THE UNIVERSITY OF CHICAGO

CHROMATIN-ENRICHED LNCRNAS (CHERNAS) ARE CELL-TYPE SPECIFIC REGULATORS OF PROXIMAL CODING GENES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF MOLECULAR GENETICS AND CELL BIOLOGY

BY

MICHAEL STEPHEN WERNER

CHICAGO, ILLINOIS

JUNE 2016

Dedicated to Scott E. Miller

"The best way to have a good idea is to have lots of ideas."

-Linus Pauling

Table of Contents

List of Figures

ACKNOWLEDGMENTS

I would like to thank Dr. Ruthenburg for his patience, mentorship, support, and for setting the 'bar' improbably high. Without question, this undeserved respect for my ability instilled the confidence I needed to complete my work. I am also grateful to the members of the Ruthenburg laboratory for their academic expertise, without which I would not have been able to perform many of the experiments hereafter described. I would also like to acknowledge the guidance of my committee members Dr. Picirilli, Dr. Staley, and Dr. Bishop to steer my project in the most useful and interesting directions. Additionally I would like to acknowledge my previous undergraduate mentors Dr. King, Dr. Price, and Dr. Schultheis (Stetson University), Dr. Grimwade and Dr. Leonard (F.I.T) and Dr. Vaughan (University of Notre Dame) for introducing me to the world of science, encouraging me to pursue my goals, and not least, dealing with my young-adult narcissism. Finally, I would like to acknowledge my family and Talia Karasov for their never-ending support.

ABSTRACT

Many long-noncoding RNAs (lncRNAs) regulate gene transcription by chromatin-associated processes such as chromatin looping, or recruitment or inhibition of chromatin-modifying enzymes. Yet to what extent the few well characterized examples are representative of the thousands of newly discovered and uncharacterized lncRNAs remains unknown. I examined the tightly-chromatin associated pool of lncRNAs in three different cell lines, and demonstrate between 50-60 pct. of lncRNAs are adhered to chromatin. I also discovered a distinct category of lncRNAs that are defined by especially tight chromatin-association. These chromatin-enriched lncRNAs (cheRNAs) are strongly correlated with the expression of neighboring genes, and transcriptional inhibition demonstrates functional enhancer roles for four cheRNAs. CheRNAs are remarkably cell-line specific, and are adjacent to genes with roles in defining the physiology of their respective tissue or cell lineages. I also demonstrate that cheRNA attachment to chromatin is mediated by stalled or active RNA polymerase, and present evidence that their evolutionary origin is linked to the spread of transposable elements. Collectively, cheRNAs are a subset of lncRNAs that may function in tissue-specific epigenetic regulation. In addition, I also described the biochemical nature of interaction between a lncRNA and its cognate protein partner WDR5 as being both specific and promiscuous, and define a plausible model to explain in vivo RNA-binding by chromatin-modifying enzymes.

INTRODUCTION

0.1 The non-coding genome, junk or functional?

After the publication of the human genome in 2001 (Venter et al., 2001; Lander et al., 2001), which revealed that humans have approximately the same amount of genes as the fruit fly or nematode, it became apparent that the seemingly more complex neurological and immune systems extant in other organisms such as ourselves cannot be explained by simply the number of genes, contrary to what had been the current thought (Lander et al., 2001). Shortly following this result, additional lines of evidence were presented that continued to raise more questions than answers (Claverie, 2005): the introduction of tiling arrays to assay RNA expression in mouse (Okazaki et al., 2002) and approximately 1 pct. of the human genome (ENCODE Project Consortium et al., 2007), and a massive sequencing effort of unbiased mouse cDNA libraries (Carninci et al., 2005), revealed that between 65-90 pct. of mammalian genomes are transcribed. If extended as a paradigm for the rest of the human genome and mammals in general, then the two surprising results of (1) only 1-2 pct of mammalian genomes consist of protein-coding genes, yet (2) the majority of mammalian genomes are transcribed, hinted at a possible new type of regulation. Could this extra genetic information, previously thought of as "junk-DNA" (Ohno, 1972) provide functionality as non-coding RNA, and perhaps account for some of the missing complexity (Mattick, 2004)?

A comparative analysis demonstrates that while there is no correlation with genome-size and developmental complexity, there is a strong correlation of the ratio of non-coding DNA to total genomic DNA (ncDNA/tgDNA) with developmental complexity (Taft and Mattick, 2004). The term "junk"-DNA was coined by Dr. Comings in 1972 (Comings, 1972), yet it was Dr. Susumu Ohno who is largely attributed with popularizing its use (Ohno, 1972). Interestingly, Dr. Ohno argued against the accumulation of additional genes for reasons that are nevertheless consistent with a functional role for non-coding RNA. His maxim was based

on the theoretical limit for the number of functional elements (i.e. genes) possible in the genome at approximately 30,000, before the deleterious mutational load would outweigh any potential beneficial role of adding a new gene (Ohno, 1972). More recent models based on regulatory overhead (notwithstanding a putative gain of function, each new gene requires its own temporal and spatial regulation) have also supported this conclusion (Gagen and Mattick, 2005). In Dr. Ohno's words "the creation of a new gene with hitherto nonexistent function is possible only if a gene becomes sheltered from the relentless pressure of natural selection," yet critically, non-coding RNA is not subject to the same mutational restraints as protein-coding genes (Mattick, 2004; Pang et al., 2006; Ponting et al., 2009; Necsulea et al., 2014). In contrast to codons for amino acids which must maintain sequence-specific fidelity and the appropriate frame relative to the start codon, only a subset of RNA positions require nucleobase fidelity to maintain a structure or protein-binding motif (Ponting et al., 2009), and even then a compensatory mutation in the affected helix could maintain base-pairing and thus sustain structure (Higgs, 1998). Nevertheless there remained controversy over whether intergenic transcribed regions could provide function, largely based on evolutionary analyses that concluded many of the these regions exhibit little to no sequence conservation (Wang et al., 2004). Indeed, as is often the case with the prevailing dogma, in this case the "junk-DNA" explanation, a significant body of work was required to demonstrate that non-coding RNA could influence gene regulation, and it would come in no small part from microRNA.

0.2 Early indications that non-coding RNA has function

The discovery that non-coding RNA could provide function is not new, however the previously identified classes such as ribosomal RNA, spliceosomal RNA, telomerase RNA, small nucleolar RNAs, transfer RNA and 7SK were largely considered ancient, and to some extent even idiosyncratic means of regulating essential cellular processes (Eddy, 2001). However, the discovery in the 90's that microRNAs lin-4 and let-7 are functional in *C. elegans* post-

embryonic development (Lee et al., 1993; Wightman et al., 1993), and present in different multicellular organisms with tissue-specific expression (Hutvágner et al., 2001; Lau et al., 2001; Lee and Ambros, 2001; Sempere et al., 2004) in the early 2000's gave significant credence to the theory that non-coding RNAs could provide tissue-specific regulation, and perhaps facilitate more complex regulatory transcriptional networks (Eddy, 2001; Kung et al., 2013). Elegant biochemical experiments described how microRNAs use the same machinery as RNAi to be processed from a larger premature transcript into a small double-stranded RNA (Hutvágner et al., 2001), which then inhibits translation of mRNAs by binding to their 3' UTRs (Filipowicz et al., 2008). This provides an example of how non-coding RNA is capable of providing function, and in a way that is distinct from proteins because they contain the same sequence-content information (nucleotides) as both mRNAs and the genome. Conceivably this could allow much faster evolution to match a cognate mRNA, or potentially a DNA locus (Mattick, 2004; Taft and Mattick, 2004). The example of microRNAs demonstrates how non-coding RNAs can be utilized to impart tissue-specific gene regulation, and was an important milestone on the way to appreciating the role of other transcribed non-coding RNA in the genome (Kung et al., 2013).

The largest category of non-coding RNA in the human genome is long non-coding RNAs (lncRNAs), ranging from between 50-95 pct. of the observed non-coding RNA species (Da Sacco et al., 2012; Iyer et al., 2015). However these RNAs, defined somewhat arbitrarily by being larger than the 200 nucleotide cutoff used for analyzing small RNAs (microRNAs, piwiRNA, snRNAs, tRNAs, snoRNAs), took longer to be appreciated (Kung et al., 2013). Some of that trepidation arose from the burden of proving that non-coding RNAs function as RNA molecules, especially with regard to lncRNAs that are potentially long enough to encode long open reading frames, and thus functional proteins (Kung et al., 2013; Rinn and Chang, 2012). However there are several characteristics that can be marshaled to argue against lncRNAs being translated and functioning as a protein. Perhaps the simplest

method is to perform an *in silico* translation of all possible frames of a transcript coupled to a BLAST search for conserved or homologous protein domains (Rinn and Chang, 2012). Additionally, for the longest open-reading frame (ORF) (AUG start to stop codon) of a given lncRNA, does the codon-bias match that of the organism from which its from? This analysis can be even more sophisticated by assessing the hexamer usage of an ORF relative to the observed hexamer frequencies, because there is a bias of certain amino-acids (and thus their corresponding codons) to be next to each other (Fickett and Tung, 1992). More simplistically, how long is the longest ORF, and how does it compare to the distribution of ORF lengths in a given organism? This simple metric is actually currently the most robust measure to separate coding from non-coding RNAs (Wang et al., 2013). Similarly, the ratio of the length of the longest ORF relative to the length of the full transcript can also be used to differentiate coding from non-coding RNA. The software program Coding Potential Assessment Tool (CPAT) aggregates many of these parameters, and collectively and efficiently provides a coding probability for any input sequence (Wang et al., 2013). However an experimental approach is always desired to compliment these analyses, such as absence of signal from mass spectrometry or ribosome-profiling data. While mass spectrometry can be useful, it tends to suffer in this respect from low signal depth and the difficulty of finding small peptides (Andrews and Rothnagel, 2014). However a targeted analysis of ribosome-profiling data, which has the advantage of high specificity and detection (Ingolia, 2104), demonstrates that lncRNAs are largely not found on ribosomes (Guttman et al., 2013).

The first lncRNA to be experimentally studied was the highly conserved and imprinted gene H19. Yet despite its high degree of nucleotide conservation, there is little conservation in any potential amino-acid reading frame, all of which contain early stop-codons (Brannan et al., 1990). Remarkably, when H19 was injected in mouse zygotes it caused lethality (Brunkow and Tilghman, 1991), and when deleted leads to an overgrowth phenotype linked to the expression of normally repressed igf2 (insulin like growth factor) (Leighton et al., 1995).

Other early functional lncRNAs are also imprinted (defined here as expressed exclusively from one parental allele), such as AIR which is required for silencing of nearby genes on the paternal allele at the igf2r (insulin-like growth factor 2 receptor) locus (Wutz et al., 1997; Sleutels et al., 2002), and perhaps the mother of all lncRNAs, XIST (Brockdorff et al., 1992; Marahrens et al., 1997; Penny et al., 1996) which is required for X-chromosome inactivation in female mammals. However the generality of these findings would take roughly a decade to be fully appreciated, largely because the mechanism of their apparent functions remained mysterious.

0.3 Model for lncRNA function: recruitment of chromatin-modifying enzymes to target genes

In the mid-2000's some of the first insights into lncRNA mechanisms connected their functions with chromatin-based mechanisms of epigenetic silencing, revealing plausible molecular roles for the "dark matter" of the genome. In 2003 Katherine Plath and colleagues demonstrated that XIST was required for polycomb repressive complex 2 (PRC2) histone methylransferase recruitment to the inactive X-chromosome, although it was not clear if XIST played a direct or indirect role in recruitment (Plath et al., 2003). Then in 2007 Dr. Howard Chang's lab at Stanford published a seminal paper describing how a lncRNA coined HOTAIR was required for recruiting PCR2 to HOXD genes (Rinn et al., 2007), and critically, that members of the PRC2 complex and CoREST were capable of binding HOTAIR *in vitro*, and could specifically Co-IP from cells, arguing that a direct interaction recruited PRC2 to a target gene, and thus established a mechanistic model for its observed function (Rinn et al., 2007; Tsai et al., 2010). Shortly thereafter it was determined XIST could also directly bind PRC2 (Zhao et al., 2008) - although more recent and more quantitative methods have largely un-raveled these results (McHugh et al., 2015; Chu et al., 2015). Nevertheless these results set in motion a series of experiments testing the requirement of lncRNAs for the recruitment

of histone-modifying enzymes to specific target genes by knockdown of lncRNAs coupled to ChIP-qPCR of the enzyme and corresponding mark, and assessment of *in vivo* binding by native Co-IP experiments. For example the lncRNA AIR (previously identified as being important for paternal imprinting at igf2r) was shown to bind the histone methyltransferase G9a (Nagano et al., 2008), and yet another imprinted lncRNA KCNQ1OT1 was capable of binding to G9a, DNMT1, and PRC2 (Mancini-Dinardo et al., 2006; Mohammad et al., 2010; Pandey et al., 2008). Importantly, knockdown of these lncRNAs resulted in loss of histone and DNA epigenetic marks at the effected genes. Because the loss of these marks effected the maintenance of repression of the underlying genes, lncRNAs became appreciated as important components of epigenetic silencing. More recent work has also implicated lncRNAs in recruiting the activating histone methyltransferases MLL/Trithorax (Wang et al., 2011; Grote et al., 2013), stimulating further investigations into the generality of lncRNAs in establishing or maintaining epigenetic information. Subsequent native immunoprecipitation experiments coupled to deep-sequencing (Khalil et al., 2009; Zhao et al., 2010) sought to identify multiple RNA species associated with histone-methyltransferases, in search of the generality of lncRNA-mediated recruitment of chromatin-modifying enzymes. While these early studies were important first-steps that advanced the field, they are fraught with the technical difficulties of isolating bona-fide *in vivo* RNA binders from random and non-specific interactions that occur in the complex and unnatural milieu of cellular or nuclear extracts (Mili and Steitz, 2004; Brockdorff, 2013), and thus have become under increasing scrutiny (Davidovich et al., 2013; McHugh et al., 2015; Chu et al., 2015).

Later experiments have attempted to mitigate this problem by using crosslinking in combination with immunoprecipitation to more rigorously wash, and thus capture *in vivo* interactions (Yang et al., 2014; Kaneko et al., 2013; Guil et al., 2012). Somewhat surprisingly, the results of these experiments have largely been inconsistent with the aforementioned model that individual lncRNAs could specifically recruit histone methyltranserases to gene targets.

For example, Kaneko et al. utilized the "gold-standard" method in the field to approach this question, PARCLIP, which incorporates a photo-activatable uracil analog into RNA that is crosslinked specifically with 365 nm wavelength radiation, and then immunoprecipitated and run-through a denaturing SDS-PAGE gel before ultimately being purified (Hafnet et al., 2010). Instead of finding discrete lncRNAs with a preference near silenced genes, the authors found that the PRC2 complex preferentially associates with RNA at the 5 prime region of active mRNA genes, that have both low occupancy of PRC2 by ChIP and that of its repressive mark H3K27me3 (Kaneko et al., 2013), despite their well established role in gene silencing (Simon and Kingston, 2009). This result is also different from a similar UV-254 crosslinking-IP experiment that demonstrated a preference for binding to introns of premature RNA (Guil et al., 2012). The conclusion was that this data represents a 'scanning' mechanism to detect whether a gene is being transcribed, and if it is, the resulting RNA can somehow inhibit (directly or indirectly) PRC2 enzymatic activity (Kaneko et al., 2013). While this model is not exactly parsimonious, PARCLIP data is hard to directly refute and thus puts into question the generality of the previously proposed model of specific recruitment (Rinn et al., 2007; Tsai et al., 2010; Rinn and Chang, 2012). Perhaps consistent with this model, other lncRNAs have been found to effect the enzymatic activity of associated proteins (Lai et al., 2013; Redon et al., 2010).

As mentioned these inconsistencies are in no small part due to the differences between native and cross-linked immunopreciptation techniques (Mili and Steitz, 2004). However an additional reason for this contradiction is that a key component of the former model was based on *in vitro* RNA binding experiments with PRC2 complex components (Brockdorff, 2013; Zhao et al., 2010, 2008; Rinn et al., 2007; Tsai et al., 2010; Wang et al., 2011). While these experiments demonstrated some modest qualitative specificity for a given RNA compared to a control RNA, more recent quantitative experiments have put these data into question. Quantitative EMSA and filter binding assays comparing different RNAs demon-

strate that the fully assembled PRC2 complex is capable of binding RNA promiscuously, and with greater affinity for longer RNAs regardless of sequence content (Davidovich et al., 2013, 2015). Still, these experiments were able to replicate a 3-8 fold affinity of PRC2 for the RepA component of XIST over other similar size-regime non-relevant RNAs depending on the buffer conditions used (Cifuentes-Rojas et al., 2014; Davidovich et al., 2015) - despite two recent separate RNA-IP mass-spectrometry experiments displaying no *in vivo* binding of PRC2 to XIST (McHugh et al., 2015; Chu et al., 2015). Additionally, luciferase reporter assays that had previously demonstrated that a two-hairpin PRC2 binding motif could reduce expression *in cis* were shown to be the result of a promoter mutation, and independent of PRC2 (Davidovich et al., 2015; Kanhere et al., 2010). Collectively, these carefully executed and controlled *in vitro* binding experiments (Davidovich et al., 2013, 2015) are actually reasonably consistent with the latter model of promiscuous binding described by the more carefully controlled *in vivo* binding experiments derived from cross-linked immunoprecipitations (Kaneko et al., 2013). However the details of how PRC2-RNA binding occurs, and exactly how it functions, potentially to inhibit PCR2 activity/deposition of H3K427me3 are still poorly understood, and to what extent these rules apply to other histone-modifying enzymes is unknown. It also important to mention that these experiments have not directly refuted the initial findings that the lncRNA HOTAIR can recruit PRC2 to specific gene targets, or other lncRNAs for other histone modifying enzymes (Nagano et al., 2008; Mohammad et al., 2010; Rinn et al., 2007; Tsai et al., 2010; Pandey et al., 2008; Wang et al., 2011; Grote et al., 2013). Moreover, despite being steeped in a mechanistic quagmire, functional experiments such as knockdown and knockout of lncRNAs clearly demonstrate that lncRNAs play an important role in histone modifications, gene regulation, and disease (Prensner et al., 2013; Chalei et al., 2014; Rinn et al., 2007; Sauvageau et al., 2013; Pandey et al., 2008; Grote et al., 2013; Wang et al., 2011)

0.4 lncRNAs frequently operate at the chromatin interface

It is also worth noting that other lncRNA-mechanisms have been discovered that act at the chromatin level. For example, lincRNA-p21 acts downstream of the oncogene p53 to facilitate further p53-induced suppression, and is required for proper p53-mediated DNA damage response (Huarte et al., 2010). This function appears to depend on a 780 nt region of lincRNA-p21 that is required to bind to the repressor protein hnRNP-K, and effects hnRNP-K localization to target genes. So similar to the proposed PRC2-recruitment model, lincRNA-p21 appears to recruit the repressive protein hnRNP-K. Importantly however, the identification of these binding events was based on multiple orthogonal experiments, including purifying proteins from biotinylated RNA Co-IP from nuclear extracts coupled to mass spectrometry, crosslinked-immunoprecipitation and RT-qPCR, and combining *in vitro* binding data with biological phenotype; mapping a short region that is (a) required for binding, and (b) required to promote p53-induced DNA damage response, thus more firmly establishing the proposed interaction (Huarte et al., 2010).

An alternative mechanism involves bridging distal enhancers to promoters of target genes via chromatin-looping. Xiang and colleagues identified a lncRNA greater than 500 kb away from the MYC gene that is transcribed from an enhancer, and is required for looping the enhancer to the MYC locus (Xiang et al., 2014). Such chromatin-looping events have been observed with other lncRNAs (Yang et al., 2013), including interactions with the Mediator complex (Lai et al., 2013) and CTCF (Li et al., 2013), and with the recently discovered class of non-coding enhancer RNAs (eRNAs) extant at enhancers defined by p300 and the histone modifications H4K4me1 and H3K27ac (Ernst et al., 2011; Lai et al., 2013; Rada-Iglesias et al., 2011; Lam et al., 2014; Lai et al., 2015). Another mechanism of lncRNA gene-regulation involves antagonizing, instead of recruiting, transcription factors to target genes. For example the lncRNA SChLAP-1 is required for certain prostate cancer metastases, and functions by inhibiting binding of the SWI/SNF chromatin-remodeling complex to target

genes (Prensner et al., 2013). Similarly, the lncRNAs Gas5 and PANDA serve as a decoys for the glucocorticoid receptor protein and NF-YA transcription factor, respectively, somehow titrating them away from target genes (Kino et al., 2010; Hung et al., 2011). In a related mechanism, the lncRNA produced from a minor promoter of the dihydrofolate reductase gene can form a complex with the major promoter and TFIIB, causing the pre-initiation complex to dissociate from the major promoter, leading to epigenetic down-regulation of dihydrofolate reductase levels (Martianov et al., 2007). Finally, lncRNAs tethered to both copies of meiotic chromosomes in fission yeast can facilitate robust pairing of homologous chromosomes (Ding et al., 2012), although fine-scale details of how this occurs remain unknown.

It is also important to mention that while the majority of mechanisms thus far discovered for lncRNAs revolve around transcriptional regulation at the chromatin interface, there a notable counterexamples. For example, the lncRNA linc-MD1 contains microRNA binding sites that serve as a "microRNA-sponge" during muscle differentiation, allowing the microRNA target mRNAs to be expressed (Cesana et al., 2011). In fact the original functional lncRNA H19 appears to bind and thus reduce the levels of soluble let-7 microRNAs (Kallen et al., 2013), bringing functional non-coding RNAs, small and large, into full circle.

The advent of cheap and increasingly deep RNA-sequencing has revealed thousands more lncRNA species in the human genome, ranging between 14,000-60,000 depending on the annotation (Cabili et al., 2011; Derrien et al., 2012; Ilott and Ponting, 2013). In addition to new annotations, new categories of non-coding RNAs have been discovered, including eRNAs (Lam et al., 2014; Andersson et al., 2014), and circular RNAs (Chen, 2016), presenting the ever greater challenge of distinguishing functional lncRNAs from transcriptional noise. Several lncRNAs have been implicated in development and disease (Chalei et al., 2014; Dinger et al., 2008; Huarte et al., 2010; Sauvageau et al., 2013; Wang et al., 2011b), and the prominent tissue-specific expression and conservation of lncRNAs as a class suggest these functions may be representative for thousands of unstudied lncRNAs (Iyer et al., 2015;

Necsulea et al., 2014; Ponting et al., 2009). If true, then non-coding RNAs could provide at least some of the missing complexity seemingly absent from the number of protein-coding genes in the human genome (Venter et al., 2001; Lander et al., 2001; Eddy, 2001).

The issue of conservation has largely abated due to the finding that lncRNA promoters are under a high degree of conservation (Guttman et al., 2009), considerations of evolutionary principles of RNA selection compared to mRNA (Quinn et al., 2016; Ponting et al., 2009; Necsulea et al., 2014), and the increasing number of mechanistic and functional studies. From the small number of mechanistic experiments described above, it appears that non-coding RNAs are capable of regulating local chromatin states, either by acting as intermediaries to recruit or block chromatin modulators (Chalei et al., 2014; Lai et al., 2013; Nagano et al., 2008; Rinn and Chang, 2012; Rinn et al., 2007) or by potentiating contacts between genes and distal enhancer-elements to promote transcriptional activation (Lai et al., 2013; Li et al., 2013; Yang et al., 2013). Yet it is still unclear whether similar chromatin-based mechanisms of gene regulation are relevant for the vast majority of unstudied lncRNAs.

0.5 Exploration of the non-coding RNA content of chromatin

To provide some insight into this issue, I sought to identify lncRNAs that are tightly-associated with chromatin, as a first-indication that they may provide similar functions to the 20-30 cases that have been explored (Rinn and Chang, 2012). In Chapter I, I employed biochemical fractionation of the nuclear compartment coupled to RNA-seq of both fractions in triplicate. Similar nuclear extraction methods have been employed to great effect in studying transcription, mRNA processing and export (Bhatt et al., 2012; Dye et al., 2006; Wuarin and Schibler, 1994). We found that the bulk of annotated lncRNAs are chromatin-enriched, suggesting widespread roles in chromatin regulation, and provide a resource for the community to explore and/or corroborate functional lncRNA mechanisms. However, Gencode and Broad lncRNA annotations accounted for only a small portion of the observed chromatin-

enriched transcripts; the majority represented a distinct subclass of lncRNAs that we term "chromatin- enriched RNA" (cheRNA). Most cheRNAs are tethered to chromatin by RNA pol II (RNAPII) and their presence correlated with neighboring gene transcriptional activity at a level similar to, or better than the current state-of-the-art active enhancer annotations (ENCODE Project Consortium et al., 2012; Ernst et al., 2011; Rada-Iglesias et al., 2011; Zentner et al., 2011). Yet cheRNAs appear distinct from recently described bi-directional transcripts that emanate from canonical active enhancers (Andersson et al., 2014; Kim et al., 2010; Wang et al., 2011a). Specifically, eRNAs are conventionally short (less than 500 nucleotides) and are found to have equal representation in both sense and anti-sense directions (Andersson et al., 2014), while cheRNAs are defined by being larger than 1000 nucleotides and are overwhelmingly unidirectional (Werner and Ruthenburg, 2015). Moreover cheRNAs exhibit transcription start site features indicative of other lncRNAs and mRNAs, such as H3K4me3, whereas eRNAs appear to be defined by the enhancer modification H3K4me1 (Andersson et al., 2014; Lam et al., 2014). Finally, eRNA positions have been mapped across the majority of human tissues, and do not appreciably (approximately 11 pct.) overlap with cheRNA loci in any of the three cell lines I have queried. Conversely, I observed that transposable elements which comprise approximately 50 pct. of the human genome, and are largely attributed with the expansion of our genome as selfish DNA elements (Giordano et al., 2007; Bourque, 2009), frequently overlap with cheRNAs (greater than 95 pct.).

Although highly chromatin-enriched (cheRNAs) exhibited a strong correlation to proximal gene expression in HEK293s (Werner and Ruthenburg, 2015), many questions remained regarding cheRNAs and their relationship to nearby genes. Do cheRNA loci have functional enhancer roles, and if so, is the RNA molecule important or is transcription an inert by-product of active enhancers (Andersson et al., 2014)? Are cheRNA-gene pairs co-regulated or independent, and are they involved in common metabolic pathways or tissue-specific gene regulation? Finally, although only 11 pct. of cheRNAs overlap bi-directional non-coding

RNA transcribed from enhancers (eRNA) (Werner and Ruthenburg, 2015), recent work suggested that a subset of 91 eRNAs were more readily detected in the chromatin-enriched RNA population, indicating some eRNAs may play similar roles in the activation of proximal immediate early genes in HeLa (Lai et al., 2015). More broadly, assigning chromatin-enrichment of nuclear transcripts in multiple cell lines should guide functional exploration of lncRNAs that display tissue-specific expression (Iyer et al., 2015; Dinger et al., 2008) and mechanistic insight for lncRNAs that have roles in disease (Sauvageau et al., 2013; Huarte, 2015). In chapter II I continued to explore cheRNAs by repeating our fractionation-sequencing method from human embryonic stem cells, K562 human myeloid leukemia cells, and K562 cells differentiated along an erythroid lineage. We found that cheRNAs are overwhelmingly cell-type specific, and that the correlation to nearby gene expression is consistent in all cell types. To test for cheRNA function we inhibited cheRNA transcription by CRISPRi (Gilbert et al., 2013, 2014) and degraded a hemin-induced cheRNA downstream of fetal-hemoglobin (HIDALGO) by anti-sense oligos, leading to a 20-60 pct. decrease in neighboring gene expression both at basal levels and during differentiation. While different cheRNAs may have different roles, these results establish at least some cheRNA molecules as cell-type specific enhancers.

We also noticed a transposable element insertion near the TSS of HIDALGO, and in a region downstream that is unique to simian primates. This is potentially intriguing because fetal-hemaglobin switching (between embryonic, fetal, and adult globins depending on developmental stage) occurs exclusively in simians. Given HIDALGOs role in activating fetal-hemoglobin, it is possible that the insertion of transposable elements in simians contributed to the evolution of this transcriptional regulation. This finding is also consistent with recent studies drawing a comparison to and potential origin of lncRNAs from transposable elements, which can provide their own promoters and regulatory regions in the form of transcription factor binding sites (Kapusta et al., 2013; Kelley and Rinn, 2012; Chuong

et al., 2016; Lynch et al., 2011, 2015). The ability of transposable elements to expand through the genome by copying and pasting themselves with reverse transcriptases and recombinases (Bourque, 2009) provides a ready source of evolutionary potential. Indeed, it appears that the spread of transposable elements carrying transcription factor binding sites adjacent to interferon response genes facilitated the evolution of the innate immune response (Chuong et al., 2016). Although not conclusive, these data provide further fodder that the expanded "junk-DNA" of the human genome and potentially other vertebrates has a role in developmental complexity (Mattick, 2004).

In chapter III I focused on the ultimate consequence of lncRNA expression and upon which it's function relies on, which is the interaction with protein-binding partners. How protein-lncRNA recognition is achieved, and untangling the specificity of binding which became an unexpected contradiction to the model of lncRNA:histone modifying enzyme recruitment (see above; Davidovich et al., 2013) is critical to fully understanding the role of lncRNAs in gene regulation. I chose to study the established and developmentally important interaction of the MLL subunit WDR5 and its cognate RNA binding-partner HOTTIP (Wang et al., 2011). We found low nanomolar affinity *in vitro* for this interaction, yet similar to other systems we found physiologically relevant binding to negative control RNAs, some of which are a thousand fold more abundant in the cell than HOTTIP, and a similar length-dependent increase in affinity as found with PRC2. Nevertheless we observed specificity *in vivo* from cross-linking immunoprecipitation experiments. We also identified two novel patches on WDR5 that bind to RNA, while also performing biophysical and biochemical experiments that cast a previously identified binding site in question (Yang et al., 2014). In an attempt to account for the apparent contradiction between *in vivo* and *in vitro* specificity I developed a simple mathematical model which includes a random probability of incorporating a small binding motif. As RNAs get longer this probability increases, thereby increasing the measured affinity for the protein. *In vivo* cognate RNA binders could contain

multiple copies of the motif, and present them through structural motifs such as hairpins. Nevertheless much work is required to test the robustness, accuracy, and generality of this model to explain RNA-protein binding data.

In conclusion, during my PhD I explored the hypothesis that non-coding RNAs regulate gene expression through chromatin-templated processes. My results show that approximately 50-60 pct. of lncRNAs are tightly associated with chromatin in multiple cells lines, hinting that they may commonly affect such mechanisms. However only the accumulation of more functional experiments can fully establish this paradigm. I also detected an extensive pool of chromatin-associated non-coding RNAs (cheRNAs) which are correlated with neighboring transcriptional activity. Perturbation of four distinct cheRNAs demonstrate enhancer function for each of their nearby genes. Subsequent experiments should focus on the contribution of cheRNAs to epigenetic memory of differentiated transcriptional programs, and examine mechanisms of cheRNA enhancer function, especially whether cheRNAs have a role in chromatin-looping. Does looping function to bridge enhancers containing activating transcription factors to the promoter of a target gene, or to facilitate recycling of polymerase back to the promoter, or does it serve both functions (O'Sullivan et al., 2004; Ansari and Hampsey, 2005; Lainé et al., 2009; Tan-Wong et al., 2009)? Additionally, functional experiments are needed to test the hypothesis that HIDALGO contributed to hemaglobin switching in primates, perhaps by introduction of luciferese reporter constructs bearing HIDALGO with and without TEs in cell lines derived from simian and prosimian apes. Moreover, combining cheRNA loci data with recent transposable element insertion could potentially provide a useful metric to identify developmentally important primate-specific enhancers - which may be especially interesting in the context of neural tissue and development. Finally, measuring fluctuations in cheRNAs during metastasis or in disease models may uncover enhancers that contribute to cancer or disease states.

Chapter 1

DISCOVERY AND CHARACTERIZATION OF CHERNAS

The content of this chapter was published as:
Werner, Michael S., and Ruthenburg, Alexander J. "Nuclear fractionation reveals thousands of chromatin-tethered noncoding RNAs adjacent to active genes." Cell Reports 12 (2015): 1089-1098.

1.1 Summary

A number of long noncoding RNAs (lncRNAs) have been reported to regulate transcription via recruitment of chromatin-modifiers or bridging distal enhancer elements to gene promoters. However, the generality of these modes of regulation and the mechanisms of chromatin attachment for thousands of unstudied human lncRNAs remain unclear. To address these questions, we performed stringent nuclear fractionation coupled to RNA-seq. We provide the first robust genome-wide identification of chromatin-associated lncRNAs, and demonstrate tethering to chromatin by RNAPII is a pervasive mechanism of attachment. We also uncovered several thousand especially chromatin-enriched RNAs (cheRNAs) that share molecular properties with known lncRNAs. Although distinct from noncoding RNAs derived from active enhancers, the production of cheRNAs is strongly correlated with the expression of neighboring protein-coding genes. This work provides an updated framework for nuclear RNA organization that includes a large chromatin-associated transcript population with enhancer-like properties, and may prove useful in de novo enhancer annotation.

1.2 Nuclear fractionation quantitatively distinguishes cis and trans-acting lncRNAs from mRNAs

Our initial aim was to identify chromatin-associated lncRNAs. We first extracted HEK293 nuclei with a buffer that effectively separates soluble and loosely bound material from the chromatin pellet, which retains tightly-bound factors (Bhatt et al., 2012; Wuarin and Schibler, 1994; Dye et al., 2006) (Figure 1.1A). We then performed RNA-seq from three biological replicates of the resulting soluble-nuclear extract (SNE) and chromatin pellet extract (CPE), yielding greater than 49 million uniquely-mapped reads from each fraction replicate. De novo assembled transcripts (Trapnell et al., 2012; Kim et al., 2013) greater than 1,000 nucleotides from the CPE were added to the latest Gencode gene annotation (Harrow et al., 2012) and each transcript was scored for its abundance in the CPE relative to SNE fractions (Figure A.1G).

We first validated our fractionation by confirming robust chromatin enrichment of two canonically chromatin associated lncRNAs, XIST and KCNQ1OT1 (Penny et al., 1996; Plath et al., 2003; Zhao et al., 2008), and soluble nuclear enrichment of the mRNAs beta-actin and GAPDH (Figure A.1A,B,D). In contrast, the trans-acting lncRNAs HOTAIR and EVF2/DLX6-AS displayed an intermediate level of solubility, consistent with proposed models that suggest both nuclear mobility and chromatin attachment (Berghoff et al., 2013; Rinn et al., 2007; Tsai et al., 2010)(Figure A.1C,D). Collectively, these data indicate biochemical fractionation of nuclei coupled to RNA-seq can distinguish mRNAs, chromatin-enriched cis-acting lncRNAs, and trans-acting lncRNAs based on their sub-nuclear compartmentalization. Applying this approach to Gencode noncoding RNAs (ncRNAs) revealed 57 pct. of lncRNAs, 52 pct. of antisense transcripts, and 28 pct. of pseudo-genes are chromatin-enriched, in contrast to 16 pct. of mRNAs. Because these values come from three biological replicates, and Cufflinks (Trapnell et al., 2013) disqualifies genes with irreproducible levels between replicates, they reflect high confidence estimates. Therefore the relative abundance

of each of these transcripts in the CPE versus SNE provides a valuable resource for further exploration of functional ncRNAs that are likely to operate at the chromatin-interface. Surprisingly, we also observed that the size of the chromatin-associated population is substantially greater than previously appreciated (Figure A.1B) (Bhatt et al., 2012) - as no previous experiment has been able to quantify the enrichment level of chromatin-associated RNAs (see below). The vast majority (approximately 90 pct.) are not present in the Gencode annotation, however an exhaustive sequencing effort that recently described 60,000 lncRNA genes provides support for approximately 3/4 of our chromatin-associated RNAs (Iyer et al., 2015).

Stripping highly abundant mRNA from the chromatin pellet with urea was critical to the discovery of novel transcripts because it effectively magnified the coverage depth of low-abundance RNA species. Indeed, far fewer reads were comprised of exons from annotated genes in the CPE relative to the SNE and whole-cell RNA sequencing (Figure A.2A). The resulting higher coverage of non-exonic portions of the transcriptome combined with the statistical power of our high-depth biological replicates enabled the first stringent assembly of chromatin-associated transcripts (see Figures A.3 for examples). These findings are distinct from previous efforts to examine chromatin-associated noncoding RNA (Djebali et al., 2012, Mondal et al., 2010) in that the fractionation method we employed efficiently isolated tightly-chromatin associated RNAs from nucleoplasmic species (Bhatt et al., 2012; Dye et al., 2006; Wuarin and Schibler, 1994) (compare Figure A.1A and I).

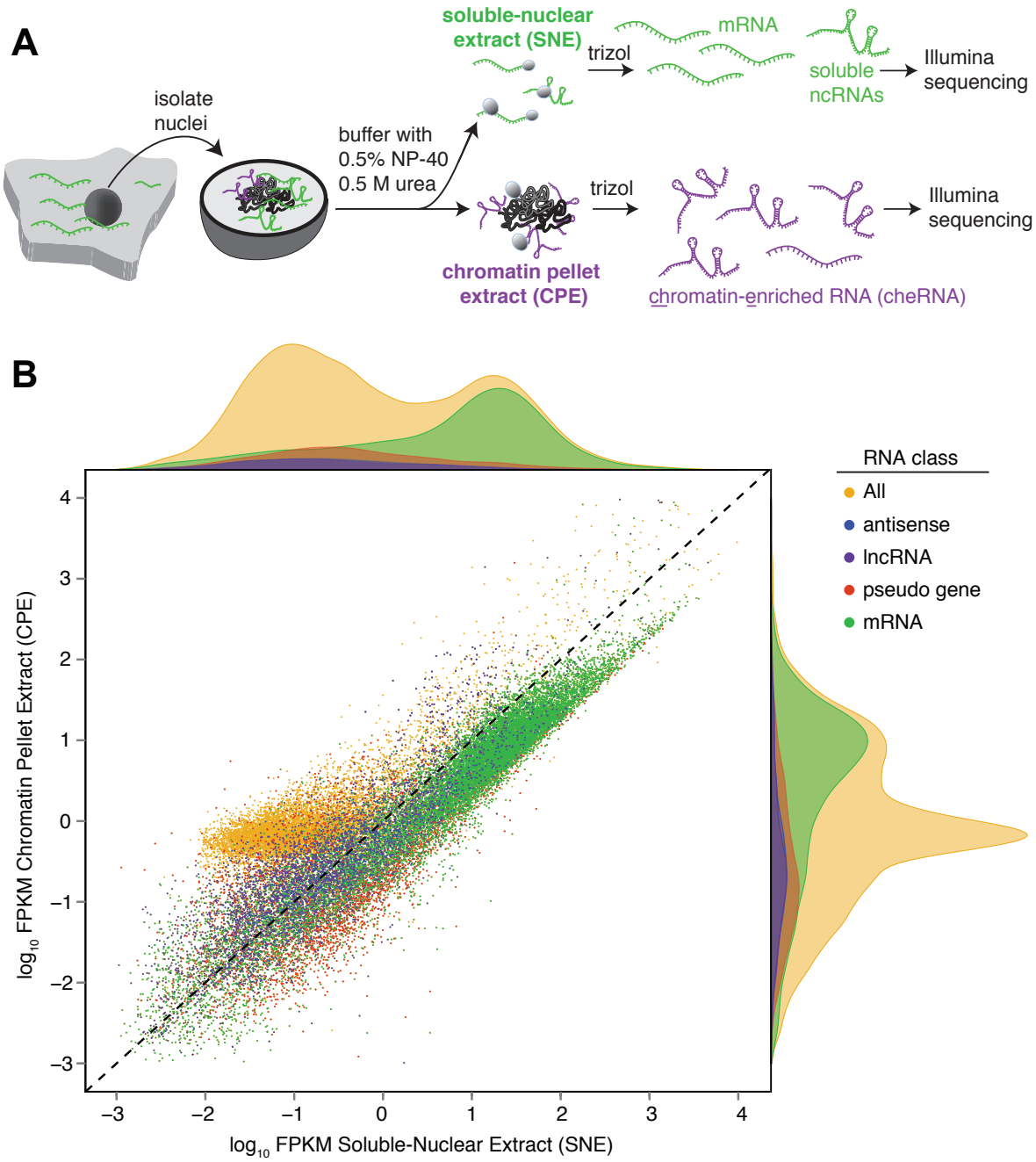


Figure 1.1: Nuclear fractionation isolates chromatin-associated RNA, which includes over half of annotated lncRNAs. (A) Depiction of the nuclear fractionation procedure adapted from Wuarin and Schibler, 1994. Purified nuclei from Human Embryonic Kidney (HEK293) cells were extracted with a forcing urea/detergent buffer to yield a soluble-nuclear extract (SNE) and chromatin pellet extract (CPE). RNA from both fractions was isolated by Trizol in triplicate, ribosome-depleted and sequenced retaining strand information. (B) Scatter plot of relative RNA abundance in each of the two fractions (Y-axis CPE, X-axis SNE) presented on a log10 FPKM-scale with densities for each category plotted as curves along each axis, and a slope of 1 indicated by the dashed line. All assembled transcripts with FPKM>0 in both fractions (32,400) are represented in orange, with Gencode (v19) categories overlaid on top: 14,674 mRNAs (protein-coding) in green, 4,829 pseudo genes in red, 2,409 antisense in blue, and 2,404 lncRNAs in purple.

1.3 Comparison to previous attempts to query chromatin-associated RNA

Mondal and colleagues sequenced RNA associated with chromatin fragments following a brief micrococcal nuclease digestion of nuclei isolated from human skin fibroblast cells (Mondal et al., 2010). However this approach to fractionate nuclei neglected the micrococcal nuclease insoluble pool of RNA that contains the bulk of chromatin modifying complexes through which many lncRNAs are thought to act (Henikoff et al., 2009; Rinn and Chang, 2012), as well as RNAs still in the act of being transcribed (Wuarin and Schibler, 1994; Bhatt et al., 2012; Dye et al., 2006; Kimura et al., 1999). The RNA pool isolated by Mondal (n=209), though from a different cell line, is more consistent with our soluble nuclear enriched pool (35/7334) versus cheRNAs (5/2691 overlap, with 4/2691 expected by chance). Notably the 7SK RNA, that was used to benchmark their method, is several-fold more abundant in our soluble nuclear versus chromatin pellet extract, as would be expected for a ribonucleoprotein complex that is readily soluble in traditional nuclear extracts (Yang et al., 2001). Moreover, this work was limited to 1.5 million reads from a single measurement that obscured strand sense, profoundly limiting the ability to assemble de novo transcripts and the statistical power afforded by biological replicates and enrichment analysis. We use a well-validated method for fractionation that uses Urea and detergent to strip loosely-bound material from chromatin, and affords cleaner bifurcation of the nuclear RNA pools (Bhatt et al., 2012; Dye et al., 2006) and technological advances in high-throughput sequencing allowed us obtain between 49-75 million uniquely mapped strand-specific sequencing reads from three replicates each of chromatin and soluble-nuclear extracts; this deeper replicate sequencing as well as more forceful chromatin extraction was crucial to reliably assemble and annotate chromatin-associated transcripts.

An ENCODE effort to sequence "chromatin RNA" from K562 cells did not include a stringet removal of soluble nuclear factors, thereby preventing adequate separation of solu-

ble from tightly chromatin-bound RNA (Djebali et al., 2012). For example, XIST, the highly abundant, tightly X-chromosome associated RNA (Brockdorff et al., 1992; Engreitz et al., 2013; Jeon and Lee, 2011; Kalantry et al., 2009; Penny et al., 1996), is no more abundant in the "chromatin fraction" than in the nucleoplasmic or total nuclear fractions (Figure A.1I). Whereas we observe a >6-fold enrichment of XIST in the CPE by RNA-seq and RT-qPCR (Figure A.1 A,D). Hence, our use of stringent nuclear fractionation coupled to high- throughput RNA sequencing is the first to unambiguously differentiate between chromatin-bound and nuclear-soluble ncRNA. Although the recent paper that pioneered this fractionation method coupled to deep sequencing in mouse cells noted the effectiveness of this fractionation by showing four known lncRNAs to be largely chromatin enriched (Bhatt et al., 2012), they did not examine the generality of this phenomenon, nor report any new chromatin-associated RNAs. Their single replicate data of approximately 2-fold lower depth lacks the power to robustly assemble chromatin-associated transcriptomes of cheRNA species, and statistically differentiate them from soluble RNA. We conclude that our work represents the first in-depth analysis of the pool of lncRNA.

1.4 Rationale for observation of novel chromatin-enriched lncRNAs

Early lncRNA discovery efforts employed microarray technology (Carninci et al., 2005; Guttman et al., 2009; Kapranov et al., 2002; Khalil et al., 2009; Rinn et al., 2007), which is limited to RNA species complementary to the chosen probes and has limited dynamic range as compared to deep sequencing (Mortazavi et al., 2008). The sub-Moore's Law scaling of next generation sequencing has steadily improved the depth of RNA-seq and coupled with advances in ab initio transcript assembly software, almost every new study using RNA-seq identifies a host of novel lncRNAs (Derrien et al., 2012; Guttman et al., 2009; 2010; Iyer et al., 2015; Necsulea et al., 2014). In our case, removing soluble mRNA from the chromatin

pellet was paramount to the annotation of novel chromatin transcripts because it increased the coverage of these low-abundance RNA species. In support, a recent meta-analysis that achieved massive depth by combining over 7,000 RNA-Seq libraries to annotate 46,000 new human lncRNAs reports 76 pct. of our cheRNAs among their set (Iyer et al., 2015). While the detection of these lncRNAs is valuable, our study adds the nuclear location (chromatin or soluble) of these transcripts, which can be utilized for subsequent mechanistic approaches. We note that the crucial distinguishing features of our cheRNAs - chromatin tethering and apparent cis-enhancer properties - are absent from this analysis. Nevertheless, this data is consistent with our argument that fractionation effectively increases sequencing depth of this pool, permitting ab initio assembly. Further possible explanations for why the cheRNAs have eluded detection apart from the very recent Iyer work include a bias against annotating monoexonic genes (Cabili et al., 2011) and the common practice of library generation from oligo-dT-selected RNA, whereas our cheRNAs are largely unspliced and lack polyadenylation (Figure 1.3).

1.5 Chromatin-enriched RNA (cheRNA)

To further investigate the chromatin-associated RNA pool we focused on intergenic transcripts that were significantly enriched in the CPE ($p<0.05$; geometric mean normalization). We term the resulting set of 2,621 transcripts chromatin-enriched RNAs (cheRNAs), noting even more robust chromatin enrichment than Gencode lncRNAs (Figure 1.2A). The majority (81 pct.) of cheRNAs were absent in RefSeq (Pruitt et al., 2014), Broad lncRNA (Cabili et al., 2011) and Gencode (Harrow et al., 2012) annotations, whereas less than 1 pct. of transcripts enriched in the soluble-nuclear pool were unique (Figure 1.2B). However 76 pct. of cheRNAs overlap transcripts detected by a recently curated amalgamation of over 7,000 RNA-seq libraries (Iyer et al., 2015).

To determine whether cheRNAs represent pervasive transcriptional noise or directed tran-

scription we analyzed their molecular properties and found them to be analogous to lncRNAs in a number of respects. CheRNAs exhibited a strong specific strand bias from their putative transcription start sites (TSSs) (Figure 1.2C, Figure A.2B), and fewer than 14 pct. of cheRNAs are located within 500 bp of coding genes, arguing that the majority do not reflect the byproducts of divergent promoters, cryptic upstream promoters or read-through transcription from upstream genes (which are typically within 500 bp of the major promoter) (Core et al., 2008; Preker et al., 2008; Seila et al., 2008). Additionally, cheRNA TSSs displayed peaks of RNAPII, as well as histone 3 lysine 27 acetylation (H3K27ac), and a bias of histone 3 lysine 4 trimethylation (H3K4me3) over monomethylation (H3K4me1) (ENCODE Project Consortium et al., 2012; Grzybowski et al., 2015) similar to those observed for lncRNAs, (Guttman et al., 2009; Rinn and Chang, 2012)(Figure 1.2D,E). We also validated the TSS at these ChIP signatures for 9/10 cheRNAs by 5'RACE (Figure A.2C, Figure A.3B-C), affording further evidence for their independent transcription.

At the molecular level, cheRNAs display relatively modest conservation compared to coding exons, yet slightly greater mean conservation than both introns and lncRNA exons (Figure 1.2F). CheRNAs also exhibit negligible coding potential (Figure 1.2G) and an analysis of ribosome profiling data from HeLa cells (Guo et al., 2010) fails to find cheRNA sequences associated with ribosomes, arguing that they are largely untranslated (Figure 1.2H). Relative to coding genes, cheRNAs are underspliced. Yet we do detect splice-junctions in approximately 1/4 of all cheRNAs comparable to lncRNAs in our data. To determine whether cheRNAs are 3' polyadenylated, a post-transcriptional processing step that is important for stability, export, and translation of mRNAs (Millevoi and Vagner, 2010), we performed RNA-seq with polyT-primed reverse transcription. As anticipated, reads that mapped to mRNAs were heavily biased towards their 3' ends. However this trend was not observed for either lncRNAs or cheRNAs (Figure 1.3A), in agreement with a previous observation that lncRNAs are over-represented in non-polyadenylated RNA sequencing libraries (Derrien et

al., 2012).

Incomplete co-transcriptional processing, as evinced by occasional splicing and lack of polyadenylation, implicated RNAPII in cheRNA chromatin-attachment. To distinguish whether cheRNAs are maintained on chromatin by ongoing transcription or if the fully processed transcripts are tethered by independent mechanisms as may be the case for XIST and other lncRNAs (Engreitz et al., 2013; Grote et al., 2013; Huarte et al., 2010; Jeon and Lee, 2011; Martianov et al., 2007; Park et al., 2002; Rinn et al., 2007), we re-examined the chromatin pellet after inhibiting RNAPII, which otherwise remains transcriptionally competent under similar preparation conditions (Core et al., 2008; Dye et al., 2006; Kimura et al., 1999). A two hour incubation with DRB, an inhibitor of RNAPII elongation (Yamaguchi et al., 1999), caused a global reduction in the chromatin-pellet abundance of mRNAs and the majority of cheRNAs (Figure 1.3B). The cis-acting lncRNA HOTTIP (Wang et al., 2011b) was also depleted in the presence of DRB suggesting it's chromatin association is also transcription dependent. In contrast, XIST, which is thought to be linked to chromatin via the YY1 transcription factor (Jeon and Lee, 2011), remained at equivalent levels (Figure 1.3C). While loss of cheRNAs in the chromatin pellet after inhibiting RNAPII argues the connection to chromatin is mediated by active transcription, we cannot rule out that the loss of signal is due to rapid degradation of cheRNAs, or a combination of both models. Also, approximately 25 pct. of cheRNAs maintained a +/-DRB ratio greater than one, suggesting that this subpopulation is adhered to chromatin by an RNAPII transcription-insensitive mechanism, or derives from another polymerase.

We also analyzed Global Run On Sequencing (GRO-seq) from the related HEK293-T cell line (Liu et al., 2013), and observed abundant nascent transcription at cheRNA loci to the exclusion of flanking regions (Figure 1.3D). Because the majority of cheRNAs are depleted after a relatively brief period (2 hours) of RNAPII inhibition, the detection of abundant nascent transcription at cheRNA loci argues that the original signal comes from nascently

transcribed RNA. The combination of (1) being un-processed (lack of polyadenylation), (2) being dependent on active transcription, and (3) exhibiting a signal of nascent transription by GRO-seq, suggests a model of polymerases tethering cheRNAs to chromatin. Intriguingly, GRO-seq coverage was disproportionally enriched at the 3' ends of cheRNAs following the peak of RNA-seq coverage, perhaps indicating temporary pausing. Collectively, these characteristics suggest that cheRNAs represent a subclass of regulated, conserved, and especially chromatin-enriched lncRNAs tethered to chromatin via ongoing or paused RNAPII transcription, thousands of which had previously escaped detection from conventional sequencing methods and the majority of annotations (Figure 1.2B, Figure A.2A).

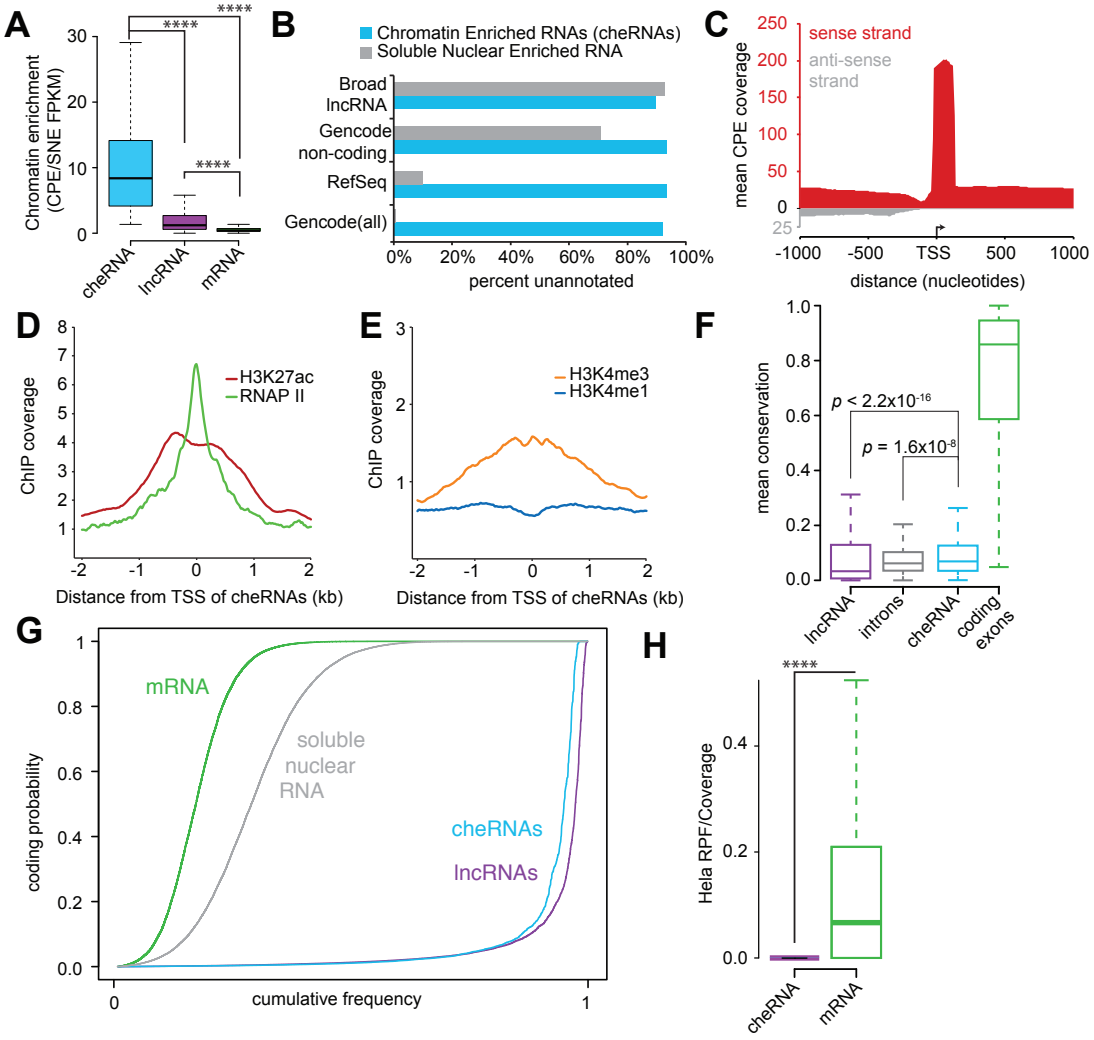


Figure 1.2: Characterization of 2,621 chromatin enriched RNAs (cheRNAs). (A) Chromatin enrichment of cheRNAs, lncRNAs, and mRNAs. (B) Fraction of unique cheRNAs compared to latest Broad lncRNA, RefSeq and Gencode annotations. (C) Average coverage of stranded RNA-seq reads from combined CPE replicates that map to a 2,000 bp window centered on the putative TSSs of cheRNAs. Coverage in the same sense as the cheRNA annotation is depicted in red and antisense reads in grey. (D) Mean ChIP-seq coverage in HEK293 cells of RNA polymerase II (RNAPII) and H3K27 acetyl (H3K27ac), and (E) H3K4me3 and H3K4me1 profiles centered at the TSS of cheRNAs. (F) Boxplot of conservation measured by mean PhastCons score of Gencode lncRNA exons, introns, cheRNAs, and mRNAs (each box spans the 25th to 75th percentile with the median indicated as a line). Comparisons of populations and p values measured by non-parametric one-sided Wilcoxon rank sum/Mann-Whitney U test. (G) Coding probability of cheRNAs (light blue), transcripts significantly enriched in the soluble-nuclear extract (grey, $p<0.05$), Gencode mRNAs (green) and lncRNAs (purple) assessed by CPAT. (H) Ratio of ribosome profiling in Hela cells to total RNA for cheRNA vs. mRNA gene loci (Guo et al., 2010).

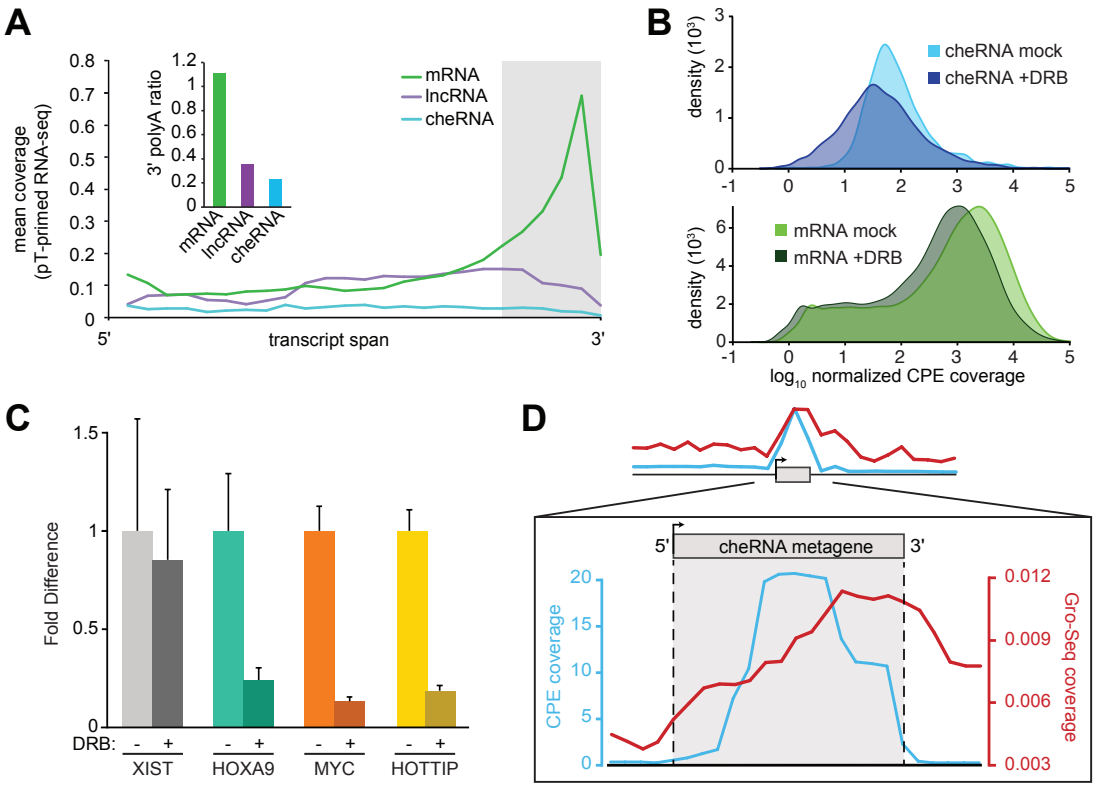


Figure 1.3: CheRNA abundance in the CPE is largely RNA pol II-dependent, and H3K4me3 demarcates their TSSs. (A) RNA-seq using polyT-primed reverse transcription, expressed as mean normalized tag counts across the length of each transcript. Inset, Mean 3'-polyA ratio, computed from (A) as the normalized tag counts mapping to the last 20 pct. divided by the first 80 pct. of each transcript. (B) The distribution of cheRNA (blue) and mRNA (green) CPE RNA-seq reads treated with the RNA polymerase II elongation inhibitor DRB (darker shade, two hour incubation at 100 microM) and 'mock' treated (lighter shade) samples for two biological replicates. All reads in a given fraction were normalized to the amount of reads that mapped to XIST, which is not affected by DRB on a two hour time scale (see C). (C) Fold difference of each indicated RNA with DRB treatment compared to DMSO only (mock) normalized to 18S rRNA. Error bars indicate S.E.M. for three independent biological replicates. (D) Average RNA-seq from the CPE (cyan) and GRO-seq (Liu et al., 2013) (red) contoured over cheRNAs reveals nascent transcription.

1.6 CheRNA transcription correlates with proximal gene expression

Inspired by several examples of lncRNAs altering the expression of neighboring genes in cis (Li et al., 2013; Mohammad et al., 2010; Wang et al., 2011b; rom et al., 2010) we investigated a similar potential function for cheRNAs. Consistent with this possibility, the sites of cheRNA transcription are closer to coding genes than expected by chance (Figure 1.4A,B), and are strongly correlated with the expression of their nearest genes in the soluble-nuclear extract and total RNA (Figure 1.4C, Figure A.3D). Remarkably, proximity to cheRNAs was more highly correlated with the expression levels of nearby genes than the presence of nearby enhancers defined by the chromatin marks H3K4me1/H3K27Ac, expressed lncRNAs, or enhancer RNAs (eRNAs) (Figure 1.4C). Gene expression as a whole decreased with distance from cheRNAs, perhaps indicating local enhancer function affecting multiple genes (Figure 1.4D), although this trend is more idiosyncratic on an individual basis (Figure A.4B).

As a representative example, we highlight cheRNA1345, which is produced from a locus 15 kb downstream of Cep135 and overlaps with an experimentally validated tissue-specific enhancer element (Figure 1.4E) (Visel et al., 2007). Peaks of H3K4me3, H3K27ac, and RNAPII decorate the TSS (confirmed by 5' RACE), which also bears a number of validated TFBS. Similar patterns are observed with other cheRNA loci (Figure A.3A-C).

Approximately two-thirds of HEK293 cheRNAs also overlapped with putative enhancers derived from nine cell lines (Ernst et al., 2011), (empirical $p<0.001$). Despite cell type mismatch, the expression of genes near HEK293 cheRNA-enhancers in each of three ENCODE cell lines queried were greater than tissue-specific 'weak' enhancers, and comparable to 'strong enhancers' (Figure A.3E). Notably, 'strong enhancers' are partially defined by high levels of H3K4me3 (Ernst et al., 2011) hinting that they might generally promote expression of cheRNAs, consistent with a recent unifying model of regulatory elements (Andersson et al., 2015).

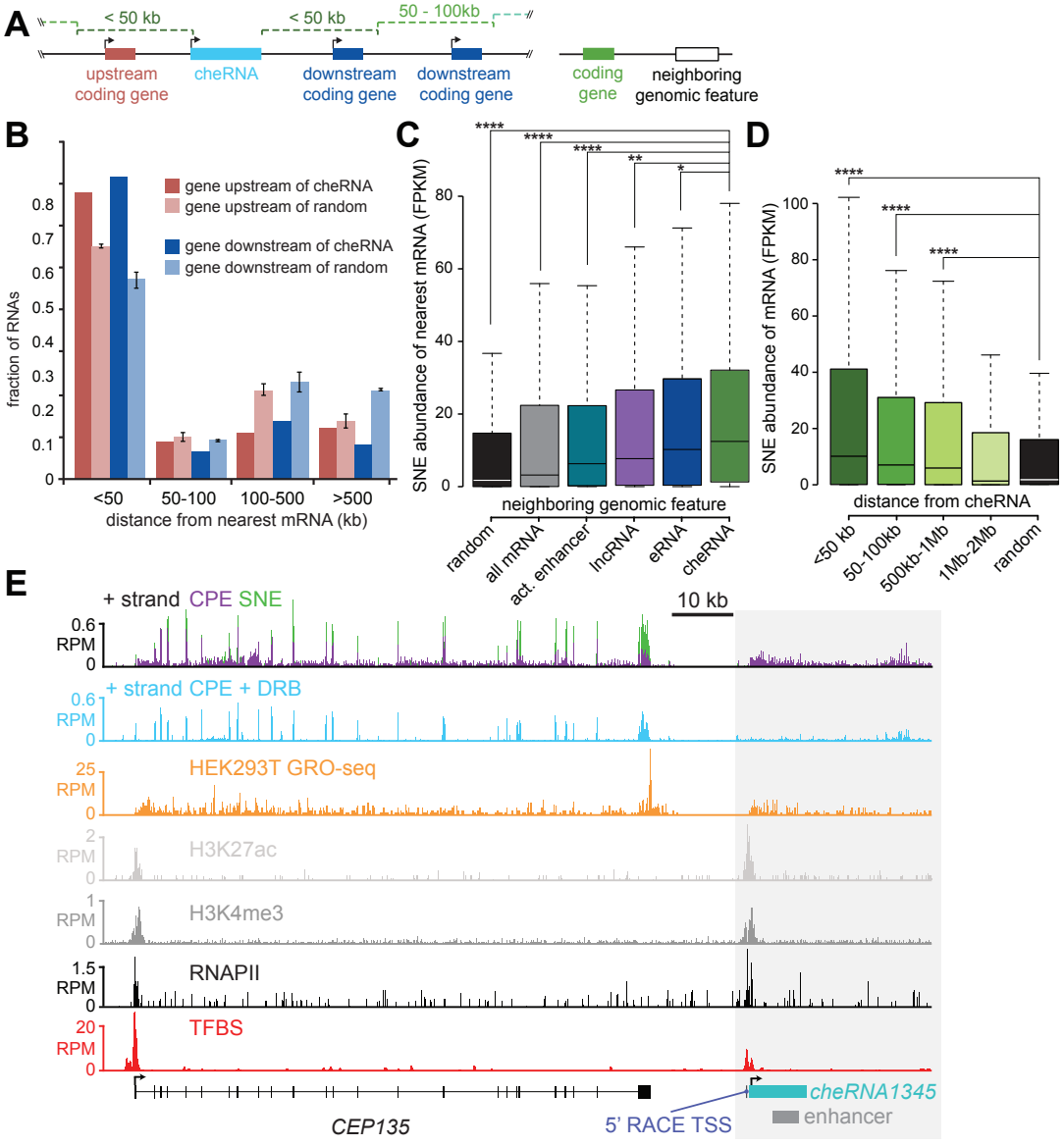


Figure 1.4: Expression of genes positively correlates with proximity to cheRNAs.(A) Schematic representation of a cheRNA locus with upstream and downstream coding genes. (B) Distribution of nearest gene distances to cheRNAs (dark columns) compared to randomly shuffled cheRNA coordinates (light columns, n=3, error bars indicate S.E.M. for three trials). (C) Comparison of SNE expression (FPKM) of the nearest genes to indicated genomic features. *p<0.05, ** p<1x10-10, ****p<2.2x10-16 computed with Mann Whitney U test. (D) Similar to (C), expression of mRNAs that fall within the indicated distances from cheRNA genes. (E) RNA-seq (SNE, green; CPE, purple) and H3K4me3, H3K27ac, RNAPII ChIP-seq contoured over the Cep135 - cheRNA1345 locus in HEK293. Experimentally validated transcription factor binding sites from multiple cell types (ENCODE Project Consortium et al., 2012) and a functional Vista Enhancer are also indicated.

1.7 Segmentation of cheRNAs based on orientation reveals important distinctions

Despite many shared properties and strong correlation to transcriptional status of neighboring genes, the set of cheRNAs is unlikely to be monolithic in function. Genes with overlapping antisense cheRNAs exhibited significantly lower expression as compared to all mRNAs (Figure A.4A), consistent with models of transcriptional interference (Callen et al., 2004). Division of cheRNAs by strand sense and orientation relative to their nearest coding genes reveals an uneven distribution, with slight skew towards shared sense (60 pct.), and a stronger bias towards being downstream (71 pct.) (Figure A.2D). CheRNAs downstream of their neighbors display even stronger expression correlation than the set as a whole, whereas the upstream cheRNAs are more weakly correlated on a whole (Figure A.4A) despite notable counterexamples(Figure A.4B). This composition raises the possibility that some cheRNAs in the same sense represent read-through from upstream genes (Iyer et al., 2015) or, to a lesser extent, cryptic initiation sites for downstream genes (Preker et al., 2008). Pervasive cheRNA biogenesis of this sort might account for the strong cis-expression correlation. Indeed, there are examples of cheRNAs spliced to proximal coding genes detectable in our data and in EST databases potentially accounting for 5 pct. of cheRNAs, comparable to the 9 pct. of lncRNAs previously noted to be cryptic UTRs of proximal coding genes (Derrien et al., 2012). However, removing these cheRNAs from our analysis does not alter the observed correlation with nearby gene expression (Figure A.4A), nor is the correlation exclusive to cheRNAs in the same sense as the nearest mRNA (Figure A.4A). Further, within 1Mb there is little correlation between distance to the nearest neighboring gene and apparent cis-activation (Figure A.4B). Collectively the vast majority of cheRNAs are independently transcribed units, and that the apparent cis-enhancer effect is independent of cheRNAs that could potentially represent read-through, cryptic introns or extended UTRs from coding genes.

1.8 Discussion

We find that the majority of known lncRNAs are chromatin-enriched, extending this property from the small set of well-studied lncRNAs (Chalei et al., 2014; Mohammad et al., 2010; Nagano et al., 2008; Rinn et al., 2007; Wang et al., 2011b) to a more general principle, thereby providing a resource for future mechanistic studies. We also observe that trans-acting lncRNAs exhibit intermediate levels of chromatin enrichment, suggesting either more labile chromatin attachment or reflecting two distinct pools of molecules: those bound to or searching for their target loci (Bond et al., 2009; Rinn et al., 2007). For a lncRNA of unknown function, the distinction between strong or intermediate enrichment may inform their mechanistic possibilities, suggesting whether cis or trans-acting lncRNA pathways are more likely.

A more holistic view of the nuclear transcriptome reveals two highly clustered populations corresponding to nuclear-soluble and chromatin-associated RNA (Figure 1.1B). While there have been hints of the existence and dimensions of this latter population (Bhatt et al., 2012; Derrien et al., 2012; Khalil et al., 2009), our findings establish this pool to be substantially larger than previously observed, consistent with hypotheses advocating a more widespread role of noncoding RNA in chromatin regulation (Bernstein and Allis, 2005; Mattick, 2004). Although many of these transcripts may be non-functional, the intergenic cheRNAs we focused on exhibited molecular properties similar to annotated lncRNAs (Derrien et al., 2012; Guttman et al., 2009; Rinn and Chang, 2012). Perhaps the most compelling case for function is our observation that proximity to a cheRNA locus was more strongly predictive of neighboring gene expression than any other class of noncoding RNA or enhancer annotation available in the HEK293 cell line (Figure 1.4C). Nevertheless, genetic perturbations and functional experiments of individual cheRNAs will be required to test their causality in proximal gene activation and examine the molecular mechanism(s) thereof. Even if these molecules themselves play no direct role in cis-enhancer function, their presence tethered

to chromatin by transcription is sufficiently predictive of cis-gene activity, that determining chromatin-enriched transcripts in other cell types may aid in annotating active cis-enhancer elements in conjunction with existing methods (Andersson et al., 2014; ENCODE Project Consortium et al., 2012; Rada-Iglesias et al., 2011; Zentner et al., 2011).

We also identify that the predominant means of tethering of cheRNAs (including hundreds of annotated lncRNAs) is through active RNAPII. Although widely speculated (Bonasio et al., 2010; Guttman and Rinn, 2012; Quinodoz and Guttman, 2014; Rinn and Chang, 2012), we are not aware of tests of this proposed chromatin association beyond a few anecdotal cases (Mao et al., 2011; Simon et al., 2011). This result is consistent with the cis-activating correlation observed, leading to a model where ongoing or paused transcription of noncoding RNAs influences the expression of proximal genes (Bonasio et al., 2010; Guttman and Rinn, 2012). Conversely, the 25 pct. of cheRNAs that remain on chromatin after pausing Pol2 transcription for two hours are compelling candidates for lncRNAs that may utilize additional methods of attachment (Jeon and Lee, 2011). Alternatively, this population could represent RNA projecting from RNAPIII or a more stably paused RNAPII, which may lose chromatin association on a longer timescale than sampled by our pharmacologic perturbation.

Given the annotated enhancer overlap and expression of nearby genes, we expected significant overlap of cheRNAs with the relatively new category of enhancer RNAs (eRNAs)(Andersson et al., 2014; Kim et al., 2010; Wang et al., 2011a). To our surprise, only 10.9 pct. of cheRNAs overlapped a compendium of eRNAs derived from the majority of human tissues (Andersson et al., 2014), compared to 6.2 pct. expected by chance. Although there appear to be functional similarities, there are several molecular characteristics that distinguish cheRNAs from the canonical definition of eRNAs. First, most eRNAs are bi-directionally transcribed from prototypical enhancers marked by the histone modifications H3K4me1 and H3K27ac (Kim et al., 2010; Wang et al., 2011a), while cheRNAs exhibited a specific strand bias (Figure 1.2C, Figure A.2B) and display H3K4 tri-methylation

over monomethylation, more typical of mRNA and lncRNA TSSs (Figure 1.2D,E). To date, only one report has described a few hundred enhancers that produced largely unidirectional eRNA, although unlike cheRNAs, most were polyadenylated and appeared to have H3K4me1/HeK4me3 ratios greater than unity (Koch et al., 2011). Further, whereas eRNAs are generally considered short (median 350 nt) (Andersson et al., 2014), CPE-specific transcripts which comprise the bulk of cheRNAs were bounded by a >1,000 nucleotide threshold to yield a median length of 2,110 nucleotides. Finally, although 58 pct. of active HEK293 eRNAs displayed chromatin enrichment, this does not appear to be a defining characteristic of eRNAs (Figure A.2E).

CheRNAs are more similar in molecular properties to the transcriptionally activating ncRNA-a subset of lncRNAs (Lai et al., 2013; rom et al., 2010), however, only a few (n=20) of these overlap cheRNAs, whereas more overlap transcripts significantly enriched in the SNE (n=338). Despite these distinctions, their apparent functional similarity is a compelling reason for evaluating potential mechanistic commonalities, and perhaps redefining the eRNA or ncRNA-a categories to include chromatin-enriched transcripts. Going forward, the challenge will be to determine how these and myriad other noncoding RNA species contribute to the complexity of transcriptional control in humans and other multicellular organisms.

Appendix A

DISCOVERY AND CHARACTERIZATION OF CHERNAS

A.1 Methods

Nuclear Fractionation

Nuclear fractionation was performed similar to (Pandya-Jones and Black, 2009; Wuarin and Schibler, 1994); 10-20 x 106 adherent HEK293 cells were grown in DMEM +10 pct. FBS, 1 pct. Penicillin/Streptomycin to 80 pct. confluence, washed in 1xPBS, and then recovered by scraping and centrifugation (500 x g, 5 min, 4C). For transcriptional inhibition experiments, 5,6-dichloro-1-?-D-ribofuranosylbenzimidazole (DRB) (Sigma Aldrich) was dissolved in DMSO to 75 mM, then added to cells for two hours at 100 microM final concentration. An equal volume of only DMSO was added to other cells as a mock control. Cell pellets were resuspended in 2.5 x volumes of Buffer A (10 mM HEPES pH 7.5, 10 mM KCl, 10 pct. glycerol, 340 mM sucrose, 4mM MgCl2, 1 mM DTT,1 x Protease Inhibitor Cocktail [1 mM PMSF, 1mM ABESF, 0.8 microM aprotinin, 20 microM leupeptin, 15 microM pepstatin A, 40 microM bestatin, 15 microM E-64]), and then an equal volume of Buffer A with 0.2 pct. (v/v) Triton X-100 was added and the mixture incubated on ice for 12 minutes to lyse cells, followed by centrifugation (1,200 x g, 5 min, 4C). The crude nuclear pellet was resuspended in 250 microl NRB (20 mM HEPES pH 7.5, 50 pct. Glycerol, 75 mM NaCl, 1 mM DTT, 1 x protease inhibitor cocktail), transferred to a microcentrifuge tube and centrifuged (500 x g, 5 min, 4C) to wash. The pellet was resuspended in 250 microl NRB, and then an equal volume of NUN buffer (20 mM HEPES, 300 mM NaCl, 1M Urea, 1 pct. NP-40 Substitute, 10 mM MgCl2, 1 mM DTT) was added and incubated 5 minutes on ice, then centrifuged (1,200 x g, 5 min, 4C). The soluble nuclear extract supernatant was transferred to another tube, and the depleted nuclear pellet was resuspended in 1 ml Buffer A to wash, transferred to another microcentrifuge tube, and centrifuged (1,200 x g, 5 min, 4C). Resulting purified

chromatin pellets were resuspended in 50 microl Buffer A. TRIzol (0.5 ml) was added to the re-suspended chromatin RNA, and to 20 pct. (v/v) of nuclear-soluble extracts for RNA extraction. The manufacturer?s protocol was followed to obtain an aqueous RNA layer, which was used as input for RNA Clean and Concentrator- 25 columns (Zymo Research). In-tube DNase digestion was performed according to the manufactures protocol, and pure chromatin and nuclear-soluble extract RNA fractions were eluted in 50 microl RNase/DNase-free water.

RT-qPCR

Reverse Transcription was performed with 100-500 ng of RNA input, 200 ng random hexamers, and transcribed with High Performance MMLV Reverse Transcriptase (Epicentre) following manufacturers protocol. After RT, RNA was hydrolyzed by adding 2 volumes of 20 mM Tris-base with 150 mM KOH and incubated 95C for 10 minutes. Sample pH was neutralized by adding equal volume 150 mM HCl, and then cDNA was diluted with TE buffer (10mM Tris-HCl pH 8.0, 1mM EDTA). Quantitative PCR was performed using SYBR Green master mix (Life technologies) on the ABI7900 (Applied Biosystems) with the indicated primers (Supplementary Table 4). For all experiments shown, error bars indicate S.E.M. for three independent biological replicates.

5' RACE

Rapid amplification of the 5'-end of cheRNAs was performed with the SMARTer RACE 5'/3' Kit (Clontech). Reverse transcription was performed with random hexamers, followed by gene-specific primer (GSP) amplification. GSP primer locations were informed by the nearest H3K4me3, H3K27Ac, and pol II peaks to our ab initio cufflinks TSS, which in some cases were a few nucleotides apart, and in other cases several kb apart. For each cheRNA, the most abundant amplicon observed by agarose gel electrophoresis after a second round of nested PCR was excised and cloned using the In-Fusion HD Cloning kit. Cloned 5' ends were then sanger-sequenced at the Functional Genomics Facility at The University of Chicago.

RNA-Seq and Bioinformatic analysis

Input RNA was converted to cDNA libraries using the TruSeq Stranded Total RNA Sample Prep kit (Illumina) with Ribo-zero gold (Epicentre). Standard amplification was performed for all samples according to the TruSeq protocol (98 for 30 sec., and 15 cycles of 98 for 10 sec., 60 for 30 sec., and 72 for 30 sec., followed by 72 for 5 minutes), and sequenced on a HighSeq2000 (lllumina). Reads were aligned to the hg19 genome assembly using Tophat2 (Kim et al., 2013; Trapnell et al., 2012). De novo chromatin fraction transcriptomes were assembled independently using Cufflinks2 (Trapnell et al., 2012), and then the three biological replicate Chromatin Pellet Extract (CPE) assemblies were merged. The three Soluble-Nuclear Extract (SNE) reads were assembled and merged relative to the Gencode (v19) annotation, and then CPE and SNE transcriptomes were added together to make a single annotation for differential expression/abundance analysis using Cuffdiff with standard options (geometric mean normalization). As the chromatin extract retains nascently transcribed RNA, we required de novo-assembled chromatin-transcripts to be unique relative to the Gencode annotation as a stringent filter to exclude transcripts derived from intragenic transcription of annotated genes. Additionally, a distribution of all CPE transcript lengths exhibited a bias of short transcripts (32 pct. <1 kb), (Figure A.1H). While many of these may represent bona fide novel transcripts, we reasoned that longer transcripts are less likely to represent spurious transcription, or incomplete transcript assembly by Cufflinks2, and thus only kept high confidence CPE transcripts >1,000 nts. This pipeline yielded 9,275 CPE-specific transcripts, and a combined 63,433 transcripts for subsequent analysis of relative abundances (Figure A.1G). Ratios of SNE and CPE FPKM enrichment were calculated for each 'gene'. In Figure 1B, all assembled transcripts with FPKM>0 in both fractions (32,400) are represented in orange, with Gencode (v19) categories overlaid on top: 14,674 mRNAs (protein-coding) in green, 4,829 pseudo genes in red, 2,409 antisense in blue, and 2,404 lncRNAs in purple. In Figure A1 there were too many reads that mapped to XIST for Cuffdiff to quantify, so the ratio given is for read coverage that is simply normalized to

total mapped reads in each fraction. For beta-actin the SNE/CPE ratio given is the sum of transcript isoform FPKMs not including an incorrectly called isoform which reflects the nearby non-coding RNA AC006483.1. For obtaining chromatin-enriched RNAs (cheRNAs), we isolated significant ($p<0.05$, Cuffdiff geometric mean) transcripts from the combined transcriptome that were enriched in the CPE. We then filtered out any transcript that overlapped with a coding gene, resulting in 2,621 cheRNAs (Figure A.1G).

For DRB experiments, two replicates of both mock (DMSO only) and DRB-treated CPE RNA were converted to cDNA as described above and sequenced with a HiSeq2000 (Illumina). Resulting reads were aligned to the hg19 genome assembly as described above, and the number of reads that mapped to cheRNAs and mRNAs in both fractions were determined by Homer (Heinz et al., 2010), and normalized by the amount of reads that mapped to XIST in each sample. RT-qPCR was normalized to 18S rRNA, whose transcription, like XIST is DRB independent on the 2-hour time scale of the experiment (Sehgal et al., 1976; Yamaguchi et al., 1999; Zandomeni et al., 1982).

Fraction overlap with Gencode exons was determined by BEDTools (Quinlan and Hall, 2010) 'intersect' and Samtools (Li et al., 2009) 'view-c' of mapped reads with a minimum quality score of 30. Fraction of unique cheRNAs determined by overlap with the indicated reference annotations using stranded (-s) 'intersectBed' with single nucleotide overlap. Strand sense of RNA-seq reads surrounding cheRNA 5' ends was determined using Homer 'annotatepeaks hist' from Forward and Reverse mapped CPE reads (q score greater than 30) to Forward and Reverse-stranded cheRNA TSS-centered windows, combined, and then plotted in excel. Conservation of cheRNAs was measured by averaging PhastCons (Siepel et al., 2005) scores across indicated RNA datasets, and coding potential was determined by the CPAT web server based on a logistic regression model for four sequenced-based linguistic features (Wang et al., 2013).

Distance and nearest gene expression was performed by first obtaining the gene ID of the

nearest mRNA to cheRNAs or relevant comparisons using 'closestBed', then obtaining their FPKMs from the SNE or HEK293 Total RNA (Memczak et al., 2013), or polyA+ RNA in other cell lines (ENCODE Project Consortium et al., 2012). Experiment Accessions: K562-ENCFF000HEY, HSMM-ENCFF000GOW, H1-ENCFF000FEM. We determined a set of putative active enhancers from HEK293 cells by overlapping H3K4me1 and H3K27ac ChIP-seq peaks using MACS2 with standard options (300 bp bandwith, n=1,900), two widely accepted chromatin hallmarks of active enhancers, and compared their nearest gene expression to cheRNAs as described above. Enhancer RNAs in HEK293 (Figure 1.4C, n=316) were obtained by overlapping eRNAs (Andersson et al., 2014) that exhibited HEK293 RNA-seq coverage with this set of histone-modification HEK293 enhancers. As predicted (Lam et al., 2014), this set of eRNAs provide a more accurate estimate of active enhancers compared to histone-modifications alone (Figure 1.4C). Comparison with ENCODE enhancers (Ernst et al., 2011) was performed with Bedtools 'intersect', and an empirical p value was determined by intersecting ENCODE enhancers with random genomic coordinates of cheRNAs (n = 1,000). ChIP-seq in HEK293 cells of H3K4me3 was obtained from our lab (Grzybowski et al., 2015), and ChIP-seq of RNA polymerase II (accession ENCSR000EZA), H3K27ac (accession ENCSR000FCH) and H3K4me1 (accession ENCSR000FCG) in HEK293 were obtained from the ENCODE database (ENCODE Project Consortium et al., 2012). Density plots centered around the TSS of cheRNAs were created with Homer software using standard options (Heinz et al., 2010). All figures with normalized ChIP-seq or RNA-seq reads as a function of chromosomal coordinate were generated using the Integrative Genome Browser (IGV) (Thorvaldsdttir et al., 2013).

poly-dT primed RNAseq

To construct a poly-dT library, total RNA from HEK293 cells was extracted using Trizol following manufacturers protocols. Strand-specific library preparation was performed similar to (Cloonan et al., 2008; Zhao et al., 2010), but with a polyT primer for first strand cDNA

synthesis. In short, a 19 nucleotide oligo-dT with an upstream Illumina adapter (5' CAA GCA GAA GAC GGC ATA CGA GCT CTT CCG ATCT-oligo-dT 3') was used for first strand synthesis. Remaining RNA was hydrolyzed in 150 mM KOH/20 mM Tris for 10 minutes at 90C, leaving only 1st strand cDNA, which was neutralized by addition of 150 mM HCl. Second strand synthesis was performed by random hexamer priming (with a barcode and another Illumina adapter sequence upstream: 5' CTT TCC CTA CAC GAC GCT CTT CCG ATC T NNN- 4bp index-NNN NNN 3'; where N=random nucleotide). Resulting ds cDNA was purified and amplified by PCR using the following primers: 5'- AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATCT -3' 5'- CAA GCA GAA GAC GGC ATA CGA GCT CTT CCG ATCT- 3' The resulting amplified library was run on a 2 pct. agarose gel, to size-select for 200-600bp fragments containing inserts plus adapters. Reads were aligned to the genome with Tophat2, and then analyzed by Homer Software as above. Read counts were normalized to the lowest and highest values for each RNA category so they were all on the same scale.

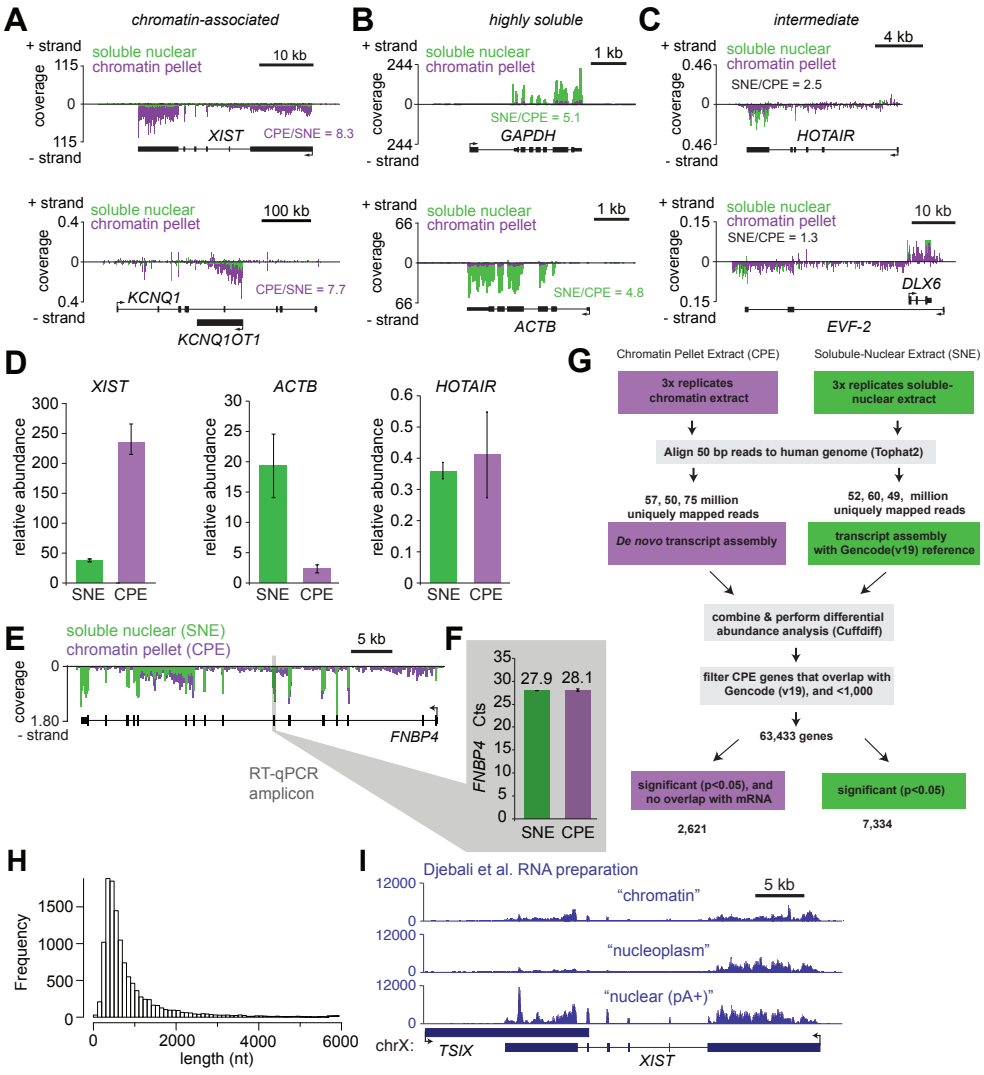


Figure A.1: Nuclear fractionation quantitatively separates chromatin-associated RNA from soluble-nuclear RNA. (A-C) Soluble-Nuclear Extract (SNE, green) and Chromatin Pellet Extract (CPE, purple) from HEK293 cells aligned to the genome and presented as coverage (normalized reads per million). Reads mapped to the forward and reverse strands are depicted above and below the axis, respectively, and ratios are depicted for each RNA. (D) RT-qPCR of indicated RNAs from CPE and SNE. Values reflect fold difference of each RNA relative to FNBP4 in each fraction, which is at equivalent levels of abundance in both SNE and CPE by (E) RNA-seq and (F) RT-qPCR. Error bars for (D, F) indicate S.E.M. for 3 biological replicates. Shaded box (E) indicates exon 7 which was targeted by RT-qPCR after equal loading of RNA from both fractions. (G) Bioinformatic pipeline of RNA-seq data. Three independent biological replicates from each fraction were mapped to the genome using Tophat2, then de-novo assembled CPE transripts (Cufflinks2) were added to the Gencode (v19) transcriptome for differential expression with Cuffdiff using standard options (Kim et al., 2013; Trapnell et al., 2013). CPE-specific transcripts that overlapped protein-coding genes, or <1,000 nucleotides were discarded to focus on high-confidence novel intergenic transcripts. (H) Histogram of CPE transcript length (bp) prior to 1,000 nt filter. (I) ENCODE sub-nuclear RNA pools in K562 cells (Djebali et al., 2012, ENCODE Project Consortium, 2012) displays a relatively even distribution of XIST between indicated samples (compare to Figure 1.1), reflecting a major difference in the enriched RNA population obtained by the more stringent fractionation method we utilized (Wuarin and Schibler, 1994, Bhat at al., 2012).

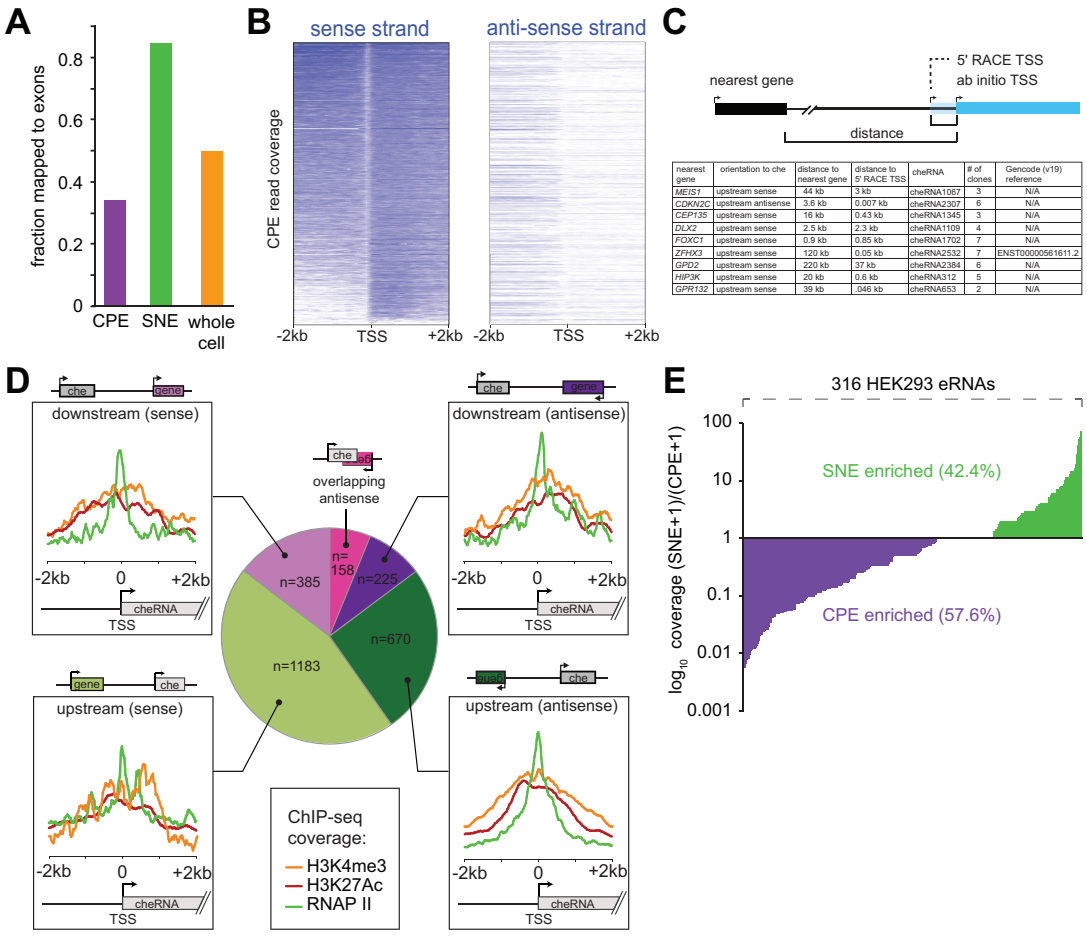


Figure A.2: CheRNAs are untranslated and eRNAs are not defined by chromatin enrichment. s(A) Fraction of RNA-seq reads (minimum quality score = 30) mapped to Gencode exons from CPE (purple), SNE (green), and total RNA (orange). While we observed a marginal increase in the number of reads from the CPE that mapped to intergenic regions (9 pct.CPE vs. 5 pct. SNE), we detected a more substantial difference in the proportion of reads that mapped specifically to exons (30 pct. of CPE vs. 85 pct. of SNE). A deeply-sequenced analysis of whole cell RNA (Memczak et al., 2013) that should capture transcripts from both pools as well as cytoplasmic RNA revealed slightly more than half of total RNA maps to exons. The lower exon coverage in the CPE appears to have contributed to the assembly of novel transcripts by effectively increasing the depth of the 10 pct. of reads that mapped to unannotated regions of the genome. (B) 4kb window of chromatin RNA-seq coverage surrounding cheRNA ab initio transcription start sites on the sense and anti-sense strands. Each row represents a cheRNA, and coverage is reflected as a heatmap (from 0=white to blue=15 FPKM) on a log2 scale. (C) 5'-RACE confirms TSS of 9/10 cheRNAs tested. RACE primers were designed with respect to ChIP-Seq profiles of Pol2 and H3K4me3 in addition to cufflinks denoted 5'-ends. Distance and orientation to nearest genes is indicated as are the number of sequenced clones in support of the 5?end. (D) Breakdown of cheRNAs by orientation and position relative to the nearest coding gene, and accompanying ChIP-seq profiles for each cheRNA subset in HEK293 cells of Pol2, H3K4me3, and H3K27ac. (E) Approximately half of eRNAs detected in our HEK293 data (n=316) exhibit robust chromatin enrichment, in contrast to cheRNAs for which this is a defining characteristic.

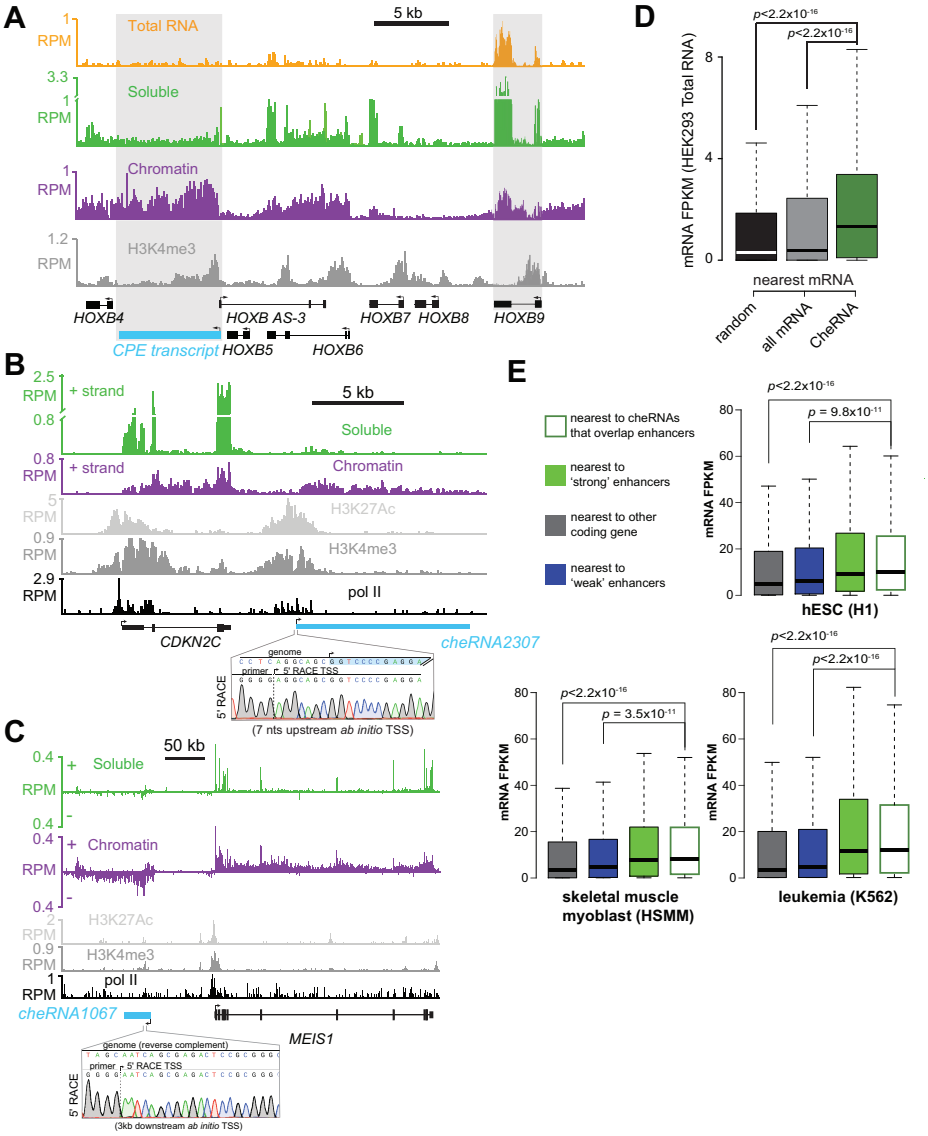


Figure A.3: CheRNA examples, correlation with expression of nearby genes in total RNA, and cheRNA-enhancer overlap is predictive of active enhancers in other cell types. (A-C) Representative genome tracks of ab initio assembled SNE and CPE transcripts, as well as total RNA (A) aligned to their chromosomal sites of origin with overlaid ChIP-seq tracks of H3K4me3 (Grzybowski et al., 2015), H3K27ac, and Pol2 from HEK293 cells (ENCODE Project Consortium et al., 2012). Note the greater apparent CPE depth due to fewer reads mapping to nearby exons compared to the soluble-nuclear extract and Total RNA. (A) Grey boxes highlight the difference between abundant exons and low abundance chromatin transcripts in each fraction and total RNA. In (B) and (C) Sanger sequencing from 5? RACE clones provides validation of predicted cheRNA transcripts start sites. (D) Similar to Figure 4C, nearest gene/mRNA expression to cheRNAs in a total RNA library (Memczak et al., 2013) compared to randomly shuffled genomic coordinates of cheRNAs, and cumulative mRNA in HEK293 cells. (E) Distribution of expression of the nearest mRNAs to ENCODE enhancers that overlap with HEK293 cheRNA loci, compared to 'weak' and 'strong' enhancers in the indicated cell lines determined by a Hidden Markov Model from patterns of histone modifications (Ernst et al., 2011). RPKM/FPKMs of genes in each cell line were obtained from polyA+ RNA (ENCODE Project Consortium, 2012).

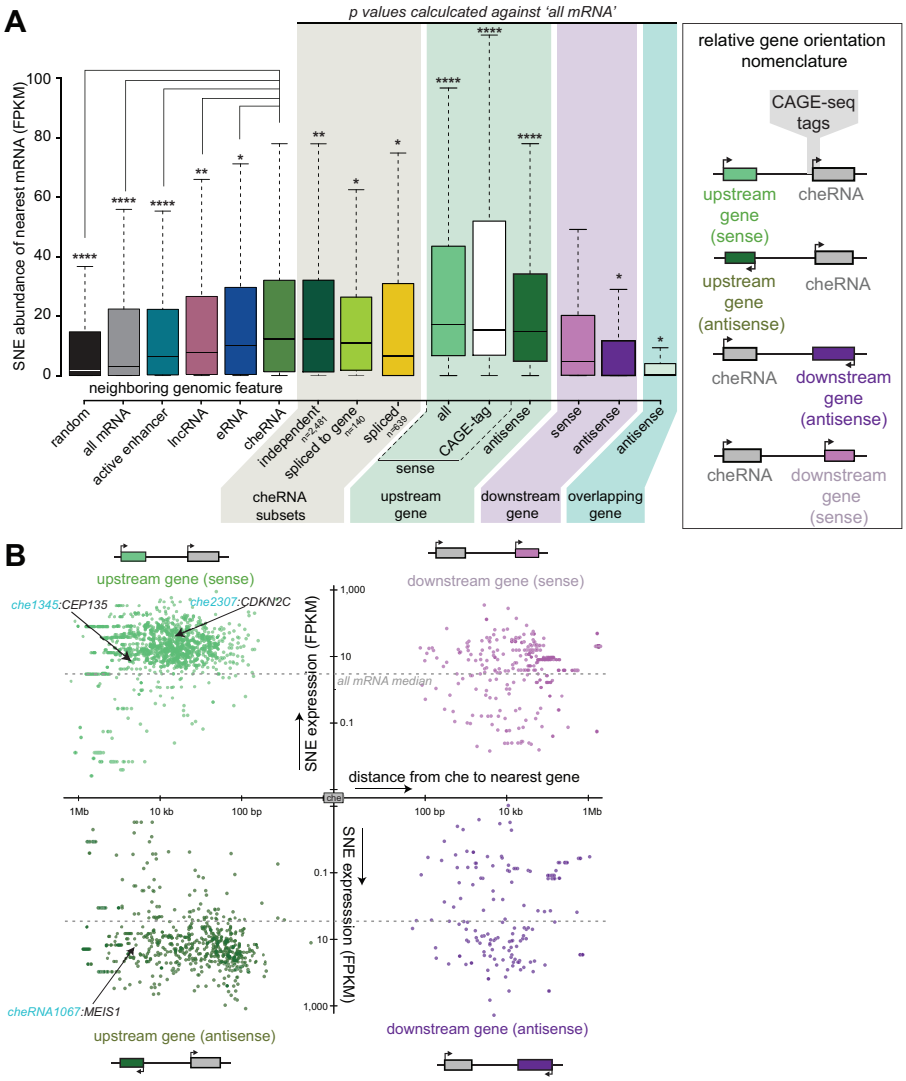


Figure A.4: Expanded analysis of nearby gene expression. (A)Similar to Figure 1.4C, but with added categories of nearest gene expression reflecting strand and orientation to cheRNAs. *p<0.05, ** p<1x10-10, ****p<2.2x10-16 values for extended categories were calculated by Mann-Whitney U test against 'all mRNA' with 'greater' set as the alternative hypothesis, except for 'antisense downstream genes' (dark purple) and 'antisense overlapping genes' (light blue) in which the alternative hypothesis was set to 'less than'. (B) Quadfurcation of cheRNA set into different orientations relative to neighboring genes reveals cis-enhancer-like effect extends out to 1 Mb away from cheRNAs, and that there are multiple highly-expressed genes downstream from cheRNAs which were obscured by the ensemble analysis in presented in (A). Dashed line indicates median expression (FPKM) of HEK293 coding genes in the SNE.

Chapter 2

CHERNAS ARE CELL-TYPE SPECIFIC AND ACTIVATE PROXIMAL CODING GENES

2.1 Summary

The advent of next generation sequencing has uncovered a wealth of transcription outside of canonical protein-coding genes, presenting the challenge of discovering which of these molecules are functional and the mechanisms by which they act. We recently reported a new class of long noncoding RNA (lncRNA) (Werner and Ruthenburg, 2015) defined by especially tight chromatin association via the intermediacy of polymerase II. The presence of chromatin-enriched RNAs (cheRNAs) is strongly predictive of nearby gene expression in HEK293 cells, suggesting a possible cis-enhancer role as has been observed with other lncRNAs (Lai et al., 2013; Wang et al., 2011; Ørom et al., 2010), yet cheRNAs appear molecularly distinct from enhancer RNAs (Lam et al., 2014). Here we address the generality, enhancer function, and role in differentiation of cheRNAs by quantitative chromatin-enrichment of nuclear RNA from human embryonic stem cells, K562 human myeloid leukemia cells, and K562 cells differentiated along an erythroid lineage. We find that cheRNAs are almost exclusively cell-type specific (less than 8 pct. shared between any two cell lines) and despite low copy number per cell, proximity to a cheRNA remains the best predictor of cis-gene expression in multiple cell types. Depletion of four candidate cheRNAs by CRISPRi or antisense oligonucleotides causes a 20-60 pct. decrease in neighboring gene expression proportional to the degree of cheRNA knockdown, establishing the cheRNA molecules as cis-enhancers. Further, erythroid differentiation by the small molecule hemin of K562 cells demonstrates 16 pct. of up-regulated genes are in proximity to a cheRNA. A hemin-induced cheRNA downstream of fetal-hemoglobin (HIDALGO) is activated prior to fetal-globin (HBG1), and is required for full induction of HBG1 upon differentiation. These results attest to the gen-

erality of cheRNA-enhancers in differentiation and lineage-specific gene regulation across multiple human cell lines.

2.2 cheRNAs are general features of the human transcriptome

To begin to address these questions and discover chromatin-eniched lncRNAs in multiple cell lines we performed biochemical fractionation of nuclei from H1 hESC and K562 cells coupled to RNA-seq (Werner and Ruthenburg, 2015; Bhatt et al., 2012; Wuarin and Schibler, 1994; Dye et al., 2006). We observed 3,293 cheRNAs in K562 cells, and 1,136 cheRNAs in H1 hESCs (Figure 2.1A,B), demonstrating the generality of cheRNAs in the human transcriptome, and providing a resource for future exploration of lncRNA mechanisms. We doped *in vitro* transcribed standards into the isolated RNA pools to establish a calibration curve for our RNA-seq data to permit facile experimental comparisons and roughly quantify cheRNA copy number per cell (Figure B.1A-C). This analysis suggests the lncRNA XIST is present at 123 (+/- 38) copies per K562 cell, in good agreement with previous estimates of murine Xist (50-200 per cell) measured by RT-qPCR and STORM microscopy (Sunwoo et al., 2015). Turning to cheRNAs, we find an average of 0.67 and 4.5 molecules per cell in H1 and K562, respectively (Figure 2.1C-D). Assuming some losses during fractionation this places most cheRNAs on the order of magnitude of 1-10 copies per cell. This scarcity explains why cheRNAs have largely evaded annotation without targeted sequencing of the chromatin-associated RNA pool and is consistent with their mechanism of attachment through RNA polymerase II (Werner and Ruthenburg, 2015). This cellular abundance measure is also similar to estimates made by single molecule FISH for lncRNA in a variety of cell types (Cabili et al., 2015; Wang et al., 2011; Hacisuleyman et al., 2014; Kretz et al., 2013).

A comparison between cheRNA species isolated from HEK293, K562, and H1 hESCs reveals striking cell-type specificity of cheRNA expression (less than 8 pct. of cheRNAs are shared between any two cell lines), perhaps indicating a role in maintaining tissue-specific

transcriptional networks (Figure 2.1C). Analogous to cheRNAs in HEK293 cells (Werner and Ruthenburg, 2015), the expression of the nearest coding genes to cheRNAs in K562 and H1 hESCs is significantly higher than genes near enhancers based on chromatin signatures (Ernst et al., 2011; Zentner et al., 2011; Rada-Iglesias et al., 2011; ENCODE Project Consortium, 2012) or eRNA loci (Andersson et al., 2014) with detectable transcription in these cell lines (Figure 2.2A). Importantly, cheRNA-proximal coding genes are also tissue specific: a gene ontology (GO) analysis demonstrates enrichment in factors important to the maintenance of their respective cell types, such as the ERK1 and ERK2 cascade in H1 hESCs (Li et al., 2007; Armstrong, 2006; Na et al., 2010) and JAK-STAT signaling in K562 (de Groot et al., 1999) (Figure B.2A). This correlation is based largely, although not exclusively, on cheRNAs downstream of their nearest mRNA, and we note there is a greater correlation for cheRNAs in the same sense as the upstream gene in H1 and K562 cells as compared to HEK293 (Figure B.3A).

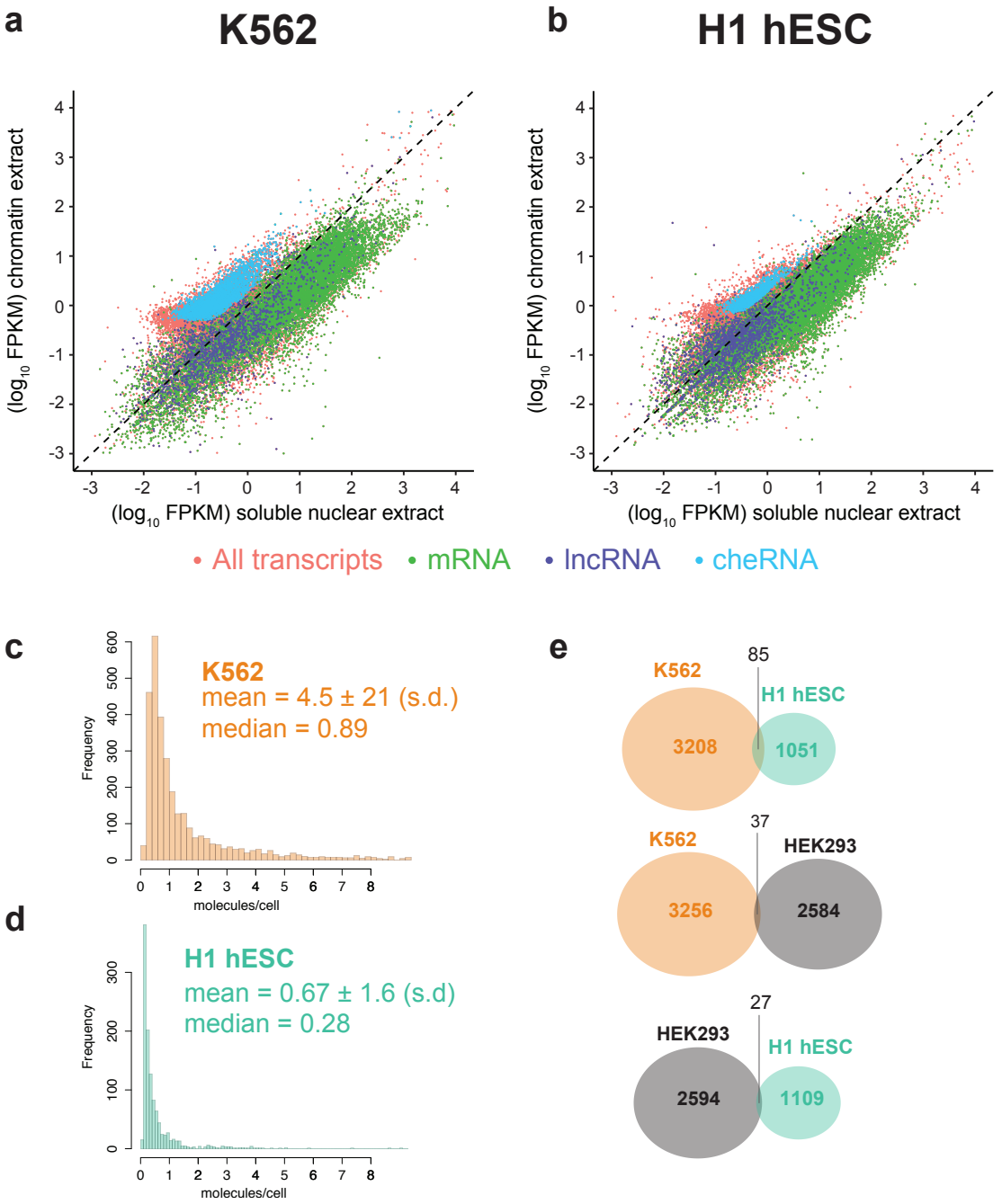


Figure 2.1: Chromatin-enriched transcripts (cheRNA) in K562 and H1 hESCs attest to the generality of this class of RNAs while displaying pronounced tissue-specificity. (A) A scatterplot of nuclear de novo assembled transcripts after forcing nuclear fractionation coupled to RNA-seq (Werner and Ruthenburg, 2015) depicts chromatin enrichment versus soluble nuclear-enrichment for K562 and H1 hESCs cells (Gencode annotation of mRNA and lncRNA, with new cheRNA species indicated in cyan, and all remaining transcripts in orange,). (B), Histogram of cheRNA molecules per cell contoured in the range of 0-8, determined by a calibration with spiked-in in vitro transcribed standards into each RNA pool prior to cDNA library preparation. (C), Overlap of cheRNAs from H1, K562 and the prior HEK293 (Werner and Ruthenburg, 2015) datasets demonstrate they are largely unique to each cell line.

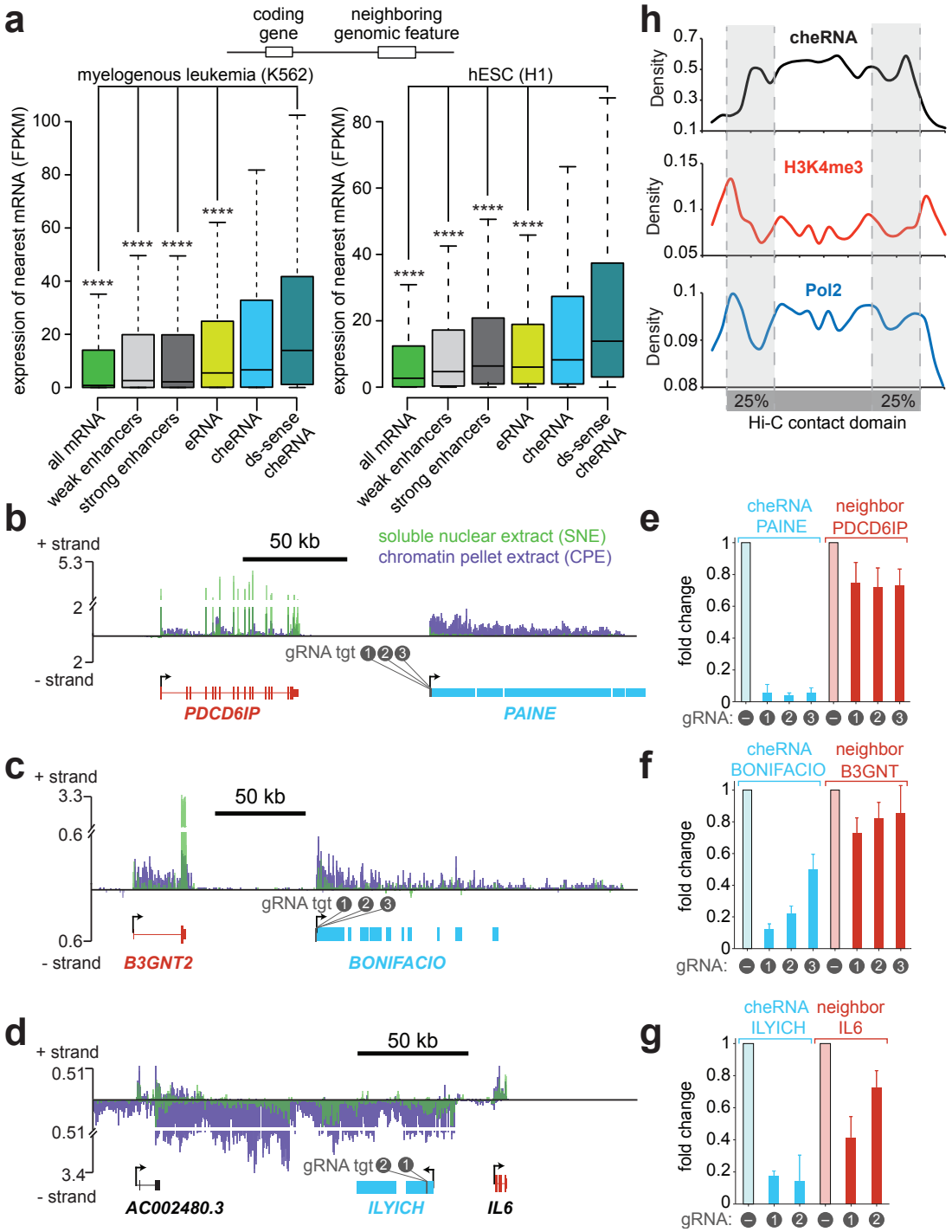


Figure 2.2: Active cheRNA loci are strong indicators of nearby gene expression in multiple cell lines and several examined display enhancer function. (A) Comparison of expression of nearest genes to cheRNAs to indicated genomic features. (B-D) Distribution of K562 chromatin pellet (purple) and soluble-nuclear extract (green) RNA overlaid on the indicated chromosomal region. (E-G) RT-qPCR for the cheRNA and its neighboring gene corresponding to panels (B-D), after CRISPRi knockdowns targeting the indicated cheRNA. Fold difference qPCR values represent delta-delta Ct relative to a non-targeting gRNA (Gilbert et al., 2014). Error bars equal propagated standard deviation from 3-4 technical replicates. (H) Average density of depth-normalized cheRNA and Pol2 and H3K4me3 ChIP-seq over a 'meta' Hi-C domain containing cheRNAs in K562 cells (Rao et al., 2014).

2.3 cheRNA loci are functional enhancers

Due to the high correlation of gene expression proximal to cheRNA loci, both of which are tissue specific, we sought to determine whether cheRNAs play a causal role in promoting local gene expression. We used CRISPRi (Gilbert et al., 2013) to inhibit transcription of two cheRNAs that are 67kb and 71kb downstream of their nearest genes, and one 19kb upstream of it's nearest gene. Several guide RNAs (gRNAs) targeting the promoter of each cheRNA reduced cheRNA levels by 50-96 pct., which led to proportional decreases in mRNA from their proximal gene (Figure 2.2). This effect was specific, as each gRNA displayed similar perturbations, and no consistent effect on the housekeeping gene GAPDH was observed (Figure B.3). What defines or delimits proximal effects of cheRNAs? Intriguingly, each of these cheRNA-gene pairs fell on the edge of a chromosome contact domain measured by Hi-C (Rao et al., 2014) (Figure B.4). More generally, approximately half (47 pct.) of K562 cheRNAs overlap a contact domain with local peaks at the contact boundaries (Figure 2.2H), perhaps playing a role in defining topologically associated domains (TADs). Although these domains can extend up to 3 megabases, the expression of the 1,766 genes within them matches that of our previous comparison (median=6.6 for both), (Figure B.3A) suggesting that some cheRNAs may affect multiple genes within transcriptionally active TADs.

To determine if cheRNAs play an active role in differentiation we performed nuclear fractionation of K562 cells induced towards an erythroid lineage using the small molecule hemin (Addya, 2004). We observed 3,407 cheRNAs in the erythroid state, and unlike comparisons between different cell lines, most (62 pct.) cheRNAs were shared with un-induced K562 cells (Figure B.5A). Notably, of 172 protein-coding genes upregulated (p less than 0.05) 48 hours after hemin treatment, 27 (16 pct.) were flanked by a cheRNA within 100kb. To better understand cheRNA-enhancer mechanisms we chose to examine one of these cheRNA-gene pairs in more detail. One of the hallmarks of definitive erythroid differentiation is the upregulation of fetal hemoglobin (HbF, HBG1/2) (Ginder, 2015), for which hemin-induction

of K562 cells has been a potent model system (Addya et al., 2004). Intriguingly, we observed abundant chromatin-enriched transcription extending 3.7kb beyond HBG1 in both un-induced and induced states, while no transcription is observed at this locus in H1 cells (Figure 2.3A and Figure B.5B). The putative cheRNA TSS exhibits chromatin features (ENCODE Project Consortium, 2012) of a potential enhancer or promoter, with peaks of transcription factor bindings sites (TFBS), H3K4me3, H3K27ac, RNAPII, and DNase hypersensitivity among others (Figure 2.3B). This downstream cheRNA, hereafter called HIDALGO "Hemin-induced cheRNA downstream of fetal hemoglobin", is induced early in erythroid differentiation, peaking between 2-4 hours after the addition of hemin (Figure 2.3C). In contrast, HBG1 induction occurs more gradually, and peaks following the initial pulse of HIDALGO (Figure 2.3C). This temporal relationship preceding the expression of a nearby gene suggests independent regulation, and has been observed with eRNAs and 'TF-promoters' and their associated genes (Arner et al., 2015).

To examine the biogenesis of HIDALGO we performed 5'/3' RACE and discovered a complex array of transcripts emanating both from the TSS of HBG1, and a location downstream that maps to a cryptic TATA box (Figure 2.3A), (TATAAG) near the initially predicted HIDALGO TSS (Figure 2.3A,B). While one of the RACE transcripts that originates from the HBG1 TSS represents in-frame read-through of HBG1 mRNA that escapes polyadenylation, two others are isoforms that are riddled with stop-codons seemingly due to alternative splicing (Figure B.6A). By virtue of incomplete processing and chromatin tethering, all of these transcripts are de facto cheRNAs. To assess the proportion of HBG1 TSS transcripts that escape polyadenylation we performed 3' RACE on HBG1, which revealed that greater than 85 pct. of transcripts are processed at the normal polyadenylation site (PAS) to become mature mRNA (Figure B.6B). We also designed a panel of qPCR primers that through deduction could differentiate between relative proportions of the observed RACE products, and determined that spliced and unspliced read-through from HBG1 (RACE products 1-2) makes up

the majority (90-95 pct.) of HIDALGO transcripts at basal levels. Hemin-treatment induces all four RACE transcripts, although the transcript emanating from the cryptic TATA box is preferentially up-regulated and comprises approx. 16 pct. of HIDALGO RNA at two-hours post-induction (Figure 2.4B, Figure B.6C). To determine if HIDALGO has a potential regulatory role we used CRISPRi with a panel of gRNAs to inhibit read-through transcription from HBG1 and initiation from the downstream cryptic TATA box (Figure 2.4B). Four different gRNAs surrounding the cryptic TATA box led to a decrease in HBG1 commensurate with the level of cheRNA knockdown (Figure 2.4C). We performed three independent biological replicates with gRNA-4 that exhibited the most robust knockdown of HIDALGO, leading to a reproducible 60 pct. reduction in HBG1 levels (Figure 2.4D). Although termination is not a predominant means of regulating transcription (Jonkers and Lis, 2015), we wondered if the observed effect could be due to the CRISPRi-KRAB fusion protein inhibiting 3'-end processing of HBG1. GAPDH is a 'housekeeping' gene not likely to be under this type of regulation, and indeed does not exhibit a downstream cheRNA. We designed two gRNAs the same distance away from the 3'-end of GAPDH as gRNA-4 is from HBG1. Neither gRNA led to a reduction in GAPDH under the same experimental conditions used to target HIDALGO (Figure B.6D). Additionally, 3'-RACE of HBG1 demonstrated that the majority of transcripts are processed immediately following the PAS, and any possible effect on the fewer than 15 pct. of transcripts that escape 3' processing could not explain the 60 pct. decrease observed.

To determine if the enhancer mechanism is mediated by RNAPII transcribing through the TFBS to loosen-up chromatin or perhaps recruit TFs, or directly involves the cheRNA transcript we used LNA-anti-sense oligos which specifically degrade RNA through RNaseH-mediated cleavage. Three different LNA oligos targeted against HIDALGO (Figure B.6E) caused a 28-42 pct. reduction in HBG1 (35 9.0 pct. mean effect, n = 3 independent replicates for each LNA), demonstrating that the RNA molecule itself has a functional role

(Figure 2.4E). To assess if HIDALGO has a role in differentiation, we made two polyclonal K562 cell-lines containing different gRNAs targeting HIDALGO, and performed a hemin-time course. In both cases inhibiting HIDALGO severely blunted hemin-induction of HBG1 (Figure 2.4F-G). The effect was most pronounced during the initial stages of induction (0-8 hours) that normally coincides with the 'burst phase' of cheRNA production (Figure 2.3C, Figure 2.4G), demonstrating that HIDALGO is important in the early activation of HBG1 during erythroid differentiation of K562 cells.

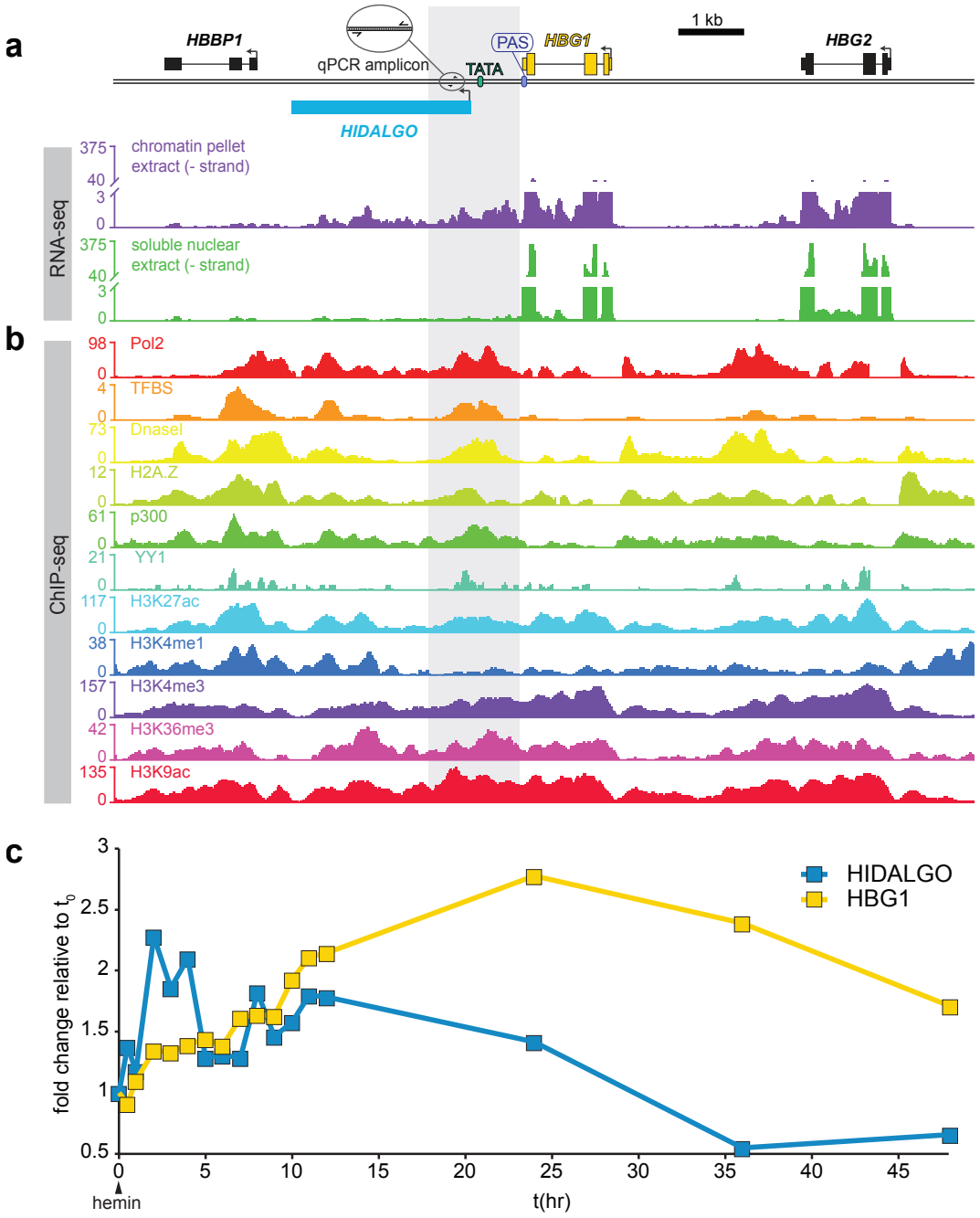


Figure 2.3: (A) Hemin-induced cheRNA downstream of fetal hemoglobin (HIDALGO) exhibits molecular hallmarks of an enhancer and is induced prior to fetal-hemoglobin (HBG1) in erythrogenesis. RNA-seq of K562 chromatin (purple) and soluble-nuclear extract (green) overlaid at the HBG1 locus and flanking regions. (B) Chromatin signatures from ChIP-seq indicate a regulatory region downstream of HBG1 where a HIDALGO, a novel cheRNA, is transcribed. (C) A timecourse experiment measuring the levels of HBG1 mRNA and HIDALGO cheRNA by RT-qPCR after the addition of 50 micromolar hemin to induce erythroid differentiation (y-axis represents fold-change by RT-qPCR relative to time-point zero and 18S RNA).

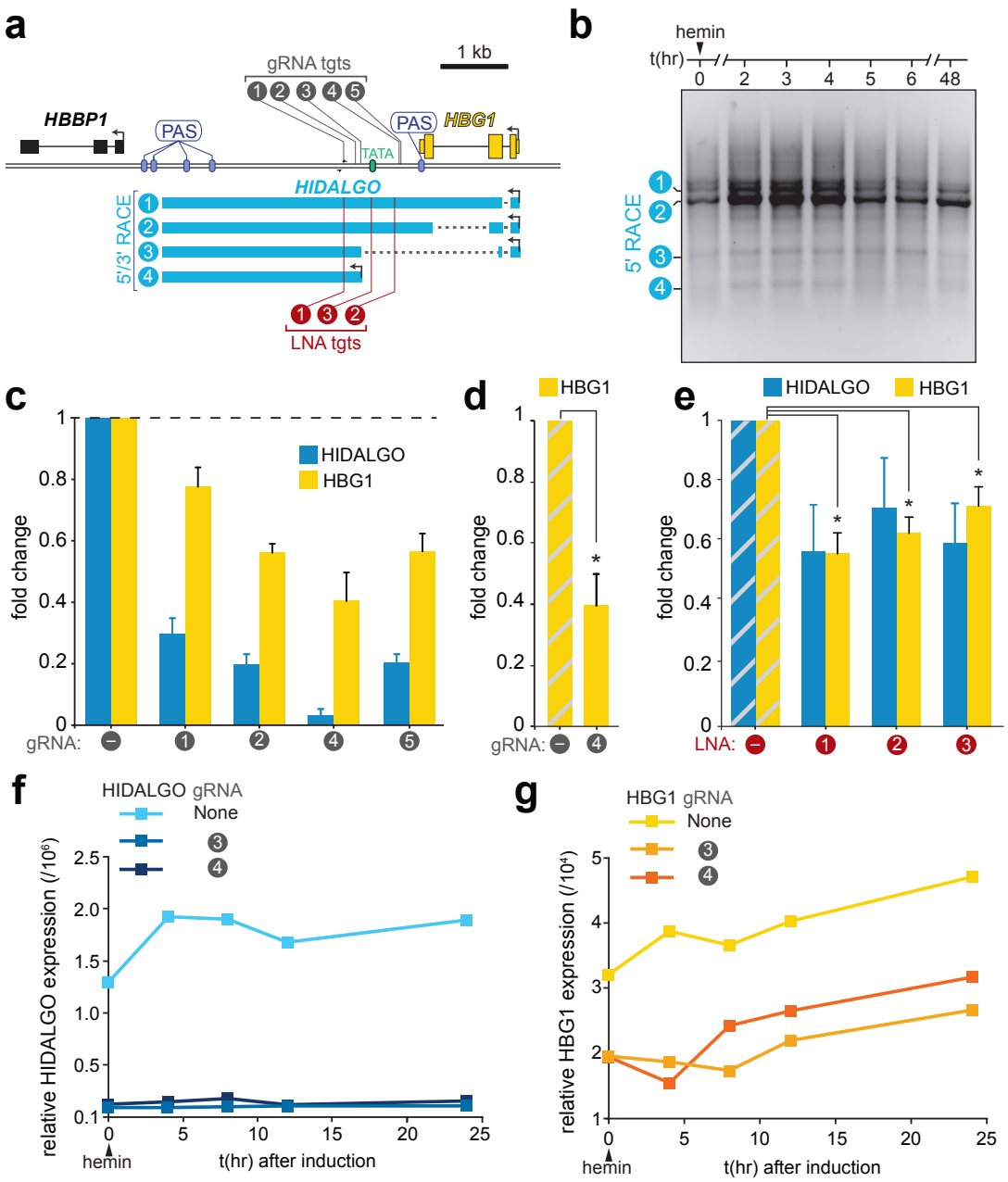


Figure 2.4: (A) Diagram of HIDALGO 5'/3' RACE products and location of CRISPRi gRNAs target sites used in (C-D), (F-G), as well as LNA AS-oligo hybridization sites of different HIDALGO transcripts. (B) 5' RACE of HIDALGO during differentiation demonstrates all transcripts undergo induction in response to hemin, but RACE product 4 at the cryptic TATA box (TATAAG) is preferentially induced. (C) Knockdown of HIDALGO by CRISPRi with three distinct gRNAs decreases HBG1 proportionally. Fold change measured by RT-qPCR relative to a non-targeting negative control gRNA ("-" gRNA) normalized to 18S RNA. Error bars represent propagated standard deviation from 3-4 technical replicates. (D) CRISPRi with gRNA-4 in 3 independent biological replicates. Error bars represent S.E.M. (E) Knockdown of HIDALGO RNA with 3 different antisense oligonucleotides (ASO) causes a decrease in HBG1 expression. Error bars represent S.EM. (n=3 ind. biological replicates). (F-G) An RT-qPCR time-course of HIDALGO and HBG1, respectively after erythrogenesis induced by hemin in dCas9-Krab K562 cell lines with or without gRNA-3 or 4 stably integrated into the genome. Y-axis represents average expression relative to 18S rRNA.

2.4 Transposable-element origin hypothesis

Finally, we were curious about the evolutionary origins of cheRNAs and examined the HIDALGO/HBG1 locus as a test case. We identified a conserved eutherian transposable element (TE) insertion near the TSS of HIDALGO, but a lack of conservation in a region just downstream (Figure B.7). The presence/absence of this latter region corresponds to the insertion of three primate-specific TEs (L1PA11, MER41A, and L1P3) which appear to have occurred in simians (old/new world monkeys and apes) but not prosimians between 35-55 MYA (Jurka et al., 2005) (Figure B.7). Curiously, fetal switching of HbF occurs in the simian but not prosimian lineage. We hypothesize that the insertion of TEs introduced or promoted the enhancer activity of HIDALGO, and speculate that this event contributed to hemaglobin switching in simian primates. Beyond HIDALGO, we note little correspondence of cheRNA loci with other genomic features such as enhancers annotated by chromatin signitures or eRNAs (approx. 10 pct.), yet between 92-98 pct. of cheRNAs overlap with class I TEs. We doubt that all of these insertions are functionally relevant, however there is a growing body of evidence that lncRNAs are often derived from TEs, and that TEs are capable of providing functional elements of gene regulation (Kapusta et al., 2013; Kelley and Rinn, 2012; Chuong et al., 2016; Lynch et al., 2011, 2015).

2.5 Discussion

CheRNAs connected to their upstream gene such as HIDALGO (estimated 12 pct. of H1 and 10 pct. of K562 cheRNAs that are less than 10kb downstream and in the same sense of their nearest gene) share strong chromatin-enrichment, independent regulation, and genetic architecture with recently described transcripts downstream of highly expressed genes (DoGs) in a neuroblastoma cell line (Vilborg et al., 2015). Notably, DoGs were identified in part by KCl induction, yet their corresponding mRNA did not undergo a commensurate increase in five

out of six studied pairs. However in this experiment the cells were collected after one hour of KCl treatment, at which point in our hemin differentiation we observe HIDALGO induction but HBG1 is not substantially induced until several hours later (Figure 2.3C). CheRNAs are not a uniform set, and future experiments should also investigate cheRNAs proximal to unexpressed genes to explore repressive functions which have been described for several lncRNAs (Pandey et al., 2008; Rinn et al., 2007; Martianov et al., 2007; Penny et al., 1996). Nevertheless, our results support a model where cheRNAs can promote cell-type identity by acting as enhancers of upstream tissue-specific genes. Polymerases that finish transcribing upstream genes but remain associated to chromatin could then initiate further rounds of downstream cheRNA transcription, serving as a feed-forward loop for stable expression as has been observed in yeast (O'Sullivan et al., 2004).

Appendix B

CHERNAS ARE TISSUE-SPECIFIC AND ACTIVATE PROXIMAL CODING GENES

B.1 Methods

Cell culture and Nuclear Fractionation H1 hESCs were grown feeder free on Matrigel (BD Bioscience)/StemPro (Invitrogen) media, and K562 cells were grown in DMEM (Gibco) containing 10 pct. FBS, 1 pct. penicillin/streptomycin. 'lus hemin' cells were treated with freshly prepared 50 microM hemin either at indicated time-points, or 48 hours before collecting for RNA-seq. Cell lysis, nuclear fractionation, and RNA isolation were performed as previously described on three biological replicates of 10 million K562 cells and H1 hESCs (Werner and Ruthenburg, 2015). Briefly, purified nuclei were extracted with a forcing buffer containing 0.5M Urea and 0.5 pct. NP-40 substitute to solubilize loosely bound factors from chromatin, and then fractionated by centrifugation. RNA from both the chromatin pellet (CPE) and soluble nuclear extract (SNE) were obtained by Trizol extraction and further purified by RNA-Clean and Concentrator columns (Zymo) with in-tube DNase digestion. In vitro transcribed RNA standards were added to purified chromatin pellet and soluble nuclear extract RNA isolates, then ribosomal RNA was depleted using Ribo-Zero Gold (Illumina), and stranded cDNA libraries were made using NEBNext Ultra DNA Library Prep Kit for Illumina, and sequenced on an Illumina Hiseq2000. K562 and H1 hESC libraries were sequenced single-end 100bp reads, while two replicates of hemin treated K562 cell libraries were sequenced with single-end 50 bp reads.

Calibrated RNA-seq Spike-in standards were in vitro transcribed with T7 polymerase, and were selected based on lack of homology to human genes and approximately similar lengths witgin the set (777-1290 nucleotides). RNA was purified with Zymo RNA-Clean and Concentrator columns, and serially diluted in a buffer containing 50 mM NaCl, 0.01 pct. NP-

40 substitute, 100 ng/microliter pUC19, 10 mM Tris-HCl ph 7.5, and 1 mM EDTA before adding to CPE and SNE RNA prior to rRNA depletion with Ribo-Zero Gold (Illumina). A simple linear regression was obtained based on the number of molecules of RNA standard added per cell number equivalent to each library (calculated from the number of cells that each extract was derived from) versus the absolute read counts from RNA-seq. The resulting equation was then used to compute the approximate molecules per cell for cheRNAs based on their absolute read counts.

5'/3' RACE RACE was performed using SMARTER 5'/3' RACE kit (Clontech) on total RNA following manufacturers protocols.

Reverse Transcription and RT-qPCR Reverse transcription was performed in 20 microl with random hexamers and MMLV-HP (epicentre) for 1 hr, 37C. RNA was then degraded by 40 40 µml 150 mM KOH, 20 mM Tris base for 10 minutes at 90 C, pH adjusted with 40 microl 150 mM HCl and 100 microl 10mM Tris, 1mM EDTA pH 8.0 buffer. Quantitative PCR was performed with 4 microl of the resulting cDNA, using SYBR Green master mix (Life Technologies) on a Bio-Rad CFX96. Relative abundance was measured with 18S rRNA as an internal reference. Fold-difference comparisons were made relative to non-targeting negative control gRNA or LNAs and the 18S internal reference. cheRNA knockdowns

knockdowns CRISPRi was performed in K562 cells with dCas9-KRAB integrated into the genome generously provided by Luke Gilbert and Jonathan Weissman (Gilbert et al., 2014). Guide RNAs (gRNAs) were cloned into a modified px300 vector containing eGFP and a modified stem loop (Chen et al., 2013)), and 4 micrograms were transfected in 6-well plates using Lipofectamine 2000 (Life Technologies) following manufacturers protocols. After two days cells were re-plated on 10 cm plates, and 4-6 days post-transfection, GFP+ cells were isolated by FACS BD Aria II/III, pelleted, and re-suspended in 500 microl Trizol. The aqueous layer from Trizol extraction was used as input for purification with RNA Clean and Concentrator columns (Zymo), and then converted into cDNA as described above. K562

polyclonal cell-lines with gRNAs incorporated into the genome were transfected as above but in a vector containing puromycin resistance. Lines with stable integrations were selected by the addition of puromycin (6.7 microg/ml) two days after transfection and passaging for two weeks under continual selection. LNA-knockdowns were performed with 50 nM LNA-FAM oligos (Exiqon) transfected with Lipofetamine 2000 for 4-6 days, and then FACS-sorted. All comparisons were made relative to a negative control gRNA (-) (Gilbert et al., 2014) or LNA-FAM (Exiqon) that was transfected in parallel but contained non-target sequences.

Bioinformatics Sequencing reads were mapped to the human genome (hg38) using tophat2 with standard options. Triplicate chromatin pellet extract transcripts were annotated using de novo transcript assembly (Trapnell et al., 2012), and unique transcripts greater than 1,000 base pairs were added to the gencode (v23) annotation. Differential abundance estimates between soluble and chromatin fractions were made by Cuffdiff using standard options and analyzed using CummeRbund. Comparison of chromatin-enrichment (chromatin/soluble nuclear extract abundance) for different classes of RNA (Fig.1 f,g) were made with the average of three replicates from each fraction using depth-normalized RNA-seq by HOMER. Density of cheRNA reads and RNAPII (ENCFF000YXN) and H3K4me3 (ENCFF000BYI) ChIP-seq data over Hi-C contact domains (Rao et al., 2014) normalized by total depth were computed using "annotate peaks" in HOMER using standard options (Heinz et al., 2010). Individual cheRNA locations relative to chromosome contacts (Extended data 3a-c) were analyzed using Juicebox (Rao et al., 2014). Comparisons of nearest gene expression were performed using 'closestBed' from bedtools (Quinlan and Hall, 2010) of gencode (v23) protein-coding genes relative to genomic features (i.e. cheRNAs, enhancers). Duplicate gene IDs were removed and soluble nuclear extract FPKMs were used as a metric of expression. All boxplots were created in R, and p values were calculated using a Wilcoxon (Mann-Whitney U) un-paired rank-sum test. ChIP-seq data from ENCODE (ENCODE Project Consortium, 2012) for figure 3 was visualized on Integrated Genomics Viewer (IGV) (Thorvaldsdóttir

et al., 2013; Robinson et al., 2011).

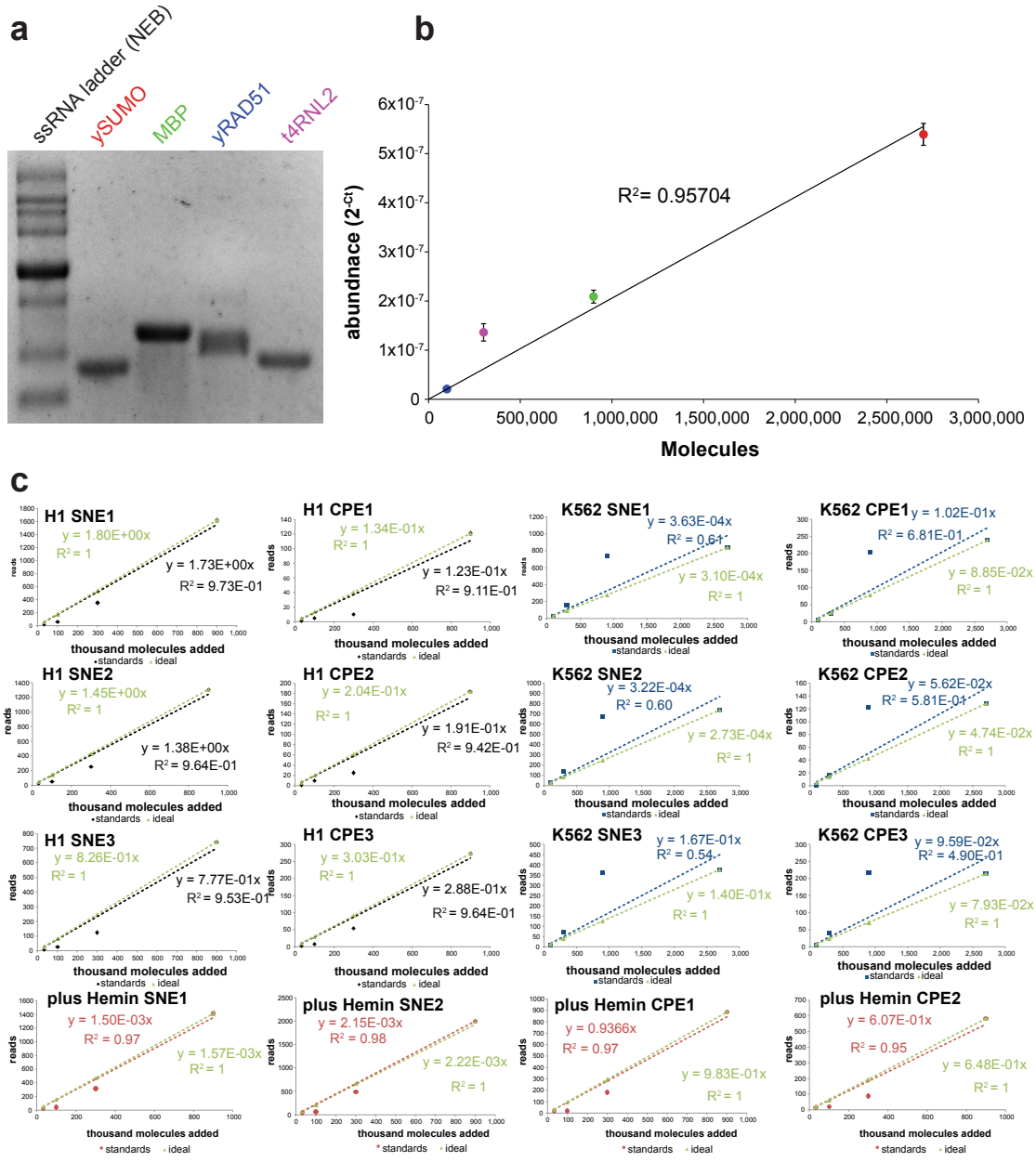


Figure B.1: In vitro transcribed RNA standards establish a calibration curve for quantitation of nuclear RNAs. a, Agarose gel of spike-in standards demonstrates purity of transcripts. b, Scatter plot of spike-in standards demonstrates a linear relationship with molecules added and abundance measured by RT-qPCR. c, Scatter plot of RNA-seq reads (y-axis) versus molecules spiked-in per cell (x-axis) recapitulates a linear relationship in all libraries. As a comparison the highest read value in each library was back calculated by the fold-dilution to a theoretical read value (green lines).

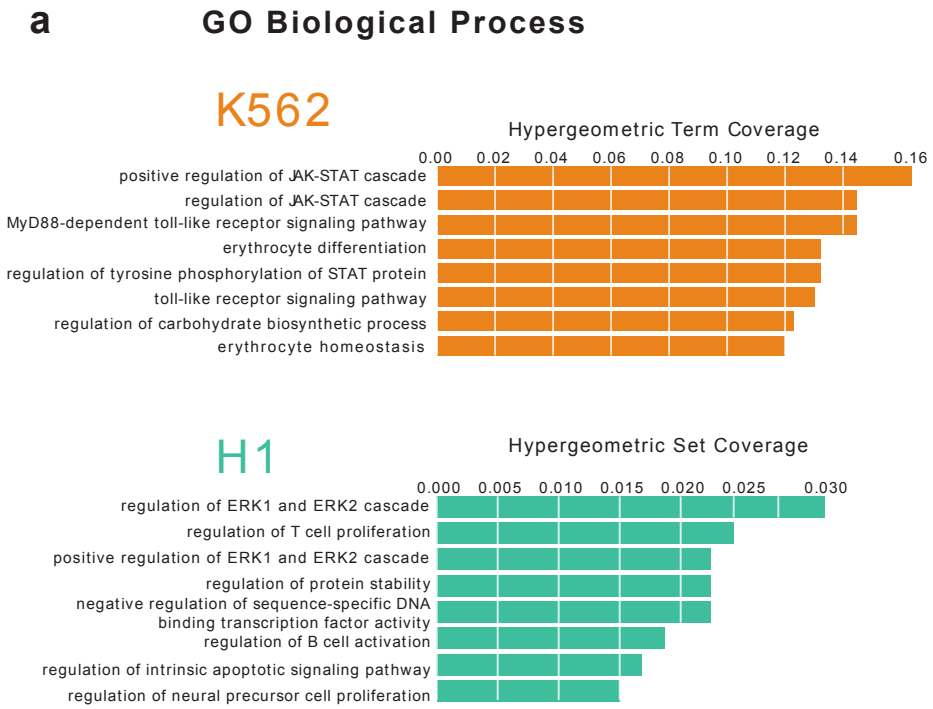


Figure B.2: Gene ontology was performed with GREAT (McLean et al., 2010) for H1 and K562 cheRNAs. The top eight categories of hypergeometric enrichment are shown for each cell line measured with raw p values and annotations containing 50-150 genes.

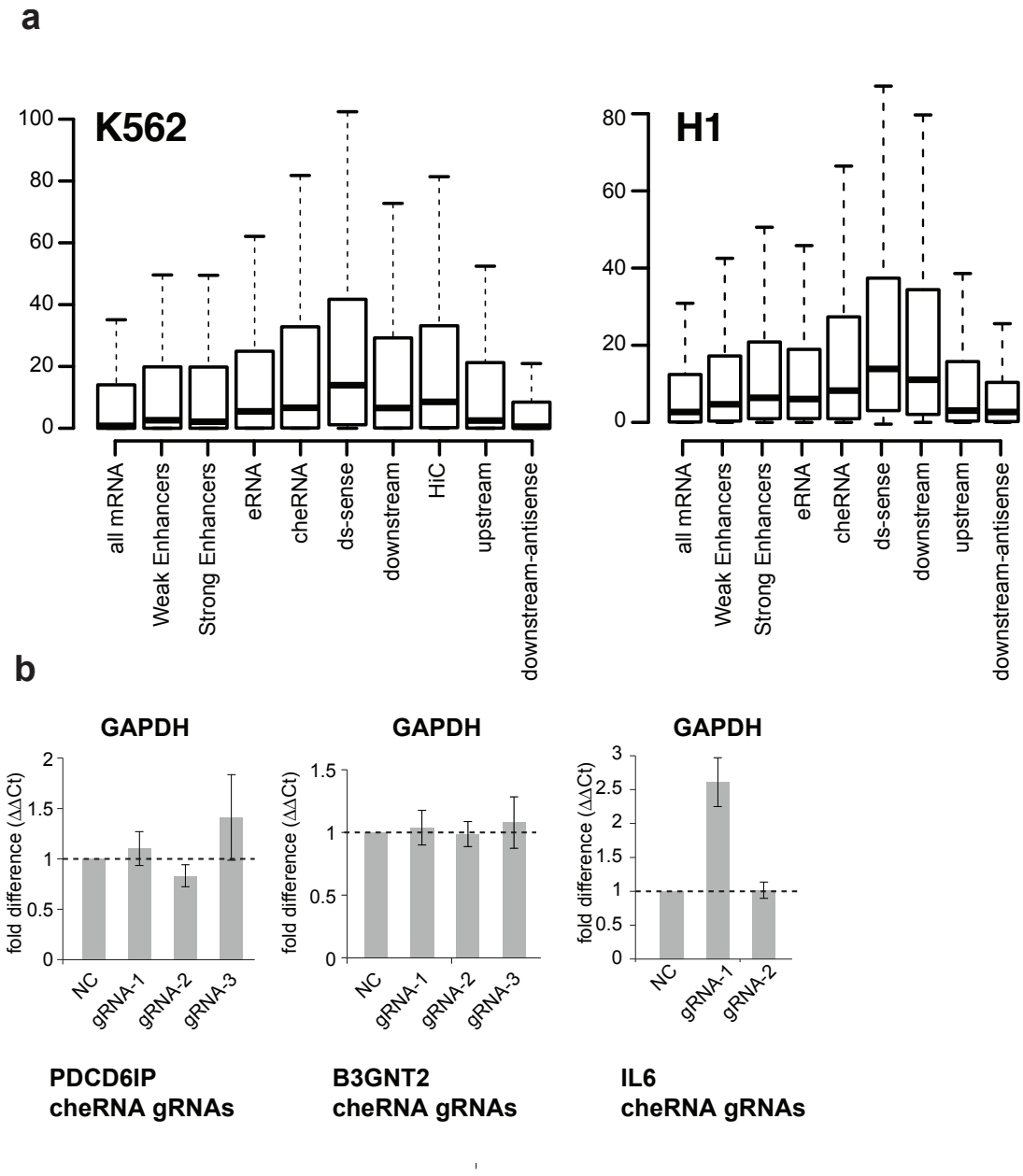


Figure B.3: Expression of proximal genes to cheRNAs by orientation and within K562 Hi-C contact domains, and demonstration of specificity in CRISPRi experiments. (A) Same as in Fig. 2.1A but further broken down by strand and orientation of cheRNAs to nearby genes. Also included are genes within Hi-C (Rao et al., 2014) contacts that overlap with cheRNAs in K562 cells. (B) RT-qPCR of GAPDH from CRISPRi experiments in Figure 2.1(E-G) demonstrates un-changed or inconsistent changes in GAPDH, in contrast to nearby genes.

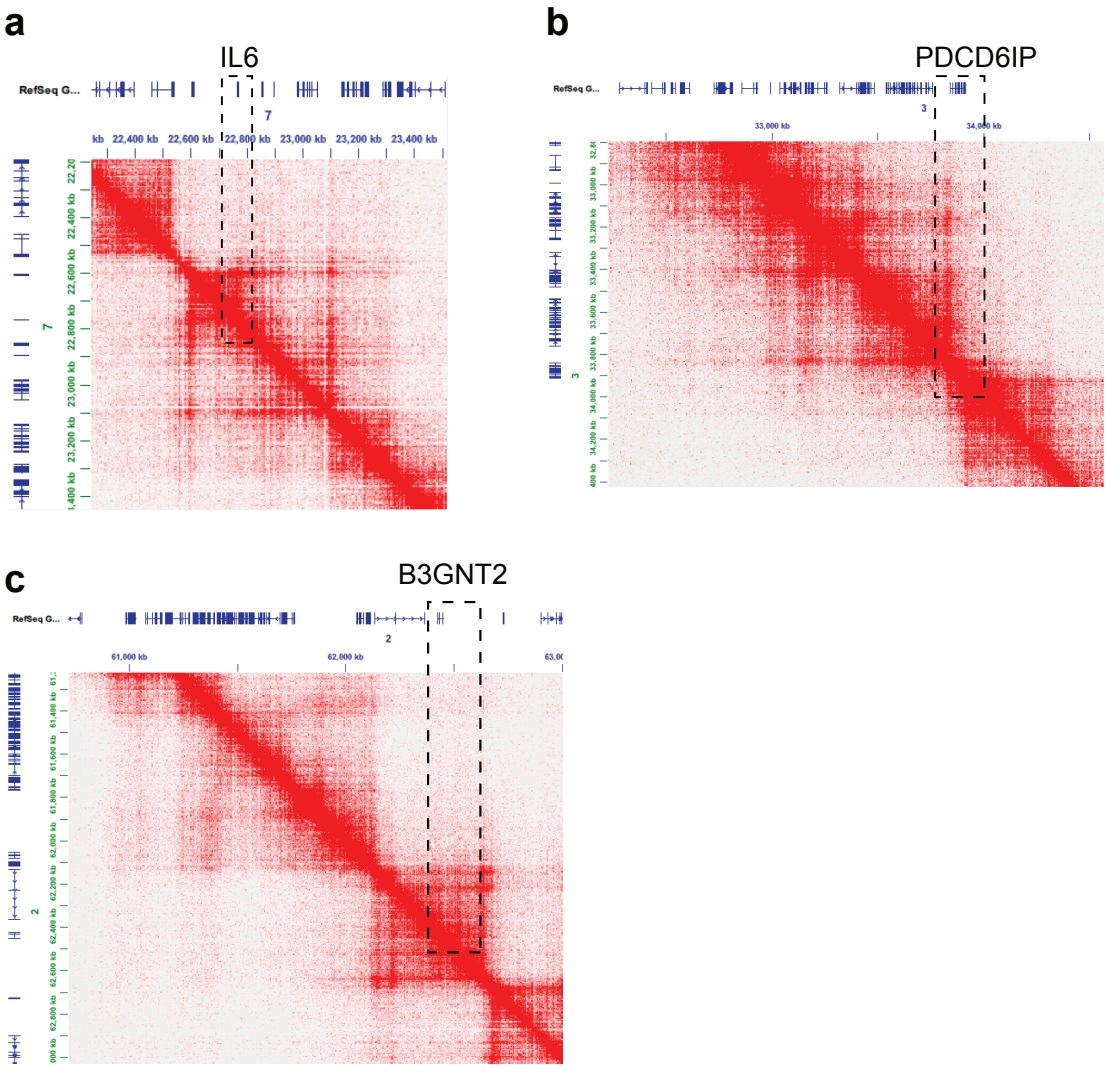


Figure B.4: Hi-C data (Rao et al., 2014) in K562 cells demonstrates cheRNA-gene pairs lie on the boundaries of topological domains. (A-C) Three-dimensional chromosome contacts measured by Hi-C in K562 cells indicated by red color-scale. X and y axis correspond to the same chromosome, and red enrichment extending out from the 1-to-1 axis represents a contact domain. CheRNA-gene pairs are indicated with dotted line boxes.

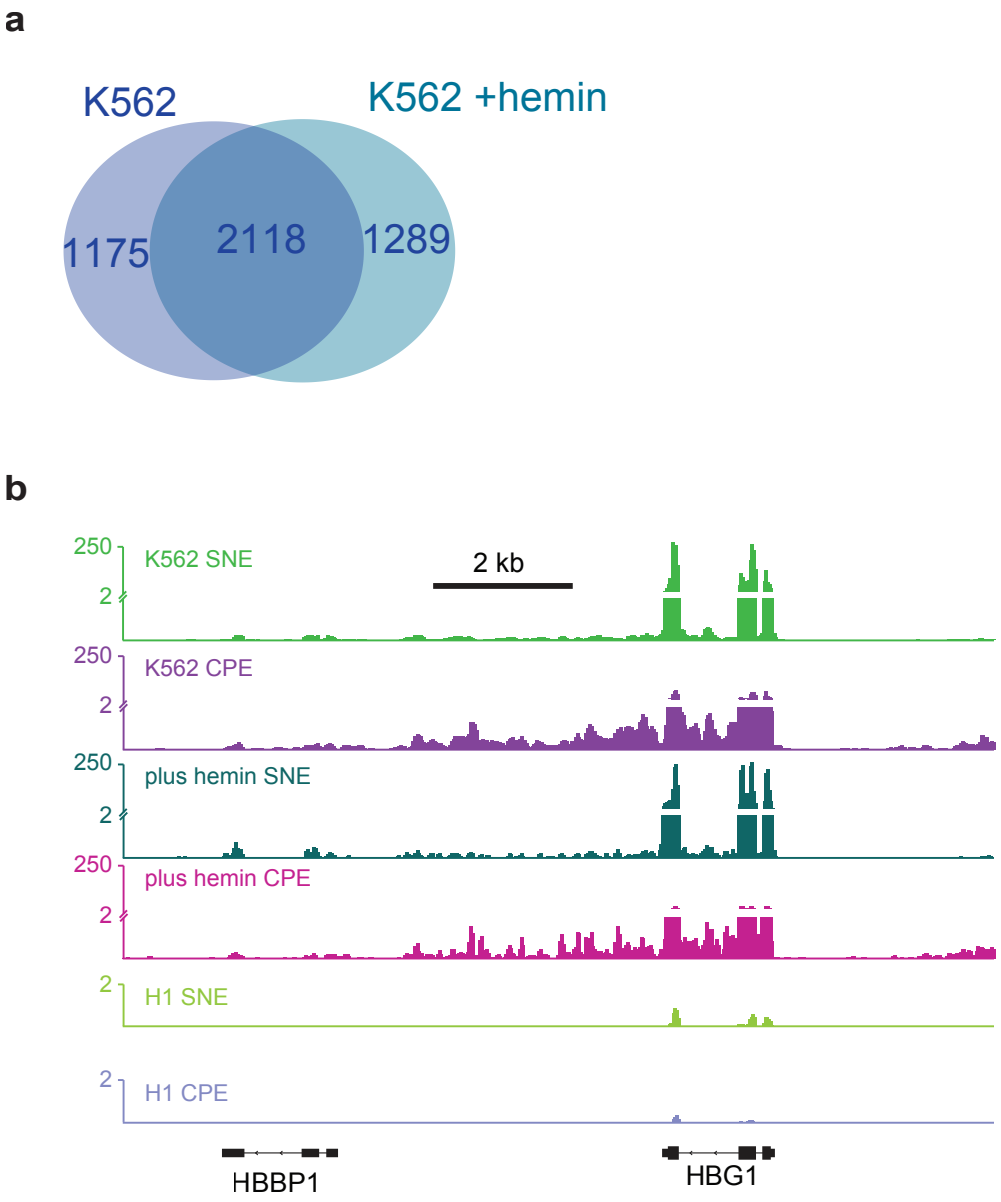


Figure B.5: Specificity of cheRNAs through differentiation and in different cell lines. (A) Overlap of cheRNAs in K562 cells in un-induced and hemin-induced states demonstrate a significant overlap, in contrast to between cell lines in general (Fig. 2.1E), and at a specific locus (B), RNA-seq of CPE (purple shades) and SNE (green shades) from K562 cells in +/- hemin conditions, and H1 hESC overlaid at the HBG1 gene - HIDALGO cheRNA locus.

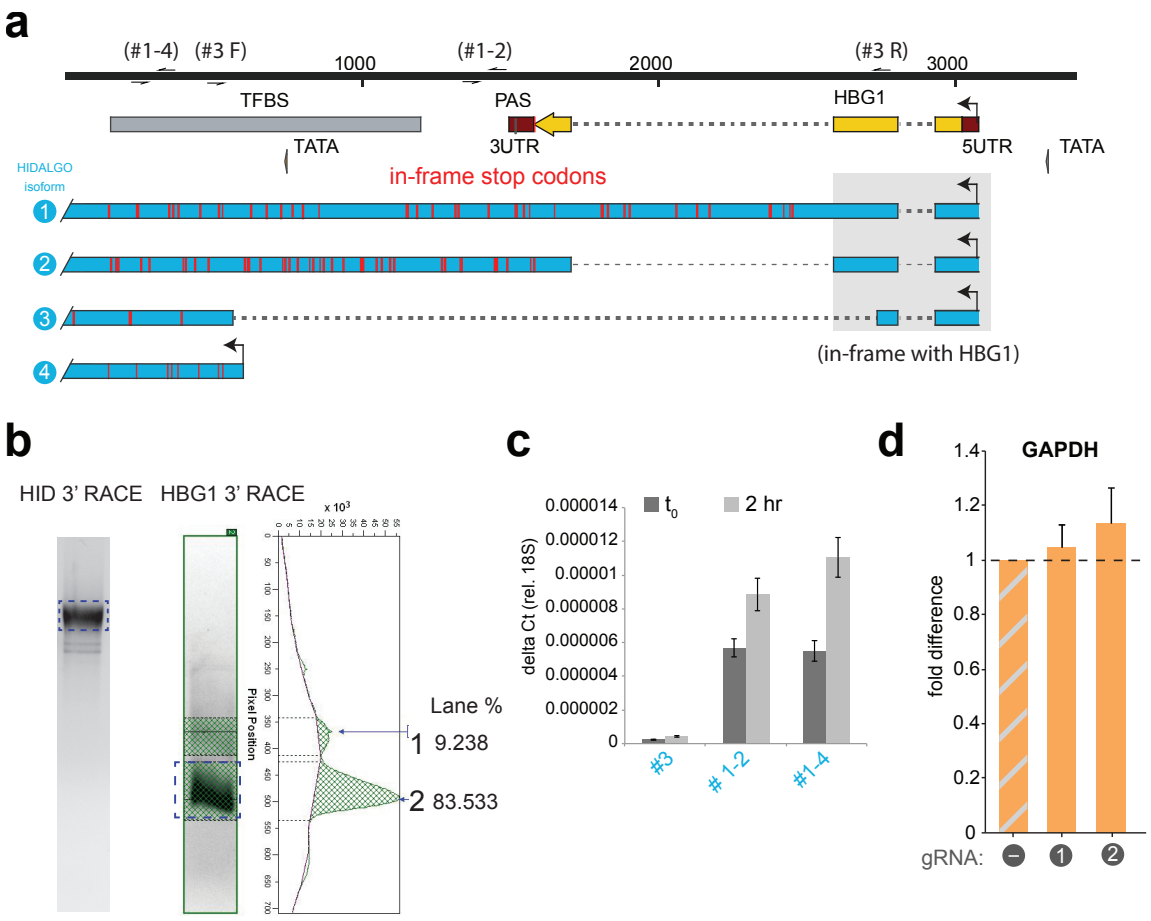


Figure B.6: Schematic of HIDALGO RACE products with indicated genetic features (PAS = polyadenylation sequence, TATA = TATA box, TFBS = transcription factor binding site). Red stripes indicate in-frame stop codons. (B) 3' RACE of HIDALGO and HBG1. Inset represents densitometry of HBG1 3' RACE run on an agarose gel and stained with ethidium bromide. (C) RT-qPCR of HIDALGO with a panel of qPCR primers demonstrates that RACE products 1-2 represent the majority of basal level expression, while an increase at two-hours after 50 microM hemin treatment is explained by RACE product 4 through deductive reasoning. (D) RT-qPCR of GAPDH after CRISPRi treatment with two gRNAs positioned the same distance away from the 3' end of GAPDH as gRNA-4 used for targeting HIDALGO (Figure 2.4A). (E) RT-qPCR of HIDALGO after treatment with LNA-antisense oligos (n = three independent biological replicates, error bars represent standard error mean).

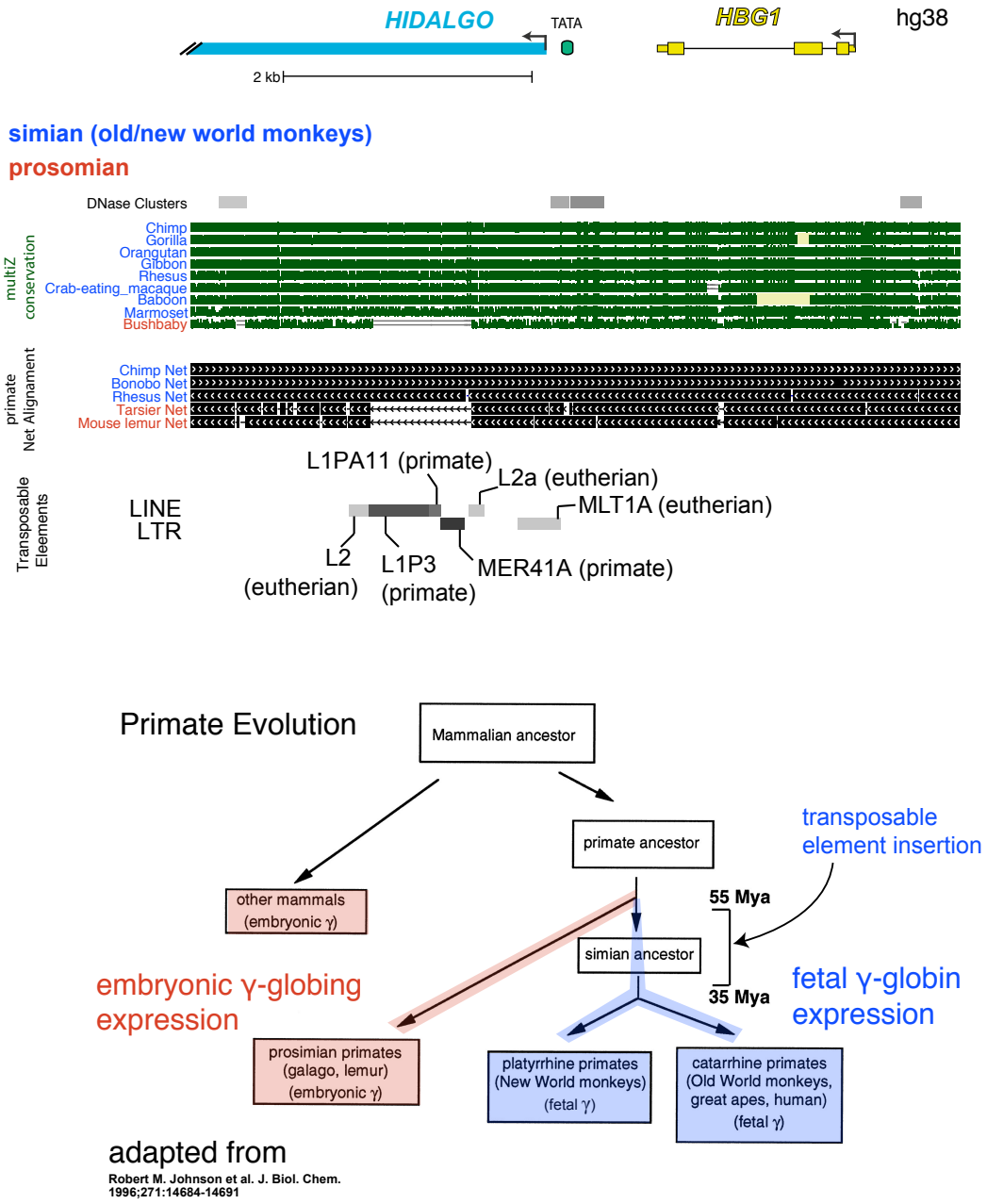


Figure B.7: HIDALGO is derived from transposable elements. (A) Phylogenetic comparison of HIDALGO promoter and adjacent region by Multiz alignment (Blanchette et al., 2004)(green histogram) and primate net-alignments (Kent et al., 2003)(black bars) in the UCSC genome browser (hg38) (http://genome.ucsc.edu/) (Kent et al., 2002; Speir et al., 2016). (B) Phylogenetic tree highlighting the introduction of transposable elements, and the relationship of fetal hemoglobin switching to the simian/prosimian divergence.

Chapter 3

BIOCHEMICAL DISSECTION OF WDR5-LNCRNA BINDING REVEALS MULTIPLE NOVEL INTERACTION SURFACES AND A MODEL FOR SPECIFICITY DESPITE LENGTH-DEPENDENT NON-SPECIFIC BINDING.

Figure 3.2B was performed by Dr. Pallavi Thaplyal. Figure 3.4B and C were performed in collaboration with Dr. Thaplyal, and Figure 3.4D was performed in collaboration with Dr. Thaplyal and Dr. Ankit Gupta.

3.1 Summary

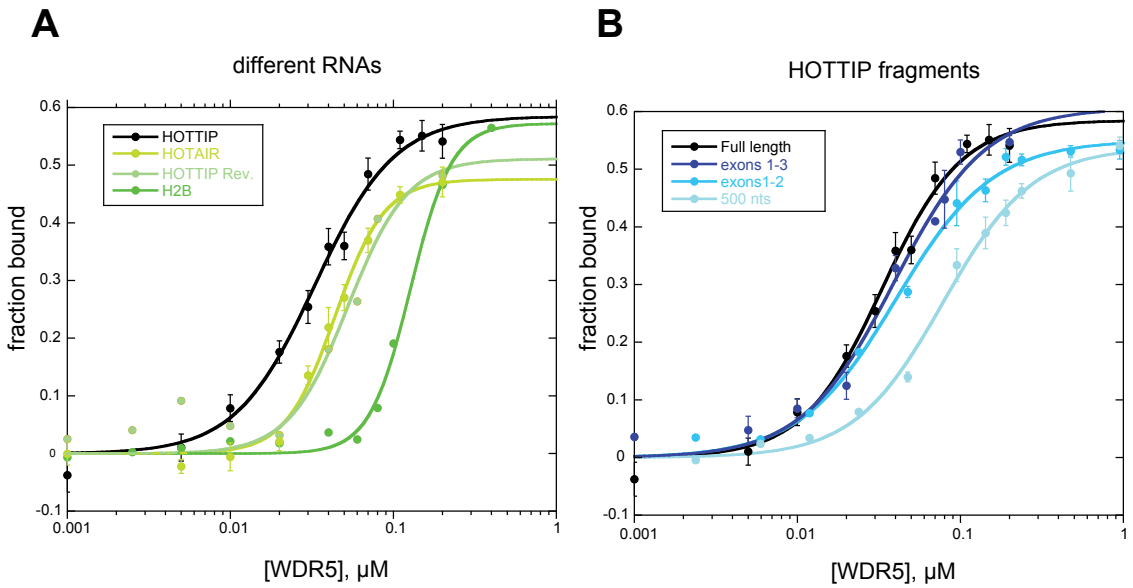
Over the last decade it has become apparent that non-coding microRNA, and more recently long non-coding RNA (lncRNA), are intricately woven into regulatory gene networks (Rinn and Chang, 2012). One of the most exciting discoveries to come out of these studies is that certain lncRNAs can recruit repressive histone-modifying complexes to specific genes (Pandey et al., 2008; Rinn et al., 2007; Martianov et al., 2007; Penny et al., 1996), and thereby affect epigenetic silencing. There are also increasing examples of parallel mechanisms that recruit the activating histone-modifying complex Mixed Lineage Lukemia (MLL) to it's target genes (Wang et al., 2011; Grote et al., 2013; Yang et al., 2014). However, although lncRNA-protein binding has received much interrogation by co-immunoprecipitation experiments (Rinn et al., 2007; Tsai et al., 2010; Grote et al., 2013; Zhao et al., 2008, 2010) it remains unclear what the molecular basis of the interactions are, and thus how to perturb them for putative pharmacological benefits. Indeed even when interrogated through in vitro studies, RNA appears to bind to the PCR2 complex relatively non-specifically in a length-dependent manner (Davidovich et al., 2013; Cifuentes-Rojas et al., 2014; Davidovich et al., 2015), and mutational studies have described binding pockets in the interior of proteins

(Yang et al., 2014) that might seemingly disrupt the fold of the protein. Specific binding of protein to RNA is a critical component of the proposed model whereby lncRNAs guide histone-modifying complexes to discreet targets. Here we describe a thorough biochemical analysis of lncRNA binding to the MLL subunit WDR5 both in vitro and vivo, and propose a simple-mathematical model to describe binding that harmonizes the current seemingly contradictory data.

3.2 WDR5 binds cognate lncRNA HOTTIP specifically, but affinity is increased by RNA length

To better understand the mechanism of selective lncRNA recruitment of MLL to target genes, we purified the WDR5 subunit that has previously been shown to bind thousands of lncRNAs (Wang et al., 2011, Yang et al., 2014), and performed quantitative filter-binding assays with in vitro transcribed RNA. As a measure of a direct binding partner we chose the lncRNA HOTTIP (Wang et al., 2014), and compared it to HOTAIR which recruits the repressive histone methyltransferase complex PRC2 (Rinn et al., 2007), histone 2B mRNA (H2B), and the reverse strand of HOTTIP. We discovered a tight affinity for HOTTIP (Kd approx. 30 nM)(Figure 3.1A). Although we demonstrated specificity for HOTTIP, we also noticed relatively tight and physiologically relevant Kds for all control RNAs tested (Figure 3.1A). For example, although H2B exhibited the weakest binding (Kd approx. 150 nM), H2B is a highly expressed mRNA present at several thousand more copies that HOTTIP, in addition to the myriad other RNA transcripts present in the cell. Furthermore, we noticed a length-dependent effect on affinity where longer RNAs provided added affinity for control RNAs (Figure 3.1A), and for varying lengths of HOTTIP (Figure 3.1B). Additionally, a pull-down of an RNA ladder (Promega), which presumably represents exclusively non-specific binding, with biotinylated WDR5 demonstrates greater pull-down of longer RNA species (Figure 3.2A). At the same time, we noticed that the interaction with HOTTIP is

resistant to salt, while control RNAs exhibit greater susceptibility (Figure 3.2B), suggesting that the interaction with HOTTIP is mediated by base-stacking contacts. Conversely, the contribution of relatively non-specific ionic interactions plays a greater role with control RNAs. Collectively, we establish using full-length HOTTIP and several non-cognate control RNAs in an in vitro system that WDR5 is capable of specific lncRNA-binding, but also exhibits non-specific length-dependent binding to RNA.



Construct	Kd
HOTTIP	28.4 ± 1.2
HOTAIR	43.9 ± 1.8
HOTTIP Rev.	52.7 ± 8.9
H2B mRNA	127.4 ± 6.4

Construct	Kd
Full length	28.4 ± 1.2
exons 1-3	37.1 ± 4.6
exons 1-2	40.3 ± 2.9
500 nts	78.2 ± 6.1

WDR5 binds specifically to HOTTIP in vitro. Equilibrium binding measurements of RNA with increasing concentrations of WDR5 were measured using filter binding on a Bio-Dot apparatus (BioRad). Fraction bound reflects intensity of P32-UTP RNA on nitrocellulose relative to total. Reactions were incubated at room temperature for 30 min. in FBA buffer (150 mM NaCl, 5 mM MgCl2, 50 mM Tris-Cl pH 7.0, 0.1 mM EDTA), and ~ 15 pM RNA. Kds were determined by fitting replicate fraction bound measurements to the hill equation: f_{max} x ([WDR5]^n)/(Kd^n+WDR5]^n); HOTTIP FL/WT WDR5 n = 21 all other combinations n = 3. Averages of individual representative experiments (n=3) are shown for display, error bars indicate standard error mean.

Figure 3.1

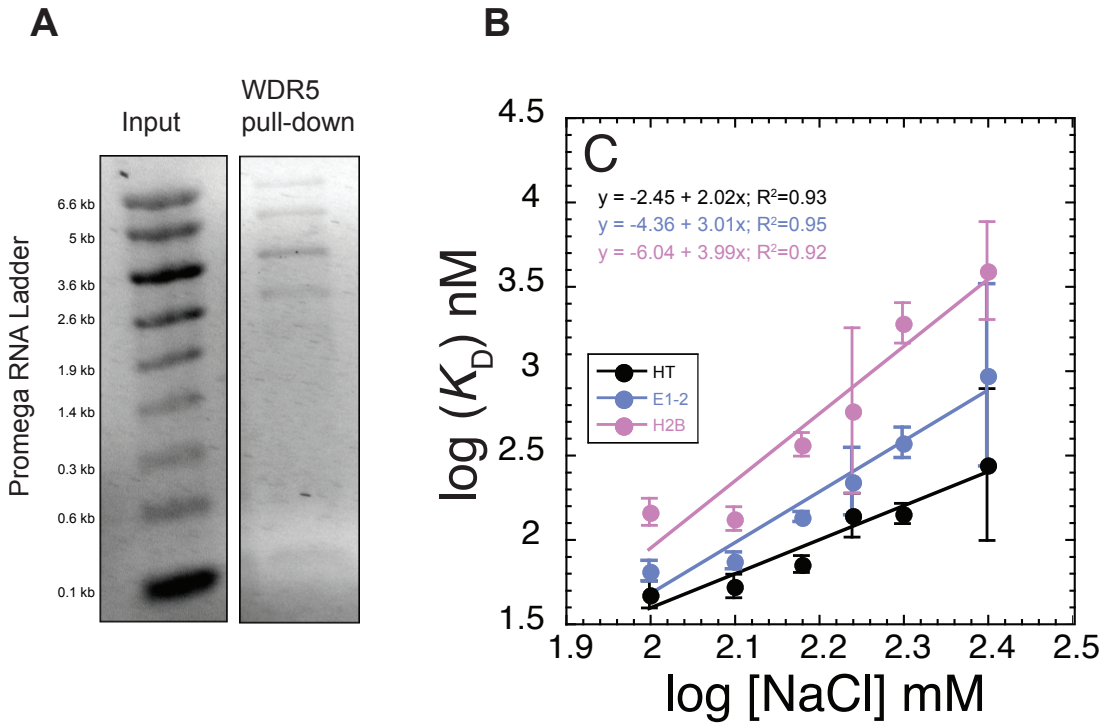


Figure 3.2: WDR5-lncRNA binding is length-dependent, but relatively salt-insensitive for cognate partner HOTTIP. (A) Biotinylated WDR5 was incubated with RNA ladder (Promega) with head-over-tail rotation for 30 minutes, then protein-RNA complexes were purified by magnetic streptavidin resin. (B) **Contributed by Dr. Pallavi Thaplyal**, A sodium chloride titration reveals differences in susceptibility to ionic-buffer strength of WDR5-RNA binding.

3.3 Identification of novel RNA-binding surfaces on WDR5

We next sought to identify a surface on WDR5 that binds to lncRNA, which once established could be perturbed to assay the contribution of lncRNA binding in general in a variety of cell contexts. A recent paper from the Chang lab (Yang et al., 2014) identified a patch on WDR5 that when mutated severally decreased RNA binding by pull-down - RT-qPCR. However this mutation is of a hydrophobic phenylalanine buried in the interior of WDR5 (Figure 3.3A), raising the possibility that the mutation disrupted WDR5 stability, and not specifically RNA-binding. First, we measured the melting temperature of the F266A mutant relative to WT WDR5. While both the mutant and WT WDR5 proteins are predominantly beta-sheet at room temperature by circular dichroism (Figure 3.3B), we discovered a striking destabilizing effect of the mutant in a denaturation assay (Figure 3.3C). The F266A shifts the melting temperature (Tm) of WDR5 from approx. 75 to 47 degrees C. Importantly, the mutant begins to unfold at 37 degrees, jeopardizing the conclusions drawn by Yang and colleagues (2014) about this mutant in an in vivo setting. To further test this hypothesis, we performed filter-binding assays on the mutant and WT proteins at room temperature and at 37 C. We find that at 37 C we recapitulate a defect in RNA binding, shifting the Kd by greater than 6-fold (Figure 3.3D), however at four degrees the F266A mutant binds as well if not better than WT (Figure 3.3E).

Seeing as the F266A mutant is called into question as an RNA-selective interface, we screened a panel of mutants at various locations on WDR5. We identified a basic patch on the side of WDR5 that when mutated at various locations inhibits HOTTIP-binding by 2-3 fold (Figure 3.3A). These interactions are likely electrostatic, and perhaps represent non-specific interactions with the RNA phosphate backbone. We also discovered a series of aromatic and hydrophobic ring residues at the top-face of WDR5 that when mutated also caused an approx. 3-fold reduction in RNA binding (Figure 3.3B). These non-polar side-chains could form pie-stacking interactions with RNA bases, which could also explain

salt-resistant binding observed with WDR5 and HOTTIP. Intriguingly, while the individual patches contribute significant but relatively modest affinity for RNA, the combination of basic-patch and top-face mutations completely inhibited RNA binding (Figure 3.3C). Importantly we demonstrated by melt-curves that all novel mutants do not significantly alter the melting temperature of WDR5 (data not shown). Finally, the position of the top-face planar residues are in close contact to several important WDR5 protein-protein interactions, including with MLL, and thus precludes targeted in vivo perturbation. However the side basic-patch is not known to be involved in other WDR5 functions, so we designed competitive inhibitor protein monobodies against this region. A titration of monobody disrupted WDR5-HOTTIP binding, and is potential tool for investigating the role of this patch in RNA binding and its molecular functions in vivo.

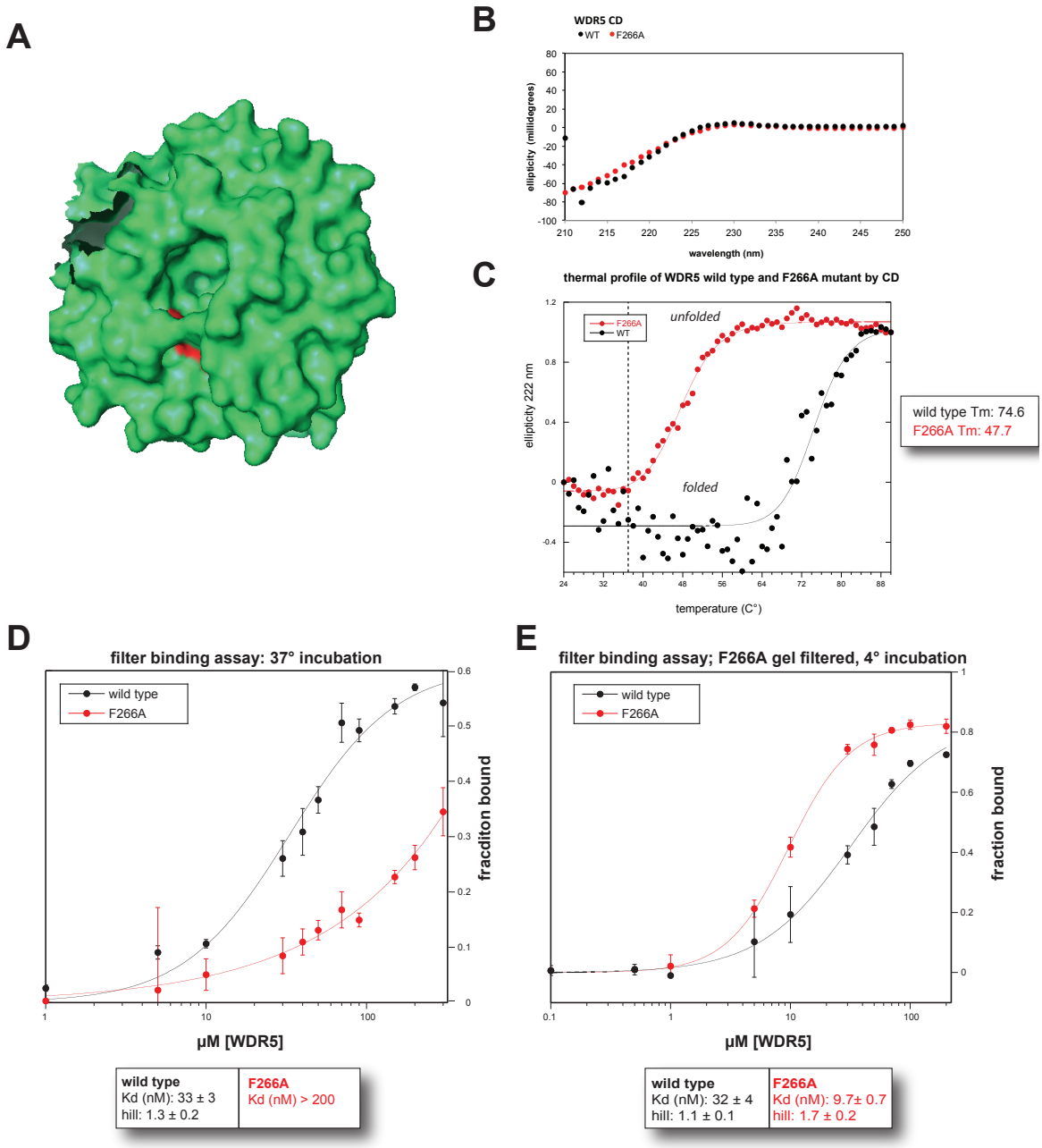


Figure 3.3: F266A mutant disrupts WDR5 protein stability. (A) F266 (red) is a hydrophobic residue buried in the interior of the WDR5. (B) Circular Dichroism demonstrates that WT and F266A mutants are predominantly beta-sheet at room temperature. (C) Melt-curve and (D-E) filter binding of WDR5 WT (black) and F266A mutant (red). In (D) binding was performed at 37 degrees C, while in (E) binding was performed at 4 degrees.

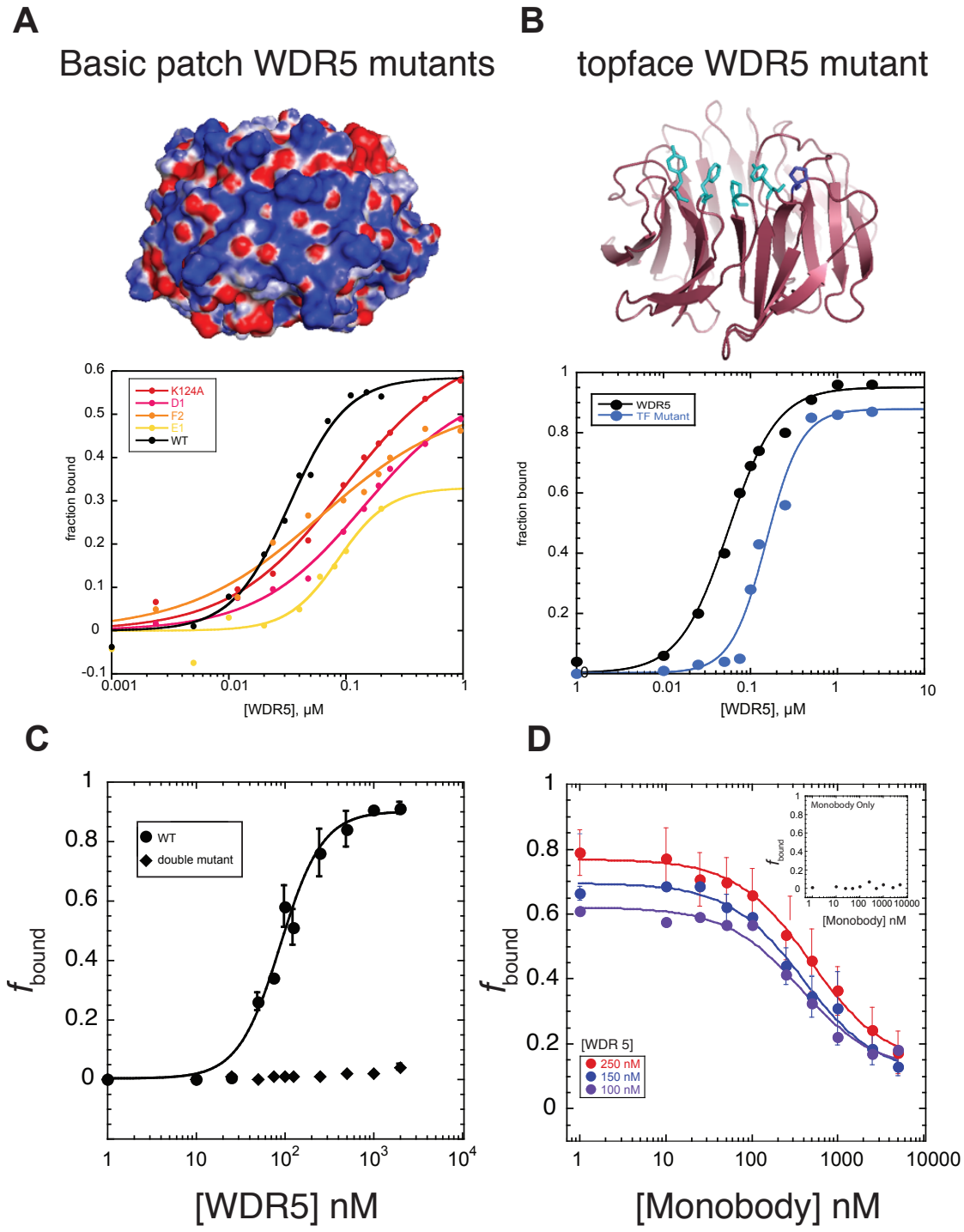


Figure 3.4: Identification of novel RNA-binding surfaces on WDR5. (A) Basic-patch (blue equals basic, red equals acidic) and (B) (**in collaboration with Dr. Pallavi Thaplyal**) top-face mutants were analyzed by filter-binding assays. (C) (**in collaboration with Dr. Pallavi Thaplyal**) A combined mutant (E1 from basic-patch and top-face mutants) completely eliminates RNA-binding. (D) Competetive inhibition of WDR5-RNA binding by increasing concentrations of an engineered monobody **(Courtesy of Dr. Shohei Koide, Dr. Ankit Gupta, and Dr. Pallavi Thaplyal)**.

3.4 In vivo RNA-binding is specific

A previous cross-linking experiment in mESC suggested that WDR5 binds many lncRNAs, however their comparison was to a F266A mutant, and involved chemical crosslinking which is non-specific to WDR5. To improve upon this method, and potentially identify RNA motifs we performed UV-crosslinking in cells doped with a 4-thiouracil analog (4SU) (PAR-CLIP). To obtain robustly purified IPs we created HEK293 cell lines with FLAG-HA-8xHis-WDR5 stably incorporated into an FRT site in the genome. This system allowed tandem purification (Figure 3.5A), and the second His-Cobalt IP allows the inclusion of denaturing buffer. A global assessment of sequenced RNA from this experiment demonstrates specificity of WDR5-RNA binders relative to total RNA (Figure 3.5B). There were 3,117 enriched clusters, and 855 crosslinking sites found for WDR5 (see Figure 3.5C for example). Subsequent analysis will attempt to decipher RNA motifs within these crosslinking clusters.

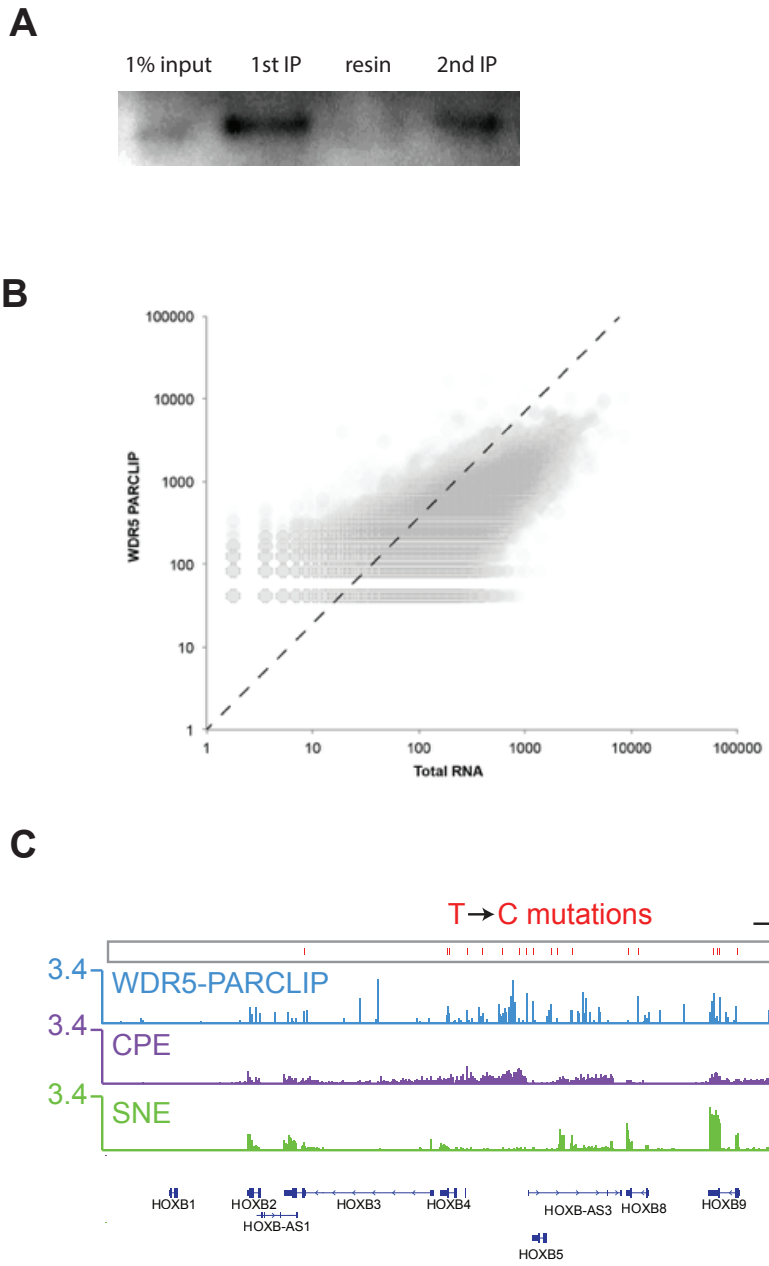


Figure 3.5: WDR5 PARCLIP. (A) Western blot of tandem duplication reveals good enrichment and conservation of protein between IPs. (B) Scatter plot of RNA abundance in total RNa (x-axis) compared to WDR5 PARCLIP (y-axis) demonstrates a right skew, and thus specificity of certain transcripts in the WDR5 IP (left of the 1-1 slope dotted line). (C) RNA-seq of WDR5 PARCLIP (blue) compared to CPE and SNE from HEK293 at the HOXB locus.

3.5 Discussion

Finally, I propose a simple model that to some extent accounts for the disparity observed between in vivo specificity and in vitro length-dependence. This model is based on the assumption that instead of a single large structured motif, histone modifying complexes bind to small linear RNA motifs through base-stacking interactions. Despite our fascination with complex RNA structure, this mode of binding appears to be the dominant mechanism utilized by RNA binding proteins in vivo (Ray et al., 2013). A small motif comprising between 3-10 nucleotides would naturally provide only a modest affinity, however multiple such motifs within the same RNA, perhaps presented in stem loops created through RNA structure, could provide significant avidity. In an in vitro setting without endogenous proteins and absent in vivo RNA structure, such small motifs could be exposed, and would randomly increase in frequency within an RNA in a length-dependent manner, in addition to non-specific ionic interactions that would also increase in frequency with longer nucleotides. A simple model that accounts for the random accumulation of small motifs of length (n) and approximate motif (Kd n) was derived and closely resembles measured Kds (Kd m) (Figure 3.6). Future exploration of crosslinking sites from WDR5 PARCLIP may substantially inform the accuracy of this model, and allow functional testing therefore.

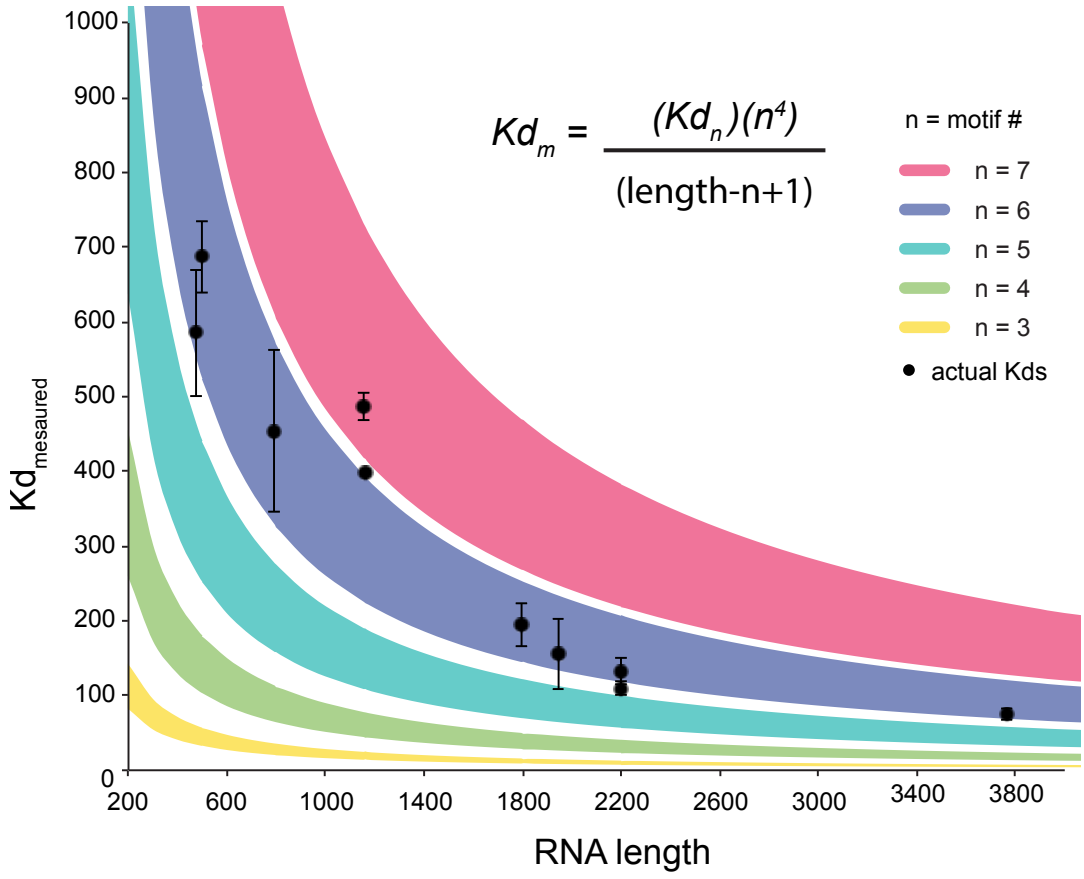


Figure 3.6: Simple model describes length-dependent RNA affinity.

Appendix C

CHERNAS ARE TISSUE-SPECIFIC AND ACTIVATE PROXIMAL CODING GENES

C.1 Methods

Protocol: General Protein Expression

Notes: The following is adapted from the pET system manual.

Steps:

Day 1 1. Transform BL21(DE3) pLysS cells or other suitable expression host, with a plasmid carrying antibiotic resistance, and your gene of interest under T7 promoter-LacI control. BL21 is a chemically competent cell line, DE3 represents that the T7 RNA polymerase is stably integrated into the host genome, and pLyse S indicates that a plasmid under chloramphenicol resistance carrying lysozyme is present (in the reverse orientation to a T7 promoter), allowing basal levels of expression and therefore inhibition of leaky T7 for tighter control of expression. Day 2 1. Grow up a 10 ml starter culture from a single transformed colony in LB+antiobitic. Grow at 37 to an O.D.600 of 0.05-0.1 (4-6 hours). If inoculating the next day, keep starter culture at 4 overnight, spin down the next day and re-suspend in fresh media. 2. Inoculate 1 Liter LB+antibiotic 1:100 with starter culture (or all of re-suspended 4 o/n starter culture), and grow 37 to O.D.600 0.3-0.6 (2-4 hours). 3. Aliquot 1ml and tape to the side of the flask as an un-induced control.

For 37 fast induction, add IPTG to 0.4-1mM final concentration and grow 3 hours

For 30 medium induction, switch flask(s) to pre-set 30 incubator and add IPTG to 0.4-1mM final concentration. Grow 5-6 hours.

For overnight slow induction, switch flask(s) to pre-set 16 incubator and let equilibrate for 30 minutes, then add IPTG to 0.4-1mM final concentration. Grow 14 hours.

4. Aliquot 1ml as an induced control. Decant culture into plastic containers that fit

swinging bucket centrifuge. Spin 11,000 x G, 4, 20 minutes (or 2x 10 minutes). Decant soup and re-suspend in 4-6 mls Lysis buffer (500 mM NaCl, 10 mM imidizole, 10 pct. glycerol, 50 mM K-phosphate buffer, pH 8.0). Transfer to a 50 ml conical, wash remaining pellet with 1-2 ml more lysis buffer and add to conical. 5. Flash freeze in liquid nitrogen and store -80. 6. To check for expression, centrifuge 1 ml +/- inductions 5 min 5,000 x g. 7. Decant and re-suspend in 50 microl Tris-Lysis buffer (Kphosphate will precipitate in SDS). Aliquot 5 microl to a separate tube, add 20 microl water + 15 microl 6x SDS. 8. Boil 5 min., centrifuge max speed 5 min. 9. Load 3 microl of supernatant into gel.

Note, if inducing biotinylation, add 50 microM biotin at induction, and optionally an additional 10-20 minute 50 microM incubation of the re-suspended pellet before flash-freezing.

Protocol: Histidine Tag Protein Purification

Notes: Phosphate is preferred over Tris when using a divalent metal column (i.e. Nickel) because Tris, as a primary amine, can compete with the his-tag for resin-immobilized divalent metals. The pH of the buffer at 8 is critical to de-protonate histidines (pKa between 6.2-6.9), and thus enable them to bind Ni(II). At pH 8, greater than 95 pct. of histidines are deprotonated. Low concentrations of BME (10 mM) are preferred over DTT because DTT is a bi-dentate sulfhydryl that can react with the Ni and the nitrolacetic acid (NTA) that bridges Ni to the resin, stripping off the metal from the column. Glycerol is a non-ionic kosmotrope that stabilizes proteins by preserving the hydration layer (water shell) surrounding them. Finally, imidazole is the functional group of histidine, so low concentrates compete with histidines on other proteins during incubation, and at high concentrations will out-compete the histidine tag on your protein for Ni on the column causing your protein to elute.

All applications should be handled on ice or at 4 C in the cold room. Once the cell pellet is thawed, avoid freezing protein until it has been verified that this does not diminish its activity.

Steps:

Lysis

1. Aliquot 50 ml of wash/lysis buffer and 50 ml of low salt buffer and supplement to 10 mM BME (7 microl liquid/10 ml buffer; dispose of tips under the fume hood), 1x protease inhibitor cocktails, and 0.5 mM PMSF (from a 100 mM stock solution in absoluate ethanol). Aliquot 18 ml of supplimented Wash/Lysis buffer to a separate tube and add 2 ml of 5 M NaCl to obtain 1 M final concentration High Salt Wash Buffer. Aliqout 18 ml of supplemented Low Salt Buffer to a separate tube and suppliment to 300 mM imidizole with 2 ml 3 M Imidazole to obtain Elution Buffer. Keep all buffers on ice or in the cold room for up to 1 month.

2. Thaw frozen cell pellet in a beaker filled with ice-water. Add 20 ml of wash/lysis buffer, and mix gently by inversion every few minutes.

3. Filter through a 100 microm nylon mesh using a filter:syringe into a 50 ml conical tube. Filtering will prevent clumps of cells from clogging the emulsiflex, and obviates the need for sonication.

4. Lyse cells by Emulsiflex:

i. Bring: 50 ml falcon tube for collection, 2 large plastic beakers (fill one with ice-water), ice bucket, Wash/Lysis buffer, and Nylon Mesh filtered sample, 70 pct. isopropanol. ii. Turn on machine: a. plug in (cord has purple tape labeled "Binder") b. Turn red knob in the back to the right c. Turn on air from building (orange "air") iii. Place coil in ice-water, and outlet tubing in a waste beaker iv. Remove top of cone/sample chamber by unscrewing v. Turn red button in the front to the right to allow green button to be pressed vi. Press green button to pump isopropanol through the sample chamber, stop by pressing the red button when it starts bubbling. vii. Unhinge sample chamber, and wash sample chamber and black gasket with di-water thoroughly. viii. Re-attach sample chamber, and fill to 'elbow' with water. ix. Push green button to pump water through. Refill when water is near bottom of sample chamber. Repeat for 3x total then hit stop button. x. Fill to elbow with wash/lysis

buffer, and pump through until "dime-size" circumference, then hit stop. Important not to let air through at this point. xi. Pour nylon mesh-filtered cells into sample chamber. xii. Push green button to pump cells through sample chamber, and when cells reach near the end of the tubing, move tubing into sample chamber to 'complete the circuit'. xiii. Once cells are comfortably moving through the loops, turn grey pressure knob (located on the front of the machine) slowly to the right. Eventually this will cause the needle to jump, continue turning the knob until each jump hits 15K psi. Count the number of jumps to 15K psi and stop at 2x the volume of the sample cells loaded (i.e. 30 mls = 60 jumps). To stop, simply turn grey knob back to the left until psi reaches baseline. xiv. Turn machine off by hitting red button, then switch outlet tubing to a collection tube on ice, and pump lysed cells into it by pressing the green button. Turn off when bubbles start emerging from tubing. xv. Wash 5x with water, and 5x with 70 pct. isopropanol. Leave 25 ml 70 pct. isopropanol after 5th wash, turn off machine by red switch at the back, turn off air, and unplug. xvi. Good to go!

Nickel Affinity purification

1. Clear Lysate by centrifuging 30,000 x g 20 minutes, 4C in Lucia's old centrifuge. Save 40 microl of lysate for gel analysis. While spinning, prepare Bio-Rad column: i. wash with diwater and allow to pass through frit a few times. ii. Add 2 ml of 50 pct. re-suspended 50 pct. Ni-NTA (1ml slurry) and allow to drain. iii. Add 5 ml wash/lysis buffer and allow to drain to equilibrate column. 2. Decant lysate soup into column, put cap on, and incubate 30 min, 4C with gentle agitation. 5. Set up 7 50 ml conical tubes; collect flow-through by unplugging frit and allowing to drain into a 50 ml collection tube. 6. Wash by adding 15 ml high-salt wash buffer, collecting wash flow-through. Repeat with 10 ml wash/lysis buffer, and then 10 ml low-salt wash buffer. 7. Place elution collection tube under column, then elute 3x 5 ml into 3 collection tubes. 8. (Optional) run SDS-PAGE of lysate, flow-through, elutions, and washes to verify lysis and affinity selection/elution. Washes and Elutions: 10 microl + 3microl 6x SDS buffer Lysate and Flow-through: 3microl + 3microl 6x SDS buffer

Pellet chip: 20 microl 6x SDS + 20 microl water, after boiling centrifuge and load 3 microl of supernatant. Resin: 15 microl of what is left in column + 10 microl 6x SDS Boil all sample 5 minutes before loading into gel

9. Nanodrop elutions, blanking with elution buffer and calculate quantity of eluted protein based on beers law:

Absorption (nm 280) = elc

e= molar absorptivity (protein-dependent) c = concentration l = path length

TAG cleavage

1. If you ran gel to determine amount of protein in each elution, combine desired elutions, and supplement to 2 mM EDTA and 3 mM DTT. Add 1:100 (w/w) TEV protease. Incubate overnight at 16C. 2. Next day, add fresh TEV protease and incubate 1hr room temperature, then run 10 microl with 3 microl 6X SDS on SDS PAGE to determine efficiency of cleavage.

FPLC

1. Load column by peristaltic pump: Need: Timer, three 50 ml conical tubes, TEV digested protein, Dilution buffer and Buffer A supplemented with BME and PMSF, razor blades, peristaltic pump, 100 mM PMSF, BME, conical rack, beaker i. Aiquot and supplement Dilution buffer and Buffer A to 10 mM BME and 1 mM PMSF, typically 30 mls of both ii. Dilute TEV-digested protein 1:3 with dilution buffer (150 mM Tris-HCl, pH 6.9, 10 pct. glycerol), eg. Add 30 mls to 10 mls of elutions. This is important to alter the pH of protein solution so that it sticks to the column, and to lower the [salt]. In this case the dilution buffer lowers the pH below the PI of the protein, effectively protonating it?s basic residues, and thus allowing it to stick to a negatively charged column (i.e. heparin). If you want to use a positively charged column, alter the dilution buffer to be a higher pH than the PI, but otherwise the method should be the same. iii. Slice holes through conical caps with razor blade to make an "x", and widen hole through center with pen. iv. Place input tubing into Buffer A through hole in cap, and output tubing into a beaker. v. Wash out storage

ethanol in tubing by turning peristaltic pump on. vi. Measure flow rate by timing volume added to a microcentrifuge tube in 1 minute. Note, previously, "5" is about 1 ml/minute. vii. Wet column head then gently screw-in. Equilibrate column with 2x column volumes Buffer A. viii. Turn peristaltic pump off, switch input tubing from Buffer A to diluted protein, and switch output tubing from collection beaker to "Flow-Through" conical. ix. Turn peristaltic pump back on to load column. Set timer based on flow-rate for 1/2 volume and ? volume to check status of flow. Very Important not to get air into column at this time, as it could both damage protein and the column. Stop peristaltic pump when there is about 500 microl remaining. x. Remove column, it is now ready for the FPLC. Wash peristaltic pump by switching input tubing to 20 pct. ethanol and output to beaker and flow through a few minutes.

2. Elute off column on FPLC Need: 150-200 ml Buffer A and B supplemented to 10 mM BME and pre-loaded column. i. Supplement Buffer A and B with BME, and place "A" and "B" suction cups in them. ii. Prepare fraction collector with open 2 ml tubes and place in notch in FPLC iii. Open AKTA software if its not already open iv. Click on Manual, Run, select method, i.e. "MW Heparin" v. Click through various windows until satisfied with protocol. Important parameters to remember:

1. Include a Pump Wash Purifier step Input A1 and B1

2. Set flow rate and pressure alarm to the specified column being used

3. Consider including a few column volumes (CV) of Buffer A before starting flow of elution buffer B vi. Start method. After Pump Wash Purifier finishes, allow a few drops of buffer A to start emerging from input tubing, then wet the column head and screw in. vii. Check that fractions are being collected correctly. viii. When fractions are done collecting switch the suction cups from Buffer A and B to 20 pct. ethanol, and (if not already included in the method) wash the column with 5 CVs, then perform a pump wash purifier to clean the system ix. Run 10 microl of fractions that have protein with 3microl of 6x SDS buffer

(boil 5' then spin down first) to check the quantity and purity of protein.

Protocol: Denaturing RIP

Buffers and Reagents

1. Buffer A: 10mM HEPES pH 7.5, 4 mM MgCl2, 10 mM KCl, 0.5 mM DTT, 340 mM Sucrose, 10 pct. glycerol, 5 mM BME, 1x Protease Inhibitor Cocktail

2. RIP Wash Buffer: 50 mM Kphosphate buffer pH 7.2, 5 mM MgCl2, 300 mM KCl, 0.01 pct. Tween-20, 1x Protease Inhibitor Cocktail, 5 mM BME

3. Denaturing Binding/Wash Buffer: 0.009 pct. Tween-20, 200 mM Kphosphate pH 7.95, 7.3 M Urea, 18 mM imidazole, 9 pct. glycerol, 0.9M NaCl

4. Proteinase K buffer: 0.25 pct. SDS, 10 mM HEPES, pH 7.5, 250 mM NaCl, 10 pct. glycerol, 0.5 pct. Tween-20. *Final pH = 6.85

5. TriZole or Phenol:Chloroform:IAA + 3M Na Acetate

6. 10 mg/ml glycogen

7. DNaseI re-suspended to 20U/microl and 1U/microl

8. Cobalt Dynabeads (Cat number 10103/4/5/D)

9. 2 ml non-adhesive microcentrifuge tubes

10. RiboLock Rnase Inhibitor 40 U/microl

11. 1M Imidazole

12. Proteinase K 10mg/ml

13. 1 and 6X SDS PAGE-loading buffer

14. BME

15. Zymo RNA Clean and Concentrate (25microg max) Columns

Supplement relevant buffers with DTT/BME, and Protease Inhibitor Cocktail directly before use. For 5 mM BME, add 3.5 microl / 10ml buffer

Notes: The exact amount of input (cells) will depend on how much material and dynabeads required, nevertheless the following is a general protocol for 20 x 106 HEK 293 cells

(2, 10 cm confluent plates).

20 x 106 cells yields 130 microg total nucleic acid content by nanodrop/HEK 293 nuclear extract input.

All washes include 1 ml of wash buffer for 3 minutes with head/tail rotation, followed by a minute magnetic separation.

Steps:

Day 1 I. Add 4SU 1. Grow up 2, 10 cm plates/experimental sample to 75 pct. confluency, then add 1microl of 1 M 4SU to each plate (100 microM final concentration) and incubate 12-18 hours 37C.

Day 2 II. Collect cell pellets 1. Decant media and wash cells with PBS. Remove all wash and place on ice tray in UV stratalinker. Cross-link with 365 nm wavelength bulbs for 0.5 J/cm2 (set to 5000 on display). 2. Add 1 ml of PBS to each plate and scrape cells with a sterile shovel, and aliquot to a conical tube on ice. Centrifuge 500 x G for 5 min., 4 C. 3. Decant soup and flash freeze cell pellet with liq. N2. Store at -80 C until processing.

Approximate time: 30 min.

Day 3and II. Obtain Nuclear extracts

Samples:

1. Measure packed cell volume (PCV) of all cell pellets to be used, and aliquot 20x volume of buffer A (if pcv = 0.2 ml, aliquot 4 mls buffer A). Supplement with protease inhibitor, and 5 mM BME.

PCV= Volume Buffer A = 20x = Protease inhibitor: 5 mM BME:

2. Aliquot 5x pcv supplemented buffer A (sBA) to a separate tube and add Triton X-100 to 0.2 pct.. 5x pcv= Triton X-100 = 5x pcv x (0.2 pct./10 pct.) =

3. Add 2.5x pcv sBA to cell pellet and allow to thaw. Re-suspend to homogeneity, measure aqueous volume, and add an equal volume of 0.2 pct. Triton X-100 sBA

2.5x pcv = aqueous volume=

4. Incubate cells on ice for 12? with occasional gentle mixing. 5. Centrifuge 1,200 x g, 4, 5 min. 6. Save cytosolic extract and freeze, decant the rest. 7. Re-suspend nuclei in 0.4 ml sBA. Aliquot 5microl to 45 microl sBA (1:10), and count with a hemocytometer.

Sample nuclei counts:

Aliquot x106 nuclei/experimental sample, and normalize all samples to the same volume with sBA. 8. Add 50 U DNase and incubate 37 C for 10 minutes. 9. Add 5M NaCl to 0.4 M NaCl final concentration, and 500 mM EDTA to 10 mM final concentration, and incubate 4 C for 30 minutes.

While extracting prepare 12CA5-dynabeads (see below)

10. Centrifuge 15,000 x g, 4 C, for 15 minutes to pellet chromatin and release soluble nuclear extract. 11. Aliquot extract to a new tube and add an equal volume of sBA to lower [salt]. 12. Aliquot 5 pct.and 1 pct. v/v input to separate tubes and save (RNA and Protein analysis, respectively). 13. Add 100 U/ml RiboLock RNase inhibitor to samples.

Approximate time: 3 hours

IV. 1st Immunoprecipitation (native) 1. Prepare 12CA5 dynabeads by aliquoting microl/sample to 2ml non-adhesive microcentrifuge tubes, and remove storage buffer. Add 1 ml 12CA5 serum and incubate 30 minutes room temp with rotation. Remove serum with magnet, and wash beads 3x with RIP wash buffer. Re-suspend beads in original bead volume PBS 2. Add 12CA5-conjugated dynabeads to nuclear extracts, and incubate 2-4 hr with rotation, 4 C. 3. Collect flow-through and 4. Wash 3 x with RIP wash buffer, including 2x tube changes at 4 C.

* While washing, Aliquot microl of cobalt-dynabeads/experimental sample and remove storage buffer with magnet particle separator.

V. 2nd Immunoprecipitation 1. Add 700 microl Denaturing Bind/Wash Buffer directly to beads, and incubate 5 minutes 50 C in shaking incubator 600 rpm, then collect denatured protein-RNA solution by magnetic partical separator (aliquot and save 10 pct. for RNA

analysis) and transfer to Cobalt-dynabeads. 2. Incubate 5 minutes with rotation. 3. Wash 4x with Denaturing Bind/Wash Buffer, include 2x tube changes. 4. Wash 1x with Proteinase K Prep. Buffer, before removing wash Aliquot and save 5 pct. for western blot analysis. 5. Re-suspend beads in original bead volume Proteinase K Buffer. 6. Add 20 mg/ml Proteinase K to 1 mg/ml final concentration. Incubate 40 minutes 50 C in shaking incubator 600 rpm. 7. Transfer elution to fresh tube. Wash beads with another bead volume of Proteinase K buffer and combine with elution.

Approximate time: 6 hours

VI. RNA clean up

1. Add equal volume Phenol:Chloroform:IAA, pH 8.0, and mix vigorously 30 seconds, let sit 1 minute room temp., then centrifuge 15, 000 x g 2 minutes 4 C. 2. Use aqueous layer as input for Zymo RNA Clean and Concentrator columns. Process according to manufacturers protocol, and include "In-tube" DNase digest. 3. Elute in 30 microl RNase-free water, and flash freeze. END DAY 3!

References

Addya, S. (2004). Erythroid-induced commitment of K562 cells results in clusters of differentially expressed genes enriched for specific transcription regulatory elements. *Physiological Genomics*, 19(1):117–130.

Addya, S., Keller, M. A., Delgrosso, K., Ponte, C. M., Vadigepalli, R., Gonye, G. E., and Surrey, S. (2004). Erythroid-induced commitment of K562 cells results in clusters of differentially expressed genes enriched for specific transcription regulatory elements. *Physiological Genomics*, 19(1):117–130.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F. O., Jørgensen, M., Andersen, P. R., Bertin, N., Rackham, O., Burroughs, A. M., Baillie, J. K., Ishizu, Y., Shimizu, Y., Furuhata, E., Maeda, S., Negishi, Y., Mungall, C. J., Meehan, T. F., Lassmann, T., Itoh, M., Kawaji, H., Kondo, N., Kawai, J., Lennartsson, A., Daub, C. O., Heutink, P., Hume, D. A., Jensen, T. H., Suzuki, H., Hayashizaki, Y., Müller, F., FANTOM Consortium, Forrest, A. R. R., Carninci, P., Rehli, M., and Sandelin, A. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461.

Andrews, S. J. and Rothnagel, J. A. (2014). Emerging evidence for functional peptides encoded by short open reading frames. *Nature Reviews Genetics*, 15(3):193–204.

Ansari, A. and Hampsey, M. (2005). A role for the CPF 3'-end processing machinery in RNAP II-dependent gene looping. *Genes & Development*, 19(24):2969–2978.

Armstrong, L. (2006). The role of PI3K/AKT, MAPK/ERK and NF signalling in the maintenance of human embryonic stem cell pluripotency and viability highlighted by transcriptional profiling and functional analysis. *Human Molecular Genetics*, 15(11):1894–1913.

Arner, E., Daub, C. O., Vitting-Seerup, K., Andersson, R., Lilje, B., Drabløs, F., Lennartsson, A., Rönnerblad, M., Hrydziuszko, O., Vitezic, M., Freeman, T. C., Alhendi, A. M. N., Arner, P., Axton, R., Baillie, J. K., Beckhouse, A., Bodega, B., Briggs, J., Brombacher, F., Davis, M., Detmar, M., Ehrlund, A., Endoh, M., Eslami, A., Fagiolini, M., Fairbairn, L., Faulkner, G. J., Ferrai, C., Fisher, M. E., Forrester, L., Goldowitz, D., Guler, R., Ha, T., Hara, M., Herlyn, M., Ikawa, T., Kai, C., Kawamoto, H., Khachigian, L. M., Klinken, S. P., Kojima, S., Koseki, H., Klein, S., Mejhert, N., Miyaguchi, K., Mizuno, Y., Morimoto, M., Morris, K. J., Mummery, C., Nakachi, Y., Ogishima, S., Okada-Hatakeyama, M., Okazaki, Y., Orlando, V., Ovchinnikov, D., Passier, R., Patrikakis, M., Pombo, A., Qin, X.-Y., Roy, S., Sato, H., Savvi, S., Saxena, A., Schwegmann, A., Sugiyama, D., Swoboda, R., Tanaka, H., Tomoiu, A., Winteringham, L. N., Wolvetang, E., Yanagi-Mizuochi, C., Yoneda, M., Zabierowski, S., Zhang, P., Abugessaisa, I., Bertin, N., Diehl, A. D., Fukuda, S., Furuno, M., Harshbarger, J., Hasegawa, A., Hori, F., Ishikawa-Kato, S., Ishizu, Y., Itoh, M., Kawashima, T., Kojima, M., Kondo, N., Lizio, M., Meehan, T. F., Mungall, C. J., Murata, M., Nishiyori-Sueki, H., Sahin, S., Nagao-Sato, S., Severin, J., de Hoon, M. J. L.,

Kawai, J., Kasukawa, T., Lassmann, T., Suzuki, H., Kawaji, H., Summers, K. M., Wells, C., FANTOM Consortium, Hume, D. A., Forrest, A. R. R., Sandelin, A., Carninci, P., and Hayashizaki, Y. (2015). Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science*, 347(6225):1010–1014.

Berghoff, E. G., Clark, M. F., Chen, S., Cajigas, I., Leib, D. E., and Kohtz, J. D. (2013). Evf2 (Dlx6as) lncRNA regulates ultraconserved enhancer methylation and the differential transcriptional control of adjacent genes. *Development*, 140(21):4407–4416.

Bhatt, D. M., Pandya-Jones, A., Tong, A.-J., Barozzi, I., Lissner, M. M., Natoli, G., Black, D. L., and Smale, S. T. (2012). Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell*, 150(2):279–290.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F. A., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., Haussler, D., and Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome research*, 14(4):708–715.

Bourque, G. (2009). Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Current Opinion in Genetics & Development*, 19(6):607–612.

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. *Molecular and Cellular Biology*, 10(1):28–36.

Brockdorff, N. (2013). Noncoding RNA and Polycomb recruitment. *RNA (New York, N.Y.)*, 19(4):429–442.

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(3):515–526.

Brunkow, M. E. and Tilghman, S. M. (1991). Ectopic expression of the H19 gene in mice causes prenatal lethality. *Genes & Development*, 5(6):1092–1101.

Cabili, M. N., Dunagin, M. C., and McClanahan, P. D. (2015). Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome ...*.

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. 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. T., 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., Schönbach, C., Sekiguchi, K., Semple, C. A. M., 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., Hayashizaki, Y., FANTOM Consortium, and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) (2005). The transcriptional landscape of the mammalian genome. *Science*, 309(5740):1559–1563.

Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., Tramontano, A., and Bozzoni, I. (2011). A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell*, 147(2):358–369.

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.

Chen, B., Gilbert, L. A., Cimini, B. A., Schnitzbauer, J., Zhang, W., Li, G.-W., Park, J., Blackburn, E. H., Weissman, J. S., Qi, L. S., and Huang, B. (2013). Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System. *Cell*, 155(7):1479–1491.

Chen, L.-L. (2016). The biogenesis and emerging roles of circular RNAs. *Nature Publishing Group*.

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(2):404–416.

Chuong, E. B., Elde, N. C., and Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science*, 351(6277):1083–1087.

Cifuentes-Rojas, C., Hernandez, A. J., Sarma, K., and Lee, J. T. (2014). Regulatory interactions between RNA and polycomb repressive complex 2. *Molecular cell*, 55(2):171–185.

Claverie, J.-M. (2005). Fewer genes, more noncoding RNA. *Science*, 309(5740):1529–1530.

Comings, D. E. (1972). The structure and function of chromatin. *Advances in human genetics*, 3:237–431.

Da Sacco, L., Baldassarre, A., and Masotti, A. (2012). Bioinformatics tools and novel challenges in long non-coding RNAs (lncRNAs) functional analysis. *International journal of molecular sciences*, 13(1):97–114.

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. *Molecular cell*, 57(3):552–558.

Davidovich, C., Zheng, L., Goodrich, K. J., and Cech, T. R. (2013). Promiscuous RNA binding by Polycomb repressive complex 2. *Nature structural & molecular biology*, 20(11):1250–1257.

de Groot, R. P., Raaijmakers, J. A., Lammers, J. W., Jove, R., and Koenderman, L. (1999). STAT5 activation by BCR-Abl contributes to transformation of K562 leukemia cells. *Blood*, 94(3):1108–1112.

Ding, D.-Q., Okamasa, K., Yamane, M., Tsutsumi, C., Haraguchi, T., Yamamoto, M., and Hiraoka, Y. (2012). Meiosis-Specific Noncoding RNA Mediates Robust Pairing of Homologous Chromosomes in Meiosis. *Science*, 336(6082):732–736.

Dinger, M. E., Amaral, P. P., Mercer, T. R., Pang, K. C., Bruce, S. J., Gardiner, B. B., Askarian-Amiri, M. E., Ru, K., Soldà, G., Simons, C., Sunkin, S. M., Crowe, M. L., Grimmond, S. M., Perkins, A. C., and Mattick, J. S. (2008). Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome research*, 18(9):1433–1445.

Dye, M. J., Gromak, N., and Proudfoot, N. J. (2006). Exon tethering in transcription by RNA polymerase II. *Molecular cell*, 21(6):849–859.

Eddy, S. R. (2001). Non—[ndash]—coding RNA genes and the modern RNA world. *Nature Reviews Genetics*, 2(12):919–929.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter,

N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C. J., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Flamm, C., Fried, C., Hackermüller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K. G., Sung, W.-K., Ooi, H. S., Chiu, K. P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M. L., Valencia, A., Choo, S. W., Choo, C. Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T. G., Brown, J. B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C. N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J. S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R. M., Rogers, J., Stadler, P. F., Lowe, T. M., Wei, C.-L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S. E., Fu, Y., Green, E. D., Karaöz, U., Siepel, A., Taylor, J., Liefer, L. A., Wetterstrand, K. A., Good, P. J., Feingold, E. A., Guyer, M. S., Cooper, G. M., Asimenos, G., Dewey, C. N., Hou, M., Nikolaev, S., Montoya-Burgos, J. I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N. R., Holmes, I., Mullikin, J. C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. A., NISC Comparative Sequencing Program, Baylor College of Medicine Human Genome Sequencing Center, Washington University Genome Sequencing Center, Broad Institute, Children's Hospital Oakland Research Institute, Batzoglou, S., Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Sidow, A., Trinklein, N. D., Zhang, Z. D., Barrera, L., Stuart, R., King, D. C., Ameur, A., Enroth, S., Bieda, M. C., Kim, J., Bhinge, A. A., Jiang, N., Liu, J., Yao, F., Vega, V. B., Lee, C. W. H., Ng, P., Shahab, A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer, M. A., Richmond, T. A., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J. C., Couttet, P., Bruce, A. W., Dovey, O. M., Ellis, P. D., Langford, C. F., Nix, D. A., Euskirchen, G., Hartman, S., Urban, A. E., and Kra... (2007). Identification and analysis of functional elements in 1human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., and Bernstein, B. E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49.

Fickett, J. W. and Tung, C.-S. (1992). Assessment of protein coding measures. *Nucleic acids research*, 20(24):6441–6450.

Filipowicz, W., Bhattacharyya, S. N., and Sonenberg, N. (2008). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nature Reviews Genetics*, 9(2):102–114.

Gagen, M. J. and Mattick, J. S. (2005). Inherent size constraints on prokaryote gene networks due to "accelerating" growth. *Theory in biosciences = Theorie in den Biowissenschaften*, 123(4):381–411.

Gilbert, L. A., Horlbeck, M. A., Adamson, B., Villalta, J. E., Chen, Y., Whitehead, E. H., Guimaraes, C., Panning, B., Ploegh, H. L., Bassik, M. C., Qi, L. S., Kampmann, M., and Weissman, J. S. (2014). Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell*, 159(3):647–661.

Gilbert, L. A., Larson, M. H., Morsut, L., Liu, Z., Brar, G. A., Torres, S. E., Stern-Ginossar, N., Brandman, O., Whitehead, E. H., Doudna, J. A., Lim, W. A., Weissman, J. S., and Qi, L. S. (2013). CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell*, 154(2):442–451.

Ginder, G. D. (2015). Epigenetic regulation of fetal globin gene expression in adult erythroid cells. *Translational Research*, 165(1):115–125.

Giordano, J., Ge, Y., Gelfand, Y., Abrusán, G., Benson, G., and Warburton, P. E. (2007). Evolutionary History of Mammalian Transposons Determined by Genome-Wide Defragmentation. *PLoS Computational Biology*, 3(7):e137–14.

Grote, P., Wittler, L., Hendrix, D., Koch, F., Währisch, S., Beisaw, A., Macura, K., Bläss, 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. *Developmental cell*, 24(2):206–214.

Guil, S., Soler, M., Portela, A., Carrère, J., Fonalleras, E., Gómez, A., Villanueva, A., and Esteller, M. (2012). Intronic RNAs mediate EZH2 regulation of epigenetic targets. *Nature structural & molecular biology*, 19(7):664–670.

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(7235):223–227.

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(1):240–251.

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. *Nature Publishing Group*, 21(2):198–206.

Higgs, P. G. (1998). Compensatory neutral mutations and the evolution of RNA. *Genetica*, 102-103(1-6):91–101.

Huarte, M. (2015). The emerging role of lncRNAs in cancer. *Nature Medicine*, 21(11):1253–1261.

Huarte, M., Guttman, M., Feldser, D., Garber, M., and Koziol, M. J. (2010). A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell*.

Hung, T., Wang, Y., Lin, M. F., Koegel, A. K., Kotake, Y., Grant, G. D., Horlings, H. M., Shah, N., Umbricht, C., Wang, P., Wang, Y., Kong, B., Langerød, A., Børresen-Dale, A.-L., Kim, S. K., van de Vijver, M., Sukumar, S., Whitfield, M. L., Kellis, M., Xiong, Y., Wong, D. J., and Chang, H. Y. (2011). Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nature genetics*, 43(7):621–629.

Hutvágner, G., McLachlan, J., Pasquinelli, A. E., Bálint, E., Tuschl, T., and Zamore, P. D. (2001). A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science*, 293(5531):834–838.

Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T. R., Prensner, J. R., Evans, J. R., Zhao, S., Poliakov, A., Cao, X., Dhanasekaran, S. M., Wu, Y.-M., Robinson, D. R., Beer, D. G., Feng, F. Y., Iyer, H. K., and Chinnaiyan, A. M. (2015). The landscape of long noncoding RNAs in the human transcriptome. *Nature genetics*, 47(3):199–208.

Jonkers, I. and Lis, J. T. (2015). Getting up to speed with transcription elongation by RNA polymerase II. *Nature Publishing Group*, 16(3):167–177.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, 110(1-4):462–467.

Kallen, A. N., Zhou, X.-B., Xu, J., Qiao, C., Ma, J., Yan, L., Lu, L., Liu, C., Yi, J.-S., Zhang, H., Min, W., Bennett, A. M., Gregory, R. I., Ding, Y., and Huang, Y. (2013). The Imprinted H19 LncRNA Antagonizes Let-7 MicroRNAs. *Molecular cell*, 52(1):101–112.

Kaneko, S., Son, J., Shen, S. S., Reinberg, D., and Bonasio, R. (2013). PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. *Nature structural & molecular biology*, 20(11):1258–1264.

Kanhere, A., Viiri, K., Araújo, C. C., Rasaiyaah, J., Bouwman, R. D., Whyte, W. A., Pereira, C. F., Brookes, E., Walker, K., Bell, G. W., Pombo, A., Fisher, A. G., Young, R. A., and Jenner, R. G. (2010). Short RNAs Are Transcribed from Repressed Polycomb Target Genes and Interact with Polycomb Repressive Complex-2. *Molecular cell*, 38(5):675–688.

Kapusta, A., Kronenberg, Z., Lynch, V. J., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M., and Feschotte, C. (2013). Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genetics*, 9(4):e1003470.

Kelley, D. and Rinn, J. (2012). Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome biology*, 13(11):R107.

Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. (2003). Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences*, 100(20):11484–11489.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome research*, 12(6):996–1006.

Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B. E., van Oudenaarden, A., Regev, A., Lander, E. S., and Rinn, J. L. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 106(28):11667–11672.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4):R36.

Kino, T., Hurt, D. E., Ichijo, T., Nader, N., and Chrousos, G. P. (2010). Noncoding RNA Gas5 Is a Growth Arrest and Starvation-Associated Repressor of the Glucocorticoid Receptor. *Science signaling*, 3(107):ra8–ra8.

Kretz, M., Siprashvili, Z., Chu, C., Webster, D. E., Zehnder, A., Qu, K., Lee, C. S., Flockhart, R. J., Groff, A. F., Chow, J., Johnston, D., Kim, G. E., Spitale, R. C., Flynn, R. A., Zheng, G. X. Y., Aiyer, S., Raj, A., Rinn, J. L., Chang, H. Y., and Khavari, P. A. (2013). Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature*, 493(7431):231–235.

Kung, J. T. Y., Colognori, D., and Lee, J. T. (2013). Long noncoding RNAs: past, present, and future. *Genetics*, 193(3):651–669.

Lai, F., Gardini, A., Zhang, A., and Shiekhattar, R. (2015). Integrator mediates the biogenesis of enhancer RNAs. *Nature*, 525(7569):399–403.

Lai, F., Orom, U. A., Cesaroni, M., Beringer, M., Taatjes, D. J., Blobel, G. A., and Shiekhattar, R. (2013). Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature*, 494(7438):497–501.

Lainé, J.-P., Singh, B. N., Krishnamurthy, S., and Hampsey, M. (2009). A physiological role for gene loops in yeast. *Genes & Development*, 23(22):2604–2609.

Lam, M. T. Y., Li, W., Rosenfeld, M. G., and Glass, C. K. (2014). Enhancer RNAs and regulated transcriptional programs. *Trends in biochemical sciences*, 39(4):170–182.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., and International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Lau, N. C., Lim, L. P., Weinstein, E. G., and Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*, 294(5543):858–862.

Lee, R. C. and Ambros, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. *Science*, 294(5543):862–864.

Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854.

Leighton, P. A., Ingram, R. S., Eggenschwiler, J., Efstratiadis, A., and Tilghman, S. M. (1995). Disruption of imprinting caused by deletion of the H19 gene region in mice. *Nature*, 375(6526):34–39.

Li, J., Wang, G., Wang, C., Zhao, Y., Zhang, H., Tan, Z., Song, Z., Ding, M., and Deng, H. (2007). MEK/ERK signaling contributes to the maintenance of human embryonic stem cell self-renewal. *Differentiation*, 75(4):299–307.

Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A. Y., Merkurjev, D., Zhang, J., Ohgi, K., Song, X., Oh, S., Kim, H.-S., Glass, C. K., and Rosenfeld, M. G. (2013). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature*, 498(7455):516–520.

Lynch, V. J., Leclerc, R. D., May, G., and Wagner, G. P. (2011). Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nature genetics*, 43(11):1154–1159.

Lynch, V. J., Nnamani, M. C., Kapusta, A., Brayer, K., Plaza, S. L., Mazur, E. C., Emera, D., Sheikh, S. Z., Grützner, F., Bauersachs, S., Graf, A., Young, S. L., Lieb, J. D., DeMayo, F. J., Feschotte, C., and Wagner, G. P. (2015). Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. *Cell reports*, 10(4):551–561.

Mancini-Dinardo, D., Steele, S. J. S., 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 & Development*, 20(10):1268–1282.

Marahrens, Y., Panning, B., Dausman, J., Strauss, W., and Jaenisch, R. (1997). Xist-deficient mice are defective in dosage compensation but not spermatogenesis. *Genes & Development*, 11(2):156–166.

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(7128):666–670.

Mattick, J. S. (2004). RNA regulation: a new genetics? *Nature Reviews Genetics*, 5(4):316–323.

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(7551):232–236.

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, 28(5):nbt.1630–9.

Mili, S. and Steitz, J. A. (2004). Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses. *RNA (New York, N.Y.)*, 10(11):1692–1694.

Mohammad, F., Mondal, T., Guseva, N., Pandey, G. K., and Kanduri, C. (2010). Kcnq1ot1 noncoding RNA mediates transcriptional gene silencing by interacting with Dnmt1. *Development*, 137(15):2493–2499.

Na, J., Furue, M. K., and Andrews, P. W. (2010). Inhibition of ERK1/2 prevents neural and mesendodermal differentiation and promotes human embryonic stem cell self-renewal. *Stem Cell Research*, 5(2):157–169.

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(5908):1717–1720.

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J. C., Grützner, F., and Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, 505(7485):635–640.

Ohno, S. (1972). So much "junk" DNA in our genome. *Brookhaven symposia in biology*, 23:366–370.

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., Schönbach, 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. M., 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(6915):563–573.

Ørom, U. A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., Guigó, R., and Shiekhattar, R. (2010). Long noncoding RNAs with enhancer-like function in human cells. *Cell*, 143(1):46–58.

O'Sullivan, J. M., Tan-Wong, S. M., Morillon, A., Lee, B., Coles, J., Mellor, J., and Proudfoot, N. J. (2004). Gene loops juxtapose promoters and terminators in yeast. *Nature genetics*, 36(9):1014–1018.

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. *Molecular cell*, 32(2):232–246.

Pang, K. C., Frith, M. C., and Mattick, J. S. (2006). Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends in Genetics*, 22(1):1–5.

Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S., and Brockdorff, N. (1996). Requirement for Xist in X chromosome inactivation. *Nature*, 379(6561):131–137.

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(5616):131–135.

Ponting, C. P., Oliver, P. L., and Reik, W. (2009). Evolution and Functions of Long Noncoding RNAs. *Cell*, 136(4):629–641.

Prensner, J. R., Iyer, M. K., Sahu, A., Asangani, I. A., Cao, Q., Patel, L., Vergara, I. A., Davicioni, E., Erho, N., Ghadessi, M., Jenkins, R. B., Triche, T. J., Malik, R., Bedenis, R., McGregor, N., Ma, T., Chen, W., Han, S., Jing, X., Cao, X., Wang, X., Chandler, B., Yan, W., Siddiqui, J., Kunju, L. P., Dhanasekaran, S. M., Pienta, K. J., Feng, F. Y., and Chinnaiyan, A. M. (2013). The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. *Nature genetics*, 45(11):1392–1398.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6):841–842.

Quinn, J. J., Zhang, Q. C., Georgiev, P., Ilik, I. A., Akhtar, A., and Chang, H. Y. (2016). Rapid evolutionary turnover underlies conserved lncRNA-genome interactions. *Genes & Development*, 30(2):191–207.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470(7333):279–283.

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., and Aiden, E. L. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, 159(7):1665–1680.

Ray, D., Kazan, H., Cook, K. B., Weirauch, M. T., Najafabadi, H. S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., Na, H., Irimia, M., Matzat, L. H., Dale, R. K., Smith, S. A., Yarosh, C. A., Kelly, S. M., Nabet, B., Mecenas, D., Li, W., Laishram, R. S., Qiao, M., Lipshitz, H. D., Piano, F., Corbett, A. H., Carstens, R. P., Frey, B. J., Anderson, R. A., Lynch, K. W., Penalva, L. O. F., Lei, E. P., Fraser, A. G., Blencowe, B. J., Morris, Q. D., and Hughes, T. R. (2013). A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177.

Redon, S., Reichenbach, P., and Lingner, J. (2010). The non-coding RNA TERRA is a natural ligand and direct inhibitor of human telomerase. *Nucleic acids research*, 38(17):gkq296–5806.

Rinn, J. L. and Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. *Annual review of biochemistry*, 81(1):145–166.

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(7):1311–1323.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26.

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.

Sempere, L. F., Freemantle, S., Pitha-Rowe, I., Moss, E., Dmitrovsky, E., and Ambros, V. (2004). Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation. *Genome biology*, 5(3):R13.

Simon, J. A. and Kingston, R. E. (2009). Mechanisms of Polycomb gene silencing: knowns and unknowns. *Nature reviews Molecular cell biology*, 10(10):697–708.

Sleutels, F., Zwart, R., and Barlow, D. P. (2002). The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature*, 415(6873):810–813.

Speir, M. L., Zweig, A. S., Rosenbloom, K. R., Raney, B. J., Paten, B., Nejad, P., Lee, B. T., Learned, K., Karolchik, D., Hinrichs, A. S., Heitner, S., Harte, R. A., Haeussler, M., Guruvadoo, L., Fujita, P. A., Eisenhart, C., Diekhans, M., Clawson, H., Casper, J., Barber, G. P., Haussler, D., Kuhn, R. M., and Kent, W. J. (2016). The UCSC Genome Browser database: 2016 update. *Nucleic acids research*, 44(D1):D717–25.

Sunwoo, H., Wu, J. Y., and Lee, J. T. (2015). The Xist RNA-PRC2 complex at 20-nm resolution reveals a low Xist stoichiometry and suggests a hit-and-run mechanism in mouse cells. *Proceedings of the National Academy of Sciences*, 112(31):E4216–E4225.

Taft, R. J. and Mattick, J. S. (2004). Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences. *Genome biology*.

Tan-Wong, S. M., Wijayatilake, H. D., and Proudfoot, N. J. (2009). Gene loops function to maintain transcriptional memory through interaction with the nuclear pore complex. *Genes & Development*, 23(22):2610–2624.

Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2):178–192.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, 7(3):562–578.

Tsai, M.-C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J. K., Lan, F., Shi, Y., Segal, E., and Chang, H. Y. (2010). Long noncoding RNA as modular scaffold of histone modification complexes. *Science*, 329(5992):689–693.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi,

K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.

Vilborg, A., Passarelli, M. C., Yario, T. A., Tycowski, K. T., and Steitz, J. A. (2015). Widespread Inducible Transcription Downstream of Human Genes. *Molecular cell*, 59(3):449–461.

Wang, J., Zhang, J., Zheng, H., Li, J., Liu, D., Li, H., Samudrala, R., Yu, J., and Wong, G. K.-S. (2004). Mouse transcriptome: Neutral evolution of —[lsquo]—non-coding—[rsquo]— complementary DNAs. *Nature*, 431(7010).

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(7341):120–124.

Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J.-P., and Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic acids research*, 41(6):e74–e74.

Werner, M. S. and Ruthenburg, A. J. (2015). Nuclear Fractionation Reveals Thousands of Chromatin-Tethered Noncoding RNAs Adjacent to Active Genes. *Cell reports*, 12(7):1089–1098.

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(5):855–862.

Wuarin, J. and Schibler, U. (1994). Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. *Molecular and Cellular Biology*, 14(11):7219–7225.

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(6652):745–749.

Xiang, J.-F., Yin, Q.-F., Chen, T., Zhang, Y., Zhang, X.-O., Wu, Z., Zhang, S., Wang, H.-B., Ge, J., Lu, X., Yang, L., and Chen, L.-L. (2014). Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus. *Cell Research*, 24(5):513–531.

Yang, L., Lin, C., Jin, C., Yang, J. C., Tanasa, B., Li, W., Merkurjev, D., Ohgi, K. A., Meng, D., Zhang, J., Evans, C. P., and Rosenfeld, M. G. (2013). lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs. *Nature*, 500(7464):598–602.

Yang, Y. W., Flynn, R. A., Chen, Y., Qu, K., Wan, B., Wang, K. C., Lei, M., and Chang, H. Y. (2014). Essential role of lncRNA binding for WDR5 maintenance of active chromatin and embryonic stem cell pluripotency. *eLife*, 3:e02046.

Zentner, G. E., Tesar, P. J., and Scacheri, P. C. (2011). Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome research*, 21(8):1273–1283.

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. *Molecular cell*, 40(6):939–953.

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(5902):750–756.