)	-1.64	0.07
Antisense mRNA	BEND6	chr6:56820040-56820056(-)	-1.47	0.07
mRNA exon	MAP2	chr2:210543331-210543381(+)	1.41	0.08
mRNA exon	KCNMA1	chr10:78629571-78629589(-)	1.86	0.08
mRNA exon	NEFH	chr22:29885590-29885903(+)	1.71	0.08
Antisense mRNA	C1GALT1C1	chrX:119760650-119760667(-)	-1.28	0.08
mRNA exon	GLRX2	chr1:193074487-193074511(-)	-1.59	0.10
mRNA exon	MEF2C	chr5:88014477-88014502(-)	1.16	0.10
mRNA exon	NEFH	chr22:29885387-29885507(+)	1.53	0.10
Antisense mRNA	ATM	chr11:108161209-108161226(+)	1.20	0.11
mRNA exon	SCD5	chr4:83550834-83550935(-)	1.51	0.12
Antisense mRNA	CD163	chr12:7651550-7651659(-)	1.61	0.12
mRNA exon	SPP1	chr4:88903663-88904132(+)	1.46	0.12
mRNA exon	TMEM123	chr11:102268541-102268559(-)	1.20	0.12
mRNA exon	WDR82	chr3:52290206-52290229(-)	1.29	0.12
Antisense mRNA	B2M	chr15:45007645-45007842(+)	1.28	0.12
mRNA exon	KCTD16	chr5:143586444-143586466(+)	1.38	0.12
mRNA exon	SQSTM1	chr5:179264066-179264095(+)	1.23	0.12
mRNA exon	QDPR	chr4:17503379-17503480(-)	1.23	0.12
mRNA exon	MAN1A2	chr1:117957358-117957443(+)	1.28	0.13
mRNA exon	EIF4G3	chr1:21268544-21268587(-)	1.22	0.14
mRNA exon	SLC17A7	chr19:49936063-49936090(-)	1.20	0.14
mRNA exon	PDZD2	chr5:32087421-32087437(+)	1.31	0.15
mRNA exon	SON	chr21:34949108-34949132(+)	1.25	0.15

Antisense mRNA	SEC16B	chr1:177929373-177929391(+)	-1.38	0.15
mRNA exon	GLUL	chr1:182353819-182353853(-)	1.37	0.16
Antisense mRNA	TTLL6	chr17:46840131-46840146(+)	-1.17	0.16
mRNA exon	CRYAB	chr11:111779396-111779441(-)	1.11	0.16
mRNA exon	KDR	chr4:55945467-55945483(-)	1.22	0.16
mRNA exon	NHP2	chr5:177576651-177576675(-)	1.07	0.16
mRNA exon	DLG2	chr11:83170812-83170846(-)	1.19	0.17
mRNA exon	ENC1	chr5:73923655-73923709(-)	-1.35	0.17
mRNA exon	CNTN1	chr12:41331379-41331447(+)	-1.34	0.17
Antisense mRNA	ICA1	chr7:8167738-8167756(+)	1.19	0.17
mRNA exon	CPE	chr4:166408644-166408696(+)	-1.17	0.17
mRNA exon	NECAP1	chr12:8242814-8242849(+)	-1.36	0.18
mRNA exon	LAP3	chr4:17609300-17609343(+)	1.32	0.18
mRNA exon	DNAJC6	chr1:65880026-65880103(+)	1.15	0.18
mRNA exon	FTH1	chr11:61735039-61735097(-)	1.24	0.21
mRNA exon	PSD3	chr8:18393376-18393440(-)	1.07	0.21
mRNA exon	SYNE1	chr6:152647191-152647214(-)	1.19	0.21
Antisense mRNA	ZMYM5	chr13:20425915-20425933(+)	-1.25	0.21
Antisense mRNA	NIPAL3	chr1:24746024-24746043(-)	-1.20	0.22
mRNA exon	RPL30	chr8:99057206-99057270(-)	0.95	0.22
mRNA exon	GNAS	chr20:57478778-57478839(+)	-1.02	0.22
mRNA exon	SRCIN1	chr17:36708098-36708114(-)	1.03	0.22
mRNA exon	MAP1B	chr5:71493868-71493952(+)	1.02	0.22
mRNA exon	OSGIN1	chr16:83994344-83994359(+)	1.23	0.22
mRNA exon	CLSTN1	chr1:9811636-9811688(-)	0.76	0.22
mRNA exon	SPOCK2	chr10:73827391-73827428(-)	-1.18	0.22
Antisense mRNA	WRAP73	chr1:3564083-3564098(+)	-1.00	0.23
Antisense mRNA	APOD	chr3:195300738-195300842(-)	0.93	0.25
mRNA exon	CPE	chr4:166418676-166418744(+)	-1.14	0.25

Similarly to the mRNA-derived smRNAs, snoRNA-derived smRNAs tended to be upregulated rather than downregulated in the AD-affected samples (**Table 4.6**). Interestingly, these transcripts largely did not overlap with the transcripts that were differentially expressed in the rRNA(-) libraries. The only snoRNA showing changes in both libraries was SNORA18, with both the full-length transcripts and the small RNA products being upregulated in AD. SNORA18 is an “orphan” snoRNA – that is, its target RNA is unknown.

Table 4.6 – Differentially expressed snoRNAs in rRNA(-) and smRNA libraries

Transcript	Log2(fold change)		Predicted target
	rRNA(-)	smRNA	
SCARNA16		0.92	Pseudouridylation of U1 snRNA U5
SNORA18	0.43	1.1	*Unknown*
SNORA26	0.39		2'O-ribose methylation of 28S rRNA A389
SNORA31		-1.44	Pseudouridylation of 28S rRNA U3713 & 18S U218
SNORA36A	-0.58		Pseudouridylation of 18S rRNA U105, U1244
SNORA53		0.91	Unknown
SNORA5A		0.91	Pseudouridylation of 18S rRNA U1625, U1238
SNORA65		0.84	Pseudouridylation of 28S rRNA U4373, U4427
SNORA68	0.35		Pseudouridylation of 28S rRNA U4393
SNORA77	0.45		Pseudouridylation of 18S rRNA U814
SNORD11		0.93	2'O-ribose methylation of 18S rRNA G509
SNORD115-39	0.42		5HT-2C mRNA
SNORD115-40	0.43		5HT-2C mRNA
SNORD115-48		-0.93	5HT-2C mRNA
SNORD117		0.64	*Unknown*
SNORD121B		0.89	2'O-ribose methylation of 28S rRNA G4607
SNORD15B		-1.15	2'O-ribose methylation of 28S rRNA A3764
SNORD36B		0.64	2'O-ribose methylation of 18S rRNA A668
SNORD36C		0.67	2'O-ribose methylation of 28S rRNA A3703
SNORD4B		-1.02	2'O-ribose methylation of 18S rRNA U121
SNORD60		0.91	2'O-ribose methylation of 28S rRNA G4340
SNORD73A		0.69	2'O-ribose methylation of 28S rRNA G1747
SNORD79	-0.67		2'O-ribose methylation of 28S rRNA A3809

Several microRNAs were differentially expressed in the smRNA libraries (Table 4.7). Downregulation of mir-132 has been shown to be implicated in derepression of FOXO3 and thus promote neuronal apoptosis in AD [160]. The totality of predicted targets of these miRs is given in **Table 4.8**.

Table 4.7 – Differentially expressed microRNAs.

microRNA	log2(Fold change)	False discovery rate
mir-412	-1.71	0.064
mir-886	1.41	0.102
mir-4326	1.50	0.115
mir-381	0.78	0.116
mir-889	0.85	0.121
mir-877	-0.88	0.159
mir-96	-1.05	0.162
mir-132	-0.76	0.162
mir-95	0.65	0.178
mir-26a-2	0.58	0.211
mir-556	-0.99	0.214
mir-425	-0.57	0.223

Table 4.8 – Experimentally validated targets of the D.E. miRNAs.

microRNA	Target gene	Description
miR-132	ARHGAP32	Rho GTPase-activating protein
miR-132	CDKN1A	cyclin-dependent kinase inhibitor 1A (p21, Cip1)
miR-132	RFX4	regulatory factor X, 4 (influences HLA class II expression)
miR-132	SIRT1	sirtuin (silent mating type information regulation 2 homolog)
mir-26a	CHFR	checkpoint with forkhead and ring finger domains
mir-26a	EZH2	enhancer of zeste homolog 2 (Drosophila)
mir-26a	PLAG1	pleiomorphic adenoma gene 1
mir-26a	PTP4A1	protein tyrosine phosphatase type IVA, member 1
mir-26a	SMAD1	SMAD family member 1
mir-26a	STRADB	STE20-related kinase adapter protein beta
mir-26a	TAF12	TAF12 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 20kDa
mir-26a	TGFBR2	transforming growth factor, beta receptor II (70/80kDa)
miR-381	LRRC4	leucine rich repeat containing 4
miR-412	ACVR1C	activin A receptor, type IC
miR-877	EFNA5	ephrin-A5
miR-877	ELF1	E74-like factor 1 (ets domain transcription factor)
miR-877	FXR2	fragile X mental retardation, autosomal homolog 2
miR-877	SCN3A	sodium channel, voltage-gated, type III, alpha subunit
miR-877	SMG5	Smg-5 homolog, nonsense mediated mRNA decay
miR-877	TP53INP2	tumor protein p53 inducible nuclear protein 2
miR-95	SNX1	sorting nexin 1
miR-96	ADCY6	adenylate cyclase 6
miR-96	AQP5	aquaporin 5
miR-96	CDKN1A	cyclin-dependent kinase inhibitor 1A (p21, Cip1)
miR-96	CELSR2	cadherin, EGF LAG seven-pass G-type receptor 2 (flamingo homolog, Drosophila)
miR-96	FOXO1	forkhead box O1
miR-96	FOXO3	forkhead box O3; forkhead box O3B pseudogene
miR-96	HTR1B	5-hydroxytryptamine (serotonin) receptor 1B
miR-96	IRS1	insulin receptor substrate 1
miR-96	KRAS	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog
miR-96	MITF	microphthalmia-associated transcription factor
miR-96	MYRIP	myosin VIIA and Rab interacting protein
miR-96	NR3C1	nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor)
miR-96	ODF2	outer dense fiber of sperm tails 2
miR-96	PRMT5	protein arginine methyltransferase 5

miR-96	RYK	RYK receptor-like tyrosine kinase

4.3.5. tRNAs are differentially expressed and processed in the AD brain

Several tRNAs, both nuclear and mitochondrial, were differentially expressed in the AD brain. In addition, we found that tRNAs also tended to be differentially processed – that is, the cleavage of tRNAs into small RNAs known as tRNA fragments, or tRFs, changes in the AD brain (Table 4.9, Table 4.10). Differential expression in the rRNA(-) libraries indicates differences in the abundance of the long tRNA transcript, whereas differential expression in the smRNA libraries indicates differential processing of the tRNA into tRNA fragments. With the exception of one tRF deriving from a pseudogenic tRNA on chr1, all the tRNA-associated small RNAs were strongly upregulated in AD. This potentially corresponds to a decrease in the activity and/or increase in cleavage of these tRNAs. In addition to the fragments, many tRNA transcripts themselves were downregulated, including most prominently $tRNA^{Cys}_{GCA}$ and the mitochondrial $mt\text{-}tRNA^{Pro}_{UGG.}$ Only in one case did we observe changes in both the tRNA transcript and its fragments: $tRNA^{Lys}_{UUU}$ shows upregulation of both. Interestingly, tRNA-associated small RNAs, mRNA-associated small RNAs, and snoRNA-associated small RNAs were the only overall-upregulated classes of small RNAs in AD.

Table 4.9 – Downregulated tRNAs and tRNA fragments in the AD brain with expression fold-changes.

Transcript	Locus	log2(FC)	p-value
tRF^{Ψ}_{GUC}	chr1:161492987(+)	-1.08	1.5E-02
$tRNA^{Cys}_{GCA}$	chr3:131947943(-)	-0.67	1.5E-06
$tRNA^{Cys}_{GCA}$	chr7:149007280(+)	-0.64	7.2E-06
mt-$tRNA^{Pro}_{UGG}$	chrM:15954(-)	-0.50	1.4E-04
$tRNA^{Arg}_{CCG}$	chr6:28849164(+)	-0.49	3.8E-04
$tRNA^{Tyr}_{GUA}$	chr14:21121257(-)	-0.45	1.2E-03
$tRNA^{Val}_{AAC}$	chr6:27618706(-)	-0.45	1.4E-03
$tRNA^{Arg}_{UCG}$	chr6:26299904(+)	-0.44	1.8E-03
$tRNA^{Tyr}_{GUA}$	chr14:21131350(-)	-0.43	2.2E-03
$tRNA^{Arg}_{CCU}$	chr17:73030525(-)	-0.43	1.9E-03
$tRNA^{Arg}_{UCG}$	chr6:26323045(+)	-0.39	4.1E-03
$tRNA^{Cys}_{GCA}$	chr15:80036996(+)	-0.39	5.0E-03
$tRNA^{Val}_{CAC}$	chr6:27248048(-)	-0.32	3.6E-03

Table 4.10 – Upregulated tRNAs and tRNA fragments in the AD brain with expression fold-changes.

Transcript	Locus	log2(FC)	p-value
tRF^{Thr}_{AGU}	chr17:8090478(+)	1.55	8.8E-04
tRF^{Ile}_{AAU}	chr6:27205343(-)	1.50	5.0E-03
tRF^{Lys}_{UUU}	chr17:8022468(+)	1.35	1.7E-03
tRF^{Ψ}_{UGC}	chr6:28601911(-)	1.32	2.2E-02
tRF^{Met}_{CAU}	chr8:124169459(-)	1.27	8.3E-03
tRF^{Asn}_{GUU}	chr1:148248113(+)	1.19	1.6E-02
tRF^{Ala}_{UGC}	chr11:50233925(-)	1.18	9.6E-03
tRF^{Thr}_{UGU}	chr6:28442320(-)	1.06	2.7E-03
tRF^{Ala}_{CGC}	chr2:157257280(+)	1.03	6.6E-03
tRF^{Phe}_{GAA}	chr12:125412379(-)	0.96	7.5E-03
tRF^{Ψ}_{AGG}	chr16:3202636(-)	0.86	2.0E-02
$tRNA^{Ser}_{CGA}$	chr17:8042198(-)	0.84	2.0E-12
tRF^{Gln}_{CUG}	chr15:66161389(-)	0.81	2.3E-02
$tRNA^{Ser}_{GCU}$	chr11:66115590(+)	0.78	2.2E-09
tRF^{Glu}_{UUC}	chr13:45491997(-)	0.76	2.8E-02
mt-$tRNA^{Asp}_{GUC}$	chrM:7516(+)	0.72	9.0E-08
$tRNA^{Glu}_{UUC}$	chr2:131094700(-)	0.63	1.4E-06
$tRNA^{Lys}_{UUU}$	chr17:8022472(+)	0.50	3.9E-04
$tRNA^{Lys}_{CUU}$	chr5:180634754(+)	0.44	2.0E-03
$tRNA^{Trp}_{CCA}$	chr6:26331671(-)	0.39	4.6E-03
$tRNA^{Asp}_{GUC}$	chr12:96429798(+)	0.39	2.5E-03
$tRNA^{Ala}_{UGC}$	chr12:125424511(+)	0.37	3.5E-03

While it is unclear what the global effects of these changes in tRNA expression and processing are, we can start by predicting their effect on the translation of specific mRNAs into proteins by finding those mRNAs which, by virtue of their coding sequences, would be most affected if the tRNAs with corresponding anticodons were perturbed. In order to separate the sets of anticodons into those whose associated tRNAs decrease in activity and those whose tRNAs increase in activity, we make a few assumptions about the nature of the expression changes in tRNAs and their associated fragments. We assume that increase of the cleavage process corresponds to an inhibition of tRNA activity, and so can combine the upregulated tRFs with the downregulated tRNAs into a set of tRNAs, and thus anticodons, whose activity is downregulated. Conversely, those tRNAs that are upregulated or whose concomitant fragments are downregulated can be presumed to be increasing in activity. Given the sets of anticodons that are uptranslated and those that are downtranslated, we can score each gene by taking the proportion of its codons that are complementary to any of the anticodons in each of the two lists. Furthermore, we limit the set of genes to those having at least one count in the brain whole transcriptome RNA-seq dataset (Table 4.11, Table 4.12.).

Table 4.11 – Top 10 brain-expressed genes predicted to be down-translated due to tRNA changes.

Gene	Description	Anticodon score	Num. codons
SRP14	Signal recognition particle 14kDa	0.198	171
PRKCG	Protein kinase C, gamma	0.190	58
VOPP1	Vesicular, overexpressed in cancer, prosurvival protein 1	0.185	65
TMEM123	Transmembrane protein 123	0.182	66
EPN2	Epsin 2	0.180	61
MLH1	MutL homolog 1	0.179	117
FAM3A	Family with sequence similarity 3, member A	0.175	63
ST3GAL5	ST3 beta-galactoside alpha-2,3-sialyltransferase 5	0.174	144
GALNT13	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 13	0.172	58
LRRCC1	Leucine rich repeat and coiled-coil domain containing 1	0.171	111

Table 4.12 - Top 10 brain-expressed genes predicted to be up-translated due to tRNA changes.

Gene	Description	Anticodon score	Num. codons
RPS25	Ribosomal protein S25	0.183	120
SRP14	Signal recognition particle 14kDa	0.164	171
RERE	Arginine-glutamic acid dipeptide (RE) repeats	0.162	105
RAI14	Retinoic acid induced 14	0.159	63
HMGN5	High mobility group nucleosome binding domain 5	0.148	603
SREK1IP1	SREK1-interacting protein 1	0.145	468
RSRC2	Arginine/serine-rich coiled-coil 2	0.143	77
CCDC91	Coiled-coil domain containing 91	0.142	380
ARGLU1	Arginine and glutamate rich 1	0.141	206
DMD	Dystrophin (DMD)	0.138	123

In order to assess the biological significance of these sets of genes, we look for enrichment of particular biological pathways in the top 1,000 genes by codon-score. Strikingly, the list of putatively down-translated genes is enriched for KEGG pathways that correspond to several neurodegenerative disorders, including Alzheimer's disease (Table 4.13). While this enrichment is largely a result of the presence in the list of mitochondrial Complex I genes, the list also includes several genes thought to be highly significant in AD pathogenesis: ADAM10 (alpha secretase), BACE1 (beta-secretase), SNCA (synuclein alpha), TF (transferrin), and PICALM.

Table 4.13 – KEGG pathways enriched for putative down-translated genes

KEGG Pathway	Adjusted p-value
Ribosome	0.0002
Oxidative phosphorylation	0.0017
Huntington's disease	0.0017
Alzheimer's disease	0.0062
Vasopressin-regulated water reabsorption	0.0062
Parkinson's disease	0.0102
Glycosphingolipid biosynthesis - ganglio series	0.0103
Glycosaminoglycan biosynthesis - chondroitin sulfate	0.0103
Epithelial cell signaling in Helicobacter pylori infection	0.0103
Lysosome	0.0127

4.3.6. Functional characterization of the differentially expressed transcripts

We combined the rRNA(-) and smRNA libraries into a list of genes based on differential expression of transcripts, of smRNA loci, and of microRNAs (by including their targets). microRNA target predictions were taken from three sources. The first is StarBase [165], which uses *Argonaute* CLIP-seq data to locate candidate miRNA target sites and then uses sequence-based prediction tools to associate microRNAs. We also included all experimentally validated predictions compiled within the mirTarBase and miRecord databases [80,164]. When a microRNA went up in expression in AD, we considered its targets as going down (repressed) and vice versa. We also included genes associated with ncRNAs found to be D.E. in the rRNA(-) libraries; for example, if an antisense transcript was D.E. then we included the sense transcript in the list. Notable observations relevant to AD pathology include downregulation of genes involved in cytoskeleton organization (DLC1, LIMA1, CEP76, RAC1, MAP4, TMSB4X, RICTOR, DST, CTNNB1) and nuclear localization (KPNA3, KPNB1, TNPO1). Notable among the upregulated categories are TGF-beta signaling genes (MAPK1, ACVR2B, E2F5, SMAD7, PPP2CB, SMAD5, RBL1, SMAD2, THBS3, ACVR1C, ACVR1), dysfunction of which has been implicated in AD [39]. In addition, apoptotic genes (MEF2C, PRKCZ, TMX1, ZAK, XIAP, STAT5B, MITF, CBX4, SOX4, FOXO1, EIF5A, FOXO3, STK17A, ITM2B, MST4, ACVR1C, PEA15, IGF1R, G2E3, MAP3K5, KRAS, SQSTM1, BCL11B, PPP2CB, RASA1, DHCR24, KCNMA1, SPPL3, BCL10, SGK3) are thought to be critical in AD pathogenesis and indeed appear to be upregulated in these data.

Table 4.14 - Top downregulated functional categories in AD.

Category	P-value
GO:0003714~transcription corepressor activity	1.8E-05
GO:0048598~embryonic morphogenesis	1.6E-04
hsa03040:Spliceosome	1.7E-02
GO:0001657~ureteric bud development	7.1E-03
GO:0016477~cell migration	3.3E-03
GO:0009967~positive regulation of signal transduction	5.9E-03
GO:0048568~embryonic organ development	4.8E-03
GO:0008139~nuclear localization sequence binding	9.4E-03
GO:0030522~intracellular receptor-mediated signaling pathway	6.1E-03
GO:0007049~cell cycle	1.2E-02
GO:0051493~regulation of cytoskeleton organization	1.0E-02
GO:0051056~regulation of small GTPase mediated signal transduction	1.0E-02
GO:0000059~protein import into nucleus, docking	5.7E-03
GO:0007584~response to nutrient	1.2E-02

Table 4.15 - Top upregulated functional categories in AD.

Category	P-value
GO:0016563~transcription activator activity	4.7E-03
GO:0030323~respiratory tube development	2.6E-05
GO:0042981~regulation of apoptosis	2.6E-06
GO:0019901~protein kinase binding	2.5E-03
GO:0045934~negative regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	1.1E-03
GO:0003714~transcription corepressor activity	6.8E-03
GO:0001944~vasculature development	6.8E-04
GO:0035295~tube development	3.5E-06
GO:0000226~microtubule cytoskeleton organization	8.1E-03
GO:0001784~phosphotyrosine binding	1.4E-02
GO:0016568~chromatin modification	4.4E-03
GO:0048598~embryonic morphogenesis	5.7E-03
hsa04350:TGF-beta signaling pathway	2.2E-04
GO:0004674~protein serine/threonine kinase activity	8.2E-04
GO:0001843~neural tube closure	9.2E-03

4.3.7. Building an integrative network

While expression data can inform us about associations between disease state and regulatory changes, their impact can be greatly increased by integrating such findings with genetics data. The recent influx in genetics results for Alzheimer's disease, along with burgeoning pathway/interaction datasets, allows us to link functional correlations with potentially causal factors in an unprecedented fashion. Here we present a network of genes that has been "seeded" by known genetic factors in AD, as well as the MAPT (tau) gene which is a key factor in the formation of the NFTs (**Figure 4.7**). It should be noted that while MAPT is not known to be a genetic factor in AD, AD is considered by many pathologists to be a *tauopathy* due to the hallmark involvement of NFTs. In addition, we have included all of the genes whose protein products are directly involved in the processing of APP: PSEN1, PSEN2, PSENEN, NCSTN, APH1A, APH1B (gamma-secretase), BACE1, BACE2 (beta-secretase), and ADAM10 (alpha-secretase). In addition to protein-protein interactions, we have also included microRNA-target interactions, the interaction between ZMYM5 and the promoter of PSEN1, interactions between gamma secretase components and APP, and interactions between SNORD115 loci and the HTR2C gene. MicroRNAs mir-132, mir-96, and mir-26a appear to be highly-connected within this network, as well as SIRT1, SDC4, SNRPF. SIRT1 is an aging-associated gene and has been found to be neuroprotective against Abeta toxicity and NFT formation in mouse brains [44]. SDC4 (syndecan-4) is involved in the healing process, and its downregulation in AD may indicate an impaired ability to recover from the damage inflicted during AD pathology [47].

Red elements were upregulated and blue elements were downregulated; triangles are miRNAs, circles are full transcripts, and squares are smRNA loci derived from mRNAs.

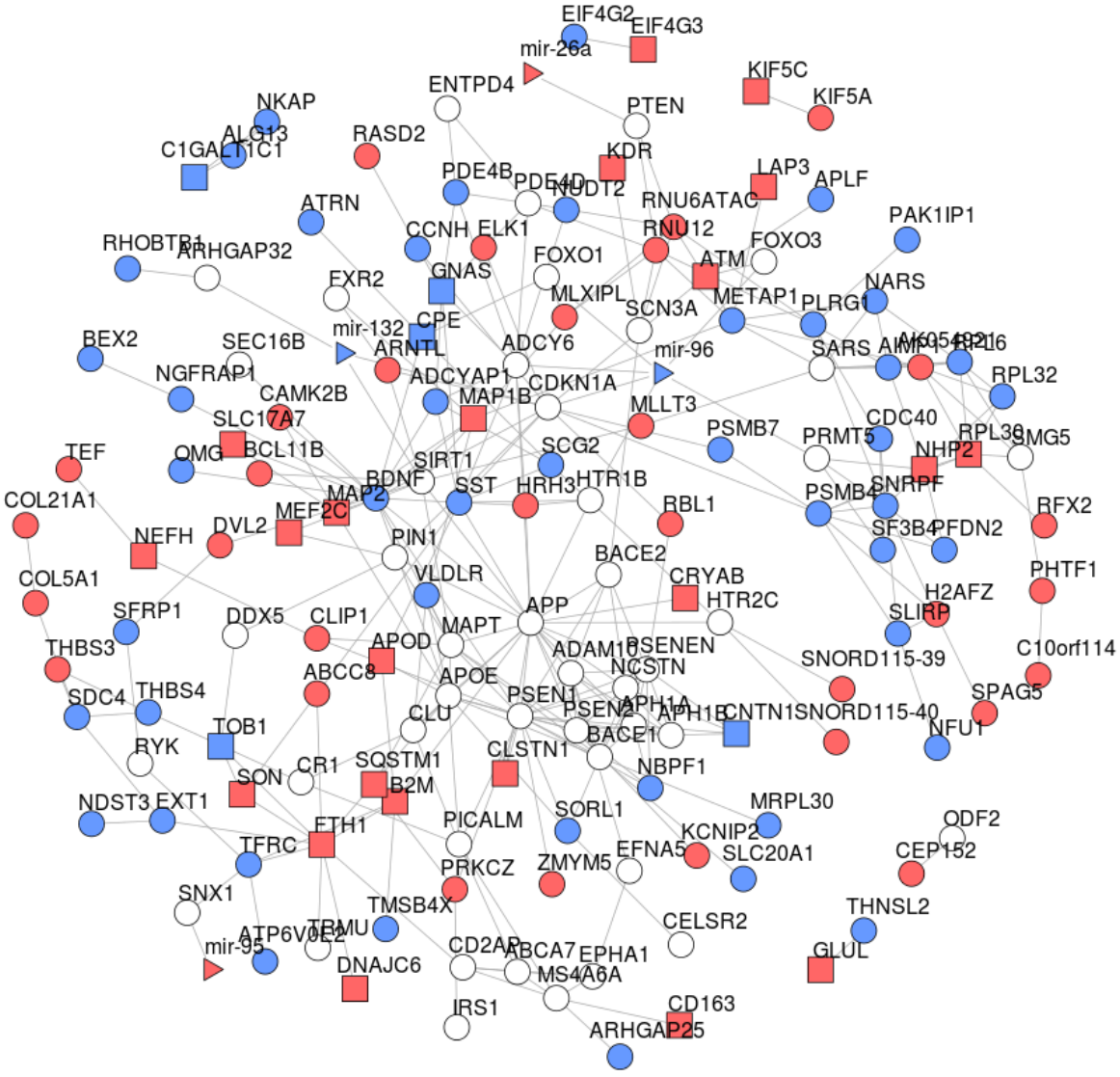

Figure 4.7 – An integrative network of AD.

4.4. Discussion

In the Alzheimer's disease state we observe several global changes in the expression of non-coding RNAs. The most striking change is that of tRNAs and a recently discovered class of RNAs comprised of tRNA fragments (tRFs). By comparing the anticodons associated with the tRNAs with the coding sequences of the human genome, we have presented a list of genes most likely to be translationally perturbed by the changes in tRNA expression. Enriched among this set of genes are Complex I members, implicating an association between tRNA dysregulation, mitochondrial function, and Alzheimer's disease. While the mitochondrial cascade hypothesis is a longstanding theory in the AD field [148], its connection to tRNAs remains largely unexplored. Another, independent, piece of evidence suggesting mitochondrial involvement in AD is the transcriptional downregulation of the mitochondrially encoded gene NADH dehydrogenase 6, yet another member of Complex I. In addition to perturbations to Complex I function, we also observe an overall decrease in mitochondrial tRNA expression. While we also see perturbations to other genes known to be involved in AD (particularly those involved in amyloid beta metabolism and related genes), our data lend great weight to the mitochondrial cascade hypothesis as an important avenue of research in AD. Another interesting gene with a high likelihood of translational changes given the tRNA changes is the signal recognition particle gene SRP14. The SRP is a ribonucleotide complex composed of the 7SL non-coding RNA as well as several proteins, including SRP14. This complex is responsible for targeting proteins to the endoplasmic reticulum (ER). This provides another potential piece of evidence that ER stress is associated with the pathology of AD, a theory that has recently gained traction [156].

By performing non-traditional RNA-sequencing we are able to elucidate the transcriptome of the AD brain in novel ways. The addition of non-coding RNAs, particularly those that modify other RNAs, helps to induce directionality to the characterization of genes associated with the disease state. As opposed to producing a simple set of correlations of gene expression amongst

an incohesive group of genes, we can now begin to generate hypotheses based on annotated regulatory connections. For example, the inclusion of microRNAs provides a “flow” to the network where microRNAs are upstream of their targets. The addition of genetically-associated loci can also provide a hint of causality where gene expression alone fails to do so. The field of genomics benefits the most when the cycle of hypothesis generating studies and functional studies flows continuously; here I have demonstrated that insights learned from small-scale functional studies can be fed back into a higher-throughput correlational study in order to sharpen its impact.

5. Conclusion

In this dissertation I have presented several methods for analyzing the non-coding transcriptome and an application of such studies to a disease: namely, Alzheimer’s disease. Given the longstanding prejudice against it in gene expression studies, the non-coding transcriptome is a treasure trove of unmined biological insights. By moving beyond simple assays of differential expression of protein-coding mRNAs we can finally begin to elucidate many previously neglected facets of the transcriptome.

In Chapter 2 I presented a method for examining noncanonical nucleotides in non-coding RNAs. Non-canonical nucleotides are just now coming of age in the field of epigenetics (DNA methylation and hydroxymethylation in particular), but the field of epitranscriptomics is only just beginning to be appreciated. These modified nucleotides are already established as biomarkers in some cancers [42,131,146] – but their incorporation into actual transcripts has yet to be studied in many disease systems. Ongoing work in the next several years will likely focus on the differences in incorporation of these modifications between disease and affected states, as well as across tissue types and even in evolutionary studies across species.

In Chapter 3 I described a method for characterizing small non-coding RNAs using a robust model that works both within and across a variety of tissue types. By beginning to describe the uncharacterized portion of the transcriptome we can begin to apply it to medicine – the most likely application of such basic biology is in the field of diagnostic biomarkers. Currently dominated by protein-based biomarkers, nucleotide-based biomarkers are likely to play a significant role in the coming years due to their greater easy of handling, greater reproducibility, and their pliability to large-scale molecular biology techniques.

In Chapter 4 I showcased a study of the non-coding transcriptome in Alzheimer's disease. By combining heretofore-uncharacterized non-coding changes in the AD transcriptome with known genetically associated loci, I was able to build a network that connected seemingly distant regions of the correlational-gene-expression network. Furthermore, the inclusion of regulatory RNAs like microRNAs and the differentially expressed minor spliceosome create an unprecedented opportunity for generating clearer directional hypotheses from gene expression data. Finally, the changes I describe in tRNA expression preferentially affect genes implicated in mitochondrial function, lending weight to the mitochondrial cascade hypothesis in AD.

6. Bibliography

[1] J. Abelson, C.R. Trotta, H. Li, tRNA Splicing, J. Biol. Chem. 273 (1998) 12685–12688.

[2] C. Addo-Quaye, T.W. Eshoo, D.P. Bartel, M.J. Axtell, Endogenous siRNA and miRNA Targets Identified by Sequencing of the Arabidopsis Degradome, Curr. Biol. 18 (2008) 758–762.

[3] P.F. Agris, Bringing order to translation: the contributions of transfer RNA anticodon-domain modifications, EMBO Rep. 9 (2008) 629–635.

[4] P.F. Agris, F.A.P. Vendeix, W.D. Graham, tRNA's wobble decoding of the genome: 40 years of modification, J. Mol. Biol. 366 (2007) 1–13.

[5] T.S. Alioto, U12DB: a database of orthologous U12-type spliceosomal introns, Nucleic Acids Res. 35 (2007) D110–115.

[6] F.W. Allen, The Biochemistry of the Nucleic Acids, Purines, and Pyrimidines, Annu. Rev. Biochem. 10 (1941) 221–244.

[7] G. Astarita, K.-M. Jung, V. Vasilevko, N.V. Dipatrizio, S.K. Martin, D.H. Cribbs, et al., Elevated stearoyl-CoA desaturase in brains of patients with Alzheimer's disease, PloS One. 6 (2011) e24777.

[8] D. Avramopoulos, Genetics of Alzheimer's disease: recent advances, Genome Med. 1 (2009) 34.

[9] J.E. Babiarz, R. Hsu, C. Melton, M. Thomas, E.M. Ullian, R. Blelloch, A role for noncanonical microRNAs in the mammalian brain revealed by phenotypic differences in Dgcr8 versus Dicer1 knockouts and small RNA sequencing, RNA N. Y. N. 17 (2011) 1489–1501.

[10] D.P. Bartel, MicroRNAs: genomics, biogenesis, mechanism, and function, Cell. 116 (2004) 281–297.

[11] B.L. Bass, RNA Editing by Adenosine Deaminases That Act on RNA, Annu. Rev. Biochem. 71 (2002) 817–846.

[12] S.L. Berger, T. Kouzarides, R. Shiekhattar, A. Shilatifard, An operational definition of epigenetics, Genes Dev. 23 (2009) 781–783.

[13] C. Bermudez-Santana, C.S. Attolini, T. Kirsten, J. Engelhardt, S.J. Prohaska, S. Steigele, et al., Genomic organization of eukaryotic tRNAs, BMC Genomics. 11 (2010) 270.

[14] L. Bertram, R.E. Tanzi, Genome-wide association studies in Alzheimer's disease, Hum. Mol. Genet. 18 (2009) R137–R145.

[15] M. Bhattacharyya, S. Bandyopadhyay, Studying the differential co-expression of microRNAs reveals significant role of white matter in early Alzheimer's progression, Mol. Biosyst. 9 (2013) 457–466.

[16] D.L. Black, B. Chabot, J.A. Steitz, U2 as well as U1 small nuclear ribonucleoproteins are involved in premessenger RNA splicing, Cell. 42 (1985) 737–750.

[17] E.M. Blalock, J.W. Geddes, K.C. Chen, N.M. Porter, W.R. Markesbery, P.W. Landfield, Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 2173–2178.

[18] M. Brameier, A. Herwig, R. Reinhardt, L. Walter, J. Gruber, Human box C/D snoRNAs with miRNA like functions: expanding the range of regulatory RNAs, Nucleic Acids Res. 39 (2011) 675–686.

[19] D. Bu, K. Yu, S. Sun, C. Xie, G. Skogerbø, R. Miao, et al., NONCODE v3.0: integrative annotation of long noncoding RNAs, Nucleic Acids Res. 40 (2012) D210–215.

[20] A. Burns, S. Iliffe, Alzheimer's disease, BMJ. 338 (2009) b158–b158.

[21] A.M. Burroughs, Y. Ando, M.J.L. de Hoon, Y. Tomaru, T. Nishibu, R. Ukekawa, et al., A comprehensive survey of 3' animal miRNA modification events and a possible role for 3' adenylation in modulating miRNA targeting effectiveness, Genome Res. 20 (2010) 1398–1410.

[22] A.M. Burroughs, Y. Ando, M.L. de Hoon, Y. Tomaru, H. Suzuki, Y. Hayashizaki, et al., Deep-sequencing of human Argonaute-associated small RNAs provides insight into miRNA sorting and reveals Argonaute association with RNA fragments of diverse origin, RNA Biol. 8 (2011) 158–177.

[23] C.A. Burtis, The determination of the base composition of RNA by high-pressure cation-exchange chromatography, J. Chromatogr. 51 (1970) 183–194.

[24] M.N. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, et al., Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses, Genes Dev. 25 (2011) 1915–1927.

[25] P. Carninci, J. Yasuda, Y. Hayashizaki, Multifaceted mammalian transcriptome, Curr. Opin. Cell Biol. 20 (2008) 274–280.

[26] T.R. Cech, The RNA Worlds in Context, Cold Spring Harb. Perspect. Biol. 4 (2012).

[27] T.R. Cech, A.J. Zaug, P.J. Grabowski, In vitro splicing of the ribosomal RNA precursor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence, Cell. 27 (1981) 487–496.

[28] C.T.Y. Chan, M. Dyavaiah, M.S. DeMott, K. Taghizadeh, P.C. Dedon, T.J. Begley, A Quantitative Systems Approach Reveals Dynamic Control of tRNA Modifications during Cellular Stress, PLoS Genet. 6 (2010) e1001247.

[29] Y. Chen, L. Jia, C. Wei, F. Wang, H. Lv, J. Jia, Association between polymorphisms in the apolipoprotein D gene and sporadic Alzheimer's disease, Brain Res. 1233 (2008) 196–202.

[30] Z. Chen, X. Duan, Ribosomal RNA Depletion for Massively Parallel Bacterial RNA-Sequencing Applications, in: Y.M. Kwon, S.C. Ricke (Eds.), High-Throughput Gener. Seq., Humana Press, 2011: pp. 93–103.

[31] H.R. Chiang, L.W. Schoenfeld, J.G. Ruby, V.C. Auyeung, N. Spies, D. Baek, et al., Mammalian microRNAs: experimental evaluation of novel and previously annotated genes, Genes Dev. 24 (2010) 992–1009.

[32] J.C. Chow, Z. Yen, S.M. Ziesche, C.J. Brown, Silencing of the mammalian X chromosome, Annu. Rev. Genomics Hum. Genet. 6 (2005) 69–92.

[33] V. Colangelo, J. Schurr, M.J. Ball, R.P. Pelaez, N.G. Bazan, W.J. Lukiw, Gene expression profiling of 12633 genes in Alzheimer hippocampal CA1: Transcription and neurotrophic factor down-regulation and up-regulation of apoptotic and pro-inflammatory signaling, J. Neurosci. Res. 70 (2002) 462–473.

[34] P.F. Crain, Preparation and enzymatic hydrolysis of DNA and RNA for mass spectrometry, Methods Enzymol. 193 (1990) 782–790.

[35] F. Crick, On Protein Synthesis, Crick Francis Protein Synth. Symp. Soc. Exp. Biol. 12 1958 138-163 Artic. 13 Images. (1958).

[36] F. Crick, Central Dogma of Molecular Biology, Nature. 227 (1970) 561–563.

[37] J. Curran, Modified nucleosides in translation, Modif. Ed. RNA ASM Press Wash. DC. (1998) 493–516.

[38] A. Czerwoniec, S. Dunin-Horkawicz, E. Purta, K.H. Kaminska, J.M. Kasprzak, J.M. Bujnicki, et al., MODOMICS: a database of RNA modification pathways. 2008 update, Nucleic Acids Res. 37 (2009) D118–121.

[39] P. Das, T. Golde, Dysfunction of TGF-beta signaling in Alzheimer's disease, J. Clin. Invest. 116 (2006) 2855–2857.

[40] C.A. Davis, M. Ares, Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in Saccharomycescerevisiae, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 3262–3267.

[41] R. Díaz-Uriarte, S.A. De Andres, Gene selection and classification of microarray data using random forest, BMC Bioinformatics. 7 (2006) 3.

[42] D. Djukovic, H.R. Baniasadi, R. Kc, Z. Hammoud, D. Raftery, Targeted serum metabolite profiling of nucleosides in esophageal adenocarcinoma, Rapid Commun. Mass Spectrom. RCM. 24 (2010) 3057–3062.

[43] D. Dominissini, S. Moshitch-Moshkovitz, S. Schwartz, M. Salmon-Divon, L. Ungar, S. Osenberg, et al., Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq, Nature. 485 (2012) 201–206.

[44] G. Donmez, The Effects of SIRT1 on Alzheimer's Disease Models, Int. J. Alzheimers Dis. 2012 (2012).

[45] I.A. Drinnenberg, G.R. Fink, D.P. Bartel, Compatibility with killer explains the rise of RNAi-deficient fungi, Science. 333 (2011) 1592.

[46] H.A. Ebhardt, H.H. Tsang, D.C. Dai, Y. Liu, B. Bostan, R.P. Fahlman, Meta-analysis of small RNA-sequencing errors reveals ubiquitous post-transcriptional RNA modifications, Nucleic Acids Res. 37 (2009) 2461–2470.

[47] F. Echtermeyer, M. Streit, S. Wilcox-Adelman, S. Saoncella, F. Denhez, M. Detmar, et al., Delayed wound repair and impaired angiogenesis in mice lacking syndecan-4, J. Clin. Invest. 107 (2001) R9–R14.

[48] S.R. Eddy, Non-coding RNA genes and the modern RNA world, Nat. Rev. Genet. 2 (2001) 919–929.

[49] G.L. Eliceiri, Small nucleolar RNAs, Cell. Mol. Life Sci. CMLS. 56 (1999) 22–31.

[50] F. Erhard, R. Zimmer, Classification of ncRNAs using position and size information in deep sequencing data, Bioinforma. Oxf. Engl. 26 (2010) i426–432.

[51] P. Fabrizio, J. Abelson, Two domains of yeast U6 small nuclear RNA required for both steps of nuclear precursor messenger RNA splicing, Science. 250 (1990) 404–409.

[52] M.A. Faghihi, M. Zhang, J. Huang, F. Modarresi, M.P. Van der Brug, M.A. Nalls, et al., Evidence for natural antisense transcript-mediated inhibition of microRNA function, Genome Biol. 11 (2010) R56.

[53] M. Fasold, D. Langenberger, H. Binder, P.F. Stadler, S. Hoffmann, DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments, Nucleic Acids Res. 39 (2011) W112–117.

[54] M. Filali, R. Lalonde, Age-related cognitive decline and nesting behavior in an APPswe/PS1 bigenic model of Alzheimer's disease, Brain Res. 1292 (2009) 93–99.

[55] A. Fire, S. Xu, M.K. Montgomery, S.A. Kostas, S.E. Driver, C.C. Mello, Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans, Nature. 391 (1998) 806–811.

[56] A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, et al., STRING v9.1: protein-protein interaction networks, with increased coverage and integration, Nucleic Acids Res. 41 (2013) D808–815.

[57] T.M. Frayling, N.J. Timpson, M.N. Weedon, E. Zeggini, R.M. Freathy, C.M. Lindgren, et al., A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity, Science. 316 (2007) 889–894.

[58] M.R. Friedländer, S.D. Mackowiak, N. Li, W. Chen, N. Rajewsky, miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades, Nucleic Acids Res. 40 (2012) 37–52.

[59] P.A. Fujita, B. Rhead, A.S. Zweig, A.S. Hinrichs, D. Karolchik, M.S. Cline, et al., The UCSC Genome Browser database: update 2011, Nucleic Acids Res. 39 (2011) D876–882.

[60] D. Gasparutto, T. Livache, A.M. Duplaa, H. Bazin, S. Favario, A. Guy, et al., [Total chemical synthesis of natural transfer RNA], Comptes Rendus Académie Sci. Sér. III Sci. Vie. 315 (1992) 1–6.

[61] J. Gebetsberger, M. Zywicki, Künzi, Andrea, N. Polacek, tRNA-Derived Fragments Target the Ribosome and Function as Regulatory Non-Coding RNA in *Haloferax volcanii*, Archaea. 2012 (2012).

[62] T. Gerken, C.A. Girard, Y.-C.L. Tung, C.J. Webby, V. Saudek, K.S. Hewitson, et al., The Obesity-Associated FTO Gene Encodes a 2-Oxoglutarate-Dependent Nucleic Acid Demethylase, Science. 318 (2007) 1469–1472.

[63] M.A. German, M. Pillay, D.-H. Jeong, A. Hetawal, S. Luo, P. Janardhanan, et al., Global identification of microRNA–target RNA pairs by parallel analysis of RNA ends, Nat. Biotechnol. 26 (2008) 941–946.

[64] M. Di Giacomo, S. Comazzetto, H. Saini, S. De Fazio, C. Carrieri, M. Morgan, et al., Multiple Epigenetic Mechanisms and the piRNA Pathway Enforce LINE1 Silencing during Adult Spermatogenesis, Mol. Cell. 50 (2013) 601–608.

[65] M.R. Green, Biochemical mechanisms of constitutive and regulated pre-mRNA splicing, Annu. Rev. Cell Biol. 7 (1991) 559–599.

[66] B.D. Gregory, R.C. O'Malley, R. Lister, M.A. Urich, J. Tonti-Filippini, H. Chen, et al., A Link between RNA Metabolism and Silencing Affecting Arabidopsis Development, Dev. Cell. 14 (2008) 854–866.

[67] H. Grosjean, R. Benne, Modification and Editing of Rna, ASM Press, 1998.

[68] H. Grosjean, L. Droogmans, M. Roovers, G. Keith, Detection of enzymatic activity of transfer RNA modification enzymes using radiolabeled tRNA substrates, Methods Enzymol. 425 (2007) 55–101.

[69] C. Guerrier-Takada, K. Gardiner, T. Marsh, N. Pace, S. Altman, The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme, Cell. 35 (1983) 849–857.

[70] R.C. Gupta, K. Randerath, Use of specific endonuclease cleavage in RNA sequencing, Nucleic Acids Res. 4 (1977) 1957–1978.

[71] M. Guttman, I. Amit, M. Garber, C. French, M.F. Lin, D. Feldser, et al., Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals, Nature. 458 (2009) 223–227.

[72] L. Habegger, A. Sboner, T.A. Gianoulis, J. Rozowsky, A. Agarwal, M. Snyder, et al., RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries, Bioinformatics. 27 (2011) 281–283.

[73] T. Hamma, A.R. Ferré-D'Amaré, Pseudouridine Synthases, Chem. Biol. 13 (2006) 1125–1135.

[74] J. Han, D. Kim, K.V. Morris, Promoter-associated RNA is required for RNA-directed transcriptional gene silencing in human cells, Proc. Natl. Acad. Sci. 104 (2007) 12422–12427.

[75] S.B. Hedges, J. Dudley, S. Kumar, TimeTree: a public knowledge-base of divergence times among organisms, Bioinforma. Oxf. Engl. 22 (2006) 2971–2972.

[76] M. Helm, C. Florentz, A. Chomyn, G. Attardi, Search for differences in post-transcriptional modification patterns of mitochondrial DNA-encoded wild-type and mutant human tRNALys and tRNALeu(UUR), Nucleic Acids Res. 27 (1999) 756–763.

[77] S.L. Hiley, J. Jackman, T. Babak, M. Trochesset, Q.D. Morris, E. Phizicky, et al., Detection and discovery of RNA modifications using microarrays, Nucleic Acids Res. 33 (2005) e2.

[78] I.L. Hofacker, W. Fontana, P.F. Stadler, L.S. Bonhoeffer, M. Tacker, P. Schuster, Fast folding and comparison of RNA secondary structures, Monatshefte Für ChemieChemical Mon. 125 (1994) 167–188.

[79] P. Hollingworth, D. Harold, R. Sims, A. Gerrish, J.-C. Lambert, M.M. Carrasquillo, et al., Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease, Nat. Genet. 43 (2011) 429–435.

[80] S.-D. Hsu, F.-M. Lin, W.-Y. Wu, C. Liang, W.-C. Huang, W.-L. Chan, et al., miRTarBase: a database curates experimentally validated microRNA–target interactions, Nucleic Acids Res. (2010).

[81] X.A. Huang, H. Yin, S. Sweeney, D. Raha, M. Snyder, H. Lin, A major epigenetic programming mechanism guided by piRNAs, Dev. Cell. 24 (2013) 502–516.

[82] J.N. Hutchinson, A.W. Ensminger, C.M. Clemson, C.R. Lynch, J.B. Lawrence, A. Chess, A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains, BMC Genomics. 8 (2007) 39.

[83] A. Jacquier, The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs, Nat. Rev. Genet. 10 (2009) 833–844.

[84] G. Jia, Y. Fu, X. Zhao, Q. Dai, G. Zheng, Y. Yang, et al., N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO, Nat. Chem. Biol. 7 (2011) 885–887.

[85] D.F. Johnson, K.S. Poksay, T.L. Innerarity, The mechanism for apo-B mRNA editing is deamination, Biochem. Biophys. Res. Commun. 195 (1993) 1204–1210.

[86] C.E. Joyce, X. Zhou, J. Xia, C. Ryan, B. Thrash, A. Menter, et al., Deep sequencing of small RNAs from human skin reveals major alterations in the psoriasis miRNAome, Hum. Mol. Genet. 20 (2011) 4025–4040.

[87] A.M. Khalil, M. Guttman, M. Huarte, M. Garber, A. Raj, D. Rivea Morales, et al., Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 11667–11672.

[88] S. Kishore, S. Stamm, The snoRNA HBII-52 Regulates Alternative Splicing of the Serotonin Receptor 2C, Science. 311 (2006) 230–232.

[89] E. Kojro, R. Postina, C. Buro, C. Meiringer, K. Gehrig-Burger, F. Fahrenholz, The neuropeptide PACAP promotes the alpha-secretase pathway for processing the Alzheimer amyloid precursor protein, FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol. 20 (2006) 512–514.

[90] J.S. Kooner, J.C. Chambers, C.A. Aguilar-Salinas, D.A. Hinds, C.L. Hyde, G.R. Warnes, et al., Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides, Nat. Genet. 40 (2008) 149–151.

[91] E.V. Koonin, A.E. Gorbalenya, K.M. Chumakov, Tentative identification of RNA-dependent RNA polymerases of dsRNA viruses and their relationship to positive strand RNA viral polymerases, FEBS Lett. 252 (1989) 42–46.

[92] A. Koulakoff, X. Mei, J.A. Orellana, J.C. Sáez, C. Giaume, Glial connexin expression and function in the context of Alzheimer's disease, Biochim. Biophys. Acta. 1818 (2012) 2048–2057.

[93] A. Kozomara, S. Griffiths-Jones, miRBase: integrating microRNA annotation and deep-sequencing data, Nucleic Acids Res. 39 (2010) D152–D157.

[94] K. Kruger, P.J. Grabowski, A.J. Zaug, J. Sands, D.E. Gottschling, T.R. Cech, Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena, Cell. 31 (1982) 147–157.

[95] P. Landgraf, M. Rusu, R. Sheridan, A. Sewer, N. Iovino, A. Aravin, et al., A mammalian microRNA expression atlas based on small RNA library sequencing, Cell. 129 (2007) 1401–1414.

[96] D. Langenberger, C.I. Bermudez-Santana, P.F. Stadler, S. Hoffmann, Identification and classification of small rnas in transcriptome sequence data, Pac. Symp. Biocomput. Pac. Symp. Biocomput. (2010) 80–87.

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

[98] F. Lantieri, J.T. Glessner, H. Hakonarson, J. Elia, M. Devoto, Analysis of GWAS top hits in ADHD suggests association to two polymorphisms located in genes expressed in the cerebellum, Am. J. Med. Genet. B Neuropsychiatr. Genet. 153B (2010) 1127–1133.

[99] Y.S. Lee, Y. Shibata, A. Malhotra, A. Dutta, A novel class of small RNAs: tRNA-derived RNA fragments (tRFs), Genes Dev. 23 (2009) 2639–2649.

[100] F. Li, P. Ryvkin, D.M. Childress, O. Valladares, B.D. Gregory, L.-S. Wang, SAVoR: a server for sequencing annotation and visualization of RNA structures, Nucleic Acids Res. 40 (2012) W59–64.

[101] F. Li, P. Ryvkin, D.M. Childress, O. Valladares, B.D. Gregory, L.-S. Wang, SAVoR: a server for sequencing annotation and visualization of RNA structures, Nucleic Acids Res. (2012).

[102] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform, Bioinformatics. 25 (2009) 1754–1760.

[103] Z. Li, C. Ender, G. Meister, P.S. Moore, Y. Chang, B. John, Extensive terminal and asymmetric processing of small RNAs from rRNAs, snoRNAs, snRNAs, and tRNAs, Nucleic Acids Res. (2012).

[104] W.S. Liang, E.M. Reiman, J. Valla, T. Dunckley, T.G. Beach, A. Grover, et al., Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons, Proc. Natl. Acad. Sci. 105 (2008) 4441–4446.

[105] D.D. Licatalosi, A. Mele, J.J. Fak, J. Ule, M. Kayikci, S.W. Chi, et al., HITS-CLIP yields genome-wide insights into brain alternative RNA processing, Nature. 456 (2008) 464–469.

[106] T.M. Lowe, S.R. Eddy, tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence, Nucleic Acids Res. 25 (1997) 0955–964.

[107] E. Lund, J.E. Dahlberg, True genes for human U1 small nuclear RNA. Copy number, polymorphism, and methylation., J. Biol. Chem. 259 (1984) 2013–2021.

[108] C. Luo, D. Tsementzi, N. Kyrpides, T. Read, K.T. Konstantinidis, Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample, PLoS ONE. 7 (2012) e30087.

[109] J.J. Mao, S. Katayama, C. Watanabe, Y. Harada, K. Noda, Y. Yamamura, et al., The relationship between alphaB-crystallin and neurofibrillary tangles in Alzheimer's disease, Neuropathol. Appl. Neurobiol. 27 (2001) 180–188.

[110] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet.journal. 17 (2011).

[111] E.S. Maxwell, M.J. Fournier, The small nucleolar RNAs, Annu. Rev. Biochem. 64 (1995) 897–934.

[112] B.J. McCarthy, J.J. Holland, Denatured DNA as a direct template for in vitro protein synthesis, Proc. Natl. Acad. Sci. 54 (1965) 880–886.

[113] P. Menzel, J. Gorodkin, P.F. Stadler, The tedious task of finding homologous noncoding RNA genes, RNA. 15 (2009) 2075–2082.

[114] T.R. Mercer, M.E. Dinger, J.S. Mattick, Long non-coding RNAs: insights into functions, Nat. Rev. Genet. 10 (2009) 155–159.

[115] K.D. Meyer, Y. Saletore, P. Zumbo, O. Elemento, C.E. Mason, S.R. Jaffrey, Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons, Cell. (2012).

[116] J.D. Mills, T. Nalpathamkalam, H.I.L. Jacobs, C. Janitz, D. Merico, P. Hu, et al., RNA-Seq analysis of the parietal cortex in Alzheimer's disease reveals alternatively spliced isoforms related to lipid metabolism, Neurosci. Lett. 536 (2013) 90–95.

[117] M.J. Moore, N.J. Proudfoot, Pre-mRNA processing reaches back to transcription and ahead to translation, Cell. 136 (2009) 688–700.

[118] S.M. Mount, I. Pettersson, M. Hinterberger, A. Karmas, J.A. Steitz, The U1 small nuclear RNA-protein complex selectively binds a 5′ splice site in vitro, Cell. 33 (1983) 509–518.

[119] A.C. Naj, G. Jun, G.W. Beecham, L.-S. Wang, B.N. Vardarajan, J. Buros, et al., Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease, Nat. Genet. 43 (2011) 436–441.

[120] P. Nissen, J. Hansen, N. Ban, P.B. Moore, T.A. Steitz, The structural basis of ribosome activity in peptide bond synthesis, Science. 289 (2000) 920–930.

[121] A. Nogalska, C. Terracciano, C. D'Agostino, W. King Engel, V. Askanas, p62/SQSTM1 is overexpressed and prominently accumulated in inclusions of sporadic inclusion-body myositis muscle fibers, and can help differentiating it from polymyositis and dermatomyositis, Acta Neuropathol. (Berl.). 118 (2009) 407–413.

[122] H.F. Noller, Ribosomal RNA and Translation, Annu. Rev. Biochem. 60 (1991) 191–227.

[123] H.F. Noller, C.R. Woese, Secondary structure of 16S ribosomal RNA, Science. 212 (1981) 403–411.

[124] F. Ozsolak, P.M. Milos, RNA sequencing: advances, challenges and opportunities, Nat. Rev. Genet. 12 (2011) 87–98.

[125] M. Pastorcic, H.K. Das, Analysis of transcriptional modulation of the presenilin 1 gene promoter by ZNF237, a candidate binding partner of the Ets transcription factor ERM, Brain Res. 1128 (2007) 21–32.

[126] Y.-F. Ren, G. Li, J. Wu, Y.-F. Xue, Y.-J. Song, L. Lv, et al., Dicer-Dependent Biogenesis of Small RNAs Derived from 7SL RNA, PLoS ONE. 7 (2012) e40705.

[127] A. Rich, U.L. RajBhandary, Transfer RNA: molecular structure, sequence, and properties, Annu. Rev. Biochem. 45 (1976) 805–860.

[128] E. Roberts, A. Sethi, J. Montoya, C.R. Woese, Z. Luthey-Schulten, Molecular signatures of ribosomal evolution, Proc. Natl. Acad. Sci. 105 (2008) 13953–13958.

[129] J. Rozenski, P.F. Crain, J.A. McCloskey, The RNA Modification Database: 1999 update, Nucleic Acids Res. 27 (1999) 196–197.

[130] Y. Saletore, K. Meyer, J. Korlach, I.D. Vilfan, S. Jaffrey, C.E. Mason, The birth of the Epitranscriptome: deciphering the function of RNA modifications, Genome Biol. 13 (2012) 1–12.

[131] S. Scorrano, L. Longo, G. Vasapollo, Molecularly imprinted polymers for solid-phase extraction of 1-methyladenosine from human urine, Anal. Chim. Acta. 659 (2010) 167–171.

[132] A.G. Seto, R.E. Kingston, N.C. Lau, The Coming of Age for Piwi Proteins, Mol. Cell. 26 (2007) 603–609.

[133] A.J. Shatkin, J.L. Manley, The ends of the affair: Capping and polyadenylation, Nat. Struct. Mol. Biol. 7 (2000) 838–842.

[134] S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, et al., dbSNP: the NCBI database of genetic variation, Nucleic Acids Res. 29 (2001) 308–311.

[135] N. Shibata, T. Nagata, S. Shinagawa, T. Ohnuma, H. Shimazaki, M. Komatsu, et al., Genetic association between APOA1 and APOD polymorphisms and Alzheimer's disease in a Japanese population, J. Neural Transm. Vienna Austria 1996. (2013).

[136] H. Siemen, D. Colas, H.C. Heller, O. Brüstle, R.A. Reijo Pera, Pumilio-2 Function in the Mouse Nervous System, PLoS ONE. 6 (2011) e25932.

[137] M.C. Siomi, K. Sato, D. Pezic, A.A. Aravin, PIWI-interacting small RNAs: the vanguard of genome defence, Nat. Rev. Mol. Cell Biol. 12 (2011) 246–258.

[138] C.M. Smith, J.A. Steitz, Classification of gas5 as a Multi-Small-Nucleolar-RNA (snoRNA) Host Gene and a Member of the 5′-Terminal Oligopyrimidine Gene Family Reveals Common Features of snoRNA Host Genes, Mol. Cell. Biol. 18 (1998) 6897–6909.

[139] H.C. Smith, M.P. Sowden, Base-modification mRNA editing through deamination — the good, the bad and the unregulated, Trends Genet. 12 (1996) 418–424.

[140] R.A. Somerville, TSE agent strains and PrP: reconciling structure and function, Trends Biochem. Sci. 27 (2002) 606–612.

[141] Y. Sonoda, H. Mukai, K. Matsuo, M. Takahashi, Y. Ono, K. Maeda, et al., Accumulation of tumor-suppressor PTEN in Alzheimer neurofibrillary tangles, Neurosci. Lett. 471 (2010) 20–24.

[142] M. Sprinzl, K.S. Vassilenko, Compilation of tRNA sequences and sequences of tRNA genes, Nucleic Acids Res. 33 (2005) D139–140.

[143] J.E. Squires, H.R. Patel, M. Nousch, T. Sibbritt, D.T. Humphreys, B.J. Parker, et al., Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA, Nucleic Acids Res. 40 (2012) 5023–5033.

[144] A. Statnikov, L. Wang, C.F. Aliferis, A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification, BMC Bioinformatics. 9 (2008) 319.

[145] G. Stefani, F.J. Slack, Small non-coding RNAs in animal development, Nat. Rev. Mol. Cell Biol. 9 (2008) 219–230.

[146] W. Struck, D. Siluk, A. Yumba-Mpanga, M. Markuszewski, R. Kaliszan, M.J. Markuszewski, Liquid chromatography tandem mass spectrometry study of urinary nucleosides as potential cancer markers, J. Chromatogr. A. 1283 (2013) 122–131.

[147] Y. Sun, X.S. Yin, H. Guo, R.K. Han, R.D. He, L.J. Chi, Elevated osteopontin levels in mild cognitive impairment and Alzheimer's disease, Mediators Inflamm. 2013 (2013) 615745.

[148] R.H. Swerdlow, J.M. Burns, S.M. Khan, The Alzheimer's disease mitochondrial cascade hypothesis, J. Alzheimers Dis. JAD. 20 Suppl 2 (2010) S265–279.

[149] R.J. Taft, C.D. Kaplan, C.S. and J.S. Mattick, Evolution, biogenesis and function of promoter-associated RNAs, Cell Cycle. 8 (2009) 2332–2338.

[150] H.M. Temin, S. Mizutani, RNA-dependent DNA polymerase in virions of Rous sarcoma virus, Nature. 226 (1970) 1211–1213.

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

[152] H. Tiedge, W. Chen, J. Brosius, Primary structure, neural-specific expression, and dendritic location of human BC200 RNA, J. Neurosci. Off. J. Soc. Neurosci. 13 (1993) 2382–2390.

[153] G. Todd, K. Karbstein, RNA takes center stage, Biopolymers. 87 (2007) 275–278.

[154] D. Tollervey, A yeast small nuclear RNA is required for normal processing of pre-ribosomal RNA., EMBO J. 6 (1987) 4169–4175.

[155] N.A. Twine, K. Janitz, M.R. Wilkins, M. Janitz, Whole Transcriptome Sequencing Reveals Gene Expression and Splicing Differences in Brain Regions Affected by Alzheimer's Disease, PLoS ONE. 6 (2011) e16266.

[156] U. Unterberger, R. Höftberger, E. Gelpi, H. Flicker, H. Budka, T. Voigtländer, Endoplasmic reticulum stress features are prominent in Alzheimer disease but not in prion diseases in vivo, J. Neuropathol. Exp. Neurol. 65 (2006) 348–357.

[157] A. Vagnoni, M.S. Perkinton, E.H. Gray, P.T. Francis, W. Noble, C.C.J. Miller, Calsyntenin-1 mediates axonal transport of the amyloid precursor protein and regulates Aβ production, Hum. Mol. Genet. 21 (2012) 2845–2854.

[158] A.P.M. Verhagen, G.J.M. Pruijn, Are the Ro RNP-associated Y RNAs concealing microRNAs? Y RNA-derived miRNAs may be involved in autoimmunity, BioEssays. 33 (2011) 674–682.

[159] M.R. Willmann, N.D. Berkowitz, B.D. Gregory, Improved genome-wide mapping of uncapped and cleaved transcripts in eukaryotes—GMUCT 2.0, Methods. (n.d.).

[160] H.-K.A. Wong, T. Veremeyko, N. Patel, C.A. Lemere, D.M. Walsh, C. Esau, et al., De-repression of FOXO3a death axis by microRNA-132 and -212 causes neuronal apoptosis in Alzheimer's disease, Hum. Mol. Genet. 22 (2013) 3077–3092.

[161] S.A. Woodson, J.G. Muller, C.J. Burrows, S.E. Rokita, A primer extension assay for modification of guanine by Ni(II) complexes, Nucleic Acids Res. 21 (1993) 5524–5525.

[162] S.K. Wyman, E.C. Knouf, R.K. Parkin, B.R. Fritz, D.W. Lin, L.M. Dennis, et al., Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity, Genome Res. 21 (2011) 1450–1461.

[163] Y. Xi, W. Li, BSMAP: whole genome bisulfite sequence MAPping program, BMC Bioinformatics. 10 (2009) 232.

[164] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao, T. Li, miRecords: an integrated resource for microRNA-target interactions, Nucleic Acids Res. 37 (2009) D105–110.

[165] J.-H. Yang, J.-H. Li, P. Shao, H. Zhou, Y.-Q. Chen, L.-H. Qu, starBase: a database for exploring microRNA–mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data, Nucleic Acids Res. 39 (2011) D202–D209.

[166] S. Yokoyama, S. Nishimura, Modified nucleosides and codon recognition, TRNA Struct. Biosynth. Funct. ASM Press Wash. DC. (1995) 207–223.

[167] I. Younis, K. Dittmar, W. Wang, S.W. Foley, M.G. Berg, K.Y. Hu, et al., Minor introns are embedded molecular switches regulated by highly unstable U6atac snRNA, eLife. 2 (2013).

[168] D.C. Youvan, J.E. Hearst, Reverse Transcriptase Pauses at N2-Methylguanine During in Vitro Transcription of Escherichia Coli 16S Ribosomal RNA, Proc. Natl. Acad. Sci. 76 (1979) 3751–3754.

[169] B. Yu, Z. Yang, J. Li, S. Minakhina, M. Yang, R.W. Padgett, et al., Methylation as a crucial step in plant microRNA biogenesis, Science. 307 (2005) 932–935.

[170] Y.T. Yu, M.D. Shu, J.A. Steitz, A new method for detecting sites of 2'-O-methylation in RNA molecules, RNA N. Y. N. 3 (1997) 324–331.

[171] J. Zhang, L. Feuk, G. Duggan, R. Khaja, S. Scherer, Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome, Cytogenet. Genome Res. 115 (2006) 205–214.

[172] Q. Zheng, P. Ryvkin, F. Li, I. Dragomir, O. Valladares, J. Yang, et al., Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in Arabidopsis, PLoS Genet. 6 (2010).

[173] Cost of Dementia Tops $157 Billion Annually in the United States | RAND.