Abstract

Development of Computational Tools to Analyze New Experimental Technologies for the Study of Noncoding RNA

Michael Rutenberg Schoenberg

2019

The advent of high throughput DNA sequencing has vastly accelerated transcriptome-wide profiling of RNA, revealing thousands of new noncoding RNA genes in humans and across the phylogenetic tree. Many of these noncoding RNAs are similar in length and processing to messenger RNAs and are referred to as long noncoding RNAs (lncRNAs). Some lncRNAs had been identified decades earlier and have genetic and biochemical evidence for function, e.g. the Xist RNA, which is the master regulator of X-chromosome inactivation in female mammals. Meanwhile, the functions (or lack thereof) of many lncRNA genes are unclear, and the detailed mechanisms of lncRNAs with known functions are also often unknown.

Beyond identification of new RNA genes, high throughput sequencing has also enabled the adaptation of biochemical methods that were traditionally read out for one target RNA at a time to a transcriptome-wide scale, while sometimes revealing new types of information or making possible the study of RNAs within complex or *in vivo* samples. This enables unprecedented characterization of the activities of both noncoding RNA genes and regulatory regions within messenger RNAs, providing potentially critical information. Each new assay brings specific analysis challenges, including data normalization, scale of interpretation, statistical overdispersion, and limited numbers of replicate experiments.

In this thesis, I have developed and applied computational and statistical methods to aid the interpretation of new technologies for the study of noncoding RNA. In the

first chapter, I review the state of the field for the study of lncRNAs and general analysis challenges presented in the interpretation of high throughput sequencing data. In the second and third chapters, I describe preliminary work in my PhD analyzing two technologies developed by collaborators: Capture Hybridization of RNA Targets (CHART) to reveal the spreading pattern of the Xist RNA across the X chromosome (ch. 2); and separation of labeled RNA populations using improved disulfide chemistry for the study of RNA dynamics (ch. 3). In the fourth chapter, I develop a new analysis method to model the statistical overdispersion of RNA chemical probing data and apply this model to investigate the contribution of variability in chemical probing data on resulting RNA secondary structure predictions. The methods described here may facilitate the use of the described technologies for integrative analysis to help distinguish candidate lncRNAs and specific regions within them for further study, as well as RNA regulatory regions in which mutations may cause disease.

Development of Computational Tools to Analyze New Experimental Technologies for the Study of Noncoding RNA

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Michael Rutenberg Schoenberg

Dissertation Director: Dr. Mark Gerstein and Dr. Matthew D. Simon

May 2019

Contents

List of Figures

Acknowledgments

There are too many people and acts of kindnesses to mention, which have helped me get to the point of finishing my PhD. Here's a small attempt at making some of those thank yous.

For the research

I couldn't have imagined what doing my PhD research would be like, but it would not have been the same without the steadfast guidance of Drs. Mark Gerstein and Matt Simon. Mark presented to me a whole world of collaborations and computational methodologies, while being incredibly patient as I explored a wide variety of projects. Matt inspired me to explore the world of noncoding RNA biology and to learn in detail about experimental techniques - even to try my hand at them. Moreover, the mentorship I received from Mark and Matt truly enabled me to do this work.

I am also grateful to my thesis committee members, Dr. Ronald Breaker and Dr. Yuval Kluger. They provided help and support all along the way, amid changes in direction and really helped me get to the finish line.

To my collaborators outside Yale, in particular Dr. Jeannie Lee, Dr. Robert Kingston, and Dr. Stefan Pinter: the work we did at the beginning of my PhD was immensely exciting and I am very grateful for the openness you had to welcoming someone new to your work, just as your were trying to bring it to completion.

To collaborators within my labs: Dr. Erin Duffy was a partner in research throughout graduate school. I so admire your tenacity and insight when confronting difficult

problems, your spirit in loving your work, and your care for those around you. Dr. Alec Sexton has been a constantly thoughtful discussion partner, a kind labmate, and extrmely generous in conducting experiments to help me do some of the critical elements of this work. Dr. Rui Fang was extremely welcoming as I joined the Simon lab and helped me get off the ground in doing experiments. Dr. Jing Zhang brought great knowledge of statistical methodology to exciting questions about RNA biology. Dr. Robert Kitchen helped me with studies of miRNA and also was immensely encouraging during my early PhD work.

Drs. Joel Rozowsky, Anurag Sethi, Gamze Gursoy, Arif Harmanci, Cristina Sisu, Prashant Emani, among many others were generous with their expertise in discussions about computational work. Similarly generous were many of my graduate student colleagues, including Dr. Shantao Li, Dr. Declan Clarke, Will Meyerson, Mengting Gu, Donghoon Lee, Hussein Mohsen, Dr. Yao Fu, Dr. Jieming Chen and so many more.

Similarly generous were all the members of the Simon lab, who helped me along as I delved into the experimental world and shared their work, advice, and laughs so generously.

For the rest of life

I couldn't have made it through graduate school without the camaraderie of my colleagues. From my wonderful 2012 BBSB cohort, to my many labmates. I wish I could say in more detail how special it has been to work with you, to share the ups and downs of science, and to become friends.

To people who inspired me so much along the way: Jan Migaki, my middle school science teacher; Dr. Ed Neuwelt, in whose lab I first learned to pipette; Dr. Brian Druker, who welcomed me into his lab and allowed me to dream that research could make a difference in people's lives; Dr. Andrew Rappe, who taught me to be an independent researcher; and to so many other teachers: thank you!

To the people who make the rest of my life possible: Alden, I can't remember life without you. Natan, I still remember our first moments of friendship. Jeremy and Alan, you're my partners in crime. To the people who supported me in grad school here: Nomi, Yoni, Meir, Maddy, Mordechai, Miriam, my wonderful neighbor Ed, and so many more: thank you!

To Sarah: I couldn't have imagined how wonderful it would be to have you in my life. I'm so grateful that I do, and I can't wait to see where life takes us.

To my family: there literally wouldn't be life without you. I am so grateful to have shared laughter, tears, adventures and so much love.

Chapter 0

Overview

This thesis contains five chapters. Chapter 1 is a reproduction of a review article that I wrote, along with Drs. Alec Sexton and Matthew Simon, describing the properties of a class of RNAs, termed long noncoding RNAs (lncRNAs), that are transcribed and processed like messenger RNAs but show little to no evidence of being translated. Our review particularly discusses the established and potential roles of lncRNAs in regulating chromatin. This chapter is related to the following published article:

- Rutenberg-Schoenberg, M, Sexton, AN, Simon, MD **(2016)**. The Properties of Long Noncoding RNAs That Regulate Chromatin. *Annu Rev Genomics Hum Genet*, 17:69-94.

Chapters 2-4 describe work analyzing technologies that combine biochemical manipulation with high throughput sequencing to study noncoding properties of RNA.

In chapter 2, I reproduce a published article focusing on the occupancy pattern of the Xist lncRNA, which is a master regulator of chromsomal dosage compensation in female mammals, across the X-chromsome. I contributed analysis to this project focusing on comparing temporal spreading patterns in different cellular contexts. This chapter is related to the following published article:

- Simon, MD, Pinter, SF, Fang, R, Sarma, K, Rutenberg-Schoenberg, M, Bowman, SK, Kesner, BA, Maier, VK, Kingston, RE, Lee, JT **(2013)**. High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. *Nature*, 504, 7480:465-469.

In chapter 3, I reproduce the following published article focusing on development of improved chemistry for isolating metabolically labeled RNA.

- Duffy, EE, Rutenberg-Schoenberg, M, Stark, CD, Kitchen, RR, Gerstein, MB, Simon, MD **(2015)**. Tracking Distinct RNA Populations Using Efficient and Reversible Covalent Chemistry. *Mol. Cell*, 59, 5:858-66.

In chapter 4, I include a draft of an article titled "Modeling overdispersion of RNA chemical probing data and application to RNA secondary structure prediction." I have this project, in collaboration with Dr. Alec Sexton, Peter Y. Wang, and under the direction of my advisors, Dr. Mark Gerstein and Dr. Matthew Simon. Work in this chapter demonstrates that simplifying assumptions often made when analyzing RNA chemical probing data do not hold true formally. As a result, more flexible distributions can provide a better fit to probing data. I further explore how variability in probing data influences RNA secondary structure predictions.

In chapter 5, I provide a very brief conclusion and suggest future directions.

In addition to the included work, I have contributed as a co-author to several other articles that are published or under review:

- Fang, R, Moss, WN, Rutenberg-Schoenberg, M, Simon, MD **(2015)**. Probing Xist RNA Structure in Cells Using Targeted Structure-Seq. *PLoS Genet.*, 11, 12:e1005668.

- Sexton, AN, Wang, PY, Rutenberg-Schoenberg, M, Simon, MD **(2017)**. Interpreting Reverse Transcriptase Termination and Mutation Events for Greater Insight into the Chemical Probing of RNA. *Biochemistry*, 56, 35:4713-4721.

- Sisu, C, Pei, B, Leng, J, Frankish, A, Zhang, Y, Balasubramanian, S, Harte, R, Wang, D, Rutenberg-Schoenberg, M, Clark, W, Diekhans, M, Rozowsky, J, Hubbard, T, Harrow, J, Gerstein, MB **(2014)**. Comparative analysis of pseudogenes across three phyla. *Proc. Natl. Acad. Sci. U.S.A.*, 111, 37:13361-6.

- Zhang J, Liu J, Lee D, Feng JJ, Lochovsky L, Lou S, Rutenberg-Schoenberg M, Gerstein M **(2019)**. RADAR: annotation and prioritization of variants in the post-transcriptional regulome of RNA-binding proteins. *BioRXiv*

Chapter 1

The Properties of Long Noncoding RNAs that Regulate Chromatin

1.1 Summary

The central topic of this thesis is the computational analysis of experiments that reveal noncoding biochemical properties of RNA. The review article below discusses the history and current state of the field studying long noncoding RNAs (lncRNAs), a class of RNAs that are transcribed and processed similarly to messenger RNAs but are not translated, with a particular focus on those lncRNAs that regulate chromatin. I led writing of this article, with contributions from Drs. Alec N. Sexton and Matthew D. Simon.

The remainder of this chapter is a reproduction, with permission, of the following publication:

Rutenberg-Schoenberg, M, Sexton, AN, Simon, MD **(2016)**. The Properties of Long Noncoding RNAs That Regulate Chromatin. *Annu Rev Genomics Hum Genet*, 17:69-94.

1.2 Abstract

Beyond coding for proteins, RNA molecules have well-established functions in the posttranscriptional regulation of gene expression. Less clear are the upstream roles of RNA in regulating transcription and chromatin-based processes in the nucleus. RNA is transcribed in the nucleus, so it is logical that RNA could play diverse and broad roles that would impact human physiology. Indeed, this idea is supported by well-established examples of noncoding RNAs that affect chromatin structure and function. There has been dramatic growth in studies focused on the nuclear roles of long noncoding RNAs (lncRNAs). Although little is known about the biochemical mechanisms of these lncRNAs, there is a developing consensus regarding the challenges of defining lncRNA function and mechanism. In this review, we examine the definition, discovery, functions, and mechanisms of lncRNAs. We emphasize areas where challenges remain and where consensus among laboratories has underscored the exciting ways in which human lncRNAs may affect chromatin biology.

1.3 LESSONS FROM EARLY DISCOVERIES OF NONCODING RNAs

Non-protein-coding RNA transcripts (ncRNAs) have been known to regulate gene expression since the advent of molecular biology. The roles of the first ncRNAs discovered [e.g., transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), and spliceosomal RNAs] involved RNA translation and RNA processing (reviewed in [Cech and Steitz, 2014]). The much broader scope of ncRNA-directed activities, including regulation of chromatin, is now widely appreciated. This understanding is built in part on discoveries in which genes for ncRNAs (rather than protein-coding genes) were identified as master regulators of certain biological phenomena. As an early example, while studying

genes responsible for developmental timing in Caenorhabditis elegans, the Ambros and Ruvkun groups identified two genes, lin-4 and lin-14, as necessary for appropriate development [Lee et al., 1993, Wightman et al., 1993]. One gene (lin-14) was a protein-coding gene, as expected; the biological activity of the other gene (lin-4, which represses lin-14) was unexpectedly caused by a short RNA molecule rather than a protein. Although this discovery was initially considered an oddity of nematode biology, similar analyses led to the realization that another gene responsible for developmental timing in these worms (let-7) functions as a small ncRNA and is conserved in other animals, including humans, opening the door to the now giant field of small ncRNA biology [Neilson and Sharp, 2008]. In mammals, these small ncRNAs are thought to function mainly in the cytoplasm (with a few exceptions, such as the Piwi-interacting RNAs). By contrast, there is general consensus that a different class of ncRNAs, the long noncoding RNAs (lncRNAs), includes RNAs that play important roles in chromatin regulation in higher eukaryotes, including mammals [Goff and Rinn, 2015, Guttman and Rinn, 2012, Vance and Ponting, 2014].

For investigators interested in lncRNAs that impact chromatin structure, the flagship example is the Xist lncRNA [Gendrel and Heard, 2014, Lee, 2009]. In female mammals, transcription from one of the two female X chromosomes is repressed, thereby balancing dosage between females (two Xs) and males (one X). The discovery of this phenomenon [LYON, 1961] led to the hypothesis that the X chromosome contains a master regulator of X chromosome inactivation (XCI). The subsequent search for this regulator uncovered the XCI center, from which a large transcript ($\sim$18 kb in mice) is expressed in a sexspecific manner [Brockdorff et al., 1992, Brown et al., 1991]. The possibility that this transcript was a protein-coding messenger RNA (mRNA) was considered, but studies eventually found that the RNA itself is the master regulator of XCI. This lncRNA was named Xist. Xist is a spliced, polyadenylated, RNA polymerase II (Pol II) transcript that is upregulated on one of

the two female X chromosomes early in development. Decades of work from multiple groups have led to consensus and a detailed understanding about the role of Xist in certain steps of XCI [Gendrel and Heard, 2014, Lee, 2009]. Xist spreads to gene-rich regions on the X chromosome through a two-step mechanism [Simon et al., 2013], eventually coating the majority of the X chromosome. Xist leads to transcriptional repression and dramatic changes to the chromatin composition, including changes to histone composition (e.g., incorporation of macroH2A), histone modification [e.g., methylation of histone H3 on lysine 27 (H3K27me) [Plath et al., 2003]], and DNA methylation. Although there is broad consensus about the importance of Xist to the initiation of XCI, the downstream events that take place on the inactive X chromosome, including the biochemical events connecting Xist to chromatin regulation and how Xist spreads in cis across an entire chromosome, are still topics of active investigation. After one X chromosome is inactivated, the identity of the inactive X chromosome is mitotically stable through future cell divisions, making XCI a classic example of chromatin-mediated epigenetic regulation.

One obvious question is why seemingly few human lncRNAs have turned up in classic genetic screens or biologically driven studies (such as those that led to the discovery of lin-4 and Xist). Explanations include early investigator bias toward coding transcripts, a bias that has recently become less pronounced thanks to a wider appreciation of the roles of ncRNAs. Furthermore, traditional mutagenesis is relatively less likely to inactivate the function of a lncRNA than the function of a coding gene, because lncRNAs lack features that can lead to inactivation of protein-coding genes (nonsense, missense, and frameshift mutations). Other challenges also might have obscured lncRNAs from these studies. For example, human lncRNAs are rarely conserved in model organisms such as flies or worms, which would otherwise allow faster and more comprehensive analyses; by one estimate, the majority of human lncRNAs are not conserved beyond primates [Derrien et al., 2012]. However, there

are well-characterized examples of lncRNAs with important biological effects that were identified in model organisms (but not found in humans).

Perhaps the best-characterized examples of lncRNA function in model organisms are the roX lncRNAs in flies, which (like Xist) are central to dosage compensation. In Drosophila melanogaster, the logic of dosage compensation is different from that in mammals: Instead of inactivating one of the two female X chromosomes, these flies balance dosage primarily by increasing transcription on the single male X chromosome [Conrad and Akhtar, 2012, Gelbart and Kuroda, 2009, Maenner et al., 2012]. Genetic screens in D. melanogaster first uncovered a complex of proteins that are required for fly dosage compensation [Belote and Lucchesi, 1980, Fukunaga et al., 1975]. Subsequently, an enhancer-trap screen identified a genomic region on the X chromosome that caused sex-specific expression of the enhancer-trap reporter [Meller et al., 1997]. Investigation of this locus led to the discovery of a non-protein-coding gene dubbed RNA on the X (roX) that coats the fly X chromosome in males, in a manner reminiscent of Xist in mammals. Meller and Rattner [Meller and Rattner, 2002] later discovered that there are two redundant roX RNAs (roX1 and roX2), at least one of which is necessary for dosage compensation and male viability. It is worth noting that this redundancy may be the reason that these lncRNA genes were missed in earlier genetic screens for regulators of dosage compensation. The roX RNAs can assemble into a chromatin-modifying complex that includes at least five proteins [male-specific lethal 1–3 (MSL1–3), males absent on the first (MOF), and maleless (MLE)] that had previously been implicated in dosage compensation [Conrad and Akhtar, 2012, Gelbart and Kuroda, 2009, Maenner et al., 2012]. This complex binds to chromatin at well-defined sites and upregulates genes on the single X chromosome. This upregulation relies on the catalytic activity of the MOF subunit, which acetylates histone H4K16, a known activating modification that can lead to chromatin decondensation. Although there are similarities between the roX RNAs in flies and Xist in mammals,

there are also important differences beyond the direction of their respective regulation (activation versus repression). The roX RNAs can rescue a roX mutant phenotype when expressed in trans from another chromosome (they will still appropriately localize to the X chromosome), whereas Xist silences in cis on the same chromosome on which it is transcribed, even if it is expressed from an autosome [Kelley et al., 1999]. Also, high-resolution binding studies have demonstrated that roX RNAs bind to well-defined punctate genomic sites [Chu et al., 2011, Simon et al., 2011], whereas similar analyses demonstrate that Xist binds much more broadly on the X chromosome [Simon et al., 2013, Engreitz et al., 2013]. Nonetheless, the roX RNAs and Xist clearly demonstrate how lncRNAs can dramatically influence chromatin biology.

There is broad consensus that the examples cited above constitute important cases where lncRNAs influence chromatin. As with any relatively new field, there are active areas of investigation where there is not yet consensus. There is still debate about what constitutes a lncRNA and how many human lncRNAs exist. Furthermore, although this review focuses on lncRNAs that influence chromatin structure, in many cases it is not clear whether any given lncRNA functions at the level of chromatin. For example, one of the first human lncRNAs discovered was H19 [Brannan et al., 1990], which is expressed from a well-studied imprinted locus and has been established as a tumor suppressor and growth suppressor [Barlow and Bartolomei, 2014]. Unlike Xist and roX lncRNAs, which directly impact chromatin regulation, H19 is exported to the cytosol, where it is processed to provide a source of a microRNA (miR-675) that suppresses growth by posttranscriptional attenuation of growth-promoting genes, including Igf1r. Only a subset of lncRNAs function by directly impacting chromatin biology, and this review addresses different properties to consider when studying these lncRNAs (Figure 1.1). Although there has been an upsurge in the number of lncRNA reports, this review focuses primarily on cases where multiple investigations have reached similar conclusions; in cases where there is not yet consensus, we focus

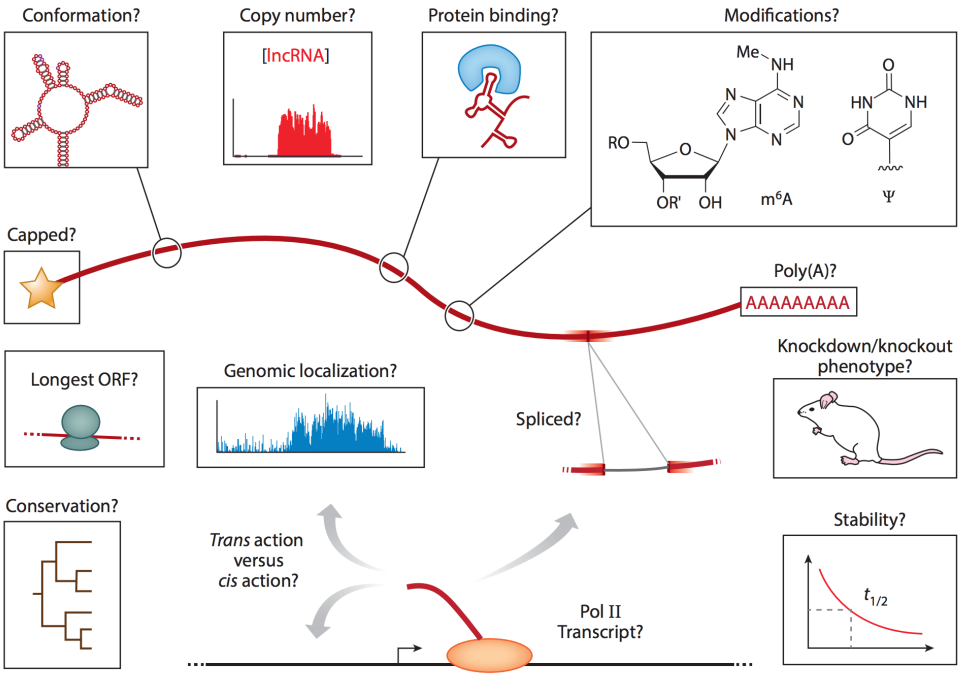

Figure 1.1: Properties of lncRNAs that act on chromatin.

Abbreviations: lncRNA, long noncoding RNA; m^{6}A,N^6-methyladenosine; Me, methylation; ORF, open reading frame; Pol II, RNA polymerase II; Ψ, pseudouridine.

on the biological and technical challenges that still need to be overcome.

1.4 WHAT MAKES AN RNA A LONG NON-CODING RNA?

lncRNAs are typically defined as RNA molecules that are at least 200 nucleotides (nt) in length and do not display potential to encode proteins [Rinn et al., 2007, Ulitsky et al., 2011] (Figure 1.1). They are generally transcribed in a regulated manner by Pol II and sometimes (but not always) processed similarly to mRNAs (e.g., they are generally capped, spliced, and polyadenylated). The lncRNA classification is problematic for three reasons. First, the current definition of lncRNA is divorced from function, in contrast to well-established families of ncRNA, such as tRNAs, rRNAs, small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), and microRNAs (miRNAs). Thus, lncRNAs are likely a collection of subclasses of RNAs with diverse functions (and in some cases without a substantial function). To address this issue, ongoing efforts are under way to tease out classification systems for lncRNAs that better reflect their biochemical and regulatory attributes [Bonasio and Shiekhattar, 2014, Cech and Steitz, 2014]. Second, the arbitrary cutoff of 200 nt may not include some RNAs that are very similar to lncRNAs. This cutoff is technically expedient (200 nt is approximately the retention of standard long RNA purification kits) [Kapranov et al., 2007] and bioinformatically expedient (we need some cutoff far from small RNAs). However, some ncRNAs that are shorter than 200 nt and/or arise from Pol I and Pol III transcription have been proposed to function in a manner similar to lncRNAs, including B1, B2, and Alu RNAs (e.g., [Mariner et al., 2008]) and promoter-associated RNAs (pRNAs) [Mayer et al., 2006, Schmitz et al., 2010], so it is reasonable to expect further refinement of this distinction.

The third problem with lncRNA classification is that the definition is negative (i.e., lncRNAs do not have substantial coding potential), but it is the positive functions of these RNA molecules (e.g., regulation of chromatin in the case of Xist and roX lncRNAs) that make them interesting. To underscore this point, some RNAs, such as the steroid receptor RNA activator 1 (SRA1) [Hube et al., 2006], have the potential to both act as a coding transcript and play a regulatory role independent of translation. Nonetheless, when trying to understand direct regulatory roles of RNAs, it is presumably easier to begin by ignoring RNAs with another obvious function (protein coding). It is difficult to confidently establish that a putative ncRNA lacks protein-coding potential. Many transcripts longer than 1,000 nt are expected to have an open reading frame (ORF—i.e., a start codon and stop codon in the same triplet reading frame) just by chance that could in principle encode a protein longer than 100 amino acids [Dinger et al., 2008, Ulitsky and Bartel, 2013]. In some cases, even much shorter ORFs can produce functional peptides (e.g., [Slavoff et al., 2013]. Several lines of evidence can help distinguish protein-coding and non-protein-coding genes. On average, ORFs in bona fide protein-coding genes display sequence conservation signals that reflect stronger selection against mutations that change the protein sequence (missense or frameshift mutations) compared with those that preserve the sequence (synonymous mutations) [Yang and Bielawski, 2000]. Furthermore, protein sequences often contain conserved structural domains with sequence similarity to parts of other proteins or have experimental support for expression in proteomics databases [Bateman et al., 2015]. Data from ribosome footprinting experiments (in which footprints of RNA protected by the ribosome are sequenced) have also contributed to our understanding of which RNAs are translated into proteins [Bazzini et al., 2014, Guttman et al., 2013]. Based on these and other metrics, a wide range of tools have been developed to help distinguish newly discovered protein-coding and lncRNA transcripts [Lin et al., 2011, Washietl et al., 2011]; this topic has

been reviewed in greater depth elsewhere [Housman and Ulitsky, 2016].

In summary, there is broad consensus that lncRNAs are transcribed and processed similarly to mRNAs and can have important functions independent of translation. Because lncRNAs are operationally defined primarily to distinguish them from other classes of RNAs, we anticipate that more positive classifications of these RNAs will become possible as we learn more about the biogenesis and function of lncRNAs.

1.5 HOW ARE NEW LONG NONCODING RNAs FOUND AND ANNOTATED?

Early examples of ncRNAs such as let-7, the roX RNAs, and Xist were found serendipitously. Now that we know such RNAs exist, it makes sense to closely examine the extensive noncoding transcription in the genome. The realization that large portions of mammalian genomes lying outside of protein-coding genes are transcribed came initially from sequencing of complementary DNA (cDNA) clones [de Hoon et al., 2015]. This work was pioneered by the Functional Annotation of the Mammalian Genome (FANTOM) consortium, which developed technology to reverse transcribe full-length transcripts and to enrich rare transcripts from pools of RNA extracted from cells, two key innovations that aided transcript discovery [Carninci et al., 2005, Okazaki et al., 2002]. Analysis of FANTOM noncoding transcripts indicated that these RNAs were slightly more conserved than average genomic regions [Ponjavic et al., 2007]. Further evidence of pervasive transcription was drawn from analysis of microarray data. Genome tiling microarrays extended beyond cDNA sequencing data to show new actively transcribed regions [Bertone et al., 2004]. Meanwhile, non-protein-coding genomic loci harboring histone modification patterns characteristic of genes were found to be robustly expressed, and analysis of this long intergenic noncoding RNA set corroborated the conservation of lncRNAs relative to the rest of the genome [Guttman et al., 2009].

This extensive noncoding transcription has been corroborated by RNA sequencing (RNA-Seq) [Cabili et al., 2011, Derrien et al., 2012, Djebali et al., 2012, Gu et al., 2012, Lister et al., 2008, Nagalakshmi et al., 2008]. There is now broad consensus that this noncoding transcription is real, but how best to annotate RNA transcripts (including lncRNAs) from these data is still an active topic of investigation.

Although RNA-Seq provides excellent information about relative RNA expression levels, annotating full-length RNA transcripts (including locations of RNA splicing sites) from short-read high-throughput sequencing data presents a challenge. Because most transcripts are much longer than sequence read lengths, it is often unclear whether distant exons are present together in a single transcript. Algorithms to address this problem [e.g., Cufflinks [Trapnell et al., 2010] and Scripture [Guttman et al., 2010]; reviewed in [Martin and Wang, 2011]] generally proceed by merging overlapping sequence reads into putative exons, connecting adjacent exons using reads mapping to splice junctions, and finally merging sets of distant exons into transcripts based on the assumption that read coverage will be uniform along a given transcript [Trapnell et al., 2010]. Using these algorithms with increased focus on RNAs with at least one splice junction, which guarantees that they contain sequences that could not arise from genomic DNA contamination, can aid annotation specificity [Cabili et al., 2011] but undoubtedly ignores some lncRNAs that are not spliced [e.g., nuclear-enriched autosomal transcript 1 (NEAT1, also known as MEN β)]. In addition to the above algorithms, new formats for short-read sequencing that maintain information about entire RNA molecules [Tilgner et al., 2015] and increases in throughput of single-molecule long-read DNA sequencing methods [e.g., sequencing in zero-mode waveguides [Eid et al., 2009] or nanopores (Oxford NanoPore) [Clarke et al., 2009]] have the potential to greatly improve lncRNA annotation because they provide a clearer picture of individual RNA transcripts. These technologies have already shown promise for discovering new transcripts and may also be useful for

RNA quantification applications [Sharon et al., 2013]. These advances will help address the important challenge of accurately annotating lncRNAs and their transcript structures. A range of projects, including Ensembl [Cunningham et al., 2015], GENCODE [Derrien et al., 2012], and RefSeq [Pruitt et al., 2014], have devoted considerable effort to annotating lncRNAs, particularly in the human genome. These annotation groups rely on cDNA sequencing, supplemented by RNA-Seq, as source information for transcripts. The GENCODE project supplements computationally predicted transcript annotations with manual expert curation based on cDNA, RNA-Seq, and other genome signals [e.g., chromatin immunoprecipitation sequencing (ChIP-Seq) of histone modifications] and validates transcripts with weak evidence by quantitative polymerase chain reaction (qPCR) [Derrien et al., 2012]. Because these large annotation projects are relatively conservative in their annotations of lncRNAs, additional efforts [e.g., LNCipedia [Volders et al., 2015] and NONCODE [Xie et al., 2014]] have sought to incorporate as much available RNA-Seq data as possible. In addition to annotations based on sequence and expression, at least one database (lncRNAdb) has been developed to catalog functional studies of lncRNAs [Quek et al., 2015].

1.6 HOW MANY LONG NONCODING RNAs ARE THERE?

There is not yet a consensus estimate of the number of bona fide lncRNAs expressed in the human genome. Numbers from the reference annotations discussed above range from tens of thousands to hundreds of thousands of lncRNA genes, with larger numbers of unique transcripts produced by alternative splicing [Derrien et al., 2012, Xie et al., 2014] (Figure 2a). Part of the complexity arises because current sequencing data do not accurately represent expression from all types of tissues, conditions, and developmental timing. Given the well-documented tissue specificity of lncR-

NAs [Cabili et al., 2011], it thus remains unclear how many lncRNA transcripts are substantially expressed only under specific conditions that are not well represented in current data. Perhaps an even larger problem when counting lncRNAs is determining how much expression should be required to constitute reliable evidence of a lncRNA. A practical answer to this question depends on the mechanisms by which these lncRNAs function. It is essential to consider what level of expression would be required to support a proposed mechanism of action. For example, if a lncRNA needs to bind another biomolecule with high stoichiometry in order to function, then the lncRNA must achieve cellular concentrations compatible with the interactor's copy number and dissociation constant. At the extreme, if a lncRNA acts as a high-affinity tether to recruit chromatin modifications to its endogenous genomic locus, one lncRNA molecule transiently expressed could conceivably be sufficient, especially if the recruited chromatin marks are epigenetically stable and able to propagate along the chromatin. On the other hand, for mechanisms in which the lncRNA behaves like a transcription factor or subunit of a chromatin-modifying complex, we would expect the lncRNA to exceed 100 copies per cell to achieve nanomolar concentrations in the nucleus [most transcription factors exceed 1,000 copies per cell [Biggin, 2011]]. Thus, there is consensus that there are at least thousands of lncRNAs, and useful estimates of the number of lncRNAs will continue to improve as we learn more about the mechanisms of the subset of lncRNAs that play functional roles.

What is the distribution of cellular lncRNA expression levels, and how many achieve concentrations similar to those of other types of chromatin-modifying machinery? It is well documented that lncRNAs on average have substantially lower expression than mRNAs (Figure 1.2c). In keeping with the discussion above, expression levels can be thought of in terms of copies per cell. Estimates for lncRNA copy numbers range from less than one copy per cell for HOTTIP [Wang et al., 2011] to tens of thousands for NEAT1 and metastasis-associated lung adenocarcinoma tran-

script 1 (MALAT1), which are some of the most abundant transcripts in the cell, on a similar level to abundant mRNAs such as those that encode ribosomal protein RPLP2 [Cabili et al., 2015]. Despite these estimates, RNA copy numbers have proven difficult to measure directly in high throughput, even when simplifying by omitting consideration of different possible transcript isoforms (e.g., from alternative splicing). The traditional gold standard for determining the concentration of an RNA in a population of cells is a quantitative northern blot experiment, in which the cellular RNA levels are compared with authentic standards. Other approaches include qPCR (in comparison with standards) and measurement of absolute copies per cell using single-molecule fluorescence in situ hybridization (smFISH) [Crosetto et al., 2015, Raj et al., 2008], which also provides information about the distribution of copy numbers in a cell population. By contrast, most high-throughput data on RNA expression come from RNASeq experiments (most frequently without internal standards), but using these RNA-Seq data to estimate copy number is challenging. Because some high-throughput smFISH studies have been conducted, it would be useful to relate those measurements of relative molecular concentration to RNA-Seq in a standardized way to the number of copies of an RNA molecule per cell. Indeed, Battich et al. [Battich et al., 2013] reported a high correlation between their RNA-Seq and smFISH measurements of nearly 1,000 RNAs (Pearson's $r = 0.45$, real scale; r = 0.84, $\log_2$ scale) (Figure 1.2c). However, making a more general statement about the relationship between RNA-Seq expression measurements and actual copies of RNAs per cell requires careful consideration [Pachter, 2014]. One important piece of information needed to connect measurements of relative molecular concentration and molecules per cell is how variable quantities of total RNA are among cells of a given type and those of different types: Human cells can vary in RNA content by as much as an order of magnitude [Marinov et al., 2014]. Also relevant are technical aspects of both RNA-Seq and smFISH quantification. RNA-Seq expres-

sion implicitly measures relative molecular concentrations, which depend on which RNAs are being quantified and thus on the chosen RNA annotation. Furthermore, short sequencing reads that map ambiguously to the genome must be assigned, usually by expectation-maximization algorithms, which are helpful but cannot overcome imperfect data [Li and Dewey, 2011, Trapnell et al., 2010]. Beyond this, RNA-Seq quantifications are often reported in reads or RNA fragments per kilobase per million reads (RPKM or FPKM, respectively) [Mortazavi et al., 2008]. These metrics accurately represent relative molecular concentration, but mean RPKMs vary among experiments depending on the relationship between length and transcript abundance. More recently, a metric called transcripts per million (TPM) has been proposed as a useful alternative to RPKM/FPKM that more consistently normalizes for transcript length and has a consistent mean between different experiments, so long as the same annotation is used ([Li and Dewey, 2011]; for a more detailed explanation, see Reference [Wagner et al., 2012] and the section 1.6.1, Metrics for Quantifying Transcripts in RNA-Seq).

As with RNA-Seq, there are technical issues to consider with smFISH quantifications. For example, Cabili and colleagues [Cabili et al., 2015, Dunagin et al., 2015] recently conducted an smFISH study of the expression of over 50 lncRNAs in three different cell lines. In contrast to mRNAs, which typically display diffuse expression, many of the lncRNAs chosen for analysis clustered in particular loci in the cell. As the unit of smFISH expression quantification is typically spots per cell, these quantifications may be underestimated for some lncRNAs, or for any RNA that has a three-dimensionally clustered expression pattern. Despite these challenges, can we use these data to estimate the number of lncRNAs in a cell? To explore how much consensus there is in lncRNA numbers when comparing across techniques, we re-quantified both coding and noncoding transcripts from the RNA-Seq data from HeLa cells and compared 791 mRNAs with the reported numbers of molecules per cell from

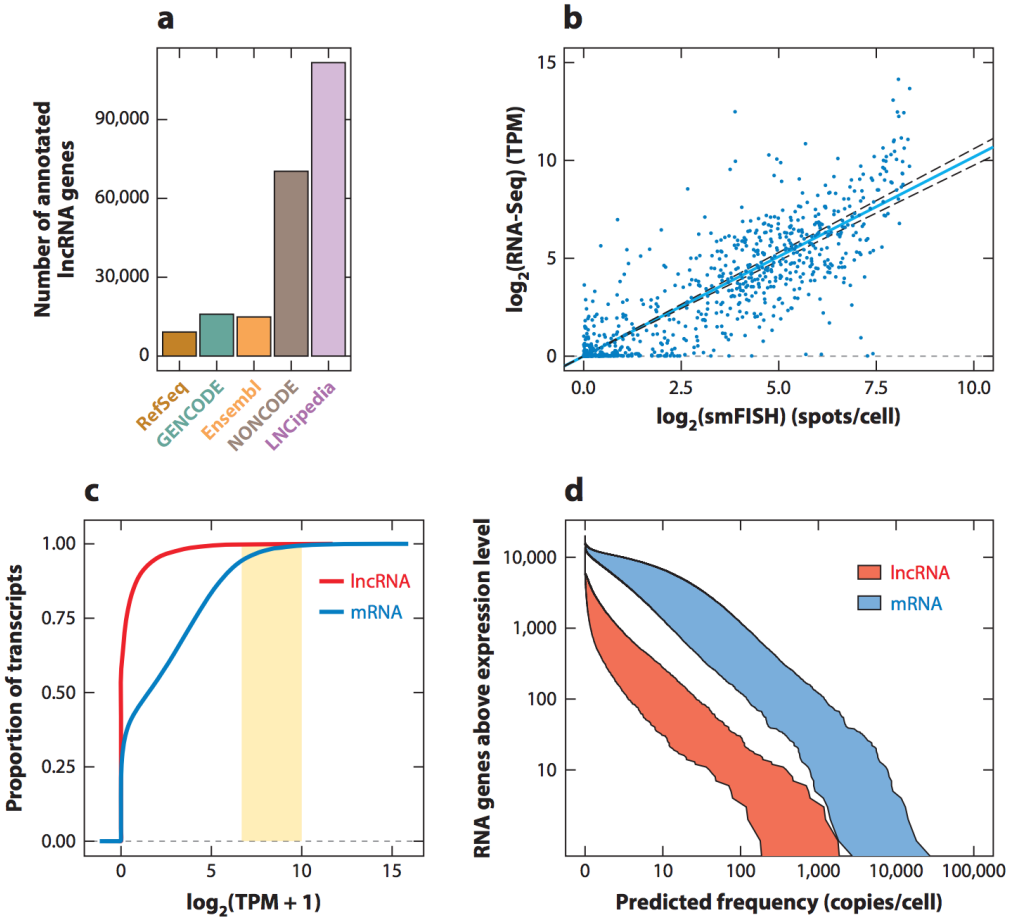

Figure 1.2: Quantifying lncRNA expression with RNA-Seq and smFISH.

(a) Estimates of the number of lncRNA genes by five annotation consortia as of November 2015. (b) Comparison of RNA-Seq and smFISH measurements of the expression of 791 mRNA genes in HeLa cells. The solid line indicates a fit by linear regression, and the dashed lines indicate its 95% confidence interval. Quantifications of RNAs annotated by the GENCODE project (version 19) were performed from polyadenylated RNA-Seq data from the ENCODE project [Djebali et al., 2012] using Kallisto [Bray et al., 2016]. (c) Cumulative distributions of expression levels measured using RNA-Seq for lncRNAs or mRNAs from the GENCODE project. (d) Predicted numbers of lncRNA and mRNA transcripts at varying absolute expression levels. Ranges of predicted copies are indicated by the shaded regions for lncRNAs (red) and mRNAs (blue). To make rough estimates of absolute RNA copy numbers, we used several regression methods to compare the relative quantifications of RNA-Seq data to absolute quantifications from smFISH data for the 791 mRNA genes shown in panel b. Abbreviations: lncRNA, long noncoding RNA; mRNA, messenger RNA; RNA-Seq, RNA sequencing; smFISH, single-molecule fluorescence in situ hybridization; TPM, transcripts per million.

smFISH experiments [Battich et al., 2013] (Figure 1.2 b). These data can provide a rough estimate that 1 molecule per cell corresponds to roughly 1–10 TPM in HeLa cells. This estimate implies that a handful to tens of lncRNAs are present at 100 or more copies per cell in HeLa cells, and that hundreds to approximately 1,000 lncRNAs are expressed with at least 1 copy on average per cell (Figure 1.2 c,d). These estimates are rough and may be relatively low, because branched DNA FISH, the smFISH technique used by Battich et al. [Battich et al., 2013], primarily detects RNA in the cytoplasm. Although most mRNAs exist predominantly in the cytoplasm, inclusion of nuclear RNA signal would increase the estimated number of RNA molecules represented by 1 TPM in RNA-Seq data.

It is also likely that the relationship between RNA copy numbers and measurements of relative concentrations from RNA-Seq vary more widely than the above estimates when considering a wider range of cell types with varying RNA content. For example, single-cell RNA-Seq of GM12878 lymphoblastoid cells using synthetic standards led to the estimate that there are 50,000 to 300,000 transcripts per cell, implying that 1 molecule per cell corresponds to ∼3–20 TPM, which in turn implies that only 0–4 lncRNAs are expressed at 100 copies per cell in this cell line [Djebali et al., 2012]. By contrast, a recent smFISH survey of lncRNAs by Cabili and colleagues [Cabili et al., 2015] found that 2–5 of ∼50 lncRNAs measured in three different cell lines (some of which were chosen for their known functions) were expressed at 100 or more copies per cell, possibly implying that more lncRNAs than we suggest are expressed at that level. Either way, these rough estimates suggest that relatively few lncRNAs achieve cellular concentrations comparable with those of transcription factors.

In summary, the human genome is thought to contain thousands to hundreds of thousands of lncRNAs, and annotation efforts are ongoing. However, careful consideration of RNA expression levels, especially considered intuitively as estimates

of molecules per cell or molecular concentrations, may help winnow this large list of genomic elements to smaller numbers that may be amenable to investigations of potential mechanistic roles and in particular cell lines.

1.6.1 METRICS FOR QUANTIFYING TRANSCRIPTS IN RNA-SEQ

RNA-Seq methods aim to measure the relative concentrations of the RNA molecules expressed in a cell. Here, we describe two metrics for quantifying this expression, reads per kilobase per million reads (RPKM) and transcripts per million (TPM); highlight factors that affect the values of both metrics; and discuss why TPM may be an improvement over RPKM. Definitions. Let t represent each transcript in the genome, which has N annotated transcripts. Also let r_t represent the reads mapping to that transcript, typically after a reassignment of ambiguously mapping reads by expectation maximization [Li and Dewey, 2011, Trapnell et al., 2010]; l_r represent the length of each read; l_t represent the length of the transcript; and R represent the total number of mapped reads in the experiment. Reads per kilobase per million reads (RPKM). RPKM is the most common metric used to describe the relative molecular concentrations of RNAs from an RNA-Seq experiment:

$$RPKM_t = r_t * \frac{10^3}{l_r} * \frac{10^6}{R}$$

Indeed, this metric correlates to relative molecular concentration (see, e.g., [Wagner et al., 2012]). However, the mean value of RPKM is not identical from experiment to experiment:

$$\langle RPKM \rangle = \frac{\sum_{t=1}^{N} RPKM_t}{N} = \frac{10^9}{R * N} \sum_{t=1}^{N} \frac{r_t}{l_t}$$

This is because the summation term $\sum_{t=1}^{N} \frac{r_t}{l_t}$ varies based on the relationship be-

tween the length and abundance of different transcripts.

Transcripts per million (TPM). TPM is a more recently proposed metric that also represents relative molecular concentration but has a consistent mean, which aids comparisons between experiments [Li and Dewey, 2011, Wagner et al., 2012]. Like RPKM, its values are dependent on which (and how many) transcripts are annotated. The underlying assumption of the TPM metric is that each read samples the proportion of the transcript that is the ratio of the read length to the length of the transcript. The number of transcripts sampled, T_t, is given by

$$T_t = \frac{r_t * l_r}{l_t}$$

TPM_t is then the ratio of the number of transcripts of the specific transcript sample to the total number of transcripts observed, T_{tot}, multiplied by 1 million:

$$TPMt = \frac{T_t * 10^6}{T_{tot}}$$

where

$$T_{tot} = \sum_{t=1}^{N} T_t$$

This gives a mean value for TPM that depends only on N, the number of total annotated transcripts encoded from the genome:

$$\langle TPM \rangle = \frac{\sum_{t=1}^{N} TPM_t}{N} = \frac{\sum_{t=1}^{N} T_t * 10^6}{T_{tot} * N} = \frac{10^6}{N}$$

1.7 WHERE ARE LONG NONCODING RNAs EXPRESSED?

One important clue to the function of a lncRNA is where it is present. Accordingly, much attention has been paid to the tissue specificity and cellular localization patterns of lncRNAs. The realization that lncRNAs are expressed more tissue specifically than mRNAs [Cabili et al., 2011, Derrien et al., 2012, Mele et al., 2015] has motivated several studies searching for roles of lncRNAs in regulating cell differentiation (e.g., [Guttman et al., 2011], [Sun et al., 2013]), although it has been aptly argued that tissue specificity is not in itself evidence for function [Ulitsky and Bartel, 2013]. Moreover, lncRNAs are widely reported to be enriched relative to mRNAs in the nucleus and on chromatin. Whether the majority of lncRNAs actually reside in the nucleus is hard to determine from RNA-Seq experiments, as this requires information about the relative quantities of RNA in different subcellular fractions. The recent smFISH survey of lncRNA expression indicated that two-thirds of the ∼50 lncRNAs studied were localized primarily to the nucleus [Cabili et al., 2015], although the selected set of RNAs chosen in this study was not intended to be representative. In addition to providing information about the localization of known RNAs, RNA-Seq experiments of chromatin fractions have been fertile ground for discovering new noncoding transcripts, which may be detected clearly only in a subcellular fraction where they are enriched [Tilgner et al., 2012, Werner and Ruthenburg, 2015]. Deeper analysis of lncRNA localization by FISH and RNA-Seq of chromatin fractions may help uncover more lncRNAs that are candidates for roles in chromatin biology. For example, the localization of roX lncRNAs on chromatin (which is particularly noticeable in salivary gland cells that produce polytene chromosomes) is the key observation that led to the hypothesis that roX lncRNAs act at the level of chromatin [Meller and Rattner, 2002].

1.8 WHAT ARE THE REGULATORY ROLES OF CHROMATIN-ACTING LONG NONCODING RNAs?

Efforts to systematically uncover the function of lncRNAs have been conducted since 2005, when Willingham et al. [Willingham et al., 2005] conducted a small hairpin RNA (shRNA) screen against ncRNAs to identify genes that modulate the activity of nuclear factor of activated T cells (NFAT) and identified the noncoding repressor of NFAT (NRON) lncRNA. Despite a decade of efforts, searching for function among the thousands of lncRNAs transcribed from mammalian genomes remains challenging. A broad screen of lncRNA knockdown in embryonic stem cells led one group of researchers to conclude that more than 90% of lncRNAs act together with chromatin-modifying machinery and have a dramatic impact on gene expression in trans, suggesting that their importance is comparable to that of transcription factors [Guttman et al., 2011]. On the other extreme, loss-of-function studies in mice have often demonstrated relatively modest roles for lncRNAs, leading to a growing consensus for a need for extra scrutiny when examining loss-of-function studies of lncRNAs [Bassett et al., 2014, Goff and Rinn, 2015]. Indeed, on occasion, reanalyses of the same data have led different groups to draw different conclusions about the potential mechanisms of regulatory effects observed upon lncRNA knockdown [Guttman et al., 2011, Tan et al., 2015]. Fetal-lethal noncoding developmental regulatory RNA (Fendrr) and megamind are two examples of lncRNAs that have been examined by multiple groups, and they highlight both the potential and challenges of studying lncRNA function.

Fendrr is transcribed antisense to the Foxf1 transcription factor gene [Grote et al., 2013], and several groups have examined its function. Insertion of a transcription termination signal in the first exon of Fendrr hindered heart and body wall development and

led to embryonic lethality in mice [Grote et al., 2013]. Introducing another element containing the Fendrr-Foxf1 locus, but with a transcription termination site in the Foxf1 gene, rescued the developmental phenotype of Fendrr transcription termination, supporting the conclusion that the original phenotype was due to loss of Fendrr lncRNA. This rescue also demonstrated that the Fendrr lncRNA can function outside its endogenous genomic locus. Others corroborated that the Fendrr locus is important to embryonic development [Sauvageau et al., 2013]. A deletion of the Fendrr locus led to embryonic lethality but caused defects in lung development as opposed to heart development, which highlights the fact that removal of subtly different combinations of ncRNA and regulatory DNA may have substantially different phenotypic consequences [Grote et al., 2013, Sauvageau et al., 2013].

The megamind lncRNA was first examined because of its conserved synteny between humans and zebrafish (in both organisms, it is expressed antisense from within the intron of *birc-6*). A hidden Markov model–based sequence search, which is much more sensitive than tools like BLAST, showed that 19 noncontiguous nucleotides of megamind were perfectly conserved among 75 copies of the RNA in 47 vertebrate species. These isolated sites of perfect conservation contrasted with low general sequence conservation in the RNA [Ulitsky et al., 2011]. Three diverse morpholino oligonucleotides that targeted megamind (or, more specifically, targeted a conserved region or one of two splice sites in the RNA) led to deformed head and brain development, consistent with the brain-specific expression of megamind RNA in both zebrafish and humans. Careful controls in this experiment included a morpholino oligonucleotide with mismatches (no phenotype was observed) and rescue of the morpholino phenotype by coinjection of zebrafish megamind RNA (or of human and mouse orthologs). Further support for this conclusion came from an independent group that identified this locus (they dubbed this RNA TUNA) based on an shRNA screen of more than 1,000 lncRNAs to identify RNAs that disrupt the pluripotency

of mouse embryonic stem cells [Lin et al., 2014]. Corroborating the findings with megamind, they found that TUNA/megamind expression is restricted to neural tissue and that knockdown of TUNA impedes differentiation of both mouse and human embryonic stem cells into neural tissue. Morpholino oligonucleotides against TUNA impeded the locomotor response of zebrafish larvae. Even with this high degree of validation across independent research groups, these results have been challenged by a study demonstrating that morpholino oligonucleotides directed against megamind lead to the same phenotype even in a megamind knockout zebrafish [Kok et al., 2015], implying that the morpholino phenotype was due to off-target effects. Additional complexity in this case arises because some organisms have multiple copies of the megamind RNA, and the two morpholino oligonucleotides tested had partial complementarity to the other RNA (76% and 68%) [Kok et al., 2015]. Regardless of how these discrepancies are eventually resolved, this example highlights both innovative approaches to uncover a conserved, functional lncRNA—megamind is one of the best-validated lncRNAs uncovered in recent years—and the challenges of building consensus about the function of newly discovered lncRNAs.

Results such as the examples above have led to an appreciation of the need for complementary approaches to uncover lncRNA function. The possibility that some lncRNAs may have few if any functions was raised when a recent analysis of another lncRNA implicated in neuronal development, Visc-2, did not provide a phenotype upon genetic ablation despite high conservation and tissue-specific expression [Oliver et al., 2015], similar to other examples reviewed previously [Nakagawa, 2016]. However, it is worth noting that the absence of a loss-of-function phenotype may be the result of compensation or redundancy (e.g., individual knockouts of roX1 and roX2 have little if any phenotype). Even when a phenotype is observed, it is worth considering that the genetic deletions of an entire lncRNA can also lead to deletion of regulatory DNA elements [Bassett et al., 2014, Goff and Rinn, 2015]. These DNA

elements can have critical impacts on gene expression, especially transcription from nearby loci. Knockdown by RNA interference or morpholinos targets RNA directly, but the specificity is not always guaranteed, and it can be difficult to predict off-target effects (see [Kok et al., 2015] and [Rossi et al., 2015] and a critical discussion in [Blum et al., 2015]).

Bassett et al. [Bassett et al., 2014] have proposed that complementation is the ideal method of demonstrating lncRNA function when using reverse genetic approaches. Inserting a lncRNA-encoding sequence at a separate genomic locus separates potential DNA regulatory elements from their targets and tests the idea that the RNA of interest alone causes the observed phenotype. Similarly, injecting or transfecting RNA into cells in which the RNA is knocked down is a promising way to show that the RNA is responsible for the knockdown phenotype [Arab et al., 2014, Ulitsky et al., 2011], although even this control has proven unreliable in some cases [Kok et al., 2015].

Because of these challenges, surgical manipulation of subelements within lncRNA loci, enabled by recent advances in using CRISPR/Cas9 and other gene editing systems [Boettcher and McManus, 2015, Hsu et al., 2014, Wright et al., 2016], will likely play an essential role in discovering new functional lncRNAs and in testing whether observed phenotypes are caused by specific RNA sequences or transcription. At the same time, ever-increasing genetic and functional genomic data will help improve hypotheses about which lncRNAs are worth prioritizing for experimental investigation. For example, ChIP-Seq data can guide study of lncRNAs that are regulated by pluripotency factors and whose knockdown induces differentiation away from the pluripotent state as well as lncRNAs that help mediate the p53 DNA damage response [Huarte et al., 2010, Sheik Mohamed et al., 2010]. Many lncRNAs have been investigated because of their association with various diseases [e.g., prostate cancer [Prensner et al., 2011, Walsh et al., 2014]; reviewed in [Lee and Bartolomei, 2013]].

The explosion of whole genome sequencing through projects such as the Cancer Genome Atlas promises to further guide investigation toward lncRNAs whose expression or sequence changes in disease states [Weinstein et al., 2013], as may extensive investigation of protein binding of RNAs through the latest phase of the ENCODE project [Dunham et al., 2012]. We anticipate that improved genetic tools and mining of genetic data will speed discovery of functional lncRNAs and help focus functional investigations of new lncRNAs.

1.9 WHAT ARE THE BIOCHEMICAL ACTIVITIES OF LONG NONCODING RNAs?

Even among lncRNAs that act on chromatin, there are diverse possibilities for where lncRNAs act, how they are associated with chromatin, and how they function biochemically. Broader discussions of ncRNA activities, including models in which they act as sinks for proteins [Yin et al., 2012] or miRNAs [Poliseno et al., 2010, Tan et al., 2015], have been published elsewhere [Tay et al., 2014]. Two extreme models of the activities of lncRNAs on chromatin are (a) that lncRNAs act locally at the sites where they are transcribed and (b) that lncRNAs are trans-acting factors (more akin to transcription factors) that can regulate well-defined sites independent of where they are expressed. As the examples below make clear, many models of lncRNA function fall between these two extremes (see also [Dimitrova et al., 2014, Huarte et al., 2010]. A landmark in the lncRNA field was the discovery of an RNA in the human HOXC locus that was expressed at the boundary between the genes that are expressed and not expressed [Rinn et al., 2007]. This lncRNA, dubbed HOX transcript antisense RNA (HOTAIR), was initially hypothesized to act locally to mark the boundary between the expressed and nonexpressed genes. Surprisingly, small interfering RNA (siRNA)–induced knockdown of HOTAIR did not

lead to changes in cis on the HOXClocus, but instead led to expression changes in the HOXD locus, which is on a different chromosome. As HOX loci are regulated by Polycomb group (PcG) machinery, an analogy was drawn to the Xist lncRNA [which eventually results in Polycomb repressive complex 2 (PRC2) recruitment to the inactive X chromosome], leading to the hypothesis that HOTAIR interacts with PRC2. This hypothesis was supported by RNA immunoprecipitation experiments demonstrating enrichment of HOTAIR upon immunoprecipitation of members of the PRC2 complex. This finding, along with the ensuing exploration of lncRNAs hypothesized to influence PRC2, has been the subject of intense debate [Brockdorff, 2013, Davidovich and Cech, 2015, Davidovich et al., 2013, Davidovich et al., 2015, Kaneko et al., 2014b, Zhao et al., 2010]. The targeting of PRC2 by lncRNAs is an attractive hypothesis both because this model would help resolve the mystery of how PRC2 is recruited to specific genomic loci in mammals [Margueron and Reinberg, 2011, Simon and Kingston, 2013] and because PRC2 was already implicated in XCI [Plath et al., 2003]. PRC2 recruitment and H3K27me3 modification are well-established hallmarks of the inactive X chromosome, as was originally shown using immunofluorescence experiments in mice [Plath et al., 2003] and humans [Chadwick and Willard, 2004], leading to examination of the hypothesis that Xist directs PRC2 activity [Zhao et al., 2008]. The conclusion that Xist directs PRC2 on the inactive X chromosome has been bolstered by allelespecific high-resolution mapping experiments comparing PRC2 localization and Xist lncRNA localization in differentiating mouse embryonic stem cells. This work demonstrated that PRC2 and Xist colocalize on gene-rich regions of the inactive X chromosome during XCI. The broad localization of PRC2 on the inactive X chromosome during XCI is qualitatively different from that observed for PRC2 enrichment elsewhere in the genome, supporting the notion that, in this case, PRC2 localization is directed by Xist. Understanding how this direction is accomplished is an active area of inves-

tigation. Current models include both direct and indirect recruitment of PRC2 by Xist (critically analyzed in [Brockdorff, 2013, Davidovich and Cech, 2015]).

The idea that lncRNAs can collaborate with chromatin-modifying complexes to regulate chromatin at numerous genomic loci has an early precedent from the roX lncRNAs. Determining the genomic localization of lncRNAs was originally accomplished by in situ hybridization [Gall and Pardue, 1969] or through biochemical hybridization capture approaches [Mariner et al., 2008] analogous to ChIP. To achieve higher resolution, the hybridization capture approaches were optimized for use in genome-wide sequencing, leading to approaches such as capture hybridization analysis of RNA targets sequencing (CHART-Seq) [Simon et al., 2011], chromatin isolation by RNA purification sequencing (ChIRP-Seq) [Chu et al., 2011], and RNA antisense purification sequencing (RAP-Seq) [Engreitz et al., 2013] that provide high-resolution analysis of genome-wide lncRNA localization [Simon, 2016]. These approaches are particularly useful when a lncRNA is hypothesized to act at genomic sites distant from its site of transcription (e.g., [Chalei et al., 2014, Hacisuleyman et al., 2014, Brown et al., 2012, Vance et al., 2014]). The original differences between these approaches include the choice of cross-linker (formaldehyde for CHART, glutaraldehyde for ChIRP, and disuccinimidyl glutarate supplemented with formaldehyde for RAP) and the choice of biotinylated capture oligonucleotides used (a few curated short DNAs for CHART, two cocktails of short DNAs that tile the RNA for ChIRP, and tiling RNAs for RAP). Since the initial reports of these techniques, there have been further modifications and some convergence between techniques [Simon, 2016]. Although there are many caveats to interpreting the results from these studies, roX2 [Chu et al., 2011, Simon et al., 2011] and Xist [Engreitz et al., 2013, Simon et al., 2013] lncRNA localization has been validated by independent approaches (Figure 1.3). From these studies, the common mechanisms for the collaboration between lncRNAs and chromatin-modifying machinery are still being developed.Many possibilities have

been considered [Bonasio and Shiekhattar, 2014, Goff and Rinn, 2015, Guttman and Rinn, 2012, Ulitsky and Bartel, 2013], including the idea that lncRNAs can act through base-paring or triplex formation with local DNA or RNA [Arab et al., 2014, Buske et al., 2012, Martianov et al., 2007, Postepska-Igielska et al., 2015, Schmitz et al., 2010]. Alternatively, as in the case of the roX lncRNAs, the specificity could be directed by proteins in the complex [Soruco et al., 2013]; even in this case, however, it is unclear what causes the target sites on the X chromosome to be favored over similar sites on autosomes.

In addition to the possibility that lncRNAs act in trans together with chromatin-modifying machinery at distant sites in the genome, lncRNAs can also act locally near their sites of transcription. It has been well established that some lncRNAs (e.g., NEAT1) assemble with protein complexes at their sites of transcription [Mao et al., 2011]. An attractive hypothesis is that lncRNAs can mark their transcription locus simply by remaining as a nascent transcript (i.e., tethered through the RNA polymerase) or by binding tightly and never dissociating. One early example of a lncRNA that was thought to employ this mechanism is Airn, a 118-kb lncRNA transcript that suppresses transcription of the antisense gene Igf2r [Wutz et al., 1997], whose promoter is 30 kb downstream of Airn. It was hypothesized that, like many lncRNAs in chromatin, Airn recruits chromatin-modifying complexes, and this hypothesis was supported by RNA immunoprecipitation assays [Nagano et al., 2008]. The functional significance of the Airn lncRNA as a mark has been called into question because the RNA itself was not found to be important for silencing of Igf2r—rather, the act of antisense transcription through the Igf2r promoter drives suppression [Latos et al., 2012]. The importance of the act of transcription rather than the RNA was determined by moving the Airn promoter and using a series of premature poly(A) termination sites [Latos et al., 2012, Stricker et al., 2008].

A similar example is the 60-kb imprinted lncRNA Kcnq1ot1, which is paternally

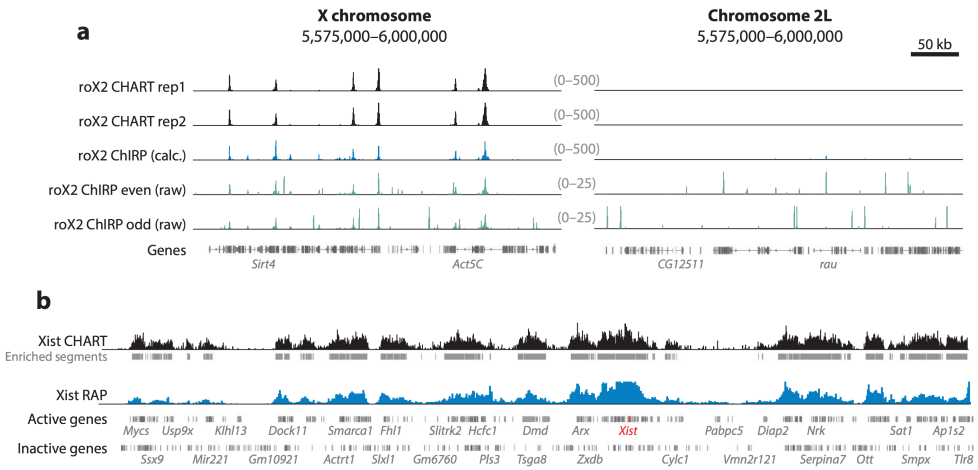

Figure 1.3: Cases of agreement between different hybridization capture approaches that reveal lncRNA genomic localization.

(a) Comparison of roX2 CHART [Simon et al., 2011] and roX2 ChIRP [Chu et al., 2011] tracks from S2 cells. The roX2 lncRNA binds at discrete sites along the X chromosome (example on left) but not autosomes (example on right) in the fly genome, as seen in both roX2 CHART and ChIRP. The raw tracks were made by reprocessing data from these two datasets using the same pipeline. Briefly, reads were aligned to the fly genome (dm6), normalized to the respective input using the spp package [Kharchenko et al., 2008], and normalized to total reads on chromosome 2L. The final ChIRP signal (roX2 ChIRP calc., downloaded from Reference 29 and converted from dm3 to dm6) was calculated from the common regions of enrichment between two independent capture oligonucleotide cocktails (even and odd). The raw signal for these distinct biochemical experiments, processed identically to the CHART signal, is shown below the combined calculated signal. Track scales are shown using a mean windowing function and are presented on the same scale for the X chromosome (left) and chromosome 2L (right). The raw ChIRP signals are shown with a lower scale to display roX2 peaks at a similar scale as the calculated track. (b) Comparison of Xist CHART [Simon et al., 2013] and Xist RAP [Engreitz et al., 2013] from differentiating female mouse embryonic stem cells, showing broad agreement of Xist localization along the X chromosome (entire chromosome shown). Tracks were reprocessed from Engreitz et al. (from 6-h retinoic-acid-induced differentiation; [Engreitz et al., 2013]) and Simon et al. (day 7 of leukemia inhibitory factor withdrawal) using the pipeline described above (but aligned to the mouse genome, mm9) [Simon et al., 2013]. Both tracks are consistent with the initial targeting of Xist to gene-rich regions on the X chromosome. Abbreviations: CHART, capture hybridization analysis of RNA targets; ChIRP, chromatin isolation by RNA purification; lncRNA, long noncoding RNA; RAP, RNA antisense purification.

expressed antisense to the potassium channel–coding gene Kcnq1 and was originally identified by associating mutations in the locus with Beckwith-Wiedemann syndrome in humans [Kanduri et al., 2006, Lee et al., 1999, Mitsuya et al., 1999, Smilinich et al., 1999]. Partly on the basis of RNA immunoprecipitation experiments, Kcnq1ot1 has been reported to interact with a variety of repressive chromatin-modifying machinery, including components of PRC2, G9a, and DNA (cytosine-5)-methyltransferase 1 (DNMT1) [Fitzpatrick et al., 2002, Pandey et al., 2008, Pandey et al., 2008]. Similar to earlier Airn studies, insertion of an early polyadenylation signal results in loss of silencing of its antisense target, Kcnq1. Additionally, although the 5′ region of Kcnq1ot1 contains a putative structured region with multiple hairpins, deleting this specific region does not result in derepression of its antisense target [Mancini-Dinardo et al., 2006]. Although it is unknown whether, as with Airn, simple overlap of the Kcnq1ot1 transcribed DNA with the transcription start site of its antisense target is sufficient to recapitulate silencing, studies have shown that transcription start site overlap and the length of the Kcnq1ot1 transcript are correlated with deposition of heterochromatin marks [Kanduri et al., 2006].

The examples above underscore the need to critically evaluate the assumption that the sequence and composition of lncRNA molecules, rather than the act of their transcription or overlapping cis-acting DNA elements, are generally important for regulation. Nonetheless, it is remarkable that most imprinted loci that have been discovered are also associated with the expression of lncRNAs (e.g., Airn, Kcnq1ot1, and H19) [Barlow and Bartolomei, 2014]. Local action is a viable hypothesis that needs to be evaluated on a case-by-case basis.

Although the origin of genomic specificity for lncRNAs that act immediately at their site of transcription is self-evident, some cis-acting RNAs are thought to regulate gene expression in cis but across longer distances. These distances can extend across thousands of base pairs or even across hundreds of megabases (in the case of Xist) of

linear DNA sequence. For longer-distance targeting, models have been proposed (for examples, see [Engreitz et al., 2013, Hacisuleyman et al., 2014, Lai et al., 2015]) in which the lncRNA acts through the three-dimensional organization of chromatin (e.g., through looping and topologically associated domains). In these models, the RNA could either drive the looping interaction or use preexisting chromatin interactions to drive genomic specificity. These models of looping interactions are reminiscent of enhancer-promoter looping interactions, and in this light, it is interesting that some lncRNAs are expressed from enhancer regions of the genome [Kim et al., 2010]. The regulatory significance of enhancer transcription and related lncRNAs is an active area of investigation [Kim et al., 2015, Orom and Shiekhattar, 2013].

As noted above, in many cases where lncRNA transcription has been observed, there is reason to wonder whether the RNA or the transcription (or perhaps neither) is important for regulation. The distinction between transcription-based models and RNA-based models of local regulation may be a false dichotomy. Another model that has recently been the focus of intense study is that local RNA, with minimal sequence requirements, could assist in binding and recruitment of proteins. This would explain the relatively low sequence specificity observed when examining the RNA-binding specificity of chromatin-modifying machinery such as PRC2 and still support a role for RNA in chromatin regulation. This model was proposed in reports focusing on PRC2 recruitment [Davidovich et al., 2013, Kaneko et al., 2014b, Kaneko et al., 2014a], and similar models have been proposed for YY1 [Sigova et al., 2015] and the targeting of Tip60-p400 by R loops [Chen et al., 2015]. These models of relatively nonspecific local RNA binding have been proposed to either stimulate or repress chromatin modifications, depending on the context.

1.10 HOW DO RNA ELEMENTS WITHIN LONG NONCODING RNAs INFLUENCE THEIR FUNCTION?

Although very little is known about the structure and mechanism of mammalian lncRNAs, substantial progress has been made using approaches to study how lncRNA elements influence processing, stability, modification, conformation, and binding properties of the lncRNA. In many cases, these developments use techniques pioneered for studying abundant RNAs (such as rRNAs), and adapting these approaches to new sequencing platforms can improve throughput and sensitivity. The biogenesis of most lncRNAs is thought to be similar to that of protein-coding mRNAs. However, there are some interesting examples where lncRNAs have different processing mechanisms that relate to their stabilities (reviewed in [Wilusz, 2016]. One example is an RNA-stabilizing element that was discovered in a lncRNA in the Kaposi's sarcoma–associated herpesvirus that accumulates to high concentrations in the lytic phase of viral infection [Conrad et al., 2007, Conrad et al., 2006]. This element forms a triple helix with the tail of the RNA [Mitton-Fry et al., 2010]. Similar stabilizing triple-helical structures were identified for the endogenous lncRNAs NEAT1 and MALAT1 [Brown et al., 2012, Wilusz et al., 2012]. Interestingly, NEAT1 and MALAT1 also have noncanonical 3′ processing pathways involving RNase P cleavage of a tRNA-like element from the 3′ end [Wilusz, 2016]. Another processing pathway that stabilizes RNAs is circularization [Gardner et al., 2012, Qian et al., 1992, Zhang et al., 2013], demonstrating a diversity of mechanisms by which lncRNA biogenesis and stability can differ from canonical mRNA processing pathways.

RNA stability, whether controlled by these noncanonical pathways or by more traditional mechanisms, is important to lncRNA biology because steady-state RNA levels are determined by both synthesis and degradation. The importance of regu-

lated RNA degradation has attracted increased attention and has motivated large studies to monitor RNA turnover transcriptome-wide. In these experiments, non-canonical nucleosides can be fed to cells, where they are integrated into new transcripts (reviewed in [Tani et al., 2012]. Improvements to chemical enrichment strategies [Duffy et al., 2015] and analysis pipelines [Rabani et al., 2014] have added to the power of these experiments. Approaches to examine RNA stability have been used to distinguish long-lived from short-lived lncRNAs, with the hypothesis that long-lived lncRNAs are less likely than short-lived lncRNAs to result from transcriptional noise [Clark et al., 2012].

Covalent modifications to individual nucleosides can also influence lncRNA stability and function. Hundreds of posttranscriptional RNA modifications have been discovered across different branches of life (mostly in tRNA and rRNA), and it is unclear how many of these play an important role in influencing mammalian lncRNAs that regulate chromatin. Thus far, genome-wide approaches have shown extensive modification of mammalian RNAs by deamination, pseudouridylation, and methylation [specifically of the exocyclic amine of adenine, forming N6-methyladenosine (m6A)] [Carlile et al., 2014, Dominissini et al., 2012, Ramaswami et al., 2012, Schwartz et al., 2014]. In the case of m^6A, this modification has been connected to regulation of RNA degradation [Wang et al., 2014].

In addition to these covalent alterations to RNA connectivity and base chemistry, extensive progress has been made in understanding noncovalent conformations of RNA. There is extensive precedent for RNA to fold into elaborate structures capable of diverse biochemical activities. Examples with structural characterization include the ribosome [Moore and Steitz, 2002], the self-splicing group II intron [Toor et al., 2008], and a diverse array of prokaryotic riboswitches [Roth and Breaker, 2009]. These examples demonstrate both the binding and the regulatory potential of various relatively short RNA structures. In general, we know little about mammalian lncRNA

structure, including the degree of conformational homogeneity, the roles of proteins in stabilizing ribonucleoprotein structure and function, and the degree of local- versus higher-order structure of lncRNAs. Early lessons from extensive in vitro probing of SRA and HOTAIR lncRNAs [Novikova et al., 2012, Somarowthu et al., 2015] have demonstrated that lncRNAs can adopt complex conformations. Chemical probing experiments have recently been extended to transcriptome-wide analysis of lncRNA in mammalian cells [Ding et al., 2014, Rouskin et al., 2014, Spitale et al., 2015b]. To address the relatively low concentrations of lncRNAs in mammalian transcriptomes, this probing can be performed in a targeted format [Kwok et al., 2013]. One such approach has led to conformational models of all elements of Xist that are predicted to be structured [Fang et al., 2015]. In the longer term, structural characterization of lncRNAs will undoubtedly provide important insight into lncRNA mechanisms and interactions. For example, in the case of the roX lncRNAs, ATP-dependent remodeling of a stem loop by the MLE RNA helicase leads to assembly of the MSL chromatin-modifying complex. Understanding RNA elements can lead to mechanistic insight into lncRNA functions on chromatin. Further biochemical characterization will provide a necessary foundation for structural characterization of lncRNAs and their complexes with chromatin-modifying proteins. Traditional techniques to reconstitute and study protein-RNA interactions include gel shift experiments, nucleotide interference mapping, and footprinting. Recent efforts have used sequencing platforms or technology to expand the throughput of these experiments to large arrays [Tome et al., 2014] or the entire transcriptome (discussed in [Baltz et al., 2012, Silverman et al., 2014]). In vivo cross-linking has been central to defining RNA-protein interactions using both protein-centric techniques [e.g., cross-linking and immunoprecipitation [Licatalosi et al., 2008]] and RNA-centric techniques [e.g., CHART mass spectrometry (CHART-MS) [West et al., 2014]]. Indeed, hybridization capture analyses have led to the discovery of new proteins that bind

lncRNAs, such as an interaction between SHARP/SPEN and Xist [Chu et al., 2015, McHugh et al., 2015, Minajigi et al., 2015].

In summary, we still know very little about the biochemistry of lncRNAs, their structures, and their interactions with chromatin proteins. Nonetheless, rapid progress in probing and enrichment techniques and the increased sensitivity of sequencing and mass spectrometry platforms provide reason for optimism.

1.11 Outlook

It is clear that lncRNAs can play important roles in regulating chromatin structure, yet we still know little about the scope of this regulation, its impact, and its mechanisms. Techniques for classifying and annotating these RNAs are still evolving. Although the discovery of classic lncRNAs that function on chromatin was largely serendipitous, most recent reports of functional lncRNAs have emerged from targeted loss-of-function studies. By contrast, one recently uncovered class of lncRNAs, named asynchronous replication and autosomal RNAs (ASARs), was discovered at loci responsible for the replication timing of individual chromosomes [Stoffregen et al., 2011]. ASARs can spread in cis along the chromatin, providing an intriguing hypothesis for how lncRNAs could be involved in coordinating the replication timing of a chromosome [Donley et al., 2015]. This hypothesis highlights one interesting theme that has emerged regarding lncRNA function in the genome: cis-regulation, but at a distance.

Increased clarity will come as the biochemical specificity of lncRNAs is connected to the functional specificity observed in vivo. Along these lines, one particularly exciting development has been the use of CRISPR/Cas9 proteins to direct RNA elements to well-defined sites in the genome [Shechner et al., 2015]. This type of tool can potentially help connect the biochemistry of individual RNA elements with their

functions on chromatin.

Beyond the discovery of new lncRNA functions, we look forward to more examples like roX and Xist lncRNAs, in which multiple laboratories corroborate each other's results. There is growing consensus about the challenges of studying the functions and mechanism of lncRNAs. Understanding these challenges is the key to developing new and better technologies to understand lncRNAs and design appropriate control experiments. We expect these advances to be instrumental in revealing the roles of lncRNAs in the regulation of chromatin biology.

Chapter 2

High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation

2.1 Summary

This chapter describes collaborative work analyzing the association pattern of the Xist RNA across the X-chromosome, its establishment during development, and its recovery process from perturbation in somatic cells. All of these analyses have the goal of understanding the Xist RNA's role in X-inactivation. Below, I describe my role in this project, referencing figures from the published journal article related to this work. I then reproduce (with permission) the full journal article journal article with slight modifications:

- Simon, MD, Pinter, SF, Fang, R, Sarma, K, Rutenberg-Schoenberg, M, Bowman, SK, Kesner, BA, Maier, VK, Kingston, RE, Lee, JT **(2013)**. High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. *Nature*, 504, 7480:465-469.

I also list all figures within this paper that I drew directly and those that were drawn using my analysis.

2.2 Description of indepedent work within collaboration

As discussed is chapter 1, the Xist RNA is the master regulator of the transcriptional inactivation of one of the two X-chromosomes in most female mammalian cells [Lee, 2009]. The Xist RNA was already known to associate with the inactive X chromosome from fluorescence in situ hybridization (FISH) experiments [Lee, 2009]. However, the association pattern was not known in greater detail. In this study, we used the then recently developed Capture Hybridization of RNA Targets (CHART-Seq) technology to map the pattern of Xist association with the inactive X-chromosome in mouse cells [Simon et al., 2011]. This technology is analogous to the widely used chromatin immunoprecipitation and sequencing (ChIP-Seq), but whereas antibodies are used to enrich DNA crosslinked to transcription factors and chemically modified histones in ChIP-Seq, biotinylated antisense oligonucleotides are used to enrich DNA crosslinked to an RNA of interest in CHART-Seq.

Within this collaborative project, I focused on comparison of Xist association patterns between different cell states, as well as reproducibility of CHART-Seq data between replicates. Unlike transcription factors and chemically modified histones [Park, 2009], as well as the roX2 RNA, which regulates chromosomal dosage compensation in *Drosophila* [Simon et al., 2011], the Xist RNA is enriched across nearly the entire X-chromosome in mouse embryonic fibroblasts, which have undergone full transcriptional inactivation 2.1c. This meant that traditional peak-calling methods to identify relatively short enriched regions ($\sim$ 100s-100,000s bp) had limited applications to comparing patterns of Xist enrichment across cellular conditions.

As an alternative approach, I compared the normalized signal in 40kb bins between conditions. Pearson correlations of different datasets provided insight into the reproducibility of replicate experiments (Figure 2.5d) as well as overall similarity between conditions (Figure 2.6a). We further were interested in visualizing areas with signal differences between conditions. To do this, we identified signal bins with $>$ 10-fold differences in signal between conditions and plotted these differences across the X-chromosome.

Our group conducted CHART-Seq experiments in three different sets of biological contexts, and I used the above methodology to compare results from each to one another. First, we conducted CHART-Seq in mouse embryonic fibroblasts, which are differentiated cells that display full inactivation of the X-chromosome. These cells display enrichment of the Xist RNA across nearly the entire X-chromosome. To investigate the establishment of Xist association with the X-chromosome during development, we performed CHART-Seq to measure Xist association in mouse embryonic fibroblasts (MEFs) at different time points after LIF withdrawal (0 days, 3 days, 7 days, and 10 days). We also measured the recovery of Xist association in MEF cells after knockoff in with locked nucleic acid (LNA) probes that had previously been shown to displace Xist from the inactive X-chromosome [Sarma et al., 2010].

For all of the above experiments, biological samples were provided by the Jeannie Lee lab and cell culture was conducted by Drs. Stefan Pinter, Kavitha Sarma and Rui Fang. Technology development for CHART-Seq experiments was conducted by Drs. Rui Fang and Matthew D. Simon and final experiments were conducted by Dr. Rui Fang. Read alignment and signal normalization to provide input for my analysis was conducted by Dr. Matthew D. Simon (see detailed methods below).

I first focused on the pattern of establishment of the X chromosome inactivation in mouse embryonic stem cells. It was notable that the bins containing 10-fold enrichment between time points (day 7 vs. day 3) and between a late time point and

differentiated cells (day 7 vs. MEF) were highly clustered within the X-chromosome and lay at the edges of domains with high Xist association signal (Figure 2.1d,e). This enabled us to conclude that in our embryonic stem cell model, Xist initially targets a set of "early" domains before spreading toward near-full coverage of the X-chromosome.

In addition to comparing stages of establishment of X inactivation to fully differentiated cells, I also examined the restults of experiments in which Xist was knocked off of the X-chromosome using antisense LNA probes and then allowed to recover. Specifically, cells were treated with each of two different LNA for 3 hours, at which point most Xist association with the inactive X-chromsome was lost by fluorescence, and then allowed to recover to an 8 hour timepoint, where Xist association was partially regained. Using the same procedure to compare Xist CHART-Seq profiles, I identified differential regions with $>$ 10-fold differences between the Xist depletion (3 hr) and recovery (8 hr) timepoints. Strikingly, even though Xist signal magnitude had not fully recovered to normal levels, enriched Xist regions were spread throughout the X-chromosome (Figure 2.3b-f). This contrasted with the establishment of Xist association with the X-chromosome in mouse embryonic stem cells, where specific domains were enriched prior to further spreading across the chromosome.

To quantify our comparison between patterns of Xist deposition patterns in development and in recovery from LNA knockoff in MEF cells, we chose to focus on comparisons of "early" and "late" domains. To make this possible, I defined late domains as those where Xist signal in MEF cells was at least 10-fold enriched over day 7 mouse embryonic stem cells. I let other X-chromosomal regions become early domains by default. Using this demarcation and my processed signal files, Dr. Rui Fang plotted figures comparing the signal in late domains relative to signal in early domains (Figure 2.3g and Figure 2.10). These plots show that "late" domains are significantly enriched in LNA recovery in MEF cells, relative to intermediate points

in Xist deposition in ES cells (day 3 and day 7).

Based on my analysis, we specifically demarcated "early" and "late" domains by defining late domains as those where Xist signal in MEF cells was at least 10-fold enriched over day 7 mouse embryonic stem cells. Other X-chromosomal regions were, by default, referred to as early domains. This demarcation of different X-chromosome regions enabled comparison of relative signal in early and late domains (Figure 2.3 g). To further illustrate the differences in Xist deposition patterns between mouse ES cells and recovery from LNA knockoff in MEF cells, I identified 10-fold different regions between MEF, d7, and LNA 8hr samples in comparison to all other samples. This analysis illustrates that regions that are enriched in ES day 7 over LNA 8 hr time points primarily lie in "early" domains and regions where LNA 8 hr time points are enriched over ES day 7 are primarily in "late" domains (shown as MEF > d7; Figure 2.11).

Within the work described below, I plotted the following figures:

- Figure 2.1 d,e
- Figure 2.3 e,f
- Figure 2.5 d
- Figure 2.6 a
- Figure 2.11

Additionally, the following figures incorporate analyses that I conducted:

- Figure 2.3 b,c,d,g
- Figure 2.6 c
- Figure 2.10

2.3 High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation

2.3.1 Abstract

The Xist long noncoding RNA (lncRNA)is essential for X-chromosome inactivation (XCI), the process by which mammals compensate for unequal numbers of sex chromosomes [Disteche, 2012, Wutz, 2011, Lee, 2012]. During XCI, Xist coats the future inactive X chromosome (Xi) [Clemson et al., 1996] and recruits Polycomb repressive complex 2 (PRC2) to the X-inactivation centre (Xic) [Zhao et al., 2008]. How Xist spreads silencing on a 150-megabase scale is unclear. Here we generate high-resolution maps of Xist binding on the X chromosome across a developmental time course using CHART-seq. In female cells undergoing XCI de novo, Xist follows a two-step mechanism, initially targeting gene-rich islands before spreading to intervening gene-poor domains. Xist is depleted from genes that escape XCI but may concentrate near escapee boundaries. Xist binding is linearly proportional to PRC2 density and H3 lysine 27 trimethylation (H3K27me3), indicating co-migration of Xist and PRC2. Interestingly, when Xist is acutely stripped off from the Xi in post-XCI cells, Xist recovers quickly within both gene-rich and gene-poor domains on a timescale of hours instead of days, indicating a previously primed Xi chromatin state. We conclude that Xist spreading takes distinct stage-specific forms. During initial establishment, Xist follows a two-step mechanism, but during maintenance, Xist spreads rapidly to both gene-rich and gene-poor regions.

2.3.2 Introduction

Xist RNA is a prototype lncRNA with global epigenetic function[Disteche, 2012, Wutz, 2011, Lee, 2012, Pontier and Gribnau, 2011]. The initiation of XCI depends on Xist [Brown et al., 1992] and loading of the Xist–PRC2 complex at a nucleation site within the Xic [Jeon and Lee, 2011]. Thereafter, Xist RNA forms a "cloud" over the X-chromosome, signalling the initiation of chromosome-wide silencing [Clemson et al., 1996]. Concurrently, PRC2 accumulates broadly along the X-chromosome [Pinter et al., 2012]. Although Xist RNA coats the Xi at cytological resolution, whether and where Xist binds at molecular resolution remains unknown. In one model, Xist targets PRC2 to the Xic, but outward spreading of PRC2 does not involve Xist. Alternatively, both nucleation and spread involve Xist, in which case Xist and PRC2 would co-migrate at a molecular scale.

2.3.3 Results and Discussion

We mapped genome-wide binding locations of Xist RNA by performing CHART-seq (capture hybridization analysis of RNA targets with deep sequencing), a technique to localize lncRNAs on chromatin using complementary oligonucleotides to enrich for DNA targets [Simon et al., 2011] (Figure 2.1a). We designed a cocktail of 11 complementary oligonucleotides for Xist CHART based on conserved or functional Xist domains [Brown et al., 1992, Brockdorff et al., 1992, Wutz et al., 2002, Sarma et al., 2010] and RNase H mapping for accessibility (Figure 2.4b,c and Extended Data Table 1). Allele-specific CHART-seq was performed at four developmental stages (Figure 2.4d): before XCI in undifferentiated female mouse embryonic stem (ES) cells (d0; ,1% of nuclei XCI positive, showing an Xist cloud or H3K27me3 focus), early-XCI (d3; ,10% positive), mid-XCI (d7; 40–50% positive), and post-XCI (mouse embryonic fibroblast (MEF) clone, $>$ 95% positive). About 600,000 sequence polymorphisms between the *Mus musculus* (mus) and *Mus castaneus* (cas)

X-chromosomes enabled > 35% allele-specific mapping to Xi and Xa (active X chromosome), respectively [Pinter et al., 2012]. Disabling the musTsix allele in the female ES cells ensured that the mus X will be Xi [Ogawa et al., 2008]. We validated results by comparing two independent capture oligonucleotide sub-mixtures and an alternative 40-oligonucleotide cocktail targeting across the length of Xist (Figure 2.2a-e and Extended Data Table 1). Regions with significant Xist enrichment localized almost exclusively to Xi (>99% X-linked, $P < 0.001$; >90% Xi-skewed, $P < 0.05$, Figure 2.5f,g,i). On autosomes, binding was minimal and of questionable significance. Enriched segments were not complementary to capture-oligonucleotides and showed minimal enrichment on Xa of d0, d3, d7 and MEF cells. Enrichment was not observed using sense control oligonucleotides (Figure 2.2a,c). These experiments excluded artefactual enrichment, validating Xist CHART-seq specificity.

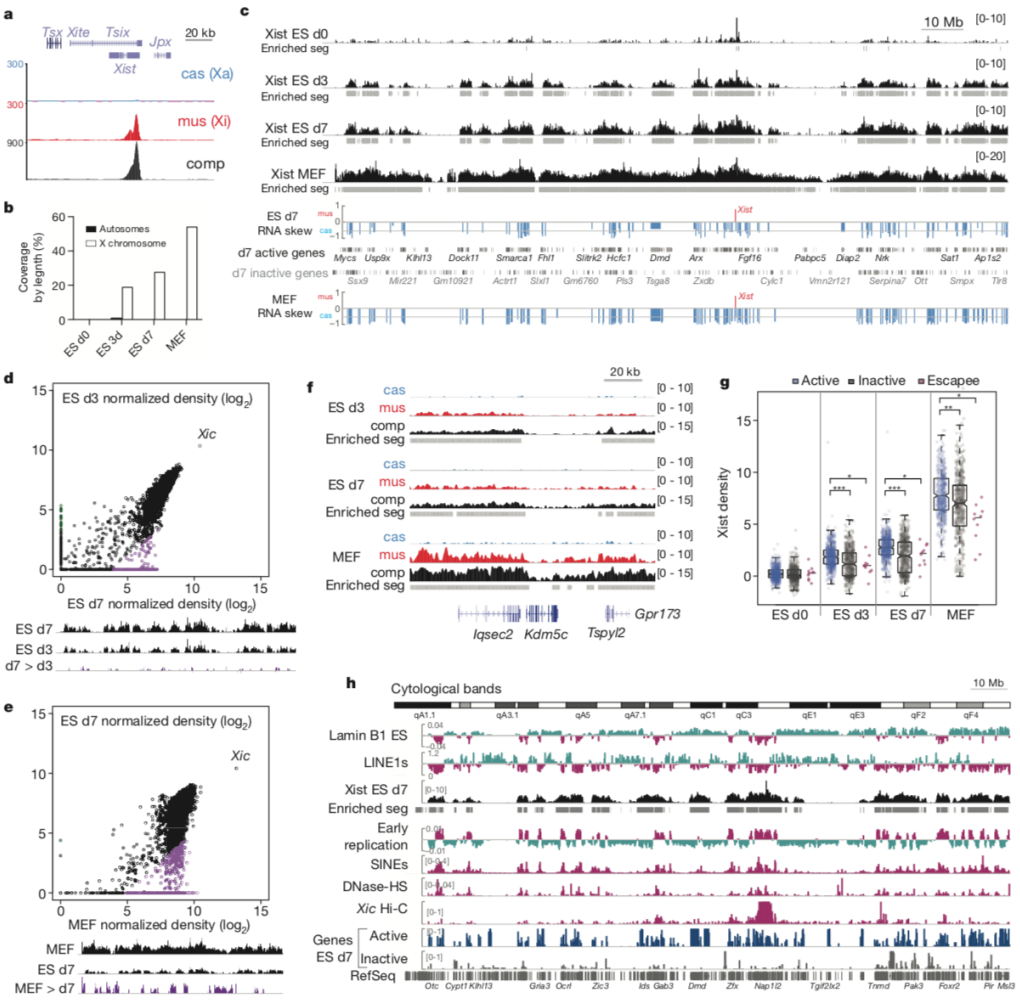

Figure 2.1: CHART-seq reveals a two-step mechanism of Xist spreading during de novo XCI.

a, Xist RNA is enriched on Xi. Normalized read densities displayed in mus, cas and composite (comp) tracks. **b,** Coverage of enriched segments on the X chromosome and the autosomes. **c,** Xist coverage at indicated time points relative to gene silencing. Enriched segments shown beneath in grey. Brackets, y-axis scale of normalized Xist density as in all figures. Xist peaks at d0 have less amplitude and density, but reflect d3 and d7 patterns, and are Xi-enriched (Figure 2.5f), consistent with initial Xist spreading to local regions, suggesting initial differentiation in a subfraction of cells. RNA-seq of d7 and MEF is shown below. Skewed allelic expression consistent with Xi-silencing (value -0.5 = threefold expression difference between Xi and Xa). **d, e,** Xist CHART signals (40-kilobase bins) from d7 correlate with d3 **(d)** and MEF **(e)** (see Figure 2.6). Regions showing more than tenfold differences after normalization are coloured purple and displayed on the X chromosome in screenshot panels below. **f,** Depletion of Xist at a representative escapee. **g,** Xist preferentially targets genes in active chromatin (H3K4me3-marked on d7). Xist densities shown for gene bodies of active (n = 532), inactive (n = 475) and escapee genes (n = 10). Medians are indicated. Individual data points overlaid on boxplot; error bars, 1.5-fold interquartile range. $^{*}p < 0.05$, $^{**}p < 10^{-8}$, $^{***}p < 2.2 \times 10^{-16}$, Mann–Whitney U tests. **h,** Xist RNA distribution from d7 cells relative to 200-kb binned chromatin features (y-scale is fraction of binned sequence unless otherwise indicated): SINEs, LINE1s (multiple LINE1 annotations from RepeatMasker included, fraction of $\frac{SINE}{LINE1}$ nucleotides in 200-kb windows; LINE1s with chosen midpoint of 0.6 to highlight anti-correlation), DNase hypersensitive sites (DNase-HS), Xic Hi-C (y-scale is normalized Hi-C signal from 40-kb Xic bin containing Xist), $\frac{active}{inactive}$ genes (classification based on presence of H3K4me3 at Xa promoters in d7 ES cells), and lamin B1 association and replication timing (y-scale is normalized microarray probe intensities).

The dominant CHART peak was in Xist exon 1 and was specific to Xi (Fig. 2.1a). A developmental time course demonstrated a progression in Xist density, with enriched segments increasing from 0.1% coverage of the X in pre-XCI cells to approximately 20% in early- and midXCI, and approximately 54% in post-XCI cells (Fig. 2.1b,c and Extended Data 2.2h). Thus, Xist RNA not only forms a cytological cloud but also binds broad swaths of the Xi at molecular resolution. Xist could either spread uniformly along the Xi or target specific regions. Intriguingly, in cells undergoing XCI (d3, d7), Xist preferentially targeted multimegabase domains (Fig. 2.1c). In post-XCI MEFs, Xist spread into intervening gene-poor regions throughout the Xi. The d3 and d7 patterns were more similar to each other than to MEF patterns (Fig. 2.1d,e and Extended Data Fig. 2.6a). Furthermore, comparative analysis identified MEF-specific domains not found during XCI (Fig. 2.1e). Despite heterogeneity in the onset of XCI in the ex vivo ES differentiation system, the highly similar d3 and d7 distributions show that Xist targets generich domains first. Extension of ES differentiation to d10 showed statistically significant filling in of gene-poor domains (Extended Data 2.6b,c), although not to the extent observed in somatic cells (MEFs). We infer that full spreading across Xi may only be achieved later in development, once differentiation into somatic lineages occurs. Thus, during de novo XCI in the embryo, Xist probably follows a two-step pattern of spreading, first targeting gene-rich clusters (hereafter, early domains) and eventually spreading to intervening gene-poor regions (late domains). Throughout the process, gene bodies of escapees [Berletch et al., 2011, Carrel and Willard, 2005] were depleted of Xist, but occasionally demonstrated Xist enrichment in flanking regions (Fig. 2.1f and Extended Data Fig. 2.7), indicating boundaries that sequester Xist and prevent spreading into neighbouring privileged escapee loci.

We investigated what might target Xist to early domains by comparisons with various chromatin features (see Methods) [Pinter et al., 2012, Splinter et al., 2011,

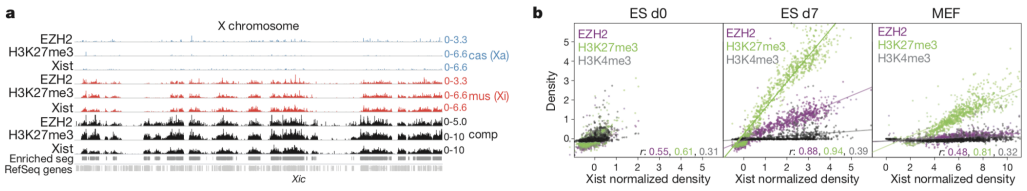

Figure 2.2: Co-spreading of Xist RNA and PRC2.

a, Normalized read densities of Xist, EZH2 and H3K27me3 on the X chromosome in d7 cells. **b,** Xist densities (200-kb bins) correlated with EZH2, H3K27me3 and H3K4me3 signals at different stages of XCI. Pearson's r displayed. EZH2/H3K27me3 R^2 values: 0.3/0.37 for d0, 0.77/0.88 for d7, and 0.23/0.66 for MEFs, respectively. H3K4me3 R^2 values: < 0.15 across all samples.

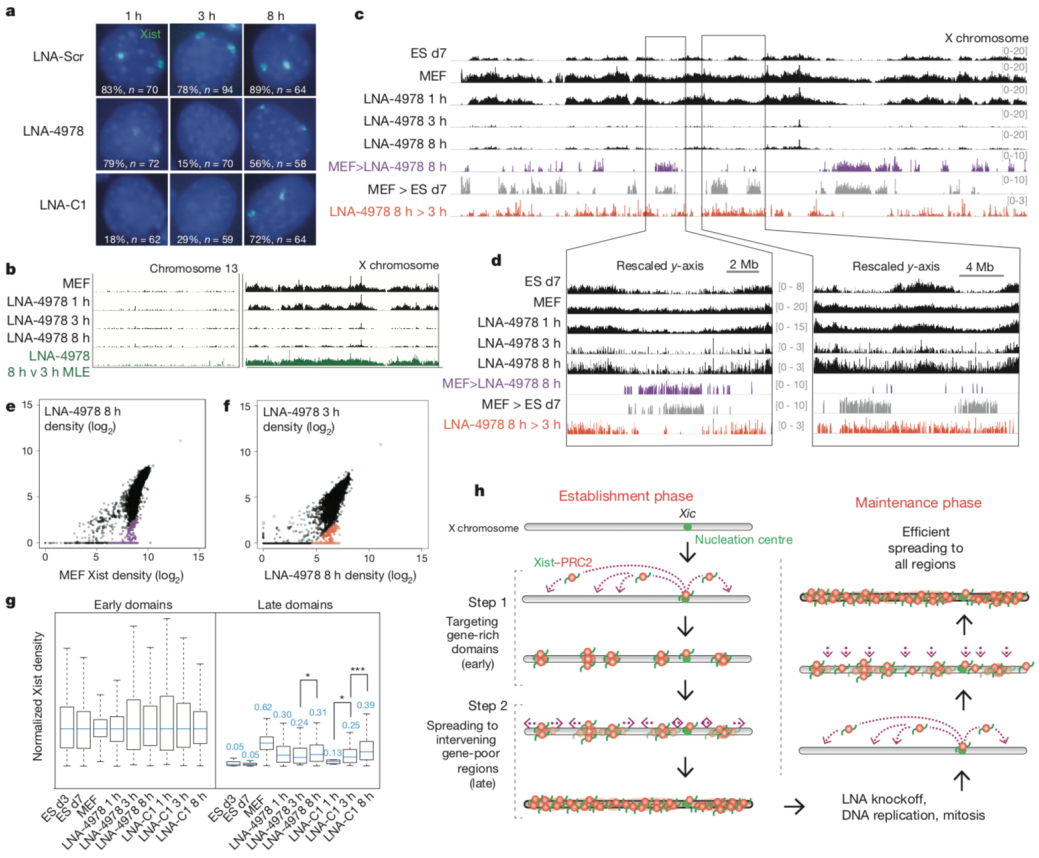

Figure 2.3: Figure 3 Xist knockoff uncovers a distinct spreading method during the maintenance phase.

a, RNA FISH shows depletion and recovery of Xist RNA (green) in MEF cells after Xist knockoff. Per cent of nuclei with Xist clouds and sample size (n) shown. Scr, scrambled LNA. **b,** Chromosome-wide recovery of Xist after LNA-4978 knockoff on chromosomes X and 13. Regions of recovery comparing 8 h over 3 h LNA-4978 were determined using a maximum likelihood enrichment (MLE) estimate. **c, d,** Xist knockoff and recovery across the X chromosome. Coloured regions show more than tenfold median-normalized differences between samples. **d,** Expanded view of one region with more late domain recovery (right) than the other (left). **e, f,** Xist CHART signals (40-kb bins) from LNA-4978 8 h correlated with MEF **(e)**; LNA-4978 3 h correlated against LNA-4978 8 h **(f)**. Regions showing more than tenfold differences after normalization are coloured as shown in d. g, Xist recovery in indicated samples, with 40-kb-binned Xist densities normalized to median levels of early domains of each sample, to determine how early and late domains recover from knockoff compared to during *de novo* XCI. Normalized median values for each sample indicated above box. $*P < 0.05; ***P < 10^{-8}$, Wilcoxon test, as in Extended Data Figs 7c and 3c. **h,** Model, distinct methods of Xist spreading during establishment and maintenance.

Dixon et al., 2012, Chadwick and Willard, 2004, Chadwick and Willard, 2003, Marks et al., 2009, Calabrese et al., 2012]. Interestingly, Xist is more likely to target genes in regions of active chromatin in ES cells. Allele-specific RNA sequencing analysis demonstrated the preference of Xist for genes that are active (for example, on the Xa and in d0 and d7 cells) and showed skewed expression in d7 ES cells and in MEFs element-1 (LINE1) ($r = -0.54$), and lamin-associated domains (LADs, $r = -0.48$) [Bickmore and van Steensel, 2013]. Xist partitioning did not correlate with cytogenetic banding on the X chromosome (2.1h) [Chadwick and Willard, 2004, Duthie et al., 1999]. LINE1s have been proposed as spreading elements25, but repetitive reads from Xist CHART-seq aligning to LINE1 were not enriched over input (Figure 2.8c). The localization of Xist showed modest positive correlation with Xic looping contacts inferred from HiC (high-throughput chromosome conformation capture) through an anchor within the Xist locus (2.1h and Extended Data Fig. 2.8b). Together, these data support a role for open chromatin in guiding Xist, with Xist coming into contact with gene-rich regions (early domains) first, and spreading secondarily to more distal gene-poor inter-regions (late domains).

Given co-nucleation of Xist and PRC2 at the Xic [Jeon and Lee, 2011], we asked whether Xist continues to associate with PRC2 during spreading. Comparison of Xist, EZH2 and H3K27me3 enrichment revealed strikingly similar chromosome profiles across time (2.2a and Extended Data Fig. 2.9a–c). By contrast, PRC2 and H3K27me3 densities on Xa did not correlate with Xist, nor did those on chromosome 13, a representative autosome (Extended Data 2.9a). Consistent with the idea that Xist directs PRC2 localization onto Xi [Zhao et al., 2008], Xist densities demonstrated an extensive linear relationship with EZH2 and its product H3K27me3 across the X chromosome in mid-XCI but not pre-XCI cells (Fig. 2.2b). Correlation with the H3K4me3 control (active mark) was poor. In MEFs, densities of H3K27me3 and Xist remained highly correlated, whereas reduced densities of PRC2 were observed during

maintenance. Interestingly, Xist densities were not necessarily greater at previously defined "PRC2 strong sites" [Pinter et al., 2012] (Figure2.9d,e); instead, Xist densities showed a general correlation with Xi-specific PRC2 enrichment (Extended Data Figs 2.8b and 2.9b,h). This supports the idea that strong sites are Xist-independent (as indeed they are present in d0 cells [Pinter et al., 2012]) and indicate that Xist and PRC2 co-migrate to new regions within the early domains on the Xi.

We then asked if localization mechanisms were inherent to Xist RNA or chromatin context. In perturbation experiments, we stripped away Xist RNA and observed recovery on the Xi of MEFs at 1 h, 3 h and 8 h. Locked nucleic acids (LNA) directed against repeat C of Xist RNA prevented nucleation and therefore spreading [Sarma et al., 2010]. RNA fluorescence in situ hybridization (FISH) showed that LNA-4978 did not overtly perturb Xist at 1 h, but led to full Xist displacement by 3 h, with Xist reassociation at 8 h (Fig. 2.3a). As reassociation requires newly synthesized Xist rather than relocalization of displaced Xist [Sarma et al., 2010], reassociation must depend on outward spreading of new RNA from the Xic, just as during XCI establishment.

Interestingly, however, CHART-seq revealed a pattern not evident cytologically by RNA FISH. At 1 h, when Xist was still visualized on Xi (Fig. 2.3a), CHART-seq demonstrated a relative loss in late domains (Fig. 2.3b-d), indicating that Xist binds more weakly to gene-poor than to gene-rich regions, and consistent with the banded pattern of Xist on the metaphase Xi observed cytologically [Duthie et al., 1999]. At 3 h, Xist was strongly depleted from both regions. At 8 h, partial recovery was evident in both regions. However, unlike spreading during de novo XCI (d3, d7), spreading of Xist during the somatic maintenance phase (MEF) did not follow a two-step process, as Xist reassociation in early and late domains occurred simultaneously (Fig. 2.3b-d). Therefore, spreading during de novo XCI was restricted to early domains and occurred on a timescale of days in the ex vivo system. In contrast, recovery and re-spreading

in post-XCI cells occurred more generally in both domains and on a timescale of hours. This quantitative difference is significant, with accumulation in late domains appearing on the same timescale as early domains during the recovery period after Xist knockoff (Fig. 2.3e-g and Extended Data Figs 2.10 and 2.11). Similar results were observed using an independent LNA, LNA-C1, targeted to a different sequence in the repeat C region and in multiple replicates (Figures 2.3a, 2.5, 2.6 , 2.10, and 2.11). Despite LNA-C1 being faster acting [Sarma et al., 2010] (Figure 2.3a), LNA-C1 and LNA-4978 treatment resulted in remarkably similar Xist knockoff and recovery on Xi.

Taken together, these data provide evidence for distinct mechanisms of Xist spreading during establishment (de novo XCI) in early embryonic cells, when spreading occurs in a two-step fashion (early to late domains), and during maintenance in somatic cells, when Xist spreads more generally into both early and late domains (Figure 2.3h). The Xi may retain an epigenetic memory of Xist [Kohlmaier et al., 2004], enabling more efficient spreading during maintenance. As Xist mostly dissociates from the Xi during mitosis [Clemson et al., 1996], epigenetic memory could facilitate the resynthesis of Xist and re-spreading in G1, and duplication of Xist patterns after DNA replication. Indeed, the continued action of Xist is essential for maintenance of XCI [Yildirim et al., 2013]. In summary, we have illuminated the mechanism by which Xist spreads on a 150-Mb scale. Comparing localization dynamics of Xist relative to other lncRNAs (E. Hasiculeyman and J. Rinn, personal communication, [Hacisuleyman et al., 2014]) and three-dimensional conformations [Engreitz et al., 2013] may prove highly informative for understanding general mechanisms of RNA-directed chromatin change.

2.3.4 Methods

Capture oligonucleotides. Capture oligonucleotides were designed based either on repetitive sequences in Xist RNA (X.A, X.C, Extended Data Fig. 1b and Extended

Data Table 1), or by using RNase H mapping [Simon et al., 2011] of functional regions of Xist [Wutz et al., 2002] to identify sites in Xist RNA available for hybridization in crosslinked chromatin extracts (Figure 2.4c and Extended Data Table 1). Regions of sensitivity were further interrogated by BLAST to identify oligonucleotides both with minimal cross-hybridization potential to other RNAs and genomic sites, as well as similar melting temperatures. Oligonucleotides were either synthesized as previously described on an Expedite Oligo synthesizer and purified using reverse phase cartridges (Poly-Pak II, Glen Research) [Simon et al., 2011], or ordered commercially (IDT, 39-biotinTEG, iSp18 spacer modified, salt free) and used without further purification.

Alternative capture oligonucleotides. Alternative capture oligonucleotides (CO40; see Extended Data Fig. 2a, c, d) were designed using the oligowiz software [Wernersson and Nielsen, 2005] limiting lengths to 22–28 nucleotides and otherwise default parameters. Mouse (mm9) transcripts were screened to minimize cross-hybridization to off-target transcripts and 40 oligonucleotides (out of 263 candidates) were picked manually to cover the length of the Xist transcript in 300–500 nucleotide intervals where possible (Extended Data Table 1). Standard unmodified desalted oligonucleotides were ordered commercially (IDT), resuspended and pooled. 39 biotinylation of pooled oligonucleotides was carried out as previously described31 using biotin-16-UTP, and biotinylated oligonucleotides recovered after a single chloroform extraction and nucleotide removal (Qiagen). Xist CHART enrichment. Clonal female MEFs [Yildirim et al., 2011] and female TsixSTOP ES cells [Ogawa et al., 2008] were cultured as previously reported [Pinter et al., 2012], including differentiation of ES cells by LIF withdrawal. Xist CHART enrichment was performed as previously reported [Simon et al., 2011] with minor modifications. Briefly, 108 cells were crosslinked initially with 1% formaldehyde for 10 min at room temperature. The crosslinking reaction was stopped by adding 0.125 M glycine. After washing 3 times with PBS, crosslinked cells were re-suspended in 10 ml sucrose buffer, dounced 20

times with a tight pestle, and kept on ice for 10 min. Nuclei were collected by centrifugation at 1,500g for 10 min on top of a cushion of 25 ml glycerol buffer. Nuclei were further crosslinked with 3% formaldehyde for 30 min at room temperature. After washing three times with PBS, nuclei were extracted once with 50 mM HEPES pH 7.5, 250 mM NaCl, 0.1 mM EGTA, 0.5% N-lauroylsarcosine, 0.1% sodium deoxycholate, 5 mM DTT, 100 U ml^{-1} SUPERasIN (Invitrogen) for 10 min on ice, and centrifuged at 400g for 5 min at 4 uC. Nuclei were resuspended in 1.5 ml 50 mM HEPES pH 7.5, 75 mM NaCl, 0.1 mM EGTA, 0.5% N-lauroylsarcosine, 0.1% sodium deoxycholate, 5 mM DTT, 100 U ml^{-1} SUPERasIN, and sonicated in microtubes using Covaris E210 sonicator at 10% duty cycle, 200 bursts per cycle, intensity 3 for 5 min. The median size of chromatin fragments was ,3 kb as determined by agarose gel electrophoresis with ethidium bromide post-staining. For each CHART enrichment, 120 ml of cleared chromatin extract was incubated overnight with 36 pmol capture oligonucleotides in a total volume of 360 ml 33 mM HEPES pH 7.5, 808 mM NaCl, 0.17% N-lauroylsarcosine, 2.5 mM DTT, 0.33% SDS, 5X Denhardt's, 5 mM EDTA, 1X protease inhibitor cocktail (Roche), 100 U ml^{-1} SUPERasIN at room temperature. The hybridized material was captured after 3 h of incubation with 240 ml MyOne streptavidin beads (Invitrogen), washed sequentially once with 30 mM HEPES pH 7.5, 240 mM NaCl, 2 M urea, 1.5 mM EDTA, 0.75 mM EGTA, 0.65% SDS, 0.75% N-lauroylsarcosine, four times with 10 mM HEPES pH 7.5, 250 mM NaCl, 2 mM EDTA, 1 mM EGTA, 0.2% SDS, 0.1% N-lauroylsarcosine and once with RNase H elution buffer (50 mM HEPES pH 7.5, 75 mM NaCl, 0.125N-lauroylsarcosine, 0.5% Triton X-100, 0.5 M urea, 10 mM DTT), and eluted by 10 ml RNase H (5 U ml^{-1}, New England Biolabs) digestion in 100 ml RNase H elution buffer for 10 min at room temperature. Eluent was subjected to crosslink reversal by treatment with SDS (1% final), proteinase K (1 mg ml^{-1} final) and Tris pH 7.5 (100 mMfinal) and heating for 1 h at 55 μC and 1–3 h at 65 μC. The enriched DNA was purified using the Qiagen

PCR purification kit per manufacturer's instructions. Prior to preparing libraries for high throughput sequencing, CHART-enriched DNA was treated with RNase cocktail (Roche) and further sheared to below 500 base pairs (Covaris E210 sonicator, at 5% duty cycle, 200 bursts per cycle, intensity 5 for 4 min total process time).

qPCR analysis and validation. The Xist CHART enrichment at several DNA loci was determined using real-time PCR (Bio-Rad iTaq Universal SYBR Green Supermix) under standard conditions for both conventional and allele-specific PCR using the primers listed in Extended Data Table 2. Enrichment values were calculated as 2DCtrelative to input. The real-time PCR experiments were from biologically independent CHART samples (that is, not the samples used for CHART-seq) as independent confirmation of the sequencing results (Extended Data Figs 2e, 4d).

Library preparation, replicates and sequencing. Sequencing libraries were either constructed by standard ChIP-seq protocols by the Yale Center for Genomic Analysis (YCGA), or as described previously33. Briefly, sequencing libraries were prepared by first repairing DNA-ends, A-tailing, ligating to universal adapters, and amplifying for 12 cycles with indexed primers. Excess adapters were removed by purification with Agencourt AMPureXP beads (Beckman Coulter) before sequencing. Sequencing was performed at the YCGA on Illumina HiSeq 2500 instruments. To confirm Xist distribution on the X chromosome, we produced biological replicates for d0, d7, MEF, and technical replicates for d3 and LNA knockoff experiments. Except for d0 where Xist CHART-seq showed mostly background signals, all replicates showed excellent positive correlation (Pearson's $r < 0.9$, Extended Data Fig. 2d). In addition, the replicates of ES d3 and d7 confirmed the specific enrichment in early domains and depletion at late domains. For LNA knockoff experiments, recovery profiles of Xist were confirmed by reCHART replicates of LNA-4798 at 3 h and 8 h. We further confirmed our LNA knockoff results with a time course for LNA C1, which targets a different sequence within the repeat C region of Xist (see Fig. 3 and

Figs 2.10 and 2.10).

Identification of early and late Xist domains. (Note: this subsection is an addendum to the published article) After observing that Xist d3 and d7 samples are much more correlated to each other (Pearson's $r = 0.92$) than to MEF (Pearson's $r = 0.71$, 0.69, respectively), we wanted to distinguish X chromosomal regions that were covered by Xist initially (by d7) from those covered only after full X inactivation (in MEFs). Using the same methodology as above, we called regions of 10x median-corrected enrichment of MEF Xist signal over d7 "late" domains and all other regions "early" domains. To assess whether Xist recovery in MEF cells treated with LNAs was similar to Xist establishment, we compared Xist density between samples in both early and late domains, normalized for the median value of the early domains for each sample (Fig. 2.3g, Fig. 2.10c). Differences in Xist late domain signal between samples were assessed using the Wilcoxon rank sum test.

RNA-seq library. RNAs greater than 200 nucleotides from d7 cells and MEFs were purified using the mirVana RNA extraction kit (Ambion), cleared of ribosomal RNA (Ribozero, Epicentre) and sheared to a median size of 200 nucleotides using the Covaris S2 sonicator. After treatment with T4 polynucleotide kinase, a commercial 5′ adenylated linker (miRNA Cloning Linker 1, IDT) was ligated to the 3′ end of RNAs using T4 RNA ligase 2 (truncated, NEB) followed by reverse transcription (SuperScriptIII, Invitrogen) using a primer (CCGATCTATTGATGGT GCCTACAG) matching the linker. After reverse transcription, RNA was hydrolysed in 10 mM Tris pH 10, 5 mM MgCl2 at 95 uC for 15 min and cDNA products greater than 100 nucleotides size selected and purified on AMPureXP. A barcoded (NNNNNN) 59 phosphorylated linker (GATCGGAAGAGCACACGTCTGAAC TCCAGTCACCNNNNNNATCTCGTATGCCGTCTTCTGCTTGddC) matching Illumina adapters was ligated to the 3′ end of the cDNA using T4 RNA ligase 1 (NEB) and directly amplified using custom forward (AATGATACGGCGACC ACCGAGATCTA-

CACTCTTTCCCTACACGACGCTCTTCCGATCTATTGAT GGTGCCTACA∗G) and reverse (CAAGCAGAAGACGGCATACGA∗G) primers (∗ denotes phosphorothioate bond at 3′ terminal nucleotides), matching the Illumina TruSeq primers. Sequencing of purified libraries was carried out on an Illumina HiSeq instrument for either paired or single-end 50 nucleotides reads. Reads were aligned allele-specifically to 129S1/SvJm (mus) and CAST/EiJ (cas) genomes using Tophat2 [Kim et al., 2013] with the "b2-sensitive" preset and otherwise default parameters (further described in 'Allele-specific alignments'). After removal of PCR duplicates, all unique reads mapping to gene bodies were summed for cas, mus and comp tracks. Read numbers over genes in the allelic tracks were used to calculate skew ($\frac{mus-cas}{mus+cas}$) and genes skewed significantly ($P < 0.01$, cumulative binomial probability) in d7 cells and MEFs were plotted (Fig. 2.1c and Extended Data Fig. 2.8a). A skew of -0.5 corresponds to three-fold difference in inferred expression between Xa and Xi, equal to 67% inactivation of the Xi gene.

LNA displacement. LNAs synthesized by Exiqon were introduced into mouse embryonic fibroblasts (MEFs) as previously described [Sarma et al., 2010]. Briefly, $2 * 10^6$ cells were resuspended in 100 μl MEF nucleofector solution with LNAs at a final concentration of 2 μM and nucleofected using a T-20 program. Fresh culture medium was added to the cells and they were collected and formaldehyde crosslinked at the time points indicated.

Allele-specific alignments. Paired-end sequencing data from CHART-seq were aligned allele-specifically as previously described [Pinter et al., 2012]. Briefly, each data set was aligned to variant CAST/EiJ and 129S1/SvJm genomes constructed using high quality polymorphisms [Keane et al., 2011] to the C57/Bl6 reference genome (mm9 build). Pairs aligning to only one variant genome and pairs aligning better to one variant genome (in number of nucleotide edits to reference) than the other were retained (allele-specific), as were pairs aligning equally well (non-allele-specific). Only

unique pairs were used for this analysis and approximately half of all pairs in the CHART-seq data provided allelic information.

Generation of normalized coverage tracks. Xist CHART-seq reads were filtered for quality and repetitive alignments; low-quality alignments and duplicate reads were removed. The resulting files were analysed using SPP software [Kharchenko et al., 2008]. In this analysis, all tags were included and the coverage generated with smoothing using 1-kb bins every 500 bp to generate input-subtracted, normalized read densities. To account for different read depths across data sets, each coverage file was scaled using the total positive read density on an autosome (chromosome 4) from the corresponding composite track (that is, the mus, cas and comp tracks were all scaled using the same factor). These data were visualized with either the IGV [Thorvaldsdottir et al., 2013] or UCSC genome browser [Kuhn et al., 2007] displaying all tracks using a mean windowing function and scales indicated in each figure. We note that other methods to generate normalized coverage files, including the generation of conservative enrichment and maximum likelihood estimates, resulted in similar distribution patterns, but did not aid in comparisons across data sets with diverse read depths. To determine regions of Xist recovery after LNA treatment, in addition to the normalization and analysis described above, we separately subjected data from 8 h LNA-4978 treatment and 8 h LNA-C1 (time points with partial recovery of Xist density) to normalization using 3 h LNA-4978 and 1 h LNA-C1 reads, respectively (the time point where most of Xist has been removed from the chromatin). The recoveries were determined using the maximum likelihood estimate function in SPP (Fig. 2.6b, Extended Data Fig. 2.10b, indicated by "MLE").

Identification of Xist-enriched segments. Significant segments of Xist enrichment were determined using Epicentre software [Huang et al., 2011] using a whole-genome semidynamic 5-kb window scan, and selecting only windows with P , 0.001 from an exact-rate ratio test. The distribution of the overlap between enriched seg-

ments and genomic features (Figure 2.5h) was determined using CEAS [Ji et al., 2006].

Meta-site and meta-gene analysis. Smoothed normalized read density for ChIP and CHART experiments generated as described above, (with the exception that these were calculated with 2,000-bp windows recorded every 50 bp to avoid aliasing difficulties) were used to calculate average density profiles using CEAS software [Ji et al., 2006].

Definition and significance estimation of allelic skew. The definition of allelic skew is based on the distribution of unique fragments (excluding PCR duplicates) in the allele-specific experimental and input tracks. Allele-specific coverage tracks were queried for a given interval and the cumulative binomial probability estimated by normal approximation from the number of effective fragments based on the interval length and the median sequenced fragment size (200 bp). Skew is then defined as ranging from -1 (fully cas) to -1 (fully mus) and as shown in Figures 2.1c and 2.8a. For example, a threefold difference between alleles is expressed as a skew of ($\pm$) 0.5.

Correlation analyses and significance estimates. Coverage densities over all tracks were tabulated for non-overlapping 40-kb, 200-kb and 1-Mb bins across the chromosome. For analysis and presentation of 40-kb bins that display differential enrichment between samples (Figures 2.1d,e and 2.1e,f, 2.10 and 2.11), bins were identified where the average median-normalized sum was greater than tenfold enriched. These enriched bins were displayed as coloured data on log2 enrichment plots. The location of these bins were identified and displayed as a separate browser track using the intensities from the enriched coverage file (Figs 2.1d,e, 2.3c,d, and 2.11). Correlations between Xist CHART samples were evaluated using Pearson's r and presented as scatterplots or heat maps. Pairwise comparisons of Xist, EZH2, H3K27me3 and H3K4me3 coverage densities were also used for linear regressions shown in scatter plots (Fig. 2.2b). For correlation of Xist seg-

ment density maps with various genomic features (Fig. 2.1h), the density of each feature was calculated for the same 40-kb, 200-kb and 1-Mb bins and Pearson'sr was determined for all pairwise comparisons and displayed as a heat map (Figure 2.8b). Dendrograms are shown where hierarchical clustering was performed based on distance matrices, and the resulting clusters were consistent across a range of bin sizes. Genomic features annotated in the mm9 reference were obtained via the UCSC table browser and included genes (RefSeq), repeats (RepeatMasker), GC%, CpG islands and conservation [Davydov et al., 2010]. In addition, peaks of DNase hypersensitivity [John et al., 2011], replication timing [Hiratani et al., 2008], lamin-association [Peric-Hupkes et al., 2010] and HiC data from the Xist locus viewpoint [Dixon et al., 2012] were queried for correlation with Xist segment density maps in this fashion. LINE1s were queried using all annotations (Figure 2.1h, Pearson's $r = -0.54$ with Xist CHART d7 ES) or single, merged annotations ($r = -0.56$). Moreover, repetitive sequences were aligned to the full murine RepeatMasker database and fraction of all hits compared between Xist d7 cells and corresponding input sample (Figure 2.8c).

Comparison of Xist spreading patterns. To assess the degree of Xist spreading and compare spreading patterns across samples, we focused on chromosomal regions where fully differentiated MEF cells had tenfold greater Xist signal (defined above) than d7 cells, which represent the intermediate stage of Xist spreading. We refer to these regions as "late" domains, and to other chromosomal regions as "early" domains. For each sample, we normalized both early and late domain signals to the median of the early domain signal.We then plotted these normalized early and late domain signals (Figures 2.3g, 2.10c), and evaluated differences between samples were assessed using the one-sided Wilcoxon test.

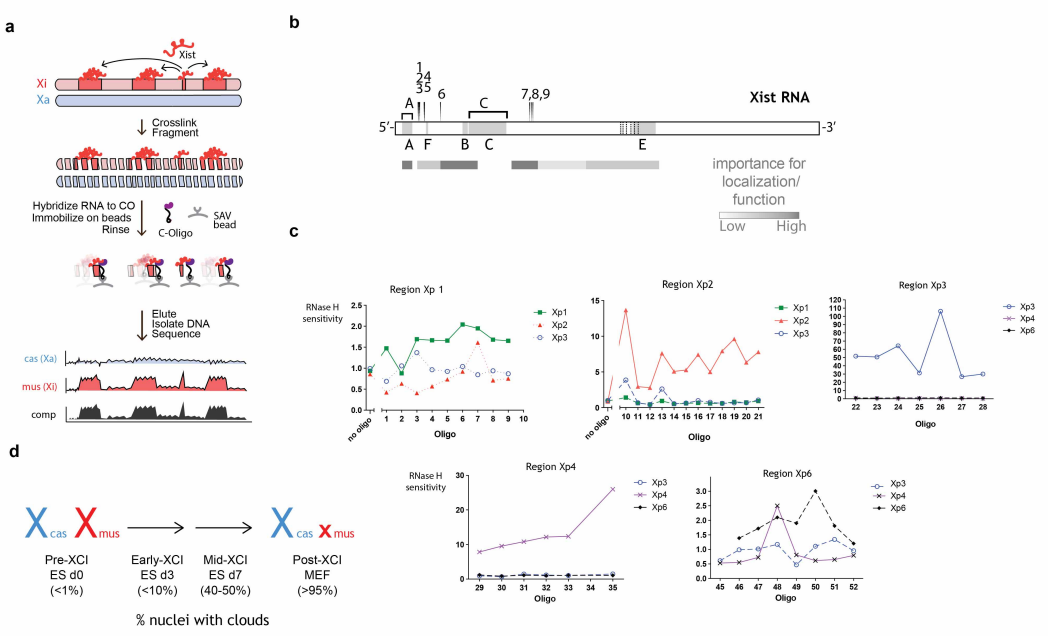

Figure 2.4: Mapping genome-wide distribution of Xist RNA at different stages of XCI using CHART-seq.

a, Experimental scheme for allele-specific analysis of Xist localization by CHART-seq. CO and C-Oligo, capture oligonucleotide; SAV, streptavidin; cas, CAST/EiJ; mus, 129S1/SvJm; comp, composite tracks. **b,** Carets above Xist schematic indicate target sites of the capture oligonucleotides (labelled 1–9, A, C; see Extended Data Table 1 for sequences). Letters below indicate the location of repeat sequences. XCI activity defined previously12. **c,** Sites available for capture-oligonucleotide hybridization were determined by RNase H mapping candidate regions of Xist RNA. The RNase H sensitivity of Xist RNA in the presence of various short DNA oligonucleotides (see Extended Data Table 1) was measured by qRT–PCR, compared to a no-oligonucleotide control and to other amplicons of Xist that are not expected to be affected by cleavage. Primers Xp1–6 are defined in Extended Data Table 2. Regions Xp1 and Xp6 demonstrated minimal sensitivity, but regions Xp2, Xp3 and Xp4 demonstrated broad sensitivity and were used to design capture oligonucleotides for CHART. **d,** Scheme for time- course allele-specific analysis in genetically marked cell lines. Approximate fractions of Xi-positive cells defined by Xist RNA-FISH or H3K27me3 staining.

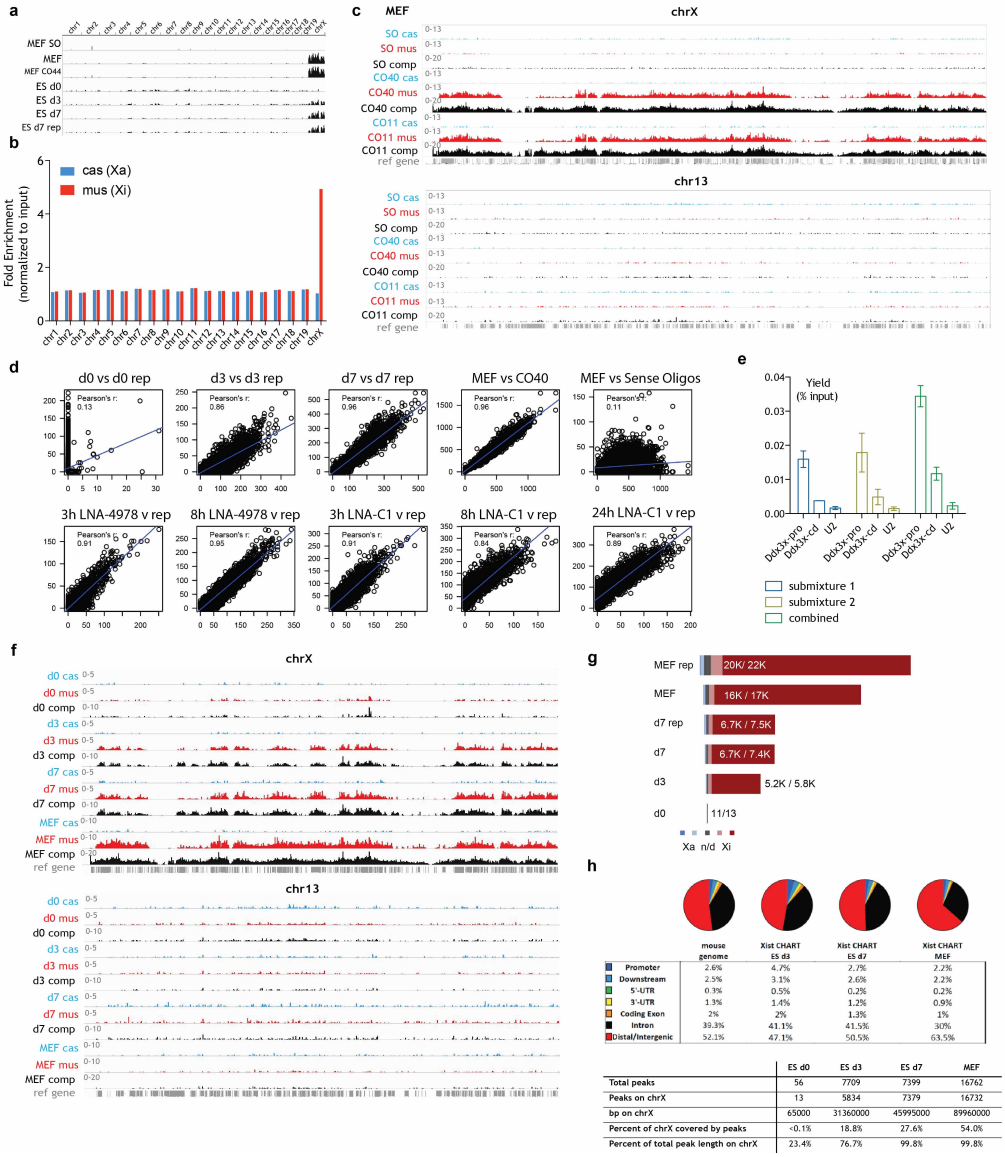

Figure 2.5: Validation and analysis of Xist CHART-seq enrichment.

Validation and analysis of Xist CHART-seq enrichment. a, The genome-wide density of seven input-normalized CHART-seq data sets used in this study based on comp reads. **b,** Allele-specific enrichment on each chromosome based on raw aligned reads relative to input (MEF). **c,** Allele-specific enrichment of control Xist CHART-seq experiments in MEF, including SO (sense oligonucleotide control) and CO40 (CHART-seq performed with alternate mix of 40 capture oligonucleotides, see Methods and Extended Data Table 1), both presented in comparison to CO11 (Extended Data Table 1, capture oligonucleotides used throughout study). Data shown for the X chromosome and a representative autosome (chromosome 13). **d,** Linear correlation analyses of Xist CHART-seq data sets, including LNA-treated samples, using the comp track showing high reproducibility. Pearson's r correlation coefficient indicated. Replicates were either biological (d0; d7; MEF) or based on replicate CHART experiments (ES d3; LNA-4978 3 h; LNA-4978 8 h; LNA-C1 3 h). **e,** Two independent sub-mixtures of capture-oligonucleotides confirm Xist CHART-seq enrichment patterns by qPCR from an independent Xist CHART experiment in MEF cells. Sub-mixture 1 is composed of capture-oligonucleotides X.1, X.3, X.5, X.7, X.9, X.A; sub-mixture 2 is composed of capture-oligonucleotides X.2, X.4, X.6, X.8, X.C (for primer locations see Figure2.7 and Extended Data Table 2). f, Allele-specific analysis of d0, d3, d7 and MEF similar to that presented in Fig. 2a. **g,** Allelic breakdown of enriched Xist segments with grey (n/d, not determined due to lack of SNPs), light blue and red (leaning towards Xa or Xi, respectively), and dark colours for significantly skewed towards Xa (blue) or Xi (red), as defined by cumulative binomial probability ($P < 0.05$) after normal approximation from effective fragments. **h,** The locations of the enriched regions compared to the mouse genome (mm9) and the overlap determined for various genomic features. Below, table summarizing peak numbers and chromosomal origin and coverage (in bp, or per cent chromosome length).

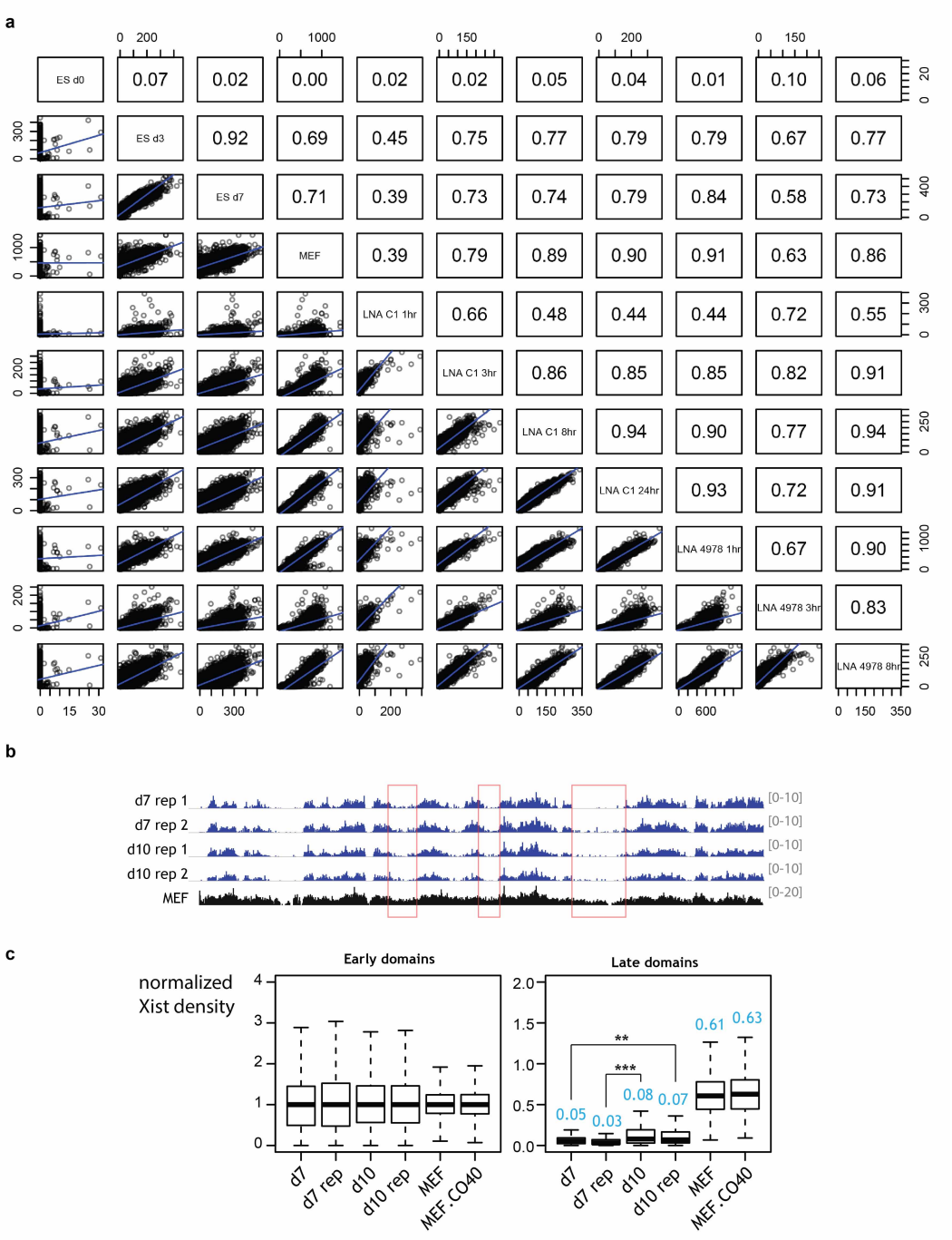

Figure 2.6: Correlation analyses of CHART-seq data sets.

Correlation analyses of CHART-seq data sets.

a, Scatter plots (below diagonal) of 40 kb-binned Xist CHART-seq comp signals across all pair-wise comparisons. Pearson's r correlation coefficients are shown in corresponding squares above the diagonal. **b,** Overview of Xist spreading during XCI. Comp tracks of Xist CHART-seq signals of d7 and d10 replicates (blue), MEF (black). **c,** Box plot of normalized Xist densities (40-kb bins) at early and late domains. Data were processed as in Fig. 2.3g and Fig. 2.10c. Normalized median values for each sample are indicated above box. $**P < 10^{-4}; ***P < 10^{-6}$, one-side Wilcoxon test. The median Xist densities in late domains relative to early domains increase moderately from d7 to d10.

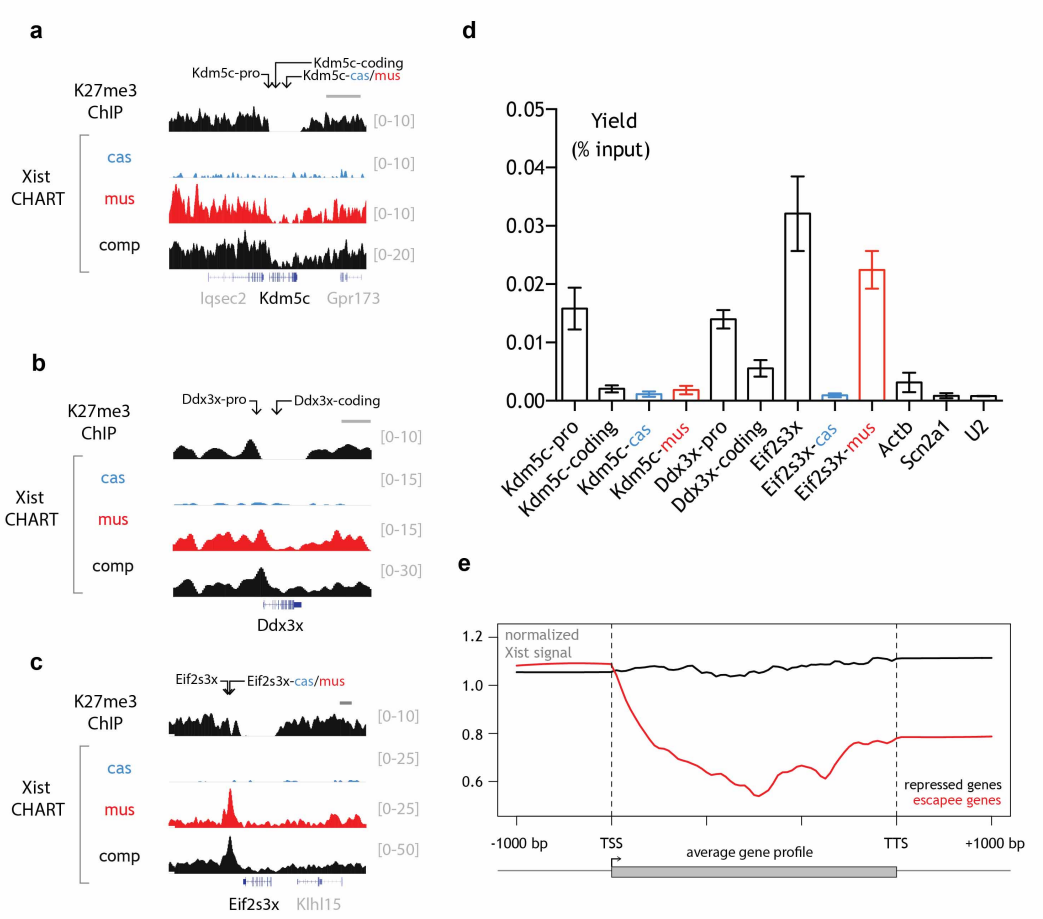

Figure 2.7: The gene bodies of escapees are depleted of Xist, but are often near peaks of Xist enrichment.

a–c, Xist distribution at Kdm5c, *Ddx3x* and *Eif2s3x* genes that escape X inactivation in MEF. cas (Xa), mus (Xi) and comp tracks of Xist and comp track of H3K27me3 ChIP-seq are shown. **d,** qPCR validation of Xist enrichment. Locations of qPCR amplicons are indicated in **a–c**. pro, promoter; coding, coding region; cas, cas (Xa)-specific; mus, mus (Xi)-specific. Allele-specific qPCR results shown in red/blue (mus/cas). Autosomal active and inactive genes were used as negative controls (*Actb, Scn2a1, U2*). Yields determined relative to input DNA. Consistent with CHART-seq results, promoter regions of *Kdm5c* and *Ddx3x* showed higher Xist signal than corresponding coding regions. Xi-specific enrichment of Xist was only observed at the 3′ region of *Eif2s3x*, but not at the coding region of *Kdm5c*. **e,** Metagene analysis of Xist density across XCI-repressed and escapee genes. Normalized composite density from Xist CHART using post-XCI (MEF) cells was smoothed (2,000-bp windows, sampled every 50 bp using SPP software) and averaged across genes on the X that are either repressed (black, defined by those that are active on the Xa but not on the Xi in MEF cells) or escape XCI (red, excluding escapee genes at the Xic). Repressed and escapee genes were determined previously. Profiles calculated using the CEAS softwarepackage with default settings.

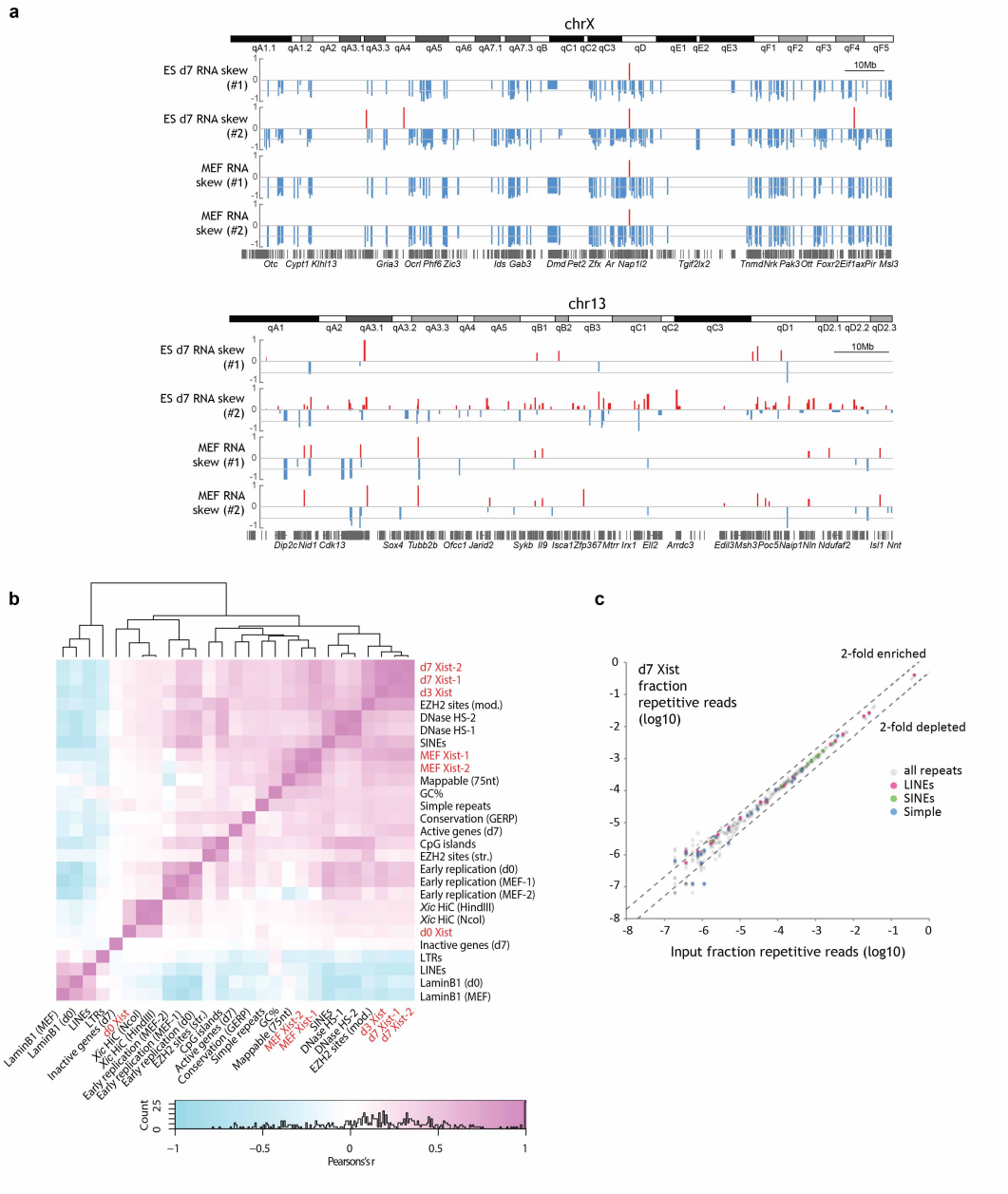

Figure 2.8: Xi-wide gene repression patterns in d7 and MEF cells, and the relationship of Xist establishment domains with various chromatin features of the X-chromosome.

Xi-wide gene repression patterns in d7 and MEF cells, and the relationship of Xist establishment domains with various chromatin features of the X-chromosome.

a, RNA-seq reads aligned allele-specifically were tabulated over gene bodies of RefSeq genes (indicated below in grey). Skew in allele-specific reads (mus-cas/mus+cas) is plotted on a range of -1 (expression fully cas-linked) to $+1$ (expression fully mus-linked). Bar chart shows allelic skew (red$>$ 0, blue$<$ 0) values over gene bodies for all genes that were significantly skewed (cumulative binomial probability). Grey lines indicate midpoint (skew = 0) for balanced expression between alleles, and -0.5, signifying threefold depletion of the mus-allele and amounting to 67% inactivation. Two replicates each are shown for d7 and MEFs. Analysis here is similar to Fig. 1c. **b,** Chromosomal organization directs Xist enrichment. Correlation matrix at 200-kb resolution, featuring significantly enriched Xist segments across XCI time course (d0, d3, d7 and MEF Xist), major repeat classes (SINEs, LINEs, LTRs, simple repeats), active and inactive genes (based on calls in d7 cells), strong and moderate EZH2 binding sites, CpG islands, CG%, conservation (GERP), DNase hypersensitivity (DNase HS), early replication timing (in male (1) and female (2) MEF and male d0 cells), Lamin B1 association (in male ES d0 and MEF), and HiC interaction frequencies in male d0 with the Xist locus (*Xic* HiC) using two restriction digests (HindIII, NcoI). Colours for positive (magenta) and negative (blue) correspond to Pearson's r values. See Methods for references to source data. c, Repetitive sequences including LINEs, SINEs and simple repeats are not significantly enriched or depleted from d7 Xist CHART DNA compared to input. Repetitively aligning reads excluded from the other analyses were re-aligned to the entire library of known repeat elements in the mouse genome (http:// www.girinst.org/repbase/). Hits in Xist CHART in d7 cells were compared to their corresponding input samples. All repeat types (grey), LINEs (pink), SINEs (green) and simple (blue) repeats are shown. Dashed lines rep-

resent twofold enrichment or depletion. The results show no enrichment of LINEs in repetitively aligning reads in the Xist CHART relative to input.

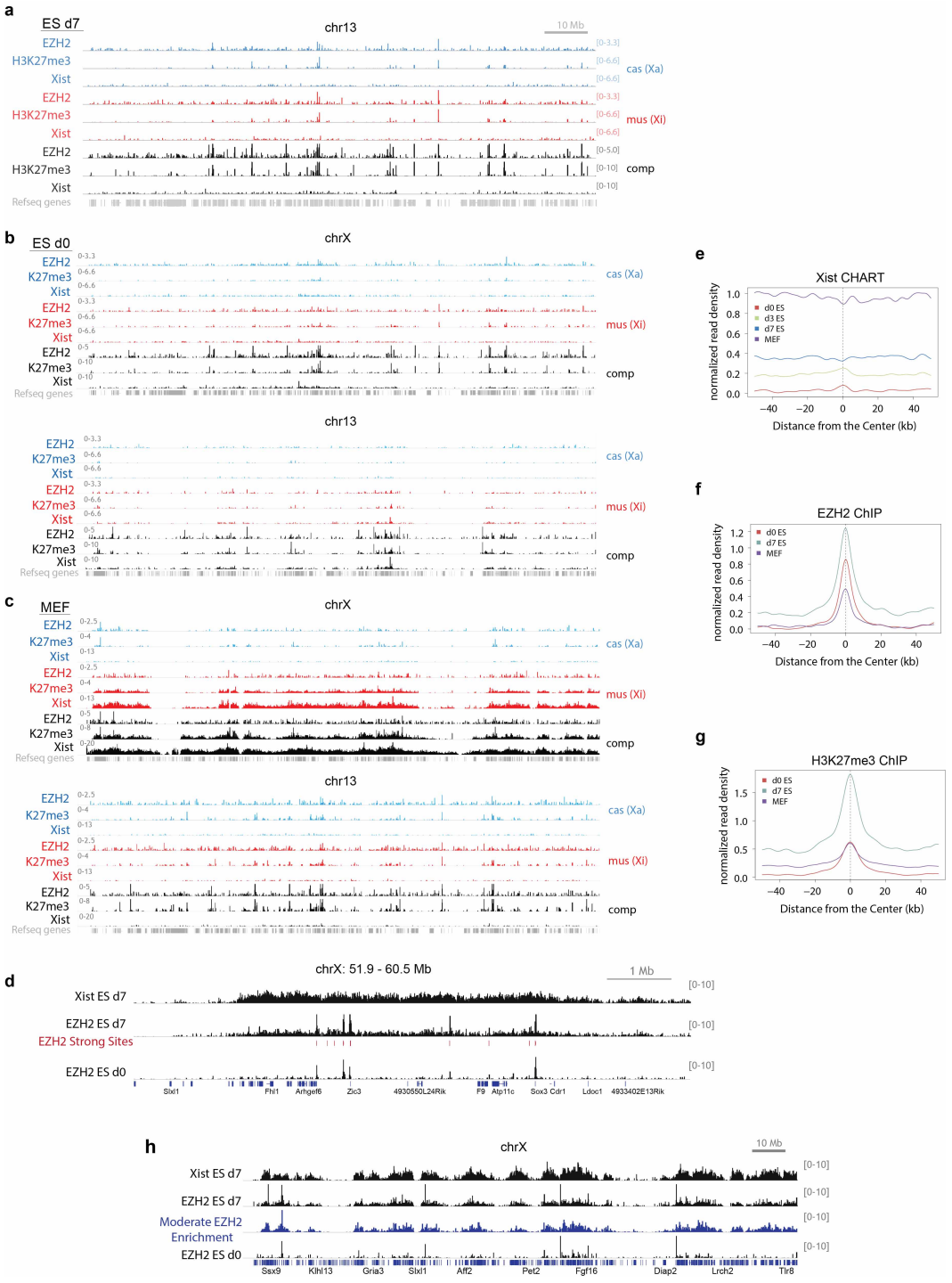

Figure 2.9: Xist binding correlates with previously identified moderate EZH2 sites.

Xist binding correlates with previously identified moderate EZH2 sites.

a, Normalized read densities of Xist, EZH2 and H3K27me3 on chromosome 13 in d7 cells shown as in Fig. 2a. **b, c,** Overview of the correlation of Xist RNA with PRC2 and H3K27me3 on the X chromosome and chromosome 13 in d0 **(b)** and MEF **(c)**. Xa and Xi allele specific and composite (comp) tracks for Xist, EZH2 and H3K27me3 are displayed as in Fig. 2a. **d,** Strong EZH2 sites have above average ChIP-seq density compared to the broad EZH2 signals on the X in d7. Comp tracks are shown. Many of the strong EZH2 sites are present before XCI in d0 cells, therefore PRC2 can bind these sites independently of Xist. **e–g**, Meta-site analysis of the average EZH2, H3K27me3 and Xist signals around strong EZH2 sites identified in d7 (ref 9). The strong enrichment of H3K27me3 signals at strong EZH2 sites are in agreement with the strong correlation of EZH2 and H3K27me3. **h,** The density plot of moderate EZH2 enrichment sites (blue) is consistent with the broad distribution of EZH2 on the X, and correlates with Xist in d7. Comp tracks are shown.

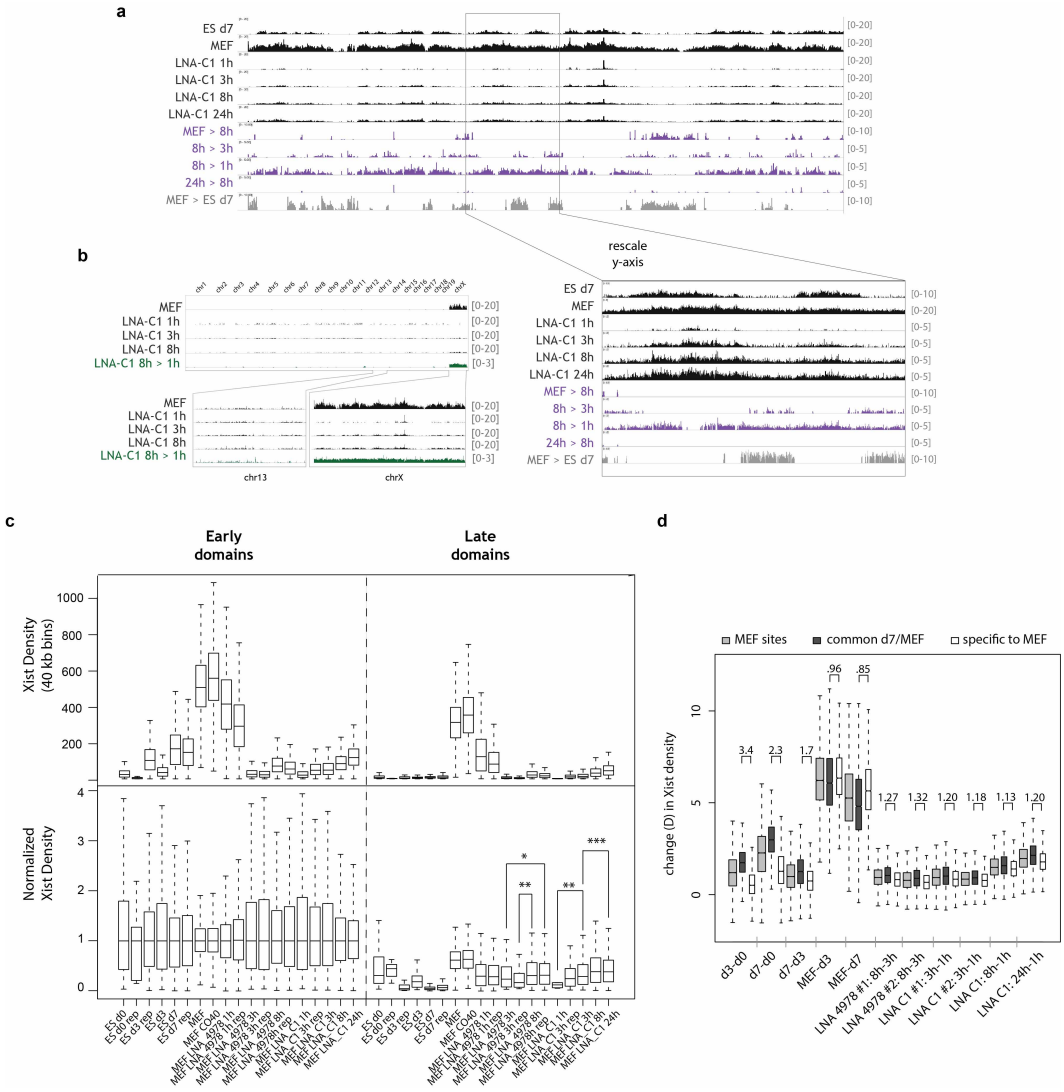

Figure 2.10: An independent LNA confirmed the chromosome-wide re-spreading of Xist.

An independent LNA confirmed the chromosome-wide re-spreading of Xist. **a,** Overview and zoom-in (bottom) of differential Xist density in MEF cells subjected to LNA-C1 treatment. Comp tracks of Xist CHART-seq signals on the chromosome X of indicated cells are shown in black. Differential regions showing >tenfold enrichment are displayed in purple or grey as in Fig. 3c, d. **b,** Genomic distribution of normalized Xist CHART densities in comparison with a maximum likelihood enrichment estimate of LNA-C1 8 h over LNA-C1 1 h CHART-seq signals (LNA-C1 8 h > 1 h, green), showing broad, chromosome-wide recovery of Xist on the X in comparison to an autosomal control. **c,** Significant increase of Xist density within late regions was observed in MEFs recovering from LNA treatment. Top, Xist density changes during XCI establishment in ES cells and in MEFs before and after LNA treatment. Boxplots of 40-kb-binned Xist CHART-seq signals of early and late domains. During ES differentiation, increased Xist density was observed within early domains where genes are enriched, but remained at low levels in late domains where gene densities are low. After LNA treatment, MEFs showed reduced Xist signals within both domains, indicating global loss of Xist coverage and partial recovery at later time points on chromosome X (LNA-4978 8 h and LNA-C1 3 h and 8 h, compared to LNA-4978 1 h and LNA-C1 1 h, respectively). Bottom, Xist recovery in indicated samples, with 40-kb-binned Xist densities normalized to median levels of early domains of each sample to determine how early and late domains recover from LNA knockoff as compared to levels found during *de novo* XCI. Normalized median values for each sample indicated above box. *P < 0.05; **P < 0.005; ***P < 0.0001, one-sided Wilcoxon test. Two independent LNAs consistently showed significant Xist recovery in late domains within hours post LNA treatment. **d,** Pattern of Xist recovery after LNA treatment with LNA-4978 and LNA-C1. Xist enriched segments (segs) in MEFs (grey, 16,760 total) were split into those common to both MEF and d7 cells (dark grey, 8,910 total) and those specific to MEFs (white, 7,850 total). Changes

in Xist density over these sites are shown for d3 - d0, d7 - d0, MEF - d7 and LNA samples (LNA-4978: 3 h stripped, 8 h recovery; LNA-C1 1 h stripped, 3 h, 8 h, 24 h recovery). Replicates indicated with #1/#2. Numerical fold-difference in median of changing Xist density between MEF-specific segs and common-segs indicated above box-plots. After LNA treatment, recovery of Xist density over MEF-specific enriched segs is close to that of common segs (only $1.2 - 1.3\times$ lower), whereas during XCI establishment Xist increase over these sites is $3.4\times$ and $2.3\times$ higher in d3 - d0 and d7 - d0, respectively. These values are highly reproducible between replicates. Width of notched box plots scaled to square root of total number of enriched segs in each group. Error bars indicate $1.5\times$ interquartile range without extending beyond $\frac{min}{max}$ data points.

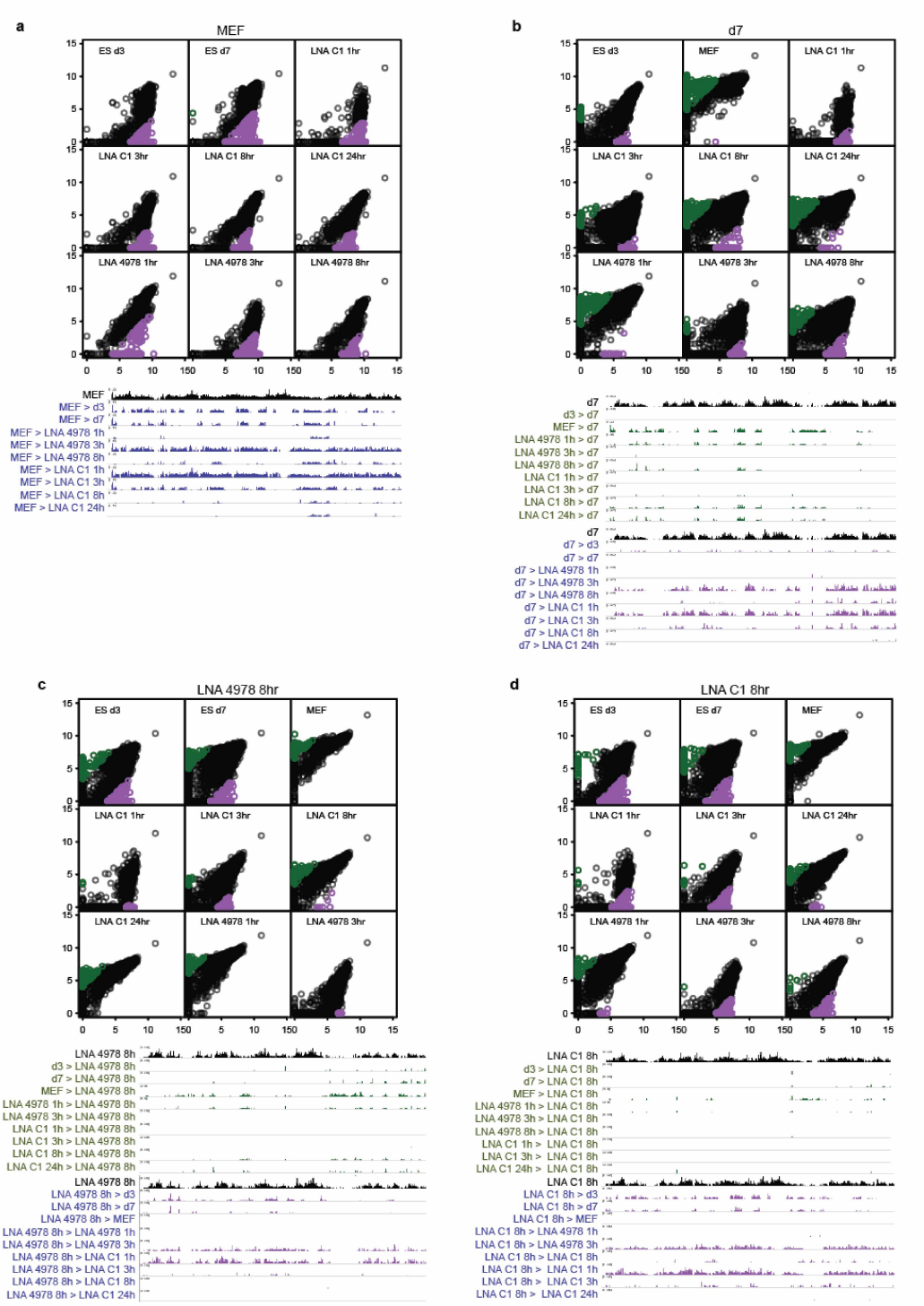

Figure 2.11: Comparison of Xist distribution post-LNA treatment with establishment and maintenance stages of XCI.

Comparison of Xist distribution post-LNA treatment with establishment and maintenance stages of XCI.

a–d, 40-kb- binned Xist CHART-seq data from comp tracks are plotted on a log_2 scale. Bins with tenfold enrichment or depletion (median corrected) of one sample versus another are coloured in purple and green, respectively. These regions of difference between samples were mapped along the X chromosome by plotting the CHART-seq signal of the enriched sample. Complete CHART-seq tracks are shown in black. Comparisons are centred on maintenance **(a),** *de novo* establishment **(b)** and recovery from LNA treatment in post-XCI cells, with knockoffs using two independent LNAs showing similar results **(c, d)**.

Chapter 3

Tracking Distinct RNA Populations Using Efficient and Reversible Covalent Chemistry

3.1 Summary

The chapter below describes collaborative work, which was led by Dr. Erin Duffy, to identify a better reagent for biochemical purification of 4-thiouridine labeled RNA. I led the computational analysis for this project, which produced a variety of specialized RNA-Seq datasets, in collaboration with Dr. Duffy and Dr. Robert R. Kitchen, under the supervision of Drs. Mark Gerstein and Matthew D. Simon. Additional experimental work for this study was conducted by Catherine D. Stark. Below, I describe my role in this project, referencing figures from the published journal article related to this work. I then reproduce (with permission) the full journal article journal article, with slight modifications:

Duffy, EE, Rutenberg-Schoenberg, M, Stark, CD, Kitchen, RR, Gerstein, MB, Simon, MD **(2015)**. Tracking Distinct RNA Populations Using Efficient and Reversible

Covalent Chemistry. *Mol. Cell*, 59, 5:858-66.

I also list all figures within this paper that I drew directly and those that were drawn using products of my analysis.

3.2 Description of indepedent work within collaboration

Metabolic labeling of RNA is a set of techniques, in which a modified nucleoside or nucleobase is fed to cells and is incorporated as a label into newly transcribed RNA. Metabolic labeling has been instrumental to the study of RNA dynamics and can help to distinguish gene regulation that occurs at the level of RNA stability or degradation from the more commonly studied regulation of transcription. To make studies of RNA dynamics using metabolic labeling possible dynamics using metabolic labeling, labeled RNA must be quantified specifically and compared to total RNA levels. This is commonly achieved by biochemical purification using specific properties of the labeled nucleotides.

One of the most common metabolic labels used to study RNA dyanmics is 4-thiouridine (s^4U). The central topic of this study is development of more efficient reversible chemistry to purify s^4U-labeled RNA for applications in the study of RNA dynamics. In previous work, s^4U-labeled RNA had been captured by reaction with HPDP-biotin, followed by purification on streptavidin beads. In this study, Dr. Erin Duffy showed that the pyridylthiol group in HPDP-biotin reacts inefficiently with s^4U, whereas a different reagent–methanethiosulfonate–reacts much more efficiently (>95% vs. <20% efficiency, Figure 3.1). Based on this finding Dr. Duffy applied MTS chemistry to metabolic labeling experiments in cultured cells, which were read out by RNA-Seq.

The first question we wanted to address was whether the increased efficiency of

MTS chemistry compared to HPDP chemistry for reaction with s^4U RNA would lead MTS-biotin to be a more efficient than HPDP-biotin for capture of s^4U-labeled RNA in cultured cells. Whereas most RNA-Seq experiments are interpreted in terms of relative quantities of different RNA molecules, this question requires absolute measurement of the proportion of a given RNA that is enriched by disulfide capture. To faciliate this analysis, we used a similar approach to Sun *et al.* (2012) and spiked in *S. pombe* RNA as a standard against which to compare the total quantity of human RNA (from cultured 293T cells) in our samples.

To map RNA-Seq reads from both s^4U-enriched and input samples, I constructed a joint genome containing both human (hg19) and *S. pombe* chromosomes. I then aligned reads using TopHat2 [Langmead and Salzberg, 2012]. I then quantified gene expression at both the gene and transcript levels using Cufflinks [Trapnell et al., 2010].

By quantifying the ratios of total mapped reads in the human and *S. pombe*, I was able to see that our MTS-bioitin enriched samples had a much higher proportion of human reads than HPDP-enriched samples, but also very low levels of human reads in samples that lacked s^4U (Figure 3.1B). Our next goal was to normalize genome-wide alignment tracks and quantifications of gene expression based on the *S. pombe* spikes. Because there is potential for cross-mapping of human and *S. pombe* reads to regions of the other genome that are conserved, we reasoned that using RNA quantifications from Cufflinks, which employ expectation maximization to reassign reads that map ambiguously based on information from specifically mapping reads, might help alleviate this problem. We thus normalized our human gene quantifications as follows:

$$FPKM_{norm} = FPKM_{raw} S_{norm}$$

where $FPKM_{norm}$ is the normalized FPKM of a human transcript or gene, $FPKM_{raw}$

is the original FPKM calculated for the sample of interest, and S_{norm} is the slope of the linear regression line of raw S. pombe gene FPKMs.

I took a similar approach to normalize genome-wide coverage tracks, with additional normalization to the total number of reads in each sequencing sample:

$$Coverage_{norm} = Coverage_{raw} S_{norm} \frac{R_{sample}}{R_{norm}}$$

where $Coverage_{norm}$ and $Coverage_{raw}$ are the normalized and raw read coverages at a given genomic position, and R_{sample} and R_{norm} are the numbers of unique reads in the sample of interest and the normalizing sample, respectively.

Normalization of coverage tracks enabled me to visualize the fact that the increased quantity of human RNA in MTS-enriched s^4U samples is seen widely across the genome (Figure 3.2C). Meanwhile background in samples with no s^4U is relatively low and uniform.

One previously known issue with RNA-Seq following s^4U enrichment with HPDP-biotin was that long transcripts enriched over short transcripts. This was interpreted to be the result of rare incorporation of s^4U into transcripts, with transcripts containing more uridines having more chances to incorporate s^4U [Sun et al., 2012]. We were thus interested in whether increasing the rate of s^4U biotinylation could help alleviate this bias. Upon examination of the genome-wide coverage tracks that I had created, we noticed that many short transcripts were depleted in HPDP-enriched samples but not in MTS-enriched samples (Figure 3.2E).

I then sought to quantify the difference in length bias between HPDP-biotin and MTS-biotin enriched s^4U RNA more broadly across the transcriptome. Previous work had mostly been conducted in yeast, which have constitutive splicing, making it clear how many uridines are in most transcripts. Alternative splicing in humans makes determining the number of uridines in transcripts more complicated. However,

taking note of work that shows that many genes express a single isoform dominantly in any given tissue or cell line [Gonzalez-Porta et al., 2013], I chose to focus on genes for which a single transcript comprised at least 90% of the total gene expression as quantified by Cufflinks. Focusing on these transcripts, I binned groups of transcripts by the number of uridines and noticed a significant relationship between number of uridines in HPDP-enriched samples but very little difference between length groups for MTS-enriched samples (Figure 3.2D). For further discussion of the relationship between s^4U incorporation, disulfide coupling, and length bias, see the appendix titled "Modeling expected yields of s^4U enrichment in metabolic labeling experiments" below (this is part of the supplemental information from our publication).

Having established that efficient disulfide coupling with MTS enables more effective enrichment of long RNA in cell culture experiments, we sought to apply MTS chemistry to enrich miRNAs. Because miRNAs have very few uridines, they would be very difficult to enrich without efficent disulfide coupling to biotin. Accordingly, the stability of these important regulatory RNAs had only been studied under conditions of transcription inhibition [Guo et al., 2015].

To study miRNA turnover, we used the RATE-Seq approach, feeding 293T cells with s^4U and observing time points toward the approach to equilibrium levels of s^4U incorporation [Neymotin et al., 2014]. To quantify miRNA samples, I used a pipeline developed by Dr. Robert R. Kitchen in the Gerstein lab, using a combination of the software sRNAbench [Rueda et al., 2015] and the Bowtie2 aligner [Langmead and Salzberg, 2012]. This pipeline maps reads in small RNA-Seq experiments hierarchically, first to rRNA, UniVec laboratory contaminants, and long RNA transcriptss and only subsequently to miRNAs and other small RNAs. This approach is helpful for filtering out contaminating reads from sources other than the intended miRNA population.

We had originally hoped to use these experiments to model miRNA half lives.

However, in analyzing quantities of synthetic spike-in RNAs, I found that these values were too noisy to produce robust absolute quantifications. However, I observed that adjacent time points correlate well throughout the RATE-Seq experiment, implying that relative quantities of different RNAs change in a consistent fashion (Figure 3.3C). This motivated analysis of differential expression between an early time point (20 minute s^4U feed) and a late time point (6 day s^4U feed), using edgeR [Robinson and Smyth, 2008, Robinson et al., 2010]. miRNAs that were enriched in the 20 minute time point were interpreted as being fast turnover, whereas RNAs enriched at the later time point were interpreted as slow turnover (Figure 3.3D). Interestingly, many of the miRNAs that were identified as fast turnover were annotated as miRNA-stars, or the less stable of the complementary miRNAs produced during the maturation process (Figure 3.3D). Strikingly, the miRNAs with the most significant differences between early and late time points display consistent changes across the time course experiment, implying that this method helps to assess the relative turnover of different miRNAs (Figure 3.3E).

Though this project was successful in validating that MTS chemistry enables better enrichment of s^4U-RNA, including the first investigation of miRNA turnover in a proliferating system. An outstanding technical question that remained was how to normalize data with spike-in controls. Subsequent experimental work in the Simon lab helped to address this challenge by recoding s^4U to cytosine analogues. This enables enrichment and spike-in free analysis [Schofield et al., 2018].

Within the work described below, I plotted the following figures:

- Figure 3.2 D
- Figure 3.3 C,D,E
- Figure 3.5 B
- Figure 3.6 A-C,E

Additionally, the following figures incorporate analyses that I conducted:

- Figure 3.2 B,C,E
- Figure 3.5 A,C-F
- Figure 3.6 D,F

3.3 Tracking Distinct RNA Populations Using Efficient and Reversible Covalent Chemistry

3.3.1 Abstract

We describe a chemical method to label and purify 4-thiouridine (s^4U)-containing RNA. We demonstrate that methanethiosulfonate (MTS) reagents form disulfide bonds with s^4U more efficiently than the commonly used HPDP-biotin, leading to higher yields and less biased enrichment. This increase in efficiency allowed us to use s^4U labeling to study global microRNA (miRNA) turnover in proliferating cultured human cells without perturbing global miRNA levels or the miRNA processing machinery. This improved chemistry will enhance methods that depend on tracking different populations of RNA, such as 4-thiouridine tagging to study tissue-specific transcription and dynamic transcriptome analysis (DTA) to study RNA turnover.

3.3.2 Introduction

RNA is continuously transcribed and degraded in a tightly regulated and transcript-specific manner. The dynamics of different RNA populations can be studied by targeted incorporation of non-canonical nucleosides. These nucleosides can provide a chemical handle for labeling and enriching RNA subpopulations. The labeling of RNA employs 5-bromouridine (5-BrU; [Tani et al., 2012], 5-ethynyluridine (5-EU;

[Jao and Salic, 2008], and 4-thiouridine (TU or s^4U; [Cleary et al., 2005, Miller et al., 2009]), which provide different vehicles for antibody detection, cycloaddition reactions, and thiol-specific reactivity, respectively. 4-thiouridine holds the advantage that labeling is covalent, unlike the antibody detection of 5-BrU, and also that the disulfide bond is reversible, unlike the click chemistry used to label 5-EU (reviewed in [Tani and Akimitsu, 2012]). Methods to enrich s^4U-incorporated RNA (s^4U-RNA) initially relied on organomercurial affinity matrices (Melvin et al., 1978), but the use of s^4U in metabolic labeling expanded after HPDP-biotin, a 2-pyridylthio-activated disulfide of biotin, was developed as a practical means to biotinylate s^4U-RNA using reversible disulfide chemistry, followed by enrichment using a streptavidin matrix [Cleary et al., 2005, Dolken et al., 2008]. The s^4U-RNAs can be eluted by reduction of the disulfide linkage and subsequently analyzed by microarray, qPCR, or deep sequencing. This modified protocol sparked a surge in techniques that use s^4U metabolic labeling. For example, half-lives of specific RNAs can be measured using s^4U metabolic labeling by quantifying the ratio of pre-existing (flow through) to newly transcribed (elution) RNA [Dolken et al., 2008]. This approach has been extended to genome-wide analysis using high-throughput sequencing (s^4U-seq; [Rabani et al., 2011]).

Combining s^4U metabolic labeling with dynamic kinetic modeling has led to the development of dynamic transcriptome analysis (DTA; [Miller et al., 2011]), and comparative dynamic transcriptome analysis (cDTA) when using S. pombe standards for normalization, which allows the determination of absolute rates of mRNA synthesis and decay [Sun et al., 2012]. Reversible transcriptional inhibition has been combined with s^4U metabolic labeling to measure transcriptional elongation rates [Fuchs et al., 2014]. Recently, s^4U metabolic labeling has been used with approach-to-equilibrium kinetics to determine absolute RNA degradation and synthesis rates based on multiple time points after s^4U labeling (RATE-seq; [Neymotin et al., 2014]).

In addition to these methods for analyzing RNA turnover, the enrichment of s^4U-RNA can also be used to determine cell-type-specific transcription (4-thiouridine tagging), which is particularly helpful for analyzing the transcriptomes of cell types that are difficult to isolate by dissection or dissociation methods [Miller et al., 2009]. As the efficient chemical modification of s^4U is central to all of these techniques, we tested the reactivity of s^4U with HPDP-biotin. Here we report that the reaction and corresponding enrichment of s^4U-RNA with HPDP are inefficient. Therefore, we developed and validated chemistry using activated disulfides to label and enrich s^4U-RNA. This chemistry increases labeling yields and decreases enrichment bias. Due to the increased efficiency of this chemistry, we were able to extend s^4U-metabolic labeling to the study of microRNAs (miRNAs), providing insight into miRNA turnover in proliferating cells without inhibition of miRNA processing pathways. Our studies expand the utility of s^4U in metabolic labeling applications and provide the foundation for clearer insight into cellular RNA dynamics through the improvement of all the methods listed above.

3.3.3 Design

We sought chemistry to enrich s^4U-RNA that satisfied several considerations. First, the chemistry should be efficient, leading to high yields of labeled s^4U residues. To maintain the advantages of reversible covalent chemistry, we focused on activated disulfide reagents, which allow reductive release after enrichment. This labeling chemistry should be rapid, minimizing time required for purification and decreasing RNA degradation during handling. Finally, the chemistry needs to be specific for s^4U and should not react with RNA that lacks thiol groups. These improvements would lead to a more robust protocol for s^4U RNA isolation. Additionally, optimized chemistry could allow the extension of labeling to small RNAs including miRNAs. Smaller RNAs are expected to be particularly sensitive to the efficiency of s^4U labeling, as

they tend to have fewer uridine residues and therefore have lower probability of successful labeling. To develop chemistry that meets the above criteria, we first used simple chemical systems to determine the reactivity of activated disulfides. We studied the specificity of labeling chemistry using synthetic RNA with and without s^4U. We used metabolic labeling experiments together with RNA sequencing (RNA-seq) to test the application of this chemistry in the context of complex RNA samples. Finally, we evaluated the use of this chemistry to study miRNA turnover, revealing fast- and slow-turnover miRNAs in proliferating cells without perturbing miRNA processing pathways.

3.3.4 Results

Optimizing Labeling Chemistry Using Free Nucleosides To examine the reactivity of s^4U-RNA with HPDP-biotin, we first studied the labeling of the s^4U nucleoside using liquid chromatography coupled to mass spectrometry (LC-MS; Figures 3.1A and 3.1B). We found biotinylation of the s^4U nucleoside with HPDP-biotin to be inefficient when using buffer conditions that are commonly used in the retrieval of s^4U-RNA [Gregersen et al., 2014]. This inefficiency stems from the forward and reverse disulfide exchange reactions (Figure 3.1A). Any disulfide formed with the electron-poor pyrimidine ring of s^4U results in a more activated product, therefore favoring the reverse rather than the forward labeling reaction. For this reason, it is not surprising that HPDP-biotin is an inefficient reagent for disulfide exchange with s^4U. Improving this chemistry would expand the utility of s^4U, improve the sensitivity of s^4U labeling, and reduce bias in s^4U-RNA enrichment.

Of the numerous activating chemistries used to make asymmetric disulfides [Jeschke, 2013, Kenyon and Bruice, 1977], thiosulfates and alkylthiosulfonates are particularly attractive (Figure 3.1C). We found that, in sharp contrast to the slow and inefficient reaction with HPDP-biotin, methylthiosulfonate-activated biotin (MTS-biotin) reacts

efficiently with s^4U, leading to >95% conversion to the mixed disulfide within just 5 minutes (Figure 3.1D). We validated this difference in s^4U reactivity between MTS reagents and 2-pyridylthio-activated disulfides using NMR (Figures 3.1E and 3.4A-C). While only a minority of s^4U reacted using 2-pyridylthio chemistry (<20%), MTS chemistry led to >95% conversion of s^4U to the mixed disulfide.

Extending MTS Labeling Chemistry to s^4U-RNA This MTS chemistry could be used to specifically fluorescently label s^4U-RNA in the context of cell extracts (Figure 3.4D). Furthermore, we found that the use of MTS-biotin leads to superior biochemical enrichment of s^4U-RNA in comparison to HPDP-biotin (compare flow through to eluent in Figures 3.1F and 3.1G) or thiosulfate-biotin (TS-biotin, Figures S1E–S1G). Importantly, MTS and HPDP chemistries are specific for enrichment of s^4U, as no significant enrichment of RNA without s^4U occurred in either case (Figures 3.1F,G and 3.4H). We therefore conclude that MTS chemistry provides a specific and highly efficient means of detecting and biochemically purifying s^4U-RNA.

We next tested the efficacy of MTS biotin as a reagent to examine newly transcribed RNA in HEK293T cells (Figure 3.2A). We treated cells with s^4U-supplemented media and reacted the isolated RNA with either HPDP-biotin (as described previously by [Gregersen et al., 2014]) or MTS-biotin. Biotinylated RNA was enriched and then analyzed by RNA-seq. To compare the RNA-seq reads across experiments, we used a normalization approach developed by Sun et al. (2012) in which the same amount of RNA from S. pombe is added to each sample prior to constructing the library for RNA-seq [Sun et al., 2012]. Consistent with our prior analysis, compared to HPDP-biotin, the use of MTS-biotin led to significantly greater normalized coverage of the human transcriptome (Figures 3.2B-C). This enrichment was reproducible across biological replicates (Pearson's $r = 0.92$, Figures 3.5A-D) and was validated by qPCR (Figures 3.5E-F). To test the specificity of MTS chemistry, we examined MTSbiotin-treated RNA from cells that had not been treated with s^4U and found substantially fewer

normalized reads than with either HPDP-biotin or MTS-biotin-enriched s^4U-RNA (Figures 3.2B-C). The result from this control experiment validated the specificity of MTS-biotin for metabolically labeled s^4U-RNA.

Alleviating Length Bias Using MTS-Biotin We next compared the distributions of enriched RNAs using MTS- and HPDP-biotin. Purification of s4 U-RNA using HPDPbiotin is reported to bias enrichment toward longer RNAs that tend to contain increasing numbers of uridines, hereafter referred to as length bias [Miller et al., 2009, Miller et al., 2011]. This bias was confirmed in our study (Figure 3.2D). While this bias can be partially mitigated statistically [Miller et al., 2009, Miller et al., 2011], more fruitful biochemical enrichment is clearly preferable, especially when examining overlapping transcript models of different sizes (e.g., spliced and unspliced, see Supplemental Information). To examine how MTS chemistry impacted the length bias in comparison with other activated disulfides, we used an in vitro transcribed RNA ladder with and without s^4U to test the relative yields of RNAs with different lengths. This analysis confirmed the presence of a length bias, and agrees well with modeling results (Figure 3.4G), demonstrating how MTS chemistry largely alleviates length bias in RNA turnover experiments. Indeed, analysis of our RNA-seq data reveals that MTS-biotin is less prone to length bias compared to HPDP-biotin (Figure 3.2D). For example, long transcripts like MALAT1 (8.7 kb) are isolated by HPDP-biotin and MTS-biotin with approximately equal efficiency, whereas shorter transcripts like SCYL1 and LTBP3 (2.3 kb and 3.4 kb, respectively, when fully spliced) are found at much greater levels in the MTS-biotin pull down (Figure 3.2E).

Studying miRNA Turnover Using MTS Chemistry Given the substantial increase in s^4U-RNA yields we observed when using MTS chemistry, we hypothesized that this chemistry could extend s^4U metabolic labeling to the study of miRNAs. The dynamics of miRNA biogenesis and degradation have gained interest because

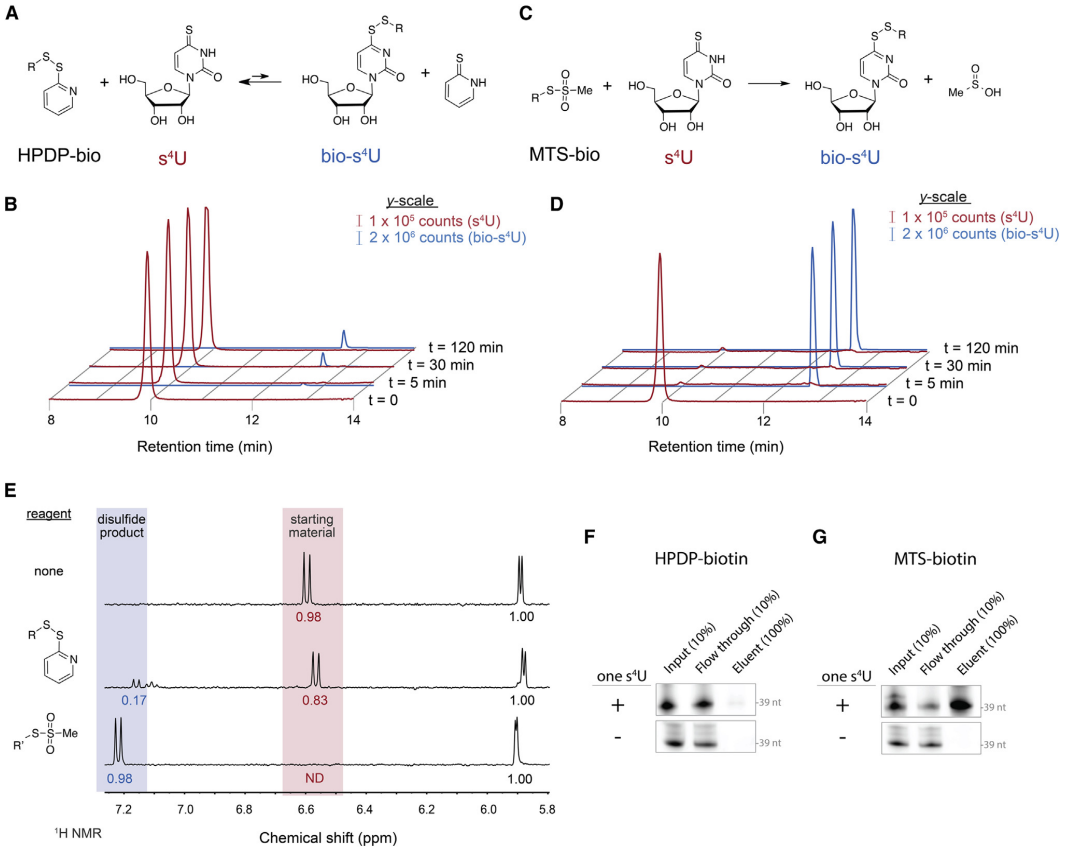

Figure 3.1: Efficient Formation of Disulfides with s^{4}U via MTS Chemistry

(A) s^4U disulfide exchange with HPDP-biotin. (B) LC-MS extracted ion chromatograms of s^4U (red) and biotin-s^4U (blue) for HPDP-biotin at the indicated reaction times. (C) s^4U disulfide exchange with MTS-biotin. (D) LC-MS chromatograms as in (B). (E) Downfield ^1H NMR spectra of (top) s^4U alone, (center) s^4U reacted with 3-[2-Pyridyldithio]propionyl hydrazide (PDPH), an HPDP-like disulfide, and (bottom) methyl-MTS. Peaks for the starting material (red shading) and products (blue shading) were integrated and normalized to the sum of the anomeric protons of s^4U and its products (5.9 ppm). For full spectra, see Figures 3.4A-C. (F and G) Enrichment of a singly thiolated 39-nt RNA by HPDP-biotin (F) or MTS-biotin (G). Fluorescently labeled 39-nt RNAs with or without a single s^4U were biotinylated with the indicated reagent and enriched on streptavidin beads, followed by urea-PAGE and fluorescence imaging. See also Figure 3.4.

disruption of miRNA homeostasis is implicated in many diseases, particularly for miRNAs that regulate progression through the cell cycle [Chang and Mendell, 2007]. Generally, miRNA turnover has been investigated by blocking transcription or by inhibiting miRNA processing, followed by analysis of miRNA stability [Bail et al., 2010, Gantier et al., 2011, Guo et al., 2015]. These approaches have demonstrated that while many miRNAs remain stable for tens of hours, there are also some miRNAs that turn over much more quickly (e.g., miR222). Extending these studies using metabolic labeling would allow the analysis of native miRNA levels in a proliferating system (unlike those studies using transcriptional block) without perturbing miRNA biogenesis or global miRNA levels (unlike studies where miRNA processing is blocked).

To investigate rates of global miRNA turnover, we treated HEK293T cells with s^4U for a range of times (Figure 3.3A) and enriched s^4U-miRNAs using MTS chemistry, followed by deep sequencing. To test whether s^4U perturbs miRNA steady-state levels, we examined miRNA levels in cells with and without s^4U treatment for 22 days, and we found high correlations in miRNA levels (Pearson's $r = 0.99$, Figure 3.6A), demonstrating that s^4U incorporation has minimal impact on miRNA levels. Our findings are consistent with previous accounts that s^4U causes minimal perturbation of longer transcripts [Gregersen et al., 2014, Hafner et al., 2010] and our own data with longer RNAs (Figure 3.6B). Consistent with our previous results and modeling, a positive control miRNA (a s^4U-miRNA spike-in added to cellular small RNAs) was enriched when using MTS-biotin but was not significantly enriched with HPDP-biotin (Figure 3.6D). We next evaluated the s^4U-miRNAs at different times after initiating s^4U treatment. We found miRNAs levels were reproducibly enriched from replicate samples (Figures 3.6C,E). Furthermore, miRNA levels in neighboring time points were most similar to each other, and those enriched at later time points (1 day, 3 days, and 6 days) approached the levels observed at steady state (22 days). As expected, the steady-state miRNA levels most closely resembled the input miRNA levels (Figure

3.3C). To determine which miRNAs turned over most quickly, we analyzed the relative distribution of enriched miRNAs at early time points (20 min) versus steady state (6 days or greater; Figure 3.3D). We identified many RNAs whose relative enrichment was significantly different from steady state at early time points and found these miRNAs displayed a consistent trend across time (Figure 3.3E). We expect fast-turnover miRNAs to be overrepresented relative to the population in early time points and slow-turnover miRNAs to be under-represented (Figure 3.3B). To evaluate this expectation, we took advantage of established properties of miRNA processing (reviewed in [Ruegger and Grosshans, 2012, Winter et al., 2009]). During miRNA biogenesis, one of the two strands from the duplex precursor generally degrades rapidly (referred to here as the miR-star), while the other strand is incorporated into the RNA-induced silencing complex (RISC) and exhibits higher stability. Therefore, we hypothesized that the miR-star sequences would be over-represented at early time points, and this hypothesis was verified: of the 52 significantly enriched and depleted miRNAs (FDR $<$ 5 3 105), about one-third of the fast-turnover miRNAs were miR-star sequences (11/30), while none of the stable miRNAs (0/22) were annotated as miR-star sequences. The fast-turnover miRNAs we identified include miRNAs that agree with previous results using transcriptional blockade (e.g., miR-222; [Guo et al., 2015]). Other miRNAs were found to be slow turnover (e.g., miR-7), and many of these are also in agreement with past studies [Bail et al., 2010, Guo et al., 2015]. In general, our results using metabolic labeling of miRNAs agree well with results from analysis of degradation after blocking miRNA production [Bail et al., 2010, Guo et al., 2015]. There are exceptions, however, such as miR-98-5p and miR191-5p, which were identified as fast-turnover miRNAs in our analysis (Figures 3.3D-E and 3.6F for qPCR validation; for a full list of fast-turnover non-star miRNAs, see Table S2), yet upon transcriptional blockade these miRNAs are stable [Bail et al., 2010, Guo et al., 2015]. While these results may be due to tissue or cell line differences, it is more likely the

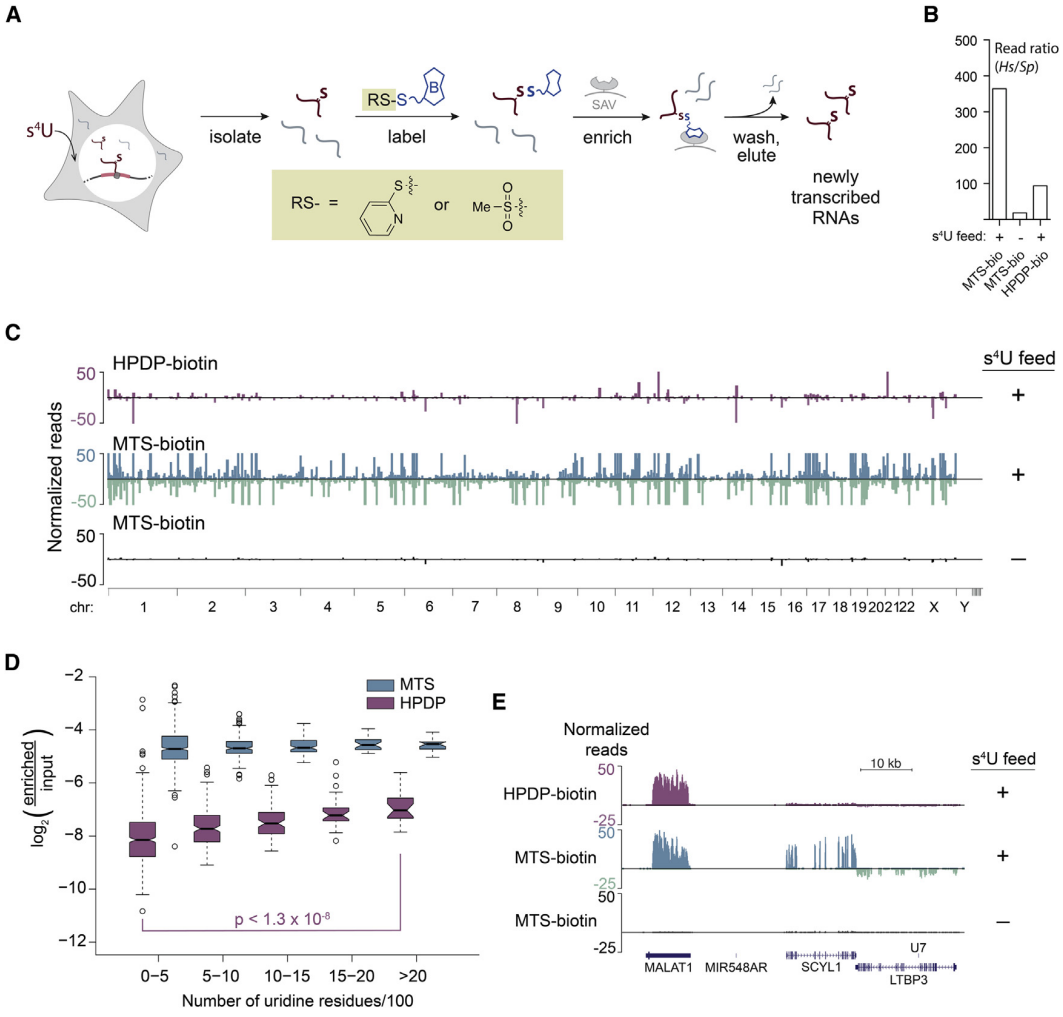

Figure 3.2: MTS-Biotin Affords Higher Specific Yields and Lower Length Bias of s^4U-RNA

(A) Schematic of s^4U metabolic labeling. 293T cells were treated with s^4U ($700\mu M$) for 1 hr, followed by total RNA extraction, biotinylation with either HPDP- or MTS-biotin, and enrichment on streptavidin-coated magnetic resin. (B) Total reads for each RNA-seq sample that mapped to the *H. sapiens* genome, divided by total number of reads that mapped to the *S. pombe* genome. (C) Whole-genome alignments of eluted samples from HPDP- or MTS-biotin enrichments. y axis indicates number of reads normalized by *S. pombe* spike-ins (see Experimental Procedures). Forward and reverse strand reads are represented as positive and negative values on the y axis, respectively. To compare coverage between samples on the same y axis scale, in some cases, read coverage exceeds the y axis upper limit in MTS-biotin (127 cases) and HPDP-biotin (4 cases). Chromosomes are indicated below the mapped reads. (D) Box plot of transcripts recovered by MTS-biotin and HPDP-biotin binned by transcript length. Blue, MTS-biotin; purple, HPDP-biotin. (E) Examples of genes enriched by HPDP- and MTS-biotin, along with a no s^4U-feed control. MALAT1 (8.7 kb), SCYL1 (2.3 kb cDNA), and LTBP3 (3.4 kb cDNA) gene architectures displayed below. See also Figure 3.5.

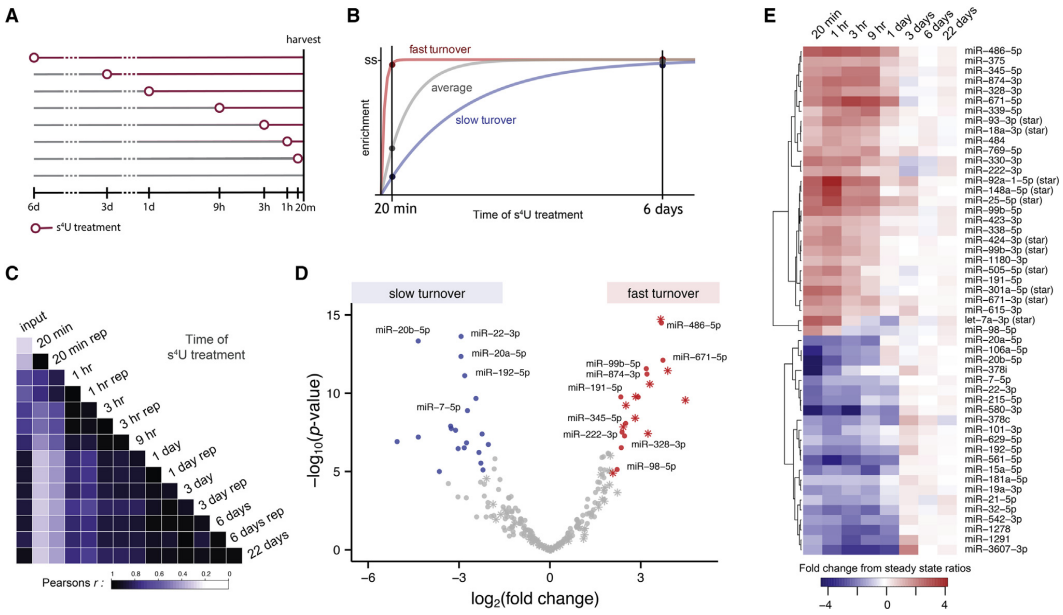

Figure 3.3: MTS Chemistry Reveals Fast- and Slow-Turnover miRNAs in miRNA RATE-Seq Experiments

(A) Schematic of s^4U treatments used in miRNA RATE-seq. (B) Cartoon of anticipated behavior of fast-turnover and slow-turnover miRNAs in comparison to average. Fast-turnover miRNAs are expected to be over- represented in the early time points, whereas slow-turnover miRNAs are depleted, relative to steady state (ss). (C) Heatmap depicting correlation coefficients (Pearson's r) between miRNA levels at different times after s^4U treatment. Replicate samples are indicated by (rep). (D) Volcano plot depicting results from a comparative analysis of miRNAs that are significantly enriched or depleted in early time points (20 min, 1 hr) relative to steady-state levels (6 and 22 days). Fast-turnover miRNAs (fold difference early time points from steady state > 4; p value $< 2 \times 10^{-5}$; Bonferroni family-wise error rate < 0.005) are colored red; slow-turnover miRNAs (fold difference early time points from steady state < 0.25; p value $< 2 \times 10^{-5}$; Bonferroni family-wise error rate < 0.005) are shown in blue. Stars indicate miRNAs defined as miRNA-stars (see Experimental Procedures); the others are indicated with circles. (E) Heatmap indicating normalized miRNA enrichment relative to steady-state level at each time point in RATE-seq for the fast- and slow-turnover miRNAs in (C). For clarity of presentation, the most significant fast-turnover miRNA in this analysis (miR-4521, log_2(fold change) $= 10.8$; p value $= 2.9 \times 10^{-40}$) has been omitted from (C) and (D) due to values exceeding the indicated scales.

faster turnover we observed for miR-98-5p and miR-191-5p is due to the cell-cycle regulation of these miRNAs [Polioudakis et al., 2015, Ting et al., 2013]. Turnover in response to progression through the cell cycle is masked when using transcriptional inhibition, but this turnover is evident using a metabolic labeling approach to study miRNA dynamics in dividing cells, underscoring one of the advantages of this improved chemistry.

3.3.5 DISCUSSION

Together, our results demonstrate that MTS-biotin is a specific reagent that can be used to efficiently label and enrich s^4URNA with higher yields and less bias than the commonly used HPDP-biotin. The dramatic improvement over existing s^4U biotinylation protocols renders MTS chemistry useful for studying dynamics of free nucleosides (Figures 3.1B, D, and E), synthetic RNAs (Figures 3.1F, G), E. coli extracts (Figure 3.4), and s^4URNA in metabolic labeling experiments (Figure 3.2). In RNA-turnover experiments, for example, the superior MTS chemistry alleviates transcript length bias, decreases the amount of starting material required, and may allow for the use of lower doses of s^4U to avoid potential toxicities that some have observed [Burger et al., 2013], but not others [Gregersen et al., 2014, Hafner et al., 2010], when metabolically labeling cells. We demonstrate the utility of this MTS chemistry using miRNA RATE-seq, which allowed us to identify fast- and slow-turnover miRNAs in proliferating cells with flux through the miRNA pathway (Figure 3.3). This advance provides the foundation for more detailed kinetic analyses of miRNA processing and turnover. More generally, applying the chemistry described herein should provide a superior means to gain insights into RNA dynamics in diverse biological systems.

3.4 Limitations

This manuscript describes improved capture of s^4U-RNA, but the enrichment will only be successful when the RNA contains sufficient levels of s^4U. In metabolic labeling experiments, incorporation of s^4U into RNA can be controlled by the concentration of s^4U during cell treatment and the time of s^4U exposure. Insufficient s^4U incorporation leads to low yields and will also favor enrichment of longer transcripts that have more uridine residues (and therefore a greater probability of s^4U incorporation). For technical considerations while performing s^4U-RNA enrichment, see Experimental Procedures and the Detailed Protocol included in the Supplemental Information.

3.5 EXPERIMENTAL PROCEDURES

Cell Lines and s^4U Metabolic Labeling HEK293T cells were cultured in high-glucose DMEM media supplemented with 10% (v/v) fetal bovine serum, and 1% (v/v) 2 mM L-glutamine. For labeling of long RNAs, cultured cells at 80% confluence were treated with 700 mM s^4U for 60 min, washed with PBS, trypsinized, and harvested. Cells were resus- pended in TRIzol reagent, flash frozen, and stored overnight at 80 C. Cell ly- sates were chloroform extracted once, and total RNA was purified by the RNeasy mini kit (QIAGEN). For miRNA labeling, cultured cells were grown for 6 days and split 1:8 on day 3. Cells were grown in the presence of 100 mM s^4U for 22days, 6days, 3days, 1day, 9hr, 3hr, 1hr, 20min, or in the absence of s^4U. On day 6, all cells were harvested using trypsin and resuspended in TRIzol reagent with exogenous s^4U-containing miRNAs (Dharmacon) and one exogenous non-s^4U miRNA (IDT). Samples were flash frozen and stored overnight at 80 C. Cell lysates were chloroform extracted once and total RNA purified by the miRvana miRNA isolation kit (Life Technologies).

Purification of s^4U-Labeled RNA Biotinylation and s^4U-RNA enrichment

with HPDP-biotin were carried out based on protocols adapted from Gregersen et al. (2014) and optimized for MTS-biotin. Reactions were carried out in a total volume of 250 ml, containing 70 mg total RNA, 10 mM HEPES (pH 7.5), 1 mM EDTA, and 5 mg MTSEA biotin- XX (Biotium) or 50 mg HPDP-biotin (Pierce) freshly dissolved in DMF (final con- centration of DMF = 20%). Reactions were incubated at room temperature for 2 hr (HPDP) or 30 min (MTS) in the dark. Following biotinylation, excess biotin reagents were removed by addition of 1 volume phenol:chloroform (Sigma), followed by vigorous mixing for 15 s, 2 min incubation at RT, and centrifugation in a Phase-Lock-Gel tube (5Prime) at 12,000 3 g for 5 min. Supernatant was removed, and RNA was precipitated with a 1:10 volume (20 ml) of 5 M NaCl and an equal volume of isopropanol (200 ml) and centrifuged at 20,000 3 g for 20 min. The pellet was washed with an equal volume of 75% ethanol. Purified RNA was dissolved in 50 ml RNase-free water and denatured at 65 C for 10 min, followed by rapid cooling on ice for 5 min. Biotinylated RNA was sepa- rated from non-labeled RNA using mMacs Streptavidin Microbeads (Miltenyi). Beads (200 ml) were added to each sample and incubated for 15 min at room temperature. In the meantime, μColumns were placed in the magnetic field of the μMacs separator and equilibrated with nucleic acid wash buffer supplied with the beads (Miltenyi). Reactions were applied to the mColumns, and flow-through was collected as the pre-existing RNA fraction. mColumns were washed twice with high-salt wash buffer (500 ml each, 100 mM Tris- HCl [pH 7.4], 10 mM EDTA, 1 M NaCl, and 0.1% Tween-20). s^4U-RNA was eluted from mColumns with 100 ml freshly prepared 100 mM DTT followed by a second elution with an additional 100 ml 5 min later. RNA was recovered from the flow-through and eluent samples using the MinElute Spin columns (QIAGEN) according to the instructions of the manufacturer. S. pombe total RNA (11 ng, a generous gift from Julien Berro) was added to each sample for downstream normalization.

s^{4}U-Seq Library Preparation and Sequencing All sequencing libraries were constructed using standard protocols by the Yale Center for Genomic Analysis (YCGA) and run on Illumina HiSeq 2500 instru- ments. Long RNA-seq was performed using 5 mg of RNA from input RNA, flow-through, or eluted fractions. Strand-specific library preparation was performed using poly-A-selected RNA collected from flow-through and eluted fractions. Samples were multiplexed using Illumina bar codes and sequenced using paired-end 2 3 75-nt cycles. For small RNA-seq, 10% input and RNA collected from eluted fractions were used for small RNA library preparation and sequenced with single-end 75-nt cycles.

Mapping and Quantification of s^{4}U-Seq Libraries For long RNA-seq, sequencing reads were aligned using Tophat2 (version 2.0.12; Bowtie2 version 2.2.3), to a joint index of the H. sapiens and S. pombe genomes (hg19 and PomBase v22) and transcriptomes (GENCODE v19 and Ensembl Fungi v22; [**?**, Kersey et al., 2014], respectively). Alignments and analyses were performed on the Yale High Performance Computing clusters. Following this, we used Cufflinks (version 2.2.1; [Trapnell et al., 2010]) to quantify annotated H. sapiens and S. pombe transcripts, using only reads that were uniquely mapped (MAPQ ¿ 20) and that aligned with up to two mismatches to the reference.

s^{4}U-Seq Normalization To compare transcript levels between samples, we normalized expression values to S. pombe spike-ins as follows:

$$FPKM_{norm} = FPKM_{raw} S_{norm}$$

where $FPKM_{norm}$ is the normalized fragments per kilobase per million reads (FPKM) of a human transcript or gene, $FPKM_{raw}$ is the original FPKM calculated for the sample of interest, and S_{norm} is the slope of the linear regression line of raw S. pombe gene FPKMs, with the normalizing sample on the y axis and the sample of

interest on the x axis (Figure 3.5B). To normalize genomic coverage tracks, we used a similar scheme:

$$Coverage_{norm} = Coverage_{raw} S_{norm} \frac{R_{sample}}{R_{norm}}$$

where $Coverage_{norm}$ and $Coverage_{raw}$ are the normalized and raw read coverages at a given genomic position, and Rsample and Rnorm are the numbers of unique reads in the sample of interest and the normalizing sample, respectively. The $\frac{R_{sample}}{R_{norm}}$ adjustment factor reflects that we are comparing raw reads instead of FPKMs. We generated stranded genomic coverage tracks using IGVTools (version 2.3.32; [Thorvaldsdottir et al., 2013]). For all analyses, we normalized to the S. pombe spike in the HPDP-biotin sample. We also accounted for the 10-fold biochemical dilution of the input samples prior to library preparation by multiplying normalized values for these samples by ten.

Assessment of Length Bias in Eluted s^{4}U-Seq RNA Because incorporation and biotinylation of s^{4}U are not perfectly efficient, especially when using HPDP-biotin, it is expected that transcripts with more uridines will be purified at rates greater than or equal to those of shorter tran- scripts. To assess length bias for each reagent, we binned transcript isoforms by numbers of uridines present and compared the fractions of total input RNA that were purified between bins using the Wilcoxon rank-sum test. To avoid noise from misassignment of reads between isoforms of individual genes, we included only the dominant isoforms of genes (¿90% of total expression) in all samples included in the analysis. We only included transcripts greater than 200 nt, since shorter transcripts were biochemically depleted in the library preparation, and removed transcripts with expression levels in the bottom quartile of the input sample.

qPCR Assays For qPCR analysis of long RNA, input or enriched RNA was

converted into cDNA with VILO reverse-transcription kit (Life Technologies). qPCR was carried out on the CFX96 real-time system (BioRad) with the iTaq Universal SYBR Green Mix. Results from all primers used were corrected for amplification efficiency. For miRNA analysis, qPCR was performed using TaqMan miRNA assays (Life Technologies) according to the instructions of the manufacturer for the following targets: hsa-miR-7, UGGAAGACUAGU GAUUUUGUUG; hsa-miR-20a, UAAAGUGCUUAUAGUGCAGGUAG; hsa- miR-98, UGAGGUAGUAAGUUGUAUUGUU; hsa-miR-99b, CACCCGUAGAA CCGACCUUGCG; hsa-miR-191, CAACGGAAUCCCAAAAGCAGCUG; hsa- miR-222, AGCUACAUCUGGCUACUGGGUCUC; EED004r, CCAUUUGUAU GUUCGGCUAACU; and EED095r, CCAUUUCGCUCGGGUGC

miRNA RATE-Seq s^4U RNA Enrichment Biotinylation and s^4U-RNA enrichment were carried out as described above (purification of s^4U-labeled RNA) with the following modifications. Excess biotinylation reagent was removed using a nucleotide cleanup kit (QIAGEN). Following enrichment, RNA was concentrated by ethanol precipitation and resuspended in 14 ml RNase-free water. After enrichment, samples were supple- mented with four synthetic miRNA standards (Dharmacon).

miRNA RATE-Seq Bioinformatic Analysis To analyze our smRNA RATE-seq data, we used a hierarchical mapping pipeline combining the sRNAbench (Rueda et al., 2014), Bowtie (Langmead et al., 2009), and Bowtie2 tools (Langmead and Salzberg, 2012). Before mapping the reads, we removed sequencing adapters, using fastx-clipper (hannonlab.cshl.edu/fastx_toolkit/). We then proceeded to use Bowtie2 to map reads first to synthetic spikes, and then to the UniVec laboratory contaminant database (www.ncbi.nlm.nih.gov/tools/vecscreen/univec/) and ribosomal RNAs from the GENCODE v19 annotation (Harrow et al., 2012). These two categories of sequences are not expected to produce reads in our miRNA libraries, except by contamination or RNA degradation. The remaining unmapped reads were then mapped using sRNAbench, first to the miRBase miRNA 21 annotation

[Kozomara and Griffiths-Jones, 2014], and then to the entire human genome (hg19). Input reads under 19 nt or with greater than one mismatch were removed from all analyses of miRNA and spike quantifications.

To perform differential expression analysis between smRNA RATE-seq time points, we used the edgeR package (version 3.2.4; [Robinson and Smyth, 2008, Robinson et al., 2010]). Specifically, we compared three early time points (both 20 min replicates and a deeply sequenced 1 hr time point) to three late time points (two 6-day replicates and a 22-day sample). miRNA read counts and dispersions were fit to a negative binomial distribution, and differential expression was evaluated using the negative binomial exact test. To correct for multiple hypothesis testing, we used the Bonferroni correction, and set a family-wise error rate of 0.005 to select differentially expressed miRNAs between early time points and the steady state.

Mass Spectrometry of s^4U Disulfide Exchange Reactions (50 ml) contained s^4U (50 mM), buffer (20 mM HEPES [pH 7.5], 1 mM EDTA), and MTS- or HPDP-biotin (5 mM) dissolved in DMF (final concentration of DMF = 5%). Aliquots were taken at designated time points and analyzed on an Agilent 6650A Q-TOF using a reverse phase column (Thermo Scientific Hypersil GOLD 3 mm, 160 3 2.1 mm) detected by electrospray ionization (positive ion mode). Chromatography conditions were established based on Su et al. (2014). Briefly, analysis was initiated with an isocratic gradient of 100% buffer A at 0.4 ml/ min for 6 min followed by a linear gradient of 0%–50% buffer B over 6 min, 50%–75% buffer B over 2 min, then an isocratic elution at 75% buffer B (buffer A: H2O in 0.1% [v/v] formic acid; buffer B: acetonitrile in 0.1% [v/v] formic acid).

NMR of s^4U Disulfide Exchange Reactions (600 ml) were performed in D2O containing 10 mM HEPES, s^4U (1 mg, 6.4 mM), and five equivalents of MeMTS or PDPH dissolved in DMF-d7 (60 ml, 10% total volume). These reactions were incubated in the dark, 2 hr for PDPH and 30 min for MeMTS. Reactions were analyzed

on an Agilent DD2 400 MHz NMR with 16 scans.

Enrichment of Singly Thiolated RNA Two fluorescently labeled RNAs were synthesized for s^4U enrichment: non- s^4U 39-nt RNA (DY647 - GGAACCGCCCG-GAUAGUGUCCUUGGGAAACCAA GUCCGGGCACCA) and one s^4U 39-nt RNA (DY547 – GGAACCGCCCGGA [s^4U]AGUGUCCUUGGGAAACCAAGUCCGGGCACCA) (Dharmacon). Bio- tinylation reactions (50 ml total) contained RNA (1 mM), 10 mM HEPES (pH 7.5), 1 mM EDTA, and 25 mM MTS- or HPDP-biotin (dissolved in DMF at 250 mM). Reactions were incubated at room temperature in the dark for 30 min or 2 hr, respectively. Following biotinylation, excess biotinylation reagents were removed with two consecutive chloroform washes, followed by purification with a nucleotide cleanup kit (QIAGEN) according to the manufac- turer's instructions. Biotinylated RNA was separated from non-labeled RNA using Dynabeads MyOne Streptavidin C1 beads (Invitrogen). Biotinylated RNA was incubated with 50 ml Dynabeads with rotation for 1 hr at room temperature in the dark. Beads were magnetically fixed and washed twice with Dynabeads high-salt wash buffer. s^4U-RNA was eluted with 100 ml of elution buffer (10 mM Tris [pH 7.4] and 100 mM DTT). Fractions were concentrated by ethanol precipitation, separated on a 12% urea-PAGE gel, and visualized by Typhoon fluorescence imager (GE).

Enrichment of an In Vitro Transcribed RNA Ladder An RNA ladder of 100–1,000 nt was transcribed in vitro using the RNA Century Plus Marker Template and Maxiscript T7 transcription kit (Invitrogen) using Cy5-CTP at a ratio of 1:1 Cy5-CTP:CTP for downstream visualization, with the option of adding s^4UTP (TriLink Biotechnologies) at a ratio of s^4UTP:UTP to the reaction. After the reaction, excess nucleotides were removed by an Illustra Microspin G-25 column (GE Healthcare Life Sciences) according to the manufacturer's instructions. RNA ladders were reacted with HPDP-, MTS-, or thiosulfonate-biotin (Biotium), following the protocol described above. Enriched samples were separated on a 5% urea-PAGE gel, stained

with GelGreen, and visualized by Typhoon fluorescence imager (GE).

Enrichment of Thiolated tRNA from E. coli E. coli WT and ΔthiI cultures were grown to mid-log phase in LB media. Strains were a generous gift from Eugene Mueller [Mueller et al., 1998]. Cells were pelleted at 3,250 3 g for 10 min at 4 C. Total RNA was purified by the mirVana miRNA isolation kit (Life Technologies). RNA pull downs were performed as above (Purification of s^4U-labeled RNA) and fractions separated on a 5% urea-PAGE gel, followed by visualization with GelGreen stain.

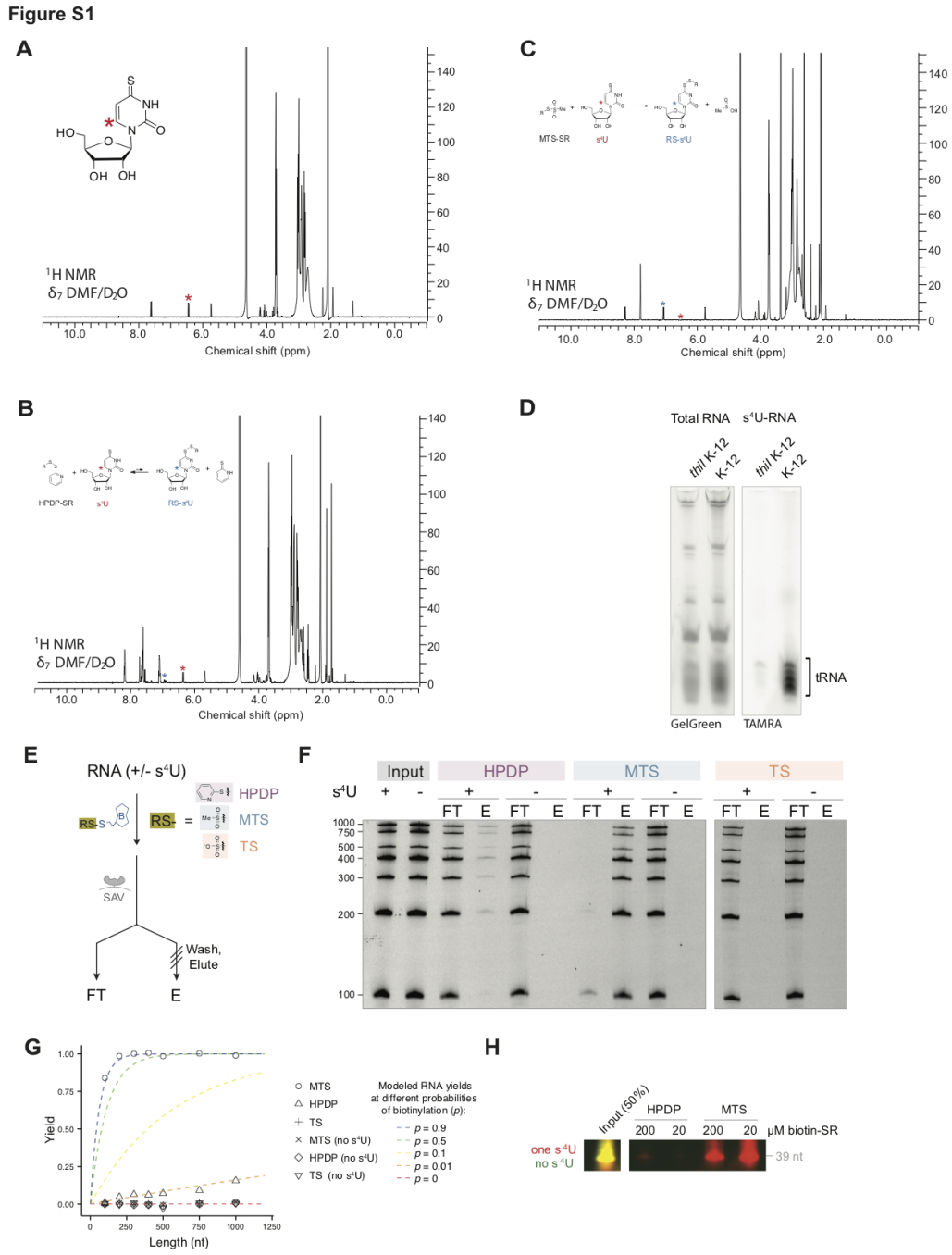

Figure 3.4: Reactivity of activated disulfides with s^{4}U and in vitro modulation of bias in MTS- and HPDP-biotin enrichments

Figure 3.4: Reactivity of activated disulfides with s^4U and in vitro modulation of bias in MTS- and HPDP-biotin enrichments

(A) ^1H NMR spectrum of s^4U alone. Peak labeled with a red "*" corresponds to the starred proton in the s^4U structure. (B) ^1H NMR spectrum of s^4U when treated with methane methylthiolsufonate (MeMTS), the same reactive disulfide of MTS-biotin. MeMTS was incubated with s^4U for 30 min and the extent of disulfide exchange was monitored by the chemical shift in proton labeled with a red "*". Peak labeled with a blue "*" represents the chemical shift upon disulfide bond formation. (C) ^1H NMR spectrum of s^4U when treated with a compound containing the same functional group of HPDP-biotin. Pyridyldithio]propionyl hydrazide (PDPH) was incubated with s^4U for 2 hr and the extent of disulfide exchange was monitored by changes in chemical shift as in (B). (D) RNA from E. coli K-12 cells was reacted with MTS-TAMRA fluorescent dye and visualized on a 5% urea-PAGE gel. K-12 cells express ThiI, an enzyme that selectively modifies U8 of tRNA to s^4U8 (Mueller et al., 1998b). RNA from a ΔthiI knockout shows little TAMRA signal (traces of unmethylated 2-thiouridine on tRNA can still react), whereas a strong TAMRA signal is present in the K-12 cells only in tRNA. Total RNA was stained with GelGreen. (E) Schematic of in vitro enrichment of s^4U-RNA using an RNA ladder. An RNA ladder was in vitro transcribed with Cy5-CTP and with or without added s^4UTP. s^4U-RNAs were enriched by reacting with disulfide-activated biotin derivatives using either HPDP, MTS, or thiolsulfonate (TS, an alternative disulfide activated biotin reagent) chemistry. (F) Input, flow-through, and elution RNAs were analyzed by urea-PAGE and visualized by Cy5 fluorescence. Band intensities were quantified using ImageJ. (G) Comparison between the yields observed in (E) and expected enrichment using models that assume different biotinylation efficiencies. In all cases modeled lines assume ratio of $s^4U/U_{total} = 0.075$ to determine the expected yield given different biotinylation efficiencies (y_{bio}) based on the equation:

$$yieldRNA = \sum_{j=0}^{N_i} [1 - (1 - y_{bio)^j}]_{p(U_i=j)})]$$

In comparison to the models results, empirical yields using the band intensities from (B) were plotted based on transcript length. (H) Effects of biotin concentration on modeled s^4U-RNA enrichment. Synthetic short RNAs (1 nM) with one s^4U residue (red) or zero s^4U residues (green) were enriched by 200 μM (comparable to 50 μg biotin in total RNA pulldown) or 20 μM HPDP- or MTS-biotin. No significant difference in enrichment was observed using these two concentrations of MTS-biotin eluent, whereas 200 μM HPDP-biotin showed 6-fold greater enrichment over 20 μM HPDP-biotin.

Figure S2

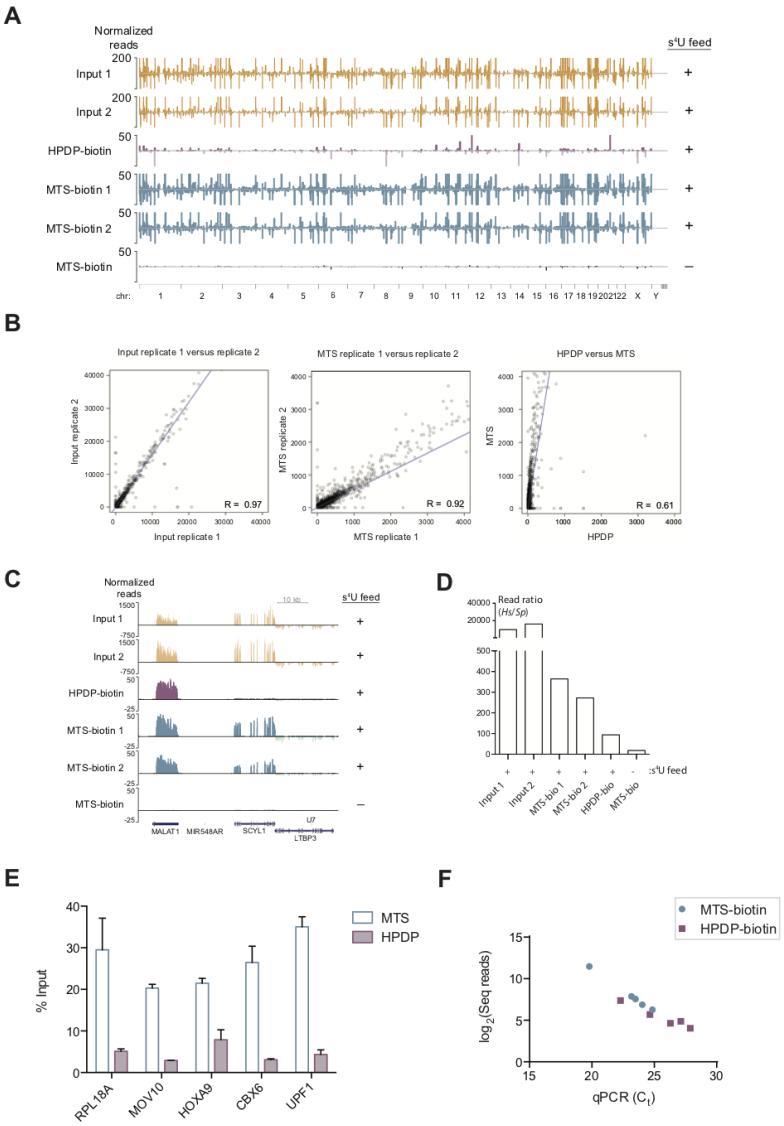

Figure 3.5: Reproducibility of MTS-biotin enrichment

Figure 3.5: Reproducibility of MTS-biotin enrichment

(A) Whole genome alignments of RNA-Seq samples as in Figure 3.2B. The y-axis indicates the number of reads normalized to total number of S. pombe aligned reads. To compare coverage between samples using the same scale on the y-axis, in many cases read coverage exceeds the y- axis upper limit in Input (135 cases), MTS-biotin (127 cases) and HPDP-biotin (4 cases). Chromosomes are indicated below the mapped reads. (B) Scatter plots and Pearson correlations of normalized FPKM values for H. sapiens transcript isoforms. Plots show Input 1 vs. Input 2 (left), MTS-biotin replicate 1 vs. HPDP-biotin (center), and MTS-biotin replicate 1 vs. MTS-biotin replicate 2 (right). (C) Example of genes enriched by HPDP-biotin and MTS-biotin as in Figure 3.2C. (D) Total reads for each RNA-Seq sample that mapped to the H. sapiens genome, normalized by total number of reads that mapped to the S. pombe genome, as in Figure 3.2D. (E) Samples enriched by MTS- or HPDP-biotin from RNA-seq submission were analyzed by qPCR using gene-specific primers for RPL18A, MOV10, HOXA9, CBX6, and UPF1 with two replicates. Ct values from qPCR were used to calculate percent input using the equation:

$$\frac{1}{2^{Ct_{sample}-Ct_{input}}}$$

where the input is the average of two replicates. Error bars indicate the mean of two technical replicates +/- SEM. (F) Ct values from qPCR were plotted against the number of reads (log2 transformed) for deep sequencing in input (triangles), MTS- (circles) and HPDP- (squares) enriched samples.

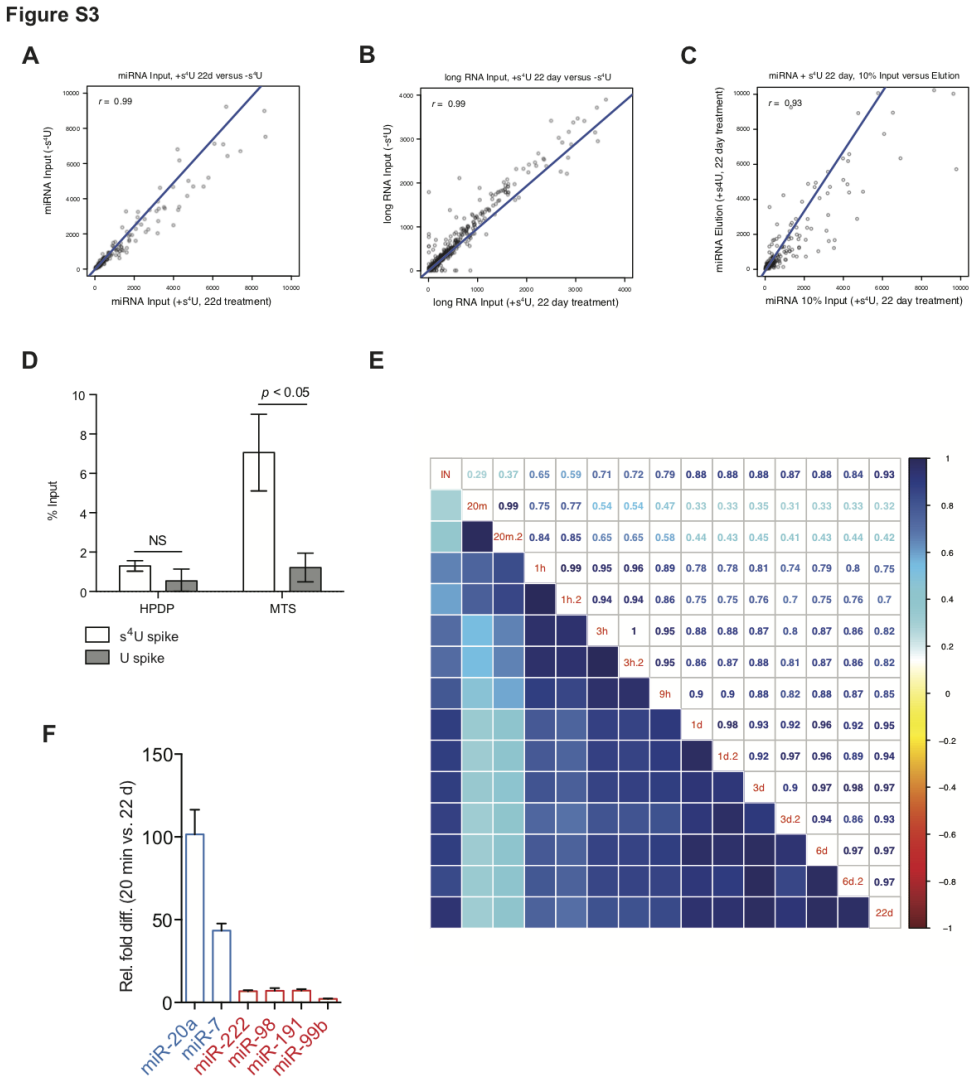

Figure 3.6: Analysis of s^{4}U metabolic labeling and enrichment for miRNA RATE-Seq

Figure 3.6: Analysis of s^4U metabolic labeling and enrichment for miRNA RATE-Seq

(A-C) Scatter plots and Pearson correlations of RNA-Seq quantifications of H. sapiens miRNA transcripts. Plots show (A) reads from miRNA isolated from cells with not s^4U treatment compared to reads from total miRNA from cells after 22 days of s^4U treatment; (B) analysis of long RNAs from the same cells as in (A); and (C) analysis of miRNA isolated from 22 day s^4U treatment (10% input) vs. MTS-biotin enriched miRNA from 22 days of s^4U treatment. (D) miRNAs enriched with HPDP- and MTS-biotin. Control miRNA spikes containing one s^4U (EED004r) or zero s^4U (EED0095r) were enriched with s^4U-miRNA samples and enrichment was detected by qPCR using the same equation as above. The s^4U-containing spike-in control was not significantly enriched over background by HPDP-biotin ($p = 0.27$), whereas the s^4U-containing spike-in control was significantly enriched by MTS-biotin ($p = 0.034$). (E) Heatmap similar to Figure 3.3C with annotated correlation coefficients (Pearson's r) between miRNA levels at different times after s^4U treatment. Replicate samples are indicated by (rep). (F) The enrichment of select miRNAs using MTS-biotin after 1 hr and 6 days s^4U treatment was validated by qPCR and quantified as fold enrichment as in Figure 3.5E. Error bars indicate the mean of three technical replicates +/- SEM.

3.6 Appendix: Modeling expected yields of s^4U enrichment in metabolic labeling experiments

The objective of the experiment described above is to determine the fraction of newly transcribed RNA (f_i) for each RNA in the cell. The output of the experiment is the yield of each RNA over input (X_i):

$$X_i = \frac{RNA_{enriched}}{RNA_{input}}$$

where X_i can be related to f_i:

$$X_i = f_i p_i(N_i)$$

Longer RNAs generally have a higher number of uridines and are therefore more likely to be captured. Specifically, the probability of capture, $p_i(N_i)$, is a function of the number of uridine residues (N_i) in the transcript. In order for $X_i \approx f_i$, we need $p_i(N_i)$ to be as close to unity as possible. We can model the length dependence of $p_i(N_i)$ as follows:

$$p_i(N_i) = \sum_{j=0}^{N_i} [1 - (1 - y_{bio})] p(U_i = j)$$

where y_{bio} is the yield of s^4U biotinylation, and U_i is the number of s^4U residues in the transcript.

If we assume that s^4U residues are randomly incorporated into newly transcribed RNA at sites of uridine, then Ui will display a binomial distribution with mean r, the ratio of s^4U to uridine ($\frac{s^4U}{U_{tot}}$) incorporated into the RNA according to the following equation:

$$p(U_i = j) = \binom{N_i}{j} r^j (1-r)^{N_i - j}$$

This model assumes that any RNA with one or more biotins will be retrieved quantitatively, which agrees well with the high affinity of streptavidin for biotin, and the observation that the flow through does not contain significant amounts of biotinylated RNA (see Figure 3.4F). We have modeled the expected yields of biotinylation at given level of s^4U incorporation ($r = 0.075$) and the results agree well with the experimental data from enrichment using an in vitro transcribed RNA ladder (Figure 3.4G), supporting the utility of this model.

Based on this model, there are two ways to decrease the length bias: (1) increase y_{bio}, the yield of conversion of s^4U to bio-s^4U, or (2) increase r, the number of s^4U's in the transcript.

This manuscript describes MTS-chemistry that dramatically increases ybio, the yield of biotinylation of each s^4U residue. In agreement with our data and these equations, this improvement leads to higher yields of s^4U-RNA (Figure 3.2B, C) and lower length bias (Figure 3.2D and Figure 3.4F). For (2), increasing the [s^4U] fed to cells can increase r. The extent to which r can be increased has practical constraints. At very high [s^4U], the nucleoside is toxic to the cells [Burger et al., 2013]. In one case [Heyn et al., 2014], it was possible to further increase incorporation by directly injecting s^4UTP into cells, which provided high enrichment of even short transcripts. However, direct injection of individual cells is not always feasible. We find that under the highest commonly used concentrations of nucleoside ([s^4U] = 700 µM), even the longest spliced transcripts are enriched at lower levels with HPDP-biotin than with MTS-biotin (Figure 3.2D). In other words, under standard incorporation rates, the difference in y_{bio} between MTS-biotin and HPDP-biotin has a significant impact on

X_i. Irrespective of the incorporation rate, it is always preferable to increase the yield of biotinylation (y_{bio}) to make more efficient use of the s^4U that has been incorporated into the labeled RNA.

It is interesting to note that according to this model, low yields of biotinylation (such as those achieved using HPDP-biotin) lead to comparative enrichment of very long transcripts (such as those containing long, unspliced introns) over moderately sized transcripts. While low sequencing coverage in the input samples prevented accurate quantitation of X_i for these long, low-abundance, unspliced transcripts, this enrichment is clearly evident in the mapped coverage, consistent with this prediction. It is also worth noting that measurements of RNA half-lives using HPDP-biotin have the potential to be accurate (provided sufficient signal-to-noise) because the length bias is constant for any given RNA over time [Neymotin et al., 2014]. The relative amounts of different transcripts and the analysis of splicing, however, may be strongly influenced by the length bias of capture. The low y_{bio} for HPDP-biotin is expected

to influence these results. An example of this effect can be seen in Figure 3.2E.

Chapter 4

Modeling of overdispersion in RNA chemical probing data and application to secondary structure prediction

4.1 Summary

This chapter describes analysis of statistical overdipsersion in RNA chemical probing data read out with high throughput sequencing. This group of techniques enables measurement of a variety of properties of RNA nucleotides and can be applied to aid RNA secondary structure prediction. The core of this chapter is a draft of a research article that is currently in preparation for submission. For that article, I conducted all computational analysis, under the joint direction of Dr. Mark Gerstein and Dr. Matthew D. Simon. Along with Dr. Simon and Dr. Alec N. Sexton, I designed a set of 60 replicate experiments to help faciliate and benchmark statistical modeling methods for RNA chemical probing data. These experiments were carried out by Dr.

Alec Sexton. Peter Y. Wang contributed a then-unpublished script for quantification of reverse transcription stops and mutations in probing data.

4.2 Modeling of overdispersion in RNA chemical probing data and application to secondary structure prediction

4.2.1 Abstract

Chemical probing techniques can reveal biologically important properties of RNA molecules—including structural context, chemical modification, and interaction with proteins—at single nucleotide resolution. This set of techniques has recently been adapted to readout with high throughput sequencing, enabling *in vivo* and transcriptome-wide studies. Despite the expanding use of chemical probing technologies, data have often been modeled using simplifying statistical assumptions that deemphasize the value of conducting replicate experiments. Here, I investigate and model overdispersion in RNA chemical probing data, demonstrating the importance of replicate experiments to biological and statistical interpretation. To facilitate this analysis, we collect novel datasets with 60 replicates that enable us to observe overdispersion more directly and to investigate the value of incremental data collection to statistical modeling. I also investigate the effects of variability of RNA chemical probing data on predictions of RNA secondary structure and apply our model to propose a quantitative metric of the contribution of uncertainty in chemical probing results to the breadth of possible predicted RNA structures.

4.2.2 Introduction

Chemical probing techniques can reveal biologically important properties of RNA molecules at single nucleotide resolution []. The applications of this versatile set of tools include studies of RNA structure, chemical modification, and interactions with proteins, all of which aid mechanistic investigations of RNA function and regulation. The utility of chemical probing experiments to study RNA biology has expanded in scope, as probes have been developed that work in vivo and techniques have been adapted to a sequencing platform for both transcriptome wide [Carlile et al., 2014, Dai et al., 2017, Ding et al., 2014, Zubradt et al., 2017, Rouskin et al., 2014, Spitale et al., 2015a] and targeted [Fang et al., 2015, Smola et al., 2015, Smola et al., 2016] analyses. The most widely used probes–e.g. dimethyl sulfate (DMS) and selective 2' hydroxyl acylating (SHAPE) reagents–aid RNA secondary structure determination by selectively modifying single-stranded and flexible nucleotides. Nucleotides modified by chemical probes are then read out by reverse transcriptase (RT), which terminates cDNA synthesis or inserts incorrect bases at chemical adducts (we refer to RT stops and mutations more generally as RT events). Comparing results of probing experiments to controls with no chemical treatment enables calculation of nucleotide reactivities, which are then converted into probabilistic constraints for RNA secondary structure prediction, or parallel inferences about other nucleotide properties [Deigan et al., 2009].

A key to the interpretation of chemical probing data is the evaluation of experimental reproducibility. We can separate this issue into two elements: measurement variability for each replicate and experimental variability between replicates. The importance of measurement variability has been acknowledged implicitly within the field. In the analysis of transcriptome-wide data, the necessity of achieving sufficient sequencing depth to obtain robust results is juxtaposed against the difficulty of collecting sufficient data for RNAs that are present in cells at low concentrations [Choudhary et al., 2016, Li et al., 2017]. As a result, many of the

initial conclusions from transcriptome-wide studies using chemical probing, or related techniques that also investigate RNA structure, relate to average properties of many transcripts [Rouskin et al., 2014, Ding et al., 2014, Mortimer et al., 2014, Zheng et al., 2010, Kertesz et al., 2010]. More recently, detailed structural modeling of a set of hundreds of mRNAs was enabled by particularly deep sequencing of SHAPE probing with a mutational readout (SHAPE-MaP) in Escherechia coli, which has a much smaller transcriptome than humans. In parallel, the traditional format of targeted probing toward RNAs of interest has also been adapted to high throughput sequencing, enabling, for example, targeted study of secondary structures across the Xist RNA for the first time [Fang et al., 2015, Smola et al., 2016].

Statistical models of chemical probing data can be used to assess the robustness of experimental observations, but existing statistical methods often make simplifying assumptions about the degree of variability that would be expected between replicate experiments [Choudhary et al., 2016, Aviran and Pachter, 2014, Li et al., 2017, Siegfried et al., 2014, Smola et al., 2015]. Ideally, if all experimental and biological conditions could be held constant between replicates, there would be a fixed probability of observing a reverse transcription stop or mutation across replicates. This would mean that models such as the Binomial distribution and the Poisson distribution, which make this assumption, would be useful for modeling probing data, and that replicates would not be needed formally. Indeed, both of these distributions are commonly used to analyze chemical probing data [Choudhary et al., 2016, Aviran and Pachter, 2014, Li et al., 2017, Siegfried et al., 2014, Smola et al., 2015]. Moreover, most analysis methods for chemical probing analysis address the results of a single replicate, or the pooled results of multiple replicates [Mustoe et al., 2018, Smola et al., 2015, Li et al., 2017]. However, there is reason to be concerned that assuming that chemical probing counts are generated from a uniform statistical process is not an ideal approach. Many forms of biological data– ranging from gene expres-

sion (RNA-Seq)[Robinson and Smyth, 2008, Robinson et al., 2010] to mutation rates in cancer genomes [Lochovsky et al., 2015]–are overdispersed relative to the binomial and Poisson distributions, precisely because the underlying probability of the event being tracked changes between replicates. Supporting the possibility that this might be the case with chemical probing data, we note that in some reports, up to eleven replicates are used to make conclusions about probing results [Carlile et al., 2014]. Moreover, a recent study used a model of overdispersion to analyze chemical probing data read out by gel electrophoresis [Vaziri et al., 2018]. However, attempts to model variability between replicates of chemical probing experiments with readout by high throughput DNA sequencing have been limited.

The question of variability in RNA chemical probing data becomes particularly important when considering the interpretation and applications of these data. One of the most common applications of chemical probing experiments is incorporation as probabilistic constraints (expressed as pseudoenergies) in thermodynamic RNA secondary structure prediction algorithms [Deigan et al., 2009, Eddy, 2014]. Recent studies have investigated the effects on structure prediction of variability in RNA melting experiments that were used to tune the nearest neighbor parameters that form the core of thermodynamic RNA secondary structure prediction algorithms [Zuber et al., 2017, Zuber et al., 2018]. This motivates parallel investigation of how variability in chemical probing experiments can also affect structure predictions.

Here we investigate the formal question of whether separate analysis of replicates is useful for analysis of chemical probing data. We then propose a modeling method for count data in chemical probing experiments, adapting methodology from the RNA-Seq field. To augment our conclusions from publicly available datasets (with small numbers of replicates) and benchmark our modeling method, we collected datasets with 60 replicates of DMS probing, using both targeted and random primers. Using this model, we then examine the effect of count variability on RNA secondary

structure predictions.

4.2.3 Results

Investigation of overdispersion in publicly available chemical probing datasets

Count data from chemical probing experiments have frequently been modeled using the binomial or Poisson distributions [Choudhary et al., 2016, Aviran and Pachter, 2014, Li et al., 2017, Siegfried et al., 2014, Smola et al., 2015]. To motivate these models, one can view the process of reverse transcription stopping (or insertion of a mutation) as a Bernouli trial, in which the reverse transcriptase will stop (or mutate, depending on the readout being used) at a given nucleotide i with some probability p_i. For the reads that reached the nucleotide of interest across an entire sequencing dataset, one can then model counts of reverse transcription events using the binomial distribution. RT events typically have very low probabilities, so when sequencing coverage is relatively high, counts would then follow the Poisson distribution, which matches the binomial distribution in the limits of a high number of trials and low probability of success. Both the binomial and Poisson distributions make the simplifying assumption that the mean probability of an RT event at each nucleotide i (p_i is constant).

To evaluate the assumption that p_i remains constant between replicates, we first performed exploratory analysis of a set of targeted SHAPE-Seq dataset for the P4-P6 helix region of the *Tetrahymena* group I intron [Loughrey et al., 2014] and its fit to the Poisson and binomial distributions. Transcription of the RNA sample and SHAPE probing for this dataset are both performed *in vitro*, making it a good candidate to meet simplifying assumptions made by the binomial and Poisson distributions.

To gain an intuitive feel for whether our sample probing dataset matches the above simplifying assumptions, we focused on the region from nucleotides 1-50 of the group I intron domain, and plotted normalized counts (see methods) for each nucleotide in a treated sample along with 95% Poisson confidence intervals around

the estimates. We then plotted normalized counts for the same nucleotides from a second replicate. Of the 50 nucleotides we examined, 44 had counts outside the range for a single replicate (2.5 outliers would be expected). To examine the level of variability in probing data across the entire group I intron domain, we plotted the normalized mean counts for two observed replicates (Figure 4.1b) and compared these to simulated replicate datasets according to the Poisson distribution (Figure 4.1c). Consistent with our initial observations, observed variability was much greater than that assumed by the Poisson model, implying that the data are overdispersed.

To investigate the overdispersion of chemical probing data more formally, we considered the p-values from the Poisson or binomial tests for observations of replicates of the group I intron SHAPE-Seq data. Since these are replicate experiments, we expect that observations from both replicates at each nucleotide come from the same distribution. If the model accurately describes the variability of the data, this would lead p-values comparing replicate observations to follow the uniform distribution. To test whether this is the case, we plot the ordered Poisson and binomial exact p-values between replicates against the quantiles of the uniform distribution (quantile-quantile plot, Figure 4.1f) and observe that the Poisson exact p-values are almost all more extreme than any expected p-value. This observation is borne out by using the Kolmogorov-Smirnov test, a standard test for whether the goodness of fit of two distributions, which shows that Poisson p-values differ greatly from the uniform distribution ($p < 2.2 * 10^{-16}$). Together, these observations show that the Poisson distribution greatly underestimates the variability in an *in vitro* SHAPE-Seq dataset of the P4-P6 domain of the *Tetrahymena* group I intron, and that these data are overdispersed.

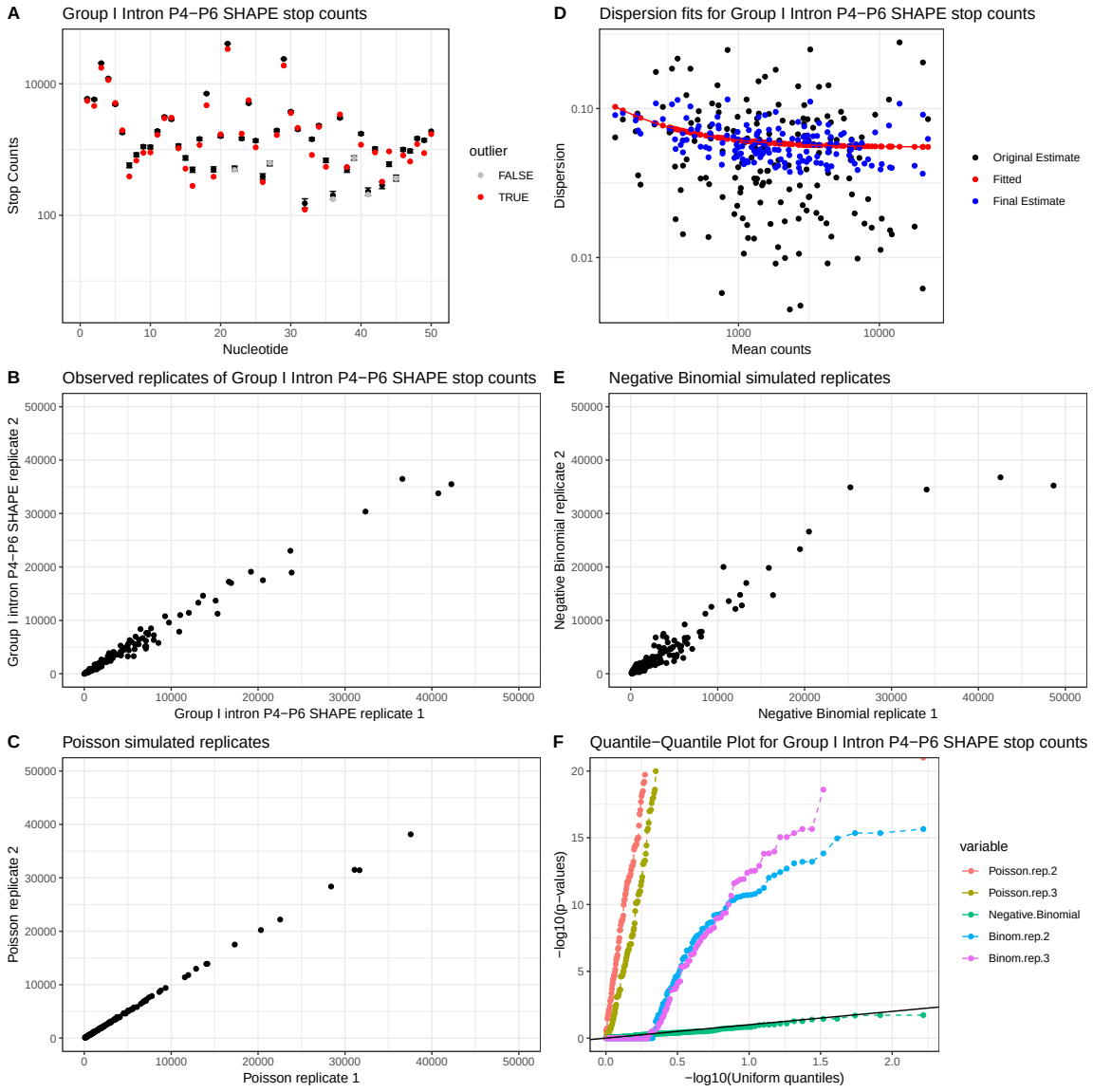

Figure 4.1: Analysis and modeling of overdispersion for *in vitro* SHAPE-Seq data for the *Tetrahymena* group I intron P4-P6 domain

Figure 4.1: Analysis and modeling of overdispersion for *in vitro* SHAPE-Seq data for the *Tetrahymena* group I intron P4-P6 domain

A) Visualization of normalized SHAPE treated counts from two replicates of SHAPE-Seq for the first 50 nucleotides of the P4-P6 domain of the *Tetrahymena* group I intron. 95% confidence intervals are drawn based upon the first replicate, and observations from the second replicate are colored based on whether they fall within the 95% confidence interval (gray) or are outliers (red). B) Normalized counts for two replicates of the P4-P6 domain of the *Tetrahymena* group I intron. C) Poisson simulated replicates of the P4-P6 domain of the *Tetrahymena* group I intron D) Trended fit for dispersion values for the P4-P6 domain of the *Tetrahymena* group I intron E) Negative binomial simulated replicates of the P4-P6 domain of the *Tetrahymena* group I intron, with parameters fit with DESeq2. F) Quantile-quantile plot for SHAPE-treated RT stop counts the group I intron SHAPE-Seq experiment. P-values of observed data against Poisson, binomial, and negative binomial models are compared to the uniform distribution on log10 scale. Two of three replicates were used to fit models and the final replicate was used for testing.

Modeling of overdispersion in chemical probing data

Having established that chemical probing data are overdispersed, we next sought to develop a more accurate way to model count data produced by chemical probing experiments that would take advantage of replicate observations. The natural choices for distributions to fit overdispersed Binomial or Poisson count data are the beta-binomial and negative binomial distributions, repsectively. In these distributions, the Binomial probability, p_i, or the Poisson mean, μ_i, are allowed to vary between observations, according to a Beta or Gamma distribution, respectively. The character of the underlying Beta or Gamma distributions enables modeling of inter-replicate variability, where no inter-replicate variability can be expressed by choosing a fixed value for these underlying distributions.

Fitting the above, more flexible distributions poses a challenge in many biological contexts, as relatively few replicates are typically conducted because of cost constraints, making it hard to make accurate variance estimates for each data point (nucleotide) individually. As we considered this problem, we noted that chemical probing techniques can be viewed largely as an extension of RNA-Seq experiments, where instead of counting reads at genes, RT events are counted at nucleotides. With the exception of chemical treatment, the key steps of the two techniques–reverse transcription, library preparation, and sequencing–are highly similar. Moreover, cost also limits the number of replicates produced for RNA-Seq experiments, and RNA-Seq data are well known to be overdispersed [Robinson and Smyth, 2007, Anders and Huber, 2010, Love et al., 2014]. We therefore considered whether we could adapt methods used for RNA-Seq analysis to model the overdispersion of chemical probing data.

To model overdispersion in chemical probing data, we chose to adapt the RNA-Seq analysis tool, DESeq2, which takes advantage of common information among many measurements (of gene expression) made in parallel to aid inference of count distributions [Love et al., 2014]. DESeq2 employs the negative binomial distribution, which

is closely relative to the Poisson distribution but contains a dispersion parameter, α, which is zero when there is no overdispersion (Poisson) but takes higher values when data are overdispersed (see Methods). DESeq2 estimates the dispersion parameter by first making estimates for each gene (or nucleotide for chemical probing), then observing a trend between mean counts and dispersion values, and finally adjusting dispersion values toward the trend (Fig 4.1d). Though DESeq2 can analyze normalized counts of any type, standard normalization for RNA-Seq is based upon the total number of reads in the experiment. In contrast, as above, we normalize input counts to the number of reads that reach the nucleotide of interest (for RT stops) or that cover the nucleotide of interest (for mutations).

We used DESeq2 to model normalized counts for our sample dataset: RT stop counts for *in vitro* SHAPE-Seq of the group I intron P4-P6 domain. We observe that as for RNA-Seq data, the dispersion parameters fit for each nucleotide trend with the mean counts (Fig 4.1d). To gain the same intuitive feel for the fit of the negative binomial models to the data, we compared simulated negative binomial replicates (Fig 4.1e) to observed replicates (Fig 4.1b) and Poisson replicates (Fig 4.1c), finding that the negative binomial replicates are much more similar to the real replciates than the Poisson replicates. To evaluate the negative binomial models produced by DESeq2 more formally, we fit a model using two replicates of the group I intron SHAPE-Seq data and computed negative binomial p-values given the model for the observations of a third replicate. As in our analysis of fit to the Binomial and Poisson distributions, if the negative binomial model matches the variability of the data, then the p-values for the third replicate relative to the model should follow the uniform distribution. We observe that our negative binomial p-values are much closer to following the uniform distribution than Poisson or Binomial p-values (Fig 4.1f), with no significant evidence that the negative-binomial p-values differ from expected uniform quantiles, in contrast to strong evidence for the binomial and Poisson distributions (Kolmogorov-Smirnov

test, Poisson and binomial: $p < 2.2 * 10^{-16}$; negative binomial: $p = 0.22$).

Comparing models of overdispersion across the range of publicly available chemical probing data

We next modeled the variety of chemical probing datasets of different types that we had found to be overdispersed relative to the Poisson and Binomial distributions. For each dataset, we fit binomial, Poisson, and negative binomial models. We used two metrics to compare the model fits: comparison of the Kolmogorov-Smirnov statistic for p-values of an outside dataset (as above) and the corrected Akaike Information Criterion (AIC_c) for a fit of each entire dataset. The Akaike Information Criterion (AIC) is a negative log likelihood-based metric penalized based on the number of parameters in the model. AIC is commonly used for evaluation of model fitting when separation of data into training and testing sets is difficult (e.g. when there are small numbers of replicates), and the corrected version of this metric adds an additional penalty when the total number of datapoints is small.

We applied these metrics to evaluate model fits across a set of publicly available chemical probing datasets conducted under different experimental conditions (*in vitro*, *in vivo*, and *ex vivo*), with different chemicals (SHAPE, DMS, and CMC (cyclohexyl-N'-(2-morpholinoethyl)carbodiimide metho-p-toluenesulfonate)), and with different readouts (RT stops and mutations). We also compared fits to chemical-treated samples to those for untreated controls. We observe by both metrics that the negative binomial model fits better than binomial and Poisson, even for untreated samples (Figure 4.3 and 4.2). With the corrected AIC metric, we also investigated different types of negative binomial model fits to determine the value of using DESeq2, in comparison to standard fitting methods, and comparing models fitting a single dispersion parameter to those using one per condition (treatment and control). We find that fitting with DESeq2 often leads to better results than other negative binomial models

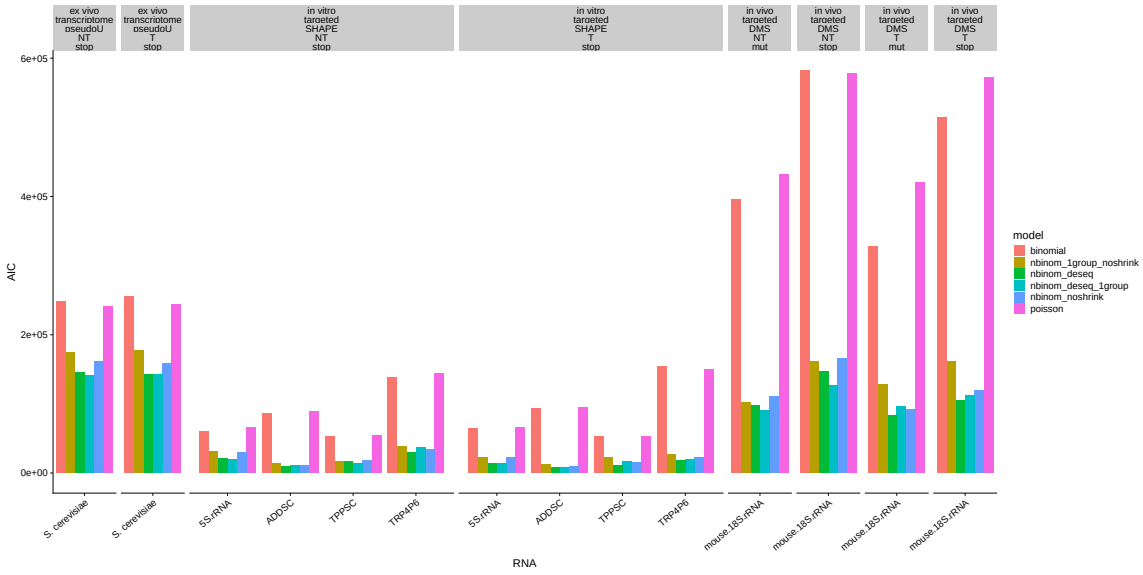

Figure 4.2: Investigation of overdispersion across datasets using corrected AIC

Analysis of model fitting to a variety of chemical probing datasets using corrected AIC.
Datasets: *in vivo*, targeted mouse 18S DMS-Seq [Sexton et al., 2017], *in vitro* SHAPE-Seq for 5S rRNA, Tetrahymena group I intron, TPP riboswitch (TPPSC), and Adenine riboswitch (ADDSC) [Loughrey et al., 2014], ex-vivo PSI-Seq in yeast [Schwartz et al., 2014]
Models:
Poisson: Poisson distribution fit to data from one condition
binomial: Binomial distribution fit to data from one condition
nbinom_deseq: Negative Binomial distribution fit using DESeq2 using treatment and control data (separate means for treatment and control, one dispersion parameter)
nbinom_noshrink: Negative Binomial distribution fit without DESeq2 shrinkage estimation, using treatment and control data.
nbinom_deseq_1group: Negative Binomial distribution fit with DESeq2, using only treatment OR control data.
nbinom_1group_noshrink: Negative Binomial distribution fit without DESeq2 shrinkage, using only treatment OR control data.

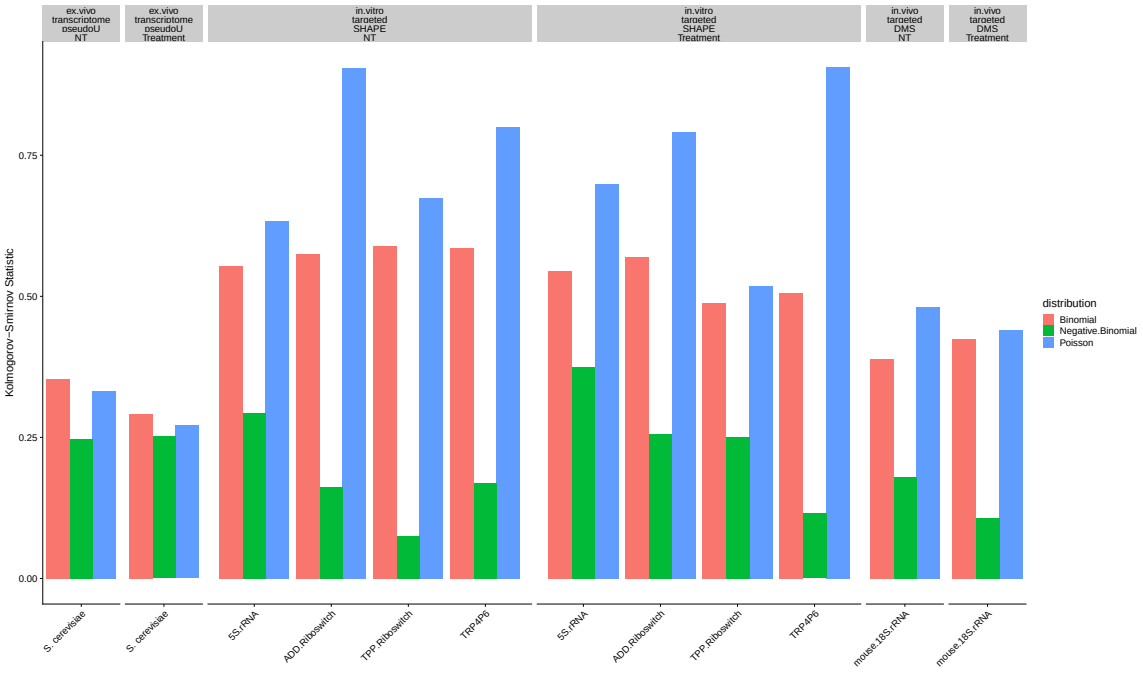

Figure 4.3: Investigation of overdispersion across datasets using Komogorov-Smirnov statistic

Analysis of model fitting to a variety of chemical probing datasets using the Kolmogorov-Smirnov statistics, when comparing p-values for a test dataset. **Datasets:** *in vivo*, targeted mouse 18S DMS-Seq [Sexton et al., 2017], *in vitro* SHAPE-Seq for 5S rRNA, Tetrahymena group I intron, TPP riboswitch (TPPSC), and Adenine riboswitch (ADDSC) [Loughrey et al., 2014], ex-vivo PSI-Seq in yeast [Schwartz et al., 2014]

using the AIC_c metric 4.2.

Modeling of overdispersion with 60 replicate reference datasets

To examine the overdispersion of chemical probing data in more detail and to assess our model more fully, we collected reference datasets with 60 biological replicates of *in vivo* DMS probing in mouse embryonic fibroblasts. All samples came from independently growing MEF cells. Reverse transcription with barcoded primers (see methods) enabled combination of groups of 20 replicates into single library preps. We collected two targeted datasets with primers specific to the mouse 7SK and GapDH RNAs, as well as an undirected experiment (random octamer primers).

We analyzed both RT stop and mutation readouts for our 60 replicate experiments in parallel, and we present results for RT stops in the mouse 7SK RNA dataset in Figure 4.4, along with corresponding analyses on other datasets in supplemental figures. To characterize the quality of our data, we plotted the correlations between RT event probability rates across the entire dataset, for both treatment and control (Figure 4.4a,b). For a sample set of nucleotides, we then visualized the Poisson and Negative Binomial distribution fits, in comparison to histograms of the real observed counts. These plots provide even clearer support for the conclusion that the negative binomial distribution fit with DESeq2 provides a better fit than the Poisson distribution (Figure 4.4c).

In addition to visualizing the overdispersion of chemical probing data more clearly and intuitively, our reference datasets enable us to evaluate our models in more detail and to investigate performance with different numbers of replicates. To do this, we chose anywhere between 2 and 30 of the 60 replicates to fit a variety of models and tested with 30 other replicates. We measured the quality of model fits using the total and median log likelihoods of our models against the test replicates (Figure 4.4c). These analyses confirm the improvement of the negative binomial model over Bino-

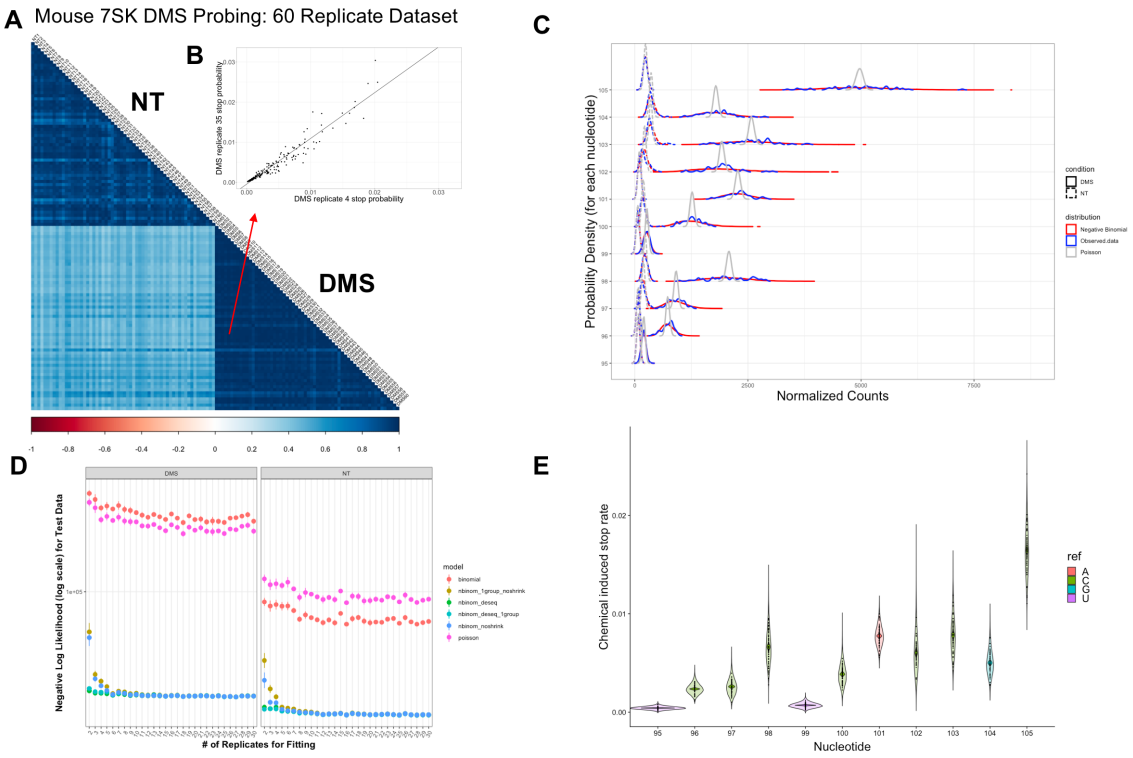

Figure 4.4: Observing and modeling overdispersion with 60 replicate datasets

Analysis of 60 biological replicate experiments of targeted *in vivo* DMS probing of the 7SK RNA in mouse embryonic fibroblasts. A) Correlations of probability of stopping at each nucleotide across the 60 replicate experiment, including treatment and control. B) Inset showing the relationship between stop rates in a sample pair of DMS-treated replicates. C) Ridgeline plots showing the distributions of real observed data, the Poisson model, and the negative binomial model, each represented using kernel density. D) Evaluation of binomial, Poisson, and negative binomial model fits, using anywhere from 2-30 replicates to fit the model and 30 different replicates to test. Model fitting to the test data is quantified as the total negative log likelihood of the test data. For the negative binomial models, estimation of one dispersion parameter per condition (treatment vs. control) is compared to fitting a single dispersion parameter across the entire experiment. Negative binomial fits with and without DESeq2 are also shown. E) Violin plots showing the inferred distributions of the chemical induced RT stop rate, $gamma_i$ for nucleotides 95-105 in the mouse 7SK RNA. Large violins represent the distribution of estimates, $gamma_{ij}$, of the chemical induced stop rate for individual replicates, while darker inset violins show the inferred distribution of $gamma_i$ with 60 observed replicates.

mial and Poisson. Moreover, when using the small numbers of replicates that are typically collected in typical experiments (e.g. [Mustoe et al., 2018, Rouskin et al., 2014, Ding et al., 2014] for fitting, the DESeq2 method models test data significantly better than negative binomial models with standard parameter fitting (this is true when modeling a single dispersion parameter α_i, or when fitting one parameter each for treatment and control: α_i^t and α_i^{nt}). Fits reach near maximum performance around 7-10 replicates (Figure 4.4c). We observe similar results for our other 60 replicate data (Figure 4.7).

To apply the inferences of our negative binomial count model to the interpretation of chemical probing experiments, it is desirable to consider not only the distributions of counts in treatment and control for a given nucleotide, but to quantify the influence of the chemical probe on the RT event rate. Others have shown that the chemical induced RT event rate, γ_i, for nucleotide i can be quantified as [Li et al., 2017]:

$$\gamma_i = \frac{p_i^t - p_i^{nt}}{1 - p_i^{nt}}$$

A key goal in modeling the data observed in chemical probing experiments is to determine the confidence of estimates of γ_i. To do this, we used our fit count distributions to infer the distribution, $p(\gamma_i \mid \mu_i^{nt}, \mu_i^t, \alpha_i, k)$, where k is the number of replicates collected. To do this, we repeatedly sampled the chosen number of replicates, k, (in this case, 60 - the number of replicates collected in the experiment) as below and then computed the chemical induced RT event rate for each sampled dataset.

$$X_i^t \sim NB(\mu_i^t, \alpha_i)$$

$$X_i^{nt} \sim NB(\mu_i^{nt}, \alpha_i)$$

We visualize our inferred distributions of, γ_i, of chemical induced RT stop rates for nucleotides 75-85 of the the mouse 7SK RNA using violin plots (Figure 4.4). We plot the inferred distributions both of individual estimates γ_{ij} from different replicates (open violins, $p(\gamma_i \mid \mu_i^{nt}, \mu_i^t, \alpha_i, k = 1)$) and the inferred distribution of γ_i from 60 replicates ($p(\gamma_i \mid \mu_i^{nt}, \mu_i^t, \alpha_i, k = 60)$). To complement, we also plot the 60 individual estimates from our observed replicate experiments. This visualization illustrates the value of collecting replicates to improve estimates of the most biochemically meaningful parameters from chemical probing experiments, particularly when data are overdispersed (Figure 4.4e).

Effects of variability in RNA chemical probing data on RNA secondary structure predictions

Having established that RNA chemical probing data are overdispersed and developed a way to model the distributions in these data, we became interested in the effect of the variability in these data on one of the most common applications of RNA chemical probing: secondary structure prediction [Eddy, 2014, Deigan et al., 2009]. Results from chemical probing data are typically incorporated into RNA secondary structure predictions as pseudoenergy constraints that are added to the existing nearest neighbor constraints, referred to as Turner rules, trained on thermodynamic melting experiments on small RNAs [Turner et al., 1988, Mathews et al., 1999, Eddy, 2014]. The incorporation of constraints from chemical probing experiments, especially those using SHAPE reagents, has enabled significantly improved predictions of the secondary structures of a variety of RNAs with known structures, including rRNAs, tRNAs, and riboswitches [Deigan et al., 2009, Watters et al., 2018, Lucks et al., 2011, Watters et al., 2018]. Because commonly used structure algorithms employ a fixed set of energy parameters to guide structure predictions [Zuker et al., 1999, Zuber et al., 2017], it is particularly important to understand how variability in these parameters may

influence prediction results. Indeed, a variety of efforts have been made to investigate how different sets of nearest-neighbor energy parameters affect RNA secondary structure predictions [Zuber et al., 2018, Zuber et al., 2017]. Similar efforts seem warranted for pseudoenergy values derived from chemical probing experiments, particularly given that in transcriptome-wide experiments, one is virtually guaranteed to collect high noise data for transcripts with low expression (due to lack of sequencing coverage, even if there is no overdispersion), in addition to more confident data for highly expressed transcripts.

To investigate the effect of the variability in RNA chemical probing data on secondary structure predictions, we followed standard methods to normalize chemical induced stop or mutation rates for each nucleotide, γ_i into reactivity values R_i (see Methods). We then incorporated these reactivity values into predictions as pseudoenergy terms using the function developed by Deigan and colleagues [Deigan et al., 2009].

$$E(R_i) = a * log(R_i + 1) + b$$

Where a and b are parameters that were set based upon empirical performance in previous studies on RNAs of known structure. We use this approach to perform both predictions of single RNA secondary structures, and we focus particularly on base pair probability matrices (BPPM) that can be predicted over the thermodynamic ensemble of structures the RNA takes on according to the model (using the McCaskill algorithm) [Mathews, 2004, McCaskill, 1990]. Base pair probability matrices complement individual structures by providing information about the overall structural landscape, which can be particularly important if there are multiple energetically accessible structures or if errors in the structure prediction parameters lead to an incorrect single predicted structure. The uniformity of the predicted structural landscape, i.e. whether the RNA is predicted to favor one structure dominantly or

samples multiple competing structures with similar energies, can be quantified by calculating the Shannon entropy of the base pair probability distributions, defined as:

$$S = \sum_{ij} -q_{ij} log(q_{ij})$$

Where q_{ij} is the predicted pairing probability for bases i and j in the pairing probability matrix and q_{ii} is the probability that a given base is unpaired. It has been established that RNA regions with low Shannon entropy are more likely to have correct predicted minimum free energy structures [Huynen et al., 1997]. Further emphasizing the importance of the base pair probability matrix, base pairs predicted to have high probabilities within the structural ensemble are more likely to be predicted correctly [Mathews, 2004]. Base pair probability matrices have also been used to help gain biochemical insights, as many RNAs of known function display a combination of low SHAPE reactivity and low Shannon entropy and RNA elements with this combination of features have been suggested to be candidate functional elements [Siegfried et al., 2014]. The above modes of interpretation of RNA base pair probability matrices rely on the assumption that the BPPM is predicted with some level of confidence, adding to our motivation to assess the influence of variability in RNA chemical probing data on these structure predictions.

As a case study for our analysis of variability in RNA chemical probing data on secondary structure predictions, we used a transcriptome-wide SHAPE-MaP dataset collected in *E. coli* [Mustoe et al., 2018]. We started by predicting BPPMs using either the SHAPE reactivities from each of two different replicate experiments, as well as using the pooled results of those experiments (Figure 4.5).

We would expect that in cases where transcripts have highly reproducible reactivity values, they would have very similar predicted BPPMs. In contrast, we

might expect more differences between predicted BPPMs when the reactivities for a given RNA are less reproducible. As a first test of this hypothesis, we plotted the reactivities and predicted BPPMs for each replicate of the SHAPE-MaP experiment for ncRNA rnpB-the RNA component of *E. coli* RNAse P-which has very high expression ($> 10^6$ mean reads per nucleotide), and the panD mRNA, which has lower expression ($\sim 10^3$ mean reads per nucleotide), but still qualified for the filtering requirements to be included in the folding analysis originally conducted on the dataset [Mustoe et al., 2018]. We plotted the BPPM using arcs that represent pairing probabilities between nucleotides, q_{ij}, colored by the value of q_{ij}. The reactivities for rnpB are highly correlated and correspondingly we see high visual similarity between BPPMs predicted from the replicate datasets (Figure 4.5). In contrast, the panD mRNA has less correlated reactivities and more differences between predicted BPPMs. Strikingly, when comparing the predicted BPPMs for panD, we see mutliple base pairs for which the pairing probability, q^1_{ij} is very high in one replicate and the value, q^2_{ij}, is barely above zero in the other. This establishes clearly that there can be cases in which the BPPMs predicted using pseudoenergy constraints from chemical probing data cannot be interpreted with the same assumptions that are often used for constraint-free predictions or predictions made with high-confidence SHAPE data.

To quantify differences between BPPMs,we computed the root mean square deviation across all elements of the matrices (see methods), a metric previously used by Zuber and colleagues to compare predicted structures with different sets of nearest neighbor constraints [Zuber et al., 2017]. Consistent with our visual observations, we see that the rnpB matrices have a much lower RMSD between replicates than the panD matrices. We then computed the RMSD between BPPMs predicted based on data from each replicate for all 186 transcripts that were included in the original publication of the dataset. We observe strong relationships between the mean read coverage across a given transcript and the RMSD between predicted BPPMs

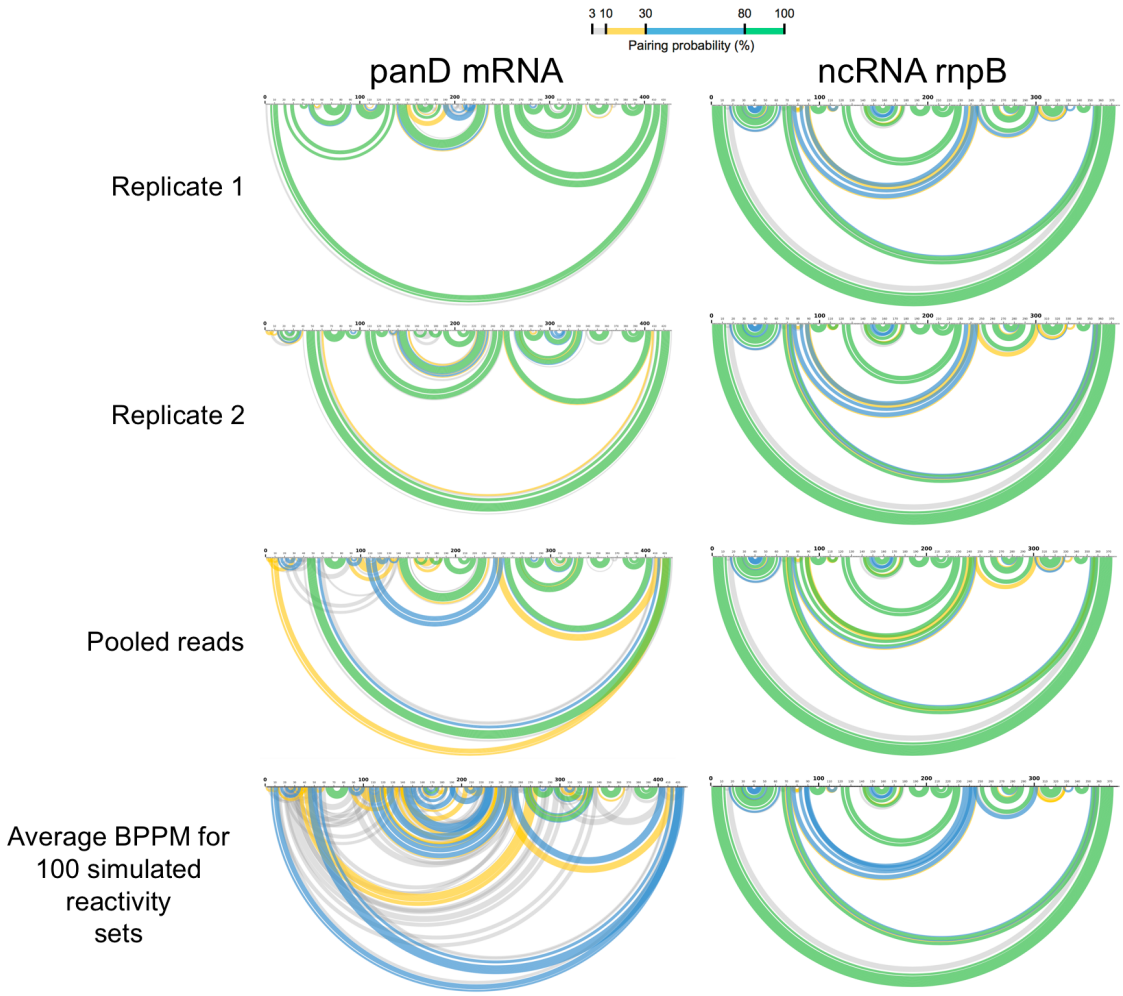

Figure 4.5: Comparison of predicted base pair probability matrices with different sets of chemical probing constraints

Predicted base pair probability matrices for the panD mRNA and ncRNA rnpB are plotted with arcs representing pairs between bases and colored by probability. Different predictions are made with the following sets of constraints, all derived from a transcriptome-wide SHAPE-MaP experiment in *E. coli*, or subsequent modeling of observed data. Row 1: Predictions made with constraints derived from replicate 1. Row 2: Predictions made with constraints derived from replicate 2. Row 3: Predictions made with constraints derived from the pooled results of both replicates. Row 4: Averaged of 100 predicted base pair probability matrices, each using a set of constraints simulated using our negative binomial model, fit with DESeq2.

(Figure 4.6a). We observed a similarly strong relationship between RMSD and the mean signal-to-noise ratio of the reactivity values across replicates, a metric suggested for use in analysis of chemical probing by Choudhary and colleagues (Figure 4.6b) [Choudhary et al., 2016].

4.2.4 Investigating the contribution of variability in chemical probing reactivity to the uncertainty in the predicted RNA thermodynamic landscapes

Having observed that variability in chemical probing data can strongly affect predictions of BPPMs, we sought to use our count models to help investigate this phenomenon. Although the sturcture prediction algorithms we use here employ a single set of thermodynamic (and pseudoenergy) parameters at a time, it is feasible to make mutliple predictions with different sets of reactivity values. Based on the negative binomial parameters we fit to the count distributions, we sampled reactivity values R_i similarly to the above sampling of γ_i (using our fit count distributions), but adding an additional normalization step (see methods). For each sample reactivity vector R^m, representing reactivity values across the RNA, we computed the base pair probability matrix, Q^m. We then averaged the individual estimates, $\vec{R}^m$, to obtain a final estimate of the BPPM, Q^{samp}.

$$Q^{samp} = \sum Q^m(\vec{R}^m)M$$

Where M is the total number of sampled sets of reactivities. Under the assumption that averaging is an effective way to summarize the predicted BPPMs with different sampled reactivity vectors, Q^{samp} represents an estimate of the BPPM for the RNA, incorporating the range of possible SHAPE reactivities for the RNA according to our negative binomial model. We computed Q^{samp} for the rnpB and panD RNAs, sam-

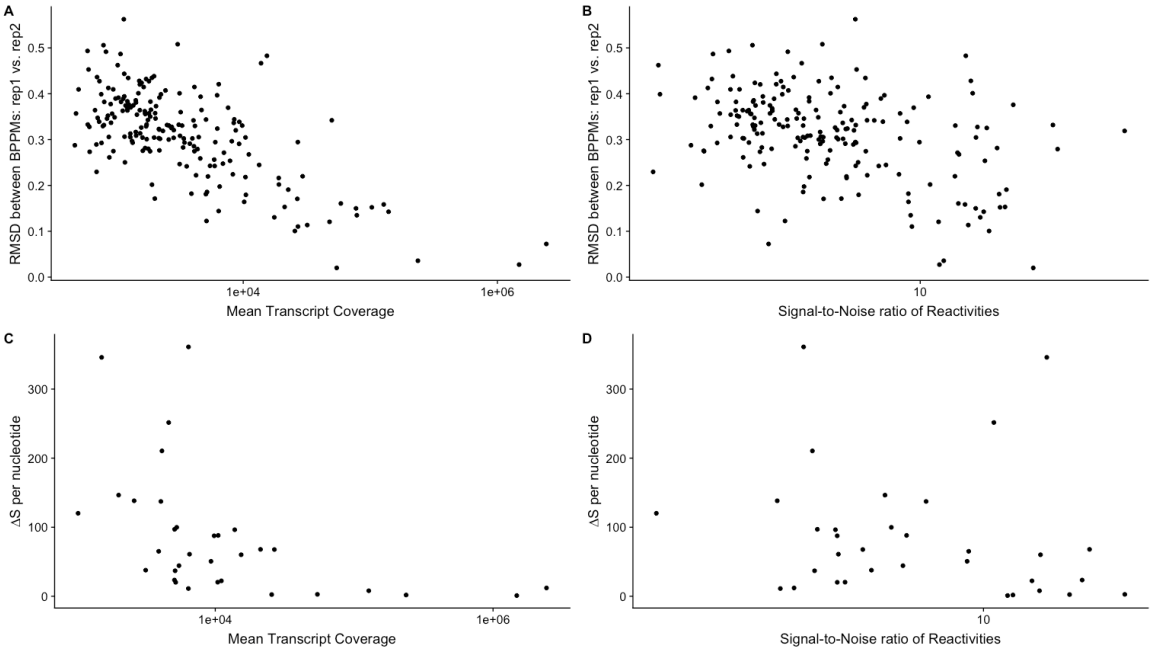

Figure 4.6: Comparing characteristics of chemical probing data to consistency of resulting predictions

A) Relationship between RMSD between predicted base pair probability matrices for different replicates and mean transcript coverage. B) Relationship between RMSD between predicted base pair probability matrices and signal-to-noise ratio of reactivity values. C) Relationship between ΔS metric of contribution of chemical probing uncertainty to breadth of predicted structural landscape and mean transcript coverage. D) Relationship between ΔS metric and signal-to-noise ratio of reactivity values.

pling over 100 reactivity sets, and compared these predicted BPPMs to the estimates of Q^1 and Q^2 from individual replicates. We see that Q^{samp} for panD has many more base pairs with intermediate predicted probabilities in the thermodynamic ensemble and many fewer base pairs with very high or very low probabilities, relative to the individual estimates Q^1 and Q^2. Correspondingly, the entropy of Q^{samp} is much greater than that of Q^1 and Q^2 for panD (191 vs. 75 bits), while the difference is much smaller for rnpB (41 vs. 40 bits). Making a prediction with a single set of reactivities leads to a clear underestimate of the predicted heterogeneity of the structural ensemble, and this effect is particularly seen when the reactivities estimates have high uncertainty.

We reasoned that the difference between the entropy of Q^{samp} and Q^{mean}, referred to as ΔS, would help to quantify the increase in diversity of structure predictions that arises from uncertainty in estimation of chemical probing reactivities.

$$\Delta S = S(Q^{samp}) - \sum_m \frac{S(Q^m)}{M}$$

We computed ΔS for all *E. coli* transcripts with greater than 1000 reads per nucleotide and with lengths below 1000 nucleotides (to aid computational efficiency), and as observed with RMSD between replicate BPPMs (Q^1 and Q^2), there is a relationship between ΔS and both read coverage and reactivity signal-to-noise ratio (Figure 4.6c,d).

4.2.5 Discussion

Here, we have investigated the overdispersion of RNA chemical probing data and the effects of this increased statistical variability on RNA secondary structure prediction. We have leveraged machinery developed for RNA-Seq analysis to fit negative binomial models to chemical probing data, despite the small number of replicates that are often collected in these experiments. Our negative binomial models fit with DESeq2

demonstrate better fits to a wide variety of chemical probing data and also to no treatment controls. We further collected reference datasets with 60 replicates, which enabled clearer demonstration of overdispersion and careful comparison of different methods of fitting the negative binomial distribution and their relative performance with small numbers of replicates.

Beyond modeling of chemical probing count data, we were interested in the effects of variability in these experiments on RNA secondary structure prediction. We first show that when chemical probing data are noisy, one can get conflicting predicted BPPM with results from different replicate experiments, even though individual predicted matrices may have low Shannon entropy and represent fairly homogeneous structural landscapes. We propose to incorporate variability in chemical probing data into BPPM predictions by averaging together individual BPPMs based on reactivities sampled based on our count distributions. These averaged BPPMs would, in principle, be compatible with structure prediction alorithms, e.g. MaxExpect, although we do not yet have any evidence to indicate whether this would be a successful approach. Based on our averaged BPPMs, we also propose that measuring the difference in Shannon entropy between the averaged BPPM and BPPMs predicted from individual reactivity sets can help indicate how much variability in chemical probing data contributes to the breadth of the predicted structural landscape.

The above approach to RNA secondary structure prediction would be compatible with other methods to model raw data from chemical probing experiments, and we believe that several improvements to our method may be possible. First, though our model captures interreplicate variability in RT event counts in chemical probing experiments, it does not adjust extreme estimates of the chemical induced RT event rate (γ_i) based on the level of count uncertainty. Other groups have proposed a variety of Bayesian approaches that address this problem in some way, but do not model overdispersion [Selega and Sanguinetti, 2016, Ledda and Aviran, 2018,

Radecki et al., 2018]. Aiding these efforts has been the use of hidden Markov models that take advantage of the fact that adjacent nucleotides often have similar structural (or other) properties. Second, in our inference of chemical induced stop rate and reactivity distributions, we make the assumption that the negative binomial parameters fit with DESeq2 are correct, rather than having their own posterior distributions with respect to the observed data. Third, our model does not account for specific experimental factors that may change between replicates. For example, it would be easy to imagine that certain samples have more total chemical treatment than others. A parameter modeling this kind of effect might be replicate specific but common to all nucleotides. Finally, we have modeled counts of RT stops and mutations separately. While this approach is common, it would be ideal to develop a way to integrate aspects of chemical probing experiments together to make inferences about RNA properties, especially because RT stops and mutations can sometimes contain orthogonal information. In addition to better modeling of probing data, we have investigated the effects of variability in chemical probing data on one approach to incorporating these data into secondary structure prediction, but many approaches exist [Eddy, 2014] and improved prediction performance might be obtained by considering the best pairing of modeling raw data and combination with existing RNA structure prediction frameworks.

Though this study highlights unrecognized variability in RNA chemical probing data, it is notable that many studies have produced important biochemical insights while making some simplifying assumptions in anlayzing probing data (e.g. [Siegfried et al., 2014]). Our analysis of RNA secondary structure prediction indicates that even though the SHAPE-Seq data are overdispersed, structure predictions can be robust to the choice of analysis method (or which replicate dataset is used for modeling) if the data collected are generally confident (e.g. due to a combination of high read coverage and high overall reproducibility between experiments). Further

improved modeling of chemical probing data, in addition to improved experimental methods, has the potential to extend the power of this technology to more different RNAs and RNA properties.

4.2.6 Supplemental Figures

4.3 Methods

4.3.1 Experimental methods

For our 60 replicate reference datasets, we included 6 nucleotide barcodes on our reverse transcription primers, between the 3' Illumina adapter and the primer sequence itself. Below is an example primer targeting the mouse GapDH gene:

CAGACGTGTGCTCTTCCGATCT **TTGACT** CATCGAAGGTGGAAGAGTGGG

We designed sets of 20 barcodes for each RT primer (N8, targeted to mouse 7SK, and targeted to mouse GapDH).

This enabled pooling of groups of 20 samples after reverse transcription. Growth of mouse embryonic fibroblasts, DMS probing, and library preparation were conducted as in [Sexton et al., 2017].

4.3.2 Sample demultiplexing, read alignment, and quantification

Demultiplexing of barcodes

Since barcodes in the 60 replicate experiment were sequenced within the core portion of the Illumina reads, we performed demultiplexing using an in-house script that required no more than 1 mismatch (all barcodes had at least 2 mismatches relative to one another).

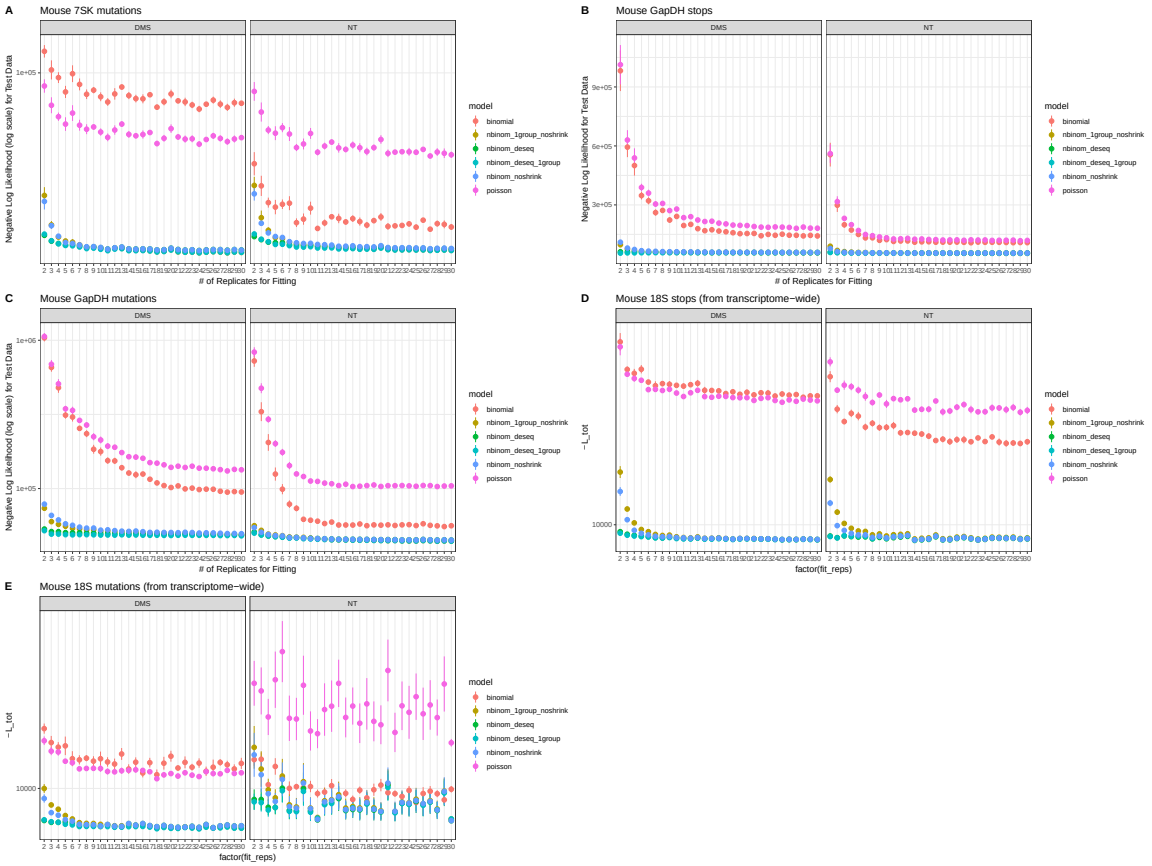

Figure 4.7: Evaluating fit of models using 60 replicate datasets

Evaluation of binomial, Poisson, and negative binomial model fits, using anywhere from 2-30 replicates to fit the model and 30 different replicates to test. Model fitting to the test data is quantified as the total negative log likelihood of the test data. For the negative binomial models, estimation of one dispersion parameter per condition (treatment vs. control) is compared to fitting a single dispersion parameter across the entire experiment. Negative binomial fits with and without DESeq2 are also shown. A) Mouse 7SK mutations B) Mouse GapDH stops C) Mouse GapDH mutations D) Mouse 18S rRNA stops (from a transcriptome-wide dataset) E) Mouse 18S rRNA mutations (from a transcriptome-wide dataset)
Models:
Poisson: Poisson distribution fit to data from one condition
binomial: Binomial distribution fit to data from one condition
nbinom_deseq: Negative Binomial distribution fit using DESeq2 using treatment and control data (separate means for treatment and control, one dispersion parameter)
nbinom_noshrink: Negative Binomial distribution fit without DESeq2 shrinkage estimation, using treatment and control data.
nbinom_deseq_1group: Negative Binomial distribution fit with DESeq2, using only treatment OR control data.
nbinom_1group_noshrink: Negative Binomial distribution fit without DESeq2 shrinkage, using only treatment OR control data.

Alignment Reads were aligned to the transcript of interest for targeted experiments and to transcriptome references for undirected experiments, and directly to desired transcripts for targeted experiemnts, using Bowtie2 [Langmead and Salzberg, 2012].

Quantification and filtering Reverse transcription events - both stops and mutations, along with normalizing read coverages - were quantified with an updated version of our previously published RTEventsCounter script [Sexton et al., 2017].

After intial quantification, for our in-house datasets and other targeted-structure-seq data, we filtered out all RT stops (but not mutations) less than 100 nt from the RT primer, because previous work in our lab had shown that many short fragments are lost, likely in the process of biochemical purification during library preparation.

4.3.3 Public datasets

We used the following public datasets in this study:

1. Transcriptome-wide SHAPE-MaP data in *E. coli*: Data were downloaded from . Reads were aligned to the reference provided by the Kevin Weeks group in their publication [Mustoe et al., 2018].

2. Mouse 18S rRNA DMS-Seq data were obtained from [Sexton et al., 2017].

3. SHAPE-Seq data: Data for four of the RNAs published by Loughrey et al. (5S rRNA, adenine riboswitch, TPP riboswitch, and group I intron) were downloaded from the RNA Mapping Database (RMDB) [Cordero et al., 2012].

4. Ψ-Seq data in yeast were downloaded from the Gene Expression Omnibus (GSE60047) [Schwartz et al., 2014].

4.3.4 Definitions of count data in RNA chemical probing

In chemical probing experiments, we count so-called events that can occur during reverse transcription, specifically either reverse transcription stops (termination of cDNA synthesis), or mutations in the resulting cDNA.

For each type of event (stops or mutations), we can view the processing of each nucleotide as a Bernouli trial, with some probability, p_i of producing the reverse transcription event of interest. Depending on the model being considered, p_i may be constant or vary within or between replicates.

Let Y_{ij} represent the reverse transcription event counts for nucleotide i in sample j, with N total nucleotides and M total samples. Each sample j can come from either chemical-treated or control RNA.

For mutations, the coverage, C_{ij} is simply the number of reads directly covering the nucleotide of interest.

For stops, the coverage is the number of reads that reads the nulceotide of interest (and therefore could have stopped at the position of interest):

$$C_{ij} = \sum_{k=i}^{N} Y_{kj}$$

4.3.5 Count normalization for modeling with Poisson-family distributions

The combination of RT event counts, Y_ij and normalizing coverage, C_ij, representing the number of reads Bernouli trials for the event type of interest, can be modeled using the Binomial and related distributions.

To enable modeling with Poisson-family distirbutions, including both Poisson and Negative Binomial, we normalized counts relative to coverage. To do this, we define pseudocounts K_{ij} that are scaled by coverage, such that every nucleotide as the same

final effective coverage D_i.

Pseudocounts:

$$K_{ij} = \frac{Y_{ij}}{C_{ij}} \frac{\sum_{k=1}^{M} C_{ik}}{M}$$

Effective coverage:

$$D_i = \frac{\sum_{k=1}^{M} C_{ik}}{M}$$

4.3.6 Count models

We can model our count data in the following ways:

1. Binomial: Reads are pooled between replicates for treatment and control and p_i^t and p_i^{nt} are computed from counts as below.

$$Y_i^t \sim Binomial(p_i^t, C_i^t)$$

$$Y_i^{nt} \sim Binomial(p_i^{nt}, C_i^{nt})$$

2. Poisson: Pseudocounts are for treatment and control are fit to separate Poisson models.

$$K_i^t \sim Poisson(\mu_i^t)$$

$$K_i^{nt} \sim Poisson(\mu_i^{nt})$$

Parameters are estimated for the Binomial and Poisson distributions using the R glm function.

3. Negative Binomial:

$$K_i^t \sim NB(\mu_i^t, \alpha_i)$$

$$K_i^{nt} \sim NB(\mu_i^{nt}, \alpha_i)$$

Parameters are estimated using DESeq2 [Love et al., 2014], which first fits parameters μ_i^t, μ_i^t, α_i^{init}, for each nucleotide independently by maximing the Cox-Reid likelihood of the data under the above negative binomial model (see [Love et al., 2014]).

Subsequently, all dispersion parameters α^{init} are fit to a trend of the form:

$$\alpha^{tr}(\mu) = \frac{a_1}{\mu} + a_0$$

Where μ_i is the mean of all counts, independent of treatment condition.

Finally, α_i^{MAP} is obtained by maximizing the sum of the Cox-Reid likelihood and a log-normal prior of the form:

$$\alpha_i \sim Normal(log(\alpha^{tr}(\mu), \sigma_d^2))$$

Where σ_d^2 is determined based upon the number of degrees of freedom in the dataset (see [Love et al., 2014] for more detail).

For comparison, in some analyses we also compare the results of the DESeq2 model fit described above with negative binomial models fit with only the results of one condition and/or without the DESeq2 shrinkage estimation procedure for the dispersion parameter.

4.3.7 Event probabilities and inference of the chemical-induced RT event rate

We can infer RT event probabilities both in individual samples and across replicates, using either counts and coverage or pseudocounts and pseudocoverage.

The event rate for each individual sample and condition is:

$$\hat{p_{ij}} = \frac{Y_{ij}}{C_{ij}} = \frac{K_{ij}}{D_i}$$

We can further compute the mean event probability in each condition (nt or t, referring to no treatment and treatment, respectively):

$$\hat{p_i^{nt}} = \sum_j \frac{Y_{ij}^{nt}}{C_{ij}^{NT}}$$

$$\hat{p_i^{t}} = \sum_j \frac{Y_{ij}^{t}}{C_{ij}^{t}} =$$

$\hat{p_i^{nt}}$ is the natural RT event rate, independent of the chemical. We can further compute an estimate of the chemical-induced RT event rate, which is the probability that an RT event occurs due to the chemical, as:

$$\hat{\gamma_i} = \frac{\hat{p_i^t} - \hat{p_i^{nt}}}{1 - \hat{p_i^{nt}}}$$

4.3.8 Model comparisons

We compared the fits of various count models with a variety of metrics.

1. Corrected Akaike Information Criterion (AIC_c) - This metric is meant to estimate the relative quality of fits to data without separating data into training and

testing sets. AIC_c is based on the more widely used AIC metric which is based on the likelihood of the data, penalized for the the number of parameters, k.

$$AIC = 2k - 2ln(\hat{L})$$

Where $\hat{L}$ is maximum value of the log likelihood function. To avoid undue influence of outliers, we set the minimum value of $\hat{L}$ for any individual data point to 10^{-100}. AIC_c is an modified version of AIC intended for when the number of data points available for fitting:

$$AIC_c = 2k - 2ln(\hat{L}) + \frac{2k^2 + 2k}{n - k - 1}$$

Where n is the number of data points used to fit the model. In our case, n would be the number of data points for a given nucleotide i. AIC_c converges to AIC when n is large.

2. Comparison of p-values for data held out in a test set – If one simulates data from a given distribution, the p-values of those data relative to the source distribution should themselves follow a uniform distribution. We test this assumption visually by making quantile-quantile plots and compare the assumption quantitatively by comparing Kolmogorov-Smirnov test results between each set of p-values computed from a given model and the uniform distribution.

3. Log-likelihood of test data – To complement the AIC metric, which does not incorporate separation of training and testing data and the KS-test, which considers the rank order of p-values, we compare the total log-likelihoods of different models on test datasets not used for fitting.

4.3.9 Using count models to infer the distribution of the chemical-induced RT event rate

Once we fit a model to observed counts (or pseudocounts) of our data, we can simulate data from the model to infer the distributions of quantities of interest under different data collection circumstances. We did this with the negative binomial model as fit with DESeq2. To sample chemical induced RT event rate values γ_i from a based on the fits to a given set of data, we repeatedly sample an identical number of replicates, k, as had been collected from the distribution of interest and then compute γ_i for each set of sampled replicates.

Sample k replicates from fit distributions:

$$K_i^{t^m} \sim NB(\hat{\mu_i^t}, \hat{\alpha_i})$$

$$K_i^{nt^m} \sim NB(\hat{\mu_i^{nt}}, \hat{\alpha_i})$$

$$\gamma_i^m = \sum_{j=1}^{k} \frac{p_{ij}^t - p_{ij}^{nt}}{1 - p_{ij}^{nt}}$$

4.3.10 Normalization of chemical induced RT event rate to generate reactivity values for RNA secondary structure prediction

The chemical induced RT event rate values $gamma_i$ for chemical probing data are often normalized to a common scale before being used to provide constraints to secondary structure prediction. This both controls for differences in overall degree of modification, and provides a consistent value to relate to structural properties.

Here we define reactivity as:

$$R_i = \frac{\gamma_i}{c}$$

Where c is a normalization factor, equal to the average of the top 10% of datapoints, after excluding any datapoints greater than the 1.5 times the interquartile range. A maximum of 10% of datapoints are allowed to be designated as outliers in normal datasets, and a maximum of 5% may be designated outliers in datasets with fewer than 100 nucleotides. This has been referred to as the "Boxplot method" [Deigan et al., 2009, Sloma and Mathews, 2015].

4.3.11 RNA secondary structure prediction

Constraints from chemical probing experiments have been incorporated into RNA secondary sturcture predictions by relating normalized reactivity values R_i at each nucleotide to energetic penalties for being paired or unpaired. These p̈seudoenergyẗerms can also be interpreted as probabilities of being paired or unpaired using the Boltzmann distribution. A common function used to relate reactivities and pseudoenergies is:

$$E_i(R_i) = a * log(R_i + 1) + b$$

Pseudoenergy terms can be added to the nearest neighbor parameters used to score structures in most thermodynamic RNA secondary structure prediction algorithms. We use the parameter values $a = 1.8$ and $b = -0.6$ that are the defaults in the RNAstructure package [**?**].

Here, we particularly focus on the results of the McCaskill algorithm, which enables calculation of the base pair probability matrix of the entire thermodynamic ensemble of an RNA of interest, under the assumptions of the thermodynamic parameters. Without describing the full details of the model, we note that we can produce an output prediction of the pairing probability matrix Q, where Q_{ij} is the probability of pairing of bases i and j in an RNA of interest as a function of the

McCaskill algorithm, additional parameters (a, b).

$$Q = F(\vec{R}, a, b, etc)$$

Where $\vec{R}$ is the vector of all reactivity values R_i across the RNA. For convenience below, we say that F_{ij} produces just the ij^{th} element of Q.

$$Q_{ij} = F_{ij}(\vec{R}, a, b, etc)$$

4.3.12 Estimation of RNA base pair probability matrices based upon posterior reactivity distributions

Though individual runs of the structure prediction algorithms that we use employ a single set of energetic parameters, the parameters themselves may be uncertain. The effects of variability in the experiments used to calculate nearest neighbor parameters on RNA secondary structure prediction have been investigated recently [Zuber et al., 2018, Zuber et al., 2017], and here we describe signifcant differences in structure predictions that can occur with different collected sets of chemical probing reactivities (see main text).

If we have a posterior distributions of all R_i, then we can sample M estimates, $\vec{R^m}$, each of which will produce a base pair probability matrix Q^m. If the posterior distributions of R_i have high variance, we may get a better estimate Q^{samp} by averaging all M predicted matrices as follows:

$$Q_{ij}^{samp} = \frac{\sum_{m=1}^{M} F_{ij}(\vec{R^m}, a, b, etc)}{M}$$

4.3.13 Metrics for comparison of base pair probability matrices

1. RMSD - Root mean squared deviation of all base pair probabilities

$$RMSD = \sqrt{\frac{\sum_{ij} (Q^1_{ij} - Q^2_{ij})^2}{N}}$$

Where Q^1 and Q^2 are two preicted base pair probability matrices for the same RNA of length N.

2. ΔS - Difference in entropy between the mean matrix of sampled reactivities, Q^{samp}, and the mean of the entropy of pairing probability matrices Q^m, predicted from individual sampled reactivities.

$$\Delta S = S(Q^{samp}) - \sum_m \frac{S(Q^m)}{M}$$

Chapter 5

Conclusion

In this thesis, I have presented an overview of the field of long noncoding RNA biology and three vignettes showing analysis of experimental technologies that seek to uncover noncoding properties of RNA, ranging from chromatin occupancy to turnover rate to molecular structure and modifications. The proliferation of technologies that consist of a combination of biochemical manipulation and readout with high throughput sequecing promises to increase the range of properties that we can measure, both to aid understanding of how known functional RNAs perform their cellular roles and to profile an increasing number of candidate RNAs whose functions may be revealed in part through integrative analysis. Beyond finding function in genuinely noncoding RNAs, noncoding properties of mRNAs can have important regulatory roles. Recent work in the Gerstein lab (to which I contributed slightly) sought to integrate a variety of transcriptome-wide assays, with a focus on CLIP-Seq assays to help identify mutations in RNA that disrupt RNA-protein interactions, or other RNA functions [Zhang et al., 2018].

An interesting question to consider in this context is: how many more functional RNAs might we expect to discover in the human (or any other eukaryotic) genome? A variety of highly conserved RNA classes that exist in eukaryotes (and throughout

life) are well known, e.g. rRNA, tRNA, snRNA, and snoRNA [Cech and Steitz, 2014]. Though these are some of the most common genes throughout the evolutionary tree, functional classes of RNA have been shown to follow a power law distribution, implying that there may be many undiscovered, but rare, functional RNAs [McCown et al., 2017]. This thinking implies that evolutionary analysis may have limits when looking for rare functional RNAs that may exist within the so-called lncRNA class. However, it is promising to consider that evolutionary analysis has been limited in its ability to help us understand the role of the Xist RNA, which appears relatively poorly conserved on a sequence level. This implies that continued biochemical and genetic study may be critical to the discovery of more functional human (and eukaryotic) RNAs.

Bibliography

[Anders and Huber, 2010] Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.

[Arab et al., 2014] Arab, K., Park, Y. J., Lindroth, A. M., Schafer, A., Oakes, C., Weichenhan, D., Lukanova, A., Lundin, E., Risch, A., Meister, M., Dienemann, H., Dyckhoff, G., Herold-Mende, C., Grummt, I., Niehrs, C. and Plass, C. (2014). Long noncoding RNA TARID directs demethylation and activation of the tumor suppressor TCF21 via GADD45A. Mol. Cell *55*, 604–614.

[Aviran and Pachter, 2014] Aviran, S. and Pachter, L. (2014). Rational experiment design for sequencing-based RNA structure mapping. RNA *20*, 1864–1877.

[Bail et al., 2010] Bail, S., Swerdel, M., Liu, H., Jiao, X., Goff, L. A., Hart, R. P. and Kiledjian, M. (2010). Differential regulation of microRNA stability. RNA *16*, 1032–1039.

[Baltz et al., 2012] Baltz, A. G., Munschauer, M., Schwanhausser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M., Wyler, E., Bonneau, R., Selbach, M., Dieterich, C. and Landthaler, M. (2012). The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. Mol. Cell *46*, 674–690.

[Barlow and Bartolomei, 2014] Barlow, D. P. and Bartolomei, M. S. (2014). Genomic imprinting in mammals. Cold Spring Harb Perspect Biol *6*.

[Bassett et al., 2014] Bassett, A. R., Akhtar, A., Barlow, D. P., Bird, A. P., Brockdorff, N., Duboule, D., Ephrussi, A., Ferguson-Smith, A. C., Gingeras, T. R., Haerty, W., Higgs, D. R., Miska, E. A. and Ponting, C. P. (2014). Considerations when investigating lncRNA function in vivo. Elife *3*, e03058.

[Bateman et al., 2015] Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Apweiler, R., Alpi, E., Antunes, R., Arganiska, J., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas, B., Chavali, G., Cibrian-Uhalte, E., Silva, A. D., De Giorgi, M., Dogan, T., Fazzini, F., Gane, P., Castro, L. G., Garmiri, P., Hatton-Ellis, E., Hieta, R., Huntley, R., Legge, D., Liu, W., Luo, J., MacDougall, A., Mutowo, P., Nightingale, A., Orchard, S., Pichler, K., Poggioli, D., Pundir, S., Pureza, L., Qi, G., Rosanoff, S., Saidi, R., Sawford, T., Shypitsyna, A., Turner, E., Volynkin, V., Wardell, T., Watkins, X., Zellner, H., Cowley, A., Figueira, L., Li, W., McWilliam, H., Lopez, R., Xenarios, I., Bougueleret, L., Bridge, A., Poux, S., Redaschi, N., Aimo, L., Argoud-Puy, G., Auchincloss, A., Axelsen, K., Bansal, P., Baratin, D., Blatter, M. C., Boeckmann, B., Bolleman, J., Boutet, E., Breuza, L., Casal-Casas, C., de Castro, E., Coudert, E., Cuche, B., Doche, M., Dornevil, D., Duvaud, S., Estreicher, A., Famiglietti, L., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Jungo, F., Keller, G., Lara, V., Lemercier, P., Lieberherr, D., Lombardot, T., Martin, X., Masson, P., Morgat, A., Neto, T., Nouspikel, N., Paesano, S., Pedruzzi, I., Pilbout, S., Pozzato, M., Pruess, M., Rivoire, C., Roechert, B., Schneider, M., Sigrist, C., Sonesson, K., Staehli, S., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Veuthey, A. L., Wu, C. H., Arighi, C. N., Arminski, L., Chen, C., Chen, Y., Garavelli, J. S., Huang, H., Laiho, K., McGarvey, P., Natale, D. A., Suzek, B. E., Vinayaka, C., Wang, Q., Wang, Y., Yeh, L. S., Yerramalla, M. S. and Zhang, J. (2015). UniProt: a hub for protein information. Nucleic Acids Res. *43*, D204–212.

[Battich et al., 2013] Battich, N., Stoeger, T. and Pelkmans, L. (2013). Image-based transcriptomics in thousands of single human cells at single-molecule resolution. Nat. Methods *10*, 1127–1133.

[Bazzini et al., 2014] Bazzini, A. A., Johnstone, T. G., Christiano, R., Mackowiak, S. D., Obermayer, B., Fleming, E. S., Vejnar, C. E., Lee, M. T., Rajewsky, N., Walther, T. C. and Giraldez, A. J. (2014). Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J. *33*, 981–993.

[Belote and Lucchesi, 1980] Belote, J. M. and Lucchesi, J. C. (1980). Male-specific lethal mutations of Drosophila melanogaster. Genetics *96*, 165–186.

[Berletch et al., 2011] Berletch, J. B., Yang, F., Xu, J., Carrel, L. and Disteche, C. M. (2011). Genes that escape from X inactivation. Hum. Genet. *130*, 237–245.

[Bertone et al., 2004] Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., Rinn, J. L., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M. and Snyder, M. (2004). Global identification of human transcribed sequences with genome tiling arrays. Science *306*, 2242–2246.

[Bickmore and van Steensel, 2013] Bickmore, W. A. and van Steensel, B. (2013). Genome architecture: domain organization of interphase chromosomes. Cell *152*, 1270–1284.

[Biggin, 2011] Biggin, M. D. (2011). Animal transcription networks as highly connected, quantitative continua. Dev. Cell *21*, 611–626.

[Blum et al., 2015] Blum, M., De Robertis, E. M., Wallingford, J. B. and Niehrs, C. (2015). Morpholinos: Antisense and Sensibility. Dev. Cell *35*, 145–149.

[Boettcher and McManus, 2015] Boettcher, M. and McManus, M. T. (2015). Choosing the Right Tool for the Job: RNAi, TALEN, or CRISPR. Mol. Cell *58*, 575–585.

[Bonasio and Shiekhattar, 2014] Bonasio, R. and Shiekhattar, R. (2014). Regulation of transcription by long noncoding RNAs. Annu. Rev. Genet. *48*, 433–455.

[Brannan et al., 1990] Brannan, C. I., Dees, E. C., Ingram, R. S. and Tilghman, S. M. (1990). The product of the H19 gene may function as an RNA. Mol. Cell. Biol. *10*, 28–36.

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

[Brockdorff, 2013] Brockdorff, N. (2013). Noncoding RNA and Polycomb recruitment. RNA *19*, 429–442.

[Brockdorff et al., 1992] Brockdorff, N., Ashworth, A., Kay, G. F., McCabe, V. M., Norris, D. P., Cooper, P. J., Swift, S. and Rastan, S. (1992). The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell *71*, 515–526.

[Brown et al., 1991] Brown, C. J., Ballabio, A., Rupert, J. L., Lafreniere, R. G., Grompe, M., Tonlorenzi, R. and Willard, H. F. (1991). A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. Nature *349*, 38–44.

[Brown et al., 1992] Brown, C. J., Hendrich, B. D., Rupert, J. L., Lafreniere, 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.

[Brown et al., 2012] Brown, J. A., Valenstein, M. L., Yario, T. A., Tycowski, K. T. and Steitz, J. A. (2012). Formation of triple-helical structures by the 3'-end sequences of MALAT1 and MENB noncoding RNAs. Proc. Natl. Acad. Sci. U.S.A. *109*, 19202–19207.

[Burger et al., 2013] Burger, K., Muhl, B., Kellner, M., Rohrmoser, M., Gruber-Eber, A., Windhager, L., Friedel, C. C., Dolken, L. and Eick, D. (2013). 4-thiouridine inhibits rRNA synthesis and causes a nucleolar stress response. RNA Biol *10*, 1623–1630.

[Buske et al., 2012] Buske, F. A., Bauer, D. C., Mattick, J. S. and Bailey, T. L. (2012). Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic data. Genome Res. *22*, 1372–1381.

[Cabili et al., 2015] Cabili, M. N., Dunagin, M. C., McClanahan, P. D., Biaesch, A., Padovan-Merhar, O., Regev, A., Rinn, J. L. and Raj, A. (2015). Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. Genome Biol. *16*, 20.

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

[Calabrese et al., 2012] Calabrese, J. M., Sun, W., Song, L., Mugford, J. W., Williams, L., Yee, D., Starmer, J., Mieczkowski, P., Crawford, G. E. and Magnuson, T. (2012). Site-specific silencing of regulatory elements as a mechanism of X inactivation. Cell *151*, 951–963.

[Carlile et al., 2014] Carlile, T. M., Rojas-Duran, M. F., Zinshteyn, B., Shin, H., Bartoli, K. M. and Gilbert, W. V. (2014). Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. Nature *515*, 143–146.

[Carninci et al., 2005] Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V. B., Brenner, S. E., Batalov, S., Forrest, A. R., Zavolan, M., Davis, M. J., Wilming, L. G., Aidinis, V., Allen, J. E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R. N., Bailey, T. L., Bansal, M., Baxter, L., Beisel, K. W., Bersano, T., Bono, H., Chalk, A. M., Chiu, K. P., Choudhary, V., Christoffels, A., Clutterbuck, D. R., Crowe, M. L., Dalla, E., Dalrymple, B. P., de Bono, B., Della Gatta, G., di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C. F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T. R., Gojobori, T., Green, R. E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T. K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S. P., Kruger, A., Kummerfeld, S. K., Kurochkin, I. V., Lareau, L. F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., McWilliam, S., Madan Babu, M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K. C., Pavan, W. J., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J. F., Ring, B. Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S. L., Sandelin, A., Schneider, C., Schonbach, C., Sekiguchi, K., Semple, C. A., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S. L., Tang, S., Taylor, M. S., Tegner, J., Teichmann, S. A., Ueda, H. R., van

Nimwegen, E., Verardo, R., Wei, C. L., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S. M., Teasdale, R. D., Liu, E. T., Brusic, V., Quackenbush, J., Wahlestedt, C., Mattick, J. S., Hume, D. A., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J. and Hayashizaki, Y. (2005). The transcriptional landscape of the mammalian genome. Science *309*, 1559–1563.

[Carrel and Willard, 2005] Carrel, L. and Willard, H. F. (2005). X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature *434*, 400–404.

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

[Chadwick and Willard, 2003] Chadwick, B. P. and Willard, H. F. (2003). Chromatin of the Barr body: histone and non-histone proteins associated with or excluded from the inactive X chromosome. Hum. Mol. Genet. *12*, 2167–2178.

[Chadwick and Willard, 2004] Chadwick, B. P. and Willard, H. F. (2004). Multiple spatially distinct types of facultative heterochromatin on the human inactive X chromosome. Proc. Natl. Acad. Sci. U.S.A. *101*, 17450–17455.

[Chalei et al., 2014] Chalei, V., Sansom, S. N., Kong, L., Lee, S., Montiel, J. F., Vance, K. W. and Ponting, C. P. (2014). The long non-coding RNA Dali is an epigenetic regulator of neural differentiation. Elife *3*, e04530.

[Chang and Mendell, 2007] Chang, T. C. and Mendell, J. T. (2007). microRNAs in vertebrate physiology and human disease. Annu Rev Genomics Hum Genet *8*, 215–239.

[Chen et al., 2015] Chen, P. B., Chen, H. V., Acharya, D., Rando, O. J. and Fazzio, T. G. (2015). R loops regulate promoter-proximal chromatin architecture and cellular differentiation. Nat. Struct. Mol. Biol. *22*, 999–1007.

[Choudhary et al., 2016] Choudhary, K., Shih, N. P., Deng, F., Ledda, M., Li, B. and Aviran, S. (2016). Metrics for rapid quality control in RNA structure probing experiments. Bioinformatics *32*, 3575–3583.

[Chu et al., 2011] Chu, C., Qu, K., Zhong, F. L., Artandi, S. E. and Chang, H. Y. (2011). Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. Mol. Cell *44*, 667–678.

[Chu et al., 2015] Chu, C., Zhang, Q. C., da Rocha, S. T., Flynn, R. A., Bharadwaj, M., Calabrese, J. M., Magnuson, T., Heard, E. and Chang, H. Y. (2015). Systematic discovery of Xist RNA binding proteins. Cell *161*, 404–416.

[Clark et al., 2012] Clark, M. B., Johnston, R. L., Inostroza-Ponta, M., Fox, A. H., Fortini, E., Moscato, P., Dinger, M. E. and Mattick, J. S. (2012). Genome-wide analysis of long noncoding RNA stability. Genome Res. *22*, 885–898.

[Clarke et al., 2009] Clarke, J., Wu, H. C., Jayasinghe, L., Patel, A., Reid, S. and Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol *4*, 265–270.

[Cleary et al., 2005] Cleary, M. D., Meiering, C. D., Jan, E., Guymon, R. and Boothroyd, J. C. (2005). Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay. Nat. Biotechnol. *23*, 232–237.

[Clemson et al., 1996] 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.

[Conrad et al., 2006] Conrad, N. K., Mili, S., Marshall, E. L., Shu, M. D. and Steitz, J. A. (2006). Identification of a rapid mammalian deadenylation-dependent decay pathway and its inhibition by a viral RNA element. Mol. Cell *24*, 943–953.

[Conrad et al., 2007] Conrad, N. K., Shu, M. D., Uyhazi, K. E. and Steitz, J. A. (2007). Mutational analysis of a viral RNA element that counteracts rapid RNA decay by interaction with the polyadenylate tail. Proc. Natl. Acad. Sci. U.S.A. *104*, 10412–10417.

[Conrad and Akhtar, 2012] Conrad, T. and Akhtar, A. (2012). Dosage compensation in Drosophila melanogaster: epigenetic fine-tuning of chromosome-wide transcription. Nat. Rev. Genet. *13*, 123–134.

[Cordero et al., 2012] Cordero, P., Lucks, J. B. and Das, R. (2012). An RNA Mapping DataBase for curating RNA structure mapping experiments. Bioinformatics *28*, 3006–3008.

[Crosetto et al., 2015] Crosetto, N., Bienko, M. and van Oudenaarden, A. (2015). Spatially resolved transcriptomics and beyond. Nat. Rev. Genet. *16*, 57–66.

[Cunningham et al., 2015] Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Giron, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Kahari, A. K., Keenan, S., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Overduin, B., Parker, A., Patricio, M., Perry, E., Pignatelli, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A.,

Wilder, S. P., Zadissa, A., Aken, B. L., Birney, E., Harrow, J., Kinsella, R., Muffato, M., Ruffier, M., Searle, S. M., Spudich, G., Trevanion, S. J., Yates, A., Zerbino, D. R. and Flicek, P. (2015). Ensembl 2015. Nucleic Acids Res. *43*, D662–669.

[Dai et al., 2017] Dai, Q., Moshitch-Moshkovitz, S., Han, D., Kol, N., Amariglio, N., Rechavi, G., Dominissini, D. and He, C. (2017). Nm-seq maps 2'-O-methylation sites in human mRNA with base precision. Nat. Methods *14*, 695–698.

[Davidovich and Cech, 2015] Davidovich, C. and Cech, T. R. (2015). The recruitment of chromatin modifiers by long noncoding RNAs: lessons from PRC2. RNA *21*, 2007–2022.

[Davidovich et al., 2015] Davidovich, C., Wang, X., Cifuentes-Rojas, C., Goodrich, K. J., Gooding, A. R., Lee, J. T. and Cech, T. R. (2015). Toward a consensus on the binding specificity and promiscuity of PRC2 for RNA. Mol. Cell *57*, 552–558.

[Davidovich et al., 2013] Davidovich, C., Zheng, L., Goodrich, K. J. and Cech, T. R. (2013). Promiscuous RNA binding by Polycomb repressive complex 2. Nat. Struct. Mol. Biol. *20*, 1250–1257.

[Davydov et al., 2010] Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A. and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput. Biol. *6*, e1001025.

[de Hoon et al., 2015] de Hoon, M., Shin, J. W. and Carninci, P. (2015). Paradigm shifts in genomics through the FANTOM projects. Mamm. Genome *26*, 391–402.

[Deigan et al., 2009] Deigan, K. E., Li, T. W., Mathews, D. H. and Weeks, K. M. (2009). Accurate SHAPE-directed RNA structure determination. Proc. Natl. Acad. Sci. U.S.A. *106*, 97–102.

[Derrien et al., 2012] 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 Guigo, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789.

[Dimitrova et al., 2014] Dimitrova, N., Zamudio, J. R., Jong, R. M., Soukup, D., Resnick, R., Sarma, K., Ward, A. J., Raj, A., Lee, J. T., Sharp, P. A. and Jacks, T. (2014). LincRNA-p21 activates p21 in cis to promote Polycomb target gene expression and to enforce the G1/S checkpoint. Mol. Cell *54*, 777–790.

[Ding et al., 2014] Ding, Y., Tang, Y., Kwok, C. K., Zhang, Y., Bevilacqua, P. C. and Assmann, S. M. (2014). In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature *505*, 696–700.

[Dinger et al., 2008] Dinger, M. E., Pang, K. C., Mercer, T. R. and Mattick, J. S. (2008). Differentiating protein-coding and noncoding RNA: challenges and ambiguities. PLoS Comput. Biol. *4*, e1000176.

[Disteche, 2012] Disteche, C. M. (2012). Dosage compensation of the sex chromosomes. Annu. Rev. Genet. *46*, 537–560.

[Dixon et al., 2012] Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

[Djebali et al., 2012] Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Roder, M.,

Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O. J., Park, E., Persaud, K., Preall, J. B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L. H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S. E., Hannon, G., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigo, R. and Gingeras, T. R. (2012). Landscape of transcription in human cells. Nature *489*, 101–108.

[Dolken et al., 2008] Dolken, L., Ruzsics, Z., Radle, B., Friedel, C. C., Zimmer, R., Mages, J., Hoffmann, R., Dickinson, P., Forster, T., Ghazal, P. and Koszinowski, U. H. (2008). High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. RNA *14*, 1959–1972.

[Dominissini et al., 2012] Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., Cesarkas, K., Jacob-Hirsch, J., Amariglio, N., Kupiec, M., Sorek, R. and Rechavi, G. (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature *485*, 201–206.

[Donley et al., 2015] Donley, N., Smith, L. and Thayer, M. J. (2015). ASAR15, A cis-acting locus that controls chromosome-wide replication timing and stability of human chromosome 15. PLoS Genet. *11*, e1004923.

[Duffy et al., 2015] Duffy, E. E., Rutenberg-Schoenberg, M., Stark, C. D., Kitchen, R. R., Gerstein, M. B. and Simon, M. D. (2015). Tracking Distinct RNA Popula-

tions Using Efficient and Reversible Covalent Chemistry. Mol. Cell *59*, 858–866.

[Dunagin et al., 2015] Dunagin, M., Cabili, M. N., Rinn, J. and Raj, A. (2015). Visualization of lncRNA by single-molecule fluorescence in situ hybridization. Methods Mol. Biol. *1262*, 3–19.

[Dunham et al., 2012] Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B. K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., Shoresh, N., Simon, J. M., Song, L., Trinklein, N. D., Altshuler, R. C., Birney, E., Brown, J. B., Cheng, C., Djebali, S., Dong, X., Dunham, I., Ernst, J., Furey, T. S., Gerstein, M., Giardine, B., Greven, M., Hardison, R. C., Harris, R. S., Herrero, J., Hoffman, M. M., Iyer, S., Kellis, M., Khatun, J., Kheradpour, P., Kundaje, A., Lassmann, T., Li, Q., Lin, X., Marinov, G. K., Merkel, A., Mortazavi, A., Parker, S. C., Reddy, T. E., Rozowsky, J., Schlesinger, F., Thurman, R. E., Wang, J., Ward, L. D., Whitfield, T. W., Wilder, S. P., Wu, W., Xi, H. S., Yip, K. Y., Zhuang, J., Pazin, M. J., Lowdon, R. F., Dillon, L. A., Adams, L. B., Kelly, C. J., Zhang, J., Wexler, J. R., Green, E. D., Good, P. J., Feingold, E. A., Bernstein, B. E., Birney, E., Crawford, G. E., Dekker, J., Elnitski, L., Farnham, P. J., Gerstein, M., Giddings, M. C., Gingeras, T. R., Green, E. D., Guigo, R., Hardison, R. C., Hubbard, T. J., Kellis, M., Kent, W., Lieb, J. D., Margulies, E. H., Myers, R. M., Snyder, M., Stamatoyannopoulos, J. A., Tenenbaum, S. A., Weng, Z., White, K. P., Wold, B., Khatun, J., Yu, Y., Wrobel, J., Risk, B. A., Gunawardena, H. P., Kuiper, H. C., Maier, C. W., Xie, L., Chen, X., Giddings, M. C., Bernstein, B. E., Epstein, C. B., Shoresh, N., Ernst, J., Kheradpour, P., Mikkelsen, T. S., Gillespie, S., Goren, A., Ram, O., Zhang, X., Wang, L., Issner, R., Coyne, M. J., Durham, T., Ku, M., Truong, T., Ward, L. D., Altshuler, R. C., Eaton, M. L., Kellis, M., Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lass-

mann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Roder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Batut, P., Bell, I., Bell, K., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H. P., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Li, G., Luo, O. J., Park, E., Preall, J. B., Presaud, K., Ribeca, P., Risk, B. A., Robyr, D., Ruan, X., Sammeth, M., Sandhu, K. S., Schaeffer, L., See, L. H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T. J., Reymond, A., Antonarakis, S. E., Hannon, G. J., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigo, R., Gingeras, T. R., Rosenbloom, K. R., Sloan, C. A., Learned, K., Malladi, V. S., Wong, M. C., Barber, G. P., Cline, M. S., Dreszer, T. R., Heitner, S. G., Karolchik, D., Kent, W., Kirkup, V. M., Meyer, L. R., Long, J. C., Maddren, M., Raney, B. J., Furey, T. S., Song, L., Grasfeder, L. L., Giresi, P. G., Lee, B. K., Battenhouse, A., Sheffield, N. C., Simon, J. M., Showers, K. A., Safi, A., London, D., Bhinge, A. A., Shestak, C., Schaner, M. R., Kim, S. K., Zhang, Z. Z., Mieczkowski, P. A., Mieczkowska, J. O., Liu, Z., McDaniell, R. M., Ni, Y., Rashid, N. U., Kim, M. J., Adar, S., Zhang, Z., Wang, T., Winter, D., Keefe, D., Birney, E., Iyer, V. R., Lieb, J. D., Crawford, G. E., Li, G., Sandhu, K. S., Zheng, M., Wang, P., Luo, O. J., Shahab, A., Fullwood, M. J., Ruan, X., Ruan, Y., Myers, R. M., Pauli, F., Williams, B. A., Gertz, J., Marinov, G. K., Reddy, T. E., Vielmetter, J., Partridge, E., Trout, D., Varley, K. E., Gasper, C., Bansal, A., Pepke, S., Jain, P., Amrhein, H., Bowling, K. M., Anaya, M., Cross, M. K., King, B., Muratet, M. A., Antoshechkin, I., Newberry, K. M., McCue, K., Nesmith, A. S., Fisher-Aylor, K. I., Pusey, B., DeSalvo,

G., Parker, S. L., Balasubramanian, S., Davis, N. S., Meadows, S. K., Eggleston, T., Gunter, C., Newberry, J., Levy, S. E., Absher, D. M., Mortazavi, A., Wong, W. H., Wold, B., Blow, M. J., Visel, A., Pennachio, L. A., Elnitski, L., Margulies, E. H., Parker, S. C., Petrykowska, H. M., Abyzov, A., Aken, B., Barrell, D., Barson, G., Berry, A., Bignell, A., Boychenko, V., Bussotti, G., Chrast, J., Davidson, C., Derrien, T., Despacio-Reyes, G., Diekhans, M., Ezkurdia, I., Frankish, A., Gilbert, J., Gonzalez, J. M., Griffiths, E., Harte, R., Hendrix, D. A., Howald, C., Hunt, T., Jungreis, I., Kay, M., Khurana, E., Kokocinski, F., Leng, J., Lin, M. F., Loveland, J., Lu, Z., Manthravadi, D., Mariotti, M., Mudge, J., Mukherjee, G., Notredame, C., Pei, B., Rodriguez, J. M., Saunders, G., Sboner, A., Searle, S., Sisu, C., Snow, C., Steward, C., Tanzer, A., Tapanari, E., Tress, M. L., van Baren, M. J., Walters, N., Washietl, S., Wilming, L., Zadissa, A., Zhang, Z., Brent, M., Haussler, D., Kellis, M., Valencia, A., Gerstein, M., Reymond, A., Guigo, R., Harrow, J., Hubbard, T. J., Landt, S. G., Frietze, S., Abyzov, A., Addleman, N., Alexander, R. P., Auerbach, R. K., Balasubramanian, S., Bettinger, K., Bhardwaj, N., Boyle, A. P., Cao, A. R., Cayting, P., Charos, A., Cheng, Y., Cheng, C., Eastman, C., Euskirchen, G., Fleming, J. D., Grubert, F., Habegger, L., Hariharan, M., Harmanci, A., Iyengar, S., Jin, V. X., Karczewski, K. J., Kasowski, M., Lacroute, P., Lam, H., Lamarre-Vincent, N., Leng, J., Lian, J., Lindahl-Allen, M., Min, R., Miotto, B., Monahan, H., Moqtaderi, Z., Mu, X. J., O'Geen, H., Ouyang, Z., Patacsil, D., Pei, B., Raha, D., Ramirez, L., Reed, B., Rozowsky, J., Sboner, A., Shi, M., Sisu, C., Slifer, T., Witt, H., Wu, L., Xu, X., Yan, K. K., Yang, X., Yip, K. Y., Zhang, Z., Struhl, K., Weissman, S. M., Gerstein, M., Farnham, P. J., Snyder, M., Tenenbaum, S. A., Penalva, L. O., Doyle, F., Karmakar, S., Landt, S. G., Bhanvadia, R. R., Choudhury, A., Domanus, M., Ma, L., Moran, J., Patacsil, D., Slifer, T., Victorsen, A., Yang, X., Snyder, M., Auer, T., Centanin, L., Eichenlaub, M., Gruhl, F., Heermann, S., Hoeckendorf, B., Inoue, D., Kellner, T., Kirchmaier, S., Mueller, C., Reinhardt, R.,

Schertel, L., Schneider, S., Sinn, R., Wittbrodt, B., Wittbrodt, J., Weng, Z., Whitfield, T. W., Wang, J., Collins, P. J., Aldred, S. F., Trinklein, N. D., Partridge, E. C., Myers, R. M., Dekker, J., Jain, G., Lajoie, B. R., Sanyal, A., Balasundaram, G., Bates, D. L., Byron, R., Canfield, T. K., Diegel, M. J., Dunn, D., Ebersol, A. K., Frum, T., Garg, K., Gist, E., Hansen, R., Boatman, L., Haugen, E., Humbert, R., Jain, G., Johnson, A. K., Johnson, E. M., Kutyavin, T. V., Lajoie, B. R., Lee, K., Lotakis, D., Maurano, M. T., Neph, S. J., Neri, F. V., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Rynes, E., Sabo, P., Sanchez, M. E., Sandstrom, R. S., Sanyal, A., Shafer, A. O., Stergachis, A. B., Thomas, S., Thurman, R. E., Vernot, B., Vierstra, J., Vong, S., Wang, H., Weaver, M. A., Yan, Y., Zhang, M., Akey, J. M., Bender, M., Dorschner, M. O., Groudine, M., MacCoss, M. J., Navas, P., Stamatoyannopoulos, G., Kaul, R., Dekker, J., Stamatoyannopoulos, J. A., Dunham, I., Beal, K., Brazma, A., Flicek, P., Herrero, J., Johnson, N., Keefe, D., Lukk, M., Luscombe, N. M., Sobral, D., Vaquerizas, J. M., Wilder, S. P., Batzoglou, S., Sidow, A., Hussami, N., Kyriazopoulou-Panagiotopoulou, S., Libbrecht, M. W., Schaub, M. A., Kundaje, A., Hardison, R. C., Miller, W., Giardine, B., Harris, R. S., Wu, W., Bickel, P. J., Banfai, B., Boley, N. P., Brown, J. B., Huang, H., Li, Q., Li, J. J., Noble, W. S., Bilmes, J. A., Buske, O. J., Hoffman, M. M., Sahu, A. D., Kharchenko, P. V., Park, P. J., Baker, D., Taylor, J., Weng, Z., Iyer, S., Dong, X., Greven, M., Lin, X., Wang, J., Xi, H. S., Zhuang, J., Gerstein, M., Alexander, R. P., Balasubramanian, S., Cheng, C., Harmanci, A., Lochovsky, L., Min, R., Mu, X. J., Rozowsky, J., Yan, K. K., Yip, K. Y. and Birney, E. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

[Duthie et al., 1999] Duthie, S. M., Nesterova, T. B., Formstone, E. J., Keohane, A. M., Turner, B. M., Zakian, S. M. and Brockdorff, N. (1999). Xist RNA exhibits a banded localization on the inactive X chromosome and is excluded from autosomal

material in cis. Hum. Mol. Genet. *8*, 195–204.

[Eddy, 2014] Eddy, S. R. (2014). Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. Annu Rev Biophys *43*, 433–456.

[Eid et al., 2009] Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J. and Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. Science *323*, 133–138.

[Engreitz et al., 2013] Engreitz, J. M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E. S., Plath, K. and Guttman, M. (2013). The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science *341*, 1237973.

[Fang et al., 2015] Fang, R., Moss, W. N., Rutenberg-Schoenberg, M. and Simon, M. D. (2015). Probing Xist RNA Structure in Cells Using Targeted Structure-Seq. PLoS Genet. *11*, e1005668.

[Fitzpatrick et al., 2002] Fitzpatrick, G. V., Soloway, P. D. and Higgins, M. J. (2002). Regional loss of imprinting and growth deficiency in mice with a targeted deletion of KvDMR1. Nat. Genet. *32*, 426–431.

[Fuchs et al., 2014] Fuchs, G., Voichek, Y., Benjamin, S., Gilad, S., Amit, I. and Oren, M. (2014). 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. Genome Biol. *15*, R69.

[Fukunaga et al., 1975] Fukunaga, A., Tanaka, A. and Oishi, K. (1975). Maleless, a recessive autosomal mutant of Drosophila melanogaster that specifically kills male zygotes. Genetics *81*, 135–141.

[Gall and Pardue, 1969] Gall, J. G. and Pardue, M. L. (1969). Formation and detection of RNA-DNA hybrid molecules in cytological preparations. Proc. Natl. Acad. Sci. U.S.A. *63*, 378–383.

[Gantier et al., 2011] Gantier, M. P., McCoy, C. E., Rusinova, I., Saulep, D., Wang, D., Xu, D., Irving, A. T., Behlke, M. A., Hertzog, P. J., Mackay, F. and Williams, B. R. (2011). Analysis of microRNA turnover in mammalian cells following Dicer1 ablation. Nucleic Acids Res. *39*, 5692–5703.

[Gardner et al., 2012] Gardner, E. J., Nizami, Z. F., Talbot, C. C. and Gall, J. G. (2012). Stable intronic sequence RNA (sisRNA), a new class of noncoding RNA from the oocyte nucleus of Xenopus tropicalis. Genes Dev. *26*, 2550–2559.

[Gelbart and Kuroda, 2009] Gelbart, M. E. and Kuroda, M. I. (2009). Drosophila dosage compensation: a complex voyage to the X chromosome. Development *136*, 1399–1410.

[Gendrel and Heard, 2014] Gendrel, A. V. and Heard, E. (2014). Noncoding RNAs and epigenetic mechanisms during X-chromosome inactivation. Annu. Rev. Cell Dev. Biol. *30*, 561–580.

[Goff and Rinn, 2015] Goff, L. A. and Rinn, J. L. (2015). Linking RNA biology to lncRNAs. Genome Res. *25*, 1456–1465.

[Gonzalez-Porta et al., 2013] Gonzalez-Porta, M., Frankish, A., Rung, J., Harrow, J. and Brazma, A. (2013). Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biol. *14*, R70.

[Gregersen et al., 2014] Gregersen, L. H., Schueler, M., Munschauer, M., Mastrobuoni, G., Chen, W., Kempa, S., Dieterich, C. and Landthaler, M. (2014). MOV10 Is a 5' to 3' RNA helicase contributing to UPF1 mRNA target degradation by translocation along 3' UTRs. Mol. Cell *54*, 573–585.

[Grote et al., 2013] Grote, P., Wittler, L., Hendrix, D., Koch, F., Wahrisch, S., Beisaw, A., Macura, K., Blass, G., Kellis, M., Werber, M. and Herrmann, B. G. (2013). The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. Dev. Cell *24*, 206–214.

[Gu et al., 2012] Gu, W., Lee, H. C., Chaves, D., Youngman, E. M., Pazour, G. J., Conte, D. and Mello, C. C. (2012). CapSeq and CIP-TAP identify Pol II start sites and reveal capped small RNAs as C. elegans piRNA precursors. Cell *151*, 1488–1500.

[Guo et al., 2015] Guo, Y., Liu, J., Elfenbein, S. J., Ma, Y., Zhong, M., Qiu, C., Ding, Y. and Lu, J. (2015). Characterization of the mammalian miRNA turnover landscape. Nucleic Acids Res. *43*, 2326–2341.

[Guttman et al., 2009] 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*, 223–227.

[Guttman et al., 2011] Guttman, M., Donaghey, J., Carey, B. W., Garber, M., Grenier, J. K., Munson, G., Young, G., Lucas, A. B., Ach, R., Bruhn, L., Yang, X., Amit, I., Meissner, A., Regev, A., Rinn, J. L., Root, D. E. and Lander, E. S. (2011).

lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature *477*, 295–300.

[Guttman et al., 2010] Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., Nusbaum, C., Rinn, J. L., Lander, E. S. and Regev, A. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat. Biotechnol. *28*, 503–510.

[Guttman and Rinn, 2012] Guttman, M. and Rinn, J. L. (2012). Modular regulatory principles of large non-coding RNAs. Nature *482*, 339–346.

[Guttman et al., 2013] Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S. and Lander, E. S. (2013). Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell *154*, 240–251.

[Hacisuleyman et al., 2014] Hacisuleyman, E., Goff, L. A., Trapnell, C., Williams, A., Henao-Mejia, J., Sun, L., McClanahan, P., Hendrickson, D. G., Sauvageau, M., Kelley, D. R., Morse, M., Engreitz, J., Lander, E. S., Guttman, M., Lodish, H. F., Flavell, R., Raj, A. and Rinn, J. L. (2014). Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. Nat. Struct. Mol. Biol. *21*, 198–206.

[Hafner et al., 2010] Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A. C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M. and Tuschl, T. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell *141*, 129–141.

[Heyn et al., 2014] Heyn, P., Kircher, M., Dahl, A., Kelso, J., Tomancak, P., Kalinka, A. T. and Neugebauer, K. M. (2014). The earliest transcribed zygotic genes are short, newly evolved, and different across species. Cell Rep *6*, 285–292.

[Hiratani et al., 2008] Hiratani, I., Ryba, T., Itoh, M., Yokochi, T., Schwaiger, M., Chang, C. W., Lyou, Y., Townes, T. M., Schubeler, D. and Gilbert, D. M. (2008). Global reorganization of replication domains during embryonic stem cell differentiation. PLoS Biol. *6*, e245.

[Housman and Ulitsky, 2016] Housman, G. and Ulitsky, I. (2016). Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of long noncoding RNAs. Biochim. Biophys. Acta *1859*, 31–40.

[Hsu et al., 2014] Hsu, P. D., Lander, E. S. and Zhang, F. (2014). Development and applications of CRISPR-Cas9 for genome engineering. Cell *157*, 1262–1278.

[Huang et al., 2011] Huang, W., Umbach, D. M., Vincent Jordan, N., Abell, A. N., Johnson, G. L. and Li, L. (2011). Efficiently identifying genome-wide changes with next-generation sequencing data. Nucleic Acids Res. *39*, e130.

[Huarte et al., 2010] Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M. J., Kenzelmann-Broz, D., Khalil, A. M., Zuk, O., Amit, I., Rabani, M., Attardi, L. D., Regev, A., Lander, E. S., Jacks, T. and Rinn, J. L. (2010). A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell *142*, 409–419.

[Hube et al., 2006] Hube, F., Guo, J., Chooniedass-Kothari, S., Cooper, C., Hamedani, M. K., Dibrov, A. A., Blanchard, A. A., Wang, X., Deng, G., Myal, Y. and Leygue, E. (2006). Alternative splicing of the first intron of the steroid receptor

RNA activator (SRA) participates in the generation of coding and noncoding RNA isoforms in breast cancer cell lines. DNA Cell Biol. *25*, 418–428.

[Huynen et al., 1997] Huynen, M., Gutell, R. and Konings, D. (1997). Assessing the reliability of RNA folding using statistical mechanics. J. Mol. Biol. *267*, 1104–1112.

[Jao and Salic, 2008] Jao, C. Y. and Salic, A. (2008). Exploring RNA transcription and turnover in vivo by using click chemistry. Proc. Natl. Acad. Sci. U.S.A. *105*, 15779–15784.

[Jeon and Lee, 2011] Jeon, Y. and Lee, J. T. (2011). YY1 tethers Xist RNA to the inactive X nucleation center. Cell *146*, 119–133.

[Jeschke, 2013] Jeschke, G. (2013). Conformational dynamics and distribution of nitroxide spin labels. Prog Nucl Magn Reson Spectrosc *72*, 42–60.

[Ji et al., 2006] Ji, X., Li, W., Song, J., Wei, L. and Liu, X. S. (2006). CEAS: cis-regulatory element annotation system. Nucleic Acids Res. *34*, W551–554.

[John et al., 2011] John, S., Sabo, P. J., Thurman, R. E., Sung, M. H., Biddie, S. C., Johnson, T. A., Hager, G. L. and Stamatoyannopoulos, J. A. (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat. Genet. *43*, 264–268.

[Kanduri et al., 2006] Kanduri, C., Thakur, N. and Pandey, R. R. (2006). The length of the transcript encoded from the Kcnq1ot1 antisense promoter determines the degree of silencing. EMBO J. *25*, 2096–2106.

[Kaneko et al., 2014a] Kaneko, S., Bonasio, R., Saldana-Meyer, R., Yoshida, T., Son, J., Nishino, K., Umezawa, A. and Reinberg, D. (2014a). Interactions between JARID2 and noncoding RNAs regulate PRC2 recruitment to chromatin. Mol. Cell *53*, 290–300.

[Kaneko et al., 2014b] Kaneko, S., Son, J., Bonasio, R., Shen, S. S. and Reinberg, D. (2014b). Nascent RNA interaction keeps PRC2 activity poised and in check. Genes Dev. *28*, 1983–1988.

[Kapranov et al., 2007] Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., Stadler, P. F., Hertel, J., Hackermuller, 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.

[Keane et al., 2011] Keane, T. M., Goodstadt, L., Danecek, P., White, M. A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., Furlotte, N. A., Eskin, E., Nellaker, C., Whitley, H., Cleak, J., Janowitz, D., Hernandez-Pliego, P., Edwards, A., Belgard, T. G., Oliver, P. L., McIntyre, R. E., Bhomra, A., Nicod, J., Gan, X., Yuan, W., van der Weyden, L., Steward, C. A., Bala, S., Stalker, J., Mott, R., Durbin, R., Jackson, I. J., Czechanski, A., Guerra-Assuncao, J. A., Donahue, L. R., Reinholdt, L. G., Payseur, B. A., Ponting, C. P., Birney, E., Flint, J. and Adams, D. J. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. Nature *477*, 289–294.

[Kelley et al., 1999] Kelley, R. L., Meller, V. H., Gordadze, P. R., Roman, G., Davis, R. L. and Kuroda, M. I. (1999). Epigenetic spreading of the Drosophila dosage compensation complex from roX RNA genes into flanking chromatin. Cell *98*, 513–522.

[Kenyon and Bruice, 1977] Kenyon, G. L. and Bruice, T. W. (1977). Novel sulfhydryl reagents. Meth. Enzymol. *47*, 407–430.

[Kersey et al., 2014] Kersey, P. J., Allen, J. E., Christensen, M., Davis, P., Falin, L. J., Grabmueller, C., Hughes, D. S., Humphrey, J., Kerhornou, A., Khobova, J., Langridge, N., McDowall, M. D., Maheswari, U., Maslen, G., Nuhn, M., Ong, C. K., Paulini, M., Pedro, H., Toneva, I., Tuli, M. A., Walts, B., Williams, G., Wilson, D., Youens-Clark, K., Monaco, M. K., Stein, J., Wei, X., Ware, D., Bolser, D. M., Howe, K. L., Kulesha, E., Lawson, D. and Staines, D. M. (2014). Ensembl Genomes 2013: scaling up access to genome-wide data. Nucleic Acids Res. *42*, D546–552.

[Kertesz et al., 2010] Kertesz, M., Wan, Y., Mazor, E., Rinn, J. L., Nutter, R. C., Chang, H. Y. and Segal, E. (2010). Genome-wide measurement of RNA secondary structure in yeast. Nature *467*, 103–107.

[Kharchenko et al., 2008] Kharchenko, P. V., Tolstorukov, M. Y. and Park, P. J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat. Biotechnol. *26*, 1351–1359.

[Kim et al., 2013] 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 et al., 2015] 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 et al., 2010] 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.

[Kohlmaier et al., 2004] Kohlmaier, A., Savarese, F., Lachner, M., Martens, J., Jenuwein, T. and Wutz, A. (2004). A chromosomal memory triggered by Xist regulates histone methylation in X inactivation. PLoS Biol. *2*, E171.

[Kok et al., 2015] Kok, F. O., Shin, M., Ni, C. W., Gupta, A., Grosse, A. S., van Impel, A., Kirchmaier, B. C., Peterson-Maduro, J., Kourkoulis, G., Male, I., DeSantis, D. F., Sheppard-Tindell, S., Ebarasi, L., Betsholtz, C., Schulte-Merker, S., Wolfe, S. A. and Lawson, N. D. (2015). Reverse genetic screening reveals poor correlation between morpholino-induced and mutant phenotypes in zebrafish. Dev. Cell *32*, 97–108.

[Kozomara and Griffiths-Jones, 2014] Kozomara, A. and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. *42*, 68–73.

[Kuhn et al., 2007] Kuhn, R. M., Karolchik, D., Zweig, A. S., Trumbower, H., Thomas, D. J., Thakkapallayil, A., Sugnet, C. W., Stanke, M., Smith, K. E., Siepel, A., Rosenbloom, K. R., Rhead, B., Raney, B. J., Pohl, A., Pedersen, J. S., Hsu, F., Hinrichs, A. S., Harte, R. A., Diekhans, M., Clawson, H., Bejerano, G., Barber, G. P., Baertsch, R., Haussler, D. and Kent, W. J. (2007). The UCSC genome browser database: update 2007. Nucleic Acids Res. *35*, D668–673.

[Kwok et al., 2013] Kwok, C. K., Ding, Y., Tang, Y., Assmann, S. M. and Bevilacqua, P. C. (2013). Determination of in vivo RNA structure in low-abundance transcripts. Nat Commun *4*, 2971.

[Lai et al., 2015] Lai, F., Gardini, A., Zhang, A. and Shiekhattar, R. (2015). Integrator mediates the biogenesis of enhancer RNAs. Nature *525*, 399–403.

[Langmead and Salzberg, 2012] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

[Latos et al., 2012] Latos, P. A., Pauler, F. M., Koerner, M. V., ?energin, H. B., Hudson, Q. J., Stocsits, R. R., Allhoff, W., Stricker, S. H., Klement, R. M., Warczok, K. E., Aumayr, K., Pasierbek, P. and Barlow, D. P. (2012). Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. Science *338*, 1469–1472.

[Ledda and Aviran, 2018] Ledda, M. and Aviran, S. (2018). PATTERNA: transcriptome-wide search for functional RNA elements via structural data signatures. Genome Biol. *19*, 28.

[Lee, 2009] Lee, J. T. (2009). Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. Genes Dev. *23*, 1831–1842.

[Lee, 2012] Lee, J. T. (2012). Epigenetic regulation by long noncoding RNAs. Science *338*, 1435–1439.

[Lee and Bartolomei, 2013] Lee, J. T. and Bartolomei, M. S. (2013). X-inactivation, imprinting, and long noncoding RNAs in health and disease. Cell *152*, 1308–1323.

[Lee et al., 1999] Lee, M. P., DeBaun, M. R., Mitsuya, K., Galonek, H. L., Brandenburg, S., Oshimura, M. and Feinberg, A. P. (1999). Loss of imprinting of a paternally expressed transcript, with antisense orientation to KVLQT1, occurs frequently in Beckwith-Wiedemann syndrome and is independent of insulin-like growth factor II imprinting. Proc. Natl. Acad. Sci. U.S.A. *96*, 5203–5208.

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

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

[Li et al., 2017] Li, B., Tambe, A., Aviran, S. and Pachter, L. (2017). PROBer Provides a General Toolkit for Analyzing Sequencing-Based Toeprinting Assays. Cell Syst *4*, 568–574.

[Licatalosi et al., 2008] Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., Clark, T. A., Schweitzer, A. C., Blume, J. E., Wang, X., Darnell, J. C. and Darnell, R. B. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456*, 464–469.

[Lin et al., 2011] Lin, M. F., Jungreis, I. and Kellis, M. (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics *27*, i275–282.

[Lin et al., 2014] Lin, N., Chang, K. Y., Li, Z., Gates, K., Rana, Z. A., Dang, J., Zhang, D., Han, T., Yang, C. S., Cunningham, T. J., Head, S. R., Duester, G., Dong, P. D. and Rana, T. M. (2014). An evolutionarily conserved long noncoding RNA TUNA controls pluripotency and neural lineage commitment. Mol. Cell *53*, 1005–1019.

[Lister et al., 2008] Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H. and Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell *133*, 523–536.

[Lochovsky et al., 2015] Lochovsky, L., Zhang, J., Fu, Y., Khurana, E. and Gerstein, M. (2015). LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. Nucleic Acids Res. *43*, 8123–8134.

[Loughrey et al., 2014] Loughrey, D., Watters, K. E., Settle, A. H. and Lucks, J. B. (2014). SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. Nucleic acids research *42*, e165–e165.

[Love et al., 2014] Love, M. I., Huber, W. and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

[Lucks et al., 2011] Lucks, J. B., Mortimer, S. A., Trapnell, C., Luo, S., Aviran, S., Schroth, G. P., Pachter, L., Doudna, J. A. and Arkin, A. P. (2011). Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). Proc. Natl. Acad. Sci. U.S.A. *108*, 11063–11068.

[LYON, 1961] LYON, M. F. (1961). Gene action in the X-chromosome of the mouse (Mus musculus L.). Nature *190*, 372–373.

[Maenner et al., 2012] Maenner, S., Muller, M. and Becker, P. B. (2012). Roles of long, non-coding RNA in chromosome-wide transcription regulation: lessons from two dosage compensation systems. Biochimie *94*, 1490–1498.

[Mancini-Dinardo et al., 2006] Mancini-Dinardo, D., Steele, S. J., Levorse, J. M., Ingram, R. S. and Tilghman, S. M. (2006). Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. Genes Dev. *20*, 1268–1282.

[Mao et al., 2011] Mao, Y. S., Sunwoo, H., Zhang, B. and Spector, D. L. (2011). Direct visualization of the co-transcriptional assembly of a nuclear body by noncoding RNAs. Nat. Cell Biol. *13*, 95–101.

[Margueron and Reinberg, 2011] Margueron, R. and Reinberg, D. (2011). The Polycomb complex PRC2 and its mark in life. Nature *469*, 343–349.

[Mariner et al., 2008] Mariner, P. D., Walters, R. D., Espinoza, C. A., Drullinger, L. F., Wagner, S. D., Kugel, J. F. and Goodrich, J. A. (2008). Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. Mol. Cell *29*, 499–509.

[Marinov et al., 2014] Marinov, G. K., Williams, B. A., McCue, K., Schroth, G. P., Gertz, J., Myers, R. M. and Wold, B. J. (2014). From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res. *24*, 496–510.

[Marks et al., 2009] Marks, H., Chow, J. C., Denissov, S., Francoijs, K. J., Brockdorff, N., Heard, E. and Stunnenberg, H. G. (2009). High-resolution analysis of epigenetic changes associated with X inactivation. Genome Res. *19*, 1361–1373.

[Martianov et al., 2007] Martianov, I., Ramadass, A., Serra Barros, A., Chow, N. and Akoulitchev, A. (2007). Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. Nature *445*, 666–670.

[Martin and Wang, 2011] Martin, J. A. and Wang, Z. (2011). Next-generation transcriptome assembly. Nat. Rev. Genet. *12*, 671–682.

[Mathews, 2004] Mathews, D. H. (2004). Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. RNA *10*, 1178–1190.

[Mathews et al., 1999] Mathews, D. H., Sabina, J., Zuker, M. and Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. *288*, 911–940.

[Mayer et al., 2006] Mayer, C., Schmitz, K. M., Li, J., Grummt, I. and Santoro, R. (2006). Intergenic transcripts regulate the epigenetic state of rRNA genes. Mol. Cell *22*, 351–361.

[McCaskill, 1990] McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers *29*, 1105–1119.

[McCown et al., 2017] McCown, P. J., Corbino, K. A., Stav, S., Sherlock, M. E. and Breaker, R. R. (2017). Riboswitch diversity and distribution. RNA *23*, 995–1011.

[McHugh et al., 2015] McHugh, C. A., Chen, C. K., Chow, A., Surka, C. F., Tran, C., McDonel, P., Pandya-Jones, A., Blanco, M., Burghard, C., Moradian, A., Sweredoski, M. J., Shishkin, A. A., Su, J., Lander, E. S., Hess, S., Plath, K. and Guttman, M. (2015). The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. Nature *521*, 232–236.

[Mele et al., 2015] Mele, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., Johnson, R., Segre, A. V., Djebali, S., Niarchou, A., Wright, F. A., Lappalainen, T., Calvo, M., Getz, G., Dermitzakis, E. T., Ardlie, K. G. and Guigo, R. (2015). Human genomics. The human transcriptome across tissues and individuals. Science *348*, 660–665.

[Meller and Rattner, 2002] Meller, V. H. and Rattner, B. P. (2002). The roX genes encode redundant male-specific lethal transcripts required for targeting of the MSL complex. EMBO J. *21*, 1084–1091.

[Meller et al., 1997] Meller, V. H., Wu, K. H., Roman, G., Kuroda, M. I. and Davis, R. L. (1997). roX1 RNA paints the X chromosome of male Drosophila and is regulated by the dosage compensation system. Cell *88*, 445–457.

[Miller et al., 2011] Miller, C., Schwalb, B., Maier, K., Schulz, D., Dumcke, S., Zacher, B., Mayer, A., Sydow, J., Marcinowski, L., Dolken, L., Martin, D. E., Tresch, A. and Cramer, P. (2011). Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. Mol. Syst. Biol. *7*, 458.

[Miller et al., 2009] Miller, M. R., Robinson, K. J., Cleary, M. D. and Doe, C. Q. (2009). TU-tagging: cell type-specific RNA isolation from intact complex tissues. Nat. Methods *6*, 439–441.

[Minajigi et al., 2015] Minajigi, A., Froberg, J., Wei, C., Sunwoo, H., Kesner, B., Colognori, D., Lessing, D., Payer, B., Boukhali, M., Haas, W. and Lee, J. T. (2015). Chromosomes. A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. Science *349*.

[Mitsuya et al., 1999] Mitsuya, K., Meguro, M., Lee, M. P., Katoh, M., Schulz, T. C., Kugoh, H., Yoshida, M. A., Niikawa, N., Feinberg, A. P. and Oshimura, M. (1999). LIT1, an imprinted antisense RNA in the human KvLQT1 locus identified by screening for differentially expressed transcripts using monochromosomal hybrids. Hum. Mol. Genet. *8*, 1209–1217.

[Mitton-Fry et al., 2010] Mitton-Fry, R. M., DeGregorio, S. J., Wang, J., Steitz, T. A. and Steitz, J. A. (2010). Poly(A) tail recognition by a viral RNA element through assembly of a triple helix. Science *330*, 1244–1247.

[Moore and Steitz, 2002] Moore, P. B. and Steitz, T. A. (2002). The involvement of RNA in ribosome function. Nature *418*, 229–235.

[Mortazavi et al., 2008] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628.

[Mortimer et al., 2014] Mortimer, S. A., Kidwell, M. A. and Doudna, J. A. (2014). Insights into RNA structure and function from genome-wide studies. Nat. Rev. Genet. *15*, 469–479.

[Mueller et al., 1998] Mueller, E. G., Buck, C. J., Palenchar, P. M., Barnhart, L. E. and Paulson, J. L. (1998). Identification of a gene involved in the generation of 4-thiouridine in tRNA. Nucleic Acids Res. *26*, 2606–2610.

[Mustoe et al., 2018] Mustoe, A. M., Busan, S., Rice, G. M., Hajdin, C. E., Peterson, B. K., Ruda, V. M., Kubica, N., Nutiu, R., Baryza, J. L. and Weeks, K. M. (2018). Pervasive Regulatory Functions of mRNA Structure Revealed by High-Resolution SHAPE Probing. Cell *173*, 181–195.

[Nagalakshmi et al., 2008] Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science *320*, 1344–1349.

[Nagano et al., 2008] Nagano, T., Mitchell, J. A., Sanz, L. A., Pauler, F. M., Ferguson-Smith, A. C., Feil, R. and Fraser, P. (2008). The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. Science *322*, 1717–1720.

[Nakagawa, 2016] Nakagawa, S. (2016). Lessons from reverse-genetic studies of lncRNAs. Biochim. Biophys. Acta *1859*, 177–183.

[Neilson and Sharp, 2008] Neilson, J. R. and Sharp, P. A. (2008). Small RNA regulators of gene expression. Cell *134*, 899–902.

[Neymotin et al., 2014] Neymotin, B., Athanasiadou, R. and Gresham, D. (2014). Determination of in vivo RNA kinetics using RATE-seq. RNA *20*, 1645–1652.

[Novikova et al., 2012] Novikova, I. V., Hennelly, S. P. and Sanbonmatsu, K. Y. (2012). Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. Nucleic Acids Res. *40*, 5034–5051.

[Ogawa et al., 2008] Ogawa, Y., Sun, B. K. and Lee, J. T. (2008). Intersection of the RNA interference and X-inactivation pathways. Science *320*, 1336–1341.

[Okazaki et al., 2002] Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., Yamanaka, I., Kiyosawa, H., Yagi, K., Tomaru, Y., Hasegawa, Y., Nogami, A., Schonbach, C., Gojobori, T., Baldarelli, R., Hill, D. P., Bult, C., Hume, D. A., Quackenbush, J., Schriml, L. M., Kanapin, A., Matsuda, H., Batalov, S., Beisel, K. W., Blake, J. A., Bradt, D., Brusic, V., Chothia, C., Corbani, L. E., Cousins, S., Dalla, E., Dragani, T. A., Fletcher, C. F., Forrest, A., Frazer, K. S., Gaasterland, T., Gariboldi, M., Gissi, C., Godzik, A., Gough, J., Grimmond, S., Gustincich, S., Hirokawa, N., Jackson, I. J., Jarvis, E. D., Kanai, A., Kawaji, H., Kawasawa, Y., Kedzierski, R. M., King, B. L., Konagaya, A., Kurochkin, I. V., Lee, Y., Lenhard, B., Lyons, P. A., Maglott, D. R., Maltais, L., Marchionni, L., McKenzie, L., Miki, H., Nagashima, T., Numata, K., Okido, T., Pavan, W. J., Pertea, G., Pesole, G., Petrovsky, N., Pillai, R., Pontius, J. U., Qi, D., Ramachandran, S., Ravasi, T., Reed, J. C., Reed, D. J., Reid, J., Ring, B. Z., Ringwald, M., Sandelin, A., Schneider, C., Semple, C. A., Setou, M., Shimada, K., Sultana, R., Takenaka, Y., Taylor, M. S., Teasdale, R. D., Tomita, M., Verardo, R., Wagner, L., Wahlestedt, C., Wang, Y., Watanabe, Y., Wells, C., Wilming, L. G., Wynshaw-Boris, A., Yanagisawa, M., Yang, I., Yang, L., Yuan, Z., Zavolan, M., Zhu, Y., Zimmer, A., Carninci, P., Hayatsu, N., Hirozane-Kishikawa, T., Konno, H., Nakamura, M., Sakazume, N., Sato, K., Shiraki, T., Waki, K., Kawai, J., Aizawa, K., Arakawa, T., Fukuda, S., Hara, A., Hashizume, W., Imotani, K., Ishii, Y., Itoh, M., Kagawa, I., Miyazaki, A., Sakai, K., Sasaki, D., Shibata, K., Shinagawa, A., Yasunishi, A., Yoshino, M., Waterston, R., Lander, E. S., Rogers, J., Birney, E. and Hayashizaki, Y. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature *420*, 563–573.

[Oliver et al., 2015] Oliver, P. L., Chodroff, R. A., Gosal, A., Edwards, B., Cheung, A. F., Gomez-Rodriguez, J., Elliot, G., Garrett, L. J., Lickiss, T., Szele, F., Green, E. D., Molnar, Z. and Ponting, C. P. (2015). Disruption of Visc-2, a Brain-Expressed Conserved Long Noncoding RNA, Does Not Elicit an Overt Anatomical or Behavioral Phenotype. Cereb. Cortex *25*, 3572–3585.

[Orom and Shiekhattar, 2013] Orom, U. A. and Shiekhattar, R. (2013). Long noncoding RNAs usher in a new era in the biology of enhancers. Cell *154*, 1190–1193.

[Pachter, 2014] Pachter, L. (2014). Estimating number of transcripts from RNA-Seq measurements (and why I believe in paywall). Bits of DNA *154*, 1190–1193.

[Pandey et al., 2008] Pandey, R. R., Mondal, T., Mohammad, F., Enroth, S., Redrup, L., Komorowski, J., Nagano, T., Mancini-Dinardo, D. and Kanduri, C. (2008). Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol. Cell *32*, 232–246.

[Park, 2009] Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. *10*, 669–680.

[Peric-Hupkes et al., 2010] Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S. W., Solovei, I., Brugman, W., Graf, S., Flicek, P., Kerkhoven, R. M., van Lohuizen, M., Reinders, M., Wessels, L. and van Steensel, B. (2010). Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. Mol. Cell *38*, 603–613.

[Pinter et al., 2012] Pinter, S. F., Sadreyev, R. I., Yildirim, E., Jeon, Y., Ohsumi, T. K., Borowsky, M. and Lee, J. T. (2012). Spreading of X chromosome inactivation via a hierarchy of defined Polycomb stations. Genome Res. *22*, 1864–1876.

[Plath et al., 2003] Plath, K., Fang, J., Mlynarczyk-Evans, S. K., Cao, R., Worringer, K. A., Wang, H., de la Cruz, C. C., Otte, A. P., Panning, B. and Zhang, Y. (2003). Role of histone H3 lysine 27 methylation in X inactivation. Science *300*, 131–135.

[Polioudakis et al., 2015] Polioudakis, D., Abell, N. S. and Iyer, V. R. (2015). MiR-191 Regulates Primary Human Fibroblast Proliferation and Directly Targets Multiple Oncogenes. PLoS ONE *10*, e0126535.

[Poliseno et al., 2010] Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W. J. and Pandolfi, P. P. (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature *465*, 1033–1038.

[Ponjavic et al., 2007] Ponjavic, J., Ponting, C. P. and Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res. *17*, 556–565.

[Pontier and Gribnau, 2011] Pontier, D. B. and Gribnau, J. (2011). Xist regulation and function explored. Hum. Genet. *130*, 223–236.

[Postepska-Igielska et al., 2015] Postepska-Igielska, A., Giwojna, A., Gasri-Plotnitsky, L., Schmitt, N., Dold, A., Ginsberg, D. and Grummt, I. (2015). LncRNA Khps1 Regulates Expression of the Proto-oncogene SPHK1 via Triplex-Mediated Changes in Chromatin Structure. Mol. Cell *60*, 626–636.

[Prensner et al., 2011] Prensner, J. R., Iyer, M. K., Balbin, O. A., Dhanasekaran, S. M., Cao, Q., Brenner, J. C., Laxman, B., Asangani, I. A., Grasso, C. S., Kominsky, H. D., Cao, X., Jing, X., Wang, X., Siddiqui, J., Wei, J. T., Robinson, D., Iyer, H. K., Palanisamy, N., Maher, C. A. and Chinnaiyan, A. M. (2011). Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat. Biotechnol. *29*, 742–749.

[Pruitt et al., 2014] Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., McGarvey, K. M., Murphy, M. R., O'Leary, N. A., Pujar, S., Rajput, B., Rangwala, S. H., Riddick, L. D., Shkeda, A., Sun, H., Tamez, P., Tully, R. E., Wallin, C., Webb, D., Weber, J., Wu, W., DiCuccio, M., Kitts, P., Maglott, D. R., Murphy, T. D. and Ostell, J. M. (2014). RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. *42*, D756–763.

[Qian et al., 1992] Qian, L., Vu, M. N., Carter, M. and Wilkinson, M. F. (1992). A spliced intron accumulates as a lariat in the nucleus of T cells. Nucleic Acids Res. *20*, 5345–5350.

[Quek et al., 2015] Quek, X. C., Thomson, D. W., Maag, J. L., Bartonicek, N., Signal, B., Clark, M. B., Gloss, B. S. and Dinger, M. E. (2015). lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. Nucleic Acids Res. *43*, D168–173.

[Rabani et al., 2011] Rabani, M., Levin, J. Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N., Amit, I. and Regev, A. (2011). Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. Nat. Biotechnol. *29*, 436–442.

[Rabani et al., 2014] Rabani, M., Raychowdhury, R., Jovanovic, M., Rooney, M., Stumpo, D. J., Pauli, A., Hacohen, N., Schier, A. F., Blackshear, P. J., Friedman, N., Amit, I. and Regev, A. (2014). High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. Cell *159*, 1698–1710.

[Radecki et al., 2018] Radecki, P., Ledda, M. and Aviran, S. (2018). Automated Recognition of RNA Structure Motifs by Their SHAPE Data Signatures. Genes

(Basel) *9*.

[Raj et al., 2008] Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. and Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. Nat. Methods *5*, 877–879.

[Ramaswami et al., 2012] Ramaswami, G., Lin, W., Piskol, R., Tan, M. H., Davis, C. and Li, J. B. (2012). Accurate identification of human Alu and non-Alu RNA editing sites. Nat. Methods *9*, 579–581.

[Rinn et al., 2007] Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., Goodnough, L. H., Helms, J. A., Farnham, P. J., Segal, E. and Chang, H. Y. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311–1323.

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

[Robinson and Smyth, 2007] Robinson, M. D. and Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. Bioinformatics *23*, 2881–2887.

[Robinson and Smyth, 2008] Robinson, M. D. and Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics *9*, 321–332.

[Rossi et al., 2015] Rossi, A., Kontarakis, Z., Gerri, C., Nolte, H., Holper, S., Kruger, M. and Stainier, D. Y. (2015). Genetic compensation induced by deleterious mutations but not gene knockdowns. Nature *524*, 230–233.

[Roth and Breaker, 2009] Roth, A. and Breaker, R. R. (2009). The structural and functional diversity of metabolite-binding riboswitches. Annu. Rev. Biochem. *78*, 305–334.

[Rouskin et al., 2014] Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. and Weissman, J. S. (2014). Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature *505*, 701–705.

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

[Ruegger and Grosshans, 2012] Ruegger, S. and Grosshans, H. (2012). MicroRNA turnover: when, how, and why. Trends Biochem. Sci. *37*, 436–446.

[Sarma et al., 2010] Sarma, K., Levasseur, P., Aristarkhov, A. and Lee, J. T. (2010). Locked nucleic acids (LNAs) reveal sequence requirements and kinetics of Xist RNA localization to the X chromosome. Proc. Natl. Acad. Sci. U.S.A. *107*, 22196–22201.

[Sauvageau et al., 2013] Sauvageau, M., Goff, L. A., Lodato, S., Bonev, B., Groff, A. F., Gerhardinger, C., Sanchez-Gomez, D. B., Hacisuleyman, E., Li, E., Spence, M., Liapis, S. C., Mallard, W., Morse, M., Swerdel, M. R., D'Ecclessis, M. F., Moore, J. C., Lai, V., Gong, G., Yancopoulos, G. D., Frendewey, D., Kellis, M., Hart, R. P., Valenzuela, D. M., Arlotta, P. and Rinn, J. L. (2013). Multiple knockout mouse models reveal lincRNAs are required for life and brain development. Elife *2*, e01749.

[Schmitz et al., 2010] Schmitz, K. M., Mayer, C., Postepska, A. and Grummt, I. (2010). Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. Genes Dev. *24*, 2264–2269.

[Schofield et al., 2018] Schofield, J. A., Duffy, E. E., Kiefer, L., Sullivan, M. C. and Simon, M. D. (2018). TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding. Nat. Methods *15*, 221–225.

[Schwartz et al., 2014] Schwartz, S., Bernstein, D. A., Mumbach, M. R., Jovanovic, M., Herbst, R. H., Leon-Ricardo, B. X., Engreitz, J. M., Guttman, M., Satija, R., Lander, E. S., Fink, G. and Regev, A. (2014). Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. Cell *159*, 148–162.

[Selega and Sanguinetti, 2016] Selega, A. and Sanguinetti, G. (2016). Trends and challenges in computational RNA biology. Genome Biol. *17*, 253.

[Sexton et al., 2017] Sexton, A. N., Wang, P. Y., Rutenberg-Schoenberg, M. and Simon, M. D. (2017). Interpreting Reverse Transcriptase Termination and Mutation Events for Greater Insight into the Chemical Probing of RNA. Biochemistry *56*, 4713–4721.

[Sharon et al., 2013] Sharon, D., Tilgner, H., Grubert, F. and Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. Nat. Biotechnol. *31*, 1009–1014.

[Shechner et al., 2015] Shechner, D. M., Hacisuleyman, E., Younger, S. T. and Rinn, J. L. (2015). Multiplexable, locus-specific targeting of long RNAs with CRISPR-Display. Nat. Methods *12*, 664–670.

[Sheik Mohamed et al., 2010] Sheik Mohamed, J., Gaughwin, P. M., Lim, B., Robson, P. and Lipovich, L. (2010). Conserved long noncoding RNAs transcriptionally regulated by Oct4 and Nanog modulate pluripotency in mouse embryonic stem cells. RNA *16*, 324–337.

[Siegfried et al., 2014] Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. and Weeks, K. M. (2014). RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nat. Methods *11*, 959–965.

[Sigova et al., 2015] Sigova, A. A., Abraham, B. J., Ji, X., Molinie, B., Hannett, N. M., Guo, Y. E., Jangi, M., Giallourakis, C. C., Sharp, P. A. and Young, R. A. (2015). Transcription factor trapping by RNA in gene regulatory elements. Science *350*, 978–981.

[Silverman et al., 2014] Silverman, I. M., Li, F., Alexander, A., Goff, L., Trapnell, C., Rinn, J. L. and Gregory, B. D. (2014). RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome. Genome Biol. *15*, R3.

[Simon and Kingston, 2013] Simon, J. A. and Kingston, R. E. (2013). Occupying chromatin: Polycomb mechanisms for getting to genomic targets, stopping transcriptional traffic, and staying put. Mol. Cell *49*, 808–824.

[Simon, 2016] Simon, M. D. (2016). Insight into lncRNA biology using hybridization capture analyses. Biochim. Biophys. Acta *1859*, 121–127.

[Simon et al., 2013] Simon, M. D., Pinter, S. F., Fang, R., Sarma, K., Rutenberg-Schoenberg, M., Bowman, S. K., Kesner, B. A., Maier, V. K., Kingston, R. E. and Lee, J. T. (2013). High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. Nature *504*, 465–469.

[Simon et al., 2011] Simon, M. D., Wang, C. I., Kharchenko, P. V., West, J. A., Chapman, B. A., Alekseyenko, A. A., Borowsky, M. L., Kuroda, M. I. and Kingston, R. E. (2011). The genomic binding sites of a noncoding RNA. Proc. Natl. Acad. Sci. U.S.A. *108*, 20497–20502.

[Slavoff et al., 2013] Slavoff, S. A., Mitchell, A. J., Schwaid, A. G., Cabili, M. N., Ma, J., Levin, J. Z., Karger, A. D., Budnik, B. A., Rinn, J. L. and Saghatelian, A. (2013). Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat. Chem. Biol. *9*, 59–64.

[Sloma and Mathews, 2015] Sloma, M. F. and Mathews, D. H. (2015). Improving RNA secondary structure prediction with structure mapping data. Meth. Enzymol. *553*, 91–114.

[Smilinich et al., 1999] Smilinich, N. J., Day, C. D., Fitzpatrick, G. V., Caldwell, G. M., Lossie, A. C., Cooper, P. R., Smallwood, A. C., Joyce, J. A., Schofield, P. N., Reik, W., Nicholls, R. D., Weksberg, R., Driscoll, D. J., Maher, E. R., Shows, T. B. and Higgins, M. J. (1999). A maternally methylated CpG island in KvLQT1 is associated with an antisense paternal transcript and loss of imprinting in Beckwith-Wiedemann syndrome. Proc. Natl. Acad. Sci. U.S.A. *96*, 8064–8069.

[Smola et al., 2015] Smola, M. J., Calabrese, J. M. and Weeks, K. M. (2015). Detection of RNA-Protein Interactions in Living Cells with SHAPE. Biochemistry *54*, 6867–6875.

[Smola et al., 2016] Smola, M. J., Christy, T. W., Inoue, K., Nicholson, C. O., Friedersdorf, M., Keene, J. D., Lee, D. M., Calabrese, J. M. and Weeks, K. M. (2016). SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells. Proc. Natl. Acad. Sci. U.S.A. *113*, 10322–10327.

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

[Soruco et al., 2013] Soruco, M. M., Chery, J., Bishop, E. P., Siggers, T., Tolstorukov, M. Y., Leydon, A. R., Sugden, A. U., Goebel, K., Feng, J., Xia, P., Vedenko, A., Bulyk, M. L., Park, P. J. and Larschan, E. (2013). The CLAMP protein links the MSL complex to the X chromosome during Drosophila dosage compensation. Genes Dev. *27*, 1551–1556.

[Spitale et al., 2015a] Spitale, J. N., Hurford, T. A., Rhoden, A. R., Berkson, E. E. and Platts, S. S. (2015a). Curtain eruptions from Enceladus' south-polar terrain. Nature *521*, 57–60.

[Spitale et al., 2015b] Spitale, R. C., Flynn, R. A., Zhang, Q. C., Crisalli, P., Lee, B., Jung, J. W., Kuchelmeister, H. Y., Batista, P. J., Torre, E. A., Kool, E. T. and Chang, H. Y. (2015b). Structural imprints in vivo decode RNA regulatory mechanisms. Nature *519*, 486–490.

[Splinter et al., 2011] Splinter, E., de Wit, E., Nora, E. P., Klous, P., van de Werken, H. J., Zhu, Y., Kaaij, L. J., van Ijcken, W., Gribnau, J., Heard, E. and de Laat, W. (2011). The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. Genes Dev. *25*, 1371–1383.

[Stoffregen et al., 2011] Stoffregen, E. P., Donley, N., Stauffer, D., Smith, L. and Thayer, M. J. (2011). An autosomal locus that controls chromosome-wide replication timing and mono-allelic expression. Hum. Mol. Genet. *20*, 2366–2378.

[Stricker et al., 2008] Stricker, S. H., Steenpass, L., Pauler, F. M., Santoro, F., Latos, P. A., Huang, R., Koerner, M. V., Sloane, M. A., Warczok, K. E. and Barlow, D. P. (2008). Silencing and transcriptional properties of the imprinted Airn ncRNA are independent of the endogenous promoter. EMBO J. *27*, 3116–3128.

[Sun et al., 2013] Sun, L., Goff, L. A., Trapnell, C., Alexander, R., Lo, K. A., Hacisuleyman, E., Sauvageau, M., Tazon-Vega, B., Kelley, D. R., Hendrickson, D. G.,

Yuan, B., Kellis, M., Lodish, H. F. and Rinn, J. L. (2013). Long noncoding RNAs regulate adipogenesis. Proc. Natl. Acad. Sci. U.S.A. *110*, 3387–3392.

[Sun et al., 2012] Sun, M., Schwalb, B., Schulz, D., Pirkl, N., Etzold, S., Lariviere, L., Maier, K. C., Seizl, M., Tresch, A. and Cramer, P. (2012). Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. Genome Res. *22*, 1350–1359.

[Tan et al., 2015] Tan, J. Y., Sirey, T., Honti, F., Graham, B., Piovesan, A., Merkenschlager, M., Webber, C., Ponting, C. P. and Marques, A. C. (2015). Extensive microRNA-mediated crosstalk between lncRNAs and mRNAs in mouse embryonic stem cells. Genome Res. *25*, 655–666.

[Tani and Akimitsu, 2012] Tani, H. and Akimitsu, N. (2012). Genome-wide technology for determining RNA stability in mammalian cells: historical perspective and recent advantages based on modified nucleotide labeling. RNA Biol *9*, 1233–1238.

[Tani et al., 2012] Tani, H., Mizutani, R., Salam, K. A., Tano, K., Ijiri, K., Wakamatsu, A., Isogai, T., Suzuki, Y. and Akimitsu, N. (2012). Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. Genome Res. *22*, 947–956.

[Tay et al., 2014] Tay, Y., Rinn, J. and Pandolfi, P. P. (2014). The multilayered complexity of ceRNA crosstalk and competition. Nature *505*, 344–352.

[Thorvaldsdottir et al., 2013] Thorvaldsdottir, H., Robinson, J. T. and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinformatics *14*, 178–192.

[Tilgner et al., 2015] Tilgner, H., Jahanbani, F., Blauwkamp, T., Moshrefi, A., Jaeger, E., Chen, F., Harel, I., Bustamante, C. D., Rasmussen, M. and Snyder,

M. P. (2015). Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. Nat. Biotechnol. *33*, 736–742.

[Tilgner et al., 2012] 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.

[Ting et al., 2013] Ting, H. J., Messing, J., Yasmin-Karim, S. and Lee, Y. F. (2013). Identification of microRNA-98 as a therapeutic target inhibiting prostate cancer growth and a biomarker induced by vitamin D. J. Biol. Chem. *288*, 1–9.

[Tome et al., 2014] Tome, J. M., Ozer, A., Pagano, J. M., Gheba, D., Schroth, G. P. and Lis, J. T. (2014). Comprehensive analysis of RNA-protein interactions by high-throughput sequencing-RNA affinity profiling. Nat. Methods *11*, 683–688.

[Toor et al., 2008] Toor, N., Keating, K. S., Taylor, S. D. and Pyle, A. M. (2008). Crystal structure of a self-spliced group II intron. Science *320*, 77–82.

[Trapnell et al., 2010] 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.

[Turner et al., 1988] Turner, D. H., Sugimoto, N. and Freier, S. M. (1988). RNA structure prediction. Annu Rev Biophys Biophys Chem *17*, 167–192.

[Ulitsky and Bartel, 2013] Ulitsky, I. and Bartel, D. P. (2013). lincRNAs: genomics, evolution, and mechanisms. Cell *154*, 26–46.

[Ulitsky et al., 2011] Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H. and Bartel, D. P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell *147*, 1537–1550.

[Vance and Ponting, 2014] Vance, K. W. and Ponting, C. P. (2014). Transcriptional regulatory functions of nuclear long noncoding RNAs. Trends Genet. *30*, 348–355.

[Vance et al., 2014] Vance, K. W., Sansom, S. N., Lee, S., Chalei, V., Kong, L., Cooper, S. E., Oliver, P. L. and Ponting, C. P. (2014). The long non-coding RNA Paupar regulates the expression of both local and distal genes. EMBO J. *33*, 296–311.

[Vaziri et al., 2018] Vaziri, S., Koehl, P. and Aviran, S. (2018). Extracting information from RNA SHAPE data: Kalman filtering approach. PLoS ONE *13*, e0207029.

[Volders et al., 2015] Volders, P. J., Verheggen, K., Menschaert, G., Vandepoele, K., Martens, L., Vandesompele, J. and Mestdagh, P. (2015). An update on LNCipedia: a database for annotated human lncRNA sequences. Nucleic Acids Res. *43*, D174–180.

[Wagner et al., 2012] Wagner, G. P., Kin, K. and Lynch, V. J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. *131*, 281–285.

[Walsh et al., 2014] Walsh, A. L., Tuzova, A. V., Bolton, E. M., Lynch, T. H. and Perry, A. S. (2014). Long noncoding RNAs and prostate carcinogenesis: the missing 'linc'? Trends Mol Med *20*, 428–436.

[Wang et al., 2011] Wang, K. C., Yang, Y. W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B. R., Protacio, A., Flynn, R. A., Gupta,

R. A., Wysocka, J., Lei, M., Dekker, J., Helms, J. A. and Chang, H. Y. (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature *472*, 120–124.

[Wang et al., 2014] Wang, X., Lu, Z., Gomez, A., Hon, G. C., Yue, Y., Han, D., Fu, Y., Parisien, M., Dai, Q., Jia, G., Ren, B., Pan, T. and He, C. (2014). N6-methyladenosine-dependent regulation of messenger RNA stability. Nature *505*, 117–120.

[Washietl et al., 2011] Washietl, S., Findeiss, S., Muller, S. A., Kalkhof, S., von Bergen, M., Hofacker, I. L., Stadler, P. F. and Goldman, N. (2011). RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. RNA *17*, 578–594.

[Watters et al., 2018] Watters, K. E., Choudhary, K., Aviran, S., Lucks, J. B., Perry, K. L. and Thompson, J. R. (2018). Probing of RNA structures in a positive sense RNA virus reveals selection pressures for structural elements. Nucleic Acids Res. *46*, 2573–2584.

[Weinstein et al., 2013] Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Chang, K., Creighton, C. J., Davis, C., Donehower, L., Drummond, J., Wheeler, D., Ally, A., Balasundaram, M., Birol, I., Butterfield, S. N., Chu, A., Chuah, E., Chun, H. J., Dhalla, N., Guin, R., Hirst, M., Hirst, C., Holt, R. A., Jones, S. J., Lee, D., Li, H. I., Marra, M. A., Mayo, M., Moore, R. A., Mungall, A. J., Robertson, A. G., Schein, J. E., Sipahimalani, P., Tam, A., Thiessen, N., Varhol, R. J., Beroukhim, R., Bhatt, A. S., Brooks, A. N., Cherniack, A. D., Freeman, S. S., Gabriel, S. B., Helman, E., Jung, J., Meyerson, M., Ojesina, A. I., Pedamallu, C. S., Saksena, G., Schumacher, S. E., Tabak, B., Zack, T., Lander, E. S., Bristow, C. A., Hadjipanayis, A., Haseley, P., Kucherlapati, R., Lee, S., Lee, E., Luquette, L. J., Mahadeshwar,

H. S., Pantazi, A., Parfenov, M., Park, P. J., Protopopov, A., Ren, X., Santoso, N., Seidman, J., Seth, S., Song, X., Tang, J., Xi, R., Xu, A. W., Yang, L., Zeng, D., Auman, J. T., Balu, S., Buda, E., Fan, C., Hoadley, K. A., Jones, C. D., Meng, S., Mieczkowski, P. A., Parker, J. S., Perou, C. M., Roach, J., Shi, Y., Silva, G. O., Tan, D., Veluvolu, U., Waring, S., Wilkerson, M. D., Wu, J., Zhao, W., Bodenheimer, T., Hayes, D. N., Hoyle, A. P., Jeffreys, S. R., Mose, L. E., Simons, J. V., Soloway, M. G., Baylin, S. B., Berman, B. P., Bootwalla, M. S., Danilova, L., Herman, J. G., Hinoue, T., Laird, P. W., Rhie, S. K., Shen, H., Triche, T., Weisenberger, D. J., Carter, S. L., Cibulskis, K., Chin, L., Zhang, J., Getz, G., Sougnez, C., Wang, M., Saksena, G., Carter, S. L., Cibulskis, K., Chin, L., Zhang, J., Getz, G., Dinh, H., Doddapaneni, H. V., Gibbs, R., Gunaratne, P., Han, Y., Kalra, D., Kovar, C., Lewis, L., Morgan, M., Morton, D., Muzny, D., Reid, J., Xi, L., Cho, J., DiCara, D., Frazer, S., Gehlenborg, N., Heiman, D. I., Kim, J., Lawrence, M. S., Lin, P., Liu, Y., Noble, M. S., Stojanov, P., Voet, D., Zhang, H., Zou, L., Stewart, C., Bernard, B., Bressler, R., Eakin, A., Iype, L., Knijnenburg, T., Kramer, R., Kreisberg, R., Leinonen, K., Lin, J., Liu, Y., Miller, M., Reynolds, S. M., Rovira, H., Shmulevich, I., Thorsson, V., Yang, D., Zhang, W., Amin, S., Wu, C. J., Wu, C. C., Akbani, R., Aldape, K., Baggerly, K. A., Broom, B., Casasent, T. D., Cleland, J., Creighton, C., Dodda, D., Edgerton, M., Han, L., Herbrich, S. M., Ju, Z., Kim, H., Lerner, S., Li, J., Liang, H., Liu, W., Lorenzi, P. L., Lu, Y., Melott, J., Mills, G. B., Nguyen, L., Su, X., Verhaak, R., Wang, W., Weinstein, J. N., Wong, A., Yang, Y., Yao, J., Yao, R., Yoshihara, K., Yuan, Y., Yung, A. K., Zhang, N., Zheng, S., Ryan, M., Kane, D. W., Aksoy, B. A., Ciriello, G., Dresdner, G., Gao, J., Gross, B., Jacobsen, A., Kahles, A., Ladanyi, M., Lee, W., Lehmann, K. V., Miller, M. L., Ramirez, R., Ratsch, G., Reva, B., Sander, C., Schultz, N., Senbabaoglu, Y., Shen, R., Sinha, R., Sumer, S. O., Sun, Y., Taylor, B. S., Weinhold, N., Fei, S., Spellman, P., Benz, C., Carlin, D., Cline, M., Craft, B., Ellrott, K., Goldman, M., Haussler,

D., Ma, S., Ng, S., Paull, E., Radenbaugh, A., Salama, S., Sokolov, A., Stuart, J. M., Swatloski, T., Uzunangelov, V., Waltman, P., Yau, C., Zhu, J., Hamilton, S. R., Getz, G., Sougnez, C., Abbott, S., Abbott, R., Dees, N. D., Delehaunty, K., Ding, L., Dooling, D. J., Eldred, J. M., Fronick, C. C., Fulton, R., Fulton, L. L., Kalicki-Veizer, J., Kanchi, K. L., Kandoth, C., Koboldt, D. C., Larson, D. E., Ley, T. J., Lin, L., Lu, C., Magrini, V. J., Mardis, E. R., McLellan, M. D., McMichael, J. F., Miller, C. A., O'Laughlin, M., Pohl, C., Schmidt, H., Smith, S. M., Walker, J., Wallis, J. W., Wendl, M. C., Wilson, R. K., Wylie, T., Zhang, Q., Burton, R., Jensen, M. A., Kahn, A., Pihl, T., Pot, D., Wan, Y., Levine, D. A., Black, A. D., Bowen, J., Frick, J., Gastier-Foster, J. M., Harper, H. A., Helsel, C., Leraas, K. M., Lichtenberg, T. M., McAllister, C., Ramirez, N. C., Sharpe, S., Wise, L., Zmuda, E., Chanock, S. J., Davidsen, T., Demchok, J. A., Eley, G., Felau, I., Ozenberger, B. A., Sheth, M., Sofia, H., Staudt, L., Tarnuzzer, R., Wang, Z., Yang, L., Zhang, J., Omberg, L., Margolin, A., Raphael, B. J., Vandin, F., Wu, H. T., Leiserson, M. D., Benz, S. C., Vaske, C. J., Noushmehr, H., Knijnenburg, T., Wolf, D., Van 't Veer, L., Collisson, E. A., Anastassiou, D., Ou Yang, T. H., Lopez-Bigas, N., Gonzalez-Perez, A., Tamborero, D., Xia, Z., Li, W., Cho, D. Y., Przytycka, T., Hamilton, M., McGuire, S., Nelander, S., Johansson, P., Jornsten, R., Kling, T. and Sanchez, J. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. *45*, 1113–1120.

[Werner and Ruthenburg, 2015] 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.

[Wernersson and Nielsen, 2005] Wernersson, R. and Nielsen, H. B. (2005). OligoWiz 2.0–integrating sequence feature annotation into the design of microarray probes. Nucleic Acids Res. *33*, W611–615.

[West et al., 2014] West, J. A., Davis, C. P., Sunwoo, H., Simon, M. D., Sadreyev, R. I., Wang, P. I., Tolstorukov, M. Y. and Kingston, R. E. (2014). The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. Mol. Cell *55*, 791–802.

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

[Willingham et al., 2005] Willingham, A. T., Orth, A. P., Batalov, S., Peters, E. C., Wen, B. G., Aza-Blanc, P., Hogenesch, J. B. and Schultz, P. G. (2005). A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. Science *309*, 1570–1573.

[Wilusz, 2016] Wilusz, J. E. (2016). Long noncoding RNAs: Re-writing dogmas of RNA processing and stability. Biochim. Biophys. Acta *1859*, 128–138.

[Wilusz et al., 2012] Wilusz, J. E., JnBaptiste, C. K., Lu, L. Y., Kuhn, C. D., Joshua-Tor, L. and Sharp, P. A. (2012). A triple helix stabilizes the 3' ends of long noncoding RNAs that lack poly(A) tails. Genes Dev. *26*, 2392–2407.

[Winter et al., 2009] Winter, J., Jung, S., Keller, S., Gregory, R. I. and Diederichs, S. (2009). Many roads to maturity: microRNA biogenesis pathways and their regulation. Nat. Cell Biol. *11*, 228–234.

[Wright et al., 2016] Wright, A. V., Nunez, J. K. and Doudna, J. A. (2016). Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. Cell *164*, 29–44.

[Wutz, 2011] Wutz, A. (2011). Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation. Nat. Rev. Genet. *12*, 542–553.

[Wutz et al., 2002] Wutz, A., Rasmussen, T. P. and Jaenisch, R. (2002). Chromosomal silencing and localization are mediated by different domains of Xist RNA. Nat. Genet. *30*, 167–174.

[Wutz et al., 1997] Wutz, A., Smrzka, O. W., Schweifer, N., Schellander, K., Wagner, E. F. and Barlow, D. P. (1997). Imprinted expression of the Igf2r gene depends on an intronic CpG island. Nature *389*, 745–749.

[Xie et al., 2014] Xie, C., Yuan, J., Li, H., Li, M., Zhao, G., Bu, D., Zhu, W., Wu, W., Chen, R. and Zhao, Y. (2014). NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic Acids Res. *42*, 98–103.

[Yang and Bielawski, 2000] Yang, Z. and Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. (Amst.) *15*, 496–503.

[Yildirim et al., 2013] Yildirim, E., Kirby, J. E., Brown, D. E., Mercier, F. E., Sadreyev, R. I., Scadden, D. T. and Lee, J. T. (2013). Xist RNA is a potent suppressor of hematologic cancer in mice. Cell *152*, 727–742.

[Yildirim et al., 2011] Yildirim, E., Sadreyev, R. I., Pinter, S. F. and Lee, J. T. (2011). X-chromosome hyperactivation in mammals via nonlinear relationships between chromatin states and transcription. Nat. Struct. Mol. Biol. *19*, 56–61.

[Yin et al., 2012] Yin, Q. F., Yang, L., Zhang, Y., Xiang, J. F., Wu, Y. W., Carmichael, G. G. and Chen, L. L. (2012). Long noncoding RNAs with snoRNA ends. Mol. Cell *48*, 219–230.

[Zhang et al., 2018] Zhang, J., Liu, J., Lee, D., Feng, J. J., Lochovsky, L., Lou, S., Rutenberg-Schoenberg, M. and Gerstein, M. (2018). RADAR: annotation and prioritization of variants in the post-transcriptional regulome of RNA-binding proteins. bioRxiv *A*.

[Zhang et al., 2013] Zhang, Y., Zhang, X. O., Chen, T., Xiang, J. F., Yin, Q. F., Xing, Y. H., Zhu, S., Yang, L. and Chen, L. L. (2013). Circular intronic long noncoding RNAs. Mol. Cell *51*, 792–806.

[Zhao et al., 2010] Zhao, J., Ohsumi, T. K., Kung, J. T., Ogawa, Y., Grau, D. J., Sarma, K., Song, J. J., Kingston, R. E., Borowsky, M. and Lee, J. T. (2010). Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol. Cell *40*, 939–953.

[Zhao et al., 2008] Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J. and Lee, J. T. (2008). Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science *322*, 750–756.

[Zheng et al., 2010] Zheng, Q., Ryvkin, P., Li, F., Dragomir, I., Valladares, O., Yang, J., Cao, K., Wang, L. S. and Gregory, B. D. (2010). Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in Arabidopsis. PLoS Genet. *6*, e1001141.

[Zuber et al., 2018] Zuber, J., Cabral, B. J., McFadyen, I., Mauger, D. M. and Mathews, D. H. (2018). Analysis of RNA nearest neighbor parameters reveals interdependencies and quantifies the uncertainty in RNA secondary structure prediction. RNA *24*, 1568–1582.

[Zuber et al., 2017] Zuber, J., Sun, H., Zhang, X., McFadyen, I. and Mathews, D. H. (2017). A sensitivity analysis of RNA folding nearest neighbor parameters identifies a subset of free energy parameters with the greatest impact on RNA secondary structure prediction. Nucleic Acids Res. *45*, 6168–6176.

[Zubradt et al., 2017] Zubradt, M., Gupta, P., Persad, S., Lambowitz, A. M., Weissman, J. S. and Rouskin, S. (2017). DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. Nat. Methods *14*, 75–82.

[Zuker et al., 1999] Zuker, M., Mathews, D. H. and Turner, D. H. (1999). Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In RNA biochemistry and biotechnology pp. 11–43. Springer.