COMPREHENSIVE COMPUTATIONAL ANALYSIS OF CHROMATIN-ENRICHED RNAS REVEAL BOTH ACTIVE AND REPRESSIVE CIS-REGULATORY NON-CODING RNAS

by

Xiangying Sun

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department/School of Biological Sciences

West Lafayette, Indiana

August 2019

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Michael Gribskov, Chair
Department of Biological Sciences
Dr. Daisuke Kihara
Department of Biological Sciences
Dr. Yuk Fai Leung
Department of Biological Sciences
Dr. Jody Banks
Department of Botany and Plant Pathology

Approved by:

Dr. Daniel Suter
Head of the Graduate Program

“That's one small step for a man, one giant leap for mankind.”

- Neil Armstrong

ACKNOWLEDGMENTS

I would like to express my greatest and the sincerest gratitude to my major advisor, Dr. Michael Gribskov, for his invaluable support and encouragement over the years. He has always been there more as a friend than a mentor. His philosophy of research and teaching instructs me to explore in the areas that I am really interested in. I am also grateful to my committee members: Dr. Daisuke Kihara, Dr. Yuk Fai Leung and Dr. Jody Banks for their insightful comments. My sincere thanks also go to Dr. Xinan Yang, Ivan P. Moskowitz and Alex Ruthenburg. They have provided precious support and collaboration to steer my research project in the most useful and interesting direction.

I also want to thank my lovely lab members and friends for the stimulating discussions, for the fun we had together and for the years of memorable company. In particular, I'm very grateful to Dr. Biaobin Jiang for his inspiring advice on my research.

Last but not the least, I want to extend my deepest love to thank my family for giving me endless love and always showing faith in me. Thanks for being my strongest support in my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

3C	chromosome conformation capture
3D-DSL	3D DNA selection and ligation
3'UTR	3’ untranslated region
4C	circular chromosome conformation capture
5C	chromosome conformation capture carbon copy
5'UTR	5’ untranslated region
6C	Combined 3C-Chip-Cloning
as-cheRNA	antisense Chromatin-Enriched RNA
as-RNA	antisense RNA
C. elegans	Caenorhabditis elegans
CAGE	Cap Analysis of Gene Expression
cheRNA	Chromatin-enriched RNA
ChIA-PET	chromatin interaction analysis by paired-end tag sequencing
ChIP–seq	chromatin immunoprecipitation (ChIP) coupled with DNA sequencing
ChromHMM	the broad Chromatin State Segmentation by Hidden Markov Model
CPC2	Coding Potential Calculator 2
CPE	Chromatin Pellet Extract
CPM	Counts Per Million
DBN	Dynamic Bayesian Network
E.coli	Escherichia coli
eRNA	Enhancer-derived RNA
ER-α	17β-oestradiol (E2)-bound oestrogen receptor α
FAIRE-seq	Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE)-seq
GRO-seq	Global Run-On sequencing
icheRNA	Intergenic Chromatin-Enriched RNA
isneRNA	Intergenic Soluble Nuclear-Extracted RNAs

LADs	Lamina-Associated Domains
lncRNA	long noncoding RNAs
miRNA	microRNA
mRNA	messenger RNA
ncDNA	non-coding DNA
ncRNA	non-coding RNA
NET-seq	native elongating transcript sequencing
NGS	next-generation sequencing
OmpF	outer membrane protein F
PEAT	paired-end analysis of TSSs
Pol II	RNA polymerase II
PRC2	Polycomb Repressive Complex 2
PRC2	Polycomb Repressive Complex 2
PRO-seq	precision run-on sequencing
RNAi	RNA interference
rRNA	ribosomal RNA
SNE	Soluble Nuclear Extract
SNE	Soluble Nuclear Extract
sneRNA	Soluble Nuclear-Extracted RNAs
snRNA	small nuclear RNA
sRNA	small RNA
TEs	Transposable elements
tgDNA	total genomic DNA
tRNA	transfer RNA
TSS	Transcript Start Site
uRNA	uridine (U)-rich RNA

ABSTRACT

Author: Sun, Xiangying. PhD
Institution: Purdue University
Degree Received: August 2019
Title: Comprehensive Computational Analysis of Chromatin-Enriched RNAs Mark Both Active and Repressive Cis-Regulative Non-coding RNAs.
Committee Chair: Michael Gribskov

Nuclear RNA-seq has revealed thousands of potentially regulatory long noncoding RNA (lncRNA). Nuclear-retained lncRNA may interact with various chromatin regulatory proteins and recruit them to cis-regulatory elements in order to regulate gene expression. We are interested in analyzing nuclear RNA-seq to identify chromatin-associated lncRNA (cheRNA) that share enhancer features and transcription-factor dependence, and are thus being indicators of cis-acting loci. Nuclear RNA-seq requires rigorous and effective pipelines that differ from the conventional pipelines used for total RNA-seq datasets, but a thorough survey of analytic pipelines for nuclear RNA-seq has not been performed.

The existing computational pipeline (Werner) has important biases. To address the flaws in Werner pipeline, we have developed three new pipelines (referred to here as Tuxedo, Concatenating, and Taco) to analyze nuclear RNA-seq datasets. In this study, we survey the four nuclear RNA-seq analytic pipelines for the cheRNA identification and use the optimal scheme to explore new structure features of cheRNA that has high cis-regulatory potential.

To evaluate the transcriptomes assembled by the four pipelines, we used RNA from K562 cells as an example. The Tuxedo pipeline assembles complete transcriptome, including 10.9k unannotated lncRNAs. Transcripts assembled by Tuxedo, compared to the other three pipelines, showed the highest fraction of ongoing transcription by Pol II, and the highest level of nascent transcription by GRO-seq, demonstrating that the Tuxedo assembled transcriptome is more concordant with the active transcription signal represented by traditional measurements.

Comparing the four pipelines, Tuxedo also outperforms the other pipelines constructing assemblies that are enriched in enhancer hallmarks. ROC analysis, using the pool of predicted

transcripts identification by all four methods, shows that Tuxedo identifies cheRNA precisely, while recapturing three known genomic features of active enhancer.

Applying the Tuxedo approach to the K562 dataset, we found that intergenic cheRNA (icheRNA) is more positively correlated with the transcription of neighboring gene than with randomly selected gene. This demonstrates, for the first time, a quantitative cis-regulatory effect of cheRNA expression. A similar analysis of FAMTOM- or ChromHMM-predicted eRNA, which is believed to have cis-regulatory enhancer effect, shows similar but weaker positive correlation. In contrast, intergenic chromatin depleted RNA (isneRNA) and neighboring gene show negative correlation.

Genomic regions with abundant H3K9me3 modification, which is usually associated with condensed, inactive chromatin regions, can be actively transcribed. IcheRNAs with high levels of H3K9me3 in the gene body are transcribed at dramatically higher levels than those with lower levels; This is seen in both the soluble nuclear extract and the chromatin pellet, indicating that icheRNA is actively transcribed from regions with high H3K9me3 modification (contrary to expectation). One hypothesis for the unexpected H3K9me3 signal around icheRNA is that the icheRNA may be embedded in condensed domains derived from mobile elements.

We observed that the TSS of antisense cheRNA (as-cheRNA) colocalized mRNA is significantly less open (measured by ATAC-seq signal), has fewer active transcription marks (POL II, H3K4me3), and has more repressive marks (H3K27me3) and PRC2 complex binding (SUZ12, EZH2), compared with random mRNA. This pattern is not observed in mRNA colocalized with antisense chromatin depleted RNA (as-sneRNA), suggesting that as-cheRNA may be cis-regulatory elements that interfere transcription of colocalized mRNAs on the opposite strand via recruiting the PRC2 complex.

Nuclear RNA-seq sheds new light on cis-regulatory elements and the Tuxedo computational pipeline can be used to analyze nuclear RNA-seq data containing both high low expression lncRNA. With our improved computational strategy, we have examined the molecular characteristics of cheRNA in greater detail than has heretofore been possible. Notwithstanding the similarity of these features to those of eRNA, our analysis finds several unique molecular characteristics that quantitatively distinguishes icheRNA and eRNA. Our evidence suggests cheRNA has diverse functions, and may interact with diverse chromatin modulators, or utilize

RNA elements to perform cell-type specific cis-regulatory roles, including transcriptional activation and repression. Our approach thus affords a straightforward approach to identifying novel regulatory lncRNA for future mechanistic evaluation.

CHAPTER 1. INTRODUCTION

1.1 Non-coding regions: the dark matter of the genome

When I taught genomics in our Fundamental Biology course, I always started by telling the students a fun fact – the human genome is not the largest eukaryotic genome on this planet. The genome of the single-celled amoebae is up to 100-fold larger than the human genome. The students were always surprised as they expected humans would need a larger genome to support their complexity. Scientists in the 20th century were also troubled by this fact when they studied the size, evolution, and function of genomes. Mirsky and Ris introduced a concept called "C-value" to describe the amount of DNA in the haploid genome of an organism. They found there was little correlation between the "C-value" and the complexity of an organism, even though there was a general increase in "C-value" with organisms from prokaryotes to vertebrates (Mirsky and Ris, 1951). Thomas later termed this puzzling observation the "C-value paradox" (Thomas 1971). In 1980, the "C-value paradox" became even more puzzling when Lewin discovered that a large portion of many genomes does not code for proteins (Kung et al., 2013). Surprisingly, the increased complexity of the genome is not reflected by an increased number of protein-coding genes; this was termed the "G-value paradox" (Taft et al., 2007), andled to the hypothesis that the organism maintains a certain amount of "junk DNA", which could consist of any sequence, to make its genome size optimal (Horner and Macgregor, 1983). At that time, the number of protein-coding genes in human genome was estimated to be in the range of 50,000-140,000 (Antequera et al., 1993; Fields et al., 1994). However, this estimate dramatically decreased after the publication of the human genome in 2001. The sequence of the human genome revealed two striking and surprising facts. First, the genome contains only 20,000-30,000 protein-coding genes, close to the number of protein-coding genes in the genomes of the invertebrate sDrosophila melanogaster and Caenorhabditis elegans; and second, a large fraction (98.8%) of the human genome is composed of non-coding DNA (Venter et al., 2001; Lander et al., 2001). Moreover, whole genome sequencing and annotation of more organisms, showed that human genome contains an even a smaller number of protein-coding genes than plants such as rice (~37,000), or protists such as Paramecium tetraurelia (~40,000) (Taft et al., 2007). It became apparent that the number of protein-coding genes does not reflect the developmental complexity of the organism. In 2004, a

comparative analysis of 85 sequenced organisms (59 bacteria, 8 archaea, and 18 eukaryotes - 7 simple eukaryotes, 1 fungus, 3 plants, 3invertebrates, 1 urochordate, and 3 vertebrates) demonstrated that the relative amount of noncoding DNA in the genome of an organism, i.e. the ratio of non-coding DNA to total genomic DNA (ncDNA/tgDNA), consistently increases with organism complexity (Taft and Mattick, 2003). A following study by the same group further showed that the distribution of intronic sequences in the genome is not random. Especially in complex organisms (e.g., mouse and human), geneswith large amounts of intronic sequence (91-100% of bases in introns are significantly enriched in genes involved in embryonic, neurological, and immune system development (Taft et al., 2007). This strongly suggested that the ubiquitous non-coding regions, previously regarded as "junk DNA", could, perhaps, be far more important than had been imagined in controlling developmental complexity of organisms.

1.2 Early discoveries of non-coding RNAs

Non-coding RNA (ncRNA) is a class of functional RNA that is transcribed from DNA, but not translated into protein. Conservative estimates from the GENCODE (v25) project showed that ncRNA is pervasively transcribed in the human genome (51.8% of the human genome is transcribed, but only 1.2% encodes proteins (Ransohoff *et al*., 2018)). Unlike protein-coding messenger RNA (mRNA), which clearly functions as the intermediate carrying genetic information from DNA to protein, ncRNA function in diverse roles.

The past decade has witnessed an explosion in the studies of ncRNA. However, before the completion of the Human Genome Project, studies of RNA were largely focused on the roles of the mRNA. While the widespread interest in ncRNA is rather recent, the discovery that ncRNA could have functions is not.

In 1955, Georges Palade identified the very first class of ncRNA: ribosomal RNA (rRNA) (Jarroux *et al*., 2017). In 1958, Francis Crick described the existence of an "adaptor" needed to mediate between the triplet genetic code and the encoded amino acids (Crick, 1958). Meanwhile, Mahlon Hoagland and Paul Zamecnik identified these "adaptors" biochemically (Hoagland *et al*., 1958). The "adaptors", which were later recognized as transfer RNA (tRNA), were the second identified class of ncRNA (Eddy, 2001). In the early 1980s, with the discovery and isolation of uridine (U)-rich RNAs (uRNAs), a new class of ncRNA called small nuclear RNAs (snRNA) was recognized

(Zieve, 1981). This class of ncRNA were later proven to be a component of the spliceosome, which is involved in the process of mRNA splicing, and is a major player in post-transcriptional RNA processing (Tam and Steitz, 1996; Sharp and Burge, 1997). At this time, the ncRNA that had been discovered were limited to housekeeping ncRNAs (including rRNAs, tRNAs and snRNAs). For many years, therefore, ncRNAs were considered only as accessory components involving in protein synthesis and their pervasive regulatory roles were overlooked.

The initial discovery that ncRNA can function as a regulatory molecule occurred in 1984, when Masayuki Inoue identified the very first regulatory ncRNA, *micF,* in *Escherichia coli* (Inouye and Delihas, 1988). The *micF* ncRNA was shown to base pair with the mRNA encoding the outer membrane protein F (OmpF), and to thereby reduce the level of OmpF. Subsequent studies confirmed that regulation of gene expression via base pairing with target mRNA is also in bacteria, supporting the widespread existence of this mechanism (De Lay *et al*., 2013). This class of prokaryotic ncRNA was designated small RNAs (sRNA). sRNA is a major class of regulatory ncRNA in prokaryotes that functions to inhibit both transcription and translation of target mRNA.

1.3 MicroRNA[1]

1.3.1 Biological background

MicroRNAs (miRNAs) are typically about 22 bases long (lengths can vary from 16-24 bases in different species) that play an important role in gene regulation in eukaryotic organisms, usually acting by targeting the mRNA for degradation, or by acting as a translation repressor (see (Catalanotto, Cogoni and Zardo 2016) for a review of nuclear functions).

---

[1] This section has been published in a peer reviewed book. Sun, X. and Gribskov, M. (2019). MicroRNA and lncRNA Databases and Analysis. Encyclopedia of Bioinformatics and Computational Biology, vol.2, pp. 165–170. Oxford: Elsevier.

The first miRNA to be discovered was *lin-4*, which was identified by Ambros *et al.* in *C. elegans* (Almeida *et al*., 2011). In the 1980s, they showed that *lin-4* acts as a negative regulator of expression of the LIN-14 protein, resulting a temporally controlled decrease in the level of LIN-14 protein starting in the first laval stage of *C. elegans* development (Ambros and Horvitz, 1987; Ambros, 1989). In 1993, this group identified and cloned two small *lin-4* transcripts of approximately 22 and 61 nucleotides in length, and showed that neither of the two transcripts encode proteins. They also showed that the *lin-4* transcripts both contain sequences complementary to a repeated sequence found in the in 3'UTR of the *lin-14* mRNA, suggesting that *lin-4* regulates translation of *lin-14* via an antisense RNA-RNA interaction (Lee *et al*., 1993). In 2000, the second miRNA *let-7* was identified in *C.elegans*, and found to play an important role in the transition from the late larval to adult cell stage (Reinhart *et al*., 2000). *Let-7* was also found to to have homologues in a variety of animal species (including vertebrates, ascidians, hemichordates, mollusks, annelids, and arthropods) (Pasquinelli *et al*., 2000). This discovery triggered greatly increased interest in miRNA, leading to the characterization of miRNAs as general regulatory elements important in development and differentiation. In 2001, Hutvagner and coworkers presented *in vivo* and *in vitro* evidence, in *D. melanogaster,* explaining the biogenesis of mature miRNA, and the RNA interference (RNAi) machinery: briefly, the *let-7* precursor is processed into a stem-loop structure by the Drosha-containing microprocessor complex and then exported to the cytoplasm to be cleaved into the mature *let-7* miRNA (Hutvagner *et al*., 2001; Jarroux *et al*., 2017). These studies, together, indicated that miRNA with regulatory function are not just isolated examples. Since then, with the advent of next-generation sequencing (NGS) technology, and development in bioinformatics methods for miRNA identification, the number of novel functional miRNA has greatly expanded (Palazzo and Lee, 2015).

MicroRNA genes are common in eukaryotic genomes; usually there are thousands of miRNA genes. For instance, in humans, the ENCODE project reported 11,000 small RNA genes (The Encode Project Consortium 2012), and about 2600 mature miRNAs are listed in miRBase (S. Griffiths-Jones 2006). MiRNA genes are often located in the introns of protein coding genes (sometimes called miRtrons), in UTRs of coding transcripts, or found as completely separate transcripts. The primary transcript (pri-miRNA) is processed to produce a 60-100 base precursor RNA by the splicing process, in the case of intron encoded miRNAs, or by Drosha (DCL1 in plants) in the case of non-intron miRNAs. In either case, the result is a precursor RNA (pre-

miRNA), with an extended base-paired hairpin structure. After export from the nucleus, the pre-miRNA is asymmetrically cleaved, near the loop of the hairpin structure by the Dicer endonuclease to produce a mature miRNA. One strand of the mature miRNA, usually with a strong preference for the strand originating from the 5' end of the pre-miRNA, referred to as the 5p strand, is loaded into the RNA-induced silencing complex, RISC, and the other, the 3p or * strand is degraded. Within the RISC, the miRNA is bound by an Argonaute (Ago) protein, and is used to locate its complementary target mRNA, usually binding in the 3' UTR. The first 2-7 bases of the miRNA, the seed sequence, are particularly important in binding to the target, although extended complementarity or the miRNA and target mRNA is common. The seed region is also of interest because miRNAs with the same seed sequence are generally assigned to the same family. The mRNA is ultimately degraded by one of several pathways once it is bound to RISC. There are many exceptions and differences from the canonical process described above, but the general aspects are highly conserved. In addition to classical miRNA, there are additional classes of regulatory small RNAs including piwi-interacting RNA (piRNA), which interact with piwi proteins, a subtype of Argonaute proteins, and appear to act primarily to repress transcription of transposable elements (for a review see (Tang 2010)), and small interfering RNAs (siRNA) which are also produced from double stranded precursors by a Dicer-like system. One of the difficulties in predicting miRNAs is that exceptions to almost every aspect of the canonical process, described above, have been found, and while the process is very similar, there are significant differences between plants, animals, and fungi.

From a computational viewpoint, the focus is usually on 1) miR discovery, identifying the miRNA genes or pri-miRNA transcripts from genome or transcriptome sequence and predicting the mature miRNA sequence from the gene/precursor sequence, and 2) predicting the mRNA targets.

1.3.2 MiRNA discovery

Mature miRNAs can be identified experimentally by isolating and sequencing small RNAs, typically 17-28 bases long. The sequences can then be mapped to the reference genome using standard short-read mapping programs such as BWA or Bowtie2. Mismatches must be allowed since miRNAs may undergo adenosine to inosine editing (Cai, *et al*. 2009). Pre-miRNAs, which are typically capped and polyadenylated, can also be identified in typical RNA-Seq experiments. In this case, the two arms of the miRNA hairpin stem are often detected as separate reads

(Kozomara and Griffiths-Jones 2013). Crosslinking-immunoprecipitation has been used to identify miRNAs bound *in viv*o to Argonaute. This should, in principle, provide much better experimental datasets, but Agarwal *et al* (Agarwal, *et al*. 2015) suggest that many putative sites miRNA binding sites may be non-functional.

Computational identification of novel miRNA genes in genomic sequence relies on a combination of sequence similarity to known mature miRNAs, typically based on Blast (Altschul, *et al*. 1997) searches, and secondary structure prediction using UnaFold (Markham and Zuker 2008), RNAFold (Mathews, *et al*. 2004), or the Vienna RNA package (Lorenz, *et al*. 2011). MiRNAs are often highly conserved between species, and the conservation of the mature miRNA should be nearly perfect, however due to "arm switching" (Griffiths-Jones, *et al*. 2011), the shift of the mature miRNA from the 5p to 3p side of the precursor hairpin, matching to just mature miRNAs can be problematic, and matching to the pre-miRNA is likely to be more reliable. The second approach to identification of miRNAs lies in detection of the long base-paired hairpin stem of the miRNA. Minimum free energy RNA secondary structure prediction methods are typically used to detect potential hairpin structures (see references, above), usually after pre-screening to restrict the analysis to only 3' UTRs, to identify sequences similar to known miRNA, or to remove protein coding sequences, structural RNAs and transposable elements. Because there are many sequences that are predicted to fold as an acceptable hairpin stem, for instance Bentwich (Bentwich, *et al*. 2005) identified about 11 million in the human genome, this basic approach is typically augmented by additional *ad hoc* criteria. These so-called context criteria examine features such as the presence of a base paired region adjacent to the precursor hairpin with a typical 2 base overhang on the 3p arm, require fewer than 4 mismatches between the 5p and 3p arms in the mature miRNA region, absence of loops in the mature miRNA region, or place additional constraints on GC-content, minimum predicted folding free energy (Zhang, *et al*. 2006) or alignment score, structural "exposure" of the seed binding region in the mRNA, exclusion of perfect inverted repeats forming the putative pre-miRNA hairpin (possible transposable element) (see, for instance, (Lucas and Budak 2012), (Meyers, *et al*. 2008)), and continuous pairing in the precursor stem. RNA sequencing data is frequently included, requiring that sequences for both the 5p and 3p arms be detected with a minimum number of reads. Most of these features have been proposed based on inspection of specific sets of predicted or validated miRNAs, and their power and generality are often unclear.

1.3.3 MiRNA target prediction

There is general consensus that complementarity of the mRNA with the miRNA seed sequence, conservation of the miRNA and mRNA target sequence across species, the predicted stability of the miRNA-mRNA duplexes, presence of multiple sites (abundance), and accessibility of the mRNA target site are among the most important features in predicting mRNA target sites (Peterson SM 2014). However, many of the same *ad hoc* features listed above may be incorporated. There are literally dozens of predictive methods (for a recent comparison see (Fan and Kurgan 2015)). The recall (fraction of known miRNA targets identified in known data) and precision (fraction of correctly predicted targets) vary widely, and offer the classic trade-off between high recall-low precision and low recall-high precision.

Many machine learning methods have been applied to miRNA discovery (see (Demirci, Baumbach and Allmer 2017) for a review). Methods aremost often trained using data from miRBase as training data. Negative datasets are usually obtained from coding regions (or equivalently, exons), or by random sampling from whole genomes. It has been suggested that an average decision tree is best, and generalizes across species. Recently, a number of groups have attempted to determine which of the many proposed features are most discriminative. Agarwal *et al*. (Agarwal, *et al*. 2015) found that 3'-UTR site abundance, predicted seed-and downstream pairing stability, the base at position 1 and 8 of the miRNA seed sequence, the base at position 8 of the target site, local UA content, predicted structural accessibility, distance from the miRNA site to the stop codon of polyA site, site conservation, ORF length, 3'UTR length, and the number of offset sites in the UTR are the most important features in identifying miRNA targets. Lopes *et al*. (Lopes, Schliep and de Carvalho 2014) found that predicted minimum free energy index, ensemble free energy, normalized number of sequence variants, normalized Shannon entropy, and normalized base-pair distance were the most important features in a random forest approach. Tran *et al*. (Tran, *et al*. 2015) found, using a boosted support vector machine approach, that the most important features are predicted folding free energy of the longest nonexact stem, maximum number of consecutive G's in the longest nonexact stem, percentage of CC and GA dinucleotides, maximum number of consecutive C's, maximum number of consecutive G's in the hairpin, percentage of G–U pairs, folding free energy normalized by hairpin size, percentage of paired U, average predicted folding free energy, percentage of unpaired–unpaired A–paired triplets, and size of bulges. As just these

three examples show, there is considerable disagreement, even today, over what features are most relevant and powerful for miRNA target identification.

1.3.4 Databases/Resources

There are many online resources related to miRNA (for a recent review see (Singh 2017). A large fraction of these have been created for a particular organism or purpose, and then not updated. Below, we give our recommendations for the most useful and reliable resources (in our opinion).

- MiRBase [(miRBase 2016) (Kozomara and Griffiths-Jones 2013)] is the original miRNA resource and still hosts the miRBase registry which provides unique names for novel miRNA genes prior to publication. Release 21 of miRBase contains 28645 entries from 223 species, including extensive annotation of functions, experimental evidence, and links to other databases.
- DIANA-TarBase [(Paraskevopoulou MD 2016, DIANA-TOOLS 2016), (Vlachos, *et al*. 2015)] focuses on experimentally validated miRNAs and includes more than half a million experimentally supported miRNA-mRNA interactions. In addition TarBase includes computational predictions made with the MicroT-CDS method.
- Plant Non-coding RNA Database [(PNRD 2016) (Yi, *et al*. 2014) focuses on all types of non-coding RNAs in plants, not just miRNAs. Since miRNAs are structurally somewhat different in plants, a plant specific resource is sometimes useful. The earlier Plant MicroRNA Database (PMRD) appeared to be inactive at the time this article was written.
- Rfam [(Rfam 2017)] contains a large amount of information about miR families, including sequences, species of occurrence, secondary structure (usually predicted), and matching motifs.

1.4 Long non-coding RNAs[1]

1.4.1 Biological background

By definition, long noncoding RNA (lncRNA) collectively refers to transcribed RNAs longer than 200 nucleotides that have low coding potential. However, the 200 nucleotide threshold is an arbitrary threshold, which was selected based on a convenient biochemical cutoff in RNA isolation protocols. *BC1*and *snaR*, for example, are examples of ncRNAs that are shorter than 200

nucleotides but still classified as lncRNAs. Therefore, in 2011, Amaral *et al*. refined this definition: lncRNAs are noncoding RNAs that may have a function as either primary or spliced transcripts, that do not encode proteins, and are neither structural RNAs families (tRNAs, snoRNAs, spliceosomal RNAs, etc.), nor processed into known classes of small RNAs, such as microRNAs (miRNAs), piwi-interacting (piRNAs) and small nucleolar RNA (snoRNAs) [1]. Clearly this is still somewhat unsatisfying as lncRNAs are primarily defined as those that do not belong to known classes.

Because of the generous definition, lncRNAs are diverse in their biogenesis, stability, sub-cellular localization, evolutionary conservation, structure and function [(Ayupe, *et al*.2015), (Johnsson, *et al*. 2014)]. LncRNAs are typically capped, spliced, and poly-adenylated. Compared with mRNAs, lncRNAs have relatively lower expression level and lower stability. They are more tissue and cell-type specific, and are often expressed in a narrower developmental time window. LncRNAs are mostly located in the nucleus, presumably to regulate gene expression at the epigenetic level, but a minority of lncRNAs are present in the cytoplasm where they regulate translation. For example, *Xist* is a well-studied lncRNA that is involved in X inactivation in placental mammals. *Xist* is localized in the nucleus, and is highly expressed from the inactivated X chromosome at the onset of X chromosome inactivation. *Xist* binds at many locations in the inactivated X chromosome (by an, as yet, not well understood process) and recruits silencing factors such as the Polycomb repressive complex 2 (PRC2) to silence X chromosome genes (Brown *et al*. 1992; Clemson *et al*. 1996).

Beyond primates, little sequence conservation is typically observed in lncRNAs, unlike mRNA. The lack of conservation in the sequence of lncRNA does not indicate a lack of common function; An increasing number of examples have shown that lncRNAs are conserved in structure rather than sequence, and that the secondary (or higher) structure of lncRNAs constitutes the main

---

[1] This section has been published in a peer reviewed book. Sun, X. and Gribskov, M. (2019). MicroRNA and lncRNA Databases and Analysis. Encyclopedia of Bioinformatics and Computational Biology, vol.2, pp. 165–170. Oxford: Elsevier.

functional unit. For example, *HOTAIR* is a trans-acting lncRNA whose sequence is poorly conserved in mammals beyond primates (Bhan and Mandal 2016). Covariance analysis of 33 mammalian *HOTAIR* sequences revealed a significant number of covarying positions and half-flips localized in all four domains of *HOTAIR*, which act to maintain a similar structure (Somarowthu, *et al*. 2015).

According to NONCODE (v 5.0, a database of lncRNAs documented in the literature), 354,855 lncRNAs have been identified in 17 species. However, the functional roles of these lncRNAs remain mostly unknown. According to lncRNAdb (a database of eukaryotic lncRNA annotations), fewer than 300 have annotated functions confirmed by overexpression or knockdown experiments. LncRNAs, in general, can either repress or activate gene expression, and have been found to be associated with cell-fate programming ((Flynn and Chang 2014)) and numerous human diseases [(Esteller 2011)]. The number of lncRNA whose mechanisms are known in detail is even more limited, less than 20. But these examples have already shown that lncRNA is involved in important biological processes, such as genomic imprinting, chromatin remodeling, post-transcriptional RNA processing, and regulation of translation. Based on our current knowledge, lncRNAs consummate their regulatory roles in 3 major ways: 1). As decoys: that is they bind to regulatory proteins and preclude their access to DNA; 2). As scaffolds: they recruit epigenetic complexes to regulate chromatin states; and 3). As guides: the lncRNA binds proteins and guides the ribonucleoprotein complex to a target [(Rinn and Chang 2012)]. *PANDA* is an example of a lncRNA decoy. It sequesters a transcription factor called NF-YA, and keeps NF-YA from binding to its target genes, thereby preventing p53-mediated apoptosis [(Hung, *et al*. 2011)]. *HOTAIR*, which is located in the HOXC cluster, is an example of a lncRNA scaffold. It can simultaneously bind PRC2 in its 5' domain and LSD1 in its 3' domain. PRC2 has the function of histone H3 lysine-27 trimethylation, and LSD1 is involved in demethylation of histone H3 at lysine 4. This combination of interactions ensures epigenetic silencing of multiple cancer related genes [(Hajjari and Salavaty 2015)]. As mentioned above, *Xist* is an example of a lncRNA guide. *Xist* recruits Polycomb 1 and 2 complexes and guides them to the X chromosome targeted for inactivation to establish and maintain its silencing. Because the mechanisms of so few lncRNAs are known in detail, it is likely that many other mechanisms will be uncovered, ultimately revealing a more complex picture of the role of lncRNAs in regulatory networks.

1.4.2 Classification of lncRNAs

In GENCODE [(GENCODE 2017)], lncRNA is classified based on its genomic location with respect to nearby protein-coding genes. This is also one of the most commonly used methods to classify lncRNAs. Initially, lncRNAs were classified as either intergenic lncRNAs or intragenic lncRNAs. The transcripts of Intergenic lncRNAs (lincRNAs) do not overlap protein coding transcripts, while intragenic lncRNAs are transcribed from regions that overlap protein coding genes and can be further classified into sense and antisense lncRNAs. Sense lncRNAs are transcribed from regions of protein-coding genes on the same strand as the mRNA. They can overlap with both introns and exons of protein-coding genes. Totally Intronic RNAs (TINs), are lncRNAs that are located entirely within intronic regions of protein-coding genes. Partially Intronic RNAs (PINs), [(Nakaya, *et al*. 2007)] are lncRNAs that partially or entirely cover the introns of protein-coding gene. Antisense lncRNAs, or Natural Antisense Transcripts (NATs), are lncRNAs transcribed from the opposite strand of protein-coding genes.

Another way to classify lncRNAs is to distinguish their roles in the regulation of gene expression, distinguishing cis-acting and trans-acting RNAs. Cis-acting lncRNAs regulate the expression of genes that are positioned at the same, or a nearby, genomic locus. They may function through transcriptional interference or chromatin modification. Promoters and enhancers are two natural targets of cis-regulatory lncRNAs, which can recruit transcription factors, or chromatin modification complexes which remodel the structure of adjacent protein coding genes, to increase transcription. Promoter lncRNAs (sometimes called bidirectional promoter lncRNAs), plncRNAs, are transcribed from regions near the transcription start site (usually within 1500 bp of the transcription start site) of protein-coding genes, whereas enhancer lncRNAs (elncRNAs) may be located up to 1 Mbp upstream or downstream of the regulated gene. Trans-acting lncRNAs can control the expression of a gene at independent loci, for example, genes on a different chromosomes.

1.4.3 Discovery and exploration of lncRNAs

Determining the nature and possible biological functions of lncRNAs has become a focus of intense research. Expression profiling is often a first step in uncovering the function of a lncRNA. Identifying differentially expressed lncRNAs in developmental stages or conditions can imply

their potential functions. Alternatively, an informatic method termed “Guilt by Association” identifies functions of lncRNAs by looking for protein-coding genes whose expression are significantly correlated with those of lncRNA [(Guttman, *et al*. 2009).

Even though some researchers have successfully identified lncRNAs using polyadenylated RNA sequencing (mRNA-Seq), total cellular RNA sequencing (total RNA-Seq) is the usually the method of choice for comprehensive expression profiling of lncRNAs. This is because some lncRNAs, particularly lncRNAs, may not be spliced or polyadenylated. By using total cellular RNA-Seq, both mRNAs and lncRNAs can be identified, regardless of whether they are polyadenylated.

In the following section, we provide a computational pipeline for the identification of lncRNAs using total RNA-Seq data.

1. An appropriate reference assembly must be identified or constructed.
    a. If using a reference genome, reads are first mapped to the genome using an intron aware mapper (*e.g.*, using Tophat2 [(Kim, *et al*. 2013)]). Transcripts from different samples are merged (*e.g.*, using cufflinks [(Trapnell, *et al*. 2012)])
    b. If not using a reference genome, reads from all samples are combined to construct a *de novo* transcript assembly (*e.g.*, using Trinity [(Grabherr, *et al*. 2011)])
2. Reads from the individual samples are separately mapped to the reference.
3. Possible protein coding transcripts are excluded by multiple filtering steps, for example, by removing
    a. annotated protein coding transcripts,
    b. transcripts with high coding potential (*e.g.*, using the Coding Potential Calculator [(Kong, *et al*. 2007)]),
    c. transcripts with highly conserved known proteins or motifs (*e.g.*, using BlastX [(Altschul, *et al*. 1997)]),
    d. transcripts that have a high rate of synonymous versus nonsynonymous substitutions (*e.g.*, using PhyloCSF [(Lin, Jungreis and M 2011)])
4. ChIP-Seq (Chromatin Immunoprecipitation Sequencing) can be used to identify lncRNAs involved in gene activation involving transcription factors, or histone modification.

1.4.4 lncRNA databases/resources

Many online resources are available for lncRNAs. As with miRNAs, the spectrum of resources rapidly changes as many databases are created for particular purposes, but not maintained over time. In Table 1, we list a few of the currently active resources. Online searches for "lncRNA database" or similar terms will typically provide an updated list of resources, and a list is also maintained on Wikipedia (see the source citation in Table 1).

Table 1 lncRNA databases and resources

Database	Species	Last Update	Description (URL http://)
lncRNAdb	69 species	23-Nov-15	Includes lnc RNAs shown to be functional by overexpression or knockdown experiments. (www.lncrnadb.org)
RNAcentral	37 species	1-Apr-17	Combines 25 well maintained ncRNA databases. Provides integrated text search, sequence similarity search, and programmatic data access. (rnacentral.org)
NONCODE	17 species	6-Sep-17	Includes lncRNAs from published literature, GenBank, and specialized Databases such as Ensembl, RefSeq, lncRNAdb and LNCipedia. Functions of lncRNA are predicted by lnc-GFP. (www.noncode.org)
LNCipedia	Human	4-May-17	Includes 146,742 annotated human lncRNAs. Provides basic transcript information, predicted secondary structure, calculated protein coding potential, and predicted microRNA binding sites. (lncipedia.org)
GreeNC	45 species	19-Sep-16	Includes lncRNAs annotated in plants and algae that are identified by using self-developed pipelines. Provides information about sequence, genomic coordinates, coding potential, and predicted folding energy. (greenc.sciencedesigners.com)

Table 1 continued

PLAR2	17 vertebrates	Unknown	Includes lncRNAs identified using self-developed pipelines. 3P-seq information are included. (www.weizmann.ac.il/Biological_Regulation/ IgorUlitsky/pipeline-lncrna-annotation-rna-seq-data-plar)
LncRNADisease	human	26-Jul-17	Includes experimentally supported lncRNA-disease association data and lncRNA interactions in various levels, including protein, RNA, miRNA, and DNA. (www.cuilab.cn/lncrnadisease)
Lnc2Cancer	human	4-Jul-16	A manually curated database that include 1488 entries of associations between 666 human lncRNAs and 97 human cancers. (www.bio-bigdata.com/lnc2cancer)

Source: Wikipedia (http://en.wikipedia.org/wiki//List_of_long_non-coding_RNA_databases).

1.5 Enhancer RNAs

1.5.1 Biological background

The past decade has witnessed an explosion in the number of identified lncRNAs, which have been proven to be significant regulators of genome architecture and gene expression. In contrast to mRNA which functions as a mediator passing genomic information from DNA to protein, lncRNAs regulate gene expression in a variety of ways. Even though only a few lncRNAs have been characterized in detail, from well-studied cases (Rinn and Chang, 2012), it is obvious that lncRNAs regulate gene expression through interaction with chromatin to form a variety of RNA, DNA and protein complexes. For instance, lincRNA-p21, which is a lncRNA activated by transcription factor p53 and HIF-1α, regulates target gene expression by binding to the repressor protein hnRNP-K to effect hnRNP-K localization on genes in the p-53 dependent apoptosis pathway (Baldassarre and Masotti, 2012). A similar example is lncRNA Meg3, which recruits Polycomb Repressive Complex 2 (PRC2) to target genes via triple-helix formation, acting as a tumor-suppressor in pancreatic neuroendocrine tumor cells (Modali, *et al*, 2015).

Enhancers are DNA regulatory elements capable of activating their cognate promoters from a variable distance (from 100 bp up to Mbs (Mora *et al.,* 2016)) to up-regulate the transcription of a target gene. This is believed to occur by forming promoter-enhancer looping interactions. Multiple lncRNAs have been demonstrated to be transcribed within enhancers regions. In 2010, a study revealed that RNA polymerase II (Pol II) recruitment to active enhancers initiates widespread transcription of ncRNA in mouse cortical neurons (Kim *et al*, 2010). These RNAs were termed enhancer-derived RNAs (eRNAs). Since then, extensive efforts have been devoted to eRNA identification in a variety of cell types and species, and to their potential functions and mechanisms. Initially, eRNAs were thought to be merely transcriptional noise caused by high concentrations of Pol II. Recent studies have confirmed that eRNAs are essential for enhancer function. In particular chromosome conformation capture (3C) (Dekker *et al*., 2002) and related techniques (*e.g.,* circular chromosome conformation capture (4C) (Simonis *et al*., 2006; Zhao *et al*., 2006), chromosome conformation capture carbon copy (5C) (Dostie *et al*., 2006), Hi-C (Lieberman-Aiden *et al*., 2009), Combined 3C-Chip-Cloning (6C) (Tiwari and Baylin, 2009), chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) (Fullwood *et al*., 2009) and 3D DNA selection and ligation (3D-DSL) (Harismendy *et al*., 2011)), have shown eRNAs to be important for stabilizing enhancer-promoter looping by interacting with cohesion/mediator proteins and facilitating enhancer function. For example, eRNA CCAT1-L, transcribed from the human 8q24 gene desert region (long region of the genome that are devoid of protein-coding gene (Ovcharenko *et al*., 2005)) upstream of the MYC oncogene, plays an important role in regulating MYC transcription by promoting the formation of a long-range chromatin loop between the MYC promoter and enhancers (Xiang *et al*, 2014). Similar examples of eRNA involvment in chromatin-loop formation have been also observed in other studies (Fullwood *et al*., 2009; Yang *et al*., 2013; Lai *et al*., 2013; Li *et al*., 2013; Ren *et al*., 2017; Meng and Bartholomew, 2018), suggesting that eRNA isa group of ncRNA that function at the chromatin level to regulate target gene expression (Kim *et al*., 2015).

Enhancer RNAs are ncRNA that function at the RNA level and are not translated into proteins. Similar to lncRNA, eRNA expression is tissue specific (Heward *et al*, 2015;). Still, eRNA possesss unique features that can be used to distinguish them from the lncRNA. The first distinctive feature is their association with specific chromatin modifications of the originating genomic regions. Genomic regions from which eRNA originates from usually have higher levels of enhancer marks (H3K27ac and H3K4me1). And unlike canonical lncRNA and mRNA, these regions are thought

to be depleted in H3K4me3 marks, unless the enhancer is highly transcribed (Meng and Bartholomew, 2018). Second, while lncRNA undergoes maturation processes such as splicing and polyadenylation, eRNA is rarely spliced or polyadenylated. Third, eRN is relatively unstable and has lower expression level, making them difficult to capturd by traditional transcriptome profiling approaches such as RNA-seq (Kim *et al*., 2015; Wang *et al*., 2018). In addition, several studies have reported that highly transcribed eRNAs are a hallmark of active enhancers, as the expression levels of eRNA is positively correlated with the activity of enhancers (Wang *et al*, 2011; Hah *et al*, 2013; Andersson, 2015). These emerging features of eRNA have greatly expanded the complexityof genomic transcriptional regulation. As eRNA marks active enhancers, targeted sequencing and bioinformatic analysis of eRNA may be a useful approach to detect enhancers and investigate biological functions of enhancers.

1.5.2 eRNAs in disease

Disease-associated SNPs and recurrent somatic cancer mutations have been identified within enhancer regions (Murakawa *et al*., 2016), and many studies have shown that eRNA is differentially transcribed in various diseases (Yao *et al*., 2015; Le *et al*., 2017; Ren *et al*., 2017; Hauberg *et al*., 2018). Here we list several studies of eRNA in neurodegenerative diseases and cancer as examples.

In 2015, Yao *et al.* identified a robust set of tissue-specific eRNAs expressed in human brain, and showed that the enhancer regions from which these eRNAs are transcribed are enriched in genetic variants associated with autism spectrum disorders (Yao *et al*., 2015). A more recent study, in 2017, found that loss of RNA Pol II binding sites in enhancer regions in Huntington's disease mouse striatum contributes to reduced transcription of eRNA, resulting in down-regulation of target genes compared with healthy individuals (Le *et al*., 2017). Another study, published in 2018, examined RNA-seq data from 537 postmortem brain samples and identified 118 differentially transcribed eRNAs associated with schizophrenia. Furthermore, a genome-wide association study of schizophrenia indicated that a genetic variant in an enhancer region alters expression of both an eRNA and its target gene, suggesting the association of schizophrenia risk variants with eRNA (Hauberg *et al*, 2018). Collectively, these examples suggest that eRNA may be valuable as diagnostic markers and therapeutic targets for human neurodegenerative diseases.

The role of eRNA as a key regulatory non-coding RNA element in cancer has also been widely appreciated. An early study in 2013 reported that, in human breast cancer cells, 17β-oestradiol (E_2)-bound estrogen receptor α (ER-α) causes a global increase in eRNA transcription in enhancers adjacent to estrogen-induced upregulated coding genes. In combination with 3D-DSL methods, upregulation of eRNA transcription induced by estrogen was found to be associated with significantly increases in corresponding enhancer-promoter interaction, indicating that eRNAs play important regulatory roles in gene expression in cancer cells (Li *et al*., 2013). Chen and colleagues (2018) introduced a comprehensive approach to detection and characterization of eRNA using RNA-seq data from 8928 tumors across 33 cancer types. This study observed global enhancer activation in most cancers compared with matched normal tissues. Moreover, they successfully identified and validated the existence of an eRNA transcribed from ethe nhancer region of PD-L1 (a major cancer immunotherapy target), suggesting a clinical ipotential for eRNA (Chen *et al*., 2018). Mutations of enhancers that are associated with cancer have also been shown to be heritable. Bal and colleagues studied six families with Bazex-Dupré-Christol syndrome and identified germline mutations in enhancer regions around oncogene ACTRT1. These mutations presumably leadi to the impairment of enhancer activity, transcription from the enhancer, and expression of ACTRT1 (Bal *et al*., 2017).

In summary, the more we study eRNA, the more evidence has collected showing that eRNA are functional molecules that act at the chromatin level. Studies have also shown that eRNA expression is a hallmark of active enhancers (Wang *et al*., 2011; Andersson, 2015), and that the expression level of eRNA is positively correlated with enhancer activity (Chen *et al*., 2018).

1.5.3 Identification of eRNAs

Currently, there is no direct way to isolate and sequence eRNA les. Still, in the past decade, several approaches have beeb applied to detect eRNA using sequencing approaches.

The first approach is to sequence the nuclear transcriptome or the total. Considering that the majority of eRNA remains in the nucleus and is not polyadenylated (Wang *et al*., 2008), poly-A+ RNA-seq is not an appropriate sequencing method. Originally, eRNAs were detected by total RNA-seq (Kim *et al*., 2010), which is still the most commonly used sequencing method in eRNA detection because of its low cost and simplicity. However, total RNA-seq basically sequences all

types of RNAs (mRNA plus multiple forms of noncoding RNA). Because the abundance of eRNA is 19-34-fold lower than that of mRNA (Liu, 2016), detection of eRNA requires higher sequencing depth than does conventional RNA-seq (Murakawa *et al*., 2016). More recently, several more sensitive sequencing methods have been used to detect eRNA, *e.g.*, cap analysis gene expression (CAGE) (Andersson *et al*., 2014), TSS-seq (Yamashita *et al*., 2011), and paired-end analysis of TSSs (PEAT) (Ni *et al*., 2010). These methods define a snapshot of the 5′ end of transcripts, which makes it possible to precisely locate the position of eRNA transcription initiation. However, these methods only work for 5' capped mature RNAs, and are biased toward detection of stable transcripts. Such approaches are not suitable for detecting actively degraded eRNA (De Santa *et al*., 2010). Recently, nascent RNA sequencing technologies, such as global nuclear run-on sequencing (GRO-seq), precision run-on sequencing (PRO-seq) and native elongating transcript sequencing (NET-seq), have been used to detect eRNA (Wang *et al*., 2011; Kwak *et al*., 2013; Mayer *et al*., 2015). GRO-seq, which is the most widely used method to measure nascent RNAs, assesses transcription from engaged Pol II by sequencing transcripts from Pol II re-initiated transcription *in vitro* (Gardini, 2017). PRO-seq is an improved method, based on GRO-seq, that =sequences the 3' end of the nascent RNA, and maps Pol II active sites with single nucleotide resolution (Mahat *et al*., 2016). NET-seq determines the 3' end of nascent Pol II bound RNAs to detect actively transcribed RNAs at single nucleotide resolution (Churchman and Weissman, 2012). Even though these methods can efficiently detect unstable nascent RNAs, including eRNAs, they all require elaborate *in vitro* experimental procedures and are relatively technically challenging.

A second approach is to annotate eRNA using classic enhancer features. For example, high levels of H3K4me1, H3K27ac, and p300 binding are epigenomic marks that have been widely used to annotate enhancers. Noncoding RNAs (sometimes intergenic RNAs) coincident with these marks are annotated as eRNA. These epigenomic marks can be measured by chromatin immunoprecipitation (ChIP) coupled with DNA sequencing (ChIP–seq) (Heintzman *et al*., 2007). Note that while these epigenomic marks are useful and informative, their levels only describe the chromatin state of genomic regions, and they are indirect indicators of enhancers. Additional novel methods are still needed to directly identify enhancers and annotate

Many targets of eRNA are expected to occur in adjacent genomic regions, *i.e.*, eRNA is often cis-regulatory. Therefore, analysis eRNA expression patterns together with that of adjacent protein-coding genes has been used to predict eRNA target. If knocking down on eRNA results in repression of a nearby protein-coding gene, the eRNA is likely to be cis-regulatory.

1.5.4 Database/resources

Presently, there is no database for eRNA. However, there are several enhancer databases and eRNA can be identified as RNA transcribed from enhancer regions. Most of current available enhancer databases are collections of predicted enhancers rather than enhancers validated *in vivo*. In addtion, these databases mainly focus on tabulating enhancer regions in the human genome. Here we introduce several enhancer databases that have been widely used in enhancer/eRNAs studies.

VISTA Enhancer Brower (https://enhancer.lbl.gov/) (Visel *et al*., 2007) is a central resource for experimentally validated human and mouse noncoding DNA fragments with enhancer activity. Candidate enhancer fragments in the human genome were first selected based on their extreme conservation in other vertebrates, or based on epigenomic evidence (ChIP-Seq identification ofof putative enhancer mark), and then were validated by *in vivo* experiments in transgenic mice. This database provides a valuable resource of experimentally validated enhancers and has been used as the gold standard in many enhancer prediction methods. However, since enhancers are specific to different developmental stages, this database has limitation that it only includes enhancers that are active at the time points that are examined by VISTA. As of 3/20/2019, this growing database contains 2963 *in vivo* tested DNA fragments, of which 1597 have enhancer activity.

FANTOM5 Human Enhancers (also called Human Transcribed Enhancer Atlas) (http://slidebase.binf.ku.dk/human_enhancers/) (Andersson *et al*., 2014) is a database describing a collection of predicted active enhancer regions defined by CAGE-based bidirectional transcription in the FANTOM5 projects. The FANTOM5 project carried out CAGE sequencing on RNAs isolated from every major human organ, over 200 cancer cell lines, 30 time courses of cellular differentiation, mouse developmental time courses, and over 200 primary cell types, making it possible to classify both cell-type-specific and ubiquitous enhancers. DNA regions that are not associated with promoters but are identified as the source of balanced transcription on both strands

(indicated by CAGE signals) are predicted to be active enhancer regions. In total, this database contains 43,011 predicted active enhancers.

ChromHMM (http://compbio.mit.edu/ChromHMM/) (Ernst and Kellis, 2012) is a Hidden Markov Model based software that integrates multiple chromatin marks, such as those from ChIP-seq, to characterize chromatin states (such as enhancer, promoter, transcribed regions, and repressed regions) for each 200-bp genomic segment. ChromHMM has been applied to 111 Roadmap primary cell lines and 16 ENCODE cell lines to predict multi-cell activity profiles for chromatin state, gene expression and regulatory motif enrichment. The correlation between these profiles has also been used to predict cell-type-specific enhancers as well as target genes. These profiles can be retrieved from the website. In addition, users can also analyze their own files to get profiles of chromatin states in other cell types following the published protocol (Ernst and Kellis, 2017).

Segway 2.0 (https://omictools.com/segway-tool) (Chan *et al*., 2018) is another chromatin state annotation software utilizing ChIP-seq or DNase-seq signals. It employs a Dynamic Bayesian Network (DBN), which takes input of the ChIP-seq or DNase-seq signals at 1-bp resolution in contrast to 200-bp resolution for ChromHMM. However, the increased resolution comes at the expense of computing efficiency. Moreover, this tool is not an open source tool and cannot be used for free.

EnhancerAtlas (http://www.enhanceratlas.org/) (Gao *et al*., 2016) is an interactive database that contains 2,534,123 predicted enhancers for 76 human cell lines and 29 tissue types. Enhancers are predicted in each cell type by summation of at least three independent high throughput experimental datasets (*e.g*., DNase-seq, Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE)-seq, eRNA, P300 binding sites, POLII binding sites, histone modifications, transcription factor binding sites and CHIA-PET) with relative weights derived from a cross-validation approach. This database also predicts enhancer targets by integrating four features (*i.e.,* enhancer and promoter activity correlation, transcription factor and promoter activity correlation, enhancer and promoter sequence co-evolution, and enhancer-to-promoter distance) using a Random Forest classifier. In addition to providing profiles of predicted enhancers, this database is also an interactive platform that allows users to (1) examine predicted enhancers in a specific genomic

regions; (2) examine predicted enhancers associated with specific genes (3) compare predicted enhancers across different cell/tissue types; (4) predict and build enhancer-gene networks.

1.6 Chromatin enriched RNAs

1.6.1 Discovery and general features of cheRNA

To provide some general insights into this mode of regulation and the mechanism of chromatin interaction for lncRNAs in nuclear, Ruthenburg and colleagues isolated lncRNAs tightly associated with chromatin in the nucleus by biochemical fractionation of the nuclear compartment followed by RNA-seq (Werner and Ruthenburg, 2015). They first extracted nuclei with a denaturing urea/detergent buffer to separate and purify soluble and loosely bound material (soluble-nuclear extract, SNE) from chromatin tightly-bound material (chromatin pellet extract, CPE). This nuclear fractionation procedure was adapted from Wuarin and Schibler (Wuarin and Schibler, 1994). RNA-seq was used to examine RNA levels in three biological replicates of SNE and CPE samples. Using *de novo* transcriptome assembly and differential expression analysis approaches, they identified 2,621 intergenic transcripts that were significantly ($p<0.05$) enriched in CPE in HEK293 cells. They termed these transcripts chromatin-enriched RNAs (cheRNAs).

They found that most of cheRNAs are tethered to chromatin by Pol II and that they are remarkably co-localized with protein-coding genes with spacing <50 kb along the chromosomal coordinate. Moreover, the presence of cheRNA is strongly correlated with the expression of the nearest protein-coding gene at a level similar to, or higher than, that of enhancers predicted by ChromHMM and FANTOM, suggesting that cheRNA are a class of RNA that functions similarly to eRNA. In addition, in a subsequent study, they demonstrated that cheRNA is transcribed in a cell-type-specific manner by profiling cheRNA in three divergent cell lines: HEK293, K562, and H1-hESC (Werner *et al*., 2017). However, there are several molecular characteristics that distinguish cheRNA from the canonical defined eRNA. First, most putative enhancer regions marked by high levels of H3K4me1 relative to H3K4me3 are bi-directionally transcribed (Lam *et al*., 2014), while transcription of cheRNA displays strand bias and high abundance of H3K4me3 marks. Second, the cheRNA identified in HEK293 have a median length of 2,110 bases, much longer than the median length of currently described eRNA (~350 bases). Finally, only a small proportion (~11%) of the FANTOM predicted eRNA loci overlap with cheRNA in any of the three

tested cell lines (HEK293, K562, H1-hESC). Despite these distinctions, the apparent functional similarity of cheRNA and eRNA provides a compelling reason to further investigate the role of cheRNA in activation of adjacent protein-coding genes.

1.6.2 Examples of cheRNA as cis-activator

To further the cis-activating function of cheRNA, Werner *et al*. used the nuclear fractionation-sequencing method to identify cheRNA in two Tier 1 cell lines (H1-hESC and K562) (Werner *et al*., 2017). In this study, they used CRISPRi to inhibit transcription of three cheRNAs that were highly expressed in the K562 nucleus. The three selected cheRNAs are: *BONIFACIO,* located 67kb downstream of the nearest protein-coding gene, *B3GNT2*; *PAINE* located 71kb downstream of its nearest protein-coding gene, *PDCD6IP;* and *ILYICH* located 19kb upstream of its nearest protein-coding gene, *IL6*. They observed that reduced transcription (60-95%)of cheRNA led to proportional decreases in the expression of the nearest protein-coding gene in two out of three cases, suggesting a model in which cheRNA acts as a transcriptional activator in cis.

In a further example, knockdown of another cheRNA *HIDALGO*, with CRISPRi decreased transcription of its nearest gene, *i.e.*, the gamma-1 fetal hemoglobin (*HBG1*) gene. To distinguish if this is the effect of read-through transcription, or of the *HIDALGO* cheRNA molecule itself playing a role, they specifically degraded *HIDALGO* RNA using antisense oligonucleotides (ASOs). They observed a significant decrease in *HBG1* transcription commensurate with the degree of *HIDALGO* knockdown, demonstrating that the cheRNA molecule itself plays the role to activate nearby gene transcription. To investigate whether *HIDALGO* plays a similar role as some eRNAs in maintaining a chromatin looping structure that facilitates contact between promoter and enhancer, Werner and colleagues performed chromatin conformation capture (3C) to study the interactions between *HIDALGO* and *HBG1*. They found that the promoter of *HBG1* forms contacts with the TSS of *HIDALGO*, and that this contact is diminished by ASO or CRISPRi depletion of *HIDALGO*. Together these results indicate that the cheRNA *HIDALGO* cis-activates *HBG1* transcription similarly to the mechanism of eRNA, that is by mediating the contact between enhancer and its target.

In addition, Werner *et al.* studied the evolution of cheRNA and found that 96% of cheRNAs in K562 cells, and 98% in H1-hESC cells, overlap with class I transposable elements (TE). Although this is not conclusive, the significant enrichment of TE suggests cheRNA may evolve from TE.

In summary, cheRNAs are operationally defined by statistically significant enrichment in chromatin after biochemical fraction of nuclei (Werner *et al*., 2017). With convincing examples, Werner *et al.* have shown that cheRNA has the potential to activate proximal protein coding genes through interaction with chromatin. However, cheRNA seems not to function using a single uniform mechanism. For example, knockdown of cheRNA *PAINE* does not lead to a decrease in expression of its adjacent protein-coding gene; the identified cheRNA *XIST* is a well-known repressor located on the X chromosome of the placental mammals that acts as a major effector of the X inactivation process. Future investigations will most certainly be needed to complete our understanding of this novel class of lncRNA.

CHAPTER 2. IDENTIFICATION OF CHROMATIN-ENRICHED RNAS USING FOUR PIPELINES

2.1 Summary

Long noncoding RNA (lncRNA) is enriched not only in the cell nucleus, but also within the chromatin-associated fraction (Quinodoz *et al*., 2014). Many nuclear lncRNAs affect coding gene expression and chromatin organization, and are important in diverse biological processes (Khalil 2009, Sun 2018). Nuclear RNA-seq has revealed thousands of potentially regulatory lncRNA, including chromatin-associated lncRNA (cheRNA) (Werner and Ruthenburg, 2015; Gayen *et al*., 2017; Werner *et al*., 2017). However, a thorough survey of analytic pipelines for nuclear RNA-seq has not been performed.

The answers to many important questions regarding nuclear RNA-seq data analyses and cheRNA identification remain elusive. First, nuclear RNA-seq library construction differs from other RNA-seq protocols (Figure 1.1a). These differences have significant consequences for the interpretation and analysis of the sequencing data (Griffith *et al*., 2015). For instance, sequencing of polyadenylated (polyA+) RNA may miss transcripts that are not usually polyadenylated, which includes many lncRNA. Total-RNA sequencing detects a higher proportion of lncRNA, but is more expensive and less efficient in quantifying coding-gene expression (Kumar *et al*., 2017). Moreover, because total RNA-seq basically sequences all types of RNAs (mRNA plus multiple forms of noncoding RNA) and the abundance of lncRNA is 19-34-fold lower than that of mRNA (Liu, 2017), detection of lncRNA requires higher sequencing depth than does conventional RNA-seq (Murakawa *et al*., 2016). For example, the numbers of detected transcripts differ when nuclear or total RNA is sequenced, with 30.0% (7.0 k out of 23.3 k) of the transcripts detected only by total RNA sequencing, and 15.9% only by nuclear (Figure 1.1b). This difference is unlikely to be simply due to sequencing depth because the median depth was 49M for four pooled total RNA samples and 33M for 22 nuclear RNA samples; the latter includes the 9 CPE and 9 SNE samples reanalyzed in this study (Table 4). Markers of transcriptional regulation including RNA polymerase II (Pol II) sites, transcription factor binding sites, cis-regulatory RNA structures, histone deacetylase, and histone enhancer hallmarks are common in the DNA corresponding to the 3700 RNAs detectable only by nuclear RNA-seq (Figure 1.1c). This observation agrees with

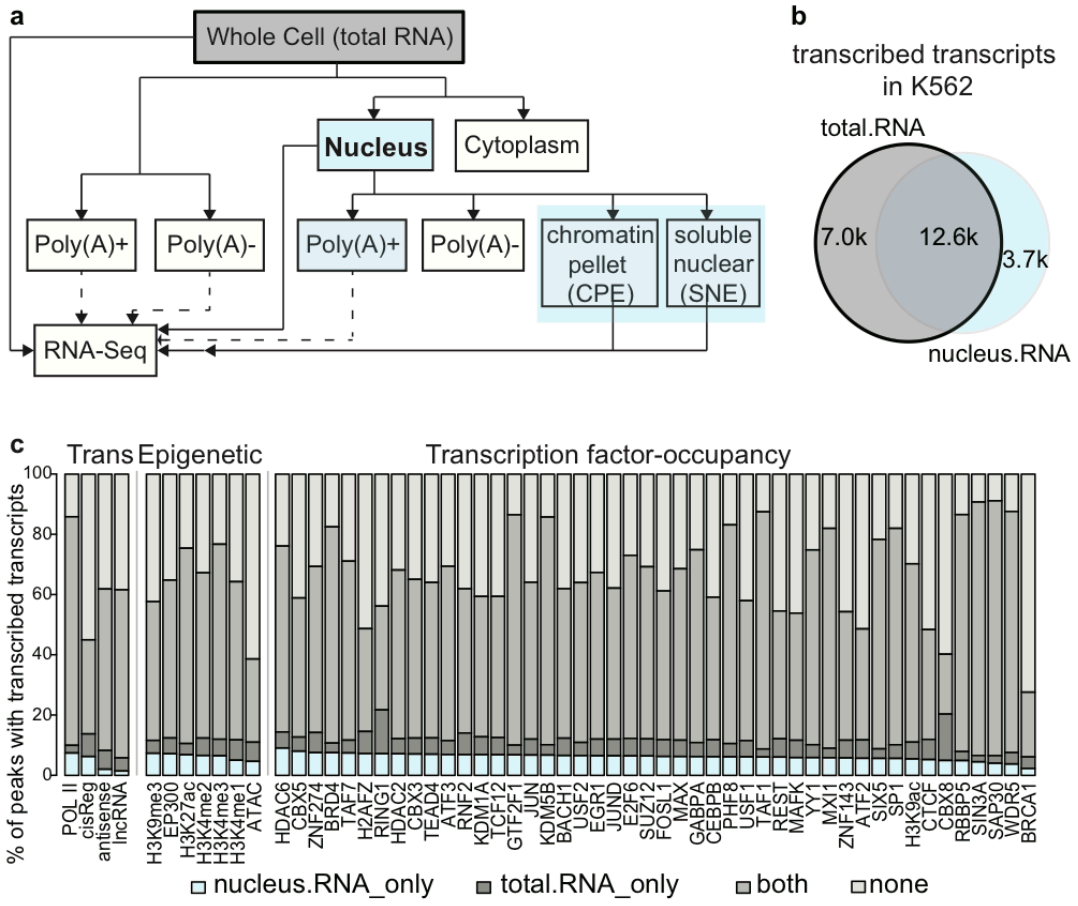


Figure 2.1 Nuclear RNA-seq sheds new insights into cis-regulatory elements.

(a) Diverse RNA-seq library strategies from parallel samples. Solid lines are the sequencing libraries (in category) analyzed in this study, and dashed lines are other available libraries. Blue color indicates the RNA-seq strategies to specifically sequence nuclear RNA. (b) Venn diagram comparing the number of predicted transcripts in two pooled RNA-seq transcriptomes. In both cases, RNA-seq data is from the K562 cell line with RNA-seq libraries from ENCODE and GEO. One transcriptome is the union of the total RNA-seq transcripts that were expressed with ENCODE transcript quantification value>0 in both replicates, in at least one of four collected samples, and the other is the union of the three types of nuclear RNA-seq transcripts (blue boxes in panel a, pooled from 22 samples (see Table 4). The latter includes either all detectable transcripts (those with non-NA values in the downloaded data) in both its replicates, or Tuxedo-assembled expressed transcripts with CPM ≥ 1. (c) Prevalence of epigenetic and transcriptional markers in nuclear and total RNA. Transcriptomic (Trans) loci were defined by the presence of ENCODE ChIP-seq peaks or similarity to anonotated Rfam families (lncRNA). Mark of interest (peaks), were compared with each transcriptome and assigned as occurring in both, only one, or neither (at least 1nt, ignoring transcript orientation). The assignment is indicated by the bar color Hallmarks are ordered according to the percentages of peaks overlapping with only the nuclear RNA transcriptome (darkest bar). CPE: Chromatin Pellet Extract; SNE: Soluble Nuclear Extract. Table 3 lists the data resources.

previous suggestions that nuclear-retained lncRNA may interact with chromatin regulatory proteins and recruit them to cis-regulatory elements in order to regulate gene expression (Quinodoz 2014, Sun 2018). Therefore, nuclear RNA-seq requires rigorous and effective pipelines different from the conventional pipelines used for total RNA-seq datasets.

In this study, we compare one published and three new analytic pipelines for nuclear RNA-seq data analysis. A newly developed pipeline, Tuxedo, outperforms the other pipelines with respect to transcriptome completeness, accuracy of cheRNA identification, and enrichment of enhancer-hallmarks at cheRNA gene regions.

2.2 Nuclear RNA-seq requires rigorous computational strategies

After the isolation of RNA and generation of sequencing libraries, a typical RNA-seq analytic workflow involves sequencing hundreds of millions of reads, alignment of reads against a reference genome or transcriptome, and downstream statistical analysis of expression. In the original method developed by Werner *et al.* (Werner and Ruthenburg, 2015; Werner *et al*., 2017), which we refer to as Werner, cheRNA was identified in the following steps: 1). chromatin pellet (CPE) transcripts were assembled using *de novo* transcript assembly with Cufflinks (Trapnell *et al*., 2010). Predicted transcripts from replicate samples were combined with Cuffmerge (Trapnell *et al*., 2010); 2). CPE predicted transcripts longer than 1000 bases were retained for further analysis and combined with soluble nuclear extract (SNE) predicted transcript assembled with Cufflinks relative to the reference GENCODE annotation (Harrow *et al*., 2012); 3). Differential abundance estimates of the combined transcript file, including replicate soluble nuclear and chromatin fractions, were made by Cuffdiff using standard options (Trapnell *et al*., 2012). This pipeline has three important biases: 1) Werner overestimates the proportion of *de novo* transcripts originating from CPE because it applies reference-guided *de novo* assembly (which can discover novel transcripts) to CPE but not to SNE fractions. 2) Werner removes transcripts shorter than 1000 bases from the analysis. LncRNA transcripts are typically shorter than (median length 592 bases) protein-coding transcripts (median 2.4k bases), and 33% of GENCODE-annotated noncoding RNA is shorter than 1000 bases long (Derrien *et al*., 2012). Removing transcripts shorter than 1,000 bases from the CPE assembly leads to significant under-detection of lncRNA. 3) In the differential expression analysis, Cuffdiff was used in Werner. However, Cuffdiff cannot do a two-

group test on RNAs that have high expression levels. For example, noncoding RNA *XIST*, which is a canonical cheRNA, was categorized as "HiDATA" and excluded from differential expression analysis by Cuffdiff. In addition, it has been shown that discarding genes that are not expressed at a biologically meaningful level in any condition (prefiltering) can increase the power for detecting differentially expression gene (Bourgon *et al*., 2010), but Werner doesn't include a prefiltering step in the differential expression analysis.

We developed three new pipelines (referred to here as Tuxedo, Concatenating, and Taco,) to analyze these datasets (Figure 2.2). The four major analytic steps in each pipeline are: sequence mapping, transcript assembly for sample, final transcriptome construction, and signature identification between CPE and SNE samples (APPENDIX B. Methods). The sequence mapping steps are the same in the three new pipelines; reads were mapped against the human genome version GRCh38.p10 using Tophat (v2.1.1) (Kim *et al*., 2013). In the transcript assembly step of the Concatenating, Tuxedo and Taco pipelines, we applied reference annotation-based transcript (RABT) assembly using Cufflinks, which assembles both annotated and *de novo* transcripta, independently for each sample and replicate. In the final transcriptome construction step, different strategies of combining the predicted transcripts were used in each pipeline. In the Concatenating pipeline, Cuffmerge (Trapnell *et al*., 2012) was used to separately merge the predicted transcripts from the three CPE replicates and three SNE replicates to produce separate CPE and SNE transcriptomes; then the CPE transcriptome and SNE transcriptome were combined to produce the final reference transcriptome used for differential expression analysis. In the Tuxedo pipeline, Cuffmerge was used to merge predicted transcripts from all CPE and SNE replicates to produce the final transcriptome. And in the Taco pipeline, we used TACO, a dynamic programming approach reported to outperform existing software tools (Niknafs *et al*., 2017), to assemble and merge all predicted transcripts from all CPE and SNE replicates to produce the reference transcriptome (more details can be found in APPENDIX B. Methods). In the last step, the identification of differential expression signatures between CEP and SNE samples, the Concatenating, Tuxedo and Taco pipelines used the model-based statistical Limma package, to identify differentially expressed transcripts. Limma has been shown to have higher precision and shorter runtimes than Cuffdiff and DESeq (Seyednasrollah *et al*., 2015). Hereafter, we evaluate the performance of the three new pipelines and Werner using the K562 dataset, which has the largest set of well described genomic features.

2.3 Identification of cheRNAs using four pipelines

We compared four pipelines, namely Werner, Concatenating, Tuxedo, and Taco, for the identification of chromatin enriched RNAs (Figure 2.2). Werner was executed by strictly following the analysis steps published in Werner's paper (Werner and Ruthenburg, 2015). Briefly, there are four steps in each pipeline: sequence mapping, transcript assembly, transcriptome construction, and signature identification.

The sequence mapping step is the same in all four pipelines. First, the reads were mapped against the human genome version GRCh38.p10 using Tophat (v2.1.1) (Kim *et al*., 2013) with default parameters for stranded RNA-seq libraries (e.g., tophat -p 8 --library-type=fr-firststrand -G gencode.v25.gtf GRCH38.genome -o CPE1 CPE1.fastq).

Strategies used in transcript assembly varied in four different pipelines. In the Werner pipeline, de novo assembly was applied only on the three biological Chromatin Pellet Extract (CPE) replicates using Cufflinks (v2.2.1) (Trapnell *et al*., 2012) (e.g., cufflinks -p 8 -u -N -library-type fr-firststrand -o cufflinks_CPE1 CPE1.bam), while reference-guided assembly (e.g., cufflinks -p 8 -u -N -library-type fr-firststrand -G gencode.v25.gtf) were applied on the three biological Soluble Nuclear Extract (SNE) replicates. In Concatenatin, Tuxedo and Taco pipelines, we independently applied the reference annotation-based transcript (RABT) assembly, which assembles both known and novel transcript. Specifically, we run the RABT assembly on the three biological CPE replicates and three biological SNE replicates by Cufflinks (Trapnell *et al*., 2012) using "cufflinks -g" option with GENCODE (v25) annotation as reference (e.g., cufflinks -u -N -library-type fr-firststrand -g gencode.v25.gtf -o cufflinks_CPE1 CPE1.bam).

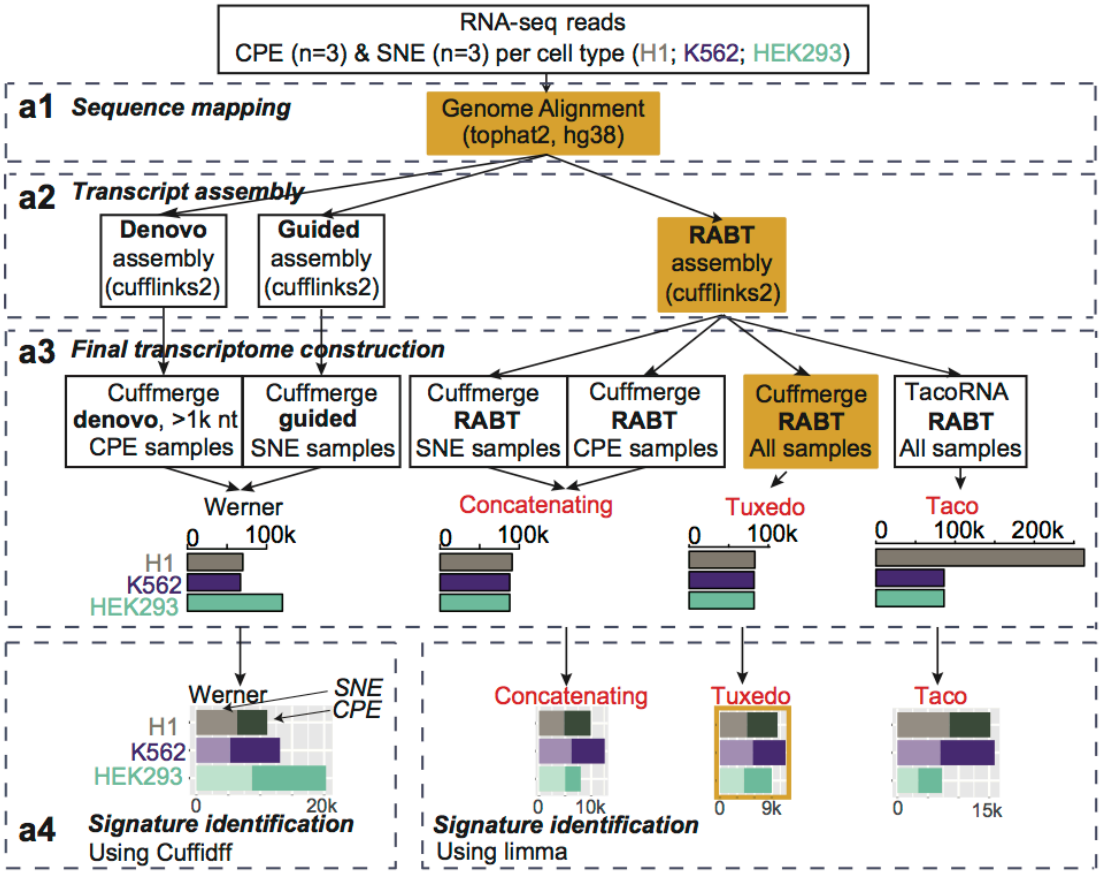


Figure 2.2 Four nuclear RNA-Seq analytic workflows.

Workflow of the four nuclear RNA-Seq analytic pipelines. Werner pipeline is executed by strictly following the analysis steps described by Werner et al. (Werner 2015). There are four major analytic steps in each pipeline are: (a1) Sequence mapping; Tophat is used to map RNA-seq reads in each sample/replicate are against the human genome version GRCh38.p10 (S Methods). (a2) Transcript assembly; Cufflinks is used to apply denovo/annotation guided/RABT (reference annotation based transcript) assembly on mapped reads (S Methods). Annotation guided assembly only assembles annotated transcripts included in provided GTF files. RABT assembly assembles both annotated transcripts and unannotated transcripts. (a3) Final transcriptome construction; Cuffmerge/TACO is used to merge assembled transcripts from all replicates/samples to construct final transcriptome GTF file (S Methods). Bar plot represents the number of "expressed" transcripts (CPM ≥ 1 in at least two samples); color indicates the assembly result in different cell line (H1: grey, K562: purple, HEK293: green). (a4) Signature identification; Cuffdiff/limma is used to identify RNAs that are differentially expressed between CPE and SNE samples (S Methods). Stacked bar plot represents the number of RNA with different abundance in the CPE/SNE samples; within each line a lighter color represents abundance in SNE and a darker color represents abundance in CPE.

Strategies used to construct transcriptome also varied in four different pipelines. In Werner, we did:

1) CPE replicates and SNE replicates were separately combined using Cuffmerge (v2.2.1) (Trapnell *et al*., 2010), resulting in both CPE transcriptome (e.g., cuffmerge -p 8 -o CPE_cuffmerge CPE_transcripts.txt) and SNE transcriptome (e.g., cuffmerge -p 8 -o SNE_cuffmerge SNE_transcripts.txt).
2) Reused transcript identifiers in the CPE transcriptome 'XLOC_' were renamed to 'CLOC_' to differentiate them from the transcript in the SNE transcriptome.
3) Only transcripts longer than 1000bp were kept in CPE transcriptome. – Note that this sub-step is specific to the Werner pipeline according to the author which may cause a bias to longer transcripts (Werner and Ruthenburg, 2015)
4) Two .bed files of CPE transcriptomes ('CLOC_') and SNE transcriptomes ('XLOC_') were obtained from their respective Cufflinks output .gtf files using gtf2bed in BEDOPS (v2.4.28) (Neph *et al*., 2012).
5) Next, we retrieved CPE-unique transcriptome using intersectBed (e.g., intersectBed -s -v -a CLOC.bed -b XLOC.bed) in bedtools (v2.26.0) (Quinlan and Hall, 2010).
6) These CPE-unique transcriptomes were then combined with SNE transcriptomes using 'cat' command to build the transcriptome for differential expression analysis.

The other three pipelines were similar to Werner except the sub-steps 1) and 3). In Concatenating, we used Cuffmerge to merge three CPE replicates (e.g., cuffmerge -p 8 -o CPE_cuffmerge CPE_transcripts.txt) and three SNE replicates (e.g., cuffmerge -p 8 -o SNE_cuffmerge SNE_transcripts.txt) separately to get the CPE transcriptome and SNE transcriptome.

In Tuxedo, we used Cuffmerge to merge all CPE and SNE replicates together to make an annotation for differential expression analysis (e.g., cuffmerge -p 8 -o tuxedo_cuffmerge ALL_transcripts.txt).

And in Taco, we used Taco (Niknafs *et al*., 2017) to merge all CPE and SNE replicates together to build the transcriptome for differential expression analysis.

In the last step to identify cheRNA signatures, Werner used Cuffdiff (v2.2.1) (Trapnell *et al*., 2010) with standard options (cuffdiff -p 8 -o cuffdiff.out --library-type fr-firststrand -L SNE,CPE -u combined_transcriptome.gtf K562_SNE1.bam, K562_SNE2.bam, K562_SNE3.bam K562_CPE1.bam, K562_CPE2.bam, K562_CPE3.bam). As a result, un-transcribed RNAs were identified as RNAs with a "NOTEST" value under "Test status" column in "gene_exp.diff" table. When contrasting expression levels in CPE samples to SNE samples, RNAs with FoldChange>1 and q_value<0.05 were identified as CPE-enriched RNA (cheRNAs) and RNAs with FoldChange<1 and q_value<0.05 were identified Soluble Nuclear-Extracted RNAs (sneRNAs).

In Concatenating, Tuxedo and Taco pipelines, we applied the same advanced computational strategy (limma) which generally showed higher precision and shortest runtimes than cuffdiff in RNA-seq data analysis (Seyednasrollah *et al*., 2015). Specifically, we did:

1) Used HTSeq (v.0.7.0) (Anders *et al*., 2015) to get the raw counts of transcripts (e.g., htseq-count -f bam -s no -m intersection-nonempty CPE1.bam tuxedo_transcriptome.gtf > CPE1_geneCounts.out).
2) Transformed the expression of RNAs from raw counts to counts per million (CPM). RNAs with CPM<1 are considered as un-transcribed. Only RNAs expressed in at least 3 out of 6 samples were retained for further analysis.
3) Normalization of RNA expression was performed by the method of trimmed mean of M-values (TMM).
4) Used limma package in R (Ritchie *et al*., 2015) to do differential expression analysis comparing CPE samples with SNE samples, per cell type. The expected FDR was estimated using the Benjamini-and-Hochberg method.
5) Transcripts having FDR<0.05 and FoldChange>1.2 were identified as chromatin enriched RNAs and transcripts having FDR<0.05 and FoldChange<0.83 were identified as chromatin depleted RNAs

2.4 Tuxedo builds a complete transcriptome for active transcripts

Lowly expressed transcripts are likely to be experimental noise (Hart et al., 2013). Unlike methods applied to coding gene profiles, in which one can define an expression cutoff for active promoters, we made an empirical decision to define predicted transcripts with CPM (counts per million) ≥ 1

as ‘expressed’ for downstream analysis. This filter resulted in an approximately log-normal distribution of expression levels and about 14 k measured transcripts per sample (Figure 2.3a), ensuring the appropriateness of model-based differential expressional analyses such as Limma (Ritchie et al., 2015).

To evaluate the transcriptomes assembled by the 4 pipelines, we used the assembly result of K562 cell data as a reference. We first compared the completeness of the transcriptomes. Transcriptomes assembled by the Tuxedo and Concatenating assemblies are very concordant. 99.8% of transcripts are the same. 84.4% of transcripts assembled by both Tuxedo and Concatenating are also assembled by Werner. This number decreases to 27.0% for Taco (Figure 2.4a). To determine the reasons for assembly inconsistency, we compared the assembly results for annotated transcripts (Figure 2.4b) and unannotated transcripts (Figure 2.4c) separately.

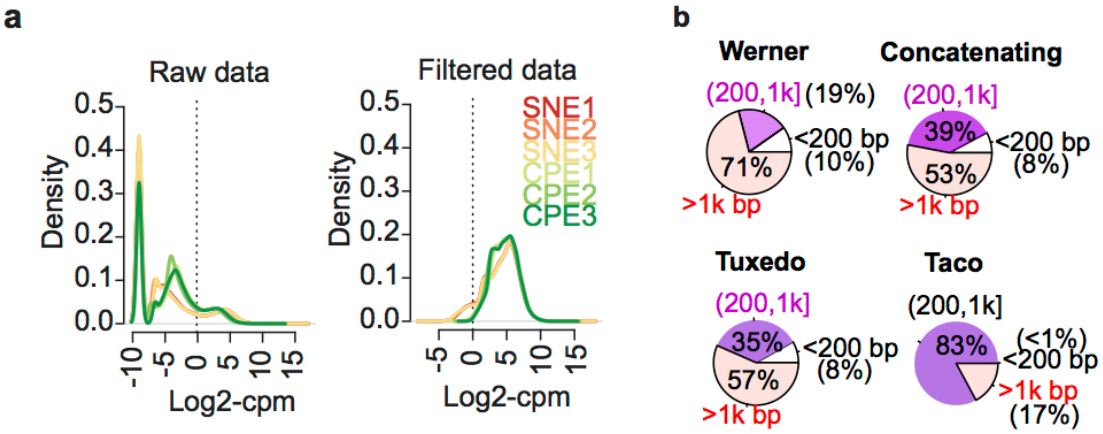


Figure 2.3 Noise filtering and transcript length.

(a) Density plot showing the consequence of filtering the lowest-expressed values by the Tuxedo pipeline. A nice bell-like shape of count distribution was observed after this filtering of noise. Line colors decode individual samples. (b) The width distribution of all transcripts built in the four pipelines showing Taco assembles relatively shorter transcripts, as 83% of its assembled RNAs are shorter than 1k bases.

The annotated transcripts assembled by Tuxedo, Concatenating and Werner pipelines are almost identical, and correspond to the set of annotated transcripts in GENCODE (v25). Taco only assembled 26.2% of annotated transcripts. This is because the Taco pipeline uses TACO instead of Cufflinks as the assembly tool. TACO only includes transcripts that have significant expression, while Cufflinks keeps all annotated transcripts when building the transcriptome. By looking at the

expression level of transcripts, we confirmed that the 40.7 k annotated transcripts that are omitted by Taco are transcripts with low expression levels in K562. Among the unannotated transcripts assembled by Tuxedo and Concatenating, 96.8 % are the same. Moreover, Tuxedo and Concatenating assembled more transcripts than Werner and Taco. The length distribution of transcripts assembled by Tuxedo and Concatenating, but not by Werner, shows that the majority of such transcripts are shorter than 1000 bases (Figure 2.4d), which is caused by removing transcripts shorter than 1000 bases from CPE samples in Werner. Taco assembled the smallest number of unannotated transcripts.

The unannotated transcripts omitted by Taco also have low expression. Even though these transcripts were lowly expressed in samples, it is still necessary to keep them in the assembled transcriptome to accurately estimate gene expression levels. We next investigated the length of assembled transcripts (Figure 2.3b, Figure 2.4e). Approximately half of the assembled transcripts have lengths between 200-1000 bases and show similar log-normal distributions for Tuxedo and Concatenating. In contrast, transcripts assembled by Werner are generally longer (71% of the assembled transcripts are longer than 1000 bases), which is another indication of the incompleteness caused by removing short transcripts. Transcripts assembled by Taco are much shorter (83% of assembled transcripts are shorter than 1000 bases). The TACO assembler employs an algorithm based on change-point detection via binary segmentation to predict transcript structure (Niknafs *et al*., 2017). This algorithm is more robust in assembly of annotated and conserved transcript such as mRNA. However, when it is applied to assembly of noncoding RNA, the TACO assembler overestimates the degree of alternative splicing and results in a large number of truncated transcripts. This is incorrect since only a small fraction of lncRNA undergo splicing (Tilgner *et al*., 2012). In summary, Tuxedo and Concatenating construct relatively complete and correctly structured transcriptomes for analysis.

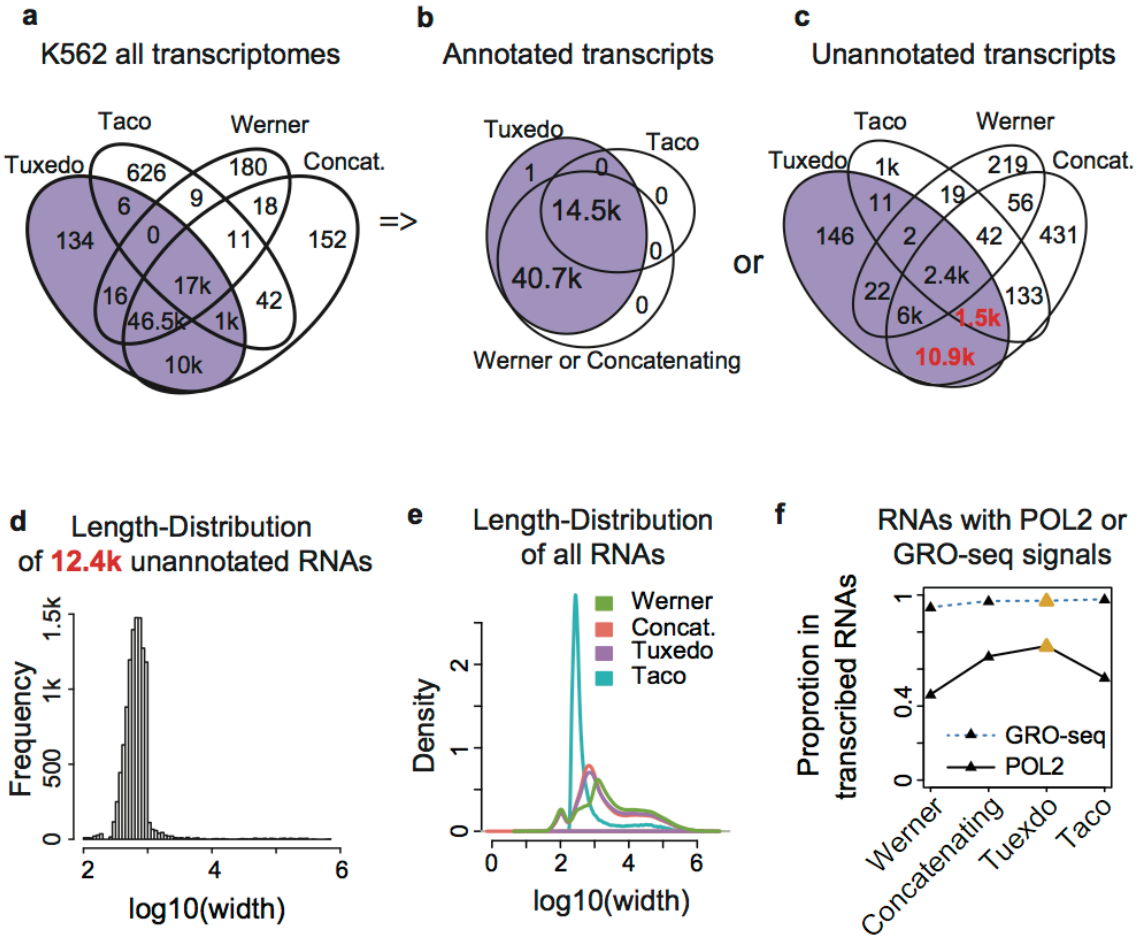


Figure 2.4 Tuxedo assembles a complete high-quality transcriptome.

(a) Overlap in predicted RNA classes in the K562 cell line. Venn-diagram showing coordinate-overlaps for all RNA predicted in the four pipelines. Numbers are calculated by the R package ChIPpeakAnno with the "findOverlapsOfPeaks" function to count number of overlapped transcripts. RNAs with an overlap of 1 base or more are considered to be overlapped. If multiple transcripts overlap in several groups, the minimal number of transcripts in any group is counted as the number of overlapping transcripts. (b) Overlap in annotated RNA in K562 cell line. Venn-diagram showing coordinate-overlaps for annotated or (c) unannotated RNA, being respectively constructed by 4 pipelines. (d) Length distribution of predicted RNA in the four pipelines. Color indicates different pipelines: Werner (green), Concatenating (Concat., red), Tuxedo (purple), Taco (blue). (e) Length distribution of unannotated RNA predicted by Tuxedo and Concatenating but not by Werner or Taco. (f) Proportion of expressed RNA (CPM ≥ 1) assembled in each pipeline that overlap (at least 1 base, same strand) by coordinate with any GRO-seq peak and POL II peak. Coordinate overlaps are calculated by using the R package GenomicRanges with the "findoverlaps" function. Expressed RNA in Tuxedo assembly has the highest proportion of overlap with peaks representing ongoing transcription (by Pol II), and nascent transcription (by GRO-seq).

Additionally, we compared the transcriptional activity of the expressed transcripts assembled by the 4 pipelines using two independent measurements: Pol II ChIP-seq and global run-on sequencing (GRO-seq) (Table 4). Expressed transcripts are defined as those having CPM ≥ 1. The expressed transcripts assembled by Tuxedo show the highest proportion of overlap with peaks representing both ongoing transcription by Pol II, and peaks representing nascent transcription by GRO-seq (Figure 2.4f), demonstrating that expressed transcripts assembled by Tuxedo are more concordant with active transcription signal represented by other methods.

2.5 Tuxedo outperforms in identifying cheRNAs

To evaluate the performance of the four pipelines in cheRNA identification, we used the set of transcripts identified by all methods as a proxy gold standard, and found Tuxedo and Concatenating outperformed Werner and Taco in the identification of both cheRNA and sneRNA (Figure 2.5a). To further check the accuracy we examined sixteen loci of known cheRNA, sneRNA, or chromatin-independent RNA (transcripts not significantly differentially expressed between CPE and SNE samples) that were previously validated in specific cell types (Werner and Ruthenburg, 2015; Werner *et al*., 2017). Tuxedo and Concatenating successfully confirmed the chromatin enrichment in all canonical cheRNAs and outperform Werner and Taco with overall positive predicted value (ppv) of 0.88 (Figure 2.5b**,** Table 2). This analysis, although possibly susceptible to threshold effects, makes up the shortage of lack of a truly gold standard in the ROC-analysis. Both analyses suggest that Tuxedo and Concatenating are better than Werner and Taco.

Because intergenic cheRNA (icheRNA**),** which are defined as cheRNA without no coordinate overlap with known coding genes, is similar to eRNA (Werner and Ruthenburg, 2015; Werner *et al*., 2017), we examined the occupancy of enhancer marks (ChIP-seq signals of EP300, H3K27ac, H3K4me1) and a repressive mark (H3K27me3) around the TSS of the 2.0 k to 6.7 k icheRNA identified by each pipeline (Figure 2.5c). In this analysis, we used ChromHMM (Ernst *et al*., 2016) predicted eRNA (Figure 2.5c, yellow) and non- enhancer RNA (Figure 2.5c**,** black) as positive and

Table 2 Canonical cheRNAs can be better identified by Tuxedo and Concatenating methods. Colored cells are validated in specific cell and used as gold standards.

16 alidations			HEK293				K562			
result	**cell**	**RNA Symbol**	Werner	Tuxedo	Concat.	Taco	Werner	Tuxedo	Concat.	Taco
CPE	HEK293	KCNQ1OT1	CPE	CPE	CPE	NA				
	HEK293	XIST	HD	CPE	CPE	CPE				
	HEK293/K562	PVT1	-	CPE	CPE	-	CPE	CPE	CPE	CPE
	K562	BONIFACIO					CPE	CPE	CPE	CPE
	K562	ILYICH					CPE	CPE	CPE	CPE
	K562	HIDALGO					-	CPE	CPE	CPE
SNE	HEK293	GAPDH	SNE	SNE	SNE	SNE				
	HEK293	ACTB	SNE	SNE	SNE	SNE				
	HEK293/K562	MYC	SNE	SNE	SNE	SNE	SNE	SNE	SNE	SNE
	K562	B3GNT2					SNE	SNE	SNE	SNE
	K562	IL6					-	-	-	NA
	K562	PDCD6IP					-	SNE	SNE	SNE
intermediate	HEK293	HOTAIR	SNE	-	-	NA				
	HEK293	DLX6-AS1, Evf-2_5p	-	CPE	CPE	NA				
correct prediction:			5	7	7	4	5	7	7	7
number of gold standarad:			8	8	8	8	8	8	8	8
PPV			0.63	0.88	0.88	0.50	0.63	0.88	0.88	0.88

NA: not tested because of low expression level
HD: not tested by Cuffdiff because of high expression level (HI-DATA)
"-": tested but not significantly differentially expressed
PPV: number of correct predictions /number of gold standard in each cell line (8 gold standard in HEK293 while 8 gold standard in K562)

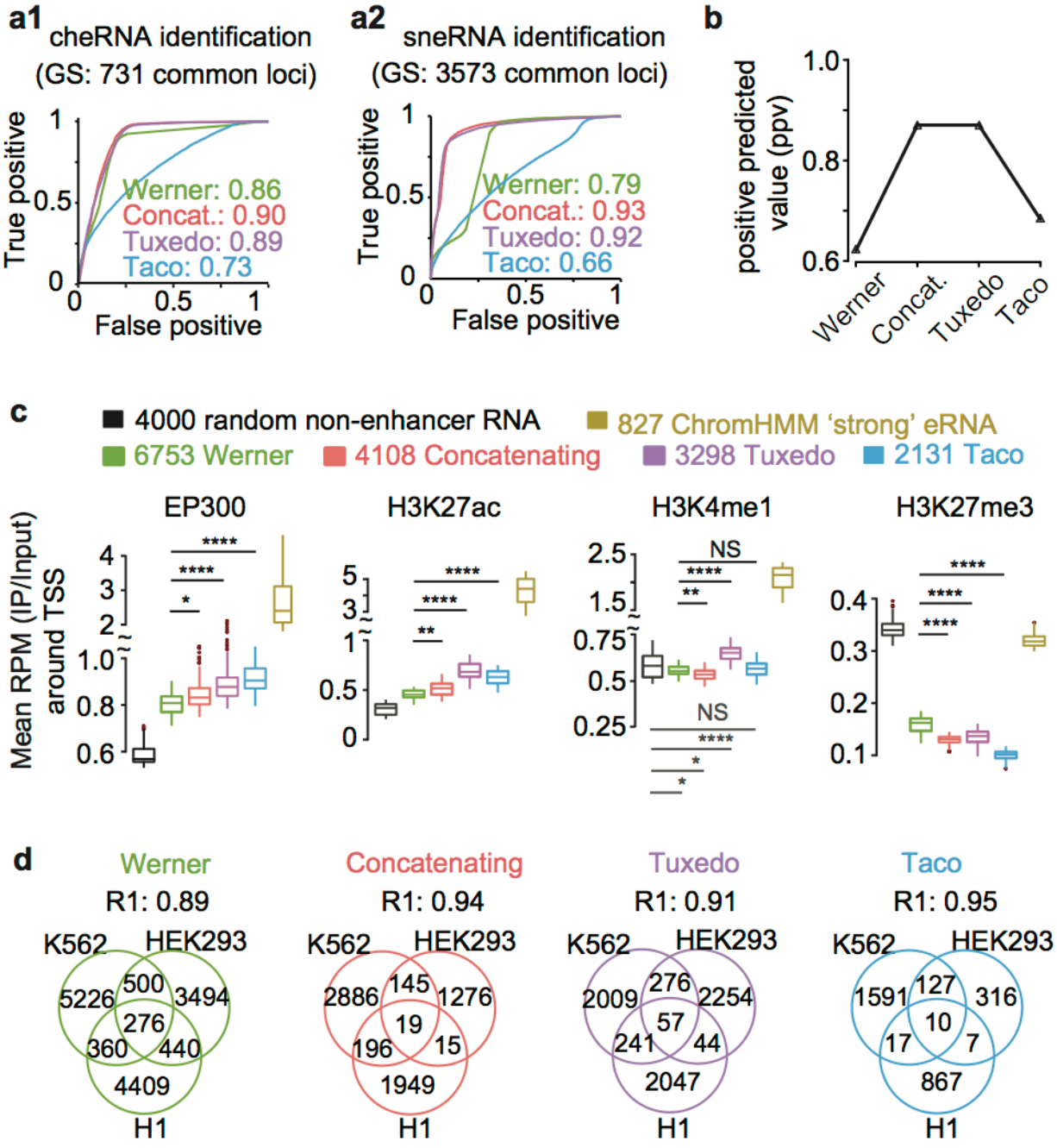
negative controls. ChromHMM-predicted eRNA is defined as intergenic RNA that overlaps (at least 1 base, same strand) with any ChromHMM predicted "strong enhancer" region and ChromHMM-predicted non-enhancer RNA is defined as transcribed RNA that has no overlap with any predicted "strong enhancer" or "weak enhancer" region. We found that the levels of enhancer

marks (EP300, H3K27ac, H3K4me1) are significantly higher around TSS of icheRNA than at the TSS of ChromHMM predicted non-enhancer RNA, while the level of repressive marks is significantly lower. Moreover, among the 4 pipelines, the icheRNA identified by Tuxedo pipeline have the highest levels of H3K27ac and H3K4me1 enhancer marks around their TSS, and relatively lower levels of repressive marks. We also noticed that the levels of enhancer marks around TSS of ChromHMM predicted eRNA are much higher than those around TSS of icheRNA. Considering that ChromHMM predicts enhancer regions based on histone modification patterns, the enhancers predicted by ChromHMM may be biased toward having high occupancy of these canonical enhancer marks. Additionally, all three new pipelines slightly improved the cell-type specificity compared to Werner, as evaluated by the proportion of tissue-specific icheRNA identified by each pipeline (represented by R1 score in Figure 2.5d).

Overall, we conclude that Tuxedo and Concatenating outperform the other two pipelines in identifying expected cheRNA, and that the Tuxedo predicted icheRNA transcripts are more highly enriched in enhancer hallmarks compared to other methods. In this sense, Tuxedo outperforms Concatenating and other pipelines in enriching enhancer hallmarks in the same cell type.

Figure 2.5 cheRNA prediction using the four pipelines in K562 cell line.

(a) Receiver operating characteristic (ROC) curves of four pipelines identifying cheRNA (a1) and sneRNA (a2). The commonly-identified 731 cheRNA or 3573 sneRNA by all four pipelines are the proxy gold standard (GS) used here. Color represents different pipeline: Werner (green), Concatenating (Concat., red), Tuxedo (purple), Taco (blue). The Tuxedo and Concatenating methods have the best performance, with AUC larger than 0.89 in both cheRNA and sneRNAs identification. (b) Average positive predicted value (ppv) in identifying sixteen canonical cheRNA/sneRNA/intermediate RNA (RNA not differentially expressed between CPE and SNE) experimentally verified in previous studies for chromatin-enrichment or depletion, using the four pipelines, respectively. Further details about these 16 loci are given in Table 2. (c) Average ChIP-seq read density around TSS (±1kb centered at TSS) of the indicated RNA classes in K562. Boxes span the lower to upper quartile boundaries, the median is indicated with solid line in each box. Color represents the icheRNA identified in four pipelines, and two control group of ChromHMM predicted RNAs in K562 (ChromHMM-predicted eRNAs as a positive control (yellow), and ChromHMM-predicted non-enhancer RNAs as a negative control (black)). ChromHMM-predicted eRNA is defined as intergenic RNA overlapped (at least 1 base, same strand) with any ChromHMM predicted "strong enhancer" region and ChromHMM-predicted non-enhancer RNA is defined as transcribed RNA that have no overlap with any ChromHMM-predicted "strong" or "weak" enhancers. (d) Fraction of cell-type-specific intergenic cheRNAs. R1 is the ratio of cell type specific RNAs versus all RNAs identified in each pipeline. Higher R1 value indicating more cell-type specific identification. Venn diagrams show the overlap of icheRNA identified in K562, HEK293 and H1-hESC cell lines by Werner (green), Concatenating (red), Tuxedo (purple) and Taco (blue). icheRNA identified by all the four pipelines except Werner (green) have high tissue-specificity (R1>0.9).


a1
cheRNA identification
(GS: 731 common loci)
True positive
False positive
Werner: 0.86
Concat.: 0.90
Tuxedo: 0.89
Taco: 0.73
a2
sneRNA identification
(GS: 3573 common loci)
True positive
False positive
Werner: 0.79
Concat.: 0.93
Tuxedo: 0.92
Taco: 0.66
b
positive predicted value (ppv)
Werner
Concat.
Tuxedo
Taco
c
4000 random non-enhancer RNA
827 ChromHMM 'strong' eRNA
6753 Werner
4108 Concatenating
3298 Tuxedo
2131 Taco
EP300
H3K27ac
H3K4me1
H3K27me3
Mean RPM (IP/Input) around TSS
NS
d
Werner
R1: 0.89
K562
HEK293
5226
500
3494
276
360
440
4409
H1
Concatenating
R1: 0.94
2886
145
1276
19
196
15
1949
Tuxedo
R1: 0.91
2009
276
2254
57
241
44
2047
Taco
R1: 0.95
1591
127
316
10
17
7
867

2.6 Discussion

Detail analysis of nuclear RNA-seq sheds new light on cis-regulatory elements (Figure 2.1c**)**. We have presented a computational pipeline, Tuxedo, for analyzing nuclear RNA-seq data containing both high low expression lncRNA (Figure 2.2). The Tuxedo pipeline makes three key computational improvements: 1) Tuxedo assembles the complete transcriptome in an unbiased way, covering both highly-expressed transcripts and lncRNA shorter than 1,000 bases. 2) Tuxedo employs an empirical threshold to distinguish between low but informative lncRNA transcription and noise. And 3) Tuxedo identifies cheRNAs precisely while recapturing three known genomic features of active enhancers. The strategies used in the Tuxedo pipeline are not restricted to cheRNA identification, and could be beneficial to nuclear RNA-seq data analyses testing broader biological hypotheses, such as to the relationship between enhancer marked and differentially expressed nuclear RNAs.

CHAPTER 3. INTERGENIC CHERNAS UNIQUELY PRESENT ERNAS FEATURES

3.1 Summary

Development of sequencing technology leads to a surprising increase in the discovery of noncoding RNA, especially miRNA and lncRNA. Unlike miRNA, the functioning mechanism of which has been well studied, most of lncRNAs are identified with no known function. Based on the studied lncRNA mechanisms, it is worth noting that lncRNA frequently functions at chromatin interface. eRNA is a subgroup of lncRNAs that are pervasively transcribed from enhancer regions and required for maintaining enhancer-promoter looping structure through chromatin interaction.

cheRNA is operationally defined by statistically significant enrichment in chromatin after biochemical fraction of nuclei. The previous work by Werner *et al*. showed that cheRNA correlates with neighboring gene transcriptional activity at a level similar to, or better than the current state-of-the-art active enhancer annotation (Werner and Ruthenburg, 2015). Perturbation of four distinct cheRNAs further suggest that cheRNA activates nearby genes through a mechanism similar with eRNA. To investigate if this similarity is widely existing among all cheRNAs, we used the suggested Tuxedo pipeline to undertake a more comprehensive examination of cheRNAs in HEK293, K562 and H1-hESC cell line. We discussed the similarity between cheRNAs and eRNAs in five aspects: 1). genomic localization; 2). coding potential; 3). RNA polyadenylation; 4). transcriptional correlation with nearby coding gene; 5). chromatin and histone signatures. We found that cheRNAs are mostly transcribed from intergenic regions. Compared to intergenic sneRNA (isneRNA), intergenic cheRNA (icheRNA) has lower coding probability, lacks polyadenylation, and its expression is more positively correlated with that of neighboring coding genes, suggesting that icheRNA rather than isneRNA is more similar to eRNA.

We also observed that icheRNA has a lower transcription level and is largely unannotated, while isneRNA is more highly transcribed and better annotated. This unbiased annotation in icheRNA and isneRNA suggest that the traditional transcriptome profiling of non-coding RNA (*e.g.* total RNA-seq) yields the broadest survey of transcripts but has limited ability to detect low expression transcripts such as those of icheRNA. Isolation and sequencing RNAs that tightly interacts with

chromatin in nuclear can identify a group of novel noncoding RNAs that has been largely overseen before.

In the aspect of chromatin signatures, we found that regions around TSS of icheRNA only show moderate level of active enhancer marks, which is consistent with previous observation by Werner *et al*. (Werner and Ruthenburg, 2015). This may indicate icheRNAs contain other RNA groups besides eRNA. Despite there are differences, icheRNA still show more apparent similarity to eRNAs than other RNA groups, which indicates that indentification of icheRNA provides a new way to annotate eRNA.

3.2 icheRNA represents a subset of noncoding RNAs de novo

Werner *et al* proposed that icheRNA is a distinct subclass of unannotated eRNA. To further examine this hypothesis, we categorized the 14k expressed nuclear RNAs detected by Tuxedo into three groups (intergenic RNA, coding-antisense RNA (labeled as “antisense RNA” in Figure 3.2a) and those that overlap mRNAs in the sense orientation (labeled as “mRNA” in Figure 3.2a).

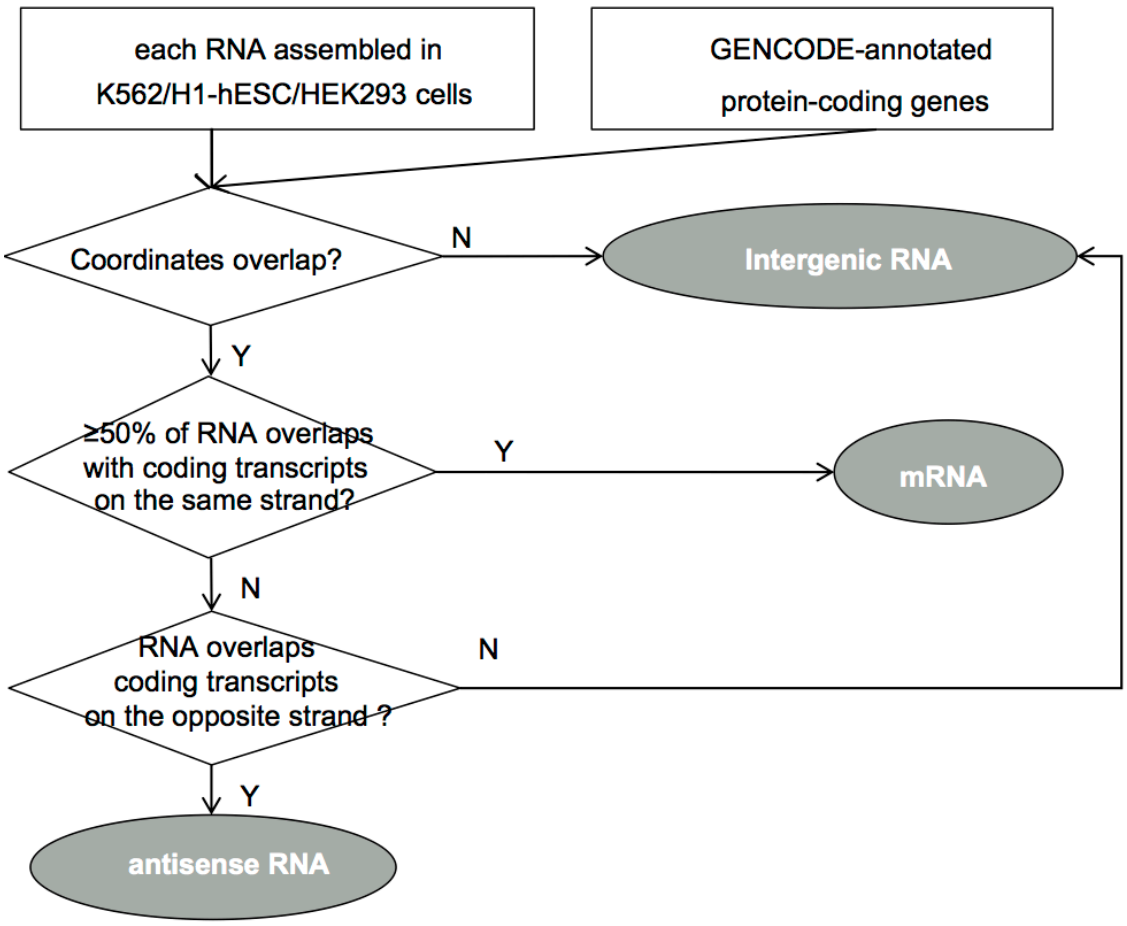


Figure 3.1 Workflow of categorizing RNA into mRNA, intergenic RNA, or antisense RNA.

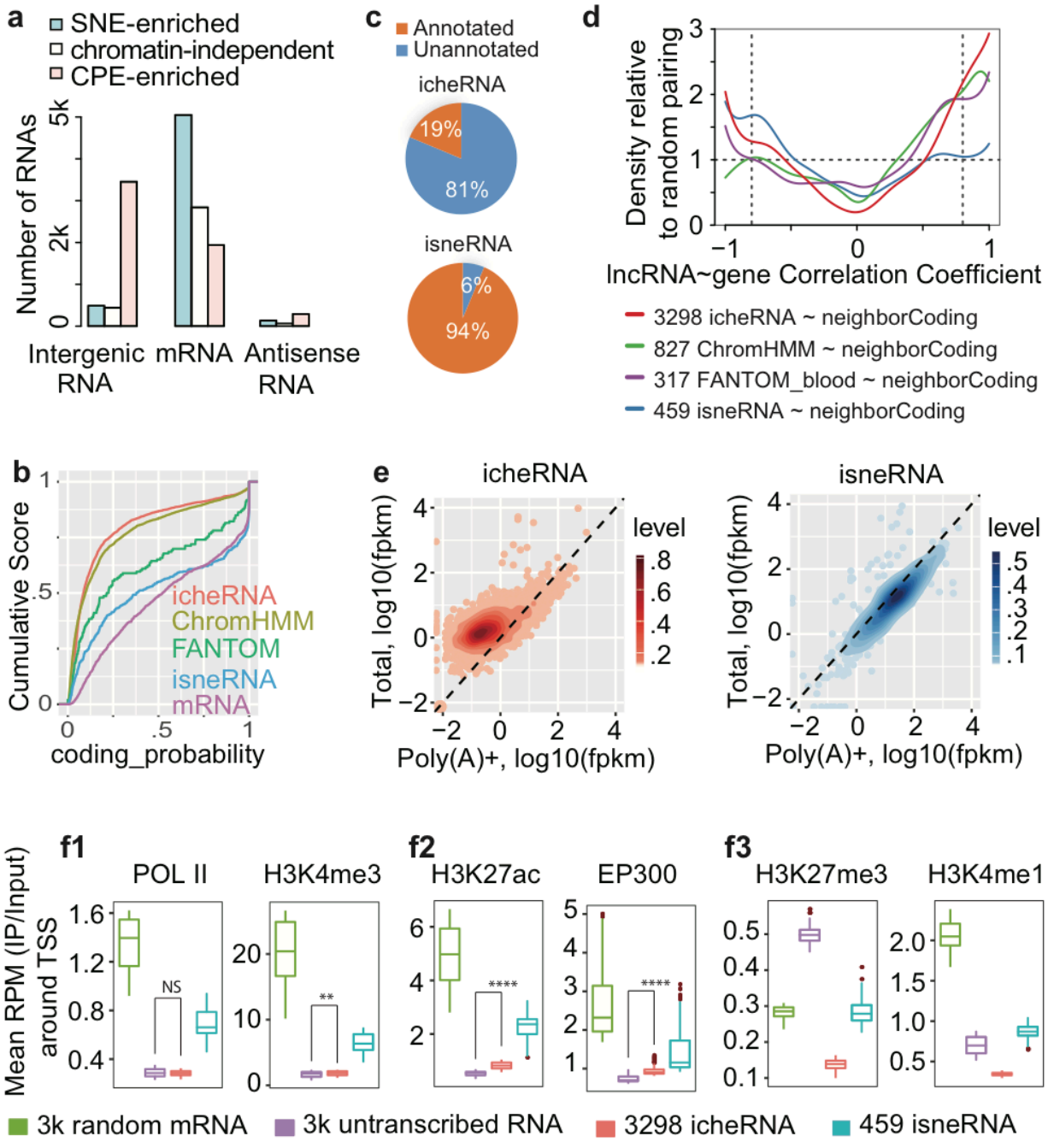
Figure 3.1 shows the workflow used to categorize the three RNA groups). A large fraction (66%) of the 5,680 identified cheRNAs are transcribed from noncoding regions (Figure 3.2a**,** pink bar). In contrast, approximate 90% of the identified 5,672 sneRNAs were mRNAs (Figure 3.2a, blue bar). Additionally, icheRNA exhibits lower coding potential (cumulative CPC2 score (Kang *et al*., 2017)) than coding genes, intergenic sneRNA (isneRNA), and intergenic RNA transcribed from ChromHMM- or FANTOM- (de Hoon *et al*., 2015) predicted enhancer regions in the same cell lines (Figure 3.1d). The coding potential of icheRNA is therefore more similar to that of ChromHMM predicted eRNA, while that of isneRNA is more similar to that of mRNA.

81% (2.7 k) of the identified 3.3 k icheRNAs are previously unannotated transcripts, in contrast to only 6% (27) of the 459 isneRNAs, (Figure 3.2c). Additionally, over half (69% of 445) of the

antisense cheRNAs are unannotated, in contrast to only 2% of 163 antisense sneRNAs. This biased annotation of noncoding RNA suggests that previously detected noncoding RNAs primarily correspond to chromatin-depleted noncoding RNA (noncoding sneRNA), and that identifying chromatin enriched RNAs from nuclear extracts can give a more balanced picture of the overall noncoding RNA population.

Figure 3.2 Known genomic features of the intergenic cheRNAs in the K562 cells.

(a) Distribution of RNA classes in fractionated libraries. Three classes of RNAs were defined based on their relative genomic locations to GENCODE (v25)-annotated protein-coding genes (Figure 3.1). Chromatin-independent RNAs refer to RNAs not differentially expressed in CPE and SNE samples. (b) Coding potential of icheRNA (red), ChromHMM predicted eRNAs (yellow), FANTOM predicted eRNAs (green), isneRNA (blue) and mRNAs (purple). Intergenic RNA overlapped (at least 1 base) with any ChromHMM/FANTOM identified enhancer region is assumed to be predicted ChromHMM/FANTOM predicted eRNAs. icheRNA (red) hold the lowest protein-coding potential. Color decoding five RNA groups. As a control, mRNAs (purple) have the highest protein-coding potential with a curve tending towards the bottom-right corner. The online tool CPC2 is used. (c) Percentage of GENCODE (v25) annotated and unannotated RNAs in icheRNA and isneRNA. (d) Pairwise Correlation of expression of RNA classes in the K562 nucleus. The Pearson correlation coefficient is shown for of each of the indicated RNA classes (icheRNA (red), isneRNA (blue), ChromHMM-predicted eRNA (green) and FANTOM predicted eRNA (purple)) and its neighboring coding genes. To pair an intergenic genomic feature with its neighboring gene, the adjacent upstream or downstream gene with the highest magnitude PCC is selected. The relative density at a certain PCC value is calculated by dividing the kernel density estimates of indicated RNA and neighboring coding gene pairs by that of indicated RNA and randomly selected coding gene pairs. (Two vertical dashed lines mark significant cutoffs of PCC values at -0.8 or 0.8). (e) Normalized expression values of fractionate RNA classes. Values are given in FPKM (Fragments Per Kilobase Million) of icheRNA (red) and isneRNA (blue) in Poly(A)+ nuclear RNA-Seq library (x-axis, GSE88339) versus nuclear total-RNA-Seq library (y-axis, GSE87982) in K562 cells. More comparisons are available in (Supplementary Figure. 3.3). (f) Average ChIP-seq read density versus input in K562 cells of RNA polymerase II (POL II), H3K4me3, H3K27ac, EP300, H3K27me3 and H3K4me1 profiles centered at promoters (±1kb centered at TSS) of randomly selected mRNAs (green), randomly selected silent RNAs (purple), icheRNA (red) and isneRNA (blue), p-values calculated by two-sided Wilcoxon rank sum test, NS $p>0.05$, * $p<0.01$, ** $p<1e-10$, **** $p<2.2e-16$. (Note that in each panel, boxes without overlaps are significantly different without showing **** for simplicity.) We randomly selected 3000 mRNAs from 9.8k transcribed mRNAs and 3000 silent RNAs from 66.9k annotated but untranscribed RNAs.


a
SNE-enriched
chromatin-independent
CPE-enriched
Number of RNAs
5k
2k
0
Intergenic RNA
mRNA
Antisense RNA
c
Annotated
Unannotated
icheRNA
19%
81%
isneRNA
6%
94%
d
Density relative to random pairing
lncRNA~gene Correlation Coefficient
3298 icheRNA ~ neighborCoding
827 ChromHMM ~ neighborCoding
317 FANTOM_blood ~ neighborCoding
459 isneRNA ~ neighborCoding
b
Cumulative Score
coding_probability
icheRNA
ChromHMM
FANTOM
isneRNA
mRNA
e
icheRNA
isneRNA
Total, log10(fpkm)
Poly(A)+, log10(fpkm)
level
f1
f2
f3
POL II
H3K4me3
H3K27ac
EP300
H3K27me3
H3K4me1
Mean RPM (IP/Input) around TSS
NS
**
****
****
3k random mRNA
3k untranscribed RNA
3298 icheRNA
459 isneRNA

3.3 icheRNA positively correlate with adjacent genes in expression

RT-PCR experiments have shown that several eRNAs are intergenic chromatin enriched RNAs (icheRNA) (Yang *et al*., 2017). Werner *et al.* also showed that protein-coding genes proximal to icheRNA have higher expression levels than those near to other expressed lncRNA, suggesting that icheRNA could predict cis-gene transcription (Werner and Ruthenburg, 2015; Werner *et al*., 2017). However, it is not clear from previous work whether higher icheRNA expression is correlated with expression of proximal protein-coding genes. To quantitatively confirm the cis-regulatory potential of icheRNA, we calculated the Pearson correlation coefficient between the expression of icheRNA and neighboring protein-coding genes, and compared it to the correlation coefficient between the expression of icheRNA and randomly selected protein-coding genes. isneRNA and neighboring protein-coding gene, ChromHMM predicted eRNA and neighboring protein-coding gene, and FANTOM predicted eRNA and neighboring protein-coding gene.

We find that icheRNA are more positively correlated with neighboring genes than with randomly selected genes (Figure 3.2d, red line shows relative density > 1 when correlation coefficient > 0.5). The same calculation for FAMTOM- or ChromHMM-predicted eRNAs, which are believed to have cis-regulatory enhancer effects, and adjacent genes in the same cell types, shows similar but weaker positive correlations. In contrast, pairs of intergenic sneRNAs (isneRNA) and neighboring genes (blue line) showed negative correlation (Figure 3.2d, blue line shows relative density > 1 when correlation coefficient < -0.5). Specifically, with a significance cutoff of correlation coefficient=0.8, 23% of the identified icheRNA transcripts, in contrast to only 11% of the isneRNA are positively correlated with proximal genes. This observation, for the first time, gives quantitative evidence for a potential cis-regulatory effect of icheRNA on adjacent genes. It also suggests that identification of icheRNA can be used as another approach to predict eRNA, comparable to approaches using ChromHMM and FANTOM database.

Transcriptional correlation analysis also displayed high relative density at correlation coefficient < -0.5 for pairs of icheRNA and neighboring coding genes (Figure 3.2d), indicating that not all icheRNAs are positively correlated with proximal protein-coding gene expression. Indeed, *XIST* is a canonical icheRNA that has a well-known repressive regulatory role, and it might be one of the icheRNAs that are negatively correlated with proximal protein-coding genes.

3.4 Polyadenylated RNA is relatively depleted in icheRNA

Most eRNAs have been reported to be unspliced and non-polyadenylated (De Santa *et al*., 2010; Kim *et al*., 2010; Lam *et al*., 2014, Kim *et al.,* 2015). To test if icheRNA are similar in this regard, we compared the relative expression (measured as Reads Per Kilobase of transcript per Million mapped reads, RPKM) of intergenic cheRNAs in nuclear Poly(A)+ RNA-seq library and nuclear total RNA-seq libraries using published datasets for K562 cells (Table 4). We observe (Figure 3.2e) lower relative abundance of icheRNA in the nuclear total-RNA-seq library than in the nuclear Poly(A)+ RNA-seq library, indicating that majority of icheRNA lack polyadenylation. A similar but weaker preference for the total-RNA-seq library was also observed for antisense cheRNAs (Figure 3.3). In contrast, all protein-coding mRNAs have equivalent expression levels in two libraries, which is consistent with the role of polyadenylation in producing mRNA in the eukaryotic cell nucleus (Guhaniyogi and Brewer, 2001). Chromatin depleted non-coding RNAs (isneRNA and antisense sneRNAs) also have similar expression levels in the two RNA-seq libraries as those of mRNAs (Figure 3.3). The patterns of polyadenylation indicate that icheRNA and isneRNA are differentially polyadenylated. With respect to polyadenylation, icheRNA is more similar to eRNA than is sneRNA, since the majority of icheRNA are not polyadenylated.

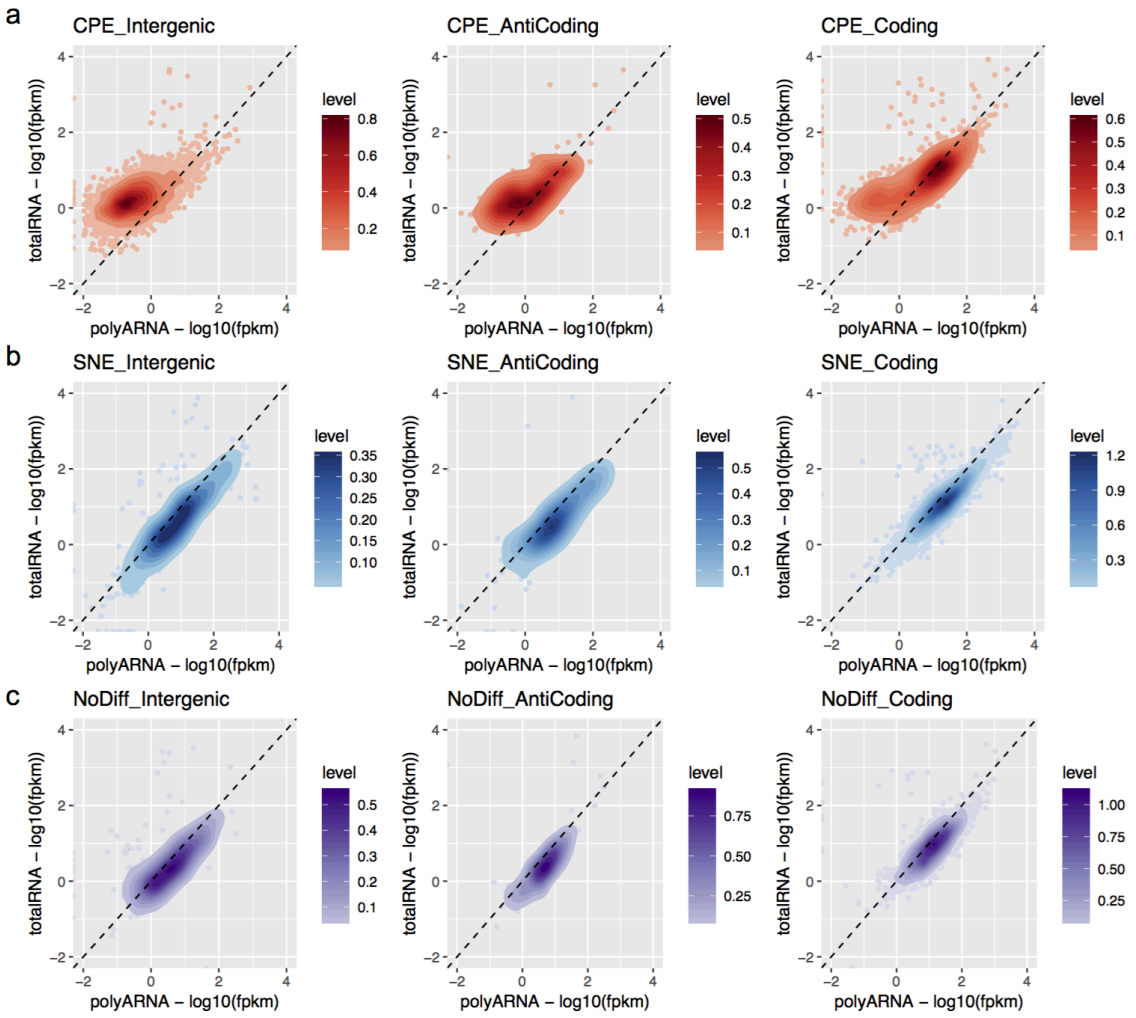


Figure 3.3 Normalized expression values of fractionate RNA classes.

Scale-density plot, comparing the expression value (in FPKM) in Poly(A)+ nuclear RNA-Seq library (x-axis) and total nuclear RNA-Seq library (y-axis) for (a) chromatin-enriched RNAs (red), (b) chromatin-depleted RNAs (blue), and (c) chromatin-independent RNAs (purple, RNAs not differentially expressed in either CPE or SNE samples) transcribed from intergenic region, region antisense to coding gene and coding gene region.

3.5 IcheRNAs and isneRNAs confer different chromatin characteristics

Histone 3 lysine 4 monomethylation (H3K4me1) and histone 3 lysine 27 acetylation (H3K27ac) have been identified as key histone modification features that mark enhancers. H3K4me1 is present at both poised and active enhancers (Dorighi *et al*., 2017), while H3K27ac uniquely marks active enhancers (Creyghton *et al*., 2010). Werner *et al*. previously observed peaks of H3K27ac

near the transcriptional start sites (TSS) of icheRNA, however, unlike prototypical eRNA, these regions did not show abundant H3K4me1 modification (Werner and Ruthenburg, 2015). To further investigate whether icheRNA have a distinct chromatin signature, we profiled the relative reads per million (RPM) of RNA polymerase II (POLII), H3K27ac, H3K4me3, H3K4me1, and H3K27me3 marks on the flanking 1 kb sequences around TSS of icheRNA, isneRNA, mRNA and unexpressed mRNA (RNAs annotated in GENCODE(v25) but not transcribed in K562) (Figure 3.2f). IcheRNA show low levels of marks associated with active transcription (POLII and H3K4me3), similar to the levels of unexpressed mRNA, and lower than those of isneRNA and mRNA (Figure 3.2f1, red and purple box). In contrast to unexpressed mRNA, icheRNA TSS flanking regions show low levels of repressive (H3K27me3) and poised enhancer (H3K4me1) marks (Figure 3.2f2**,** red and purple box), but are enriched in active enhancer (H3K27ac and EP300) marks (Figure 3.2f3**,** red and purple box). Note that in addition to being enriched at enhancer regions, H3K27ac and EP300 are also pervasively found near TSS of actively transcribed regions. icheRNA TSS thus have a chromatin profile that is distinctly different from those of mRNA, isneRNA, and unexpressed mRNA, suggesting that significantly different modes of regulation may be controlling icheRNA expression.

In summary, icheRNA and isneRNA differ in many respects. In addition to the enrichment of specific epigenetic marks near the TSS, icheRNA has lower coding probability, lacks polyadenylation, and its expression is more positively correlated with that of neighboring coding genes. Overall icheRNA is more similar to eRNA, while isneRNA is more similar to mRNA. The similarity of icheRNA to eRNA, as defined by ChromHMM and FANTOM predictions, suggests that icheRNA identification may provide a useful independent approach to predicting eRNA.

3.6 Discussion

Operationally, cheRNA is defined by its statistically significant enrichment in chromatin after biochemical fractionation of nuclei. With our improved computational strategy, we have examined the molecular characteristics of cheRNAs in greater detail than has heretofore been possible. We find that, first, cheRNAs are more likely to be transcribed from noncoding regions, while sneRNAs are mostly transcribed from protein-coding regions. Second, icheRNA has a lower transcription level and is largely unannotated, in contrast to isneRNA which is more highly transcribed and

better annotated. Traditional transcriptome profiling of non-coding RNA, using techniques such as total RNA-seq, yields the broadest survey of transcripts, but has limited ability to detect low expression transcripts such as those of icheRNA. Thus, previous analyses of noncoding RNA primarily focused on noncoding RNA with relatively high transcription levels (*e.g.*, isneRNA and as-sneRNA). In contrast, sequencing and identifying chromatin enriched RNAs in a nuclear extract more sensitively identifies low expression noncoding RNAs that previously have been ignored by conventional sequencing and analysis methods. Third, we have shown that icheRNA, in contrast isneRNA, is mostly non-coding, non-polyadenylated, and positively correlated with the expression of neighboring coding genes (Figures 3.2a-3.2e). The above analysis also shows that icheRNAs possess stronger positive correlation with adjacent protein-coding genes in expression, compared with ChromHMM- and FANTOM-predicted eRNAs, suggesting that separating intergenic RNAs into chromatin-enriched and chromatin-depleted groups can be used as another approach to predict eRNAs, comparable to approaches applied in ChromHMM and FANTOM database.

Notwithstanding the similarity of these features to those of eRNA, icheRNA has several unique molecular characteristics that distinguish it. For example, icheRNA is generally longer than eRNA (median length of icheRNA is ~4,400 bases; eRNA is ~350 bases, Andersson *et al*., 2014)) and icheRNA shows only modest coincidence with enhancer marks (H3K27ac, H3K4me1 and EP300) that are used to canonically define eRNA (Figure 3.2f). Moreover, some canonical icheRNA (*e.g.*, *XIST*) are known to be repressive regulators rather than activators as is eRNA. Combining all this evidence, we conclude that icheRNA and eRNAs are two distinct non-coding RNA groups that overlap. Despite there are differences, icheRNA still show more apparent similarity to eRNAs than other RNA groups.

CHAPTER 4. CIS-REGULATORY POTENTIAL OF TWO CHERNAS SUBSETS

4.1 Summary

To provide more insights into the features of cheRNAs, in this chapter, we explore two new potential cis-regulatory functions of subsets of cheRNAs: intergenic cheRNAs transcribed from genes in condensed chromatin (marked by H3K9me3), and cheRNA transcript antisense to coding genes.

Firstly, we found that regions around TSS of icheRNAs are depleted with H3K9me3 chromatin mark, however, the DNA regions transcribing icheRNA body show high H3K9me3 level. This revealed an unexpected association between chromatin-based RNA and H3K9me3, a chromatin mark associated with closed/repressed chromatin. To further explore this association, we separate icheRNA into two groups: icheRNA with H3K9me3 mark and icheRNA without H3K9me3 mark. We compared the chromatin signatures around DNA regions related to the two groups. We found that icheRNA transcribed from H3K9me3 marked regions show elevated transcriptional activity and higher levels of enhancer marks compared to icheRNA transcribed from regions without H3K9me3 marks. Moreover, we measured the H3K9me3 levels across canonical icheRNA transcribed regions and found that three previously identified icheRNA (*HIDALGO*, *ILYICH*, *BONIFACIO*) with validated positive activator functions show relatively higher H3K9me3 levels than the only icheRNA with a known repressive role (*XIST*). Together, these evidences suggest that DNA regions transcribing icheRNA, even with high levels of H3K9me3 modification, can be actively transcribed and may have the potential to indicate active enhancer region.

Secondly, we discussed one possible origin for the unexpected H3K9me3 signal around icheRNA. Unlike H3K27me3, H3K9me3 is more global and permanent, and are frequently associated with constitutive heterochromatin regions. For example, H3K9me3 has been found to be enriched at Lamina-Associated Domains (LADs), which are genomic regions in close contact with the nuclear lamina. These regions are termed as Lamina-Associated Domains (LADs) (van Steensel and Belmont, 2017). We showed that icheRNAs are overrepresented in LADs than other RNA groups (48% of icheRNAs are transcribed from LADs in contrast to only 12% for other RNAs). Moreover,

considering H3K9me3 is also associated with transposable elements (TE) repression (He et al., 2019), we examined the enrichment of TEs among icheRNAs and icheRNAs with H3K9me3 marks in K562 cell. We found that 82% of icheRNAs and 96% of icheRNAs with H3K9me3 marks overlap with class 1 TEs. In conclusion, one hypothesis for the unexpected H3K9me3 signal around icheRNA is that the icheRNA may be embedded in condensed domains derived from mobile elements.

Lastly, we explored the cis-regulatory potential of antisense cheRNA (as-cheRNA). We observed that antisense RNA (asRNA) accumulates preferentially in the nucleus associating with chromatin. Similar with icheRNA, as-cheRNA lacks annotation. By examining the transcriptional activity of the colocalized mRNA on the opposite strand, we observed that the TSS of antisense cheRNA (as-cheRNA) colocalized mRNA is significantly less open (measured by ATAC-seq signal), has fewer active transcription marks (POL II, H3K4me3), and has more repressive marks (H3K27me3) and PRC2 complex binding (SUZ12, EZH2), compared with random mRNA. Moreover, this pattern is not observed in mRNA colocalized with antisense chromatin depleted RNA (as-sneRNA). Even though still not conclusive, this unique pattern observed only in as-cheRNA suggests as-cheRNA to be cis-acting repressor that interfere transcription of colocalized mRNAs on the opposite strand via recruiting the PRC2 complex.

4.2 IcheRNA with H3K9me3 across transcript body is prone to present active cis-regulation

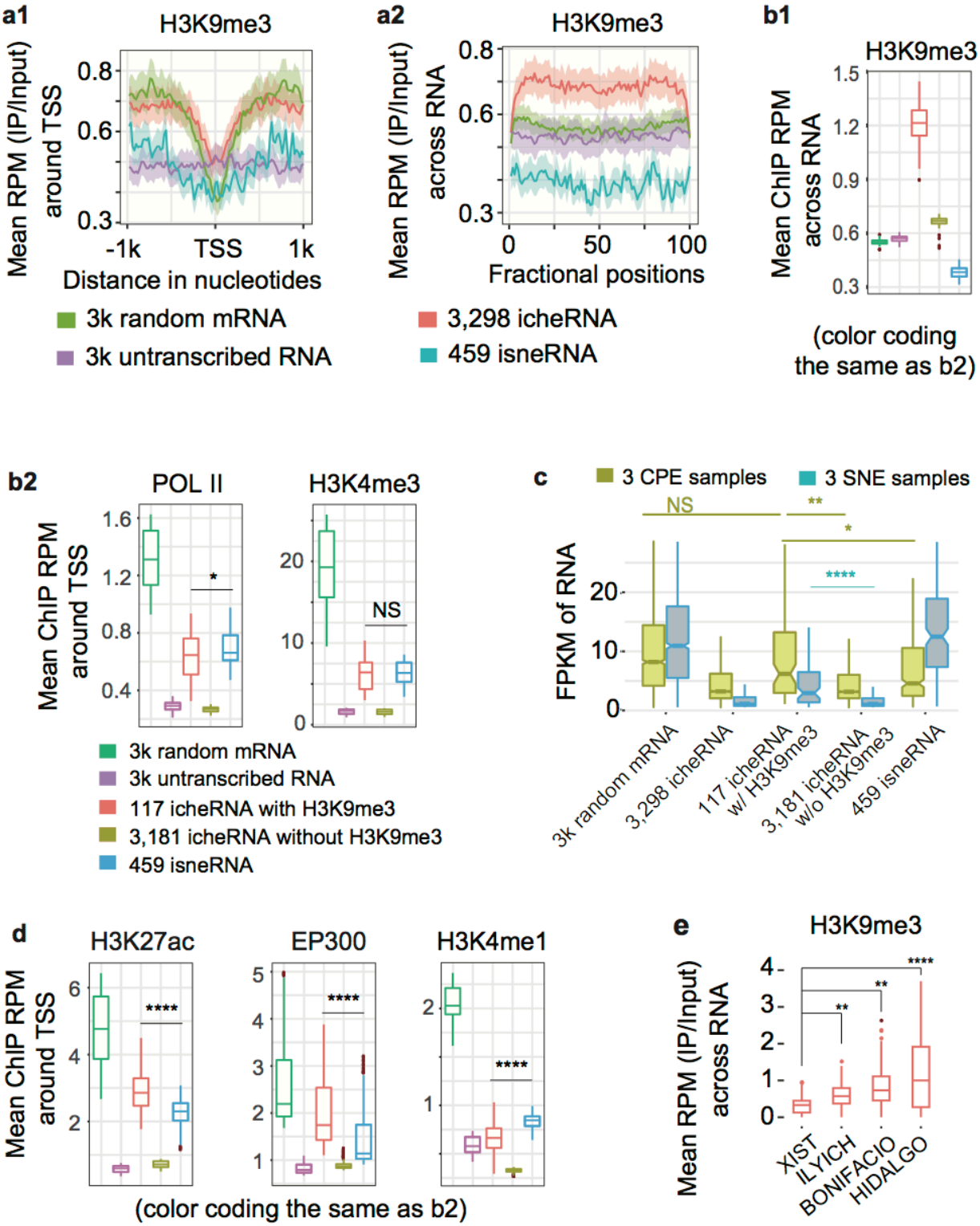
Histone 3 lysine 9 trimethylation (H3K9me3) is associated with constitutive heterochromatin, and has been shown to mark transcriptionally repressed regions that are mutually exclusive with H3K27me3 marked repressive regions (Kouzarides, 2007; Hublitz *et al*., 2009; Zhang *et al*., 2015; Becker *et al*., 2016). We find that the levels of H3K9me3 near actively transcribed icheRNA and mRNA TSS (Figure 4.1a1, red line and green line) are much higher than near transcriptionally silenced regions (DNA regions near to unexpressed mRNA) (Figure 4.1a1, purple line). In addition, H3K9me3 profiles at actively transcribed regions are quite different from those at transcriptionally silent regions: H3K9me3 modification is low near the TSS of transcribed RNA (icheRNA, isneRNA, and mRNA) (Figure 4.1a1, red line, blue line and green line) but not depleted around TSS of unexpressed mRNA (Figure 4.1a1, purple line). It has been suggested, for coding transcripts, that H3K9me3 at the promoter is repressive, whereas H3K9me3 across the mRNA

transcript body is activatory (Kouzarides, 2007). The pattern we observe is similar, and when combined with the previous observation that high levels of H3K9me3 modification are present in some active genes (Barski 2007), it suggests that high H3K9me3 levels do not necessarily indicate transcriptional repression; H3K9me3 modification at the TSS region is more strongly associated with transcriptional silencing, in contrast, H3K9me3 at gene body regions can be actively transcribed.

We also note that H3K9me3 levels within the DNA region of transcribed icheRNA is substantially higher than near other transcribed RNA (*e.g.*, mRNA and isneRNA) (Figure 4.1a2). To further investigate the effect of H3K9me3 on icheRNA transcription, we separated DNA regions transcribing icheRNA into high H3K9me3 (at least 1 peak of H3K9me3 mark near the transcribed icheRNA) and low H3K9me3 (no H3K9me3 mark near the transcribed icheRNA) groups. These groups are labeled as "icheRNA with H3K9me3" and "icheRNA without H3K9me3", respectively in Figure 4.1b. DNA regions in the "icheRNA with H3K9me3" have significantly higher levels of H3K9me3 modification than those in the "icheRNA without H3K9me3" group (Figure 4.1b1). Furthermore, chromatin signatures associated with active transcription (POL2, H3K4me3) (Figure 4.1b2), as well as transcription levels in both CPE and SNE samples (Figure 4.1c), are strikingly elevated in the "icheRNA with H3K9me3" group compared to the "icheRNA without H3K9me3" group, indicating that icheRNA are more actively transcribed from regions with high H3K9me3 modification. It also reinforces the evidence indicating that regions with abundant H3K9me3 modification can be actively transcribed.

Figure 4.1 icheRNA with H3K9me3 signal concur chromatin modification patterns of active enhancers.

(a) Average H3K9me3 ChIP-seq read density versus input in K562 cells (a1) at promoters (±1kb centered at TSS) of, or (a2) across regions transcribing, randomly selected mRNAs (green), randomly selected silent RNAs (purple), icheRNA (red) and isneRNA (blue). (b) Average ChIP-seq read density versus input in K562 cells of (b1) H3K9me3 profiles across regions transcribing, or (b2) POL II and H3K4me3 profiles at promoters (±1kb centered at TSS) of, randomly selected mRNAs (green), randomly selected silent RNAs (purple), icheRNA coincident with H3K9me3 marks (icheRNA with H3K9me3, red), icheRNA without H3K9me3 (yellow) and isneRNA (blue). (c) Normalized expression values in FPKM in chromatin pallet extract (CPE, red boxes) and soluble nuclear extract (SNE, blue) of K562 cells for randomly selected mRNA, icheRNA, icheRNA with H3K9me3 (icheRNA w/ H3K9me3), icheRNA without H3K9me3 (icheRNA w/o H3K9me3) and isneRNA. (d) Average ChIP-seq read density in K562 cells of active enhancer marks (H3K27ac and EP300) and poised enhancer mark (H3K4me1) profiles at promoters (±1kb centered at TSS) of randomly selected mRNAs (green), randomly selected silent RNAs (purple), icheRNA with H3K9me3 (red), icheRNA without H3K9me3 (yellow) and isneRNA (blue). (e) Average H3K9me3 ChIP-seq read density versus input in K562 cells across regions transcribing four canonical cheRNAs. The four cheRNAs were ordered according to their known transcriptomic regulatory functions, from the repressor (XIST) on the left to other three cis-activators (ILYICH, BONIFACIO, HIDALGO) on the right. p-values calculated by two-sided Wilcoxon rank sum test, NS $p>0.05$, * $p<0.01$, ** $p<1e-10$, **** $p<2.2e-16$.


a1
H3K9me3
Mean RPM (IP/Input) around TSS
0.8
0.6
0.3
-1k
TSS
1k
Distance in nucleotides
3k random mRNA
3k untranscribed RNA
a2
H3K9me3
Mean RPM (IP/Input) across RNA
0.8
0.6
0.3
0
50
100
Fractional positions
3,298 icheRNA
459 isneRNA
b1
H3K9me3
Mean ChIP RPM across RNA
1.5
1.2
0.9
0.6
0.3
(color coding the same as b2)
b2
POL II
H3K4me3
Mean ChIP RPM around TSS
1.6
1.2
0.8
0.4
*
20
10
0
NS
3k random mRNA
3k untranscribed RNA
117 icheRNA with H3K9me3
3,181 icheRNA without H3K9me3
459 isneRNA
c
3 CPE samples
3 SNE samples
FPKM of RNA
NS
**
*
****
20
10
0
3k random mRNA
3,298 icheRNA
117 icheRNA w/ H3K9me3
3,181 icheRNA w/o H3K9me3
459 isneRNA
d
H3K27ac
EP300
H3K4me1
Mean ChIP RPM around TSS
6
4
2
****
5
4
3
2
1
****
2
1
****
(color coding the same as b2)
e
H3K9me3
Mean RPM (IP/Input) across RNA
4
3
2
1
0
**
**
****
XIST
ILYICH
BONIFACIO
HIDALGO

Our previous analysis showed that icheRNA possesses features similar to eRNA, however, the TSS of icheRNA show only moderately higher levels of enhancer marks compared to unexpressed mRNA, and lower levels than TSS of isneRNA. We measured the levels of enhancer marks (H3K27ac, EP300 and H3K4me1) around TSS of "icheRNA with H3K9me3" (Figure 4.1d, red box). We found that levels of active enhancer marks (H3K27ac and EP300) around TSS of "icheRNA with H3K9me3" are significantly higher than at the TSS of "icheRNA without H3K9me3" (Figure 4.1d, yellow box) and TSS of isneRNA (Figure 4.1d, purple box), indicating that icheRNA with H3K9me3 marks shows high levels of active enhancer marks near the TSS, but all icheRNA do not. Moreover, we measured the H3K9me3 levels across canonical icheRNA transcribed regions and found that three previously identified icheRNA (*HIDALGO*, *ILYICH*, *BONIFACIO*) with validated positive activator functions show relatively higher H3K9me3 levels than the only icheRNA with a known repressive role (*XIST*) (Figure 4.1e). These examples reinforce the hypothesis that DNA regions transcribing icheRNA, even with high levels of H3K9me3 modification, may act as enhancers.

4.3 Possible origin for H3K9me3 signal around icheRNA

In metazoan cell nuclei, hundreds of large chromatin domains, termed Lamina-Associated Domains (LADs), have found to be in close contact with the nuclear lamina. LADs are enriched in histone modification of H3K9me2 and H3K9me3, modifications that are typical of heterochromatin (van Steensel and Belmont, 2017). A study on a 1 Mb LAD encompassing the human HBB loci showed that knockdown of H3K9me3 by depletion of the two H3K9me3 methyltransferases Suv39H1 and Suv39H2 caused detachment of the LADs and nuclear lamina, suggesting that H3K9me3 modification contributes to anchoring LADs to nuclear lamina (Bian *et al*., 2013). Considering H3K9me3 is a chromatin mark associated with closed/repressed chromatin, gene transcription from H3K9me3 enriched LADs is expected to be repressed. However, the unexpected association between icheRNA and high levels of H3K9me3 chromatin marks suggests that icheRNA genes may be embedded in, and actively transcribed from, condensed LADs. Indeed, we find that 48% of icheRNAs are transcribed from LADs (greater than chance expectation, empirical $p < 2.2e-16$), in contrast, only 12% of other RNAs are transcribed from LADs. Moreover, agree with the previous hypothesis by Werner *et al*. that TEs may provide an evolutionary origin to chromatin enriched RNAs (Werner *et al*., 2017), we noticed that 82% of icheRNAs and 96% of

icheRNAs with H3K9me3 chromatin marks in K562 overlap with class 1 TEs. Together, these observations suggest that icheRNA may represent a group of RNAs transcribed from condensed chromatin domains derived from mobile elements, and that the transcription of these domains is regulated in a cell-specific way.

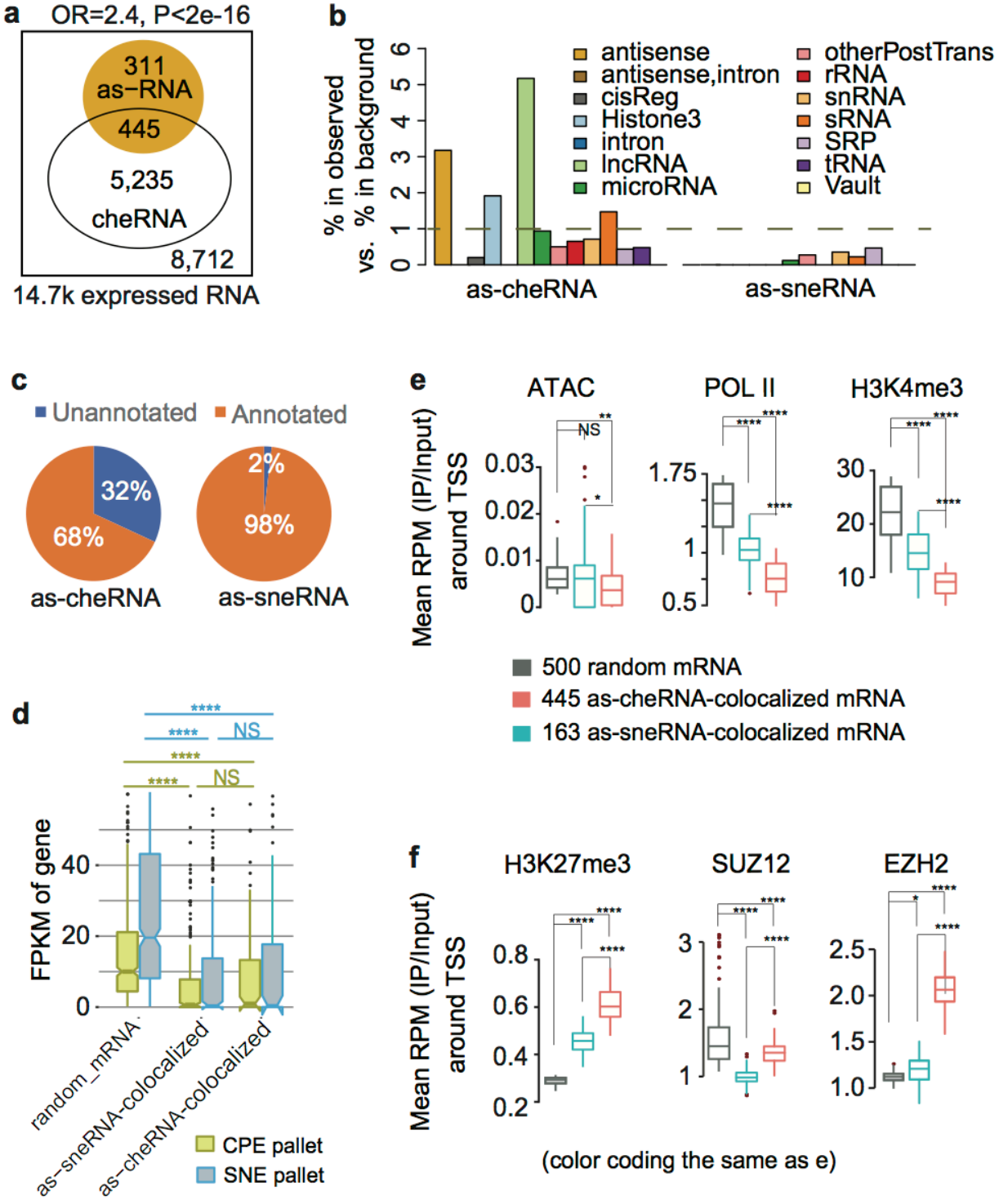
4.4 Antisense cheRNAs (as-cheRNA) concur local mRNA repression

Antisense RNA (asRNA) complementary to protein-coding transcript(s) has been shown to interfere with transcription of mRNA on the opposite strand (Tufarelli *et al*., 2003). Consistent with this, asRNA accumulates preferentially in the nucleus associating with chromatin, we observe that almost (59%) of the identified 756 asRNAs in K562 cell nucleus are chromatin enriched and only 22% are chromatin depleted (Figure 3.2a), indicating a significant enrichment of asRNA in the chromatin pellet (Figure 4.2a). Moreover, we notice that about one third of the chromatin enriched asRNAs (as-cheRNA) are unannotated while almost all chromatin depleted asRNAs (antisense sneRNA, as-sneRNA) are annotated (Figure 4.2c), suggesting that many as-cheRNA are completely novel.

Regulatory mechanisms involving asRNA range from simple transcriptional interference through competing for RNA Pol II (Shearwin *et al*., 2005) to regulation of epigenomic modifications (Kotake *et al*., 2011; Bhan and Mandal, 2014). A current hypothesis suggests that asRNA is acts in gene regulation at the chromatin level by recruiting epigenetic regulators, *e.g.,* polycomb repressive complex 2 (PRC2), to its corresponding sense mRNA to induce histone methylation and gene repression (Magistri *et al*., 2012; Latgé *et al*., 2018). Inspired by this hypothesis, we investigated a similar potential function for both as-cheRNA and as-sneRNA. Functional RNA molecules often exhibit secondary structures that are better conserved than their sequences (Kalvari 2018), we first interrogated the equence based predicted secondary structure of as-cheRNA and as-sneRNA in comparison to known RNA families in Rfam. Rfam collects multiple-sequence alignment-based families of RNA secondary structural motifs (Kalvari *et al*., 2018). The motif sizes are generally less than 400 bases long Figure 4.3b), much shorter than the asRNA in the assembled transcriptome. We annotated each asRNA as belonging to a Rfam family if it fully covered a Rfam family motif. We then calculated, for each Rfam family, a) the fraction of as-cheRNA/as-sneRNA annotated to this Rfam family (the fraction in observation); b) the fraction of

Figure 4.2 as-cheRNAs indicate local mRNA silencing.

(a) Venn diagram showing the enrichment of cheRNA among asRNA. Fisher's exact test is used to estimate the odds ratio and p-value to quantify the strength of enrichment. Odds ration larger than 1 and p-value less than 0.05 indicate significant enrichment. (b) Enrichment of 14 Rfam ncRNA secondary structure family among as-cheRNA (left sub-panel) and sneRNAs (right sub-panel). The dashed line indicates a RR-score of 1. An RR-score larger than 1 indicates that as-cheRNA/as-sneRNA is overrepresented in the selected Rfam family. (c) Percentage of GENCODE (v25) annotated (orange) and unannotated (blue) RNA in as-cheRNA and as-sneRNA. (d) Normalized expression values in FPKM in chromatin pallet extract (CPE, yellow) and soluble nuclear extract (SNE, blue) of K562 cells for randomly selected mRNA, as-sneRNA and as-cheRNA. (e) Average ATAC-Seq read density and ChIP-seq read density of histone marks representing active transcription (POLII and H3K4me3) versus input in K562 cells at promoters (±1kb centered at TSS) of randomly selected mRNA (grey), as-cheRNA antisense overlapped mRNA (as-cheRNA-colocalized mRNA, red) and as-sneRNA antisense overlapped mRNA (as-sneRNA-colocalized mRNA, blue). (f) Average ChIP-seq read density of repressive histone mark (H3K27me3) and two PRC2 subunits (SUZ12 and EZH2) versus input in K562 cells at promoters (±1kb centered at TSS) of randomly selected mRNA (grey), as-cheRNA antisense overlapped mRNA (as-cheRNA-colocalized mRNA, red) and as-sneRNA antisense overlapped mRNA (as-sneRNA-colocalized mRNA, blue). p-values are calculated using a two-sided Wilcoxon rank sum test, NS $p>0.05$, * $p<0.01$, ** $p<1e-10$, **** $p<2.2e-16$.


a
OR=2.4, P<2e-16
311
as-RNA
445
5,235
cheRNA
8,712
14.7k expressed RNA
b
% in observed
vs. % in background
0 1 2 3 4 5 6
antisense
antisense,intron
cisReg
Histone3
intron
lncRNA
microRNA
otherPostTrans
rRNA
snRNA
sRNA
SRP
tRNA
Vault
as-cheRNA
as-sneRNA
c
Unannotated
Annotated
32%
68%
2%
98%
as-cheRNA
as-sneRNA
d
FPKM of gene
0
20
40
****
****
NS
****
****
NS
random_mRNA
as-sneRNA-colocalized
as-cheRNA-colocalized
CPE pallet
SNE pallet
e
ATAC
POL II
H3K4me3
Mean RPM (IP/Input)
around TSS
0
0.01
0.02
0.03
0.5
1
1.75
10
20
30
NS
500 random mRNA
445 as-cheRNA-colocalized mRNA
163 as-sneRNA-colocalized mRNA
f
H3K27me3
SUZ12
EZH2
Mean RPM (IP/Input)
around TSS
0.2
0.4
0.6
0.8
1
2
3
1.0
1.5
2.0
2.5
(color coding the same as e)

all assembled RNA annotated to this Rfam family (the fraction in background); and c) the ratio of the fraction in observation over the fraction background (RR-score). An RR-score larger than 1 indicates that as-cheRNA/as-sneRNA is overrepresented in the selected Rfam family. Among fourteen major RNA structural groups in the Rfam database (v13, hg38) (Figure 4.3a), three structural groups (Histone 3, lncRNA, and antisense) are significantly overrepresented in as-cheRNA (**Figure 7b**, two or more folds). In particular, the overrepresentation of the antisense structure group among as-cheRNA suggests that the function of as-cheRNA, rather than that of as-sneRNA, is likely to be structure-based.

We then measured the transcription level in CPE and SNE of mRNA that antisense overlaps with as-cheRNA and as-sneRNA. We find that the transcription of both as-cheRNA-colocalized mRNA and as-sneRNA-colocalized mRNA are relatively low compared to that of random mRNA (Figure 4.2d), suggesting a negative correlation between the transcription of sense and antisense RNA. Even though both as-cheRNA-colocalized mRNA and as-sneRNA-colocalized mRNA are shown to be repressed at similar levels, the chromatin features and histone patterns around the TSS of the two mRNA groups are significantly different. The TSS of as-cheRNA-colocalized mRNA (Figure 4.2e-4.2f, red box) are significantly less open (measured by Encode ATAC-seq signal), have fewer active transcription marks (POL2, H3K4me3), but have more repressive marks (H3K27me3) and show higher PRC2 complex binding (SUZ12, EZH2) compared with random mRNA (Figure 4.2e-4.2f, black box). This pattern was not observed in as-sneRNA-colocalized mRNAs (Figure 4.2e-4.2f, blue box). Altogether, this suggests that as-cheRNA and as-sneRNA may cis-repress gene transcription through different mechanisms. As-cheRNA may be cis-regulatory elements that repress transcription of colocalized mRNAs on the opposite strand via recruiting the PRC2 complex to specific genomic loci.

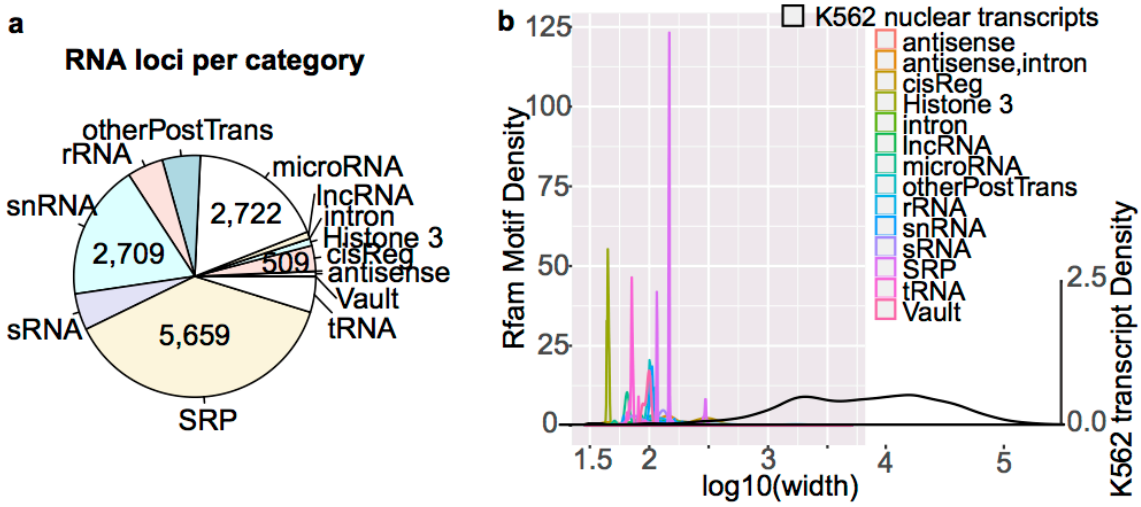


Figure 4.3 Fourteen major RNA structural groups in the Rfam database (v13, hg19).

(a) Pi plot showing the proportion in RNA structural motifs per group; (b) histogram of RNA structural motif widths compared to nuclear RNA-seq transcriptome (Tuxedo). Color decoding the fourteen major RNA structural groups.

4.5 Discussion

IcheRNA transcribed from H3K9me3 marked regions are more actively transcribed and more highly associated with elevated levels of enhancer marks than icheRNA without H3K9me3 marks. This observation indicates that H3K9me3 not only marks actively transcribed regions, but that it may also mark potential enhancer regions. The association between H3K9me3 and enhancers was also previously suggested by Zhu *et al* (2012), who described the widespread presence of H3K9me3 at enhancer flanking regions. They also showed anecdotal examples in which regulating H3K9me3 levels at the enhancers of Mdc and Il12b, affected Mdc and Il12b transcription in dendritic cells and macrophages, suggesting that H3K9me3 plays an important role in regulating enhancer activity (Zhu *et al*., 2012). If it can be verified that the regulatory role of H3K9me3 is a common feature of many enhancers, icheRNA coincident with H3K9me3 marks may prove a very effective predictor for chromatin-based eRNA, and may be a powerful approach to predicting novel enhancer regions.

Antisense RNA (asRNA) is another class of noncoding RNA that has been shown to have cis regulatory functions. Consistent with previous knowledge, our analysis confirms that asRNA is more abundant in the nuclear chromatin enriched pellet than in soluble nuclear pellet. Similar to

isneRNA and icheRNA, almost all as-sneRNAs are annotated, while a large fraction of as-cheRNAs lack annotation. This further suggests that sequencing RNAs abundant in the nuclear chromatin pellet can identify many novel noncoding RNAs. Despite the fact that both as-cheRNA and as-sneRNA show negative correlations in transcription level with their corresponding sense mRNA, the chromatin pattern around the TSS of as-cheRNA-colocalized mRNA and as-sneRNA-colocalized mRNA are quite different. Regions around the TSS of as-cheRNA-colocalized mRNA are less open and lack active transcription marks, but have high level of H3K27me3 and PRC2 binding, suggesting that as-cheRNA may regulate sense mRNA transcription in cis acting as a guide RNA for regulatory complexes that modify the target chromatin. Even though this investigation of as-cheRNA is still preliminary, it provides some testable hypotheses for asRNA function.

CHAPTER 5. SUMMARY

5.1 Challenges

The scientific discipline of genetics is founded upon Gregor Mendel's experimental work on peas. However, his intention was not to offer a general law of inheritance. Instead, his purpose was only to find out a law of the development of hybrids in plants (Gayon, 2016). So does the original goal of the Human Genome Project, which was launched with an aim to determine the sequence of human genome and make the map of the genes to facilitate the study of inherited diseases. The Human Genome Project accomplished this goal very well. Besides, it also revealed the importance of noncoding regions in the genome and leaded to a surprising increase in the study of ncRNA. When the sequence of the human genome was published in 2001, it showed that 99% of the human genome will not be translated into proteins. It was later shown that these noncoding regions are pervasively transcribed into ncRNAs. Since then, ncRNAs have been characterized in many species and were shown to be involved in processes such as development and pathologies, revealing a new layer of regulation in eukaryotic cells (Jarroux *et al*., 2017). Before 1999, the number of discovered ncRNAs in mammalian organisms was only less than 300. By 2004, this number increased dramatically to 5000. Most of these newly identified ncRNAs are miRNAs and putative lncRNAs with unknown function (Hüttenhofer *et al*., 2005).

Unlike miRNA, the mechanism of which has been well studied and understood, the understanding of lncRNA functioning mechanisms is still limited to the few individual examples. By definition, lncRNA is functional RNA molecule with a length of more than 200 nts that does not encode protein. This is a very broad definition, making this RNA group contains a variety of RNAs that function differently. LncRNA is also found to be less conserved, expressed at lower levels than mRNA, and show high level of cell and developmental specificity. All these features make lncRNA hard to be systematically identified and studied.

From the few known examples of lncRNAs, it’s worth to notice that lncRNA frequently functions at chromatin interface in nuclear. This feature distinguishes lncRNA from other noncoding RNAs that function by base pairing. Inspired by this, our collaborator Werner *et al*. employed biochemical fractionation of the nuclear compartment coupled to RNA-seq to identify lncRNAs

that are tightly associated with chromatin, termed as cheRNA (Werner and Ruthenburg, 2015). From perturbation of four distinct cheRNAs, they demonstrated that cheRNA positively correlates with neighbor gene expression, which indicates similarity with eRNA.

This study aims to provide a more holistic view of the nuclear noncoding transcriptome. We started with designing and surveying four computational strategies for nuclear RNA-seq data analysis. We showed that a new pipeline (Tuxedo) outperforms in assembly of both highly expressed mRNAs and lowly expressed lncRNAs. Besides, Tuxedo pipelines identifies cheRNAs with higher accuracy than the original pipeline. With this improved pipeline, we identified two highly clustered populations corresponding to nuclear-soluble RNA (sneRNA) and chromatin-associated RNA (cheRNA) in K562, HEK293 and H1-hESC cell lines. We characterized and compared the genomic features of cheRNA and sneRNA, and found that these two RNA groups are distinct in many aspects. CheRNAs are mostly transcribed from intergenic regions, in contrast, sneRNAs are mostly transcribed from protein coding regions. Compared to intergenic sneRNA (isneRNA), intergenic cheRNA (icheRNA) has lower coding probability, lacks polyadenylation, and its expression is more positively correlated with that of neighboring coding genes, suggesting that icheRNA rather than isneRNA is more similar to eRNA. We also observed that DNA regions transcribing icheRNA are abundant with H3K9me3 modification. In addition, we found that icheRNA transcribed from regions with high level of H3K9me3 modification show elevated transcriptional activity and higher levels of enhancer marks compared to icheRNA transcribed from regions with low level of H3K9me3. This unexpected association between chromatin-based RNA and high level of H3K9me3 suggests that DNA regions transcribing icheRNA, even methylated by H3K9me3, can be actively transcribed and may have the potential to indicate active enhancer region. Following this, we proposed one hypothesis for the origin of H3K9me3 signal around icheRNA transcribed region, which is icheRNA may be embedded in condensed domains derived from mobile elements. In the end, we explored a potential cis-repressive function for as-cheRNA. We showed that as-cheRNA appears to inhibit colocalized mRNA on the opposite strand through a mechanism of recruiting PRC complex to specific genomic loci.

In summary, quantitative identification of chromatin-enriched nuclear RNA provides a powerful way to profile the nuclear transcriptional landscape, especially to profile the noncoding transcriptome. The computational pipeline presented here provides researchers with a reliable

approach to identifying cheRNA, and studying cell-type specific gene regulators. Although the cheRNA is unlikely to be monolithic in function, icheRNA, especially icheRNA with high levels of H3K9me3 marks, may act as a transcriptional cis-activator similar to eRNA. In contrast, as-cheRNA may interact with diverse chromatin modulators to cis-repress transcription. With the Tuxedo pipeline, the future challenge will be refining the functional mechanisms of this noncoding RNA class through exploring their regulatory roles, which are involved in diverse molecular and cellular processes in human and other organisms.

5.2 Future work

CheRNAs are not a uniform set. However, based on our analysis results, icheRNA transcribed from H3K9me3 modified regions is more prone to enhancer cis gene transcription, and as-cheRNA appear to recruit PRC2 complex to colocalized mRNA on the opposite strand to repress gene transcription. Considering the current evidences are majorly from computational analysis, the future work should be focused on performing experiments to validate these cis-regulatory potential in vivo. Here I list five questions as a guide line for related future work:

(1). Does in vivo experiment (perturbation experiment) also supports that icheRNA transcribed from H3K9me3 marked region positively correlates with neighbor coding gene in transcription?

(2). Does alteration of H3K9me3 level at icheRNA transcribed region will affect transcriptional activity of icheRNA?

(3). If the answer to the question (1) is yes, does the alteration of H3K9me3 level at icheRNA transcribed region will affect the activatory function of icheRNA?

(4). Does repression of as-cheRNA in vivo will increase the transcription of colocalized sense mRNA?

(5). Does repression of as-cheRNA in vivo will decrease the binding of PRC2 complex on colocalized sense mRNA?

Besides performing related experiments to provide in vivo supports, the future work can also be extended to identification of cheRNA in other cell lines and organisms to facilitate the study of regulatory ncRNA that involves in disease development and cellular process.

APPENDIX A. DATASETS

RNA-seq raw datasets (in HEK293, H1 and K562 cell lines) were obtained from the NCBI Short-Read Archive (SRA) (Table 3). For the K562 cells, nuclear RNA sequencing, the ChIP-seq of multiple histone marks and transcription factors and ATAC-seq datasets were downloaded from ENCODE data portal (Tables 3-4). The noncoding RNA family were defined by Rfam (v13) (Kalvari *et al*., 2018). While mutiple resource IDs are available, we downloaded it from the ENCODE by ENCODE_ACCESSION IDs. While only hg19 landscape is available, we liftover hg19 landscape to the hg38 landscape.

Table 3 Genomic landscapes re-analyzed in Figure 2.1c.

Feature type	marker	ID	# of peaks
Histone	ATAC	GSM1782764	65,009,291
	ATAC	GSM1782765	
	EP300	ENCFF755HCK	28,757
	H3K27ac	ENCFF038DDS	52,334
	H3K4me1	ENCFF159VKJ	108,229
	H3K4me2	ENCFF118PIE	66,293
	H3K4me3	ENCFF148POZ	118,763
	H3K4me3	ENCFF616DLO	
	H3K4me3	ENCFF909PMV	
	H3K4me3	ENCFF961SPZ	
	H3K9me3	ENCFF371GMJ	5,584
Transcriptomic	rfam antisense		97
	rfam cisReg		647
	rfam lncRNA		138
	POLR2A	ENCFF099NYA	169,631
	POLR2A	ENCFF182YZG	
	POLR2A	ENCFF285MBX	
	POLR2A	ENCFF668VIK	
	POLR2A	ENCFF730DLS	
	POLR2A	ENCFF881ONC	
TF-binding	ATF2	ENCFF803FHN	46,737
	ATF3	ENCFF467WOR	7,875
	BACH1	ENCFF543FNN	4,707
	BRCA1	ENCFF652NES	815
	BRD4	ENCFF806CQB	8,493

TF-binding	CBX3	ENCFF951BQB	26,789
	CBX5	ENCFF403TAE	7,022
	CBX8	ENCFF210GJE	4,697
	CEBPB	ENCFF321KQD	71,925
	CEBPB	ENCFF813LOW	
	CTCF	ENCFF119XFJ	
	CTCF	ENCFF396BZQ	200,637
	CTCF	ENCFF519CXF	
	CTCF	ENCFF843VHC	
	E2F6	ENCFF417DTI	51,227
	E2F6	ENCFF533GSH	
	EGR1	ENCFF175VSS	
	EGR1	ENCFF375RDB	103,204
	EGR1	ENCFF561OGS	
	FOSL1	ENCFF087MFG	8,194
	GABPA	ENCFF124HAC	15,818
	GTF2F1	ENCFF478HYJ	5,219
	H2AFZ	ENCFF921IKK	100,908
	H3K9ac	ENCFF306MNO	172,520
	H3K9ac	ENCFF558JOB	
	HDAC2	ENCFF363GSV	
	HDAC2	ENCFF618YRQ	48,253
	HDAC2	ENCFF741IMY	
	HDAC6	ENCFF295GBP	1,570
	JUN	ENCFF032UMW	
	JUN	ENCFF167WUZ	
	JUN	ENCFF394CEC	50,782
	JUN	ENCFF672LKE	
	JUN	ENCFF881AVX	
	JUND	ENCFF213EYD	47,477
	KDM1A	ENCFF483BRD	80,331
	KDM1A	ENCFF796VMI	
	KDM5B	ENCFF668XLN	22,315
	MAFK	ENCFF893SCL	26,862
	MAX	ENCFF618VMC	97,737
	MAX	ENCFF900NVQ	
	MXI1	ENCFF243QTL	8,988
	PHF8	ENCFF952YDR	28,768
	RBBP5	ENCFF666PCE	24,374
	REST	ENCFF023ZUW	63,662
	REST	ENCFF290ESJ	
	RING1	ENCFF779XNE	1,077

TF-bidning	RNF2	ENCFF349MSP	69,849
	RNF2	ENCFF462AZY	
	RNF2	ENCFF741CLJ	
	RNF2	ENCFF820LKT	14,223
	SAP30	ENCFF103RHL	15,822
	SIN3A	ENCFF407VGB	
	SIN3A	ENCFF802JAN	3,590
	SIX5	ENCFF247LOF	14,782
	SP1	ENCFF452LDK	2,454
	SUZ12	ENCFF856HYC	19,263
	TAF1	ENCFF453TIB	685
	TAF7	ENCFF852NOL	45,012
	TCF12	ENCFF912LXU	
	TCF12	ENCFF952JIK	36,110
	TEAD4	ENCFF547MLB	21,382
	USF1	ENCFF717KGR	3,542
	USF2	ENCFF425FVY	6,630
	WDR5	ENCFF985TIE	51,788
	YY1	ENCFF024TJO	
	YY1	ENCFF635XCI	
	YY1	ENCFF953BTB	29,840
	ZNF143	ENCFF700GZI	3,440
	ZNF274	ENCFF323AWS	
	ZNF274	ENCFF498VQZ	

Table 4

Publicly accessible omics datasets analyzed in this study

DATASET_NAME	SEQ TYPE	CELL LINE	GENOME	GEO/ENCODE ACCESSION
HEK293_CPE1	RNA-seq	HEK293	HG38	GSM1623143
HEK293_CPE2	RNA-seq	HEK293	HG38	GSM1623144
HEK293_CPE3	RNA-seq	HEK293	HG38	GSM1623145
HEK293_SNE1	RNA-seq	HEK293	HG38	GSM1623140
HEK293_SNE2	RNA-seq	HEK293	HG38	GSM1623141
HEK293_SNE3	RNA-seq	HEK293	HG38	GSM1623142
H1_CPE1	RNA-seq	H1-hESC	HG38	GSM2208157
H1_CPE2	RNA-seq	H1-hESC	HG38	GSM2208158
H1_CPE3	RNA-seq	H1-hESC	HG38	GSM2208159
H1_SNE1	RNA-seq	H1-hESC	HG38	GSM2208160
H1_SNE2	RNA-seq	H1-hESC	HG38	GSM2208161
H1_SNE3	RNA-seq	H1-hESC	HG38	GSM2208162
K562_CPE1	RNA-seq	K562	HG38	GSM2208147
K562_CPE2	RNA-seq	K562	HG38	GSM2208148

K562_CPE3	RNA-seq	K562	HG38	GSM2208149
K562_SNE1	RNA-seq	K562	HG38	GSM2208150
K562_SNE2	RNA-seq	K562	HG38	GSM2208151
K562_SNE3	RNA-seq	K562	HG38	GSM2208152
K562_total_RNA-seq_1	RNA-seq	K562	HG38	ENCFF010QAI
K562_total_RNA-seq_2	RNA-seq	K562	HG38	ENCFF345SBQ
K562_total_RNA-seq_3	RNA-seq	K562	HG38	ENCFF509AOR
K562_total_RNA-seq_4	RNA-seq	K562	HG38	ENCFF745GPL
K562_nuclear_polyA_RNA-seq_1	RNA-seq	K562	HG38	ENCLB278NDX
K562_nuclear_polyA_RNA-seq_2	RNA-seq	K562	HG38	ENCLB538THW
K562_nuclear_total_RNA-seq_1	RNA-seq	K562	HG38	ENCLB873LMQ
K562_nuclear_total_RNA-seq_2	RNA-seq	K562	HG38	ENCLB645CDM
K562 GRO-Seq	GRO-seq		HG19	GSM1480325
K562_ATAC_1	ATAC-seq	K562	HG19	GSM1782764
K562_ATAC_2	ATAC-seq	K562	HG19	GSM1782765
K562_POL2_1	CHIP-seq	K562	HG38	ENCFF730DLS
K562_POL2_2	CHIP-seq	K562	HG38	ENCFF668VIK
K562_POL2_3	CHIP-seq	K562	HG38	ENCFF285MBX
K562_POL2_4	CHIP-seq	K562	HG38	ENCFF182YZG
K562_POL2_5	CHIP-seq	K562	HG38	ENCFF099NYA
K562_POL2_6	CHIP-seq	K562	HG38	ENCFF881ONC
K562_H3K27me3_1	CHIP-seq	K562	HG38	ENCFF049HUP
K562_H3K27me3_2	CHIP-seq	K562	HG38	ENCFF031FSF
K562_H3K27ac	CHIP-seq	K562	HG38	ENCFF038DDS
K562_H3K9me3	CHIP-seq	K562	HG38	ENCFF371GMJ
K562_H3K4me3_1	CHIP-seq	K562	HG38	ENCFF961SPZ
K562_H3K4me3_2	CHIP-seq	K562	HG38	ENCFF909PMV
K562_H3K4me3_3	CHIP-seq	K562	HG38	ENCFF616DLO
K562_H3K4me3_4	CHIP-seq	K562	HG38	ENCFF148POZ
K562_H3K4me1	CHIP-seq	K562	HG38	ENCFF159VKJ
K562_EP300	CHIP-seq	K562	HG38	ENCFF755HCK

APPENDIX B. METHODS

Calculating numbers of coordinate-overlaps

The numbers of coordinate-overlapped transcripts (shown in Figure 2.4a-2.4c) are calculated by using the R package ChIPpeakAnno (Zhu *et al*., 2010; Zhu, 2013) with the "findOverlapsOfPeaks" function. Transcripts with a coordinate-overlapping of 1bp or more on the same strand are considered to be overlapped. If one transcript in one set is (or multiple transcripts are) overlapped with multiple transcripts in the other set, the number of overlapped transcripts is counted as the minimal number of involved transcripts in any of the two groups. The venn diagrams shown in Figure 2.4a-2.4c are plotted using the R package ChIPpeakAnno with the "makeVennDiagram".

Calculating proportions of transcripts coincident with GRO-seq/POL II signals

The K562 POLL II "bed narrowPeak" files in GRCh38 are downloaded from ENCODE. GRO-seq "bigwig" files in hg19 are downloaded from GEO (Edgar *et al*., 2002) and a liftover of the hg19 annotations to GRCh38.p10 were then generated using an online tool called Batch Coordinate Conversion (liftOver) in UCSC genome browser (Kent *et al*., 2002) (Table 4). Transcripts overlapped 1bp or more with GRO-seq/POL II peaks by coordinates are defined as transcripts coincident with GRO-seq/POL II signals. Overlapping between transcripts and GRO-seq/POL II peak regions are done by using the R package GenomicRanges (Lawrence *et al*., 2013) with the "findoverlaps" function.

Categorize transcripts into mRNA, intergenic RNA (iRNA), and antisense RNA (as-RNA)

We categorized the assembled RNAs into three sub groups based on their relative genomic locations to GENCODE (v25)-annotated protein-coding genes (Figure 3.1). We firstly overlapped the coordinates of all assembled RNAs with GENCODE annotated protein-coding genes by using the "findOverlaps" function in R package GenomicRanges (v1.32.3) (Lawrence *et al*., 2013). Those assembled-RNAs that were not overlapped with any protein-coding genes were categorized as intergenic RNAs (iRNAs). The RNAs overlapping with protein-coding genes on the same strand were spitted into two sub-groups: those with an overlapped region accounts for at least 50% of the assembled RNA region were categorized as 'mRNAs'; and the others were categorized as iRNAs.

Finally, the assembled RNAs whose coordinates overlapped with protein-coding genes on the opposite strand were identified as antisense RNAs. Among those antisense RNAs, the ones that overlapped with protein-coding promoters (1000 bp windows around TSS of genes) were further categorized as antisense RNAs at 5UTR; other antisense RNAs were then categorized as antisense RNAs at 3UTR.

Coding probability calculation

The coding probability of RNA transcripts was calculated using Coding Potential Calculator 2 (CPC2) (Kang *et al*., 2017). CPC2 assessed coding probability by employing a support vector machine model based on four sequence intrinsic features: Fickett TESTCODE score of DNA sequences (Fickett, 1982), open reading frame (ORF) length, ORF integrity, and isoelectric point.

AUC analysis

AUC analysis was performed using the ROCR (v1.0-7) package in R (Sing *et al*., 2005). The commonly identified 731 cheRNAs or 3573 sneRNAs by all four pipelines were used as gold standard to calculate the accuracy of prediction in AUC analysis.

Chromatin states analysis and comparison

When comparing chromatin states of interested loci, we used ChIP-seq signals directly from BAM files instead of the published peak files for better sensitivity. Files meeting the following criteria were included in the analysis: (1). Format = Bam; (2) Genome version = GRCh38; (3). Output type = alignments.

To compare different chromatin features and chromatin accessibility, the metagene analysis was performed at either body regions or promoter regions (±1kb of TSS) of RNAs using the Bioconductor package metagene (v2.14.0) (Noguchi *et al*., 2006). When comparing ChIP-Seq signals using the downloaded bam files (which may ignore the ChIP-seq input control) with metagene analyses, we input not only the bam file for a histone mark but also its input control. Briefly, three steps were performed for meta-gene analysis:

1) The read coverages of all selected regions were extracted from BAM files and normalized to reads per million aligned (RPM) using the Bioconductor package metagene.
2) We divided each interested region into 100 equally-sized bins and calculated the averaged RPM within each bin.
3) Metagene profiles were plotted in the format of a ribbon plot or a box plot. If plotted in a ribbon plot, lines represent averaged RPM and ribbons represent the 95% confidence interval of the mean calculated using 1000 bootstraps; If plotted in a box plot, each box represents the distribution of averaged RPM at each bin.
4) To statistically compare two averaged RPM distributions, two-sided Wilcoxon rank sum test was performed to calculate p-value.

Retrieving ChromHMM predicted enhancer-driven RNAs (eRNAs)

To retrieve ChromHMM predicted eRNAs in K562 cell line, we downloaded the broad Chromatin State Segmentation by Hidden Markov Model from ENCODE (Broad ChromHMM) (Ernst and Kellis, 2012) profile in hg19 for K562 cell line from ENCODE (http://genome.ucsc.edu/encode/downloads.html). A map of these downloaded hg19 annotations to GRCh38.p10 was then conducted using an online tool called Batch Coordinate Conversion (liftOver) in the UCSC genome browser (Kent *et al*., 2002). In this work ChromHMM-predicted eRNAs were defined as intergenic RNAs overlap (at least 1 base) with ChromHMM-predicted "Strong enhancer" regions.

Retrieving FANTOM profiles

To retrieve FANTOM-predicted eRNAs in the K562 cell line, we downloaded the FANTOM-predicted enhancer regions in hg19 (ubiquitous_enhancers_cells.bed.txt) from FANTOM5 consortium (http://slidebase.binf.ku.dk/human_enhancers/presets) (Andersson *et al*., 2014). A liftover of the hg19 annotations to GRCh38.p10 for the downloaded profile were then generated using an online tool called Batch Coordinate Conversion (liftOver) in UCSC genome browser (Kent *et al*., 2002). FANTOM-predicted eRNAs were defined as intergenic RNAs overlap (at least 1 base) with FANTOM-predicted enhance regions.

Relative density of correlation between intergenic RNAs and neighbor coding genes

We calculated the pairwise Pearson correlation coefficient (PCC) between the intergenic RNA and protein-coding gene. We tested five types of intergenic RNA-gene groups: the icheRNA with random protein-coding gene pairs; the icheRNA with neighbor protein-coding gene pairs (icheRNA:neighborCoding); ChromHMM-predicted eRNAs with neighbor protein-coding genes pairs (ChromHMM-neighborCoding), FANTOM-predicted eRNAs with neighbor protein-coding genes pairs (FANTOM-neighborCoding) and the isneRNA with neighbor protein-coding gene pairs (isneRNA:neighborCoding).

The PCC of each intergenic RNA-gene pair was calculated based on their expression levels across all CPE and SNE samples of three cell types (K563, HEK293, H1-hESC). To pair an intergenic RNA with its neighbor protein-coding gene out of its nearest upstream and downstream genes on the same strand, the one with the highest absolute PCC value is selected. A significant cutoff of PCC values was set at -0.8 or 0.8, respectively for the negative or positive correlation. Kernel density is estimated for each intergenic RNA-gene pair group. Relative density for each intergenic RNA and neighboring protein-coding gene pairs group (e.g. icheRNA:neighborCoding) is calculated in the way of dividing the kernel density estimates of indicated intergenic RNA and neighboring protein-coding gene pairs group (e.g. icheRNA:neighborCoding) by the kernel density of icheRNA and randomly selected coding gene pairs group.

RNA structural analysis based on the Rfam annotations

Each annotating family in Rfam (v13, hg38) is represented by a multiple-sequence alignment, a consensus secondary structure, and a covariance model (Kalvari *et al*., 2018); and we grouped one or more annotating families into a super-family according to their function as well proportions in the above noncoding transcriptome. The homologous ncRNA sequences in each super-family were generally less than 400 bp (2.6 on the log10-scale, Fig 9), much shorter than the ncRNAs in the assembled transcriptome. Therefore, only when a ncRNA transcript fully covers an annotating family sequence, we annotated this ncRNA transcript with a Rfam super-family.

To assess the probabilities of a Rfam super-family (i) among a set of ncRNAs of interest (t), we calculated the ratio of ratios (RR) using Formula 1.

$$RR(t,r) = \frac{|t\ overlap\ r|}{|t|} \Big/ \frac{|t\ overlap\ T|}{|T|}, \qquad \text{Formula 1}$$

where T={t} is the collection of all noncoding transcripts in the transcriptome, and |.| is the number of transcripts meeting a condition.

This RR score was calculated for each Rfam-family for its frequency within a ncRNA set (t) versus its global frequency. Therefore, an RR-score above 2 indicates a ncRNA set (t)-selective RNA structural family.

polyA RNA-seq and total RNA-seq analysis

To compare expression levels of nuclear RNAs in different RNAseq libraries, we downloaded the raw sequencing datasets of K562 nuclear polyA RNA-seq (GSE88339) and nuclear total RNA-seq (GSE87982) from ENCODE data portal (Table 4). Reads were aligned to the human genome version GRCh38.p10 using Tophat (v2.1.1) (Kim *et al*., 2013) (e.g., tophat -p 8 --library-type=fr-firststrand -G gencode.v25.gtf GRCH38.genome -o polyA1 polyA1.fastq). Then the Fragments Per Kilobase Million (FPKM) of RNA transcripts were calculated using Cufflinks (v2.2.1) (Trapnell *et al*., 2012) (e.g., cufflinks -p 8 -u -N -library-type fr-firststrand -o FPKM_polyA1 -G gtf polyA1.bam). When making dot plots in Figures 5 and Figure 6, only expressed RNAs (with CPM>1) were plotted.

ChIP-seq peak signal

ChIP-seq peak signals were downloaded from ENCODE as “bed narrowPeak” files (Table 4). When one sample includes several replicates, we used the “bed narrowPeak” file with the Irreproducible Discovery Rate (IDR) values thresholded at the optimization precision (“optimal idr thresholded peaks”). When multiple samples are available and collected for one mark, we generated a union signal which was the pool of ChIP-seq peaks identified at least once from biological replicates. All files were downloaded with GRCh38 mapping assembly.

These bed/wig files generated from ENCODE used a score associated with each peak (enriched interval) which is the mean signal value across the interval. (Note that a broad region with moderate enrichment may deviate from the background more significantly than a short region with

high signal.) The input control information is on the same page where the bed/wig/bam file is download.

Calculating proportions of transcripts overlapping with LADs

The genomic coordinates of human (hg19) fibroblast LADs are downloaded from ENCODE (http://compbio.med.harvard.edu/modencode/webpage/lad/human.fibroblast.DamID.hg19.bed) and a liftover of the hg19 annotations to GRCh38.p10 were then generated using an online tool called Batch Coordinate Conversion (liftOver) in UCSC genome browser (Kent *et al*., 2002). Transcripts embedded in LADs are defined if more than 50% of the transcript overlaps with genomic coordinates of LADs. Overlapping between genomic coordinates of transcripts and LADs is done by using the R package GenomicRanges (Lawrence *et al*., 2013) with the "findoverlaps" function.

Calculating proportions of transcripts overlapping with class 1 TEs

The annotation of class 1 TEs in human (GRCh38.p10) is downloaded from RepeatMasker (http://www.repeatmasker.org/species/hg.html) (Yang *et al*., 2004). Transcripts overlapping with class 1 TEs are defined if the sequence of the transcript contains at least one sequence of class 1 TEs.

APPENDIX C. SOURCE FILE FOR TUXEDO PIPELINE

```
##############################################################################
# 1. Genome Alignment using Tophat
##############################################################################

#######         K562_CPE1
tophat --rg-sample SRR3703288 --rg-id SRR3703288 --library-type=fr-firststrand \
--segment-length 50 --segment-mismatches 2 --no-coverage-search --keep-fasta-order -p 32 \
-o /homeDir/GSE83531_cheRNA/tophat/SRR3703288 \
/homeDir/Reference_Sequences/Human/FASTA/GRCh38.primary_assembly.genome \
/homeDir/GSE83531_cheRNA/FastQ/SRR3703288.fastq.gz

#######         K562_CPE2
tophat --rg-sample SRR3703289 --rg-id SRR3703289 --library-type=fr-firststrand \
--segment-length 50 --segment-mismatches 2 --no-coverage-search --keep-fasta-order -p 32 \
-o /homeDir/GSE83531_cheRNA/tophat/SRR3703289 \
/homeDir/Reference_Sequences/Human/FASTA/GRCh38.primary_assembly.genome \
/homeDir/GSE83531_cheRNA/FastQ/SRR3703289.fastq.gz

#######         K562_CPE3
tophat --rg-sample SRR3703290 --rg-id SRR3703290 --library-type=fr-firststrand \
--segment-length 50 --segment-mismatches 2 --no-coverage-search --keep-fasta-order -p 32 \
-o /homeDir/GSE83531_cheRNA/tophat/SRR3703290 \
```

```
/homeDir/Reference_Sequences/Human/FASTA/GRCh38.primary_assembly.genome \
/homeDir/GSE83531_cheRNA/FastQ/SRR3703290.fastq.gz

#######         K562_SNE1
tophat --rg-sample SRR3703291 --rg-id SRR3703291 --library-type=fr-firststrand \
--segment-length 50 --segment-mismatches 2 --no-coverage-search --keep-fasta-order -p 32 \
-o /homeDir/GSE83531_cheRNA/tophat/SRR3703291 \
/homeDir/Reference_Sequences/Human/FASTA/GRCh38.primary_assembly.genome \
/homeDir/GSE83531_cheRNA/FastQ/SRR3703291.fastq.gz

#######         K562_SNE2
tophat --rg-sample SRR3703292 --rg-id SRR3703292 --library-type=fr-firststrand \
--segment-length 50 --segment-mismatches 2 --no-coverage-search --keep-fasta-order -p 32 \
-o /homeDir/GSE83531_cheRNA/tophat/SRR3703292 \
/homeDir/Reference_Sequences/Human/FASTA/GRCh38.primary_assembly.genome \
/homeDir/GSE83531_cheRNA/FastQ/SRR3703292.fastq.gz

#######         K562_SNE3
tophat --rg-sample SRR3703293 --rg-id SRR3703293 --library-type=fr-firststrand \
--segment-length 50 --segment-mismatches 2 --no-coverage-search --keep-fasta-order -p 32 \
-o /homeDir/GSE83531_cheRNA/tophat/SRR3703293 \
/homeDir/Reference_Sequences/Human/FASTA/GRCh38.primary_assembly.genome \
/homeDir/GSE83531_cheRNA/FastQ/SRR3703293.fastq.gz
```

```
#############################################################################
# 2. clean bam files using samtools to remove reads mapped mitochondria chromosome
#############################################################################

#######         K562_CPE1
samtools idxstats /homeDir/GSE83531_cheRNA/All_bam/SRR3703288_accepted_hits.bam \
| cut -f 1 | grep 'chr' | grep -v 'chrM' | \
xargs samtools view -b
/homeDir/GSE83531_cheRNA/All_bam/SRR3703288_accepted_hits.bam > \
/homeDir/GSE83531_cheRNA/cleaned_bam/SRR3703288_clean.bam
samtools index /homeDir/GSE83531_cheRNA/cleaned_bam/SRR3703288_clean.bam

#######         K562_CPE2
samtools idxstats /homeDir/GSE83531_cheRNA/All_bam/SRR3703289_accepted_hits.bam \
| cut -f 1 | grep 'chr' | grep -v 'chrM' | \
xargs samtools view -b
/homeDir/GSE83531_cheRNA/All_bam/SRR3703289_accepted_hits.bam > \
/homeDir/GSE83531_cheRNA/cleaned_bam/SRR3703289_clean.bam
samtools index /homeDir/GSE83531_cheRNA/cleaned_bam/SRR3703289_clean.bam

#######         K562_CPE3
samtools idxstats /homeDir/GSE83531_cheRNA/All_bam/SRR3703290_accepted_hits.bam \
| cut -f 1 | grep 'chr' | grep -v 'chrM' | \
```

```
xargs samtools view -b
/homeDir/GSE83531_cheRNA/All_bam/SRR3703290_accepted_hits.bam > \
/homeDir/GSE83531_cheRNA/cleaned_bam/SRR3703290_clean.bam
samtools index /homeDir/GSE83531_cheRNA/cleaned_bam/SRR3703290_clean.bam

#######         K562_SNE1
samtools idxstats /homeDir/GSE83531_cheRNA/All_bam/SRR3703291_accepted_hits.bam \
| cut -f 1 | grep 'chr' | grep -v 'chrM' | \
xargs samtools view -b
/homeDir/GSE83531_cheRNA/All_bam/SRR3703291_accepted_hits.bam > \
/homeDir/GSE83531_cheRNA/cleaned_bam/SRR3703291_clean.bam
samtools index /homeDir/GSE83531_cheRNA/cleaned_bam/SRR3703291_clean.bam

#######         K562_SNE2
samtools idxstats /homeDir/GSE83531_cheRNA/All_bam/SRR3703292_accepted_hits.bam \
| cut -f 1 | grep 'chr' | grep -v 'chrM' | \
xargs samtools view -b
/homeDir/GSE83531_cheRNA/All_bam/SRR3703292_accepted_hits.bam > \
/homeDir/GSE83531_cheRNA/cleaned_bam/SRR3703292_clean.bam
samtools index /homeDir/GSE83531_cheRNA/cleaned_bam/SRR3703292_clean.bam

#######         K562_SNE3
samtools idxstats /homeDir/GSE83531_cheRNA/All_bam/SRR3703293_accepted_hits.bam \
```

```
| cut -f 1 | grep 'chr' | grep -v 'chrM' | \
xargs samtools view -b
/homeDir/GSE83531_cheRNA/All_bam/SRR3703293_accepted_hits.bam > \
/homeDir/GSE83531_cheRNA/cleaned_bam/SRR3703293_clean.bam
samtools index /homeDir/GSE83531_cheRNA/cleaned_bam/SRR3703293_clean.bam

#############################################################################
# 3. RABT assembly using Cufflinks
#############################################################################

#######         K562_CPE1
cufflinks -o /homeDir/GSE83531_cheRNA/cufflinks_rabt/SRR3703288 \
-p 32 --library-type fr-firststrand \
-g
/homeDir/Reference_Sequences/Human/Annotation/gencode.v25.primary_assembly.annotation.
gtf \
-u /homeDir/GSE83531_cheRNA/cleaned_bam/SRR3703288_clean.bam

#######         K562_CPE2
cufflinks -o /homeDir/GSE83531_cheRNA/cufflinks_rabt/SRR3703289 \
-p 32 --library-type fr-firststrand \
```

```
-g
/homeDir/Reference_Sequences/Human/Annotation/gencode.v25.primary_assembly.annotation.
gtf \
-u /homeDir/GSE83531_cheRNA/cleaned_bam/SRR3703280_clean.bam

#######         K562_CPE3
cufflinks -o /homeDir/GSE83531_cheRNA/cufflinks_rabt/SRR3703290 \
-p 32 --library-type fr-firststrand \
-g
/homeDir/Reference_Sequences/Human/Annotation/gencode.v25.primary_assembly.annotation.
gtf \
-u /homeDir/GSE83531_cheRNA/cleaned_bam/SRR3703290_clean.bam

#######         K562_SNE1
cufflinks -o /homeDir/GSE83531_cheRNA/cufflinks_rabt/SRR3703291 \
-p 32 --library-type fr-firststrand \
-g
/homeDir/Reference_Sequences/Human/Annotation/gencode.v25.primary_assembly.annotation.
gtf \
-u /homeDir/GSE83531_cheRNA/cleaned_bam/SRR3703291_clean.bam

#######         K562_SNE2
cufflinks -o /homeDir/GSE83531_cheRNA/cufflinks_rabt/SRR3703292 \
```

```
-p 32 --library-type fr-firststrand \
-g
/homeDir/Reference_Sequences/Human/Annotation/gencode.v25.primary_assembly.annotation.
gtf \
-u /homeDir/GSE83531_cheRNA/cleaned_bam/SRR3703292_clean.bam

#######         K562_SNE3
cufflinks -o /homeDir/GSE83531_cheRNA/cufflinks_rabt/SRR3703293 \
-p 32 --library-type fr-firststrand \
-g
/homeDir/Reference_Sequences/Human/Annotation/gencode.v25.primary_assembly.annotation.
gtf \
-u /homeDir/GSE83531_cheRNA/cleaned_bam/SRR3703293_clean.bam

##############################################################################
# 4. Build transcriptome using Cuffmerge
##############################################################################

cuffmerge -o /homeDir/GSE83531_cheRNA/cuffmerge_rabt_k562/ -p 32 \
-g
/homeDir/Reference_Sequences/Human/Annotation/gencode.v25.primary_assembly.annotation.
gtf \
-s /homeDir/Reference_Sequences/Human/FASTA/GRCh38.primary_assembly.genome.fa \
```

```
/homeDir/GSE83531_cheRNA/cufflinks_rabt/che_sne_K562.txt

##############################################################################
# 5. Get geneCounts using htseq
##############################################################################

#######         K562_CPE1
htseq-count -f bam -s reverse -m intersection-nonempty \
/homeDir/GSE83531_cheRNA/cleaned_bam/K562_cleaned_bam/SRR3703288_clean.bam \
/homeDir/GSE83531_cheRNA/cuffmerge_rabt_k562/merged.gtf > \
/homeDir/GSE83531_cheRNA/HTseq_ruthenburg_1000_K562/SRR3703288_geneCounts.out

#######         K562_CPE2
htseq-count -f bam -s reverse -m intersection-nonempty \
/homeDir/GSE83531_cheRNA/cleaned_bam/K562_cleaned_bam/SRR3703289_clean.bam \
/homeDir/GSE83531_cheRNA/cuffmerge_rabt_k562/merged.gtf > \
/homeDir/GSE83531_cheRNA/HTseq_ruthenburg_1000_K562/SRR3703289_geneCounts.out

#######         K562_CPE3
htseq-count -f bam -s reverse -m intersection-nonempty \
/homeDir/GSE83531_cheRNA/cleaned_bam/K562_cleaned_bam/SRR3703290_clean.bam \
/homeDir/GSE83531_cheRNA/cuffmerge_rabt_k562/merged.gtf > \
/homeDir/GSE83531_cheRNA/HTseq_ruthenburg_1000_K562/SRR3703290_geneCounts.out
```

```
#######         K562_SNE1

htseq-count -f bam -s reverse -m intersection-nonempty \

/homeDir/GSE83531_cheRNA/cleaned_bam/K562_cleaned_bam/SRR3703291_clean.bam \

/homeDir/GSE83531_cheRNA/cuffmerge_rabt_k562/merged.gtf > \

/homeDir/GSE83531_cheRNA/HTseq_ruthenburg_1000_K562/SRR3703291_geneCounts.out


#######         K562_SNE2

htseq-count -f bam -s reverse -m intersection-nonempty \

/homeDir/GSE83531_cheRNA/cleaned_bam/K562_cleaned_bam/SRR3703292_clean.bam \

/homeDir/GSE83531_cheRNA/cuffmerge_rabt_k562/merged.gtf > \

/homeDir/GSE83531_cheRNA/HTseq_ruthenburg_1000_K562/SRR3703292_geneCounts.out


#######         K562_SNE3

htseq-count -f bam -s reverse -m intersection-nonempty \

/homeDir/GSE83531_cheRNA/cleaned_bam/K562_cleaned_bam/SRR3703293_clean.bam \

/homeDir/GSE83531_cheRNA/cuffmerge_rabt_k562/merged.gtf > \

/homeDir/GSE83531_cheRNA/HTseq_ruthenburg_1000_K562/SRR3703293_geneCounts.out


##############################################################################

# 6. Differential expression analysis using limma package in R

##############################################################################

###### Reference: https://www.bioconductor.org/help/workflows/RNAseq123/
```

```
library(limma)
library(edgeR)

###### read htseq geneCounts into a DGE matrix used in edgeR
htseqToDGE <- function(files_dir){
  library(edgeR)
  files <- list.files(path=files_dir, pattern="geneCounts")
  files <- sort(files, decreasing=T)
  files_fullPath <- sapply(files, function(x) paste0(files_dir,x))
  #########
  raw_DGE <- readDGE(files_fullPath, header=F)
  list <- c("no_feature", "ambiguous", "too_low_aQual", "not_aligned", "alignment_not_unique")
  for(i in list){
    tmp <- grep(i, rownames(raw_DGE$counts))
    if(length(tmp)>0) raw_DGE$counts <- raw_DGE$counts[-tmp,]
  }
  colnames(raw_DGE) <- sapply(files, function(x) unlist(strsplit(x, "_", fixed=T))[1])
  group <- as.factor(c("SNE", "SNE", "SNE", "CPE", "CPE", "CPE"))
  raw_DGE$samples$group <- group
  return(raw_DGE)
}

###### filtered raw counts, normalize raw counts and make density plots and barplots
```

```
filterNormPlot <- function(raw_DGE, method, cpm_cutoff){
  library(edgeR)
  library(RColorBrewer)

  # plot log-cpm of raw data
  par(mfrow=c(2,2))
  nsamples <- ncol(raw_DGE)
  col <- brewer.pal(nsamples, "RdYlGn")
  samplenames <- colnames(tuxedo_DGE)
  lcpm <- cpm(raw_DGE, log=TRUE)
  plot(density(lcpm[,1]), col=col[1], lwd=2, las=2, ylim=c(0,0.5),
       main="", xlab="")
  title(main=paste0("A. ",method," Raw data"), xlab="Log2-cpm")
  abline(v=0, lty=3)
  for (i in 2:nsamples){
    den <- density(lcpm[,i])
    lines(den$x, den$y, col=col[i], lwd=2)
  }
  legend("topright", samplenames, text.col=col, bty="n")

  # filter by cpm
  cpm <- cpm(raw_DGE)
  keep.exprs <- rowSums(cpm>cpm_cutoff)>=3
```

```
filtered_DGE <- raw_DGE[keep.exprs, keep.lib.sizes=FALSE]
dim(filtered_DGE)
filtered_lcpm <- cpm(filtered_DGE, log=TRUE)
plot(density(filtered_lcpm[,1]), col=col[1], lwd=2, las=2, ylim=c(0,0.5),
    main="", xlab="")
title(main=paste0("B. ", method, " Filtered data"), xlab="Log2-cpm")
abline(v=0, lty=3)
for (i in 2:nsamples){
  den <- density(filtered_lcpm[,i])
  lines(den$x, den$y, col=col[i], lwd=2)
}
legend("topright", samplenames, text.col=col, bty="n")

# plot unnormalized data
boxplot(filtered_lcpm, las=2, col=col, main="")
title(main=paste0("C. Example: ",method," Unnormalized filtered data"),ylab="Log2-cpm")

# normalization
norm_DGE <- calcNormFactors(filtered_DGE, method = "TMM")
norm_DGE$samples$norm.factors
norm_lcpm <- cpm(norm_DGE, log=TRUE)
boxplot(norm_lcpm, las=2, col=col, main="")
title(main=paste0("D. Example: ",method," Normalised filtered data"),ylab="Log2-cpm")
```

```
  return(norm_DGE)
}

limmaFitDE <- function(norm_DGE, method){
  library(limma)
  group <- norm_DGE$samples$group
  design <- model.matrix(~0+group)
  contr.matrix <- makeContrasts(CPEvsSNE = groupCPE-groupSNE, levels = colnames(design))

  par(mfrow=c(1,2))
  v <- voom(norm_DGE, design, plot=TRUE)
  vfit <- lmFit(v, design)
  vfit <- contrasts.fit(vfit, contrasts=contr.matrix)
  efit <- eBayes(vfit)
  plotSA(efit, main=paste0(method, " Final model: Mean???variance trend"))
  return(efit)
}

tuxedo_path <- "HTseq_tuxedo/"
tuxedo_DGE <- htseqToDGE(tuxedo_path)

pdf(file="limma_filter_norm_plot_2.pdf", width=10, height=8)
tuxedo_normDGE <- filterNormPlot(tuxedo_DGE, "Tuxedo", 1)
```

```
dev.off()

tuxedo_efit <- limmaFitDE(tuxedo_normDGE, "Tuxedo")

summary(decideTests(tuxedo_efit, p.value = 0.05, lfc=log2(1.2)))

tuxedo_limmaRES <- topTreat(tuxedo_efit, sort="none",number=Inf)
tuxedo_limmaRES$decision <- decideTests(tuxedo_efit, p.value = 0.05, lfc=log2(1.2))
```

REFERENCES

Almeida, M.I., Reis, R.M. and Calin, G.A. (2011). MicroRNA history: discovery, recent applications, and next frontiers. Mutat Res 717, 1-8.

Agarwal, V., Bell, G.W., Nam, J.W., and Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. eLIFE 4, e05005.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25, 3389–3402.

Ambros, V. (1989). A hierarchy of regulatory genes controls a larva-to-adult developmental switch in C. elegans. Cell 57, 49-57.

Ambros, V. and Horvitz, H.R. (1987). The lin-14 locus of Caenorhabditis elegans controls the time of expression of specific postembryonic developmental events. Genes Dev 1, 398-414.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166-169.

Andersson, R. (2015). Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. Bioessays 37, 314-23.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F.O., Jørgensen, M., Andersen, P.R., Bertin, N., Rackham, O., Burroughs, A.M., Baillie, J.K., Ishizu, Y., Shimizu, Y., Furuhata, E., Maeda, S., Negishi, Y., Mungall, C.J., Meehan, T.F., Lassmann, T., Itoh, M., Kawaji, H., Kondo, N., Kawai, J., Lennartsson, A., Daub, C.O., Heutink, P., Hume, D.A., Jensen, T.H., Suzuki, H., Hayashizaki, Y., Müller, F., Forrest, A.R.R., Carninci, P., Rehli, M., and Sandelin, A. (2014). An atlas of active enhancers across human cell types and tissues. Nature 507, 455-461.

Antequera, F. and Bird, A. (1993). Number of CpG islands and genes in human and mouse. Proc Natl Acad Sci USA 90, 11995-9.

Ayupe, A.C., Tahira, A.C., Camargo, L., Beckedorff, F.C., Verjovski-Almeida, S., and Reis, E.M. (2015). Global analysis of biogenesis, stability and sub-cellular localization of lncRNAs mapping to intragenic regions of the human genome. RNA Biol 12, 877–892.

Bal, E., Park, H.S., Belaid-Choucair, Z., Kayserili, H., Naville, M., Madrange, M., Chiticariu, E., Hadj-Rabia, S., Cagnard, N., Kuonen, F., Bachmann, D., Huber, M., Le Gall, C., Côté, F., Hanein, S., Rosti, R.Ö., Aslanger, A.D., Waisfisz, Q., Bodemer, C., Hermine, O., Morice-Picard, F., Labeille, B., Caux, F., Mazereeuw-Hautier, J., Philip, N., Levy, N., Taieb, A., Avril, M.F., Headon, D.J., Gyapay, G., Magnaldo, T., Fraitag, S., Crollius, H.R., Vabres, P., Hohl, D., Munnich, A., and Smahi, A. (2017). Mutations in ACTRT1 and its enhancer RNA elements lead to aberrant

activation of Hedgehog signaling in inherited and sporadic basal cell carcinomas. Nat Med 23, 1226-1233.

Baldassarre, A. and Masotti, A. (2012). Long non-coding RNAs and p53 regulation. Int J Mol Sci 13, 16708–16717.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell 129, 823-837.

Becker, J.S., Nicetto, D., and Zaret, K.S. (2016). H3K9me3-Dependent Heterochromatin: Barrier to Cell Fate Changes. Trends Genet 32, 29-41.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., Sharon, E., Spector, Y., and Bentwich, Z. (2005). Identification of hundreds of conserved and nonconserved human microRNAs. Nat Genet 37, 766–770.

Bhan, A. and Mandal, S.S. (2016). Estradiol-induced transcriptional regulation of long non-coding RNA, HOTAIR. Methods Mol Biol 1366, 395–412.

Bhan, A., and Mandal, S.S. (2015). LncRNA HOTAIR: A master regulator of chromatin dynamics and cancer. Biochim Biophys Acta 1856, 151-164.

Bian, Q., Khanna, N., Alvikas, J., and Belmont, A.S. (2013). beta-Globin cis-elements determine differential nuclear targeting through epigenetic modifications. J Cell Biol 203, 767-783.

Bourgon, R., Gentleman, R., and Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. Proc Natl Acad Sci U S A 107, 9546-9551.

Brown, C.J., Hendrich, B.D., Rupert, J.L., Lafrenière, R.G., Xing, Y., Lawrence, J., and Willard, H.F. (1992). The human XIST gene: Analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell 71, 527–542.

Cai, Y., Yu, X., Hu, S., and Yu, J. (2009). A brief review on the mechanisms of miRNA regulation. Genomics Proteomics Bioinformatics 7, 147–154.

Catalanotto, C., Cogoni, C., and Zardo, G. (2016). MicroRNA in control of gene expression: MicroRNA in control of gene expression. Int J Mol Sci 17, 1712

Chan, R.C.W., Libbrecht, M.W., Roberts, E.G., Bilmes, J.A., Noble, W.S. and Hoffman, M.M. (2018). Segway 2.0: Gaussian mixture models and minibatch training. Bioinformatics 34, 669–671.

Chen, H., Li, C., Peng, X., Zhou, Z., Weinstein, J.N., Cancer Genome Atlas Research Network and Liang, H. (2018). A Pan-Cancer Analysis of Enhancer Expression in Nearly 9000 Patient Samples. Cell 173, 386-399

Churchman, L.S. and Weissman, J.S. (2012). Native elongating transcript sequencing (NET-seq). Curr Protoc Mol Biol Chapter 4, Unit 4.14.1-17.

Clemson, C.M., McNeil, J.A., Willard, H.F., and Lawrence, J.B. (1996). XIST RNA paints the inactive X chromosome at interphase: Evidence for a novel RNA involved in nuclear/chromosome structure. J Cell Biol 132, 259–275.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., Boyer, L.A., Young, R.A., and Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A 107, 21931-21936.

Crick, F.H. (1958). On protein synthesis. Symp Soc Exp Biol 12, 138-63.

de Hoon, M., Shin, J.W., and Carninci, P. (2015). Paradigm shifts in genomics through the FANTOM projects. Mamm Genome 26, 391-402.

De Lay, N., Schu, D.J. and Gottesman, S. (2013). Bacterial small RNA-based negative regulation: Hfq and its accomplices. J Biol Chem 288, 7996–8003.

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H., Ragoussis, J., Wei, C.L., and Natoli, G. (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol 8, e1000384.

Dekker, J., Rippe, K., Dekker, M. and Kleckner, N. (2002). Capturing chromosome conformation. Science 295, 1306-11.

Demirci, M.D.S., Baumbach, J., and Allmer, J. (2017). On the performance of pre-microRNA detection algorithms. Nat Commun 8, 330.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J.B., Lipovich, L., Gonzalez, J.M., Thomas, M., Davis, C.A., Shiekhattar, R., Gingeras, T.R., Hubbard, T.J., Notredame, C., Harrow, J., and Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res 22, 1775-1789.

DIANA-TOOLS. (2016). DIANA-TOOLS. Available at: diana.imis.athena-innovation.gr (accessed 21.10.17).

Dorighi, K.M., Swigut, T., Henriques, T., Bhanu, N.V., Scruggs, B.S., Nady, N., Still, C.D., 2nd, Garcia, B.A., Adelman, K., and Wysocka, J. (2017). Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. Mol Cell 66, 568-576 e564.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., Green, R.D., and Dekker, J. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res 16, 1299-309.

Eddy, S.R. (2001). Non-coding RNA genes and the modern RNA world. Nat Rev Genet 2, 919-29.

Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30, 207-210.

Ernst, J. and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc 12, 2478-2492.

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat Methods 9, 215-216.

Esteller, M. (2011). Non-coding RNAs in human disease. Nat Rev Genet 12, 861–874.

Fan, X. and Kurgan, L. (2015). Comprehensive overview and assessment of computational prediction of microRNA targets in animals. Brief Bioinform 16, 780–794.

Fickett, J.W. (1982). Recognition of protein coding regions in DNA sequences. Nucleic Acids Res 10, 5303-5318.

Fields, C., Adams, M.D., White, O. and Venter, J.C. (1994). How many genes in the human genome? Nature genetics Vol.7, p.345-346

Flynn, R.A. and Chang, H.Y. (2014). Long noncoding RNAs in cell-fate programming and reprogramming. Cell Stem Cell 14, 752–761.

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

Fullwood, M.J. and Ruan, Y. (2009). ChIP-based methods for the identification of long-range chromatin interactions. J Cell Biochem 107, 30–39.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., Chew, E.G.Y., Huang, P.Y.H., Welboren, W.J., Han, Y., Ooi, H.S., Ariyaratne, P.N., Vega, V.B., Luo, Y., Tan, P.Y., Choy, P.Y., Wansa, K.D.S.A., Zhao, B., Lim, K.S., Leow, S.C., Yow, J.S., Joseph, R., Li, H., Desai, K.V., Thomsen, J.S., Lee, Y.K., Karuturi, R.K.M., Herve, T., Bourque, G., Stunnenberg, H.G., Ruan, X., Cacheux-Rataboul, V., Sung, W.K., Liu, E.T., Wei, C.L., Cheung, E., and Ruan, Y. (2009). An oestrogen-receptor-α-bound human chromatin interactome. Nature 462, 58-64.

Gao, T., He, B., Liu, S., Zhu, H., Tan, K. and Qian, J. (2016). “EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. Bioinformatics 32, 3543-3551.

Gardini, A. (2017). Global Run-On Sequencing (GRO-Seq). Methods Mol Biol 1468, 111-20.

Gayen, S., and Kalantry, S. (2017). Chromatin-enriched lncRNAs: a novel class of enhancer RNAs. Nat Struct Mol Biol 24, 556-557.

Gayon, J. (2016). From Mendel to epigenetics: History of genetics. C R Biol 339, 225-30.

GENCODE, 2017. GENCODE. Available at: https://www.gencodegenes.org/ (accessed 21.10.17).

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29, 644–652.

Griffith, M., Walker, J.R., Spies, N.C., Ainscough, B.J., and Griffith, O.L. (2015). Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. PLoS Comput Biol 11, e1004393.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., and Enright, A.J. (2006). miRBase: MicroRNA sequences, targets and gene nomenclature. Nucleic Acids Res 34, D140–D144.

Griffiths-Jones, S., Hui, J.H., Marco, A., and Ronshaugen, M. (2011). MicroRNA evolution by arm switching. EMBO Rep 12, 172–177.

Guhaniyogi, J., and Brewer, G. (2001). Regulation of mRNA stability in mammalian cells. Gene 265, 11-23.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., Cabili, M.N., Jaenisch, R., Mikkelsen, T.S., Jacks, T., Hacohen, N., Bernstein, B.E., Kellis, M., Regev, A., Rinn, J.L., and Lander, E.S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458, 233–237.

Hah, N., Murakami, S., Nagari, A., Danko, C.G. and Kraus, W.L. (2013). Enhancer transcripts mark active estrogen receptor binding sites. Genome Res 23, 1210-23.

Hajjari, M. and Salavaty, A. (2015). HOTAIR: An oncogenic long non-coding RNA in different cancers. Cancer Biol Med 12, 1–9.

Harismendy, O., Notani, D., Song, X., Rahim, N.G., Tanasa, B., Heintzman, N., Ren, B., Fu, X.D., Topol, E.J., Rosenfeld, M.G., and Frazer, K.A. (2011). 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. Nature 470, 264-8.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J.M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A.,

Reymond, A., Gerstein, M., Guigó, R., and Hubbard, T.J. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 22, 1760-1774.

Hart, T., Komori, H.K., LaMere, S., Podshivalova, K., and Salomon, D.R. (2013). Finding the active genes in deep RNA-seq gene expression studies. BMC Genomics 14, 778.

Hauberg, M.E., Fullard, J.F., Zhu, L., Cohain, A.T., Giambartolomei, C., Misir, R., Reach, S., Johnson, J.S., Wang, M., Mattheisen, M., Børglum, A.D., Zhang, B., Sieberts, S.K., Peters, M.A., Domenici, E., Schadt, E.E., Devlin, B., Sklar, P., Roeder, K., Roussos, P., and CommonMind Consortium. (2018). Differential activity of transcribed enhancers in the prefrontal cortex of 537 cases with schizophrenia and controls. Mol Psychiatry, doi: 10.1038/s41380-018-0059-8.

He, J., Fu, X., Zhang, M., He, F., Li, W., Abdul, M.M., Zhou, J., Sun, L., Chang, C., Li, Y., Liu, H., Wu, K., Babarinde, I.A., Zhuang, Q., Loh, Y.H., Chen, J., Esteban, M.A., and Hutchins, A.P. (2019). Transposable elements are regulated by context-specific patterns of chromatin marks in mouse embryonic stem cells. Nat Commun 10, 34.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., Wang, W., Weng, Z., Green, R.D., Crawford, G.E., and Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39, 311-8.

Hendrix, D., Levine, M., and Shi, W. (2010). miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data. Genome Biol 11, R39.

Heward, J.A., Roux, B.T. and Lindsay, M.A. (2015). Divergent signalling pathways regulate lipopolysaccharide-induced eRNA expression in human monocytic THP1 cells. FEBS Lett 589, 396-406.

Hoagland, M.B., Stephenson, M.L., Scott, J.F., Hecht, L.I. and Zamecnik, P.C. (1958). A soluble ribonucleic acid intermediate in protein synthesis. J Biol Chem 231, 241-57.

Horner, H.A. and Macgregor, H.C. (1983). C value and cell volume: their significance in the evolution and development of amphibians. J Cell Sci 63, 135-46.

Hublitz, P., Albert, M., and Peters, A.H. (2009). Mechanisms of transcriptional repression by histone lysine methylation. Int J Dev Biol 53, 335-354.

Huda, A., and Jordan, I.K. (2009). Analysis of transposable element sequences using CENSOR and RepeatMasker. Methods Mol Biol 537, 323-336.

Hung, T., Wang, Y., Lin, M.F., Koegel, A.K., Kotake, Y., Grant, G.D., Horlings, H.M., Shah, N., Umbricht, C., Wang, P., Wang, Y., Kong, B., Langerød, A., Børresen-Dale, A.-L., Kim, S.K., van de Vijver, M., Sukumar, S., Whitfield, M.L., Kellis, M., Xiong, Y., Wong, D.J., and Chang, H.Y. (2011). Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. Nat Genet 43, 621–629.

Hüttenhofer, A., Schattner, P. and Polacek, N. (2005). Non-coding RNAs: hope or hype? Trends Genet 21, 289-97.

Hutvágner, G., McLachlan, J., Pasquinelli, A.E., Bálint, E., Tuschl, T. and Zamore, P.D. (2001). A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. Science 293, 834-8.

Inouye, M. and Delihas, N. (1988). Small RNAs in the prokaryotes: a growing list of diverse roles. Cell 53, 5-7.

Jarroux, J., Morillon, A. and Pinskaya, M. (2017). History, Discovery, and Classification of lncRNAs. Adv Exp Med Biol 1008, 1-46.

Johnsson, P., Lipovich, L., Grandér, D., and Morris, K.V. (2014). Evolutionary conservation of long non-coding RNAs; sequence, structure, function. Biochim Biophys Acta 1840, 1063–1071.

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R., Bateman, A., Finn, R.D., and Petrov, A.I. (2018). Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res 46, D335-D342.

Kang, Y.J., Yang, D.C., Kong, L., Hou, M., Meng, Y.Q., Wei, L., and Gao, G. (2017). CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res 45, W12-W16.

Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermüller, J., Hofacker, I.L., Bell, I., Cheung, E., Drenkow, J., Dumais, E., Patel, S., Helt, G., Ganesh, M., Ghosh, S., Piccolboni, A., Sementchenko, V., Tammana, H., and Gingeras, T.R. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316, 1484–1488.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res 12, 996-1006.

Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B.E., van Oudenaarden, A., Regev, A., Lander, E.S., and Rinn, J.L. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci U S A 106, 11667-11672.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14, R36.

Kim, T.K., Hemberg, M., and Gray, J.M. (2015). Enhancer RNAs: a class of long noncoding RNAs synthesized at enhancers. Cold Spring Harb Perspect Biol 7, a018622.

Kim, T.K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P.F., Kreiman, G., and Greenberg, M.E. (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature 465, 182-187.

Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., and Gao, G. (2007). CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res 35, W345–W349.

Kotake, Y., Nakagawa, T., Kitagawa, K., Suzuki, S., Liu, N., Kitagawa, M., and Xiong, Y. (2011). Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene. Oncogene 30, 1956-1962.

Kouzarides, T. (2007). Chromatin modifications and their function. Cell 128, 693-705.

Kozomara, A. and Griffiths-Jones, S. (2013). miRBase: Annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 42, D68–D73.

Kumar, A., Kankainen, M., Parsons, A., Kallioniemi, O., Mattila, P., and Heckman, C.A. (2017). The impact of RNA sequence library construction protocols on transcriptomic profiling of leukemia. BMC Genomics 18, 629.

Kung, J.T.Y., Colognori, D. and Lee, J.T. (2013). Long noncoding RNAs: past, present, and future. Genetics 193, 651-69.

Kwak, H., Fuda, N.J., Core, L.J. and Lis, J.T. (2013). Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. Science 339, 950-3.

Lai, F., Orom, U.A., Cesaroni, M., Beringer, M., Taatjes, D.J., Blobel, G.A. and Shiekhattar, R. (2013). Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature 494, 497-501.

Lam, M.T., Li, W., Rosenfeld, M.G., and Glass, C.K. (2014). Enhancer RNAs and regulated transcriptional programs. Trends Biochem Sci 39, 170-182.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N.,

Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J., Szustakowki, J., and International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. Nature 409, 860-921.

Latge, G., Poulet, C., Bours, V., Josse, C., and Jerusalem, G. (2018). Natural Antisense Transcripts: Molecular Mechanisms and Implications in Breast Cancers. Int J Mol Sci 19.

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for computing and annotating genomic ranges. PLoS Comput Biol 9, e1003118.

Le Gras, S., Keime, C., Anthony, A., Lotz, C., De Longprez, L., Brouillet, E., Cassel, J.C., Boutillier, A.L., and Merienne, K. (2017). Altered enhancer transcription underlies Huntington's disease striatal transcriptional signature. Sci Rep 7, 42875.

Lee, R.C., Feinbaum, R.L. and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75, 843-54.

Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A.Y., Merkurjev, D., Zhang, J., Ohgi, K., Song, X., Oh, S., Kim, H.S., Glass, C.K., and Rosenfeld, M.G. (2013). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature 498, 516-20.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A., Lander, E.S., and Dekker, J. (2009), "Comprehensive mapping of long-range interactions reveals folding principles of the human genome", Science (New York, N.Y.), U.S. National Library of Medicine, 9 October, available at: https://www.ncbi.nlm.nih.gov/pubmed/19815776/ (accessed 25 April 2019).

Lin, M.F., Jungreis, I. and Kellis, M. (2011). PhyloCSF: A comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics 27, i275–i282.

Liu, F. (2017). Enhancer-derived RNA: A Primer. Genomics Proteomics Bioinformatics 15, 196-200.

Lopes, I.O.N., Schliep, A., and de Carvalho, A.C.P. (2014). The discriminant power of RNA features for pre-miRNA recognition. BMC Bioinformatics 15, 124.

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

Lucas, S.J. and Budak, H. (2012). Sorting the wheat from the chaff: Identifying miRNAs in genomic survey sequences of Triticum aestivum chromosome 1AL. PLOS One 7, e40859.

Magistri, M., Faghihi, M.A., St Laurent, G., 3rd, and Wahlestedt, C. (2012). Regulation of chromatin structure by long noncoding RNAs: focus on natural antisense transcripts. Trends Genet 28, 389-396.

Mahat, D.B., Kwak, H., Booth, G.T., Jonkers, I.H., Danko, C.G., Patel, R.K., Waters, C.T., Munson, K., Core, L.J., and Lis, J.T. (2016). Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat Protoc 11, 1455-76.

Markham, N.R. and Zuker, M. (2008). UNAFold: Software for nucleic acid folding and hybridization. Methods Mol Biol 453, 3–31.

Mathews, D.H., Disney, M.D., Childs, J.L., Schroeder, S.J., Zuker, M., and Turner, D.H. (2004). Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc Natl Acad Sci, USA 101, 7287–7292.

Mayer, A., di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., Sandstrom, R., Stamatoyannopoulos, J.A., and Churchman, L.S. (2015). Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. Cell 161, 541-554.

Meng, H. and Bartholomew, B. (2018). Emerging roles of transcriptional enhancers in chromatin looping and promoter-proximal pausing of RNA polymerase II. J Biol Chem 293, 13786-13794.

Meyers, B.C., Axtell, M.J., Bartel, B., Bartel, D.P., Baulcombe, D., Bowman, J.L., Cao, X., Carrington, J.C., Chen, X., Green, P.J., Griffiths-Jones, S., Jacobsen, S.E., Mallory, A.C., Martienssen, R.A., Poethig, R.S., Qi, Y., Vaucheret, H., Voinnet, O., Watanabe, Y., Weigel, D., and Zhu, J.-K. (2008). Criteria for annotation of plant MicroRNAs. Plant Cell 20, 3186–3190.

miRBase. (2016). miRBase. Available at: http://www.mirbase.org/ (accessed 21.10.17).

Mirsky, A.E. and Ris, H. (1951). The desoxyribonucleic acid content of animal cells and its evolutionary significance. J Gen Physiol 34, 451-62.

Modali, S.D., Parekh, V.I., Kebebew, E. and Agarwal, S.K. (2015). Epigenetic regulation of the lncRNA MEG3 and its target c-MET in pancreatic neuroendocrine tumors. Mol Endocrinol 29, 224-37.

Mora, A., Sandve, G.K., Gabrielsen, O.S. and Eskeland, R. (2016). In the loop: promoter-enhancer interactions and bioinformatics. Brief Bioinform 17, 980–995.

Murakawa, Y., Yoshihara, M., Kawaji, H., Nishikawa, M., Zayed, H., Suzuki, H., Fantom Consortium, and Hayashizaki, Y. (2016). Enhanced Identification of Transcriptional Enhancers Provides Mechanistic Insights into Diseases. Trends Genet 32, 76-88.

Nakaya, H.I., Amaral, P.P., Louro, R., Lopes, A., Fachel, A.A., Moreira, Y.B., El-Jundi, T.A., da Silva, A.M., Reis, E.M., and Verjovski-Almeida, S. (2007). Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. Genome Biol 8, R43.

Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S., Sandstrom, R., Humbert, R., and Stamatoyannopoulos, J.A. (2012). BEDOPS: high-performance genomic feature operations. Bioinformatics 28, 1919-1920.

Ni, T., Corcoran, D.L., Rach, E.A., Song, S., Spana, E.P., Gao, Y., Ohler, U., and Zhu, J. (2010). A paired-end sequencing strategy to map the complex landscape of transcription initiation. Nat Methods 7, 521-7.

Niknafs, Y.S., Pandian, B., Iyer, H.K., Chinnaiyan, A.M., and Iyer, M.K. (2017). TACO produces robust multisample transcriptome assemblies from RNA-seq. Nat Methods 14, 68-70.

Noguchi, H., Park, J., and Takagi, T. (2006). MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Res 34, 5623-5630.

Ovcharenko, I., Loots, G.G., Nobrega, M.A., Hardison, R.C., Miller, W. and Stubbs, L. (2005). Evolution and functional classification of vertebrate gene deserts. Genome Res 15, 137–145.

Palazzo, A.F. and Lee, E.S. (2015). Non-coding RNA: what is functional and what is junk? Front Genet 6, 2.

Paraskevopoulou, M.D., Vlachos, I.S., and Hatzigeorgiou, A.G. (2016). DIANA-TarBase and DIANA suite tools: Studying experimentally supported microRNA targets. Curr Protoc Bioinformatics 55, 12.14.1–12.14.18.

Pasquinelli, A.E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M.I., Maller, B., Hayward, D.C., Ball, E.E., Degnan, B., Müller, P., Spring, J., Srinivasan, A., Fishman, M., Finnerty, J.,

Corbo, J., Levine, M., Leahy, P., Davidson, E., and Ruvkun, G. (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature 408, 86-9.

Peterson, S.M., Thompson, J.A., Ufkin, M.L., Sathyanarayana, P., Liaw, L., and Congdon, C.B. (2014). Common features of microRNA target prediction tools. Front Genet 18, 5–23.

PNRD, 2016. PNRD. Available at: http://structuralbiology.cau.edu.cn/PNRD/.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842.

Quinodoz, S., and Guttman, M. (2014). Long noncoding RNAs: an emerging link between gene regulation and nuclear organization. Trends Cell Biol 24, 651-663.

Ransohoff, J.D., Wei, Y. and Khavari, P.A. (2018). The functions and unique features of long intergenic non-coding RNA. Nat Rev Mol Cell Biol 19, 143-157.

Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature 403, 901-6.

Ren, C., Liu, F., Ouyang, Z., An, G., Zhao, C., Shuai, J., Cai, S., Bo, X., and Shu, W. (2017). Functional annotation of structural ncRNAs within enhancer RNAs in the human genome: implications for human disease. Sci Rep 7, 15518.

Rfam, 2017. Rfam. Available at: http://rfam.xfam.org/ (accessed 21.10.17).

Rinn, J.L. and Chang, H.Y. (2012). Genome regulation by long noncoding RNAs. Annu Rev Biochem 81, 145-66.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43, e47.

Seyednasrollah, F., Laiho, A., and Elo, L.L. (2015). Comparison of software packages for detecting differential expression in RNA-seq studies. Brief Bioinform 16, 59-70.

Sharp, P.A. and Burge, C.B. (1997). Classification of introns: U2-type or U12-type. Cell 91, 875-9.

Shearwin, K.E., Callen, B.P., and Egan, J.B. (2005). Transcriptional interference--a crash course. Trends Genet 21, 339-345.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet 38, 1348-54.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics 21, 3940-3941.

Singh, N.K. (2017). MicroRNAs databases: Developmental methodologies, structural and functional annotations. Interdiscip Sci 9, 357–377.

Sloan, C.A., Chan, E.T., Davidson, J.M., Malladi, V.S., Strattan, J.S., Hitz, B.C., Gabdank, I., Narayanan, A.K., Ho, M., Lee, B.T., Rowe, L.D., Dreszer, T.R., Roe, G., Podduturi, N.R., Tanaka, F., Hong, E.L., and Cherry, J.M. (2016). ENCODE data at the ENCODE portal. Nucleic Acids Res 44, D726-732.

Somarowthu, S., Legiewicz, M., Chillón, I., Marcia, M., Liu, F., and Pyle, A.M. (2015). HOTAIR forms an intricate and modular secondary structure. Mol Cell 58, 353–361.

Sun, Q., Hao, Q., and Prasanth, K.V. (2018). Nuclear Long Noncoding RNAs: Key Regulators of Gene Expression. Trends Genet 34, 142-157.

Taft, R.J. and Mattick, J.S. (2003). Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences. Genome Biology 5, P1

Taft, R.J., Pheasant, M. and Mattick, J.S. (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. Bioessays 29, 288-99.

Tang, F. (2010). Small RNAs in mammalian germline: Tiny for immortal. Differentiation 79, 141–146.

Tarn, W.Y. and Steitz, J.A. (1996). Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns. Science 273, 1824-32.

The Encode Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 91–100.

Tilgner, H., Knowles, D.G., Johnson, R., Davis, C.A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T.R., and Guigo, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res 22, 1616-1625.

Tiwari, V.K. and Baylin, S.B. (2009). Combined 3C-ChIP-cloning (6C) assay: a tool to unravel protein-mediated genome architecture. Cold Spring Harb Protoc 2009, pdb.prot5168.

Tran, V.T., Tempel, S., Zerath, B., Zehraoul, F., and Tahi, F. (2015). miRBoost: Boosting support vector machines for microRNA precursor classification. RNA 21, 775–785.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31, 46-53.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7, 562-578.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28, 511-515.

Tufarelli, C., Stanley, J.A., Garrick, D., Sharpe, J.A., Ayyub, H., Wood, W.G., and Higgs, D.R. (2003). Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. Nat Genet 34, 157-165.

van Steensel, B., and Belmont, A.S. (2017). Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. Cell 169, 780-791.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Miklos, G.L.G., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V.D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gu, Z., Guan, P., Heiman, T.J., Maureen E. Higgins, R.R.J., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z.Y., Wang, A., Wang, X., Jian Wang, M.H.W., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S.C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Robert Rodriguez, Y.H.R., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Vech, C., Wang, G., Wetter, J., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigó, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Parris Caulk, Y.H.C., Coyne, M., Dahlke, C., Mays, A.D., Dombroski, M., Donnelly, M., Ely, D., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W.,

McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., Zhu, X., Aderonke Awe, Sukyee Tse, Sherita Williams, and Shiva Esparham. (2001). The Sequence of the Human Genome. Science 291, 1304-51.

Visel, A., Minovitsky, S., Dubchak, I. and Pennacchio, L.A. (2007). VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic Acids Res 35, D88-92.

Vlachos, I.S., Paraskevopoulou, M.D., Karagkouni, D., Georgakilas, G., Vergoulis, T., Kanellos, I., Anastasopoulos, I.-L., Maniou, S., Karathanou, K., Kalfakakou, D., Fevgas, A., Dalamagas, T., and Hatzigeorgiou, A.G. (2015). DIANA-TarBase v7.0: Indexing more than half a million experimentally supported miRNA:mRNA interactions. Nucleic Acids Res 43, D153–D159.

Wang, D., Garcia-Bassets, I., Benner, C., Li, W., Su, X., Zhou, Y., Qiu, J., Liu, W., Kaikkonen, M.U., Ohgi, K.A., Glass, C.K., Rosenfeld, M.G., and Fu, X.D. (2011). Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature 474, 390-4.

Wang, J., Zhao, Y., Zhou, X., Hiebert, S.W., Liu, Q. and Shyr, Y. (2018). Nascent RNA sequencing analysis provides insights into enhancer-mediated gene regulation. BMC Genomics 19, 633.

Wang, X., Arai, S., Song, X., Reichart, D., Du, K., Pascual, G., Tempst, P., Rosenfeld, M.G., Glass, C.K., and Kurokawa, R. (2008). Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. Nature 454, 126-30.

Werner, M.S., and Ruthenburg, A.J. (2015). Nuclear Fractionation Reveals Thousands of Chromatin-Tethered Noncoding RNAs Adjacent to Active Genes. Cell Rep 12, 1089-1098.

Werner, M.S., Sullivan, M.A., Shah, R.N., Nadadur, R.D., Grzybowski, A.T., Galat, V., Moskowitz, I.P., and Ruthenburg, A.J. (2017). Chromatin-enriched lncRNAs can act as cell-type specific activators of proximal gene transcription. Nat Struct Mol Biol 24, 596-603.

Wuarin, J. and Schibler, U. (1994). Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. Mol Cell Biol 14, 7219-25.

Xiang, J.F., Yin, Q.F., Chen, T., Zhang, Y., Zhang, X.O., Wu, Z., Zhang, S., Wang, H.B., Ge, J., Lu, X., Yang, L., and Chen, L.L. (2014). Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus. Cell Res 24, 513-31.

Yamashita, R., Sathira, N.P., Kanai, A., Tanimoto, K., Arauchi, T., Tanaka, Y., Hashimoto, S.I., Sugano, S., Nakai, K., and Suzuki, Y. (2011). Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. Genome Res 21, 775-89.

Yang, L., Lin, C., Jin, C., Yang, J.C., Tanasa, B., Li, W., Merkurjev, D., Ohgi, K.A., Meng, D., Zhang, J., Evans, C.P., and Rosenfeld, M.G. (2013). lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs. Nature 500, 598-602.

Yang, X.H., Nadadur, R.D., Hilvering, C.R., Bianchi, V., Werner, M., Mazurek, S.R., Gadek, M., Shen, K.M., Goldman, J.A., Tyan, L., Bekeny, J., Hall, J.M., Lee, N., Perez-Cervantes, C., Burnicka-Turek, O., Poss, K.D., Weber, C.R., de Laat, W., Ruthenburg, A.J., and Moskowitz, I.P. (2017). Transcription-factor-dependent enhancer transcription defines a gene regulatory network for cardiac rhythm. Elife 6.

Yao, P., Lin, P., Gokoolparsadh, A., Assareh, A., Thang, M.W.C. and Voineagu, I. (2015). Coexpression networks identify brain region-specific enhancer RNAs in the human brain. Nat Neurosci 18, 1168-74.

Yi, X., Zhang, Z., Ling, Y., Xu, W., and Su, Z. (2014). PNRD: A plant non-coding RNA database. Nucleic Acids Res 43, D982–D989.

Zhang, B.H., Pan, X.H., Cox, S.B., and Anderson, T.A. (2006). Evidence that miRNAs are different from other RNAs. Cell Mol Life Sci 63, 246–254.

Zhang, T., Cooper, S., and Brockdorff, N. (2015). The interplay of histone modifications - writers that read. EMBO Rep 16, 1467-1481.

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K.S., Singh, U., Pant, V., Tiwari, V., Kurukuti, S., and Ohlsson, R. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet 38, 1341-7.

Zhu, L.J. (2013). Integrative analysis of ChIP-chip and ChIP-seq dataset. Methods Mol Biol 1067, 105-124.

Zhu, L.J., Gazin, C., Lawson, N.D., Pages, H., Lin, S.M., Lapointe, D.S., and Green, M.R. (2010). ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics 11, 237.

Zhu, Y., van Essen, D., and Saccani, S. (2012). Cell-type-specific control of enhancer activity by H3K9 trimethylation. Mol Cell 46, 408-423.

Zieve, G.W. (1981). Two groups of small stable RNAs. Cell 25, 296-7.