Modulation of lineage-specific cell differentiation by long non-coding RNAs

Juan Alvarez
B.A. Molecular Biology
Princeton University, 2009

Submitted to the Department of Biology
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2015

Author…………………………………………………………………………………………

Department of Biology
January 13, 2015

Certified by……………………………………………………………………………………

Harvey F. Lodish, PhD
Professor of Biology
Thesis Supervisor

Certified by……………………………………………………………………………………

Alexander van Oudenaarden, PhD
Professor of Biology
Thesis Supervisor

Accepted by…………………………………………………………...………………………………

Michael Hemann, PhD
Professor of Biology
Graduate Committee Co-Chair

Modulation of lineage-specific cell differentiation by long non-coding RNAs

Juan Alvarez

Abstract

Mammalian genomes comprise thousands of non-protein-coding genes. These can produce small non-coding RNAs (such as rRNAs and tRNAs), as well as long non-coding RNAs (lncRNAs), which are >200nt and resemble mRNAs in their biogenesis. Although the functions of the vast majority of lncRNAs remain unknown, many are tissue- and developmental stage-specific, suggesting roles in lineage-specific development.

We generated deep transcriptome surveys from differentiating mouse red blood cells, and implemented a computational strategy for *de novo* lncRNA discovery to comprehensively catalog erythroid-expressed lncRNAs. We found >100 previously unannotated loci, many of which are erythroid-specific and are induced by key erythroid transcription factors during differentiation. We exploited these features to select 12 candidates for loss-of-function studies, and found that depleting 10 out of 12 impaired red cell maturation, inhibiting cell size reduction and subsequent enucleation.

To study how lncRNAs regulate erythropoiesis, we focused on EC6, an unpolyadenylated lncRNA needed for silencing neighboring loci encoding NF-kB activators. De-repression of these genes upon EC6 knockdown leads to activation of NF-kB and other immune pathways that antagonize erythropoiesis, resulting in impaired proliferation and elevated apoptosis during differentiation. We showed that EC6 is retained in chromatin and binds the nuclear matrix factor hnRNP U, which may enable co-localization with its targets to mediate their repression.

Extending our work to a different lineage, we reconstructed transcriptomes from distinct mouse adipose tissues and identified ~1500 lncRNAs. These included many brown fat-specific loci induced during differentiation which are targets of key adipogenic factors. Inhibiting one of them, lnc-BATE1, compomised brown adipocyte development, impairing activation of brown fat genes, mitochondrial biogenesis, and thermogenic function. We showed that lnc-BATE1 acts *in trans* and binds hnRNP U, which is also required for proper brown adipocyte maturation.

This work demonstrates that lncRNAs modulate lineage-specific cell differentiation by promoting or suppressing competing gene expression programs controlling cell fate.

Thesis Supervisors: Harvey F. Lodish, Alexander van Oudenaarden
Titles: Professors of Biology

Acknowledgements

The wonderful, exhilarating, and by all means extraordinary quest that this thesis embodies would not have been possible without the guidance, encouragement and remarkable assistance that I was fortunate to receive from a number of people. Profuse thanks are due:

To my outstanding undergraduate mentors, Jacques Fresco, David Botstein, and Amy Caudy, for stirring me into coming to MIT and for continuing to provide me with invaluable advice, critical perspective, and unconditional friendship to this day.

To the MIT Biology graduate program, for investing on my education and allowing me to pursue my interests with minimal constraints, and to the graduate committee, for supporting me each step along my journey. I am particularly indebted to our wonderful administrative and support staff, and especially to Betsey Walsh, for indispensable guidance in all imaginable ways.

To my colleagues Bernardo Pando, Miaoqing Fang, Eric Wang, Jason Merkin, and Joshua Arribere, for encouraging me to find my own voice in research, and to Chris Burge, and Wendy Gilbert, for enlisting me in a number of great research adventures during my rotations.

To my advisor Alexander van Oudenaarden, for welcoming me into a new home with a unique lab environment full of tremendously talented and supportive colleagues who doubled as true a-list mentors and collaborators. Such exceptional team was of course nucleated by Alexander's infectious enthusiasm for quantitative, rigorous, and elegant systems biology research. Through his unwavering support and by truly leading by example while offering side-by-side, soft-spoken guidance, Alexander's mentorship greatly enabled me to become a more mature and seasoned researcher.

To Dong hyun Kim, Lenny Teytelman and Gregor Neuert, for taking the time to provide me with vital training in experimental and computational biology, as well as with detailed knowledge about the scientific literature, and for encouraging me -both directly and via their example- to be persistent and pursue my research interests even in the face of repeated setbacks.

To Stefan Semrau and Nikolai Slavov, with whom I collaborated in more projects that I can possibly count, always inspired by their boundless dedication, creativity and passion for research. To them, but also to Christoph Engert, for their frank friendship, and of course for their complicity in organizing those parties that made our names known around town.

To all the other members of the van Oudenaarden lab, especially Sandy Klemm, for critical research guidance and discussions, and to them, as well as to Monica Wolf, for making the lab such a livable, fun place to be.

To my second advisor, Harvey Lodish, who offered me a new lab home over an elevator ride one fateful morning after class. Over the years I have learned that one can hardly speak of Harvey without recurring to superlative praise. His well-earned wisdom, deep-rooted sense of practicality and inconceivable encyclopedic knowledge exist only to balance his tremendous sensibility, unparalleled intuition and astonishing generosity, the witnessing of which has been one of those life-changing experiences shaping my graduate years and beyond. I am eternally indebted to Harvey's for his all-around faith in my abilities and for the kind of support, encouragement and canny advice that one hopes to count on for a lifetime.

To Wenqian Hu, who in many ways has been my third advisor, for essentially taking me under his wing and including me on important projects where I could contribute or greatly benefit from. All this while treating me like an equal and fully trusting my judgment, research intuition and abilities. Our enduring research partnership has been most productive and has truly flourished into a lasting friendship for which I am most thankful.

To Dave Bartel, Laurie Boyer, and John Rinn for taking the time to be an incredible thesis committee. I have been most lucky to count on your valuable scientific advice, research intuition, and career guidance, which certainly turned me into a more rigorous and critical investigator.

To Jiahai Shi and Marko Knoll, for valuable training, critical advice on experiments, sharing of reagents and providing direct assistance with experiments, without which much of my experimental work would certainly not have thrived. For this I am most thankful, as well as for their wonderful mentorship and advice on life and research.

To Lei Sun and Bingbing Yuan, whose mentorship, research collaboration and frank friendship have been such a formative component of my graduate experience, one that transcends disciplines, working hours and even continents.

To all the other members of the Lodish lab, especially Hojun Li, for critical research guidance and discussions, and to them, as well as to Claire Mitrokostas and Mary Anne Donovan, for keeping the lab running and making it such a great working environment.

To the excellent staff at the Koch and Whitehead Institutes' flow cytometry, genome technology, bioinformatics, and information technology cores, especially Patti Wisniewski, George Bell, Stephen Goldman, and Jennifer Love, whose world-class support has been most indispensable to my research adventures.

To a number of MIT and visiting students, especially Alec Garza-Galindo, Staphany Park, and Austin Gromatzky, who so importantly contributed to my research throughout the years. I have been most fortunate to be able to help mentor such incredibly talented minds.

To my funding sources, which include the National Institutes of Health, National Heart, Lung, and Blood Institute, the National Institutes of Health, National Cancer Institute Physical Sciences in Oncology Initiative, and the National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, for making my research possible.

To my friends at MIT, especially Dave Hall, Bill Hesse, Leah Schmidt, Andy Yuan, and Sr. Moll, for helping me keep the work-life balance through their great friendship, unflinching support, and the most thought-provoking and stimulating late-night conversations, without which I could have never found such contentment while adapting to this town I call home now.

To my family, especially my mother, source of all my love, and my sister, eternal flame of my heart, for always believing in me and for their patience, support and courage in helping me pursue my dreams away from them over so many years now. I am incredibly grateful for their unconditional love and for their pushing me to constantly strive for improvement by always shooting for the stars.

To Katie Villa, my shining beacon of light and center of my life, without whose interminable support and love this work and life would not be possible, I am eternally indebted to your fufiness, and can hardly wait to start a family with you and walk the long and winding road by your side.

And finally, to Max, taken away from me so prematurely, yet to whom I owe the bulk of my scientific curiosity and passion for the life sciences, this and of course all my work is dedicated to you.

Table of Contents

List of Figures

Introduction

Chapter 1: Global discovery of erythroid long non-coding RNAs reveals novel regulators of red blood cell development

Chapter 2: Control of red blood cell development by the long non-coding RNA EC6

Chapter 3: De novo reconstruction of adipose tissue-specific transcriptomes reveals novel long non-coding RNA regulators of brown adipocyte development and physiology

Outlook and future directions

Introduction

Parts of this chapter were first published as:

Hu W, **Alvarez-Dominguez JR**, Lodish HF. 2012. Regulation of mammalian cell differentiation by long non-coding RNAs. *EMBO Rep* **13**: 971-983.

Alvarez-Dominguez JR, Hu W, Lodish HF. 2013. Regulation of Eukaryotic Cell Differentiation by Long Non-coding RNAs. in *Molecular Biology of Long Non-coding RNAs* (eds. AM Khalil, J Coller), pp. 15-67. Springer Science, New York.

Alvarez-Dominguez JR, Hu W, Gromatzky AA, Lodish HF. 2014a. Long noncoding RNAs during normal and malignant hematopoiesis. *International journal of hematology* **99**: 531-541.

The transfer of information from DNA to proteins is mediated by both RNA and protein components. Historically, our understanding of how these components function stems from a model proposed by Jacob and Monod over half a century ago (Jacob and Monod 1961). According to this model, *structural* genes are transcribed into mRNA that acts as a template for protein synthesis, and this process is controlled by the products of *regulator* genes. The biochemical nature of these regulatory products was unclear at the time, but evidence that these could be RNA or protein was widely discussed then. In the 40 years that followed, a dominant view emerged of proteins as the main regulators, partly facilitated by their ease of detection and experimental manipulation compared to RNA, which is less abundant and more unstable. However, recent improvements in our ability to sequence entire genomes and detect their RNA outputs increasingly suggested greater roles for RNA regulators than previously anticipated.

The initial sequencing of various eukaryotic genomes about a decade ago resulted in the surprising realization that the number of protein-coding genes does not appear to vary significantly across metazoans, despite significant differences in developmental complexity. In contrast, the proportion of non-coding DNA (including introns) does appear to increase, after accounting for varying ploidy, with developmental complexity (Mattick 2004; Taft et al. 2007). This led some to hypothesize that increasing amounts of RNA regulators, originating from some of these non-coding DNA regions, could have played a major role in giving rise to the diversity of cell differentiation programs that underlie development of multicellular organisms (Prasanth and Spector 2007; Amaral and Mattick 2008). Alternatively, greater proportions of non-coding DNA could merely reflect greater tolerance for non-functional DNA acquisition among increasingly complex organisms, which can result from small effective population sizes where there is little fitness cost to excess genomic load, as seen for metazoans and specially humans

(Charlesworth 2009; Palazzo and Gregory 2014). Distinguishing between these possibilities thus required evidence that non-coding DNA regions can be transcribed into functional species that act as RNA regulators.

Evidence that non-coding DNA regions are indeed transcribed accumulated over the past ten years as the focus shifted from sequencing genomes to cataloguing their transcriptomes. We now know that for every eukaryote examined the majority of the genome is capable of being transcribed, albeit across a wide range of expression levels (Kapranov et al. 2007b; Jacquier 2009; Djebali et al. 2012). Only a fraction of the RNA species detected could be recognized as associated with messenger RNA production, however, or as previously known classes of small non-coding RNAs (such as ribosomal, transfer, and splicing-associated RNAs) (Bertone et al. 2004; Carninci et al. 2005; Carninci et al. 2006; Birney et al. 2007; Kapranov et al. 2007a; Guttman et al. 2009). This raised the possibility that some of the newly identified transcribed regions may actually encode novel classes of functional RNAs.

Of the newly identified RNA species, those longer than 200 nucleotides that seem to have little to no protein-coding capacity have been termed long non-coding RNAs (Reviewed in Wilusz et al. 2009). lncRNAs typically resemble messenger RNAs in being capped, polyadenylated and spliced, and many are present at similar levels as mRNAs. Of all RNA classes, lncRNAs are among the least well-understood. Many of them are differentially expressed across tissues, developmental stages, and physiological states (Guttman et al. 2010; Cabili et al. 2011; Derrien et al. 2012). A few dozen have been functionally characterized in mammals, and several have been implicated in important processes such as X chromosome inactivation, imprinting, maintenance of pluripotency, lineage commitment, and apoptosis (Penny et al. 1996; Sleutels et al. 2002; Kino et al. 2010; Sheik Mohamed et al. 2010; Guttman et al. 2011). An emerging theme among known lncRNA functions has thus been the regulation of cell fate, which has lent support to the notion that distinct collections of lncRNAs help orchestrate the development of distinct tissues.

In this thesis, I present our contributions to the *de novo* identification and functional characterization of tissue-specific lncRNAs that modulate tissue-specific developmental programs. First, I describe our work on the global discovery and characterization of a comprehensive catalog of erythroid lncRNAs, demonstrating via loss-of-function studies that diverse types of lncRNAs are essential for the proper generation of mature red cells. Second, I describe the identification and characterization of adipose tissue-selective lncRNAs, establishing one of them as a novel regulator of brown adipocyte development and physiology. Third, I describe our work on elucidating the mechanisms by which two of these lncRNAs regulate cell differentiation. Finally, I discuss emerging principles of lncRNA function and evaluate how they provide a framework for the integration of lncRNAs into known regulatory networks of mammalian cell differentiation.

To provide context, I begin by providing a brief history of technical approaches for lncRNA discovery and characterization. Next, I discuss evidence supporting a role for lncRNAs as lineage-specific developmental regulators, highlighting select examples that illustrate advances in our understanding of how lncRNAs contribute to a diversity of cell differentiation processes. Finally, I summarize the specific contributions of this thesis toward discovery and functional characterization of lncRNA modulators of lineage-specific cell development.

Finding and identifying lncRNAs

About 30 years after the notion of messenger RNA was established, the first lncRNA to be identified as such was described in the context of mouse embryonic development (Pachnis et al. 1988; Brannan et al. 1990). H19 was identified as a product of RNA Polymerase II, enriched in fetal liver and in cardiac and skeletal muscle, which becomes strongly repressed after birth. H19 was capped and polyadenylated but contained no large open reading frame (ORF) for translation. Rather, it contained only small sporadic ORFs that were not evolutionary conserved, could not template translation *in vivo,* and did not produce detectable polypeptides. Shortly after, many more examples of this novel type of RNA were characterized in diverse eukaryotes, including Xist in mouse and human (Brockdorff et al. 1992; Brown et al. 1992), meiRNA in yeast (Watanabe and Yamamoto 1994), and roX1 in flies (Meller et al. 1997).

Over the following decade, the development of constantly improving technologies for transcriptome analysis propelled new efforts to detect and characterize lncRNAs at a global scale. The advent of entire genome sequences precipitated a number of collaborative efforts to survey their full transcriptional output (Tjaden et al. 2002; Yamada et al. 2003; Bertone et al. 2004; Stolc et al. 2004; Carninci et al. 2005; Stolc et al. 2005; David et al. 2006; Li et al. 2006; Birney et al. 2007; Nagalakshmi et al. 2008; Wilhelm et al. 2008). These efforts drove the rapid adaptation of classic expression profiling techniques into large-scale approaches of ever-increasing throughput, as occurred for CAGE (Cap analysis of Gene Expression) (Shiraki et al. 2003), microarrays (Selinger et al. 2000) and cDNA sequencing (Mortazavi et al. 2008).

A surprising outcome of surveying multiple eukaryotic transcriptomes, regardless of the technical approach, was that only a fraction of the detected transcripts could be recognized as protein-coding or as previously characterized classes of small non-coding RNA (such as rRNA, tRNA, snoRNA, microRNA or piRNA). This generated much excitement over the potential biological functions of thousands of newly discovered RNAs (Kapranov et al. 2007a; Amaral et al. 2008; Berretta and Morillon 2009; Jacquier 2009; Mercer et al. 2009). Since the number of uncharacterized loci was comparable to that of known protein-coding loci, it was also speculated that an increase in the number of the former along the eukaryotic lineage may explain large differences in developmental complexity among eukaryotes with otherwise comparable numbers of protein-coding genes and protein families (Mattick 2004; Prasanth and Spector 2007; Amaral et al. 2008). However, concerns that most newly discovered RNAs were mere by-products of transcription of known genes, or intergenic insertions of transposon or random sequence transcribed at background levels that are neutrally evolving and confer little or no selective advantage, were also raised (Brosius 2005; Struhl 2007; van Bakel et al. 2010). Expansion of such non-functional elements, in turn, could be due to accumulation of genomic insertions and deleterious alterations over the evolution of metazoans, which exhibit slow generation times (where excess genomic load may not hinder replication) and small effective population sizes (where small deleterious effects may not impact fitness) (Poole 2004; Gregory 2005; Palazzo and Gregory 2014).

Preliminary clues about the functional importance of newly discovered RNAs first emerged for those that were relatively abundant and longer than 200nt (putative lncRNAs), as they presented the clearer opportunity for detailed characterization. First, analysis of their global primary sequence conservation showed evidence of evolutionary constraint (Pheasant and Mattick 2007; Ponjavic et al. 2007). Second, expression profiling revealed that many exhibit dynamic and cell-type specific expression patterns during development (Blackshaw et al. 2004; Stolc et al. 2004; Inagaki et al. 2005; Ravasi et al. 2006; Dinger et al. 2008a). Third, individual lncRNA candidates were found to localize to specific subcellular structures (Brown et al. 1992;

Mercer et al. 2008b; Nagano et al. 2008; Clemson et al. 2009; Redrup et al. 2009; Sasaki et al. 2009; Sunwoo et al. 2009).

Considering that both the expression and conservation of putative lncRNAs were much poorer than those of protein-coding genes, however, doubts about their origin and biological relevance persisted. One technical concern was that, given the propensity of reverse transcriptase for spurious second-strand production during first-strand cDNA synthesis, catalogs of putative lncRNAs could be plagued by spurious antisense transcripts (Perocchi et al. 2007; Ozsolak and Milos 2011). Another important concern was that many of the newly identified RNAs were simply biological noise (Huttenhofer et al. 2005; Ponjavic et al. 2007; Struhl 2007), non-functional byproducts of the transcription of neighboring protein-coding or regulatory loci (including enhancers) (Struhl 2007; Ebisuya et al. 2008; De Santa et al. 2010; Kim et al. 2010). Clearly, additional evidence was needed to distinguish biologically relevant lncRNA candidates from technical or biological noise.

A strategy devised by Guttman and colleagues to address these issues was to focus only on intergenic regions showing evidence of stable expression, as assayed by a signature of chromatin marks associated with active Pol II transcription (Guttman et al. 2009). This signature consisted of a short stretch of H3K4me3, marking Pol II initiation, followed by a longer stretch of H3K36me3, marking the region of Pol II elongation. The strategy identified in 4 mouse cell types about 1500 intergenic lncRNA (lincRNA) loci that were 5kb or greater in length and did not overlap known protein-coding genes, microRNAs or endogenous siRNAs. Their products were polyadenylated and mainly multiexonic transcripts with little or no protein-coding potential and evidence of 5’ capping. This subset of mouse lncRNAs indeed showed higher expression and conservation than previous collections, and a number of them were putatively associated with various developmental processes through correlative analysis of tissue expression patterns. Extending the approach to human yielded about 1800 human lincRNAs (Khalil et al. 2009).

There were important limitations to this approach for comprehensive discovery of *bona fide* lncRNAs, however. Not all loci transcribed by Pol II are marked by this K4-K36 signature; a subsequent study in mouse found that ~25% of lincRNA and mRNA transcripts identified by RNA-seq alone are not (Guttman et al. 2010), and in human the number appears to be greater (Cabili et al. 2011). Conversely, not all regions with a detectable K4-K36 domain correspond to gene bodies; some correspond to actively transcribed enhancers (De Santa et al. 2010; Cabili et al. 2011). Moreover, close examination of lncRNA catalogs indicate that a substantial fraction of these transcripts originate from intragenic and intergenic enhancers (Cabili et al. 2011; Kowalczyk et al. 2012; Marques et al. 2013; Alvarez-Dominguez et al. 2014b). Moreover, it is possible that some lncRNAs are transcribed by RNA polymerase III (see (White 2011) for discussion) and thus may lack these particular chromatin marks.

Subsequent studies now employ a combination of strategies for the reliable identification of stably-expressed lncRNAs (Guttman et al. 2010; Cabili et al. 2011; Ulitsky et al. 2011; Derrien et al. 2012). Detection and assembly of *de novo* lncRNA transcript models is most frequently conducted by RNA-seq alone. Then, evidence of independent full-length transcriptional units is sought by augmenting these models with evidence of transcript boundaries from orthogonal approaches. For example, transcriptional start sites can be determined directly through CAGE analysis or inferred from H3K4 marks. Similarly, the 3’ ends can be mapped by poly(A)-position profiling or inferred by computational detection of motifs for poly(A) addition. The use of paired-end sequencing reads can also enable assessment of whether lncRNAs and adjacent protein-coding genes share the same primary transcript. Constantly improving combination strategies are thus being used to obtain increasingly reliable collections of lncRNA genes in various organisms, resulting in a rapidly growing number of lncRNAs with recognized functions (see http://www.lncrnadb.org (Amaral et al. 2011) for a comprehensive database).

Excluding functional protein-coding capacity

A distinctive feature of lncRNAs is that they lack functional protein-coding capacity. This is typically taken to mean that they do not produce stable proteins. To determine if this is true, the gold standard is to assess if polypeptides are produced from any open reading frame from lncRNA collections (Banfai et al. 2012; Slavoff et al. 2013). However, due to technical difficulties, such as the detection of putative low-abundance polypeptides, or the absence of corresponding antibodies, the coding capacity of a newly identified RNA transcript is usually determined indirectly by computational and biochemical approaches (see (Dinger et al. 2008b; Guttman and Rinn 2012) for review).

Computationally, large-scale evaluation of coding potential can be done by examining candidate transcripts for presence and conservation of ORFs, by looking for homology to known protein domains, and by scrutinizing putative ORFs for known biases in codon usage within a species or in frequency of codon substitution throughout evolution.

The presence of ORFs in a transcript is a necessary but not sufficient qualification for protein-coding capacity. A putative ORF may occur purely by chance in any stretch of sequence, with the probability scaling with sequence length (Dinger et al. 2008b). To distinguish functional from spurious ORFs, early large-scale studies employed an empirical ORF cutoff of 100aa, consistent with the observation that >95% of proteins annotated in public databases at the time were >100aa in length (Okazaki et al. 2002; Frith et al. 2006). This was problematic, however, for two main reasons. First, it misclassified *bona fide* lncRNAs known to be functional at the time. Indeed, human XIST and H19 do contain ORFs as long as 172 and 256 amino acids, respectively, but these are not evolutionary conserved and fail to template polypeptide synthesis *in vivo* (Brannan et al. 1990; Brockdorff et al. 1992). Such ORFs may be spurious or alternatively be a vestige of former coding capacity (Ponting et al. 2009); in the case of Xist, it is believed that the lncRNA evolved from genes that formerly encoded proteins (Duret et al. 2006). Second, an 100aa ORF cutoff misclassified known protein-coding genes producing extremely short yet functional peptides, such as the 11aa peptide-encoding *tarsal-less* gene (Galindo et al. 2007). We now know that functional short peptide-encoding genes are widespread throughout the eukaryotic lineage (see (Andrews and Rothnagel 2014) for review), even though their ORFs can be substantially smaller than those occurring by chance in long sequences.

Clearly, additional ORF features need to be evaluated to discriminate functional from spurious ORFs. Known protein-coding ORFs typically display organism-specific differences in the frequency of occurrence of synonymous codons. The absence of such codon usage bias from a putative ORF can thus be used to argue that it is unlikely to represent a canonical functional ORF. Evolutionary analysis can also be used to evaluate functional coding potential (see (Lin et al. 2008) for review). Coding regions are under purifying selection to retain synonymous over non-synonymous codon substitutions to preserve their function. Non-coding regions, in contrast, experience no such selection and thus typically exhibit similar frequencies of synonymous and non-synonymous substitutions. The absence of a codon substitution bias inferred from the multi-species alignment of a putative ORF sequence can thus be used as evidence against functional coding capacity (Lin et al. 2011). However, approaches based on evolutionary analysis may fail to identify newly evolved functional ORFs. To address this, methods that do not require cross-species comparisons should be considered (Dinger et al. 2008b), together with direct inspection of homology in known protein domain databases.

Collectively, integrative computational strategies can be a powerful way of testing the coding potential of large collections of lncRNA candidates. Those candidates that pass computational tests, however, ultimately require experimental support of their non-coding status, as some may represent exceptions that defy the assumptions of these tests.

Biochemically, non-coding status implies that candidate transcripts are not associated with actively translating ribosomes. This can be experimentally tested by examining if they are present on polysomes through polysome fractionation analysis (Warner et al. 1963). This approach employs sucrose density gradients and ultracentrifugation to fractionate cell lysates. RNA transcripts associated with ribosomes predominantly sediment with the greatest velocity, whereas non-ribosome-associated transcripts remain at the top of the gradient. Care should be taken when interpreting these outcomes, however. If a transcript remains at the top of the gradient, it can be either a non-coding transcript or a translationally repressed protein-coding one. Conversely, if a transcript sediments with a higher velocity through the gradient, it only implies that the transcript is associated with large particles, which can be ribosomes or other large complexes that sediment with equal velocity. Specific disruption of translation, such as treatment with compounds that inhibit translation elongation, is required to discriminate between these two possibilities.

An alternative approach to polysome fractionation analysis is ribosome footprint profiling. This method relies on deep sequencing of RNA fragments protected from RNase digestion to infer the occupancy and density of ribosomes along sequences at high resolution (Ingolia et al. 2009; Ingolia et al. 2011). However, fragments in ribosome profiling libraries may also result from RNase protection by other, similarly-sized RNP complexes or by stable RNA secondary structures. Clearly, additional features need to be evaluated to obtain true evidence of translation from ribosome footprint profiling data.

True RNase-protected footprints generated by translating 80S ribosomes should present a three-nucleotide periodicity, reflective of the triplet nature of the genetic code. This feature can be used to accurately identify, among sequences covered by ribosome profiling fragments, those that are actively translated in discrete reading frames (Michel et al. 2012; Bazzini et al. 2014). Translation by 80S ribosomes restricted to a single, discrete ORF should also yield RNase-protected fragments that drop sharply at the end of the ORF, as seen for messenger RNAs. Accordingly, metrics such as the ribosome release score or the disengagement score have been developed to evaluate restriction of fragments from ribosome profiling libraries to a discrete ORF without downstream enrichment (Chew et al. 2013; Guttman et al. 2013). These metrics have provided evidence that most lncRNAs in zebrafish and mouse that are covered by ribosome profiling fragments do not present clear termination evidence in any single ORF, distinguishing them from known mRNAs and from 3’UTRs (which are largely devoid of such fragments). However, the same distinction is drawn for most mRNA transcript leaders, which are not expected to produce stable proteins, suggesting that true but unproductive translation of multiple, continuous ORFs occurs both within lncRNAs and within mRNA transcript leaders.

Alternatively, the frame-unbiased footprints within these units may simply arise from non-ribosomal sources.

A recent strategy devised by Ingolia and colleagues to discriminate true ribosome footprints from non-ribosomal sources on individual transcripts was to compare their fragment size distribution with that cumulatively seen for known protein-coding sequences (which should mostly consist of true 80S footprints) (Ingolia et al. 2014). This approach revealed that footprints from most lncRNAs and from most mRNA transcript leaders are generated by translating 80S ribosomes, which is further supported by their co-purification with tagged 60S subunits, tri-nucleotide periodicity, consistency with early AUG initiation, and size change in response to elongation inhibitors. These footprints are not restricted to single, discrete ORFs, however, suggesting widespread non-canonical translation of overlapping reading frames within lncRNAs and mRNA transcript leaders. Such form of translation is not expected to yield stable proteins with adaptive cellular functions, consistent with the general lack of conserved ORFs in these units, but production of unstably folded peptides alone, or their necessary surveillance, may play biological roles. Alternatively, non-canonical translation could simply reflect biological noise, i.e. unproductive outcomes of background imprecision by the translation machinery.

Even if the reads from a ribosome profiling experiment are filtered for true translating 80s footprints, the experiment by itself cannot assess whether these ribosomes actively make stable peptides. Some functionally characterized lncRNAs, including H19 and GAS5, do associate with ribosomes as part of their processing, but do not produce stable proteins (Li et al. 1998; Smith and Steitz 1998). Thus, as with polysome fractionation, additional experimental evidence is needed to obtain true evidence of translation status.

The translation status of a transcript can be supported by its localization within cells, which can be determined by detection *in situ* (e.g. via RNA in situ fluorescence hybridization), or in cell homogenates corresponding to nuclear and cytoplasmic fractions. RNAs predominantly resident in the nucleus, such as Xist, are strong candidates to be non-coding, as translation occurs in the cytoplasm. One caveat of these studies, however, is that they only reveal the steady-state localization of the transcript. If the RNA is rapidly shuttling between nucleus and cytoplasm, or efficiently degraded only in one of the two compartments, information obtained from its steady-state localization may be misleading (see (Grunwald et al. 2011) for review).

It is worth noting that even if a polypeptide is in fact produced from an RNA transcript, this alone does not rule out its possible function as an RNA regulator. Examples of transcripts with dual functions as messenger and RNA regulator have indeed been described from bacteria to man (Chooniedass-Kothari et al. 2004; Kloc et al. 2005; Hube et al. 2006; Jenny et al. 2006; Wadler and Vanderpool 2007). Alternatively, a functional lncRNA may be translated into non-functional peptides if its gene has come under recent selective pressure to lose or gain functional coding capacity (Dinger et al. 2008b; Ulitsky et al. 2011). Importantly, evolutionary transitions between coding and non-coding status can be specific to a given species or phylogenetic lineage (Duret et al. 2006; Ulitsky et al. 2011), complicating the use of cross-species preservation of coding capacity as evidence against non-coding function.

In cases where a transcript's function is actually known, its functional coding capacity can be directly tested by using frame-shift mutations to disrupt putative ORFs and assessing whether function is compromised. If function is independent of putative ORFs, a strong claim can be made that functionally the transcript is indeed non-coding. Ultimately, combination strategies augmenting computational analyses with detailed biochemical experiments will be needed to convincingly determine whether or not the increasing number of putative lncRNAs identified by large-scale studies function or not as RNA regulators.

Characterizing lncRNA properties

lncRNAs likely comprise a variety of families with diverse properties and functions, much like protein-coding genes. Preliminary efforts to define these subclasses have mainly focused on the genomic positioning of lncRNA loci. Based on these criteria, lncRNAs can be classified as intergenic, antisense to protein-coding genes, or overlapping known non-coding regions (such as enhancers, introns of protein-coding genes, or known small ncRNA loci). Historically, efforts to characterize lncRNAs have mainly focused on intergenic ones, as they are easier to unambiguously identify and perturb than the other subclasses. Global characterization of other types of lncRNAs has thus been generally lagging.

Recent studies have laid conceptual frameworks for annotating the structural, conservation and expression features of diverse types of lncRNAs (Guttman et al. 2010; Cabili et al. 2011; Derrien et al. 2012). Structurally, lncRNAs have exons of comparable size to those of mRNAs but tend to have fewer of them, resulting in shorter transcript lengths and fewer isoforms. Splicing of lncRNAs occurs through canonical splice sites, but appears to be a less efficient process than for mRNAs (Derrien et al. 2012; Tilgner et al. 2012). At their termini, lncRNAs show clear evidence of 5’ capping and 3’ polyadenylation, but to a lower extent than mRNAs expressed at similar levels (Guttman et al. 2009; Derrien et al. 2012). These observations may be limited, however, by technical difficulties in accurately defining full transcriptional units and in retrieving reads spanning splice sites, due in part to the relatively short reads of current sequencing technologies or to assembly errors (Cabili et al. 2011; Ozsolak and Milos 2011).

Conservation analyses have revealed that, in general, intergenic lncRNAs show distinctive evidence of purifying selection in their primary sequence (Guttman et al. 2009; Khalil et al. 2009; Marques and Ponting 2009; Cabili et al. 2011; Ulitsky et al. 2011). However, while sequence conservation across lincRNA promoters is comparable with that of mRNAs, it is significantly lower across exons. This may be explained by the fact that lncRNA and protein-coding genes are subject to fundamentally distinct selective constraints. Protein-coding genes are under pressure to preserve the polypeptide information continuously encoded in their exons. lncRNA genes, however, may experience selective pressure to preserve secondary structure information, which can be more sparsely encoded than polypeptide information and thus tolerate greater sequence change (Washietl et al. 2005; Maenner et al. 2010; He et al. 2011; Parker et al. 2011; Schorderet and Duboule 2011; Novikova et al. 2012). Alternatively, selection may act to preserve only discontinuous, short regulatory sequences within lncRNA units (Duret et al. 2006; Marques and Ponting 2009), or simply to maintain their overall genomic position, span and orientation (Ponting et al. 2009; Cabili et al. 2011; Ulitsky et al. 2011). These considerations may confound assessment of purifying selection at lncRNA loci as well as identification of orthologs across species, especially if global vs. local sequence alignment methods are used.

Approaches that integrate conservation of secondary structure or of synteny to lncRNA discovery and characterization have been recently developed (Stanke et al. 2008; Gorodkin and Hofacker 2011; Ulitsky et al. 2011). It is worth pointing out that a few functionally characterized lincRNAs do show strong conservation of primary sequence from zebrafish to human (Guttman et al. 2009; Ponting et al. 2009; Sheik Mohamed et al. 2010; Ulitsky et al. 2011). Evolutionary conservation, however, may not be a requirement for functionality. Indeed, many functional lncRNAs appear rapidly evolving among eukaryotes, and a substantial number appear restricted to the primate lineage (Pollard et al. 2006; Amaral and Mattick 2008; Dinger et al. 2008a; Marques and Ponting 2009; Derrien et al. 2012). On the other hand, primary sequence conservation alone may not constitute sufficient evidence of RNA-based function, as it may simply reflect presence of ancestral regulatory or pseudogenized elements within the underlying DNA sequence.

In terms of expression, on average lncRNAs appear to be expressed at lower levels but in a more tissue- and cell type-specific manner than mRNAs (Guttman et al. 2010; Cabili et al. 2011; Derrien et al. 2012). The latter feature could confound the former, however, in studies that profile impure samples containing a mixture of diverse cell types, as transcripts restricted to a single rare cell type may escape detection even if they are highly expressed in that one cell type. Alternatively, low expression levels determined by ensemble measurements may simply reflect high cell-to-cell variability in the synthesis or degradation of highly expressed but short-lived transcripts, or they may reflect high expression during short time windows of the cell or metabolic cycles averaged over unsynchronized cell populations. Measuring expression within single cells or conducting bulk assays in homogenous cell populations should help address these issues.

Ultimately, lncRNAs may be better classified by their functions and mechanisms than by properties like genomic positioning, biogenesis or localization. However, this will require identification of lncRNA families with coherent themes of biological function and shared modes of action, which is simply intractable until enough examples of functionally characterized lncRNAs acting in similar ways in a variety of biological processes begin to accumulate.

lncRNAs as lineage-specific regulators

The seemingly exquisite spatial and temporal lncRNA expression patterns in metazoans suggest that some lncRNAs may function to help specify cell fate during development. Alternatively, such patterns may be a by-product of the tissue- and developmental stage-specific activity of neighboring protein-coding or non-coding regulatory units, or of entire chromosomal domains. Hence, experimental evidence in the form of targeted perturbations is needed to characterize the specific functions of lncRNAs during development.

Over the past two decades, such detailed perturbation studies have been undertaken for a few prototypical lncRNAs, implicating them in modulation of specific developmental processes. For example, Xist plays a well-characterized essential role in X-chromosome inactivation in female mammals via epigenetic silencing (see (Lee 2011) for review), and H19 regulates growth during embryogenesis via imprinting of the maternal Igf2 allele (see (Gabory et al. 2010) for review). For the vast majority of lncRNAs identified by recent large-scale studies, however, their potential roles in development remain to be experimentally determined.

Several interesting observations suggest that pursuing such studies may be worthwhile. First, the non-coding proportion of the transcriptome appears to increase with developmental complexity, suggesting that ncRNA regulators, and among them lncRNAs, may have contributed to the emergence of diverse gene expression programs underlying differentiation of specialized cell types during metazoan development (Mattick 2004; Prasanth and Spector 2007; Amaral and Mattick 2008; Mercer et al. 2009; Pauli et al. 2011). Second, given that lncRNAs as a class show greater tissue-specificity than protein-coding mRNAs, it seems conceivable that distinct collections of lncRNAs modulate the developmental programs of distinct tissues. Third, dysregulation of lncRNAs has been observed under many pathological conditions including cancer, heart disease and Alzheimer's disease (Reviewed in Wapinski and Chang 2011), suggesting that abnormal expression of some of these transcripts may contribute to the development of pathophysiological cellular states.

Importantly, recent studies have shown lncRNAs to be capable of regulating gene expression via diverse mechanisms (**Fig.1**). For example, lncRNAs can function as molecular scaffolds that recruit chromatin modifiers to target genes *in cis* or *in trans* and thereby modulate their expression (see (Schmitt and Paro 2006; Koziol and Rinn 2010) for review). In addition, lncRNAs can also modulate post-transcriptional events such as mRNA splicing (Tripathi et al. 2010), translation (Beltran et al. 2008; Carrieri et al. 2012; Yoon et al. 2012), and degradation (Gong and Maquat 2011). Furthermore, some lncRNAs can impair the function of specific microRNAs and thus indirectly enhance stability of the mRNAs normally downregulated by these miRNAs (Franco-Zorrilla et al. 2007; Cesana et al. 2011; Karreth et al. 2011; Salmena et al. 2011). Detailed mechanistic examples of how lncRNAs regulate gene expression have been summarized in recent reviews (Wang and Chang 2011; Guttman and Rinn 2012; Rinn and Chang 2012). Such regulatory capacities thus render lncRNAs as likely important players in the modulation of cell differentiation programs.

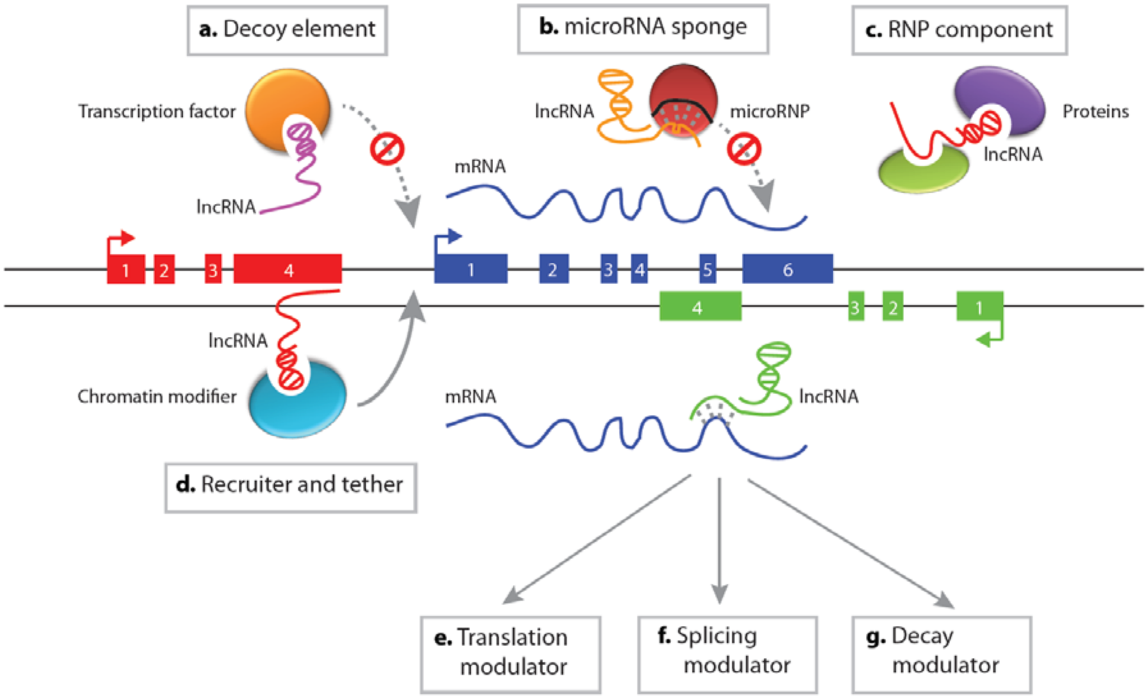


Figure 1. Mechanisms of lncRNA function.

(**A**) Some lncRNAs act as decoy elements to transcription factors, titrating them away from their DNA targets.

(**B**) Others work as decoys at the post-transcriptional level, titrating microRNA effector complexes away from their mRNA targets via target site mimics. These target site mimics lack sequence features needed for proper transcript degradation, having the net effect of 'sponging' microRNA effector complexes.

(**C**) Many lncRNAs bind specific combinations of proteins, such as chromatin modifiers or TFs, thus serving as scaffold elements to assemble specific RNP complexes.

(**D**) Recruitment and tethering of chromatin modifying complexes to their DNA targets *in cis* has also emerged as a well-characterized function for a number of lncRNAs. Not depicted is recruitment *in trans*.

A few lncRNAs appear to directly modulate post-transcriptional processing of their mRNA targets, including their translation (**E**), splicing (**F**) and decay (**G**).

Adapted from (Alvarez-Dominguez et al. 2013).

Over the past few years, growing numbers of loss-of-function and gain-of-function studies have greatly expanded the number of eukaryotic lncRNAs linked to cell differentiation processes (see (Wilusz et al. 2009; Hu et al. 2012; Fatica and Bozzoni 2014) for review). In multicellular eukaryotes these include, but are not limited to, progenitor cell self-renewal, apoptosis, and differentiation of pluripotent or lineage-specific progenitors during embryogenesis or during mature tissue homeostasis (**Fig.2**). In the next sections, we discuss select examples of lncRNAs implicated in the regulation of various cell differentiation processes, using as a guide the life cycle of multicellular organisms, from embryogenesis to adult tissue homeostasis. In particular, we focus on examples illustrating recent advances in our growing understanding of the mechanisms by which lncRNAs contribute to the development of cell lineages in mammals.

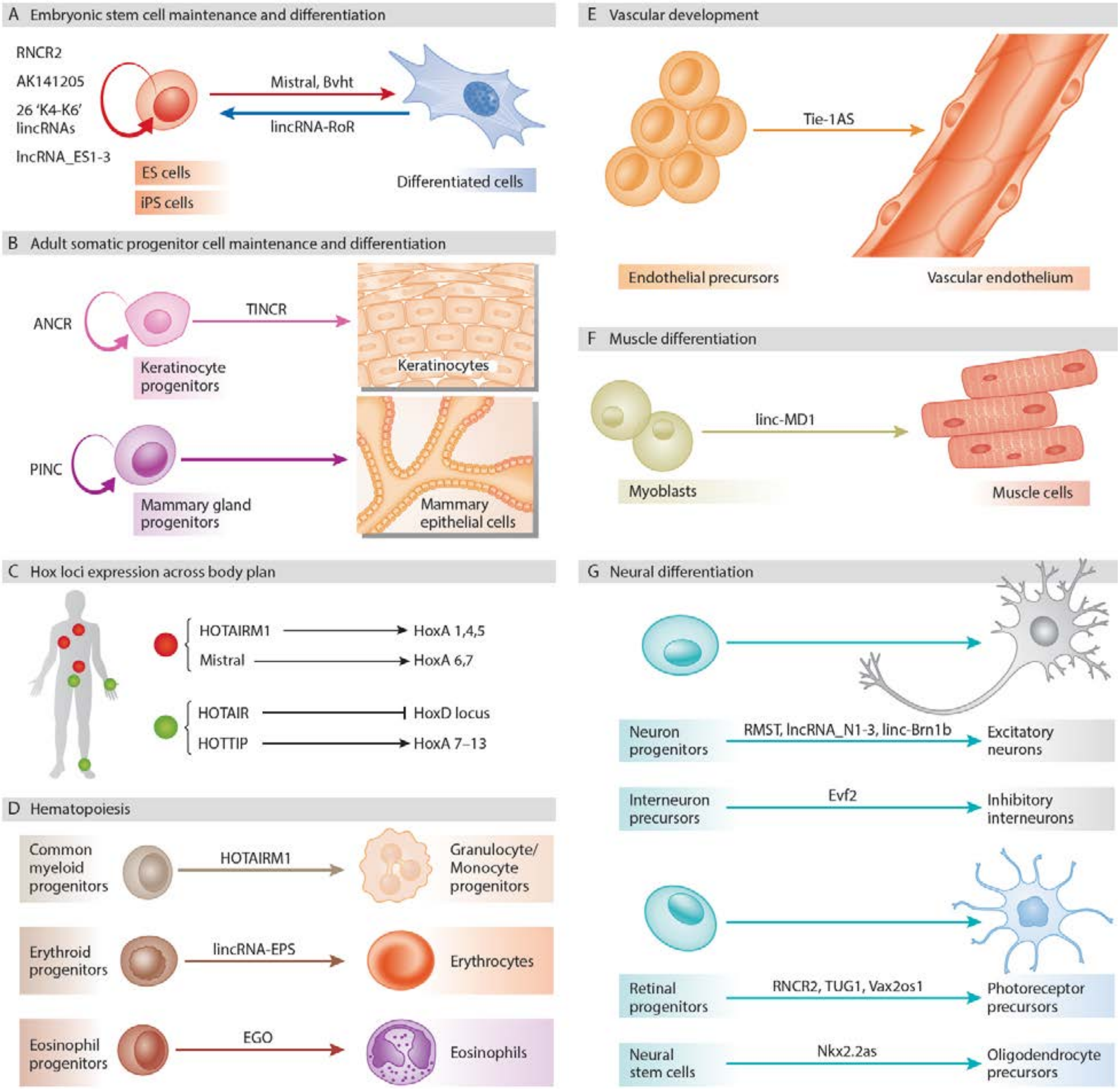


Figure 2. Regulation of mammalian cell differentiation by lncRNAs.

(**A**) Many lncRNAs are required for maintenance of embryonic stem cell pluripotency. Others favor differentiation into specialized lineages, while yet others contribute to dedifferentiation of specialized cells into iPS cells.

(**B**) lncRNAs are also important for the maintenance or differentiation of adult progenitor cells of the epidermal lineage.

(**C**) Several lncRNAs transcribed from Hox clusters regulate transcription of Hox genes, contributing to their distinct expression across cells from distinct anatomical positions.

(**D-F**) lncRNAs have also been associated with development of cells from the hematopoietic (D), vascular (E), and muscle (F) lineages.

(**G**) Many lncRNAs are differentially expressed and specifically localized across neural tissues. Many of them modulate differentiation of neural progenitors into excitatory, inhibitory or retinal photoreceptor neurons, whereas others promote oligodendrocyte differentiation.

Adapted from (Hu et al. 2012).

lncRNAs in embryonic stem cell maintenance and differentiation

The fusion of sex gametes in metazoans begins the process of embryogenesis, whereby an embryo is produced from the fertilized egg. The early stages of this process give rise to pluripotent embryonic cells, which have the developmental plasticity of differentiating into all derivatives of the primary germ layers (ectoderm, endoderm, and mesoderm). In culture these pluripotent cells can generate embryonic stem (ES) cells that also have the capacity to produce all the cell types in an organism through division and differentiation. Maintaining pluripotency of ES cells requires specific transcriptional regulation mediated by key transcription factors, such as Oct4, Sox2, and Nanog (see (Young 2011) for review). In addition to these protein regulators, lncRNAs are also involved in modulating ES cell fate.

In a study in mouse ES cells, Lipovich and coworkers focused on four highly conserved lncRNAs bound by Oct4 and Nanog (Sheik Mohamed et al. 2010). Inhibition or misexpression of two of these, RNCR2 and AK14205, caused exit from the pluripotent state as evidenced by loss of pluripotency markers, upregulation of lineage-specific ones, cell proliferation, and morphology. These effects were accompanied by altered levels of Oct4 and Nanog themselves, suggesting that lncRNAs act in the regulatory networks that control ES cell pluripotency.

This possibility was examined at a larger scale by a study focusing on 147 lincRNAs identified in mouse ES cells by the K4-K36 chromatin signature (Guttman et al. 2009; Guttman et al. 2011). For about 90% of the lincRNAs tested, inhibition by shRNAs resulted in significant changes in the ES cell gene expression program. Importantly, 26 of them were specifically implicated in the maintenance of the pluripotent state, as assayed after knockdown by loss of pluripotency markers and cell morphology. Another 30 lincRNAs were also implicated in repressing specific differentiation programs, although their loss of function alone was not sufficient to elicit differentiation. Importantly, expression of most of these lincRNAs is regulated by diverse combinations of ES cell-specific transcription factors, including Oct4, Sox2, Nanog and Klf4. Furthermore, many of these lincRNAs bind diverse combinations of chromatin regulatory proteins, potentially giving rise to specific RNP complexes.

Roles of lncRNAs in ES cell maintenance and differentiation appear conserved in human. Recent work focusing on differentiation of human ES cells into neurons identified three transcripts, lncRNA_ES1-3, that act in maintaining the pluripotent state (Ng et al. 2012). Knockdown of these lncRNAs by siRNA impairs pluripotency, as indicated by downregulation of pluripotency markers and upregulation of lineage markers. As with the 26 ‘K4-K36’ mouse lincRNAs, lncRNA_ES1-3 physically bind chromatin modifiers of the Polycomb group. Surprisingly, they also appear to bind the pluripotency-associated transcription factor Sox2, suggesting that lncRNAs may also act as scaffolds for combinations of chromatin modifiers and transcription factors.

Collectively, these results implicate lncRNAs in the regulatory networks maintaining ES cell identity, potentially by assembling regulatory complexes of chromatin modifiers and/or transcription factors. However, the coding capacity of all of these transcripts was only evaluated computationally, and subsequent studies have called into question the non-coding status of some of these genes (Ingolia et al. 2011). Thus, additional experimental evidence is still needed to verify that such loci function as non-coding RNA regulators.

Some lncRNAs are also involved in inducing ES cell pluripotency via reprogramming of somatic cells. Induced pluripotent stem (iPS) cells can be derived from terminally differentiated somatic cells by ectopic expression of key ES cell transcription factors such as Oct4, Nanog, Sox2, and c-Myc (see (Stadtfeld and Hochedlinger 2010) for review). Such cellular reprogramming is accompanied by extensive global remodeling of the epigenome (Hanna et al. 2010). Loewer et al. found that several lincRNAs contribute to this process of dedifferentiation (Loewer et al. 2010). Comparison of lincRNAs expressed in iPS cells versus those expressed in ES cells identified 10 that are specifically enriched in the former. These lincRNAs also appear regulated by the pluripotency-associated master transcription factors Oct4 and Nanog, suggesting a functional role in the generation of iPS cells. In particular, inhibition of one such lincRNA, lincRNA-RoR, leads to a 2- to 8-fold decrease in iPS colony formation. This effect appears to be mediated by impaired growth and elevated apoptosis via p53. Conversely, over-expression results in a ~ 2.5-fold increase in cellular reprogramming, a modest yet significant effect. These observations indicate that lncRNAs can modulate transcriptional programs associated with inducing or maintaining ES cell pluripotency, and that their impact on these processes can range from essential to subtle but detectable.

lncRNAs are also involved in modulating ES cell differentiation towards specific lineages. This can be induced by treatment with retinoic acid (RA), which results in downregulation of pluripotency markers and activation of lineage-specific ones. This process is mediated by epigenetic repressors, belonging to the Polycomb group, and by epigenetic activators, belonging to the Trithorax group. A component of the latter, the H3K4 methyltransferase MLL1, interacts with lncRNA Mistral during activation of lineage-associated gene expression (Bertani et al. 2011). Mistral is an unspliced and polyadenylated 798nt transcript upregulated during RA-induced ES cell differentiation. Knockdown of Mistral by siRNAs results in attenuated expression of key transcription factors that promote differentiation along the mesoderm lineage. This effect appears mediated by recruitment of the MLL1 epigenetic activator to these TF loci via direct physical interaction with its methyltransferase domain.

Another lncRNA, *Braveheart* (*Bvht*) is needed for differentiation of ES cell-derived mesoderm toward a cardiac cell fate (Klattenhoff et al. 2013). *Bvht* is an ES cell-expressed multiexonic ~590nt transcript that fails to template peptide synthesis *in vitro* and is selectively retained in the developing heart. Inhibition of *Bvht* by shRNAs is not required for self-renewal of ES cells, but instead impairs their capacity to differentiate into cardiac tissue capable of spontaneous contractility in culture. Accordingly, *Bvht* is required for maintaining the cardiac cell fate in cultured primary neonatal cardiomyocytes. Gene expression analysis revealed that *Bvht* acts *in trans* to promote activation of the core cardiovascular gene network. *Bvht* directly interacts with the repressive Polycomb group component SUZ12, and its loss of function is accompanied by persistence of H3K27me3 at the promoters of the critical genes in this core network. Importantly, epistasis experiments indicated that *Bvht* functions upstream of the master cardiac TF MesP1, suggesting that they function in the same pathway to direct cardiovascular cell fate commitment. Hence, epigenetic modulation of gene expression via lncRNA cofactors plays direct roles during both ES cell pluripotency and differentiation.

lncRNAs in the regulation of embryogenesis

Differentiation of proliferating ES cells into early embryos requires precise temporal and spatial execution of diverse gene expression programs. The capacity of lncRNAs to modulate expression of target genes predicts their involvement in executing these programs. Indeed, lncRNAs are essential to some of the earliest developmental programs during embryogenesis.

In order to equalize the dosage of X-linked genes between the sexes, early female mammalian embryos inactivate expression from one of the two copies of the X chromosome. This is achieved through epigenetic silencing of most of the chromosome mediated by a regulatory network of lncRNAs (see (Lee 2011) for review). The best characterized of these is Xist, a polyadenylated, nuclear transcript with multiple spliced isoforms that can reach ~18 to 19kb in length in mouse and human. Xist is exclusively expressed by the inactive X, from a region called the X inactivation center (Xic), and is required for its silencing. After being transcribed Xist remains tethered to the Xic, an effect mediated by the YY1 RNA/DNA binding protein (Jeon and Lee 2011). Tethered Xist in turn recruits, via a structured RNA domain, the PRC2 chromatin repressive complex (Zhao et al. 2008), which facilitates formation of heterochromatin via the histone modification H3K27me3. Xist and PRC2 co-migrate to spatially proximal gene-dense regions of active chromatin within the X chromosome, leading to their PRC2-mediated silencing (Plath et al. 2003; Zhao et al. 2008; Engreitz et al. 2013; Simon et al. 2013). These regions are repositioned into a growing heterochromatin compartment, effectively bringing new sites into close contact with the Xic for further proximity transfer of the Xist-PRC2 complex. By this mechanism, Xist spreads over a ~150 Mb scale to silence most of the genes in the inactive X chromosome. Thus, the Xist lncRNA is essential for epigenetic silencing of the X chromosome during mammalian embryogenesis. Attesting to its importance during development, paternally-inherited loss of Xist is lethal due to lack of X-inactivation in extra-embryonic tissues (Marahrens et al. 1997). Moreover, deregulation of Xist in females causes blood cancer (Yildirim et al. 2013) (see below).

Xist function is conserved in all placental mammals, despite limited global conservation of its primary sequence (Wutz 2011). Moreover, marsupial mammals appear to have independently evolved the same function through an unrelated lncRNA (Grant et al. 2012). Remarkably, as with mammals, flies also utilize lncRNA regulators for sex chromosome dosage compensation during embryogenesis (see (Conrad and Akhtar 2011) for review).

Several other lncRNAs modulate X inactivation through their regulation of Xist expression (Lee 2011). For example, Tsix is transcribed antisense to the Xist locus but in the active X, and its expression is anticorrelated with that of Xist. Transcription of Tsix leads to stable silencing of Xist *in cis* via recruitment of the DNA methyltransferase DNMT3A to the Xist promoter. Both Xist and Tsix are themselves regulated by lncRNAs Jpx and Xite, respectively. Jpx is required for Xist upregulation *in trans* at the inactive X, whereas Xite favors stable Tsix expression *in cis* at the active X. Importantly, loss of Tsix or of Jpx is female-lethal (Lee 2000; Tian et al. 2010). These examples illustrate how a cascade of lncRNA interactions helps establish epigenetic states that in turn specify and maintain developmental fate.

Imprinting to ensure monoallelic expression is another developmental process mediated by lncRNAs during embryogenesis (see (Barlow 2011) for review). This is typically achieved via epigenetic modification of promoter elements. For example, H19 controls embryonic imprinting of the maternal allele encoding the growth-regulator Igf2 (Gabory et al. 2010). H19 is a 2.3kb lncRNA transcribed by Pol II that undergoes capping and polyadenylation, can be found in both nucleus and cytoplasm, and contains only small ORFs that do not template protein synthesis *in vivo*. During embryogenesis, H19 is transcribed from the maternal allele and regulates growth via imprinting of Igf2 *in cis* and by serving as the primary transcript of miR-675, which downregulates the receptor of the Igf2 ligand expressed from the paternal allele (Igf1r) (Cai and Cullen 2007; Gabory et al. 2010; Keniry et al. 2012). Accordingly, deletion of the maternal H19 allele causes embryonic overgrowth due to increased Igf2 dosage (Leighton et al. 1995; Ripoche et al. 1997). Thus, H19 regulates growth during embryogenesis by controlling Igf2 dosage. In addition, H19 is reactivated in various cancers where it may influence tumor growth.

As with Xist, expression from the H19 locus is itself regulated by various other lncRNAs. Growth control is also regulated at the level of the Igf2 receptor (Igf2r), which is itself imprinted by another lncRNA, Air. Transcribed from the paternal allele in the second intron of the Igf2r locus, Air is essential for *cis*-imprinting of several genes on the paternal chromosome in a tissue-specific manner (Sleutels et al. 2002). Air acts *in cis* to silence the paternal allele of Igfr2 via transcriptional interference and also Slc22a3 and Slc22a2 via recruitment of the G9a histone methyltransferase (Sleutels et al. 2002; Sleutels et al. 2003; Nagano et al. 2008; Latos et al. 2012), Similarly, lncRNA Kcnq1ot1, a ~90kb transcript expressed from the paternal allele, directs epigenetic silencing of multiple neighboring genes (Mancini-Dinardo et al. 2006; Pandey et al. 2008). Kcnq1ot1 recruits both G9a and PRC2 to exert *in cis* repression of its targets, analogous to the mode of action of Xist, H19, and Air.

lncRNAs can also modulate differentiation of nascent mesoderm toward specialized cell fates during embryogenesis (Grote et al. 2013; Sauvageau et al. 2013). *Fendrr* is a nuclear ~2.4kb multiexonic lncRNA divergently transcribed ~1.25kb upstream of Foxf1 that is specifically expressed in the caudal end of the lateral plate mesoderm of mid-gestation embryos and remains enriched in the developing respiratory and digestive tracts and later in the mature adult lung. Loss of *Fendrr* by replacing its first exon with a strong transcriptional terminator causes heart and body wall defects that lead to early embryonic lethality. Importantly, these defects can be rescued by introduction of a functional *Fendrr* transgene, demonstrating RNA-based function. Genetic deletion of *Fendrr* exons 2-6 did not alter Mendelian ratios during early embryogenesis, on the other hand, but resulted in defects in lung and heart maturation that led to perinatal lethality due to respiratory failure. Such discrepancy in phenotypic outcomes may be due to the different genetic perturbation strategies used. In the case of the transcription terminator insertion strategy, it is possible that disruption of promoter-proximal regulatory elements affecting both Fendrr and divergent Foxf1 led to the earlier severe developmental phenotype, consistent with altered Foxf1 expression under this but not the genetic deletion strategy. Mechanistically, *Fendrr* binds to members of both the Polycomb and Trithorax histone modifying complexes. Accordingly, *Fendrr* loss of function is accompanied by altered promoter chromatin marks and altered expression of transcription factors controlling mesoderm differentiation. Thus, these findings highlight an essential role for *Fendrr* in directing proper differentiation of lateral plate mesoderm-derived tissues.

Collectively, the examples above illustrate how epigenetic control mediated by lncRNAs plays an essential role during embryo growth and development.

Regulation of Hox gene expression and body plan patterning by lncRNAs

Body plan patterning in developing metazoan embryos is regulated by the Hox family of genes (see (Mallo et al. 2010) for review). These genes encode transcription factors that regulate a variety of developmental loci by binding to their regulatory elements via a protein domain known as the homeodomain. The gene expression programs specified by these loci in turn determine the body plan during embryogenesis. Precise temporal and spatial expression of Hox genes and accurate maintenance of their expression patterns are thus essential for animal development and cell fate determination. Consequently, Hox genes are subject to intensive transcriptional and post-transcriptional regulation (Pearson et al. 2005; Yekta et al. 2008). In addition to transcription factors and microRNAs, the Hox gene clusters also encode hundreds of lncRNAs (Lipshitz et al. 1987; Rinn et al. 2007), many of which play important roles in modulating Hox gene expression.

Hox genes were first identified in the fruit fly *Drosophila melanogaster* through mutations affecting segmental identities along the posterior-anterior body plan (Lewis 1978). Characterization of the function and regulation of the full range of fly Hox genes over the next decade led to the discovery of both the Polycomb and Trithorax groups of epigenetic regulators (see (Ringrose and Paro 2004) for review). These complexes regulate Hox loci by maintaining their repressed or active transcription states, respectively, through cell division cycles. They achieve this by establishing repressed or active chromatin throughout cis-regulatory elements called Polycomb response elements (PREs). Close examination of PREs revealed that these elements are actually transcribed, and that the resulting lncRNAs exert regulatory functions (see (Schmitt and Paro 2006) for review). Forcing transcription through silent PREs during embryogenesis switches their epigenetic state and leads to developmental abnormalities due to Hox gene misexpression. The same phenotype is observed when transcription from active PREs is disrupted. Thus, production of lncRNAs mediates the epigenetic state at Hox loci PREs. In fact, the lncRNAs themselves appear to recruit Polycomb/Trithorax complexes to PREs, by remaining tethered to them and physically binding these complexes. These observations have led to a model whereby Polycomb/Trithorax regulators find their chromatin targets via direct interaction with the lncRNAs tethered to them (Hekimoglu and Ringrose 2009).

As with flies, regulation of Hox genes in mammals involves regulatory lncRNA components. There are 39 Hox genes in mammals, grouped into four chromosomal loci (HOXA to HOXD) that are expressed along the anterior-posterior axis of the body in a manner collinear with their genomic position from 3’ to 5’ of the cluster. Rinn et al. identified a 2.2 kb lncRNA called HOTAIR that can repress the HOXD locus *in trans* (Rinn et al. 2007; Tsai et al. 2010). HOTAIR is transcribed antisense to protein-coding genes at the HOXC cluster in cells with posterior and distal positional identities. Its knockdown results in upregulation of genes residing in the HOXD cluster, the strongest effect being a ~2-fold increase in HOXD10 expression. Such activation is accompanied by loss of epigenetic silencing as assayed by reduction in levels of H3K27me3. Repression of HOXD genes by HOTAIR is mediated by direct recruitment of PRC2 and of another chromatin modifying complex containing LSD1, a lysine demethylase which primarily targets H3K4. This role is mediated by structural domains at the 5' and 3' ends of HOTAIR, consistent with greater evolutionary constraint on their inferred secondary structure than on their primary sequence (He et al. 2011). Thus, HOTAIR acts to repress transcription of the HoxD locus via physical recruitment of chromatin modifiers *in trans*.

Deleting a large chromosomal region including HOTAIR, eight HoxC genes, two microRNAs, and several other lncRNAs in mouse does not alter overall body plan and only results in modest de-repression of HoxD genes (Suemori and Noguchi 2000; Schorderet and Duboule 2011). However, targeted deletion of only HOTAIR indeed affects HoxD gene expression in developing anterior and distal skeletal structures and leads to homeotic transformations of the spine and malformation of distal skeletal structures (Li et al. 2013), similar to targeted deletion of individual HoxC genes (Suemori et al. 1995; Saegusa et al. 1996). Such discrepancies in phenotypic outcomes again highlight the critical importance of the genetic perturbation strategy of choice. In the case of the gross chromosomal deletion strategy, compensatory effects may be elicited by perturbation of proximal regulatory elements or antagonistic genes.

Recently, HOTAIR has also been implicated in disease, as it is found overexpressed in a wide variety of cancers (Gutschner and Diederichs 2012). In breast and colorectal cancer, for example, HOTAIR appears to modulate tumor invasiveness by enhancing PRC2-mediated repression of genes that suppress metastasis (Gupta et al. 2010; Kogo et al. 2011). Therefore, HOTAIR plays a critical role during both development and disease by helping specify gene expression programs via epigenetic modulation. Because HOTAIR recruits not only a Polycomb/Trithorax complex but also an unrelated chromatin modifier, this example laid the ground for an expanded model of lncRNAs as platforms for the assembly of specific combinations of broad-acting chromatin modifiers (Koziol and Rinn 2010; Tsai et al. 2010)

In addition to repressing transcription via Polycomb, Hox lncRNAs can also facilitate transcriptional activation via Trithorax. Three lncRNAs from the HoxA cluster, HOTTIP, Mistral and HOTAIRM1, have such capacity (Zhang et al. 2009; Bertani et al. 2011; Wang et al. 2011). HOTTIP resides in the 5’ tip of the HoxA locus. Although poorly expressed, this ~3.7kb lncRNA can be specifically detected at distal/posterior sites in the embryo. The positive correlation between HOTTIP expression and that of its neighbors at the HoxA locus suggests that HOTTIP modulates their activity. Consistent with this notion, inhibition of HOTTIP by siRNA results in 30-80% reduction in the expression of the HoxA7-13 genes in a manner inversely proportional to their distance from HOTTIP. This reduction is associated with appearance of repressive H3K27me3 and disappearance of active H3K4me3 marks, accompanied by decreased occupancy of the Trithorax WDR5/MLL1 complex. Biochemical analysis revealed that WDR5 can specifically interact with HOTTIP and that this interaction causes target gene activation only when HOTTIP is physically proximal, as indicated by tethering experiments. Hence, HOTTIP helps maintain the active epigenetic state of the HoxA locus, and this effect depends on both direct association with the WDR5/MLL complex and immediate physical proximity. This is supported by detection of endogenous chromatin interactions between HOTTIP and target loci as assessed by chromosome conformation capture, and by the fact that its low copy number (<1 copy per cell measured by single-molecule RNA FISH) would limit significant activity *in trans*. To study HOTTIP function *in vivo*, Wang and colleagues injected retroviruses carrying shRNAs into the upper limb buds of early chicken embryos (Wang et al. 2011). Knockdown caused decreased expression of HoxA10-13, as expected, and this effect was most pronounced at the distal edge of developing limb buds, where the 5' HoxA genes are most prominently expressed. Remarkably, by late embryonic stages this results in up to ~20% reduction in distal limb bones, which exhibit notably abnormal morphology. Such phenotypes mirror those of mice lacking 5' HoxA genes, indicating that HoxA lncRNAs also contribute to organismal development by affecting HoxA gene expression.

Cells at anterior and proximal locations of the body plan express genes at the 3' end of the HoxA locus instead of genes at the 5' tip. This is again mediated by lncRNA activators, such as Mistral and HOTAIRM1. Mistral recruits the WDR5/MLL1 complex to activate expression of its neighbors HoxA6 and HoxA7 (see above). As with HOTTIP, recruitment of the WDR5/MLL1 complex by Mistral can result in chromosome conformation changes that contribute to gene activation during cell differentiation. Transcription of the remaining 3' HoxA genes, HoxA1-5, is influenced by HOTAIRM1 through an analogous mechanism. HOTAIRM1 was first characterized in the context of hematopoiesis (see below).

The examples above clearly indicate that, much like proteins and microRNAs, lncRNAs play important roles in repressing and activating Hox genes. Thus, regulation by lncRNAs can contribute to the precise temporal and spatial control of genes that specify the body plan in metazoans.

lncRNAs in neural cell differentiation and brain development

The development of neural tissues during embryogenesis also involves a variety of cell differentiation processes executed under exquisite temporal and spatial control. Formation of the vertebrate central nervous system (CNS) alone involves the generation of millions of neurons with distinct gene expression programs conferring distinct molecular and physiologic properties. There are two broad types of cells in this system: neurons and glia cells. These are generated from neural stem cells, which can be isolated from adult brain or derived from ES cells. As with developing mesoderm, lncRNAs are active in the developing CNS and play key roles during neural fate specification.

The first clue about the importance of lncRNAs in neurogenesis came from the observation that hundreds of them are specifically expressed in the CNS in both fruit flies and mice (Inagaki et al. 2005; Mercer et al. 2008b). These include members of all known lncRNA families, such as intronic, antisense and intergenic lncRNAs. In the mouse brain, detection by RNA FISH revealed that many lncRNAs are expressed in specific neural cell types, neuroanatomical regions, and subcellular compartments. Such expression specificity suggested that some of these lncRNAs may modulate the development or function of specific neural cell types. Consistent with this, transcriptome profiling during neurogenesis revealed that many lncRNAs are differentially expressed during mouse neuronal-glial fate specification and during oligodendrocyte lineage maturation (Mercer et al. 2010).

Functional studies have now revealed critical roles for several lncRNAs in modulating neural cell development. For example, Evf2 has been functionally implicated in hippocampus development (Feng et al. 2006; Bond et al. 2009; Berghoff et al. 2013). Evf2 is a multiexonic and polyadenylated transcript expressed from an ultraconserved enhancer located between the Dlx5 and Dlx6 loci. These loci encode homeodomain-containing transcription factors with critical roles in inhibitory interneuron differentiation and migration. Loss of Evf2 by transcriptional terminator insertion in mice results in activation of Dlx5 and Dlx6, leading to reduced numbers of GABAergic interneurons, which in turn compromises synaptic inhibition in the early postnatal hippocampus and dentate gyrus that persists in the adult. Mechanistically, Evf2 inhibits CpG DNA methylation of the ultraconserved enhancer to effectively modulate competition between the transcriptional activator of Dlx5, Dlx1/2, and its repressor, Mecp2, in favor of Mecp2. Importantly, loss of this effect in mutant mice can be rescued by Evf2 expression form a separate transgene, indicating a function *in trans.* Conversely, Evf2 appears to repress Dlx6 *in cis* via antisense transcription, enabling differential control of adjacent genes that share proximal regulatory elements. Hence, Evf2 plays a critical role in the formation of GABA-dependent neuronal circuitry in the developing hippocampus by modulating expression of key transcription factors in control of the GABAergic interneuron cell fate.

Another lncRNA that functions in the specification of a particular type of neurons is linc-Brn1b. This gene is located ~10kb downstream of Brn1 (also known as Pou3f3), a transcription factor involved in CNS development. It encodes a ~3kb multiexonic lncRNA enriched in the nucleus that is expressed within embryonic neural progenitors of the telencephalon and remains enriched in the upper layers of the developing cerebral cortex and later in the mature somatosensory cortex and primary visual cortex. Genetic deletion of linc-Brn1b causes ~50%

loss of Brn1 protein and results in decreased proliferation of cortical progenitors in the subventricular zone, leading to reduced numbers of cortical projection neurons of layer II-IV. Accordingly, mice deleted for linc-Brn1b present abnormal organization of the somatosensory cortex. Thus, although its mechanism is unknown, linc-Brn1b plays an essential role in cortical lamination by modulating proper generation of different types of projection neurons.

Analogously, lncRNAs also modulate glial cell fate specification. The lncRNA Nkx2.2AS participates in neurogenesis by favoring differentiation of neural stem cells along the oligodendrocyte lineage (Tochitani and Hayashizaki 2008). Nkx2.2AS is a cytoplasmic transcript transcribed antisense to Nkx2.2, a master transcription factor of oligodendrocyte differentiation. Overexpression of Nkx2.2AS in cultured primary neural stem cells increased expression of Nkx2.2 by about 30% and resulted in a modest increase in the formation of oligodendrocytes. Thus, Nkx2.2AS appears to favor the oligodendrocyte cell fate by enhancing Nkx2.2 expression, although no *in vivo* loss-of-function evidence has emerged.

lncRNAs also play important roles during retinal cell development (see (Rapicavoli and Blackshaw 2009) for review). For example, the nuclear-retained lincRNA RNCR2 becomes specifically enriched in retinal progenitor cells during embryogenesis. Knockdown by shRNAs resulted in differentiation of progenitor cells towards non-retinal cell lineages, such as amacrine cells, suggesting that RNCR2 is involved in retinal cell fate specification. The same effect was observed by mislocalization of RNCR2 to the cytoplasm, via fusion with an IRES-controlled GFP transgene, indicating that correct cellular localization of the lncRNA is important for its function. RNCR2 seems to specifically interact with the SF1 splicing factor through conserved repeat sequences that resemble intron branch point motifs (Tsuiji et al. 2011). Binding of

RNCR2 to SF1 *in vitro* can inhibit splicing complex formation, suggesting that RNCR2 may function by regulating splicing efficiency.

Another lncRNA regulator of retinal development is TUG1, a ~6.7kb spliced and polyadenylated transcript that localizes to both nucleus and cytoplasm and is conserved throughout mammals. TUG1 is directly activated by Taurine, the master regulator of rod photoreceptor production. Downregulation of TUG1 by RNAi leads to disrupted photoreceptor formation due to impaired migration into the outer nuclear layer and increased apoptosis. Accordingly, TUG1 is directly activated by p53 upon DNA damage and acts to repress a range of cell cycle genes via association with PRC2 (Guttman et al. 2009; Khalil et al. 2009). Analogously, Meola et al. reported that overexpression of the lncRNA Vax2os1 inhibits retinal progenitor cell proliferation (Meola et al. 2012). Vax2os1is selectively expressed in the developing retina, and it appears to function through impairment of cell cycle progression and increased apoptosis.

Regulation by lncRNAs has also been studied during differentiation of human ES cells towards neuronal progenitor cells and ultimately neurons (Ng et al. 2012). About 35 lncRNAs were found to be upregulated during terminal neuronal differentiation by this strategy. Knockdown by siRNAs of 4 of these, RMST and lncRNA_N1-3, resulted in global gene expression changes and impairment of neuronal differentiation. Mechanistically, 3 of these lncRNAs appear to act in the regulation of chromatin state, as they reside predominantly in the nucleus and bind the PRC2 complex.

These examples indicate that as a group of gene expression regulators, lncRNAs play important yet diverse roles in neuronal differentiation both in culture and *in vivo*. The involvement of many lncRNAs in epigenetic control, and the fact that a large proportion of primate- and human-specific lncRNAs seem specifically enriched in the brain, predict that some might also be involved in maintaining proper neuronal function during complex physiological processes, such as long-term memory formation, sensory processing or behavioral patterns (Mercer et al. 2008a; Anguera et al. 2011; Lipovich et al. 2012).

Functional roles of lncRNAs during CNS development appear conserved from zebrafish to human. A recent study of hundreds of lincRNAs in the zebrafish *Danio rerio*, including 29 with detectable human orthologs, found 2 required for normal development of both brain and retina (Ulitsky et al. 2011). The first one, Cyrano, is a ~4.5kb polyadenylated transcript conserved in mouse and humans that is expressed in brain, notochord and subsequently spinal cord. Knockdown of Cyrano by antisense morpholinos caused small heads and eyes due in part to defects in neural tube opening and loss of retinal neuroD-positive cells. Remarkably, these defects could be rescued by ectopic expression of either mature zebrafish cyrano or its human or mouse orthologs. Cyrano harbors a 26nt sequence highly conserved throughout vertebrates that mirrors a microRNA-7 binding site, suggesting that it might exert its function through microRNA regulation. The second lincRNA, Megamind, is a ~2.4kb transcript antisense to an intron of birc6 that is specifically enriched in the brain. Knockdown of Megamind resulted in abnormal nervous system development such as smaller heads and eyes, enlarged brain ventricles (hydrocephalia) and loss of Neuro-D positive cells in the retina. As with Cyrano, Megamind and its brain-specific expression are conserved in mouse and human, and its loss of function phenotype was rescued by either the zebrafish transcript or its human or mouse orthologs. Hence, lncRNA sequences, expression patterns and functions during neural development appear conserved from zebrafish to human.

lncRNAs during muscle differentiation

Muscle cell differentiation is a highly coordinated developmental program executed during both embryogenesis and adult tissue homeostasis. Many key transcription factors and microRNAs controlling gene expression during muscle differentiation, growth, and morphogenesis have been characterized in both *in vitro* tissue culture and *in vivo* mouse models (see (Braun and Gautel 2011) for review). In addition to these components, lncRNAs are also active regulators of muscle development.

A number of lncRNAs are differentially expressed during differentiation of myoblasts into myotubes (Sunwoo et al. 2009). A study by Cesana et al. characterized a muscle-specific one, linc-MD1, which regulates the action of two microRNAs important for muscle development, microRNA-133 and microRNA-135 (Cesana et al. 2011). linc-MD1 is an alternatively spliced polyadenylated transcript activated by the key myogenic transcription factor MyoD during myoblast differentiation. It hosts microRNA-206 in one intron and microRNA-133b in one exon, and unlike lncRNAs that regulate epigenetic modification, it resides in the cytoplasm. Inspection of its sequence revealed highly conserved binding sites for both microRNA-133 and microRNA-135. Functional studies indicated that linc-MD1 can “sponge” these microRNAs and thus indirectly upregulate their mRNA targets, including Mef2c and Maml1, which are required for normal myogenesis. Accordingly, linc-MD1 inhibition compromises muscle differentiation, as assayed by reduced myogenic marker accumulation. Overexpression of a mutated linc-MD1 transcript from which microRNA-133b cannot be released, on the other hand, results in increased marker expression, indicating that microRNA-host lncRNAs can have independent regulatory functions. The alternative fates of linc-MD1, processing into miR-133b vs. life as a cytoplasmic RNA regulator, are controlled by HuR, a known myogenic regulator, which favors the latter (Legnini et al. 2014). Thus, linc-MD1 plays a role in fine-tuning the activity of miRNAs important for the muscle differentiation program.

This example illustrates a recently proposed model whereby RNA transcripts can indirectly modulate each other by competing for the available pool of common microRNA regulators (Rubio-Somoza et al. 2011). Interestingly, linc-MD1 appears downregulated in Duchenne muscular dystrophy myoblasts, and rescuing its levels via ectopic expression partially restores normal myogenesis in culture.

The opposite pattern is observed for another lncRNA, DBE-T, which is inactive in normal muscle cells but becomes activated in facioscapulohumeral muscular dystrophy (FSHD) (Cabianca et al. 2012). FSHD is caused by shortening of a 3.3kb repeat tract called D4Z4. Under normal conditions, the D4Z4 repeat array is epigenetically silenced by Polycomb, resulting in a repressive chromatin state that leads to silencing of FSHD genes via long-range interactions. Under FSHD, shortening of the D4Z4 repeat array causes loss of Polycomb silencing and facilitates transcription of the upstream DBE-T locus. Activated DBE-T lncRNA in turn recruits the Trithorax group protein Ash1L to the FSHD locus and coordinates de-repression of FSHD genes through long-range chromatin interactions. Thus, transcription of DBE-T mediates an epigenetic switch at the FSHD locus via direct recruitment of chromatin remodeling complexes. Interestingly, the FSHD locus shares several sequence features with Drosophila Polycomb/Trithorax response elements, which are also epigenetically switchable by virtue of lncRNA transcription (see above). Therefore, roles for lncRNAs in driving epigenetic switches at

Polycomb/Trithorax target elements in control of nearby gene expression are conserved from flies to human.

lncRNAs can also mediate mRNA decay processes active during muscle differentiation. Using C2C12 myoblasts as an *in vitro* culture system, Gong et al. observed that two mRNA decay pathways, Staufen1-mediated mRNA decay (SMD) and nonsense-mediated mRNA decay (NMD), contribute to muscle differentiation by regulating the abundance of target mRNAs (Gong et al. 2009). Certain polyadenylated and cytoplasmic lncRNAs, termed ½-sbsRNAs, seem to trigger SMD by imperfect base-paring to the 3' UTR of select target mRNAs through shared Alu repeat elements. These lncRNA-mRNA interactions can recruit Staufen1, the key SMD effector, and lead to degradation of the mRNA (Gong and Maquat 2011). ½-sbsRNAs are broadly expressed throughout human tissues, suggesting a ubiquitous role in mRNA decay. Hence, some cytoplasmic lncRNAs are able to modulate mRNA stability through the SMD pathway. The examples of linc-MD1 and ½-sbsRNAs provide evidence that in addition to regulating chromatin modification in the nucleus, some lncRNAs can also modulate microRNA activity and mRNA stability in the cytoplasm.

lncRNAs and maintenance of adult tissue homeostasis

Several lncRNAs have been discovered in the context of mature tissue homeostasis. Kretz et al. demonstrated that a lncRNA antagonizes differentiation of keratinocyte progenitors within epidermal tissue, which is typically renewed in a weekly basis (Kretz et al. 2012). Using global transcriptome sequencing, they identified lncRNAs expressed during terminal differentiation of human keratinocytes, adipocytes and osteoblasts. Among >1000 dynamically expressed lncRNAs, they focused on one, ANCR, that showed reduced expression upon differentiation of all three cell types. ANCR is an intergenic 855nt transcript that hosts both an intronic microRNA and an intronic snoRNA, present in the pre-processed but not the mature transcript. Depleting mature ANCR by siRNAs in keratinocyte progenitors resulted in upregulation of the epidermal differentiation program, including induction of epidermal markers, in the absence of differentiation stimuli. The same effects were observed upon knockdown of ANCR in *ex vivo* regenerated epidermal tissue that recapitulates normal epidermis organization, where ANCR loss led to ectopic differentiation in the progenitor-rich basal compartment.

An analogous role is fulfilled by another lncRNA, PINC, which is enriched in progenitor cells within the mammary gland (Ginger et al. 2001; Ginger et al. 2006; Shore et al. 2012). PINC is an alternatively spliced and polyadenylated transcript that can be found in the nucleus or the cytoplasm depending on the cell cycle stage, enriched in luminal and alveolar progenitors within the mammary gland. Physiologically, PINC is upregulated throughout pregnancy and becomes depleted during late pregnancy and early lactation, when alveolar cells undergo terminal differentiation into milk-producing cells. Accordingly, PINC is activated *in vivo* by stimulation of the mammary gland with estrogen and progesterone, and becomes downregulated *in vitro* when immortalized mammary epithelial cells are induced to differentiate by treatment with lactogenic hormones. In these cells, inhibition of PINC by siRNAs affected survival by limiting their cell cycle progression in the absence of differentiation stimuli, whereas in the presence of such stimuli it favored differentiation along the alveolar lineage. PINC overexpression, on the other hand, blocked alveolar differentiation. These effects seem mediated by repressing expression of genes associated with alveologenesis via direct association with PRC2, likely through the coordinately expressed PRC2 subunit RbAp46. Thus, as with ANCR, PINC acts to prevent adult lineage-determined progenitors from differentiating, likely via epigenetic repression of gene expression.

lncRNAs can also promote differentiation of lineage-determined progenitors. One such example is TINCR, a conserved ~3.7kb multiexonic lncRNA that is predominantly cytoplasmic and which, contrary to ANCR, is depleted in human keratinocyte progenitors but becomes highly induced upon their differentiation in culture (~20-25 molecules per cell) (Kretz et al. 2013). Depletion of TINCR by RNAi in regenerated epidermal tissue did not affect its organization, but disrupted expression of key mediators of epidermal maturation, resulting in impaired epidermal barrier formation. Mechanistically, TINCR binds to the key SMD effector Staufen1 as well as to a range of mRNAs associated with epidermal barrier formation via a 25-nucleotide sequence motif. Importantly, inhibition of Staufen1 phenocopied TINCR inhibition and showed ~50% overlap of differentially regulated genes, indicating that both bind to and functionally stabilize mRNAs required for proper epidermal tissue maturation. Thus, as with stem cells, lncRNAs play key roles in modulating the balance between self-renewal and differentiation of adult somatic progenitors.

Morphogenetic differentiation is another developmental process crucial for proper adult tissue homeostasis. Regulation by lncRNAs has been documented in two important morphogenetic processes: the epithelial-to-mesenchymal transition (EMT) and the formation of the vascular endothelium.

EMT is essential during embryogenesis for formation of mesoderm and the neural tube, and during epithelial cancer formation it is associated with elevated proliferation and metastasis. During EMT, epithelial cells that normally adhere to one another in ordered layers via E-cadherin revert to a migratory and undifferentiated fate characteristic of mesenchymal cells. Beltran et al. found that an antisense lncRNA to Zeb2, Zeb2NAT, acts as a positive regulator of EMT (Beltran et al. 2008). Zeb2 is normally inactive in epithelial cells, and its activation along with that of Snail and Zeb1 can lead to EMT via downregulation of E-cadherin. Zeb2NAT upregulation by Snail1 causes Zeb2 activation via an unusual mechanism. The Zeb2NAT lncRNA appears to directly bind the Zeb2 pre-mRNA to prevent splicing of an intron containing an internal ribosome entry site. Retention of this site is in turn required for efficient translation of Zeb2 and thus for activation of the EMT differentiation program. Interestingly, Snail1 also represses E-cadherin by binding to its promoter, thus promoting EMT both directly and indirectly via Zeb2NAT-mediated translation of Zeb2.

In an analogous example, Li et al described an antisense lncRNA, Tie-1AS, which seems to play a role during formation of the vascular endothelium, the inner lining of blood vessels (Li et al. 2010). Tie-1AS is an evolutionary conserved, ~800nt lncRNA transcribed antisense to Tie-1, which encodes a cell surface tyrosine kinase receptor for angiopoietin ligands. Tie-1AS appears to regulate the mRNA levels of Tie-1 by formation of a Tie-1 and Tie1-AS RNA duplex. Transient transfection of Tie-AS disrupts vascular tube formation both in zebrafish *in vivo* and in human vascular endothelial progenitors in culture. Accordingly, the ratio of Tie-1 mRNA vs. Tie-1AS lncRNA seems altered in pathological human vascular samples. This study suggests that modulation of Tie-1 levels by Tie-AS may be required for proper maintenance of vascular endothelial cells. However, loss-of-function experiments are needed to further clarify the physiological role of this antisense lncRNA. Hence, as with Zeb2NAT, direct interaction of antisense lncRNA Tie-1AS with its target mRNA may act to modulate somatic tissue morphogenesis during development and potentially during disease.

Modulation of hematopoiesis by lncRNAs

Hematopoiesis, the developmental process by which mature blood cells are generated from primary progenitors, is essential in all animals. In healthy humans, about two million erythrocytes must be generated every second to replace those lost by senescence, and overall numbers need to be maintained within a narrow physiological range. All of the hematopoietic effector cells (erythrocytes, myelocytes, and lymphocytes) derive from hematopoietic stem cells (HSCs) through a cascade of cell lineage specification, proliferation, and differentiation events. Hematopoietic multipotent and lineage-determined progenitor cells can be readily isolated using cell surface markers and have been extensively studied, making the hematopoietic system one of the best paradigms for studying cell lineage specification and differentiation in mammals (see (Orkin and Zon 2008) for review). In addition to well-characterized transcription factors and microRNAs, recent evidence indicates that lncRNAs also modulate hematopoiesis (**Fig.3**).

The capacity of lncRNAs to modulate self-renewal of embryonic and adult somatic stem cells predicts that they too may act in the circuitry controlling the HSC state. Li and colleagues recently described the first example of lncRNA-mediated maintenance of adult HSC quiescence (Venkatraman et al. 2013). The lncRNA H19, which contributes to growth control during embryogenesis, remains active in long-term HSCs and is gradually downregulated in short-term HSCs and multipotent progenitors. Genetic deletion of H19 from the maternal allele results in increased HSC activation and proliferation and impairs repopulating ability. As expected (see above), this effect is mediated by de-repression of maternal Igf2 expression and by increased Igf1r translation, resulting in increased signaling through the Igf1r. Accordingly, overexpressing the H19-derived miR-675 restores proper levels of Igf1r protein, and concomitant deletion of the Igf1r locus partially rescues the H19 knockout phenotype. Thus, H19 promotes HSC quiescence by regulating the Igf2-Igfr1 pathway at the transcriptional and posttranscriptional levels.

Interestingly, several other imprinted lncRNAs are upregulated in HSCs or in other adult stem cells relative to their differentiated progeny (Berg et al. 2011; Venkatraman et al. 2013). These include Air and the small RNA hosts from the Dlk1-Dio3 imprinted region Rian and Gtl2.

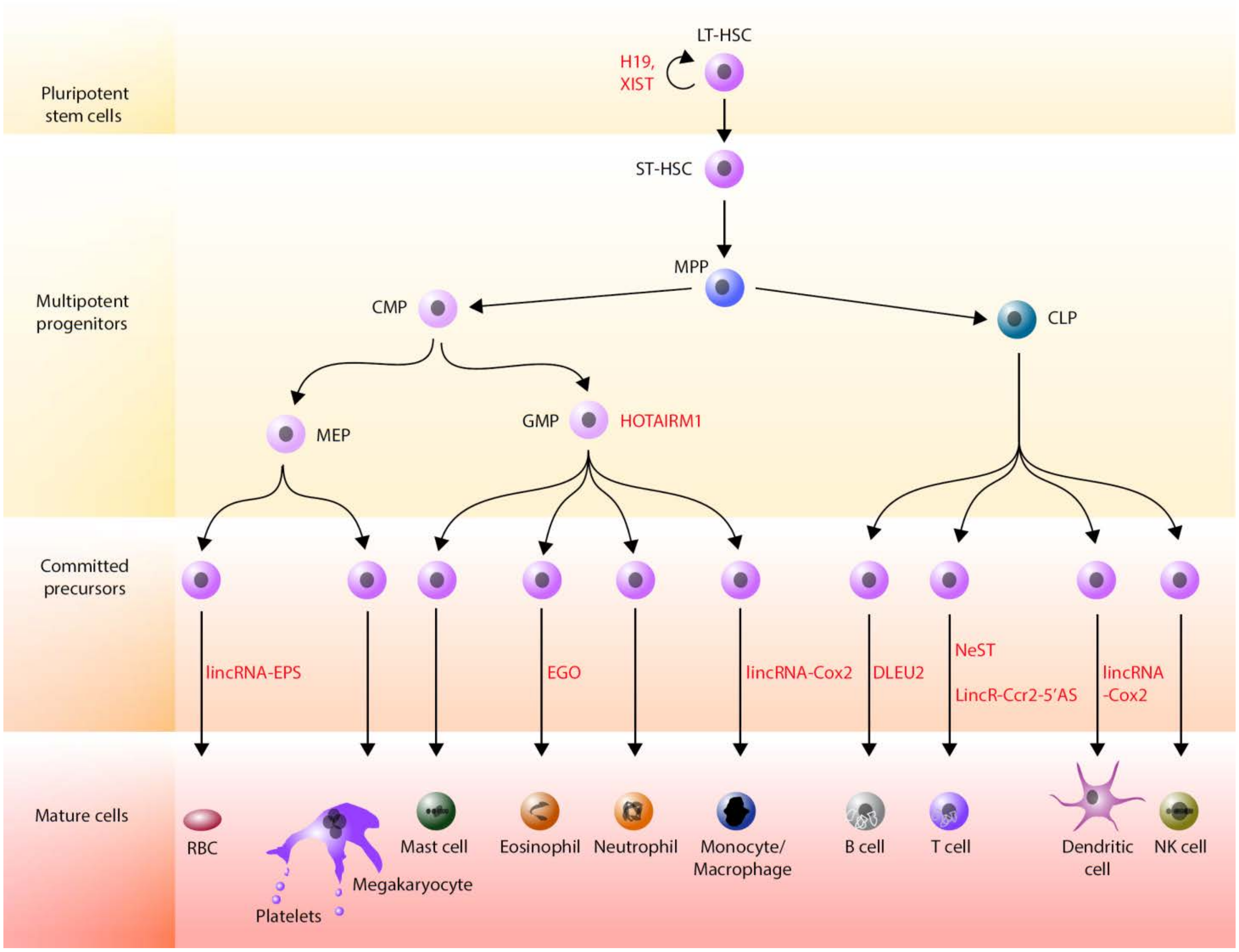


Figure 3. lncRNAs in blood cell development.

lncRNAs required for blood cell development, as determined through loss-of-function studies, are depicted next to the stage of hematopoietic development affected by their inhibition. LT-HSC, long-term hematopoietic stem cell; ST-HSC, short-term hematopoietic cell; MPP, multipotent progenitor; CMP, common myeloid progenitor; CLP, common lymphoid progenitor; MEP, megakaryocyte/erythroid progenitor; GMP; granulocyte/monocyte progenitor; RBC, red blood cell; NK cell, natural killer cell. Adapted from (Alvarez-Dominguez et al. 2014a)

Another well-known lncRNA plays an important role in proper modulation of HSC self-renewal and differentiation. Lee and colleagues have recently reported that regulation of HSC X chromosome dosage by the Xist lncRNA is causally linked to blood cancer (Yildirim et al. 2013) (**Fig.4B**). Conditional deletion of Xist in mouse HSCs, after the occurrence of X-inactivation, was lethal for both homozygous (Xist$^{-/-}$) and heterozygous (Xist$^{-/+}$) mutant females (at 100% penetrance) but not for their male counterparts. Deceased mutant females exhibited massive splenomegaly and extramedullary hematopoiesis, associated with hyperproliferation of all hematopoietic lineages, with myeloid cells outproliferating those of the other lineages. Bone marrow dysfunction was also observed, including myelofibrosis, myeloproliferation and myelodysplasia, leading to chronic myelomonocytic leukemia and erythroleukemia, thus recapitulating human myeloproliferative neoplasm and myelodysplastic syndrome (MPN/MDS). Wild-type mice transplanted with Xist$^{-/-}$ bone marrow developed MPN/MDS, whereas Xist$^{-/-}$ mice transplanted with wild-type bone marrow did not, indicating a cell-autonomous HSC defect. Accordingly, Xist$^{-/-}$ mice displayed impaired HSC maturation and loss of long-term HSCs. Mechanistically, these effects were mediated by widespread X reactivation leading to genome-wide changes including upregulation or downregulation of oncogenes or tumor suppressors implicated in MPN and MDS, respectively, including the key myeloid transcription factor Gata1. Thus, Xist is required for both establishment and long-term maintenance of proper X dosage *in vivo*, and its loss of function leads to blood cancer. Interestingly, selective loss of the inactive X and duplication of the active one are also frequently seen in breast and ovarian malignancies.

In the lymphoid lineages, roles for lncRNAs were first proposed by early observations of lymphoid-specific lncRNAs that are dynamically regulated during T-cell differentiation and

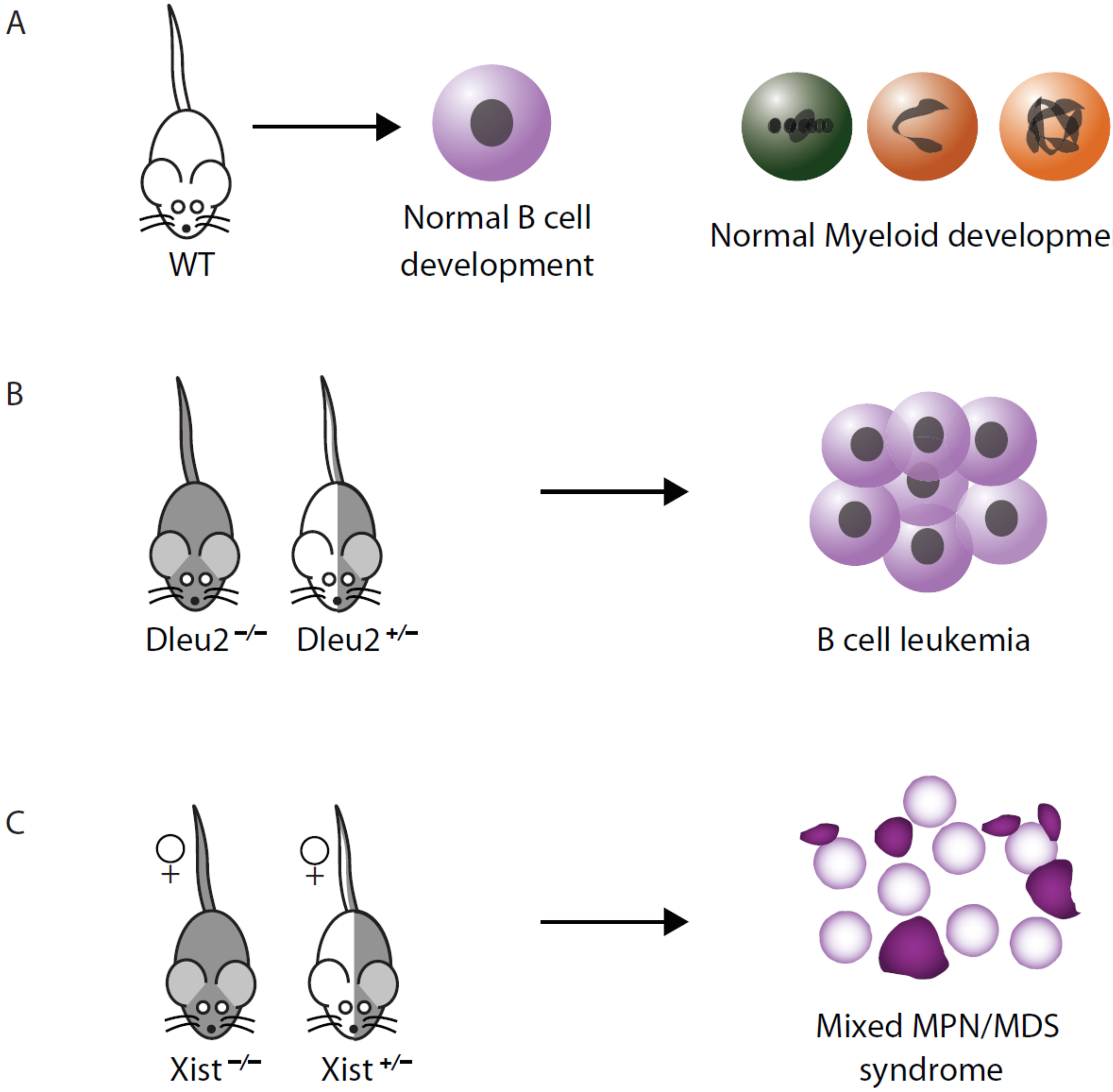


Figure 4. Mouse knockout models link lncRNAs to blood cancer pathogenesis.

(**A**) Wild-type mice develop B cells and myeloid cells normally.

(**B**) Dleu2−/− and Dleu2−/+ mice fail to properly regulate cell cycle progression and apoptosis of B cells, developing a B cell chronic lymphocytic leukemia reminiscent of human CLL that significantly reduces lifespan.

(**C**) Xist−/− and Xist−/+ female mice fail to maintain proper X dosage, developing a deficiency in HSC maturation that leads to a lethal mixed myeloproliferative neoplasm and myelodysplastic (MPN/MDS) syndrome.

Adapted from (Alvarez-Dominguez et al. 2014a).

activation (Liu et al. 1997; Haasch et al. 2002; Pang et al. 2009). Recent studies have now provided evidence for the importance of several lncRNAs in immune cell function.

The lncRNA NeST (Tmevpg1) modulates the ability of mice to respond to viral and bacterial infections (Collier et al. 2012; Gomez et al. 2013). The NeST locus was identified through a forward genetic screen as a susceptibility locus for sensitivity to pathogenesis of Thelier's virus in mice (Bureau et al. 1992; Vigneau et al. 2001; Vigneau et al. 2003). The NeST locus encodes an lncRNA specifically expressed by the T_H1 subset of helper T cells. Endogenous or ectopic expression of NeST regulates the degree of inflammation induced by infecting pathogens, such as Thelier's virus or Salmonella. Mechanistically, NeST regulates expression of the cytokine IFN-γ, critical for innate and adaptive immunity, by specifically interacting with WDR5, a Trithorax group component, in $CD8^+$ T cells. Thus, NeST acts in immune effector cells to regulate the outcome of viral or bacterial infections by epigenetically activating expression of the IFN-γ locus via direct interactions with chromatin modifiers.

Fitzgerald and colleagues recently characterized lincRNA-Cox2, which is dramatically upregulated downstream of signaling by the Toll-like receptors (TLRs) 1 and 2 in mouse bone marrow-derived dendritic cells and macrophages (Guttman et al. 2009; Carpenter et al. 2013). lincRNA-Cox2 encodes several RNA isoforms that do not associate with ribosomes. Suppressing them by shRNAs leads to upregulation of several IFN-stimulated genes, including Ccl5 and Ccrl. Overexpressing lincRNA-Cox2, on the other hand, results in severe attenuation of Ccl5 and overexpression of TLR-induced interleukin 6. lincRNA-Cox2 is found in both the nucleus and the cytoplasm and interacts with heterogeneous nuclear RNPs A/B and A2/B2 to achieve inhibitory activity. Thus, like NeST, lincRNA-Cox2 also acts during inflammatory signaling by modulating expression of immune response genes via interactions with regulatory complexes.

In another recent study, Hu et al. globally profiled lncRNA expression during the differentiation of naïve CD8$^+$ T cells into various helper T cell subsets, and functionally characterized LincR-Ccr2-5'AS, a T_H2-specific gene activated by the transcription factor Gata3 (Hu et al. 2013). LincR-Ccr2-5'AS is located between the genes encoding the chemokine receptors Ccr3 and Ccr2 and is coregulated with them. Depleting LincR-Ccr2-5'AS with shRNAs in T_H2 cells downregulated expression of the nearby Ccr1, Ccr2, Ccr3 and Ccr5 genes, without affecting their chromatin architecture, impairing their ability to migrate to the lungs *in vivo*. The global gene expression response of T_H2 cells depleted for LincR-Ccr2-5'AS significantly overlaps with that following GATA3 depletion, suggesting a functional link between the two during T cell differentiation and immune function. These examples document critical roles for lncRNAs in modulating immune responses by associating with regulatory complexes to specify immune gene expression programs.

Attesting to the importance of lncRNAs in the immune system, one of them, Dleu2, has been causally linked to B cell chronic lymphocytic leukemia (CLL), the most common adult B cell-derived cancer (**Fig.4C**). CLL is associated with deletion of 13q14, found at >50% frequency in both CLL and CD5$^+$ monoclonal B cell lymphocytosis (MBL), and at lower frequencies in CD5$^-$ B cell-derived malignancies and T cell lymphomas (Liu et al. 1995; Rosenwald et al. 1999; Dohner et al. 2000; Rawstron et al. 2008). The minimal deleted region (MDR) within 13q14 is a ~110 kb region comprising the DLEU2 lncRNA, which hosts miR-15a and miR-16-1. Mice deleted for the entire MDR or only for miR-15a/16-1 in all cells (or only in B cells) displayed MBL, developed CLL at moderate penetrance, and in some cases progressed into diffuse large B cell lymphoma, thus recapitulating human CLL (Klein et al. 2010). However, disease penetrance was 42% in *MDR*$^{-/-}$ mice vs. 26% in *miR-15a/16-1*$^{-/-}$ mice, which translated in shortened lifespan in the former but not the latter. Moreover, ectopic expression of miR-15a/16-1 in human CLL cells rescued their over-proliferation but not their resistance to apoptosis. Thus, loss of DLEU2 leads to CLL at least partially through loss of the cell cycle inhibitory microRNAs 15a and 16-1, but additional roles of DLEU2 contributing to a more aggressive disease course remained unexplained. Recent studies have now linked DLEU2 to the *in cis* repression of gene neighbors that positively regulate NF-kB, whose activation in CLL cells has been shown to prevent apoptosis (Mertens and Stilgenbauer 2012; Garding et al. 2013). Thus, DLEU2 acts as a tumor suppressor by regulating both cell cycle progression and NF-kB signaling in B cells, although the precise mechanisms for the latter role remain elusive.

In the myeloid lineage, the first lncRNA to be described was EGO, a conserved gene transcribed antisense to ITPR1 that modulates the development of eosinophils (Wagner et al. 2007). These blood cells play roles in immune responses against parasites and in allergic diseases such as asthma. EGO is normally expressed in human CD34+ HSCs and becomes upregulated during their differentiation into eosinophils. The EGO transcript is noncoding, as it does not associate with ribosomes. Knockdown of EGO by siRNAs in cultured CD34+ progenitors impaired the expression of genes critical for eosinophil development, including major basic protein and eosinophil derived neurotoxin. Thus, EGO can contribute to eosinophilopoiesis by enhancing the expression of genes needed for this process.

lncRNAs are also implicated in specification of the granulocyte lineage from common myeloid progenitors. Zhang et al. have studied a lincRNA (HOTAIRM1) in the HOXA cluster that is upregulated during retinoic acid–induced granulocytic differentiation of myeloid progenitor cells (Zhang et al. 2009; Zhang et al. 2014). Transcribed from the HOXA1/2 intergenic region, HOTAIRM1 is about 500nt in length and does not associate with ribosomes. It exhibits coordinated expression with HoxA1 and HoxA2 along the body plan, suggesting that it might be involved in maintaining their active state. Knockdown of HOTAIRM1 by shRNAs inhibits RA-induced HoxA1 and HoxA4 activation and impedes cell cycle exit, compromising granulocytic maturation. This effect may be mediated through its interaction with various chromatin modifiers (Guttman et al. 2011). Hence, HOTAIRM1 modulates myelopoiesis by regulating cell cycle progression via modulation of neighboring genes at the HoxA locus.

Prior to the work described in this thesis, our group characterized a lncRNA that plays an essential role in red blood cell development (Hu et al. 2011). lincRNA-EPS is specifically enriched in erythroid cells and becomes strongly induced during terminal differentiation (~20-30 molecules per cell). It encodes a ~2.5kb capped and polyadenylated transcript that is alternatively spliced and resides in the nucleus. lincRNA-EPS knockdown by shRNAs blocked proliferation of erythroid precursors in culture and resulted in elevated apoptosis. Conversely, ectopic expression protected them from apoptosis triggered by erythropoietin starvation. These effects are mediated by a highly conserved region in the 3’ terminal exon of lincRNA-EPS, which is sufficient for its anti-apoptotic activity. Importantly, disrupting the putative short ORFs within the transcript does not alter its function. Mechanistically, lincRNA-EPS appears to regulate apoptosis by repressing expression of a number of pro-apoptotic proteins, most prominently the caspase activating adaptor protein Pycard. Thus, lincRNA-EPS is an erythroid-specific lncRNA that modulates the balance between pro- and anti-apoptotic signaling during development of the erythroid lineage.

Collectively, these examples illustrate that lncRNAs fulfill diverse regulatory functions that shape the development of hematopoietic cells of different lineages during both health and disease. Such functional capacities suggest that lncRNA dysregulation may be a major factor contributing to lineage-specific blood disorders associated with developmental deficiencies.

lncRNAs in the modulation of adipogenesis

Adipogenesis, the development of mature adipocytes from pre-adipocyte precursors, has been one of the most intensively studied cell differentiation processes, given the prevalence of obesity and other metabolic disorders. Two main types of adipose lineages exist –white adipocytes, which store chemical energy in triglycerides, and brown adipocytes, which instead consume energy via uncoupled respiration to generate heat and protect against cold and obesity. Like hematopoiesis, formation of white and brown adipocytes involves a cascade of cell lineage specification, proliferation, and differentiation events governed by transcription factors, chromatin modifiers, and microRNAs (see (Rosen and Spiegelman 2014) for review).

The first lncRNA found to play a role in adipocyte physiology was the steroid receptor RNA activator SRA (see (Colley and Leedman 2011) for review). The SRA locus encodes a large number of isoforms found in the nucleus and the cytoplasm, some of which are able to encode a conserved SRAP protein, rendering it a bifunctional gene. Genetic deletion of SRA in mice results in elevated insulin sensitivity and renders them resistant to obesity and glucose intolerance induced by a high-fat diet. Accordingly, non-coding SRA has been shown to bind to and co-activate a number of nuclear steroid hormone receptors, including Pparγ, a key adipogenic regulator. At the same time, SRAP protein has also been shown to perform co-activator function with nuclear receptors, including Esrra, another adipogenic modulator. Thus, SRA may employ both RNA and protein components acting on overlapping pathways to modulate adipocyte differentiation and metabolism.

Prior to the work described in this thesis, our group used global gene expression analysis to more broadly survey lncRNAs active during adipogenesis. This led to the identification of 175 lncRNAs that are differentially expressed during differentiation of both white and brown adipocytes *in vivo* and in culture (Sun et al. 2013). In white adipocytes, the promoters of about a third of these genes were bound by key adipogenic transcription factors PPARγ and C/EBPα. Knockdown of 10 of these lncRNAs by siRNAs impaired white adipocyte differentiation to various extents, as assayed by differentiation markers and global gene expression analysis. Accordingly, these were designated as lncRNAs regulated in adipogenesis (RAP) 1–10. Detailed characterization of the one with the strongest phenotype, lncRAP-1, revealed an intriguing role in nuclear architecture organization. lncRAP-1 is a conserved intergenic locus in the X chromosome that escapes X inactivation. It encodes a large number of isoforms that are nuclear-retained and chromatin-associated and contain multiple 156bp repeats (Hacisuleyman et al. 2014). Accordingly, lncRAP-1 was renamed functional intergenic repeating RNA element (Firre). Interestingly, Firre localizes across a 5 Mb domain around its site of transcription and is in close proximity to five distinct trans-chromosomal loci, four of which encode proteins involved in adipogenesis and energy metabolism. Mechanistically, Firre binds heterogeneous nuclear ribonucleoprotein U via its 156 bp repeats, and this interaction is needed for co-localization of the loci in other chromosomes contacted by Firre. Thus, Firre modulates adipogenesis by bringing adipogenic factors together in spatial proximity via trans-chromosomal interactions mediated by a nuclear matrix protein, potentially enabling their co-regulation.

Contributions of this thesis

When we began this work, we reasoned that development of specialized tissues must be modulated by their own specific collections of lncRNAs. This was suggested by three notions that were emerging at the time: (i) non-coding genomic regions are pervasively transcribed and have expanded during evolution along with increases in developmental complexity; (ii) lncRNAs can modulate developmental processes, as known for Xist, H19 and a few other lncRNAs that were well-characterized; and (iii) many lncRNAs are highly tissue specific, recalling developmental TFs.

To explore our hypothesis, three main challenges lied ahead: (i) finding these tissue-specific lncRNAs, (ii) testing their impact on development, and (iii) characterizing how they work. The following chapters detail how we met these challenges by focusing on two well-studied lineages –red blood cells and adipocytes– which led to the following contributions:

(1) **Global lncRNA discovery by de novo assembly of lineage-specific transcriptomes.** This work demonstrated that lineage-specific transcriptomes harbor hundreds of lncRNAs waiting to be discovered.

(2) **Integrative annotation of lncRNAs during cell differentiation.** This work enabled identification of erythroid and adipose lncRNAs whose specific characteristics suggested roles in the development or functioning of these tissues.

(3) **Prioritizing lncRNAs from large catalogs for functional study.** This work provided ranked predictions of which lncRNAs were most likely to be functional in these tissues.

(4) **Functional characterization of novel lineage-specific lncRNA regulators.** This work demonstrated that the development and physiology of the erythroid and adipose lineages is modulated by lncRNAs specific to those tissues.

(5) **Mechanisms of lncRNA function during lineage-specific developmental programs.** This work demonstrated that lncRNAs modulate lineage-specific cell differentiation by partnering with ubiquitous regulatory protein complexes to promote or suppress competing gene expression programs controlling cell fate.

Thesis overview

This thesis studies the role of lncRNAs during lineage-specific cell differentiation. Chapter1 describes work in which I utilized deep transcriptome surveys to catalog lncRNAs active during mouse red blood cell development, and showed via loss-of-function assays that they participate in the regulatory circuitry underlying erythropoiesis. Chapter 2 describes work in which I focused on one lncRNA required for red cell maturation and characterized its molecular function. Chapter 3 details similar work undertaken to catalog lncRNAs active across different mouse fat depots and elucidate the function of one that is required for brown adipocyte development and function. The workflow and approaches described in these chapters provide a basis for the study of tissue-specific lncRNAs and the mechanisms by which they contribute to the development of individual cell lineages. In the concluding chapter, I synthesize emerging themes of lncRNA function during cell differentiation, and discuss future steps and considerations towards probing the ultimate contribution of lncRNA-based regulation to *in vivo* organismal development.

Appendices A, B, and C contain supplementary information for Chapters 1, 2, and 3, respectively. Appendix D contains a publication describing our efforts to distinguish between protein-coding and non-coding RNA transcripts based on experimental evidence of translation and bioinformatics analyses in yeast and in mammalian cells. Appendix E contains a manuscript in preparation in which we implicate an lncRNA mapping to a neuroblastoma susceptibility locus in the pathogenesis of this disease.

References

Alvarez-Dominguez JR, Hu W, Gromatzky AA, Lodish HF. 2014a. Long noncoding RNAs during normal and malignant hematopoiesis. *International journal of hematology* **99**: 531-541.

Alvarez-Dominguez JR, Hu W, Lodish HF. 2013. Regulation of Eukaryotic Cell Differentiation by Long Non-coding RNAs. in *Molecular Biology of Long Non-coding RNAs* (eds. AM Khalil, J Coller), pp. 15-67. Springer Science, New York.

Alvarez-Dominguez JR, Hu W, Yuan B, Shi J, Park SS, Gromatzky AA, van Oudenaarden A, Lodish HF. 2014b. Global discovery of erythroid long noncoding RNAs reveals novel regulators of red cell maturation. *Blood* **123**: 570-581.

Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS. 2011. lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res* **39**: D146-151.

Amaral PP, Dinger ME, Mercer TR, Mattick JS. 2008. The eukaryotic genome as an RNA machine. *Science* **319**: 1787-1789.

Amaral PP, Mattick JS. 2008. Noncoding RNA in development. *Mamm Genome* **19**: 454-492.

Andrews SJ, Rothnagel JA. 2014. Emerging evidence for functional peptides encoded by short open reading frames. *Nat Rev Genet* **15**: 193-204.

Anguera MC, Ma W, Clift D, Namekawa S, Kelleher RJ, 3rd, Lee JT. 2011. Tsx produces a long noncoding RNA and has general functions in the germline, stem cells, and brain. *PLoS Genet* **7**: e1002248.

Banfai B, Jia H, Khatun J, Wood E, Risk B, Gundling WE, Jr., Kundaje A, Gunawardena HP, Yu Y, Xie L et al. 2012. Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res* **22**: 1646-1657.

Barlow DP. 2011. Genomic imprinting: a mammalian epigenetic discovery model. *Annu Rev Genet* **45**: 379-403.

Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, Vejnar CE, Lee MT, Rajewsky N, Walther TC et al. 2014. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J* **33**: 981-993.

Beltran M, Puig I, Pena C, Garcia JM, Alvarez AB, Pena R, Bonilla F, de Herreros AG. 2008. A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. *Genes Dev* **22**: 756-769.

Berg JS, Lin KK, Sonnet C, Boles NC, Weksberg DC, Nguyen H, Holt LJ, Rickwood D, Daly RJ, Goodell MA. 2011. Imprinted Genes That Regulate Early Mammalian Growth Are Coexpressed in Somatic Stem Cells. *PLoS One* **6**.

Berghoff EG, Clark MF, Chen S, Cajigas I, Leib DE, Kohtz JD. 2013. Evf2 (Dlx6as) lncRNA regulates ultraconserved enhancer methylation and the differential transcriptional control of adjacent genes. *Development* **140**: 4407-4416.

Berretta J, Morillon A. 2009. Pervasive transcription constitutes a new level of eukaryotic genome regulation. *EMBO Rep* **10**: 973-982.

Bertani S, Sauer S, Bolotin E, Sauer F. 2011. The noncoding RNA Mistral activates Hoxa6 and Hoxa7 expression and stem cell differentiation by recruiting MLL1 to chromatin. *Mol Cell* **43**: 1040-1046.

Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S et al. 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**: 2242-2246.

Birney E Stamatoyannopoulos JA Dutta A Guigo R Gingeras TR Margulies EH Weng Z Snyder M Dermitzakis ET Thurman RE et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**: 799-816.

Blackshaw S, Harpavat S, Trimarchi J, Cai L, Huang H, Kuo WP, Weber G, Lee K, Fraioli RE, Cho SH et al. 2004. Genomic analysis of mouse retinal development. *PLoS Biol* **2**: E247.

Bond AM, Vangompel MJ, Sametsky EA, Clark MF, Savage JC, Disterhoft JF, Kohtz JD. 2009. Balanced gene regulation by an embryonic brain ncRNA is critical for adult hippocampal GABA circuitry. *Nat Neurosci* **12**: 1020-1027.

Brannan CI, Dees EC, Ingram RS, Tilghman SM. 1990. The product of the H19 gene may function as an RNA. *Mol Cell Biol* **10**: 28-36.

Braun T, Gautel M. 2011. Transcriptional mechanisms regulating skeletal muscle differentiation, growth and homeostasis. *Nat Rev Mol Cell Biol* **12**: 349-361.

Brockdorff N, Ashworth A, Kay GF, McCabe VM, Norris DP, Cooper PJ, Swift S, Rastan S. 1992. The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell* **71**: 515-526.

Brosius J. 2005. Waste not, want not--transcript excess in multicellular eukaryotes. *Trends Genet* **21**: 287-288.

Brown CJ, Hendrich BD, Rupert JL, Lafreniere RG, Xing Y, Lawrence J, Willard HF. 1992. The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* **71**: 527-542.

Bureau JF, Montagutelli X, Lefebvre S, Guenet JL, Pla M, Brahic M. 1992. The interaction of two groups of murine genes determines the persistence of Theiler's virus in the central nervous system. *Journal of Virology* **66**: 4698-4704.

Cabianca DS, Casa V, Bodega B, Xynos A, Ginelli E, Tanaka Y, Gabellini D. 2012. A long ncRNA links copy number variation to a polycomb/trithorax epigenetic switch in FSHD muscular dystrophy. *Cell* **149**: 819-831.

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**: 1915-1927.

Cai XZ, Cullen BR. 2007. The imprinted H19 noncoding RNA is a primary microRNA precursor. *Rna-a Publication of the Rna Society* **13**: 313-316.

Carninci P Kasukawa T Katayama S Gough J Frith MC Maeda N Oyama R Ravasi T Lenhard B Wells C et al. 2005. The transcriptional landscape of the mammalian genome. *Science* **309**: 1559-1563.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**: 626-635.

Carpenter S, Aiello D, Atianand MK, Ricci EP, Gandhi P, Hall LL, Byron M, Monks B, Henry-Bezy M, Lawrence JB et al. 2013. A Long Noncoding RNA Mediates Both Activation and Repression of Immune Response Genes. *Science* **341**: 789-792.

Carrieri C, Cimatti L, Biagioli M, Beugnet A, Zucchelli S, Fedele S, Pesce E, Ferrer I, Collavin L, Santoro C et al. 2012. Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature* **491**: 454-457.

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

Charlesworth B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* **10**: 195-205.

Chew GL, Pauli A, Rinn JL, Regev A, Schier AF, Valen E. 2013. Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. *Development* **140**: 2828-2834.

Chooniedass-Kothari S, Emberley E, Hamedani MK, Troup S, Wang X, Czosnek A, Hube F, Mutawe M, Watson PH, Leygue E. 2004. The steroid receptor RNA activator is the first functional RNA encoding a protein. *FEBS Lett* **566**: 43-47.

Clemson CM, Hutchinson JN, Sara SA, Ensminger AW, Fox AH, Chess A, Lawrence JB. 2009. An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Mol Cell* **33**: 717-726.

Colley SM, Leedman PJ. 2011. Steroid Receptor RNA Activator - A nuclear receptor coregulator with multiple partners: Insights and challenges. *Biochimie* **93**: 1966-1972.

Collier SP, Collins PL, Williams CL, Boothby MR, Aune TM. 2012. Cutting Edge: Influence of Tmevpg1, a Long Intergenic Noncoding RNA, on the Expression of Ifng by Th1 Cells. *J Immunol* **189**: 2084-2088.

Conrad T, Akhtar A. 2011. Dosage compensation in Drosophila melanogaster: epigenetic fine-tuning of chromosome-wide transcription. *Nat Rev Genet* **13**: 123-134.

David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM. 2006. A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci U S A* **103**: 5320-5325.

De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, Muller H, Ragoussis J, Wei CL, Natoli G. 2010. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol* **8**: e1000384.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res* **22**: 1775-1789.

Dinger ME, Amaral PP, Mercer TR, Pang KC, Bruce SJ, Gardiner BB, Askarian-Amiri ME, Ru K, Solda G, Simons C et al. 2008a. Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res* **18**: 1433-1445.

Dinger ME, Pang KC, Mercer TR, Mattick JS. 2008b. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput Biol* **4**: e1000176.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F et al. 2012. Landscape of transcription in human cells. *Nature* **489**: 101-108.

Dohner H, Stilgenbauer S, Benner A, Leupolt E, Krober A, Bullinger L, Dohner K, Bentz M, Lichter P. 2000. Genomic aberrations and survival in chronic lymphocytic leukemia. *N Engl J Med* **343**: 1910-1916.

Duret L, Chureau C, Samain S, Weissenbach J, Avner P. 2006. The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science* **312**: 1653-1655.

Ebisuya M, Yamamoto T, Nakajima M, Nishida E. 2008. Ripples from neighbouring transcription. *Nat Cell Biol* **10**: 1106-1113.

Engreitz JM, Pandya-Jones A, McDonel P, Shishkin A, Sirokman K, Surka C, Kadri S, Xing J, Goren A, Lander ES et al. 2013. The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome. *Science* **341**: 767-U233.

Fatica A, Bozzoni I. 2014. Long non-coding RNAs: new players in cell differentiation and development. *Nat Rev Genet* **15**: 7-21.

Feng J, Bi C, Clark BS, Mady R, Shah P, Kohtz JD. 2006. The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes Dev* **20**: 1470-1484.

Franco-Zorrilla JM, Valli A, Todesco M, Mateos I, Puga MI, Rubio-Somoza I, Leyva A, Weigel D, Garcia JA, Paz-Ares J. 2007. Target mimicry provides a new mechanism for regulation of microRNA activity. *Nature Genetics* **39**: 1033-1037.

Frith MC, Forrest AR, Nourbakhsh E, Pang KC, Kai C, Kawai J, Carninci P, Hayashizaki Y, Bailey TL, Grimmond SM. 2006. The abundance of short proteins in the mammalian proteome. *PLoS Genet* **2**: e52.

Gabory A, Jammes H, Dandolo L. 2010. The H19 locus: role of an imprinted non-coding RNA in growth and development. *Bioessays* **32**: 473-480.

Galindo MI, Pueyo JI, Fouix S, Bishop SA, Couso JP. 2007. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol* **5**: e106.

Garding A, Bhattacharya N, Claus R, Ruppel M, Tschuch C, Filarsky K, Idler I, Zucknick M, Caudron-Herger M, Oakes C et al. 2013. Epigenetic upregulation of lncRNAs at 13q14.3 in leukemia is linked to the In Cis downregulation of a gene cluster that targets NF-kB. *PLoS Genet* **9**: e1003373.

Ginger MR, Gonzalez-Rimbau MF, Gay JP, Rosen JM. 2001. Persistent changes in gene expression induced by estrogen and progesterone in the rat mammary gland. *Mol Endocrinol* **15**: 1993-2009.

Ginger MR, Shore AN, Contreras A, Rijnkels M, Miller J, Gonzalez-Rimbau MF, Rosen JM. 2006. A noncoding RNA is a potential marker of cell fate during mammary gland development. *Proc Natl Acad Sci U S A* **103**: 5781-5786.

Gomez JA, Wapinski OL, Yang YW, Bureau JF, Gopinath S, Monack DM, Chang HY, Brahic M, Kirkegaard K. 2013. The NeST long ncRNA controls microbial susceptibility and epigenetic activation of the interferon-gamma locus. *Cell* **152**: 743-754.

Gong C, Kim YK, Woeller CF, Tang Y, Maquat LE. 2009. SMD and NMD are competitive pathways that contribute to myogenesis: effects on PAX3 and myogenin mRNAs. *Genes Dev* **23**: 54-66.

Gong C, Maquat LE. 2011. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature* **470**: 284-288.

Gorodkin J, Hofacker IL. 2011. From structure prediction to genomic screens for novel non-coding RNAs. *PLoS Comput Biol* **7**: e1002100.

Grant J, Mahadevaiah SK, Khil P, Sangrithi MN, Royo H, Duckworth J, McCarrey JR, VandeBerg JL, Renfree MB, Taylor W et al. 2012. Rsx is a metatherian RNA with Xist-like properties in X-chromosome inactivation. *Nature* **487**: 254-258.

Gregory TR. 2005. Synergy between sequence and size in large-scale genomics. *Nature reviews Genetics* **6**: 699-708.

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

Grunwald D, Singer RH, Rout M. 2011. Nuclear export dynamics of RNA-protein complexes. *Nature* **475**: 333-341.

Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL et al. 2010. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**: 1071-1076.

Gutschner T, Diederichs S. 2012. The Hallmarks of Cancer: A long non-coding RNA point of view. *Rna Biol* **9**.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**: 223-227.

Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, Young G, Lucas AB, Ach R, Bruhn L et al. 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**: 295-300.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C et al. 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**: 503-510.

Guttman M, Rinn JL. 2012. Modular regulatory principles of large non-coding RNAs. *Nature* **482**: 339-346.

Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES. 2013. Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins. *Cell* **154**: 240-251.

Haasch D, Chen YW, Reilly RM, Chiou XG, Koterski S, Smith ML, Kroeger P, McWeeny K, Halbert DN, Mollison KW et al. 2002. T cell activation induces a noncoding RNA transcript sensitive to inhibition by immunosuppressant drugs and encoded by the proto-oncogene, BIC. *Cell Immunol* **217**: 78-86.

Hacisuleyman E, Goff LA, Trapnell C, Williams A, Henao-Mejia J, Sun L, McClanahan P, Hendrickson DG, Sauvageau M, Kelley DR et al. 2014. Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nat Struct Mol Biol* **21**: 198-206.

Hanna JH, Saha K, Jaenisch R. 2010. Pluripotency and cellular reprogramming: facts, hypotheses, unresolved issues. *Cell* **143**: 508-525.

He S, Liu S, Zhu H. 2011. The sequence, structure and evolutionary features of HOTAIR in mammals. *BMC Evol Biol* **11**: 102.

Hekimoglu B, Ringrose L. 2009. Non-coding RNAs in Polycomb/Trithorax regulation. *Rna Biol* **6**: 129-137.

Hu G, Tang Q, Sharma S, Yu F, Escobar TM, Muljo SA, Zhu J, Zhao K. 2013. Expression and regulation of intergenic long noncoding RNAs during T cell development and differentiation. *Nat Immunol*.

Hu W, Alvarez-Dominguez JR, Lodish HF. 2012. Regulation of mammalian cell differentiation by long non-coding RNAs. *EMBO Rep* **13**: 971-983.

Hu W, Yuan B, Flygare J, Lodish HF. 2011. Long noncoding RNA-mediated anti-apoptotic activity in murine erythroid terminal differentiation. *Genes Dev* **25**: 2573-2578.

Hube F, Guo J, Chooniedass-Kothari S, Cooper C, Hamedani MK, Dibrov AA, Blanchard AA, Wang X, Deng G, Myal Y et al. 2006. Alternative splicing of the first intron of the steroid receptor RNA activator (SRA) participates in the generation of coding and noncoding RNA isoforms in breast cancer cell lines. *DNA Cell Biol* **25**: 418-428.

Huttenhofer A, Schattner P, Polacek N. 2005. Non-coding RNAs: hope or hype? *Trends Genet* **21**: 289-297.

Inagaki S, Numata K, Kondo T, Tomita M, Yasuda K, Kanai A, Kageyama Y. 2005. Identification and expression analysis of putative mRNA-like non-coding RNA in Drosophila. *Genes Cells* **10**: 1163-1173.

Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJ, Jackson SE, Wills MR, Weissman JS. 2014. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep* **8**: 1365-1379.

Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**: 218-223.

Ingolia NT, Lareau LF, Weissman JS. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**: 789-802.

Jacob F, Monod J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**: 318-356.

Jacquier A. 2009. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat Rev Genet* **10**: 833-844.

Jenny A, Hachet O, Zavorszky P, Cyrklaff A, Weston MD, Johnston DS, Erdelyi M, Ephrussi A. 2006. A translation-independent role of oskar RNA in early Drosophila oogenesis. *Development* **133**: 2827-2833.

Jeon Y, Lee JT. 2011. YY1 tethers Xist RNA to the inactive X nucleation center. *Cell* **146**: 119-133.

Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL et al. 2007a. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**: 1484-1488.

Kapranov P, Willingham AT, Gingeras TR. 2007b. Genome-wide transcription and the implications for genomic organization. *Nat Rev Genet* **8**: 413-423.

Karreth FA, Tay Y, Perna D, Ala U, Tan SM, Rust AG, DeNicola G, Webster KA, Weiss D, Perez-Mancera PA et al. 2011. In vivo identification of tumor- suppressive PTEN ceRNAs in an oncogenic BRAF-induced mouse model of melanoma. *Cell* **147**: 382-395.

Keniry A, Oxley D, Monnier P, Kyba M, Dandolo L, Smits G, Reik W. 2012. The H19 lincRNA is a developmental reservoir of miR-675 that suppresses growth and lgf1r. *Nature Cell Biology* **14**: 659-665.

Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A et al. 2009. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **106**: 11667-11672.

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**: 182-187.

Kino T, Hurt DE, Ichijo T, Nader N, Chrousos GP. 2010. Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Sci Signal* **3**: ra8.

Klattenhoff CA, Scheuermann JC, Surface LE, Bradley RK, Fields PA, Steinhauser ML, Ding H, Butty VL, Torrey L, Haas S et al. 2013. Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. *Cell* **152**: 570-583.

Klein U, Lia M, Crespo M, Siegel R, Shen Q, Mo T, Ambesi-Impiombato A, Califano A, Migliazza A, Bhagat G et al. 2010. The DLEU2/miR-15a/16-1 cluster controls B cell proliferation and its deletion leads to chronic lymphocytic leukemia. *Cancer Cell* **17**: 28-40.

Kloc M, Wilk K, Vargas D, Shirato Y, Bilinski S, Etkin LD. 2005. Potential structural role of non-coding and coding RNAs in the organization of the cytoskeleton at the vegetal cortex of Xenopus oocytes. *Development* **132**: 3445-3457.

Kogo R, Shimamura T, Mimori K, Kawahara K, Imoto S, Sudo T, Tanaka F, Shibata K, Suzuki A, Komune S et al. 2011. Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. *Cancer research* **71**: 6320-6326.

Kowalczyk MS, Hughes JR, Garrick D, Lynch MD, Sharpe JA, Sloane-Stanley JA, McGowan SJ, De Gobbi M, Hosseini M, Vernimmen D et al. 2012. Intragenic enhancers act as alternative promoters. *Mol Cell* **45**: 447-458.

Koziol MJ, Rinn JL. 2010. RNA traffic control of chromatin complexes. *Curr Opin Genet Dev* **20**: 142-148.

Kretz M, Siprashvili Z, Chu C, Webster DE, Zehnder A, Qu K, Lee CS, Flockhart RJ, Groff AF, Chow J et al. 2013. Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature* **493**: 231-235.

Kretz M, Webster DE, Flockhart RJ, Lee CS, Zehnder A, Lopez-Pajares V, Qu K, Zheng GX, Chow J, Kim GE et al. 2012. Suppression of progenitor differentiation requires the long noncoding RNA ANCR. *Genes Dev* **26**: 338-343.

Latos PA, Pauler FM, Koerner MV, Senergin HB, Hudson QJ, Stocsits RR, Allhoff W, Stricker SH, Klement RM, Warczok KE et al. 2012. Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. *Science* **338**: 1469-1472.

Lee JT. 2000. Disruption of imprinted X inactivation by parent-of-origin effects at Tsix. *Cell* **103**: 17-27.

-. 2011. Gracefully ageing at 50, X-chromosome inactivation becomes a paradigm for RNA and chromatin control. *Nat Rev Mol Cell Biol* **12**: 815-826.

Legnini I, Morlando M, Mangiavacchi A, Fatica A, Bozzoni I. 2014. A feedforward regulatory loop between HuR and the long noncoding RNA linc-MD1 controls early phases of myogenesis. *Mol Cell* **53**: 506-514.

Leighton PA, Ingram RS, Eggenschwiler J, Efstratiadis A, Tilghman SM. 1995. Disruption of imprinting caused by deletion of the H19 gene region in mice. *Nature* **375**: 34-39.

Lewis EB. 1978. A gene complex controlling segmentation in Drosophila. *Nature* **276**: 565-570.

Li K, Blum Y, Verma A, Liu Z, Pramanik K, Leigh NR, Chun CZ, Samant GV, Zhao B, Garnaas MK et al. 2010. A noncoding antisense RNA in tie-1 locus regulates tie-1 function in vivo. *Blood* **115**: 133-139.

Li L, Liu B, Wapinski OL, Tsai MC, Qu K, Zhang J, Carlson JC, Lin M, Fang F, Gupta RA et al. 2013. Targeted Disruption of Hotair Leads to Homeotic Transformation and Gene Derepression. *Cell Rep*.

Li L, Wang X, Stolc V, Li X, Zhang D, Su N, Tongprasit W, Li S, Cheng Z, Wang J et al. 2006. Genome-wide transcription analyses in rice using tiling microarrays. *Nat Genet* **38**: 124-129.

Li YM, Franklin G, Cui HM, Svensson K, He XB, Adam G, Ohlsson R, Pfeifer S. 1998. The H19 transcript is associated with polysomes and may regulate IGF2 expression in trans. *J Biol Chem* **273**: 28247-28252.

Lin MF, Deoras AN, Rasmussen MD, Kellis M. 2008. Performance and scalability of discriminative metrics for comparative gene identification in 12 Drosophila genomes. *PLoS Comput Biol* **4**: e1000067.

Lin MF, Jungreis I, Kellis M. 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**: i275-282.

Lipovich L, Dachet F, Cai J, Bagla S, Balan K, Jia H, Loeb JA. 2012. Activity-dependent Human Brain Coding/Non-coding Gene Regulatory Networks. *Genetics*.

Lipshitz HD, Peattie DA, Hogness DS. 1987. Novel transcripts from the Ultrabithorax domain of the bithorax complex. *Genes Dev* **1**: 307-322.

Liu AY, Torchia BS, Migeon BR, Siliciano RF. 1997. The human NTT gene: Identification of a novel 17-kb noncoding nuclear RNA expressed in activated CD4(+) T cells. *Genomics* **39**: 171-184.

Liu Y, Hermanson M, Grander D, Merup M, Wu X, Heyman M, Rasool O, Juliusson G, Gahrton G, Detlofsson R et al. 1995. 13q deletions in lymphoid malignancies. *Blood* **86**: 1911-1915.

Loewer S, Cabili MN, Guttman M, Loh YH, Thomas K, Park IH, Garber M, Curran M, Onder T, Agarwal S et al. 2010. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat Genet* **42**: 1113-1117.

Maenner S, Blaud M, Fouillen L, Savoye A, Marchand V, Dubois A, Sanglier-Cianferani S, Van Dorsselaer A, Clerc P, Avner P et al. 2010. 2-D structure of the A region of Xist RNA and its implication for PRC2 association. *PLoS Biol* **8**: e1000276.

Mallo M, Wellik DM, Deschamps J. 2010. Hox genes and regional patterning of the vertebrate body plan. *Dev Biol* **344**: 7-15.

Mancini-Dinardo D, Steele SJ, Levorse JM, Ingram RS, Tilghman SM. 2006. Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. *Genes Dev* **20**: 1268-1282.

Marahrens Y, Panning B, Dausman J, Strauss W, Jaenisch R. 1997. Xist-deficient mice are defective in dosage compensation but not spermatogenesis. *Gene Dev* **11**: 156-166.

Marques AC, Hughes J, Graham B, Kowalczyk MS, Higgs DR, Ponting CP. 2013. Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. *Genome biology* **14**: R131.

Marques AC, Ponting CP. 2009. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol* **10**: R124.

Mattick JS. 2004. RNA regulation: a new genetics? *Nat Rev Genet* **5**: 316-323.

Meller VH, Wu KH, Roman G, Kuroda MI, Davis RL. 1997. roX1 RNA paints the X chromosome of male Drosophila and is regulated by the dosage compensation system. *Cell* **88**: 445-457.

Meola N, Pizzo M, Alfano G, Surace EM, Banfi S. 2012. The long noncoding RNA Vax2os1 controls the cell cycle progression of photoreceptor progenitors in the mouse retina. *RNA* **18**: 111-123.

Mercer TR, Dinger ME, Mariani J, Kosik KS, Mehler MF, Mattick JS. 2008a. Noncoding RNAs in Long-Term Memory Formation. *Neuroscientist* **14**: 434-445.

Mercer TR, Dinger ME, Mattick JS. 2009. Long non-coding RNAs: insights into functions. *Nat Rev Genet* **10**: 155-159.

Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS. 2008b. Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci U S A* **105**: 716-721.

Mercer TR, Qureshi IA, Gokhan S, Dinger ME, Li G, Mattick JS, Mehler MF. 2010. Long noncoding RNAs in neuronal-glial fate specification and oligodendrocyte lineage maturation. *BMC Neurosci* **11**: 14.

Mertens D, Stilgenbauer S. 2012. CLL and deletion 13q14: merely the miRs? *Blood* **119**: 2974-2975.

Michel AM, Choudhury KR, Firth AE, Ingolia NT, Atkins JF, Baranov PV. 2012. Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res* **22**: 2219-2229.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**: 621-628.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**: 1344-1349.

Nagano T, Mitchell JA, Sanz LA, Pauler FM, Ferguson-Smith AC, Feil R, Fraser P. 2008. The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* **322**: 1717-1720.

Ng SY, Johnson R, Stanton LW. 2012. Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. *EMBO J* **31**: 522-533.

Novikova IV, Hennelly SP, Sanbonmatsu KY. 2012. Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. *Nucleic Acids Res* **40**: 5034-5051.

Okazaki Y Furuno M Kasukawa T Adachi J Bono H Kondo S Nikaido I Osato N Saito R Suzuki H et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**: 563-573.

Orkin SH, Zon LI. 2008. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* **132**: 631-644.

Ozsolak F, Milos PM. 2011. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* **12**: 87-98.

Pachnis V, Brannan CI, Tilghman SM. 1988. The structure and expression of a novel gene activated in early mouse embryogenesis. *EMBO J* **7**: 673-681.

Palazzo AF, Gregory TR. 2014. The case for junk DNA. *PLoS Genet* **10**: e1004351.

Pandey RR, Mondal T, Mohammad F, Enroth S, Redrup L, Komorowski J, Nagano T, Mancini-Dinardo D, Kanduri C. 2008. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol Cell* **32**: 232-246.

Pang KC, Dinger ME, Mercer TR, Malquori L, Grimmond SM, Chen WS, Mattick JS. 2009. Genome-Wide Identification of Long Noncoding RNAs in CD8(+) T Cells. *J Immunol* **182**: 7738-7748.

Parker BJ, Moltke I, Roth A, Washietl S, Wen J, Kellis M, Breaker R, Pedersen JS. 2011. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. *Genome Res* **21**: 1929-1943.

Pauli A, Rinn JL, Schier AF. 2011. Non-coding RNAs as regulators of embryogenesis. *Nat Rev Genet* **12**: 136-149.

Pearson JC, Lemons D, McGinnis W. 2005. Modulating Hox gene functions during animal body patterning. *Nat Rev Genet* **6**: 893-904.

Penny GD, Kay GF, Sheardown SA, Rastan S, Brockdorff N. 1996. Requirement for Xist in X chromosome inactivation. *Nature* **379**: 131-137.

Perocchi F, Xu ZY, Clauder-Munster S, Steinmetz LM. 2007. Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. *Nucleic Acids Research* **35**.

Pheasant M, Mattick JS. 2007. Raising the estimate of functional human sequences. *Genome Res* **17**: 1245-1253.

Plath K, Fang J, Mlynarczyk-Evans SK, Cao R, Worringer KA, Wang H, de la Cruz CC, Otte AP, Panning B, Zhang Y. 2003. Role of histone H3 lysine 27 methylation in X inactivation. *Science* **300**: 131-135.

Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A et al. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443**: 167-172.

Ponjavic J, Ponting CP, Lunter G. 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* **17**: 556-565.

Ponting CP, Oliver PL, Reik W. 2009. Evolution and functions of long noncoding RNAs. *Cell* **136**: 629-641.

Poole AM. 2004. Is all that junk really regulatory RNA? *Nature Reviews Genetics* **5**.

Prasanth KV, Spector DL. 2007. Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. *Genes Dev* **21**: 11-42.

Rapicavoli NA, Blackshaw S. 2009. New meaning in the message: noncoding RNAs and their role in retinal development. *Dev Dyn* **238**: 2103-2114.

Ravasi T, Suzuki H, Pang KC, Katayama S, Furuno M, Okunishi R, Fukuda S, Ru K, Frith MC, Gongora MM et al. 2006. Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res* **16**: 11-19.

Rawstron AC, Bennett FL, O'Connor SJ, Kwok M, Fenton JA, Plummer M, de Tute R, Owen RG, Richards SJ, Jack AS et al. 2008. Monoclonal B-cell lymphocytosis and chronic lymphocytic leukemia. *N Engl J Med* **359**: 575-583.

Redrup L, Branco MR, Perdeaux ER, Krueger C, Lewis A, Santos F, Nagano T, Cobb BS, Fraser P, Reik W. 2009. The long noncoding RNA Kcnq1ot1 organises a lineage-specific nuclear domain for epigenetic gene silencing. *Development* **136**: 525-530.

Ringrose L, Paro R. 2004. Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. *Annu Rev Genet* **38**: 413-443.

Rinn JL, Chang HY. 2012. Genome regulation by long noncoding RNAs. *Annu Rev Biochem* **81**: 145-166.

Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E et al. 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**: 1311-1323.

Ripoche MA, Kress C, Poirier F, Dandolo L. 1997. Deletion of the H19 transcription unit reveals the existence of a putative imprinting control element. *Genes Dev* **11**: 1596-1604.

Rosen ED, Spiegelman BM. 2014. What we talk about when we talk about fat. *Cell* **156**: 20-44.

Rosenwald A, Ott G, Krumdiek AK, Dreyling MH, Katzenberger T, Kalla J, Roth S, Ott MM, Muller-Hermelink HK. 1999. A biological role for deletions in chromosomal band 13q14 in mantle cell and peripheral t-cell lymphomas? *Genes Chromosomes Cancer* **26**: 210-214.

Rubio-Somoza I, Weigel D, Franco-Zorilla JM, Garcia JA, Paz-Ares J. 2011. ceRNAs: miRNA target mimic mimics. *Cell* **147**: 1431-1432.

Saegusa H, Takahashi N, Noguchi S, Suemori H. 1996. Targeted disruption in the mouse Hoxc-4 locus results in axial skeleton homeosis and malformation of the xiphoid process. *Dev Biol* **174**: 55-64.

Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP. 2011. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* **146**: 353-358.

Sasaki YT, Ideue T, Sano M, Mituyama T, Hirose T. 2009. MENepsilon/beta noncoding RNAs are essential for structural integrity of nuclear paraspeckles. *Proc Natl Acad Sci U S A* **106**: 2525-2530.

Sauvageau M, Goff LA, Lodato S, Bonev B, Groff AF, Gerhardinger C, Sanchez-Gomez DB, Hacisuleyman E, Li E, Spence M et al. 2013. Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife* **2**: e01749.

Schmitt S, Paro R. 2006. RNA at the steering wheel. *Genome Biology* **7**.

Schorderet P, Duboule D. 2011. Structural and functional differences in the long non-coding RNA hotair in mouse and human. *PLoS Genet* **7**: e1002071.

Selinger DW, Cheung KJ, Mei R, Johansson EM, Richmond CS, Blattner FR, Lockhart DJ, Church GM. 2000. RNA expression analysis using a 30 base pair resolution Escherichia coli genome array. *Nat Biotechnol* **18**: 1262-1268.

Sheik Mohamed J, Gaughwin PM, Lim B, Robson P, Lipovich L. 2010. Conserved long noncoding RNAs transcriptionally regulated by Oct4 and Nanog modulate pluripotency in mouse embryonic stem cells. *RNA* **16**: 324-337.

Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T et al. 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A* **100**: 15776-15781.

Shore AN, Kabotyanski EB, Roarty K, Smith MA, Zhang Y, Creighton CJ, Dinger ME, Rosen JM. 2012. Pregnancy-Induced Noncoding RNA (PINC) Associates with Polycomb Repressive Complex 2 and Regulates Mammary Epithelial Differentiation. *PLoS Genet* **8**: e1002840.

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

Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, Karger AD, Budnik BA, Rinn JL, Saghatelian A. 2013. Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nature chemical biology* **9**: 59-64.

Sleutels F, Tjon G, Ludwig T, Barlow DP. 2003. Imprinted silencing of Slc22a2 and Slc22a3 does not need transcriptional overlap between Igf2r and Air. *EMBO J* **22**: 3696-3704.

Sleutels F, Zwart R, Barlow DP. 2002. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* **415**: 810-813.

Smith CM, Steitz JA. 1998. Classification of gas5 as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5'-terminal oligopyrimidine gene family reveals common features of snoRNA host genes. *Mol Cell Biol* **18**: 6897-6909.

Stadtfeld M, Hochedlinger K. 2010. Induced pluripotency: history, mechanisms, and applications. *Genes Dev* **24**: 2239-2263.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**: 637-644.

Stolc V, Gauhar Z, Mason C, Halasz G, van Batenburg MF, Rifkin SA, Hua S, Herreman T, Tongprasit W, Barbano PE et al. 2004. A gene expression map for the euchromatic genome of Drosophila melanogaster. *Science* **306**: 655-660.

Stolc V, Samanta MP, Tongprasit W, Sethi H, Liang S, Nelson DC, Hegeman A, Nelson C, Rancour D, Bednarek S et al. 2005. Identification of transcribed sequences in Arabidopsis thaliana by using high-resolution genome tiling arrays. *Proc Natl Acad Sci U S A* **102**: 4453-4458.

Struhl K. 2007. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat Struct Mol Biol* **14**: 103-105.

Suemori H, Noguchi S. 2000. Hox C cluster genes are dispensable for overall body plan of mouse embryonic development. *Dev Biol* **220**: 333-342.

Suemori H, Takahashi N, Noguchi S. 1995. Hoxc-9 mutant mice show anterior transformation of the vertebrae and malformation of the sternum and ribs. *Mechanisms of development* **51**: 265-273.

Sun L, Goff LA, Trapnell C, Alexander R, Lo KA, Hacisuleyman E, Sauvageau M, Tazon-Vega B, Kelley DR, Hendrickson DG et al. 2013. Long noncoding RNAs regulate adipogenesis. *Proc Natl Acad Sci U S A* **110**: 3387-3392.

Sunwoo H, Dinger ME, Wilusz JE, Amaral PP, Mattick JS, Spector DL. 2009. MEN epsilon/beta nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. *Genome Res* **19**: 347-359.

Taft RJ, Pheasant M, Mattick JS. 2007. The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays* **29**: 288-299.

Tian D, Sun S, Lee JT. 2010. The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. *Cell* **143**: 390-403.

Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, Djebali S, Curado J, Snyder M, Gingeras TR, Guigo R. 2012. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res* **22**: 1616-1625.

Tjaden B, Saxena RM, Stolyar S, Haynor DR, Kolker E, Rosenow C. 2002. Transcriptome analysis of Escherichia coli using high-density oligonucleotide probe arrays. *Nucleic Acids Res* **30**: 3732-3738.

Tochitani S, Hayashizaki Y. 2008. Nkx2.2 antisense RNA overexpression enhanced oligodendrocytic differentiation. *Biochem Biophys Res Commun* **372**: 691-696.

Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, Watt AT, Freier SM, Bennett CF, Sharma A, Bubulya PA et al. 2010. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol Cell* **39**: 925-938.

Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E, Chang HY. 2010. Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**: 689-693.

Tsuiji H, Yoshimoto R, Hasegawa Y, Furuno M, Yoshida M, Nakagawa S. 2011. Competition between a noncoding exon and introns: Gomafu contains tandem UACUAAC repeats and associates with splicing factor-1. *Genes Cells* **16**: 479-490.

Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. 2011. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147**: 1537-1550.

van Bakel H, Nislow C, Blencowe BJ, Hughes TR. 2010. Most "dark matter" transcripts are associated with known genes. *PLoS Biol* **8**: e1000371.

Venkatraman A, He XC, Thorvaldsen JL, Sugimura R, Perry JM, Tao F, Zhao M, Christenson MK, Sanchez R, Yu JY et al. 2013. Maternal imprinting at the H19-Igf2 locus maintains adult haematopoietic stem cell quiescence. *Nature* **500**: 345-+.

Vigneau S, Levillayer F, Crespeau H, Cattolico L, Caudron B, Bihl F, Robert C, Brahic M, Weissenbach J, Bureau JF. 2001. Homology between a 173-kb region from mouse chromosome 10, telomeric to the Ifng locus, and human chromosome 12q15. *Genomics* **78**: 206-213.

Vigneau S, Rohrlich PS, Brahic M, Bureau JF. 2003. Tmevpg1, a candidate gene for the control of Theiler's virus persistence, could be implicated in the regulation of gamma interferon. *Journal of Virology* **77**: 5632-5638.

Wadler CS, Vanderpool CK. 2007. A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide. *Proc Natl Acad Sci U S A* **104**: 20454-20459.

Wagner LA, Christensen CJ, Dunn DM, Spangrude GJ, Georgelas A, Kelley L, Esplin MS, Weiss RB, Gleich GJ. 2007. EGO, a novel, noncoding RNA gene, regulates eosinophil granule protein transcript expression. *Blood* **109**: 5191-5198.

Wang KC, Chang HY. 2011. Molecular mechanisms of long noncoding RNAs. *Mol Cell* **43**: 904-914.

Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, Lajoie BR, Protacio A, Flynn RA, Gupta RA et al. 2011. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**: 120-124.

Wapinski O, Chang HY. 2011. Long noncoding RNAs and human disease. *Trends Cell Biol* **21**: 354-361.

Warner JR, Knopf PM, Rich A. 1963. A multiple ribosomal structure in protein synthesis. *Proc Natl Acad Sci U S A* **49**: 122-129.

Washietl S, Hofacker IL, Lukasser M, Huttenhofer A, Stadler PF. 2005. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat Biotechnol* **23**: 1383-1390.

Watanabe Y, Yamamoto M. 1994. S. pombe mei2+ encodes an RNA-binding protein essential for premeiotic DNA synthesis and meiosis I, which cooperates with a novel RNA species meiRNA. *Cell* **78**: 487-498.

White RJ. 2011. Transcription by RNA polymerase III: more complex than we thought. *Nat Rev Genet* **12**: 459-463.

Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bahler J. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**: 1239-1243.

Wilusz JE, Sunwoo H, Spector DL. 2009. Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev* **23**: 1494-1504.

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

Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, Southwick AM, Wu HC, Kim C, Nguyen M et al. 2003. Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science* **302**: 842-846.

Yekta S, Tabin CJ, Bartel DP. 2008. MicroRNAs in the Hox network: an apparent link to posterior prevalence. *Nat Rev Genet* **9**: 789-796.

Yildirim E, Kirby JE, Brown DE, Mercier FE, Sadreyev RI, Scadden DT, Lee JT. 2013. Xist RNA Is a Potent Suppressor of Hematologic Cancer in Mice. *Cell* **152**: 727-742.

Yoon JH, Abdelmohsen K, Srikantan S, Yang X, Martindale JL, De S, Huarte M, Zhan M, Becker KG, Gorospe M. 2012. LincRNA-p21 Suppresses Target mRNA Translation. *Mol Cell*.

Young RA. 2011. Control of the embryonic stem cell state. *Cell* **144**: 940-954.

Zhang X, Lian Z, Padden C, Gerstein MB, Rozowsky J, Snyder M, Gingeras TR, Kapranov P, Weissman SM, Newburger PE. 2009. A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster. *Blood* **113**: 2526-2534.

Zhang X, Weissman SM, Newburger PE. 2014. Long intergenic non-coding RNA HOTAIRM1 regulates cell cycle progression during myeloid maturation in NB4 human promyelocytic leukemia cells. *Rna Biol* **11**: 777-787.

Zhao J, Sun BK, Erwin JA, Song JJ, Lee JT. 2008. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322**: 750-756.

Chapter 1: Global discovery of erythroid long non-coding RNAs reveals novel regulators of red blood cell development

Parts of this chapter were first published as:

Alvarez-Dominguez JR, Hu W, Yuan B, Shi J, Park SS, Gromatzky AA, van Oudenaarden A, Lodish HF. 2014. Global discovery of erythroid long noncoding RNAs reveals novel regulators of red cell maturation. *Blood* **123**: 570-581.

Author contributions: J.R.A.-D. designed, performed, and presented the analysis of sequencing reads from RNA-seq, ChIP-seq, DNase-seq, CAGE-seq, poly(A)-seq and ChIA-PET experiments (Figures 1-7 and Figures S4, S6-S9 and S13); designed, performed and presented bioinformatics analyses (Figure 1 and Figures S1-S3 and S5); performed, analyzed and presented data from RNA-FISH experiments (Figure 5 and Figure S10); analyzed and presented qPCR data from tissue expression, knockdown and cell fractionation experiments (Figures 5-7 and Figures S10 and S12); analyzed and presented flow cytometry data from knockdown experiments (Figure 6); designed the website, and wrote the manuscript. W.H. designed and performed RNA-seq experiments (Figures 1, 2, 4, 5 and 7), and performed, analyzed and presented data from tissue expression and cell fractionation experiments (Figure 5 and Figure S10). W.H., J.S., and A.A.G. designed, performed, analyzed and presented data from knockdown experiments (Figures 6 and 7 and Figures S11 and S12). B.Y. performed analysis of sequencing reads from RNA-seq experiments (Figures 1-3). B.Y. and S.S.P. performed and contributed design of and presentation of bioinformatics analyses (Figure 1 and Figures S1-S5). J.R.A.-D., W.H., A.v.O. and H.F.L. contributed to study design and revised the manuscript.

Acknowledgments: We thank Laurie Boyer, David Bartel, Nikolai Slavov, Lei Sun, and members of the Lodish lab, especially Marko Knoll, for helpful discussions and critical comments on this manuscript. We also thank the MIT and Whitehead Institute flow cytometry cores and the Whitehead Institute genome technology core for technical support and Alec Garza-Galindo for help with design and implementation of the online resource accompanying this study. W.H. is a Merck Fellow of the Life Sciences Research Foundation and is supported by a Pathway to Independence Award (1K99HL118157-01) from the National Institutes of Health, National Heart, Lung, and Blood Institute. This research was supported by the National Institutes of Health, National Cancer Institute Physical Sciences Oncology Center at the Massachusetts Institute of Technology (U54CA143874) and by the grants from National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases DK068348 and from the National Institutes of Health, National Heart, Lung, and Blood Institute, 5P01 HL066105.

Erythropoiesis is regulated at multiple levels to ensure the proper generation of mature red cells under multiple physiological conditions. To probe the contribution of long non-coding RNAs (lncRNAs) to this process, we examined >1 billion RNA-Seq reads of polyadenylated and non-polyadenylated RNA from differentiating mouse fetal liver red blood cells, and identified 655 lncRNA genes including not only intergenic, antisense and intronic but also pseudogene and enhancer loci. Over 100 of these genes are previously unrecognized and highly erythroid-specific. By integrating genome-wide surveys of chromatin states, transcription factor occupancy, and tissue expression patterns, we identify multiple lncRNAs that are dynamically expressed during erythropoiesis, show epigenetic regulation and are targeted by key erythroid transcription factors GATA1, TAL1 or KLF1. We focus on 12 such candidates and find that they are nuclear-localized and exhibit complex developmental expression patterns. Depleting 10 out of 12 candidates reproducibly inhibited red cell enucleation, leading to the accumulation of immature but terminally differentiated erythroblasts. Our study provides an annotated catalog of erythroid lncRNAs, readily available through an online resource, and shows that diverse types of lncRNAs participate in the regulatory circuitry underlying erythropoiesis.

Introduction

Red blood cell development is a highly coordinated process essential throughout the lifetime of all mammals. In healthy humans, ~2 million erythrocytes need to be generated every second to replace those lost by senescence, and overall numbers need to be maintained within a narrow physiological range. The cells forming the erythroid lineage derive from a small population of pluripotent stem cells, which reside within the fetal liver or the adult bone marrow, via a cascade of cell lineage specification, proliferation, and differentiation events (reviewed in (Hattangadi et al. 2011). The earliest progenitors committed to the erythroid lineage are the slowly proliferating burst-forming unit erythroids (BFU-E). These undergo limited self-renewal and differentiate through the mature stage into the rapidly proliferating colony-forming unit erythroids (CFU-E). CFU-E precursors in turn divide 3-5 times over 2-3 days as they differentiate and undergo drastic changes, such as expulsion of the nucleus and other organelles, leading up to formation of mature erythrocytes.

Each stage in the production of red blood cells is regulated by a specific network of signaling factors and downstream effectors, and disruption of these networks leads to disease (Cantor and Orkin 2002; Kerenyi and Orkin 2010). The main short-term signaling hormone is erythropoietin (Epo), a cytokine that stimulates terminal proliferation and differentiation of CFU-E precursors mainly via the JAK/STAT signaling pathway. Recent studies have characterized many components of the complex networks activated downstream of Epo signaling during erythropoiesis. These include a variety of transcription factors, chromatin modifiers and microRNAs (Hattangadi et al. 2011). microRNAs have proven to be important modulators of critical aspects of erythropoiesis, such as lineage commitment, progenitor proliferation and terminal differentiation, suggesting that other types of ncRNA may also play important roles.

Long non-coding RNAs (lncRNAs) are transcripts longer than 200 nucleotides without functional protein-coding capacity. Large-scale studies indicate that these RNAs are pervasively transcribed in mammalian cells (Bertone et al. 2004; Carninci et al. 2005; Birney et al. 2007). Based on their genomic region of origin, lncRNAs can be classified as intergenic (lincRNAs), antisense to other genes (alncRNAs), intron-overlapping with protein-coding genes (ilncRNAs), small RNA (sRNA) hosts (shlncRNAs), enhancer-derived (elncRNAs), or pseudogene-derived (plncRNAs). Efforts to characterize lncRNAs so far have largely focused on lincRNAs, given that as non-overlapping transcriptional units they are readily identifiable and experimentally tractable (Ulitsky and Bartel 2013). Globally, lincRNAs are expressed at lower levels but in a more cell type-specific manner than mRNAs (Cabili et al. 2011), suggesting roles in lineage-specific development or in specialized cellular functions. Indeed, several lincRNAs have been implicated in modulating mammalian cell differentiation (Hu et al. 2012). The relative contributions of the full panorama of lncRNA classes to the same developmental process, however, remain poorly understood.

Here, we comprehensively characterize the landscape of lncRNAs expressed during red blood cell development *in vivo.* We use RNA-Seq to survey the poly(A)+ and poly(A)- RNA transcriptomes of differentiating E14.5 mouse fetal liver erythroid cells and identify 655 lncRNAs of various classes, including 132 previously unannotated loci with erythroid-restricted expression. We uncover ~100 lncRNAs with dynamic expression and chromatin patterns during differentiation, many of which are targeted by key erythroid TFs GATA1, TAL1 or KLF1. These include novel erythroid-specific lncRNAs found in the nucleus that show striking patterns of developmental stage specificity and are often conserved in human. Depleting 12 such candidates with shRNAs revealed critical roles in the transition from terminally differentiated erythroblasts to mature enucleated erythrocytes. Our data and workflow provide a roadmap for the identification of lncRNAs with roles in erythropoiesis, which can be readily implemented through an online resource (http://lodishlab.wi.mit.edu/data/lncRNAs/). Overall, our study provides a comprehensive catalog of erythroid lncRNAs and reveals several novel modulators of erythropoiesis.

Methods

Cell isolation, culture and terminal differentiation assays

Mouse fetal liver erythroid cell purification, culture and differentiation were conducted as described previously (Zhang et al. 2003; Hattangadi et al. 2010).

RNA-Seq and analysis

Total RNA was isolated from mouse fetal liver TER119+ or TER119- cells using the QIAGEN miRNeasy Kit. Ribosomal RNA was depleted using the Epicentre Ribo-Zero Gold Kit. Strand-specific sequencing libraries were prepared as described (Borodina et al. 2011) from the total, poly(A)+ or poly(A)- RNA fractions and sequenced using the Illumina HiSeq2000 platform. Paired-end RNA-Seq reads were mapped to the mouse genome (mm9 version) using TopHat (Trapnell et al. 2009) and transcripts were assembled *de novo* using Cufflinks (Trapnell et al. 2010). We also examined poly(A)+ RNA-Seq reads from purified BFU-E, CFU-E and TER119+ cells (Flygare et al. 2011), and from 30 cell and tissue types from the mouse ENCODE consortium (Stamatoyannopoulos et al. 2012) (supplemental Table 3). Gene-level expression for all datasets was quantified as fragments per kilobase of exon model per million mapped fragments (FPKM) using Cufflinks based on our *de novo* gene models. Differential gene expression was determined using DESeq (Anders and Huber 2010) with an FDR threshold of 5%. Further details can be found in supplemental Methods. RNA-Seq data from this study have been deposited at the National Center for Biotechnology Information Gene Expression Omnibus (repository number GSE52126).

lncRNA identification and classification

Once a *de novo* transcriptome from TER119+ and TER119- cells was assembled, to identify reliable lncRNA models we considered only multi-exonic transcripts and ran them through the following filters: (1) size selection, (2) empirical read coverage threshold, (3) known protein domain filter, (4) predicted coding potential threshold, and (5) overlap with known mRNA exon annotations filter (see Supplemental methods for specific details). To assign lncRNAs to specific classes, we examined their overlap with annotated genes from the Ensembl (Flicek et al. 2012), RefSeq (Pruitt et al. 2012), and UCSC Genome Browser (Dreszer et al. 2012) databases or with enhancers annotated in E14.5 fetal liver cells (Shen et al. 2012).

Single-molecule RNA FISH and analysis

RNA FISH was performed as described (Raj et al. 2008). Fluorescence microscopy, image acquisition and image analysis methods were previously published (Neuert et al. 2013). Further details can be found in supplemental Methods.

Retroviral transduction

Purified erythroid progenitors were transduced by MSCV-based retroviruses following previously described protocols (Hattangadi et al. 2010).

Flow cytometry and analysis

For all flow cytometry experiments, we gated on transduced cells (GFP+ subpopulation), for phenotypic analysis. The procedures for immunostaining and flow cytometry analysis of erythroid differentiation and enucleation were described previously (Ji et al. 2008; Hattangadi et al. 2010). Average cell size was quantified as the mean of the distribution of forward scatter pulse area measurements.

Results

Global discovery of lncRNAs expressed in fetal liver and erythroid cells

The mouse fetal liver is the primary site of erythropoiesis between embryonic days 12-16. To catalog lncRNAs expressed during fetal erythropoiesis *in vivo*, we used high-throughput sequencing to survey the long RNA transcriptome of E14.5 fetal liver cells, and characterized that of the erythroid lineage subpopulation. In brief, we used previously developed methods (Zhang et al. 2003; Hattangadi et al. 2010) to purify fetal liver TER119-positive cells, representing a pure population of differentiating hemoglobinizing erythroblasts, and TER119-negative cells, enriched for erythroid progenitors (~90%) but also containing cells from other hematopoietic lineages and niche cells (Zhang et al. 2003). We generated strand-specific, paired-end 100bp RNA-Seq reads of total RNA depleted of rRNA from both cell populations, and of poly(A)+ and poly(A)- RNA from TER119+ cells. In addition, we examined RNA-Seq reads of poly(A)+ RNA from FACS-purified fetal liver burst-forming unit erythroid (BFU-E) progenitors, colony-forming unit erythroid (CFU-E) progenitors, and TER119+ erythroblasts (Flygare et al. 2011). Using TopHat (Trapnell et al. 2009), we mapped in total >1 billion RNA-Seq reads to the mouse genome (supplemental Table 1), thus surveying the fetal erythroid differentiation transcriptome at unprecedented resolution. Transcripts were reconstructed *de novo* from these data using Cufflinks (Trapnell et al. 2010), and compared with Ensembl (Flicek et al. 2012), RefSeq (Pruitt et al. 2012), and UCSC (Dreszer et al. 2012) gene annotations.

To identify lncRNAs with high confidence, we considered only multiexonic transcripts and discarded any that overlap with known mRNA exons in the same strand or that have predicted coding potential based on three orthogonal approaches (Figure 1A; see supplemental Methods). First, used the phylogenetic codon substitution frequency (pCSF) metric (Lin et al. 2011) to filter out transcripts under evolutionary pressure to preserve synonymous amino acid codons. Second, we discarded any transcript with ORFs similar to those of known proteins or protein domains from the Pfam (Finn et al. 2010) or Refseq databases. Third, we used the coding potential calculator (CPC) metric (Kong et al. 2007) to remove any transcript with characteristic coding features, independent of their conservation.

Our stringent strategy yielded 800 lncRNAs from 655 loci, all of which have no predicted functional coding capacity (supplemental Figures 1A and B). In total, the transcriptome from TER119+ and TER119- cells contained 9512 known mRNA genes (~92%), 655 lncRNA genes (~6%), and 209 genes that have unclear coding capacity based on our criteria and were thus discarded from further analysis (supplemental Figure 1C). About 42.5% of the known mouse coding genome is expressed in erythroid cells, consistent with analogous observations in individual human cell lines (Djebali et al. 2012). Importantly, we identified 194 lncRNAs from 132 loci that were previously unannotated (Figure 1B), likely missed by previous databases due to their erythroid-specific expression (see below).

To classify the repertoire of fetal liver lncRNAs, we examined their overlap with annotated genes (Dreszer et al. 2012; Flicek et al. 2012; Pruitt et al. 2012) or enhancers (Shen et al. 2012) (Figure 1C) and systematically assigned them to lncRNA classes (supplemental Figure 1D and supplemental Methods). Our strategy identified 299 lincRNAs, 153 alncRNAs, 92 ilncRNAs, 52 elncRNAs, 27 shlncRNAs and 3 plncRNAs (Figure 1D and supplemental Table 2), as well as 29 lncRNAs that could not be classified. The majority of our intergenic, antisense, intronic, sRNA-hosting and pseudogene lncRNAs that are found in Ensembl are annotated as such (supplemental Figure 1F), thus validating our strategy. For enhancer lncRNAs, as expected (Heintzman et al. 2009; Creyghton et al. 2010) we found enrichment around the transcriptions start site (TSS) for H3K27Ac and H3K4me1 over H3K4me3, as well as for serine 5 phosphorylated RNA Pol II, in E14.5 fetal liver cells (supplemental Figure 1E).

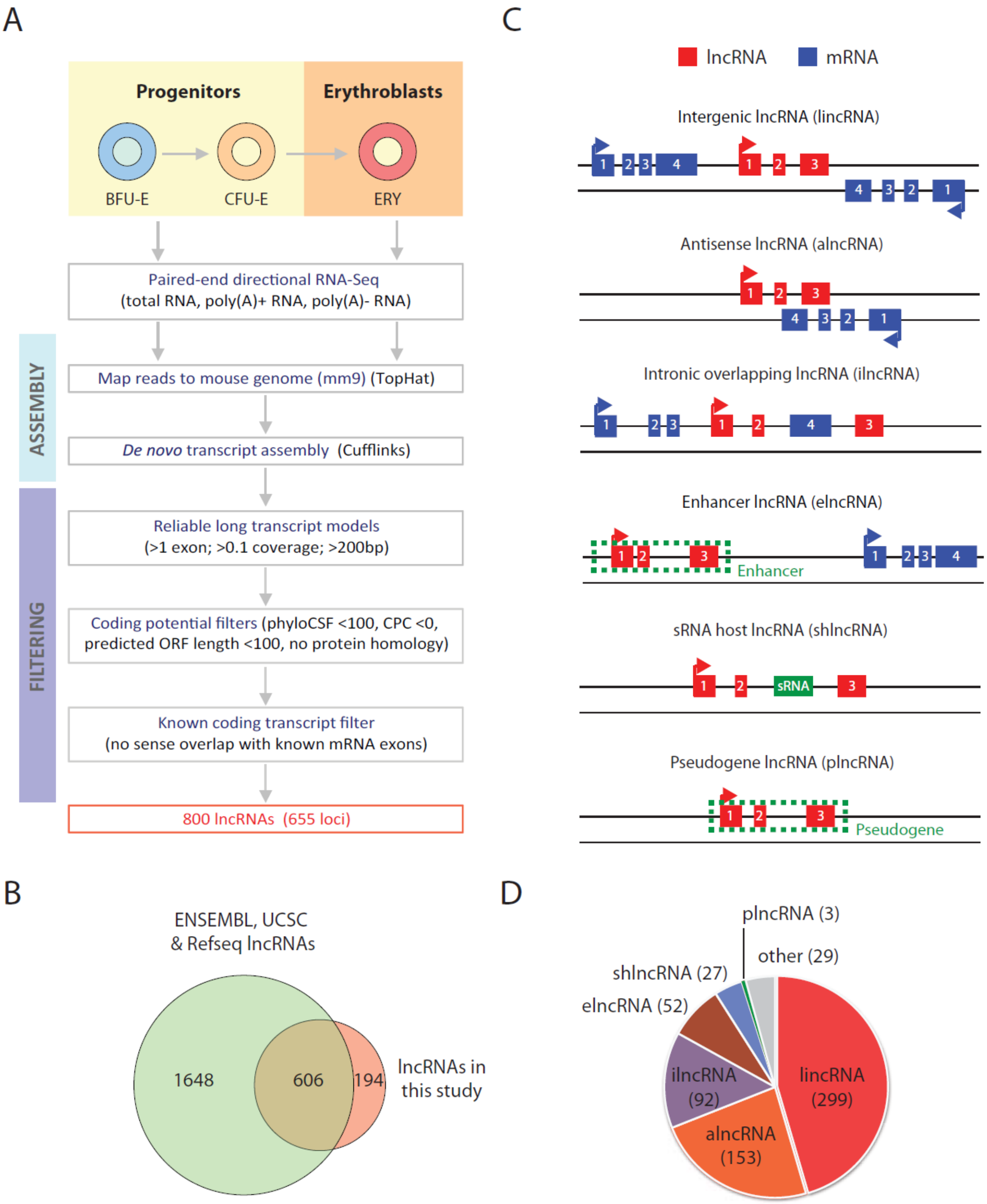

A
Progenitors
Erythroblasts
BFU-E
CFU-E
ERY
Paired-end directional RNA-Seq
(total RNA, poly(A)+ RNA, poly(A)- RNA)
ASSEMBLY
Map reads to mouse genome (mm9) (TopHat)
De novo transcript assembly (Cufflinks)
FILTERING
Reliable long transcript models
(>1 exon; >0.1 coverage; >200bp)
Coding potential filters (phyloCSF <100, CPC <0,
predicted ORF length <100, no protein homology)
Known coding transcript filter
(no sense overlap with known mRNA exons)
800 lncRNAs (655 loci)
C
🟥 lncRNA
🟦 mRNA
Intergenic lncRNA (lincRNA)
Antisense lncRNA (alncRNA)
Intronic overlapping lncRNA (ilncRNA)
Enhancer lncRNA (elncRNA)
Enhancer
sRNA host lncRNA (shlncRNA)
sRNA
Pseudogene lncRNA (plncRNA)
Pseudogene
B
ENSEMBL, UCSC
& Refseq lncRNAs
1648
606
194
lncRNAs in
this study
D
plncRNA (3)
other (29)
shlncRNA (27)
elncRNA (52)
ilncRNA
(92)
lincRNA
(299)
alncRNA
(153)

Figure 1. Identification of lncRNAs expressed in fetal liver and erythroid cells.

(**A**) Workflow for lncRNA discovery. See text and supplemental Methods for details.

(**B**) Overlap between lncRNAs annotated in Ensembl, UCSC or RefSeq databases and lncRNAs identified in this study.

(**C**) Definitions of different classes of lncRNAs based on their genomic region of origin.

(**D**) Distribution of 655 lncRNAs expressed in fetal liver into different lncRNA classes.

Structural features of fetal liver and erythroid lncRNAs

The majority of fetal liver-expressed lncRNAs are capped, RNA Pol II transcripts, as evidenced by specific enrichment for CAGE tags (Carninci et al. 2006; Faulkner et al. 2009) and for RNAPII occupancy around the TSS in >80% of them (Figure 2A). Consistent with previous studies (Guttman et al. 2009; Marques and Ponting 2009; Guttman et al. 2010; Cabili et al. 2011; Derrien et al. 2012), our lncRNAs exhibit 1-2 orders of magnitude lower expression levels than mRNAs, except for shlncRNAs, whose expression range spans six orders of magnitude (Figure 2B). In addition, lncRNAs are significantly enriched in the poly(A)- fraction relative to mRNAs ($p<10^{-8}$- $p<10^{-15}$, Kolmogorov-Smirnov [KS] test) (Figure 2C), especially intronic, sRNA-hosting and pseudogene lncRNAs. For 40-60% of lncRNAs detected in at least one of our poly(A)+ datasets we found specific enrichment for poly(A)-seq tags (Derti et al. 2012) (Figure 2A), supporting our annotation of 3' ends.

Structurally, lncRNAs have fewer exons than mRNAs and are thus generally shorter, except for shlncRNAs, which often resemble mRNAs in length due to larger exons (Figures. 2D and E). shlncRNAs also exhibit 2-3 times more isoforms than other lncRNAs (Figure 2F), consistent with intron retention or exon trimming during sRNA processing. Given our sequencing depth and experimental support of transcript boundaries, these class-specific traits are unlikely to be artifacts of incomplete transcript detection or assembly.

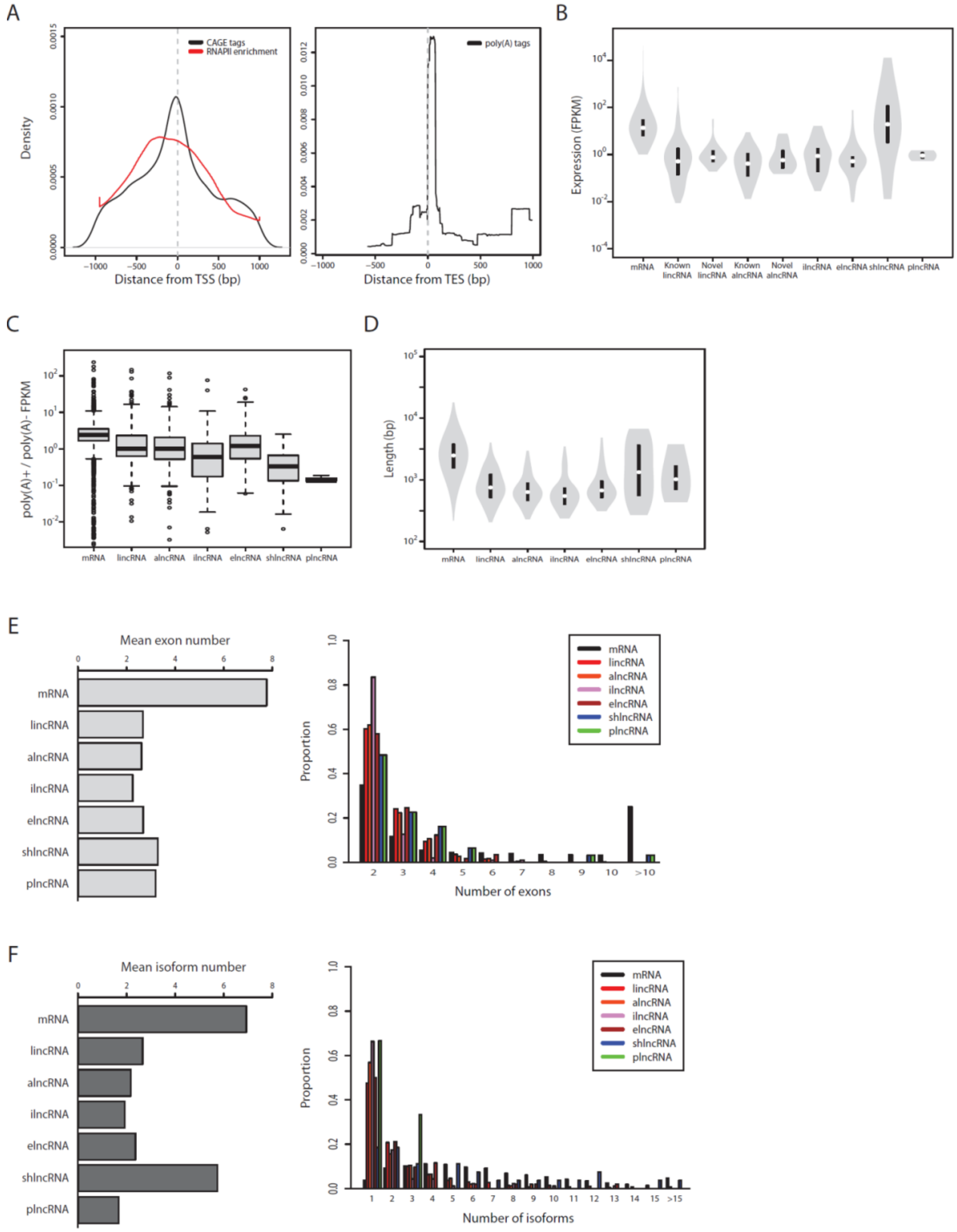


Figure 2. Structural features of fetal liver-expressed lncRNAs.

(**A**) (Left) density of CAGE tags and RNA Pol II enrichment within lncRNA TSS ± 1 kb regions overlapped by these marks (>80% across lncRNA classes). (Right) density of poly(A) sequencing tags within lncRNA TES ± 1 kb regions overlapped by these tags (40-60% across lncRNA classes).

(**B**) Violin plots of gene-level expression (FPKM) distributions for mRNAs and lncRNAs.

(**C**) Ratio of gene-level expression values (FPKM) in poly(A)+ vs. poly(A)- RNA fractions for mRNAs and lncRNAs.

(**D**) Violin plots of length distributions for mRNA and lncRNA transcripts.

(**E**) Number of exons in mRNA or lncRNA transcripts. Mean (left panel) and distribution (right panel) of number of exons per transcript for each transcript type are shown. Transcripts with >10 exons are pooled together in the last distribution category.

(**F**) Number of isoforms in mRNA or lncRNA genes. Mean (left panel) and distribution (right panel) of number of isoforms per locus for each gene type are shown. Transcripts with >15 isoforms are pooled together in the last category.

Widespread conservation of fetal liver-expressed lncRNAs

We reasoned that if fetal liver-expressed lncRNAs are functionally relevant for cellular development, they should be evolutionary conserved. To test this, we looked for sequence conservation at the DNA level across 30 vertebrate genomes as measured by phastCons (Siepel et al. 2005) (Figure 3A). We find that promoter conservation across all lncRNA classes is essentially indistinguishable from that of mRNAs, whereas lncRNA exons are generally less conserved in primary sequence than mRNA exons, but more so than size-matched control intergenic regions. We note that the exons shlncRNAs tend to be better conserved than those of other lncRNA families ($p<10^{-5}$-$p<10^{-3}$, Wilcoxon test), consistent with a capacity to host widely conserved sRNAs. To further investigate the evolutionary trajectories of our lncRNAs, we systematically searched for orthologous genomic regions across 19 vertebrate genomes via pairwise alignments using BLAT (Kent 2002) (Figure 3B; see supplemental Methods). For 80-95% of mouse lncRNAs of all classes we identified putative orthologous regions in at least one other vertebrate genome (median 5-9 orthologs), including 52-71% conserved between mouse and human, and a handful conserved from zebrafish to man. Consistent with previous observations (Ulitsky et al. 2011; Derrien et al. 2012), most lncRNA loci identified in the mouse fetal liver appear to have emerged among mammals, and a considerable proportion of them (13-32%) appear specific to the rodent lineage. Interestingly, lincRNAs seem to be the fastest-evolving type of lncRNA, whereas shlncRNAs are widely conserved throughout mammals (Figure 3C).

To obtain evidence that our putative orthologous lncRNA loci are expressed, we examined a catalog of vertebrate transcripts syntenically mapped to the mouse genome by TransMap (Zhu et al. 2007) (Figure 3D). We found evidence of orthologous expression for 18-

32% of lncRNAs across different families except for shlncRNAs, for which 87% had an expressed ortholog in another species (Figure 3E). This may be due not just to the higher conservation of shlncRNAs but also to their higher expression, which facilitates detection and thus cross-species mapping. Expressed orthologs fell predominantly among the better profiled genomes, such as rat and human, and thus may improve with greater coverage of other species.

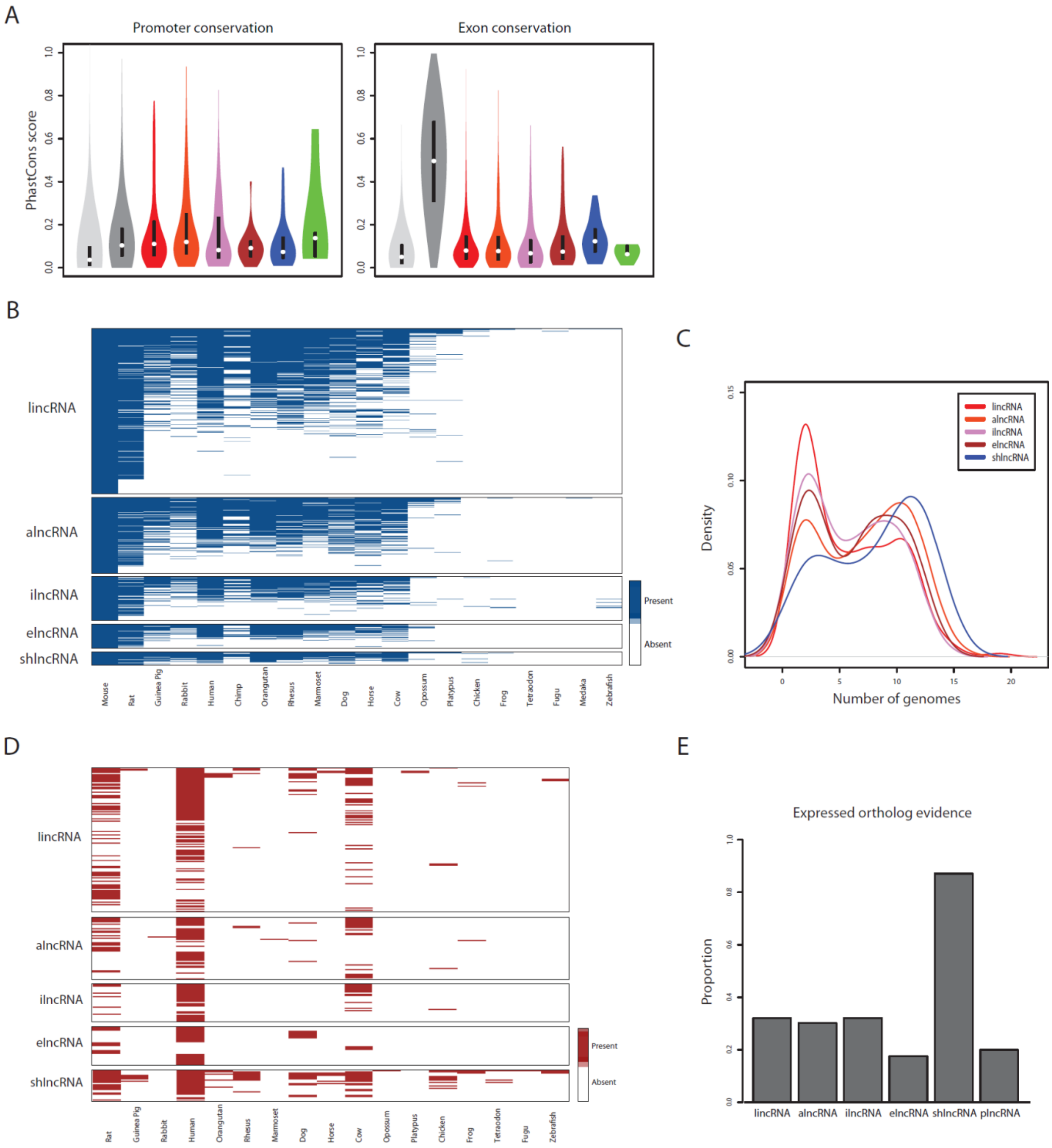


Figure 3. Conservation of lncRNAs expressed in murine fetal liver.

(**A**) Sequence conservation of mRNA and lncRNA promoters (left) and exons (right), across 30 vertebrate genomes as measured by PhastCons. For each lncRNA transcript, a control was generated by shuffling its exon-intron structure to a randomly chosen region of intergenic space within the same chromosome (n=1000 random shuffles).

(**B**) Orthologous genomic regions of mouse lncRNAs identified across 19 vertebrate genomes (see supplemental Methods). For each genome, detected orthologs are indicated in blue.

(**C**) Breadth of lncRNA ortholog conservation across 20 vertebrate genomes.

(**D**) Expressed orthologous transcripts of mouse lncRNAs identified across 17 vertebrate genomes. Expressed orthologs were identified from a catalog of vertebrate transcripts mapped to the mouse genome by TransMap (see Supplemental methods). For each genome, detected expressed orthologs are indicated in red.

(**E**) Mouse lncRNAs with evidence of at least one expressed ortholog in another species.

High tissue specificity among fetal liver and erythroid lncRNAs

Our high-resolution transcriptomic survey of TER119+ and TER119- cells covers genes from erythroid cells and from minor cell populations of other lineages. Moreover, genes expressed at basal levels in fetal liver cells may be more characteristic of other tissues. To globally examine tissue specificity, we quantified the expression of fetal liver-expressed genes across a compendium of 30 primary cell and tissue types purified, sequenced and analyzed using common guidelines for the mouse ENCODE consortium (Stamatoyannopoulos et al. 2012) (Figure 4; supplemental Table 3). For each gene, we scored the specificity of its expression in a given tissue as the fraction of the total expression across tissues that it represents, equivalent to its fractional expression level. Based on these scores, we find exquisite patterns of tissue-specific enrichment for both mRNAs and lncRNAs, including genes highly specific to each of the hematopoietic lineages examined (Figure 4A). As expected (Cabili et al. 2011; Derrien et al. 2012), lncRNAs show greater tissue specificity than mRNAs ($p<10^{-15}$, Kolmogorov-Smirnov (KS) test), which holds true across lncRNA classes and matched expression ranges (supplemental Figures 2A and B). However, there were notable differences in tissue specificity patterns among lncRNA families (Supplemental Figures 2A and C). For example, elncRNAs are the only class enriched in both fetal erythroblasts and in the adult bone marrow, suggesting that they originate from enhancers active at both fetal and adult stages.

We next focused on the subset of genes showing erythroid-specific expression, which we defined as tissue-restricted genes having fetal erythroblasts as the tissue with the maximal tissue specificity score (highlighted in Figure 4A; see Methods). These genes comprised 7-8% of fetal liver-expressed mRNAs and lncRNAs, respectively. Previously unannotated lncRNAs are enriched in this group relative to annotated ones (supplemental Figure 2D), highlighting the importance of our focus on erythroid cells. In contrast, broadly expressed lncRNAs are mostly depleted or expressed at background levels in fetal erythroblasts (supplemental Figure 2E). By way of example, shown in Figure 4B are four lncRNAs -shlncRNA-EC6, elncRNA-EC1, lincRNA-EC9, and alncRNA-EC3, that we focus on later on because of their erythroid specificity, promoter targeting by erythroid TFs, high expression in erythroblasts, and induction during terminal differentiation (see below).

elncRNA-EC1, lincRNA-EC9, and alncRNA-EC3 are expressed in erythroblasts but not in the closely related megakaryocyte or megakaryocyte-erythroid progenitor or in other tissues examined (Figure 4B). In contrast, the shlncRNA-EC6 locus, encoding the known lncRNA DLEU2, is broadly expressed but it is transcribed from a different promoter in erythroblasts vs. other cell types, including closely related lineages (Figure 4B). Interestingly, processing of this erythroid-restricted isoform, presumably to release microRNAs 16-1 and 15a from the poly(A)+ precursor, generates mature poly(A)+ and poly(A)- transcripts of similar stability, as evidenced by comparable levels in their respective RNA fractions (see also supplemental Figure 7A).

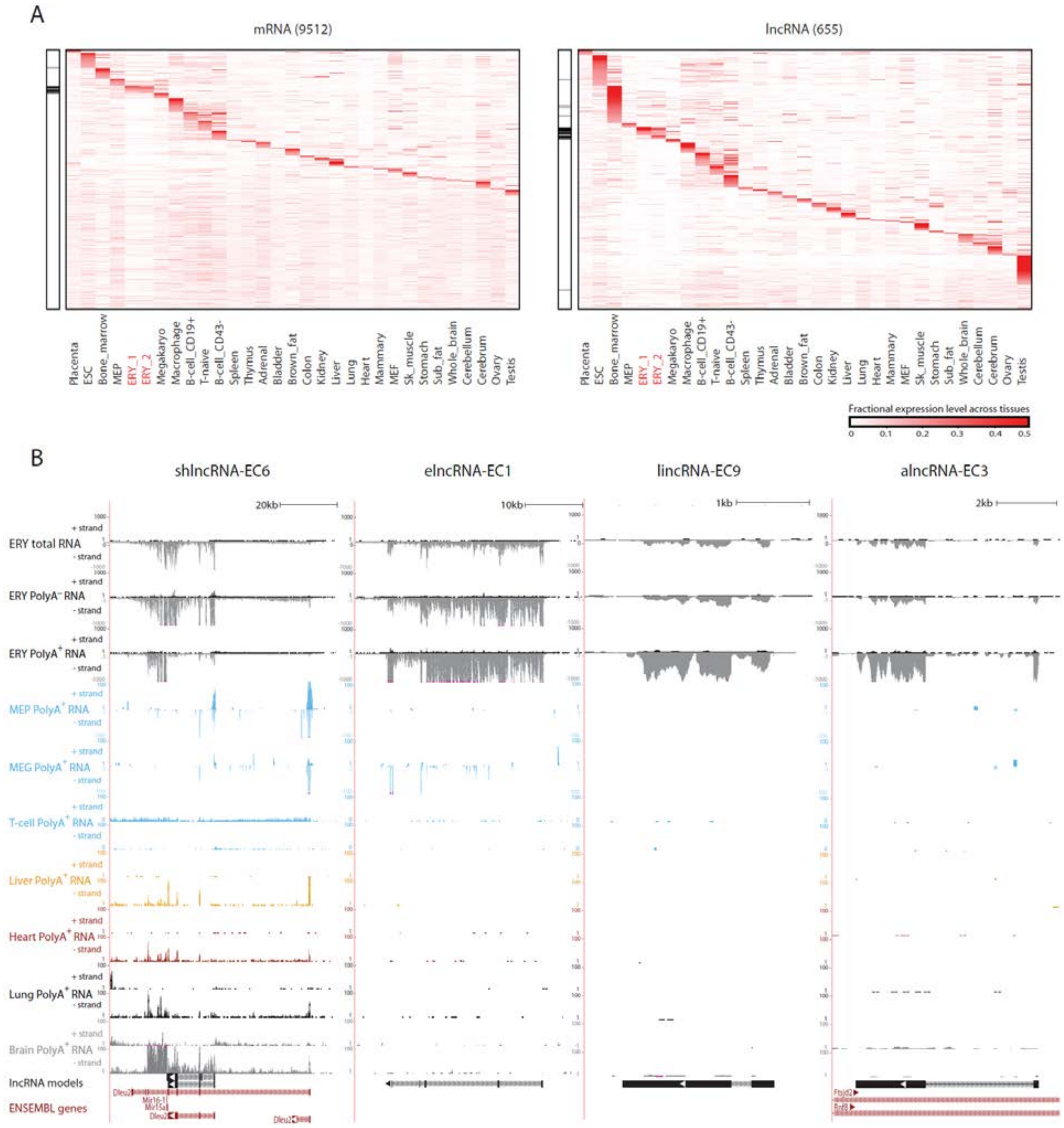


Figure 4. Tissue specificity of fetal liver and erythroid lncRNAs.

(**A**) Relative abundance of mRNA and lncRNA genes (rows) expressed in fetal liver across 30 primary cell and tissue types from the mouse ENCODE consortium (columns). Color intensity represents the fractional gene-level expression across all tissues examined. ERY_1 and ERY_2 (red) are fetal liver TER119+ erythroblast replicates. Tissue expression was quantified based on gene models from our *de novo* assembly using Cufflinks. Black bars in the left panels highlight empirically defined erythroid-restricted genes.

(**B**) Examples of erythroid-enriched lncRNA loci. These loci were selected based on their expression, regulation and tissue specificity features (see text). Images from the UCSC Genome Browser depict RNA-Seq signal as the density of mapped strand-specific RNA-Seq reads. The plus strand (transcribed left to right) and minus strand (transcribed right to left) are denoted to the left of the tracks. Tracks 1-6 show in black the RNA-Seq signal of total, poly(A)- or poly(A)+ RNA from fetal liver TER119+ erythroblasts (ERY). Tracks 7-12 depict in light blue the RNA-Seq signal of poly(A)+ RNA from other hematopoietic cells: adult megakaryocyte-erythroid progenitors (MEP), fetal megakaryocytes (MEG), and adult T-naïve cells (T-cell), all from the ENCODE consortium. Tracks 13-20 show the RNA-Seq signal of poly(A)+ RNA from other tissues from the ENCODE consortium: adult liver (yellow), adult heart (red), adult lung (black) and E14.5 whole brain (gray). The bottom tracks depict lncRNA transcript models inferred by *de novo* assembly using Cufflinks (black), and Ensembl gene annotations (red). Left-to-right arrows indicate transcripts in the plus strand; right-to-left arrows indicate transcripts in the minus strand. Note that all lncRNA transcripts shown are transcribed in the minus strand.

lncRNAs are dynamically regulated during erythropoiesis

To examine the regulation of lncRNAs during erythropoiesis, we focused on the subset of fetal liver-expressed genes that are reliably detected in FACS-purified BFU-Es, CFU-Es or TER119+ erythroblasts. We considered only genes expressed in both replicates of at least 1 stage from BFU-Es to erythroblasts, resulting in 275 erythroid-expressed lncRNAs (supplemental Table 4), and then used DESeq (Anders and Huber 2010) to identify differentially expressed ones ($p<0.05$, DESeq test). Validating our approach, we identified 728 differentially expressed mRNAs increasing >2-fold between progenitors and erythroblasts that as expected encode proteins enriched for erythroid-specific roles (supplemental Figure 3A). We also identified 96 lncRNAs of various classes that are differentially expressed during erythropoiesis (Figure 5A). These comprised mostly lncRNAs that become strongly induced as progenitors differentiate into erythroblasts, as exemplified by the four lncRNAs examined in Figure 4B (Figure 5B). Importantly, we find that lncRNAs exhibit greater expression variability than mRNAs through the stages of differentiation, with novel erythroid-enriched lncRNAs being more dynamically expressed than previously annotated ones (supplemental Figures 3B and C). The regulated expression and high differentiation-stage specificity of these lncRNAs suggests their potential involvement in red blood cell development.

Coordination of lncRNA expression and chromatin dynamics during erythropoiesis

To investigate how differentially expressed lncRNAs are regulated at the chromatin level, we examined their histone modification and RNA Pol II occupancy profiles, as determined by ChIP-Seq in fetal liver erythroid-progenitor cells and TER119+ erythroblasts (Wong et al. 2011).

Histone marks included those associated with active gene promoters (H3K4me2, H3K4me3, H4K16Ac and H3K9Ac), with repressed ones (H3K27me3) and with transcription elongation along gene bodies (H3K36me3, H3K79me2). As expected (Wong et al. 2011; Derrien et al. 2012), similar histone mark distributions are found around the TSS of mRNA and lncRNA loci, but not control intergenic regions, in both progenitors and erythroblasts (supplemental Figure 4). We find that, as seen with mRNAs, quantitative changes in the levels of RNA Pol II and chromatin activation marks (within 2 kb of the TSS) or elongation marks (within gene bodies) correlate with changes in lncRNA expression (supplemental Figure 5). In contrast, the repressive H3K27me3 mark was uncorrelated with expression changed and was generally depleted near active TSSs. The most predictive activation and elongation histone marks were H3K4me2 and H3K79me2, respectively (Pearson's $r=0.71$, $p<10^{-12}$ and $r=0.46$, $p<10^{-4}$, Fisher's exact test), shown for the highly induced lncRNAs in Figure 5B. By contrast, repressive H3K27me3 marking was generally depleted near active TSSs, as exemplified at the shlncRNA-EC6 locus, where the repressed, distal promoter is marked by H3K27me3. Notably, H3K27me3 marking of this promoter is already established at the committed progenitor stage (Figure 5B).

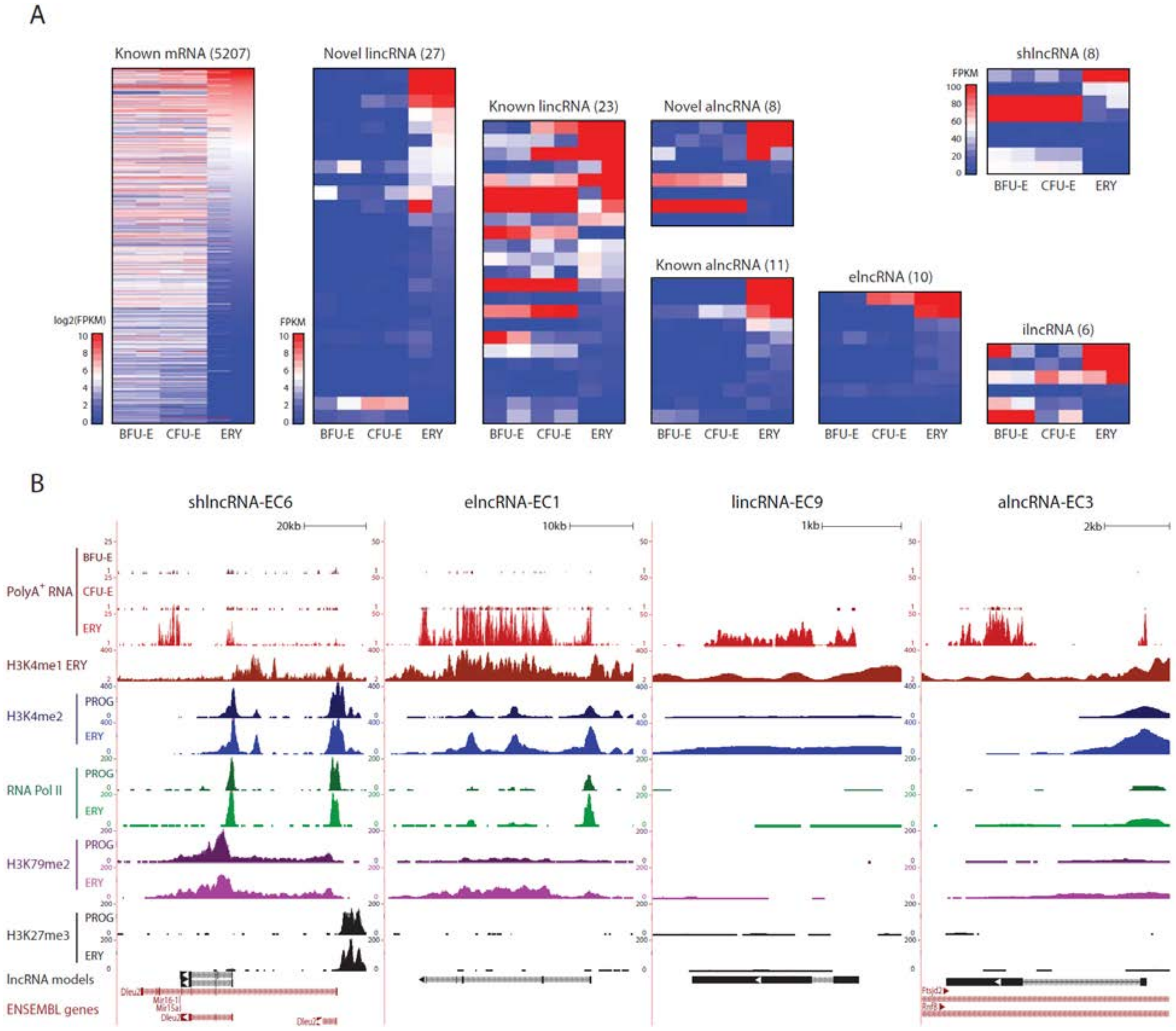


Figure 5. Dynamic expression patterns of lncRNAs during erythroid differentiation.

(**A**) Abundance of mRNAs and lncRNAs that are differentially expressed during erythropoiesis, as determined by DESeq at a 5% false discovery threshold. Shown are absolute gene expression estimates (FPKM) from poly(A)+ RNA-Seq of FACS-purified BFU-Es, CFU-Es and TER119+ erythroblasts (ERY) (2 replicates each), based on gene models from our *de novo* assembly using Cufflinks.

(**B**) Examples of differentially expressed lncRNA loci, the same RNAs as in Figure 2B. Images from the UCSC Genome Browser depict RNA-Seq signal as the density of mapped RNA-Seq reads and ChIP-Seq signal as the density of processed signal enrichment. Tracks 1-3 show in red the non-strand-specific RNA-Seq signal of poly(A)+ RNA from FACS-purified fetal liver BFU-Es, CFU-Es and TER119+ erythroblasts (ERY). Tracks 4-12 depict the ChIP-Seq signal for: H3K4me1, a chromatin mark enriched in promoter and enhancer regions, in ERY (dark red); H3K4me2, associated with transcriptional activation, in erythroid progenitor-enriched fetal liver cells (PROG) and ERY (dark and light blue); serine 5 phosphorylated RNA Pol II, enriched at the TSS of active genes, in PROG and ERY (dark and light green); H3K79me2, associated with transcriptional elongation, in PROG and ERY (dark and light purple); and H3K27me3, associated with transcriptional repression, in PROG and ERY (black). The bottom tracks depict lncRNA transcript models and Ensembl gene annotations as in Figure 2B.

lncRNAs are targets of core erythroid transcription factors

We reasoned that if differentially expressed lncRNAs play roles during erythropoiesis, they should be targeted by erythroid-important TFs. To test this, we examined genome-wide maps of GATA1, TAL1 and KLF1 occupancy determined by ChIP-Seq in fetal liver TER119+ erythroblasts (Pilon et al. 2011; Wu et al. 2011). Binding sites for these factors, inferred by MACS (Zhang et al. 2008) (empirical FDR<0.05), were intersected with the promoter-proximal regions (TSS ± 1 kb) of differentially expressed lncRNAs or mRNAs. As expected (Wontakal et al. 2012), all three factors co-occupied the promoters of 278 mRNA genes differentially expressed during erythropoiesis, including 136 that are upregulated >2-fold and encode proteins with erythroid-specific roles (Figure 6A; supplemental Figure 6A), thus validating our approach. We then found that 60 out of 96 differentially expressed lncRNAs are indeed bound at their promoters by GATA, TAL1 or KLF1 in erythroblasts (Figure 6A). We also found that promoter-proximal co-occupancy by GATA1 and TAL1 for mRNAs and lncRNAs is significantly associated with gene induction ($p<10^{-15}$ and $p<10^{-9}$, respectively, Wilcoxon test) and promoter H3K4me2 marking ($p<10^{-15}$ and $p<10^{-4}$, respectively, KS test) (Figure 6B; supplemental Figure 6B), extending previous observations of mRNAs (Yu et al. 2009; Wu et al. 2011; Wontakal et al. 2012). In contrast, proximal binding by KLF1 alone seems to be a poor predictor of gene expression change (Figure 6B). Binding peaks for GATA1, TAL1 and KLF1 are shown for our lncRNA models in Figure 6C. The coincidence of these peaks with DNAse I hypersensitive sites (Figure 6C) and with RNA pol II binding and active chromatin marks (Figure 5B) in these genes strongly supports their specific regulation by these factors. Importantly, both TF binding events and chromatin architecture are conserved in the human K562 erythroleukemia cell line for the DLEU2 human ortholog and for the putative ortholog of elncRNA-EC1, inferred by local alignment and synteny (supplemental Figure 6D; see supplemental Methods). For these two genes, we obtained direct evidence of regulation by GATA1 by confirming their time-dependent activation after GATA1 restoration in the mouse G1E-ER4 cell line (Weiss et al. 1997; Welch et al. 2004) (supplemental Figure 6C).

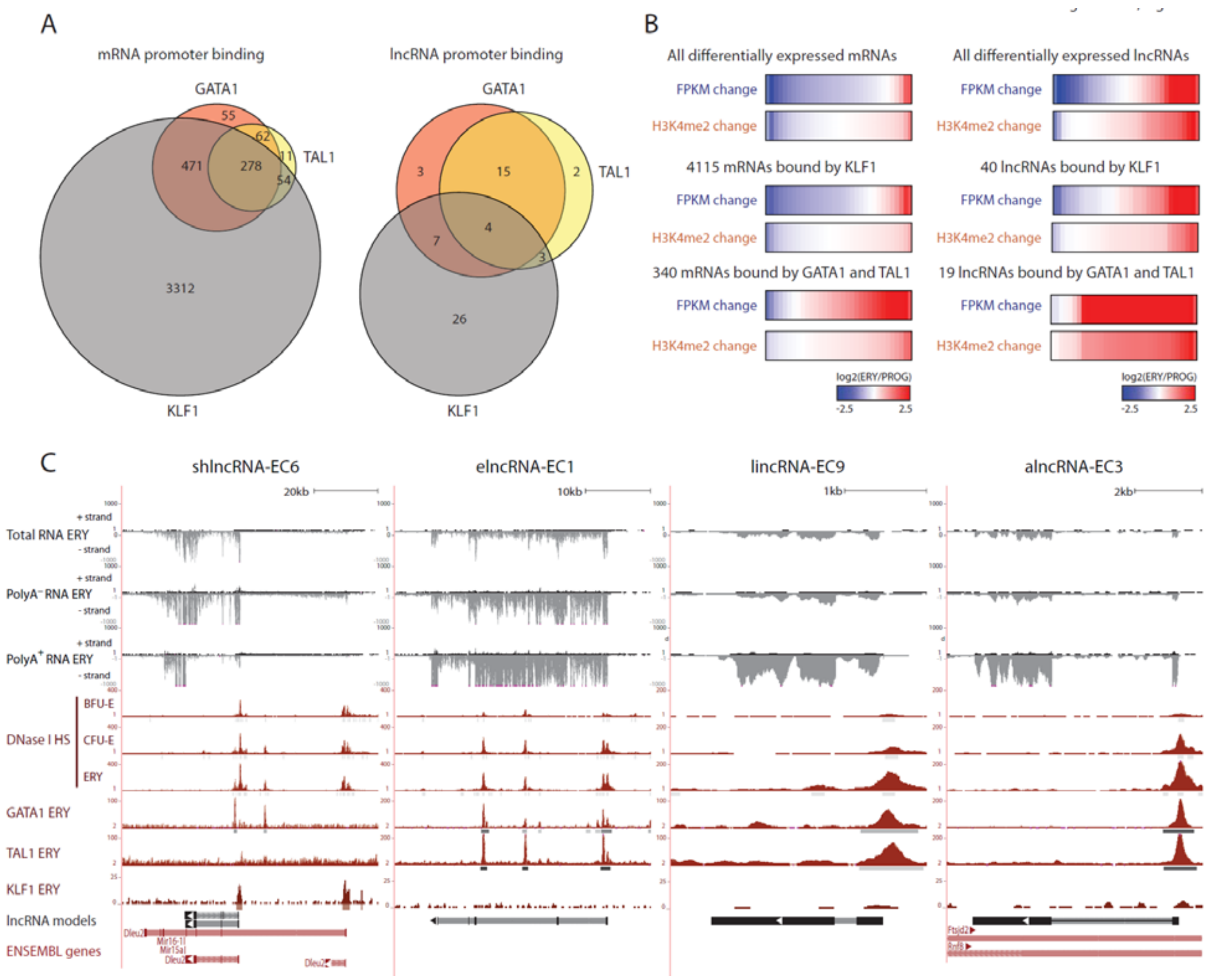


Figure 6. lncRNAs are targeted by core erythroid transcription factors.

(**A**) Binding of GATA1, TAL1 and KLF1 transcription factors within promoter-proximal regions (TSS ± 1 kb) of mRNAs (left) and lncRNAs (right) that are differentially expressed during erythropoiesis (see text).

(**B**) Changes in expression and promoter-proximal (TSS ± 1 kb) H3K4me2 levels for all differentially expressed mRNA or lncRNA genes, for the subset bound by KLF1 or for those bound by both GATA and TAL1. Changes are shown as the log2 ratio of the levels in TER119+ erythroblasts (ERY) to the levels in erythroid progenitor-enriched fetal liver cells (PROG).

(**C**) Examples of differentially expressed lncRNA loci that are bound proximally by GATA1, TAL1 or KLF1, the same RNAs as in Figures 2B and 3B. Images from the UCSC Genome Browser depict RNA-Seq signal as the density of mapped RNA-Seq reads, DNAse I hypersensitivity (HS) signal as the density of mapped sequencing tags, and ChIP-Seq signal as the density of processed signal enrichment. Tracks 1-6 show in black the strand-specific RNA-Seq signal in the plus strand or minus strand (denoted to the left of the tracks) of total, poly(A)-

or poly(A)+ RNA from fetal liver TER119+ erythroblasts (ERY). Tracks 7-9 depict in red the signal for DNAse I HS, associated with open chromatin, in BFU-Es, CFU-Es and ERY. Tracks 10-12 show in red the ChIP-Seq signal for GATA1, TAL1 and KLF1, respectively, in ERY. Peaks of signal enrichment are shown in grey under the DNAse I HS tracks (determined by I-max, empirical FDR<1%) and under the GATA1, TAL1 and KLF1 tracks (determined by MACS, empirical FDR<5%). The bottom tracks depict lncRNA transcript models and Ensembl gene annotations as in Figure 2B.

Validation of candidate lncRNAs reveals nuclear localization and complex developmental patterns

To select candidates for functional studies, we devised a strategy that integrates the experimental and computational analyses described above to stringently identify lncRNAs likely to play roles in erythropoiesis (see supplemental Methods). Briefly, we focused on differentially expressed lncRNAs with independent evidence of active transcription. Next, we required that they be targets of GATA1, TAL1 or KLF1 or that their expression be erythroid-specific. Finally, we ranked them first by their relative fold increase in expression between progenitors and erythroblasts and then by their absolute expression in erythroblasts. This strategy yielded 6 lincRNAs, 4 alncRNAs, 2 elncRNAs and 1 shlncRNA as the top candidate modulators of erythropoiesis. The expression, regulation and conservation features of these candidates are summarized in Figures 7A and B, and are shown in detail in supplemental Figure 7. Ranked 4th among them is LincRNA-EPS, which we previously found to promote erythroid differentiation by preventing apoptosis (Hu et al. 2011), thus validating our approach. Several of the new candidates are previously unannotated lncRNAs that are induced to similar levels as LincRNA-EPS during terminal differentiation. These include erythroid-specific ones that are co-targeted by GATA1 and TAL1 (Figure 6C; supplemental Figures 7B, G and J), but also more broadly expressed ones that are targeted by KLF1 instead (supplemental Figures 7E, F and I). We note that lincRNA-EC4, not targeted proximally by any of these factors, has been independently characterized recently (Lincred1 in (Tallack et al. 2012)) and shown to be targeted distally by KLF1. Overall, the expression and regulation features of our candidates in erythroblasts are generally absent from megakaryocytes (Stamatoyannopoulos et al. 2012) (Figure 7A), suggesting erythroid-specific functions.

To validate our candidates, we used quantitative real-time PCR (qPCR) to measure expression across hematopoietic lineages purified from fetal liver or adult bone marrow (Figure 7C). Consistent with our RNA-Seq data, we confirmed hematopoietic tissue specificity for all of them except for lincRNA-EC4 (also enriched in brain, heart and kidney). Strikingly, most of our lincRNAs are enriched in fetal liver but not adult erythroblasts (Figure 7C). By contrast, alncRNA-EC3 and the two elncRNAs examined are enriched in both fetal and adult erythroblasts. A third pattern was apparent for shlncRNA-EC6, lincRNA-EC8 and the remaining 3 alncRNAs, which are expressed during both fetal and adult hematopoiesis but are only erythroid-enriched at the fetal stage. Thus, lncRNA expression can be highly specific to developmental stage, even within the same cell lineage.

We next used single-molecule RNA fluorescence *in situ* hybridization (smFISH) (Raj et al. 2008) to visualize lncRNA transcripts in FACS-purified fetal liver TER119+ erythroblasts (Figure 7D). These experiments revealed our candidates to be predominantly nuclear, which we confirmed by cellular fractionation followed by qPCR (supplemental Figure 8). smFISH also indicated mean lncRNA levels of ~1-10 transcripts per cell, consistent with previous measurements in mouse and human (Khalil et al. 2009; Wang et al. 2011).

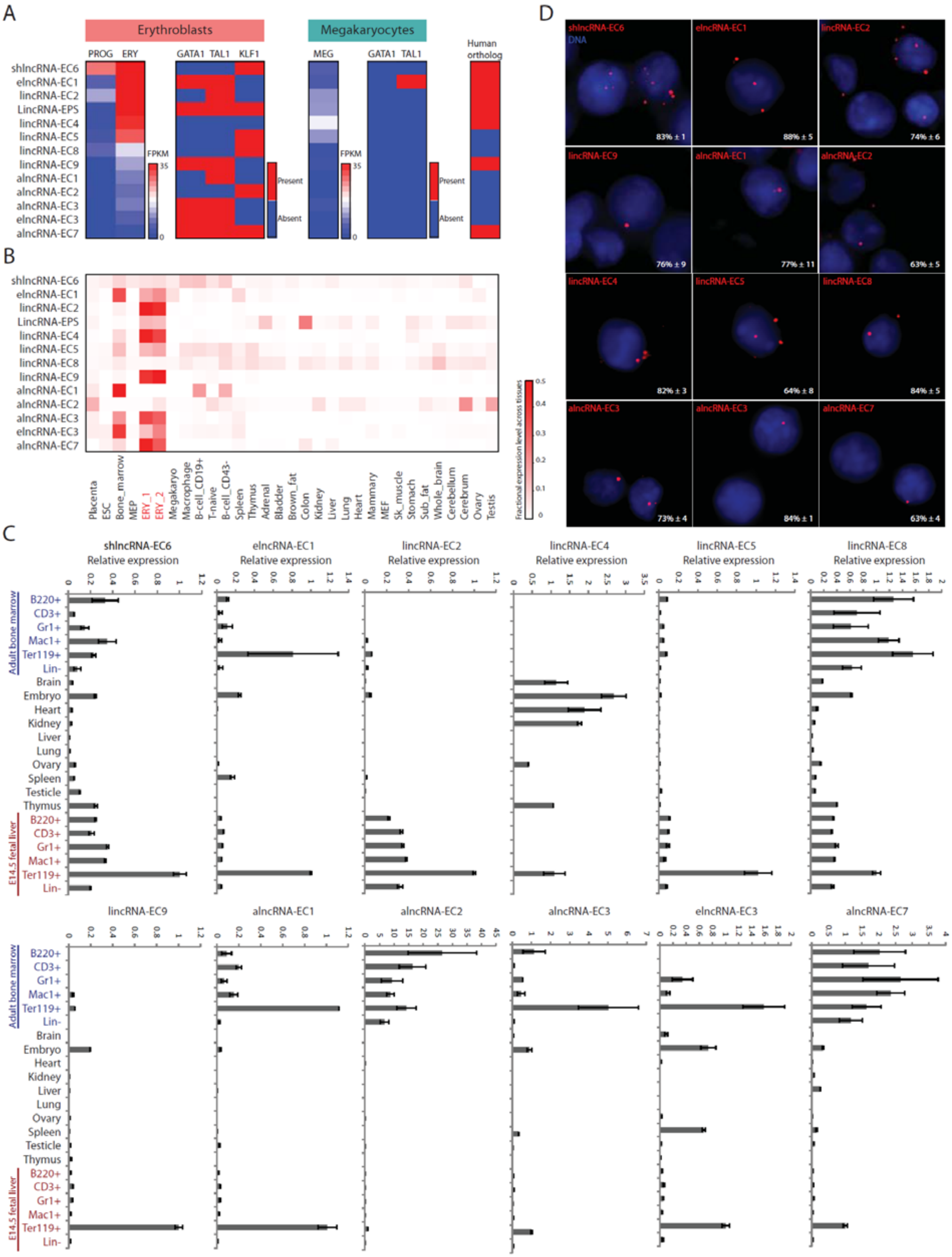

A
Erythroblasts
Megakaryocytes
PROG
ERY
GATA1
TAL1
KLF1
MEG
GATA1
TAL1
Human ortholog
shlncRNA-EC6
elncRNA-EC1
lincRNA-EC2
LincRNA-EPS
lincRNA-EC4
lincRNA-EC5
lincRNA-EC8
lincRNA-EC9
alncRNA-EC1
alncRNA-EC2
alncRNA-EC3
elncRNA-EC3
alncRNA-EC7
FPKM
35
0
Present
Absent
B
Placenta
ESC
Bone_marrow
MEP
ERY_1
ERY_2
Megakaryo
Macrophage
B-cell_CD19+
T-naive
B-cell_CD43-
Spleen
Thymus
Adrenal
Bladder
Brown_fat
Colon
Kidney
Liver
Lung
Heart
Mammary
MEF
Sk_muscle
Stomach
Sub_fat
Whole_brain
Cerebellum
Cerebrum
Ovary
Testis
Fractional expression level across tissues
0.5
0.4
0.3
0.2
0.1
0
C
shlncRNA-EC6
elncRNA-EC1
lincRNA-EC2
lincRNA-EC4
lincRNA-EC5
lincRNA-EC8
Relative expression
lincRNA-EC9
alncRNA-EC1
alncRNA-EC2
alncRNA-EC3
elncRNA-EC3
alncRNA-EC7
Adult bone marrow
B220+
CD3+
Gr1+
Mac1+
Ter119+
Lin-
Brain
Embryo
Heart
Kidney
Liver
Lung
Ovary
Spleen
Testicle
Thymus
E14.5 fetal liver
B220+
CD3+
Gr1+
Mac1+
Ter119+
Lin-
D
shlncRNA-EC6
DNA
83% ± 1
elncRNA-EC1
88% ± 5
lincRNA-EC2
74% ± 6
lincRNA-EC9
76% ± 9
alncRNA-EC1
77% ± 11
alncRNA-EC2
63% ± 5
lincRNA-EC4
82% ± 3
lincRNA-EC5
64% ± 8
lincRNA-EC8
84% ± 5
alncRNA-EC3
73% ± 4
alncRNA-EC3
84% ± 1
alncRNA-EC7
63% ± 4

Figure 7. Selection and validation of lncRNA targets.

(**A**) Summary of expression, regulation and conservation features of the top candidate lncRNA modulators of erythropoiesis (see text). Expression: shown are absolute gene expression estimates (FPKM) from RNA-Seq of total RNA from erythroid progenitor-enriched fetal liver cells (PROG) and TER119+ erythroblasts (ERY) or of poly(A)+ RNA from primary megakaryocytes (MEG), quantified as in Figure 3. Regulation: heatmaps represent whether promoter-proximal binding by GATA1, TAL1 of KLF1, analyzed as in Figure 4, is seen in ERY or MEG. Conservation: heatmap represents whether an orthologous region, identified by local alignment and synteny, is found in the human genome (see supplemental Methods for details).

(**B**) Relative abundance of the top lncRNA candidates across 30 mouse primary tissue and cell types from ENCODE, determined as in Figure 2. Color intensity represents the fractional expression level across all tissues examined. ERY_1 and ERY_2 (red) are TER119+ fetal liver erythroblast biological replicate experiments.

(**C**) Relative expression of the top lncRNA candidates across mouse organs and cells of different tissues and developmental stages, as determined by qPCR. Expression levels were normalized to those of 18S rRNA, and fold-changes were calculated relative to fetal TER119+ erythroblast levels. Data are shown as mean ± s.e.m (n = 3).

(**D**) Detection of individual lncRNA transcripts by single-molecule RNA FISH. Shown are maximum z-stack projections of fluoresce microscopy images of fixed TER119+ erythroblasts hybridized to singly-labeled RNA FISH probes. lncRNA molecules are pseudocolored red and DAPI-stained nuclei are pseudocolored blue. For each panel, the mean ± s.e.m (n = 2) percent of nuclear-localized transcripts is shown at the bottom right corner.

lncRNAs of multiple classes regulate red cell maturation

To conduct loss-of-function experiments, we generated three shRNAs for each lncRNA with the exception of alncRNA-EC1, where only one shRNA was possible due to extensive repeats. For alncRNAs, shRNAs were designed to target regions that either do not overlap the transcript on the opposite strand or only overlap its introns. We introduced each shRNA via retroviral transduction into lineage-negative fetal liver cells, which are enriched for erythroid progenitors (Flygare et al. 2011), and then cultured them in erythropoietin-containing media to induce *ex vivo* terminal proliferation and differentiation (Zhang et al. 2003). We confirmed efficient knockdown of all candidates (mean 27.5-99.3% knockdown per shRNA) (supplemental Figure 9). Flow cytometric analysis was then performed to evaluate three hallmarks of erythropoiesis: expression of the TER119 marker, cell size reduction and subsequent enucleation (Ji et al. 2008). Knockdown of each lncRNA candidate severely impaired enucleation, as evidenced by a mean 20-85% reduction in enucleation efficiency relative to scrambled shRNA, reproducible across separate shRNAs (Figure 8B, all $p<10^{-16}$, Student's t-Test). Cells inhibited for these lncRNAs did induce TER119, an early event during terminal differentiation (Figure 8A), but exhibited greater cell size relative to control (11-26% greater average cell size, all $p<0.05$, Student's t-Test; Figure 8C), consistent with retention of immature but terminally differentiated erythroblasts. These results indicate that our lncRNAs exert their functions after TER119 activation but before terminal red cell maturation. The precise events regulated by the lncRNAs are the subject of Chapter 3 of this thesis.

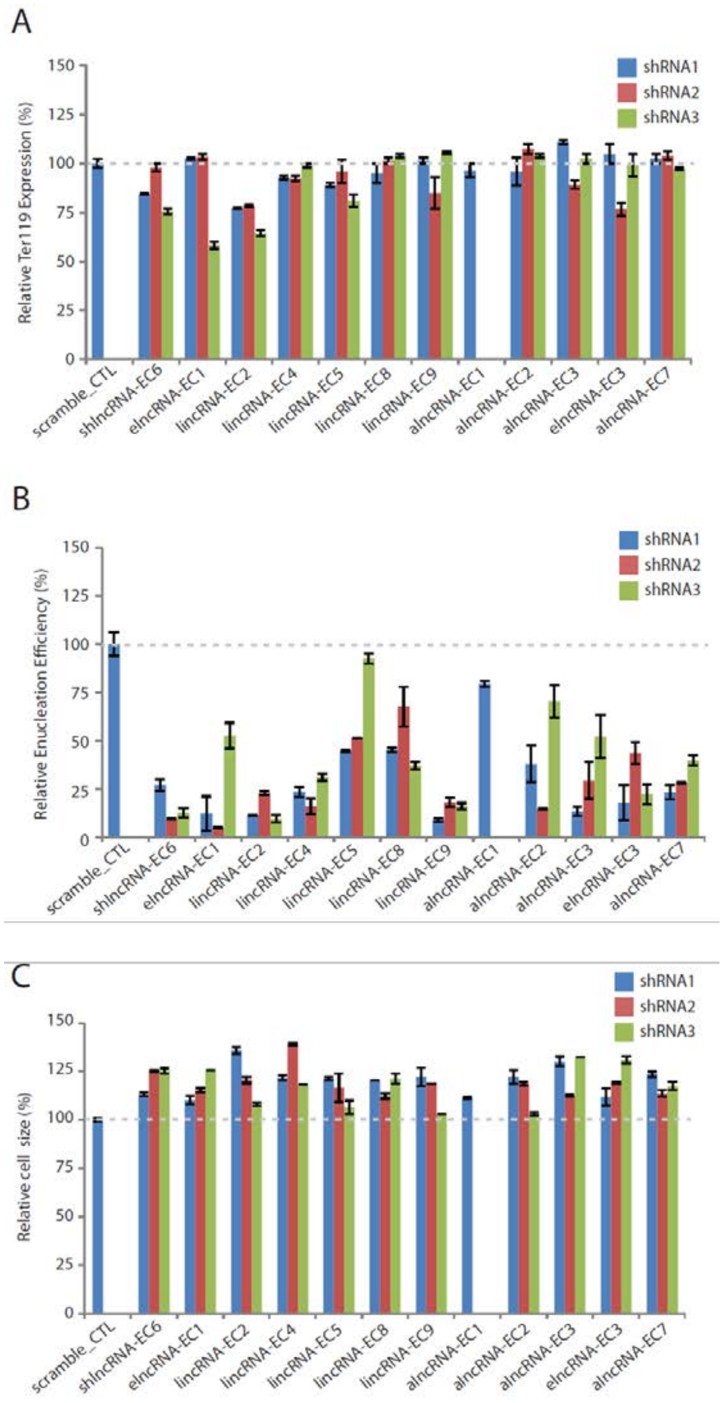


Figure 8. Modulation of red cell maturation by multiple types of lncRNAs.

(**A**) Relative expression of the early erythroid differentiation marker TER119 in erythroid progenitor-enriched fetal liver cells transduced with retroviral vectors encoding control or lncRNA-targeting shRNAs and induced to differentiate in culture. Expression levels were determined by qPCR, normalized to those of 18S rRNA, and are shown as percentage of the levels in the control shRNA experiment (dotted gray line). Data are mean ± s.e.m (n = 2).

(**B**) Relative cell size of cells treated as in (A). Average cell sizes were determined by flow cytometry (see Methods) and are shown as percentage of the values for the control shRNA experiment (dotted gray line). Data are mean ± s.e.m (n = 2).

(**C**) Relative enucleation efficiency of cells treated as in (A). Enucleation efficiency was determined by flow cytometry (see Methods) and is shown as percentage of the values for the control shRNA experiment (dotted gray line). Data are shown as mean ± s.e.m (n = 2).

Integrating lncRNAs into erythroid differentiation gene networks

Our functional studies indicate that lncRNAs are important components of the circuitry controlling red blood cell development, together with transcription factors and chromatin modifiers. We therefore sought to globally integrate erythroid differentiation lncRNAs and protein-coding factors into a functionally coherent gene network, which we inferred from weighted tissue co-expression analysis (Figure 9A; see supplemental Methods). To focus on genes important for erythropoiesis, we considered only lncRNAs and mRNAs that are bound at their promoters by both GATA1 and TAL1 and whose expression is upregulated >1.5-fold between progenitors and TER119+ erythroblasts. This resulted in a network of 200 protein-coding and 38 lncRNA genes representing the core program activated by GATA1 and TAL1 during terminal erythroid differentiation.

We find that lncRNAs are highly integrated components of the GATA1/TAL1 transcriptional network, which includes several members of the intergenic, antisense and enhancer subclasses. These lncRNAs cluster with cohorts of co-expressed coding genes in discrete modules that contain key transcription factors. For example, the core erythroid module comprises TAL1 and KLF1 as well as erythroid-specific proteins EPOR, FECH, ART4 and ANK1 (Figure 9B), recapitulating a known erythroid transcriptional circuit (Hattangadi et al. 2011). In addition to proteins, several intergenic lncRNAs are densely interconnected within this module, including lincRNA-EC9, which forms a tight module with KLF1, TAL1 and FECH. Other modules contain lncRNAs, TFs and additional proteins that are upregulated during erythropoiesis but that are also prominent in other myeloid lineages, such as that organized around ZNFX1 (Figure 9B), suggesting broad-acting myeloid networks. Thus, our work indicates that lncRNAs of diverse genomic origins are densely connected components of the transcriptional networks specified by key TFs that are needed for lineage-specific development.

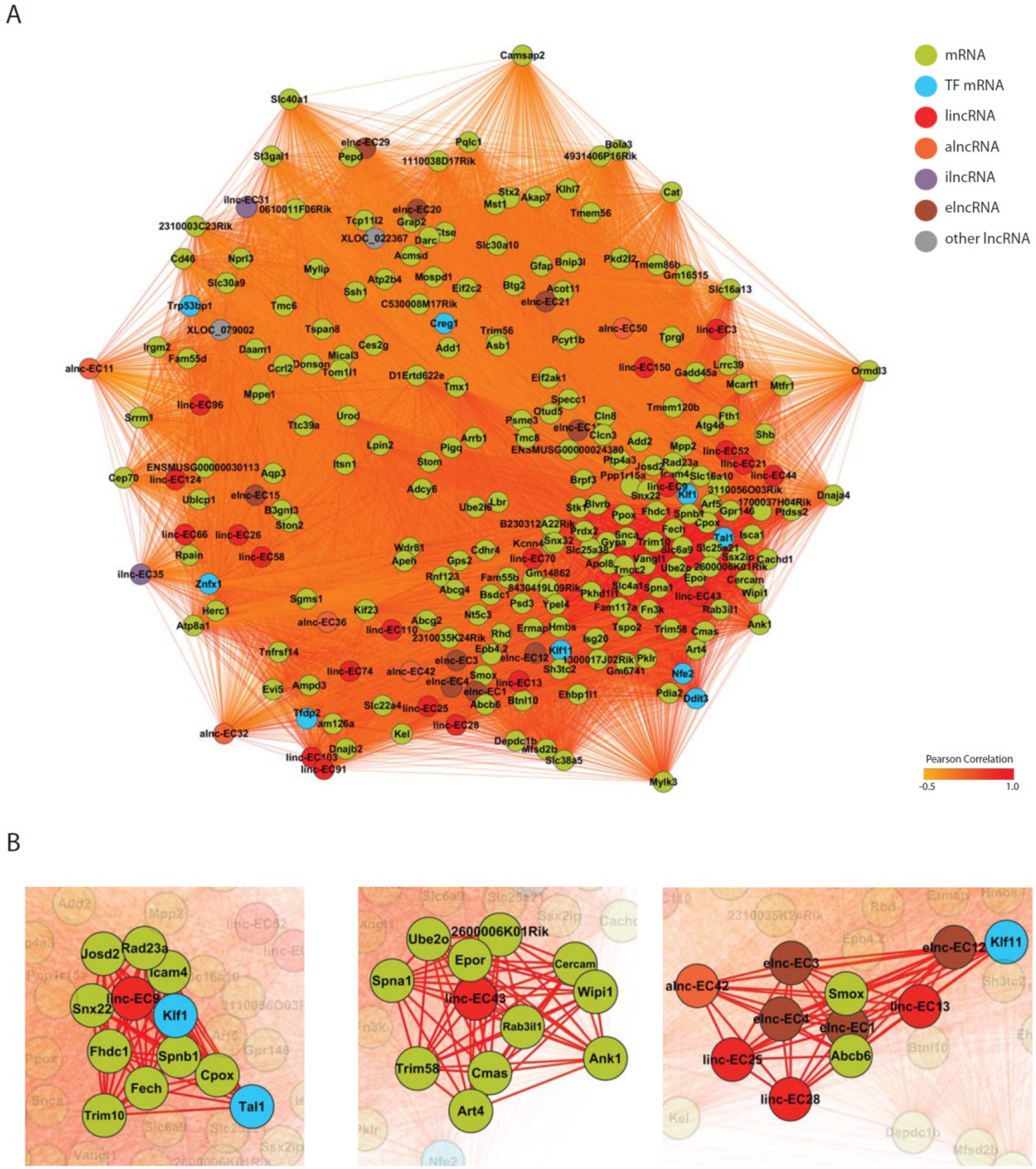


Figure 9. Network of GATA1 and TAL1-activated erythroid differentiation genes.

(**A**) Network of mRNAs and lncRNAs bound by GATA and TAL1 within 1 kb of the TSS and induced >2-fold during erythroid differentiation (see supplemental Methods). Nodes represent genes and edges represent their pairwise correlations in expression across 30 mouse tissues.

(**B**) Examples of network modules containing erythroid-specific lncRNAs, transcription factors and other proteins.

lncRNAs in hematological disease-associated regions

Because the majority of erythroid differentiation lncRNAs were not tested for function, we sought evidence of phenotypic relevance in their human orthologs. We identified 35 lncRNAs whose orthologs map within regions linked to specific traits or diseases by published GWAS studies (Hindorff et al. 2009). These included 10 lncRNAs overlapping 16 trait- or disease-related SNPs within their introns or exons. Importantly, in a handful of cases these SNPs were associated with hematological phenotypes. For example, elncRNA-EC4 is transcribed from an enhancer region upstream of RCOR1 in both mice and humans (Figure 10). RCOR1 encodes a transcriptional co-repressor critical for erythroid and megakaryocytic differentiation (Saleque et al. 2007; Laurent et al. 2009; Yao et al. 2014). In mouse red blood cells, elncRNA-EC4 is bound at its promoter by GATA1 and TAL1 and its expression is coordinated with that of RCOR1 (Figure 10A). In human, the orthologous lncRNA is located within a ~35kb linkage disequilibrium region bound at multiple sites by GATA1, TAL1, and PU.1, and overlaps a SNP associated with platelet count variation (rs11628318, odds ratio = 2.57, $p<10^{-9}$) (Gieger et al. 2011) (Figures 10B and C). rs11628318[A] is part of a binding motif recognized by TBP in the K562 erythroleukemia cell line, and may affect activation of the enhancer-derived lncRNA and hence that of its neighbor RCOR1. Thus, elncRNA-EC4 may be an additional modulator of erythropoiesis and megakaryopoiesis with phenotypic consequence.

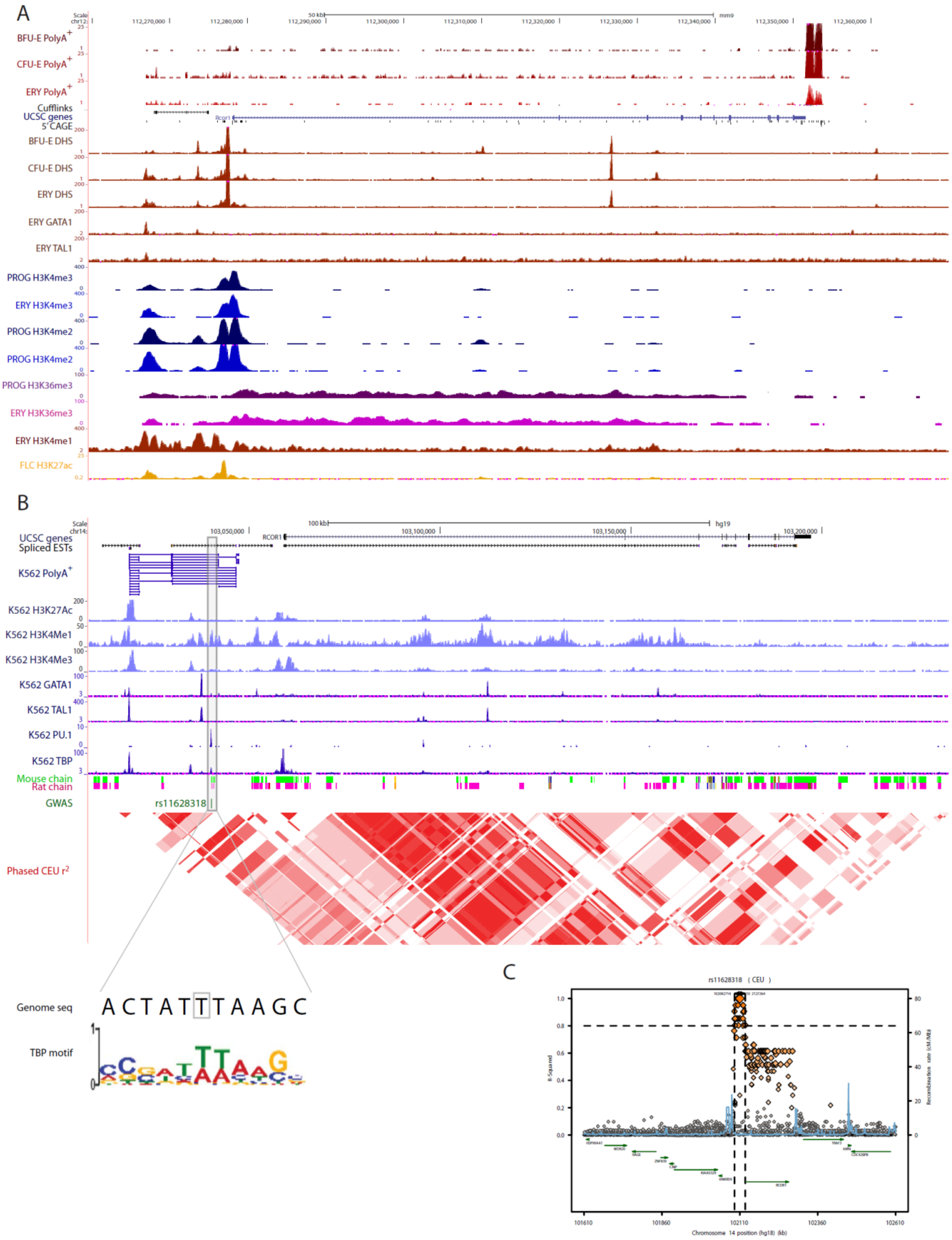

A
Scale
chr12:
50 kb
mm9
112,270,000
112,280,000
112,290,000
112,300,000
112,310,000
112,320,000
112,330,000
112,340,000
112,350,000
112,360,000
BFU-E PolyA+
CFU-E PolyA+
ERY PolyA+
Cufflinks
UCSC genes
5'CAGE
Rcor1
BFU-E DHS
CFU-E DHS
ERY DHS
ERY GATA1
ERY TAL1
PROG H3K4me3
ERY H3K4me3
PROG H3K4me2
PROG H3K4me2
PROG H3K36me3
ERY H3K36me3
ERY H3K4me1
FLC H3K27ac
B
Scale
chr14:
100 kb
hg19
103,050,000
103,100,000
103,150,000
103,200,000
RCOR1
UCSC genes
Spliced ESTs
K562 PolyA+
K562 H3K27Ac
K562 H3K4Me1
K562 H3K4Me3
K562 GATA1
K562 TAL1
K562 PU.1
K562 TBP
Mouse chain
Rat chain
GWAS
rs11628318
Phased CEU r2
Genome seq
A C T A T T T A A G C
TBP motif
1
0
C
rs11628318 (CEU)
102092714
2127264
1.0
0.8
0.6
0.4
0.2
0.0
R-Squared
80
60
40
20
0
Recombination rate (cM/Mb)
HSP90AA1
WDR20
RAGE
ZNF839
CINP
KIAA0329
ANKRD9
RCOR1
TRAF3
AMN
CDC42BPB
101610
101860
102110
102360
102610
Chromosome 14 position (hg18) (kb)

Figure 10. elncRNA-EC4 is located within an intergenic region associated with platelet count variation.

(**A**) UCSC Genome Browser track displays the landscape of transcription, chromatin accessibility, TF occupancy and histone modifications at the region encoding elncRNA-EC4 and RCOR1 in mouse as in Figures 5B and 6C. The last track displays the ChIP-Seq signal for H3K27Ac, associated with active promoters and enhancers, in fetal liver cells (yellow).

(**B**) The elncRNA-EC4 - RCOR1 region is conserved in human. UCSC Genome Browser track displays RNA-Seq and ChIP-Seq signal in K562 cells as in (A). Shown at the top are UCSC gene models (light blue), spliced ESTs from Genbank (blank), and transcript models based on our detection of orthologous genomic regions from local alignment and synteny to the mouse genome (dark blue; see supplemental Methods). Tracks 4-6 display in light blue the ChIP-Seq signal for H3K27Ac, H3K4me1 and H3K4me3. Tracks 7-10 show in dark blue the ChIP-Seq signal for GATA1, TAL1, PU.1 and TBP (enriched at promoters and enhancers). Tracks 11-12 show UCSC chained sequence alignments between mouse and human and between rat and human. Track 13 displays correlation scores between SNPs from the CEU HapMap population in a region spanning a SNP associated with platelet count variation by a genome-wide association study (GWAS; rs11628318, green). Bottom: magnified view of the 11 nucleotides surrounding the SNP ([T/A] marked by gray box) and the webLogo view of the TBP binding motif that overlaps it. (**C**) Regional association plot of SNPs from the CEU HapMap population within a broad chromosomal domain centered at rs11628318. A narrow peak (boxed) identifies SNPs significantly associated with rs11628318 specifically within the elncRNA-EC4 locus.

Discussion

Red blood cell development involves a hierarchy of well-defined cell differentiation states, making it an ideal system to identify lineage- and stage-specific regulators. Ours is the first study to catalog the repertoire of lncRNAs active during erythropoiesis, including over 100 previously unannotated lncRNA genes that are often erythroid-restricted. We comprehensively characterized these RNAs by their tissue specificity, expression patterns, chromatin state and TF binding *in vivo*, and integrated these features to select candidates for functional studies. Remarkably, most lncRNAs selected this way proved critical for the proper maturation of erythroblasts into specialized erythrocytes. Thus, our study provides a roadmap for the efficient identification of lncRNAs with roles in erythropoiesis, which can be implemented through a useful online resource (http://lodishlab.wi.mit.edu/data/lncRNAs/) where users can select lncRNAs based on their expression and regulation features to discover functional ones.

Our comprehensive characterization of erythroid lncRNAs revealed that diverse patterns in structural, conservation, regulation, tissue and developmental expression traits delineate different lncRNA families. sRNA hosting lncRNAs are widely conserved and expressed broadly across mouse tissues at similar levels than mRNAs, whereas intergenic and enhancer lncRNAs are more rapidly evolving and are highly cell type-specific. In addition, only enhancer lncRNAs show consistent expression between fetal and adult erythropoiesis, and only antisense and intronic lncRNAs appear significantly coordinated in expression with protein-coding genes near their site of transcription. Despite these differences, the majority of lncRNAs share common biogenesis features, namely Pol II transcription, 5’ capping, 3’ polyadenylation and alternative splicing, and lncRNAs of all kinds show dynamic and stage-specific patterns of expression during erythropoiesis (except pseudogene lncRNAs).

Recent work has recognized greater cell type specificity in lncRNAs vs. mRNAs (Cabili et al. 2011; Derrien et al. 2012). Indeed, many lncRNAs expressed in the mouse fetal liver exhibit strong specificity for one or two of the 30 cell types examined. Surprisingly, we also find high developmental stage-specificity of lncRNAs. We highlight lincRNAs EC-2, 4 and 9, which we find are essentially required for erythrocyte maturation in the fetal liver but are absent in the adult bone marrow. Strikingly, the tissue specificity of certain lncRNAs appears to be lost between the fetal and adult stages of development. Thus, we find crucial differences in the lncRNA programs deployed for the same cell lineage at different stages of development. One explanation for the developmental stage-specific deployment of lncRNAs may be their capacity to mount robust yet short-lived responses to dynamic developmental and environmental cues.

Analysis of how lncRNAs are regulated during erythropoiesis revealed that in fact expression and chromatin dynamics are similarly coordinated for the various types of lncRNAs during terminal differentiation, and that this is in part achieved through common targeting by erythroid TFs GATA1, TAL1 and KLF1. This may exist to synchronize expression of functionally coherent lncRNA/mRNA modules needed at different stages of differentiation. Our construction of a core network of genes activated by GATA1 and TAL1 during differentiation revealed that indeed mRNAs and lncRNAs of various types are highly interconnected in discrete co-expression modules. These modules provide a starting point for associating *trans*-acting lncRNAs with potential mRNA functional partners. Of note, because we only considered genes bound at promoter-proximal regions by GATA1 and TAL1, the complexity and interconnectedness of the erythropoiesis GATA1/TAL1 transcriptional circuitry are likely much greater than inferred.

As predicted by their regulated expression patterns and tissue specificity, diverse types of lncRNAs play critical roles during erythropoiesis. Depletion of 5 lincRNAs, 4 alncRNAs, 2 elncRNAs and 1 shlncRNA severely impaired erythrocyte maturation. Surprisingly, none abolished expression of TER119, an early differentiation marker, indicating that these RNAs are needed during late stages of maturation, when highly specialized processes such as cell size reduction, chromatin condensation and enucleation take place. Thus, our work demonstrates that lncRNAs of various genomic origins can regulate erythrocyte output, contributing to a deeper understanding of the molecular networks driving erythropoiesis that become mutated in disease. Collectively, these insights highlight lncRNAs as potential therapeutic targets that may be potentially exploited for efficient *in vitro* production of mature red blood cells.

References

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11**: R106.

Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S et al. 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**: 2242-2246.

Birney E Stamatoyannopoulos JA Dutta A Guigo R Gingeras TR Margulies EH Weng Z Snyder M Dermitzakis ET Thurman RE et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**: 799-816.

Borodina T, Adjaye J, Sultan M. 2011. A strand-specific library preparation protocol for RNA sequencing. *Methods Enzymol* **500**: 79-98.

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**: 1915-1927.

Cantor AB, Orkin SH. 2002. Transcriptional regulation of erythropoiesis: an affair involving multiple partners. *Oncogene* **21**: 3368-3376.

Carninci P Kasukawa T Katayama S Gough J Frith MC Maeda N Oyama R Ravasi T Lenhard B Wells C et al. 2005. The transcriptional landscape of the mammalian genome. *Science* **309**: 1559-1563.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**: 626-635.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**: 21931-21936.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res* **22**: 1775-1789.

Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, Babak T. 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Res* **22**: 1173-1183.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F et al. 2012. Landscape of transcription in human cells. *Nature* **489**: 101-108.

Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, Kuhn RM, Meyer LR, Wong M, Sloan CA, Rosenbloom KR et al. 2012. The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res* **40**: D918-923.

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T et al. 2009. The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* **41**: 563-571.

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K et al. 2010. The Pfam protein families database. *Nucleic Acids Res* **38**: D211-222.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S et al. 2012. Ensembl 2012. *Nucleic Acids Res* **40**: D84-90.

Flygare J, Rayon Estrada V, Shin C, Gupta S, Lodish HF. 2011. HIF1alpha synergizes with glucocorticoids to promote BFU-E progenitor self-renewal. *Blood* **117**: 3435-3444.

Gieger C Radhakrishnan A Cvejic A Tang W Porcu E Pistis G Serbanovic-Canic J Elling U Goodall AH Labrune Y et al. 2011. New gene functions in megakaryopoiesis and platelet formation. *Nature* **480**: 201-208.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**: 223-227.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C et al. 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**: 503-510.

Hattangadi SM, Burke KA, Lodish HF. 2010. Homeodomain-interacting protein kinase 2 plays an important role in normal terminal erythroid differentiation. *Blood* **115**: 4853-4861.

Hattangadi SM, Wong P, Zhang L, Flygare J, Lodish HF. 2011. From stem cell to red cell: regulation of erythropoiesis at multiple levels by multiple proteins, RNAs, and chromatin modifications. *Blood* **118**: 6258-6268.

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW et al. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**: 108-112.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**: 9362-9367.

Hu W, Alvarez-Dominguez JR, Lodish HF. 2012. Regulation of mammalian cell differentiation by long non-coding RNAs. *EMBO Rep* **13**: 971-983.

Hu W, Yuan B, Flygare J, Lodish HF. 2011. Long noncoding RNA-mediated anti-apoptotic activity in murine erythroid terminal differentiation. *Genes Dev* **25**: 2573-2578.

Ji P, Jayapal SR, Lodish HF. 2008. Enucleation of cultured mouse fetal erythroblasts requires Rac GTPases and mDia2. *Nat Cell Biol* **10**: 314-321.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656-664.

Kerenyi MA, Orkin SH. 2010. Networking erythropoiesis. *J Exp Med* **207**: 2537-2541.

Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A et al. 2009. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **106**: 11667-11672.

Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G. 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* **35**: W345-349.

Laurent B, Randrianarison-Huetz V, Kadri Z, Romeo PH, Porteu F, Dumenil D. 2009. Gfi-1B promoter remains associated with active chromatin marks throughout erythroid differentiation of human primary progenitor cells. *Stem Cells* **27**: 2153-2162.

Lin MF, Jungreis I, Kellis M. 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**: i275-282.

Marques AC, Ponting CP. 2009. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol* **10**: R124.

Neuert G, Munsky B, Tan RZ, Teytelman L, Khammash M, van Oudenaarden A. 2013. Systematic identification of signal-activated stochastic gene regulation. *Science* **339**: 584-587.

Pilon AM, Ajay SS, Kumar SA, Steiner LA, Cherukuri PF, Wincovitch S, Anderson SM, Mullikin JC, Gallagher PG, Hardison RC et al. 2011. Genome-wide ChIP-Seq reveals a dramatic shift in the binding of the transcription factor erythroid Kruppel-like factor during erythrocyte differentiation. *Blood* **118**: e139-148.

Pruitt KD, Tatusova T, Brown GR, Maglott DR. 2012. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* **40**: D130-135.

Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. 2008. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods* **5**: 877-879.

Saleque S, Kim J, Rooke HM, Orkin SH. 2007. Epigenetic regulation of hematopoietic differentiation by Gfi-1 and Gfi-1b is mediated by the cofactors CoREST and LSD1. *Mol Cell* **27**: 562-572.

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV et al. 2012. A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**: 116-120.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034-1050.

Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, Groudine M, Bender M, Kaul R, Canfield T et al. 2012. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol* **13**: 418.

Tallack MR, Magor GW, Dartigues B, Sun L, Huang S, Fittock JM, Fry SV, Glazov EA, Bailey TL, Perkins AC. 2012. Novel roles for KLF1 in erythropoiesis revealed by mRNA-seq. *Genome Research* **22**: 2385-2398.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105-1111.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511-515.

Ulitsky I, Bartel DP. 2013. lincRNAs: Genomics, Evolution, and Mechanisms. *Cell* **154**: 26-46.

Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. 2011. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147**: 1537-1550.

Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, Lajoie BR, Protacio A, Flynn RA, Gupta RA et al. 2011. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**: 120-124.

Weiss MJ, Yu CN, Orkin SH. 1997. Erythroid-cell-specific properties of transcription factor GATA-1 revealed by phenotypic rescue of a gene-targeted cell line. *Molecular and Cellular Biology* **17**: 1642-1651.

Welch JJ, Watts JA, Vakoc CR, Yao Y, Wang H, Hardison RC, Blobel GA, Chodosh LA, Weiss MJ. 2004. Global regulation of erythroid gene expression by transcription factor GATA-1. *Blood* **104**: 3136-3147.

Wong P, Hattangadi SM, Cheng AW, Frampton GM, Young RA, Lodish HF. 2011. Gene induction and repression during terminal erythropoiesis are mediated by distinct epigenetic changes. *Blood* **118**: e128-138.

Wontakal SN, Guo X, Smith C, MacCarthy T, Bresnick EH, Bergman A, Snyder MP, Weissman SM, Zheng D, Skoultchi AI. 2012. A core erythroid transcriptional network is repressed by a master regulator of myelo-lymphoid differentiation. *Proc Natl Acad Sci U S A* **109**: 3832-3837.

Wu W, Cheng Y, Keller CA, Ernst J, Kumar SA, Mishra T, Morrissey C, Dorman CM, Chen KB, Drautz D et al. 2011. Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration. *Genome Res* **21**: 1659-1671.

Yao HL, Goldman DC, Nechiporuk T, Kawane S, McWeeney SK, Tyner JW, Fan G, Kerenyi MA, Orkin SH, Fleming WH et al. 2014. Corepressor Rcor1 is essential for murine erythropoiesis. *Blood* **123**: 3175-3184.

Yu M, Riva L, Xie HF, Schindler Y, Moran TB, Cheng Y, Yu DN, Hardison R, Weiss MJ, Orkin SH et al. 2009. Insights into GATA-1-Mediated Gene Activation versus Repression via Genome-wide Chromatin Occupancy Analysis. *Molecular Cell* **36**: 682-695.

Zhang J, Socolovsky M, Gross AW, Lodish HF. 2003. Role of Ras signaling in erythroid differentiation of mouse fetal liver cells: functional analysis by a flow cytometry-based novel culture system. *Blood* **102**: 3938-3946.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.

Zhu J, Sanborn JZ, Diekhans M, Lowe CB, Pringle TH, Haussler D. 2007. Comparative genomics search for losses of long-established genes on the human lineage. *PLoS Comput Biol* **3**: e247.

Chapter 2: Control of red blood cell development by the long non-coding RNA EC6

This chapter represents a manuscript in preparation by the following authors:

Juan R. Alvarez-Dominguez, Wenqian Hu, Marko Knoll, and Harvey F. Lodish

Author contributions: J.R.A.-D. designed, performed, and presented the analysis of sequencing reads from RNA-seq, ChIP-seq, DNase-seq, Hi-C, and ChIA-PET experiments (Figures 2, 4-6, and Figures S1, S3, and S4); designed, performed, analyzed, and presented data from knockdown experiments, RNA-FISH experiments, RNA-seq experiments, and from enucleation, proliferation and apoptosis assays (Figures 3-6 and Figure S2); designed, performed, and presented bioinformatics analyses (Figure 6 and Figure S5); analyzed and presented qPCR data from knockdown and tissue expression experiments (Figures 1 and 2), and wrote the manuscript. W.H. designed and performed knockdown experiments and tissue expression experiments (Figures 1 and 2). M.K. performed, analyzed, and presented data from immunoprecipitation experiments (Figure 7). J.R.A.-D., W.H., and H.F.L. designed the study and revised the manuscript.

Acknowledgments: We thank David Bartel, Laurie Boyer, John Rinn, and members of the Lodish lab, especially Jiahai Shi, for helpful discussions and critical comments on this study. We also thank the MIT and Whitehead Institute flow cytometry cores and the Whitehead Institute genome technology core for technical support. W.H. is a Merck Fellow of the Life Sciences Research Foundation and is supported by a Pathway to Independence Award (1K99HL118157-01) from the National Institutes of Health, National Heart, Lung, and Blood Institute. This research was supported by the grants from National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases DK068348 and from the National Institutes of Health, National Heart, Lung, and Blood Institute, 5P01 HL066105.

Introduction

Having demonstrated the functional importance of lncRNAs to erythropoiesis, we wanted to determine how lncRNAs achieve these effects.

We focused our work on our top candidate, shlncRNA-EC6, which comprises an erythroid-specific isoform of DLEU2 induced by Gata1 during terminal differentiation. Processing of this isoform in mouse or in human generates a poly(A)- transcript enriched in chromatin, which we termed EC6. Knockdown of EC6 in cultured erythroid progenitors leads to elevated apoptosis and severely inhibits their proliferation during terminal differentiation. EC6 is retained at its site of transcription, which is in physical proximity to three adjacent protein-coding genes via promoter-promoter chromatin interactions. These neighbors are selectively downregulated during development of the erythroid lineage and encode activators of NF-kB signaling, which antagonizes erythropoiesis. Accordingly, depletion of EC6 leads to specific de-repression of these genes and is accompanied by activation of NF-kB signaling among other pathways of immune cell development. Mechanistically, EC6 interacts with the nuclear matrix factor hnRNP U, which potentially enables co-localization with its targets to mediate their repression. Thus, EC6 promotes erythroid differentiation by suppressing alternative immune developmental programs.

This work demonstrates that lncRNAs can contribute to erythroid lineage-specific development by modulating competing gene expression programs controlling cell fate.

Results

Identification of erythroid lncRNAs affecting neighboring gene expression

To explore how lncRNAs regulate erythropoiesis, for the 10 functional lncRNAs characterized in Chapter 1 we examined the expression of their nearest or overlapping neighbor by qPCR following knockdown of the lncRNA. Seven of these lncRNAs had no effect on their closest neighbor’s expression upon knockdown (Figure 1A). In contrast, inhibiting shlncRNA-EC6/DLEU2 caused upregulation of SPRYD7/CLLD6, residing ~45 kb away (Figure 1B). No function is known for the SPRYD7 protein, although a role in NF-kB signaling has been proposed (Garding et al. 2013). Two other lncRNAs were linked to promoting expression of their gene neighbors during erythroid differentiation. Depleting the enhancer-derived elncRNA-EC3 leads to ~35-70% loss of KIF2A expression, which resides ~40 kb away (Fig 1C). KIF2A is a kinesin motor involved in microtubule dynamics that is required for normal mitotic progression (Debernardi et al. 1997; Homma et al. 2003), but no specific role during erythropoiesis has been described. Similarly, inhibiting the enhancer-derived alncRNA-EC7 causes >80% depletion of neighboring BAND3/SLC4A1 ~10 kb away (Figure 1D). BAND3 is a major anion exchanger of the erythrocyte membrane, and its mutation can lead to hemolytic anemias.(Jarolim et al. 1992; Bruce et al. 2005). These data suggest that a subset of lncRNA regulators of erythropoiesis can act to promote or suppress expression of neighboring genes during differentiation, consistent with their correlated or anticorrelated expression patterns (Figures 1B-D). In the next sections, we focus on shlncRNA-EC6/DLEU2, given its relevance to chronic lymphocytic leukemia (see Introduction), and propose a model for its molecular function based on mechanistic insights gained from detailed characterization and functional studies.

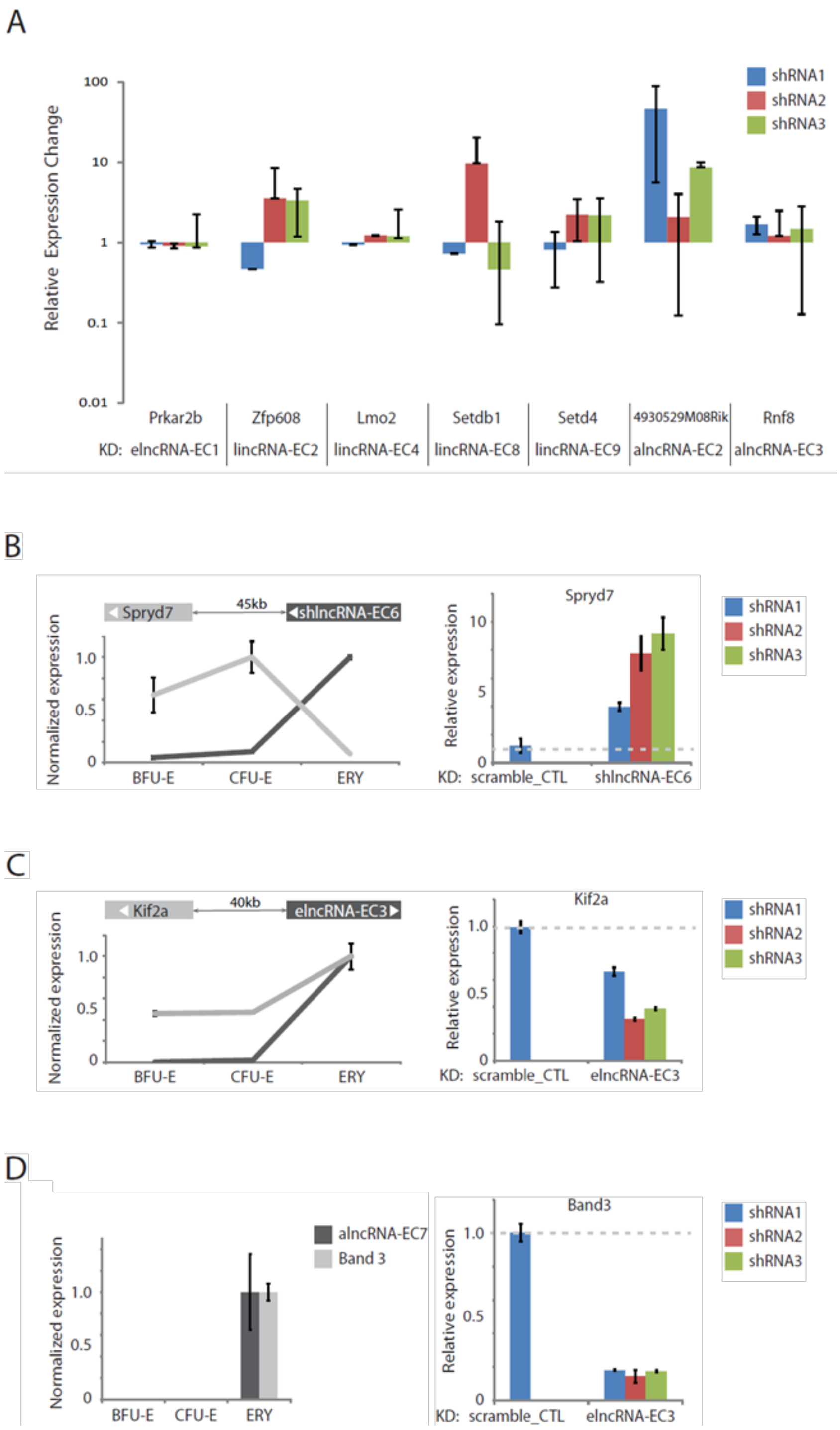


Figure 1. Impact of candidate erythroid lncRNA inhibition on neighboring gene expression.

(**A**) Relative change in the expression of the closest mRNA neighbors of elncRNA-EC1, lincRNAs EC2, EC4, EC8 and EC9, and alncRNAs EC2 and EC3 upon shRNA-mediated depletion of the lncRNA, as determined by qPCR. Values were normalized to those of 18S rRNA, and fold-changes were computed relative to the scramble shRNA control. Data are mean ± s.e.m (n = 3).

(**B**) SPRYD7 (light gray) is anticorrelated in expression with its neighbor shlncRNA-EC6 (dark gray) during erythropoiesis. Depletion of shlncRNA-EC6 with separate shRNAs in *ex vivo*-differentiated TER119+ erythroblasts results in reproducible upregulation of SPRYD7 relative to scramble shRNA control (data are mean ± s.e.m, n = 3).

(**C**) KIF2A (light gray) is coordinated in expression with neighboring elncRNA-EC3 (dark gray) during erythroid differentiation. Inhibiting elncRNA-EC3 with separate shRNAs leads to reduced expression of KIF2A relative to scramble shRNA control (data are mean ± s.e.m, n = 3).

(**D**) Band3 expression is coordinated with that of neighboring alncRNA-EC7 during differentiation. Inhibiting alncRNA-EC7 with shRNAs abolishes expression of Band3 relative to scramble shRNA control (data are mean ± s.e.m, n = 3).

EC6 is an erythroid-selective DLEU2 isoform activated by GATA1

shlncRNA-EC6 was identified as the top candidate modulator of erythropoiesis based on its abundance, tissue specificity, regulation, and induction during differentiation (Chapter 1). Indeed, shlncRNA-EC6 is upregulated ~20-fold during differentiation (to FPKM >45) and is highly specific to differentiated red blood cells (Figures 2A and B). The shlncRNA-EC6 locus encodes the known lncRNA DLEU2, which maps to a critical region at chromosomal band 13q14.3 whose deletion is causally linked to the pathogenesis of B cell chronic lymphocytic leukemia (CLL) and of other immune cell malignancies (Introduction). DLEU2 hosts microRNAs 15a and 16-1, but a function independent of microRNA generation is suggested by the fact that its knockout or ectopic expression shows a stronger cancer phenotype compared to miR-15a/16-1 knockout or misexpression (Lerner et al. 2009; Klein et al. 2010; Lia et al. 2012), and by the fact that rare cases of CLL exist where the 13q14 deletion does not encompass the microRNAs (Ouillette et al. 2008; Ouillette et al. 2011).

The DLEU2 locus is broadly transcribed across tissues into multiple isoforms that differ in splicing patterns and can be generated from two alternative promoters, here referred to as Dleu2 long and Dleu2 short isoforms (Figure 2D). In erythroblasts, only the proximal promoter is active, as the distal one is marked by H3K27me3 from the early committed-progenitor stage (Figure 2F). RNA-seq-based *de novo* transcript reconstruction indicated that the erythroid-specific DLEU2 isoform (shlncRNA-EC6) predominantly corresponds to a Dleu2 short variant with a retained intron (Figures 2D and S1), here referred to solely as EC6. Examination of chromatin accessibility around EC6 revealed two sites at the proximal promoter region that become progressively accessible during differentiation and are bound by GATA1 in differentiated erythroblasts (Figure 2E), correlating with potent upregulation of EC6 selectively at the terminal differentiation stage. In contrast, no other regulatory sites are formed *de novo* during differentiation. We directly assessed GATA1's role in EC6 activation by examining EC6 expression across a timecourse of restored, estradiol-inducible GATA1 expression in the mouse G1E-ER4 cell line, and verified dose-dependent activation (Figure 2C). Thus, EC6 consists of a novel DLEU2 isoform activated by GATA1 selectively in red blood cells.

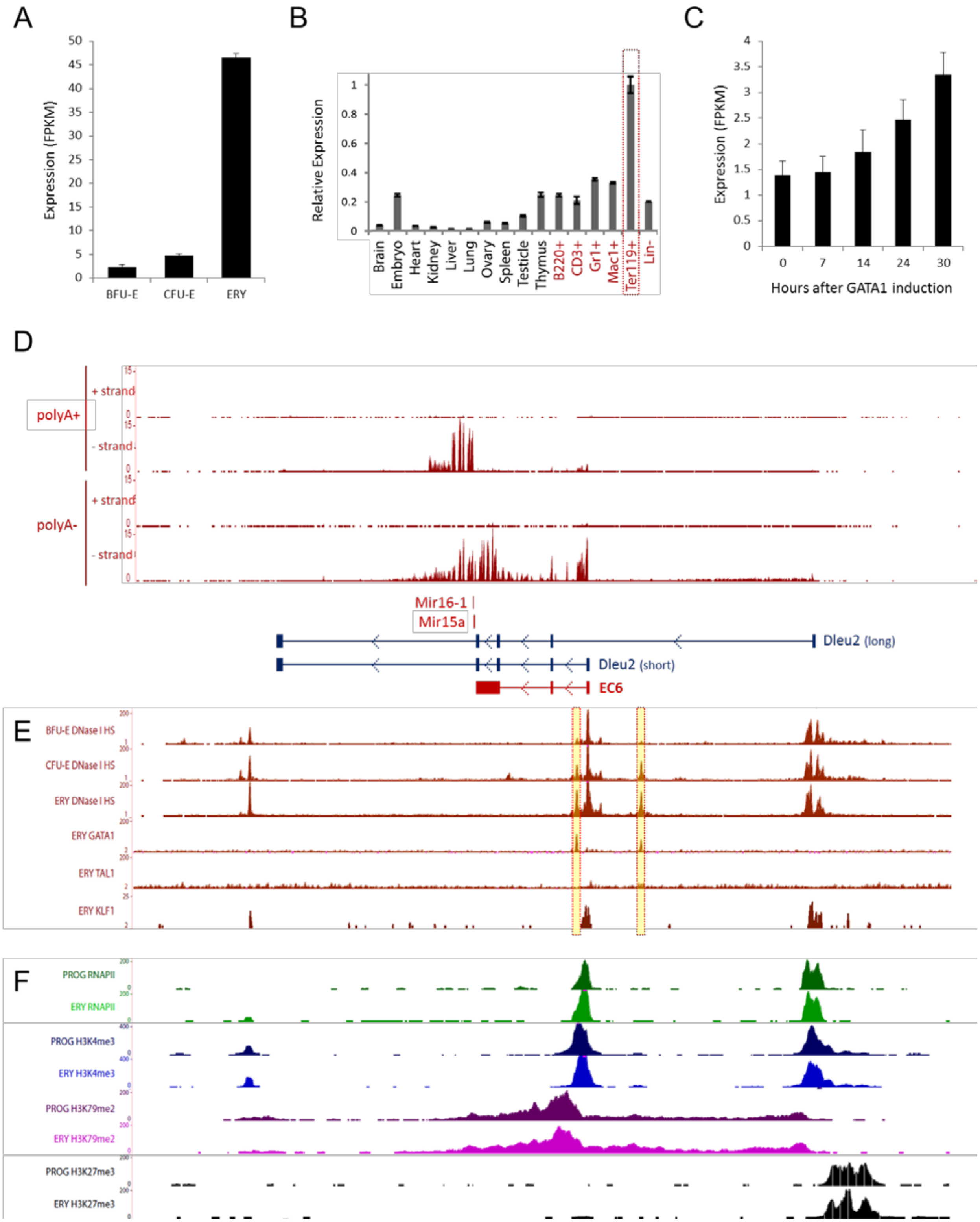

A
Expression (FPKM)
0
5
10
15
20
25
30
35
40
45
50
BFU-E
CFU-E
ERY
B
Relative Expression
0
0.2
0.4
0.6
0.8
1
Brain
Embryo
Heart
Kidney
Liver
Lung
Ovary
Spleen
Testicle
Thymus
B220+
CD3+
Gr1+
Mac1+
Ter119+
Lin-
C
Expression (FPKM)
0
0.5
1
1.5
2
2.5
3
3.5
4
0
7
14
24
30
Hours after GATA1 induction
D
polyA+
+ strand
- strand
polyA-
+ strand
- strand
15
0
Mir16-1
Mir15a
Dleu2 (long)
Dleu2 (short)
EC6
E
BFU-E DNase I HS
CFU-E DNase I HS
ERY DNase I HS
ERY GATA1
ERY TAL1
ERY KLF1
200
25
2
1
F
PROG RNAPII
ERY RNAPII
PROG H3K4me3
ERY H3K4me3
PROG H3K79me2
ERY H3K79me2
PROG H3K27me3
ERY H3K27me3
200
400
0

Figure 2. EC6 is and erythroid-selective DLEU2 isoform activated by GATA1.

(**A**) Gene-level quantification of the shlncRNA-EC6 locus by RNA-seq in BFU-E, CFU-E and TER119+ erythroblasts (ERY) (n = 2 replicates).

(**B**) Relative expression of shlncRNA-EC6 across mouse organs and cells of different tissues and developmental stages, as determined by qPCR. Expression levels were normalized to those of 18S rRNA, and fold-changes were calculated relative to fetal TER119+ erythroblast levels. Data are shown as mean ± s.e.m (n = 3 replicates).

(**C**) Gene-level quantification of shlncRNA-EC6 by RNA-seq in G1ER cells induced to differentiate at various time points during induction.

(**D-F**) DLEU2 locus map. Shown are RNA-Seq signal as the density of mapped RNA-Seq reads, DNAse I hypersensitivity (HS) signal as the density of mapped sequencing tags, and ChIP-Seq signal as the density of processed signal enrichment. (D) Tracks show in red the strand-specific RNA-Seq signal in the plus or minus strands (denoted to the left of the tracks) of poly(A)- or poly(A)+ RNA from fetal liver TER119+ erythroblasts (ERY). Depicted at the bottom are relevant lncRNA and miRNA transcripts at the locus. (E) Tracks show in red the signal for DNAse I HS, associated with open chromatin, in BFU-E and CFU-E progenitors or ERY, and the ChIP-Seq signal for GATA1, TAL1 and KLF1, respectively, in ERY. (F) Tracks depict the ChIP-Seq signal for: serine 5 phosphorylated RNA Pol II, enriched at the TSS of active genes, in erythroid progenitor-enriched fetal liver cells (PROG) and ERY (dark and light green); H3K4me3, associated with transcription activation, in PROG and ERY (dark and light blue); H3K79me2, associated with transcriptional elongation, in PROG and ERY (dark and light purple); and H3K27me3, associated with transcriptional repression, in PROG and ERY (black).

Inhibition of EC6 blocks proliferation and leads to apoptosis during terminal erythroid differentiation

To conduct loss-of-function studies, we designed, cloned, and validated triplicate shRNAs targeting separate regions of EC6. shRNA-expressing vectors were introduced into lineage-negative fetal liver cells, enriched for erythroid progenitors (Flygare et al. 2011), via retroviral transduction, followed by culture in maintenance medium to allow shRNA expression and subsequent induction of differentiation in erythropoietin-containing media (Hattangadi et al. 2010). Between 30-50% knockdown of EC6 was achieved in differentiation day 2 TER119+ erythroblasts (Figure 3A), which had little effect on miR-15a/16-1 levels (Figure 3B), as expected from their co-transcriptional processing. EC6 knockdown strongly inhibited proliferation during terminal differentiation, resulting in a ~3-fold reduction in cell number relative to control shRNA (Figure 3C).

We reasoned that impaired proliferation upon EC6 depletion could be likely due to limited cell survival, as erythropoiesis is highly sensitive to competing pro- and anti-apoptotic signals and erythroid precursors normally undergo apoptosis in the absence of adequate differentiation signals (Hattangadi et al. 2011; Hu et al. 2011). To test this hypothesis, we measured the apoptotic state of day 1 EC6 KD cells via Annexin-V staining, and confirmed a ~2- to 3-fold increase in the apoptotic cell number relative to controls (Figures 3D and S2A). Consistent with this phenotype, EC6 depletion severely inhibited red cell maturation, as evidenced by ~75-90% reduction in enucleation efficiency (Figure 3E and S2B), leading to the accumulation of immature but terminally differentiated erythroblasts, which are larger and nucleated (Figures 3F-G and S2C). Thus, EC6 contributes to red cell maturation by promoting cell survival.

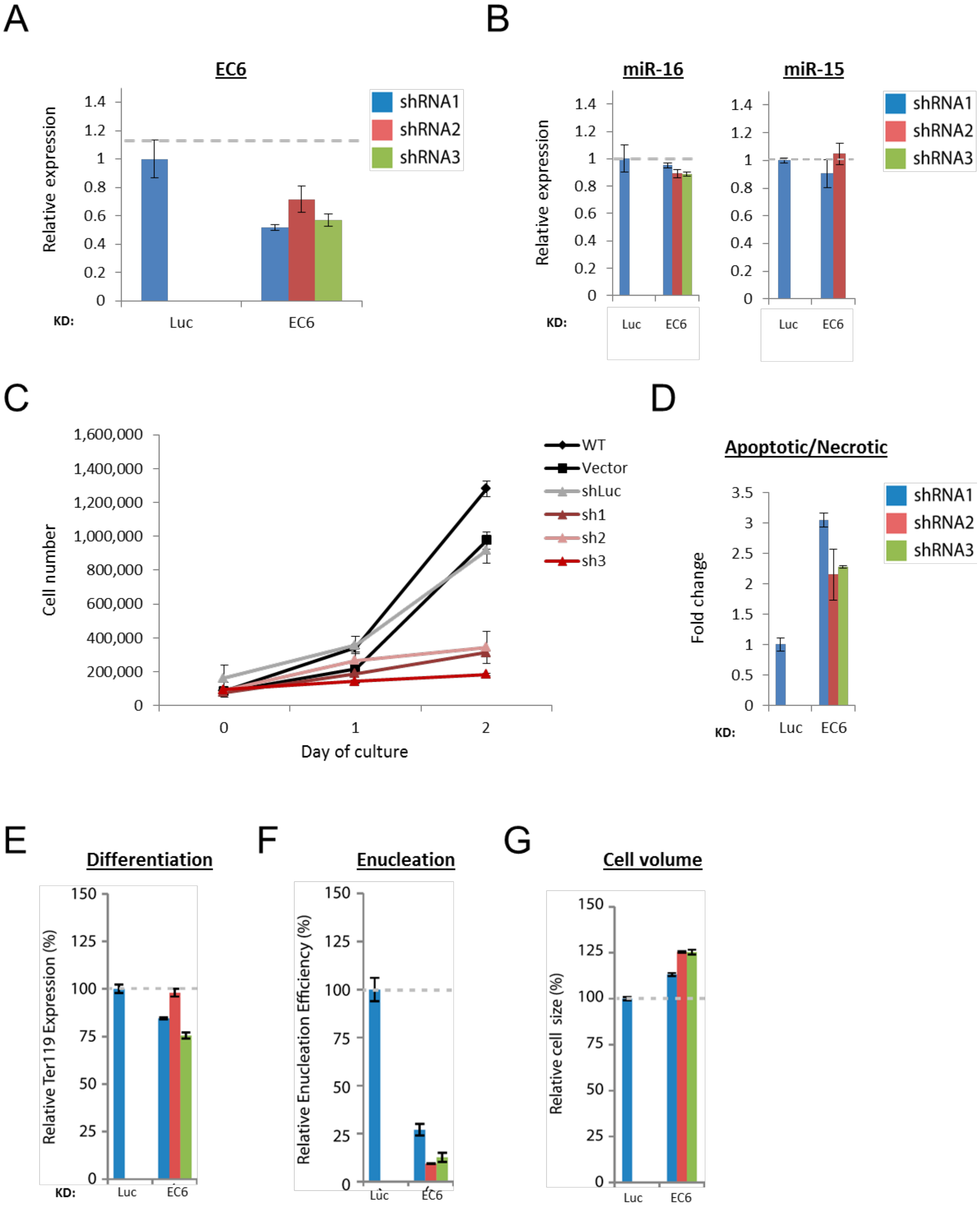


Figure 3. Inhibition of EC6 blocks proliferation and leads to apoptosis during terminal erythroid differentiation.

(**A**) Relative expression of EC6 in TER119+ fetal liver cells upon depletion by shRNAs. Erythroid progenitor-enriched fetal liver cells were transduced by retroviral vectors encoding shRNAs targeting different transcript regions or scramble shRNA control, and effectively transduced cells were induced to differentiate in culture and analyzed for lncRNA expression by qPCR. Values were normalized to those of 18S rRNA, and fold-changes were computed relative to the scramble shRNA control. Data are mean ± s.d. (n = 3 replicates).

(**B**) As in (A) but for miR-16-1 and mirR-15a.

(**C**) Cell counts measured at 24 hours and 48 hours after induction of differentiation for EC6-depleted, control-depleted, vector-only treated or wild-type (WT) cells (n ≥2 replicates).

(**D**) Fraction of apoptotic and necrotic cells for EC6-depleted or control-depleted cells assayed by Annexin V staining.

(**E-G**) Relative expression of the differentiation marker TER119 (E), cell size (F) and enucleation efficiency (G) of erythroid progenitor-enriched fetal liver cells transduced with retroviral vectors encoding control or lncRNA-targeting shRNAs and induced to differentiate in culture. Expression levels were determined by qPCR, normalized to those of 18S rRNA, and are shown as percentage of the levels in the control shRNA experiment (dotted gray line). Data are mean ± s.e.m (n = 2 replicates).

EC6 is a polyA- lncRNA retained in chromatin near its site of transcription

Our global survey of poly(A)+ and poly(A)- RNA from TER119+ erythroblasts illuminated the complex biogenesis of EC6 (Figure 4A). Cleavage of a poly(A)+ precursor RNA, presumably to release microRNAs 16-1 and 15a, leaves behind both EC6, the 5’ poly(A)- remaining transcript, and a 3’ poly(A)+ fragment that has similar stability, as evidenced by similar levels in their respective RNA fractions (Figure 4A). Because DLEU2 is conserved between human and mouse, we verified that it is also processed into 5’ poly(A)- and 3’ poly(A)+ fragments in human erythroleukemia K562 cells (Figures 4B and 4D; Supplemental Experimental Procedures). K562 cells have additionally been examined for both poly(A)- and poly(A)+ RNA content across cytosol, nuclear and subnuclear compartments as part of the human ENCODE project (Djebali et al. 2012). We therefore quantified levels of EC6 and of the 3’ fragment in their respective RNA fractions across cellular compartments, and found that while the latter is enriched in the cytosol (Figure 4C), EC6 is almost exclusively (~94%) nuclear (Figure 4E) and specifically enriched within chromatin (Figure 4F).

Chromatin-retained regulatory lncRNAs, such as Xist and Firre (Brockdorff et al. 1992; Brown et al. 1992; Hacisuleyman et al. 2014), can be retained at their site of transcription to act on physically proximal targets (*cis* regulation), in which case they present focal nuclear localization. Others, like HOTAIR and lincRNA-EPS (Rinn et al. 2007; Hu et al. 2011), can diffuse from their transcription site to act on physically distal targets (*trans* regulation), thereby presenting diffuse nuclear localization. To distinguish between these two possibilities, we used single-molecule RNA FISH to independently target the introns of EC6, which mark its site of transcription, and its exons, which map the location of mature transcripts (Figure 4G). These experiments revealed predominantly nuclear and focal localization of EC6 in physical proximity to its site of transcription (Figure 4G), suggesting a *cis* mechanism of action. Thus, EC6 is a polyA- lncRNA retained near its site of transcription.

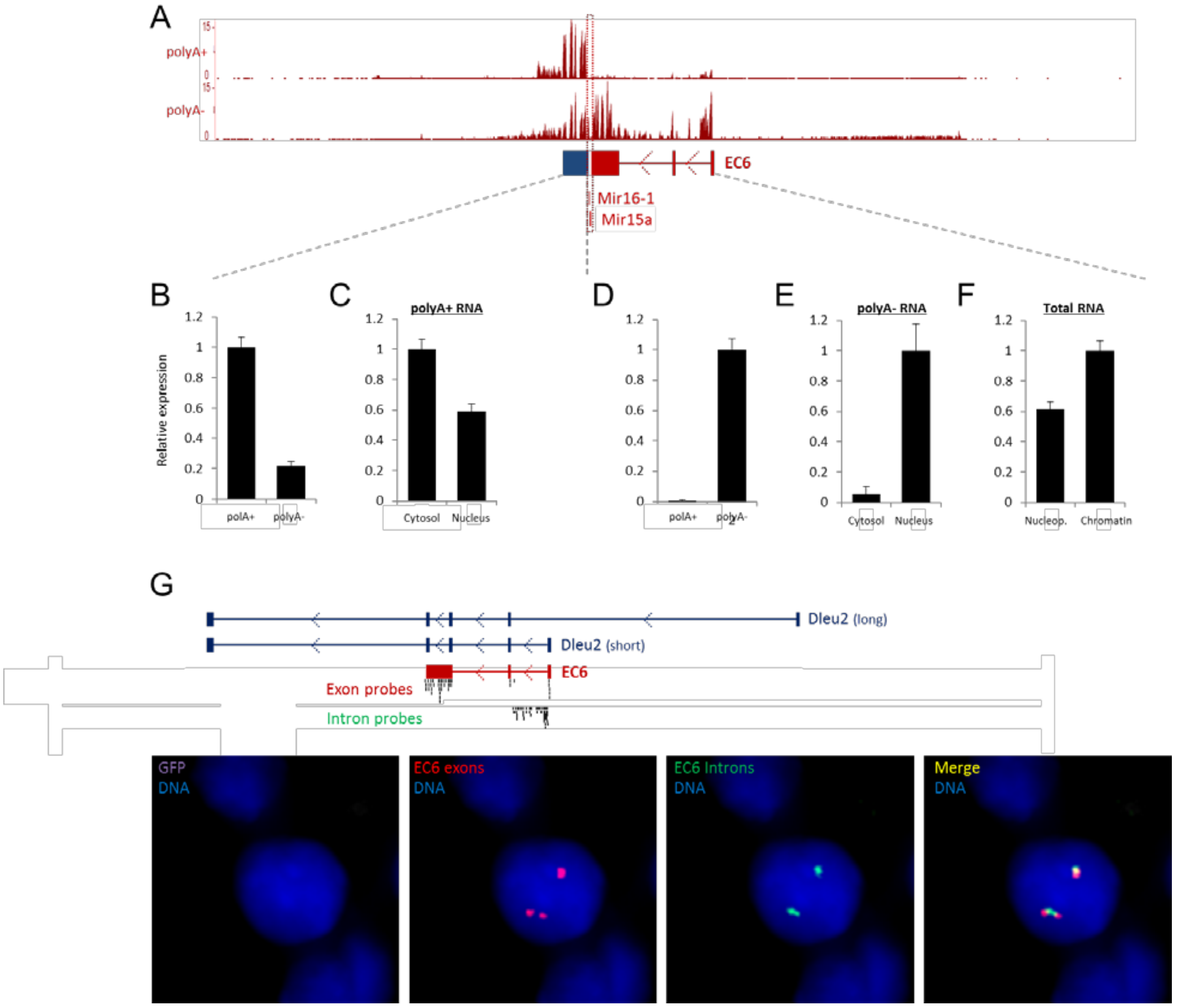


Figure 4. EC6 is a polyA- lncRNA retained in chromatin near its site of transcription.

(**A**) DLEU2 locus map as in Figure 2E.

(**B**) Relative expression of the EC6 precursor 3'fragment indicated in blue in (A) across poly(A)+ and poly(A)- RNA fractions of K562 cells.

(**C**) As in B but for poly(A)+ RNA from cytosol and nucleus cellular fractions.

(**D**) Relative expression of EC6 in poly(A)+ and poly(A)- RNA fractions of K562 cells.

(**E**) As in (D) but for poly(A)- RNA from cytosol and nucleus cellular fractions.

(**F**) As in (D) but for total RNA from nucleoplasm and chromatin fractions.

(**G**) (Top) Relevant lncRNA and miRNA transcripts at the DLEU2 locus. RNA FISH probes targeting EC6 exons and introns are indicated. (Bottom) Dual-color single-molecule RNA FISH images of EC6 exons (red) and introns (green). Shown are maximum z-stack projections of fluoresce microscopy images of fixed TER119+ erythroblasts hybridized to singly-labeled RNA FISH probes. DAPI-stained nuclei are pseudocolored blue and the GFP fluorescence channel is shown as a control for background fluorescence and signal specificity.

EC6 acts to repress physically proximal protein-coding gene targets

To determine which loci are in physical proximity to the site of EC6 transcription, we examined long-range chromatin interactions anchored at the site surveyed by paired-end tagging-based chromatin interaction analysis (ChIA-PET) in K562 cells (Li et al. 2012; Heidari et al. 2014) (Figures 5A and S4). We analyzed ChIA-PET data for interactions associated with RNA polymerase II, the looping factors CTCF and Cohesin, and the regulatory marks H3K4me1/2/3 and H3K27ac. We found only intra-chromosomal interactions contained within a 1.4 Mb domain, confined within the boundaries of a topologically associating domain that is preserved across cell types and between mouse and human (Figure S3). ChIA-PET for RNAPII, looping factors and regulatory marks consistently revealed promoter-promoter interactions between the two alternative DLEU2 promoters and five protein-coding genes, the farthest of which is located ~0.8 Mb away from EC6 (Figures 5A and S4). These genes include SPRYD7, which we previously showed to be de-repressed upon EC6 knockdown (Figure 1B), and thus represent candidate EC6 targets. Accordingly, all five genes are strongly suppressed upon terminal differentiation (Figure 5B).

Suppression of genes physically contacted by the EC6 locus could be mediated by the retained EC6 lncRNA or be an indirect effect of widespread mRNA downregulation during terminal erythropoiesis (Wong et al. 2011). We reasoned that since EC6 is restricted to erythroid cells, genes specifically suppressed by it should be explicitly silenced in erythroid vs. other cell types. We thus ranked the expression level of the five interacting loci across 30 mouse cell and tissue types surveyed using standardized methods for the mouse ENCODE project (Stamatoyannopoulos et al. 2012) (Figure 5C). This analysis revealed that only three of them - EBPL, SPRYD7, and DLEU5- are explicitly silent in red blood cells, suggesting specific EC6 targets. Accordingly, we verified that transduction of shRNAs targeting EC6 in erythroid precursors followed by induction of differentiation resulted in specific de-repression of these genes but not of KPNA3 or RNASEH2B (Figure 5D). To directly test if EC6 localizes to these loci, we conducted RNA FISH experiments independently targeting the exons of EC6, mapping its location within cells, and the introns of SPRYD7 or of RNASEH2B, which mark their sites of transcription. As expected, EC6 localized to the locus of its target SPRYD7 (Figure 5E) but not to that of non-target RNASEH2B, despite its physical proximity in the linear chromosome (Figure 5F). These results demonstrate that EC6 localizes to physically proximal gene targets and mediates their repression during terminal erythroid differentiation.

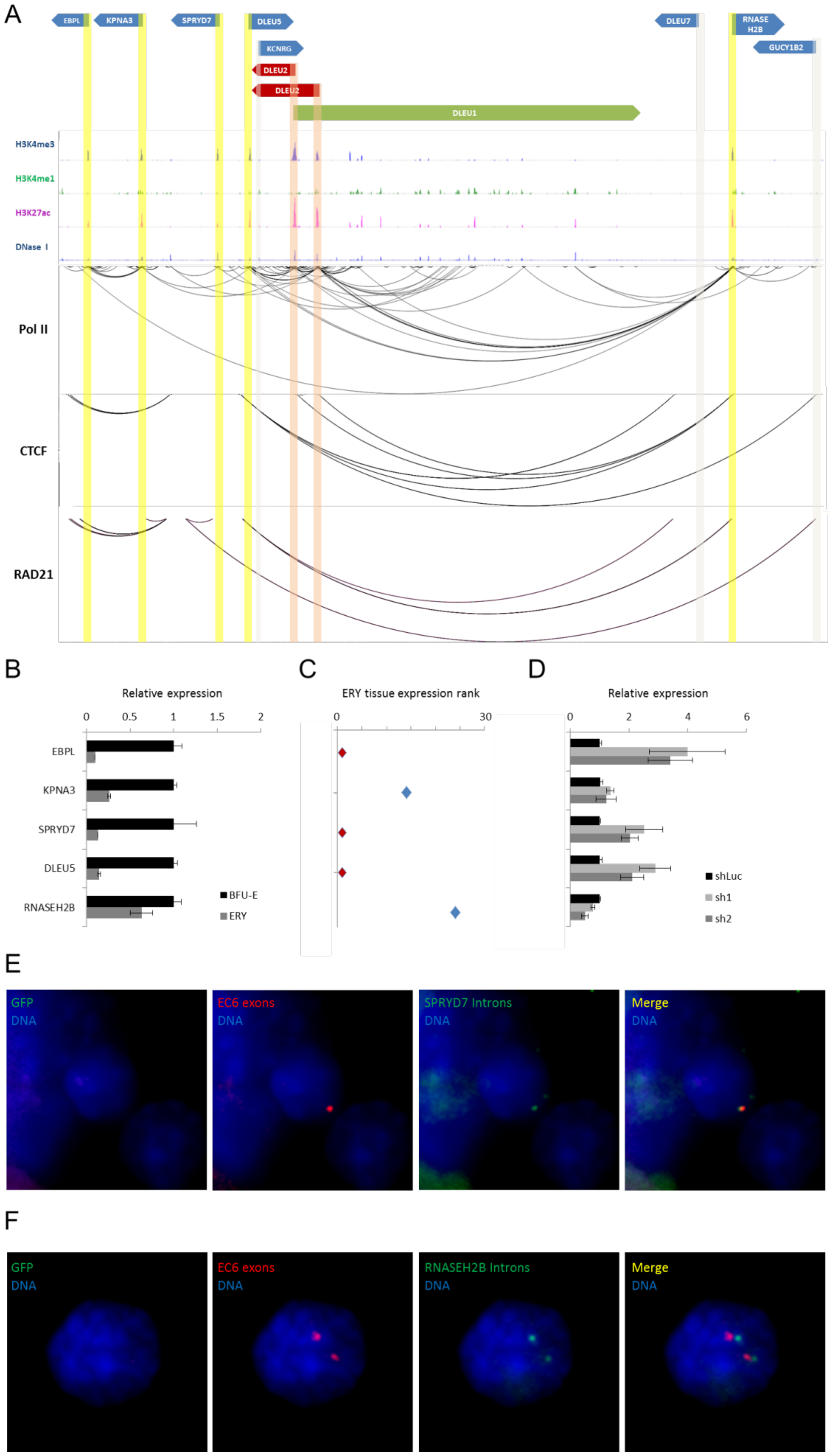

A
EBPL
KPNA3
SPRYD7
DLEU5
KCNRG
DLEU2
DLEU2
DLEU1
DLEU7
RNASE H2B
GUCY1B2
H3K4me3
H3K4me1
H3K27ac
DNase I
Pol II
CTCF
RAD21
B
Relative expression
0
0.5
1
1.5
2
EBPL
KPNA3
SPRYD7
DLEU5
RNASEH2B
BFU-E
ERY
C
ERY tissue expression rank
0
30
D
Relative expression
0
2
4
6
shLuc
sh1
sh2
E
GFP
DNA
EC6 exons
DNA
SPRYD7 Introns
DNA
Merge
DNA
F
GFP
DNA
EC6 exons
DNA
RNASEH2B Introns
DNA
Merge
DNA

Figure 5. EC6 acts to repress physically proximal protein-coding gene targets.

(**A**) Locus map of a 1.4 Mb domain within 13q14 containing chromatin-chromatin interactions between the DLEU2 locus and several neighboring protein-coding genes. Top tracks depict chromatin modifications associated with promoters (H3K4me3, blue) and enhancers (H3K4me1, green; H3K27ac, purple), as well as DNase I hypersensitive sites marking open chromatin (blue), in K562 cells. The bottom tracks depict chromatin interactions associated with binding of RNA Pol II, the looping factor CTCF or the cohesion subunit Rad21 in K562 cells, determined by ChIA-PET. Highlighted in red are the two DLEU2 alternative promoters, in yellow are promoters of genes active in K562 cells, and in gray promoters of genes that are inactive.

(**B**) Gene-level quantification of the relative expression change between BFU-E progenitors and TER119+ erythroblasts (ERY) for the indicated genes, determined by RNA-seq (n = 2 replicates).

(**C**) Rank of erythroid cell expression level across a collection of 30 cell and tissue types for the indicated genes, determined by RNA-seq. Colored in read are genes showing their lowest expression ranking in erythroid cells.

(**D**) Relative expression of the indicated genes in TER119+ fetal liver cells upon EC6 depletion by shRNAs. Erythroid progenitor-enriched fetal liver cells were transduced by retroviral vectors encoding shRNAs targeting different transcript regions or scramble shRNA control, and effectively transduced cells were induced to differentiate in culture and analyzed for lncRNA expression by qPCR. Values were normalized to those of 18S rRNA, and fold-changes were computed relative to the scramble shRNA control. Data are mean ± s.d. (n = 2 replicates).

(**E**) As in Figure 4G but for EC6 exons (red) and the introns of the indicated genes (green).

EC6 promotes erythroid differentiation by suppressing immune developmental programs

To determine how suppression of proximal genes enables EC6 to promote erythroid cell survival, we first turned to the known functions of its targets. DLEU5 (RFP2/TRIM13/RNF77 in human) is an E3 ubiquitin ligase that has been linked to sensitizing cells to apoptosis via caspase-8 activation (Joo et al. 2011; Tomar et al. 2012; Tomar et al. 2013) and serves as a potent inducer of NF-kB signaling (Matsuda et al. 2003; Garding et al. 2013). In contrast, no functions are known for SPRYD7 (C13ORF1/CLLD6 in human) and EBPL, although SPRYD7 has been found to be stabilized by DLEU5 and its depletion leads to impaired NK-kB inducibility (Garding et al. 2013). Interestingly, loss of DLEU2 or of neighboring KPNA3, DLEU5, or DLEU7, which underlies CLL pathogenesis, is linked to dysregulation of NF-kB signaling (see (Sampath and Calin 2010; Mertens and Stilgenbauer 2012) for commentary), which is known to antagonize erythroid differentiation (Zhang et al. 1998; Liu et al. 2003).

To gain further clues that may explain the impact of EC6 inhibition on erythroid cell physiology, we conducted RNA-seq of shRNA-expressing erythroblasts after 24 hours of culture with erythropoietin-containing differentiation media, and identified 996 differentially expressed genes ($P<0.05$, DESeq), comprising 616 upregulated and 380 downregulated relative to control KD cells (Figure 6A). Upregulated genes were enriched for roles in promoting apoptosis and immune cell development and function (Figure 6B top and S5A), and are generally repressed during normal erythroid differentiation (Figure 6C top). In contrast, downregulated genes were enriched for general functions in cell growth and proliferation (Figure 6B bottom and S5B), and comprise both differentiation-repressed and differentiation-induced genes (Figure 6C bottom). These gene expression changes are consistent with elevated apoptosis and blocked proliferation in EC KD cells, and further reveal a role for EC6 in suppression of immune developmental programs.

We next reasoned that the mechanism by which EC6 counteracts immune cell programming may involve suppression of key immune regulators. To test this hypothesis, we used Ingenuity Pathway Analysis (http://www.ingenuity.com) to identify plausible networks of upstream regulators which may explain gene upregulation in EC6 KD cells (Table S1; Supplemental Experimental Procedures). This analysis identified a network involving P53 and multiple NF-kB complex members as the top mechanistic network whose activation best explains upregulation of gene in the dataset (180 genes; $P < 10^{-17}$, Fisher's test) (Figure 6D). Supporting this notion, independent gene set enrichment analysis identified loci activated by these factors as significantly upregulated upon EC6 inhibition (Figure S5C). Together, these findings suggest that EC6 silences a cluster of physically adjacent N-kB activators to suppress NF-kB signaling and favor differentiation along the erythroid lineage.

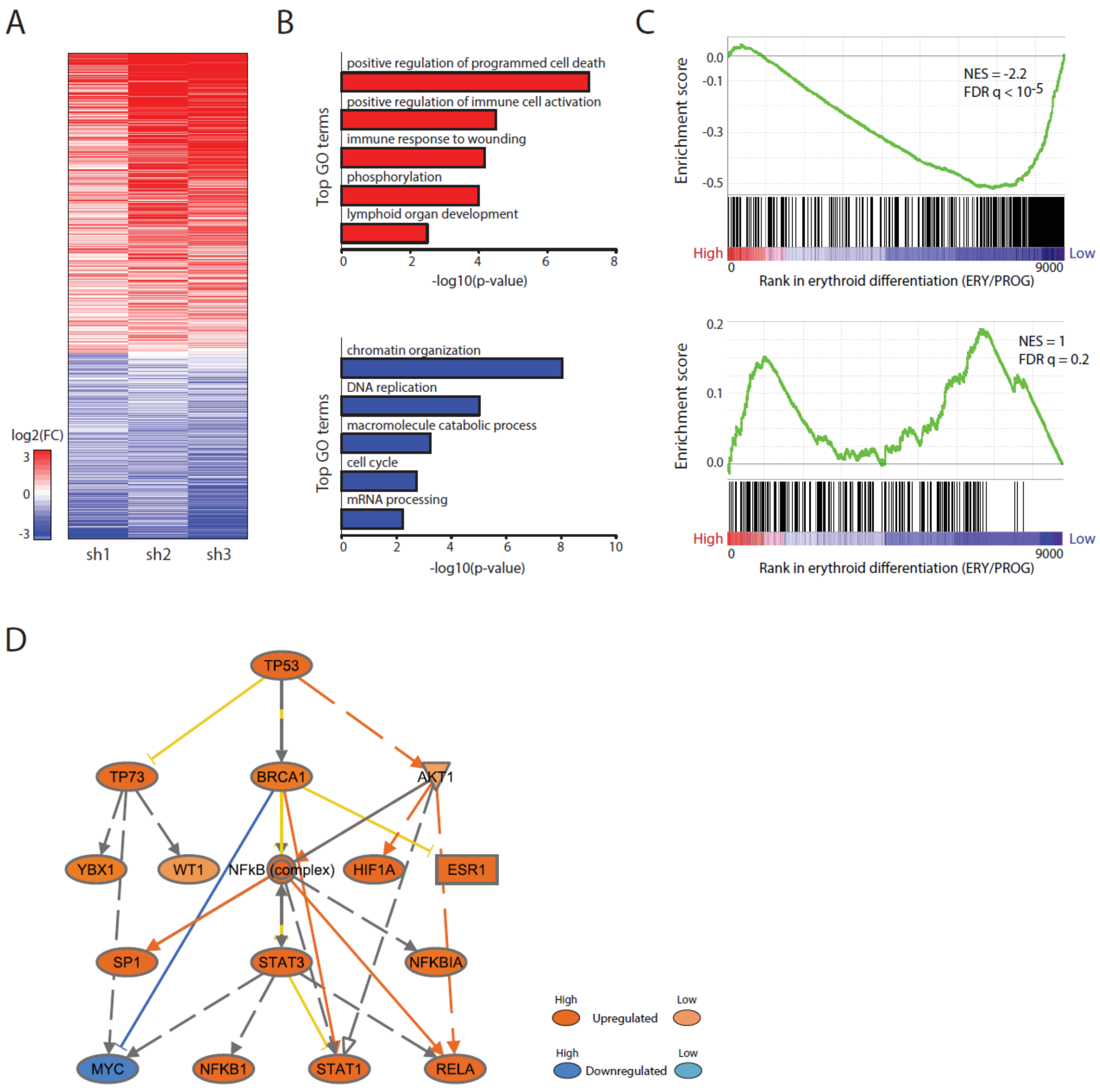


Figure 6. EC6 suppresses immune cell programs via repression of NF-kB signaling.

(**A**) Expression change of 996 genes that are differentially expressed (P <0.05, DESeq) in cultured red blood cells upon shRNA-mediated inhibition of EC6. Changes are log2 expression (FPKM) ratios over control shRNA.

(**B**) Top 5 non-redundant gene ontology (GO) biological process terms enriched (P <0.05, Fisher's test) among mRNA genes that show significantly higher (top) or lower (bottom) expression upon shRNA-mediated inhibition of EC6 relative to control.

(**C**) Gene set enrichment analysis (GSEA) showing overlap between genes upregulated (top) or downregulated (bottom) upon EC6 knockdown and the erythroid differentiation gene signature published previously (Alvarez-Dominguez et al. 2014). NES, normalized enrichment score; FDR, false discovery rate.

(**D**) Network diagram depicting the top mechanistic network whose inhibition best explains (P $<10^{-17}$, Fisher's test) genes upregulated (P <0.05, DESeq) upon EC6 knockdown. Arrows indicate direct transcriptional activation, and blocked lines indicate direct transcriptional repression. Lines colored in blue or yellow indicate that the predicted inhibition of the upstream regulator is consistent or inconsistent with the state of the downstream molecule, respectively, whereas those in gray generated no prediction. See Supplemental Experimental Procedures for molecule shapes details.

EC6 interacts with hnRNP U

lncRNAs have been found to silence other genes via recruitment of general chromatin modifying or DNA methylating complexes, or via direct recruitment or eviction of sequence-specific transcriptional repressors or activators, respectively (Wang and Chang 2011; Guttman and Rinn 2012). To study how EC6 silences its targets, we first examined the chromatin and methylation landscape of the region (Figure S5). We could not find notable deposition of the repressive mark H3K27me3, other than at the DLEU2 distal promoter, however, and did not observe changes in DNA methylation patterns during the course of differentiation. We thus hypothesized that EC6 might instead play a role in facilitating the local chromatin conformation that we originally found to be nucleated at its transcription site. Accordingly, we tested EC6 for binding to known chromatin organization factors by conducting RNA immunoprecipitation experiments in mouse erythroleukemia cells, which activate EC6 upon induction of differentiation (Figure S6). We detected a strong and specific interaction with the nuclear matrix factor hnRNP U, but not with subunits of the Mediator or Cohesin looping complexes or with a subunit of the PRC2 chromatin modifying complex (Figure 7). Interestingly, hnRNP U is known to modulate nuclear architecture via interaction with lncRNAs such as Xist and Firre (Hasegawa et al. 2010; Hacisuleyman et al. 2014). Thus, we speculate that EC6 interacts with the nuclear organization factor hnRNP U potentially to bring its targets into physical proximity and mediate their repression by as-yet-unknown factors, a possibility that warrants further investigation.

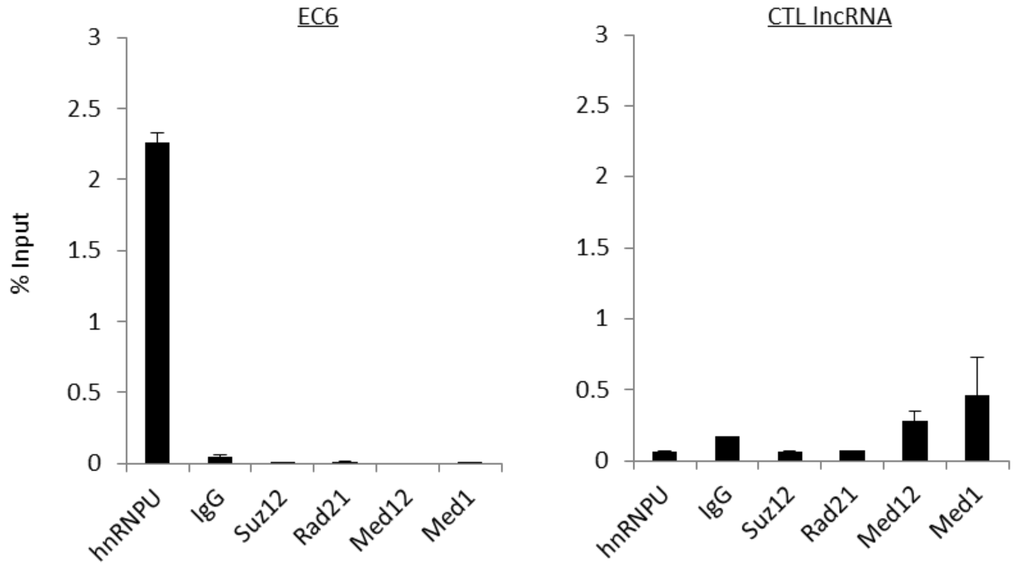


Figure 7. EC6 interacts with hnRNP U.

Association between EC6 or control lncRNA and the indicated proteins in the nucleus of mouse erythroleukemia cells collected after 4 days of differentiation, assessed by native RNA immunoprecipitation followed by qPCR. Data are mean ± s.e.m (n = 3 replicates).

Discussion

Erythropoiesis is tightly modulated to meet physiological demand for red blood cells throughout an organism's lifetime. The transcriptional networks in control of red blood cell development are thus highly responsive to changes in environmental cues, but how they integrate these signals to modulate the balance between progenitor self-renewal, apoptosis, proliferation and ability to differentiate is not well understood. Here, we demonstrate that the long non-coding RNA EC6 is a critical modulator of these outcomes, presenting an example of an lncRNA needed for productive commitment to the erythroid fate and suppression of alternative ones.

Our finding that DLEU2 produces an erythroid-restricted isoform added a new layer of complexity to this intensely studied locus. Monoallelic and biallelic deletions of 13q14, the region containing DLEU2, are found in 55% and 16% of all CLL patients, respectively (Liu et al. 1995; Rosenwald et al. 1999; Dohner et al. 2000; Rawstron et al. 2008). These deletions can span vastly varying lengths but share in common loss of a minimal ~30kb region containing DLEU2 and miR-15a/16-1. Both DLEU2 and the microRNAs play roles in CLL pathogenesis, as indicated by the fact that deletion of only the microRNAs, or of only DLEU2, leads to lymphoproliferative disease, while deletion of both has a synergistic effect leading to a more aggressive disease in mice and in humans (Ouillette et al. 2008; Klein et al. 2010; Ouillette et al. 2011). Much larger deletions comprising DLEU2, miR-15a/16-1 and neighboring genes KCNRG, DLEU5, DLEU7, and RNASEH2B present an even worse disease progression and are embryonic lethal in homozygous mouse models (Lia et al. 2012), suggesting that CLL involves several related genes localized next to each other that act in similar cancer pathways. In fact, most genes at 13q14 (KPNA3, SPRYD7, DLEU5, miR-15a/16-1 and DLEU7) have been functionally implicated in NF-kB signaling. Accordingly, it has been recently proposed that Dleu2 acts *in cis* to downregulate NF-kB modulators at the 13q14 cluster (Garding et al. 2013).

Here, we find tantalizing evidence that the erythroid-specific, GATA1-induced Dleu2 variant EC6 also acts to repress genes at the 13q14 cluster in erythroid cells. These cells express cluster genes EBPL, KPNA3, SPRYD7, DLEU5 and miR-15a/16-1 but not KCNRG or DLEU7, and selectively silence EBPL, SPRYD7 and DLEU5 compared to 30 other cell types examined. Inhibition of EC6 results in EBPL, SPRYD7, and DLEU5 de-repression and leads to activated NF-kB signaling, as evidenced by global gene expression analysis and inference of upstream regulators. These effects sensitize cells to apoptosis and present a block to proliferation, however, opposite to the phenotype of lymphoid CLL cells displaying activated NF-kB. This discrepancy might be explained by opposite outcomes of NF-kB signaling under different cellular contexts. NF-kB signaling is central to homeostasis of blood cell lineages and plays a pivotal role during inflammation as a lymphocyte pro-survival factor (Siebenlist et al. 2005).In lymphoid CLL cells, NF-kB is activated downstream of B-cell receptor signaling by signals from the microenvironment and is thought to contribute to apoptosis resistance(Hewamana et al. 2008; Herishanu et al. 2011). In T cells, however, NF-kB activation downstream of the T-cell receptor plays a prominent pro-apoptotic role during negative T-cell selection (Jimi et al. 2008). In the erythroid lineage, NF-kB activity is high in early committed progenitors but becomes repressed during differentiation downstream of Epo receptor signaling, leading to de-repression of NF-E2 and subsequent activation of erythroid-specific genes (Liu et al. 2003). In the absence of Epo signaling, NF-kB activity remains high and erythroid precursors progressing through the CFU-E stage normally undergo apoptosis (Hattangadi et al. 2011). Thus, fine-tuning of the NF-kB circuit modulates blood cell differentiation and apoptosis with varying outcomes under different cellular contexts and potentially under different stages of normal or malignant development (Chen et al. 2009).

Our results argue that in erythroid cells, failure to repress signaling through NF-kB upon EC6 inhibition leads to improper activation of immune response and development programs and limits activation of erythroid-specific genes, blocking proliferation and sensitizing cells to P53-mediated apoptosis. These effects are opposite to what would be expected from downregulation of microRNAs 15a and 16-1, which are strong NF-kB inducers and instead block proliferation of erythroid cells and sensitize them to apoptosis when overexpressed (Sankaran et al. 2011; Garding et al. 2013). Thus, EC6 functions in NF-kB modulation independent of microRNA generation, apparently through direct repression of EBPL, SPRYD7, and DLEU5. Roles for EBPL and SPRYD7 in NF-kB signaling modulation remain correlative, however, and further studies are needed to clarify their specific contributions to the EC6 KD phenotype and erythropoiesis.

Mechanistically, our data suggest a model whereby EC6 is retained at its site of transcription and brings other genes in the 13q14 domain into physical proximity to mediate their repression. Our work further reveals that such co-regulation is likely mediated by promoter-promoter chromatin contacts between 13q14 genes and the DLEU2 site of transcription. Whether these interactions are facilitated by the lncRNAs themselves or established independently remains unclear. Our finding that EC6 binds the nuclear matrix factor hnRNP U suggests a possible factor mediating these interactions. However, since hnRNP U is a general RNA processing factor, the functional impact of its interaction with EC6 remains to be determined. One possibility is that hnRNP U binds EC6 transcripts at their site of transcription and mediates their co-localization to spatially proximal chromatin sites to enable their co-repression. Thus, 13q14 lncRNAs and hnRNP U may act to organize a local repressive compartment of critical importance for 13q14 silencing in diverse blood lineages. We find no evidence that repression within such compartment occurs via chromatin or DNA modification, warranting further investigation into the repressive factors involved.

Methods

Cell isolation, culture and terminal differentiation assays

Mouse fetal liver erythroid cell purification, culture and differentiation were conducted as described (Zhang et al. 2003; Hattangadi et al. 2010).

RNA-Seq analysis

We examined strand-specific deep RNA sequencing of total RNA depleted of ribosomal RNA isolated from E14.5 FL TER119-positive and -negative cells, and of the poly(A)+ and poly(A)- RNA fractions of TER119+ cells (Alvarez-Dominguez et al. 2014). In addition, we examined RNA-seq reads of poly(A)-selected RNA from FACS-purified primary FL BFU-Es and CFU-Es progenitors and TER119+ erythroblasts (Flygare et al. 2011). Analysis details can be found in the Supplemental Experimental Procedures.

Single-molecule RNA FISH and analysis

RNA FISH was performed as described(Raj et al. 2008). Fluorescence microscopy, image acquisition and image analysis methods were previously published (Neuert et al. 2013; Alvarez-Dominguez et al. 2014). Oligonucleotide probe sets are available upon request.

Retroviral transduction

Purified erythroid progenitors were transduced by MSCV-based retroviruses following previously described protocols(Hattangadi et al. 2010).

Flow cytometry and analysis

For all flow cytometry experiments, we gated on transduced cells (GFP+ subpopulation), for phenotypic analysis. The procedures for immunostaining and flow cytometry analysis of erythroid differentiation, enucleation and cell size were described previously (Ji et al. 2008; Hattangadi et al. 2010; Alvarez-Dominguez et al. 2014).

Apoptosis assays

Annexin V assays were performed using flow cytometry as described previously (Hu et al. 2011).

RNA immunoprecipitation

RNA immunoprecipitation was done as described (Rinn et al. 2007). Briefly, 4-day differentiated mouse erythroleukemia cells were pelleted by centrifugation at 500g for 5min at 4°C. $1x10^7$ cells were re-suspended in 2ml 1X PBS, then lysed in nuclear isolation buffer (2ml nuclear isolation buffer + 6ml water, premixed) for 20 minutes. Nuclei were pelleted by centrifugation at 2,500 g for 15 minutes. Supernatant was discarded (cytosolic fraction) and nuclei were re-supended in 1ml RIP buffer containing the HALT protease and phosphatase inhibitor (Thermo scientific), split into two fractions, and mechanically sheared using a dounce homogenizer with 20 strokes. Nuclear membrane and debris were pelleted by centrifugation at 13,000 rpm for 10 minutes at 4°C. The supernatant was pre-cleared by adding 30μl slurry of protein A/G beads (Santa Cruz, sc-2003) and incubation for 2 hours at 4°C on a rotator. Beads were removed by centrifugation at 2500 g for 1 minute and 10% of the supernatant was removed to a new tube (10% input) and the rest was incubated with antibodies to hnRNP U (abcam, ab20666), IgG (abcam, ab37415), Suz12 (abcam, ab12073), Rad21 (abcam, ab9263), Med12 (Bethyl, A300-774A), or Med1 (Bethyl, A300-793A) for 3h at 4°C on a rotator. Then 60µl slurry of protein A/G beads were added for 2h at 4°C on a rotator. Beads were pelleted by centrifugation at 2500 rpm for 30s and washed 3 times in 500µl RIP for 10 minutes each followed by one wash with 1X PBS. For the isolation of RNA, the beads were re-suspended in 1mL TRIzol after the last wash step and isolated according to the manufacturer's instructions. The RNA pellet was re-suspended in 10µl dH_2O and was directly used for reverse transcription using random hexamers and SuperScript II (Invitrogen). Analysis was done by qPCR as described (Wong et al. 2011).

References

Alvarez-Dominguez JR, Hu W, Yuan B, Shi J, Park SS, Gromatzky AA, van Oudenaarden A, Lodish HF. 2014. Global discovery of erythroid long noncoding RNAs reveals novel regulators of red cell maturation. *Blood* **123**: 570-581.

Brockdorff N, Ashworth A, Kay GF, McCabe VM, Norris DP, Cooper PJ, Swift S, Rastan S. 1992. The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell* **71**: 515-526.

Brown CJ, Hendrich BD, Rupert JL, Lafreniere RG, Xing Y, Lawrence J, Willard HF. 1992. The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* **71**: 527-542.

Bruce LJ, Robinson HC, Guizouarn H, Borgese F, Harrison P, King MJ, Goede JS, Coles SE, Gore DM, Lutz HU et al. 2005. Monovalent cation leaks in human red cells caused by single amino-acid substitutions in the transport domain of the band 3 chloride-bicarbonate exchanger, AE1. *Nat Genet* **37**: 1258-1263.

Chen SS, Raval A, Johnson AJ, Hertlein E, Liu TH, Jin VX, Sherman MH, Liu SJ, Dawson DW, Williams KE et al. 2009. Epigenetic changes during disease progression in a murine model of human chronic lymphocytic leukemia. *Proc Natl Acad Sci U S A* **106**: 13433-13438.

Debernardi S, Fontanella E, De Gregorio L, Pierotti MA, Delia D. 1997. Identification of a novel human kinesin-related gene (HK2) by the cDNA differential display technique. *Genomics* **42**: 67-73.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F et al. 2012. Landscape of transcription in human cells. *Nature* **489**: 101-108.

Flygare J, Rayon Estrada V, Shin C, Gupta S, Lodish HF. 2011. HIF1alpha synergizes with glucocorticoids to promote BFU-E progenitor self-renewal. *Blood* **117**: 3435-3444.

Garding A, Bhattacharya N, Claus R, Ruppel M, Tschuch C, Filarsky K, Idler I, Zucknick M, Caudron-Herger M, Oakes C et al. 2013. Epigenetic upregulation of lncRNAs at 13q14.3 in leukemia is linked to the In Cis downregulation of a gene cluster that targets NF-kB. *PLoS Genet* **9**: e1003373.

Guttman M, Rinn JL. 2012. Modular regulatory principles of large non-coding RNAs. *Nature* **482**: 339-346.

Hacisuleyman E, Goff LA, Trapnell C, Williams A, Henao-Mejia J, Sun L, McClanahan P, Hendrickson DG, Sauvageau M, Kelley DR et al. 2014. Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nat Struct Mol Biol* **21**: 198-206.

Hasegawa Y, Brockdorff N, Kawano S, Tsutui K, Tsutui K, Nakagawa S. 2010. The matrix protein hnRNP U is required for chromosomal localization of Xist RNA. *Dev Cell* **19**: 469-476.

Hattangadi SM, Burke KA, Lodish HF. 2010. Homeodomain-interacting protein kinase 2 plays an important role in normal terminal erythroid differentiation. *Blood* **115**: 4853-4861.

Hattangadi SM, Wong P, Zhang L, Flygare J, Lodish HF. 2011. From stem cell to red cell: regulation of erythropoiesis at multiple levels by multiple proteins, RNAs, and chromatin modifications. *Blood* **118**: 6258-6268.

Heidari N, Phanstiel DH, He C, Grubert F, Jahanbani F, Kasowski M, Zhang MQ, Snyder MP. 2014. Genome-wide map of regulatory interactions in the human genome. *Genome Res* **24**: 1905-1917.

Herishanu Y, Perez-Galan P, Liu D, Biancotto A, Pittaluga S, Vire B, Gibellini F, Njuguna N, Lee E, Stennett L et al. 2011. The lymph node microenvironment promotes B-cell receptor signaling, NF-kappaB activation, and tumor proliferation in chronic lymphocytic leukemia. *Blood* **117**: 563-574.

Hewamana S, Alghazal S, Lin TT, Clement M, Jenkins C, Guzman ML, Jordan CT, Neelakantan S, Crooks PA, Burnett AK et al. 2008. The NF-kappaB subunit Rel A is associated with in vitro survival and clinical disease progression in chronic lymphocytic leukemia and represents a promising therapeutic target. *Blood* **111**: 4681-4689.

Homma N, Takei Y, Tanaka Y, Nakata T, Terada S, Kikkawa M, Noda Y, Hirokawa N. 2003. Kinesin superfamily protein 2A (KIF2A) functions in suppression of collateral branch extension. *Cell* **114**: 229-239.

Hu W, Yuan B, Flygare J, Lodish HF. 2011. Long noncoding RNA-mediated anti-apoptotic activity in murine erythroid terminal differentiation. *Genes Dev* **25**: 2573-2578.

Jarolim P, Palek J, Rubin HL, Prchal JT, Korsgren C, Cohen CM. 1992. Band 3 Tuscaloosa: Pro327----Arg327 substitution in the cytoplasmic domain of erythrocyte band 3 protein associated with spherocytic hemolytic anemia and partial deficiency of protein 4.2. *Blood* **80**: 523-529.

Ji P, Jayapal SR, Lodish HF. 2008. Enucleation of cultured mouse fetal erythroblasts requires Rac GTPases and mDia2. *Nat Cell Biol* **10**: 314-321.

Jimi E, Strickland I, Voll RE, Long M, Ghosh S. 2008. Differential role of the transcription factor NF-kappaB in selection and survival of CD4+ and CD8+ thymocytes. *Immunity* **29**: 523-537.

Joo HM, Kim JY, Jeong JB, Seong KM, Nam SY, Yang KH, Kim CS, Kim HS, Jeong M, An S et al. 2011. Ret finger protein 2 enhances ionizing radiation-induced apoptosis via degradation of AKT and MDM2. *European journal of cell biology* **90**: 420-431.

Klein U, Lia M, Crespo M, Siegel R, Shen Q, Mo T, Ambesi-Impiombato A, Califano A, Migliazza A, Bhagat G et al. 2010. The DLEU2/miR-15a/16-1 cluster controls B cell proliferation and its deletion leads to chronic lymphocytic leukemia. *Cancer Cell* **17**: 28-40.

Lerner M, Harada M, Loven J, Castro J, Davis Z, Oscier D, Henriksson M, Sangfelt O, Grander D, Corcoran MM. 2009. DLEU2, frequently deleted in malignancy, functions as a critical host gene of the cell cycle inhibitory microRNAs miR-15a and miR-16-1. *Exp Cell Res* **315**: 2941-2952.

Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J et al. 2012. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**: 84-98.

Lia M, Carette A, Tang HY, Shen Q, Mo TW, Bhagat G, Dalla-Favera R, Klein U. 2012. Functional dissection of the chromosome 13q14 tumor-suppressor locus using transgenic mouse lines. *Blood* **119**: 2981-2990.

Liu JJ, Hou SC, Shen CK. 2003. Erythroid gene suppression by NF-kappa B. *J Biol Chem* **278**: 19534-19540.

Matsuda A, Suzuki Y, Honda G, Muramatsu S, Matsuzaki O, Nagano Y, Doi T, Shimotohno K, Harada T, Nishida E et al. 2003. Large-scale identification and characterization of human genes that activate NF-kappaB and MAPK signaling pathways. *Oncogene* **22**: 3307-3318.

Mertens D, Stilgenbauer S. 2012. CLL and deletion 13q14: merely the miRs? *Blood* **119**: 2974-2975.

Neuert G, Munsky B, Tan RZ, Teytelman L, Khammash M, van Oudenaarden A. 2013. Systematic identification of signal-activated stochastic gene regulation. *Science* **339**: 584-587.

Ouillette P, Collins R, Shakhan S, Li J, Li C, Shedden K, Malek SN. 2011. The prognostic significance of various 13q14 deletions in chronic lymphocytic leukemia. *Clinical cancer research : an official journal of the American Association for Cancer Research* **17**: 6778-6790.

Ouillette P, Erba H, Kujawski L, Kaminski M, Shedden K, Malek SN. 2008. Integrated genomic profiling of chronic lymphocytic leukemia identifies subtypes of deletion 13q14. *Cancer research* **68**: 1012-1021.

Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. 2008. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods* **5**: 877-879.

Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E et al. 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**: 1311-1323.

Sampath D, Calin GA. 2010. Coding and noncoding: the CLL mix. *Blood* **115**: 3858-3859.

Sankaran VG, Menne TF, Scepanovic D, Vergilio JA, Ji P, Kim J, Thiru P, Orkin SH, Lander ES, Lodish HF. 2011. MicroRNA-15a and -16-1 act via MYB to elevate fetal hemoglobin expression in human trisomy 13. *Proc Natl Acad Sci U S A* **108**: 1519-1524.

Siebenlist U, Brown K, Claudio E. 2005. Control of lymphocyte development by nuclear factor-kappaB. *Nature reviews Immunology* **5**: 435-445.

Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, Groudine M, Bender M, Kaul R, Canfield T et al. 2012. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biology* **13**.

Tomar D, Prajapati P, Sripada L, Singh K, Singh R, Singh AK, Singh R. 2013. TRIM13 regulates caspase-8 ubiquitination, translocation to autophagosomes and activation during ER stress induced cell death. *Biochimica et biophysica acta* **1833**: 3134-3144.

Tomar D, Singh R, Singh AK, Pandya CD, Singh R. 2012. TRIM13 regulates ER stress induced autophagy and clonogenic ability of the cells. *Biochimica et biophysica acta* **1823**: 316-326.

Wang KC, Chang HY. 2011. Molecular mechanisms of long noncoding RNAs. *Mol Cell* **43**: 904-914.

Wong P, Hattangadi SM, Cheng AW, Frampton GM, Young RA, Lodish HF. 2011. Gene induction and repression during terminal erythropoiesis are mediated by distinct epigenetic changes. *Blood* **118**: e128-138.

Zhang J, Socolovsky M, Gross AW, Lodish HF. 2003. Role of Ras signaling in erythroid differentiation of mouse fetal liver cells: functional analysis by a flow cytometry-based novel culture system. *Blood* **102**: 3938-3946.

Zhang MY, Sun SC, Bell L, Miller BA. 1998. NF-kappaB transcription factors are involved in normal erythropoiesis. *Blood* **91**: 4136-4144.

Chapter 3: De novo reconstruction of adipose tissue-specific transcriptomes reveals novel long non-coding RNA regulators of brown adipocyte development and physiology

This chapter represents a manuscript in preparation by the following authors:

Juan R. Alvarez-Dominguez[*], **Zhiqiang Bai**[*], Bingbing Yuan, Dan Xu, Kinyui Alice Lo, Nikolai Slavov, Shuai Chen, Harvey F. Lodish, Lei Sun

Author contributions: J.R.A.-D. designed, performed, and presented the analysis of sequencing reads from RNA-seq, ChIP-seq, CAGE-seq, poly(A)-seq, and ribosome profiling experiments (Figures 1,2,4,5 and Figures S1, S2, S6 and S7); designed, performed and presented bioinformatics analyses (Figures 1 and S1); performed, analyzed and presented data from RNA-FISH experiments (Figure 3 and Figure S3); analyzed and presented qPCR data from tissue expression, differentiation and cell fractionation experiments (Figure 3); analyzed and presented data from microscopy experiments (Figure 4); designed the website, and wrote the manuscript. Z.B. and L.S. performed RNA-seq, tissue expression, differentiation, and cell fractionation experiments (Figures 1, 2, 4, and 5); Z.B., D.X, K.A.L. and S.C. performed, analyzed and presented data from RACE, knockdown, microscopy, extracellular flux, cold exposure, differentiation, overexpression, rescue, RNA IP and pull-down experiments (Figures 4-7 and Figures S3-S5 and S7). B.Y. performed analysis of sequencing reads from RNA-seq experiments and contributed design of and presentation of bioinformatics analyses (Figures 1-2 and Figures S1 and S8). N.S. contributed analysis of mass spectrometry data. J.R.A.D., Z.B., L.S. and H.F.L. designed the study and revised the manuscript.

Acknowledgments: We thank members of the Lodish lab, especially Marko Knoll, for reagents and for helpful discussions and critical comments on this manuscript. We also thank the Whitehead Institute genome technology core for technical support and Andy Nutter-Upham for help with website design and implementation. This research was supported by the Singapore NRF fellowship (NRF-2011NRF-NRFF 001-025) to L.S and by grants from National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases DK047618, DK068348 and from the National Institutes of Health, National Heart, Lung, and Blood Institute, 5P01 HL066105 to HFL.

Brown adipose tissue (BAT) protects against obesity by promoting energy expenditure via uncoupled respiration. To uncover BAT-specific long non-coding RNAs (lncRNAs), we used RNA-seq to reconstruct de novo transcriptomes of mouse brown, inguinal white, and epididymal white fats and identified ~1500 lncRNAs, including 127 BAT-restricted loci induced during differentiation that are often targeted by key regulators PPARγ, C/EBPα and C/EBPβ. One of them, lnc-BATE1, is required for establishment and maintenance of BAT identity and thermogenic capacity. lnc-BATE1 functions *in trans* upstream of key BAT-selective regulators to selectively promote the BAT gene program. We show that lnc-BATE1 binds heterogeneous nuclear ribonucleoprotein U and that both are required for brown adipogenesis. Further, we demonstrate a role for lnc-BATE1 in promoting browning of white adipocytes. Our work provides an annotated catalog for the study of fat depot-selective lncRNAs, available online, and establishes lnc-BATE1 as a novel regulator of BAT development and physiology.

Introduction

Brown adipose tissue (BAT), a specialized mammalian organ for energy expenditure and heat generation, is an attractive therapeutic target for obesity. BAT is densely packed with mitochondria expressing high levels of Uncoupling protein 1 (Ucp1), which facilitates proton leakage to uncouple respiration from ATP synthesis. In rodents, BAT is activated by overfeeding as a physiological response to limit weight gain (Rothwell and Stock 1979). Mice deficient in BAT activity are susceptible to obesity and diabetes (Lowell et al. 1993; Hamann et al. 1996; Feldmann et al. 2009), while mice with increased BAT activity or increased numbers of brown adipocytes within their white fat are healthy and lean (Chiang et al. 2009; Seale et al. 2011; Bostrom et al. 2012).In humans, recent studies have demonstrated the presence of active BAT among adults (Nedergaard et al. 2007; Cypess et al. 2009; van Marken Lichtenbelt et al. 2009; Virtanen et al. 2009). Human BAT activity correlates positively with resting metabolic rate and negatively with body mass index (Cypess et al. 2009; Saito et al. 2009), suggesting that its function may contribute to body weight variability among individuals. Understanding the mechanisms underlying BAT development is thus an area of immense interest.

Previous studies have revealed many protein regulators of BAT development (Kajimura et al. 2010; Villarroya and Vidal-Puig 2013). We and others have shown that microRNAs, such as miR-193b, miR-133 and mIR-155, can also regulate BAT lineage determination and browning of white fat *in vitro* and *in vivo* (Mori et al. 2010; Sun et al. 2011; Trajkovski et al. 2012; Chen et al. 2013; Sun and Trajkovski 2014). Identifying novel RNA regulators of BAT development thus represents an attractive opportunity for finding new therapeutic targets against obesity.

Long non-coding RNAs (lncRNAs) are increasingly recognized as an additional layer of regulation during cell development and disease (Hu et al. 2012; Troy and Sharpless 2012; Fatica and Bozzoni 2013; Alvarez-Dominguez et al. 2014a). We previously showed that a set of lncRNAs common to white and brown adipocytes are essential for adipogenesis (Sun et al. 2013). One of them, lnc-RAP1 (Firre), is exclusively nuclear and interacts with the nuclear matrix factor hnRNP-U to mediate trans-chromosomal interactions between loci encoding known adipogenic factors (Hacisuleyman et al. 2014). Roles for BAT-selective lncRNAs in the specific regulation of BAT development and physiology, however, remain largely unexplored.

Here, we integrate genome-wide surveys of transcription by ultra-deep RNA-seq and chromatin state by ChIP-seq to comprehensively characterize the panorama of lncRNAs active in mouse brown, inguinal white and epididymal white adipose tissues (BAT, iWAT and eWAT, respectively). We uncover >1000 previously unannotated lncRNA genes, including 127 with BAT-restricted expression, many of which are induced during BAT differentiation and are targeted by key adipogenic regulators PPARγ, C/EBPα and C/EBPβ. We focus on one of them, lnc-BATE1, and demonstrate its requirement for the proper development and maintenance of mature brown adipocytes capable of thermogenesis, as well as for white adipocyte browning. lnc-BATE1 acts *in trans* to selectively promote the core BAT gene program and binds hnRNP-U, which is also required for brown adipogenesis, suggesting a model for how it contributes to BAT development and physiology. Our workflow provides a roadmap for the discovery of fat depot-selective lncRNAs contributing to development and function of specific adipocyte lineages, which can be readily implemented through an online resource (https://sites.google.com/site/sunleilab/data/lncrnas).

Results

Global Discovery of Adipose lncRNAs

Our previous work on lncRNAs important for white and brown adipogenesis was limited to existing gene annotations (Sun et al. 2013), which suffer from incompleteness and inaccuracy. To better define lncRNAs active in adipose *in vivo*, including those restricted to different fat depots, we set out to reconstruct *de novo* the transcriptome of primary mouse BAT, iWAT and eWAT (Figure 1A). We performed paired-end sequencing of long poly (A)-selected RNAs from each tissue and mapped about half a billion reads to the mouse genome (Table S1). We then used Cufflinks (Trapnell et al. 2010) to assemble gene and transcript models and to quantify their expression. As a measure of quality, we examined expression estimates for genes annotated by Ensembl (Flicek et al. 2014) and confirmed the high precision and reproducibility of our data (Figures S1A and S1B).

As many as 30% of the transcribed genomic bases in these samples mapped outside of all presently annotated loci (Figure S1C), presenting a large opportunity for gene discovery. To define lncRNA models with high confidence, we focused on transcripts with evidence of at least one splicing event that do not intersect known mRNA exons in the same strand, and implemented a stringent pipeline to evaluate their coding capacity (Figure 1A, Supplemental Experimental Procedures). This analysis classified the BAT, iWAT and eWAT transcriptome into 13342 known mRNA genes, 1535 lncRNA genes, and 566 genes of unclear coding potential based on our criteria. Our lncRNAs do not appear to encode peptides, no matter how small, as evidenced by mass spectrometry, by ribosome profiling, and by computational assessment of coding capacity (Figures 1B, S1D and S1E). We further confirmed our ability to delineate authentic lncRNA transcripts by finding a specific enrichment for 5’ CAGE and 3’ poly(A) tags at their transcription start and end sites, respectively (Figure 1C). Importantly, 1237 lncRNA transcripts from 1032 loci do not intersect Ensembl, RefSeq or UCSC annotations, highlighting the necessity of our *de novo* reconstruction approach. Overall, ~90% of our lncRNAs are supported by at least one other source of unbiased experimental evidence in addition to RNA-seq (Figure S1G; Supplemental Experimental Procedures), globally validating our lncRNA predictions.

Analysis of the properties of adipose lncRNAs revealed that they are globally lower-expressed than mRNAs, yet share the same marks of active transcription at their promoters (Figures 1C, 1D and S1H), consistent with being independent Pol II transcripts. About half of the lncRNAs originate from active enhancer elements, as defined by a high H3K4me1/H3K4me3 ratio, agreeing with recent findings (Natoli and Andrau 2012). As characteristic of mouse (Guttman et al. 2009; Guttman et al. 2010) and human lncRNAs (Cabili et al. 2011; Derrien et al. 2012), adipose lncRNAs have fewer exons and are thus shorter than mRNAs, and they show higher primary sequence conservation in promoters than in exons (Figures S1I-S1L). Importantly, 297 out of 1535 lncRNA genes are detectable (FPKM >0) in only one of the three adipose tissues examined (Figure 1E), despite comparable coverage across samples (Figure S1F), indicating substantial depot-restricted expression. About a third of these depot-specific loci are exclusive to BAT and resemble genes encoding key BAT-intrinsic proteins, as illustrated by lnc-BATE1 (Figure 1F), a lncRNA that we focus on later because of its remarkable BAT specificity and induction during brown adipogenesis (see below). Thus, we provide a comprehensive catalog of *bona fide* and mostly unannotated adipose lncRNAs (Table S2), many of which may contribute to development or function of distinct adipocyte lineages.

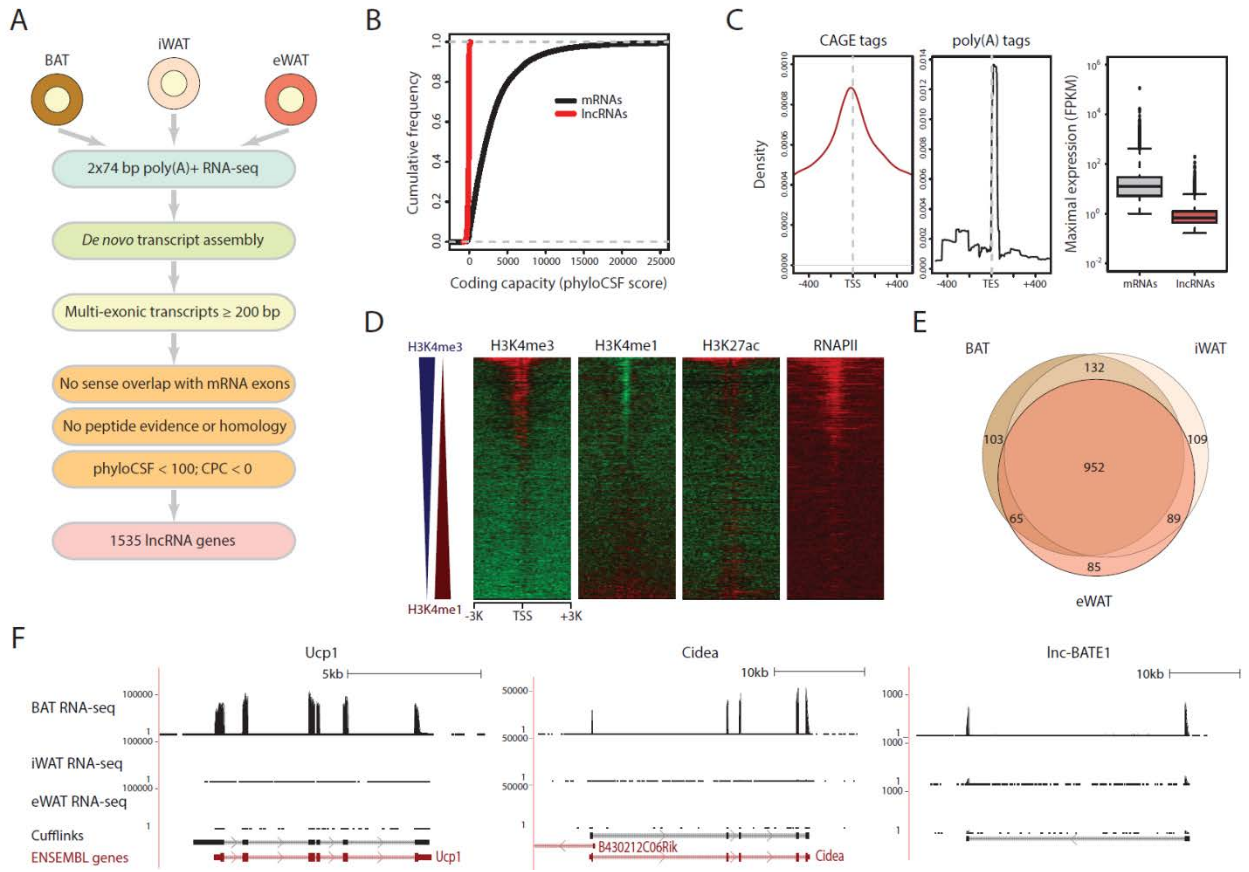


Figure 1. Global Discovery of Adipose Tissue lncRNAs.

(**A**) Pipeline for lncRNA discovery. See text and Supplemental Experimental Procedures for details.

(**B**) Coding capacity of adipose tissue-expressed mRNAs and lncRNAs as estimated by phyloCSF (Lin et al. 2011).

(**C**) Density of CAGE tags (left) and poly(A) tags (center) within 1 kb of lncRNA transcription start sites (TSS) or end sites (TES), respectively. (Right) Box plots of maximal gene-level expression distributions for adipose-expressed mRNAs (maximal FPKM >1) and lncRNAs (maximal FPKM >0.1).

(**D**) Evidence of histone marking, open chromatin and RNA Pol II binding within TSS ± 3kb regions of adipose-tissue expressed lncRNAs. Histone marks enriched at promoters (H3K4me3), enhancers (H3K4me1), and active promoters or enhancers (H3K427ac) in BAT from ENCODE (Stamatoyannopoulos et al. 2012) are shown, as well as binding of serine 5 phosphorylated RNA Pol II (RNAPII) in cultured brown adipocytes (Lee et al. 2013). Color intensity represents the log2 signal enrichment over input. Heat maps are sorted by the difference in enrichment for H3K4me3 and H3K4me1, depicted by blue and red triangles to the left, respectively.

(**E**) Overlap between the number of lncRNAs detected (FPKM >0) in BAT, iWAT and eWAT.

(**F**) Examples of BAT-restricted mRNAs and lncRNAs. UCSC genome browser tracks depict RNA-seq signal for poly(A)+ RNA from BAT, iWAT and eWAT as density of mapped reads. The bottom tracks depict *de novo* transcript models by Cufflinks and Ensembl gene annotations. Left-to-right arrows indicate transcripts in the plus strand; right-to-left arrows indicate transcripts in the minus strand.

Adipose Tissue-specific lncRNAs and their Regulation

To examine the tissue specificity of adipose-expressed lncRNAs, we profiled their expression across a panel of 30 primary tissues from the mouse ENCODE project (Stamatoyannopoulos et al. 2012) (Figure 2A). We developed an algorithm (detailed in Supplemental Experimental Procedures) to score the specificity of each gene to each tissue by its fractional expression level. By this measure, we find greater tissue specificity among lncRNAs than among mRNAs (Figure S2A), consistent with previous studies (Cabili et al. 2011; Derrien et al. 2012). To identify depot-specific lncRNAs, we used an empirical threshold to define tissue-restricted genes and selected those with an adipose subtype as the tissue of maximal specificity (Supplemental Experimental Procedures). This strategy yielded 127 BAT-, 81 iWAT-, and 240 eWAT-specific lncRNAs (Figure S2B and Table S2). Thus, we also find greater fat subtype specificity among lncRNAs (~30%) than among protein-coding genes (7%). This is illustrated by lnc-BATE1 (Figure S2C), which is highly abundant in BAT but not in any of the other tissue types examined, including the lineage-related skeletal muscle (Timmons et al. 2007; Kajimura et al. 2009).

To investigate the regulatory basis for fat depot-restricted lncRNA expression, we first examined published global occupancy maps of PPARγ, a master adipogenic TF, assessed by ChIP-seq in primary BAT and eWAT (Rajakumari et al. 2013). We found that PPARγ targets the promoters of 754 (~50%) adipose lncRNAs in BAT or in eWAT, as evidenced by binding events inferred by MACS (Zhang et al. 2008) within their TSS ± 3kb regions (Figure S2D). Importantly, BAT-selective lncRNAs were enriched for BAT-specific PPARγ promoter binding events (Figure 2B), as exemplified by lnc-BATE10 (Figure 2C), whereas eWAT-selective lncRNAs (such as lnc-eWATE5) were enriched for eWAT-specific ones, a mode of regulation shared with key depot specific proteins (Figures 2B and 2C). Among depot-specific lncRNAs whose promoters are bound by PPARγ in both tissues, we still found quantitatively richer PPARγ binding in their tissue of selective expression (Figures S2D and S2E).

We then focused on lncRNAs active in BAT, for which profiles of expression, histone modification and TF binding during the course of brown adipogenesis in culture are available (Lee et al. 2013; Sun et al. 2013) (Figure S2F). As expected from their subtype-specific regulation, BAT-selective lncRNAs were specifically enriched for induction during brown adipogenesis, with 49 (38%) induced >2-fold as precursors differentiate into brown adipocytes (Figure 2D). lncRNA activation was reflected at the chromatin level and was associated with binding of C/EBPα, C/EBPβ and PPARγ early during differentiation (Figures 2E, S2F and S2G). The most predictive activation event was C/EBPα targeting, most of which represented new binding events at differentiation day 2, while co-targeting by C/EBPα, C/EBPβ and PPARγ was associated with the strongest induction levels (Figure S2G). These findings characterize multiple BAT-selective lncRNAs that are targeted by common adipogenic TFs, often in a BAT-specific manner, and show dynamic regulation during differentiation.

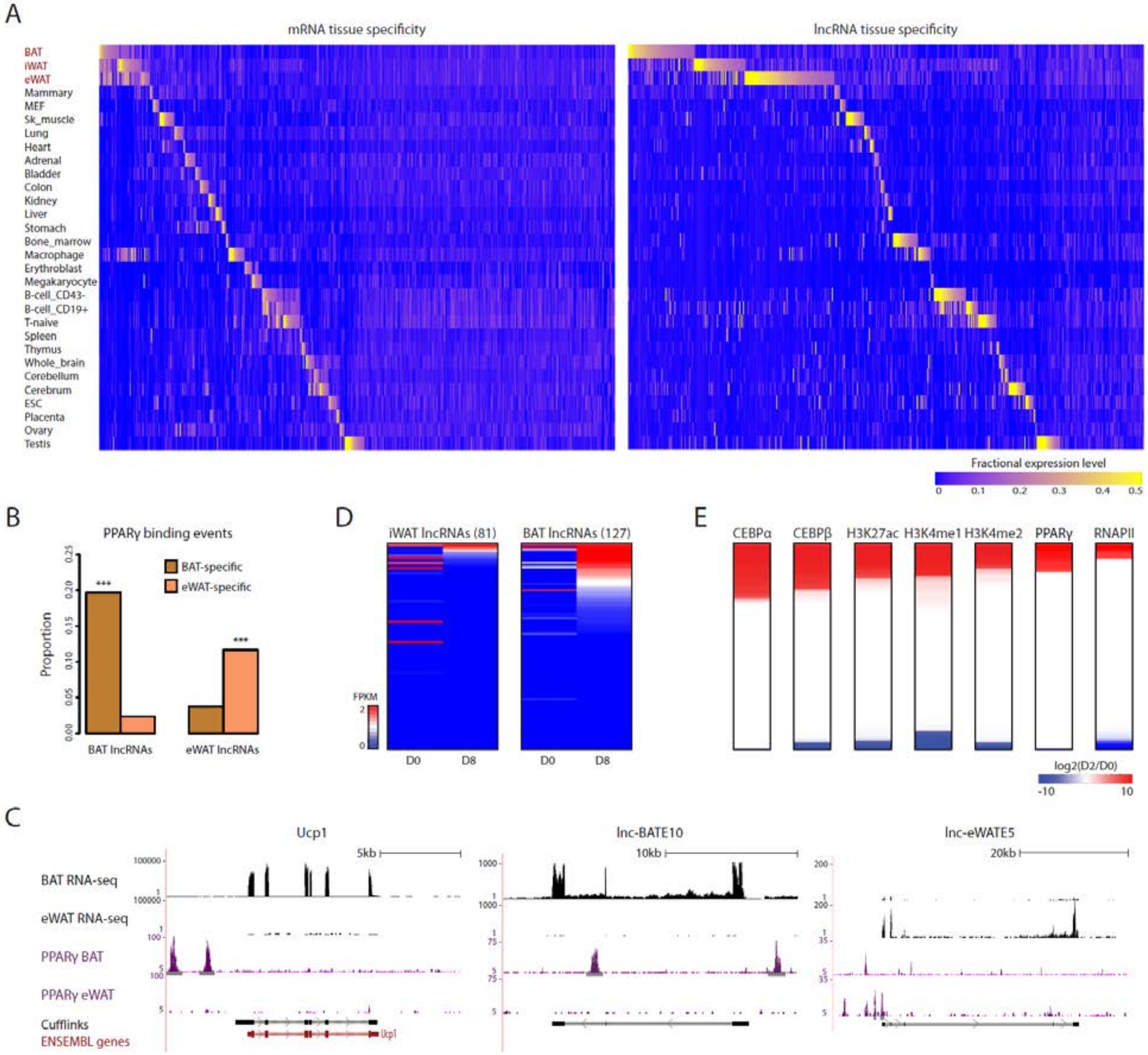


Figure 2. Adipose tissue-specific lncRNAs and their regulation.

(**A**) Abundance of adipose-expressed mRNAs (13342) and lncRNAs (1535) across 30 tissues from ENCODE, based on our *de novo* gene models. Color intensity represents the fractional expression across all the tissues examined. Rows are ordered based on empirical thresholding to distinguish tissue-restricted from broadly-expressed genes (Supplemental Experimental Procedures).

(**B**) Proportion of BAT-specific and eWAT-specific lncRNAs with promoter-proximal (TSS ± 3kb) BAT- or eWAT-specific PPARγ binding (Rajakumari et al. 2013), as determined by peaks of ChIP-seq signal enrichment. ***p <0.001 (Kolmogorov-Smirnov test).

(**C**) Examples of BAT- and eWAT-restricted lncRNAs showing BAT- or eWAT-specific PPARγ promoter-proximal binding, respectively. Ucp1, a BAT-restricted mRNA locus targeted by PPARγ specifically in BAT, is shown for comparison. UCSC genome browser tracks depict RNA-seq signal for poly(A)+ RNA from BAT and eWAT as density of mapped reads (black) and ChIP-seq signal for PPARγ binding in BAT and eWAT as density of processed signal enrichment (purple). Peaks of signal enrichment are shown in gray under the ChIP-seq tracks. The bottom tracks depict *de novo* transcript models by Cufflinks and Ensembl gene annotations as in Figure 1F.

(**D**) Expression dynamics of BAT-specific and iWAT-specific lncRNAs during brown adipogenesis in culture. Shown are abundance estimates (FPKM) from poly(A)+ RNA-seq of cultured brown pre-adipocytes (D0) and cultured brown adipocytes (D8) (Sun et al. 2013), based on our *de novo* gene models.

(**E**) Dynamic changes in promoter-proximal chromatin marking and transcription factor binding among BAT-specific lncRNAs during brown adipogenesis in culture. Shown are changes in ChIP-signal for binding of C/EBPα, C/EBPβ, PPARγ, and RNA Pol II, as well as H3K27ac, H3K4me1, and H3K4me2 marking, between immortalized brown pre-adipocytes before (D0) and after (D2) adipogenic induction (Lee et al. 2013). Changes are shown as the log2 ratio of normalized read counts within TSS ± 3kb regions.

Validation of BAT-selective lncRNAs

To focus our validation efforts, we ranked candidate lncRNAs by their BAT specificity score, differential expression during brown adipogenesis, and BAT expression level as estimated by RNA-seq. We selected the top 40 candidates and independently assessed their BAT selectivity by qPCR. For 38 out of 40 lncRNAs, we confirmed that their expression in BAT was significantly higher than the average expression across 12 major mouse organs, with 26 showing the highest absolute levels in BAT (Figure 3A). We also monitored their expression during brown adipocyte differentiation in culture using qPCR, and found that all 40 candidates were upregulated (Figure 3B). Next, to examine their subcellular distribution, we isolated RNA from the cytoplasmic and nuclear fractions of differentiated primary brown adipocytes and quantified their expression by qPCR (Figure 3C). Most of our candidates (27 out of 40) were enriched in the nucleus, with four of them closely resembling the 47S pre-rRNA at >90% nuclear retention, consistent with previous observations (Derrien et al. 2012; Alvarez-Dominguez et al. 2014b). Others, including lnc-BATE1, were similarly present in the nucleus and in the cytoplasm. These results demonstrate the specific involvement of lncRNAs in the brown adipocyte developmental program, suggesting predominant roles in the nucleus.

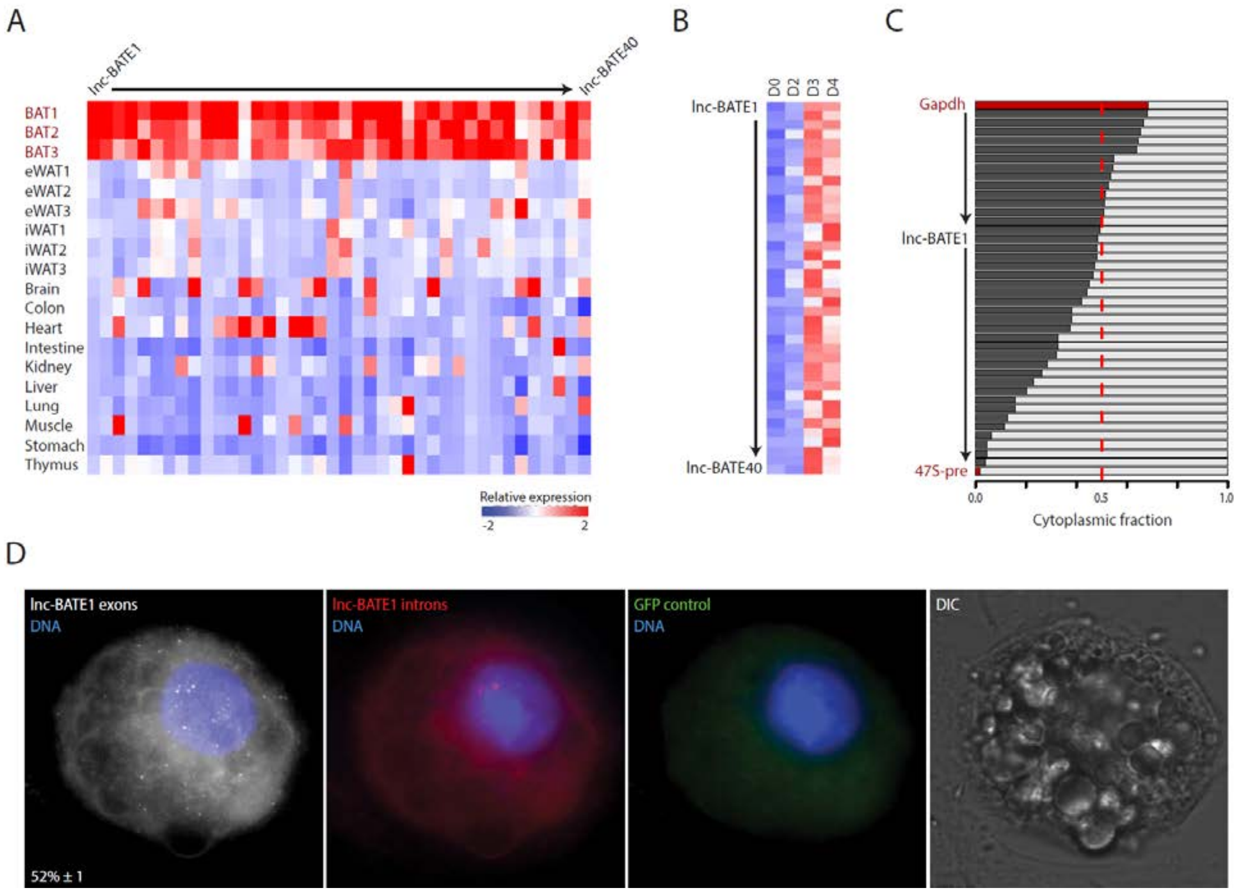


Figure 3. Validation of BAT-selective lncRNAs.

(**A**) Expression of 40 lncRNAs evaluated in BAT, eWAT and iWAT (n =3) and across 10 primary tissue samples by qPCR. Color intensity represents column mean-centered expression.

(**B**) Upregulation of 40 BAT lncRNAs during brown adipocyte differentiation. Expression values during a 4-day differentiation timecourse of cultured mouse pre-adipocytes were determined by qPCR (n =3). Color intensity represents row mean-centered expression.

(**C**) Subcellular localization of 40 BAT lncRNAs. The relative proportion of cytoplasmic (black) and nuclear (gray) expression was assessed by qPCR (n =3). Gapdh mRNA and 47S pre-rRNA represent predominantly cytoplasmic and predominantly nuclear controls, respectively. Rows are ordered from highest to lowest cytoplasmic fraction.

(**D**) Detection of individual lnc-BATE1 transcripts by single-molecule RNA FISH. Shown are maximum z-stack projections of fluorescence microscopy images. lncRNA molecules and DNA staining are pseudocolored as indicated at the top left corner of each panel. Shown at the bottom left panel corner for lnc-BATE1 exons is the mean ± SEM (n=3) percent of nuclear-localized transcripts. GFP control indicates background fluorescence measured in the GFP channel. DIC indicates imagining in the differential interference contrast channel.

lnc-BATE1 is Required for Brown Adipocyte Development, Function, and Maintenance

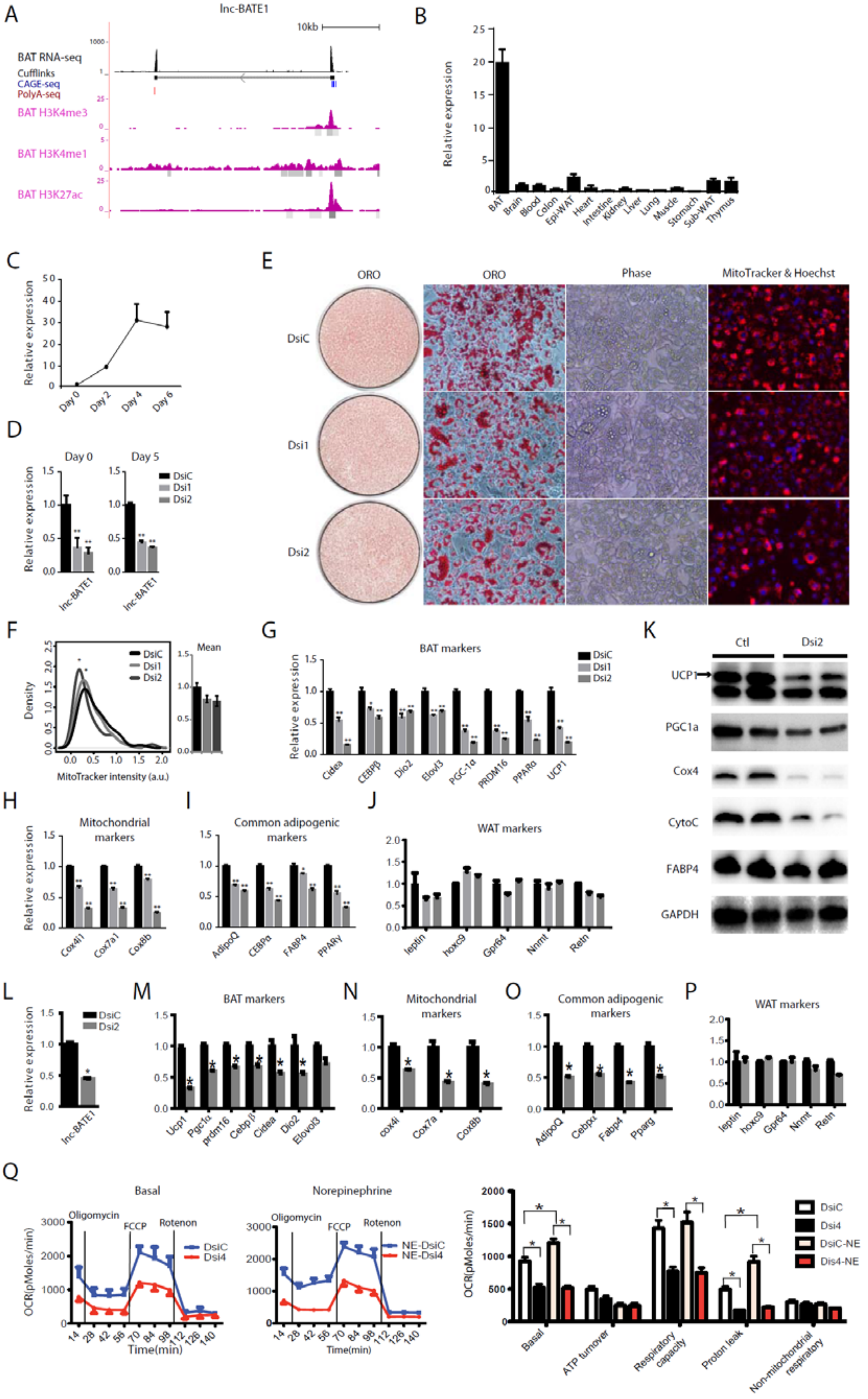
Our ranking of adipose lncRNAs by their abundance, regulation and depot-selectivity identified lnc-BATE1 as a top candidate modulator of brown adipogenesis. lnc-BATE1 is an independent intergenic locus targeted by C/EBPα, C/EBPβ and PPARγ that gives rise to polyadenylated transcripts spliced from two exons (Figures 4A and S2H), coincident with the RefSeq gene NR_077224. 5’ and 3’ RACE revealed 3 transcript variants with slightly different transcription start sites and a common termination site (Figures S3A and S3B). lnc-BATE1 is equally distributed between cytosol and nucleus, as evidenced by cell fractionation and by single-molecule RNA FISH, which additionally indicated mean levels of 18 ± 2 transcripts per cell (Figures 3C, 3D and S3C). Consistent with RNA-seq data, lnc-BATE1 is highly specific to BAT and upregulated 30-fold during differentiation (Figures 4B and 4C).

To investigate the function of lnc-BATE1, we designed Dicer-substrate siRNAs (DsiRNAs) and transfected them into primary brown pre-adipocytes, followed by induction of differentiation. Over 70% knockdown was achieved at differentiation day 0, and about 60% remained at day 5 (Figure 4D). Depleting lnc-BATE1 resulted in very limited changes in lipid accumulation and cell morphology during differentiation (Figure 4E), but significantly reduced mRNA levels of all brown fat markers examined, including Cidea, C/EBPβ, Dio2, Elovl3, PGC1α, PRDM16, PPARα, and Ucp1 (Figure 4G), as well as mitochondrial markers Cox4i, Cox7a and Cox8b (Figure 4H), and, to a lesser extent, of common adipogenic markers AdipoQ, C/EBPα, Fabp4, and PPPARγ (Figure 4I). Knockdown of lnc-BATE1 using traditional siRNAs or retroviral shRNAs targeting different transcript regions showed very similar phenotypes (Figures S3D-J), which further correlated with knockdown efficiency, indicating that the molecular phenotype of lnc-BATE1 depletion is unlikely due to RNAi off-target effects.

In contrast to its dramatic effects on BAT gene expression, lnc-BATE1 depletion did not affect expression of WAT markers such as Lep, HoxC9, Gpr64, Nnmt and Retn (Harms et al. 2014) (Figure 4J), suggesting a preferential influence on BAT-selective genes, which is further supported by global gene expression analysis (see below; Figure 6D). Western Blot data further confirmed reduced protein levels of BAT-selective genes (Ucp1, Pgc1a) and mitochondrial markers (Cox4, CytoC).

Inhibition of BAT-intrinsic genes upon lnc-BATE1 loss could be due to preferential disruption of the BAT gene program or be an indirect effect of poor cell differentiation. To distinguish between these two possibilities, we depleted lnc-BATE1 in mature brown adipocytes using an electroporation method that resulted in ~60% knockdown (Figure 4L). We observed no significant changes in cell morphology at 72 hours post-transfection (not shown), but found a significant reduction in BAT, mitochondrial, and common adipogenic markers (Figures 4M-O) but not WAT markers (Figure 4P), consistent with the phenotypes observed upon lnc-BATE1 depletion at the beginning of differentiation. Thus, lnc-BATE1 is essential for the selective establishment of the BAT gene expression program in developing brown adipocytes and its maintenance in mature ones.

lnc-BATE1 inhibition also affected mitochondrial biogenesis, as indicated by decreased MitoTracker staining (Figures 4E and 4F), downregulation of mitochondrial markers (Figures 4H and S3J), and loss of Ucp1 protein (Figure 4K). To directly examine whether lnc-BATE1 knockdown alters cellular respiration, we performed extracellular flux analysis and measured cellular oxygen consumption in the presence and absence of the adrenergic agent norepinephrine (NE) to stimulate thermogenesis (Figure 4Q). In the absence of NE, depleting lnc-BATE1 led to markedly decreased oxygen consumption rates (OCR), attributed to lower basal and maximal respiratory capacity and to markedly lower proton leakage. Upon NE treatment, basal oxygen consumption and proton leakage were clearly increased in control cells but not in lnc-BATE1-depleted cells. These data demonstrate that lnc-BATE1 is essential for thermogenic function.



A
lnc-BATE1
10kb
BAT RNA-seq
Cufflinks
CAGE-seq
PolyA-seq
BAT H3K4me3
BAT H3K4me1
BAT H3K27ac
B
Relative expression
BAT
Brain
Blood
Colon
Epi-WAT
Heart
Intestine
Kidney
Liver
Lung
Muscle
Stomach
Sub-WAT
Thymus
C
Relative expression
Day 0
Day 2
Day 4
Day 6
D
Day 0
Day 5
DsiC
Dsi1
Dsi2
lnc-BATE1
E
ORO
ORO
Phase
MitoTracker & Hoechst
DsiC
Dsi1
Dsi2
F
Density
MitoTracker intensity (a.u.)
Mean
G
BAT markers
Cidea
CEBPβ
Dio2
Elovl3
PGC-1α
PRDM16
PPARα
UCP1
K
Ctl
Dsi2
UCP1
PGC1a
Cox4
CytoC
FABP4
GAPDH
H
Mitochondrial markers
Cox4i1
Cox7a1
Cox8b
I
Common adipogenic markers
AdipoQ
CEBPα
FABP4
PPARγ
J
WAT markers
leptin
hoxc9
Gpr64
Nnmt
Retn
L
DsiC
Dsi2
lnc-BATE1
M
BAT markers
Ucp1
Pgc1α
prdm16
Cebp β
Cidea
Dio2
Elovol3
N
Mitochondrial markers
cox4i
Cox7a
Cox8b
O
Common adipogenic markers
AdipoQ
Cebpα
Fabp4
Pparg
P
WAT markers
leptin
hoxc9
Gpr64
Nnmt
Retn
Q
Basal
Norepinephrine
Oligomycin
FCCP
Rotenon
DsiC
Dsi4
NE-DsiC
NE-Dsi4
OCR(pMoles/min)
Time(min)
DsiC
Dsi4
DsiC-NE
Dis4-NE
Basal
ATP turnover
Respiratory capacity
Proton leak
Non-mitochondrial respiratory

Figure 4. lnc-BATE1 is required for brown adipocyte differentiation.

(**A**) Locus map of lnc-BATE1. UCSC genome browser track 1 depicts BAT poly(A)+ RNA-seq signal as density of mapped reads. Track 2 depicts *de novo* transcript models by Cufflinks; right-to-left arrow indicates transcript in the minus strand. Tracks 3-4 display RNA 5'capping and 3'polyadenylation sites as evidenced by CAGE tags (blue) and poly(A) tags (red), respectively; only tags from the strand of transcription are shown. Tracks 5-7 display ENCODE BAT ChIP-seq signal from H3K4me3, H3K4me1 and H3K27ac marks, respectively, as density of processed signal enrichment; peaks of signal enrichment are shown in gray under each track.

(**B**) Expression of lnc-BATE1 across 14 mouse tissues assessed by qPCR.

(**C**) Expression of lnc-BATE1 during the course of brown adipocyte differentiation in culture assessed by qPCR.

(**D**) Expression of lnc-BATE1 in cultured brown adipocytes transfected with DsiRNA control (DsiC) or DsiRNAs targeting lnc-BATE1 (Dsi1 and Dsi2) and collected for qPCR at differentiation days 0 and 5.

(**E**) Representative images of DsiRNA-treated cultured brown adipocytes at differentiation day 5 labelled with Oil red O (ORO, red) or MitoTracker® Deep Red FM (red) plus Hoechst (blue), respectively.

(**F**) Quantification of integrated density signal of MitoTracker® fluorescence in individual cells from (E). Signal distributions are shown to the left and their mean values to the right.

(**G-J**) Expression of BAT markers (G), mitochondrial markers (H), common adipogenic markers (I), and WAT markers (J) in DsiRNA-treated cultured day 5 brown adipocytes.

(**K**) Protein levels of BAT, mitochondrial and common adipogenic markers assessed by western blot on cell lysates from DsiRNA-treated cultured day 5 brown adipocytes.

(**L**) Expression of lnc-BATE1 measured by qPCR in mature brown adipocytes transfected with DsiRNA control (DsiC) or DsiRNA targeting lnc-BATE1 (Dsi2).

(**M-P**) Expression of BAT markers (M), mitochondrial markers (N), common adipogenic markers (O), and WAT markers (P) in DsiRNA-treated mature brown adipocytes.

(**Q**) Representative metabolic flux curves from cultured DsiRNA-treated cultured day 5 brown adipocytes in the presence and absence of norepinephrine (left). Oxygen consumption rate data represent measurements from 10 wells ± s.e.m. and are normalized by protein concentration (right). qPCR data were normalized by the mRNA level of housekeeping gene RPL23. Error bars are mean ± s.e.m., n=3. **P* ≤0.05, ***P* ≤0.01.

lnc-BATE1 Stimulates White Fat Browning

To determine whether lnc-BATE1 is induced during browning of subcutaneous white fat, we exposed 12-week old mice to 4^{o}C for one week, harvested inguinal WAT to enrich for beige adipocytes, and performed qPCR to examine lnc-BATE1 expression. We found that lnc-BATE1 is upregulated 3-4 fold during cold-induced browning (Figure 5A), suggesting a role in adaptive thermogenesis. To test this, we used retroviral shRNAs to infect primary inguinal white pre-adipocytes, followed by induction of differentiation in the absence or presence of norepinephrine. Similar to the phenotypes seen in brown adipocytes, loss of lnc-BATE1 in inguinal white adipocytes results in limited effects on lipid accumulation and cell morphology (not shown), but leads to impaired expression of the examined BAT, mitochondrial and, to a lesser extent, common adipogenic markers (Figures 5B and 5C). In contrast, 6 out of 7 WAT-selective genes were not downregulated and, in fact, 4 were significantly upregulated (Figure 5D). In the presence of norepinephrine, we further found that induction of thermogenic genes Ucp1 and PGC1α is blunted by lnc-BATE1 depletion (Figure 5E).

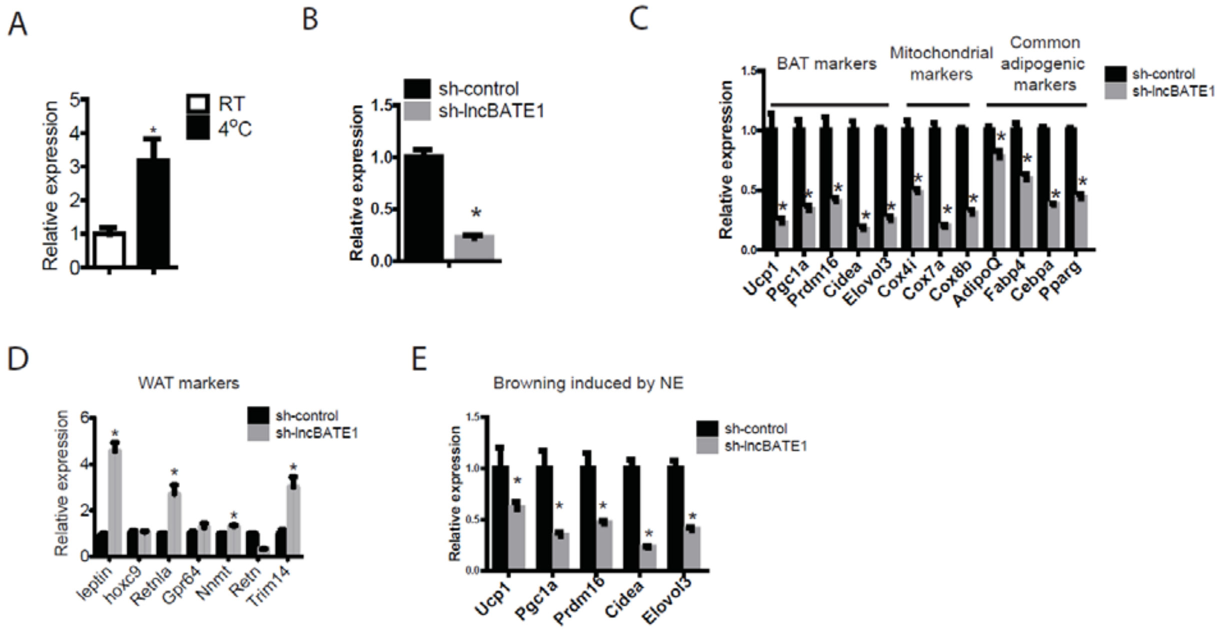


Figure 5. lnc-BATE1 plays an important role during browning of white adipocytes.

(**A**) Induction of lnc-BATE1 expression during cold-induced browning of subcutaneous white fat.

(**B-D**) Inhibition of lnc-BATE1 in inguinal white adipocytes (B) impairs expression of key BAT, mitochondrial, and common adipogenic markers (C) but not WAT markers (D).

(**E**) Inhibition of lnc-BATE1 in inguinal white adipocytes impairs norepinephrine-induced thermogenic gene expression. Error bars are s.e.m., n =3. *P ≤0.05, **P ≤0.01.

lnc-BATE1 is Necessary but not Sufficient for Brown Adipogenesis

To study the impact of lnc-BATE1 gain-of-function on brown adipogenesis, we cloned all three isoforms into a retroviral vector which we transduced into brown pre-adipocytes followed by induction of differentiation. We could not observe any significant changes in lipid accumulation, cell morphology (not shown) or enhancement of BAT marker gene expression, whether using standard or 10-fold diluted differentiation cocktail (Figures S4A and S4B), indicating that ectopic expression of lnc-BATE1 is not sufficient to stimulate brown adipocyte development. Since lnc-BATE1 is essential for white adipocyte browning, we asked whether its gain-of-function is sufficient to promote browning. Overexpression of lnc-BATE1 in primary inguinal and epididymal white pre-adipocytes followed by differentiation induction did not result in any significant change in BAT-selective genes, however (Figures S4C-E). Finally, we examined whether lnc-BATE1 functions in brown adipocyte lineage determination from myoblast progenitors by ectopically expressing lnc-BATE1 in C2C12 myoblasts followed by induction of differentiation, but did not find significant changes in cell morphology (not shown) or in expression of all the myogenic markers examined (Figure S4F). Thus, lnc-BATE1 is essential but not sufficient for brown adipocyte development and function.

lnc-BATE1 Regulates the Core Gene Network of Brown Adipocyte Differentiation

To gain further insights into lnc-BATE1 function from global gene expression analysis, we performed RNA-seq in DsiRNA-treated brown adipocytes at two differentiation time points, and identified 1,014 differentially expressed genes ($P < 0.05$, DESeq) comprising 781 enriched and 233 depleted in lnc-BATE1-inhibited cells relative to control cells (Figure 6A). Higher-expressed genes were enriched for general functions in cell division, cell adhesion and signaling processes that are normally downregulated during adipogenesis (Figures 6B top and S5A), whereas lower-expressed ones comprised genes specifically associated with brown adipocyte differentiation, as well as mitochondrial biogenesis and function, that fail to be activated upon loss of lnc-BATE1 (Figures 6B bottom and S5B). Gene set enrichment analysis (Subramanian et al. 2005) of lnc-BATE1 KD depleted genes further demonstrated a highly significant overlap with the brown adipocyte differentiation gene signature published previously (Figure 5C; (Sun et al. 2013)). These results indicate that lnc-BATE1 promotes a genetic program associated with brown adipogenesis.

Suppression of brown adipogenesis upon lnc-BATE1 loss could be due to suppression of genes important for adipogenesis in general. To test this possibility, we exploited our tissue specificity scoring strategy to define groups of BAT-specific, WAT-specific, and common adipogenic protein-coding genes for which we studied the impact of lnc-BATE1 KD on expression levels (Figures 6D and S5C; see Supplemental Experimental Procedures). We found that inhibiting lnc-BATE1 has no effect on WAT-selective genes and a more profound effect on BAT-selective vs. common adipogenic factors, such that significant downregulation is predominantly observed for BAT-selective genes (~15%) vs. common adipogenic factors (~5%). Thus, lnc-BATE1 selectively promotes a BAT-specific genetic program.

We reasoned that the mechanism by which lnc-BATE1 selectively promotes the global BAT-specific program may be through stimulation of key BAT-selective transcription factors. To test this hypothesis, we used Ingenuity Pathway Analysis to identify upstream regulators that may be responsible for global gene downregulation in lnc-BATE1 KD cells (Supplemental Experimental Procedures and Table S3). This analysis identified PGC1α, ESRRα, PPARα, and

PPARγ as the top transcription modulators whose inhibition best explains the downregulated genes (P $<10^{-14}$ – P $<10^{-5}$, Fisher's test) (Figure 6E). Supporting this notion, independent gene set enrichment analysis identified genes activated by these factors as significantly depleted upon lnc-BATE1 inhibition (empirical P $<10^{-5}$; Figure S5D). Of note, of these core upstream regulators only PGC1α was significantly depleted by differentiation day 3, indicating that it is among the earliest regulators suppressed in lnc-BATE1 KD cells. Accordingly, we found a significant overlap between lnc-BATE1 KD depleted genes and the published gene signature from concurrent genetic loss of PGC1α and depletion of PGC1β (Figure S5E; (Uldry et al. 2006)). Together, these data suggest that lnc-BATE1 functions upstream of key BAT-selective regulators to activate the core gene network associated with brown adipocyte differentiation and function.

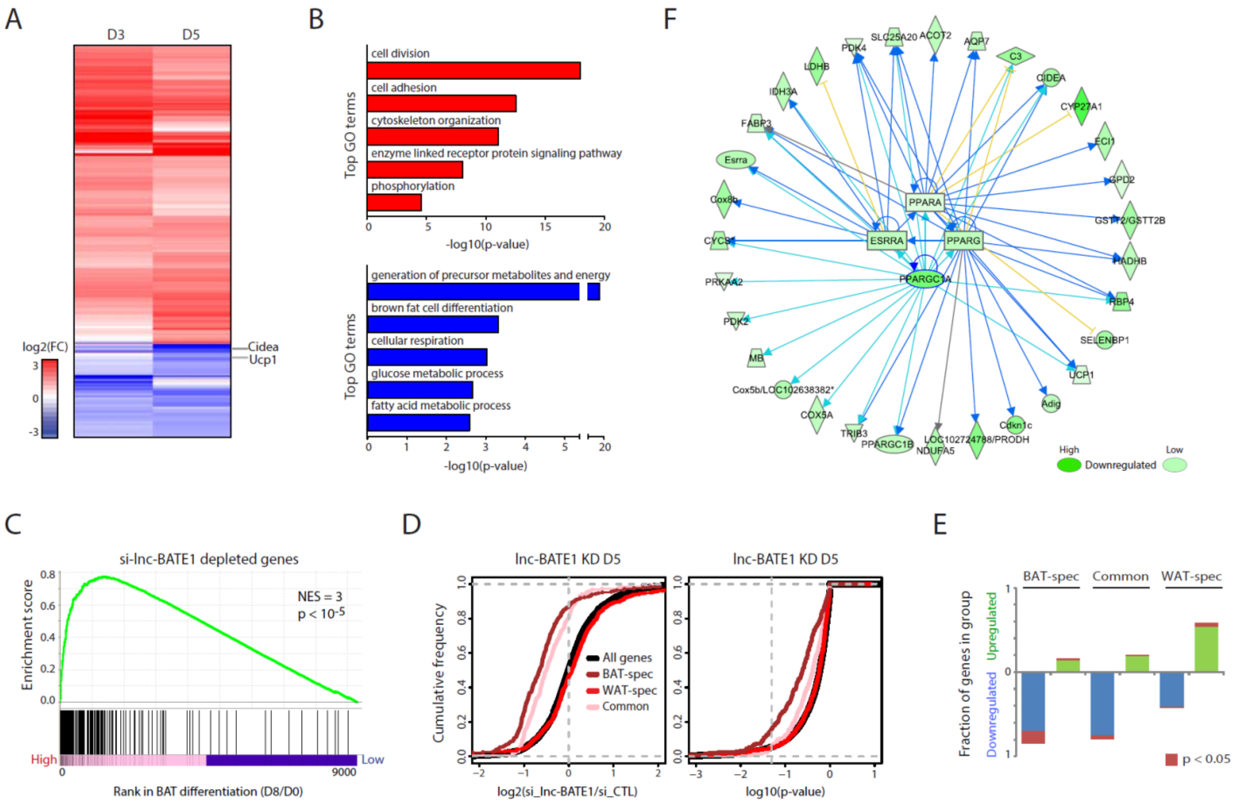


Figure 6. lnc-BATE1 regulates the core gene network of brown adipocyte differentiation.

(**A**) Expression change of *1,014* mRNAs that are differentially expressed *(P <0.05, DESeq)* in cultured brown adipocytes upon lnc-BATE1 KD, collected at differentiation days 3 (D3) and 5 (D5). Changes are log2 expression (FPKM) ratios over control siRNA.

(**B**) Top 5 non-redundant gene ontology (GO) biological process terms enriched (P <0.05, Fisher's test) among mRNA genes that show significantly higher (top) or lower (bottom) expression *(P <0.05, DESeq)* upon lnc-BATE1 KD relative to control.

(**C**) Gene set enrichment analysis for overlap between genes depleted upon lnc-BATE1 KD and the BAT differentiation gene signature published previously (Sun et al. 2013). NES, normalized enrichment score; p, empirical p-value.

(**D**) Cumulative density distributions of expression changes (left) and p-values for these changes (right) for all expressed protein-coding genes and for BAT-specific, WAT-specific and common adipogenic genes in lnc-BATE1 siRNA-treated cultured day 5 brown adipocytes. Changes are log2 expression (FPKM) ratios relative to control siRNA. The 0.05 p-value significance threshold is indicated by a vertical dashed gray line.

(**E**) Proportion of BAT-specific, WAT-specific and common adipogenic genes that are upregulated (log2 expr. change vs. control > 0) or downregulated (log2 expr. change vs. control <0) in lnc-BATE1 siRNA-treated cultured day 5 brown adipocytes.

(**F**) Network diagram depicting the top upstream transcription regulators whose inhibition best explains genes downregulated *(P <0.05, DESeq)* upon lnc-BATE1 KD, along with their known direct targets. Arrows indicate direct transcriptional activation, and blocked lines indicate direct transcriptional repression. Lines colored in blue or yellow indicate that the predicted inhibition of the upstream regulator is consistent or inconsistent with the state of the downstream molecule, respectively, whereas those in gray generated no prediction. Highlighted blue lines correspond to PGC-1α relationships. See Supplemental Experimental Procedures for molecule shapes details.

lnc-BATE1 Functions *in trans*

lncRNAs can function *in cis* or *in trans* via diverse mechanisms during cell differentiation (Wang and Chang 2011; Guttman and Rinn 2012; Hu et al. 2012; Fatica and Bozzoni 2013). To distinguish between these two possibilities, we analyzed the expression of genes neighboring lnc-BATE1 within a ~1.75Mb window (Figure S6A). We found no correlation in the tissue expression patterns of these genes with lncBAT-1 expression (Figure S6B), and showed that their levels are unaffected by lncBAT-1 depletion (Figures S6C and S6D), indicating that lnc-BATE1 does not act *in cis* to regulate its neighbors.

To investigate if lnc-BATE1 functions *in trans*, we tested whether the defects in brown adipocyte differentiation elicited by its depletion could be rescued by ectopically-expressed lnc-BATE1 that escapes DsiRNA targeting. To this end, we constructed an exogenous mutated lnc-BATE1 (lnc-BATE1_Exo) with a 4nt mutation at the DsiRNA2 targeting site designed to abolish knockdown (Figure 7A), and transduced it or a GFP control into brown pre-adipocytes prior to DsiRNA transfection and subsequent induction of differentiation (Figure 7B). RNA was extracted at differentiation day 4, and lnc-BATE1 expression was examined using primers specific to endogenous or exogenous variants or common to both (Figure 7C). Introduction of lnc-BATE1_Exo, which localized to both nucleus and cytoplasm, increased total lnc-BATE1 levels by over 5-fold (Figure 7D). Subsequent addition of DsiRNA2 significantly depleted levels of endogenous lnc-BATE1 but not of lnc-BATE1_Exo, as expected (Figure 7E). Confirming results in Figure 4, lnc-BATE1 knockdown by DsiRNA2 resulted in decreased expression of 8 BAT markers and to a lesser extent of 4 common adipogenic markers in control cells infected with GFP (Figures 6F and 6G top panels). Ectopic expression of Dsi2RNA-resistant lnc-BATE1_Exo, however, rescued the expression of half of the examined BAT markers, including

Dio2, Elovl3, PPARα and UCP1, and of the common adipogenic factors C/EBPα and PPARγ (Figures 7F and 7G bottom panels). These results demonstrate RNA-based function and indicate that lnc-BATE1 can act *in trans* to modulate brown adipocyte development.

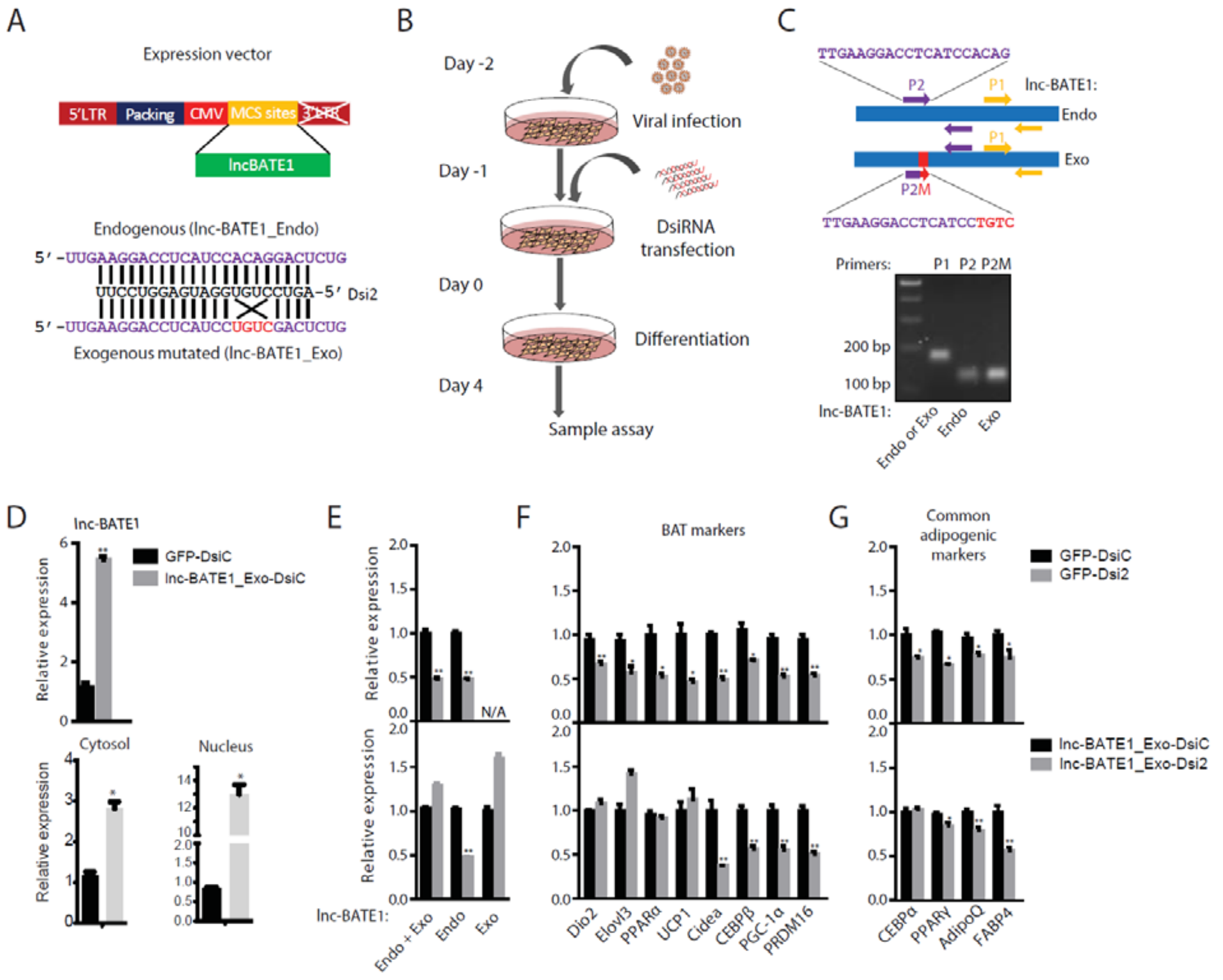


Figure 7. Exogenous siRNA-resistant lnc-BATE1 partially rescues gene suppression in brown adipocytes depleted of endogenous lnc-BATE1.

(**A**) Construction of an exogenous siRNA-resistant lnc-BATE1 mutant (lnc-BATE1_Exo) from the endogenous transcript (lnc-BATE1_Endo).

(**B**) Schematic illustration of procedure used for rescue experiments.

(**C**) Design of qPCR primer pairs and agarose gel image of the resulting PCR products. Lane 2: lnc-BATE1_Endo or _Exo amplified by P1 primer pair; lane 3: lnc-BATE1_Endo amplified by P2 primer pair; lane 4: lnc-BATE1_Exo amplified by P2M primer pair.

(**D**) Expression (top) and localization (bottom) of total lnc-BATE1 in brown adipocytes infected with GFP control viruses or with lnc-BATE1_Exo viruses prior to transfection with control DsiRNA (DsiC).

(**E-G**) Expression of endogenous or exogenous lnc-BATE1 (E), brown adipocyte markers (F) and general adipogenic markers (G) in brown adipocytes infected with GFP control virus or with lnc-BATE1_Exo virus prior to transfection with control DsiRNA (DsiC) or DsiRNA against lnc-BATE1 (Dsi2). Error bars are s.e.m., n =3. **P* ≤0.05, ***P* ≤0.01.

lnc-BATE1 Interacts with hnRNPU

lncRNAs are thought to function by binding proteins to form functional ribonucleoprotein complexes (Rinn and Chang 2012). To gain mechanistic insights into how lnc-BATE1 functions, we sought to identify its protein partners by an in vitro biotin-RNA pull-down assay with nuclear and cytosolic lysates (Experimental Procedures). Co-precipitated proteins were resolved on a SDS-PAGE gel followed by silver staining to identify bands specific to lnc-BATE1 and subject them to mass spectrometry. However, we were unable to observe any differentially stained band relative to control RNA (not shown), suggesting that lnc-BATE1's protein partners are either of low abundance or co-migrate with other proteins of similar molecular weight that mask the signal.

We next sought to examine by RNA immunoprecipitation (RIP) specific proteins known to interact with lncRNAs in adipocytes. Our previous study demonstrated that the nuclear matrix factor hnRNP U is required for the proper localization of *Firre*, a lncRNA essential for white adipocyte differentiation, to four genomic loci encoding adipogenic factors (Sun et al. 2013; Hacisuleyman et al. 2014). Interestingly, we found a putative lnc-BATE1 hnRNP U binding site (Figure S7), inferred from motif analysis of hnRNP U CLIP-seq binding data in human cells (Huelga et al. 2012), suggesting that hnRNP U may be a protein partner of lnc-BATE1.

Before probing whether lnc-BATE1 and hnRNP U interact, we first asked whether hnRNP U contributes to brown adipocyte development. We used siRNAs to inhibit hnRNP U in brown pre-adipocytes and found that its depletion significantly impaired lipid droplet accumulation (Figure 8A) and BAT marker gene expression (Figure 8B) indicating that it is essential for brown adipocyte differentiation. We then performed RIP using an antibody against hnRNP U, and detected a specific interaction with lnc-BATE1 (Figures 8C and 8D). In contrast, RIP against SUZ12, a subunit of the PRC2 complex that binds a wide range of lncRNAs non-specifically (Davidovich et al. 2013; Kaneko et al. 2013; Cifuentes-Rojas et al. 2014), did not display any enrichment for lnc-BATE1. These data were confirmed by in vitro biotin-RNA pull-down using western blots to examine hnRNP U enrichment (Figure 8E). The well-established binding of androgen receptor (AR) 3’UTR RNA to HuR protein (Yeap et al. 2002) served as a positive control for these experiments and, at the same time, as a negative control for hnRNP U binding. As expected, hnRNP U and HuR were enriched by lnc-BATE1 and AR 3’UTR RNA, respectively, whereas the housekeeping Gapdh protein was not (Figure 8E). These findings demonstrate a specific and direct interaction between lnc-BATE1 and hnRNP U, suggesting that they form a functional ribonucleoprotein complex to regulate brown adipocyte development.

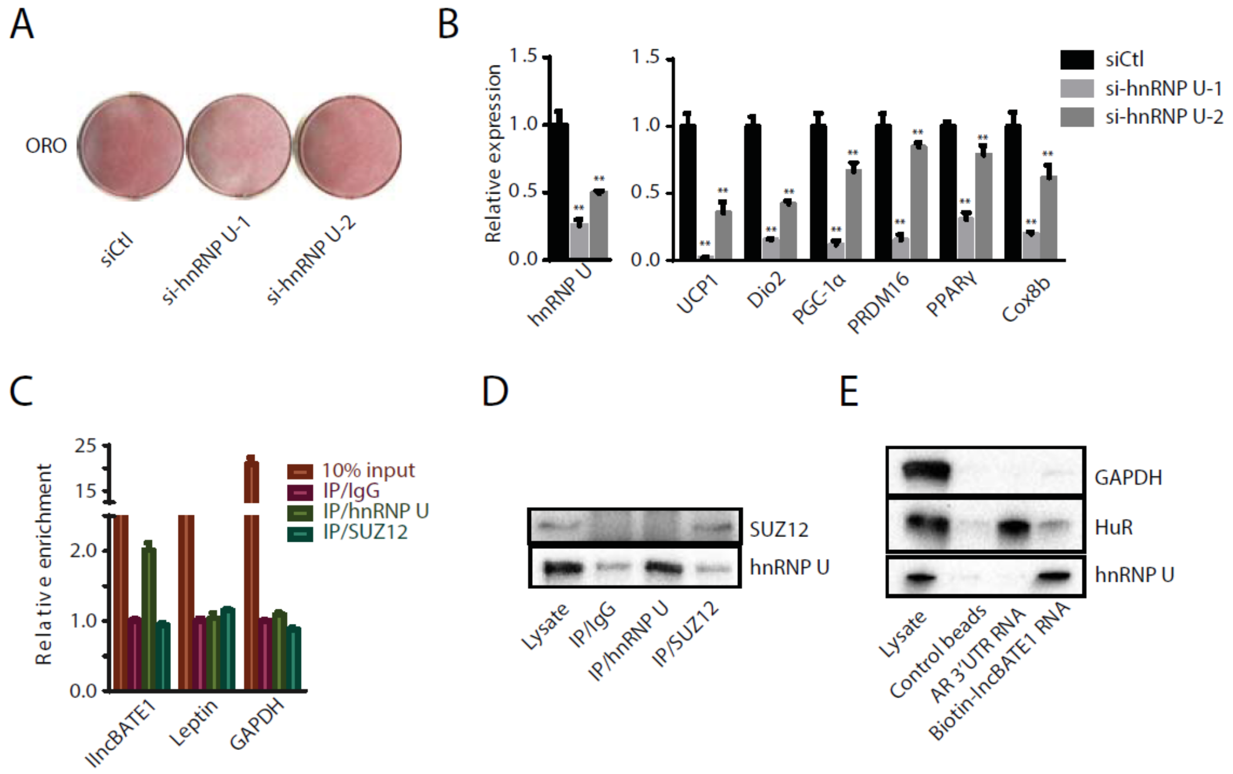


Figure 8. lnc-BATE1 interacts with hnRNP U, which is required for brown adipocyte differentiation.

(**A**) Oil red O staining of brown adipocytes differentiated in culture upon siRNA-mediated hnRNP U knockdown.

(**B**) Expression of hnRNP U and marker genes in cultured brown adipocytes following hnRNP U targeting by siRNAs, quantified by qPCR.

(**C-D**) Association between endogenous lnc-BATE1 and hnRNP U in the nucleus of cultured brown adipocytes. RNA immunoprecipitation (RIP) enrichment was assessed as RNA associated to hnRNP U or Suz12 relative to IgG control by qPCR (C) or Western blot (D).

(**E**) lnc-BATE1 and hnRNP U specifically interact *in vitro*. Western blots for biotin-RNA pull-down showing specific interaction between lnc-BATE1 and hnRNP U but not GAPDH or HuR protein, which specifically interacts with androgen receptor (AR) 3’UTR RNA instead. Error bars are s.e.m., n=3. *$P \leq 0.05$, **$P \leq 0.01$.

Discussion

Elucidating factors governing the development of distinct types of fat is crucial for finding new opportunities to treat metabolic disorders. In particular, regulators that selectively promote brown adipogenesis are of key interest as potential therapeutic targets for obesity. Long non-coding RNAs are rapidly emerging as important tissue-specific developmental modulators (Hu et al. 2012; Fatica and Bozzoni 2013), yet little was known about their specific contributions to BAT development and physiology (Zhao et al. 2014). Here, we present the first comprehensive catalog of lncRNAs active across different adipose tissue types, including ~450 that are highly fat depot-specific, providing a valuable resource for the discovery of lncRNAs with adipocyte lineage-specific functions. This resource is available online (https://sites.google.com/site/sunleilab/data/lncrnas) and can be used to identify functional lncRNAs based on their adipose tissue expression, specificity and regulation features, as illustrated by our work showing that lnc-BATE1, a lncRNA chosen based on these features, is required for the brown adipocyte phenotype. We find that lnc-BATE1 selectively promotes the core brown adipocyte gene program by acting *in trans* upstream of key BAT-specific regulators.

lncRNAs are often activated in a highly cell type- and stage-selective manner (Cabili et al. 2011; Alvarez-Dominguez et al. 2014b). However, our previous study (Sun et al. 2013) identified only a few with strong preference for brown vs. white adipocytes, which we attributed to incompleteness of lncRNA catalogs at the time. Our present catalog now includes hundreds of adipose subtype-specific lncRNAs, which we find are targeted by common adipogenic TFs but in a depot-specific manner, much like key depot-specific protein regulators. The cofactors needed to confer such adipocyte lineage selectivity to broadly expressed TFs remain to be investigated, but in BAT may include known players such as PGC-1α, PRDM16 and EBF2 (Kajimura et al. 2010; Rajakumari et al. 2013).

A different type of thermogenic adipocytes, termed “beige” or “brite”, have been shown to form within white fat depots, in response to cold stress or other stimuli, but share many components of the BAT gene program (Petrovic et al. 2010; Schulz et al. 2011; Wu et al. 2012). We find that lnc-BATE1 is upregulated during cold-induced beige adipocyte expansion and its loss selectively impairs BAT gene expression during white adipogenesis and browning, indicating a broad requirement for thermogenic programming in distinct adipocyte lineages. In contrast, loss of lnc-BATE1 can lead to significant upregulation of WAT-selective genes during white adipogenesis, suggesting that lnc-BATE1 acts not only to sustain a thermogenic phenotype but to suppress WAT-selective programming.

Mechanisms of lncRNA function often involve partnering with proteins such as chromatin modifiers and RNA binding factors (Wang and Chang 2011; Guttman and Rinn 2012). For instance, hnRNP U is responsible for localization of lncRNAs Xist and Firre to the subcellular domains where they exert their functions (Hasegawa et al. 2010; Hacisuleyman et al. 2014). We find that lnc-BATE1 directly interacts with hnRNP U, with a binding specificity well above the threshold of promiscuous RNA binding set by PRC2. Although hnRNP U is ubiquitously expressed across different cell types, we find that it is required for brown adipocyte differentiation, suggesting the possibility that it interacts with lnc-BATE1 to form a functional ribonucleoprotein complex and exert its function in a cell type-specific manner. hnRNP U participates in many aspects of RNA metabolism, however, including splicing, localization, and transport (Han et al. 2010), and so the functional impact of its specific interaction with lnc-

BATE1 on brown adipocyte development and function warrants further investigation. Importantly, lnc-BATE1 is present at many copies per cell in both the nucleus and the cytoplasm, suggesting that it may interact with additional cytosolic protein or RNA partners. Together, our work provides a basis for the study of adipose tissue-selective lncRNAs and demonstrates their importance as BAT-specific regulators, which may be exploited for selective stimulation of BAT development for therapeutic use.

Methods

Tissue isolation and Cell Culture

Primary fat tissues were isolated from 8-week-old B/C mice, and primary brown and white pre-adipocytes were isolated from 3~4-week-old mice and differentiated in culture as described (Sun et al. 2011). 293T cells and C2C12 myoblasts were maintained in DMEM plus 10% or 20% FBS, respectively. C2C12 cells were differentiated in DMEM with 2% horse serum.

RNA-seq and Analysis

Total RNA from BAT, iWAT and eWAT samples was isolated using a Qiagen kit. Sequencing libraries were prepared as described (Sun et al. 2011) and sequenced on the Illumina HiSeq2000 platform. Paired-end reads were mapped to the mouse genome (mm9 version) using TopHat (Trapnell et al. 2009), and *de novo* transcript models were constructed using Cufflinks (Trapnell et al. 2010). Gene expression (FPKM) was quantified by Cufflinks based on *de novo* transcript models, for each fat type, for previously published primary and cultured adipocyte samples, and for 30 cell and tissue types from ENCODE (Stamatoyannopoulos et al. 2012) (see Supplemental Experimental Procedures for further details). RNA-seq data from this study have been deposited at the National Center for Biotechnology Information Gene Expression Omnibus (accession number GSE*).

lncRNA Knockdown by RNAi

Pre-adipocytes at ~80% confluence were transfected with 100nM siRNAs or DsiRNAs. 6~8 hours later, cells were recovered in full culture medium, grown to confluence, and induced to differentiate as described (Sun et al. 2013). For shRNA-mediated knockdown, cells at ~60% confluence were infected with shRNA retroviruses and induced to differentiate 48h post-infection. siRNA knockdown in mature brown adipocytes was performed as described (Rajakumari et al., 2013) (see Supplemental Experimental Procedures for further details).

Plasmid and Retroviral Transduction

lncRNA expression plasmids or shRNA viral plasmids were co-transfected with retroviral packaging vector pCL-Eco into 293T cells using FuGENE6 (Promega), and viruses were collected at 48h and 72h post-transfection, respectively, to infect pre-adipocytes supplemented with 5µg/ml polybrene. Cells were induced to differentiate 48h post-infection and collected for downstream analysis at indicated times.

lncRNA Cloning

To ectopically express lnc-BATE1, 3 different variants were cloned into a modified pSIREN-RetroQ-ZsGreen Vector (Clontech). To make an exogenous mutated lnc-BATE1, mutated nucleotides were introduced into the longest isoform by PCR amplification of overlapping products harboring mutated nucleotides. lnc-BATE1 shRNA plasmids were made by inserting annealed oligos into pMKO vector. Constructs with correct inserts were confirmed by sequencing.

Extracellular Flux Analysis

Primary BAT cells seeded in an X-24 cell culture plate were transfected with DsiRNA targeting lnc-BATE or control DsiRNA and induced to differentiate as described (Sun et al. 2013). 5-day differentiated cells were then applied to an Extracellular Flux Analyzer (Seahorse bioscience) for analysis of oxygen consumption rate according to the manufacturer’s instructions.

RNA Immunoprecipitation

4-day differentiated brown adipocytes were trypsinized, washed, resuspended in hypotonic buffer, kept on ice for 15min, and their nuclei released using a glass dounce homogenizer and pelleted. The supernatant was collected as the cytosolic fraction, and the pellet was resuspended in 2ml lysis buffer (25mM hepes, 150mM KCl, 5mM $MgCl_2$, 1mM DTT, protease inhibitor cocktail) and sheared by dounce homogenizer. Pelleted debris was discarded, and nuclear lysate was supplemented with RNase inhibitor (300U/ml final conc.). 30ul Protein A/G beads (SC2003, Santa Cruz) were incubated with 5ug IgG or indicated antibody in 200ul lysis buffer for 30min, and antibody-bound beads were then washed twice in lysis buffer, followed by incubation with 500ul nuclear lysate for 3h at 4°C. After 4 washes with lysis buffer supplemented with 0.5% NP-40 and 40U/ml RNase inhibitor, 20% of beads were kept for western blot and the rest used for RNA extraction (see Supplemental Experimental Procedures for further details).

RNA Pull-down

Biotin-labeled lnc-BATE1 and androgen receptor 3'UTR RNA were *in vitro* using a MEGAscript kit (Life Technologies) Biotinylated RNAs were purified with a NucAway spin column as described (Tsai et al. 2010). Magnetic Dynabeads M-280 streptavidin beads (Life Technologies) were pre-treated with 0.1 M NaOH and washed with 0.1M NaCl, and 50ul beads were then incubated with 30pmol biotin-labeled lnc-BATE1 or control RNA in binding buffer (1M NaCl, 5mM Tris) for 30min at room temperature. Biotinylated RNA-bound beads were then washed with binding buffer and incubated with brown adipocyte nuclear lysate for 3h at 4°C. Beads were then washed 4 times with wash buffer (25mM hepes, 75mM KCl, 5mM $MgCl_2$, 1mM DTT, protease inhibitor cocktail, 40U/ml RNase inhibitor), and RNA-bound proteins were released by boiling beads in sample buffer for 5min at 95°C. Protein enrichment was examined by western blot using specific antibodies (see Supplemental Experimental Procedures for further details).

References

Alvarez-Dominguez JR, Hu W, Gromatzky AA, Lodish HF. 2014a. Long noncoding RNAs during normal and malignant hematopoiesis. *International journal of hematology* **99**: 531-541.

Alvarez-Dominguez JR, Hu W, Yuan B, Shi J, Park SS, Gromatzky AA, van Oudenaarden A, Lodish HF. 2014b. Global discovery of erythroid long noncoding RNAs reveals novel regulators of red cell maturation. *Blood* **123**: 570-581.

Bostrom P, Wu J, Jedrychowski MP, Korde A, Ye L, Lo JC, Rasbach KA, Bostrom EA, Choi JH, Long JZ et al. 2012. A PGC1-alpha-dependent myokine that drives brown-fat-like development of white fat and thermogenesis. *Nature* **481**: 463-468.

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**: 1915-1927.

Chen Y, Siegel F, Kipschull S, Haas B, Frohlich H, Meister G, Pfeifer A. 2013. miR-155 regulates differentiation of brown and beige adipocytes via a bistable circuit. *Nat Commun* **4**: 1769.

Chiang SH, Bazuine M, Lumeng CN, Geletka LM, Mowers J, White NM, Ma JT, Zhou J, Qi N, Westcott D et al. 2009. The protein kinase IKKepsilon regulates energy balance in obese mice. *Cell* **138**: 961-975.

Cifuentes-Rojas C, Hernandez AJ, Sarma K, Lee JT. 2014. Regulatory interactions between RNA and polycomb repressive complex 2. *Molecular cell* **55**: 171-185.

Cypess AM, Lehman S, Williams G, Tal I, Rodman D, Goldfine AB, Kuo FC, Palmer EL, Tseng YH, Doria A et al. 2009. Identification and importance of brown adipose tissue in adult humans. *N Engl J Med* **360**: 1509-1517.

Davidovich C, Zheng L, Goodrich KJ, Cech TR. 2013. Promiscuous RNA binding by Polycomb repressive complex 2. *Nat Struct Mol Biol* **20**: 1250-1257.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res* **22**: 1775-1789.

Fatica A, Bozzoni I. 2013. Long non-coding RNAs: new players in cell differentiation and development. *Nat Rev Genet* **15**: 7-21.

Feldmann HM, Golozoubova V, Cannon B, Nedergaard J. 2009. UCP1 ablation induces obesity and abolishes diet-induced thermogenesis in mice exempt from thermal stress by living at thermoneutrality. *Cell Metab* **9**: 203-209.

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S et al. 2014. Ensembl 2014. *Nucleic Acids Res* **42**: D749-755.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**: 223-227.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C et al. 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**: 503-510.

Guttman M, Rinn JL. 2012. Modular regulatory principles of large non-coding RNAs. *Nature* **482**: 339-346.

Hacisuleyman E, Goff LA, Trapnell C, Williams A, Henao-Mejia J, Sun L, McClanahan P, Hendrickson DG, Sauvageau M, Kelley DR et al. 2014. Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nat Struct Mol Biol* **21**: 198-206.

Hamann A, Flier JS, Lowell BB. 1996. Decreased brown fat markedly enhances susceptibility to diet-induced obesity, diabetes, and hyperlipidemia. *Endocrinology* **137**: 21-29.

Han SP, Tang YH, Smith R. 2010. Functional diversity of the hnRNPs: past, present and perspectives. *The Biochemical journal* **430**: 379-392.

Harms MJ, Ishibashi J, Wang W, Lim HW, Goyama S, Sato T, Kurokawa M, Won KJ, Seale P. 2014. Prdm16 is required for the maintenance of brown adipocyte identity and function in adult mice. *Cell Metab* **19**: 593-604.

Hasegawa Y, Brockdorff N, Kawano S, Tsutui K, Tsutui K, Nakagawa S. 2010. The matrix protein hnRNP U is required for chromosomal localization of Xist RNA. *Dev Cell* **19**: 469-476.

Hu W, Alvarez-Dominguez JR, Lodish HF. 2012. Regulation of mammalian cell differentiation by long non-coding RNAs. *EMBO Rep* **13**: 971-983.

Huelga SC, Vu AQ, Arnold JD, Liang TY, Liu PP, Yan BY, Donohue JP, Shiue L, Hoon S, Brenner S et al. 2012. Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell Rep* **1**: 167-178.

Kajimura S, Seale P, Kubota K, Lunsford E, Frangioni JV, Gygi SP, Spiegelman BM. 2009. Initiation of myoblast to brown fat switch by a PRDM16-C/EBP-beta transcriptional complex. *Nature* **460**: 1154-1158.

Kajimura S, Seale P, Spiegelman BM. 2010. Transcriptional control of brown fat development. *Cell Metab* **11**: 257-262.

Kaneko S, Son J, Shen SS, Reinberg D, Bonasio R. 2013. PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. *Nat Struct Mol Biol* **20**: 1258-1264.

Lee JE, Wang C, Xu S, Cho YW, Wang L, Feng X, Baldridge A, Sartorelli V, Zhuang L, Peng W et al. 2013. H3K4 mono- and di-methyltransferase MLL4 is required for enhancer activation during cell differentiation. *Elife* **2**: e01503.

Lin MF, Jungreis I, Kellis M. 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**: i275-282.

Lowell BB, V SS, Hamann A, Lawitts JA, Himms-Hagen J, Boyer BB, Kozak LP, Flier JS. 1993. Development of obesity in transgenic mice after genetic ablation of brown adipose tissue. *Nature* **366**: 740-742.

Mori M, Nakagami H, Rodriguez-Araujo G, Nimura K, Kaneda Y. 2010. Essential role for miR-196a in brown adipogenesis of white fat progenitor cells. *PLoS Biol* **10**: e1001314.

Natoli G, Andrau JC. 2012. Noncoding Transcription at Enhancers: General Principles and Functional Models. *Annual Review of Genetics, Vol 46* **46**: 1-19.

Nedergaard J, Bengtsson T, Cannon B. 2007. Unexpected evidence for active brown adipose tissue in adult humans. *American journal of physiology Endocrinology and metabolism* **293**: E444-452.

Petrovic N, Walden TB, Shabalina IG, Timmons JA, Cannon B, Nedergaard J. 2010. Chronic peroxisome proliferator-activated receptor gamma (PPARgamma) activation of epididymally derived white adipocyte cultures reveals a population of thermogenically competent, UCP1-containing adipocytes molecularly distinct from classic brown adipocytes. *J Biol Chem* **285**: 7153-7164.

Rajakumari S, Wu J, Ishibashi J, Lim HW, Giang AH, Won KJ, Reed RR, Seale P. 2013. EBF2 determines and maintains brown adipocyte identity. *Cell Metab* **17**: 562-574.

Rinn JL, Chang HY. 2012. Genome regulation by long noncoding RNAs. *Annu Rev Biochem* **81**: 145-166.

Rothwell NJ, Stock MJ. 1979. A role for brown adipose tissue in diet-induced thermogenesis. *Nature* **281**: 31-35.

Saito M, Okamatsu-Ogura Y, Matsushita M, Watanabe K, Yoneshiro T, Nio-Kobayashi J, Iwanaga T, Miyagawa M, Kameya T, Nakada K et al. 2009. High incidence of metabolically active brown adipose tissue in healthy adult humans: effects of cold exposure and adiposity. *Diabetes* **58**: 1526-1531.

Schulz TJ, Huang TL, Tran TT, Zhang H, Townsend KL, Shadrach JL, Cerletti M, McDougall LE, Giorgadze N, Tchkonia T et al. 2011. Identification of inducible brown adipocyte progenitors residing in skeletal muscle and white fat. *Proc Natl Acad Sci U S A* **108**: 143-148.

Seale P, Conroe HM, Estall J, Kajimura S, Frontini A, Ishibashi J, Cohen P, Cinti S, Spiegelman BM. 2011. Prdm16 determines the thermogenic program of subcutaneous white adipose tissue in mice. *J Clin Invest* **121**: 96-105.

Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, Groudine M, Bender M, Kaul R, Canfield T et al. 2012. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biology* **13**.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**: 15545-15550.

Sun L, Goff LA, Trapnell C, Alexander R, Lo KA, Hacisuleyman E, Sauvageau M, Tazon-Vega B, Kelley DR, Hendrickson DG et al. 2013. Long noncoding RNAs regulate adipogenesis. *Proc Natl Acad Sci U S A* **110**: 3387-3392.

Sun L, Trajkovski M. 2014. MiR-27 orchestrates the transcriptional regulation of brown adipogenesis. *Metabolism: clinical and experimental* **63**: 272-282.

Sun L, Xie H, Mori MA, Alexander R, Yuan B, Hattangadi SM, Liu Q, Kahn CR, Lodish HF. 2011. Mir193b-365 is essential for brown fat differentiation. *Nat Cell Biol* **13**: 958-965.

Timmons JA, Wennmalm K, Larsson O, Walden TB, Lassmann T, Petrovic N, Hamilton DL, Gimeno RE, Wahlestedt C, Baar K et al. 2007. Myogenic gene expression signature establishes that brown and white adipocytes originate from distinct cell lineages. *Proc Natl Acad Sci U S A* **104**: 4401-4406.

Trajkovski M, Ahmed K, Esau CC, Stoffel M. 2012. MyomiR-133 regulates brown fat differentiation through Prdm16. *Nat Cell Biol* **14**: 1330-1335.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511-515.

Troy A, Sharpless NE. 2012. Genetic "lnc"-age of noncoding RNAs to human disease. *J Clin Invest* **122**: 3837-3840.

Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E, Chang HY. 2010. Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**: 689-693.

Uldry M, Yang W, St-Pierre J, Lin J, Seale P, Spiegelman BM. 2006. Complementary action of the PGC-1 coactivators in mitochondrial biogenesis and brown fat differentiation. *Cell Metab* **3**: 333-341.

van Marken Lichtenbelt WD, Vanhommerig JW, Smulders NM, Drossaerts JM, Kemerink GJ, Bouvy ND, Schrauwen P, Teule GJ. 2009. Cold-activated brown adipose tissue in healthy men. *N Engl J Med* **360**: 1500-1508.

Villarroya F, Vidal-Puig A. 2013. Beyond the sympathetic tone: the new brown fat activators. *Cell Metab* **17**: 638-643.

Virtanen KA, Lidell ME, Orava J, Heglind M, Westergren R, Niemi T, Taittonen M, Laine J, Savisto NJ, Enerback S et al. 2009. Functional brown adipose tissue in healthy adults. *N Engl J Med* **360**: 1518-1525.

Wang KC, Chang HY. 2011. Molecular mechanisms of long noncoding RNAs. *Mol Cell* **43**: 904-914.

Wu J, Bostrom P, Sparks LM, Ye L, Choi JH, Giang AH, Khandekar M, Virtanen KA, Nuutila P, Schaart G et al. 2012. Beige adipocytes are a distinct type of thermogenic fat cell in mouse and human. *Cell* **150**: 366-376.

Yeap BB, Voon DC, Vivian JP, McCulloch RK, Thomson AM, Giles KM, Czyzyk-Krzeska MF, Furneaux H, Wilce MC, Wilce JA et al. 2002. Novel binding of HuR and poly(C)-binding protein to a conserved UC-rich motif within the 3'-untranslated region of the androgen receptor messenger RNA. *J Biol Chem* **277**: 27183-27192.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.

Zhao XY, Li S, Wang GX, Yu Q, Lin JD. 2014. A long noncoding RNA transcriptional regulatory circuit drives thermogenic adipocyte differentiation. *Mol Cell* **55**: 372-382.

Outlook and future directions

Parts of this chapter were first published as:

Hu W, **Alvarez-Dominguez JR**, Lodish HF. 2012. Regulation of mammalian cell differentiation by long non-coding RNAs. *EMBO Rep* **13**: 971-983.

Alvarez-Dominguez JR, Hu W, Lodish HF. 2013. Regulation of Eukaryotic Cell Differentiation by Long Non-coding RNAs. in *Molecular Biology of Long Non-coding RNAs* (eds. AM Khalil, J Coller), pp. 15-67. Springer Science, New York.

Alvarez-Dominguez JR, Hu W, Gromatzky AA, Lodish HF. 2014a. Long noncoding RNAs during normal and malignant hematopoiesis. *International journal of hematology* **99**: 531-541.

As observed in multiple differentiation systems, lineage-specific development involves concurrent activation of one fate and repression of others. Our work suggests that regulatory lncRNAs are key tools deployed by cells to exert this programming. Overall, our findings indicate that distinct collections of lncRNAs become active in distinct lineages and partner with ubiquitous regulatory protein complexes to activate or repress lineage-specific developmental programs. This notion has also emerged from a number of parallel studies described in previous sections, most of which began to accumulate as ours developed.

The choice of RNP-based regulation vs. solely protein-based modes or regulation may be related to the versatility with which RNA molecules can interact with DNA, with other RNAs, and with protein complexes (Dethoff et al. 2012; Geisler and Coller 2013). Such versatility is enabled by specific features of primary sequence, secondary structure, and genomic positioning that experience different types of selective pressure during evolution (Marques and Ponting 2009; Ulitsky et al. 2011; Derrien et al. 2012). Deployment of lineage-specific complexes made of only proteins would require each of these proteins to evolve a separate surface binding domain per each distinct binding partner in each distinct tissue. lncRNAs, in contrast, can rapidly evolve and shuffle combinations of protein-binding domains amid otherwise unconstrained sequence, such that their tissue-specific expression can bring together different complexes in different lineages to give rise to lineage-specific gene expression programs. Accordingly, exploiting lncRNAs as flexible scaffolds for cell type-specific RNP regulatory complexes may have precipitated a rapid expansion in the repertoire of developmental programs during metazoan evolution (Prasanth and Spector 2007; Amaral and Mattick 2008).

While there is now tantalizing evidence for this model, it is far from proven. Key questions remain about how do lncRNAs achieve selective binding of protein, DNA or RNA partners in vivo, how can they bring these partners to target locations in trans, and what features determine whether they act solely in cis, in trans, or both? And ultimately, how can RNAs present at low steady-state copy numbers productively capture enough protein partners to regulate entire gene networks? Over the next sections, we present our own perspectives on some of these important questions.

Molecular Mechanisms of lncRNA Regulators of Cell Differentiation

lncRNAs can modulate gene expression via diverse mechanisms (Wang and Chang 2011; Guttman and Rinn 2012; Moran et al. 2012). They frequently function in partnership with regulatory proteins such as chromatin modifiers, transcription factors and RNA decay factors (Wang and Chang 2011; Guttman and Rinn 2012; Ulitsky and Bartel 2013). Of those lncRNAs currently implicated in cell differentiation processes, many seem to direct gene expression through recruitment of chromatin modifiers. This is consistent with multiple observations that chromatin modifiers, such as PRC2, can associate with a diversity of noncoding transcripts (Khalil et al. 2009; Zhao et al. 2010; Guttman et al. 2011; Derrien et al. 2012). Interestingly, one major, though not exclusive, function of lncRNA during development is to promote, in a cell-type specific manner, assembly of select combinations of ubiquitously-expressed chromatin modifiers in target genomic regions, thereby exerting epigenetic control with exquisite spatial and temporal precision. However, key questions remain about how specific binding to chromatin modifier partners is achieved *in vivo*, what sequence properties enable lncRNAs to target these partners to specific areas in the genome, and what role does local chromatin conformation play in modulating these interactions.

An expanding toolbox of molecular approaches is rapidly becoming available to address these and other questions about lncRNA molecular mechanisms. As illustrated in Chapters 2 and 3, investigating these typically begins by first asking whether a lncRNA can act in *cis* or in *trans*. Distinguishing between these modes of action *a priori* is important as they require different experimental strategies to elucidate mechanism. lncRNAs that act *in trans* are amenable to ectopic expression studies, providing an opportunity for loss-of-function rescue via expression of a transgene from an ectopic site or via its direct introduction as exogenous RNA. *cis*-acting lncRNAs, in contrast, call for perturbation strategies that preserve their relative proximity to their targets and sometimes their relative orientation, chromatin state and chromatin conformation.

To determine *cis* vs. *trans* regulation, a good first step is to assess subcellular localization: *cis*-acting lncRNAs will be predominantly nuclear, whereas *trans*-acting can be nuclear, cytoplasmic, or equally distributed across both compartments. Cellular fractionation followed by RNA detection can be a cost-effective method to broadly distinguish between these possibilities. In addition, direct lncRNA visualization by RNA FISH can provide additional information: *cis*-acting lncRNAs like Xist, Firre or EC6 display focal nuclear localization around their site of transcription, whereas *trans*-acting lncRNAs display diffuse localization throughout the nucleus or cytoplasm or both. Moreover, RNA FISH can provide high-resolution mapping of lncRNA localization to even smaller subcellular structures informative of function, such as the nucleolus, paraspeckles, or other granule RNA structures (Yamashita et al. 1998; Kloc et al. 2005; Nagano et al. 2008; Clemson et al. 2009; Sasaki et al. 2009; Sunwoo et al. 2009). In combination with other methods, such as DNA-FISH, immunofluorescence, or fluorescent protein tagging, RNA-FISH can also be used to detect lncRNAs in specific chromosomes or in regions of silent or active chromatin (Redrup et al. 2009; Reinius et al. 2010; Sexton et al. 2012), and can also be used to examine multimerization potential and co-localization with specific RNA or protein partners (Khalil et al. 2009; Chakraborty et al. 2012).

Importantly, several powerful assays have been recently developed to determine in an unbiased, high-throughput manner the genomic binding sites of lncRNAs (Chu et al. 2011; Simon et al. 2011; Engreitz et al. 2013), their RNA targets (Kretz et al. 2013; Engreitz et al. 2014), and their associated proteins (Kretz et al. 2013; West et al. 2014). These and other assays will greatly facilitate the exploration of lncRNA mechanisms within cell differentiation systems. Judging by the constant development and broad application of these assays, we predict that such exploration will greatly advance in the coming years.

Integrating lncRNAs to Known Regulatory Networks of Cell Differentiation

Differentiation programs are exquisitely controlled at every stage by complex networks that respond to varying developmental and environmental signals. The examples discussed in this thesis argue that lncRNAs are likely to be integrated as key components of these regulatory networks, on par with transcription factors, chromatin modifiers and microRNAs. Precisely how lncRNAs should be integrated can be answered by first exploring their regulatory relationships with other components (Figure 1).

As exemplified in Chapters 1-3, expression of lineage-specific lncRNAs modulating lineage-specific developmental programs is indeed controlled by the key transcription factors directing those programs. Interestingly, some lncRNAs are reported to physically bind transcription factors (Willingham et al. 2005; Kino et al. 2010; Ng et al. 2012; Wang et al. 2014), suggesting that mutual modulation between lncRNAs and transcription factors is possible.

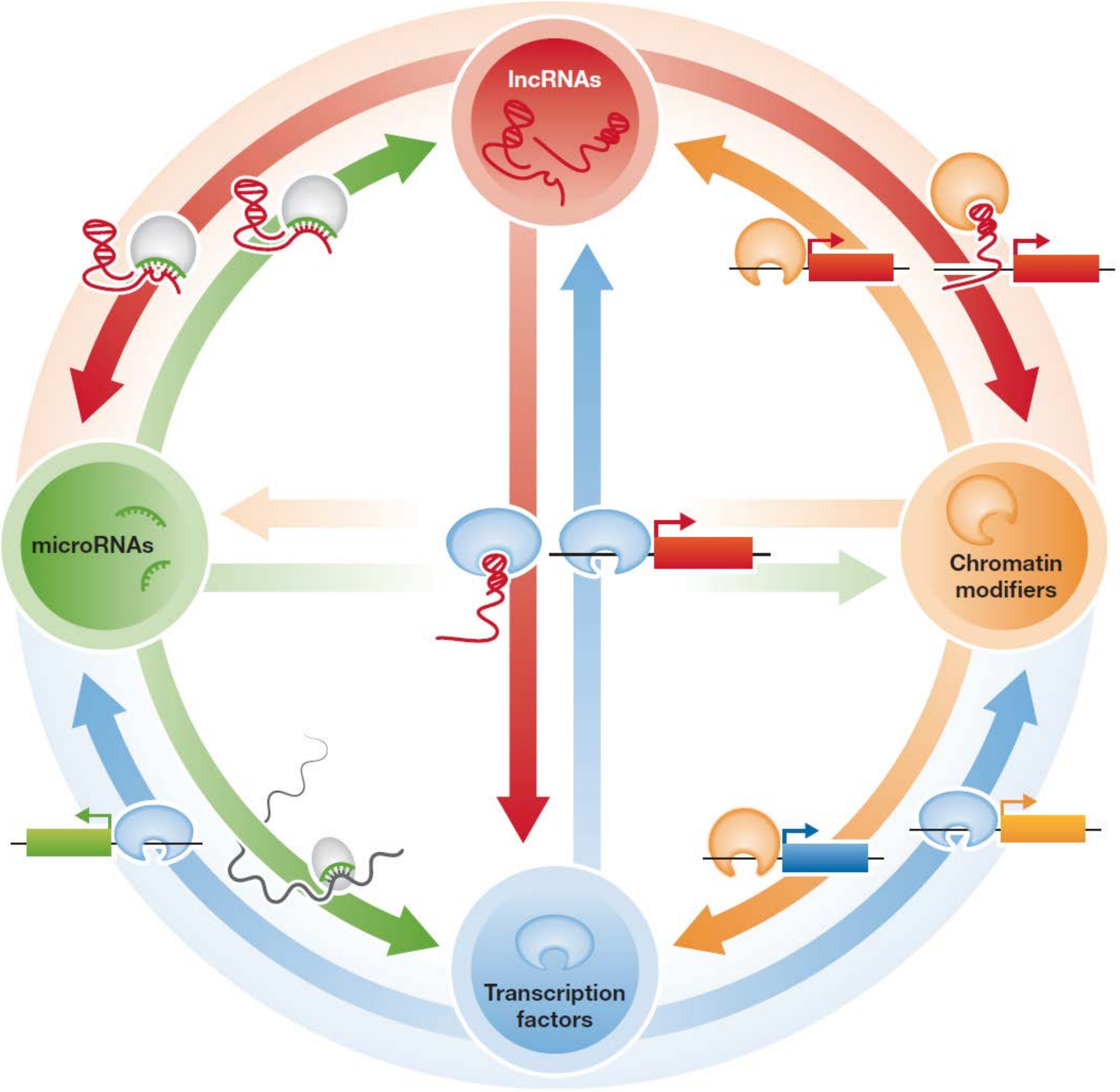


Figure 1. Integrating lncRNAs into known regulatory networks of cell differentiation.

Integrating lncRNA functions with those of microRNAs, TFs and chromatin modifiers during cell differentiation will require exploring their mutual regulatory relationships. Examples of some of these relationships are depicted. lncRNAs may regulate microRNAs or TFs as target site decoys, and they may also associate with chromatin modifiers as structural components within RNP complexes or as guides and tethers to their chromatin targets. microRNAs post-transcriptionally regulate transcripts from TF, chromatin modifier or lncRNA loci by base-pairing to short stretches within their sequences. TFs control transcription of all the other regulators by directly binding to their promoters. Similarly, chromatin modifiers enforce epigenetic states influencing expression from all the other network components. Not depicted are regulatory relationships between microRNAs and chromatin modifier components.

Further progress in identifying the global binding sites of key transcription factors during cell differentiation, as well as the protein interactome of lncRNAs, will be of great help in reconstructing regulatory networks involving lncRNAs. Simply intersecting such datasets with transcriptome profiling along developmental processes will be of great use in identifying lncRNAs likely to function in those processes.

Our present understanding of the relationship between lncRNAs and chromatin modifiers is governed by the constant observation of functionally productive physical associations between these factors. In fact, the prevalence of such functional partnerships throughout eukaryotes, as evidenced by the many examples presented in the Introduction, has changed our understanding of how chromatin modifiers themselves operate. This is best illustrated in the case of Polycomb group proteins, which are now believed to recognize their target loci not through interactions with DNA but through interactions with RNA tethered to the DNA (Schmitt and Paro 2006; Hekimoglu and Ringrose 2009; Zhao et al. 2010). A growing body of evidence now suggests that this model might extend to several other classes of epigenetic modifiers (Koziol and Rinn 2010; Tsai et al. 2010; Spitale et al. 2011; Guttman and Rinn 2012). Thus, lncRNAs may be integrated into regulatory networks involving chromatin modifiers by serving as structural components, guides, and/or physical tethers. However, care should be placed in assuming such functions. Physical association by itself does not prove function and can be rather promiscuous (Zhao et al. 2010; Davidovich et al. 2013; Kaneko et al. 2013; Cifuentes-Rojas et al. 2014), such that detailed studies including structure-function mapping are required for demonstrating functional relevance of lncRNA-chromatin modifier associations.

Several studies have also proposed that certain lncRNAs and microRNAs can regulate each other at the post-transcriptional level (Franco-Zorrilla et al. 2007; Cesana et al. 2011;

Karreth et al. 2011; Salmena et al. 2011; Ulitsky et al. 2011). However, it remains unclear how individual lncRNAs which only contribute a small fraction of the total target site abundance for a given miRNA can possibly influence enough miRNA molecules to affect overall miRNA networks. In addition, global identification of lncRNA targets of microRNAs remains in its earliest stages (Jeggari et al. 2012). Identifying microRNAs and lncRNAs with complementary expression patterns during cell differentiation may thus generate candidate lncRNA-microRNA regulatory pairs to be tested in detail for integration into regulatory networks. Such studies may not only serve to define such networks, but also to expand our understanding on how they contribute to development.

In comparing the role of lncRNAs with those of other factors involved in cell differentiation processes, it is important to note that, as with microRNAs, the biological effects of many lncRNAs can be rather mild, with mild changes in the expression of target loci upon lncRNA perturbation seen frequently. This may be in part due to limitations in achieving efficient knockdown of lncRNAs by RNAi approaches, and may be solved by their direct genetic perturbation, which is now greatly facilitated by emerging applications of the CRISPR-Cas9 system for targeted gene disruption, repression and activation (see (Hsu et al. 2014) for review and (Gilbert et al. 2014; Konermann et al. 2014) for recent developments). Alternatively, it may be that lncRNAs primarily act to tune target gene expression, much like microRNAs. Genetic models of *in vivo* lncRNA function may thus be required to discriminate between these two possibilities, as discussed in the next section.

Compared to transcription factors, chromatin modifiers, and microRNA regulators, lncRNAs seem to employ a wider diversity of molecular mechanisms to modulate their targets, including functions at the level of transcription, translation and stability (Figure 1). Therefore, it may not be surprising that during cell differentiation lncRNAs may cooperate with, or sequester away, any of the other regulatory components to ensure tuning of genetic circuits at both the transcriptional and post-transcriptional levels (Redrup et al. 2009; Keniry et al. 2012).

In vivo functions of lncRNAs

Although perturbation of many lncRNAs results in phenotypic changes during differentiation of *in vitro* cultured cells, our knowledge of the *in vivo* functions of lncRNAs remains limited. Several lncRNA-altered animals have been generated to bridge this gap in knowledge. Pioneering studies in non-mammalian vertebrate models have established essential developmental roles for conserved lncRNAs. For example, knockdown of lincRNAs Cyrano and Megamind severely impact CNS development in zebrafish, and such deficiencies can be rescued with their mouse and human orthologs (Ulitsky et al. 2011). Similarly, knockdown of HOTTIP in chicken embryos results in shortening and bending of distal bones (Wang et al. 2011).

In vivo developmental phenotypes from mouse knockout models, however, have only recently been forthcoming (Bond et al. 2009; Klein et al. 2010; Anguera et al. 2011b; Grote et al. 2013; Li et al. 2013; Sauvageau et al. 2013; Yildirim et al. 2013). These include lncRNAs like Fendrr whose early developmental roles renders them required for life, as well as lncRNAs whose roles during later, lineage-specific developmental processes render them critical for proper tissue physiology in mature animals. For example, mice deleted for Evf2 are delayed in forming GABAergic interneurons during early hippocampus development and thus exhibit compromised synaptic inhibition capacity during adulthood (Bond et al. 2009). Similarly, male mice deleted for the X-linked Tsx, which show reduced fertility due to elevated apoptosis during spermatogenesis, also display enhanced hippocampal short-term memory (Anguera et al. 2011a). Milder phenotypes have also been described for H19 and Air, which regulate embryonic and early post-natal growth. Deleting H19, which mediates maternal imprinting of the growth regulator Igf2, results in embryonic weight increases of 10-20% (Leighton et al. 1995; Ripoche et al. 1997; Wutz et al. 2001). Similarly, deleting Air, which is required for paternal imprinting of the Igf2 receptor Igf2r, changes embryonic weight by about 20% (Wutz et al. 2001).

It is important to consider several potential caveats of the study of lncRNA function *in vivo,* however. For example, investigation of *in vivo* lncRNA models under informative physiological conditions may be crucial to elucidating context-dependent phenotypes, as seen for MALAT1 (Gutschner et al. 2013). Moreover, the observation of *in vivo* phenotypes can heavily depend on the strategy used for deleting an lncRNA, as illustrated by the case of the lncRNA HOTAIR (Rinn et al. 2007; Li et al. 2013). Strategies that perturb adjacent or overlapping genes in addition to the lncRNA can give rise to confounding or even compensatory effects that mask phenotype. Similarly, loss of regulatory DNA sites along with lncRNA deletion makes it difficult to dissect the individual contributions of the lncRNA vs. those of its underlying DNA sequence (notably, the same applies to investigating protein-coding gene function). As with protein-coding genes, these considerations call for use of orthogonal approaches to investigate lncRNA function, such as systematic deletion mapping studies or use of rescue experiments that demonstrate RNA-based function, such as those presented in Chapter 3. Recently developed techniques for efficient targeted genetic perturbation, such as the CRISPR-Cas9 system, can be of great help in this respect, and are currently being employed to study the physiological functions of the lncRNAs described in Chapters 2 and 3. Such *in vivo* studies will be necessary to properly establish the contributions of EC6 and lnc-BATE1 to the development and physiology of the erythroid and brown adipose tissue systems.

References

Anguera MC, Ma W, Clift D, Namekawa S, Kelleher RJ, 3rd, Lee JT. 2011a. Tsx produces a long noncoding RNA and has general functions in the germline, stem cells, and brain. *PLoS Genet* **7**: e1002248.

Anguera MC, Ma WY, Clift D, Namekawa S, Kelleher RJ, Lee JT. 2011b. Tsx Produces a Long Noncoding RNA and Has General Functions in the Germline, Stem Cells, and Brain. *Plos Genetics* **7**.

Bond AM, Vangompel MJ, Sametsky EA, Clark MF, Savage JC, Disterhoft JF, Kohtz JD. 2009. Balanced gene regulation by an embryonic brain ncRNA is critical for adult hippocampal GABA circuitry. *Nat Neurosci* **12**: 1020-1027.

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

Chakraborty D, Kappei D, Theis M, Nitzsche A, Ding L, Paszkowski-Rogacz M, Surendranath V, Berger N, Schulz H, Saar K et al. 2012. Combined RNAi and localization for functionally dissecting long noncoding RNAs. *Nat Methods* **9**: 360-362.

Chu C, Qu K, Zhong FL, Artandi SE, Chang HY. 2011. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol Cell* **44**: 667-678.

Cifuentes-Rojas C, Hernandez AJ, Sarma K, Lee JT. 2014. Regulatory interactions between RNA and polycomb repressive complex 2. *Molecular cell* **55**: 171-185.

Clemson CM, Hutchinson JN, Sara SA, Ensminger AW, Fox AH, Chess A, Lawrence JB. 2009. An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Mol Cell* **33**: 717-726.

Davidovich C, Zheng L, Goodrich KJ, Cech TR. 2013. Promiscuous RNA binding by Polycomb repressive complex 2. *Nat Struct Mol Biol* **20**: 1250-1257.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res* **22**: 1775-1789.

Dethoff EA, Chugh J, Mustoe AM, Al-Hashimi HM. 2012. Functional complexity and regulation through RNA dynamics. *Nature* **482**: 322-330.

Engreitz JM, Pandya-Jones A, McDonel P, Shishkin A, Sirokman K, Surka C, Kadri S, Xing J, Goren A, Lander ES et al. 2013. The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome. *Science* **341**: 767-U233.

Engreitz JM, Sirokman K, McDonel P, Shishkin AA, Surka C, Russell P, Grossman SR, Chow AY, Guttman M, Lander ES. 2014. RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent Pre-mRNAs and chromatin sites. *Cell* **159**: 188-199.

Franco-Zorrilla JM, Valli A, Todesco M, Mateos I, Puga MI, Rubio-Somoza I, Leyva A, Weigel D, Garcia JA, Paz-Ares J. 2007. Target mimicry provides a new mechanism for regulation of microRNA activity. *Nat Genet* **39**: 1033-1037.

Geisler S, Coller J. 2013. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat Rev Mol Cell Biol*.

Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, Guimaraes C, Panning B, Ploegh HL, Bassik MC et al. 2014. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159**: 647-661.

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

Gutschner T, Hammerle M, Eissmann M, Hsu J, Kim Y, Hung G, Revenko A, Arun G, Stentrup M, Gross M et al. 2013. The Noncoding RNA MALAT1 Is a Critical Regulator of the Metastasis Phenotype of Lung Cancer Cells. *Cancer research* **73**: 1180-1189.

Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, Young G, Lucas AB, Ach R, Bruhn L et al. 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**: 295-300.

Guttman M, Rinn JL. 2012. Modular regulatory principles of large non-coding RNAs. *Nature* **482**: 339-346.

Hekimoglu B, Ringrose L. 2009. Non-coding RNAs in Polycomb/Trithorax regulation. *Rna Biol* **6**: 129-137.

Hsu PD, Lander ES, Zhang F. 2014. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**: 1262-1278.

Jeggari A, Marks DS, Larsson E. 2012. miRcode: a map of putative microRNA target sites in the long non-coding transcriptome. *Bioinformatics* **28**: 2062-2063.

Kaneko S, Son J, Shen SS, Reinberg D, Bonasio R. 2013. PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. *Nat Struct Mol Biol* **20**: 1258-1264.

Karreth FA, Tay Y, Perna D, Ala U, Tan SM, Rust AG, DeNicola G, Webster KA, Weiss D, Perez-Mancera PA et al. 2011. In vivo identification of tumor- suppressive PTEN ceRNAs in an oncogenic BRAF-induced mouse model of melanoma. *Cell* **147**: 382-395.

Keniry A, Oxley D, Monnier P, Kyba M, Dandolo L, Smits G, Reik W. 2012. The H19 lincRNA is a developmental reservoir of miR-675 that suppresses growth and lgf1r. *Nature Cell Biology* **14**: 659-665.

Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A et al. 2009. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **106**: 11667-11672.

Kino T, Hurt DE, Ichijo T, Nader N, Chrousos GP. 2010. Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Sci Signal* **3**: ra8.

Klein U, Lia M, Crespo M, Siegel R, Shen Q, Mo T, Ambesi-Impiombato A, Califano A, Migliazza A, Bhagat G et al. 2010. The DLEU2/miR-15a/16-1 cluster controls B cell proliferation and its deletion leads to chronic lymphocytic leukemia. *Cancer Cell* **17**: 28-40.

Kloc M, Wilk K, Vargas D, Shirato Y, Bilinski S, Etkin LD. 2005. Potential structural role of non-coding and coding RNAs in the organization of the cytoskeleton at the vegetal cortex of Xenopus oocytes. *Development* **132**: 3445-3457.

Konermann S, Brigham MD, Trevino AE, Joung J, Abudayyeh OO, Barcena C, Hsu PD, Habib N, Gootenberg JS, Nishimasu H et al. 2014. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature*.

Koziol MJ, Rinn JL. 2010. RNA traffic control of chromatin complexes. *Curr Opin Genet Dev* **20**: 142-148.

Kretz M, Siprashvili Z, Chu C, Webster DE, Zehnder A, Qu K, Lee CS, Flockhart RJ, Groff AF, Chow J et al. 2013. Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature* **493**: 231-235.

Leighton PA, Ingram RS, Eggenschwiler J, Efstratiadis A, Tilghman SM. 1995. Disruption of imprinting caused by deletion of the H19 gene region in mice. *Nature* **375**: 34-39.

Li L, Liu B, Wapinski OL, Tsai MC, Qu K, Zhang J, Carlson JC, Lin M, Fang F, Gupta RA et al. 2013. Targeted Disruption of Hotair Leads to Homeotic Transformation and Gene Derepression. *Cell Rep*.

Marques AC, Ponting CP. 2009. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol* **10**: R124.

Moran VA, Perera RJ, Khalil AM. 2012. Emerging functional and mechanistic paradigms of mammalian long non-coding RNAs. *Nucleic Acids Res*.

Nagano T, Mitchell JA, Sanz LA, Pauler FM, Ferguson-Smith AC, Feil R, Fraser P. 2008. The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* **322**: 1717-1720.

Ng SY, Johnson R, Stanton LW. 2012. Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. *EMBO J* **31**: 522-533.

Redrup L, Branco MR, Perdeaux ER, Krueger C, Lewis A, Santos F, Nagano T, Cobb BS, Fraser P, Reik W. 2009. The long noncoding RNA Kcnq1ot1 organises a lineage-specific nuclear domain for epigenetic gene silencing. *Development* **136**: 525-530.

Reinius B, Shi C, Hengshuo L, Sandhu KS, Radomska KJ, Rosen GD, Lu L, Kullander K, Williams RW, Jazin E. 2010. Female-biased expression of long non-coding RNAs in domains that escape X-inactivation in mouse. *BMC Genomics* **11**: 614.

Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E et al. 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**: 1311-1323.

Ripoche MA, Kress C, Poirier F, Dandolo L. 1997. Deletion of the H19 transcription unit reveals the existence of a putative imprinting control element. *Genes Dev* **11**: 1596-1604.

Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP. 2011. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* **146**: 353-358.

Sasaki YT, Ideue T, Sano M, Mituyama T, Hirose T. 2009. MENepsilon/beta noncoding RNAs are essential for structural integrity of nuclear paraspeckles. *Proc Natl Acad Sci U S A* **106**: 2525-2530.

Sauvageau M, Goff LA, Lodato S, Bonev B, Groff AF, Gerhardinger C, Sanchez-Gomez DB, Hacisuleyman E, Li E, Spence M et al. 2013. Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife* **2**: e01749.

Schmitt S, Paro R. 2006. RNA at the steering wheel. *Genome Biology* **7**.

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. 2012. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**: 458-472.

Simon MD, Wang CI, Kharchenko PV, West JA, Chapman BA, Alekseyenko AA, Borowsky ML, Kuroda MI, Kingston RE. 2011. The genomic binding sites of a noncoding RNA. *Proc Natl Acad Sci U S A* **108**: 20497-20502.

Spitale RC, Tsai MC, Chang HY. 2011. RNA templating the epigenome: long noncoding RNAs as molecular scaffolds. *Epigenetics* **6**: 539-543.

Sunwoo H, Dinger ME, Wilusz JE, Amaral PP, Mattick JS, Spector DL. 2009. MEN epsilon/beta nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. *Genome Res* **19**: 347-359.

Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E, Chang HY. 2010. Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**: 689-693.

Ulitsky I, Bartel DP. 2013. lincRNAs: Genomics, Evolution, and Mechanisms. *Cell* **154**: 26-46.

Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. 2011. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147**: 1537-1550.

Wang KC, Chang HY. 2011. Molecular mechanisms of long noncoding RNAs. *Mol Cell* **43**: 904-914.

Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, Lajoie BR, Protacio A, Flynn RA, Gupta RA et al. 2011. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**: 120-124.

Wang P, Xue Y, Han Y, Lin L, Wu C, Xu S, Jiang Z, Xu J, Liu Q, Cao X. 2014. The STAT3-binding long noncoding RNA lnc-DC controls human dendritic cell differentiation. *Science* **344**: 310-313.

West JA, Davis CP, Sunwoo H, Simon MD, Sadreyev RI, Wang PI, Tolstorukov MY, Kingston RE. 2014. The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. *Mol Cell* **55**: 791-802.

Willingham AT, Orth AP, Batalov S, Peters EC, Wen BG, Aza-Blanc P, Hogenesch JB, Schultz PG. 2005. A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* **309**: 1570-1573.

Wutz A, Theussl HC, Dausman J, Jaenisch R, Barlow DP, Wagner EF. 2001. Non-imprinted Igf2r expression decreases growth and rescues the Tme mutation in mice. *Development* **128**: 1881-1887.

Yamashita A, Watanabe Y, Nukina N, Yamamoto M. 1998. RNA-assisted nuclear transport of the meiotic regulator Mei2p in fission yeast. *Cell* **95**: 115-123.

Yildirim E, Kirby JE, Brown DE, Mercier FE, Sadreyev RI, Scadden DT, Lee JT. 2013. Xist RNA Is a Potent Suppressor of Hematologic Cancer in Mice. *Cell* **152**: 727-742.

Zhao J, Ohsumi TK, Kung JT, Ogawa Y, Grau DJ, Sarma K, Song JJ, Kingston RE, Borowsky M, Lee JT. 2010. Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol Cell* **40**: 939-953.

Appendix A: Supplementary information for chapter 1

Parts of this work were first published as supplementary information for:

Alvarez-Dominguez JR, Hu W, Yuan B, Shi J, Park SS, Gromatzky AA, van Oudenaarden A, Lodish HF. 2014. Global discovery of erythroid long noncoding RNAs reveals novel regulators of red cell maturation. *Blood* **123**: 570-581.

Note: Supplementary Tables 2 and 4 have been omitted for space considerations

Supplemental Methods

Annotation data sources

The mouse July 2007 (NCBI37/mm9) genome assembly was used throughout the study. Ensembl transcript structures and annotations were obtained from Ensembl version 67 (http://useast.ensembl.org/info/data/ftp/). RefSeq, UCSC and TransMap transcript structures and annotations, as well as annotations of repetitive elements, were obtained from the UCSC genome browser (July, 2012). Enhancer annotations from E14.5 fetal liver cells (Shen et al. 2012) were downloaded from the Ren lab website (http://chromosome.sdsc.edu/mouse/download.html).

RNA-Seq data sources

We used previously published poly(A)-selected RNA-Seq reads from purified fetal liver BFU-E, CFU-E and TER119+ cells (Flygare et al. 2011) deposited in the Gene Expression Omnibus (GEO, accession number GSE26086). poly(A)-selected RNA-Seq reads from 30 primary cell and tissue types (supplemental Table 3) aligned to the mouse genome (mm9 version) were downloaded from the mouse ENCODE portal (http://genome.ucsc.edu/ENCODE/downloadsMouse.html). poly(A)-selected RNA-Seq reads from G1E-ER4 cells were accessed through the PSU genome browser (http://main.genome-browser.bx.psu.edu) (March 2013). poly(A)-selected RNA-Seq reads from K562 cells aligned to the human genome (hg19 version) from the human ENCODE consortium were accessed through the UCSC genome browser (March 2013).

ChIP-Seq data sources

We examined previously published density maps of ChIP-Seq signal enrichment for histone modifications (H3K4me3, H3K4me2, H4K16Ac, H3K9Ac, H3K27me3, H3K36me3, H3K79me2) and for serine 5 phosphorylated RNA polymerase II in mouse fetal liver erythroid TER119-negative and TER119-positive cells (Wong et al. 2011) deposited in GEO (accession number GSE32111). Similarly, density maps of processed ChIP-Seq signal enrichment for H3K4me1, H3K4me3 and H3K27Ac in E14.5 fetal liver cells (Stamatoyannopoulos et al. 2012) were downloaded from the mouse ENCODE portal (http://genome.ucsc.edu/ENCODE/downloadsMouse.html). Peaks of processed ChIP-Seq signal enrichment for GATA1 and TAL1 (Wu et al. 2011) and for KLF1 (Pilon et al. 2011) in TER119+ erythroblasts, inferred by the MACS algorithm (Zhang et al. 2008) with an FDR threshold of 5%, were obtained from the mouse ENCODE portal (http://genome.ucsc.edu/ENCODE/downloadsMouse.html) and from the PSU genome browser (http://main.genome-browser.bx.psu.edu), respectively. Density maps of processed ChIP-Seq signal enrichment for H3K4me1, H3K4me3, H3K36me3, H3K27ac and for RNA POL II, GATA1, TAL1 and p300 occupancy in K562 cells were accessed through the human ENCODE consortium tracks in the UCSC genome browser (March 2013).

Additional data sources

Mapped CAGE tags from the FANTOM3 and FANTOM4 projects (Carninci et al. 2005; Ravasi et al. 2010) were downloaded from the FANTOM website (http://fantom.gsc.riken.jp/4/download/GenomeBrowser/ucsc/mm9/). Mapped poly(A)-sequencing tags from the Merck Research Laboratories (Derti et al. 2012) were downloaded from the UCSC genome browser (July 2012). Density maps of sequencing tags from DNAse I hypersensitive sites in mouse BFU-Es, CFU-Es, and TER119+ erythroblasts and in human K562 cells were accessed through the mouse and human ENCODE tracks in the UCSC genome browser (November 2012). ChIA-PET interactions and clusters in K562 cells were accessed through the WashU Epigenome Browser (http://epigenomegateway.wustl.edu/).

RNA-Seq and analysis

E14.5 mouse fetal liver cells were separated into TER119+ and TER119- fractions via magnetic-assisted cell sorting. Total RNA was isolated from these cells using the QIAGEN miRNeasy Kit according to the manufacturer’s instructions. Ribosomal RNA was depleted from 4 ug total RNA using the Ribo-Zero Gold Kit from Epicentre. Strand-specific sequencing libraries were generated following a previously described protocol (Borodina et al. 2011) from the total RNA (TER119- and TER119+ cells) or from the poly(A)+ and poly(A)- fractions (TER119+ cells). The latter fractions were separated using the Solexa kit (Illumina) according to the manufacturer’s instructions. cDNA fragments of 400-600 bp from these libraries were selected by gel purification and then sequenced on a Illumina HiSeq2000 sequencer. The resulting directional 100 bp paired-end reads were quality-checked with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and low quality reads and adapters were removed using the fastq_quality_trimmer (“-t 20” parameter) and the fastx_clipper tools from FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Only perfect paired reads were mapped to mm9 using TopHat v1.3.2 (Trapnell et al. 2009) (default parameters and "--min-anchor 5"). To construct transcript models, we took the mapped reads from total RNA data and performed *de novo* assembly using Cufflinks v1.3.0 (Trapnell et al. 2010) (default parameters and "-p 4 --min-frags-per-transfrag 0") considering annotations from RefSeq, Ensembl and UCSC to maximize assembly accuracy. To retain only reliable transcript models, all transcripts were required to have >1 exons, be >200bp in length, have <50% repeat-masked sequence, and exhibit average base-level read coverage >0.1 (empirical threshold). Transcript models meeting these criteria were then used for quantifying expression in the total RNA data (as well as in every other dataset examined in this study). Gene-level expression was estimated as fragments per kilobase of exon model per million mapped fragments (FPKM) using Cufflinks. To identify genes expressed during erythroid differentiation, we selected all mRNAs and lncRNAs from the TER119- and TER119+ data that are expressed at >0 FPKM in both replicates of at least one of the BFU-E, CFU-E or TER119+ differentiation stages. To quantify differential gene expression, mapped reads were assigned to the transcript models derived from our *de novo* analysis using HTSeq-count (http://www-huber.embl.de/users/anders/HTSeq/doc/index.html) with the "intersection_strict" mode. Normalization and differentially expressed genes between TER119+ and TER119- cells were then analyzed with the DESeq R Bioconductor package (Anders and Huber 2010). DESeq controls for the variation in the number of reads obtained across samples by conducting count-based normalization. After normalization, fold changes and their significance (p-values), indicating differential expression, are determined using a model based on the negative binomial distribution. To select a significance threshold, we benchmarked the DESeq results against a set of 475 protein-coding genes known to be induced during terminal erythroid differentiation (Wong et al. 2011). Based on this analysis, we considered all genes with a p-value < 0.05 to be differentially expressed. To examine gene expression in poly(A)+ RNA-Seq data from purified BFU-E, CFU-E and TER119+ cells (Flygare et al. 2011), the non-directional 36 bp paired-end reads were mapped to mm9 by TopHat guided by the gene models assembled *de novo* from the total RNA-Seq data, using the "–segment-length 15" parameter due to their short length. Gene-level expression for these reads, and for already-mapped poly(A)+ RNA-Seq reads from 30 mouse ENCODE cell and tissue types (Stamatoyannopoulos et al. 2012) (supplemental table 3), were quantified by Cufflinks based on the gene models assembled *de novo* from the total RNA-Seq data.

lncRNA identification pipeline

After filtering transcript models to retain only reliable ones (see above), we implemented the following strategy to identify long non-coding RNAs:

1. To remove transcripts with known protein domains, for each transcript we retrieved the longest ORF in all three possible frames using the Sixpack tool from EMBOSS (Rice et al. 2000), and then used HMMER3 (Finn et al. 2011) to query the Pfam A and Pfam B databases (downloaded from ftp://ftp.sanger.ac.uk/pub/databases/Pfam/current_release/ on Nov 2012) with default parameters. Any transcript with a significant Pfam hit (E-value > 0.001) was excluded from further analysis.
2. Repeat-masked transcripts were also blasted against the human, rat and mouse RefSeq databases separately using Blastx (Gish and States 1993). Transcripts mapping to all three protein databases with an E-value <0.0001 were removed.

3. We used PhyloCSF (Lin et al. 2011) to filter out transcripts under evolutionary pressure to preserve synonymous amino acid codons. The PhyloCSF score of a given transcript indicates how much more probable its alignment across 29 mammalian genomes is under a model of protein-coding sequence evolution than under a non-coding model. We calculated PhyloCSF scores with "—removeRefGaps --frames=3 --orf=ATGStop" parameters and discarded any transcript with a score >100, which corresponds to a 9.3% false negative rate and a 9.7% false positive rate using RefSeq mRNAs and RefSeq lincRNAs as reference.
4. We used the Coding Potential Calculator (Kong et al. 2007) to exclude transcripts with characteristic coding features independent of their conservation. The CPC score of a given transcript indicates its distance to a classification as protein-coding based on significant similarity to sequence features of known protein-coding transcripts learned via support vector machine learning. We calculated CPC scores using default parameters and discarded any transcript with a CPC score >0, which corresponds to "coding" or "weakly coding" classifications.
5. Transcripts with a predicted ORF longer than 100 aa were removed.
6. We used BEDTools to intersect our *de novo* transcript models with transcript models from the RefSeq, UCSC and Ensembl databases, and discarded any transcript overlapping at least 1 bp in the same strand with any known mRNA exon.

The same strategy was used to curate lncRNA annotations from the Ensembl, UCSC and RefSeq catalogs.

lncRNA classification pipeline

To classify lncRNAs into known categories (see supplemental Figure 1D), we used BEDTools to intersect them with known annotations and implemented the following strategy:

1. lncRNAs that do not intersect any gene annotation from Ensembl, RefSeq or UCSC or any region from our enhancer dataset were classified as novel “intergenic lncRNAs”.
2. lncRNAs that have a transcription start site (TSS) within 1 kb of an enhancer annotated in E14.5 fetal liver cells (Shen et al. 2012) were classified as “enhancer lncRNAs”.
3. lncRNAs that overlap in the opposite strand with any region (intronic or exonic) of an annotated transcript model (Ensembl, RefSeq or UCSC), but do not overlap any known annotation in the same strand, were classified as novel “antisense lncRNAs”.
4. lncRNAs whose uniquely-mapped reads overlap with an annotated pseudogene transcript model in the same strand were classified as “pseudogene lncRNAs”. Pseudogene models were curated from pseudogenes annotated by the Vertebrate Genome Annotation database (Vega) (Ashurst et al. 2005) after removing those overlapping RefSeq NR_ transcripts.
5. lncRNAs that overlap with an annotated “intergenic lncRNA” transcript model in the same strand were classified as known “intergenic lncRNAs”. Intergenic lncRNA models were curated from Ensembl “lincRNA” annotations and from RefSeq non-coding RNAs after removing transcripts overlapping with RefSeq coding, UCSC coding, Vega pseudogene or Ensembl miRNA, misc_RNA, rRNA, tRNA, snoRNA, or snRNA annotations.
6. lncRNAs that overlap with an annotated sRNA in the same strand were classified as “sRNA host lncRNAs”. sRNA models were curated from Ensembl miRNA, misc_RNA, rRNA, tRNA, snoRNA, and snRNA annotations.

7. lncRNAs that overlap with an annotated "antisense lncRNA" transcript model in the same strand were classified as known "antisense lncRNAs". Antisense lncRNA models were curated from Ensembl "antisense" annotations.
8. lncRNAs that overlap with an annotated protein-coding transcript model (Ensembl, RefSeq or UCSC) in the same strand (but do not overlap any of its exons) were classified as "intron overlapping lncRNAs".
9. lncRNAs ascribed to more than one class were inspected manually and resolved based on the evidence for each class with the following priority: sRNA host > enhancer lncRNA > intronic lncRNA > pseudogene lncRNA > antisense lncRNA > intergenic lncRNA.
10. lncRNAs that could not be placed in any category were classified as "other lncRNAs"

CAGE and polyA-Seq analysis

We used BEDTools (Quinlan and Hall 2010) to intersect mapped CAGE tags with the inferred TSS of our lncRNA models as well as to intersect mapped poly(A)-sequencing tags from the Merck Research Laboratories (Derti et al. 2012) with the proposed lncRNA transcription end sites.

Conservation analysis

PhastCons conservation scores from a 30-way vertebrate genome alignment seeded with the mouse genome (Blanchette et al. 2004) were downloaded from the UCSC genome browser. For each lncRNA transcript, a computational control was generated by shuffling its exon coordinates to a randomly chosen region within the same chromosome with no known Ensembl, UCSC or

RefSeq gene annotations. Average PhastCons scores were obtained by dividing the aggregate PhastCons scores along exons or promoter-proximal regions (TSS ± 1 kb) by the total length of the genomic area covered by the scores. To identify orthologs of mouse lncRNAs, we implemented a two-step strategy. First, we used the LiftOver tool from the UCSC browser (default parameters and “Min ratio of alignment blocks or exons that must map: 0.1”), which is based on blastz (Schwartz et al. 2003), to map the mm9 exon coordinates to each of 19 vertebrate genomes via pairwise alignments, and retained only multi-exonic >200 bp alignments. Second, for those lncRNAs for which no ortholog could be found during the first step, were re-analyzed them using BLAT(Kent 2002) (default parameters and “-minIdentity=50”) against each genome, considering only the best hit and retaining only multi-exonic >200 bp alignments. To identify expressed orthologs of mouse fetal liver-expressed lncRNAs, we used BEDTools to intersect them with UCSC, RefSeq, mRNA or EST transcripts from 19 vertebrate genomes mapped to the mouse genome by TransMap (Zhu et al. 2007) via pairwise syntenic blastz alignments (downloaded from the UCSC genome browser), and considered any lncRNA having an exonic overlap in the same strand with a TransMap transcript to have an expressed ortholog.

Tissue expression analysis

We retrieved poly(A)+ RNA-Seq reads from a panel of mouse ENCODE primary cell and tissue types (supplemental table 3) mapped to the mouse genome (mm9 version) and quantified gene-level expression (FPKM) using Cufflinks based on the *de novo* gene models assembled from our total RNA-Seq data. Specificity of expression of a given gene to a given tissue was scored as the fraction of the total expression across tissues that it represents (i.e. its fractional expression level). Tissue specificity scores ranged from 3.308263e-07 to 0.988. To distinguish broadly-expressed from tissue-restricted genes, we benchmarked these scores against select housekeeping and erythroid-restricted mRNAs (Kingsley et al. 2013). Based on this analysis, we chose an empirical cutoff of 0.1 (3-fold higher than the background expectation of 0.032 for uniformly expressed genes) and designated genes scoring above this threshold as tissue-restricted. Tissue-restricted genes having fetal liver TER119+ erythroblasts as the tissue with the maximal expression specificity score were considered enriched for erythroid specificity. Genes expressed >1 FPKM (mRNAs) or >0.1 FPKM (lncRNAs) in erythroblasts whose average expression value across all other tissues is also >1 FPKM (mRNAs) or >0.1 FPKM (lncRNAs), respectively, were considered to be broadly expressed.

Gene ontology analysis

Gene lists were analyzed for enrichment of Gene Ontology (GO) terms using DAVID (Huang da et al. 2009b; Huang da et al. 2009a). Only GO Biological Process and Molecular Function terms (GOTERM_BP_FAT and GOTERM_MF_FAT) were considered. To identify the most significant non-redundant GO terms in a gene list, we grouped annotation terms into non-redundant clusters using the Functional Annotation Clustering tool, and then selected the most significant term in each of the top clusters ranked by their enrichment score. Only GO terms enriched in our lists with a Benjamini-Hochberg adjusted p-value <0.05 are reported.

ChIP-Seq analysis

Chromatin marks at promoter-proximal regions (TSS ± 1 kb) or along gene bodies (TSS to TES span) were assessed by computing the average processed ChIP-Seq signal enrichment within a 2kb region centered on the TSS or along the region between the TSS and TES of each gene examined, respectively. The mean expression and chromatin mark enrichment of each gene differentially expressed gene are reported. To determine promoter-proximal targeting by GATA1, TAL1 or KLF1, genome-wide ChIP-Seq maps of estimated binding peaks were examined for overlap of at least 1 bp with the TSS ± 1 kb region of each gene analyzed.

Selection of lncRNA candidates

We used the following criteria to prioritize erythroid lncRNAs for functional studies:

1. To select for robust transcript models, we retained only transcripts with independent support of the TSS by CAGE tags or by overlap with a RNAPII enrichment peak.
2. To pick actively transcribed loci, we required that they overlap with at least one active histone mark peak within 1 kb of the TSS and with at least one elongation histone mark peak along the gene body.
3. To identify lncRNAs that may be important for erythropoiesis, we required that they be targeted by at least one of the key erythroid TFs GATA1, TAL1 or KLF1, or that they count TER119+ erythroblasts as the cell type of maximal expression specificity.
4. To uncover lncRNAs with likely roles in the production of mature erythroblasts, we ranked them based on their relative fold change in expression between fetal liver erythroid progenitor-enriched cells and TER119+ erythroblasts, and then based on their absolute expression level at the TER119+ stage.

Single-molecule RNA FISH and analysis

Oligonucleotide DNA probes (20 nt each) tiling the exonic regions of target transcripts were designed using the online designer at http://www.singlemoleculefish.com (version 3.0), with a minimum spacing of 2 nt and a target GC content of 45%. The probe designer applies stringent cutoffs to maximize probe specificity by masking specific regions of the mouse genome, including repetitive and low complexity regions. Probes were synthesized with an amine group at the 3'end (Biosearch Technologies), coupled to Alexa fluor 594 (Invitrogen) or Cy5 (GE Amersham) and purified on an HPLC column. Fetal liver TER119+ erythroblasts were purified by FACS using a FITC-conjugated antibody (BD Biosciences) and fixed in 1-2 ml of 3.7% formaldehyde, 1x PBS for 10 minutes at room temperature, and permeabilized in 70% ethanol for at least 16hr. Single-molecule RNA FISH was performed as described previously (Raj et al. 2008). For hybridization to DNA probes, cells were rehydrated in wash buffer containing 25% (v/v) formamide and 2x SSC for 5min, and 50µl of hybridization solution, containing labeled DNA probes in 25% (v/v) formamide, 2x SSC, 1mg/ml BSA, 10mM Vanadyl-ribonucleoside complex, 0.5mg/ml E. coli tRNA and 0.1 g/ml dextran sulfate, were added to the sample and incubated overnight at 30°C. Optimal probe concentrations were determined empirically for each probe set by running a dilution series. Before imaging, cells were washed twice in 25% (v/v) formamide and 2x SSC for 30min, with 5ng/ml DAPI added after the first wash for nuclear counterstaining. Hybridized cells were immobilized in chambered cover glasses (Lab-Tek) coated with Cell-Tak tissue adhesive (BD) prior to imaging. Fluorescence microscopy and image acquisition and analysis were conducted as described (Neuert et al. 2013). For imaging, 200µl of an oxygen-scavenging solution, containing 10mM Tris (pH 7.5), 2x SSC and 0.4% glucose supplemented with 74µg/ml glucose oxidase, 74µg/ml catalase, and 2mM Trolox, were added to the immobilized cells. Images were taken with a Nikon TI-E inverted fluorescence microscope using a 100x oil-immersion objective, custom filters designed to distinguish between the different fluorophores used and a Photometrics Pixis 1024 CCD camera (Princeton Instruments) managed by the MetaMorph software (Molecular Devices, Downington, PA). Stacks of images were taken automatically with 0.3µm between z-slices in the DAPI, AF594 and Cy5 channels. For each biological replicate, at least 3 fields of view covering ~100 cells were imaged. For image processing, the maximum projection of DAPI image z-stacks was used to identify individual cells using an in-house edge detection algorithm. Connected regions spanning longer than the expected range of cell sizes were rejected. The AF594 or Cy5 channels were then used to detect diffraction-limited spots representing individual RNA transcripts. A Laplacian filter was applied on each z-stack to enhance particle signals, and a fixed pixel intensity threshold was used to detect individual spots in each plane. Optimal thresholds were determined empirically for each probe set. Nuclear vs. cytoplasmic localization was determined manually by visual inspection of merged DAPI and AF594 or Cy5 images. For image presentation, maximum-project AF5943 or Cy5 images (pseudo-colored red) were merged with maximum-project DAPI images (pseudo-colored blue). Enhanced contrast in the DAPI channel was used to emphasize nuclear counterstaining boundaries.

Additional bioinformatics analyses

Statistical tests, correlation analyses and plots were implemented in R (http://www.R-project.org/) with default parameters, unless stated otherwise. Pearson’s product-moment correlation test between paired samples was implemented using the *cor.test* function in R. Heatmaps were produced using the heatmap.2 function of the *gplots* package (http://CRAN.R-project.org/package=gplots).

References

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11**: R106.

Ashurst JL, Chen CK, Gilbert JG, Jekosch K, Keenan S, Meidl P, Searle SM, Stalker J, Storey R, Trevanion S et al. 2005. The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res* **33**: D459-465.

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**: 708-715.

Borodina T, Adjaye J, Sultan M. 2011. A strand-specific library preparation protocol for RNA sequencing. *Methods Enzymol* **500**: 79-98.

Carninci P Kasukawa T Katayama S Gough J Frith MC Maeda N Oyama R Ravasi T Lenhard B Wells C et al. 2005. The transcriptional landscape of the mammalian genome. *Science* **309**: 1559-1563.

Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, Babak T. 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Res* **22**: 1173-1183.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39**: W29-37.

Flygare J, Rayon Estrada V, Shin C, Gupta S, Lodish HF. 2011. HIF1alpha synergizes with glucocorticoids to promote BFU-E progenitor self-renewal. *Blood* **117**: 3435-3444.

Gish W, States DJ. 1993. Identification of protein coding regions by database similarity search. *Nat Genet* **3**: 266-272.

Huang da W, Sherman BT, Lempicki RA. 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**: 1-13.

-. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**: 44-57.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656-664.

Kingsley PD, Greenfest-Allen E, Frame JM, Bushnell TP, Malik J, McGrath KE, Stoeckert CJ, Palis J. 2013. Ontogeny of erythroid gene expression. *Blood* **121**: e5-e13.

Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G. 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* **35**: W345-349.

Lin MF, Jungreis I, Kellis M. 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**: i275-282.

Neuert G, Munsky B, Tan RZ, Teytelman L, Khammash M, van Oudenaarden A. 2013. Systematic identification of signal-activated stochastic gene regulation. *Science* **339**: 584-587.

Pilon AM, Ajay SS, Kumar SA, Steiner LA, Cherukuri PF, Wincovitch S, Anderson SM, Mullikin JC, Gallagher PG, Hardison RC et al. 2011. Genome-wide ChIP-Seq reveals a dramatic shift in the binding of the transcription factor erythroid Kruppel-like factor during erythrocyte differentiation. *Blood* **118**: e139-148.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. 2008. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods* **5**: 877-879.

Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, Tan K, Akalin A, Schmeier S, Kanamori-Katayama M, Bertin N et al. 2010. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* **140**: 744-752.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**: 276-277.

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. 2003. Human-mouse alignments with BLASTZ. *Genome Res* **13**: 103-107.

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV et al. 2012. A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**: 116-120.

Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, Groudine M, Bender M, Kaul R, Canfield T et al. 2012. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol* **13**: 418.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105-1111.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511-515.

Wong P, Hattangadi SM, Cheng AW, Frampton GM, Young RA, Lodish HF. 2011. Gene induction and repression during terminal erythropoiesis are mediated by distinct epigenetic changes. *Blood* **118**: e128-138.

Wu W, Cheng Y, Keller CA, Ernst J, Kumar SA, Mishra T, Morrissey C, Dorman CM, Chen KB, Drautz D et al. 2011. Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration. *Genome Res* **21**: 1659-1671.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.

Zhu J, Sanborn JZ, Diekhans M, Lowe CB, Pringle TH, Haussler D. 2007. Comparative genomics search for losses of long-established genes on the human lineage. *PLoS Comput Biol* **3**: e247.

Supplemental Figure Legends

Supplemental Figure 1. Identification and classification of fetal liver-expressed lncRNAs

(A) Maximum predicted ORF length (across all possible reading frames) of mRNAs and lncRNAs expressed in mouse TER119- and TER119+ fetal liver cells.

(B) Coding capacity of mRNAs and lncRNAs as estimated by PhyloCSF scores,(Lin et al. 2011) which reflect evolutionary pressure to preserve synonymous amino acid codons.

(C) Distribution of Ensembl annotations for genes classified as unclear coding potential based on our coding potential filters.

(D) Workflow for classification of lncRNA transcripts based on overlap with Ensembl transcript annotations or with enhancers annotated in E14.5 fetal liver cells.(Shen et al. 2012) Parallel “if” steps are not mutually exclusive, but parallel “if/else” steps are. Transcripts assigned to more than one category, and loci encoding transcripts assigned to different categories, were resolved manually as described in Supplemental methods.

(E) Average enrichment over elncRNA TSS ± 1 kb regions for histone marks associated with active enhancers (H3K4me1 and H3K27Ac) or with active promoters (H3K4me3) in E14.5 fetal liver cells, and for RNA Pol II in TER119+ fetal liver cells. Color intensity represents processed ChIP-Seq signal enrichment. Heatmaps are sorted by the H3K27Ac values.

(F) Distribution of Ensembl annotations for each class of lncRNA transcript identified. Novel transcripts are not found in either the Ensembl, Refseq or UCSC databases.

Supplemental Figure 2. Tissue specificity of fetal liver and erythroid lncRNAs

(A) Relative abundance of various classes of lncRNA genes (rows) expressed in fetal liver across 30 mouse primary cell and tissue types from the ENCODE consortium (columns). Color intensity represents the fractional gene-level expression across all tissues examined. ERY_1 and ERY_2 (red) are fetal liver TER119+ erythroblast replicates. Black bars in the left panels highlight erythroid-restricted genes.

(B) Distribution of maximal tissue specificity scores of fetal liver-expressed mRNA and lncRNA genes across matched expression ranges. Scores represent the fractional expression level across 30 mouse tissues as in (A).

(C) Patterns of tissue specificity among mRNAs and lncRNA genes. For each tissue, color intensity represents the proportion of genes with that tissue as their tissue of maximal expression specificity. Tissue specificity is calculated by the fractional expression level across all tissues as in (A).

(D) Proportion of mRNA and lncRNA genes with fetal liver TER119+ erythroblasts as their tissue of maximal expression specificity.

(E) Violin plots of erythroid expression enrichment distributions for broadly expressed mRNAs or lncRNAs, defined as mRNAs with >1 FPKM in fetal liver TER119+ erythroblasts and >1 average FPKM across all other tissues or lncRNAs with >0.1 FPKM in fetal liver TER119+ erythroblasts and >0.1 average FPKM across all other tissues, respectively. Erythroid expression enrichment represents the ratio of erythroblast expression to the mean expression across all other tissues.

Supplemental Figure 3. Expression features of mRNAs and lncRNAs during erythroid differentiation

(A) Top 10 non-redundant GO terms enriched among mRNA genes that are differentially expressed during erythropoiesis and are upregulated >2-fold in expression between BFU-E or CFU-E progenitors and TER119+ erythroblasts. Differential expression was determined by DESeq at a 5% false discovery threshold. GO term enrichment was determined by DAVID at a 5% false discovery threshold.

(B) Dynamic expression of mRNA and lncRNA genes during erythropoiesis. For each gene, the average coefficient of variation in gene-level expression (FPKM) through the BFU-E, CFU-E and TER119+ stages of differentiation is shown.

(C) Fraction of differentially expressed mRNA and lncRNA genes, determined by DESeq at a 5% false discovery threshold.

Supplemental Figure 4. Chromatin signatures around the TSS of mRNAs and lncRNAs that are dynamically expressed during erythroid differentiation

(A) Scaled density of enrichment signal for chromatin modifications within TSS ± 1kb regions of differently expressed mRNA or lncRNA transcripts in erythroid progenitor-enriched fetal liver cells. For each lncRNA transcript, a control was generated by mapping its exon-intron structure to a randomly chosen region of intergenic space and analyzing its TSS ± 1kb chromatin mark enrichment.

(B) Same as in (A) but for TER119+ erythroblasts.

Supplemental Figure 5. Correlation between gene expression and chromatin modification changes of lncRNAs that are dynamically expressed during erythroid differentiation

(A) Change in gene-level expression vs. change in mean RNA Pol II or chromatin mark enrichment within 1kb of the TSS of lncRNAs that are differentially expressed during erythroid differentiation. Changes are shown as the log2 ratio of the levels in TER119+ erythroblasts to the levels in erythroid progenitor-enriched fetal liver cells. Correlation values (Pearson's r) are shown in the top left corners.

Supplemental Figure 6. Targeting of lncRNAs by key erythroid transcription factors

(A) Top 5 non-redundant GO terms enriched in mRNA genes co-occupied by GATA1, TAL1 and KLF1 within TSS ± 1kb promoter regions that are differentially expressed during erythropoiesis and are induced >2-fold in expression between BFU-E or CFU-E progenitors and

TER119+ erythroblasts. GO term enrichment was determined by DAVID at a 5% false discovery threshold.

(B) Examples of mRNA loci co-targeted by GATA1 and TAL1 in TER119+ erythroblasts. UCSC Genome Browser tracks display raw density maps of strand-specific RNA-Seq reads of total RNA from erythroid progenitor-enriched fetal liver cells (PROG) or TER119+ fetal liver erythroblasts (ERY), and density maps of processed signal enrichment for ChIP-Seq against H3K4me2 in PROG or ERY and against GATA1 or TAL1 in ERY. Locus names are indicated above each panel and UCSC annotations are shown in the middle track. Gray panels highlight sites of GATA1 and TAL1 co-binding and H3K4me2 marking near the TSS.

(C) Time course of transcriptional activation of shlncRNA-EC6 and elncRNA-EC1 after restoration of GATA1 in the mouse G1E-ER4 cell line. UCSC Genome Browser tracks display raw density maps of strand-specific RNA-Seq reads of poly(A)+ RNA from G1-ER4 cells at various time points after treatment with Estradiol. Shown at the bottom are lncRNA transcript models based on our *de novo* assembly using Cufflinks, as well as Ensembl annotations.

(D) Conservation of shlncRNA-EC6 and elncRNA-EC1 regulatory features and chromatin architecture in the human K562 cell line. Tracks from the UCSC Genome Browser show density maps of normalized signal enrichment for RNA-Seq reads of poly(A)+ RNA from K562 cells, raw density maps of sequencing tags from DNase I hypersensitive (HS) sites assayed in K562, and density maps of processed signal enrichment for ChIP-Seq against H3K4me3, H3K36me3, H3K4me1, H3K27Ac, RNA Pol II, p300, GATA1 and TAL1 in K562. Shown at the bottom are lncRNA transcript models based on our detection of orthologous genomic regions from local alignment and synteny to the mouse genome, as well as UCSC annotations.

Supplemental Figure 7. Expression, regulation and conservation features of the top 12 candidate lncRNA modulators of erythropoiesis

(A-L) UCSC Genome Browser tracks display the landscape of transcription, chromatin accessibility, TF occupancy, histone modification and sequence conservation of the loci encoding the lncRNA candidates listed in Figure 5A. Shown are raw density maps of RNA-Seq reads of poly(A)+ RNA from FACS-purified BFU-Es, CFU-Es or TER119+ erythroblasts (ERY), raw density maps of strand-specific RNA-Seq reads of total RNA from erythroid progenitor-enriched fetal liver cells (PROG) or ERY, lncRNA transcript models based on our *de novo* assembly using Cufflinks, UCSC annotations, CAGE tag clusters, raw density maps of sequencing tags from DNase I hypersensitive (HS) sites assayed in BFU-E, CFU-Es or ERY, and density maps of processed signal enrichment for ChiP-Seq against GATA1, TAL1 or KLF1 in ERY and against RNAPII or chromatin modifications associated with transcriptional activation (H3K4me3, H3K4me2, H416Ac, H3K9Ac), elongation (H3K79me2, H3K36me3), repression (H3K27me3) or enhancer activity (H3K4me1) in PROG or ERY and H3K27Ac in E14.5 fetal liver cells. The last track displays density of phastCons scores of conservation across 19 placental mammalian genomes aligned to the mouse genome.

Supplemental Figure 8. Subcellular localization and single-cell transcript counts of the top 12 candidate lncRNA modulators of erythropoiesis

(A) Average percent of nuclear- and cytoplasmic-localized transcripts of lncRNA candidates, determined by single-molecule RNA FISH in fixed TER119+ fetal liver erythroblasts. Data are mean ± s.e.m (n = 2).

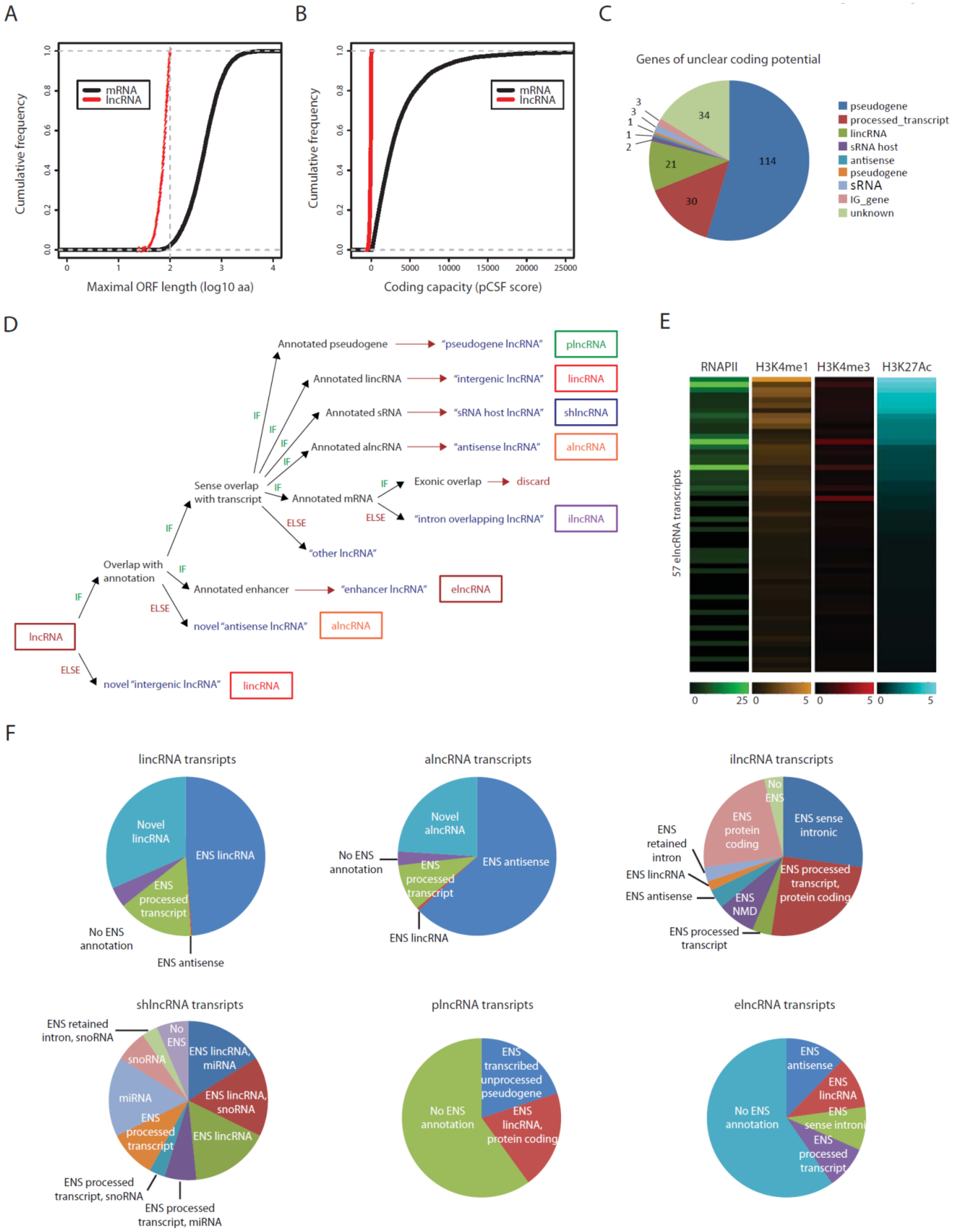
(B) Relative expression level of lncRNA candidates in the nuclear and cytoplasmic fractions of TER119+ fetal liver erythroblasts, determined by qPCR analysis of RNA extracted from these fractions. Values were normalized to those of 18S rRNA, and fold-changes were computed relative to the nuclear fraction levels. Shown as positive controls are 47S pre-rRNA, a nuclear species, and alpha-hemoglobin mRNA, a cytoplasmic species. Data are mean ± s.e.m (n = 3).

(C) Box-and-whisker plots for distributions of average per-cell transcript counts across microscopy images of lncRNA candidates in individual fixed TER119+ fetal liver erythroblasts expressing >0 lncRNA transcripts, determined by single-molecule RNA FISH. Median values are indicated by the thick horizontal bars. Circles represent outliers (>1.5 x upper quartile or <1.5 x lower quartile values). Cell counts are given by n. Data are from 2 biological replicate experiments.

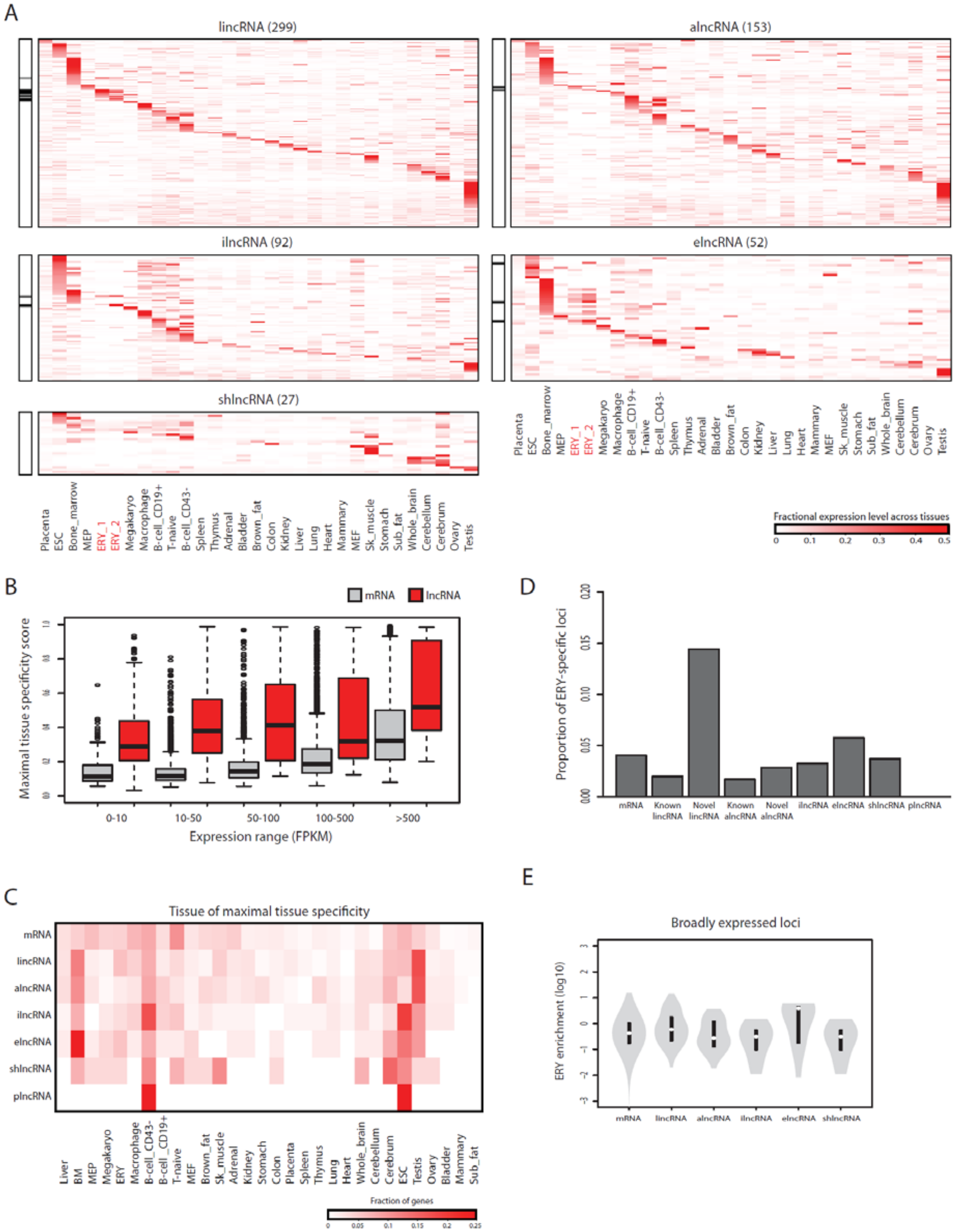
Supplemental Figure 9. Inhibition of the top 12 candidate lncRNA modulators of erythropoiesis

Relative expression of lncRNA candidates in TER119+ fetal liver cells upon depletion by shRNAs. Erythroid progenitor-enriched fetal liver cells were transduced by retroviral vectors encoding shRNAs targeting different transcript regions or scramble shRNA control, and effectively transduced cells were induced do differentiate in culture and analyzed for lncRNA expression by qPCR. Values were normalized to those of 18S rRNA, and fold-changes were computed relative to the scramble shRNA control. Data are mean ± s.d. (n = 3).

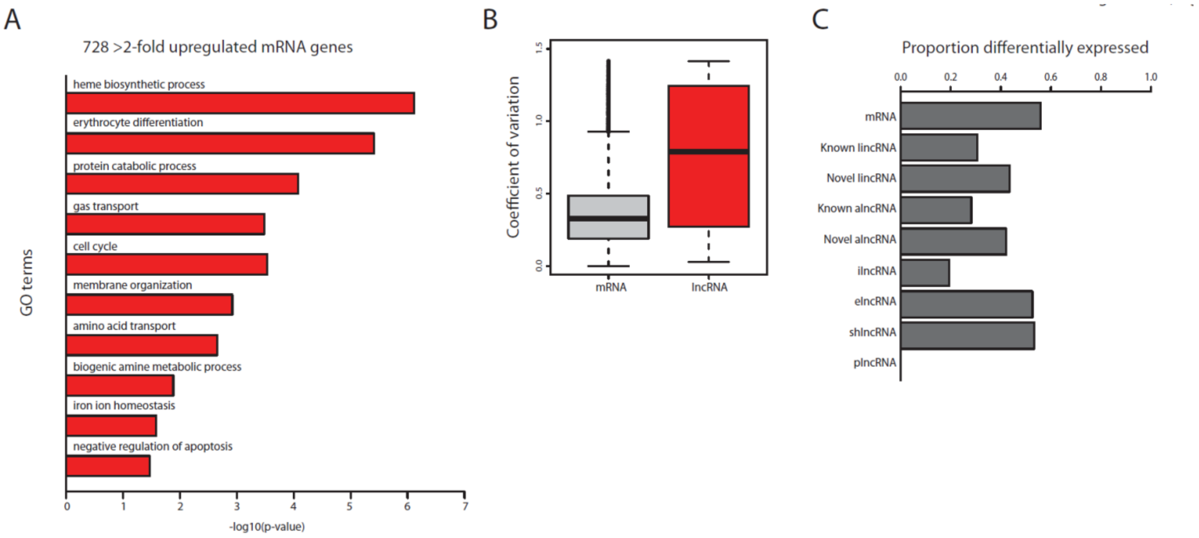
Supplemental Figures



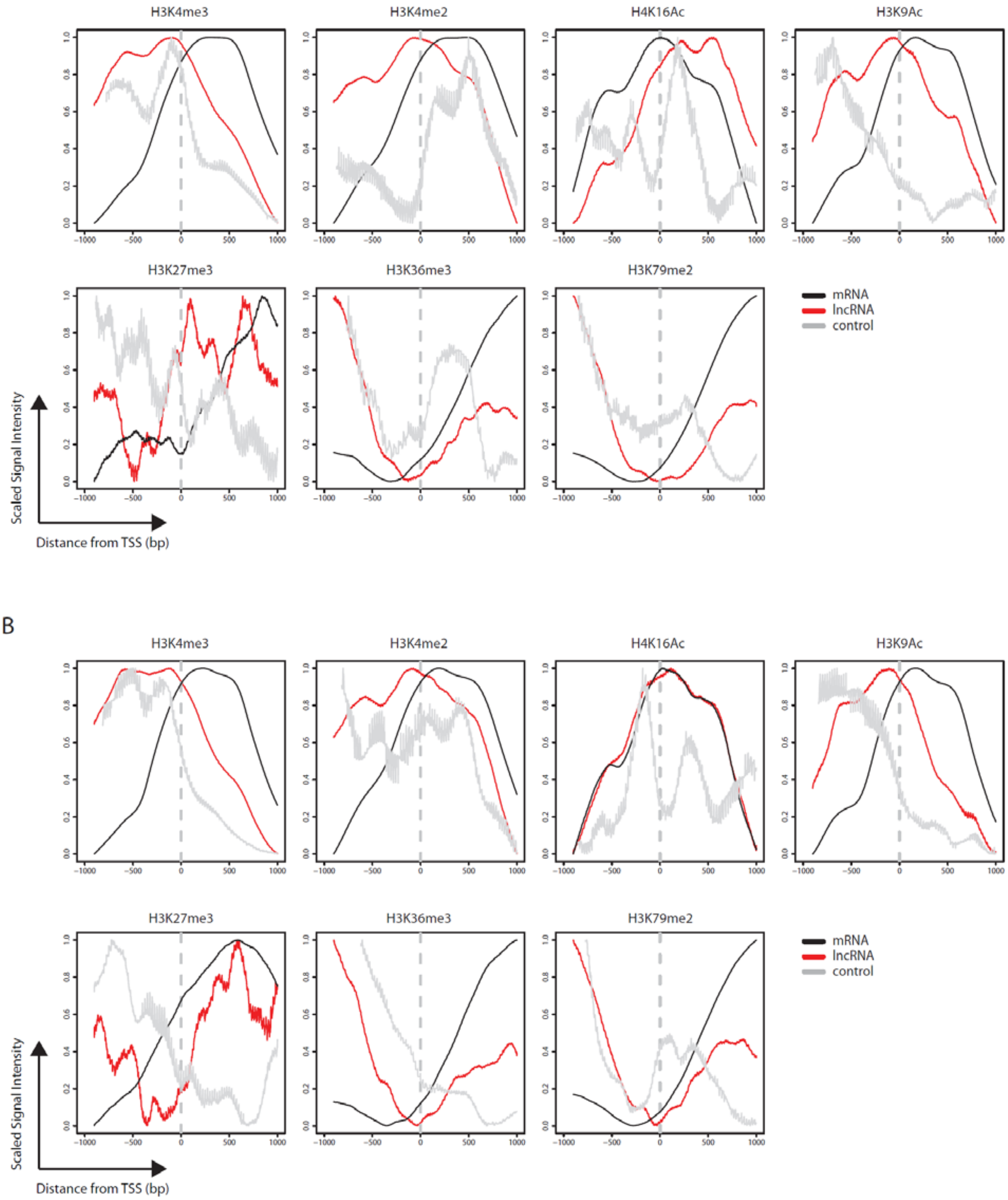
Supplemental Figure 1. Identification and classification of fetal liver-expressed lncRNAs



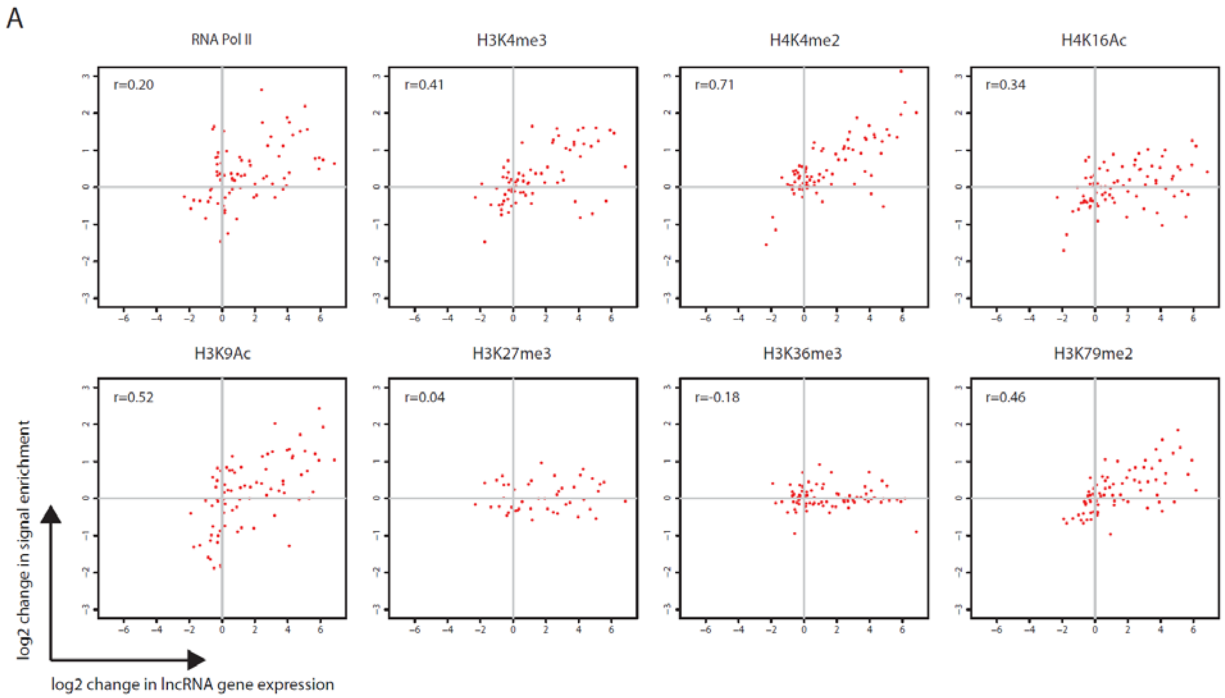
Supplemental Figure 2. Tissue specificity of fetal liver and erythroid lncRNAs



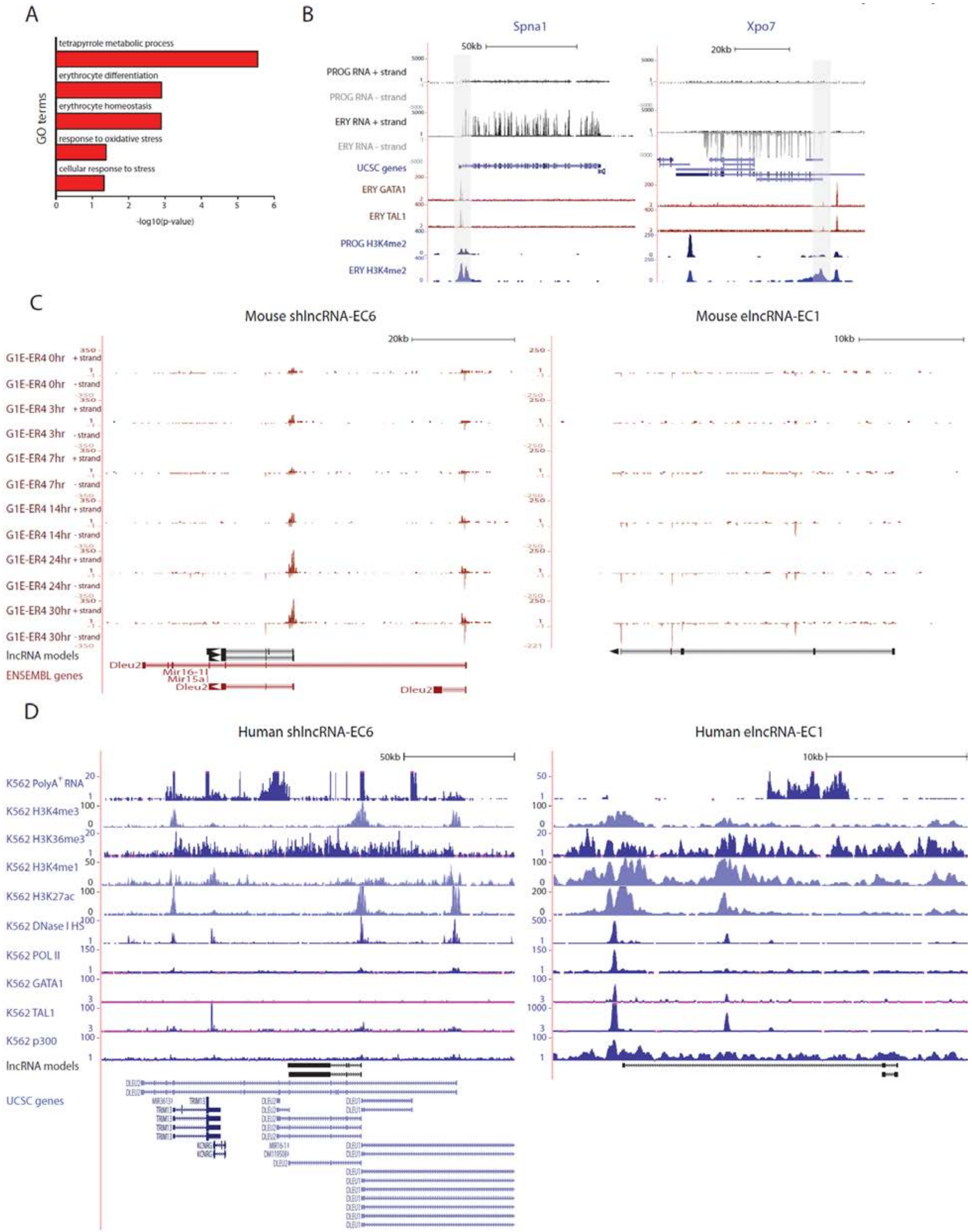
Supplemental Figure 3. Expression features of mRNAs and lncRNAs during erythroid differentiation



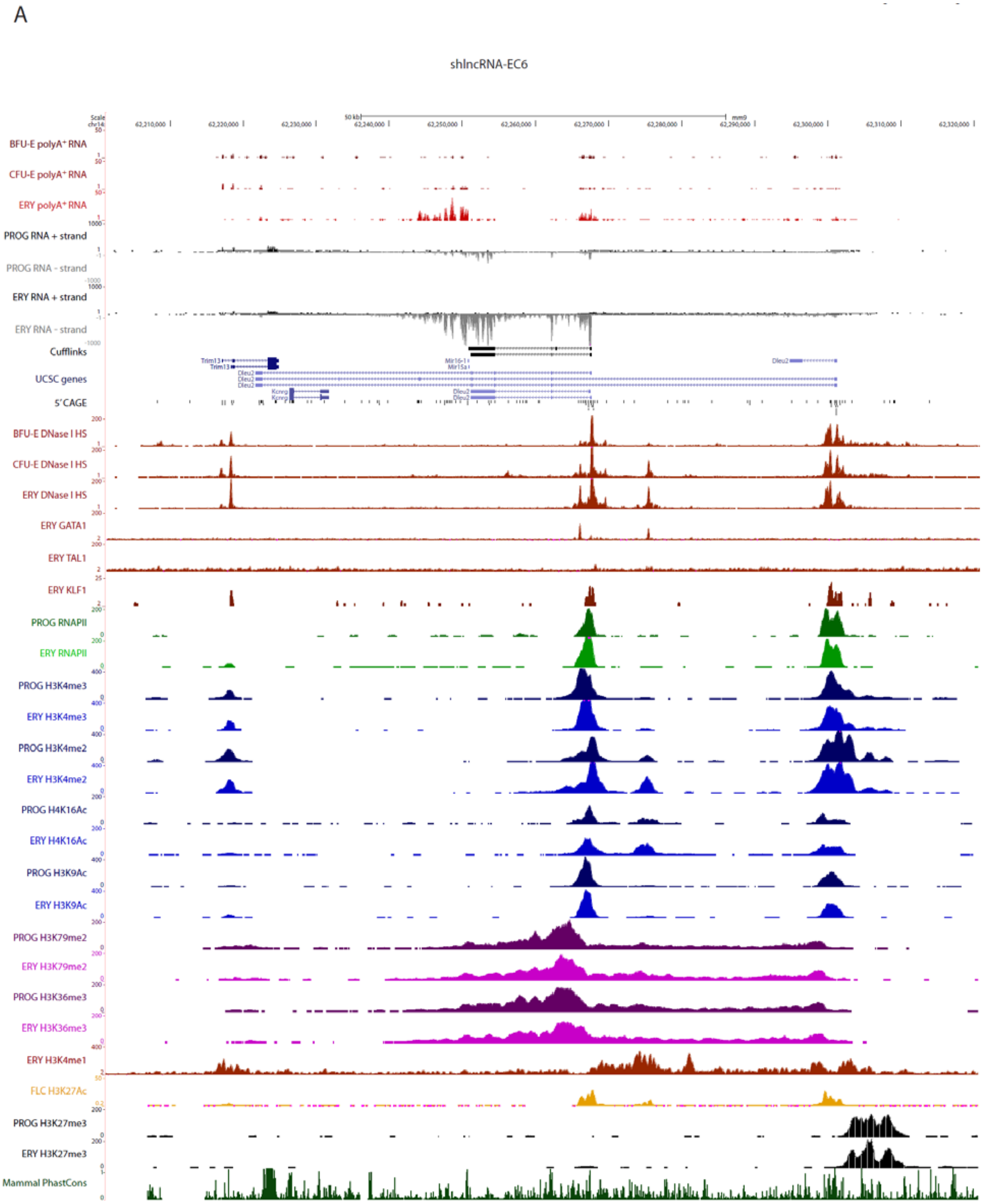
Supplemental Figure 4. Chromatin signatures around the TSS of mRNAs and lncRNAs that are dynamically expressed during erythroid differentiation



Supplemental Figure 5. Correlation between gene expression and chromatin modification changes of lncRNAs that are dynamically expressed during erythroid differentiation

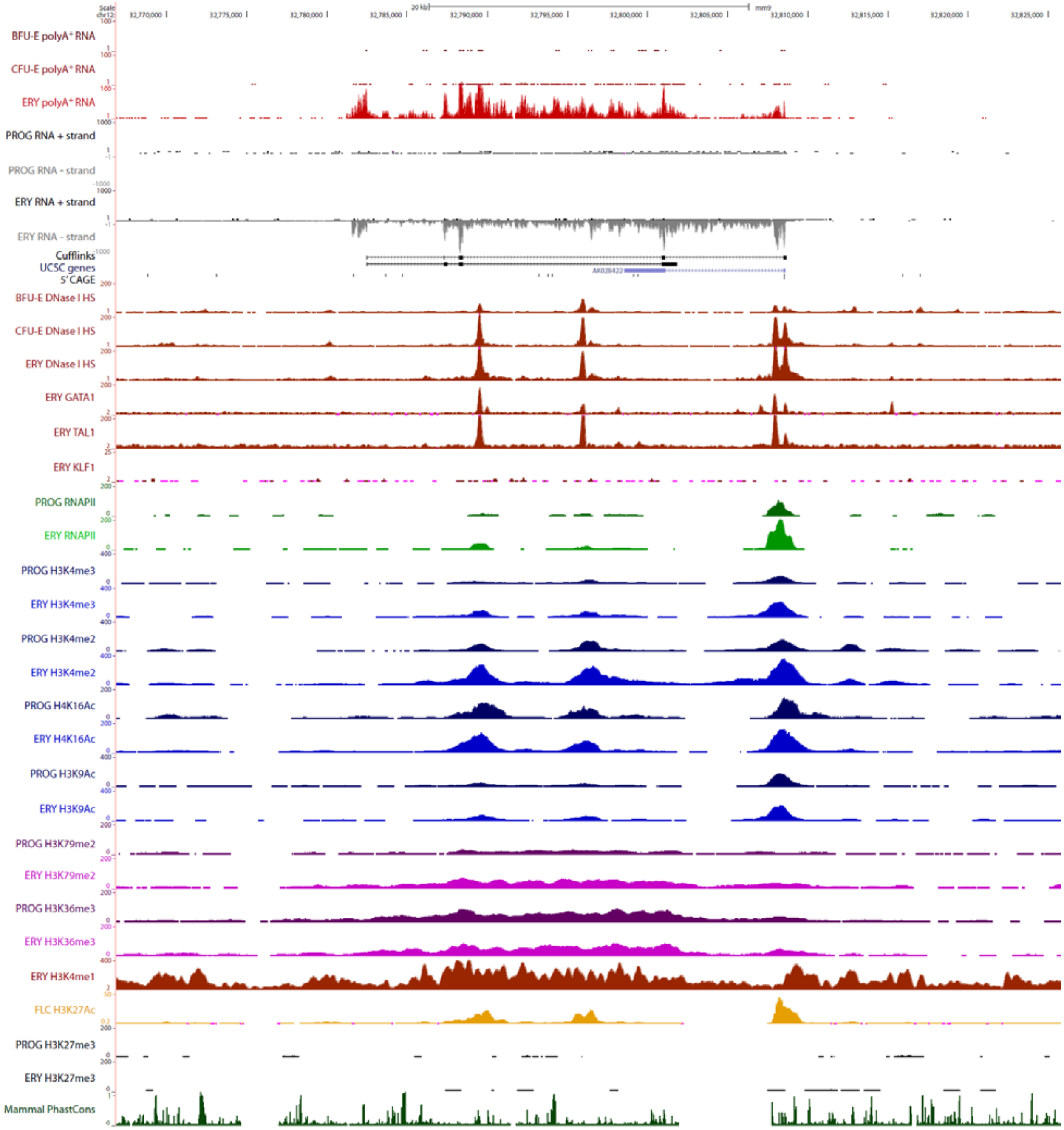


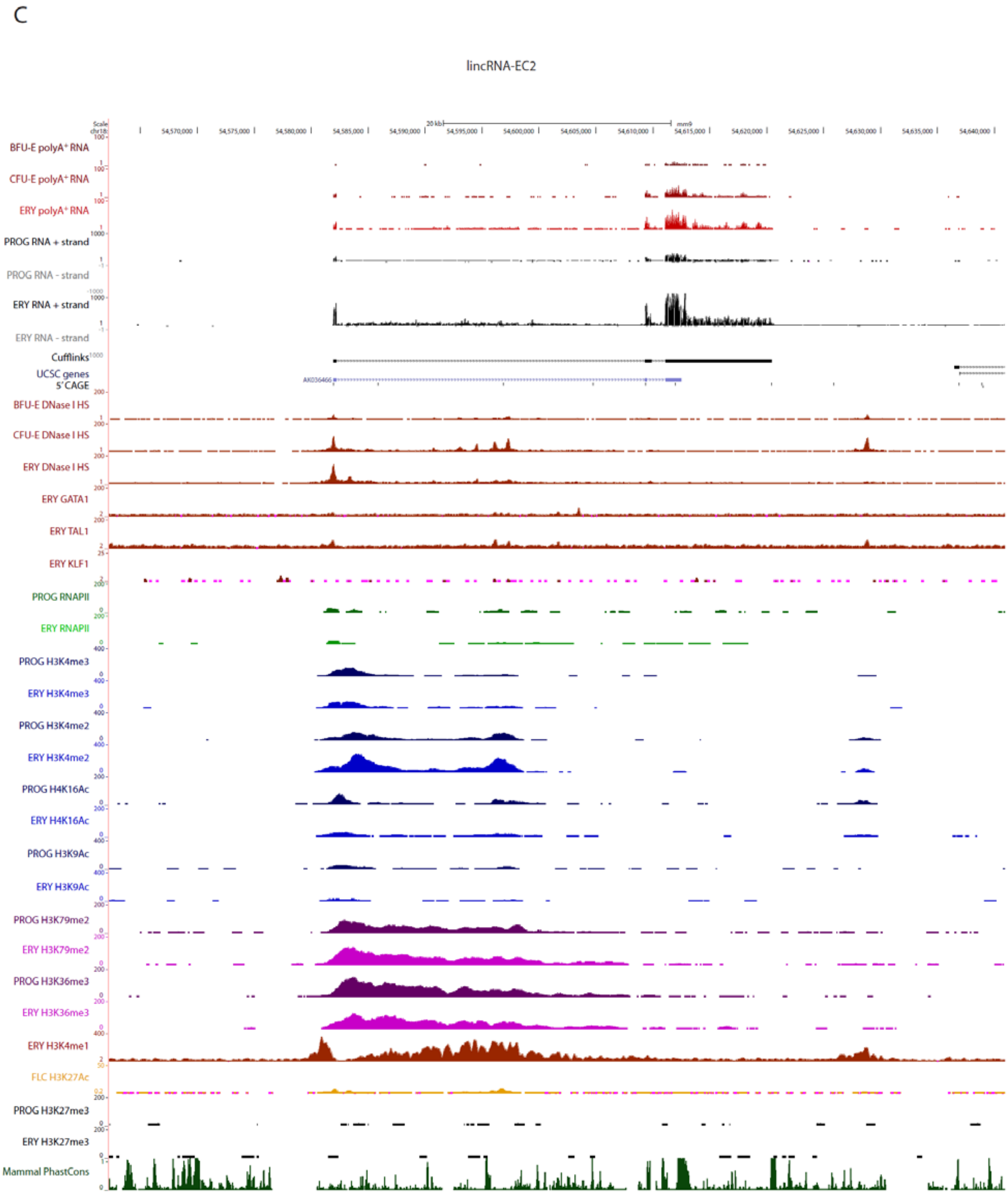
Supplemental Figure 6. Targeting of lncRNAs by key erythroid transcription factors

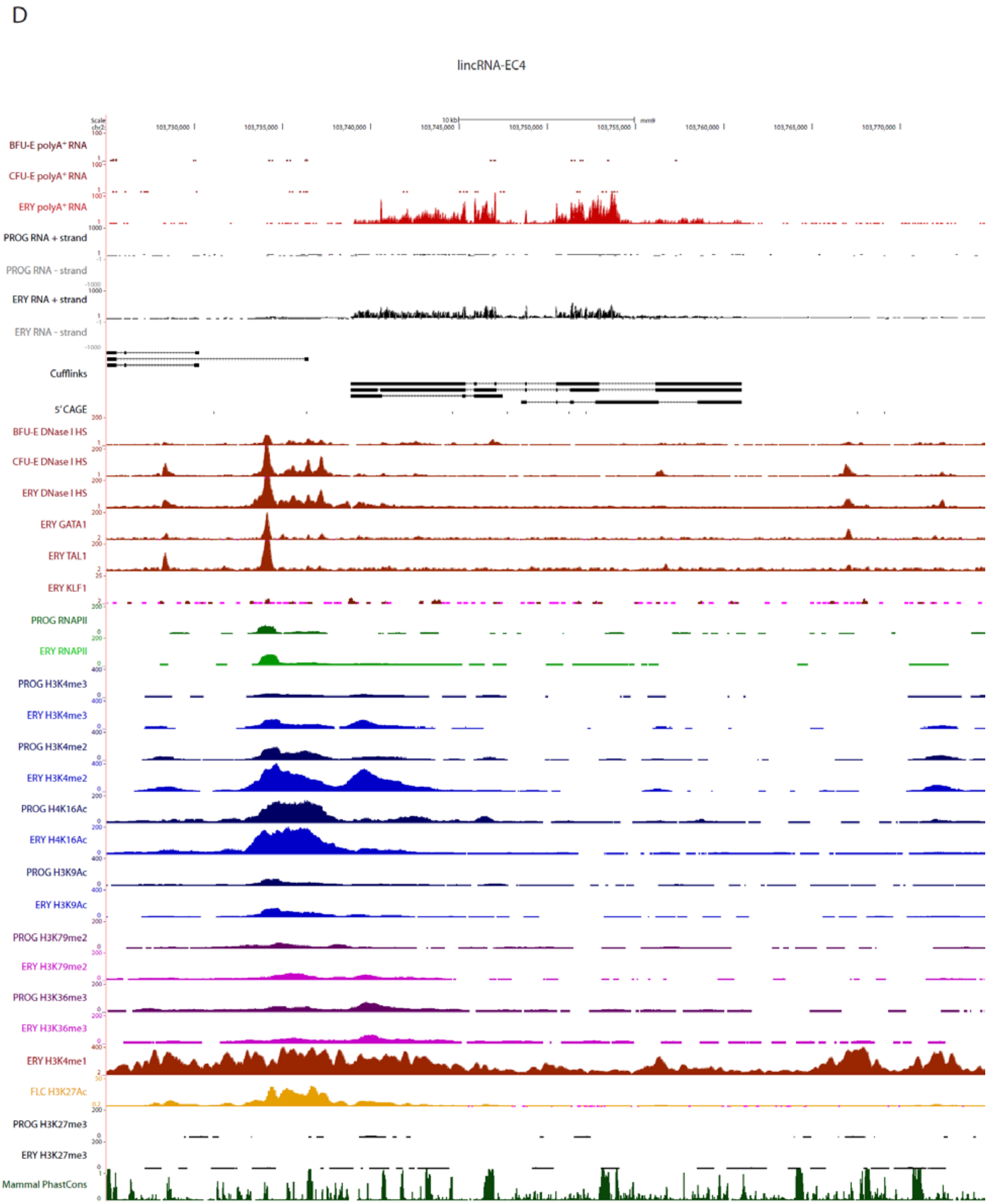

A
shlncRNA-EC6
Scale
chr14:
50 kb
mm9
62,210,000
62,220,000
62,230,000
62,240,000
62,250,000
62,260,000
62,270,000
62,280,000
62,290,000
62,300,000
62,310,000
62,320,000
BFU-E polyA+ RNA
CFU-E polyA+ RNA
ERY polyA+ RNA
PROG RNA + strand
PROG RNA - strand
ERY RNA + strand
ERY RNA - strand
Cufflinks
UCSC genes
5'CAGE
Trim13
Trim13
Dleu2
Dleu2
Dleu2
Kcnrg
Kcnrg
Mir16-1
Mir15a
Dleu2
Dleu2
Dleu2
BFU-E DNase I HS
CFU-E DNase I HS
ERY DNase I HS
ERY GATA1
ERY TAL1
ERY KLF1
PROG RNAPII
ERY RNAPII
PROG H3K4me3
ERY H3K4me3
PROG H3K4me2
ERY H3K4me2
PROG H4K16Ac
ERY H4K16Ac
PROG H3K9Ac
ERY H3K9Ac
PROG H3K79me2
ERY H3K79me2
PROG H3K36me3
ERY H3K36me3
ERY H3K4me1
FLC H3K27Ac
PROG H3K27me3
ERY H3K27me3
Mammal PhastCons

B

elncRNA-EC1

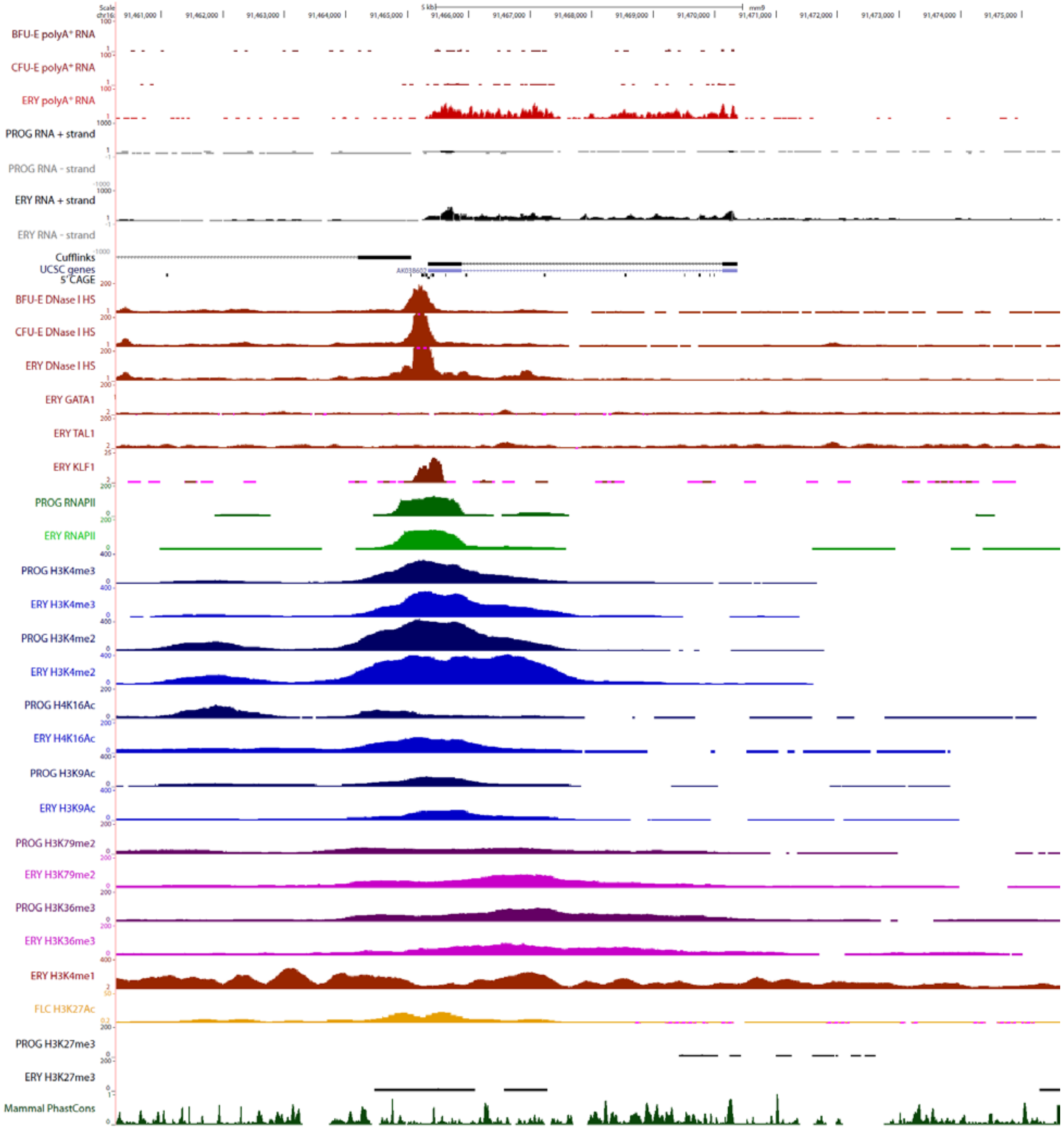


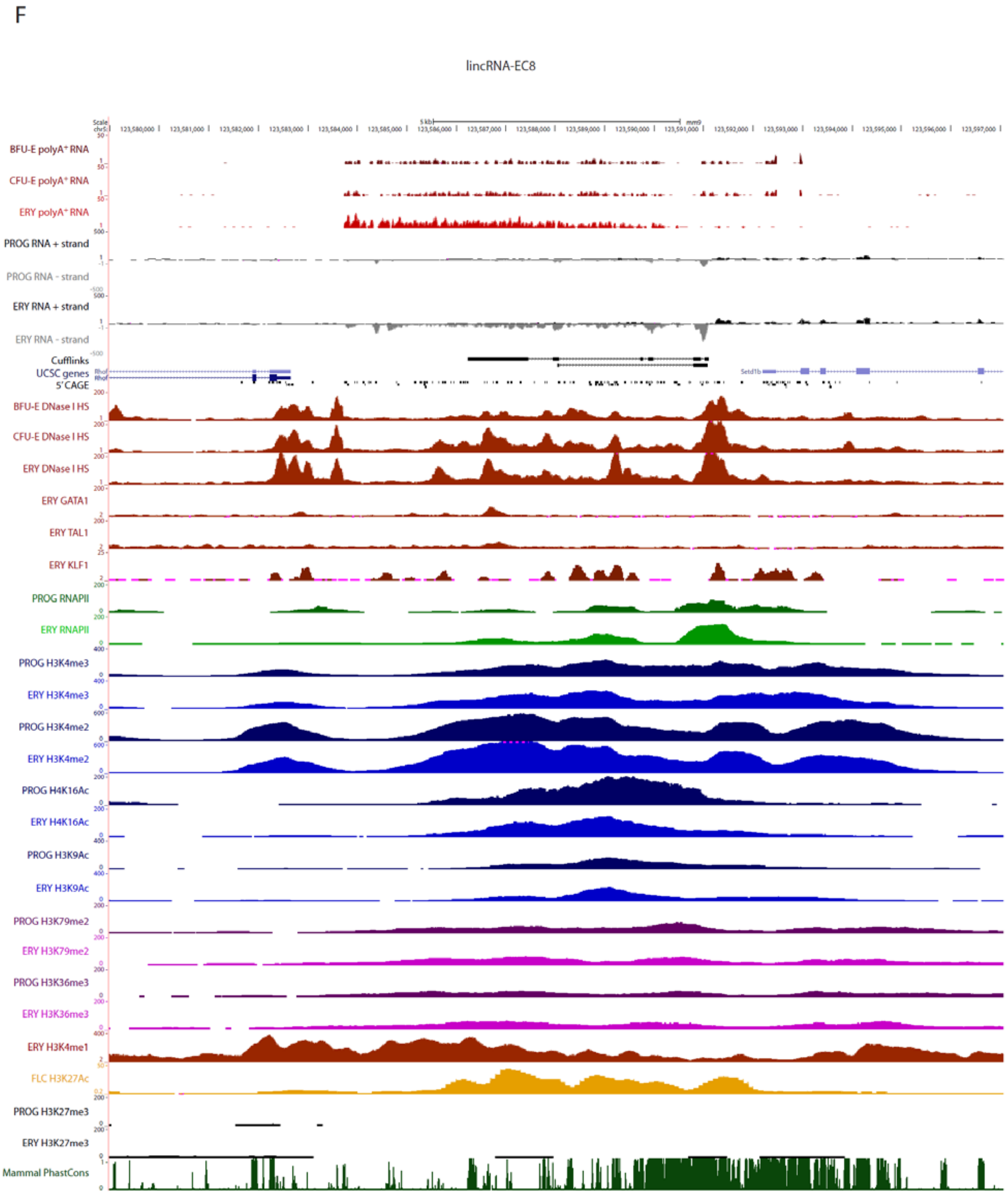

C
lincRNA-EC2
Scale
chr18:
20 kb
mm9
54,570,000
54,575,000
54,580,000
54,585,000
54,590,000
54,595,000
54,600,000
54,605,000
54,610,000
54,615,000
54,620,000
54,625,000
54,630,000
54,635,000
54,640,000
BFU-E polyA+ RNA
CFU-E polyA+ RNA
ERY polyA+ RNA
PROG RNA + strand
PROG RNA - strand
ERY RNA + strand
ERY RNA - strand
Cufflinks
UCSC genes
5′CAGE
AK036466
BFU-E DNase I HS
CFU-E DNase I HS
ERY DNase I HS
ERY GATA1
ERY TAL1
ERY KLF1
PROG RNAPII
ERY RNAPII
PROG H3K4me3
ERY H3K4me3
PROG H3K4me2
ERY H3K4me2
PROG H4K16Ac
ERY H4K16Ac
PROG H3K9Ac
ERY H3K9Ac
PROG H3K79me2
ERY H3K79me2
PROG H3K36me3
ERY H3K36me3
ERY H3K4me1
FLC H3K27Ac
PROG H3K27me3
ERY H3K27me3
Mammal PhastCons


D
lincRNA-EC4
Scale
chr2:
10 kb
mm9
103,730,000
103,735,000
103,740,000
103,745,000
103,750,000
103,755,000
103,760,000
103,765,000
103,770,000
BFU-E polyA⁺ RNA
CFU-E polyA⁺ RNA
ERY polyA⁺ RNA
PROG RNA + strand
PROG RNA - strand
ERY RNA + strand
ERY RNA - strand
Cufflinks
5' CAGE
BFU-E DNase I HS
CFU-E DNase I HS
ERY DNase I HS
ERY GATA1
ERY TAL1
ERY KLF1
PROG RNAPII
ERY RNAPII
PROG H3K4me3
ERY H3K4me3
PROG H3K4me2
ERY H3K4me2
PROG H4K16Ac
ERY H4K16Ac
PROG H3K9Ac
ERY H3K9Ac
PROG H3K79me2
ERY H3K79me2
PROG H3K36me3
ERY H3K36me3
ERY H3K4me1
FLC H3K27Ac
PROG H3K27me3
ERY H3K27me3
Mammal PhastCons

E

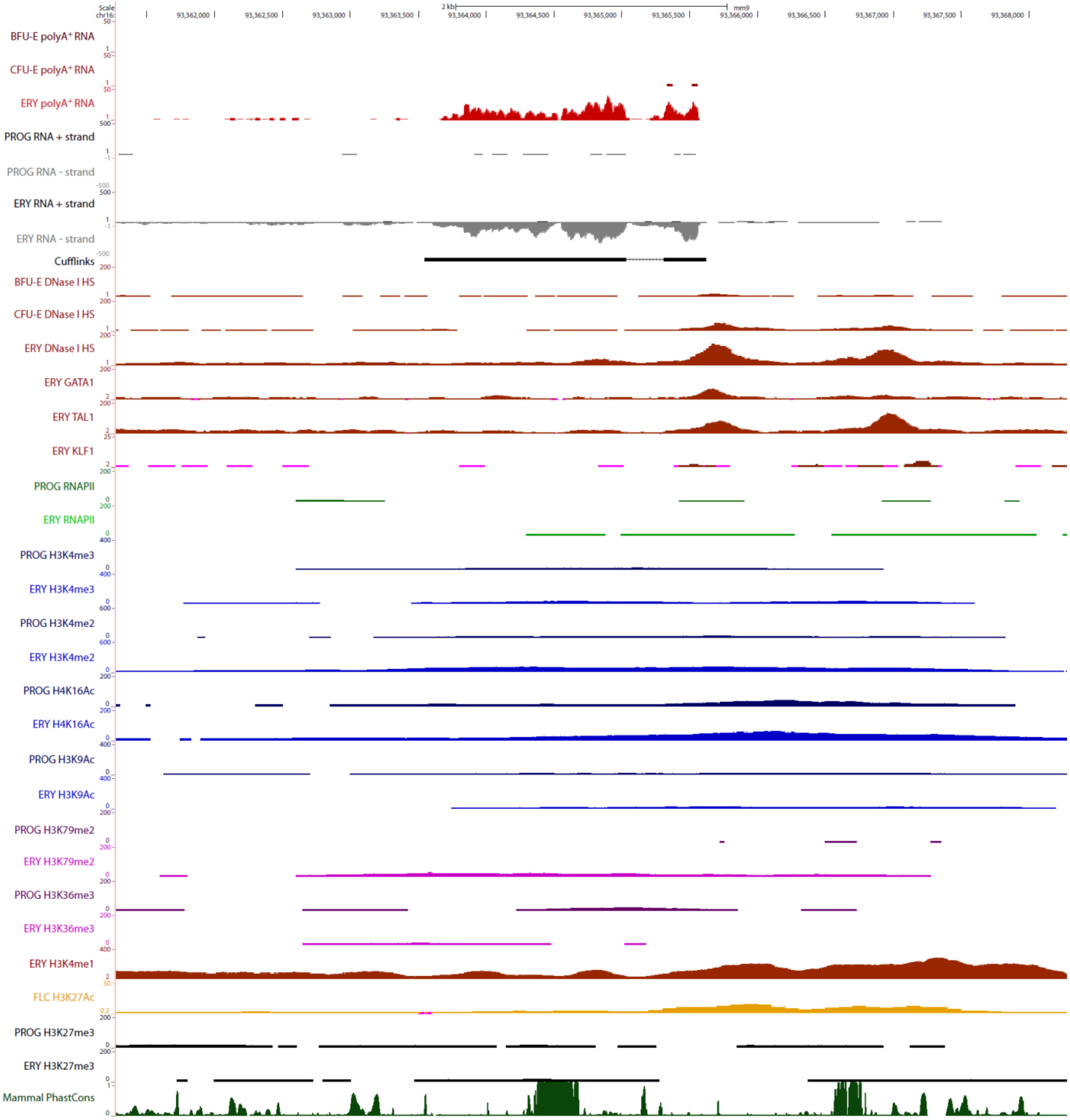
lincRNA-EC5

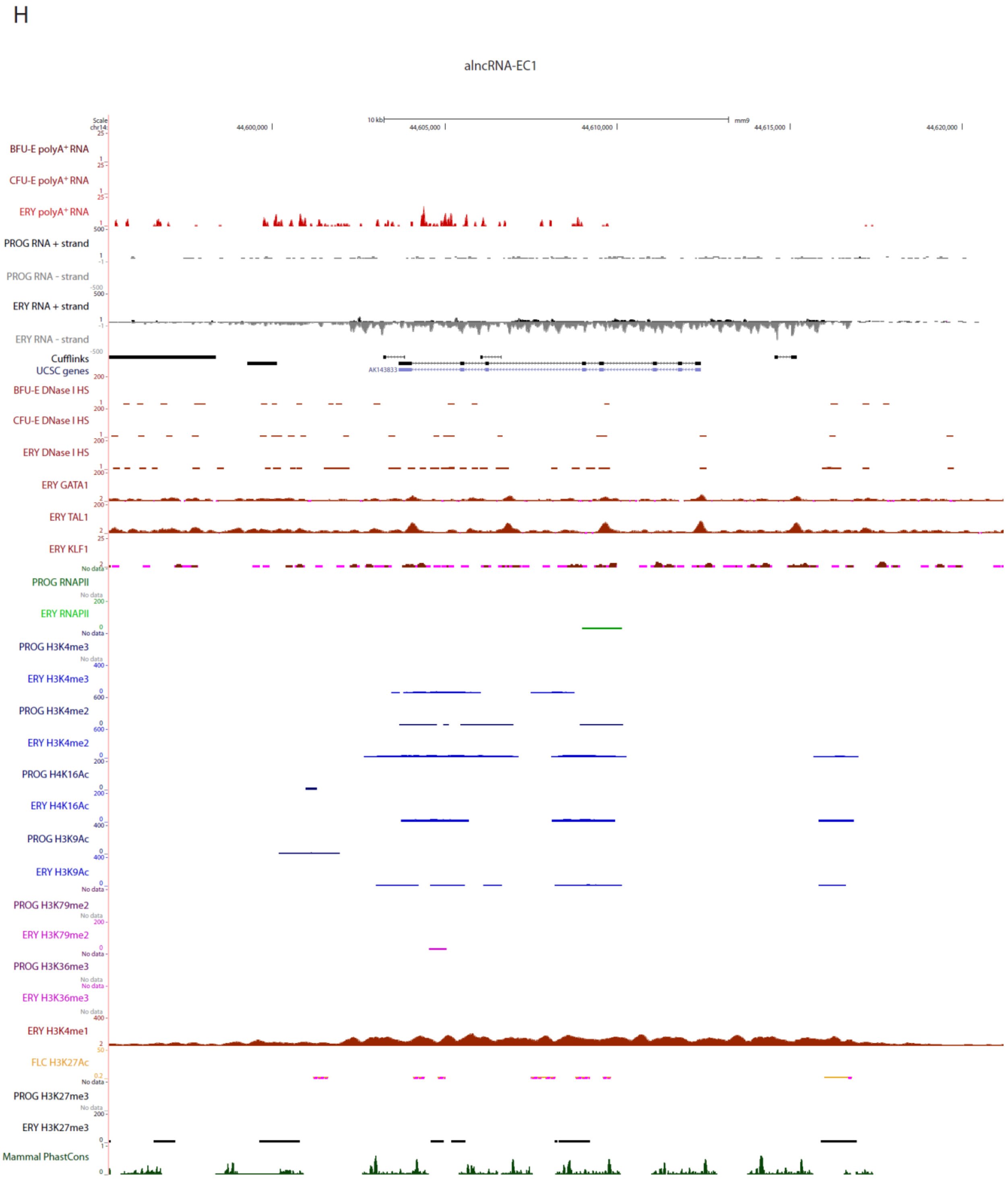



F
lincRNA-EC8
Scale
chr5:
5 kb
mm9
123,580,000
123,581,000
123,582,000
123,583,000
123,584,000
123,585,000
123,586,000
123,587,000
123,588,000
123,589,000
123,590,000
123,591,000
123,592,000
123,593,000
123,594,000
123,595,000
123,596,000
123,597,000
BFU-E polyA⁺ RNA
CFU-E polyA⁺ RNA
ERY polyA⁺ RNA
PROG RNA + strand
PROG RNA - strand
ERY RNA + strand
ERY RNA - strand
Cufflinks
UCSC genes
Rhof
Rhof
Setd1b
5'CAGE
BFU-E DNase I HS
CFU-E DNase I HS
ERY DNase I HS
ERY GATA1
ERY TAL1
ERY KLF1
PROG RNAPII
ERY RNAPII
PROG H3K4me3
ERY H3K4me3
PROG H3K4me2
ERY H3K4me2
PROG H4K16Ac
ERY H4K16Ac
PROG H3K9Ac
ERY H3K9Ac
PROG H3K79me2
ERY H3K79me2
PROG H3K36me3
ERY H3K36me3
ERY H3K4me1
FLC H3K27Ac
PROG H3K27me3
ERY H3K27me3
Mammal PhastCons

G

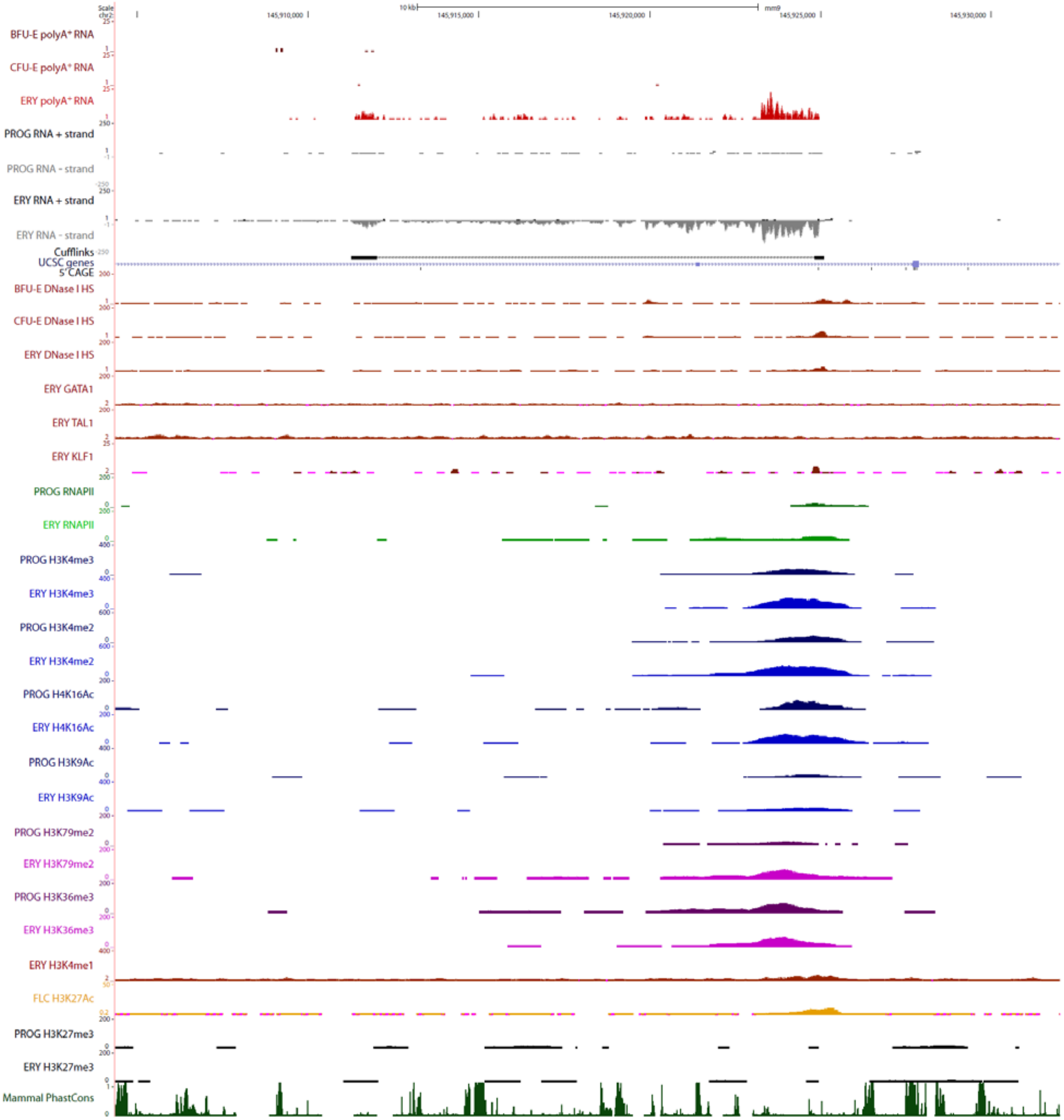
lincRNA-EC9




H
alncRNA-EC1
Scale
chr14:
10 kb
mm9
44,600,000
44,605,000
44,610,000
44,615,000
44,620,000
BFU-E polyA⁺ RNA
CFU-E polyA⁺ RNA
ERY polyA⁺ RNA
PROG RNA + strand
PROG RNA - strand
ERY RNA + strand
ERY RNA - strand
Cufflinks
UCSC genes
AK143833
BFU-E DNase I HS
CFU-E DNase I HS
ERY DNase I HS
ERY GATA1
ERY TAL1
ERY KLF1
PROG RNAPII
ERY RNAPII
PROG H3K4me3
ERY H3K4me3
PROG H3K4me2
ERY H3K4me2
PROG H4K16Ac
ERY H4K16Ac
PROG H3K9Ac
ERY H3K9Ac
PROG H3K79me2
ERY H3K79me2
PROG H3K36me3
ERY H3K36me3
ERY H3K4me1
FLC H3K27Ac
PROG H3K27me3
ERY H3K27me3
Mammal PhastCons
No data

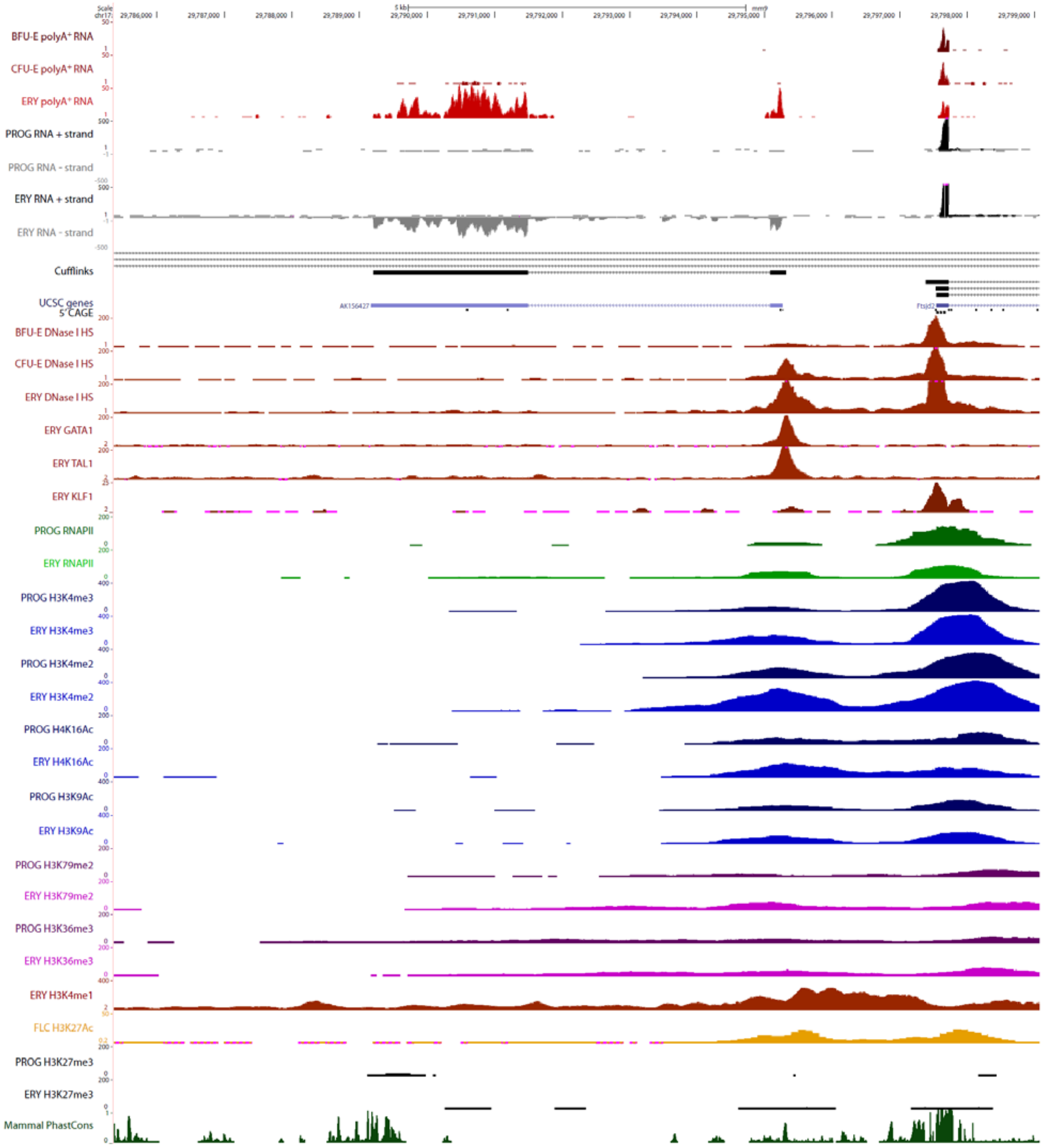
I

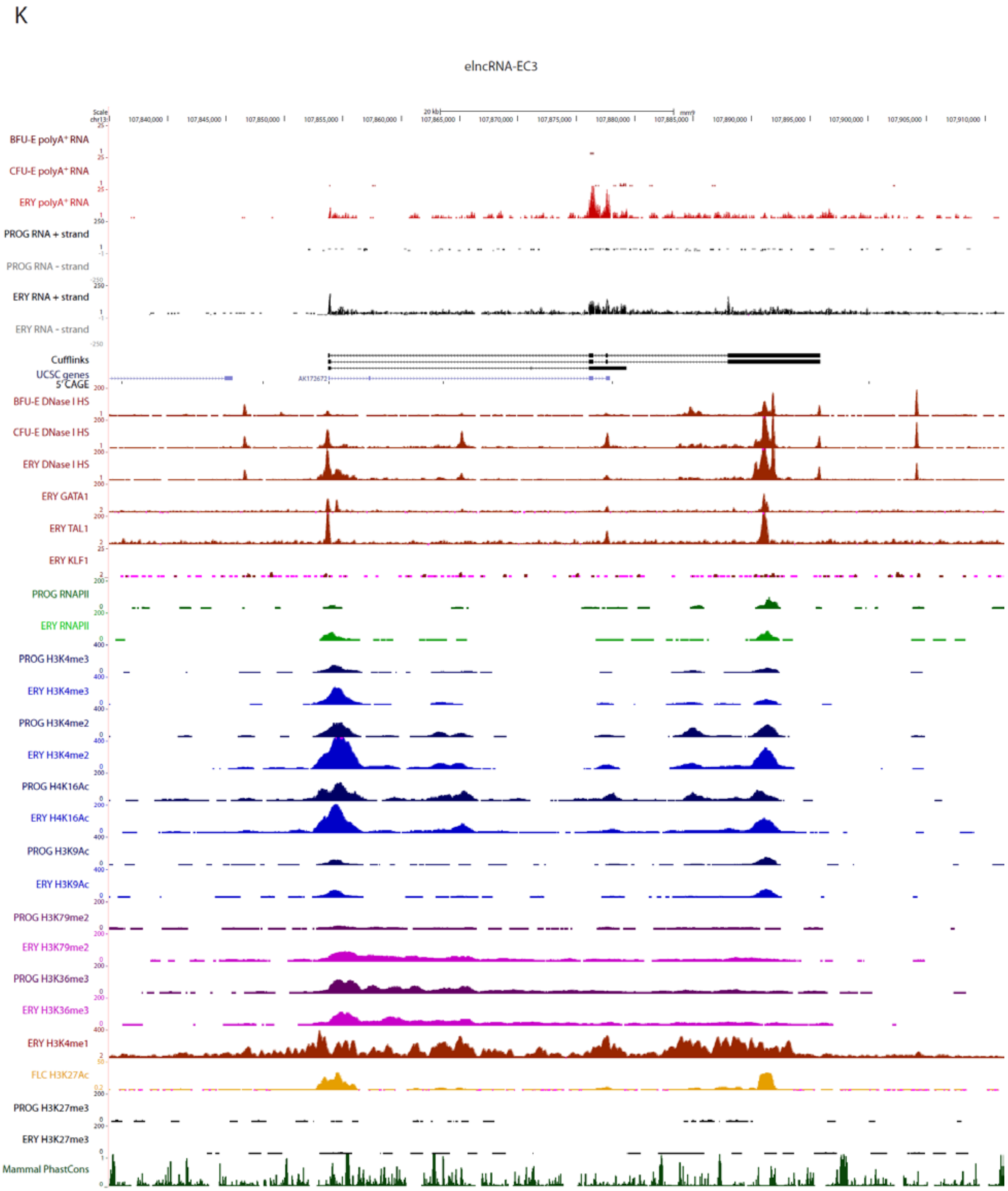
alncRNA-EC2

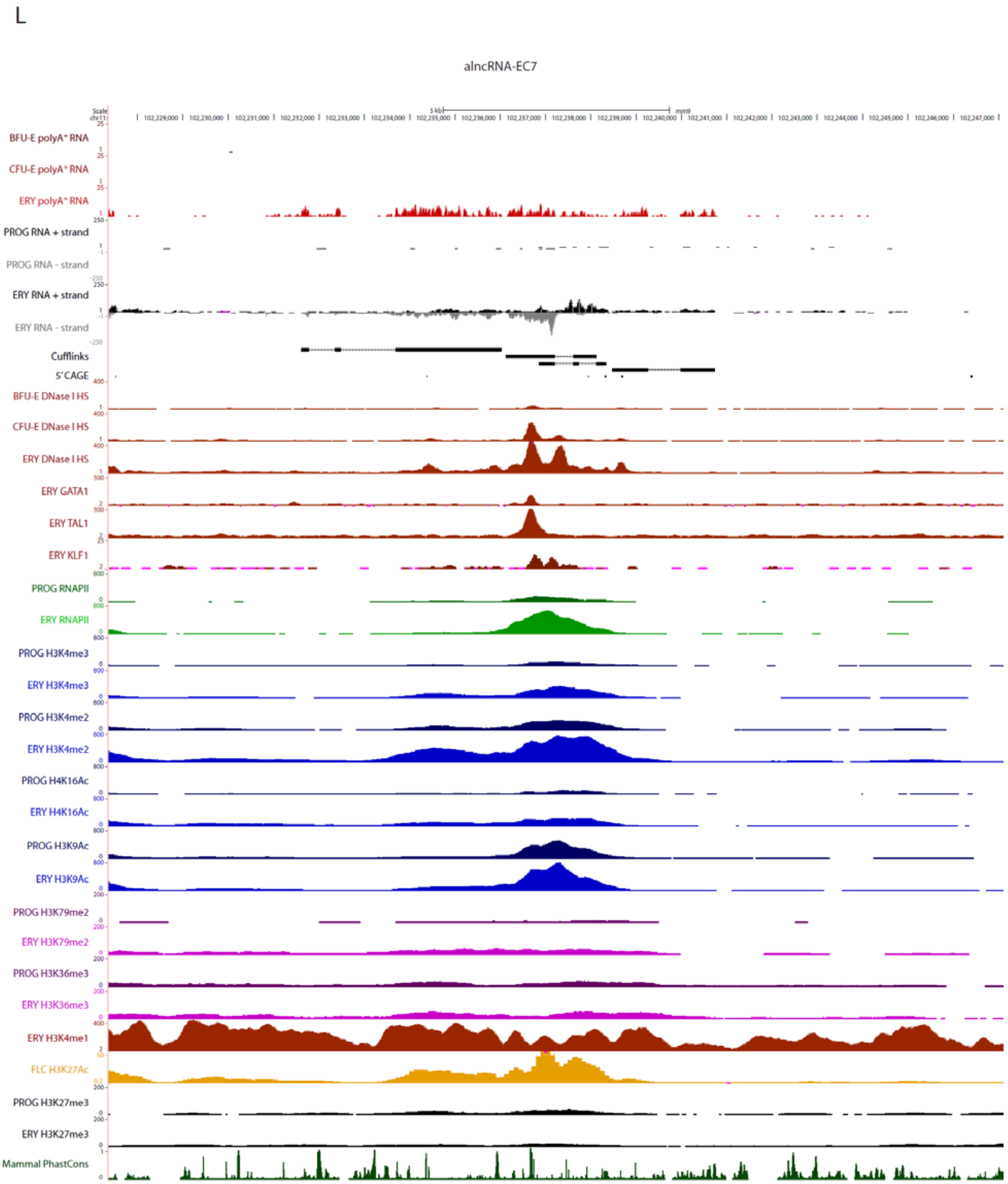


J

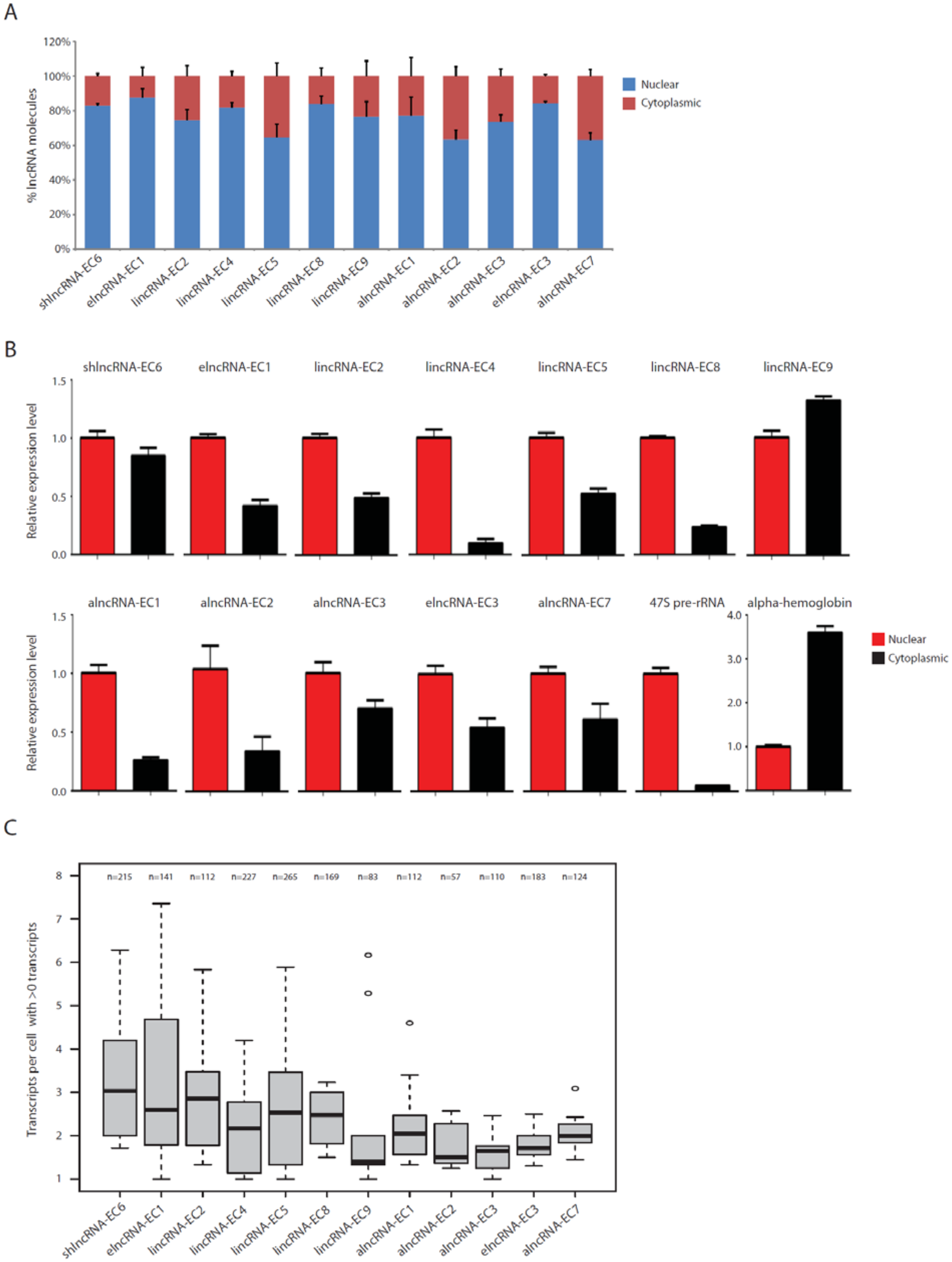
alncRNA-EC3



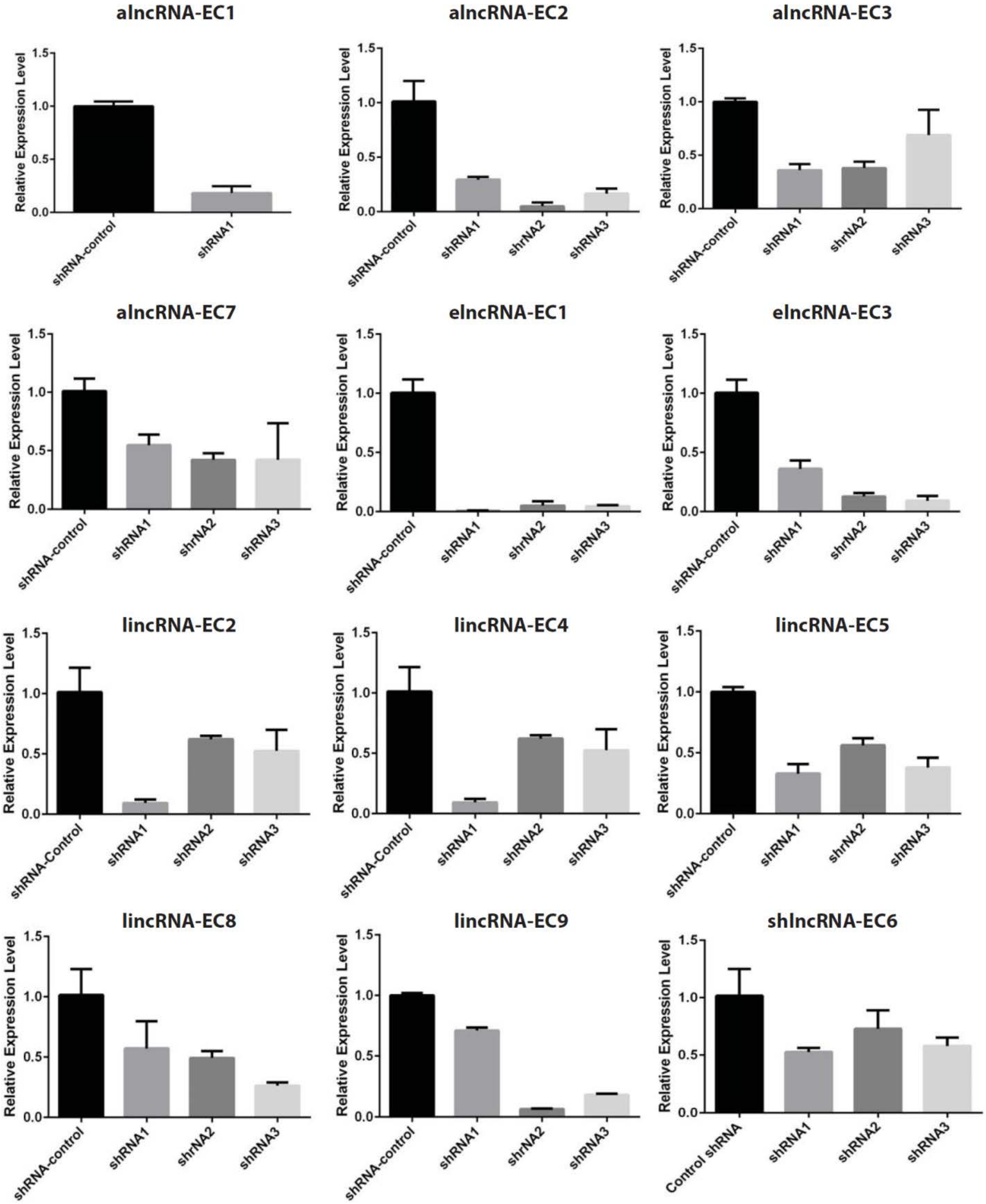

K
elncRNA-EC3
Scale
chr13:
20 kb
mm9
107,840,000
107,845,000
107,850,000
107,855,000
107,860,000
107,865,000
107,870,000
107,875,000
107,880,000
107,885,000
107,890,000
107,895,000
107,900,000
107,905,000
107,910,000
BFU-E polyA+ RNA
CFU-E polyA+ RNA
ERY polyA+ RNA
PROG RNA + strand
PROG RNA - strand
ERY RNA + strand
ERY RNA - strand
Cufflinks
UCSC genes
5'CAGE
AK172672
BFU-E DNase I HS
CFU-E DNase I HS
ERY DNase I HS
ERY GATA1
ERY TAL1
ERY KLF1
PROG RNAPII
ERY RNAPII
PROG H3K4me3
ERY H3K4me3
PROG H3K4me2
ERY H3K4me2
PROG H4K16Ac
ERY H4K16Ac
PROG H3K9Ac
ERY H3K9Ac
PROG H3K79me2
ERY H3K79me2
PROG H3K36me3
ERY H3K36me3
ERY H3K4me1
FLC H3K27Ac
PROG H3K27me3
ERY H3K27me3
Mammal PhastCons



Supplemental Figure 7. Expression, regulation and conservation features of the top 12 candidate lncRNA modulators of erythropoiesis



Supplemental Figure 8. Subcellular localization and single-cell transcript counts of the top 12 candidate lncRNA modulators of erythropoiesis



Supplemental Figure 9. Inhibition of the top 12 candidate lncRNA modulators of erythropoiesis

Supplemental Tables

Sample	Platform	Library	Read length	Mapped reads	Reference
BFU-E	Illumina_GA2x	Long polyA(+), non-directional	2x36bp	26,831,176	Flygare et al., 2011
CFU-E	Illumina_GA2x	Long polyA(+), non-directional	2x36bp	31,662,710	Flygare et al., 2011
Ter119+	Illumina_GA2x	Long polyA(+), non-directional	2x36bp	30,007,165	Flygare et al., 2011
Ter119-	Illumina_HiSeq_2000	Long total RNA minus rRNA, directional	2x100bp	196,350,605	This study
Ter119+	Illumina_HiSeq_2000	Long total RNA minus rRNA, directional	2x100bp	252,864,867	This study
Ter119+	Illumina_HiSeq_2000	Long poly(A)+ RNA, directional	2x100bp	300,637,701	This study
Ter119+	Illumina_HiSeq_2000	Long poly(A)- RNA, directional	2x100bp	228,930,450	This study

Supplemental Table 1. List of RNA-Seq datasets used in this study

Sample	Platform	Library	Read length
Adrenal	Long polyA(+), directional	2x76D	Illumina_GA2x
Brown Adipose Tissue	Long polyA(+), directional	1x36	Illumina_HiSeq_2000
B-cell_(CD19+)	Long polyA(+), directional	1x50	ABI SOLiD
B-cell_(CD43-)	Long polyA(+), directional	1x50	ABI SOLiD
Bladder	Long polyA(+), directional	2x101D	Illumina_HiSeq_2000
Bone Marrow Derived Macrophage	Long polyA(+), directional	1x30	Illumina_HiSeq_2000
Bone Marrow	Long polyA(+), directional	1x30	Illumina_HiSeq_2000
Cerebellum	Long polyA(+), directional	2x101D	Illumina_HiSeq_2000
Cerebrum	Long polyA(+), directional	1x50	ABI SOLiD
Colon	Long polyA(+), directional	2x76D	Illumina_GA2x
Erythroblast Rep1	Long polyA(+), directional	2x99D	Illumina_HiSeq_2000
Erythroblast Rep2	Long polyA(+), directional	2x99D	Illumina_HiSeq_2000
Embryonic Stem Cell	Long polyA(+), directional	1x30	Illumina_HiSeq_2000
Heart	Long polyA(+), directional	2x76D	Illumina_GA2x
Kidney	Long polyA(+), directional	2x76D	Illumina_GA2x
Liver	Long polyA(+), directional	2x76D	Illumina_GA2x
Lung	Long polyA(+), directional	2x76D	Illumina_GA2x
Mammary Gland	Long polyA(+), directional	2x76D	Illumina_GA2x
Mouse Embryonic Fibroblast	Long polyA(+), directional	1x36	Illumina_GA2
Megakaryocyte	Long polyA(+), directional	2x99D	Illumina_HiSeq_2000
Megaryocyte Erythroid Progenitor	Long polyA(+), directional	2x99D	Illumina_HiSeq_2000
Ovary	Long polyA(+), directional	2x76D	Illumina_GA2x
Placenta	Long polyA(+), directional	2x101D	Illumina_HiSeq_2000
Skeletal Muscle	Long polyA(+),	1x50	ABI SOLiD

	directional		
Spleen	Long polyA(+), directional	2x76D	Illumina_GA2x
Stomach	Long polyA(+), directional	2x76D	Illumina_GA2x
Subcutaneous Fat Pad	Long polyA(+), directional	2x76D	Illumina_GA2x
Testis	Long polyA(+), directional	2x76D	Illumina_GA2x
Thymus	Long polyA(+), directional	2x76D	Illumina_GA2x
T-Naive	Long polyA(+), directional	1x50	ABI SOLiD
Whole Brain	Long polyA(+), directional	2x101D	Illumina_HiSeq_2000

Supplemental Table 3. List of mouse ENCODE cell and tissue types used in this study

Appendix B: Supplementary information for chapter 2

This work represents a manuscript in preparation by the following authors:

Juan R. Alvarez-Dominguez, Wenqian Hu, Marko Knoll, and Harvey F. Lodish

Supplemental Experimental Procedures

Data sources

The mouse July 2007 (NCBI37/mm9) genome assembly was used throughout the study. Ensembl transcript structures and annotations were obtained from Ensembl version 67 (http://useast.ensembl.org/info/data/ftp/). RefSeq and UCSC transcript structures and annotations were obtained from the UCSC genome browser (July, 2012). We analyzed poly(A)-selected RNA-Seq reads from 30 primary cell and tissue types (Alvarez-Dominguez et al. 2014)aligned to the mouse genome (mm9 version). RNA-Seq reads of poy(A)+ and poly(A)- RNA isolated from whole-cell, cytosol and nucleus, as well as total RNA isolated from whole-cell, nucleoplasm and chromatin fractions of K562 cells, aligned to the human genome (hg19 version), were accessed through the UCSC genome browser (March 2013). We also examined previously published density maps of ChIP-Seq signal enrichment for histone modifications (H3K4me3, H3K4me2, H4K16Ac, H3K9Ac, H3K27me3, H3K36me3, H3K79me2) and for serine 5 phosphorylated RNA polymerase II in mouse fetal liver erythroid TER119-negative and TER119-positive cells(Wong et al. 2011) deposited in GEO (accession number GSE32111). Density maps of processed ChIP-Seq signal enrichment for GATA1 and TAL1(Wu et al. 2011), and for KLF1(Pilon et al. 2011), in TER119+ erythroblasts, were obtained from the mouse ENCODE portal (http://genome.ucsc.edu/ENCODE/downloadsMouse.html) and from the PSU genome browser (http://main.genome-browser.bx.psu.edu), respectively. Density maps of sequencing tags from DNAse I hypersensitive sites in mouse BFU-Es, CFU-Es, and TER119+ erythroblasts were accessed through the mouse and human ENCODE tracks in the UCSC genome browser

(November 2012). K562 ChIA-PET interactions associated with Pol II and CTCF, and associated with Pol II, RAD21, H3K4me1/2/3 and H3K27Ac (Heidari et al. 2014) were accessed through the WashU Epigenome Browser (http://epigenomegateway.wustl.edu/) and downloaded from GEO (accession# GSE59395; Oct 2014), respectively. Hi-C and topological domain data in human and mouse embryonic stem cells, and in the mouse cortex (Dixon et al. 2012) were downloaded from GEO (accession#GSE35156; Oct 2014) and visualized through the WashU Epigenome Browser (http://epigenomegateway.wustl.edu/).

RNA-seq analysis

Gene-level and isoform-level expression was estimated from the mapped reads from RNA-seq data as fragments per kilobase of exon model per million mapped fragments (FPKM) using Cufflinks based on previously assembled gene and isoform models (Alvarez-Dominguez et al. 2014). To quantify differential gene expression, mapped reads were assigned to pre-defined transcript models using HTSeq-count (http://www-huber.embl.de/users/anders/HTSeq/doc/index.html) with the "union" mode and differentially expressed genes were identified using the DESeq R Bioconductor package(Anders and Huber 2010). DESeq controls for the variation in the number of reads obtained across samples by conducting count-based normalization. After normalization, fold changes and their significance (*p*-values), indicating differential expression, are determined using a model based on the negative binomial distribution. To select a significance threshold, we benchmarked the DESeq results against a set of 475 protein-coding genes known to be induced during terminal erythroid differentiation(Wong et al. 2011). Based on this analysis, we considered all genes with a p-value < 0.05 to be differentially expressed.

Gene Ontology analysis

Gene lists were analyzed for enrichment of Gene Ontology (GO) terms using DAVID (Huang da et al. 2009b; Huang da et al. 2009a). Only Biological Process terms (GOTERM_BP_FAT) were considered. To identify the top non-redundant GO terms, we grouped them using the Functional Annotation Clustering tool, and then selected the most significant informative term (P <0.05) within each of the top clusters, ranked by their enrichment score.

Ingenuity Pathway Analysis

Significantly upregulated genes (P <0.05, DESeq) EC6-depleted cells vs. control cells were analyzed for discovery of regulatory networks using Ingenuity Pathway Analysis (IPA, Ingenuity Systems, Inc., Redwood City, CA). Only experimentally observed regulatory relationships were considered for network generation. We used the Upstream Regulator Analysis tool to identify upstream regulators that may be responsible for the gene expression changes in the dataset. This analysis seeks to identify upstream regulators and predict whether they are activated or inhibited given the observed expression changes of their downstream targets, without taking into account expression of the upstream regulators themselves. We focused our analysis on transcription regulators for which an activation state prediction could be generated, and ranked them based on the p-value generated by IPA for their overlap with the expected causal effects on their targets.

We then inspected the mechanistic network predicted for each upstream regulator. Molecule shapes in the network indicate: ligand-dependent nuclear receptor (rectangle), transcription regulator (ellipse), enzyme (rhombus), transporter (trapezoid), kinase (inverted triangle), and other (circle).

Gene set enrichment analysis

Gene set enrichment analysis (GSEA (Subramanian et al. 2005)) was performed using default parameters and “-metric log2_Ratio_of_Classes”, “-permute gene_set”, “-nperm 10000”. We focused our analysis on curated or pre-ranked gene sets with nominal empirical FDR q-value <0.05. Enrichment of curated gene sets was analyzed using protein-coding genes differentially expressed (P <0.05, DESeq) in EC6-depleted cells vs. control as the expression dataset. The sets of genes significantly upregulated or downregulated in EC6 KD cells were analyzed for enrichment against the erythroid differentiation gene signature previously published (Alvarez-Dominguez et al. 2014), pre-ranked by the log2 expression change between differentiated erythroblasts and lineage negative progenitors.

Supplemental Figure Legends

Supplemental Figure 1. Locus map of DLEU2 and corresponding isoforms

Shown is RNA-Seq signal as the density of mapped strand-specific RNA-Seq reads. Tracks show in red the strand-specific RNA-Seq signal in the plus or minus strands (denoted to the left of the tracks) of poly(A)- and poly(A)+ RNA from fetal liver TER119+ erythroblasts. The bottom tracks depict relevant *de novo* transcript models by Cufflinks and UCSC gene annotations. Left-to-right arrows indicate transcripts in the plus strand; right-to-left arrows indicate transcripts in the minus strand.

Supplemental Figure 2. Flow cytometry analysis of EC6 KD cells

Erythroid progenitor-enriched fetal liver cells were transduced with retroviral vectors encoding control or EC6-targeting shRNAs and induced to differentiate in culture. (A-C) Representative flow cytometry analysis of Annexin V / propidium iodide staining for apoptotic / necrotic cells (A), TER119 and DAPI staining for enucleated cells (B) and their size distributions (C).

Supplemental Figure 3. Topological architecture of 13q14 assayed by HiC

Locus map of a topological domain encompassing the 13q14 region. Top tracks depict the relevant genes in the region. Bottom tracks depict normalized chromatin interaction frequencies assayed by HiC in IMR90 cells (top) and H1 cells (bottom), displayed as 2D heatmaps (Dixon et al. 2012).

Supplemental Figure 4. Chromatin interaction analysis at 13q14 by ChIA-PET

Locus map of a 1.4 Mb domain within 13q14 containing chromatin-chromatin interactions between the DLEU2 locus and several neighboring protein-coding genes. Top tracks depict the relevant genes in the region. Bottom tracks depict chromatin interactions associated with binding of the indicated factors or with the indicated chromatin modifications in K562 cells, determined by ChIA-PET (Heidari et al. 2014).

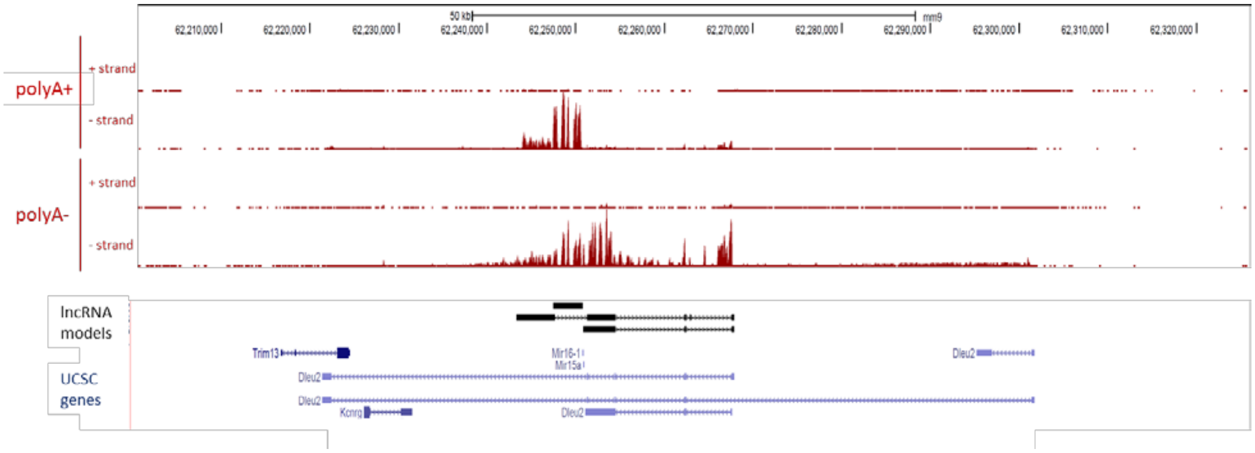
Supplemental Figure 5. Gene set enrichment analysis of genes differentially expressed upon EC6 KD

(A) Genes upregulated in EC6 KD cells are involved in promoting apoptosis and developmental programs associated with lymphoid cell development and function. Graphical data represent enrichment scores across the genome-wide transcriptional profile (~9000 genes). Genes that are higher-expressed in EC6 KD cells are presented in red, whereas lower-expressed ones are shown in blue. Number of members in the gene set, normalized enrichment scores and their FDR values are indicated.

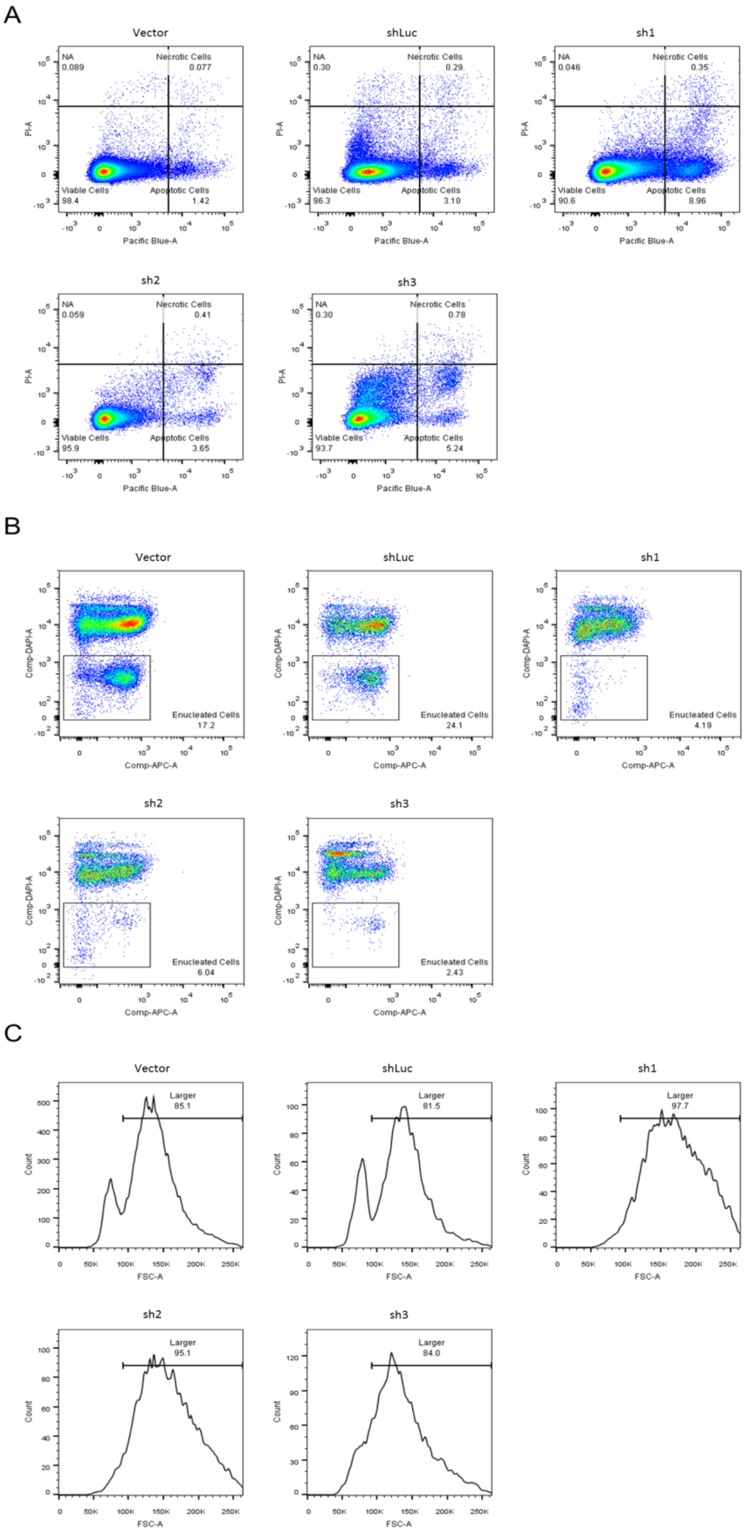
(B) Genes downregulated in EC6 KD cells are enriched for general roles in cell growth and proliferation. Expression estimates and GSEA analysis as in (A) but for downregulated genes.

(C) Genes that promote apoptosis downstream of P53 and NF-kB signaling are significantly upregulated in EC6 KD cells. GSEA analysis as in (A).

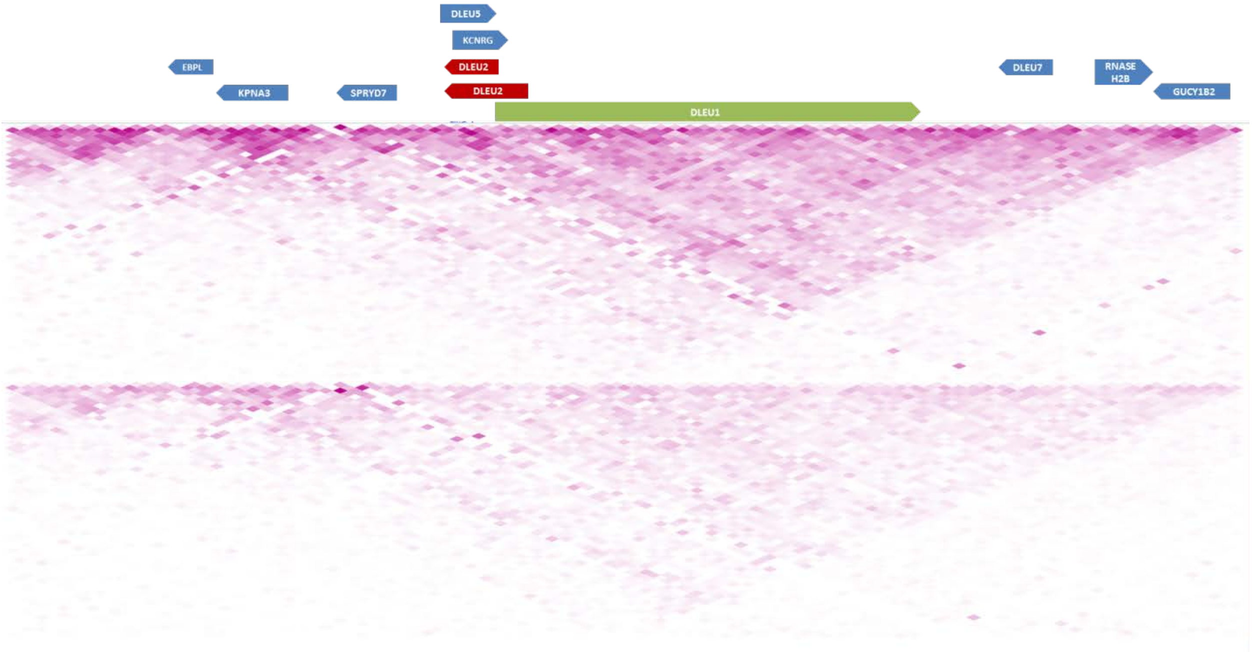
Supplemental Figures



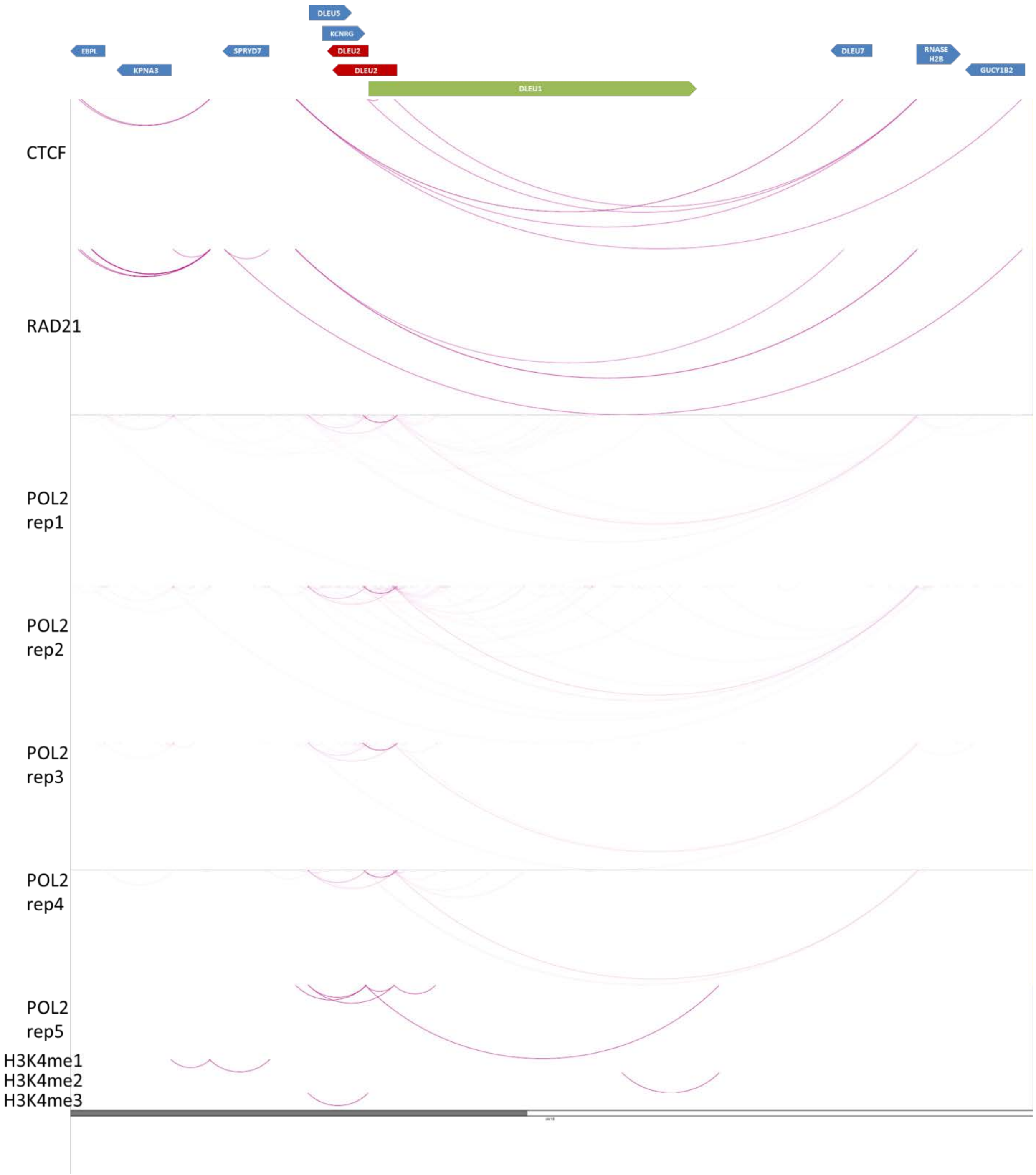
Supplemental Figure 1. Locus map of DLEU2 and corresponding isoforms



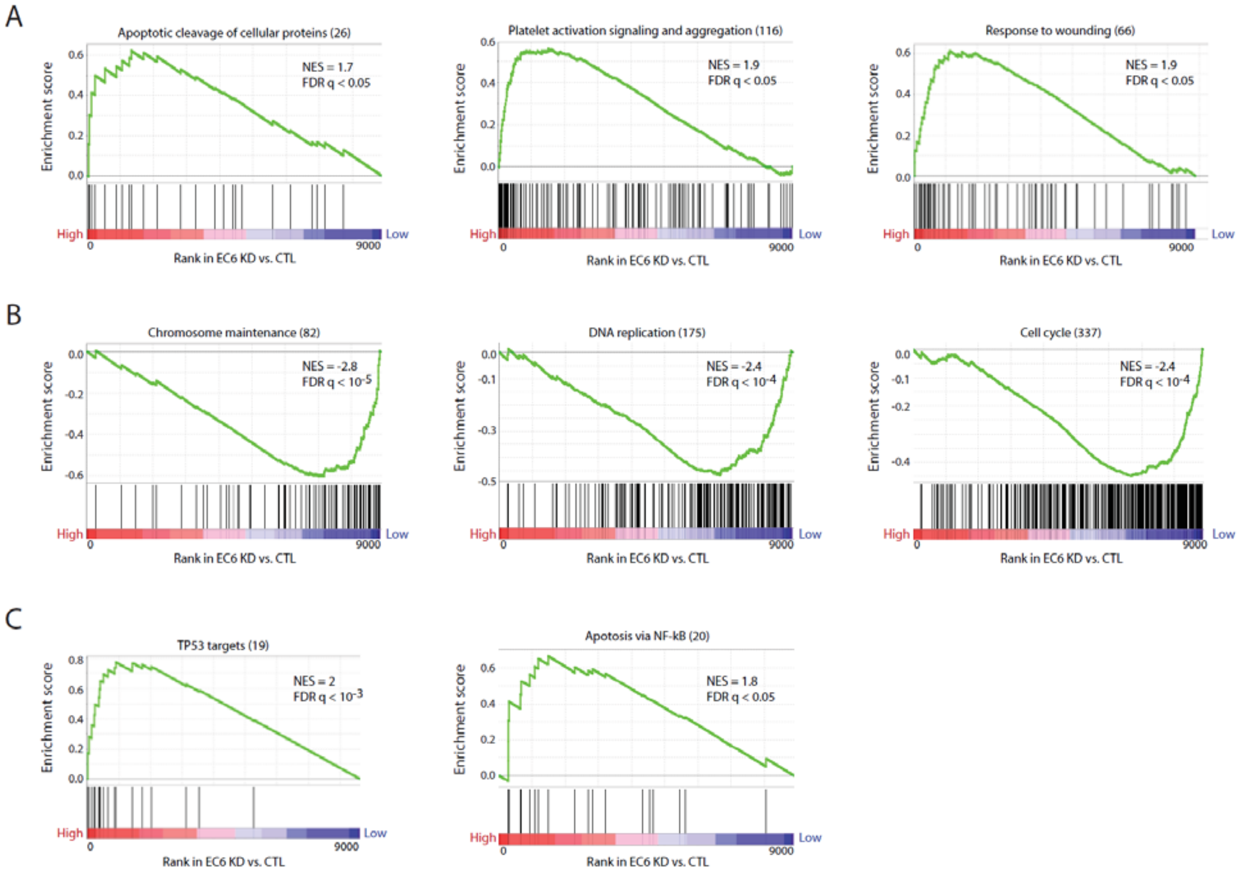
Supplemental Figure 2. Flow cytometry analysis of EC6 KD cells



Supplemental Figure 3. Topological architecture of 13q14 assayed by HiC



Supplemental Figure 4. Chromatin interaction analysis at 13q14 by ChIA-PET



Supplemental Figure 5. Gene set enrichment analysis of genes differentially expressed upon EC6 KD

Mechanistic Network	Upstream Regulator	Predicted Activation State	Activation z-score	p-value of overlap
AKT1,BRCA1,ESR1,HIF1A,MYC,NFkB (complex),NFKB1,NFKBIA,RELA,SP1,STAT1,STAT3,TP53,TP73,WT1,YBX1	TP53	Activated	4.893	3.45E-18
IL12 (complex),IL4,MYC,NFkB (complex),NFKB1,REL,RELA,STAT1,STAT3,STAT4,STAT6	STAT4	Activated	5.067	5.34E-12
CREBBP,GATA1,GFI1B,HDAC1,IRF8,JUN,RUNX1,SPI1	GATA1	Activated	2.933	1.30E-10
APP,AR,CEBPA,CEBPB,EGR1,ESR1,ESR2,FOXO1,HDAC1,HEXIM1,HIF1A,JUN,MYC,NFkB (complex),NFKB1,RARA,RELA,SP1,STAT1,STAT3,TP53,VEGFA	SP1	Activated	3.659	8.69E-10
CEBPB,HIF1A,IFNG,IRF8,NFkB (complex),NFKB1,NFKBIA,REL,RELA,SP1,STAT1,STAT3,TNF,TP53,TP63	RELA	Activated	2.899	6.03E-08

Supplemental Table 1. Top 5 mechanistic networks identified by IPA

References

Alvarez-Dominguez JR, Hu W, Yuan B, Shi J, Park SS, Gromatzky AA, van Oudenaarden A, Lodish HF. 2014. Global discovery of erythroid long noncoding RNAs reveals novel regulators of red cell maturation. *Blood* 123: 570-581.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* 11: R106.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485: 376-380.

Heidari N, Phanstiel DH, He C, Grubert F, Jahanbani F, Kasowski M, Zhang MQ, Snyder MP. 2014. Genome-wide map of regulatory interactions in the human genome. *Genome Res* 24: 1905-1917.

Huang da W, Sherman BT, Lempicki RA. 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37: 1-13.

-. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4: 44-57.

Pilon AM, Ajay SS, Kumar SA, Steiner LA, Cherukuri PF, Wincovitch S, Anderson SM, Mullikin JC, Gallagher PG, Hardison RC et al. 2011. Genome-wide ChIP-Seq reveals a dramatic shift in the binding of the transcription factor erythroid Kruppel-like factor during erythrocyte differentiation. *Blood* 118: e139-148.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102: 15545-15550.

Wong P, Hattangadi SM, Cheng AW, Frampton GM, Young RA, Lodish HF. 2011. Gene induction and repression during terminal erythropoiesis are mediated by distinct epigenetic changes. *Blood* 118: e128-138.

Wu W, Cheng Y, Keller CA, Ernst J, Kumar SA, Mishra T, Morrissey C, Dorman CM, Chen KB, Drautz D et al. 2011. Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration. *Genome Res* 21: 1659-1671.

Appendix C: Supplementary information for chapter 3

This work represents supplementary information for a manuscript in preparation by the following authors:

Juan R. Alvarez-Dominguez[*], **Zhiqiang Bai**[*], Bingbing Yuan, Dan Xu, Kinyui Alice Lo, Nikolai Slavov, Shuai Chen, Harvey F. Lodish, Lei Sun

Note: Supplementary Tables 2 and 4 have been omitted for space considerations

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Data sources

The mouse July 2007 (NCBI37/mm9) genome assembly was used throughout the study. Ensembl transcript structures and annotations were obtained from Ensembl version 67 (http://useast.ensembl.org/info/data/ftp/). RefSeq and UCSC transcript structures and annotations were obtained from the UCSC genome browser (March, 2013). We analyzed previously published poly(A)+ RNA-seq reads from primary brown and white adipocytes and from cultured brown adipocytes and pre-adipocytes (Sun et al. 2013), deposited in the Gene Expression Omnibus (GEO; accession number GSE29898). Mouse ENCODE (Stamatoyannopoulos et al. 2012) mapped reads from RNA-seq of poly(A)-selected RNA in 30 primary cells and tissues, from ChIP-seq for histone modifications (H3K4me3, H3K4me1, H3K27ac) in BAT, and from DNase I hypersensitivity in genital fat pad, were downloaded from the mouse ENCODE portal (http://genome.ucsc.edu/ENCODE/downloadsMouse.html). We also analyzed previously published mapped reads from ChIP-seq for histone modifications (H3K4me1, H3K4me2, H3K27ac), transcription factors (CEBPα, CEBPβ, PPARγ) and for serine 5 phosphorylated RNA polymerase II in brown adipocytes derived from immortalized brown pre-adipocytes (Lee et al. 2013), as well as ChIP-seq reads for PPARγ binding in primary BAT and eWAT (Rajakumari et al. 2013). Mapped CAGE tags from the FANTOM3, FANTOM4 and FANTOM5 projects (Carninci et al. 2005; Ravasi et al. 2010; Consortium et al. 2014) were downloaded from the FANTOM website (http://fantom.gsc.riken.jp/4/download/GenomeBrowser/ucsc/mm9/). Mapped poly(A)-seq tags from the Merck Research Laboratories (Derti et al. 2012) were downloaded from the UCSC genome browser (March 2013).

RNA-seq and analysis

Total RNA from primary mouse interscapular brown adipose tissue, epididymal fat pad, and subcutaneous fat pad was isolated using a QIAGEN kit. Sequencing libraries of poly(A)+ RNA from these samples were prepared using the Solexa kit (Illumina) according to the manufacturer's instructions and sequenced on a Illumina HiSeq2000 sequencer. The resulting 74 bp paired-end reads were quality-checked with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and low quality reads were removed using the fastq_quality_trimmer ("-t 20 -l 25" parameters) from the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Only reads in proper pairs were mapped to mm9 using TopHat v.2.0.4.12 (Trapnell et al. 2009) with default parameters and "--min-anchor 5". The resulting junction files from the BAT, iWAT and eWAT samples were then concatenated and used for another run of TopHat using the --no-novel-juncs parameter. The resulting mapped reads were used to construct *de novo* transcript models using Cufflinks v.2.02 (Trapnell et al. 2010), and the transcript models from each sample were merged using the Cuffmerge utility. We verified that repeating this analysis using only uniquely-mapping reads (86-95% of all reads across our samples) retained 99% of the Cufflinks models identified as lncRNAs by our pipeline, indicating that they do not derive from promiscuous alignment of reads belonging to some other known loci. Transcript- and gene-level expression was quantified by Cufflinks using the Cuffmerge adipose-merged transcriptome, for each fat type and for 30 cell and tissue types from ENCODE. Gene expression in published poly(A)+ RNA-seq data from primary brown and white adipocytes and from cultured brown adipocytes and pre-adipocytes was quantified by Cufflinks using the *de novo* transcript models. Precision of expression estimates was evaluated via resampling using the RPKM_saturation utility from RSeQC (Wang et al. 2012).

lncRNA identification

To retain only reliable transcript models, we considered only multi-exonic transcripts >200bp in length and identified lncRNAs using the following strategy:

1. We implemented a minimal read coverage threshold of >=3 in at least one of the samples. This threshold optimizes the sensitivity and specificity of full length vs. partial length transcript identification based on benchmarking against RefSeq protein-coding or UCSC non-coding gene annotations (Cabili et al. 2011).We verified that coverage value distributions were unbiased across our samples, and obtained similar results using FPKM instead of coverage as a detection threshold.
2. We used BEDTools to intersect *de novo* transcript models with existing transcript models from RefSeq, UCSC and Ensembl, and discarded any transcript overlapping at least 1 bp in the same strand with any annotated mRNA exon, necessary to filter out incompletely assembled mRNAs, unannotated sense-overlapping mRNA isoforms or unannotated 5' or 3' mRNA extensions. We verified our capacity to infer correct strand directionality in our *de novo* transcript models by confirming that 100% of transcripts classified as protein-coding that could be unambiguously assigned to ENSEMBL annotated mRNAs were assigned to the correct strand.
3. For every candidate we retrieved the longest ORF in all three possible frames, using the Sixpack tool from EMBOSS (Rice et al. 2000), and then used HMMER3 (Finn et al.

2011) to query the Pfam A and Pfam B databases (downloaded from ftp://ftp.sanger.ac.uk/pub/databases/Pfam/current_release/ on Nov 2012) with default parameters and discard any transcript with a significant hit (E-value > 0.001). Repeat-masked transcripts were also blasted against the mouse RefSeq protein databases using Blastx (Gish and States 1993), and transcripts mapping with an E-value <0.0001 were removed.

4. We used PhyloCSF (Lin et al. 2011) to filter out any transcript under evolutionary pressure to preserve synonymous amino acid codons, as judged from its sequence alignment across 29 mammalian genomes. We calculated PhyloCSF scores using "--removeRefGaps --frames=3 --orf=ATGStop" parameters and discarded any transcript with a score >100, corresponding to a 9.3% false negative rate and a 9.7% false positive rate using RefSeq mRNAs and RefSeq lincRNAs as reference.
5. We used the Coding Potential Calculator (Kong et al. 2007) to exclude transcripts with characteristic protein-coding features, independent of their conservation. We calculated CPC scores using default parameters and discarded any transcript with a CPC score >0, corresponding to "coding" or "weakly coding" classifications.
6. We discarded any remaining transcript encoding a peptide (see Mass spectrometry analysis).

Ribosome profiling analysis

We assessed the protein-coding capacity of the lncRNAs in our catalog using the ribosome release criteria (Guttman et al. 2013) based on previously published ribosome profiling data from mouse 3T3 cells (Shalgi et al. 2013). Ribo-seq and RNA-seq reads from untreated 3T3 cells aligned to mm9 were downloaded from GEO (accession number GSE32060).We then used these reads to compute the ribosome release score using the RRS program (http://lncrna.caltech.edu/software/RRS.jar; default parameters and –n "true") for known mRNA ORFs or for all predicted ORFs within our lncRNAs or within mRNA 5' UTRs or 3' UTRs as annotated by Ensembl. Predicted ORFs in all three possible frames were identified using the FindORFs utility (http://lncrna.caltech.edu/software/FindORFs.jar; default parameters and –c "false").

Mass spectrometry analysis

To filter out lncRNAs from our catalog encoding proteins, we searched for peptides predicted from their sequences in a previously published deep mass-spectrometry shotgun coverage of murine brown fat (Huttlin et al. 2010). Mass/charge spectra were analyzed using MaxQuant v.1.4.1.2 (Cox and Mann 2008). Searches were performed against *in silico* translated peptides from the lncRNA sequences, against the MaxQuant common contaminants database, and against a database comprising all sequences from the mouse Uniprot/Swiss-Prot database. *In silico* translated peptides were retrieved using Sixpack. All searches were run on a Windows server 2008 64bit operating system with 64 CPU blades and 256 GB of RAM using the following general parameters. Parent ion mass tolerance was set to 20ppm, mass tolerance for MS/MS ions was set to 0.02 Da for HCD and to 0.6 Da for CID spectra, minimal peptide length was specified at 6 amino acids, and peptide charge state was limited to +7. Searches had trypsin enzyme specificity, allowing 2 missed cleavages. Asn and Gln deamidation and Met oxidation were included as variable modifications in the search parameters. All peptides were filtered at 1 %

FDR. Our search identified 21265 high confidence unique peptides mapping to 3275 mouse Swissprot proteins and 12379 shared peptides mapping to thousands of protein groups. This coverage of the mouse proteome validates both the dataset and the ability of our search to identify peptides present in murine brown fat. We found only 4 peptides whose posterior error probability (PEP) did not exceed 5 %, corresponding to *in silico* translated ORFs from lncRNAs in our list that were discarded from further consideration. The small number of identified peptides and their relatively large PEP indicate that our strategy for identifying RNA sequences that are not translated has succeeded.

ChIP-seq analysis

Density maps of signal from ChIP-seq for histone modifications, transcription factors, and for serine 5 phosphorylated RNA polymerase II in cultured brown adipocytes and pre-adipocytes (Lee et al. 2013) were retrieved from GEO (accession number GSE50466). We used UCSC utilities and BEDTOOLS to reconstruct mapped reads from these files, and visualized their enrichment at lncRNA promoter-proximal regions (TSS ± 3 kb) using NGSPLOT (Shen et al. 2014). Read counts at these regions were calculated using HTSeq-count (http://www-huber.embl.de/users/anders/HTSeq/doc/index.html) with the “union” mode, and normalized using DESeq (Anders and Huber 2010). ChIP-seq reads for PPARγ binding in primary BAT and eWAT (Rajakumari et al. 2013) and peak coordinates for BAT- or eWAT-specific binding events were downloaded from GEO (accession number GSE43763), and reads were aligned to mm9 using bowtie2 (Langmead and Salzberg 2012) with default parameters. We used BEDTOOLS to intersect BAT- and eWAT-specific lncRNA TSS ± 3 kb regions with BAT- and eWAT-specific peaks, defined by Rajakumari et al. by performing peak calling for one depot sample as foreground and the other sample as background. To identify common PPARγ binding peaks, we used mapped reads from BAT or eWAT as input to the peak-calling algorithm MACS (Zhang et al. 2008) using default parameters, the resulting peaks were then pooled and overlapping peaks merged using BEDTOOLS, and any peaks overlapping BAT- and eWAT-specific peaks were discarded. Enrichment of PPARγ at the resulting peaks was quantified by HTSeq-count ("union" mode), and normalized using DESeq.

Global lncRNA validation analysis

To globally validate the lncRNAs identified by our RNA-seq pipeline, we sought independent evidence of expression from the following orthogonal experimental sources:

1. EST and CAGE sequencing tags: we used BEDTools to intersect lncRNA exons with overlapping EST or CAGE tags in the same strand, downloaded from the UCSC database (EST track) or from the FANTOM website (FANTOM4 tag clusters).
2. RNAP II ChIP-seq: we intersected lncRNA TSS ± 3kb regions with binding peaks for RNAP II in day 2 cultured brown adipocytes (Lee et al. 2013), inferred by MACS using default parameters.
3. Histone marks ChIP-seq: we intersected lncRNA TSS ± 3kb regions with enrichment peaks for histone modifications (H3K4me3, H3K4me1, H3K27ac) in BAT, downloaded from ENCODE (Stamatoyannopoulos et al. 2012).

Evolutionary conservation analysis

PhastCons conservation scores based on a 30-way genome alignment seeded with the mouse genome (Blanchette et al. 2004) were downloaded from the UCSC genome browser database for mRNA or lncRNA exons, TSS + 1 kb promoter regions, and introns. The conservation score for the exons, TSS + 1 kb promoter region or introns of a given gene was then obtained by aggregating the PhastCons scores along the region and dividing by the region's total length.

Tissue specificity analysis

Specificity of the expression of a given gene to a given tissue was scored as the fraction of the gene's cumulative expression across tissues represented in that tissue (i.e. its fractional expression level). Tissue specificity scores ranged from 1e-10 to 0.98. We chose an empirical cutoff of 0.15 that optimally distinguishes between known tissue-restricted mRNAs and known uniformly expressed ones and between known BAT-enriched mRNAs and general adipogenic markers, and designated genes scoring above this threshold as tissue-specific. Genes designated as tissue-specific that have BAT, iWAT or eWAT as the tissue with the maximal specificity score were considered BAT-, iWAT- or eWAT-specific, respectively. WAT-specific genes comprised both iWAT- and eWAT-specific genes pooled together. To identify adipose-specific genes common to both BAT and WAT (common adipogenic genes), for each gene we calculated the mean expression value across all non-adipose tissues, and subtracted it from the expression value in each of BAT, iWAT, and eWAT. We then retained only genes that for each adipose tissue were at least 75 FPKM units above the mean expression value across all non-adipose tissues, an empirical threshold that optimally classified AdipoQ, PPARγ, C/EBPα, FABP4 and

Lpl, but not known BAT or WAT markers, as common adipogenic. A small number of genes classified as common adipogenic that were also classified as BAT- or WAT-specific by their tissue specificity scores were kept only in the common adipogenic group. We verified remarkable separation of BAT and WAT tissue specificity score distributions in the BAT- and WAT-specific gene groups, as well as near-identical BAT and WAT distributions in the common adipogenic group.

Single-molecule RNA FISH

Two sets of 34 and 48 DNA 20 nt oligonucleotide probes uniquely mapping along the exons or introns of lnc-BATE1, respectively, were designed using the online designer at http://www.singlemoleculefish.com (version 3.0), with a minimum spacing of 2nt and a target GC content of 45%, applying maximum stringency cutoffs to maximize probe specificity and avoid repetitive and low complexity regions. The probe sequences are available upon request. Probes were synthesized with an amine group at the 3'end (Biosearch Technologies) and coupled to Alexa fluor 594 (Invitrogen) or Cy5.5 (GE Amersham). Fluorophore-coupled probes were ethanol precipitated and purified on an HPLC column. Brown adipocytes cultured in chambered cover glasses (Lab-Tek) were fixed with 1-2 ml of 3.7% (v/v) para-formaldehyde, 1x PBS for 10 minutes at room temperature, and permeabilized by incubating at 4°C in 70% ethanol for at least 16hr. Single-molecule RNA FISH was performed in the chambered cover glasses as described previously (Raj et al. 2008). For hybridization to DNA probes, cells were rehydrated in wash buffer containing 25% (v/v) formamide and 2x SSC for 5min, and then 100μl of hybridization solution, containing labeled DNA probes (2 ng/μl final concentration) in 25% (v/v) formamide, 2x SSC, 1mg/ml BSA, 10mM Vanadyl-ribonucleoside complex, 0.5mg/ml E. coli tRNA and 0.1 g/ml dextran sulfate, were added to the sample and incubated overnight at 37°C. Before imaging, cells were washed twice in 25% (v/v) formamide and 2x SSC for 30min at 37°C, with 5ng/ml DAPI added for the second wash for nuclear counterstaining. Fluorescence microscopy, image acquisition and analysis were conducted as described previously (Neuert et al. 2013; Alvarez-Dominguez et al. 2014). For imaging, 200µl of an oxygen-scavenging solution, containing 10mM Tris (pH 7.5), 2x SSC and 0.4% glucose supplemented with 74µg/ml glucose oxidase, 74µg/ml catalase, and 2mM Trolox, were added to the adherent cells. Images were taken with a Nikon TI-E inverted fluorescence microscope using a 100x oil-immersion objective, custom filters designed to distinguish between different fluorophores, and a Photometrics Pixis 1024 CCD camera (Princeton Instruments) managed by the MetaMorph software (Molecular Devices, Downington, PA). Stacks of images were taken automatically with 0.3µm between z-slices in the Differential Interference Contrast (DIC), DAPI, GFP, AF594 and Cy5 channels. For each biological replicate, at least 10 fields view were imaged. For image processing, the maximum projection of DAPI image z-stacks was merged with the DIC z-slice of maximum contrast and the composite image was used to identify individual cells. AF594 or Cy5 images were compared to GFP control images to detect diffraction-limited spots representing individual RNA transcripts using fixed pixel intensity thresholds. Nuclear vs. cytoplasmic localization was determined manually by visual inspection of merged DAPI and AF594 or Cy5 images. For image presentation, enhanced contrast in the DAPI channel was used to emphasize nuclear counterstaining boundaries.

cDNA synthesis and quantitative real-time PCR

Total RNA from tissue or cell samples was isolated as mentioned above. cDNA was made with random or oligo(dT) primers using M-MLV (Promega). Sybr Green based qPCR was performed in an Applied Biosystems 7900HT Fast Real-time PCR System, using RPL23 as an internal control for normalization.

5’ and 3’ RACE

5’ and 3’ RACE were performed using a FirstChoice RLM-RACE Kit (Life technologies) according to the manufacturer’s instructions. The resulting PCR products were separated on a 1% agarose gel. All visible bands were recovered and cloned into a pGEM-T easy vector. The transcription start and end sites of lncBAT-1 were determined by sequencing >10 colonies with inserts from each band.

Oil-Red-O, Hoechst and Mitotracker Staining

ORO staining was performed as described (Sun et al. 2011). For co-staining of Hoechst and Mitotracker, 5-day differentiated brown adipocytes were stained with 100mM Mitotracker Red FM and 1:5000 dilution of Hoechst at 37°C for 40min. Cell images were captured on an Epson perfection v700 photo scanner for Oil-Red-O staining of whole wells, a Nikon digital sight DS-U3 system for Oil-Red-O cell staining with bright field imaging, and a LECIA DMI3000 B Inverted Microscope for Hoechst and Mitotracker co-staining. For Mitotracker cell fluorescence quantification, the 32-bit full range (0-255) values of images were background-subtracted and analyzed using ImageJ (Schneider et al. 2012). Cells were outlined manually and their area, integrated density and mean gray values measured. Distributions of integrated signal density values were compared using a paired t-test for statistical significance. For image presentation, background-subtracted, 32-bit fluorescence and Hoechst channel images were merged using a linear Look up Table.

Western blot

Cells were collected and lysed in RIPA buffer (50mM Tris-Cl.ph7.4, 1% Triton X-100, 150mM NaCl, 1Mm EDTA, 1mM PMSF, protease inhibitor cocktail). Protein samples were separated on a 4-15% TGX gel (Bio-Rad) and transferred onto a PVDF membrane. The membrane was blocked with 3% BSA in 1XTBST at room temperature for 2h, incubated with a primary antibody overnight at 4°C, and then incubated with a horseradish peroxidase-conjugated secondary antibody for 2h at room temperature. Specific bands were revealed with chemiluminescence substrates and recorded with a ChemDoc MP Image System (Bio-Rad). Primary antibody against Ucp1, hnRNP U, SUZ12, GAPDH were purchased from Abcam. Anti-HuR antibody was purchased from Santa Cruz Biotechnology.

lncRNA Knockdown in mature brown adipocytes

Electroporation of DsiRNA against lnc-BATE1 in mature brown adipocytes was performed in a lonzal 4D-Nucleofactor system using the Amaxa SE cell line 4D-Nucleofactor X Kit with some program modifications. Briefly, 2X106 differentiated BAT cells were trypsinized, washed with PBS, and re-suspended in electroporation buffer (82ul nucleofactor solution + 18ul supplemental solution). 100 pmol of DsiRNA was added to cells and gently mixed. Cells were transferred to the electroporation cuvette and electroporated using 3T3-L1 (undifferentiated) program with pulse code CA133. After electroporation, cells were incubated at 37 °C for 10 minutes before adding fresh medium and transferring to a culture plate. Growth medium was changed 12 hours after electroporation.

Analysis of RNA-seq from lnc-BATE1 knockdown cells

Paired-end 100bp reads from siRNA_lnc-BATE1 or siRNA_Control samples were mapped to mm9 using TopHat v.2.0.4.12 (Trapnell et al. 2009) with default parameters and "--min-anchor 5". Counts of reads mapping within gene models from Ensembl were obtained using HTSeq-count (http://www-huber.embl.de/users/anders/HTSeq/doc/index.html) with the "union" mode, and normalized using DESeq (Anders and Huber 2010). Only genes for which a read count could be estimated in each of the samples were considered. Differentially expressed genes ($P < 0.05$ fold-change between lnc-BATE1 and control siRNA, DESeq) in either differentiation stage were used for downstream analysis. Similar downstream analysis results were obtained by using Cuffdiff v2.1.1 instead of DESeq for differential gene expression analysis.

Gene Ontology analysis

Gene lists were analyzed for enrichment of Gene Ontology (GO) terms using DAVID (Huang da et al. 2009b; Huang da et al. 2009a). Only Biological Process and Molecular Function terms (GOTERM_BP_FAT and GOTERM_MF_FAT) were considered. To identify the top non-redundant GO terms, we grouped them using the Functional Annotation Clustering tool, and then selected the most significant informative term (P <0.05) within each of the top clusters, ranked by their enrichment score.

Ingenuity Pathway Analysis

Differentially expressed protein-coding genes (P <0.05, DESeq) with lower expression in lnc-BATE1-depleted cells vs. control were analyzed for discovery of regulatory networks using Ingenuity Pathway Analysis (IPA, Ingenuity Systems, Inc., Redwood City, CA). Only experimentally observed direct regulatory relationships were considered for network generation. We used the Upstream Regulator Analysis tool to identify upstream regulators that may be responsible for the gene expression changes in the dataset. This analysis seeks to identify upstream regulators and predict whether they are activated or inhibited given the observed expression changes of their downstream targets, without taking into account expression of the upstream regulators themselves. We focused our analysis on transcription regulators for which an activation state prediction could be generated, and ranked them based on the p-value generated by IPA for their overlap with the expected causal effects on their targets. We then generated a regulatory network based on the known relationships between the top candidate regulators actually present in the dataset and the known target molecules in the dataset used to identify them. The measured expression changes of both regulators and targets in lnc-BATE1-depleted cells vs. control cells were then overlaid on the network. Molecule shapes in the network indicate: ligand-dependent nuclear receptor (rectangle), transcription regulator (ellipse), enzyme (rhombus), transporter (trapezoid), kinase (inverted triangle), and other (circle).

Gene set enrichment analysis

Gene set enrichment analysis (GSEA (Subramanian et al. 2005)) was performed using default parameters and "-metric log2_Ratio_of_Classes", "-permute gene_set", "-nperm 5000". We focused our analysis on curated or pre-ranked gene sets with nominal P <0.05 and empirical FDR <0.25. Enrichment of curated gene sets was analyzed using protein-coding genes differentially expressed (P <0.05, DESeq) in lnc-BATE1-depleted cells vs. control as the expression dataset. The set of genes significantly depleted in lnc-BATE1 KD cells were analyzed for enrichment against the BAT differentiation gene signature previously published (Sun et al. 2013), pre-ranked by the log2 expression change between differentiation days 8 and 0, and against the published gene signature of concurrent genetic deletion of PGC1α and shRNA KD of PGC1β (Uldry et al. 2006), pre-ranked by the log2 expression change between WT and PGC1α KO PGC1β KD. Expression values for the WT and concurrent PGC1α KO PGC1β KD datasets were normalized using GEO2R (http://www.ncbi.nlm.nih.gov/geo/geo2r/).

RNA immunoprecipitation

4-day differentiated primary brown adipocytes grown in 15cm plates were trypsinized, washed, re-suspended in pre-chilled hypotonic buffer, and kept on ice for 15 minutes. Nuclei were released using a glass dounce homogenizer with 10 strokes and pelleted by centrifugation at 2,500g for 15min at 4°C. The supernatant was collected as cytosolic fraction. The pellet was re-suspended in 2ml lysis buffer (25mM hepes, 150mM KCl, 5mM $MgCl_2$, 1mM DTT, protease inhibitor cocktail), and sheared by dounce homogenizer with 50~60 strokes. Nuclear membrane and debris were pelleted and discarded. RNase inhibitor was supplemented in nuclear lysate to a final concentration of 300U/ml before immediate use or storage at -80°C. 30ul Protein A/G beads (SC2003, Santa Cruz) were incubated with 5ug control IgG or indicated antibody in 200ul lysis buffer for 30min at room temperature. Antibody-bound beads were then washed twice in lysis buffer, followed by incubation with 500ul nuclear lysate for 3h at 4°C with continuous rotation. After 4 washes with lysis buffer supplemented with 0.5% NP-40 and 40U/ml RNase inhibitor, 20% of beads were kept for western blot and the rest used for RNA extraction. Bead-associated RNA was co-precipitated with glycogen and re-suspended in 10ul RNase free H_2O. cDNA synthesis and qPCR were then performed as described above.

RNA pull-down

Biotin-labeled lnc-BATE1 RNA and androgen receptor (AR) 3’UTR RNA were *in vitro* transcribed from PCR fragments harboring a 5’ T7 promoter sequence using a MEGAscript kit (Life Technologies) according to the manufacturer’s instructions, with a ratio between biotin-CTP and CTP of 1:20. Biotinylated RNAs were further purified with NucAway spin column and re-folded as described previously (Tsai et al. 2010). Magnetic Dynabeads M-280 streptavidin beads (Life Technologies) were pre-treated with 0.1 M NaOH and washed with 0.1M NaCl, and 50ul beads were then incubated with 30pmol biotin-labeled lnc-BATE1 or control RNA in RNA binding buffer (1M NaCl, 5mM Tris) for 30min at room temperature. Biotinylated RNA-bound beads were then washed with RNA binding buffer and incubated with brown adipocyte nuclear lysate for 3h at 4°C with continuous rotation. Supernatant and 10% beads were kept for RNA extraction to examine RNA stability during the pull-down assay. Beads were then washed 4 times with wash buffer (25mM hepes, 75mM KCl, 5mM $MgCl_2$, 1mM DTT, protease inhibitor cocktail, 40U/ml RNase inhibitor), and RNA-bound proteins were released by boiling beads in sample buffer for 5 minutes at 95 °C. Protein enrichment was then examined by western blot using specific antibodies.

Additional computational methods

Signal density maps were generated using BEDTOOLS (Quinlan and Hall 2010) and visualized in the UCSC genome browser. Statistical tests and plots were implemented in R (http://www.R-project.org/) with default parameters unless stated otherwise. Expression heatmaps were generated using the heatmap.2 function of the *gplots* R package (http://CRAN.R-project.org/package=gplots).

SUPPLEMENTAL REFERENCES

Alvarez-Dominguez JR, Hu W, Yuan B, Shi J, Park SS, Gromatzky AA, van Oudenaarden A, Lodish HF. 2014. Global discovery of erythroid long noncoding RNAs reveals novel regulators of red cell maturation. *Blood* **123**: 570-581.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11**: R106.

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**: 708-715.

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**: 1915-1927.

Carninci P Kasukawa T Katayama S Gough J Frith MC Maeda N Oyama R Ravasi T Lenhard B Wells C et al. 2005. The transcriptional landscape of the mammalian genome. *Science* **309**: 1559-1563.

Consortium F, the RP, Clst, Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Lassmann T, Itoh M et al. 2014. A promoter-level mammalian expression atlas. *Nature* **507**: 462-470.

Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**: 1367-1372.

Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, Babak T. 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Res* **22**: 1173-1183.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39**: W29-37.

Gish W, States DJ. 1993. Identification of protein coding regions by database similarity search. *Nat Genet* **3**: 266-272.

Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES. 2013. Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins. *Cell* **154**: 240-251.

Huang da W, Sherman BT, Lempicki RA. 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**: 1-13.

-. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**: 44-57.

Huttlin EL, Jedrychowski MP, Elias JE, Goswami T, Rad R, Beausoleil SA, Villen J, Haas W, Sowa ME, Gygi SP. 2010. A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143**: 1174-1189.

Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G. 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* **35**: W345-349.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357-359.

Lee JE, Wang C, Xu S, Cho YW, Wang L, Feng X, Baldridge A, Sartorelli V, Zhuang L, Peng W et al. 2013. H3K4 mono- and di-methyltransferase MLL4 is required for enhancer activation during cell differentiation. *Elife* **2**: e01503.

Lin MF, Jungreis I, Kellis M. 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**: i275-282.

Neuert G, Munsky B, Tan RZ, Teytelman L, Khammash M, van Oudenaarden A. 2013. Systematic identification of signal-activated stochastic gene regulation. *Science* **339**: 584-587.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. 2008. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods* **5**: 877-879.

Rajakumari S, Wu J, Ishibashi J, Lim HW, Giang AH, Won KJ, Reed RR, Seale P. 2013. EBF2 determines and maintains brown adipocyte identity. *Cell Metab* **17**: 562-574.

Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, Tan K, Akalin A, Schmeier S, Kanamori-Katayama M, Bertin N et al. 2010. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* **140**: 744-752.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**: 276-277.

Schneider CA, Rasband WS, Eliceiri KW. 2012. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* **9**: 671-675.

Shalgi R, Hurt JA, Krykbaeva I, Taipale M, Lindquist S, Burge CB. 2013. Widespread regulation of translation by elongation pausing in heat shock. *Molecular cell* **49**: 439-452.

Shen L, Shao N, Liu X, Nestler E. 2014. ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics* **15**: 284.

Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, Groudine M, Bender M, Kaul R, Canfield T et al. 2012. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biology* **13**.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**: 15545-15550.

Sun L, Goff LA, Trapnell C, Alexander R, Lo KA, Hacisuleyman E, Sauvageau M, Tazon-Vega B, Kelley DR, Hendrickson DG et al. 2013. Long noncoding RNAs regulate adipogenesis. *Proc Natl Acad Sci U S A* **110**: 3387-3392.

Sun L, Xie H, Mori MA, Alexander R, Yuan B, Hattangadi SM, Liu Q, Kahn CR, Lodish HF. 2011. Mir193b-365 is essential for brown fat differentiation. *Nat Cell Biol* **13**: 958-965.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105-1111.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511-515.

Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E, Chang HY. 2010. Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**: 689-693.

Uldry M, Yang W, St-Pierre J, Lin J, Seale P, Spiegelman BM. 2006. Complementary action of the PGC-1 coactivators in mitochondrial biogenesis and brown fat differentiation. *Cell Metab* **3**: 333-341.

Wang L, Wang S, Li W. 2012. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**: 2184-2185.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.

SUPPLEMENTAL FIGURE LEGENDS

Figure S1. Identification and characterization of adipose tissue lncRNAs and their properties

(A) Precision analysis of expression estimates for adipose tissue-expressed genes annotated by Ensembl (43% expressed >1 FPKM in at least one adipose tissue). Box plots depict how the BAT expression level calculated by resampling a series of subsets from the total RNA reads deviates from that estimated using all of the reads, measured as the percent relative error. Transcripts were sorted by their expression value and divided into expression quartiles (Q1-Q4).

(B) Reproducibility analysis of expression estimates for adipose tissue-expressed Ensembl genes as in (A). Correlation of expression levels between BAT replicates from this study (left), between merged BAT replicates from this study and a previously published dataset (Sun et al. 2013) (center), and between iWAT from this study and its previously published counterpart (right).

(C) Distribution of RNA-seq alignment locations relative to RefSeq mRNA annotations.

(D) Maximum predicted ORF length (across all 3 possible reading frames) of adipose-tissue expressed mRNAs and lncRNAs.

(E) Cumulative density distribution of the ribosome release score (Guttman et al. 2013) across adipose tissue-expressed known mRNA ORFs or over any ORF within their 5'UTRs, 3'UTRs, or within lncRNA exons, calculated from RNA-seq and Ribo-seq of mouse 3T3 cells (Shalgi et al. 2013).

(F) Distribution of read coverage values across our adipose tissue samples.

(G) Detection rates for lncRNAs identified in our study. For each lncRNA (rows), detection by the indicated experimental criterion (columns) is indicated in black.

(H) Average profiles of ChIP-seq signal enrichment for histone marks, open chromatin, and RNA Pol II binding within TSS ± 3kb regions of adipose tissue-expressed lncRNAs as in Fig.1F. Color coding for each profile is shown within.

(I) Transcript length distributions for adipose tissue-expressed lncRNAs and mRNAs

(J) Distribution of number of exons per adipose tissue-expressed mRNA or lncRNA transcript

(K) Number of isoforms per adipose-expressed mRNA or lncRNA locus.

(L) Sequence conservation of mRNA or lncRNA promoters (left), exons (center) and introns (right) across 30 vertebrate genomes as measured by Phastcons.

Figure S2. Tissue specificity and regulation of adipose lncRNAs

(A) Distribution of maximal tissue specificity scores for adipose tissue-expressed mRNA or lncRNA genes. ***P <0.001 (Kolmogorov-Smirnov test).

(B) Abundance of BAT-, iWAT- and eWAT-specific lncRNAs across 30 tissues from ENCODE shown as in Fig.2A.

(C) Locus map of lnc-BAT1, a BAT-restricted lncRNA. UCSC genome browser tracks depict poly(A)+ RNA-seq signal (as density of mapped reads in the indicated tissues), *de novo* transcript models by Cufflinks, and Ensembl gene annotations as in Fig.1F.

(D) PPARγ binding in BAT and in eWAT within TSS ± 3kb regions of adipose-expressed lncRNAs. Color intensity represents normalized ChIP-seq signal. Heat maps are sorted in descending order of signal enrichment in the first profile.

(E) Distribution of PPARγ ChIP-seq read counts in BAT or eWAT within peaks of signal enrichment intersecting TSS ± 3kb regions of both BAT- and eWAT-specific lncRNAs. Read counts were quantified by DESeq and common peaks were determined by MACS.

(F) ChIP-seq signal for histone marks, open chromatin and RNA Pol II binding in cultured brown adipocytes within TSS ± 3kb regions of adipose tissue-expressed lncRNAs. Shown are day 0 (D0) and day 2 (D2) time points of brown adipogenesis from immortalized brown pre-adipocytes. Color intensity represents normalized ChIP-seq signal. Heat maps are sorted in descending order of signal enrichment in the first profile.

(G) Heat maps of BAT-specific lncRNA gene-level expression (FPKM) from poly(A)+ RNA-seq of cultured brown adipocytes as in Fig.2D (left) and ChIP-seq signal as in (F) (right).

(H) Locus map of lnc-BAT1 showing activation and TF binding during differentiation. Track 1 depicts poly(A)+ RNA-seq signal as density of mapped reads in cultured brown pre-adipocytes (D0) and cultured brown adipocytes (D8); track 2 depicts *de novo* transcript models by Cufflinks, with right-to-left arrow indicating transcript in the minus strand; tracks 3-5 display PPARγ, CEBPα, CEBPβ ChIP-seq signal as density of processed signal enrichment at days 0 (D0) and 2 (D2) of differentiation from immortalized brown pre-adipocytes in culture.

Figure S3. **Depleting lnc-BATE1 by siRNA or shRNA leads to significant down-regulation of key marker genes**

(A) Determination of 5’ and 3’ ends by RACE reveals 3 variants of lnc-BATE1 in brown adipocytes. Agarose gel image shows 5’RACE and 3’RACE PCR products as indicated.

(B) Gene structure of lnc-BATE1 in mouse genome and its 3 variants.

(C-E) Expression of lnc-BATE1 (C), common adipogenic markers (D) and brown adipocyte markers (E) in 3-days differentiated brown adipocytes transfected with siRNAs as indicated.

(F-I) Expression of lnc-BATE1 (F), brown adipocyte markers (G), common adipogenic markers (H) and mitochondrial markers (I) in 5-days differentiated brown adipocytes infected with shRNAs as indicated.

Error bars are s.e.m., n =3. **P* ≤0.05, ***P* ≤0.01.

Figure S4. **Ectopic expression of lnc-BATE1 does not stimulate brown adipocyte differentiation or determination**

(A-B) Expression of lnc-BATE1 and BAT marker genes in brown adipocytes differentiated for 5 days under standard conditions (A) or conditions of 10-fold reduced differentiation cocktail (B).

(C-E) Overexpression of lnc-BATE1 is not sufficient to promote browning in cultured inguinal white adipocytes in the absence (C) or presence of norepinephrine (NE) (D) or in cultured epididymal white adipocytes.

(F) Expression of lnc-BATE1 and myogenic markers in C2C12 myoblasts differentiated for 6 days.

Error bars are s.e.m., n =3

Figure S5. **Analysis of global gene expression changes upon lnc-BATE1 depletion.**

(A) Genes enriched in lnc-BATE1-depleted cells are normally downregulated during brown adipogenesis. Expression estimates (log2 FPKM) in cultured brown adipocytes at differentiation days 0 (D0) and 8 (D8) are shown for genes expressed significantly higher (P <0.05, DESeq) upon lnc-BATE1 inhibition relative to control (left). GSEA identified genes involved in cell cycle progression, cell adhesion, and various signaling processes as significantly enriched within this group (right). Graphical data represent enrichment scores across the genome-wide transcriptional profile (~9000 genes). Genes that are higher-expressed in lnc-BATE1-depeleted cells are presented in red, whereas lower-expressed ones are shown in blue. Number of members in the gene set, normalized enrichment scores and their empirical P values are indicated.

(B) Genes depleted in lnc-BATE1-inhibited cells are normally upregulated during brown adipogenesis. Expression estimates and GSEA analysis as in (A) but for lower-expressed genes.

(C) Cumulative density distributions of expression changes (left) and p-values for these changes (right) for all expressed protein-coding genes and for BAT-specific, WAT-specific and common adipogenic genes in siRNA-treated cultured day 3 brown adipocytes. Changes are log2 expression (FPKM) ratios over control siRNA. The 0.05 p-value significance threshold is indicated by a vertical dashed gray line.

(D) Genes targeted by PGC1α, ESRRα, and PPARγ are significantly depleted in lnc-BATE1-inhibited cells. GSEA analysis as in (B).

(E) Genes downregulated in lnc-BATE1 KD cells are significantly inhibited upon concurrent genetic loss of PGC1α and depletion of PGC1β (Uldry et al. 2006). GSEA analysis as in (B).

Figure S6. **Depletion of lncBATE1 does not affect expression of neighboring genes**

(A) Schematic illustration of neighboring genes flanking the lnc-BATE1 locus.

(B) Expression of lnc-BATE1 and neighboring genes across different tissues shown as in Fig.2A.

(C-D) Expression of neighboring genes in 3-days and 5-days differentiated brown adipocytes transfected with DsiRNA2, assessed by RNA-seq (C) and qPCR (D). Error bars are s.e.m., n =3.

Figure S7. **Prediction of a potential hnRNP U binding site in lnc-BATE1 RNA**

Positional weight matrix motif of hnRNP U binding sites identified from previous CLIP-Seq data (Huelga et al., 2012) is displayed in the top panel. Putative hnRNP U binding site in lncBATE1 is displayed in the bottom panel.

Supplemental Figures

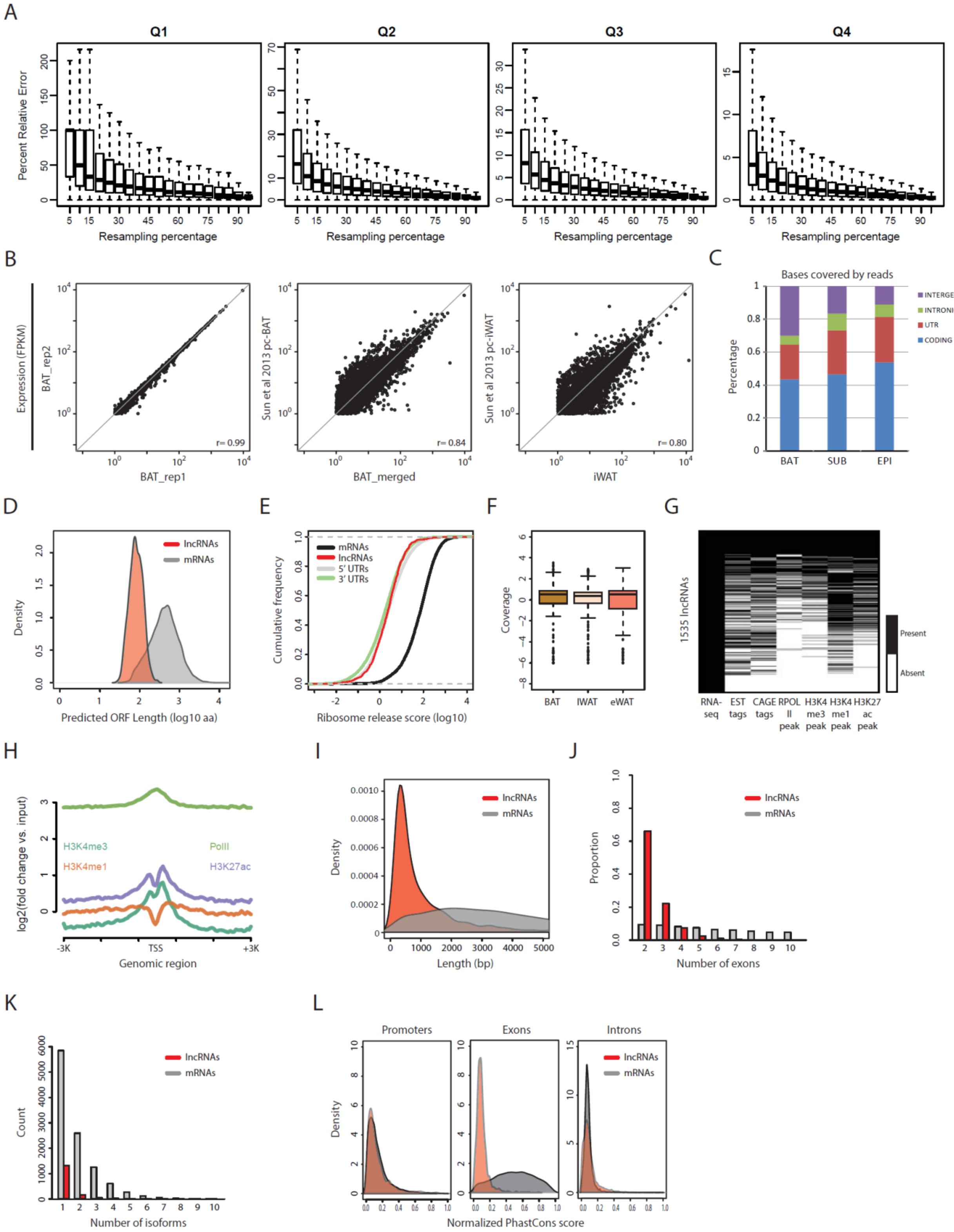


Figure S1. Identification and characterization of adipose tissue lncRNAs and their properties

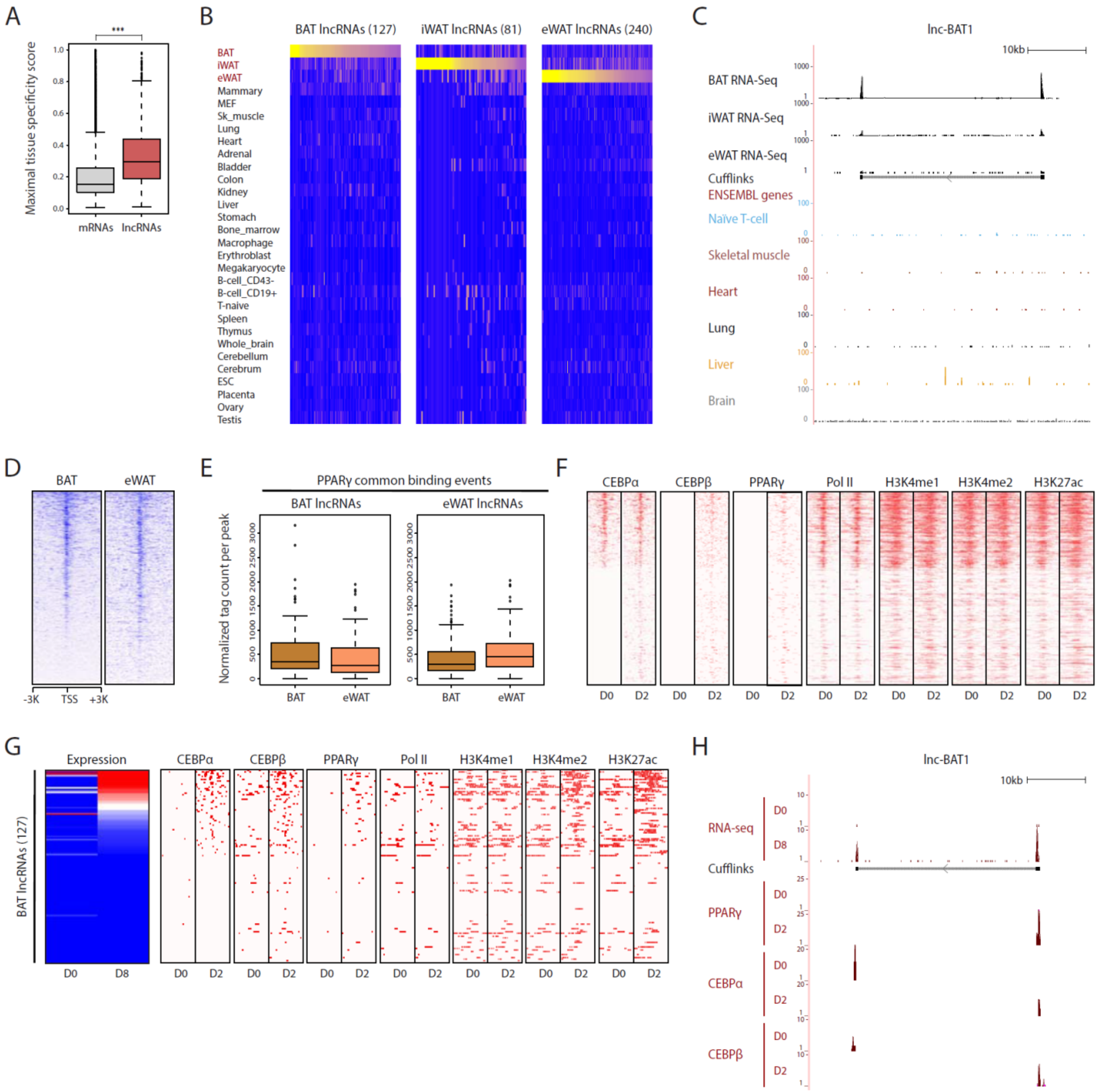


Figure S2. Tissue specificity and regulation of adipose lncRNAs

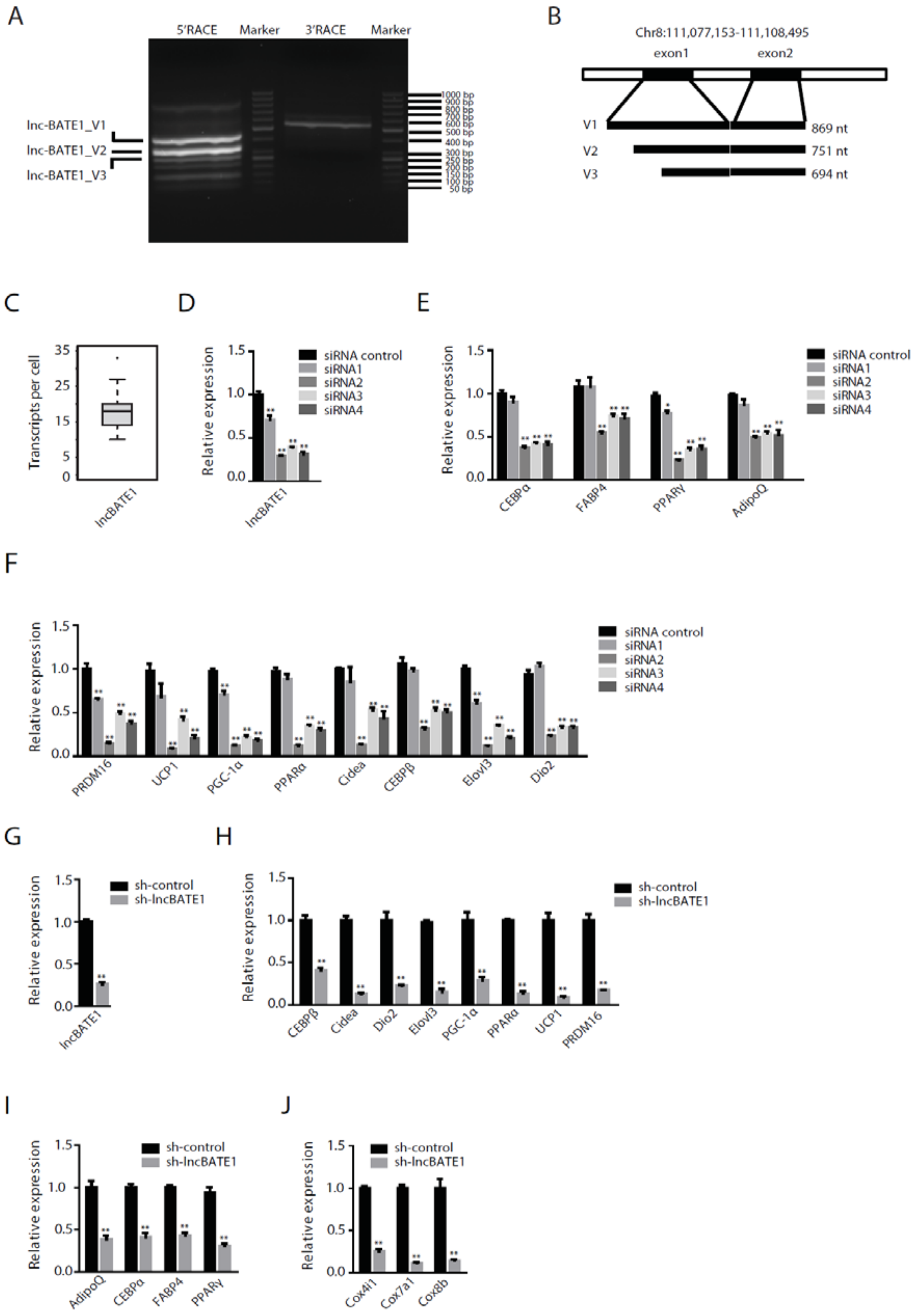


Figure S3. Depleting lnc-BATE1 by siRNA or shRNA leads to significant down-regulation of key marker genes

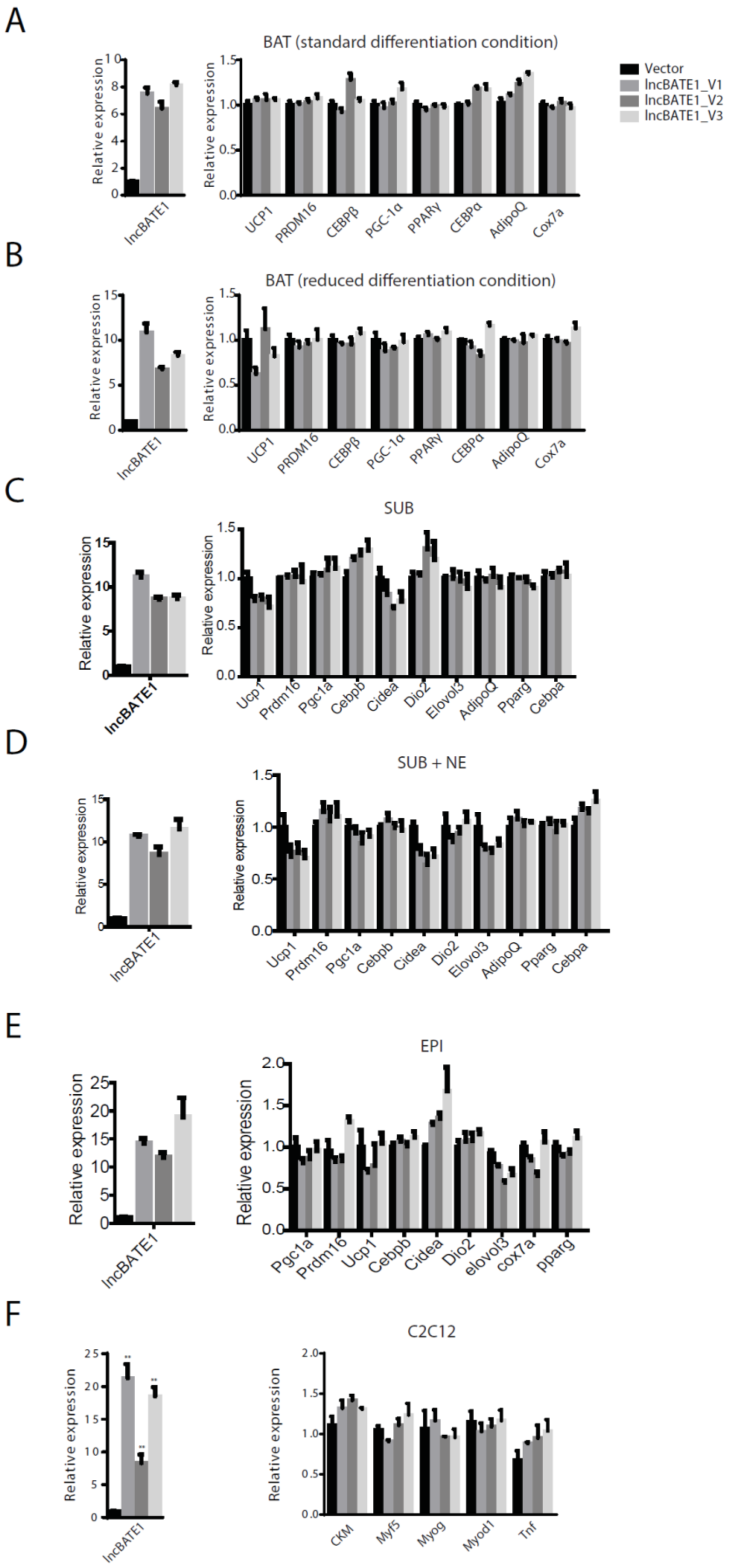


Figure S4. **Ectopic expression of lnc-BATE1 does not stimulate brown adipocyte differentiation or determination**

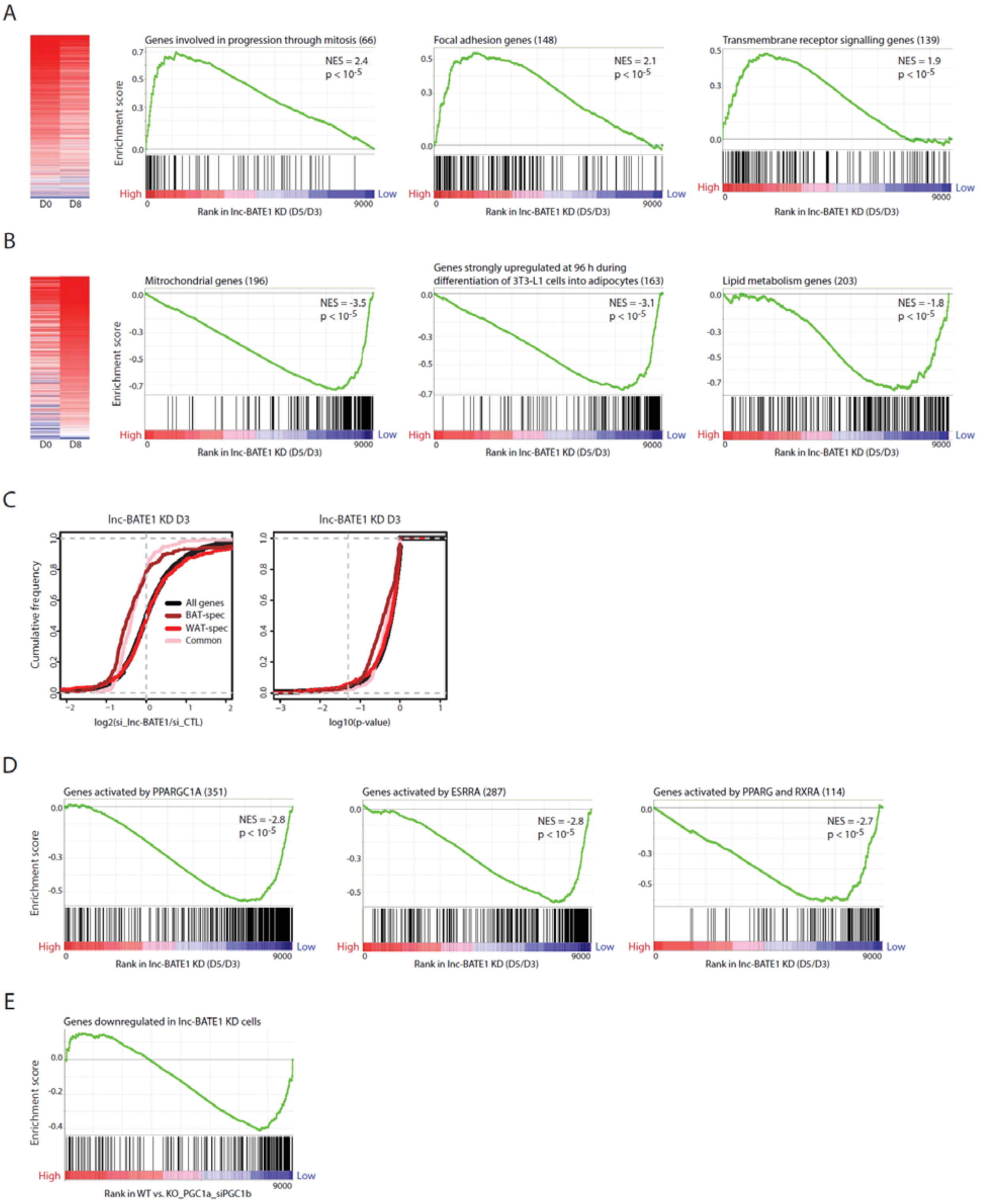


Figure S5. **Analysis of global gene expression changes upon lnc-BATE1 depletion.**

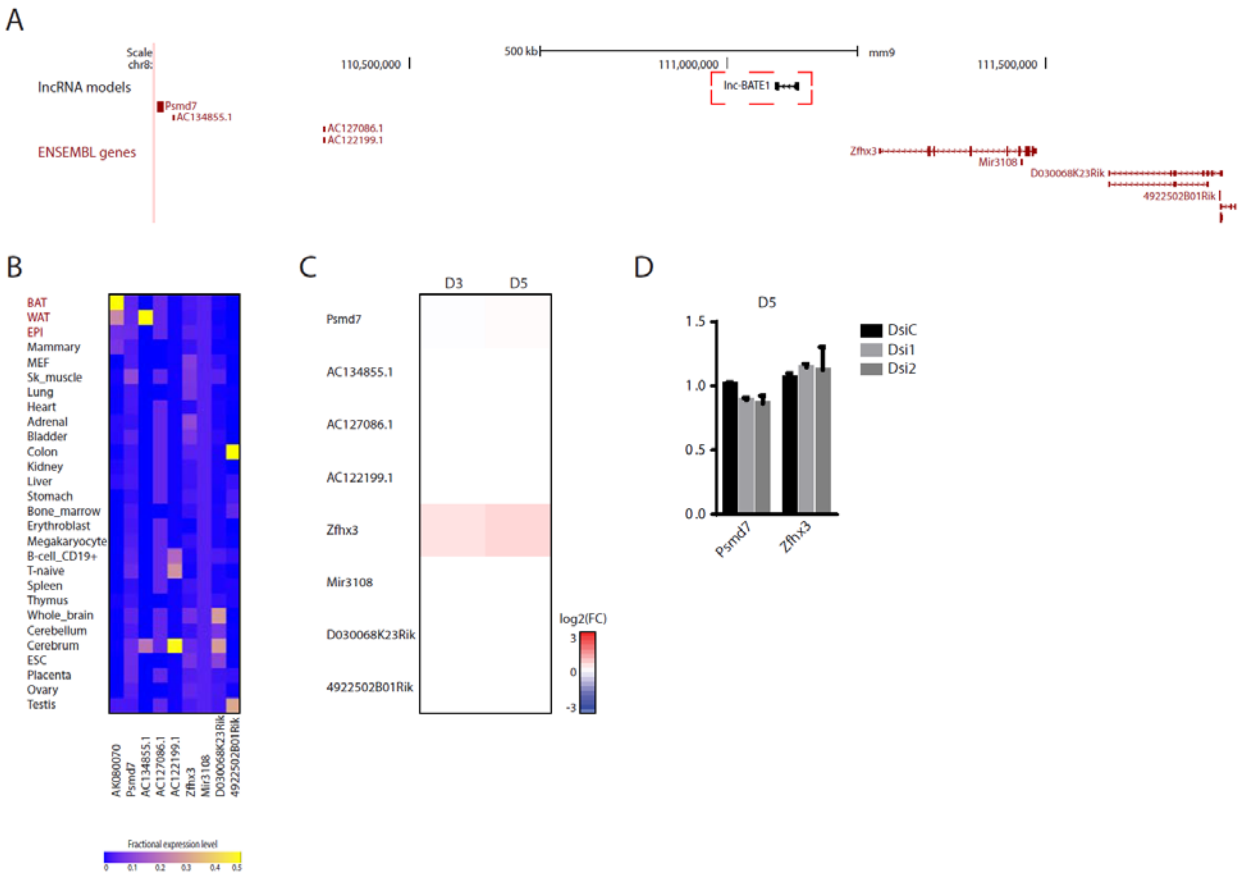


Figure S6. **Depletion of lncBATE1 does not affect expression of neighboring genes**

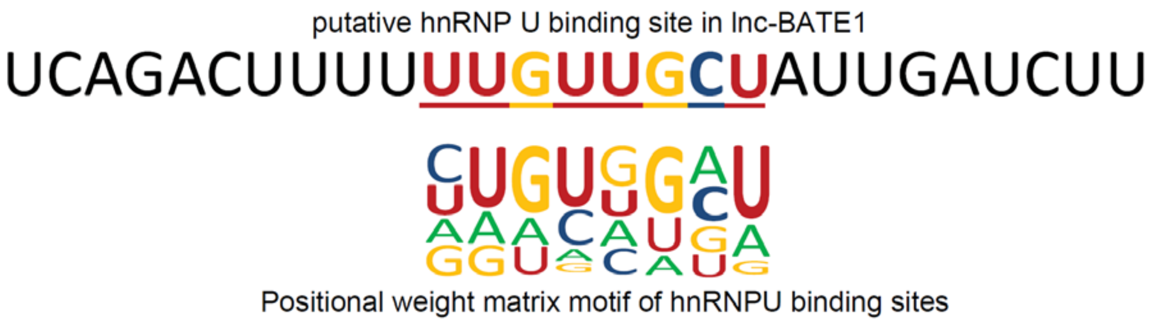

putative hnRNP U binding site in lnc-BATE1
UCAGACUUUUUUGUUGCUAUUGAUCUU
Positional weight matrix motif of hnRNPU binding sites

Figure S7. **Prediction of a potential hnRNP U binding site in lnc-BATE1 RNA**

Supplemental Tables

Sample	Platform	Library	Read length	Mapped reads	Reference
BAT_d0	Illumina_GA2x	Long polyA(+)	1x36bp	8,354,876	Sun et al 2013
BAT_d8	Illumina_GA2x	Long polyA(+)	1x36bp	8,042,012	Sun et al 2013
iWAT_d0	Illumina_GA2x	Long polyA(+)	1x36bp	8,846,613	Sun et al 2013
iWAT_d8	Illumina_GA2x	Long polyA(+)	1x36bp	10,071,021	Sun et al 2013
BAT_rep1	Illumina_HiSeq_2000	Long polyA(+)	2x74bp	110,266,711	This study
BAT_rep2	Illumina_HiSeq_2000	Long polyA(+)	2x74bp	108,609,089	This study
iWAT	Illumina_HiSeq_2000	Long polyA(+)	2x74bp	107,693,884	This study
eWAT	Illumina_HiSeq_2000	Long polyA(+)	2x74bp	121,906,117	This study
BAT_d3_NC	Illumina_HiSeq_2000	Long polyA(+)	2x100bp	33,207,127	This study
BAT_d5_NC	Illumina_HiSeq_2000	Long polyA(+)	2x100bp	38,250,329	This study
BAT_d3_siRNA	Illumina_HiSeq_2000	Long polyA(+)	2x100bp	41,100,611	This study
BAT_d5_siRNA	Illumina_HiSeq_2000	Long polyA(+)	2x100bp	45,530,104	This study

Table S1. List of RNA-seq datasets used in this study.

Upstream Regulator	Predicted Activation State	Activation z-score	p-value of overlap	Target molecules in dataset
PPARGC1A	Inhibited	-4.093	4.05E-15	C3,CIDEA,COX5A,Cox5b/LOC102638382,CYCS,Esrra,FABP3,IDH3A,MB,PD
ESRRA	Inhibited	-2.397	3.27E-08	Cox8b,CYCS,Esrra,FABP3,IDH3A,LDHB,PDK4,PPARA,PPARGC1A,SLC25A
PPARA	Inhibited	-2.726	7.30E-07	ACOT2,AQP7,C3,CIDEA,CYP27A1,ECI1,FABP3,GPD2,GSTT2/GSTT2B,HAD
PPARG	Inhibited	-2.796	7.30E-07	Adig,AQP7,C3,Cdkn1c,CIDEA,FABP3,HADHB,LOC102724788/PRODH,NDUF
HNF4A	Inhibited	-2.202	8.35E-06	ACOT13,ACOX2,ADCK3,ALKBH7,APH1A,C10orf10,C3,CBS/LOC102724560, 5,NDUFS4,PANK1,PDCD4,PDK2,PDK4,PPARA,PPARGC1A,PPP1R3C,SLC2

Table S3. Top 5 upstream regulators of genes depleted in lnc-BATE1-inhibited cells

Appendix D: Translation of small open reading frames within unannotated RNA transcripts in Saccharomyces cerevisiae

This work was first published as:

Smith, J., **Alvarez-Dominguez, J.R.**, Kline, N., Huynh, N., Geisler, S., Hu, W., Coller, J., and Baker, K.E. (2014). Translation of small open reading frames within unannotated RNA transcripts in Saccharomyces cerevisiae. *Cell Reports* **7**, 1858-1866.

Author contributions: J.R.A.-D. designed, performed, and presented the analysis of sequencing reads from RNA-seq, Ribo-seq, and CLIP-seq experiments in mouse embryonic stem cells (Figure 4 and Figure S4); contributed design of and presentation of bioinformatics analyses (Figure S2); performed, analyzed, and presented the data from RNA-FISH experiments in yeast cells (not shown), and contributed to study design and manuscript revisions.

Acknowledgments: The authors thank Tim Nilsen for critical evaluation of the manuscript and members of the K.E.B. and J.C. labs for helpful insight into this work. This research was funded by the National Institute of General Medical Sciences (GM080465 to J.C. and GM095621 to K.E.B.) and the National Science Foundation (NSF1253788 to K.E.B.).

OPEN ACCESS
CellPress

Translation of Small Open Reading Frames within Unannotated RNA Transcripts in *Saccharomyces cerevisiae*

Jenna E. Smith,[1] Juan R. Alvarez-Dominguez,[2] Nicholas Kline,[1] Nathan J. Huynh,[1] Sarah Geisler,[1,3] Wenqian Hu,[2] Jeff Coller,[1] and Kristian E. Baker[1,*]
[1]Center for RNA Molecular Biology, Case Western Reserve University, Cleveland, OH 44106, USA
[2]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA
[3]Present address: Department of Biosystems Science and Engineering, Eidgenössische Technische Hochschule Zürich, 4058 Basel, Switzerland
*Correspondence: keb22@case.edu

http://dx.doi.org/10.1016/j.celrep.2014.05.023

SUMMARY

High-throughput gene expression analysis has revealed a plethora of previously undetected transcripts in eukaryotic cells. In this study, we investigate >1,100 unannotated transcripts in yeast predicted to lack protein-coding capacity. We show that a majority of these RNAs are enriched on polyribosomes akin to mRNAs. Ribosome profiling demonstrates that many bind translocating ribosomes within predicted open reading frames 10–96 codons in size. We validate expression of peptides encoded within a subset of these RNAs and provide evidence for conservation among yeast species. Consistent with their translation, many of these transcripts are targeted for degradation by the translation-dependent nonsense-mediated RNA decay (NMD) pathway. We identify lncRNAs that are also sensitive to NMD, indicating that translation of noncoding transcripts also occurs in mammals. These data demonstrate transcripts considered to lack coding potential are bona fide protein coding and expand the proteome of yeast and possibly other eukaryotes.

INTRODUCTION

The recent advent of high-throughput DNA sequencing technologies has led to the detection of a plethora of novel RNA transcripts and the revelation that vast regions of the genome once thought to be transcriptionally silent are, in fact, actively engaged by RNA polymerases (Bernstein et al., 2012). Although some of these RNA products arguably represent transcriptional noise, a growing body of evidence suggests that many may have bona fide function in the cell. In particular, long noncoding RNAs (lncRNAs) have emerged as important regulators of gene expression, with established roles in epigenetic modification of chromatin, transcriptional control, and mRNA regulation post-transcriptionally (Geisler and Coller, 2013).

lncRNAs are classified based on transcript size (>200 nucleotides [nt] in length) and as lacking computationally predicted protein coding regions of significant size and/or conservation (Derrien et al., 2012). The general assumption that lncRNAs are not translated is, however, at odds with their striking similarity to protein-coding mRNAs. Specifically, most lncRNAs are products of RNA polymerase II and harbor 5′ methyl-guanosine caps and 3′ termini of polyadenosine residues (Guttman et al., 2009)—key features promoting the efficient translation of mRNA. Indeed, investigation into a role for lncRNAs as templates for protein synthesis has suggested that these transcripts may associate with the cellular translation machinery. Polyribosome purification and genome-wide ribosome profiling have shown that lncRNAs cofractionate with and/or bind ribosomes (Ingolia et al., 2011; Chew et al., 2013; Brar et al., 2012; van Heesch et al., 2014). The predictive value of ribosome profiling to define protein-coding potential has, however, been recently challenged (Guttman et al., 2013), and the overall contribution to the proteome of peptides generated from translation of lncRNA is suggested to be low (Bánfai et al., 2012). Therefore, it remains unclear how widespread the translation of predicted noncoding RNAs may be and what percentage of lncRNAs function strictly as regulatory RNA.

Similar to metazoa, budding yeast *Saccharomyces cerevisiae* has been shown to express an extensive repertoire of novel transcripts (David et al., 2006; Nagalakshmi et al., 2008). Study of a limited number of RNAs in this class has implicated them in controlling gene expression generally through transcriptional regulation or interference (Geisler and Coller, 2013); however, like lncRNAs, the function of most unannotated transcripts in yeast and the extent of their biological role in the cell remain unknown. In this study, we investigate hundreds of previously unannotated transcripts in yeast and provide strong evidence that many of these RNAs possess protein-coding capacity. Specifically, we find unannotated RNAs associate with polyribosomes to extents similar to mRNA and that they encode small open reading frames (ORFs) bound by ribosomes. Consistent with their translation, we observe a significant percentage of these RNAs are sensitive to nonsense-mediated RNA decay

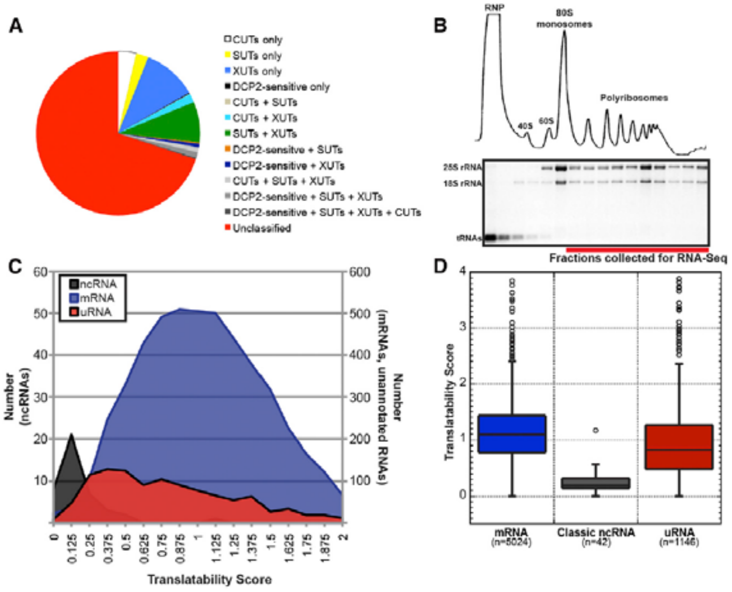


Figure 1. Yeast uRNAs Cosediment with Polyribosomes

(A) Overlap between uRNAs identified in this study and stable unannotated transcripts (SUTs), RNAs targeted for degradation by the ribonucleases RRP6 (CUTs) or XRN1 (XUTs), and DCP2-sensitive lncRNAs (see the Supplemental Experimental Procedures).

(B) Polysome analysis of yeast cell lysates. Top: UV trace after sedimentation through sucrose gradients. Bottom: ethidium bromide stain of RNA isolated from each gradient fraction. RNA for Polysome-seq pooled from fractions indicated.

(C) Translatability score (FPKM$_{polysomes}$/FPKM$_{steady-state}$ based on averages of expression from two biological replicates) for characterized ncRNAs, mRNAs, and uRNAs.

(D) Distribution of translatability scores as in (C) for each class of RNA. Box includes 25th to 75th percentiles; whiskers indicate ± 1.5 IQRs, with outliers indicated by circles.

See also Figure S1, Table S1, and Table S2.

(NMD), a translation-dependent process. Similarly, we calculate that a subset of mammalian lncRNA is sensitive to NMD, indicating that these transcripts are also substrates for translation. Together, our data expand the coding capacity of the yeast genome beyond the current annotation and suggest expression of dozens of short polypeptides from transcripts previously predicted to lack coding potential.

RESULTS

Hundreds of Unannotated and Previously Unclassified RNA Transcripts Are Expressed in *S. cerevisiae*

We performed genome-wide gene expression analysis using RNA-seq to generate a global map of transcripts expressed in yeast. Whole-cell, steady-state RNA from wild-type cells was ribosomal RNA-depleted and used to construct strand-specific cDNA libraries that were analyzed with Illumina HiSeq to produce ~11–22 million uniquely mapped sequence reads (Table S1 and Figure S1A). Reads mapping to annotated features of the Ensembl sacCer2 *Saccharomyces* genome confirmed expression of 5,066 protein-coding mRNA and classic noncoding RNA transcripts (ncRNA; e.g., snRNA, snoRNA). The remainder of reads mapped to unique and unannotated loci (see the Supplemental Experimental Procedures; Roberts et al., 2011) revealing expression of 1,146 transcripts with a length greater than 200 nt, herein referred to as unannotated RNAs (uRNAs; Table S2). A number of uRNAs are expressed from loci corresponding to transcripts previously described by our group as DCP2-sensitive, long noncoding RNAs (Geisler et al., 2012) or RNAs previously described as either stable unannotated transcripts (Xu et al., 2009) or RNAs targeted for degradation by the ribonucleases RRP6 or XRN1 (Xu et al., 2009; van Dijk et al., 2011; Figure 1A; see the Supplemental Experimental Procedures). The remainder of uRNAs (~800) lack previous classification and include transcripts expressed from intergenic regions of the genome and antisense to annotated protein-coding genes.

A Majority of uRNAs Associate with Polyribosomes Akin to mRNA

The yeast genome has been exhaustively annotated for protein coding capacity, and the uRNAs we identified by RNA-seq are predicted to lack protein-coding potential. Recent studies, however, have uncovered unexpected associations between predicted noncoding RNA and the translation machinery, leading us to directly assess whether uRNAs in yeast are, in fact, noncoding. To evaluate the translational status of uRNAs, we used polyribosome analysis to enrich translation complexes and their associated RNA by sedimentation of cell lysates through sucrose gradients. Gradient fractions corresponding to polysomes were pooled (Figure 1B) and isolated RNA analyzed with RNA-seq to provide a genome-wide view of polyribosome-associated RNA (i.e., Polysome-seq). The ~23 million mapped reads (Table S1 and Figure S1B) were compared to RNA-seq data generated from total RNA to generate a translatability score representing the relative ratio of polysome association for every cellular transcript.

As anticipated, classic ncRNAs were generally excluded from polyribosomes as represented by low translatability scores (Figures 1C and 1D; mean, 0.24 ± 0.19 SD). In contrast, protein-coding mRNAs spanned a large range of translatability, reflecting differences in translation efficiency as well as different rates of cotranslational degradation (mean, 1.12 ± 0.49 SD; Hu et al., 2009). Importantly, 98.98% of mRNA exhibited a translatability score greater than the mean score for classic ncRNA, demonstrating that polyribosome analysis provides an effective

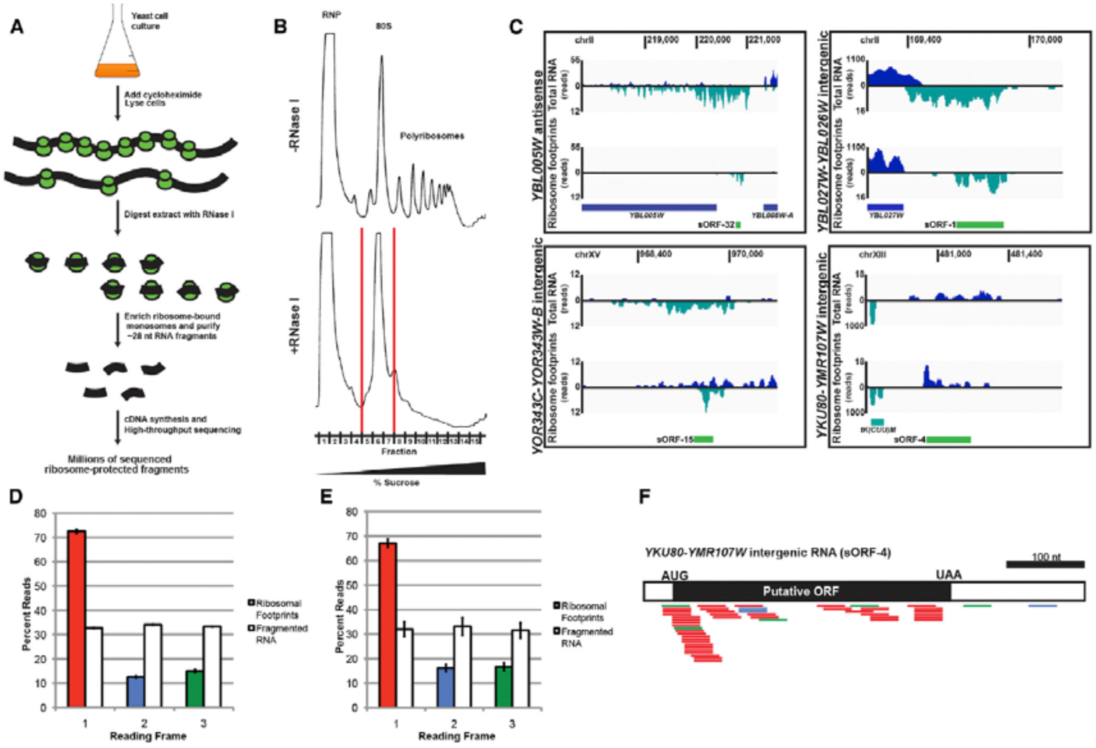


Figure 2. Ribosome Profiling Provides Evidence for Translation of uRNAs

(A) Schematic of ribosome profiling protocol.

(B) Representative UV trace of polyribosome gradients from cell lysates without (−) or with (+) RNase I treatment. Fractions encompassing the collapsed 80S peak following RNase I-treatment collected for analysis are indicated.

(C) RNA-seq and ribosome footprints for sample uRNAs. Watson strand (navy); Crick strand (teal). Annotated genes (navy or teal bars) and putative sORFs delineated by ribosome footprints (green bars) are indicated.

(D) Fraction of 28 nt ribosome-protected fragments (RPFs) mapping to each of three frames for annotated mRNAs. Two biological replicates of each WT and *upf1Δ* ribosome footprints were analyzed as four independent samples and single replicates of each WT and *upf1Δ* fragmented RNA were analyzed as two independent samples. Data are mean ± SEM.

(E) Fraction of 28 nt RPFs mapping to each of 3 frames for the 61 uRNAs demonstrating ribosome phasing (where ≥50% of RPFs mapped to a single frame). For each uRNA, the +1 frame was retrospectively classified. Each uRNA was teated as a single replicate; data shown as mean ± SEM.

(F) Shows 28 nt RFPs mapping to *YKU80-YMR107W* intergenic uRNA demonstrate phasing and delineate an ORF within AUG start and UAA stop codons. RPFs colored based on frame to which they map as in (D) and (E).

See also Figure S2 and Table S1.

biochemical method to characterize the association of RNAs with the translation machinery. Analysis of the translatability score for uRNAs revealed a wide range of association with the translation machinery similar to that of mRNAs (mean, 0.98 ± 0.79 SD; Figures 1C and 1D), although with a distinct distribution pattern that cannot simply be attributed to differences in RNA length (Figure S1C). Critically, > 95% of uRNAs have a translatability score greater than the mean for classic ncRNA, highlighting a significant distinction between well-characterized noncoding RNAs and transcripts predicted to be nonprotein coding. These data reveal that uRNAs demonstrate a varying degree of association with ribosomes and provide preliminary evidence that many uRNAs in yeast engage the translation machinery.

Ribosome Profiling Reveals Short ORFs within uRNAs

We performed ribosomal profiling to corroborate the association of uRNA with the translation machinery and define—at nucleotide resolution—the nature of interaction between each uRNA and 80S ribosomes (Ingolia et al., 2009; Figure 2A). To minimize recovery of nonribosome-bound, nuclease-protected RNA fragments that can arise by this procedure (Guttman et al., 2013), RNase-digested cell lysates were subject to sucrose gradient centrifugation, and the broadened 80S gradient fractions resulting from collapse of polyribosomes were exclusively selected (Figure 2B). High-throughput sequencing of 80S-bound material derived from wild-type cells generated ~3–6 million mapped nonribosomal RNA reads (Table S1). Analysis of ribosome-protected fragments revealed >50% of uRNAs detected in this analysis (185 of 331) bound ribosomes at levels ≥ 10% of expressed transcript levels (see Supplemental Experimental Procedures). Importantly, when reads generated by ribosome profiling were compared to RNA-seq reads of fragmented total RNA prepared in parallel, the resulting footprinting score correlated strongly with translatability scores calculated from Polysome-seq (Figure S2A).

In addition to validating that a large fraction of uRNAs in yeast is ribosome bound, analysis of the distribution of ribosome-protected fragments along uRNAs revealed two striking observations. First, the coverage area of fragments aligning to individual uRNAs was small, suggesting that these transcripts encode polypeptides of limited size (Figure 2C). Indeed, the average size of nuclease-protected regions on uRNAs was 365 nt, significantly smaller than annotated yeast coding regions, which average 1,344 nt (Figure S2B). Second, the distribution of ribosome-protected fragments mapped predominantly proximal to the uRNA 5′ end, consistent with the scanning model of translation for mRNAs (Figure 2C; Kozak, 1989). Moreover, the distribution of 80S-protected RNA resulted, in some cases, in long regions of downstream RNA that do not appear to associate with ribosomes (discussed below).

To further resolve the protein coding potential for uRNAs based on ribosome profiling, we analyzed nuclease-protected fragments exactly 28 nt in length for their ability to predict periodicity—fragments that align to a single reading frame due to the 3 nt translocation of the ribosome along the RNA in vivo (Ingolia et al., 2009). Analysis of 28 nt reads mapping to annotated protein-coding genes demonstrated that >70% corresponded to the +1 frame position (Figure 2D), confirming codon-triplet phasing and a strong bias toward in-frame footprints as compared to fragmented input RNA. Strikingly, for uRNAs with sufficient 28-mer footprints, 61 of 80 transcripts had footprints mapping predominantly to a single frame (see Supplemental Experimental Procedures; Figure 2E). Moreover, for 53 of these, the ribosome-protected fragments clearly demarcated at least one reading frame flanked by canonical AUG initiation and translation termination codons (e.g., Figure 2F). Metagene analysis of ribosome footprints along mRNAs and uRNAs confirm the annotation of yeast coding regions and predicted ORFs, respectively (Figures S2C and S2D). Importantly, ORFs predicted to be encoded within uRNAs are small—between 10 and 100 amino acids—and will be referred to herein as short ORFs (sORFs; Table S3).

Evidence for Expression of sORFs Encoded within uRNAs

Several pieces of evidence indicated that a subset of unannotated transcripts expressed in yeast are polyribosome-associated, enriched for 80S ribosome binding within a subregion of the transcript, and harbor translocating ribosomes seemingly engaged in protein synthesis. Inspection of sORF-containing uRNA expression indicated that these transcripts are present at levels equivalent to many mRNAs encoding short polypeptides (Figure S3), suggesting that the putative protein products encoded by uRNAs may be present at physiologically relevant levels and play important biological roles in the cell. To verify that sORFs predicted by ribosome profiling can be translated in vivo, we epitope-tagged three individual sORFs at their chromosomal loci by homologous recombination (Longtine et al., 1998; Figure 3A). A polypeptide of the expected size was detected from one of these and was dependent upon insertion of the epitope in the correct predicted reading frame (Figure 3B), demonstrating sORF translation under endogenous conditions.

To avoid alteration of the genomic locus downstream of the sORF that occurs as a consequence of chromosomal gene tagging, we cloned DNA encoding five intergenic uRNAs and inserted sequences encoding an epitope tag precisely upstream of the predicted stop codon (Figure 3C). Using this approach, we observed peptide products from two predicted sORFs (Figure 3D). Importantly, uRNA transcription is driven by endogenous promoter elements within the cloned DNA and expressed transcripts harbor native leader and 3′ untranslated region (3′ UTR) sequences. These data provide clear evidence for in vivo translation of sORFs from uRNA predicted to lack protein coding potential.

sORFs Are Conserved within Fungal Species

As a means to evaluate if polypeptides encoded by sORFs have biological significance, we examined the level of evolutionary conservation within yeast. Importantly, ten species spanning >100 million years of evolution across 12 distinct clades were evaluated (Kurtzman and Robnett, 2003), with the expectation that conservation of peptides amid such significant genetic divergence is indicative of selective pressure to maintain sORF expression. Comparison of peptide sequences predicted from uRNAs revealed that 39 sORFs exhibited varying levels of conservation within closely related species (with 20 sORFs displaying conservation between >1 species; Figure 3E and Table S4). Homologs for six of the most conserved polypeptides were detected within at least one fungal species outside of the *Saccharomyces* sensu stricto genus, with three of these found in strains predicted to diverge from *S. cerevisiae* >100 million years ago. Importantly, 12 sORFs exhibited a bias toward synonymous mutation, demonstrating conservation at the level of peptide sequence that is not a consequence of conserved nucleotide sequence elements (Table S4; Zhang et al., 2006). Finally, sORFs for 14 uRNAs are encoded within conserved genomic regions identified by phastCons (Table S4; Siepel et al., 2005). Together, these data reveal evolutionary pressure to maintain expression of a subset of sORFs within yeast species and argue that the encoded polypeptides have important biological functions in the cell.

Numerous uRNAs Are Targets of Nonsense-Mediated RNA Decay

Our mapping of ribosome-protected fragments revealed that the region of 80S coverage on many uRNAs was limited and concentrated proximal to the transcript 5′ end. Moreover, for a number of uRNAs, the predicted sORF was followed downstream by an extended stretch of unprotected RNA. Based on observations in yeast and metazoa implicating 3′ UTR length in targeting mRNA to rapid decay by the nonsense-mediated RNA decay pathway (NMD; Muhlrad and Parker, 1999; Singh et al., 2008), we hypothesized that a subset of yeast uRNAs might also be targeted

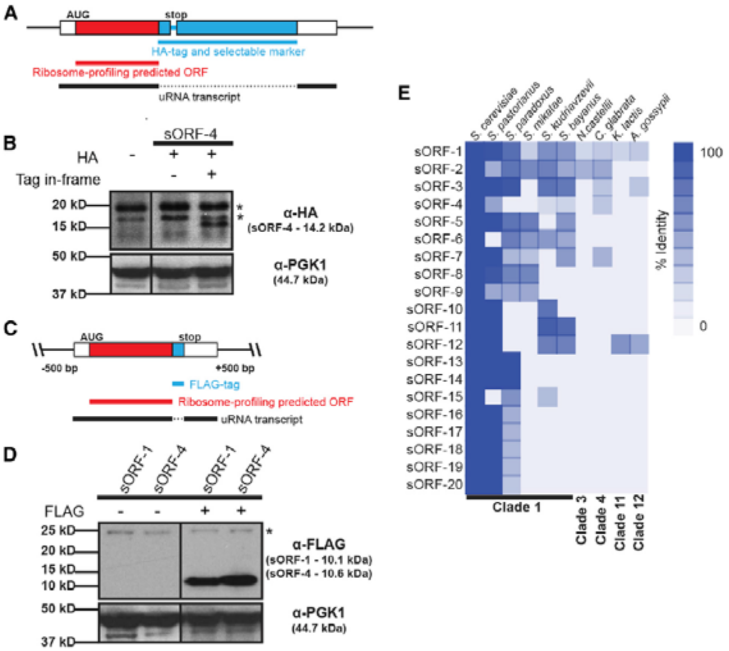


Figure 3. Evidence for Expression and Conservation of sORFs

(A) Epitope tagging of putative sORFs at their endogenous chromosomal locus by homologous recombination. Solid black line represents uRNA defined by RNA-seq.

(B) Western blot analysis detects the translation product of chromosomally tagged sORF-4. Signal is specific to in-frame tag and corresponds to molecular weight for the chimeric peptide. Asterisk indicates a nonspecific signal. PGK1 serves as loading control.

(C) Genomic DNA flanking mapped uRNAs was cloned and the putative sORF epitope tagged at its C terminus. Solid black line represents uRNA defined by RNA-seq.

(D) Western blot detects translation of yeast sORF-1 and sORF-4. Signal is specific for epitope-tagged sORF and corresponds to expected molecular weight for each chimeric peptide. Asterisk indicates a nonspecific signal. PGK1 serves as loading control.

(E) Conservation of sORFs among divergent yeast species. Putative peptides encoded by sORFs were identified in other yeast species based on six-frame translation using TBLASTN. Percent identical residues relative to full-length putative peptide indicated. Top 20 most conserved candidates shown.

See also Figure S3, Table S3, and Table S4.

by NMD. Importantly, sensitivity of uRNAs to NMD would serve to provide additional evidence that these transcripts engage actively translocating ribosomes because NMD is strictly a translation-dependent process (Maquat, 2004).

To determine whether uRNA are sensitive to the NMD pathway, we performed RNA-seq on steady-state RNA isolated from cells deficient in the NMD pathway (due to deletion of *UPF1* encoding a key component of the NMD machinery; Leeds et al., 1991; Table S1). Comparison of RNA levels between wild-type and *upf1Δ* cells revealed 192 of 1,146 uRNAs (16.8%) increased in abundance ≥2-fold in the absence of NMD (see the Supplemental Experimental Procedures; Figures 4A and 4B), several of which were verified experimentally with Northern blot analysis (Figures 4C and S4A). Although increased steady-state abundance in the absence of UPF1 does not differentiate direct versus indirect substrates of the NMD pathway, we found that NMD-sensitive uRNAs associated with polyribosomes to a similar extent as that observed for NMD-sensitive protein-coding mRNA (Figure S4B) and demonstrated dramatically higher average translatability scores compared to NMD-insensitive uRNAs (Figure S4C). Moreover, for individual transcripts, increased ribosome footprints were observed for NMD-sensitive uRNAs in the absence of UPF1, including *ICR1*, a characterized noncoding transcript previously shown to be sensitive to NMD (Toesca et al., 2011; Figure 4D, Table S1). The sensitivity of a subset of uRNAs to NMD and enhanced ribosome association in the absence of UPF1 provide further support that numerous uRNAs in yeast encode sORFs engaged by actively translating ribosomes.

We observed that the average length of RNA protected by ribosome footprints, although short, was not significantly different among uRNAs that were NMD-sensitive versus insensitive. In contrast, the length of RNA downstream of the ribosome-protected region was significantly longer for NMD-sensitive transcripts compared to those that did not respond to inactivation of the NMD pathway (891 nt ± 64 SEM versus 287 nt ± 50 SEM; Figure 4E). These findings are consistent with the observation that mRNAs in yeast with 3′ UTR lengths greater than 300 nt are efficiently targeted to NMD (Kebaara and Atkin, 2009), and provide a mechanistic explanation by which only a subset of ribosome-associated uRNAs are sensitive to NMD.

Sensitivity of lncRNA to NMD Indicates Translation of "Noncoding" Transcripts in Mammals

As a means to evaluate whether predicted nonprotein-coding transcripts in higher eukaryotes also encode sORFs that are translated, we evaluated recent genome-wide gene expression and UPF1 protein binding data gathered from mouse embryonic stem cells (mESC; Hurt et al., 2013). Our analysis identified 519 annotated mRNAs whose expression increased >1.5-fold in cells inhibited for NMD versus control cells (of 13,043 expressed protein-coding genes; ~4%), many of which correspond to previously characterized NMD targets (Hurt et al., 2013). Strikingly, 46 transcripts classified as lncRNAs also increased >1.5-fold upon inhibition of NMD (of 265 lncRNA; Figure 4F). Consistent with these transcripts being direct targets for NMD, UPF1 binding sites were enriched 9.6-fold on these RNAs over

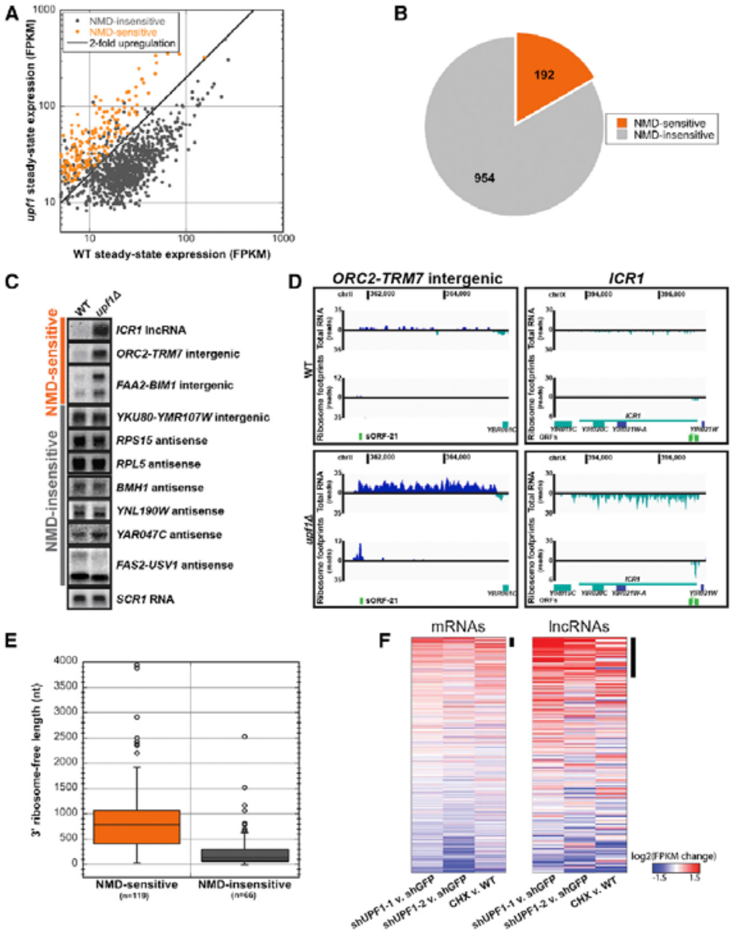


Figure 4. uRNAs Are Subject to Translation-Dependent Nonsense-Mediated RNA Decay

(A) uRNA expression levels (FPKM) in wild-type (WT) versus *upf1Δ* measured with RNA-seq reveals sensitivity to NMD. NMD-sensitive uRNAs exhibit ≥2-fold increase in steady-state levels in *upf1Δ* (statistically significant at an false discovery rate < 0.05 by Cuffdiff analysis; orange).

(B) Fraction of uRNAs showing sensitivity to NMD.

(C) Northern blot analysis of steady state RNA from WT and *upf1Δ* cells shows uRNAs and lncRNA *ICR1* predicted by RNA-seq to be regulated by NMD. Representative *SCR1* loading control is shown.

(D) Sequence coverage for NMD-sensitive uRNA or lncRNA *ICR1* in WT (top) or NMD-deficient (*upf1Δ*) cells (bottom). Data are presented as in Figure 2C.

(E) Length distribution of downstream ribosome-free regions for NMD-sensitive and -insensitive uRNAs. Box includes 25th to 75th percentiles; whiskers indicate ± 1.5 IQRs, with outliers indicated by circles.

(F) Change in mRNA and lncRNA expression in each of three NMD inhibition experiments in mESCs (shRNA UPF1-1, shRNA UPF1-2, and cycloheximide [CHX] treatment; Hurt et al., 2013). Changes are log2 expression (FPKM) ratios over control, averaged over two replicates. Potential NMD targets, defined as genes derepressed >1.5-fold, are highlighted (black bar).

See also Figure S4.

NMD-insensitive lncRNAs (Figures S4D and S4E). These data provide evidence that a number of mammalian lncRNAs lacking predicted protein-coding potential are engaged in active translation.

In addition to a similar proportion of uRNAs and lncRNAs being sensitive to perturbations in the NMD pathway (16.8% and 17.4% in yeast and mESC, respectively), we observed that ribosome footprints are enriched specifically on NMD-sensitive lncRNAs upon NMD inhibition compared to NMD-insensitive transcripts, and that these 80S ribosome-protected fragments map proximal to the transcript 5′ end (Figure S4F). Based on the observation that 3′ UTR length also plays a role in targeting transcripts to NMD in mammalian cells (Singh et al., 2008), unprotected RNA downstream of putative coding regions within lncRNAs likely contributes to the sensitivity of these RNAs to NMD, and suggests a common mechanism by which such transcripts are subject to regulation by this cellular RNA surveillance pathway.

DISCUSSION

Our analysis of the global landscape of expressed transcripts in yeast revealed hundreds of previously uncharacterized RNAs that do not map to annotated, protein-coding gene loci. We show by a number of means, including polyribosome analysis, ribosome profiling, and NMD sensitivity, that many of these unannotated transcripts are associated and/or actively engaged with translating ribosomes. Moreover, periodicity observed for a subset of ribosome-protected fragments facilitated precise demarcation of ORFs utilized by the translation machinery in vivo, providing heightened evidence for translation of defined short polypeptides encoded within a number of yeast uRNAs.

We demonstrate that a significant fraction of yeast uRNAs is sensitive to NMD, a translation-dependent surveillance pathway generally described to target mRNA. Moreover, analysis of published genome-wide expression data in mESC cells revealed a similar percentage of mammalian lncRNAs are also sensitive to NMD. Targeting of individual or subsets of predicted noncoding RNA to NMD has been previously observed in various organisms, including yeast (Thompson and Parker, 2007; Toesca et al., 2011), plants (Kurihara et al., 2009), and human cells (Tani et al., 2013), and these data argue that predicted noncoding RNAs are present in the cell cytoplasm and, contrary to expectations, engage the translation machinery. Our ribosome profiling data extend these observations and not only predict short protein-coding sequences within these transcripts, but also reveal extended regions of RNA downstream of predicted sORFs that are unprotected by 80S ribosomes. Importantly, these ribosome-free regions mimic long 3′ UTRs that commonly target mRNA to NMD in yeast and metazoa and provide a mechanistic explanation for how uRNAs (and lncRNAs) are targeted by this specialized decay pathway.

Whereas sensitivity to NMD provides compelling evidence supporting translation of sORFs encoded within uRNAs, we note that the accelerated degradation of transcripts targeted by NMD would reduce steady-state levels of these uRNAs and effectively dampen expression of any polypeptide encoded by the predicted sORF. Biologically, the sensitivity of uRNAs to NMD may serve to ensure that these transcripts maintain a primary role as functional RNA molecules, either in the nucleus as regulators of transcriptional events or in the cytoplasm as modulators of mRNA and/or protein function. Alternatively, the degradation of uRNAs by NMD may provide a unique means to regulate sORF expression, allowing robust accumulation of small polypeptides under conditions when NMD efficiency is reduced or inactivated (Huang and Wilkinson, 2012).

At present, we have demonstrated expression of polypeptides from two conserved yeast sORFs; however, a biological function for these and other predicted sORF translation products remains unclear. Notwithstanding, roles for small polypeptides in cellular function are well documented (Andrews and Rothnagel, 2014). In yeast, mating pheromones are 12 and 13 amino acids in length, and the large ribosomal protein L41 required for 25S rRNA folding is 25 amino acids long. Systematic analysis of annotated yeast mRNAs encoding small ORFs (<100 codons) revealed dozens that are important for cell growth under various conditions (Kastenmayer et al., 2006). Recently, functional small peptides have been found that are expressed from predicted nonprotein-coding RNAs in flies (Galindo et al., 2007; Magny et al., 2013) and zebrafish (Pauli et al., 2014), and short polypeptides derived from lncRNAs have been detected in human cells (Slavoff et al., 2013). We predict, therefore, that a number of sORFs identified in this study will express peptide products with important biological roles in yeast.

Transcriptome analysis of polyribosome-associated RNA revealed a large percentage of uRNAs associated with polysomes, similar to that observed for mRNAs. The distribution of uRNA association with polysomes was, however, clearly distinct from that of mRNA. We attribute this difference to functional heterogeneity within the class of uRNAs as compared to mRNAs. In contrast to mRNAs whose primary role is as templates for protein synthesis, uRNAs identified in our analysis include transcripts that we demonstrate are translated and ones for which there is limited association with the translational machinery. It will be of interest to evaluate uRNAs with low translatability scores for function as RNA regulators and accumulation in various compartments within the cell. Indeed, demonstration of several uRNAs of this type using single molecule fluorescence in situ hybridization suggests that these transcripts are enriched in the nucleus (data not shown). A likely role for these uRNAs is as regulators of gene expression through chromatin modification or influencing transcriptional events.

As a model eukaryote, *S. cerevisiae* has been the focus of extensive gene expression analysis and the proverbial guinea pig for many large-scale genomic and transcriptomic experimental studies. Because of this attention, the yeast genome has been described in exquisite detail and is currently annotated to express 6,380 protein-coding transcripts. As technologies measuring gene expression at finer resolution are developed or honed, previously undetected transcripts will continue to be uncovered. Our work adds to a number of recent studies identifying expression of RNA transcripts in yeast predicted to lack protein coding potential. Although a majority of these RNAs (and similar noncoding RNAs in metazoa) lack characterized function in the cell, we show here that a number encode predicted sORFs exploited by the translational machinery for the expression of small polypeptides, some of which demonstrate evolutionary conservation. Our present findings reveal additional protein coding capacity within the yeast genome, but it will not be unexpected to learn that the remarkable complexity that continues to be uncovered in this single-celled eukaryote will also be found hidden in the genomes of other, more complex organisms, including humans.

EXPERIMENTAL PROCEDURES

Yeast Culture and Standard Methods

Cells were grown under standard conditions, unless otherwise noted. Yeast strains, plasmids, and oligonucleotides are listed in Table S5. RNA isolation, and Northern and western blot analyses were performed as previously described (Geisler et al., 2012). Epitope-tagged sORFs were generated using homologous recombination (Longtine et al., 1998) or standard molecular cloning strategies.

Total RNA Library Preparation

Five micrograms of DNase I-treated whole-cell RNA was depleted of rRNA using Epicenter Human/Mouse/Rat RiboZero rRNA Removal Kit. Strand-specific, random-primed cDNA libraries were generated by the CWRU Genome and Transcriptome Sequencing Core using the Epicenter ScriptSeq v2 RNA-seq Library Preparation Kit.

Polysome-Associated RNA Library Preparation

Yeast whole-cell lysates were subjected to polyribosome analysis on a 15%–45% (w/w) sucrose gradient. RNA was extracted from fractions containing polyribosomes and pooled. Five micrograms of RNA was used to prepare libraries as described above.

Ribosome Profiling Library Preparation

Isolation and sequencing of ribosome-protected RNA fragments was performed based on the described protocol (Ingolia et al., 2012), with modifications as described in the Supplemental Experimental Procedures. For fragmented total RNA libraries, whole-cell RNA was purified, DNase-treated, and rRNA depleted as for the total RNA library preparation. RNA was fragmented with base as described (Ingolia, 2010), and 26–34 nt fragments gel-purified and used for library preparation.

RNA Sequencing

cDNA libraries were sequenced on the Illumina HiSeq platform. Details of sequencing data analysis can be found in the Supplemental Experimental Procedures.

Analysis of mESC Data

RNA-seq, Ribo-seq, and CLIP-seq data generated by Hurt et al. (2013) were downloaded from the Gene Expression Omnibus (GSE41785). Details of data analysis can be found in the Supplemental Experimental Procedures.

ACCESSION NUMBERS

The NCBI BioProject accession number for the RNA-seq and ribosome profiling data presented in this paper is PRJNA245106.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, four figures, and five tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2014.05.023.

ACKNOWLEDGMENTS

The authors thank Tim Nilsen for critical evaluation of the manuscript and members of the K.E.B. and J.C. labs for helpful insight into this work. This research was funded by the National Institute of General Medical Sciences (GM080465 to J.C. and GM095621 to K.E.B.) and the National Science Foundation (NSF1253788 to K.E.B.).

Received: February 4, 2014
Revised: April 3, 2014
Accepted: May 9, 2014
Published: June 12, 2014

REFERENCES

Andrews, S.J., and Rothnagel, J.A. (2014). Emerging evidence for functional peptides encoded by short open reading frames. Nat. Rev. Genet. *15*, 193–204.

Bánfai, B., Jia, H., Khatun, J., Wood, E., Risk, B., Gundling, W.E., Jr., Kundaje, A., Gunawardena, H.P., Yu, Y., Xie, L., et al. (2012). Long noncoding RNAs are rarely translated in two human cell lines. Genome Res. *22*, 1646–1657.

Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M.; ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Brar, G.A., Yassour, M., Friedman, N., Regev, A., Ingolia, N.T., and Weissman, J.S. (2012). High-resolution view of the yeast meiotic program revealed by ribosome profiling. Science *335*, 552–557.

Chew, G.-L., Pauli, A., Rinn, J.L., Regev, A., Schier, A.F., and Valen, E. (2013). Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs. Development *140*, 2828–2834.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W., and Steinmetz, L.M. (2006). A high-resolution map of transcription in the yeast genome. Proc. Natl. Acad. Sci. USA *103*, 5320–5325.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789.

Galindo, M.I., Pueyo, J.I., Fouix, S., Bishop, S.A., and Couso, J.P. (2007). Peptides encoded by short ORFs control development and define a new eukaryotic gene family. PLoS Biol. *5*, e106.

Geisler, S., and Coller, J. (2013). RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. Nat. Rev. Mol. Cell Biol. *14*, 699–712.

Geisler, S., Lojek, L., Khalil, A.M., Baker, K.E., and Coller, J. (2012). Decapping of long noncoding RNAs regulates inducible genes. Mol. Cell *45*, 279–291.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 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*, 240–251.

Hu, W., Sweet, T.J., Chamnongpol, S., Baker, K.E., and Coller, J. (2009). Co-translational mRNA decay in *Saccharomyces cerevisiae*. Nature *461*, 225–229.

Huang, L., and Wilkinson, M.F. (2012). Regulation of nonsense-mediated mRNA decay. Wiley Interdiscip Rev RNA *3*, 807–828.

Hurt, J.A., Robertson, A.D., and Burge, C.B. (2013). Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay. Genome Res. *23*, 1636–1650.

Ingolia, N.T. (2010). Genome-wide translational profiling by ribosome footprinting. Methods Enzymol. *470*, 119–142.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223.

Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell *147*, 789–802.

Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M., and Weissman, J.S. (2012). The ribosome profiling strategy for monitoring translation *in vivo* by deep sequencing of ribosome-protected mRNA fragments. Nat. Protoc. *7*, 1534–1550.

Kastenmayer, J.P., Ni, L., Chu, A., Kitchen, L.E., Au, W.-C., Yang, H., Carter, C.D., Wheeler, D., Davis, R.W., Boeke, J.D., et al. (2006). Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae*. Genome Res. *16*, 365–373.

Kebaara, B.W., and Atkin, A.L. (2009). Long 3′-UTRs target wild-type mRNAs for nonsense-mediated mRNA decay in *Saccharomyces cerevisiae*. Nucleic Acids Res. *37*, 2771–2778.

Kozak, M. (1989). The scanning model for translation: an update. J. Cell Biol. *108*, 229–241.

Kurihara, Y., Matsui, A., Hanada, K., Kawashima, M., Ishida, J., Morosawa, T., Tanaka, M., Kaminuma, E., Mochizuki, Y., Matsushima, A., et al. (2009). Genome-wide suppression of aberrant mRNA-like noncoding RNAs by NMD in *Arabidopsis*. Proc. Natl. Acad. Sci. USA *106*, 2453–2458.

Kurtzman, C.P., and Robnett, C.J. (2003). Phylogenetic relationships among yeasts of the '*Saccharomyces* complex' determined from multigene sequence analyses. FEMS Yeast Res. *3*, 417–432.

Leeds, P., Peltz, S.W., Jacobson, A., and Culbertson, M.R. (1991). The product of the yeast UPF1 gene is required for rapid turnover of mRNAs containing a premature translational termination codon. Genes Dev. *5* (12A), 2303–2314.

Longtine, M.S., McKenzie, A., 3rd, Demarini, D.J., Shah, N.G., Wach, A., Brachat, A., Philippsen, P., and Pringle, J.R. (1998). Additional modules for versatile and economical PCR-based gene deletion and modification in Saccharomyces cerevisiae. Yeast *14*, 953–961.

Magny, E.G., Pueyo, J.I., Pearl, F.M., Cespedes, M.A., Niven, J.E., Bishop, S.A., and Couso, J.P. (2013). Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. Science *341*, 1116–1120.

Maquat, L.E. (2004). Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. Nat. Rev. Mol. Cell Biol. *5*, 89–99.

Muhlrad, D., and Parker, R. (1999). Aberrant mRNAs with extended 3′ UTRs are substrates for rapid degradation by mRNA surveillance. RNA *5*, 1299–1307.

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

Pauli, A., Norris, M.L., Valen, E., Chew, G.-L., Gagnon, J.A., Zimmerman, S., Mitchell, A., Ma, J., Dubrulle, J., Reyon, D., et al. (2014). Toddler: an embryonic signal that promotes cell movement via Apelin receptors. Science *343*, 1248636.

Roberts, A., Pimentel, H., Trapnell, C., and Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics *27*, 2325–2329.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050.

Singh, G., Rebbapragada, I., and Lykke-Andersen, J. (2008). A competition between stimulators and antagonists of Upf complex recruitment governs human nonsense-mediated mRNA decay. PLoS Biol. *6*, e111.

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

Tani, H., Torimura, M., and Akimitsu, N. (2013). The RNA degradation pathway regulates the function of GAS5 a non-coding RNA in mammalian cells. PLoS ONE *8*, e55684.

Thompson, D.M., and Parker, R. (2007). Cytoplasmic decay of intergenic transcripts in *Saccharomyces cerevisiae*. Mol. Cell. Biol. *27*, 92–101.

Toesca, I., Nery, C.R., Fernandez, C.F., Sayani, S., and Chanfreau, G.F. (2011). Cryptic transcription mediates repression of subtelomeric metal homeostasis genes. PLoS Genet. *7*, e1002163.

van Dijk, E.L., Chen, C.L., d'Aubenton-Carafa, Y., Gourvennec, S., Kwapisz, M., Roche, V., Bertrand, C., Silvain, M., Legoix-Né, P., Loeillet, S., et al. (2011). XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. Nature *475*, 114–117.

van Heesch, S., van Iterson, M., Jacobi, J., Boymans, S., Essers, P.B., de Bruijn, E., Hao, W., Macinnes, A.W., Cuppen, E., and Simonis, M. (2014). Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. Genome Biol. *15*, R6.

Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Münster, S., Camblong, J., Guffanti, E., Stutz, F., Huber, W., and Steinmetz, L.M. (2009). Bidirectional promoters generate pervasive transcription in yeast. Nature *457*, 1033–1037.

Zhang, Z., Li, J., Zhao, X.-Q., Wang, J., Wong, G.K.-S., and Yu, J. (2006). KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics *4*, 259–263.

Cell Reports, Volume 7

Supplemental Information

Translation of Small Open Reading Frames within Unannotated RNA Transcripts in *Saccharomyces cerevisiae*

Jenna E. Smith, Juan R. Alvarez-Dominguez, Nicholas Kline, Nathan J. Huynh, Sarah Geisler, Wenqian Hu, Jeff Coller, and Kristian E. Baker

Smith et al., Figure S1, related to Figure 1

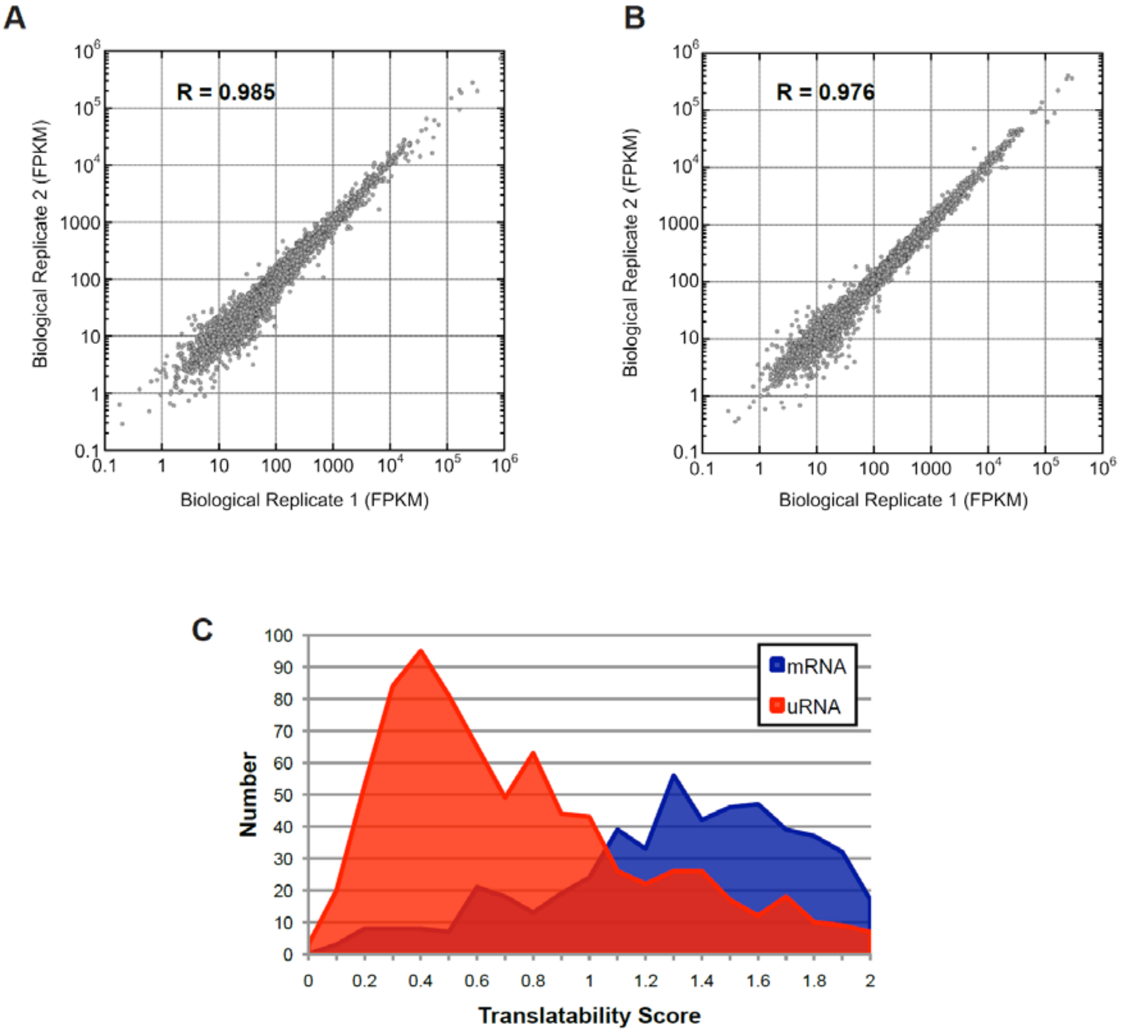


Figure S1, related to Figure 1. RNA-seq provides evidence that uRNAs associate with polyribosomes. (A) Regression analysis for wildtype biological replicates of RNA-seq for total RNA or (B) Polysome-seq for polysome-associated RNA. Correlation coefficients indicated. (C) Translatability score distribution for osubset of mRNAs and uRNAs 200-500 nt in length. mRNA n=571 (11.4% of total analyzed in Figure 1C; blue); uRNA n=824 (71.9% of total analyzed in Figure 1C; red).

Smith et al., Figure S2, related to Figure 2

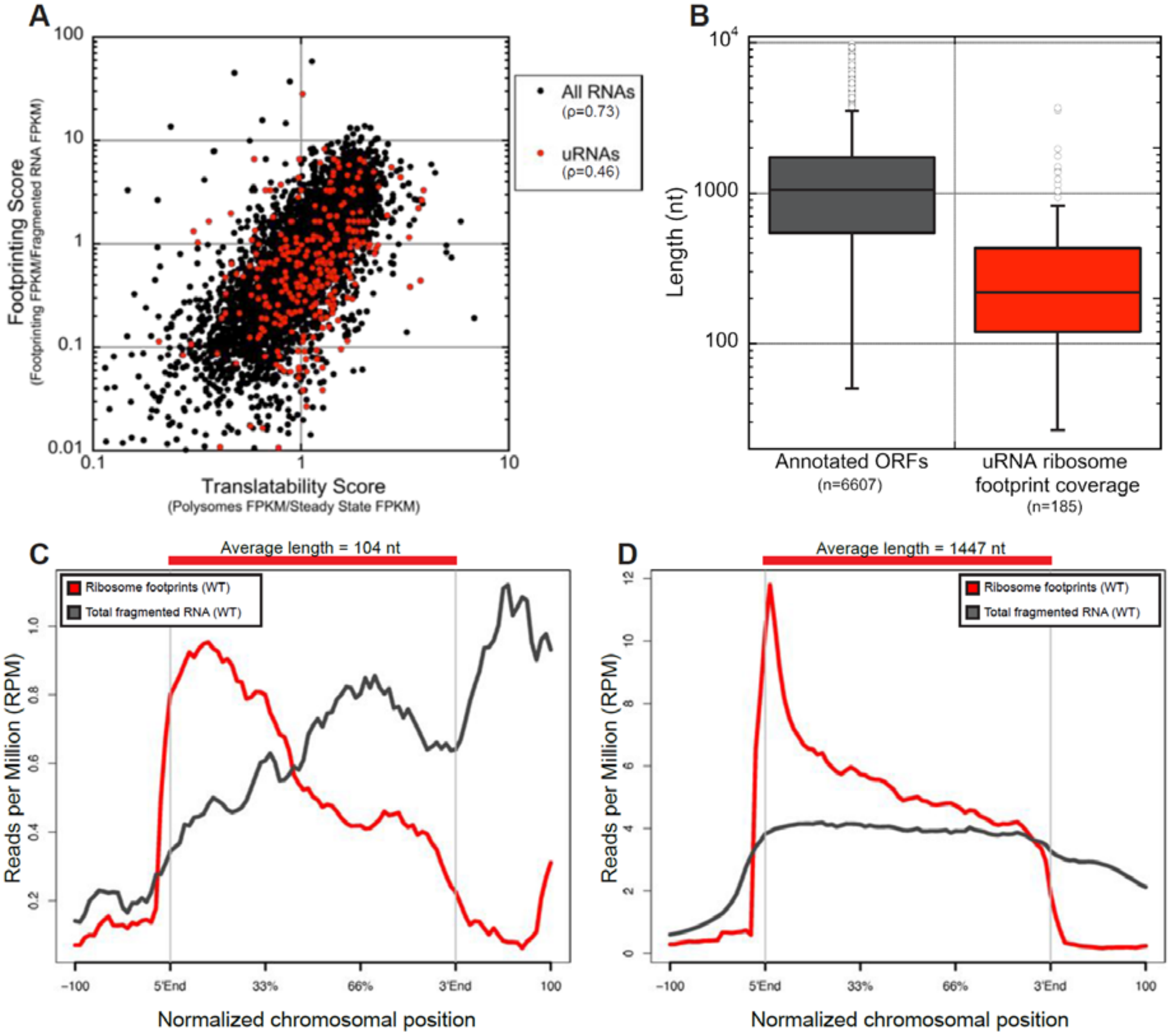


Figure S2, related to Figure 2. Ribosome footprinting predicts short open-reading frames encoded within uRNAs. (A) Comparison of polysome RNA-seq and ribosome profiling. The Translatability Score calculated from polysome RNA-seq and Footprinting Score calculated by ribosome footprinting were compared for all RNAs for which a score could be calculated for both assays. All RNAs (black) and uRNAs (red) are shown. Spearman rank correlation coefficients indicated. (B) Boxplot of the average length of yeast annotated ORFs (including verified, uncharacterized, and dubious ORFs) compared to the average size of the region covered by ribosome footprints for uRNAs. Box includes 25-75 percentiles; whiskers indicate ± 1.5 IQRs, with outliers indicated by circles. (C) and (D) Metagene plot of average sequencing coverage of ribosome footprints (red) or total fragmented RNA (gray) mapped along all predicted sORF coding regions (C) (n=43) or annotated mRNA CDS (D) (n=5017), plus 100 nucleotides flanking either end. 5'End indicates predicted or annotated start codon position; 3'End indicates predicted or annotated stop codon position. Red bar demarcates putative sORF metagene (C) or mRNA ORF metagene (D), with the average size indicated. Data are mean reads per million.

Smith et al., Figure S3, related to Figure 3

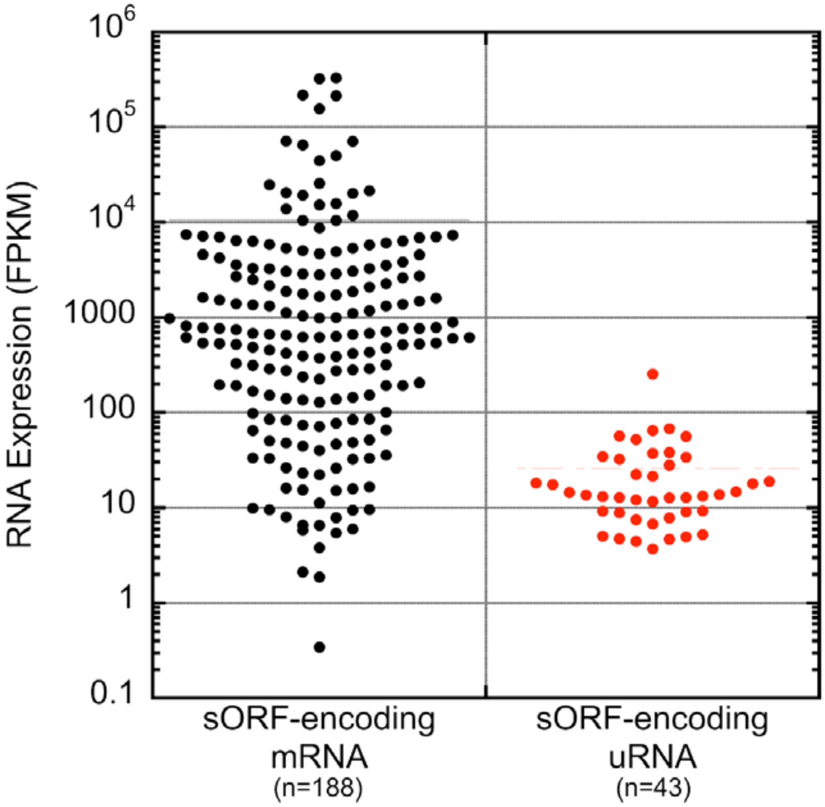


Figure S3, related to Figure 3. sORF-encoding uRNA expression is within range of mRNAs encoding short ORFs. Dot plot displaying expression levels (FPKM) for annotated sORF-encoding mRNAs (black), or uRNAs containing putative sORFs (red). Only mRNAs with an average expression of >0 in WT and >10 in *upf1Δ*, (expression thresholds for all analyses) are included.

Smith et al., Figure S4, related to Figure 4

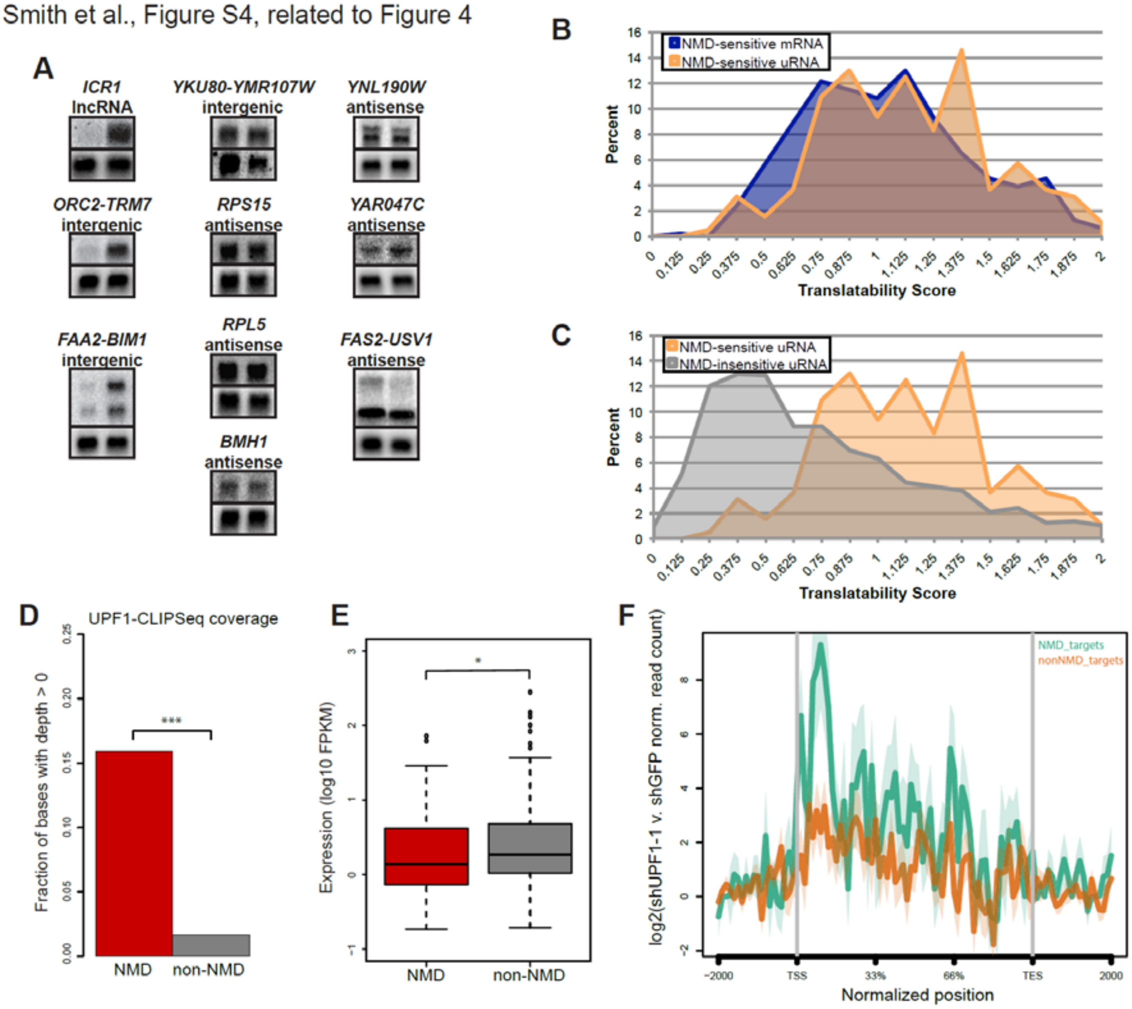


Figure S4, related to Figure 4. uRNAs are sensitive to translation-dependent nonsense-mediated RNA decay. (A) Northern blot analysis of steady state RNA from WT (lane 1) and *upf1Δ* cells (lane 2). Panels are identical to Figure 4C but include their respective *SCR1* loading control (bottom panels). (B) Translatability Score distribution for characterized mRNAs (blue) and uRNAs (orange) sensitive to NMD. (C) Translatability Score distribution for uRNAs sensitive to NMD (orange) vs. insensitive to NMD (gray). (D) Fraction of bases with coverage by UPF1 CLIP-seq tags for lncRNA transcripts sensitive to NMD (red) vs. NMD-insensitive lncRNAs (gray). UPF1 CLIP-seq tags are combined from three independent wild-type experiments. ***p<0.001 (Kolmogorov-Smirnov test). (E) Box-and-whisker plot of the distribution of gene-level expression values (FPKM) for lncRNAs sensitive to NMD (red) vs. those insensitive to NMD (gray). Expression values represent the average expression in wild-type cells from two independent experiments. *p<0.05 (Kolmogorov-Smirnov test). (F) Metagene plot of average change in the density of ribosome profiling reads mapping along the transcript body of intergenic lncRNAs sensitive to NMD (green) vs. those not sensitive to NMD (orange) following UPF1 depletion by shRNA. Changes are shown as log2 ratios of normalized Ribo-Seq read count in the shUPF1-1 experiment to that in the shGFP control. (TSS) transcription start site; (TES) transcription end site. Data are mean +/- SEM.

SUPPLEMENTAL TABLES

Table S1, related to Figures 1, 2, and 4. High-throughput sequencing statistics.

Sample	Replicate	# of Reads	Mapped reads		rRNA reads		non-rRNA mapped reads	
			#	%	#	% (of mapped reads)	#	% (of mapped reads
WT steady state	1	13,803,030	11,096,422	80.4	-	-	-	-
	2	31,213,437	22,617,117	72.5	-	-	-	-
upf1Δ steady state	1	15,855,617	12,523,774	79.0	-	-	-	-
	2	31,154,574	21,890,286	70.3	-	-	-	-
WT polysomes	1	31,342,788	22,141,454	70.6	-	-	-	-
	2	33,619,004	23,239,241	69.1	-	-	-	-
upf1Δ polysomes	1	25,964,168	17,821,788	68.6	-	-	-	-
	2	25,784,511	17,782,505	69.0	-	-	-	-
WT steady state fragmented RNA	-	7,341,479	5,507,344	75.0	526,798	9.6	4,980,546	90.4
upf1Δ steady state fragmented RNA	-	7,912,713	5,759,010	72.8	219,344	2.9	5,539,666	96.2
WT ribosome profiling	1	28,404,464	26,963,688	94.9	20,725,760	75.1	6,237,928	23.1
	2	22,793,695	21,682,577	95.1	17,916,898	80.8	3,765,679	17.4
upf1Δ ribosome profiling	1	21,727,393	20,603,069	94.8	16,440,899	77.9	4,162,170	20.2
	2	23,351,093	21,834,783	93.5	16,885,425	75.6	4,949,358	22.7

Biological replicates of steady-state RNA-seq, Polysome-seq, and ribosome profiling were assayed. Total number of reads before mapping to the sacCer2 yeast reference genome, following mapping (see Supplemental Experimental Procedures), and percentage mapped reads listed for all datasets. Additionally, total number and percentage of rRNA reads, and number and percentage non-rRNA mapped reads listed for ribosome profiling.

Table S2, related to Figure 1. List of uRNAs investigated in this study. All uRNAs identified in this study are listed. For each uRNA, a unique locus identifier, the chromosomal coordinates, strand of expression, and proximal annotated features are listed. uRNAs are referred to in figures based on the "Orientation to reference annotations" column. [a]Coordinates as defined by Cufflinks and based on sacCer2 genome assembly. [b]All coordinates listed in ascending order; for Crick (-) strand, 5' end originates at the second coordinate. (See Excel file)

Table S3, related to Figure 2. List of putative sORFs identified by ribosome profiling.

sORF Number	Encompassing uRNA	sORF Coordinates[ab]	Putative peptide
sORF-1	XLOC_000132-	chrII:169873-169637	MQRVRIGQWVYDMEAIHRSDSHECPKRTCGNQTLNPGSIIKEIQYKKYRYILFPPIAANQFTPGCSEYIPILHDAIKD*
sORF-2	XLOC_000464-	chrIII:309884-309747	MEPPIFIILTILSKLTFDKDDEQTTRRSKALEYCSVQPLSSNIAI*
sORF-3	XLOC_002595+	chrXII:675730-675861	MRSLKCSILMPFFSQIFSVSILAWDALSNKTLLTRSNAKVEVN*
sORF-4	XLOC_002893+	chrXIII:480923-481186	MISMEAINNFIKTAPKHDYLTGGVHHSGNVDVLQLSGNKEDGSLVWNHTFVDVDNNVVAKFEDALEKLESLHRRSSSSTGNEEHANV*
sORF-5	XLOC_000429+	chrIII:242629-242685	MLFHGFVSSKGLSVVPQQ*
sORF-6	XLOC_000768-	chrIV:916135-916022	MNLNRATEKTGRHNKKFISCMTYVTVNYIPNKLYIEK*
sORF-7	XLOC_002334-	chrXI:513546-513430	MMYFTRMERPQESTKQNMMLKYHLFSNIHIASILSYAV*
sORF-8	XLOC_002919-	chrXIII:619370-619098	MTEMTLLPKKKSISIHSSKYSGIQLSWFENLFSSSEEITSPSRLLTTLLTILFRYESSFFSPKILTVLNPIPKCGYEPILMASETFVDSS*
sORF-9	XLOC_003873+	chrXVI:777577-777669	MLVINISQLKEIKLKYSTKTYFRIKFLGGT*
sORF-10	XLOC_001697+	chrVII:902532-902762	MGCISVSHILKLSNSPSKKRDLITRLNQRVTSGCLDLHVSSKKLINPSEKAAAEIKSARVSWSSQTRYCCFFNLFS*
sORF-11	XLOC_002196+	chrXI:66560-66604	MMEMDSGCDCVVKM*
sORF-12	XLOC_002899+	chrXIII:502246-502341	MRCLIPLPKWNANRYAAEPLATHWDHLGIFS*
sORF-13	XLOC_000636+	chrIV:525039-525113	MFKNSKNLSINPGNDADKASFSDP*
sORF-14	XLOC_000636+	chrIV:525082-525150	MRIRRASPTHEHNRSSSVLVVG*
sORF-15	XLOC_003583-	chrXV:969893-969765	MKRKKRNVIIGMMPFVALTSTTSIYDNGNMCDQSELCSCHLH*
sORF-16	XLOC_003728+	chrXVI:286664-286765	MDMHYVGHMTLIVQVQLMYRLKRLRLLIYNAIY*
sORF-17	XLOC_000364+	chrIII:39662-39805	MNIHKIKIICMFLIYRYTIKHFQEFVCRLFARDLIKVLPTKINLYFF*
sORF-18	XLOC_000631+	chrIV:513023-513172	MLVKDYSIGYTEFTRTPPQGYRNLHKGIDISNISSNIVIFLFILCVVSH*
sORF-19	XLOC_003307+	chrXV:38888-39046	MLSLTFISPTLSQIFDHVIYMLFSNQVINFTELKVAAAENSHLIILYIDPTY*
sORF-20	XLOC_003329+	chrXV:96656-96823	MTCTYIVVNVNMSGLPMVQMKEKFVFGCTELLRIEEGLIFHTCFYVLVLASNHPY*
sORF-21	XLOC_000204+	chrII:362898-362939	MSRQLVTLLLVYT*
sORF-22	XLOC_000631+	chrIV:512693-512740	MYILNVYALNTIDDF*
sORF-23	XLOC_001229+	chrV:285208-285264	MPAVLMVNPCPISLRHFL*
sORF-24	XLOC_001689+	chrVII:875763-875801	MSLVFIQPHFIF*

sORF-25	XLOC_003019+	chrXIII:873056-873103	MPYITNTAEATMSTV*
sORF-26	XLOC_003128+	chrXIV:270232-270270	MKYYKKFINFLN*
sORF-27	XLOC_001405+	chrVII:18942-19028	MFNIFAMIHSRFCPALNFTPTLRVHSVR*
sORF-28	XLOC_000405+	chrIII:169068-169193	YLHSRYRHINIKILNIETSFRLRFGCRKPKMQFKWVNNLNG*
sORF-29	XLOC_000841+	chrIV:1206674-1206796	MLPRLLKHVGIEINYHLLTSIYITSILSYTVLEDDANDEK*
sORF-30	XLOC_001222+	chrV:268931-268963	MYGCARHSSS*
sORF-31	XLOC_003621+	chrXV:1003992-1004036	MLVLNMSIQSLVVH*
sORF-32	XLOC_000144-	chrII:220889-220800	MSFPYEHAQAKNLPEILLYYKKKEMNLSK*
sORF-33	XLOC_000337-	chrII:792024-791974	MSFQRLKLLKTALFVY*
sORF-34	XLOC_000617-	chrIV:471496-471440	MNTVTINKAGLTEHVGSG*
sORF-35	XLOC_000803-	chrIV:1019114-1019028	MSWKLLHFLPFEISSYMYMLTLLTSKMS*
sORF-36	XLOC_001190-	chrV:177560-177402	MVFIRDCVMNSFHYVNEPRSNVNQRTCGQESYNYLFNEPKEITCSRLGVNLF*
sORF-37	XLOC_001277-	chrV:432186-432139	MQSRNKKQIAISSPL*
sORF-38	XLOC_001365-	chrVI:159009-158926	MALAIRTRMHNRQDIIGMQQLKLLLML*
sORF-39	XLOC_001409-	chrVII:17702-17652	MSNSTILENNTIVRCI*
sORF-40	XLOC_001678-	chrVII:811263-811213	MRIYTLHTYYRKHCYF*
sORF-41	XLOC_001678-	chrVII:811200-811162	MANERSSFIPLS*
sORF-42	XLOC_002811-	chrXIII:311954-311892	MHLRDKVKLPSYSCKRNLYF*
sORF-43	XLOC_003248-	chrXV:5661-5560	MIRMSWWPCITNIGSSRGLETLNAVRMDIYIGD*
sORF-44	XLOC_003248-	chrXV:6801-6511	MHRIPCQNVFCYNTFHEANWGYYLHLSGFHICFLDNSFNSSIVVRMAMRIYYGTHRFLRTMSIVKFKRLLGSLCGKKRINDYQRSITFDYSHIRDI*
sORF-45	XLOC_003866-	chrXVI:781588-781517	MVSLDSLLVLLKKNGILFLDNVS*
sORF-46	XLOC_003609-	chrXV:1064363-1064313	MNGRKNCSFEECFSGR*
sORF-47	XLOC_003356-	chrXV:326433-326329	MNLHVFIKLIEMEFSSSSRSVFCRLCSYDAKRLK*

All putative sORFs defined based on phased ribosome footprints (See Supplemental Experimental Procedures), the uRNA by which they are encoded, the chromosomal coordinates for the sORF, and the putative peptide sequence are presented. [a]All chromosomal coordinates based on sacCer2 genome annotation. [b]For Crick strand, coordinates are listed in descending (5'-3') order.

Table S4, related to Figure 3. **Conservation of sORFs in other yeast species.** For each sORF, the phastCons log-odds score for previously identified conserved elements (Siepel et al., 2005), TBLASTN results (percent identical residues relative to full-length putative peptide, E-values, and bit scores), and Ka/Ks ratios (Zhang et al., 2006) are presented. [a]Bold indicates phastCons conserved element completely overlaps sORF. [b]Italics indicates that conserved element may be influenced by gene antisense to sORF uRNA. [c]N/A indicates no conserved element corresponds to sORF locus. [d]Percent identity score of 0 indicates no alignment found at E<10. [e]For all comparisons where BLAST produced no match, "-" is recorded. [f]N.D. = not determined; nucleotide sequences show 100% alignment. [g]N.S. = not significant; value not reported due to Fisher's p-value >0.05. [h]Bold indicates Ka/Ks ratio supports purifying selection. [i]All data reported as for *S. pastorianus*. (See Excel file)

Table S5, related to Experimental Procedures. List of strains, oligos, and plasmids.

Name	Description[a]	Notes	Reference
yKB154	MATa, *ura3*, *leu2*, *his3*, *met15*	Wild-type	EUROSCARF
yKB146	MATa, *ura3*, *leu2*, *his3*, *met15*, *upf1*::KAN	*upf1Δ*	EUROSCARF
yKB596	MATa, *ura3*, *leu2*, *his3*, *met15*, sORF-4-HA::HIS3	sORF-4 plus 3xHA C-terminal tag	This study
yKB597	MATa, *ura3*, *leu2*, *his3*, *met15*, sORF-4-HA::HIS3	sORF-4 plus out-of-frame 3xHA C-terminal tag	This study
oJC1348	GTCATGCTCCTTTTTATGGGTTCTCGTCGTAATAATCCTG	*ORC2-TRM7* intergenic uRNA oligo probe	This study
oJC1352	ACCTGAAAGAGACGCCTTGTATCTTCTATAGGTCAACTAG	*FAA2-BIM1* intergenic uRNA oligo probe	This study
oJC1917	GTTATTCTATTCTTGAGCAGGCACTTTTAGGGTTGGGCAA	*ICR1* ncRNA oligo probe	This study
oJC1981	GTATGGTTCCATACTAAACTACCATCTTCTTTATTGCCGC	*YKU80-YMR107W* intergenic uRNA oligo probe	This study
oJC306	GTCTAGCCGCGAGGAAGG	*SCR1* ncRNA oligo probe	This study
oJC1984	GAAATGTCCACTGAAGATTTCGTCAAGTTGGCCCC	*RPS15* antisense uRNA reverse PCR primer to make template for asymmetric PCR	This study
oKB707	CTGGAACAATGATCATGTTTCTCATGTGGGTTCTGACTGGAGC	*RPS15* antisense uRNA forward PCR primer to make template for asymmetric PCR	This study
oKB708	AGCTAGAGTTAGAAGAAGATTTGCCCGTG	*RPS15* antisense uRNA asymmetric PCR primer for Northern probe	This study
oJC1989	CAACACAAGGCCAAGTACAACACTCCAAAGTACAGATTGG	*RPL5* antisense uRNA reverse PCR primer to make template for asymmetric PCR	This study
oKB713	CAACTTCTTCAACACCCTTGTAAGTTTCGTCCAAACC	*RPL5* antisense uRNA forward PCR primer to make template for asymmetric PCR	This study
oKB714	CTGTCAAATCATCTCTTCTACCATCACTGGTG	*RPL5* antisense uRNA asymmetric PCR primer for Northern probe	This study
oJC1991	GTCGAGTTGATTTGTTCGTACCGTTCGAAGATTGAGACCG	*BMH1* antisense uRNA reverse PCR primer to make template for asymmetric PCR	This study
oKB711	AGAGAAGTTAAGAGCCAAACCTAGACGGATTGGGTGAG	*BMH1* antisense uRNA forward PCR primer to make template for asymmetric PCR	This study
oKB712	AACTAACTAAGATCTCCGACGATATTTTGTCCG	*BMH1* antisense uRNA asymmetric PCR primer for Northern probe	This study
oKB702	CGTAAGAACAATGCCGCCCCTGGTCCATCTAATTTCAACT	*YNL190W* antisense uRNA reverse PCR primer to make template for asymmetric PCR	This study
oKB717	CACTTTTGCACAAGCACACGTAAACACATAGTAGTCGAAATAG	*YNL190W* antisense uRNA forward PCR primer to make template for asymmetric PCR	This study
oKB718	CCATAAAATTGTTTGGTGTTACCGCTGGTAG	*YNL190W* antisense uRNA asymmetric PCR primer for Northern probe	This study
oKB700	GTTCCTCGATCGACTAGTGCCATTCAATGAGATAAGGAGT	*YAR047C* antisense uRNA reverse PCR primer to make template for asymmetric PCR	This study
oKB720	GAGCAGAGGTTAGCTCCGTCTCAACCAATTTTGTAC	*YAR047C* antisense uRNA forward PCR primer to make template for asymmetric PCR	This study
oKB721	AGTATAGTAAGATATAATCCCACTAACGATTAGCGAGTG	*YAR047C* antisense uRNA asymmetric PCR primer for Northern probe	This study
oKB748	CTTCCAGAGCGCCAGCATCGATCATAGCTG	*FAS2-USV1* antisense uRNA forward PCR primer to make template for asymmetric PCR	This study
oKB750	CTGGTGGGTTTACTATTACTGTCGCTAGAAAATACTTACAAACTCGCTG	*FAS2-USV1* antisense uRNA reverse PCR primer to make template for asymmetric PCR	This study
oKB749	GGACTACCATCTGGTAGACAAGATGGTG	*FAS2-USV1* antisense uRNA asymmetric PCR primer for Northern probe	This study
oKB688	/5Phos/AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGC/iSp18/CACTCA/iSp18/TTCAGACGTGTGCTCTTCCGATCTATTGATGGTGCCTACAG	RT primer for ribosome profiling	Ingolia *et al.*, 2012

oKB689	AATGATACGGCGACCACCGAGATCTACAC	PCR amplification of ribosome profiling cDNA libraries, forward primer	Ingolia *et al.*, 2012
oKB690	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCG	PCR amplification of ribosome profiling cDNA libraries, reverse primer, Index #1	Ingolia *et al.*, 2012
oKB691	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTGCTCTTCCG	PCR amplification of ribosome profiling cDNA libraries, reverse primer, Index #2	Ingolia *et al.*, 2012
oKB692	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCG	PCR amplification of ribosome profiling cDNA libraries, reverse primer, Index #3	Ingolia *et al.*, 2012
oKB693	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCGG	PCR amplification of ribosome profiling cDNA libraries, reverse primer, Index #4	Ingolia *et al.*, 2012
oKB694	CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTCAGACGTGTGCTCTTCCG	PCR amplification of ribosome profiling cDNA libraries, reverse primer, Index #5	Ingolia *et al.*, 2012
oKB695	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTGCTCTTCCG	PCR amplification of ribosome profiling cDNA libraries, reverse primer, Index #6	Ingolia *et al.*, 2012
oKB769	CTCATCCTCATCCACAGGCAATGAAGAACACGCTAACGTTCGGATCCCCGGGTTAATTAA	Forward PCR primer to amplify 3xHA-His3 product from pFA6a-3HA-His3MX6, with gene-specific sequences for chromosomal tagging of sORF-4	This study
oKB770	CTTATTTCTCACATCATTATGAAGTGACTCCCCTCGGTTAGAATTCGAGCTCGTTTAAAC	Reverse PCR primer to amplify 3xHA-His3 product from pFA6a-3HA-His3MX6, with gene-specific sequences for chromosomal tagging of sORF-4	This study
oKB784	CTCATCCTCATCCACAGGCAATGAAGAACACGCTAACGTCGGATCCCCGGGTTAATTAA	Forward PCR primer to amplify 3xHA-His3 product from pFA6a-3HA-His3MX6, with gene-specific sequences for chromosomal tagging of sORF-4 out-of-frame	This study
oKB789	TTCCTTACGGAACCCAAGTGTG	Forward PCR primer to amplify *YBL027W-YBL026W* intergenic uRNA +/- 500 bp, for generation of pKB561	This study
oKB790	TTACTGTATCTACATCGGGATACTAATAGTAC	Reverse PCR primer to amplify *YBL027W-YBL026W* intergenic uRNA +/- 500 bp, for generation of pKB561	This study
oKB791	ATACCAATTTTACACGATGCCATAAAGGACGATTATAAAGATGATGATGATAAATAGACAAGCTACGTTGAAACAAGAACCCGC	Forward PCR primer to insert 1XFLAG tag at C-terminus of sORF-1 in *YBL027W-YBL026W* intergenic uRNA in pKB561, for generation of pKB565	This study
oKB792	GCGGGTTCTTGTTTCAACGTAGCTTGTCTATTTATCATCATCATCTTTATAATCGTCCTTTATGGCATCGTGTAAAATTGGTAT	Reverse PCR primer to insert 1XFLAG tag at C-terminus of sORF-1 in *YBL027W-YBL026W* intergenic uRNA in pKB561, for generation of pKB565	This study
oKB797	GGTACTTCCGCTAATAGACTACAAAC	Forward PCR primer to amplify *YKU80-YMR107W* intergenic uRNA +/- 500 bp, for generation of pKB562	This study
oKB798	GTTCGTACTTCCTTCTGAGCAG	Reverse PCR primer to amplify *YKU80-YMR107W* intergenic uRNA +/- 500 bp, for generation of pKB562	This study
oKB799	TCCACAGGCAATGAAGAACACGCTAACGTTGATTATAAAGATGATGATGATAAATAACCGAGGGGAGTCACTTCATAATGATGT	Forward PCR primer to insert 1XFLAG tag at C-terminus of sORF-4 in *YKU80-YMR107W* intergenic uRNA in pKB562, for generation of pKB566	This study
oKB800	ACATCATTATGAAGTGACTCCCCTCGGTTATTTATCATCATCATCTTTATAATCAACGTTAGCGTGTTCTTCATTGCCTGTGGA	Reverse PCR primer to insert 1XFLAG tag at C-terminus of sORF-4 in *YKU80-YMR107W* intergenic uRNA in pKB562, for generation of pKB566	This study
yEpLac181	2μ, LEU2	Parental vector used to construct pKB561, pKB562, pKB565, pKB566	Gietz and Sugino, 1988
pKB561	*YBL027W-YBL026W* intergenic uRNA +/- 500 bp		This study
pKB562	*YKU80-YMR107W* intergenic uRNA +/- 500 bp		This study
pKB565	*YBL027W-YBL026W* intergenic uRNA +/- 500 bp + C-terminal FLAG		This study
pKB566	*YKU80-YMR107W* intergenic uRNA +/- 500 bp + C-terminal FLAG		This study
pFA6a-3HA -His3MX6		Vector used to construct chromosomal 3xHA-tagged sORF loci	Longtine et al., 1998

All yeast strains, oligonucleotides, and plasmids used to generate data or constructs presented in this study are provided. Description (including nucleotide sequences of oligonucleotides), notes regarding the context of their use, and source for all reagents are provided. [a]All oligonucleotide sequences are listed from 5'-3'.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Yeast Culture

Cells were grown at 30 °C in synthetic medium plus 2% glucose and appropriate amino acids at 250 RPM to mid-log phase, unless otherwise noted. Yeast strains, plasmids, and oligonucleotides (Integrated DNA Technologies) can be found in Table S5.

Total RNA Library Preparation

Whole-cell RNA was isolated using glass-bead cell lysis and phenol extraction as previously described (Geisler et al., 2012). 5 µg of DNase I-treated (Roche 04716728001) whole-cell RNA was depleted of rRNA using the Human/Mouse/Rat RiboZero rRNA removal kit (Epicentre MRZH11124). Small RNAs were excluded using RNA Clean and Concentrator-5 spin columns (Zymo R1015), substituting 26.6% ethanol final concentration at steps 1-2 of the manufacturer's recommended protocol to enhance removal of RNAs <200 nt (data not shown). Strand-specific, random-primed cDNA libraries were generated by the CWRU Genome and Transcriptome Sequencing Core, using the ScriptSeq v2 RNA-Seq Library Preparation Kit (Epicentre SSV21106) and ScriptSeq Index PCR Primers. Libraries were prepared for biological replicates of WT and *upf1Δ* strains.

Polyribosome Analysis

Yeast cultures were grown to mid-log phase, treated with 100 µg/mL cycloheximide (CHX), harvested immediately by centrifugation, and cell pellets flash frozen on dry ice. Lysis was carried out at 4 °C. Cell pellets were lysed in polysome lysis buffer (10 mM Tris, pH 7.4, 100 mM NaCl, 30 mM $MgCl_2$, 100 µg/mL CHX, 1 mM DTT) by mechanical disruption using glass beads. Cell debris was removed by centrifugation through an 18 Ga puncture hole for 2 minutes at 2000 RPM, and the resulting lysate was pre-cleared at 29,000 RPM for 10 minutes in a Beckman TLA-120.2 rotor. Lysate was treated on ice with 1% Triton X-100 for 5 minutes. 10 units (OD_{260}) of lysate were added to a 15-45% (w/w) sucrose gradient (buffer 50 mM Tris acetate, pH 7.0, 50 mM NH_4Cl, 12 mM $MgCl_2$, 1 mM DTT) prepared using a Biocomp gradient maker. Gradients were centrifuged for 2:26 hr at 41,000 RPM in a Beckman Sw-41Ti rotor. Gradients were fractionated, and RNA was precipitated and extracted as described previously (Sweet et al., 2012). 5 µg of RNA was used to prepare polysome-seq libraries as described above for total RNA libraries.

RNA-Seq and Polysome-Seq Sequencing and Analysis

Sequencing and mapping: cDNA libraries prepared from total and polysome-associated RNA were sequenced on the Illumina HiSeq2000 platform at the Institute for Integrative Genome Biology High-Throughput Sequencing Core at the University of California, Riverside on a single-end, 100 cycle flow cell. On the Galaxy platform (usegalaxy.org; Goecks et al., 2010; Blankenberg et al., 2010; Giardine et al., 2005), the sequencing data FASTQ files were run through the "NGS: QC and manipulation/ FASTQ Groomer" tool. "NGS: QC and manipulation/Compute quality statistics" was used to compute quality scores and 1 low-quality nucleotide was trimmed from the right end of all reads during mapping. Reads were mapped to the sacCer2 genome on Galaxy with "NGS: Mapping/Map with Bowtie for Illumina" (Langmead et al., 2009) using a SOAP-like alignment policy to allow 2 mismatches over the entire length of the read (-v 2), and excluding any read that did not map uniquely to the genome (-m 1).

Identification of unannotated RNAs: Reads were assembled into transcripts using Cufflinks v2.1.1 (Trapnell et al., 2010) with bias correction and multi-read correction, using reference annotation-based transcript assembly (RABT; Roberts et al., 2011) to identify unannotated transcripts (-GTF-guide -b -u --library-type ff-firststrand; all other parameters default). The sacCer2 Ensembl Genes annotation downloaded from the UCSC genome table browser was used as a guide during transcript assembly (genome.ucsc.edu/cgi-bin/hgTables?command=start). RABT assembles reads into transcripts, and then compares assembled transcripts to the reference genome annotation to identify transcripts significantly different from transcripts predicted by the annotation (Roberts et al., 2011), allowing the identification of novel transcripts that map to regions of the sacCer2 genome lacking annotated features.

Using the default --overlap-radius option, unique transcripts must be separated by at least 50 basepairs from annotated transcripts to prevent merging either at the Cufflinks step or following Cuffmerge step. Transcripts <200 nucleotides or with a coverage <1 read per million were filtered from the dataset. A master annotation compiling RNAs detected in all datasets was generated with Cuffmerge (Cufflinks v2.1.1). Notably, Cuffmerge includes a step that filters transcripts likely to be artifacts including possible polymerase run-on fragments, or transcripts within 2 kilobases downstream of a reference transcript (Trapnell et al., 2012).

For classification of RNAs: "mRNAs" include any gene annotated with a YXXNNNX systematic name; "known ncRNAs" include C/D box snoRNAs (*snR18, snR65, snR4, snR71, snR76, snR45, snR63, snR128, snR190, snR70*), H/ACA box snoRNAs (*snR46, snR30, snR44, snR34, snR11, snR49, snR81, snR8, snR5, snR161, snR43, snR189, snR84, snR80, snR37, snR42, snR85, snR86, snR191, snR9, snR36, snR35, snR31*), spliceosomal RNAs (*snR14, snR7-L, snR19, LSR1*), U3 snoRNA *snR17b*, telomerase RNA *TLC1*, signal recognition particle 7S RNA *SCR1*, RNase MRP *NME1*, and RNase P component *RPR1*; uRNAs include all assembled transcripts that were not assigned and did not align to a reference annotation, excluding those mapping to mitochondrial DNA. Any assembled transcript spanning more than one annotated chromosomal feature was excluded from all downstream analysis.

Quantification of expression: Expression (FPKM) was calculated using Cuffdiff (Cufflinks v2.1.1; Trapnell et al., 2013) with bias correction and multi-read correction, providing biological replicates for analysis, with the master annotation generated by Cuffmerge above as a reference (-b -u --library-type ff-firststrand). Any RNAs which did not have an average expression in RNA-seq datasets of FPKM ≥10 in *upf1Δ*, and FPKM >10 in wild-type for ncRNAs, were excluded from further analysis.

Comparison to previous ncRNA transcripts: Comparison of uRNAs to previous transcript annotations (DCP2-sensitive, Geiser et al. 2012; SUTs, CUTs, Xu et al., 2009; or XUTs, van Dijk et al., 2011) was performed by manually comparing the published chromosomal coordinates from each of these 4 classes of transcripts to the coordinates of uRNAs defined by Cufflinks analysis. If a uRNA overlapped a previously classified ncRNA >50%, or vice versa, the uRNA was categorized as being identical to or overlapping a member of that class. In many cases, ncRNAs have already been previously classified in more than one category. For example, when XUTs were described, 543 were identified as also being SUTs and 183 were identified as also being CUTs (van Dijk et al., 2011); this ambiguity between classes is reflected in the fact that many uRNAs overlap ncRNAs in more than one class. Additionally, in some cases uRNA annotations spanned adjacent but non-overlapping ncRNAs, also resulting in grouping of the uRNA into more than one class; however, in these cases the uRNA transcript isoform described here is likely to have distinct stability characteristics from overlapping ncRNAs.

Translatability Score calculation: For each detected RNA, we calculated the ratio of RNA-Seq reads associated with polysomes (Polysome-seq data) relative to reads from total RNA at steady-state (RNA-seq data), to calculate the Translatability Score ($FPKM_{polysomes}/FPKM_{steady\text{-}state}$). Sequencing datasets were normalized for *PGK1* and *RPL41A* mRNAs to have a translatability score of 1. All graphical representations of translatability score data are presented as histograms of the number of RNAs per bin, generated with 40 bins from scores 0-5.

Identification of NMD-sensitive RNAs: RNAs were identified as upregulated in *upf1Δ* by comparing WT and *upf1Δ* total RNA samples using Cuffdiff with parameters as described above. Upregulated transcripts were required to be statistically significant at an FDR of <0.05 and show a ≥2-fold average increase in expression.

Ribosome Profiling Library Preparation

Isolation and sequencing of ribosome-protected RNA fragments was performed based on the described protocol (Ingolia et al., 2012), with the following modifications. Yeast cultures were grown in synthetic dextrose medium plus amino acids to mid-log phase, treated with 100 μg/mL CHX, harvested immediately by centrifugation, and cell pellets flash frozen on dry ice. Lysis was carried out at 4 °C. Cell pellets were lysed in polysome lysis buffer (10 mM Tris, pH 7.4, 100 mM NaCl, 30 mM $MgCl_2$, 100 µg/mL CHX, 1 mM DTT) by mechanical disruption using glass beads. Cell debris was removed by centrifugation through an 18 Ga puncture hole for 2 minutes at 2000 RPM, and the resulting lysate was pre-cleared at 14,000 RPM for 10 minutes in tabletop centrifuge. Lysates were treated with 1% Triton X-100 for 5 minutes. 12.5 units (OD_{260}) of lysate were treated with 188 U RNase I (Invitrogen AM2294) in 250 µL at 24 °C for 1hr. Lysates were loaded onto a 15-45% (w/w) sucrose gradient, centrifuged, and fractionated as described for polysome analysis above.

RNA was precipitated from fractions containing the 80S monosome peak with 2 volumes of 95% ethanol at -80 °C overnight, and centrifuged for 30 minutes at 13,200 RPM to collect RNA. RNA was resuspended in LET (25 mM Tris, pH 8.0, 100 mM LiCl, 20 mM EDTA) plus 1% SDS, and extracted once each with an equal volume of phenol/ LET, phenol/chloroform/LET, and chloroform. RNA was precipitated with 300 mM NaCl, 1.5 µL GlycoBlue, and >1 volume isopropanol for 30 minutes on dry ice. RNA was collected by centrifugation at 13,200 RPM for 30 minutes at 4 °C, air dried, and resuspended in 10 mM Tris, pH 8.0. RNA from all monosome fractions for each sample was pooled, and 5 µg aliquots depleted of ribosomal RNA using the Human/Mouse/Rat RiboZero rRNA removal kit (Epicentre MRZH11124). Each rRNA-depleted sample was purified through RNA Clean and Concentrator-5 spin columns (Zymo R1015), substituting 60% ethanol at steps 1-2 of the manufacturer's recommended protocol to facilitate purification of small RNAs.

Size-selection of 26-34 nt fragments of RNA was carried out by electrophoresis on a 15% denaturing polyacrylamide gel, excision, and gel purification as described (Ingolia et al., 2012). 2 aliquots per sample were pooled, and a second ribosomal RNA depletion was performed using the Epicentre Human/Mouse/Rat RiboZero kit (eliminating the 50 °C incubation step) and Zymo RNA Clean and Concentrator-5 spin columns to purify RNA as above. RNA was dephosphorylated, a 3' linker ligated, first-strand cDNA synthesized, and cDNA circularized as in described protocol, (Ingolia et al., 2012). cDNA libraries were amplified with 12-14 cycles of PCR with indexed primers (see Table S5).

To generate fragmented RNA control libraries, whole-cell RNA was purified, DNase-treated, and ribosomal RNA removed as described for the RNA-seq library preparation. RNA was fragmented with base as described (Ingolia, 2010) and fragments of 26-34 nt were gel purified and used for library preparation as described above for ribosome footprinting libraries. Libraries were prepared for biological replicates of WT and *upf1Δ* strains for ribosome footprinting, or a single replicate of each strain for the fragmented RNA control.

Ribosome Profiling/Fragmented RNA Sequencing and Analysis

Sequencing and mapping: cDNA libraries prepared for total fragmented RNA or ribosome footprints were sequenced on the Illumina HiSeq2500 platform at the Institute for Integrative Genome Biology High-Throughput Sequencing Core at the University of California, Riverside on a single-end, 50 cycle flow cell. Using the Galaxy platform (usegalaxy.org; Goecks et al., 2010; Blankenberg et al., 2010; Giardine et al., 2005), the sequencing data FASTQ files were run through the "NGS: QC and manipulation/FASTQ Groomer" tool. "NGS: QC and manipulation/Compute quality statistics" was used to compute quality scores which indicated high-quality sequencing across the length of the reads. Data processing was carried out in Galaxy as described (Ingolia et al., 2012). Briefly, the sequencing adaptor was clipped from the 3' end of each read with "NGS: QC and manipulation/Clip," and any reads without a clipped adaptor or that were <25nt in length after clipping were discarded. The clipped read was trimmed to nucleotides 2-50 with "NGS:QC and manipulation/Trim sequences." Reads were mapped to the sacCer2 yeast genome on Galaxy with "NGS: Mapping/Map with Bowtie for Illumina" (Langmead et al., 2009) using a SOAP-like alignment policy allowing 1 mismatch (-v 1), reporting 1 alignment per read (-k 1), and discarding any reads aligning to more than 16 locations in the genome (-m 16). rRNA reads (any reads mapping to chrXII: 451,000-468,999) were identified and removed using the "Filter and Sort/Select" tool.

Modifying uRNA coordinates: The 5' and 3' termini of all uRNA detectable by total RNA fragmentation were manually demarcated. The most inclusive 5' and 3' terminus among the uRNA boundaries annotated by Cufflinks and the manual annotation of the total fragmented RNA was identified. These updated uRNA transcript boundaries were converted into GTF format, and combined with the sacCer2 Ensembl Genes annotation for use as the reference annotation for quantification of ribosome profiling sequencing data (see below). This adjustment ensured that quantification of ribosome footprinting and total fragmented RNA sequencing was inclusive of the largest isoform of each transcript identified between this sequencing dataset and the RNA-seq sequencing dataset which initially defined uRNA coordinates.

Quantification of ribosome footprint coverage: FPKMs were obtained using Cuffdiff (Cufflinks v2.1.1; Trapnell et al., 2013) with bias correction and multi-read correction, providing biological replicates for analysis where possible, using the reference GTF file described above in "Modifying uRNA coordinates" (-b -u --library-type ff-firststrand). RNAs with poor coverage in the total fragmented RNA datasets (FPKM = 0 in WT or FPKM <10 in *upf1Δ*) were excluded from footprinting score analysis. 331 uRNAs met this filtering cutoff.

Calculation of footprinting score: To calculate the footprinting score, for each RNA we determined the ratio of ribosome footprinting reads relative to reads from total fragmented RNA ($FPKM_{footprints}/FPKM_{fragments}$). Sequencing datasets were normalized for *PGK1* and *RPL41A* mRNAs to have a footprinting score of 1. To compare the translation of RNAs as measured by both the translatability score and footprinting score, for all RNAs with a score >0, a Spearman rank correlation coefficient was calculated.

Because the absence of ribosome footprints could be either due to a true failure to associate with the translation machinery, or insufficient depth of our ribosome profiling, we classify uRNAs showing sufficient evidence of ribosome association (footprinting score in WT > 0 and footprinting score in *upf1Δ* >0.1) as showing evidence of being ribosome-bound in this assay, and make no conclusions about the absence of ribosome footprinting data. 185 uRNAs (of 331 analyzed) demonstrated ribosome association by these cutoffs.

Demarcation of ribosome-free regions: For uRNAs, the region covered by ribosome footprints was manually demarcated based on visualization of ribosome footprinting sequencing reads in the IGV genome browser (www.broadinstitute.org/igv; Robinson et al., 2011). Footprint occupancy regions were annotated to be representative of the ribosome footprint profile and include >75% of ribosome footprint sequencing reads. In cases where ribosome footprints fell marginally outside the uRNA boundaries, the 5' or 3' ribosome-free size was set as "0". Only those uRNAs meeting the expression cutoffs defined above in "Quantification of ribosome footprint coverage" and with sufficient evidence of ribosome association as described in "Calculation of footprinting score" were included in this analysis (n=185).

Assignment of phasing frames for mRNAs: To establish phasing of ribosome footprints along annotated mRNAs, individual sequencing datasets (ribosome profiling or total fragmented RNA) were filtered to include only reads of 27 nucleotides; these reads represent reads that were 28 nucleotides prior to trimming 1 nucleotide from the 5' end during mapping, and represent reads which predictably demonstrate ribosome occupancy (Ingolia et al., 2009). Using a custom script, each nucleotide position within all annotated CDS (based on the Ensembl sacCer2 Genes annotation downloaded from the UCSC genome table browser; genome.ucsc.edu/ cgi-bin/hgTables?command=start) was assigned a frame as follows: a position -11 of the first nucleotide of the AUG start codon was assigned an in-frame "+1", as well every third nucleotide thereafter through -16 of the last position of the CDS; a position -10 of the first nucleotide of the AUG start codon was assigned "+2", as well as every third nucleotide thereafter through -15 of the last position of the CDS; a position -9 of the first nucleotide of the AUG start codon was assigned "+3", as well as every third nucleotide thereafter through -14 of the last position of the CDS. The sequencing datasets were cross-referenced to this nucleotide frame definition such that the frame number corresponding to the start position of the sequencing read indicated the frame to which the read aligned; reads not aligning to a CDS were assigned a frame of "0" and not further analyzed. For each dataset, the percentage of reads assigned to each frame was calculated. Graphed data represent the average percentage of reads aligned to each frame for 4 replicates of ribosome footprinting data, and 2 replicates of fragmented RNA data, +/- SEM. Scripts were written using Python.

Assignment of phasing frames for uRNAs: Using a custom script, each nucleotide position within all uRNAs defined in this study was assigned a frame as described above, with the exception that reference points were the transcript start and end position, rather than a CDS start and stop position. Sequencing datasets containing only 27-nucleotide reads (described above) were compared to the uRNA nucleotide frame definition as above, to assign each read to a corresponding frame. The total number of 27-mer reads aligning to each uRNA was determined, combining all 4 ribosome footprinting datasets (two WT biological replicates and two *upf1Δ* biological replicates) or both fragmented RNA datasets (one WT biological replicate and one *upf1Δ* biological replicate); any uRNA with less than 10 combined 27-mer ribosome footprinting reads was discarded from this analysis. For all uRNAs with at least 10 combined ribosomal footprinting reads (n=80), the percentage of reads aligning to each frame was determined. Any uRNAs demonstrating at least 50% of reads aligning to a single frame was considered to show evidence of translation-dependent phasing, and this frame was arbitrarily set to frame +1. Graphed data represent the average percentage of reads aligning to each frame for an individual phased uRNA +/- SEM, and include a total of 61 uRNAs that demonstrated phasing. Scripts were written using Python.

Identification of sORFs: uRNAs demonstrating phasing of ribosome footprints were individually examined to determine if a putative translated ORF could be identified based on the frame to which the ribosome footprinting sequencing reads aligned. This identification required an in-frame canonical AUG start codon near the 5' end of ribosome footprints (often centered within the P site of the most 5' footprinting read), and the putative ORF was extended through the first in-frame stop codon following this AUG. All such putative ORFs encoding peptides of at least 10 residues constitute our class of sORFs. In some cases more than one utilized sORF was identified per uRNA. In one case (sORF-28), no canonical start codon could be identified despite strong evidence for phased ribosome footprints throughout the region; in this case, the codon within the P site of the most 5' ribosome footprint was considered the first codon for this sORF.

Data visualization: Snapshots of ribosome profiling read coverage were obtained using the IGV genome browser (www.broadinstitute.org/igv; Robinson et al., 2011). Metagene plots of ribosome footprint coverage were generated using ngsplot (https://code.google.com/p/ngsplot/), providing 6-column BED files of sORF or mRNA CDS regions (default parameters and -R bed -FL 30 -SE 0 -L 100).

Northern Analysis of Steady-State RNA

Whole-cell RNA was isolated using glass-bead lysis followed by a phenol/chloroform extraction (Geisler et al., 2012). 40 μg of whole-cell RNA was separated by agarose gel electrophoresis on a 1.4% agarose gel with 5.92% formaldehyde. RNA was transferred to a Hybond-N nylon membrane (GE Healthcare RPN303N) and immobilized with UV crosslinking. Membranes were washed in 0.1X SSC/0.1% SDS for 1 hour at 65 °C, incubated for 1 hour in hybridization buffer (10X Denhardt's solution, 6X SSC, 0.1% SDS), and probed overnight in hybridization buffer with either 5' ^{32}P end-labelled DNA oligonucleotides or α-^{32}P CTP probes generated by asymmetric PCR (Rio et al., 2011; see Table S5) at individually optimized temperatures, to detect the RNA of interest. Excess probe was washed from membrane three times for 15 minutes with 6X SSC/ 0.1% SDS at individually optimized temperatures. Membrane was exposed to a storage phosphor screen (Molecular Dynamics), and developed using a GE Typhoon 9400 Variable Mode Imager (Amersham Biosciences).

Generation of HA-tagged sORF Strains

Chromosomal tagging of sORFs at their endogenous loci was performed using standard homologous recombination methods (Longtine et al., 1998). This approach results in incorporation of the 3xHA tag at the C-terminus of the sORF immediately followed by *ADH1* terminator sequences, and incorporation of a downstream selectable marker to facilitate screening of clones. sORFs were selected based on high expression of the uRNA, strong evidence of ribosome footprint phasing, and intergenic genomic location. Yeast genome sequences retrieved from the Saccharomyces Genome Database (www.yeastgenome.org) were used to determine gene-specific sequences to target knock-in of the 3xHA tag and selectable marker to the correct locus. These sequences were designed to insert the 3xHA tag immediately upstream of the predicted stop codon. The 3xHA tag was inserted either in-frame with the putative sORF, or out-of-frame as a control to demonstrate frame-dependent expression of the 3xHA tag. Incorporation of the 3xHA tag was confirmed by Sanger sequencing for each locus. See Table S5 for primers and plasmids used to generate strains.

Generation of FLAG-tagged sORF Plasmids

Based on the yeast genome sequences retrieved from the Saccharomyces Genome Database (www.yeastgenome.org), the genomic region encompassing several uRNAs containing putative sORFs (sORFs were selected based on high expression of the uRNA, strong evidence of ribosome footprint phasing, and intergenic genomic location) plus and minus ~500 bp was amplified by PCR with Phusion High-Fidelity DNA Polymerase (NEB M0530S), which produces a blunt-end PCR product. PCR products were ligated at 16 °C overnight into yEpLac181 previously digested with *Sma*I blunt-end restriction enzyme (NEB R0141S) using T4 DNA Ligase (Roche 10 481 220 001). Ligated plasmids were transformed into calcium chloride competent XL1-Blue *Escherichia coli*, plated on 2% agar Luria broth plates plus 100 µg/mL ampicillin, and individual clones screened by restriction digest and sequencing to confirm ligation of the appropriate insert. 1X FLAG tag (DYKDDDDK) was added in-frame to the C-terminus of each putative sORF, immediately upstream of the putative stop codon, using a single round of site-directed mutagenesis PCR with Phusion High-Fidelity DNA Polymerase. PCR product was treated with *Dpn*I restriction enzyme (NEB R0176S) to digest any methylated template. Plasmid was transformed into XL1-Blue *E. coli* as above, and clones screened by sequencing to confirm the in-frame insertion of the FLAG sequence. All plasmids were transformed into WT yeast using a standard lithium acetate transformation protocol. Transformed strains were subsequently maintained and grown in selective media lacking leucine.

Protein Isolation and Western Blot Analysis

WT yeast cultures containing either 1) chromosomal 3xHA-tagged sORFs, or 2) plasmids containing uRNAs encoding a putative sORF with or without a C-terminal FLAG-tag, were grown and treated with proteasome inhibitor MG-132 as described (Liu et al., 2007), and flash frozen on dry ice. Cell pellets were heated in 5*M* urea at 95 °C for 2 minutes, then lysed by mechanical disruption with glass beads by vortexing for 5 minutes. Solution A was added to lysates (125 mM Tris-HCl, pH 6.8, 2% SDS), followed by vortexing for 1 minute and heating to 95 °C for 2 minutes. Glass beads and cellular debris were cleared from lysates by centrifugation at 13,200 RPM for 4 minutes. Equivalent OD units (A_{260}) of lysate in 1X SDS sample buffer (125 mM Tris-HCl, pH 6.8, 2% SDS, 100 mM DTT, 10% glycerol, 0.05% bromphenol blue) were separated on NuPAGE Novex 4-12% Bis-Tris gels (Life Technologies NP0321BOX) by electrophoresis in 1X MOPS SDS running buffer (50 mM MOPS, 50 mM Tris base, 0.1% SDS, 1 mM EDTA, pH 7.7). Proteins were transferred to an Immobilon-P PVDF transfer membrane (Millipore IPVH15150) in 1X western transfer buffer (25 mM Tris base, 192 mM glycine, 20% methanol) at 4 °C by electroblotting at 250 mA for 2 hours. Membrane was blocked in blocking buffer (5% milk powder in 1X TBS/0.1% Tween-20) overnight at 4 °C. Membrane was incubated with primary antibodies (rabbit polyclonal α-FLAG 1:10,000 [Sigma F7425], mouse monoclonal α-HA 1:5,000 [Covance MMS-101P], or mouse monoclonal α-PGK1 1:10,000 [Invitrogen 459250]) and secondary antibodies (goat α-rabbit IgG HRP 1:5000 [Pierce 31460] or goat α-mouse IgG HRP 1:5000 [Santa Cruz sc-2005]) in blocking buffer for 1 hour. Between each incubation, membrane was washed with 1X TBS/0.1% Tween-20 3 times for 15 minutes. Signal was detected by chemiluminescence using Blue Ultra Autorad film (GeneMate F-2029).

Conservation of sORF Peptides in Other Fungi

BLAST analysis: A custom database to be used for BLAST search was generated with NCBI BLAST tool formatdb V2.2.29+, with the following genomes: from the *Saccharomyces* Genome Database (yeastgenome.org): *Saccharomyces bayanus* strain S23-6C, *Saccharomyces kudravzevii* strain IFO1802, *Saccharomyces mikatae* strain IFO1815, *Saccharomyces paradoxus* strain NRRL Y-17217, *Saccharomyces pastorianus* strain Weihenstephan 34/70, and 33 *Saccharomyces cerevisiae* strains (standard laboratory strain S228C, AWRI1631, AWRI796, BY4742, BY4741, CBS7960, CEN.PK, CLIB215, CLIB324, CLIB382, EC1118, EC9-8, FL100, FostersB, FostersO, JAY291, Kyokai7, LalvinQA23, M22, PW5, RM11-1a, Sigma1278b, T7, T73, UC5, VIN13, VL3, W303, Y10, YJM269, YJM789, YPS163, ZTW1); from the NCBI Genome database (www.ncbi.nlm.nih.gov/genome): *Naumovozyma castellii* strain CBS 4309 (assembly ASM23723v1), *Candida glabrata* strain CBS138 (assembly ASM253v2), *Kluyveromyces lactis* strain NRRL Y-1140 (assembly ASM251v1), and *Ashbya gossypii* strain ATCC 10895 (assembly ASM9102v4).

Putative sORF peptides were provided as query for TBLASTN against our custom curated database. TBLASTN was run using BLAST v2.2.29+, with the E-value threshold set to 10 and all other parameters default. Results in which the subject sequence contained a termination codon that interrupted the peptide were filtered from the dataset. The number of identical residues relative to the length of the query was used to calculate percent identity. In many cases, local regions of high-identity alignment were reported that did not extend across the entire query length; for these the number of identical residues relative to the full length of the query was used to calculate percent identity. Only the hit with the highest percentage of identical residues relative to the full length of the query for each species is reported. Data is only reported for non-*S. cerevisiae* alignments. See Table S4.

PhastCons conserved elements: Conserved elements across 7 yeast species (*Saccharomyces. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. castelli*, and *S. kluyveri*) have previously been identified using phastCons and reported (Siepel et al., 2005), and are accessible in the UCSC genome browser (http://genome.ucsc.edu/) using the "Most Conserved" track. Using the sacCer2 *S. cerevisia*e genome assembly, we report the log-odds score of conserved elements that partially or completely overlap each putative sORF. If more than one conserved element overlapped an sORF, the log-odds score for the element displaying the highest degree of overlap is reported. See Table S4.

Calculation of the Ka/Ks ratio: The Ka/Ks ratio (ω; the relative rate of nonsynonymous to synonymous mutations along a conserved sequence), was calculated for putative sORFs using the Ka/Ks_Calculator (Zhang et al., 2006), with the method of model averaging. For each ratio, the sacCer2 reference genome sequence was compared to the nucleotide sequence corresponding to the highest-identity TBLASTN result for each species, as reported in Table S4. Ka/Ks ratios were only calculated if 1) a TBLASTN alignment was reported, and 2) the aligned nucleotide sequence corresponding to the TBLASTN peptide hit did not align 100% with the reference genome nucleotide sequence (marked as "N.D."). Only Ka/Ks ratios with a Fisher's p-value <0.05 are reported.

Analysis of Mammalian Data

Data sources: Ensembl transcript structures and annotations for the mouse July 2007 (NCBI37/mm9) genome assembly were obtained from Ensembl version 67 (http://useast.ensembl.org/info/data/ftp/). Transcript models assembled from poly(A)-selected RNA of mESCs (Guttman et al., 2010) were collected from the Scripture portal (http://www.broadinstitute.org/software/scripture/). RNA-Seq and Ribo-Seq data for shRNA-, cycloheximide- or control-treated mESCs, as well as CLIP-Seq data for UPF1 binding were downloaded from the Gene Expression Omnibus (GSE41785; Hurt et al., 2013).

RNA-Seq analysis: Paired-end directional reads were quality-checked with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and mapped to mm9 using TopHat v2.0.8 (Trapnell et al., 2009; default parameters and --solexa1.3-quals

--library-type fr-firststrand --min-anchor 5 -r 170). Gene-level expression was estimated as fragments per kilobase of exon model per million mapped fragments (FPKM) using Cufflinks v2.1.1 (Trapnell et al., 2010; default parameters and --min-frags-per-transfrag 0 --compatible-hits-norm --min-isoform-fraction 0.0) considering gene annotations from Ensembl v67 and lincRNA annotations from mESCs (Guttman et al., 2010).

Ribo-Seq analysis: Non-directional reads were quality-checked with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and trimmed from the 3' end using fastx_clipper (http://hannonlab.cshl.edu/fastx_toolkit/index.html) to generate 30 nt fragments. Trimmed fragments were then aligned to rRNA annotations from Ensembl v67 using Bowtie v2.1.0 (Langmead et al., 2009; default parameters and --seedlen=23), and non-rRNA reads were then mapped to mm9 with TopHat (default parameters and --solexa1.3-quals --min-anchor 5 --no-novel-juncs) considering gene annotations from Ensembl v67 and lincRNA annotations from mESCs (Guttman et al., 2010). The fold-change in the normalized density of Ribo-Seq reads mapping along the transcript body of intergenic lincRNAs in shUPF1 v. shGFP libraries was calculated and visualized using ngsplot (https://code.google.com/p/ngsplot/; default parameters and -R genebody -F rnaseq,lincRNA -FL 30).

CLIP-Seq analysis: UPF1 CLIP-Seq reads that were previously processed (trimmed and subtracted from overlapping amplified IgG CLIP-Seq reads) and mapped uniquely to mm9 or to a splice junction database allowing 2-nt mismatches (Hurt et al., 2013) were directly analyzed for their overlap with gene bodies. Data from 3 replicate libraries (two RNAse A- and one RNAse I-treated libraries) were combined prior to analysis. Coverage along gene transcripts was computed using BEDTools (Quinlan and Hall, 2010).

Defining a set of lncRNAs: Starting with Ensembl v67 annotations, we considered only genes annotated as “lincRNA”, “non-coding” or “antisense” that had no transcript annotated as “ambiguous_orf”. We then incorporated putative lincRNA transcript models (Guttman et al., 2010) from loci that are reliably active in mESCs (Guttman et al., 2011) provided that they were not annotated in Ensembl v67 already. Finally, we computed the ribosome release score (Guttman et al., 2013) with RRS (http://guttmanlab.caltech.edu/software/RRS.jar; default parameters) based on the shGFP RNA-Seq and Ribo-Seq libraries for both putative lncRNAs curated from Ensembl v67 and from the mESC lincRNA collection. Any gene with a transcript having an RRS score >10 was excluded from further analysis.

Gene expression analysis: To identify mRNAs and lncRNAs reliably expressed in mESCs, we considered only genes that are expressed at FPKM>0.1 in each shRNA-, cycloheximide- or control-treated RNA-Seq library and are expressed at FPKM>1 in at least one of these libraries. This strategy yielded 13043 and 265 mESC-expressed mRNAs and lncRNAs, respectively.

Defining a set of NMD targets: To reliably identify NMD-targeted genes, we assessed the consistency in the response across the three NMD inhibitory treatments (shUPF1-1, shUPF1-2 and CHX). Consistency was determined by taking the geometric mean of the fold-change in FPKM in treated samples versus controls (shUPF1-1 v. shGFP, shUPF1-2 v. shGFP, and CHX v. WT, each averaged over 2 replicates). Genes with a geometric mean >1.5 were designated as consistent NMD targets, based on benchmarking against a set of known mRNA isoforms targeted by NMD.

Additional bioinformatics analyses: Computational analyses were conducted using custom scripts in Python, Perl and R. Statistical tests and plots were implemented in R, and heatmaps were produced using the *gplots* R package (http://CRAN.Rproject.org/package=gplots).

SUPPLEMENTAL REFERENCES

Ashurst, J.L., Chen, C.K., Gilbert, J.G., Jekosch, K., Keenan, S., Meidl, P., Searle, S.M., Stalker, J., Storey, R., Trevanion, S., et al. (2005). The Vertebrate Genome Annotation (Vega) database. Nucleic Acids Res *33*, D459-465.

Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a web-based genome analysis tool for experimentalists. Curr. Protoc. Mol. Biol. *89*, 19.10.1-19.10.21.

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W.J., and Nekrutenko, A. (2005). Galaxy: a platform for interactive large-scale genome analysis. Genome Res. *15*, 1451-1455.

Gietz, R.D., and Sugino, A. (1988). New yeast-Escherichia coli shuttle vectors constructed with in vitro mutagenized yeast genes lacking six-base pair restriction sites. Gene *74*, 527-534.

Goecks, J., Nekrutenko, A., Taylor, J., and The Galaxy Team. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computation research in the life sciences. Genome Biol. *11*, R86.

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., et al. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature *477*, 295-300.

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

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

Liu, C., Apodaca, J., Davis, L.E., and Rao, H. (2007). Proteasome inhibition in wild-type yeast *Saccharomyces cerevisiae* cells. Biotechniques, *42*, 158-162.

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

Rio, D.C., Ares, M. Jr., Hannon, G.J., and Nilsen, T.W. (2010). RNA: A Laboratory Manual (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press).

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24-26.

Sweet, T., Kovalak, C., and Coller, J. (2012). The DEAD-box protein Dhh1 promotes decapping by slowing ribosome movement. PLoS Biol. *10*, e1001342.

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

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

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

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

Wang, L., Wang, S., and Li, W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics *28*, 2184-2185.

Appendix E: CASC15 is a tumor suppressor lncRNA at the 6p22 neuroblastoma susceptibility locus

This work represents a manuscript in preparation by the following authors:

Mike R. Russell, Annalise Penikis, Derek Oldridge, **Juan R. Alvarez-Dominguez**, Lee McDaniel, Maura Diamond, Olivia Padovan, Pichai Raman, Yimei Lee, Jun Wei, Shile Zhang, Janahan Gnanchandran, Robert Seeger, Shahab Asgharzadeh, Javed Khan, Sharon Diskin, John M. Maris and Kristina A. Cole.

Author contributions: J.R.A.-D. designed, performed, and presented the analysis of sequencing reads from RNA-seq, ChIP-seq, CAGE-seq, poly(A)-seq, and DNase-seq experiments (Figures 2, 3, and Figure S2); contributed design of and presentation of bioinformatics analyses (Figure S2), and contributed to study design and manuscript revisions.

Acknowledgments: The authors would like to acknowledge the Children's Oncology Group (U10---CA98543) for providing blood and tumor specimens from neuroblastoma patients..

THE LONG NON-CODING RNA CASC15 IS A NEUROBLASTOMA SUPPRESSOR GENE AT THE 6P22 GENOME-WIDE ASSOCIATION STUDY-DEFINED SUSCEPTIBILITY LOCUS

Mike R. Russell[1], Annalise Penikis[1], Derek Oldridge[1], Juan R. Alvarez-Dominguez[2,3], Lee McDaniel[1], Maura Diamond[1], Olivia Padovan[4], Pichai Raman[1,5], Yimei Lee[1], Jun Wei[6], Shile Zhang[6], Janahan Gnanchandran[7], Robert Seeger[7], Shahab Asgharzadeh[7], Javed Khan[6], Sharon Diskin[1,8], John M. Maris[1,8,9] and Kristina A. Cole[1,8,9,a]

[1]Division of Oncology, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104; [2]Whitehead Institute for Biomedical Research, [3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, 02142, USA, [4]Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104; [5]Center for Biomedical Informatics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104; [6]Oncogenomics Section, Pediatric Oncology Branch, Center for Cancer Research, National Cancer Institute, 37 Covent Drive, Bethesda, Maryland; [7]Department of Pediatrics, Division of Hematology-Oncology, Children's Hospital Los Angeles and Saban Research Institute, University of Southern California, Los Angeles, CA 90027; [8]Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104-4318; and [9]The Abramson Family Cancer Research Institute, University of Pennsylvania School of Medicine, Philadelphia, PA 19104.

Running title: CASC15 is a neuroblastoma suppressor gene

Keywords: Neuroblastoma, CASC15, LINC00340, FLJ22536

[a] Corresponding author:
Dr. Kristina A. Cole
Children's Hospital of Philadelphia
3501 Civic Center Blvd.
Philadelphia, PA 19104, USA.
colek@email.chop.edu

ABSTRACT

We previously identified 6p22 as a neuroblastoma susceptibility locus using a genome-wide association study (GWAS) design, but the mechanisms underlying tumorigenesis at this locus remained elusive. Fine mapping, demonstrates that these highly significant single nucleotide polymorphisms (SNPs) reside within a long intergenic non-coding RNA (*LINC00340*) termed *cancer-associated susceptibility candidate 15* (*CASC15*), and include at least one risk allele able to impact enhancer function. A distinct short isoform identified by RNA sequencing (*CASC15-S*), was highly associated with advanced disease and patient survival probability (p= 3.2x10^{-6}). *CASC15-S* depletion in human neuroblastoma–derived cell lines increased cellular growth (p<0.001) and migratory capacity (p=0.006), suggestive of a tumor suppressive function. Gene expression analysis following *CASC15-S* ablation demonstrated significant downregulation of neuroblastoma-specific markers with concomitant increases in cell adhesion and extracellular matrix transcripts. These data suggest that *CASC15-S* regulates neural growth and differentiation pathways, and that dysregulation of *CASC15-S* contributes to the initiation and progression of neuroblastoma.

A cancer of the developing autonomic nervous system, neuroblastoma is the most common malignancy diagnosed in the first year of life and often lethal in older children; accounting for approximately 10% of all pediatric cancer mortality[1-3]. While the majority of low-risk neuroblastoma patients are cured with surgery alone, high-risk patients frequently die from disease progression despite highly intensive chemoradiotherapy. Neuroblastomas are thought to develop from cells of the peripheral nervous system committed to the sympathicoadrenal lineage (autonomic nervous system) [1-4]. Because malignant transformation can occur at any point during sympathetic development, tumors may arise from various stages of neural crest lineage (most commonly the adrenal gland), contributing to the hallmark heterogeneity observed in this disease[4]. To address the etiology of sporadic neuroblastoma, we conducted the first pediatric cancer genome-wide association study (GWAS), leading to the identification of numerous validated susceptibility loci in several populations[5-12]. Moreover, we showed that many of these susceptibility alleles are specifically associated with either high-risk or low-risk disease features as well as patient outcomes. The majority of these SNPs act in *cis* to influence expression of protein coding genes at these loci, and several of these transcripts, such as *LMO1*, *BARD1* and *LIN28B,* appear to play an oncogenic role in established tumors [5-12].

The first identified neuroblastoma GWAS signal, and the one that remains most significant, is contained within a 94.2kb linkage disequilibrium (LD) block on chromosome 6p22.3. Like other subsequently identified loci, we observed a highly significant association with neuroblastoma susceptibility and clinically aggressive presentation. Children homozygous for risk alleles at 6p22.3 are predisposed to neuroblastoma development (odds ratio: 1.97, 95% confidence interval: 1.58-2.45), more likely to have metastatic stage 4 disease (p=0.02), *MYCN* amplification (p=0.006) and suffer disease relapse (p=0.01)[5]. At the time of this finding, however, this locus was devoid of any annotated genes with protein coding potential, impeding further characterization of this region in neuroblastoma initiation.

Recent data obtained from whole genome sequencing has illustrated far fewer protein-coding genes than previously predicted; however it is now clear that as much as 70% of the genome is transcribed into products other than traditional protein-coding mRNAs[13,14]. Although many of these transcriptionally active loci produce RNA species involved in translation (i.e. ribosomal and transfer RNAs), several other RNA classes have been functionally validated as bona-fide regulatory molecules. The recently identified long non-coding RNAs (lncRNAs), defined as RNA species >200nt in length and lacking an open reading frame, have been increasingly implicated in a wide variety of cellular functions[15]. LncRNAs share several transcriptional features in common with mRNAs - they are often spliced, demonstrate RNA polymerase II occupancy, contain a 5' methylguanosine cap, and are commonly (though not always) polyadenylated[16,17]. Although lncRNA function is highly context dependent, they commonly play a prominent role in the spatiotemporal regulation of gene expression during developmental processes[18-20], and therefore exhibit a tendency to be located proximal to developmentally critical protein-coding genes[21]. Indeed, several lncRNAs reside near protein-coding genes known to regulate lineage commitment in neural crest cells[22], serving as an attractive hypothesis to explain the etiology of embryonal cancers such as neuroblastoma.

As might be expected, lncRNAs have been increasingly implicated in a variety of oncogenic processes through association with epigenetic complexes and modification of chromatin accessibility - ultimately influencing gene expression[23-26]. To date, there are no reports concerning the role of lncRNAs in the initiation and progression of solid pediatric neoplasms, despite the fact that many childhood cancers are fundamentally defects of normal human development[27]. Here we describe a novel, uncharacterized lncRNA, *CASC15* (previously known as *LINC00340*) identified through a robust neuroblastoma GWAS signal. This lncRNA specifically localizes to several brain regions, is differentially expressed between low- and high-risk disease, and correlates with poor overall survival in neuroblastoma patients. Initial characterization of this transcript indicates that it functions as tumor suppressor in neuroblastoma, with depletion resulting in increased neuroblastoma cell growth and migratory capacity. Moreover, this lncRNA regulates gene pathways involved in cellular adhesion, migration, as well as modulating neural-specific differentiation genes.

RESULTS

Fine mapping and identification of CASC15 isoform expression

The initial discovery-phase neuroblastoma GWAS, consisting of 1,032 neuroblastoma cases and 2,043 healthy aged-matched controls, identified three common polymorphisms clustered on chromosome 6p22.3 that are highly associated with aggressive disease and a significantly increased risk of neuroblastoma development [4]. This region is contained in a linkage dissociation (LD) block containing the *LINC00340* and *LOC729177* genes and flanked upstream by *SOX4*. To map this region with finer detail, we extended the number of subjects to 2,101 cases and 4,202 genetically matched controls, and performed regional imputation [10, 28]. This refined our 94.2kb LD block to a 30kb region, identifying 32 SNPs in high linkage disequilibrium ($r^2 >$ 0.8, p = 8.26x10^{-10} – 1.88x10^{-15}) localized to an intronic region of an annotated gene locus formerly titled *FLJ22536 / LINC00340* and more recently renamed *cancer associated susceptibility candidate 15* (*CASC15*; **Fig. 1a, Supplementary Table 1**). The linkage disequilibrium (LD) structure for the Northern European (CEU) population supports this 30-kb region of interest (**Fig. 1b**); however, as has been shown for other neuroblastoma loci, refinement of this locus using an African American cohort was not possible due to the substantially different LD structure observed in this population [11]. To further demonstrate the importance of this region in neuroblastoma tumorigenesis, we performed differential analyses of noncoding RNA expression between 220 high-risk and 30 low-risk primary tumors for which we had microarray expression data (**Table 1**). Indeed, we found that *CASC15* exhibited a 4.4-fold lower level of expression in high-risk disease compared to low-risk (p = 1.0x10^{-17}) - the most significant finding from this experiment and in support of the signal obtained in our neuroblastoma GWAS.

However, several computationally predicted lncRNAs map to this locus, including multiple *CASC15* isoforms and an overlapping lncRNA, *CASC14* (formerly LOC729177) transcribed from the antisense strand (**Supplementary Fig. 1a**). To identify the products expressed from this locus in neuroblastoma, we first utilized RNA sequencing (RNASeq) and RNA-paired end tagged (RNA-PET) reads available from the ENCODE project (https://genome.ucsc.edu/ENCODE/) from SK-N-SH and SK-N-BE2 neuroblastoma cell lines. These data demonstrated the existence of two capped and polyadenylated nuclear

CASC15 transcripts, a long (hg19, chr6:21,666,675-22,194,616; Ensembl: CASC15-003) and a short isoform (hg19, chr6:22,146,883-22,194,616; Ensembl: CASC15-004) (**Fig. 2a, Supplementary Figs. 1b, 2a**). These noncoding transcripts are highly conserved in vertebrates, and readily detected in several brain regions and neuroblastoma cell lines (**Supplementary Fig. 2a-d)**, with putative promoter regions separated by 480kb suggestive of independent transcriptional regulation. We next examined RNA expression data from a panel of 16 primary human tissues as part of the Illumina Human Body Map project, where we found that the short *CASC15* isoform was expressed in abundance in the brain, but at modest to low levels in most other tissues (**Fig 2b**). We subsequently utilized RNA sequencing data from 108 primary neuroblastoma tumors generated as part of the NCI TARGET initiative, using non-overlapping, transcript-specific reads to investigate the expression of lncRNAs transcribed from this locus. We augmented these findings with our own strand-specific RNA sequencing data from three neuroblastoma cell lines and two primary tumors, confirming nearly complete alignment to the plus strand (avg. 92.6%, min. 86.9%, max 98.6%). Together, these RNA sequencing data identify the short *CASC15* isoform as the predominant transcript expressed from this locus in neuroblastoma, with expression values 20- to 40-fold higher than *CASC14* and full-length *CASC15,* respectively (**Fig. 2c**). Lastly, RNA sequencing data identified a third unspliced transcript, spanning part of exon 1 of the short *CASC15* isoform through a downstream noncoding element (*hs1335*) with a validated enhancer function in the developing murine neural tube, further supporting the role of this locus in neural development[29].

To experimentally validate our RNASeq data we performed 5' and 3' RACE for CASC15, subsequently cloning and sequencing these transcripts from two neuroblastoma cell lines and fetal brain tissue. We verified the sequence of both the 12-exon 1.9kb *CASC15* transcript (NR_015410.1, Ensembl: CASC15-003) and the 4-exon short (1.2kb) variant (Ensembl: CASC15-004), hereafter referred to as *CASC15-S*. *CASC15-S* resembles a known cDNA clone (GenBank: AK094718), containing a unique first exon, yet sharing its remaining sequence with the last three exons of *CASC15* (**Fig. 2a, Supplementary Fig. 3a**). Despite several attempts, we were unable to isolate a 5'-capped product for the intron-less transcript (Ensembl: CASC15-006) predicted to overlap exon 1 of CASC15-S and extend to the noncoding enhancer element (*hs1335*). Indeed, although the presence of this isoform is indicated on the SK-N-SH RNASeq track (**Fig. 2a**), it is absent in RNA-PET data (**Supplementary Fig. 1b),** supporting this result. Finally, RNA fluorescent *in situ* hybridization (RNA-FISH) experimentally confirmed the presence, relative abundance and cellular localization of these transcripts in both NB-69 and NGP neuroblastoma cells; using non-overlapping, strand-specific probes to label *CASC15*, *CASC15-S*, *CASC14*, *SOX4* and *hs1335*. These studies revealed exclusively nuclear *CASC15-S* and *hs1335* transcripts (**Fig. 2d**), predominantly cytoplasmic *SOX4* localization (**Supplementary Fig. 3b**) and virtually no *CASC14* or *CASC15* expression (**Supplementary Fig**. **3c,d**). Taken together, these experimental results confirm our predictive bioinformatic data identifying *CASC15-S* as a bona-fide lncRNA transcript in neuroblastoma.

To better understand how common genetic variation at this locus impacts neuroblastoma risk, we attempted to associate our previously published, highly significant polymorphism, rs6939340 ($p = 1.67 \times 10^{-14}$, odds ratio: 1.80, 95% confidence interval: 1.55 – 2.1) with expression of gene products at this locus. However, we were unable to demonstrate a significant correlation with risk genotype and transcript expression levels in either our 250 primary neuroblastomas (**Supplementary Fig. 4a**) or a representative panel of 20 neuroblastoma cell lines (**Supplementary Fig. 4b**). Furthermore, all three of our previously published polymorphisms lie in regions devoid of DNase hypersensitivity or other epigenetic marks indicative of transcriptional activity. In fact, eQTL analysis of all imputed genotypes with CASC15-S expression failed to reach significance after multiple comparison testing, leading us to postulate that polymorphisms contributing marginal effects may aggregately impact function at this locus. We therefore took advantage of our genome wide imputation data to investigate other highly significant polymorphisms lying within putative regulatory regions. As has been shown in other post-GWAS follow-up studies[39], we employed the following workflow (**Fig. 3a**) to narrow the field of potentially impactful polymorphisms: (1) we first chose polymorphisms with a GWAS *p*-value $< 1 \times 10^{-10}$, resulting in 32 candidates. (2) We further refined field by selecting only those SNPs within regions of DNaseI hypersensitivity (suggesting open chromatin) and H3K27 acetylation marks (indicative of enhancer activity) leaving us with four candidates: rs1543310, rs6905441, rs9295534 and rs9368402. (3) Lastly, we looked for SNPs with evolutionary conservation, resulting in a candidate polymorphism, rs9295534 ($p = 3.51 \times 10^{-12}$, odds ratio: 1.63, 95% confidence interval: 1.4 – 1.89) upstream of CASC15-S that localizes to an expanse of regulatory chromatin and dense transcription factor binding sites in several cell lines (**Fig. 3b, Supplementary Fig. 5)**. This region exhibits enhancer activity, evidenced by H3K27Ac CHIP-Seq data in several fetal tissues available from the NIH Roadmap Epigenomics Mapping Consortium (http://www.roadmapepigenomics.org/) (**Fig. 3c**). We next verified rs9295534 genotypes in Chp134 (homozygous non-risk) and Lan5 (homozygous risk) neuroblastoma cells by Sanger sequencing of a 1.5kb region encompassing this SNP, and subsequently cloned risk and non-risk fragments from these lines. To assess the impact of rs9295534 genotype on transcriptional activity, we inserted risk (A/A) and non-risk (T/T) fragments into luciferase reporter constructs upstream of a minimal promoter. Results from these experiments demonstrated significantly decreased transcriptional reporter activity following insertion of the risk genotype fragment (**Fig. 3d**), suggesting this region possesses an enhancer-like function that is disrupted following the inclusion of a neuroblastoma risk allele.

CASC15-S is differentially expressed in neuroblastoma and highly associated with disease outcome.

Because the rs9295534 homozygous risk genotype, by virtue of its linkage with rs6939340, is associated with high-risk neuroblastoma and poor survival [3], and also demonstrates decreased transcriptional activity in a minimal promoter assay, one would expect that patients with high risk/poor survival would have low CASC15S expression. To confirm this, we again turned to our primary neuroblastoma tumors with exon-based gene expression (n=250), where we observed significantly lower *CASC15-S* expression in high-risk/stage 4 patient tumors compared to low-risk stage 1 patient tumors (**Fig. 4a**). This finding was independent of *MYCN* mRNA expression, a known major oncogenic driver in neuroblastoma tumors (p = 0.29, **Supplementary Fig. 6a**). In addition, patients with tumors enriched in *CASC15-S* expression exhibited superior survival when compared to patients harboring neuroblastoma with low *CASC15-S* expression, both when low- and high-risk patients were included in the analysis (adj. p = 3.2x10^{-06}, **Fig. 3b)** but also when comparing expression from only high-risk patients (adj. p = 0.002, **Supplementary Fig. 6b**). Furthermore, this finding remained significant (p = 0.0084) after multivariate analysis adjusting for clinical factors such as age (as a continuous variable), MYCN amplification, and 1p/11q LOH status. Although expressed at much lower levels than *CASC15-S*, the long isoform of *CASC15* demonstrated a similar pattern in patient tumors (**Supplementary Fig. 6c-e**). Taken together, these data indicate that low *CASC15-S* expression correlates with a more aggressive phenotype in neuroblastoma, leading to poor overall survival.

In order to understand the contribution of *CASC15-S* in the initiation and progression of neuroblastoma, we next performed gene set enrichment analysis (GSEA) restricted to only high-risk neuroblastoma samples (n=220) where we utilized a 1.9-fold difference in median *CASC15-S* expression to define 146 low- and 74 high-expressing *CASC15-S* samples (**Fig. 4c**). The top differential gene expression profile that emerged from these analyses (Asgharzadeh neuroblastoma poor survival down, normalized enrichment score = -2.69, nominal *p*-value < 0.0001, FDR *q*-value < 0.0001) indicates that tumors with high *CASC15-S* expression are enriched in expression of genes known to be down-regulated in poor-outcome neuroblastoma, suggesting that *CASC15-S* exerts a protective effect and results in less aggressive disease, even within this subgroup of high-risk cases (**Fig. 4d**).

CASC15 depletion in neuroblastoma cell lines enhances proliferation and invasive capabilities

Having demonstrated a clinically relevant association with patient outcome for *CASC15-S* in neuroblastoma patients, we next sought functional validation for a role in tumorigenesis. We first assessed *CASC15-S* levels by quantitative RT-PCR across a well-characterized panel of neuroblastoma cell lines (n=21) where we observed differential expression similar to our primary tumor panel (**Fig. 5a**). Although the long isoform of *CASC15* was also detected in neuroblastoma cells, expression levels were again much lower and did not correlate with expression of *CASC15-S* (**Supplementary Fig. 6f**). To investigate the functional role of *CASC15-S* in neuroblastoma, we initially generated siRNA constructs targeting the 3' end of the gene, thereby simultaneously depleting both *CASC15* isoforms. Depletion of *CASC15/CASC15-S* resulted in a highly reproducible increase in neuroblastoma proliferation as evidenced by real-time cell growth and cell viability assays (**Fig. 5b**). We next recapitulated these results by selective depletion of only *CASC15*-S - targeting its unique first exon in several neuroblastoma lines (**Fig. 5c, Supplementary Fig. 7a,b**). As might be expected, depletion of full-length *CASC15* (targeting exon 6) or *CASC14* had no observable impact on cell growth or viability; validating our initial findings of *CASC15-S* as the functional isoform in neuroblastoma (**Supplementary Figs. 7c-d, 8a-d)**. We subsequently derived neuroblastoma cell lines stably depleted of *CASC15-S* (**Supplementary Fig. 8e**) and found that these cells also exhibited a substantial increase in cellular proliferation identical to what we observed in our transient siRNA-based *CASC15-S* depletion experiments (**Fig. 4d, Supplementary Figs. 7e-f, 9)**. Furthermore, rescue experiments, conducted by ectopically expressing *CASC15-S* in these cells, were able to revert the accelerated growth (**Fig. 5e**). Microscopic examination revealed overt morphological changes in shCASC15-S SK-N-BE2 cells; including striking changes in cell shape and size, with a resultant three-fold increase in cell area (659.5 ± 50.2μm^2 in control vs. 2024 ± 211.1 μm^2 in shCASC15 cells, $p < 0.0001$) (**Fig. 5f, g**). Furthermore, both SK-N-BE2 (**Fig. 6a, b**) and SK-N-SH neuroblastoma cells (not shown) stably depleted of *CASC15-S* exhibited an increased migratory capacity and invasiveness as evidenced by wound-healing assays ($p = 0.0006$, SK-N-SH and $p < 0.0001$, SK-N-BE2). Taken together, these data suggest that loss of *CASC15-S* promotes increased cellular growth and a more migratory phenotype in neuroblastoma.

CASC15-S regulates a subset of genes involved in neural crest development.

To better characterize the phenotypic changes we observed following *CASC15-S* depletion, we surveyed gene expression signatures of neuroblastoma cells depleted of *CASC15-S*. We transiently depleted *CASC15-S* and assessed gene expression changes at 48 hours in SK-N-SH cells using Affymetrix exon-level arrays. We observed substantial upregulation of several known cell adhesion genes, most notably entactin (*NID1*, $p = 3.7x10^{-4}$) and activated leukocyte cell adhesion molecule (*ALCAM*, $p = 1.05x10^{-7}$). Ingenuity pathway analyses (IPA) demonstrated highly significant upregulation of pathways involved in cell migration, proliferation and metastasis (**Fig. 6c**). We next analyzed SK-N-BE2 neuroblastoma cells stably silenced for *CASC15-S* and compared them to control vector transfected cells on Affymetrix Human Transcriptome 2.0 arrays, resulting in 362 differentially regulated genes. Analyses of these data demonstrated significant downregulation of several proneural gene family members involved in neurogenesis and differentiation, including neurogenic differentiation 1 (*NEUROD1,* p=$1.1x10^{-4}$), neural precursor cell-expressed, developmentally down-regulated gene 9 (*NEDD9,* p=$5.8x10^{-4}$) and neurogenin 2

(*NEUROG2, p*=5.3x10^{-4}). Taken together, these data suggest that loss of *CASC15-S* leads shifts the neuroblastoma gene expression away from a well-differentiated neural phenotype and promotes increased expression of cellular adhesion and migratory genes - a finding consistent with our phenotypic and morphological observations.

DISCUSSION

High-risk neuroblastoma remains a major challenge due to a relative paucity of somatic mutations hindering the development of targeted therapies[1]. The identification of mutations in *ALK* and *PHOX2B* has helped explain the origin of familial neuroblastoma; however, an understanding of the basis of sporadic disease has only recently begun to come into focus[34,35]. Here we identify and demonstrate the involvement of *CASC15-S* in the development and progression of neuroblastoma, illustrating the first long noncoding RNA identified in an embryonal cancer through a genome-wide association signal. Because this signal mapped to a genomically complex region, we used RNA-Seq to detect and quantify the relevant transcripts expressed at the 6p22.3 gene locus, revealing predominant expression of *CASC15-S* in neuroblastoma tumors. Despite this robust association, however, eQTL analyses correlating risk genotypes with transcript expression failed to reach statistical significance - a likely consequence of an underpowered patient data set and/or additional mechanisms capable of impacting expression (such as additional SNPs affecting expression, post-translational modification/degradation, etc.). Fine mapping of this region using genome-wide imputation identified a highly significant polymorphism (rs9295534) localizing to the closest upstream enhancer of *CASC15-S*; evidenced by open chromatin, H3K4me1 and H3K27Ac marks. Moreover, expression of the rs9295534 risk allele disrupts the enhancer function of this region in minimal promoter reporter assays. These findings suggest that, while we cannot definitively demonstrate a causative SNP based on traditional eQTL analyses, that genotype can indeed impact transcriptional ability at this locus.

A growing body of work supports a defined role of lncRNAs as spatiotemporal-specific regulators of gene expression critical for ensuring proper differentiation during development. The predominant expression of *CASC15-S* in brain (but not other tissues), the derivation of several cDNA clones from brain regions, and its proximity to a validated enhancer element (hs1335) strongly suggest that this lncRNA is uniquely involved in neural tube development. The role of lncRNA-mediated tumorigenesis in embryonal cancers provides a logical hypothesis to explain the etiology of neuroblastoma tumors, which are typically devoid of activating somatic mutations[1], and a preliminary understanding of how this transcript functions in neuroblastoma biology can be proposed from the functional and expression data we demonstrate here. *CASC15-S* expression in neuroblastoma tumors strongly correlates with disease stage and overall survival, and patients with lower *CASC15-S* expression have reduced expression of genes typically lost in poor outcome neuroblastoma. Conversely, patient tumors with high *CASC15-S* levels are enriched in expression of these genes, demonstrating a protective role for *CASC15-S*. Furthermore, ablation of *CASC15-S* in neuroblastoma cell lines increases cellular proliferation and morphological changes indicative of a less differentiated, more aggressive phenotype. Indeed, we found that *CASC15-S* depleted neuroblastoma cells possess increased migratory capacity, thereby complimenting gene expression analyses that demonstrate upregulation of pathways involved in cellular adhesion and migration, a concomitant decrease in apoptotic pathways and silencing of neural development and neuroblastoma-specific gene signatures. Taken together, these changes in cellular phenotype and gene signatures suggest that *CASC15-S* is responsible for maintaining a more differentiated and benign cell state, with *CASC15-S* loss leading to a poorly differentiated phenotype and expression of genes associated with transformed cells. On the other hand, depletion of the other lncRNAs overlapping this transcript, *CASC14* and *CASC15*, do not exhibit a similar phenotypic effect in these neuroblastoma cell lines and assays. Taken together, these data indicate that *CASC15-S* is a bona-fide functional lncRNA transcript in neuroblastoma.

In conclusion, this work characterizes the genomic region responsible for our highly significant GWAS signal on chromosome 6p22.3 and identifies a novel long noncoding RNA implicated in neuroblastoma initiation and progression. We propose the following working model to describe how this lncRNA influences neuroblastoma tumorigenesis. Our loss-of-function experiments indicate that *CASC15-S* is responsible for the regulation of gene families involved in neural development and differentiation, and that loss of these regulatory mechanisms leads to an increase in cellular proliferation and migration. Therefore, an inappropriate reduction of *CASC15-S* expression during neural crest development would have negative consequences for proper cellular lineage commitment and predispose these cells to undergo malignant transformation. Because expression of the rs9295534 homozygous risk allele results in attenuated transcriptional activity at this locus, *CASC15-S* transcript levels are subsequently reduced, as illustrated by a trend toward lower expression in homozygous risk patients. This *CASC15-S* deficiency consequently results in enhanced neuroblastoma growth and migratory capacity in these cells, manifesting as high-stage disease and a poor overall survival. Taken together, these data demonstrate a convincing role for this lncRNA in the etiology and tumorigenesis of high-risk neuroblastoma.

METHODS

Genome Wide Association Study (GWAS) and imputation. In an effort to refine the association signal and search for a causal variant at the 6p22 locus, we performed genotype imputation in a previously described European ancestry cohort of 2,101 neuroblastoma cases and 4,202 controls[10] using the 1000 Genomes Phase I Release 3 as a reference.

Genotyping: Members of the discovery cohort were genotyped using three different Illumina array technologies: Illumina Infinium II HumanHap550 version 1, Illumina Infinium HumanHap550 version 3, and Human Quad610 BeadChip. Optical density spectrophotometry and pico-green assays were used to assess genotype quality. 750 ng of blood DNA was taken from each subject and amplified 1000-1500 fold. Subsequently, DNA was fragmented to 300-600 base pairs, precipitated, suspended, and hybridized onto the appropriate Illumina BeadCHIP. A 50 bp probe sequence was designed to hybridize adjacent to a single nucleotide polymorphism (SNP) site via single base extension (SBE). After primer hybridization, a single hapten labeled dideoxynucleotide was used to extend the nucleic acids. Haptens were detected with a multilayer immunohistochemical sandwich assay followed by Illumina BeadArray Reader scanning at two wavelengths. Intensity values were calculated for each bead type, then loaded into Illumina GenomeStudio, normalized, and clustered. Samples with a genotype call rate below 95% were removed from the analysis.

Quality Control: The Human Quad610, HumanHap550 v1, and HumanHap550 v3 contained 620901, 555352, and 561466 markers. A subset of 518435 SNPs, common to all three assays, was extracted from each array to produce a uniform set of SNPs. Because strands are sometimes different between these arrays, when combining the arrays, an in-house software tool was used to flip SNPs to the plus strand. Consequently a single ped file containing 518435 plus strand SNPs across 6303 subjects was created. Because there were no A/T or C/G SNPs in the 518435 common SNPs, it was possible to determine strand unambiguously. Prior to imputation, the data were filtered to have a Hardy-Weinberg equilibrium p-value below 0.001, a genotype call rate above 99%, and a minor allele frequency (MAF) greater than 1%.

Imputation and Statistical Tests: SHAPEIT v2.r790 was used to phase the quality-controlled set of SNPs on chromosome six between 22000 kbp and 22210 kbp (Human genome build GRCh37/hg19) using recombination data from the 1000 genomes phase I December 9, 2013 release, downloaded from the IMPUTE2 website (https://mathgen.stats.ox.ac.uk/impute/data_download_1000G_phase1_integrated_SHAPEIT2_16-06-14.html). Phased genotype data was next fed into IMPUTE2 version 2.3.1, again using the 1000 genomes phase I December 9, 2013 release, to impute genotypes in the locus. Imputed SNPs were tested for significance using snptest 2.4.1. Prior to plotting, a filtered set of SNPs (info score > 0.8, MAF > 0.01) was prepared and fed into an in-house copy of LocusZoom 1.2 for visualization. The in-house copy of LocusZoom was customized to utilize the same 1000 genome phase I December 9, 2013 release as used in SHAPEIT and IMPUTE2 in order to maintain homogeneity of data. Finally, HaploView 4.2 was used to visualize the LD structure, which was then merged with the LocusZoom data using Inkscape 0.48.

Neuroblastoma Patient Datasets. The neuroblastoma RNAseq, SNP profiling and HuEx datasets are part of the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) initiative, supported by NCI Grant U10 CA98543. This specific grant is a collaboration of the Children's Oncology Group, Children's Hospital of Philadelphia, Children's Hospital of Los Angeles and the National Cancer Institute. The low-level sequence data have been deposited in the Sequence Read Archive (SRA) at the National Center for Biotechnology Information (NCBI), and are further accessible through the database of genotypes and phenotypes (dbGAP, http://www.ncbi.nlm.nih.gov/gap) under the accession number phs000218. The gene expression and copy number data, as well as clinical information on the NBL cases studied, available via the TARGET Data Matrix (http://target.nci.nih.gov/dataMatrix/TARGET_DataMatrix.html).

5'/3' Rapid Amplification of cDNA Ends (RACE). 5' and 3' RACE was performed via the First Choice RLM RACE kit (Ambion) using 10ug of RNA obtained from fetal brain or NB69 neuroblastoma cells following the manufacturers protocol. Specificity for *CASC15-S* was achieved by nested PCR using the following gene specific primer (GSP) pairs:

5' RACE: Outer GSP: 5'-CTAGCCCATCAGTTCCTTCG -3'
5' RACE: Inner GSP: 5'-TTCACCCTGTCCTCCAAGTC-3'
3' RACE: Outer GSP: 5'-TGGTTACCTGAGCTGCTCCT-3'
3' RACE: Inner GSP: 5'-CTCAGCCAGTGCAACACAAC-3'

Gene products were cloned into a pCR4-TOPO vector for sequencing.

RNA Sequencing (RNASeq). PolyA selected RNA libraries obtained from 108 high-risk neuroblastoma patients as part of the NCI TARGET project were prepared using TruSeq v3 (Illumina) for RNA sequencing on Illumina HiSeq2000 sequencers. The 101bp paired-end reads generated were aligned to the hg19 build of the human reference genome using TopHat v2.2.0. The HTSeq package (v0.6.1) was used to map aligned reads to the VISTA-annotated enhancer region, *hs1335*, as well as all transcripts annotated in Ref-Seq (v66) and/or the UCSC Genome Browser. Transcript expression values were normalized and quantified using the metric of reads per kilobase per million reads (RPKM). For isoform-specific quantitation of *CASC15* and *CASC15-S*, only exons and exon-exon junctions that were unique to each isoform were used in computing RPKM.

RNA Fluorescent *In Situ* Hybridization (RNA-FISH). Fluorescently labeled, non-overlapping oligonucleotide probes (20-mers) were designed to tile RNA transcripts, including *CASC14 (LOC729177)*, *SOX4, hs1335, CASC15* and *CASC15-S*. Probes were then divided into odd and even pools and hybridized to neuroblastoma cells. Images were obtained using a fluorescence microscope at 40x magnification, and colocalization of overlapping fluorescence signal from even and odd pools was used to confirm transcript-specific hybridization. All image processing was carried out using ImageJ.

Microarray Expression Data. Primary neuroblastoma tumor RNA (n=250) was isolated at diagnosis and hybridized to Affymetrix Human Exon 1.0ST arrays as part of the TARGET consortium. For gene knockdown experiments, SK-N-SH neuroblastoma cells were transfected in triplicate with 50nM siRNA as described below and subsequently hybridized to Affymetrix Human Exon 1.0ST arrays at the nucleic acids and protein core (NAPCORE) facility at the Children's Hospital of Philadelphia. SK-N-BE2 cells stably silenced for *CASC15* were hybridized to Affymetrix Human Transcriptome Array 2.0 arrays. For all experiments, data was processed using Robust Multi-array Average (RMA) background correction with quantile normalization and subjected to differential expression (> 2.0 fold change, ANOVA p < 0.05, false discovery rate < 0.05 as cutoff parameters) using Partek Genomics Suite v6.6. All other statistical tests were conducted using GraphPad Prism 6 software.

Gene Set Enrichment and Ingenuity Pathways Analyses. GSEA was conducted using the expression data from the 250 Affymetrix Human Exon 2.0ST arrays described above. *GSEA:* Gene set enrichment software (v2.1.0) was obtained from the Broad Institute (http://www.broadinstitute.org/gsea/index.jsp). CASC15-S high and low groups were compared across all datasets (c1-c6) using 1000 gene_set permutations with the default parameters (weighted enrichment, Signal2Noise ranking, etc.). *IPA:* Ingenuity pathway analysis (v21249400) was conducted using the full gene list of differential expression obtained from the Partek analysis (see Microarray expression data). For these analyses, a fold-change > 2.0-fold, p-value < 0.05 and a FDR < 0.05 were used as a cutoff.

Quantification of CASC15-S in neuroblastoma cells. Quantification of RNA transcripts was performed on a panel of neuroblastoma, medulloblastoma and adult cancer lines using Taqman PCR. For quantification of *CASC15*-S, we developed a custom primer/probeset spanning exons 1-2 and consisting of the following primers:

forward: GCTGTCGACGAAGGAACTGAT
reverse: GTCCAAGTCAAAAGTCTCATCCAAGA

Primer/probe sets for *CASC15 (LINC00340), CASC14 (LOC729177)* and *SOX4* are commercially available (Life Technologies). Quantification was normalized to the geometric mean of housekeeping genes *TBP, GUSB* and *HPRT-1*.

Cell growth and siRNA assays. Constructs targeting *GAPDH* and *PLK-1* (Thermo Scientific), as well as *CASC15* (n271797, n271792), *CASC14* (s59455) (Life Technologies) were transfected into neuroblastoma cells in triplicate, using 50nM of siRNA (Thermo Scientific) in 0.1-0.2% DharmaFECT 1 (Thermo Scientific). To specifically target the short *CASC15* isoform, we developed custom siRNA constructs targeting only first exon of *CASC15-S* and without overlap of *CASC14*.

CASC15-S #1: Sense – 5'-AGAGGAACUGUAAAUUGUAtt-3', Antisense – 5'-UACAAUUUACAGUUCCUCUat-3'

CASC15-S #2: Sense – 5'-GGAUAAACAAGGUAUCUGAtt-3', Antisense – 5'-UCAGAUACCUUGUUUAUCCat-3'

For CASC15-S addback experiments, plasmids containing the cDNA for either CASC15-S or GFP were transfected in triplicate in a 96-well plate using 250ng of DNA and 3ul of Lipofectamine 2000.

Cell growth assays were conducted using the xCELLigence real-time growth (RT-CES, ACEA Biosciences) and/or Cell-TiterGlo (Promega) assays according to manufacturer protocols.

shRNA, expression and lentiviral constructs. siRNA constructs (listed above) were used to created double-stranded shRNA constructs and were subsequently cloned into the pLenti-GFP DEST lentiviral vector (Addgene). 5x10^6 293T cells were transfected with 15ug of the pLenti transfer vector, 15ug of pLP1, 6ug of pLP2 and 3ug of pVSV-G vectors. Lentiviral particles were collected at 48 and 72 hours post transfection. Lentiviral transduction of neuroblastoma cells was carried out overnight at 37C using 3µg/ml of polybrene.

Wound Healing Assays. Scratch assays were carried out on Be2 and SK-N-SH cells stably depleted of *CASC15* plated at 85% confluence in 60mm dishes, and were scratched using a sterile 200µl pipette tip. Cells were photographed at regular intervals using a previously calibrated 5x light microscope (Nikon). Assessment of cell migration was carried out by measuring scratch closure as a percentage of initial scratch size in ImageJ, and was compared to control cells using a linear regression function in GraphPad Prism 6.

Cell Culture. Neuroblastoma cell lines were obtained from the neuroblastoma cell line bank maintained at the Children's Hospital of Philadelphia. Cell line identity is routinely confirmed via AmpFLSTR Identifiler (Applied Biosystems), last done in November 2013. Non-neuroblastoma cell lines were purchased from ATCC and all cell lines are routinely tested for mycoplasma. All cell lines are maintained in basal media (either RPMI1640 or DMEM) supplemented with 10% FBS and 1% gentamycin and cultured at 5% CO2.

Statistics. Where appropriate, group comparisons were determined with a two-sided t-test and Spearman correlation testing using Graphpad Prism.

For Kaplan Meier analysis: optimal cutoff was determined by employing a scanning approach to the Kaplan-Meier method by iteratively splitting the ordered genes expression values across samples into two groups and calculating the p-value by the Mantel-Haenszel log-rank test. The lowest p-value corresponds to the optimal breakpoint. A Benjamini-Hochberg correction was applied to reflect the presence of multiple hypotheses testing.

For multivariate analyses, a Cox Proportional Hazard model was used to evaluate the effect of each gene expression on overall survival, adjusting for clinical factors such as age (as a continuous variable), MYCN amplification, and 1p/11q LOH status.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the Children's Oncology Group (U10---CA98543) for providing blood and tumor specimens from neuroblastoma patients.

REFERENCES

1. Maris JM. Recent advances in neuroblastoma. (2010) *N. Engl. J. Med*. 362 (23): 2202–11. PMID: 20558371.
2. Brodeur GM. (2003) Neuroblastoma: biological insights into a clinical enigma. *Nat Rev Cancer* . 3(3):203–16. PMID: 12612655
3. Cheung N-K V, Dyer M. (2013) Neuroblastoma: developmental biology, cancer genomics and immunotherapy. *Nat Rev Cancer*. 13(6): 397–411. PMID: 23702928
4. Takahashi Y, Sipp D, Enomoto H. Tissue interactions in neural crest cell development and disease. (2013) *Science*. 341(6148): 860–3. PMID: 23970693
5. Maris JM, Mosse YP, Bradfield JP, Hou C, Monni S, Scott RH, et al. (2008) Chromosome 6p22 locus associated with clinically aggressive neuroblastoma. *N. Engl. J. Med*. 358 (24): 2585–93. PMID: 18463370.
6. Capasso M, Devoto M, Hou C, et al. (2009) Common variations in BARD1 influence susceptibility to high-risk neuroblastoma. *Nat Genet* 41: 718-23.
7. Diskin SJ, Hou C, Glessner JT, et al. (2009) Copy number variation at 1q21.1 associated with neuroblastoma. *Nature* 459: 987-91.
8. Nguyen le B, Diskin SJ, Capasso M, et al. (2011) Phenotype restricted genome-wide association study using a gene-centric approach identifies three low-risk neuroblastoma susceptibility loci. *PLoS Genet* 7: e1002026.
9. Wang K, Diskin SJ, Zhang H, Attiyeh EF, Winter C, Hou C, et al. (2011) Integrative genomics identifies LMO1 as a neuroblastoma oncogene. *Nature*. 469 (7329): 216–20. PMID: 21124317.
10. Diskin SJ, Capasso M, Schnepp RW, Cole KA, Attiyeh EF, Hou C, et al. (2012) Common variation at 6q16 within HACE1 and LIN28B influences susceptibility to neuroblastoma. *Nat. Genet*. 44 (10): 1126–30. PMID: 22941191.
11. Latorre V, Diskin SJ, Diamond MA, et al. (2012) Replication of Neuroblastoma SNP Association at the BARD1 Locus in African-Americans. *Cancer Epidemiology, Biomarkers & Prevention* 21: 658-63.
12. Capasso M, Diskin SJ, Totaro F, et al. (2013) Replication of GWAS-identified neuroblastoma risk loci strengthens the role of BARD1 and affirms the cumulative effect of genetic variations on disease susceptibility. *Carcinogenesis.* 34: 605-11.

13. Manolio T a, Collins FS, Cox NJ, Goldstein DB, Hindorff L a, Hunter DJ, et al. (2009) Finding the missing heritability of complex diseases. *Nature*. 461 (7265): 747–53. PMID: 19812666.

14. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, et al. (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*. 306: 2242–46. PMID: 1103388.

15. Carninci P, Kasukawa T, Katayama S, Gough J, FrithMC, et al. (2005) The transcriptional landscape of the mammalian genome. *Science*. 309: 1559–63 21. PMID: 16141072.

16. Mercer T, Dinger M, Mattick J. (2009) Long non-coding RNA insights into functions. *Nat. Rev. Genet*. 10 (March): 155–9. PMID: 19188922.

17. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. (2012) *Annu. Rev. Biochem*. 81: 145–66. PMID: 22663078.

18. Lee JT. Epigenetic Regulation by Long Noncoding RNAs. (2012) *Science*. 338 (6113): 1435–9. PMID: 23239728.

19. Mercer TR, Mattick JS. Structure and function of long noncoding RNAs in epigenetic regulation. (2013) *Nat. Struct. Mol. Biol.* 20(3): 300–7. PMID: 23463315.

20. Chu C, Qu K, Zhong F, Artandi S, Chang H. (2011) Genomic maps of lincRNA occupancy reveal principles of RNA chromatin interactions. *Mol. Cell.* 44(4):667–78.

21. Ponjavic J, Oliver PL, Lunter G, Ponting CP (2009) Genomic and transcriptional co-localization of protein-coding and long non- coding RNA pairs in the developing brain. *PLoS Genet* 5:e1000617. PMID: 19696892.

22. Knauss JL, Sun T. (2013) Regulatory mechanisms of long noncoding RNAs in vertebrate central nervous system development and function. *Neuroscience*. 235:200–14. PMID: 23337534.

23. Gibb EA, Brown CJ, Lam WL. (2011) The functional role of long non-coding RNA in human carcinomas. *Mol Cancer*; 10:38. 11. PMID: 21489289.

24. Prensner JR, Chinnaiyan AM. (2011) The emergence of lncRNAs in cancer biology. *Cancer Discovery*. 1(5): 391–407. PMID: 22096659.

25. Prensner JR, Iyer MK, Sahu A, Asangani I a, Cao Q, Patel L, et al. (2013) The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. *Nat. Genetics*; 45(11): 1392–8. PMID: 24076601.

26. Gutschner T, Hämmerle M, Eissmann M, Hsu J, Kim Y, Hung G, et al. (2013) The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Res.* Feb 1, 73(3): 1180–9. PMID: 23243023.

27. Federico S, Brennan R, Dyer M. Childhood Cancer and Developmental Biology: A Crucial Partnership. (2011) *Top. Dev. Biol*. (901): 1–10. PMID: 21295682.

28. B. N. Howie, P. Donnelly, and J. Marchini (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics* 5(6): e1000529.

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

30. Ørom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, et al. (2010) Long noncoding RNAs with enhancer-like function in human cells. *Cell.* 143(1): 46–58. PMID: 20887892.

31. Guil, S., Soler, M., Portela, A., Carre` re, J., et al.. (2012). Intronic RNAs mediate EZH2 regulation of epigenetic targets. *Nat. Struct. Mol. Biol*. 19, 664–670.

32. Gyurján I, Sonderegger B, Naef F, Duboule D. (2011) Analysis of the dynamics of limb transcriptomes during mouse development. *BMC Dev Biol*. Jan; 11:47. PMID: 3160909.

33. 1. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, et al. (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*. 148(1-2):84–98. PMID: 3339270.

34. Mossé YP, Laudenslager M, Longo L, Cole K a, Wood A, Attiyeh EF, et al. (2008) Identification of ALK as a major familial neuroblastoma predisposition gene. *Nature*. Oct; 455(7215): 930–5.

35. Mosse YP, Laudenslager M, Khazi D, et al. (2004) Germline PHOX2B mutation in hereditary neuroblastoma. *Am J Hum Genet.*; 75: 727-30. PMCID: PMC1182065.

36. Delaneau, O., Marchini, J. & Zagury, J.F. (2011) A linear complexity phasing method for thousands of genomes. *Nat Methods* 9, 179-81.

37. Howie, B.N., Donnelly, P. & Marchini, J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 5, e1000529.

38. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 39, 906-13.

39. Zhang X, Bailey SD, Lupien M. (2014) Laying a solid foundation for Manhattan--'setting the functional basis for the post-GWAS era'. *Trends Genet.* 30(4): 140-9.

FIGURE LEGENDS

Figure 1. Fine mapping of 6p22 identifies CASC15-S as a candidate cis-acting neuroblastoma susceptibility gene. (**a**) Regional association plot of single nucleotide polymorphisms obtained from genome wide imputation (n=2,817 neuroblastoma cases, n=7,473 controls) identifying a narrow peak of 32 single nucleotide polymorphisms (SNPs) with *p*-values < $1x10^{-10}$ ($p = 4.67x10^{-10} – 4.81x10^{-17}$) residing within the lncRNA *CASC15*. (**b**) Linkage dissociation structure of this region in Northern European (CEU) population demonstrates a signal that overlaps with these SNPs, refining our initial LD block of 94.2kb down to 30kb. (SNP annotations: color = representative of LD r^2 value; □ = predicted coding or 3′ UTR region; **Δ** = nonsynonymous; * = tfbccons, conserved motif at transcription factor binding site, ☒ = mcs44placental, highly conserved region in placental mammals).

Figure 2. CASC15-S is the predominant lncRNA isoform expressed in neuroblastoma. (**a**) Graphical representation of *CASC15* lncRNA transcripts observed to originate from this locus via RNASeq and confirmed by 5'/3' RACE. This locus includes two predominant transcripts on the sense strand: a 1.9kb *CASC15* transcript (current RefSeq for this lncRNA) and a novel shorter (1.2kb) transcript, *CASC15-S* that shares the last 3 exons with the long isoform of *CASC15*. Epigenetic data from the ENCODE project supports active transcription (RNA PolII occupancy) and an enhancer like function (H3K27Ac, H3K4Me3) for this region, including a proximal VISTA enhancer element (hs1335) [31]. RNASeq, RNA PolII and P300 marks were taken from SK-N-SH cells, enhancer tracks used were from MGG8 human glioblastoma stem cells. [48] (**b**) Quantification of normalized CASC15 isoform expression across a panel of 16 normal primary tissues (Human Body Map project) indicates predominant expression of CASC15-S in brain. (**c**) RNA sequencing performed on 108 primary neuroblastoma tumors analyzed for unique isoform expression provides supporting evidence for the predominance of a short *CASC15* isoform, expressed at 20-fold over *CASC14* and 40-fold the expression levels of full-length *CASC15* (**d**) RNA fluorescent in-situ hybridization

(RNA-FISH) was conducted in NGP neuroblastoma cells for several transcripts known to exist at this locus. To ensure probe specificity, odd and even pools of fluorescently labeled oligonucleotide probes were used to tile *CASC15-S* and *VISTA hs1335*. The yellow fluorescence observed in the "overlay" panels was obtained from overlap of the *hs1335* and *CASC15-S* fluorescent signals and indicate nuclear colocalization of these two transcripts (40x magnification).

Figure 3. *CASC15-S* expression is highly associated with neuroblastoma patient outcome. (**a**) A filtration strategy was used to refine imputed polymorphisms at the 6p22.3 locus for functional SNPS, resulting in four highly significant polymorphisms within regions of regulator chromatin and possessing putative enhancer activity. Further filtering based on evolutionary conservation yielded rs9295534. (**b**) Evidence from glioblastoma cells that rs9295534 overlaps the closest enhancer site to *CASC15-S*, as evidenced by DNaseI sensitivity (DS), H3K4me1 and H3K27ac marking. Light blue line indicates the location of rs9295534. (**c**) The imputed polymorphism rs9295534 was chosen for further characterization due to an exceptionally low p-value (p=3.51x10^{-12}) and encapsulation within a region of H3K27Ac marks in several fetal tissues. (**d**) 1500bp risk or non-risk fragments inserted into a luciferase reporter vector upstream of a minimal promoter (pGL4.23) demonstrated significantly lower transcriptional activity with the risk genotype fragment. Luciferase activity was normalized using a contransfected Renilla luciferase under the control of the CMV promoter (pGL4.75).

Figure 4. (**a**) Clinically annotated primary neuroblastoma tumors (n=250) obtained at diagnosis hybridized to Affymetrix Human Exon 1.0ST microarrays that high-risk stage 4 neuroblastomas (n=220) demonstrate significantly lower expression of *CASC15-S* than low-risk stage 1 tumors (n=30). (**b**) Kaplan-Meier curve demonstrating significantly poorer overall survival for children with tumors expressing low levels of *CASC15-S* (n=163 for group "low", n=87 for group "high", *adj. p* = 3.2x10^{-06}). (**c**) Relevant metrics for selection of high-risk patient tumors used for differential gene expression analyses between patients with high (n=74) and low (n=146) *CASC15-S* levels. Median survival was significantly increased in patients with high levels of *CASC15-S.* (**d**) The most highly significantly regulated pathway using GSEA was identified to be a set of genes downregulated in poor outcome neuroblastoma (Asgharzadeh neuroblastoma poor survival down). Patients with high *CASC15-S* expression demonstrated enrichment of these genes, suggesting that *CASC15-S* acts in a protective manner. (*** $p < 0.0001$).

Figure 5. Depletion of *CASC15-S* increases the aggressiveness of neuroblastoma cells. (**a**) *CASC15-S* expression was investigated in a panel of neuroblastoma cell lines (n=21) and was normalized relative to the geometric mean of *GUSB*, *HPRT* and *TBP* housekeeping genes. *CASC15-S* demonstrated a wide range of expression across neuroblastoma cell lines. (**b**) SK-N-BE2 neuroblastoma cells transiently transfected with siRNA targeting an exon common to both *CASC15* and *CASC15-S* isoforms, or (**c**) specifically targeting only the unique exon of *CASC15-S,* show a significant increase in cellular proliferation. (**d**) Stable depletion of *CASC15-S* in SK-N-BE2 cells was achieved with lentiviral transduction of shRNA, and recapitulated the increased growth observed with transient knockdown. (**e**) Forced ectopic expression of *CASC15-S* cDNA was able to rescue the growth characteristics of SK-N-BE2 shCASC15-S cells, reverting their growth pattern to that of wild-type cells (**f**) Morphological observation of SK-N-BE2 cells stably depleted of *CASC15-S* showed that cells were substantially larger than control cells (scale bar = 100µm). (**g**) Cell area was measured in biological triplicate (n=10 for each replicate) and quantified in ImageJ, where the area of shCASC15 cells was found to be 3.1-fold increased over controls (*** p < 0.001).

Figure 6. *CASC15-S* regulates a subset of genes involved in neural differentiation and neuroblastoma tumorigenesis. (**a**) SK-N-BE2 cells constitutively depleted of *CASC15-S* demonstrated an increased migratory capacity in wound healing assays (t=24h). (**b**) Linear regression comparison of wound closure at regular intervals demonstrates a clear enhancement of migration in silenced cells (EV = Empty Vector). (**c**) Most significantly altered pathways in SK-N-SH and BE2 neuroblastoma cells following depletion of *CASC15-S* and subjected to Ingenuity pathway analysis (Ingenuity® Systems, www.ingenuity.com). Pathways shown were the top gene signatures to arise from differential analysis, and indicate activation of cellular programs of proliferation, migration and metastasis (shown in red), as well as downregulation of several pathways known to modulate neural-specific development (shown in green).

Table 1. Differential expression analysis of long noncoding RNAs between high- and low-risk neuroblastomas. Gene expression analysis was conducting using 220 high- and

30 low-risk primary neuroblastoma tumors, focusing on differences in lncRNA expression. The top differentially regulated lncRNA was CASC15 (shown here as LINC00340), which was significantly lower in high-risk disease (ANOVA, p = 3.6x10^{-17}).

SUPPLEMENTARY FIGURE LEGENDS

Supplementary Figure 1. Supporting evidence for multiple transcripts mapping to chromosome 6p22.3. (**a)** Graphical representation of CASC14 and the multiple *CASC15* isoforms annotated in the Ensembl Genome Browser. The long (CASC15-003) and short isoforms of *CASC15* (CASC15-004) are highlighted in red and blue, respectively. (**b**) RNA paired-end tagged transcripts in SK-N-SH neuroblastoma cells demonstrate the existence of capped and polyadenylated CASC15 transcripts. *CASC15-S* reads are denoted by the red arrows and were found in both the nucleus and cytosol.

Supplementary Figure 2. Support for *CASC15-S* as a long noncoding transcript. (**a**) Structural features of capping and polyadenylation by as evidenced by CAGE and PolyA-Seq, and conservation by 100 vertebrate multiple sequence alignment and by placental mammal chain alignments from the UCSC browser. (**b**) *CASC15-S* exemplifies non-coding status by coding-potential assessment tool (CPAT) analysis. (**c**) Nuclear localization by quantification of RNA-PET in the SK-N-SH cell line. (**d**) Expression for CASC15 isoforms across 51 ENCODE cell lines.

Supplementary Figure 3. (**a**) Results of Sanger sequencing from 5' and 3'-RACE products of *CASC15-S*. This isoform matches a cDNA clone (FLJ37399, clone BRAMY2027587) derived from amygdala, but is truncated at the 3' end of the gene to form a 1181nt product. Exons 1-4 are shown as different colors. (**b**) RNA fluorescent in-situ hybridization (RNA-FISH) conducted in NGP cells demonstrates predominantly nuclear SOX4. NB-69 cells labeled with probes full-length CASC14 (**c**) and/or CASC15 (**d**) demonstrated transcripts with very low or no observable levels (40x magnification).

Supplementary Figure 4. Observed lack of correlation between rs6939340 genotype and expression of 6p22 gene products. (**a**) HuEx data from patient samples (n=250) demonstrates that there is not a significant correlation between risk and non-risk genotypes and expression of *CASC15*, *CASC15-S*, *CASC14* or *SOX4*. We also observed no significant correlation using either of the two previously published SNPs: rs4712653 and rs9295536 (**b**) Neuroblastoma cell line genotypes for rs6939340. Cell line expression of gene products at 6p22.3 did not correlate with genotype (see figure 2c).

Supplementary Figure 5. Supporting data for rs9295534 as a candidate SNP. Evidence from the UCSC genome browser to indicate active transcription, shown by DNase clustering and overlap of multiple transcription factor binding sites encompassing rs9295534 (denoted by light blue line).

Supplementary Figure 6. Additional expression data for CASC15 isoforms in neuroblastoma cell lines and patients. (**a**) Expression of CASC15-S is not significantly different between MYCN amplified and non-amplified tumors (*p* = 0.29) (**b**) Overall survival of only high-risk patients was compared between *CASC15-S* expression groups (n=146 for group low, n=74 for group high, *adj. p* = 0.002). (**c**) Full length CASC15 (n=12 unique probes) expression from patient tumors (n=250) hybridized to Affymetrix Human Exon 1.0ST arrays demonstrates significantly less expression of *CASC15* compared to *CASC15-S* (n=1 unique probe). (**d**) Similar to what was observed with *CASC15-S*, low levels of full-length *CASC15* correlate with poor overall survival (*p* = 0.007). (**e**) Low expression of full length *CASC15* correlates significantly with stage 4 disease (p<0.0001). (**f**) Taqman-based expression levels of full-length *CASC15* in a panel of 26 neuroblastoma cell lines.

Supplementary Figure 7. Additional cell lines showing increased neuroblastoma cell growth following *CASC15-S* depletion. (**a**) IMR-5 and (**b**) SK-N-SH cells were transfected with 50nM siRNA specific for *CASC15-S* and monitored for real time cell growth. siPLK-1, was used as a positive control. (**c**) SK-N-SH and (**d**) NGP neuroblastoma cells were stably depleted of *CASC15-S* and exhibited a robust increase in cell growth compared to control counterparts. Conversely, depletion of *CASC14* has no effect on the growth of neuroblastoma cells. (**e**) Kelly and (**f**) Ebc1 neuroblastoma cells were transfected with siRNA specific for *CASC14* (s59455, targeting only exon 3), *CASC15*, *PLK1* or a non-targeting control. Little to no change in growth kinetics was observed in si*CASC14*-transfected cells. (***p<0.0001)

Supplementary Figure 8. Depletion of the long isoform of *CASC15* does not impact neuroblastoma cell viability. (**a and c**) NB-16 and SK-N-AS neuroblastoma cells were transfected in triplicate with 50nM siRNA specific for the long isoform of *CASC15* (n271792), *GAPDH*, *PLK-1* or a scrambled control. Cell viability was assessed via the Cell TiterGlo Assay (Promega) 72 hours after transfection. No change in cell viability was observed despite substantial depletion of full length *CASC15* (**b and d**). (**e**) shCASC15-S cells demonstrate substantial depletion of *CASC15-S* levels. SK-N-BE2, NB-16 and SK-N-SH cells were assayed for CASC15-S by Taqman PCR compared to their empty vector counterparts.

Supplementary Figure 9. Cell viability assays for neuroblastoma cell lines stably depleted of CASC15-S. (**a**) SK-N-BE2, (**b**) SK-N-SH and (c) NGP cells plated in triplicate at 8x10^3 cells/well in a 96-well plate were cultured for 72h and then read using the Cell TiterGlo cell viability assay (Promega). Neuroblastoma cells stably depleted of CASC15 showed a significant increase in cell viability and cell number as measured by intracellular ATP.

Supplementary Table 1. List of imputation results at the 6p22.3 region sorted by significance. Four candidate polymorphisms based on DNaseI hypersensitivity are highlighted in yellow, rs9295536 is highlighted using red font. Column descriptions are listed on sheet 2.

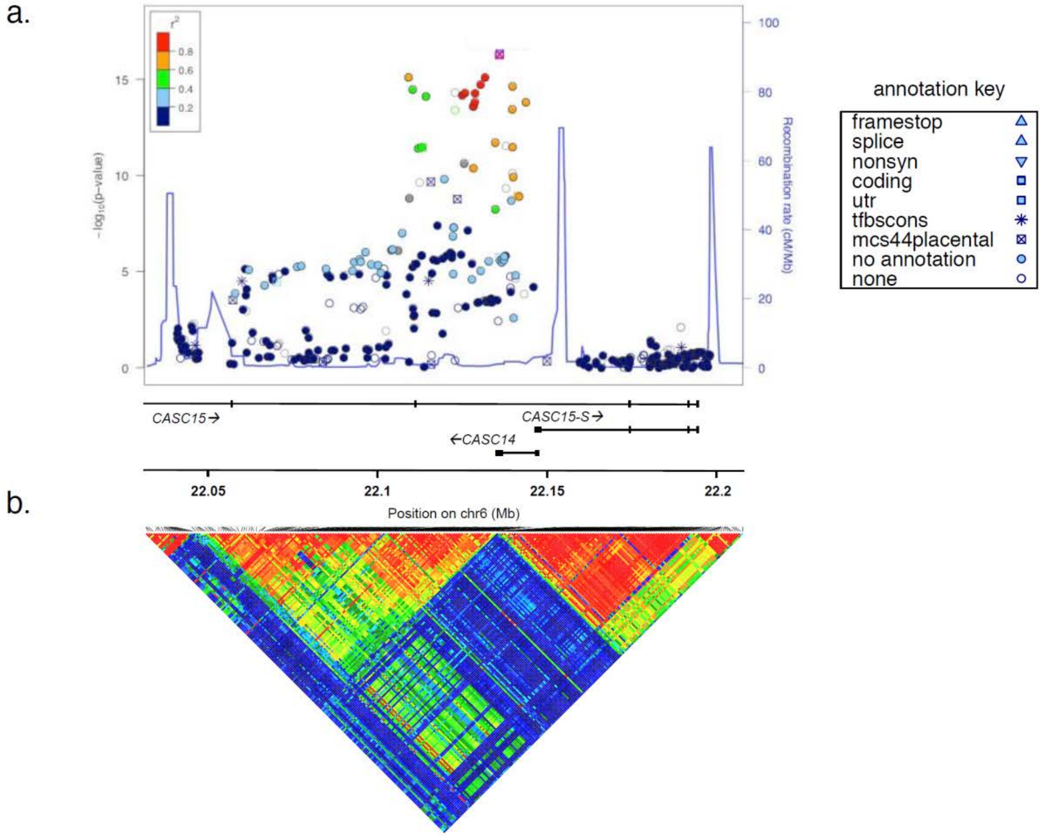

a.
b.
r²
0.8
0.6
0.4
0.2
15
10
5
0
-log₁₀(p-value)
100
80
60
40
20
0
Recombination rate (cM/Mb)
CASC15→
CASC15-S→
←CASC14
22.05
22.1
22.15
22.2
Position on chr6 (Mb)
annotation key
framestop
splice
nonsyn
coding
utr
tfbscons
mcs44placental
no annotation
none

Figure 1

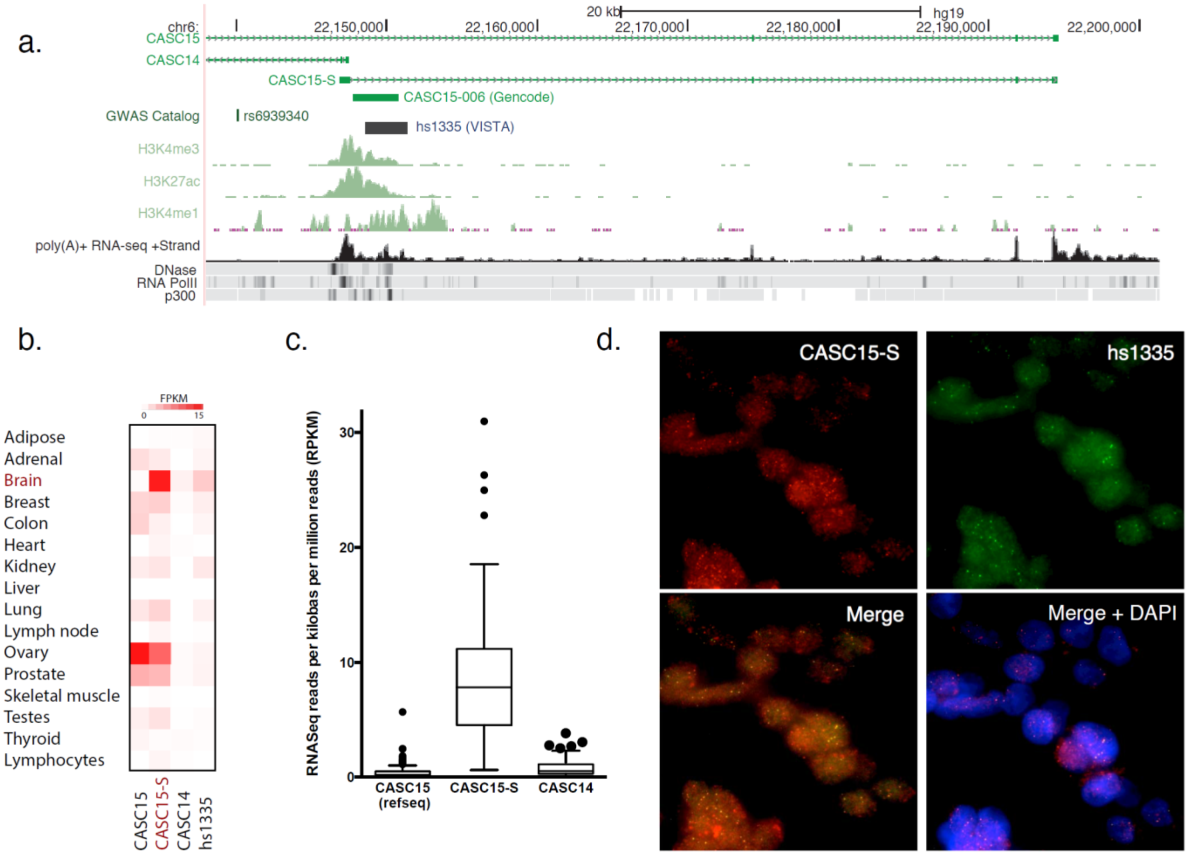

a.
20 kb
hg19
chr6:
22,150,000
22,160,000
22,170,000
22,180,000
22,190,000
22,200,000
CASC15
CASC14
CASC15-S
CASC15-006 (Gencode)
GWAS Catalog
rs6939340
hs1335 (VISTA)
H3K4me3
H3K27ac
H3K4me1
poly(A)+ RNA-seq +Strand
DNase
RNA PolII
p300
b.
FPKM
0
15
Adipose
Adrenal
Brain
Breast
Colon
Heart
Kidney
Liver
Lung
Lymph node
Ovary
Prostate
Skeletal muscle
Testes
Thyroid
Lymphocytes
CASC15
CASC15-S
CASC14
hs1335
c.
RNASeq reads per kilobas per million reads (RPKM)
30
20
10
0
CASC15 (refseq)
CASC15-S
CASC14
d.
CASC15-S
hs1335
Merge
Merge + DAPI

Figure 2

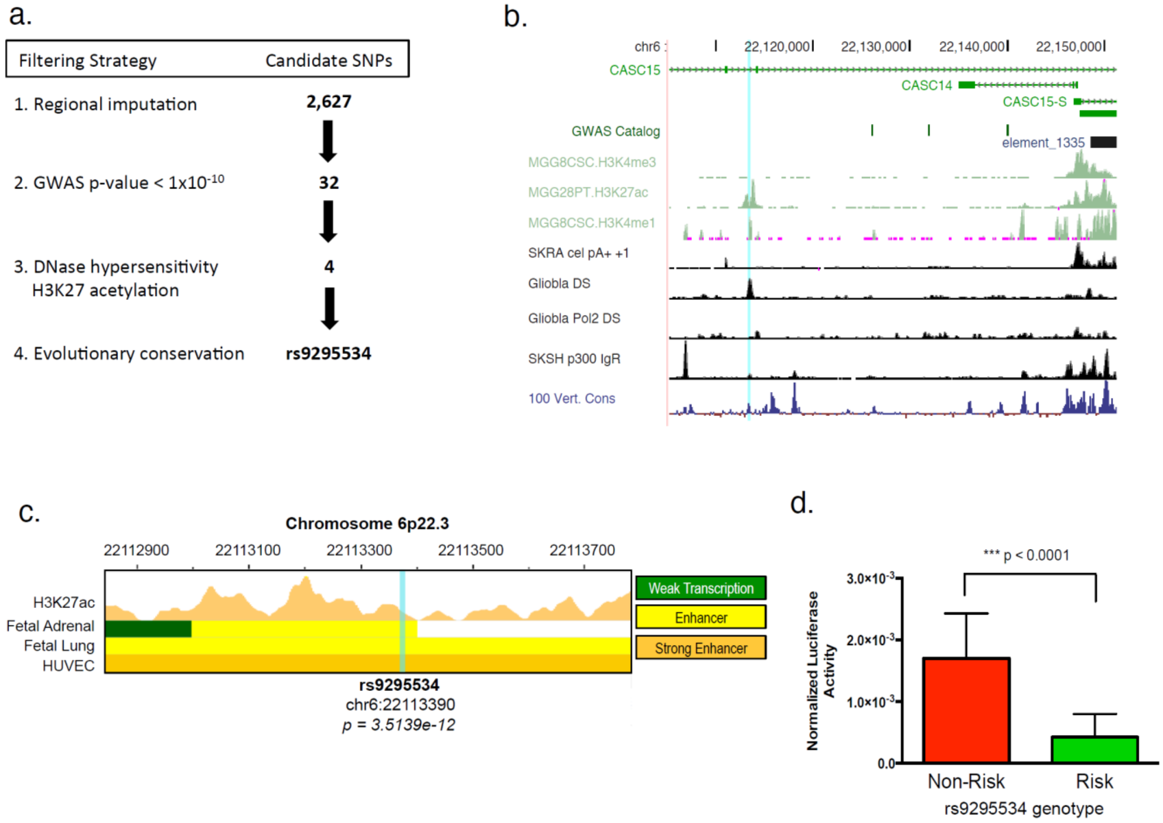
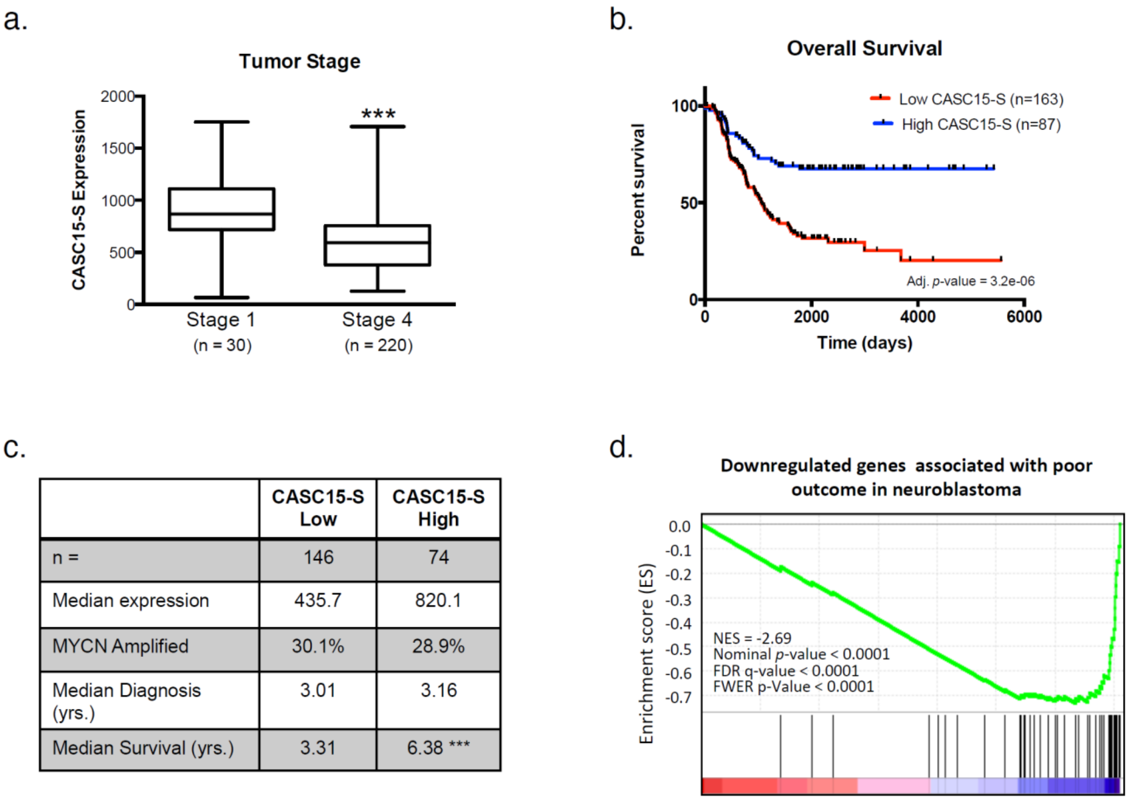

a.
Filtering Strategy
Candidate SNPs
1. Regional imputation
2,627
2. GWAS p-value < 1x10-10
32
3. DNase hypersensitivity H3K27 acetylation
4
4. Evolutionary conservation
rs9295534
b.
chr6:
22,120,000
22,130,000
22,140,000
22,150,000
CASC15
CASC14
CASC15-S
GWAS Catalog
element_1335
MGG8CSC.H3K4me3
MGG28PT.H3K27ac
MGG8CSC.H3K4me1
SKRA cel pA+ +1
Gliobla DS
Gliobla Pol2 DS
SKSH p300 IgR
100 Vert. Cons
c.
Chromosome 6p22.3
22112900
22113100
22113300
22113500
22113700
H3K27ac
Fetal Adrenal
Fetal Lung
HUVEC
Weak Transcription
Enhancer
Strong Enhancer
rs9295534
chr6:22113390
p = 3.5139e-12
d.
*** p < 0.0001
3.0×10-3
2.0×10-3
1.0×10-3
0.0
Normalized Luciferase Activity
Non-Risk
Risk
rs9295534 genotype

Figure 3



	CASC15-S Low	CASC15-S High
n =	146	74
Median expression	435.7	820.1
MYCN Amplified	30.1%	28.9%
Median Diagnosis (yrs.)	3.01	3.16
Median Survival (yrs.)	3.31	6.38 ***

Figure 4

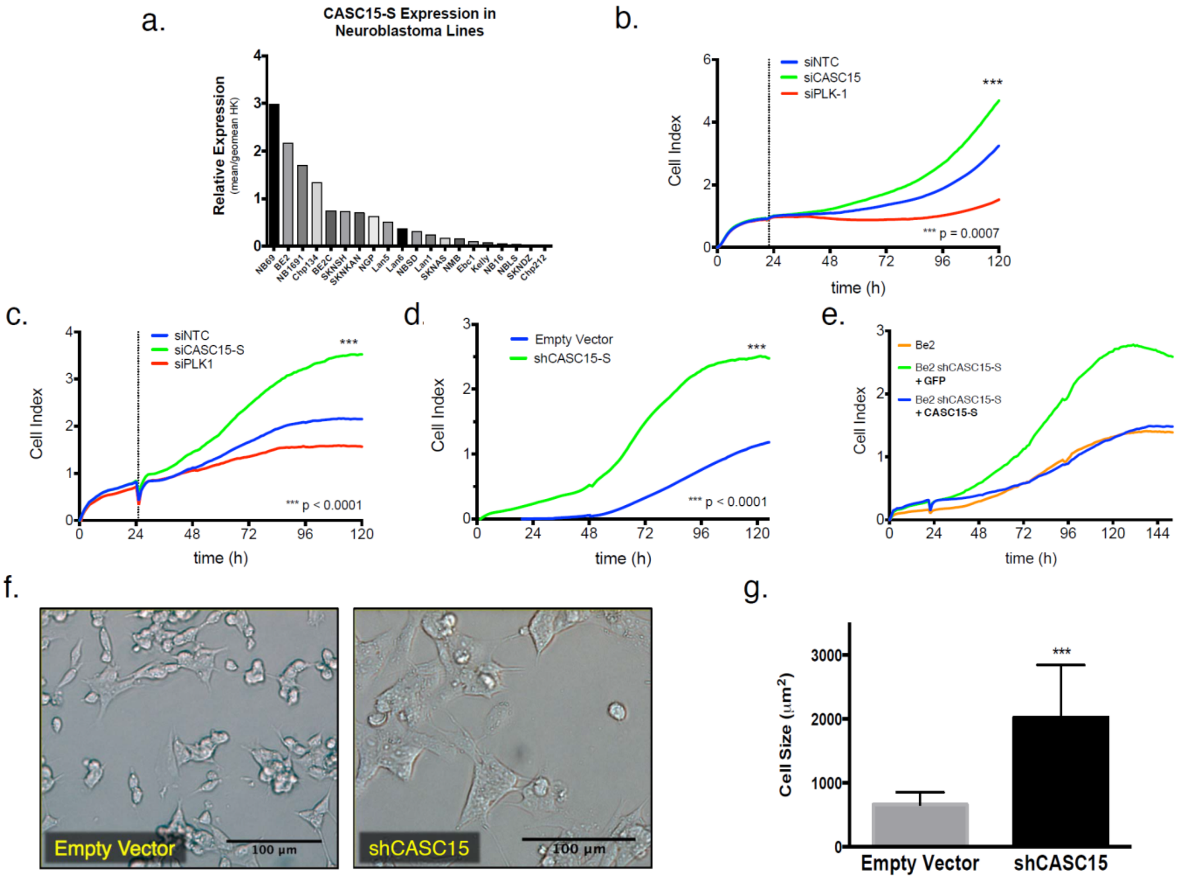
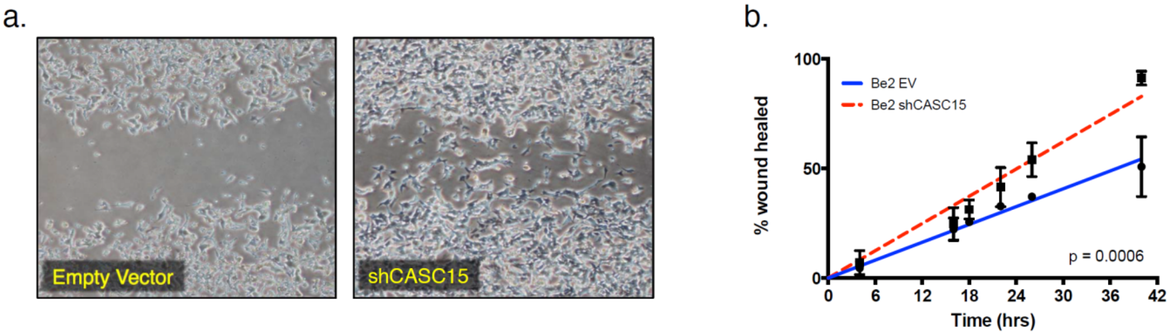

a.
CASC15-S Expression in Neuroblastoma Lines
Relative Expression (mean/geomean HK)
0
1
2
3
4
NB69
BE2
NB1691
Chp134
BE2C
SKNSH
SKNKAN
NGP
Lan5
Lan6
NBSD
Lan1
SKNAS
NMB
Ebc1
Kelly
NB16
NBLS
SKNDZ
Chp212
b.
siNTC
siCASC15
siPLK-1
***
Cell Index
0
2
4
6
*** p = 0.0007
0
24
48
72
96
120
time (h)
c.
siNTC
siCASC15-S
siPLK1
***
Cell Index
0
1
2
3
4
*** p < 0.0001
0
24
48
72
96
120
time (h)
d.
Empty Vector
shCASC15-S
***
Cell Index
0
1
2
3
*** p < 0.0001
0
24
48
72
96
120
time (h)
e.
Be2
Be2 shCASC15-S + GFP
Be2 shCASC15-S + CASC15-S
Cell Index
0
1
2
3
0
24
48
72
96
120
144
time (h)
f.
Empty Vector
100 µm
shCASC15
100 µm
g.
Cell Size (µm²)
0
1000
2000
3000
***
Empty Vector
shCASC15

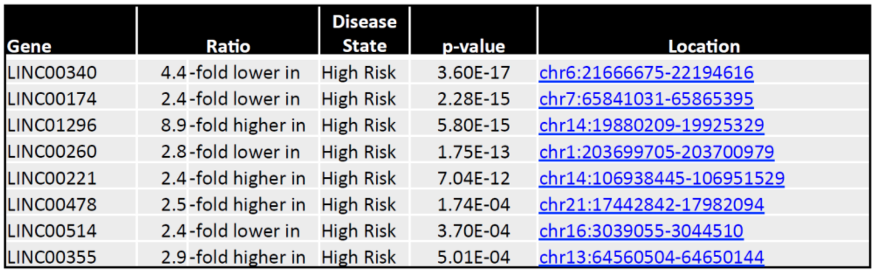
Figure 5



c.

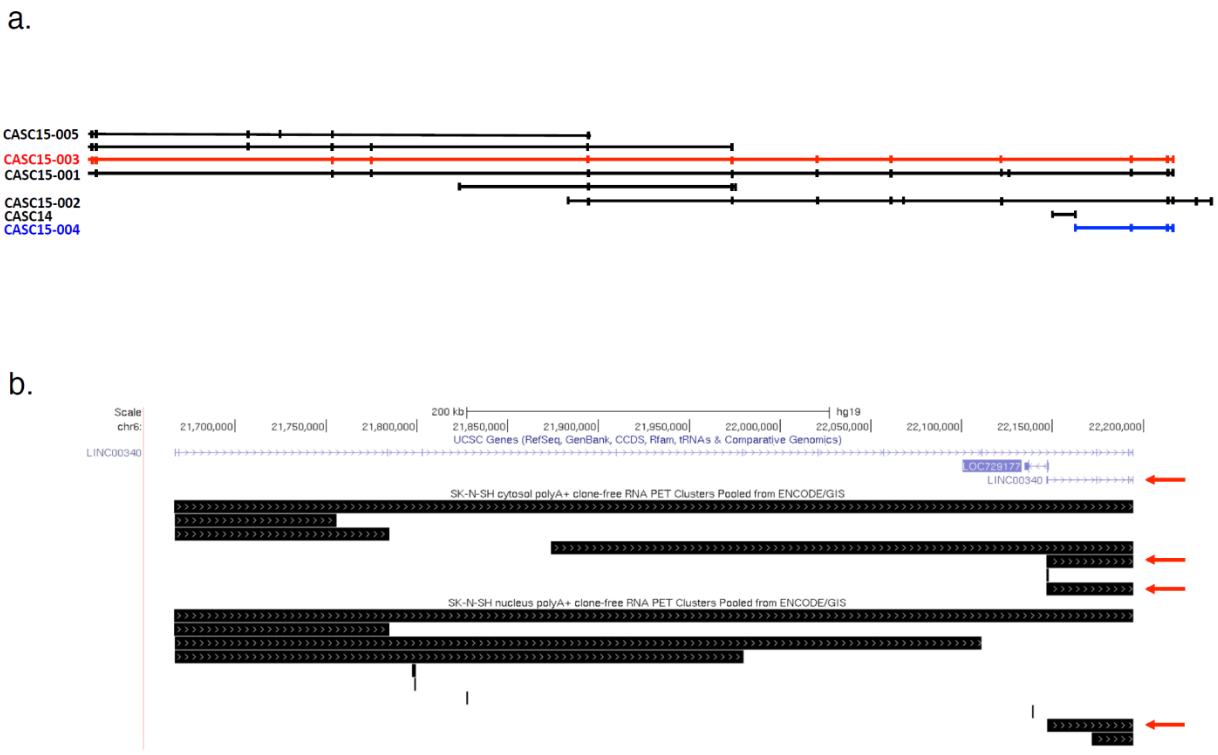
Diseases or Functions	p-Value	Activation z-score	# Molecules
cell movement	2.65x10^{-14}	5.283	113
migration of cells	2.84x10^{-14}	4.71	105
homing of cells	2.07x10^{-10}	4.36	42
angiogenesis	2.30x10^{-11}	3.718	50
development of blood vessel	3.92x10^{-13}	3.657	59
vasculogenesis	2.66x10^{-14}	3.338	57
cell movement of tumor cell lines	1.63x10^{-11}	2.972	55
proliferation of connective tissue cells	2.39x10^{-12}	2.762	48
metastasis	1.44x10^{-12}	2.081	51
proliferation of tumor cells	1.07x10^{-12}	0.44	59
organization of cytoskeleton	7.55x10^{-7}	-2.187	37
organization of cytoplasm	2.06x10^{-6}	-2.195	38
extension of cellular protrusions	2.42x10^{-4}	-2.202	9
migration of neural stem cells	5.06x10^{-8}	-2.219	5
differentiation of neurons	8.31x10^{-4}	-2.303	13
migration of neurons	5.65x10^{-8}	-2.369	15
flux of Ca2+	5.35x10^{-3}	-2.395	9
migration of stem cells	1.75x10^{-6}	-2.433	6
ion homeostasis of cells	2.71x10^{-3}	-2.577	15
outgrowth of neurites	1.35x10^{-4}	-2.801	15

Figure 6

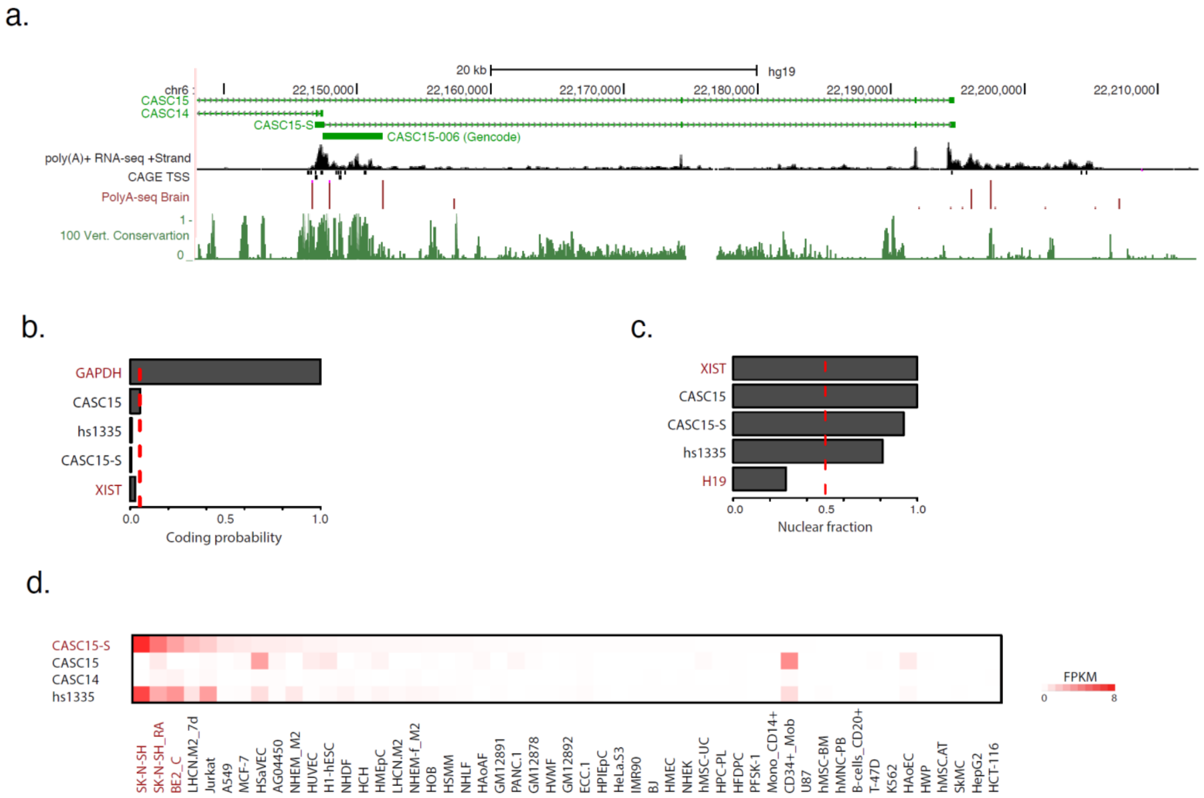


Gene	Ratio	Disease State	p-value	Location
LINC00340	4.4-fold lower in	High Risk	3.60E-17	chr6:21666675-22194616
LINC00174	2.4-fold lower in	High Risk	2.28E-15	chr7:65841031-65865395
LINC01296	8.9-fold higher in	High Risk	5.80E-15	chr14:19880209-19925329
LINC00260	2.8-fold lower in	High Risk	1.75E-13	chr1:203699705-203700979
LINC00221	2.4-fold higher in	High Risk	7.04E-12	chr14:106938445-106951529
LINC00478	2.5-fold higher in	High Risk	1.74E-04	chr21:17442842-17982094
LINC00514	2.4-fold lower in	High Risk	3.70E-04	chr16:3039055-3044510
LINC00355	2.9-fold higher in	High Risk	5.01E-04	chr13:64560504-64650144

Table 1



Supplementary Figure 1

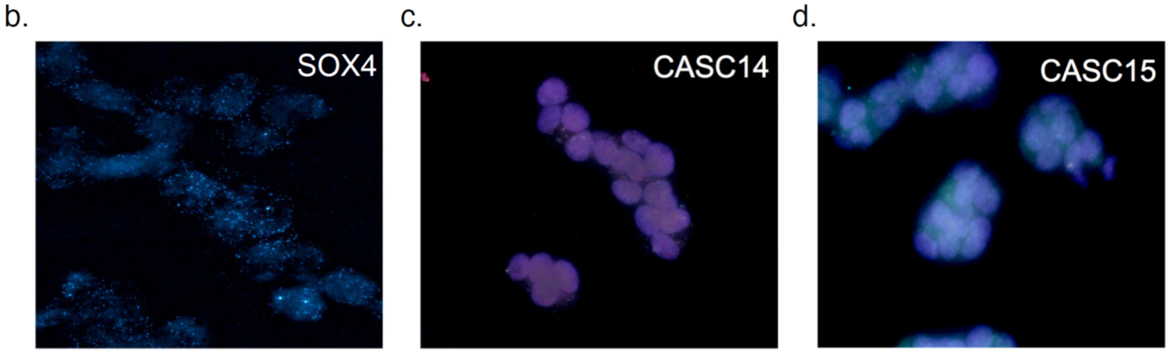


Supplementary Figure 2

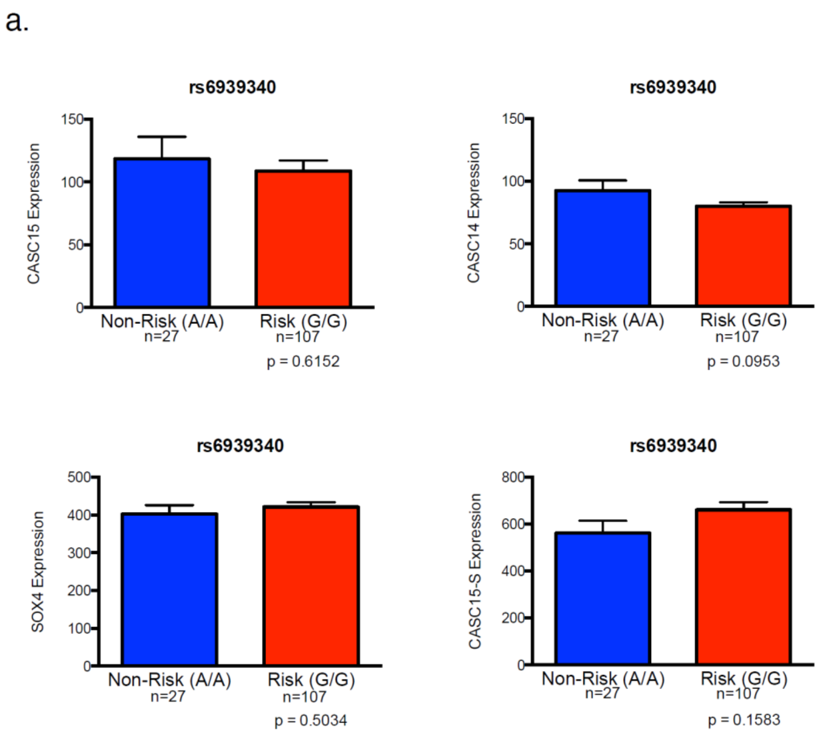
a.

CASC15-S Observed Sequence (1181nt)

cttttcctcacagctccacggctgcatctccgtgggcacgcaagcttcccctggttacctgagctgctcctgccgtctcccgcctgggcttcgccgtggtgcacccgatcccggaatcgtgcgtctgcgccct
gcgaaagaaggacctgctggcggagctccggccggggtctcctgcctcgcagctgggcgaggggacttggaggacagggtgaagctgcagaagacctggggtgggatggctagagaggacgcc
aaggactggggaaggggaagttaggaataccttacatccaatgcccacccgtgctccgcagggcaagggcagccgtcgcctcggccgcgtgcacccagctcaggctgttcccagggatttagtctgg
ggggacaacccatggcgagatgtggtggcattttacctcagagtggagctgaagatggataaacaaggtatctgatgtatctgcctgagaaggcagagctggagaaaggcggagcgagggagcgc
gtgaaaagaaagagatgccgaatgccgggtgattgtctgccgcttgctgtgcactctcattctagggaggatggcataatttataacccagcacatggatagaggaactgtaaattgtagtctggatgtcc
ccggcgctgtcgacgaaggaactgatgggctaggctcacaggcagaaggaattttccttgtcttggatgagacttttgacttggacttttgggttaagttctggagaccagaaggccaaaatcaaaagtat
gggcaggcttgatttctttagaagactccagcggagaactgtgtctccttgcttctgattctacatctccatccatgggccactgtttcagcaacctcagccagtgcaacacaacctcagccaagaagagtat
gcagagaaaggagtcccctacctgccacaaaactgttgtctgaaaactgtctcatattgtctcaagttgtcattcattgtgaattagacctgtttaacatgtaatctgcaacatgcttcactgtctaattttccaga
gcccctcatataaggaactgtattattggtataatcatcatggtgaagaagttggtatgtgggggagagatgacagaaacagagagtaagtcagagctggctgcctgacagataaaaaggaaatgacc
aaaaaaaaaaaaaaa



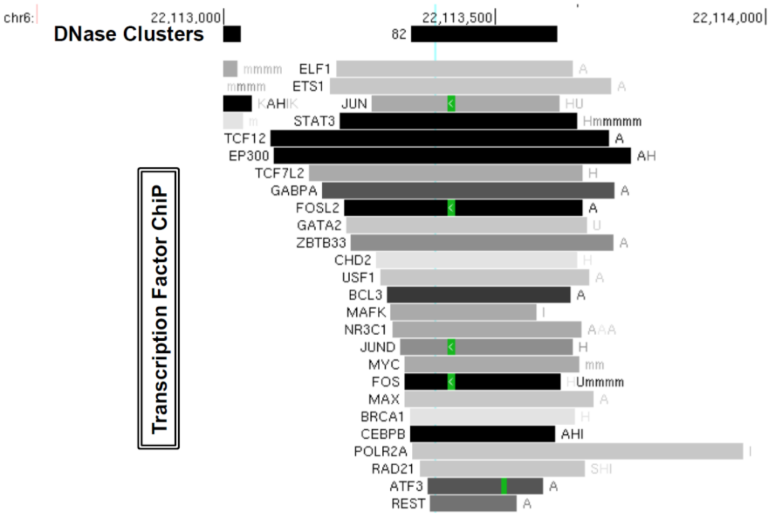
Supplementary Figure 3



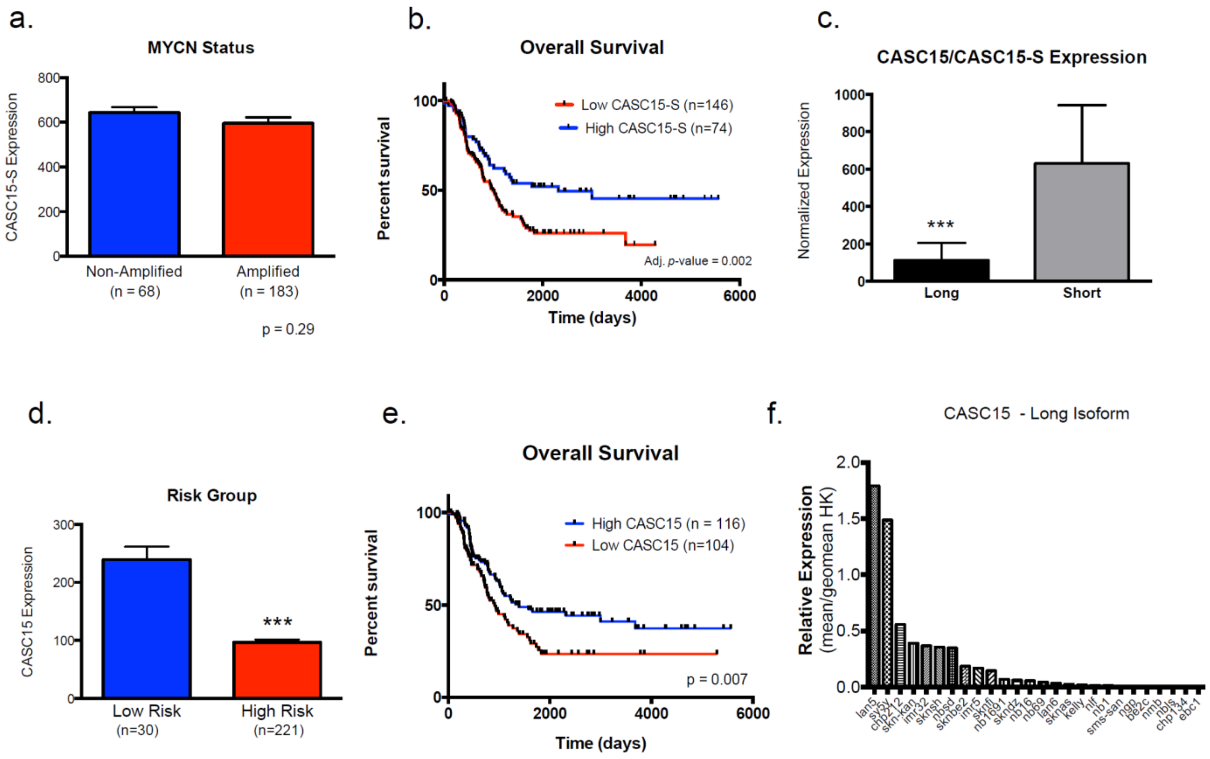
b.

Cell Line	Genotype for rs6939340
CHP-134	A/A
NB-16	A/A
SH-SY5Y	A/A
SK-N-SH	A/A
BE-2C	A/G
CHLA-150	A/G
CHLA-255	A/G
CHLA-79	A/G
CHP-100	A/G
CHP-212	A/G
IMR-32	A/G
IMR-5	A/G
KELLY	A/G
NB-1691	A/G
NB-1771	A/G
SK-N-AS	A/G
SK-N-BE2	A/G
SMS-SAN	A/G
CHLA-136	G/G
CHP-902-R	G/G
LAN-5	G/G
NB-EBc1	G/G
NB-1	G/G
NB-1643	G/G
NB-LS	G/G
NB-SD	G/G
NGP	G/G
NLF	G/G
NMB	G/G
SK-N-DZ	G/G
SK-N-FI	G/G
SMS-KAN	G/G
SMS-KCN	G/G

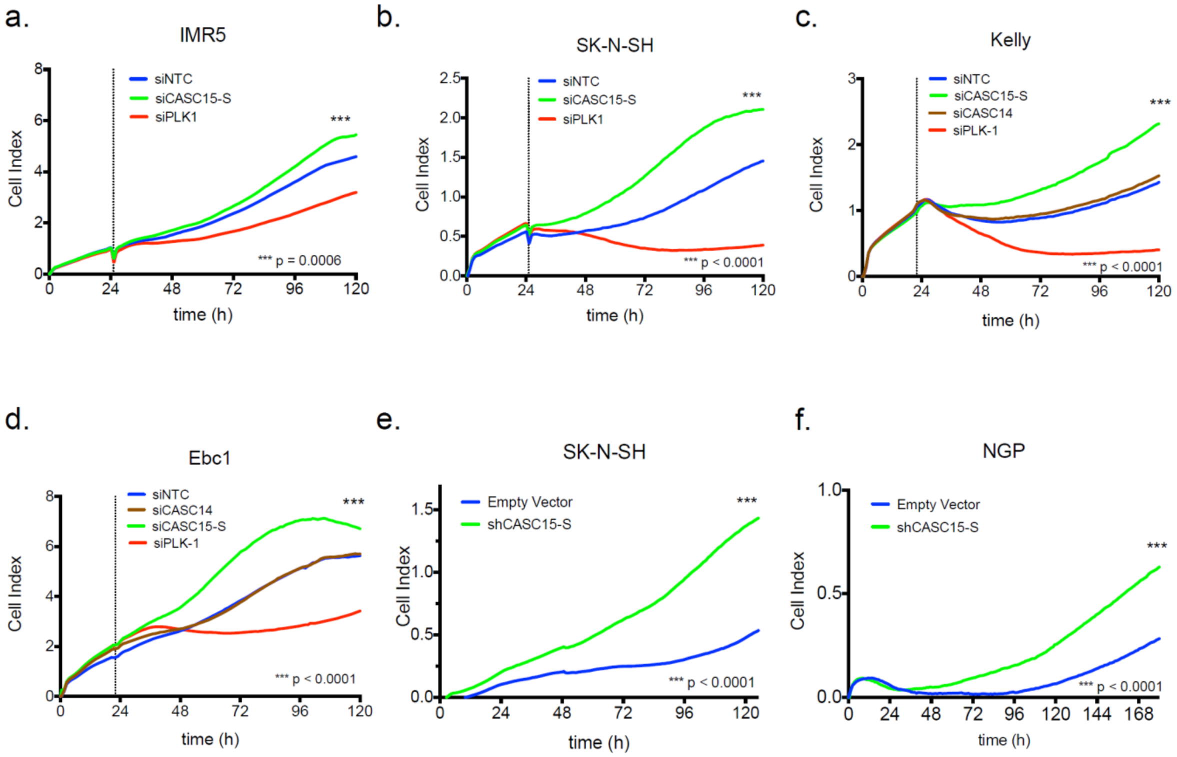
Supplementary Figure 4



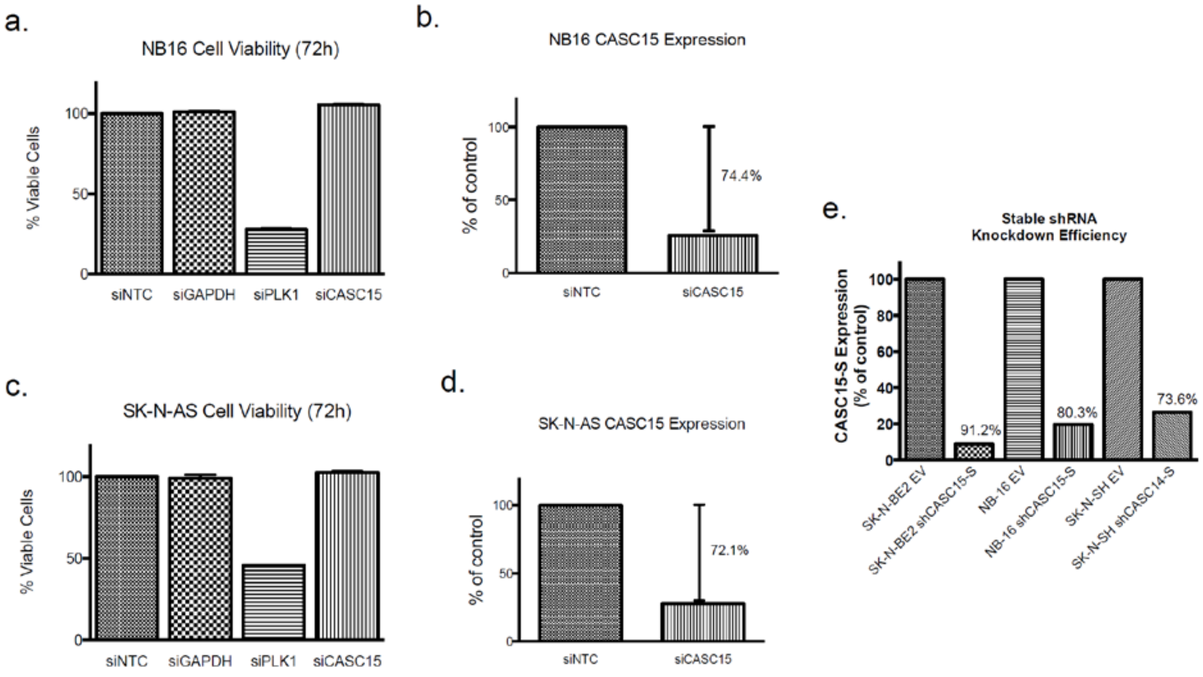
Supplementary Figure 5



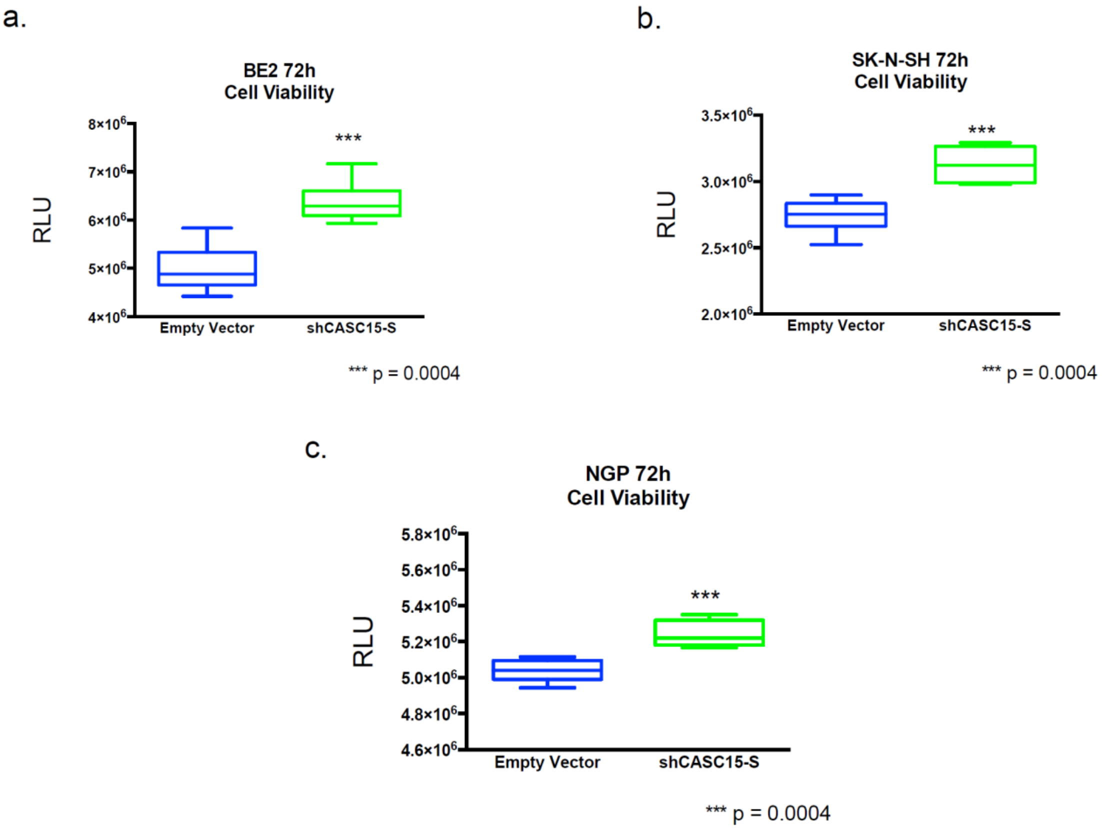
Supplementary Figure 6



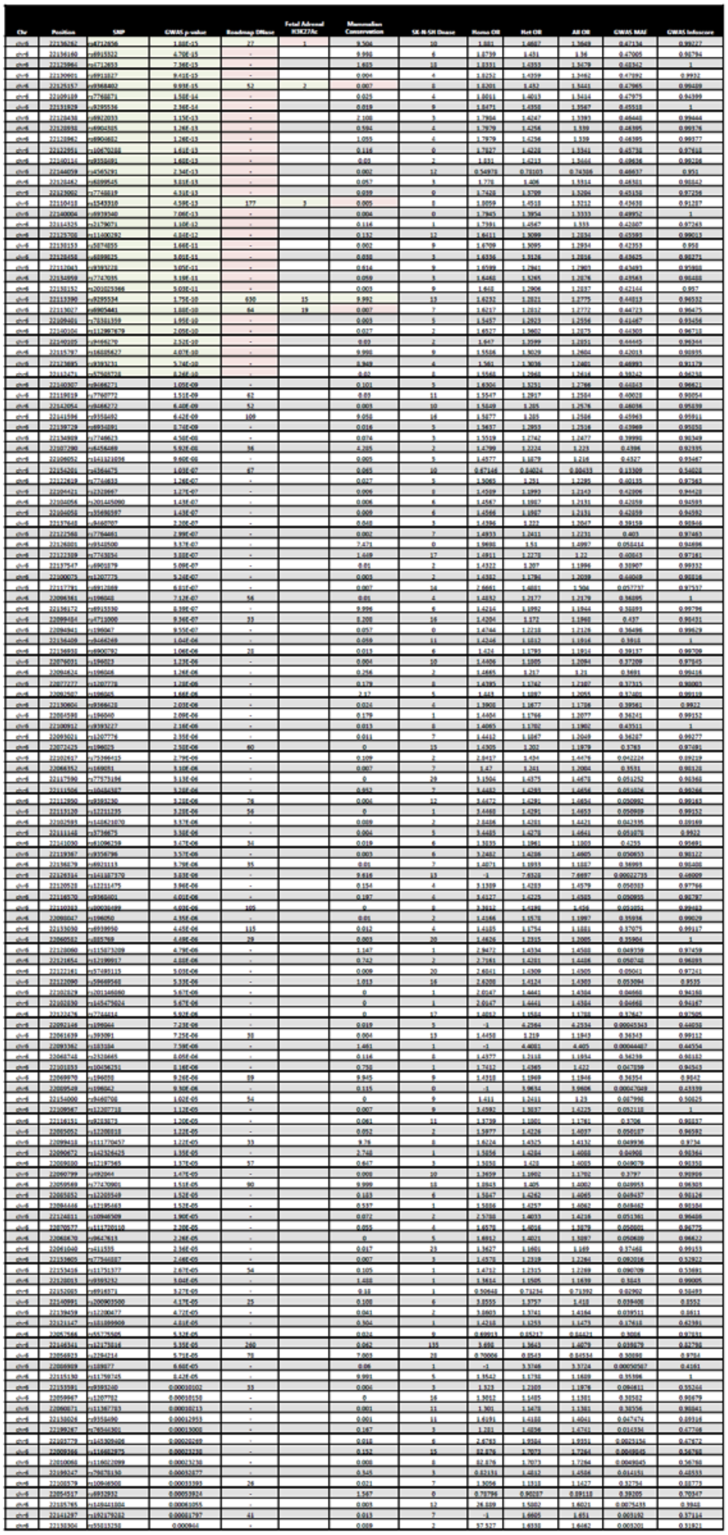
Supplementary Figure 7



Supplementary Figure 8



Supplementary Figure 9



Chr	Position	SNP	GWAS p-value	Roadmap DNase	Fetal Adrenal H3K27Ac	Mammalian Conservation	SK-N-SH Dnase	Homo OR	Het OR	All OR	GWAS MAF	GWAS Infoscore
chr6	22136262	rs4712656	1.88E-15	27	1	9.504	10	1.881	1.4667	1.3669	0.47134	0.99227
chr6	22136160	rs6915322	4.70E-15	-		9.998	6	1.8739	1.451	1.36	0.47005	0.98794
chr6	22129864	rs4712653	7.36E-15	-		1.605	10	1.8331	1.4333	1.3479	0.48347	1
chr6	22130001	rs6911827	9.41E-15	-		0.004	4	1.8252	1.4359	1.3462	0.47092	0.9932
chr6	22125157	rs9368402	9.93E-15	52	2	0.007	8	1.8201	1.452	1.3441	0.47065	0.99409
chr6	22109189	rs7768871	1.58E-14	-		0.025	4	1.8011	1.4013	1.3414	0.47975	0.94399
chr6	22131828	rs9295536	2.36E-14	-		0.019	9	1.8471	1.4358	1.3567	0.45518	1
chr6	22128430	rs6922033	1.15E-13	-		2.108	3	1.7904	1.4247	1.3393	0.46448	0.99444
chr6	22128898	rs6904305	1.26E-13	-		0.584	4	1.7879	1.4258	1.339	0.46395	0.99376
chr6	22128962	rs6904602	1.26E-13	-		1.055	4	1.7879	1.4256	1.339	0.46395	0.99377
chr6	22122951	rs10670288	1.61E-13	-		0.116	0	1.7827	1.4228	1.3341	0.45738	0.97618
chr6	22140114	rs9358491	1.60E-13	-		0.03	2	1.831	1.4213	1.3444	0.49636	0.99208
chr6	22144059	rs4565291	2.34E-13	-		0.002	12	0.54978	0.78103	0.74308	0.48637	0.951
chr6	22128462	rs6899545	3.81E-13	-		0.057	3	1.778	1.406	1.3314	0.46381	0.98842
chr6	22123002	rs7748819	4.31E-13	-		0.039	0	1.7428	1.3709	1.3204	0.45158	0.97256
chr6	22110418	rs1543910	4.58E-13	177	3	0.005	8	1.8059	1.4518	1.3212	0.43638	0.91287
chr6	22140004	rs6939340	7.08E-13	-		0.004	0	1.7945	1.3954	1.3333	0.49952	1
chr6	22114325	rs2179071	1.10E-12	-		0.116	1	1.7391	1.4567	1.333	0.42807	0.97265
chr6	22125700	rs11400282	4.84E-12	-		0.132	12	1.6411	1.3089	1.2834	0.45589	0.99013
chr6	22130153	rs5874855	1.66E-11	-		0.002	9	1.6709	1.3095	1.2934	0.42353	0.958
chr6	22128458	rs6899825	3.01E-11	-		0.038	3	1.6336	1.3126	1.2816	0.45625	0.98271
chr6	22112043	rs9393228	3.05E-11	-		0.614	9	1.6599	1.2941	1.2903	0.43483	0.95808
chr6	22134959	rs7747035	3.18E-11	-		0.059	3	1.6468	1.3265	1.2876	0.43563	0.98488
chr6	22130152	rs201025366	5.03E-11	-		0.003	9	1.648	1.2906	1.2837	0.42144	0.957
chr6	22113390	rs9295534	1.75E-10	630	15	9.992	13	1.6232	1.2821	1.2775	0.44013	0.96532
chr6	22113027	rs6905441	1.88E-10	64	19	0.007	7	1.6217	1.2812	1.2772	0.44723	0.96475
chr6	22109401	rs78301359	1.95E-10	-		0.003	5	1.5457	1.2923	1.2556	0.41467	0.93456
chr6	22140104	rs112987679	2.05E-10	-		0.027	2	1.6527	1.3602	1.2875	0.44303	0.96718
chr6	22140105	rs9466270	2.52E-10	-		0.03	2	1.647	1.3599	1.2851	0.44445	0.96344
chr6	22115797	rs16885627	4.07E-10	-		9.998	9	1.5506	1.3028	1.2604	0.42013	0.98895
chr6	22123885	rs9393231	5.74E-10	-		8.969	4	1.561	1.3036	1.2601	0.40893	0.91179
chr6	22112471	rs57568728	8.26E-10	-		0.02	8	1.5568	1.2968	1.2414	0.38242	0.94238
chr6	22140307	rs9466271	1.05E-09	-		0.101	5	1.6304	1.3251	1.2766	0.44843	0.96621
chr6	22119019	rs7760772	1.51E-09	62		0.03	11	1.5547	1.2917	1.2584	0.40028	0.98054
chr6	22142054	rs9466272	6.40E-09	52		0.003	10	1.5849	1.285	1.2576	0.48036	0.95039
chr6	22141596	rs9358492	6.42E-09	109		9.050	16	1.5877	1.285	1.2508	0.45963	0.95911
chr6	22139729	rs6934091	8.74E-09	-		0.016	5	1.5637	1.2953	1.2538	0.43969	0.95058
chr6	22134909	rs7748623	4.50E-08	-		0.074	3	1.5519	1.2742	1.2477	0.39990	0.98349
chr6	22107290	rs6456489	5.92E-08	36		4.205	2	1.4799	1.2224	1.223	0.4386	0.92335
chr6	22106052	rs141121098	9.60E-08	-		0.005	5	1.4577	1.1879	1.216	0.4327	0.93467
chr6	22154201	rs6564475	1.03E-07	67		0.045	10	0.67146	0.84024	0.80633	0.13309	0.56028
chr6	22122619	rs7748833	1.26E-07	-		0.027	5	1.5065	1.251	1.2295	0.40135	0.97563
chr6	22104421	rs2328887	1.27E-07	-		0.006	8	1.4509	1.1993	1.2143	0.42006	0.94428
chr6	22104056	rs201445090	1.43E-07	-		0.006	6	1.4567	1.1887	1.2131	0.42059	0.94589
chr6	22104058	rs35690597	1.43E-07	-		0.009	6	1.4566	1.1887	1.2131	0.42059	0.94592
chr6	22137648	rs9460707	2.20E-07	-		0.048	3	1.4386	1.222	1.2047	0.39159	0.98846
chr6	22122568	rs7764461	2.99E-07	-		0.002	7	1.4933	1.2411	1.2231	0.403	0.97463
chr6	22126001	rs9348500	3.57E-07	-		7.471	0	1.9698	1.51	1.4997	0.058414	0.94696
chr6	22122309	rs7743034	3.88E-07	-		1.449	17	1.4811	1.2278	1.22	0.40043	0.97161
chr6	22137547	rs6901879	5.09E-07	-		0.01	2	1.4322	1.207	1.1996	0.38907	0.99332
chr6	22100075	rs1207775	5.24E-07	-		0.003	2	1.4382	1.1786	1.2038	0.44048	0.98836
chr6	22117791	rs6912889	6.81E-07	-		0.007	16	2.6661	1.4881	1.506	0.057737	0.97537
chr6	22096361	rs196040	7.12E-07	56		0.01	4	1.4832	1.2177	1.2179	0.36095	1
chr6	22136172	rs6915330	8.38E-07	-		9.996	6	1.4214	1.1892	1.1944	0.38893	0.99796
chr6	22099484	rs4711000	9.36E-07	33		8.208	16	1.4204	1.172	1.1968	0.437	0.98431
chr6	22094941	rs196047	9.55E-07	-		0.057	0	1.4744	1.2218	1.2128	0.36496	0.99629
chr6	22156409	rs9464249	1.04E-06	-		0.059	11	1.4246	1.1812	1.1916	0.3818	1
chr6	22136938	rs6900792	1.06E-06	28		0.013	6	1.424	1.1789	1.1914	0.39137	0.99709
chr6	22079031	rs196023	1.23E-06	-		0.004	10	1.4406	1.1805	1.2094	0.37209	0.97845
chr6	22084624	rs196048	1.26E-06	-		0.256	2	1.4665	1.217	1.21	0.3691	0.99416
chr6	22077277	rs1207778	1.28E-06	-		0.179	8	1.4395	1.1742	1.2107	0.37315	0.98003
chr6	22082907	rs196045	1.66E-06	-		2.17	5	1.443	1.1887	1.2055	0.37480	0.99119
chr6	22130604	rs9366628	2.03E-06	-		0.024	4	1.3908	1.1677	1.1708	0.39581	0.9922
chr6	22084588	rs196040	2.09E-06	-		0.179	1	1.4404	1.1766	1.2077	0.36241	0.99152
chr6	22100912	rs9393227	2.18E-06	-		0.013	8	1.4085	1.1782	1.1902	0.43511	1
chr6	22093021	rs1207776	2.35E-06	-		0.011	7	1.4412	1.1867	1.2049	0.36207	0.99277
chr6	22072425	rs196025	2.50E-06	60		0	15	1.4305	1.202	1.1979	0.3783	0.97481
chr6	22102817	rs75396415	2.79E-06	-		0.109	2	2.8417	1.424	1.4479	0.042224	0.89219
chr6	22068352	rs169051	3.10E-06	-		0.007	7	1.47	1.241	1.2004	0.3531	0.98128
chr6	22117590	rs77573196	3.13E-06	-		0	29	3.1504	1.4375	1.4678	0.051252	0.98368
chr6	22111506	rs10484307	3.20E-06	-		0.952	7	3.4482	1.4293	1.4656	0.051026	0.98266
chr6	22112850	rs9393230	3.20E-06	76		0.004	12	3.4472	1.4281	1.4654	0.050892	0.98183
chr6	22113120	rs12211335	3.28E-06	56		0	1	3.4468	1.4291	1.4653	0.050909	0.99152
chr6	22102989	rs140621070	3.37E-06	-		0.009	2	2.8486	1.4201	1.4421	0.042935	0.89169
chr6	22111148	rs3736675	3.38E-06	-		0.004	5	3.4485	1.4278	1.4641	0.051078	0.9922
chr6	22141030	rs61096259	3.47E-06	34		0.019	6	1.3835	1.1961	1.1808	0.4255	0.95691
chr6	22119367	rs9356796	3.57E-06	-		0.003	6	3.2482	1.4208	1.4605	0.050653	0.98122
chr6	22136679	rs6921113	3.79E-06	35		0.01	7	1.4071	1.1933	1.1887	0.36993	0.98408
chr6	22126314	rs141187370	3.83E-06	-		9.616	13	-1	7.6528	7.6697	0.00022735	0.46009
chr6	22120528	rs12211475	3.96E-06	-		0.154	4	3.1309	1.4203	1.4579	0.050383	0.97766
chr6	22116570	rs9368401	4.01E-06	-		0.187	4	3.4127	1.4225	1.4585	0.050855	0.98787
chr6	22110383	rs80030489	4.03E-06	105		0	8	3.3812	1.4188	1.456	0.051051	0.99482
chr6	22090047	rs196050	4.35E-06	-		0.01	2	1.4166	1.1578	1.1997	0.35036	0.99029
chr6	22133030	rs6939950	4.45E-06	115		0.012	4	1.4185	1.1754	1.1881	0.37075	0.99117
chr6	22060582	rs885769	4.48E-06	29		0.003	20	1.4626	1.2315	1.2005	0.35904	1
chr6	22128080	rs115873209	4.79E-06	-		1.147	1	2.9472	1.4334	1.4588	0.049359	0.97459
chr6	22121654	rs12199917	4.88E-06	-		0.742	2	2.7161	1.4281	1.4408	0.050748	0.96093
chr6	22122161	rs57489115	5.03E-06	-		0.009	20	2.6041	1.4309	1.4505	0.05041	0.97241
chr6	22122090	rs59669568	5.33E-06	-		1.013	16	2.6208	1.4124	1.4389	0.053094	0.9535
chr6	22102829	rs201346080	5.67E-06	-		0	1	2.0147	1.4441	1.4304	0.04666	0.94180
chr6	22102830	rs145479024	5.67E-06	-		0	1	2.0147	1.4441	1.4304	0.04668	0.94167
chr6	22122476	rs7744414	5.92E-06	-		0	17	1.4012	1.1584	1.1788	0.37647	0.97505
chr6	22092146	rs196044	7.23E-06	-		0.019	5	-1	4.2964	4.2534	0.00045343	0.44698
chr6	22061639	rs393091	7.25E-06	38		0.004	13	1.4458	1.219	1.1943	0.34343	0.99112
chr6	22093362	rs183184	7.38E-06	-		1.481	1	-1	4.4081	4.405	0.00044407	0.44554
chr6	22068748	rs2328665	8.05E-06	-		0.116	8	1.4377	1.2118	1.1934	0.34238	0.98182
chr6	22101033	rs10456251	8.16E-06	-		0.758	1	1.7412	1.4365	1.422	0.047039	0.94943
chr6	22069970	rs196058	9.26E-06	89		9.945	9	1.4318	1.1969	1.1946	0.36354	0.9842
chr6	22089549	rs196042	9.30E-06	-		0.115	0	-1	3.9634	3.9606	0.00047049	0.43339
chr6	22154000	rs9460708	1.02E-05	54		0	9	1.411	1.2411	1.23	0.087998	0.50825
chr6	22109567	rs12207718	1.12E-05	-		0.007	9	3.4582	1.3837	1.4225	0.052118	1
chr6	22116151	rs9283875	1.20E-05	-		0.061	11	1.3739	1.1801	1.1781	0.3706	0.98837
chr6	22085052	rs12208018	1.22E-05	-		0.052	2	1.5877	1.4226	1.4037	0.050187	0.96582
chr6	22099418	rs111770457	1.22E-05	33		9.78	8	1.6224	1.4325	1.4132	0.049936	0.9734
chr6	22090672	rs142326425	1.35E-05	-		2.748	1	1.5856	1.4284	1.4088	0.04908	0.98364
chr6	22088080	rs12187945	1.37E-05	57		0.647	3	1.5858	1.428	1.4085	0.049079	0.98358
chr6	22090799	rs692044	1.47E-05	-		0.008	10	1.3659	1.1802	1.1702	0.3797	0.98908
chr6	22059589	rs77470901	1.51E-05	90		9.999	18	1.8843	1.405	1.4002	0.049953	0.96309
chr6	22085052	rs12209549	1.52E-05	-		0.103	6	1.5847	1.4262	1.4065	0.049437	0.98126
chr6	22084446	rs12195463	1.52E-05	-		0.537	1	1.5886	1.4257	1.4062	0.049482	0.98104
chr6	22124811	rs10946309	1.90E-05	-		0.072	2	2.5788	1.4033	1.4216	0.051361	0.96408
chr6	22070577	rs111720118	2.20E-05	-		0.055	4	1.6578	1.4016	1.3879	0.050001	0.96775
chr6	22060670	rs9647613	2.26E-05	-		0	5	1.6912	1.4021	1.3897	0.050609	0.96622
chr6	22061040	rs411535	2.36E-05	-		0.017	23	1.3627	1.1601	1.169	0.37468	0.99153
chr6	22153605	rs77546887	2.46E-05	-		0.007	3	1.4578	1.2319	1.2264	0.092016	0.52822
chr6	22153416	rs11751377	2.67E-05	54		0.105	1	1.4712	1.2315	1.2289	0.090709	0.53891
chr6	22128013	rs9393232	3.04E-05	-		1.488	1	1.3614	1.1505	1.1639	0.3843	0.99005
chr6	22152085	rs6916371	3.27E-05	-		0.18	1	0.50648	0.71234	0.71392	0.02902	0.58499
chr6	22140991	rs200903500	4.17E-05	25		0.108	6	3.8555	1.3757	1.418	0.039408	0.8552
chr6	22139459	rs12200477	4.72E-05	-		0.041	2	3.8803	1.3741	1.4164	0.039511	0.8811
chr6	22121147	rs181099909	4.81E-05	-		0.504	1	1.4218	1.1253	1.1479	0.17618	0.62381
chr6	22057566	rs55775505	5.32E-05	-		0.024	9	0.69913	0.85217	0.84471	0.3008	0.97931
chr6	22146341	rs12175814	5.35E-05	240		0.062	135	3.498	1.3643	1.4079	0.038879	0.82798
chr6	22054823	rs2294214	5.71E-05	78		7.003	28	0.70006	0.8543	0.84534	0.30898	0.9784
chr6	22088909	rs189077	6.60E-05	-		0.06	1	-1	3.3746	3.3724	0.00050507	0.4161
chr6	22115130	rs11750745	8.42E-05	-		9.991	5	1.3542	1.1738	1.1609	0.35306	1
chr6	22153981	rs9393240	0.00010102	33		0.004	3	1.323	1.2108	1.1976	0.096811	0.55244
chr6	22059967	rs1207702	0.00010158	-		0	16	1.3012	1.1485	1.1301	0.38582	0.98679
chr6	22060871	rs11387783	0.00010213	-		0.001	11	1.301	1.1478	1.1301	0.38556	0.98841
chr6	22130026	rs9358490	0.00012953	-		0.001	11	1.6191	1.4188	1.4041	0.047474	0.89316
chr6	22109267	rs76544301	0.00013008	-		0.167	3	1.203	1.4856	1.4741	0.014334	0.47748
chr6	22103779	rs145309606	0.00020249	-		0.018	6	2.6763	1.9584	1.9351	0.0025154	0.67672
chr6	22009366	rs116682875	0.00023238	-		0.152	15	82.876	1.7073	1.7264	0.0049045	0.56788
chr6	22010060	rs116022099	0.00023238	-		0.008	8	82.876	1.7073	1.7264	0.0049045	0.56788
chr6	22109247	rs79878130	0.00032877	-		0.545	3	0.82131	1.4812	1.4506	0.014151	0.48533
chr6	22100579	rs10946900	0.00033393	26		0.021	7	1.3056	1.1318	1.1427	0.32794	0.88773
chr6	22054517	rs6932932	0.00053924	-		1.567	8	0.70796	0.90207	0.89118	0.39205	0.70347
chr6	22105765	rs149441004	0.00061055	-		0.003	12	26.889	1.5002	1.6021	0.0075433	0.9948
chr6	22141297	rs192179262	0.00081797	41		0.013	7	-1	1.4605	1.651	0.003192	0.37114
chr6	22130304	rs55813258	0.000944	-		0.009	2	57.527	1.4938	1.6462	0.003201	0.51821

Supplementary Table 1