The mechanism and function of pervasive noncoding transcription in the mammalian genome

by
Xuebing Wu

B.S. Control Science and Engineering, Tsinghua University, Beijing (2007)
M.S. Bioinformatics, Tsinghua University, Beijing (2009)

Submitted to the Program in Computational and Systems Biology
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
[September 2014]
August 2014

Signature redacted

Signature of Author: _____
Computational and Systems Biology Ph.D. program
August 15, 2014

Signature redacted

Certified by: _____
Phillip A. Sharp
Institute Professor of Biology
Thesis supervisor

Signature redacted

Certified by: _____
Christopher B. Burge
Professor of Biology and Biological Engineering
Thesis supervisor

Signature redacted

Accepted by: _____
Christopher B. Burge
Professor of Biology and Biological Engineering
Director of Computational and Systems Biology Ph.D. program

The mechanism and function of pervasive noncoding transcription in the mammalian genome

by
Xuebing Wu

Abstract

The vast majority of the mammalian genome does not encode proteins. Only 2% of the genome is exonic, yet recent deep survey of human transcripitome suggested that 75% of the genome is transcribed, including half of the intergenic regions. Such pervasive transcription typically leads to short-lived, low-copy number noncoding RNAs (ncRNAs). We are starting to understand the biogenesis and mechanisms regulating the noncoding transcription. However, it is still unclear what's the functional impact of pervasive transcription and the ncRNAs at the level of the genome, the cell, and the organism.

A large fraction of ncRNAs in cells is generated by divergent transcription that occurs at the majority of mammalian gene promoters. RNA polymerases transcribe divergently on opposite strands, producing precursor mRNAs (pre-mRNAs) on one side and promoter upstream antisense RNAs (uaRNAs) on the other side. Like typical products of pervasive transcription, uaRNAs are relatively short and unstable as compared to pre-mRNAs, suggesting there are mechanisms suppressing uaRNA transcription and enforcing promoter directionality.

We describe the U1-PAS axis, a mechanism that enhances gene transcription but suppresses noncoding transcription. Two RNA processing signals, the U1 signal, or 5' splice site sequences recognized by U1 snRNP during splicing, and polyadenylation signal (PAS), differentially mark the two sides of gene transcription start site (TSS), ensuring the generation of full-length mRNA but inducing early termination of uaRNAs. The U1-PAS axis also suppresses pervasive transcription on the antisense strand of genes, as well as intergenic transcription.

Transcription is a mutagenic process that could accelerate evolution. We uncover a link between pervasive transcription and genome evolution. Specifically, transcription-induced mutational bias in germ cells could strengthen the U1-PAS axis, which in turn enhances transcription, thus forming a positive feedback loop, which eventually drives new gene origination, and facilitates genome rearrangements.

Tools to directly interfere with transcription with specificity are necessary to understand the function of noncoding transcription, especially when the RNA product is rapidly degraded or nonfunctional. The newly emerged CRISPR-Cas9 system provides the opportunity to target any desired locus. We comprehensively characterize the binding specificity of Cas9 in the mouse genome. We find that Cas9 specificity varies dramatically but in a predictable manner, depending on the seed sequence and chromatin accessibility. Our results will facilitate Cas9 target design and enable genome manipulation with high precision.

Thesis supervisors:
Phillip A. Sharp, Institute Professor of Biology
Christopher B. Burge, Professor of Biology and Biological Engineering

Acknowledgements

First and foremost, I want to thank my advisor Phil Sharp. I have been extremely lucky to become a Sharpie. I had absolutely no pipetting experience before I rotated in the lab, and Phil was brave enough to give me this opportunity. As a mentor Phil is remarkably knowledgeable, supportive, professional, and unexpectedly accessible. Phil always encourages me to learn new skills in experimental biology, which I found to be extremely enjoyable and helpful for my research.

I'm also very grateful to Chris Burge, my co-advisor and director of the CSB program. Beyond the invaluable insights on science and critiques of presentations in lab meetings, Chris also helped me to settle down when I arrived at US. I also want to thank him for welcoming me to his lab and giving me the opportunity to learn from a group of very talented people.

I want to thank Dave Bartel and Feng Zhang for sharing their invaluable insights and also the strong support on my scientific career. I would also like to thank Rick Young, Laurie Boyer, and Tyler Jacks who have served on my thesis committee, for their time and for being so encouraging and supportive.

My thanks also go to the lab. Many thanks to Arvind who taught me how to pipette and everything I need to start my own experiments. To Jeremy, Deyin, Paul, Jesse, and Tim, who are always there to help. In particular I want to thank my collaborator Albert, from whom I have learned so much. To Andrea, for the hard and exciting work we did together. To my baymates, Sidi and Margaret, for the friendship and for organizing the birthday parties. To Margarita for great administrative support and for the help in scheduling numerous meetings. To Mary for keeping the lab running smoothly and for forgiving the many mistakes I have made as I learn to do experiments. To all other lab members for making the lab such a great place to work.

To the CSB 2009 classmates, Zi, Vikram, Chris, Anna and Adrian for celebrating my first birthday at US, and for walking through the grad school journey together. To Bonnielee Whang and Jacquie Carota for the great administrative support from the CSB program.

I'm also indebted to my mentors and colleagues in Tsinghua University, especially Rui Jiang, Michael Zhang, Shao Li, and Xuegong Zhang for the guidance on my first research project. My thanks also go to Yanda Li for introducing me to the field of genomics.

Finally, I'm grateful to my parents for both nature and nurture, and my two sisters for their support to me and to the family over the years. Most importantly, my deepest gratitude goes to my wife, Qifang, for all her love and support, for her hard work in raising two kids. Without her I can achieve nothing. She made me who I am today. To my two sons, Charles and Tim, for making everyday fun.

Table of Contents

Chapter 1: Introduction

In this chapter, I describe known and speculated mechanisms and functions for pervasive noncoding transcription, with a focus on divergent transcription.

Overview

The vast majority of the genome is noncoding. Of the 3 billion DNA letters that make up the human genome, less than 2% encode proteins. Intriguingly, 75% of the human genome is actively transcribed (Djebali et al., 2012), producing large number of noncoding RNAs with unknown functions. Two outstanding questions arise: how does the cell deal with such pervasive transcription and what's the potential function of the transcription itself or the RNA product.

To address these questions, I focus on a major class of noncoding transcription events, divergent transcription from active promoters (Seila et al., 2008). Divergent transcription generates promoter upstream antisense transcripts that are typically short and unstable, characteristics of pervasive noncoding transcription products.

I begin the thesis with an overview of pervasive transcription and divergent transcription, as well as known and potential mechanisms that could suppress noncoding transcription and possible functions of noncoding transcription. I then describe a novel mechanism, the U1-PAS axis, for suppressing divergent antisense transcription in the promoter region, which also functions as a general mechanism for limiting pervasive noncoding transcription throughout the genome. Further consideration of this process from an evolutionary perspective suggests that the interplay between transcription and the U1-PAS axis could shape genome evolution by driving new gene origination.

In an effort to introduce precise perturbation to probe potential functions of noncoding transcription, I characterize the target specificity of the RNA-guided DNA targeting CRISPR-Cas9 system, which will facilitate the design of highly specific guide RNAs in various applications, including targeted transcription interference.

Finally, I discuss a hypothesis regarding the regulation of the U1-PAS during mouse embryonic development.

Pervasive transcription

Pervasive transcription refers to the widespread transcription activity in the genome. The discovery that the majority of the human genome is transcribed was made after the sequencing of the human genome (Lander et al., 2001), the development of high-throughput technologies such as microarray (Schena et al., 1995) and next-generation sequencing (NGS) (Shendure and Ji, 2008), as well as large-scale integrative studies (Birney et al., 2007; Carninci et al., 2005; Djebali et al., 2012; He et al., 2008; Katayama et al., 2005). The collaborative analysis of the 1% of the human genome in the pilot ENCODE project reported that 74% of the studied region show evidence of transcription with support from at least two technologies (Birney et al., 2007). The same conclusion holds when the entire genome is investigated by RNA sequencing (RNA-seq) (Djebali et al., 2012). Given that less than 2% of the genome are exonic, most of the pervasive transcription activity gives rise to noncoding RNA that has no representation in the proteome, and has been considered to be the "dark matter" of the genome (Johnson et al., 2005).

Regarding the biogenesis of pervasive transcription, most noncoding transcription activity seems to associate with genes (van Bakel et al., 2010), such as promoters, terminators, and antisense strand within genes. In addition, those truly intergenic transcripts fall within open chromatin, which likely represent transcription at distal regulatory elements of genes such as enhancers. The connection to genes suggests potential regulatory role for these noncoding transcription activities. Alternatively, the association with open chromatin and the low abundance of the noncoding RNAs raise the possibility that these noncoding transcription events are simply "transcriptional noise" or the byproducts of spillover of the transcription machinery from genes into nearby accessible chromatin regions.

Five major classes of noncoding RNAs in the mammalian genome have been identified: 1) promoter upstream antisense transcripts from promoter divergent transcription (Core et al., 2008; Seila et al., 2008), 2) enhancer RNAs (eRNAs) from transcribed enhancers, typically also divergent (Kim et al., 2010), 3) large intergenic noncoding RNAs (lincRNAs) defined by H3K4me3 and H3K36me3 chromatin signatures (Guttman et al., 2009), 4) antisense transcripts overlapping with genes (He et al., 2008; Katayama et al., 2005), and 5) primary transcripts of small RNAs, such as microRNA and piRNA. These definitions are not mutually exclusive, as there are instances of overlap between eRNAs and lincRNAs, as well as between lincRNAs and small RNA precursors.

Similarly widespread transcription activity has been observed in other species, including the single cell eukaryote yeast. Similar to the situation in

mammals, most noncoding transcription originates from accessible chromatin, and in particular nucleosome free regions (NFRs) at the two ends of genes (Neil et al., 2009; Xu et al., 2009). In yeast the noncoding RNAs are classified based on the degradation pathway involved, including: 1) cryptic unstable transcripts (CUTs) that are mostly degraded by Rrp6, the 3'-5' nuclear exosome (Davis and Ares, 2006; Houalla et al., 2006; Wyers et al., 2005), 2) stable unannotated transcripts (SUTs) that are less sensitive to Rrp6 (Neil et al., 2009; Xu et al., 2009), 3) Xrn1-sensitive unstable transcripts (XUTs) that are stabilized when the cytoplasmic 5'-3' exoribonuclease Xrn1 is depleted (van Dijk et al., 2011), and 4) Nrd1-unterminated transcripts (NUTs) whose transcription termination depends on Nrd1 (Schulz et al., 2013). Again these categories are not mutually exclusive, especially between NUTs and the other three classes.

Divergent transcription

Divergent transcription at promoters

Divergent transcription generates the majority of long noncoding RNAs in mouse and human embryonic stem cells (ESCs) (Sigova et al., 2013). The Sharp lab (Seila et al., 2008) and the Lis lab (Core et al., 2008) independently reported genome-wide evidence for widespread divergent transcription activities at protein-coding gene promoters in human and mouse cell lines, which are further supported by another report focusing on 1% of the human genome (Preker et al., 2008). Specifically, in addition to the transcription of genes that makes messenger RNA

precursors, transcription also occurs upstream of protein-coding genes in the antisense direction from the gene promoters generating noncoding RNAs.

Two different assays provide complementary support for divergent transcription activity: nuclear run-on (Core et al., 2008) and small RNA cloning (Seila et al., 2008). Both are augmented by deep sequencing to generate unbiased genome-wide evidence of transcription activity. Core et al develop the global run-on-sequencing (GRO-seq) assay, which relies on *in vitro* extension of engaged polymerase on the 3' end of nascent RNA. Cloning and sequencing of the 3' end of nascent transcripts provides a good surrogate of polymerase occupancy *in vivo*. GRO-seq data in primary human lung fibroblast (IMR90) reveals two divergent peaks of polymerase flanking gene transcription start sites at 77% of active genes, suggesting extensive antisense transcription upstream of gene TSS. The sense peak is around 50 bps downstream of the gene TSS, whereas the antisense peak is around 250 bps upstream of the gene TSS. Seila et al cloned and sequenced small RNAs in mouse ESCs, and similarly observed pileup of small RNAs around 250 bps upstream gene TSS from about 67% of active promoters, especially CpG rich promoters. In addition, two distinct ChIP-seq peaks are detected flanking gene TSS for both RNA polymerase II and H3K4me3, indicating divergent transcription initiation.

Both studies provide unbiased evidence showing that the majority (67%-77%) of mammalian gene promoters are divergently transcribed. Another report (Preker et al., 2008) published in the same issue of Science shows that within the ENCODE pilot project region (Birney et al., 2007), which is about 1% of the human genome, RNA could be observed about 0.5 to 2.5kb upstream of gene TSS upon

depletion of Rrp40, a subunit of nuclear exosome, a 3'-5' RNA-degradation machine (Houseley et al., 2006; Schmid and Jensen, 2008). This suggests the existence of transcription activities in noncoding regions, and in addition demonstrates that the products of such transcription activities are substrates of a nuclear exosome, partially explaining why divergent transcription had not been previously reported.

The antisense transcripts generated by divergent transcription are short and unstable, compared to the long and stable mRNA precursors generated on the sense side. The GRO-seq signal on the antisense side quickly diminishes to background level beyond 1-2kb of the TSS, as compared to more than 10kb of signal on the mRNA side (Core et al., 2008). Also, the antisense region lacks dimethylation at lysine 79 on histone H3 (H3K79me2), a histone modification associated with transcription elongation in genes (Nguyen and Zhang, 2011). Moreover, the exosome-sensitive RNAs enrich in the 0.5 to 2.5 kb region upstream the TSS, suggests a median length of 1kb for these RNAs.

Subsequently divergent transcription has been observed in yeast (Neil et al., 2009; Xu et al., 2009) and worm (Kruesi et al., 2013), but much less frequent in fly (Core et al., 2012).

The 3' ends of upstream antisense transcripts

The upstream antisense transcript contains a 5' cap, but the nature of the 3' end remains elusive. Two groups perform detailed characterization of a few divergent transcripts (Flynn et al., 2011; Preker et al., 2011). Flynn et al examines the promoters of four genes that generate divergent transcripts in mouse ESCs (Flynn et al., 2011), whereas Preker et al analyzes a few genes in human HeLa cells

and HEK293 cells (Preker et al., 2011). In both reports, the upstream antisense RNAs (uaRNAs) or promoter upstream transcripts (PROMPTs) are found to contain a 5' cap, like mRNA. However, in contrast to PROMPTs, which are claimed to contain a 3' poly-A tail or at least oligo-A tail (a few adenosine at the 3' end), the uaRNAs examined by Flynn rarely contain a non-templated stretch of Adenosine, and have very heterogeneous 3' ends, i.e. terminating at multiple locations. One caveat for the Flynn et al study is that the lack of poly-A tail, as well as the heterogeneous 3' ends of the RNAs cloned, could represent degradation intermediates of polyadenylated RNAs with defined 3' ends. However, it is also worth noting that the evidence for Preker to claim that PROMPTs contain a poly-A tail is that the cDNA was prepared from reverse transcription with oligo-dT as primers without affinity purification of poly-A tail containing RNAs (Preker et al., 2008). Oligo-dT priming is known to copy RNA containing internal A-stretches thus truncated non-polyadenylated RNAs could be reverse transcribed into cDNA (Nam et al., 2002). Moreover, non-polyadenylated RNAs could self-prime reverse transcriptase generating cDNA from total RNA in the absence of exogenous primers (Frech and Peterhans, 1994; Moison et al., 2011).

Many uaRNAs have poly-A tails associated with polyadenylation signals. We sequenced the 3' ends of poly-A selected RNAs in mouse ESCs, and found that many uaRNAs contain poly-A tails associated with canonical cleavage and polyadenylation sequence motifs (Almada et al., 2013). Similar results are observed in human HeLa cells (Ntini et al., 2013). In addition, polyadenylated upstream antisense RNAs in human HEK293 cells (Martin et al., 2012) showed similar patterns of occupancy by core factors involved in the canonical 3' end cleavage and polyadenylation

machinery (Almada et al., 2013; Ntini et al., 2013). Together these data demonstrate that at least some uaRNAs have defined 3' ends likely generated by the same pathway as mRNA 3' end maturation.

However, it is unclear globally, what fraction of uaRNAs contain a polyA tail, since we were selecting poly-A RNAs. It is still possible that a fraction of uaRNAs contain no poly-A tails. The presence of a poly-A tail in uaRNA is intriguing, given that two of the most important functions of the poly-A tail is to enhance the stability of the RNA and to facilitate its translation, yet uaRNA is neither stable nor translated. It is likely that a significant fraction of uaRNAs are actually generated without a poly-A tail, given that there are many polyadenylation-independent mechanisms leading to the termination of uaRNAs (See section "Mechanisms for suppressing noncoding transcription").

The stability of upstream antisense transcripts

Some upstream antisense transcripts are rapidly degraded. Flynn et al estimated that uaRNAs from the four genes examined have an average half-life of 18 minutes, and only accumulate to 1-4 copies per cell, approximately one copy per DNA template (Flynn et al., 2011). RNAi knockdown of various subunits of the nuclear exosome leads to several fold stabilization of the uaRNAs (Flynn et al., 2011; Ntini et al., 2013; Preker et al., 2008), suggesting that nuclear exosome is responsible for removal of some of the uaRNAs/PROMPTs. Whether other degradation pathways are involved is unclear.

It's unclear whether most polyadenylated uaRNAs are also short-lived and degraded by exosome. None of the reports of exosome sensitive uaRNAs/PROMPTs

include a poly-A selection step to enrich polyadenylated RNAs. Preker et al measured RNA level by microarray hybridization of cDNA generated by oligo-dT primed reverse transcription (Preker et al., 2008). As noted above, internal priming of oligo-dT primers, as well as self-priming, could reverse transcribe non-polyadenylated RNAs, thus it is unclear whether the RNA species sensitive to exosome knockdown were polyadenylated or not. The recent RNA-seq data from the same lab (Ntini et al., 2013) is depleted of ribosomal RNA but again not poly-A selected. The exosome-sensitive uaRNAs from four divergent promoters examined by Flynn et al lack poly-A tails (Flynn et al., 2011). Only in an artificial case, where a PROMPT is cloned into a plasmid between a CMV promoter and a SV40 poly-A site, is poly-A RNA shown to be sensitive to exosome knockdown (Ntini et al., 2013). Given the lack of comprehensive data on endogenous genes, and the role of poly-A in enhancing RNA stability, polyadenylated uaRNA could be stable and not degraded by exosome. In fact, we have sequenced poly-A selected RNAs from mouse ESCs after depleting the exosome subunit Exosc5, and observe no increase of polyadenylated uaRNAs in either aggregated signal over all promoters, or at the four promoters where non-polyadenylated uaRNAs had been shown to increase by 2-4 fold (unpublished data).

Divergent transcription at enhancers

Two years after the discovery of divergent transcription at protein-coding gene promoters, in 2010, genome-wide data suggested that thousands of mouse neuron activity regulated enhancers, marked by H3K4me1 and bound by Pol II, are transcribed divergently, generating noncoding enhancer RNAs (or eRNAs) on both

sides of the enhancer (Kim et al., 2010). The lack of directionality seems to be even stronger at enhancers than at promoters. These eRNAs are typically short, and can only be detected in total RNA sequencing but not polyA selected RNA, suggesting a lack of poly-A tail, although later studies do find polyadenylated eRNAs (Djebali et al., 2012). The presence of eRNAs is associated with activation of nearby genes, suggesting enhancer transcription is a mark of active enhancers (Hah et al., 2013; Kim et al., 2010). In fact, in 2003, the enhancer for the major histocompatibility complex (MHC) class II, also called locus control region (LCR), was also found to be transcribed divergently when the gene is activated (Masternak et al., 2003). However, it is still unknown whether enhancer transcription is the cause or consequence of gene activation.

Subsequently, the presence of divergent short transcripts has been used to identify active enhancers (Melgar et al., 2011), which are also marked by H3K27ac (Creyghton et al., 2010). Recently, the FANTOM5 consortium performed cap analysis of gene expression (CAGE) in hundreds of human tissues and cell types, and identified 43,000 enhancers showing divergent transcription activities (Andersson et al., 2014). Unlike CpG-rich divergent promoters, these enhancer regions are CpG-poor, suggesting divergent transcription is not a unique feature of CpG promoters.

Mechanisms for suppressing noncoding transcription

Despite the widespread transcription activity in the genome, transcription outside genes, including both intergenic transcription and antisense transcription, is typically unproductive yielding very low copy numbers of RNA. Even those more

stable polyadenylated noncoding RNAs identified from deep transcriptome sequencing only accumulate to less than 10% of average mRNAs (Cabili et al., 2011; Djebali et al., 2012; Sigova et al., 2013). The low abundance of noncoding transcripts could potentially be explained by three broad types of mechanisms: lack of features enhancing transcription, the presence of features suppressing transcription, and selective degradation of the transcripts. I will discuss all three types of mechanisms and they are roughly ordered by the stage of the transcription cycle they affect, although sometimes multiple stages or unknown stages are affected. It is important to note that, although much has been learned by genome-scale analysis, the molecular details of many of the mechanisms remains unclear.

Unidirectional promoter elements

Divergent noncoding transcription is mainly associated with CpG promoters in mammals but also non-CpG promoters in species such as yeast and worm, yet in all of these species divergent noncoding transcription is largely absent from promoters containing TATA-box elements (Core et al., 2012; Kruesi et al., 2013; Seila et al., 2008), consistent with previous reports that TATA-box specifies the orientation of transcription (Xie et al., 2001). Similarly in Drosophila, most gene promoters contain other directional elements and are transcribed unidirectionally (Core et al., 2012).

It is unclear how these unidirectional promoter elements prevent noncoding transcription on the opposite orientation. One possible model is the competition for transcription machinery. Factors binding to TATA-box elements and other promoter unidirectional elements may have higher affinity for other components of the

preinitiation complex, including the RNA polymerases, thus preventing the opposite orientation from forming transcription complex. In addition, these directional elements may trigger or facilitate some other pathways described below, such as chromatin assembly and remodeling, deposition of specific histone modifications, or gene looping.

Chromatin remodeling and histone modifications

Densely positioned nucleosomes are thought to limit promoter accessibility and repress transcription. In yeast, noncoding RNA transcription typically initiates near the 5' end and 3' end of genes that are typically nucleosome free regions (NFR) (Neil et al., 2009). In budding yeast, the chromatin remodeling complex Isw2 increases nucleosome occupancy in intergenic regions and suppresses cryptic noncoding antisense transcription at the 3' end of three genes (Whitehouse et al., 2007) and hundreds of cryptic transcripts from nucleosome free regions (Yadon et al., 2010). Similarly, in fission yeast, the loss of the chromatin remodelers Hrp1 and Hrp3 leads to increased noncoding transcription in both centromeric regions and gene regions, without affecting overall gene expression (Hennig et al., 2012; Pointner et al., 2012; Shim et al., 2012). In addition to chromatin remodelers, histone chaperone proteins have also been implicated in suppressing noncoding transcription. For example, depletion of Spt6 in both budding yeast (Cheung et al., 2008) and fission yeast (DeGennaro et al., 2013) results in elevated antisense transcription. Another histone chaperone Hira in fission yeast suppresses transcription from some coding genes, LTRs, and antisense regions (Anderson et al., 2009). Many of these factors are broadly involved in chromatin packing and their

depletion may simply increase the accessibility of the DNA in general and not necessarily in promoter regions.

More recently, a large-scale mutant screen in budding yeast identified the chromatin assembly complex CAF-I as a negative regulator of divergent noncoding transcription (Marquardt et al., 2014). Mutation in subunits of the CAF-I complex increases the level of over a thousand nascent divergent noncoding transcripts, although overall the magnitude of increase is modest. On the other hand, H3K56 acetylation and Swi/Snf complex work together to enhance divergent noncoding transcription by promoting rapid nucleosome turnover.

In addition to H3K56 acetylation mentioned above, derepression of H4 acetylation by depleting Rpd3 small (Rpd3S) H4 deacetylation complex in budding yeast leads to a four-fold increase in nascent antisense transcription, although mostly at the 3' end of genes (Churchman and Weissman, 2011). In mammals, upstream antisense regions of divergent promoters lack elongation marks such as H3K79me2, although it is unclear whether this modification is the cause or consequence of productive transcription elongation (Seila et al., 2008), and if it is causal, what is the upstream event that specifies the mark on gene regions but not intergenic regions during transcription.

The fate of RNA degradation could also be coded as a chromatin state, which could be passed through cell cycles. In fission yeast, the histone variant H2A.Z, a variant preferentially deposited at the 5' end of genes, works together with the RNAi factor Ago1 or heterochromatin factor Clr4 (homolog of mammalian methyltransferase SUV39H) to target nuclear exosome to convergent genes to

remove read-through antisense transcripts (Zofall et al., 2009). The histone deacetylase (HDAC) Clr6 also targets exosome to degrade antisense transcripts at euchromatic loci as well as centromeric regions (Nicolas et al., 2007). Whether similar mechanisms exist in other species is unclear.

Gene loops

For a number of genes in budding yeast such as FMP27 and SEN1, the two ends of the gene, i.e. the promoter and the terminator, juxtapose to form a chromatin loop (Ansari and Hampsey, 2005; O'Sullivan et al., 2004). Gene loops are mediated by the physical interaction between transcription initiation factors such as TFIIB and 3' end processing factor, such as Ssu72, and requires an active promoter and a functional poly-A site (Ansari and Hampsey, 2005; O'Sullivan et al., 2004; Singh and Hampsey, 2007). It has been proposed that a gene loop might facilitate recycling of the transcription machinery on the same gene. Tan-Wong et al recently reports that when Ssu72 is deleted, in addition to disrupting the gene loop, more promoter upstream antisense transcripts are produced (Tan-Wong et al., 2012). Inactivating other 3' end processing factors such as Pta1, Rna14, and Rna15, or replacing the poly-A site with non-poly-A termination signal, or inactivating TFIIB also leads to elevated promoter noncoding RNA, suggesting a role for gene looping in suppressing divergent antisense transcription. Elevated Pol II ChIP signals immediately upstream of gene promoters suggests Ssu72 likely suppresses transcription initiation.

It is unclear how frequently the gene looping mechanism is used in yeast. Microarray profiling identified 605 promoters (~10% of yeast genes) producing more antisense transcripts upon Ssu72 mutation (Tan-Wong et al., 2012). It is also unclear whether these promoters are also controlled by gene loops, or Ssu72 itself has a role in restricting noncoding transcription.

It is not known whether gene looping is used in higher species such as mammals. The authors show that in human 293 cells, an artificial construct of the beta-globin gene fused to a SV40 late poly-A site display gene loop conformation, and when the poly-A site is mutated, the gene loop is disrupted and the level of promoter divergent noncoding RNA increases (Tan-Wong et al., 2012). However, it remains to be seen whether the endogenous beta-globin locus works this way. In addition, several comprehensive chromatin interaction mapping efforts in human or mouse have found far more enhancer loops than gene loops (Kagey et al., 2010; Li et al., 2012; Sanyal et al., 2012).

Pausing and release

Shortly after initiation, RNA polymerase II pauses 20 to 100 bps downstream transcription start site (TSS) largely caused by negative elongation factor NELF and DSIF (Adelman and Lis, 2012). Release of Pol II depends on the positive elongation factor P-TEFb, which phosphorylates pausing factors and the C-Terminal domain (CTD) of Pol II at serine 2 position, and is necessary and sufficient for paused Pol II to enter productive elongation.

Rahl et al demonstrate that pausing factors DSIF and NELF co-occupy with Pol II genome-wide, including divergent antisense regions, suggesting pausing does

occur in those regions (Rahl et al., 2010). One potential model to limit noncoding transcription is the inefficient recruitment of P-TEFb to escape pausing. Two earlier studies reject this model based on a few genes examined (Flynn et al., 2011; Preker et al., 2011). Flynn et al show that in mouse ESCs, the four uaRNAs tested respond similarly to mRNA to knockdown of negative elongation factor NELF and DSIF, confirming Pol II indeed also pauses when transcribing uaRNA. Inhibiting P-TEFb by flavopiridol for an hour leads to 4 to 5-fold reduction in uaRNA, compared to 8 to 12-fold reduction of nascent mRNA transcript, suggesting uaRNA transcription depends on P-TEFb, although to a smaller extent. Within 30 minutes of withdraw of flavopiridol, both uaRNA and nascent mRNAs go back to normal levels, although intermediate time points are not taken and so it is hard to compare the kinetics of recovery. These data suggest that both uaRNA and mRNA are regulated by P-TEFb, likely through release of Pol II pausing.

The transition of the phosphorylation status of Pol II CTD is correlated with release of pausing. Immediately after transcription initiation, serine 5 of Pol II CTD is phosphorylated, and then serine 2 is phosphorylated with entrance to productive elongation. Preker et al showed that in the two human genes examined and also from genome-wide ChIP data in human CD4+ T-cells, the ratio of serine 2 phosphorylated Pol II ChIP signal over total Pol II is comparable within 2kb of gene TSS, suggesting serine 2 phosphorylation, thus pausing release, is not limiting divergent transcription (Preker et al., 2011). However, another genome-wide Pol II ChIP data set shows the opposite (Rahl et al., 2010). Rahl et al report that although total Pol II or serine 5 phosphorylated Pol II occupy the promoter-proximal region

(by ChIP-seq), serine 2 phosphorylated Pol II only accumulates in gene body, and is absent from the antisense noncoding region, implying the lack of pausing release.

This discrepancy could due to different quantification approaches (ser2/total ratio or ser2 only), or different antibodies (ser2 specificity), or cell types. GRO-seq is another way to measure Pol II occupancy independent of antibody. More recently, Jonkers et al perform GRO-seq in mouse ESCs with and without flavopiridol treatment (Jonkers et al., 2014). By 12.5 minutes, flavopiridol dramatically reduces Pol II in the gene body, whereas much weaker if any decrease is seen in the promoter divergent region, suggesting that either P-TEFb is less active in the antisense direction, or additional factors are blocking efficient elongation in the antisense direction that are independent of P-TEFb or Pol II phosphorylation, such as DNA superhelical tension (see next section). Nonetheless, the above three studies in mouse ESCs (Flynn et al., 2011; Jonkers et al., 2014; Rahl et al., 2010) argue that P-TEFb has less an impact on antisense transcription as compared to sense direction.

The next question is whether P-TEFb differentially recruited to the two sides of the TSS, and if so, how does it know the gene direction from the antisense direction. A fraction of P-TEFb is sequestered in the 7SK snRNP complex, which contains the abundant small nuclear noncoding RNA 7SK, and proteins such as HEXIM that inactivates the P-TEFb kinase domain (Peterlin and Price, 2006). The mechanism for the release of the inactivated P-TEFb from the silencing complex and its recruitment to active promoters have recently been identified (Ji et al., 2013). It turns out some SR proteins, especially SC35/SRSF2, are components of the 7SK

complex, which recruits the 7SK complex to promoter-proximal nascent RNAs containing Exonic Splicing Enhancer (ESE) elements. The binding of SC35 to nascent RNA triggers the release of P-TEFb from the 7SK complex and elongation of transcription. ESEs are very short and their degenerate binding motifs that occur frequently in the genome. However, it is generally C+G rich (SSNG, S=C or G), and highly enriched in the first exons of coding genes as compared to the upstream antisense region, which could be the molecular basis for differential recruitment and activation of P-TEFb on the two sides of gene TSS. Consistent with this, SC35, CDK9, HEXIM, and Pol II all showed much stronger ChIP signals in the sense direction as compared to the antisense side. SC35 CLIP signals are also much higher on the gene side (Ji et al., 2013). Other factors have also been implicated in the recruitment of P-TEFb, including the oncogene c-Myc (Rahl et al., 2010), the bromodomain protein BRD4 (Patel et al., 2013), and an elongation factor ELL (Byun et al., 2012). It is unclear whether these factors show any bias for sense and antisense transcription.

Interestingly, depletion of the noncoding RNA 7SK leads to increased divergent transcription at 2,000 promoters in mouse ESCs, although whether this is through P-TEFb is unclear (Castelo-Branco et al., 2013). The increase of divergent transcription upon 7SK knockdown largely disappears with treatment of P-TEFb inhibitor flavopiridol or bromo and extra terminal (BET) bromodomain inhibitor I-BET151 (BET proteins recruit P-TEFb to acetylated histones and activate of transcription (Dawson et al., 2011)). This is expected if decrease of 7SK silencing snRNP releases P-TEFb to activate divergent transcription. However, intriguingly,

the corresponding sense genes are not affected (Castelo-Branco et al., 2013), ruling out a simple model of P-TEFb release upon 7SK knockdown. Genome-wide, only 438 genes are upregulated, and 30 genes downregulated upon 7SK knockdown. In contrast, about two thousand genes show elevated read-through transcription of the poly-A site, implying a role of 7SK in enhancing 3' end cleavage and polyadenylation, and / or promoting degradation of read-through transcripts.

Overall, current data suggests that pausing indeed occurs during divergent antisense transcription and P-TEFb is required for optimal expression of the antisense transcripts, yet the dependence of uaRNAs on P-TEFb is much weaker compared to mRNA.

DNA superhelical tension

Transcription generates positive and negative supercoiling before and after the polymerase, respectively (Kouzine et al., 2004, 2008, 2013, 2014). Positive supercoiling needs to be resolved by topoisomerase, otherwise it impedes polymerase elongation. On the other hand, negative supercoiling facilitates DNA melting, transcription initiation and elongation, and could explain extensive divergent transcription in the promoter regions that accumulate negative supercoiling due to gene transcription (Seila et al., 2008, 2009).

Further, if no topoisomerase is recruited or activated at transcribing noncoding regions, transcription in these regions is likely to be limited. However, there is little data suggesting topoisomerase activity is preferentially recruited to genes but not noncoding regions. The Levens' group recently identified a positive feedback loop between Topoisomerase I (Top1) and RNA polymerase II

phosphorylated at serine 2 of the C-terminal domain (CTD) (personal communication). Phosphorylated Pol II super-activates Top1, which relaxes the positive supercoiling ahead of translocating Pol II, facilitating transcription elongation. On the other hand, Top1 is inactive near the promoters where Pol II is hypophosphorylated, thus the template DNA remains negatively supercoiled to facilitate DNA melting and transcription re-initiation.

If indeed phosphorylation of Pol II CTD is much less efficient in the antisense direction, as discussed above, the difference could be amplified by the positive feedback loop between Top1 activity and Pol II phosphorylation, and lead to drastic differences in Top1 activity and elongation rate. Therefore, although the negative supercoiling in promoter caused by Pol II on the gene side might facilitate the initiation of antisense transcription, the positive supercoiling caused by the antisense polymerase itself might not be resolved, leading to inefficient elongation.

Interestingly, inhibition of Top1 activity in HCT116 cells by a specific drug camptothecin (CPT) leads to a dramatic increase of divergent antisense transcripts, specifically at intermediately active CpG island promoters, but not inactive promoters or super-active promoters, or promoters without CpG island (Marinello et al., 2013). The enhanced divergent transcription indicates Top1 normally suppresses divergent transcription at intermediately active promoters, presumably through relaxing the negative supercoiling in the promoter region, and Top1 activity may not be strong enough to release all the tension caused by super-active gene transcription. In a few genes, inhibiting P-TEFb activity by siRNA knockdown of CDK9 (P-TEFb subunit) or treating with CDK9 inhibitor DRB, both partially reduce

the effect of CPT, indicating that both inefficient pausing release and superhelical tension contribute to the lack of productive elongation in the antisense direction.

R-loop formation

R-loops are three strand structures formed during transcription. The nascent RNA folds back to form RNA:DNA hybrid and displaces the other DNA strand (Aguilera and García-Muse, 2012). R-loop formation has a negative impact on transcription. If sufficient RNase H is present, the hybrid RNA strand will be destroyed, reducing the level of the transcript. If the RNase H concentration is low, the structure will be stable and could block subsequent rounds of transcription, and as well signal for DNA repair activities leading to chromosome instability.

The antisense direction of the divergent promoter might be prone to R-loop formation. First, the promoter region is typically negative supercoiled, and negative supercoiling facilitates R-loop formation, since negative supercoiling facilitates DNA melting and RNA hybridization to the DNA (Roy et al., 2010). Secondly, divergent transcripts typically lack splicing signals, and splicing factors could suppress R-loop formation, presumably by binding to the RNA and preventing its hybridization with DNA (Li and Manley, 2006, 2005; Paulsen et al., 2009).

Lack of splicing signals

In most eukaryotes especially mammals, most protein-coding genes contain introns and are spliced; in contrast, splicing is either missing or very inefficient in noncoding transcripts (Tilgner et al., 2012). Splicing and transcription are tightly coupled (Lenasi and Barboric; Luco and Misteli, 2011; Luco et al., 2011; Perales and

Bentley, 2009). For example, the presence of introns generally increases RNA abundance (Brinster et al., 1988; Choi et al., 1991; Nott et al., 2003; Palmiter et al., 1991), which could due to increased transcription, more efficient polyadenylation and nuclear export, or more stable transcripts. One mechanism for transcription regulation is through 5' splice site and U1 snRNP, which can enhance transcription at multiple stages, including initiation, elongation, and termination.

U1 snRNA and the first 5' splice site recruits general transcription factors. In eukaryotes the first 5' splice site is typically within 200 bps of the transcription start site, making it well positioned to regulate transcription initiation. Furger et al showed that mutating promoter-proximal 5' splice site in a retroviral gene construct reduces its nascent transcription by three fold in HeLa cells, which could be partially rescued by a complementary mutation in U1 snRNA (Furger et al., 2002).

Subsequently U1 snRNA was shown to co-purify from HeLa cell extract with TFIIH, a general transcription initiation factor implicated in Pol II promoter escape and transcription re-initiation (Kwek et al., 2002). In *in vitro* experiments, the interaction between U1 snRNP and TFIIH enhances transcription initiation by stimulating abortive initiation and by enhancing transcription re-initiation in a 5' splice site dependent manner (Kwek et al., 2002). Later it was shown that *in vivo*, 5' splice sites recruit TFIID, TFIIB, and TFIIH to promoters, and TFIID and TFIIH components are also specifically recruited to 5' splice sites (Damgaard et al., 2008). Together these studies demonstrated a role for promoter-proximal U1 snRNP and 5' splice site in directly enhancing transcription initiation.

Splicing of the first intron deposits active marks of transcription. In addition to interacting directly with initiation factors, the first 5' splice site can also modulate the epigenetic state near promoters and help to recruit initiation factors indirectly (Bieberstein et al., 2012). Marks of transcription initiation, H3K4me3 and H3K9ac, peak at the first 5' splice site, especially when the first exon is small, i.e. the 5' splice site is promoter-proximal. In one gene examined, inhibiting splicing by either spliceostatin A (SSA) or 3' splice site mutation reduces H3K4me3 in the first intron, suggesting splicing helps to deposit the mark. SSA treatment also inhibits global nascent transcription by two fold as determined by metabolic labeling. Interestingly, this study also showed that when the first exon is long, i.e. the 5' splice site is far away from TSS, antisense transcription is more likely to initiate in the first exon.

Splicing factors enhance transcription elongation. U snRNPs and splicing factors have also been implicated in enhancing transcription elongation (Fong and Zhou, 2001; Lin et al., 2008). Affinity purification of an elongation factor TAT-SF1 from 293T cell extract identifies specific components of U1 snRNP and U2 snRNP (Fong and Zhou, 2001). The purified TAT-SF1-snRNP complex stimulates transcription as well as splicing *in vitro*. The stimulatory effect on transcription is likely mediated by the interaction of TAT-SF1 with elongation factor P-TEFb and the interaction of snRNPs with the nascent transcripts. In addition, SC35, an SR family splicing factor, has also been found to enhance transcription (Lin et al., 2008). Depletion of SC35 in MEFs leads to accumulation of RNA polymerase II in the body of a subset of genes, likely due to the inefficient recruitment of P-TEFb and reduced CTD Ser2 phosphorylation (Lin et al., 2008). U1 snRNP could also enhance

transcription elongation via a very different mechanism: suppressing premature cleavage and polyadenylation of the transcript, which triggers transcription termination. This mechanism will be discussed in detail in the next section.

Splicing enhances RNA stability. It's been shown *in vitro* that functional splicing signals protect pre-mRNA from nuclear degradation (Hicks et al., 2006), likely because the spliced RNA is packaged into and thus protected by splicing factors and exon junction complexes (Singh et al., 2012). In contrast, inefficient splicing of many fission yeast pre-mRNAs triggers a polyadenylation-dependent nuclear decay pathway involving poly(A) binding protein Pab2 and nuclear exosome component Rrp6 (Lemieux et al., 2011). Noncoding RNAs are generally not spliced or spliced inefficiently (Tilgner et al., 2012) and this may render them susceptible to degradation.

Polyadenylation-coupled transcription termination and RNA degradation

Premature termination. Termination of most eukaryotic gene transcription is triggered by recognition of the cleavage and polyadenylation signal (PAS) in the nascent RNA, which is mainly an AAUAAA or similar sequences. There are on average two AAUAAA sites per kb, much higher than the 0.7 sites per kb assuming random combination of nucleotides in typical mammalian genomes with 40% G+C content. Shortly after transcription starts, those putative PAS may induce cleavage and polyadenylation and subsequent transcription termination, leading to short RNAs that may be prematurely processed and are rapidly turned over by RNA surveillance pathways in the nucleus. Intronic PASs are suppressed by U1 snRNP

binding to near by 5' splice sites (Berg et al., 2012; Kaida et al., 2010) and potentially also secondary structures or other mechanisms. Interestingly, AATAAA sites are specifically depleted on the sense strand of genes, compared to flanking intergenic regions, suggesting evolutionary selection against premature termination signals in genes (Almada et al., 2013; Glusman et al., 2006). Moreover, AATAAA sites are enriched on the antisense strand of genes, compared to the sense strand of genes, or even intergenic regions (Almada et al., 2013), suggesting selection for premature termination of antisense transcription events.

Inefficient polyadenylation. Transcripts processed with weak PAS may be preferentially targeted for degradation. Inefficient cleavage and / or polyadenylation leading to reduced mRNA stability has been reported in human and other species (Batt et al., 1994). PAS associated with upstream antisense cleavage sites are significantly weaker, i.e. less consensus in sequence, than those associated with annotated gene ends (Almada et al., 2013). Sequence composition analysis reveals that, compared to annotated 3' end of genes, upstream antisense cleavage sites are less U-rich but more A-rich upstream of the cleavage site (Almada et al., 2013). In human the U-rich upstream element is recognized by the CPSF component Fip1, which stimulates poly(A) polymerase. A yeast strain defective in Fip1 leads to poorly polyadenylated pre-mRNAs which are rapidly depleted by nuclear exosome (Saguez et al., 2008). In addition, the A-rich region around the PAS in upstream antisense cleavage sites may recruit nuclear poly(A) binding protein Pabpn1 (mammalian homolog of fission yeast Pab2), as suggested by the observation that human PABPN1 can be recruited to PAS (Jenal et al., 2012). Given that the yeast

homolog Pab2 physically interacts with exosome and can be recruited to genes co-transcriptionally (Lemay et al., 2010), it is possible that the Pabpn1-exosome complex is preferentially recruited to PAS in uaRNAs, leading to rapid degradation of uaRNAs.

Poly-A tail length. Several studies have shown that at least some uaRNA/PROMPTs are polyadenylated in a way similar to mRNA. However, it is unclear whether polyA tails are of the normal length. Abnormal length of the poly-A tail could lead to RNA decay. Previously Jensen et al showed that in budding yeast, block of nuclear export leads to mRNA hyperadenylation and accumulation at the site of transcription (Jensen et al., 2001). This is likely because export factors help to disassemble the 3' end processing complex and release the transcript (Qu et al., 2009). Splicing may play a role in releasing mRNA by recruiting export complex (Rigo and Martinson, 2009). In the cases of noncoding RNAs, which are usually poorly spliced, it's likely that by default the nuclear export complex to disassemble the 3' end processing complex is inefficiently recruited, leading to hyperadenylation and accumulation at the site of transcription. Recently a pathway has been described to link hyperadenylated nuclear RNA to exosome degradation (Bresson and Conrad, 2013), which depends on the nuclear poly-A binding protein PABPN1. PABPN1 has previously been shown to interact with exosome as well, both in human (Beaulieu et al., 2012) and yeast (Lemay et al., 2010).The association of exosome with chromatin may be part of the mechanism or the consequence. Two high-throughput assays have been developed to measure the poly-A tail length

genome-wide, which opens the door to investigate if indeed uaRNAs are hyperadenylated (Chang et al., 2014; Subtelny et al., 2014).

Nrd1-dependent termination and degradation

In yeast, cryptic unstable transcripts (CUTs) contain a high density of Nrd1 and Nab3 binding sites, which recruits Sen1 helicase to terminate transcription, and then the Trf4p/Air2p/Mtr4p polyadenylation (TRAMP) complex is activated to add short A-tails to the RNA, which is subsequently degraded by the exosome (Arigo et al., 2006; Thiebaut et al., 2006; Vasiljeva and Buratowski, 2006). Nuclear depletion of Nrd1 protein leads to drastic up-regulation of promoter noncoding RNAs (Schulz et al., 2013). The same pathway is unlikely to function in mammals. First, there is no clear homolog of the yeast Nrd1 and Nab3 in mammals. BLAST search identified Scaf4 and hnRNP-C as the mouse proteins with most similar amino acid sequences to yeast Nrd1 and Nab3 proteins, respectively. However, the sequence similarity is very low (10-30% of the protein sequences can be aligned with a max identify of 20-30%), and there is essentially no literature on the functions of mouse or human Scaf4, and hnRNP-C is primarily involved in pre-mRNA processing and has not been implicated in transcription termination and RNA degradation. Second, although the putative homolog of TRAMP subunits do form a complex, the complex is restricted to nucleoli, thus involved in rRNA degradation rather than promoter antisense transcripts (Lubas et al., 2011).

Lubas et al identify in human 293 cells the trimeric Nuclear Exosome Targeting (NEXT) complex, containing the RNA helicase MTR4, the Zn-knuckle protein ZCCHC8, and the putative RNA binding protein RBM7 (Lubas et al., 2011).

The NEXT complex is required for the degradation of PROMPTs by exosome. More recently, Andersen et al identified the CBCN complex, consisting of the Cap-binding complex (CBC) and the NEXT complex, as well as two other proteins: Arsenite resistance protein 2 (ARS2) and the zinc-finger protein ZC3H18 (Andersen et al., 2013). RIP followed by tiling analysis found that ARS2 binds preferentially to PROMPTs, suggesting ARS2 might direct preferential degradation of PROMPTs. However, siRNA knockdown of ARS2 yields minor changes in the level of selected PROMPTs. In addition, given that the exosome is a 3' to 5' exoribonuclease, it might have difficulty gaining access to the 3' end of the transcript when recruited to the 5' end cap of RNA.

RNA length or promoter proximity dependent degradation

PROMPTs/uaRNAs differ significantly from mRNA precursors in terms of transcript length. The average length of PROMPTs/uaRNAs is about 1kb, as compared to ~25 kb of nascent mRNA. Ntini et al recently show that in a gene construct, the RNA become less sensitive to exosome when the length increases, by moving the poly-A sites away from the transcription start site (Ntini et al., 2013). Similar data has been reported in a HIV1 construct (Andersen et al., 2012). Globally, comparing total RNA-seq in exosome depleted cells to wild type cells, the stabilization of PROMPTs peaks at about 700 bps upstream the TSS, then gradually decreases to background level as it gets further away from TSS, suggesting that longer transcripts are less susceptible to exosome. The mechanism underlying this length or distance dependent degradation is unclear. Here I propose two potential

models. First, exosome is enriched at the promoter region, therefore the 3' ends of shorter RNAs are closer to exosome, as compared the 3' ends of nascent mRNA that are ~25kb away from the promoter. Consistent with this model, the majority of the exosome ChIP-seq peaks are near active promoters in fly (Lim et al., 2013). The localization of exosome in promoter regions could also help to remove premature termination products from genes. Second, mechanisms exist to measure RNA length and sort them to different pathways (McCloskey et al., 2012). For example, hnRNP-C, the putative mammalian homolog of yeast Nab3, can sort small snRNA and mRNA to different nuclear export pathways based on the RNA length (McCloskey et al., 2012). The same or similar mechanisms might lead to preferential retention and / or degradation of short RNAs from pervasive transcription.

RNA Pol II CTD tyrosine 1 phosphorylation

Two recent studies suggested a potential role of tyrosine 1 phosphorylation (Tyr1P) in promoting the degradation of divergent noncoding transcripts (Descostes et al., 2014; Hsin et al., 2014). Descostes et al showed that in human cells, Tyr1P localizes to promoters, and behaves more like promoter-proximal marks Ser5P and Ser7P, as compared to the promoter-distal mark Ser2P (Descostes et al., 2014). In a subset of genes (~1000), Tyr1P showed enrichment relative to Ser5P and Ser7P, in promoter proximal polymerase and Tyr1P preferentially co-localized with promoter associated antisense small RNAs. Tyr1P is also a better mark for enhancers as compared to Ser5P and Ser7P, an interesting observation given that divergent transcription also occur at enhancers and is similarly unproductive.

Interestingly, in a separate study Hsin et al showed that in chicken DT40 cells expressing a mutant Pol II CTD where all but one Tyr were mutated to Phe (thus abolishing Tyr1P), over 90% of the ~120 genes with altered promoter antisense transcripts showed up-regulation of antisense RNAs, as assayed by 3' end polyA RNA sequencing (Hsin et al., 2014). Further analysis suggested that the increase in antisense RNAs is neither due to global down-regulation of exosome, nor increased transcription at those loci, leading to a potential role of Tyr1P in promoting rapid turnover of promoter antisense RNAs. Although promising, it is intriguing to notice that mutating all Serine 2 or Serine 5 in the CTD also showed predominant upregulation of promoter antisense RNAs, although to a lesser extent.

Summary

Most of the mechanisms described have only been reported in one species, such as fission yeast, budding yeast, or mammals. It is unclear whether most of these mechanisms are shared across species, or unique mechanisms evolve during evolution. It is also likely that different classes of noncoding RNA are more susceptible to certain pathways, and multiple pathways work together to suppress the activity of non-intended transcription or the resultant deleterious RNAs. The relative contribution of each mechanism in suppressing noncoding transcription is also unclear.

Functional consequences of pervasive noncoding transcription

Almost ten years after the discovery of widespread transcription activity in the mammalian genome (Birney et al., 2007; Carninci et al., 2005), it's still unclear whether most noncoding transcription and the resultant RNAs are functional or simply "noise" produced by the transcription machinery. Several challenges remain. First, the RNAs produced by pervasive transcription are large in number, diverse in sequences, and much less conserved than coding genes. Second, most noncoding RNAs are present in cells at very low copy numbers, making it difficult to measure or manipulate them. Third, many RNAs might only function under specific conditions, such as unknown developmental stages or external stress.

Despite these challenges, progress has been made by characterization of individual noncoding RNAs or classes of lncRNAs and uncovering diverse mechanisms of how the RNA or the act of transcription might impact specific genes, or the genome as a whole, or the species during evolution.

Regulating genes by noncoding RNAs

RNA is a versatile molecule, which could function in a variety of ways, by using sequence, secondary structure, or enzymatic activity from the 3D structure. It is thus not surprising that a variety of very diverse mechanisms have been uncovered through the detailed study of a small number of noncoding RNAs (Guttman and Rinn, 2012; Rinn and Chang, 2012; Wang and Chang, 2011; Wilusz et al., 2009).

Recruit or target chromatin complex to silence genes. This is the most well characterized role of noncoding RNAs, especially lincRNAs (Guttman et al., 2011;

Khalil et al., 2009). Many long noncoding RNAs, such as Xist and HOTAIR, have been shown to interact with chromatin remodeling complexes, such as Polycomb repressive complex 2 (PRC2), which then silence large number of target genes (Lee, 2012). However, it is still unclear whether the ncRNA or protein factors in the complex encode the target specificity.

Mediate enhancer-promoter interactions. Enhancer RNAs produced from active enhancers have been reported to mediate long-range interactions between enhancer and promoter (Lai et al., 2013; Lam et al., 2013; Li et al., 2013; Melo et al., 2013; Mousavi et al., 2013). It remains to see whether this is true for the majority of enhancer transcription, that the RNA itself is functional.

Post-transcriptional regulation. A few long noncoding RNAs are abundant and exported to cytoplasm to regulate mRNA. Again a variety of mechanisms are used, such as by acting as decoy or sponge of microRNAs (Ebert and Sharp, 2010a, 2010b; Ebert et al., 2007; Johnsson et al., 2013; Poliseno et al., 2010; Salmena et al., 2011), by inducing Staufen-mediated mRNA decay by pairing to 3' UTR repeats (Gong and Maquat, 2011; Wang et al., 2013), or by modulating translation through pairing to mRNA (Carrieri et al., 2012; Yoon et al., 2012).

Regulating genes by the act of transcription

Interference with gene transcription. Cis-interference with gene transcription can occur in two different ways. First, direct collision of the transcription machinery. Polymerase initiating from the noncoding region could travel into genes, knocking off transcription factors bound at gene promoters or transcribing polymerases in the gene body. This has been demonstrated in yeast (Prescott and

Proudfoot, 2002). This also seems to be the mechanism for the silencing of the imprinted Igf2r gene, by the overlapping transcription of the long noncoding RNA Airn in the absence of repressive chromatin in the promoter (Latos et al., 2012). Second, the act of noncoding transcription can change local chromatin to activate or silence nearby genes. For example in yeast, transcription at the noncoding locus SRG1 pushes nucleosomes to the downstream promoter of SER3, reduces the size of the nuclear free region and inactivates SER3 (Hainer et al., 2011).

Protect gene from transcription interference. The act of divergent transcription at promoters may serve as boundary elements preventing read-through interference from upstream promoters, including the promoter of an upstream gene or intergenic ncRNA, or alternative promoters of the same gene. A third of all mouse promoters are within genes or less than 2kb downstream of another gene, thus could potentially be impacted by such read through transcription. Given the pervasive intergenic transcription in mammalian genomes, an even larger fraction of genes are under risk of such interference. Divergent transcription activity drives Pol II moving convergently with respect to upstream read-through transcription, which will block its elongation due to transcription collision followed by subsequent RNA Pol II degradation (Hobson et al., 2012), therefore preventing upstream transcription from interfering with downstream gene transcription. In this regard, divergent promoters are similar to some B2 SINE elements, which are capable of initiating bidirectional Pol II and Pol III transcription, and act as chromatin domain boundary during mouse organogenesis (Lunyak et al., 2007).

Maintain permissive chromatin for gene transcription. Divergent transcription at promoter regions could maintain a permissive chromatin environment for gene transcription. The antisense Pol II could help to evict nucleosomes, generate negative supercoiling in the promoter and facilitate transcription initiation, and even the recycling of polymerase.

Reduce gene transcription noise. The stochastic nature of diffusion of transcription factors and polymerase searching for binding sites along the genome, and the random opening and shutting of chromatin, both introduce noise in transcription (Sanchez et al., 2013). Divergent transcription within the close proximity of gene promoters should reduce fluctuations by maintaining both the open chromatin state in the promoter region, and a pool of transcription machines that are accessible for gene transcription. Consistent with this, yeast genes associated with bidirectional promoters have significantly lower level of noise in gene expression (Wang et al., 2011).

Sliding to find target genes. Divergent transcription from enhancer regions may be one of the mechanisms underlying enhancer-promoter communications, namely the sliding/scanning/tracking model proposed in the 1980s (Bulger and Groudine, 1999). In this model, enhancer complex scans the entire region between enhancer and promoter to find the promoter and activate gene transcription. This model can explain many features of enhancer functions; such as it usually activates the nearest gene regardless the distance, and can be blocked by insulator or transcription terminator between enhancer and promoter. The presence of divergent transcription from enhancers, could explain another important feature of

enhancer function: orientation independence, i.e. enhancer can activate nearby genes regardless its orientation with respect to the gene. In fact the scanning model predicts enhancer divergent transcription, because the enhancer complex has no mechanism to determine which direction to search for the target gene, so for it to function it needs to initiate transcription and scan on both sides.

Genome organization and integrity

Besides regulating specific genes, noncoding transcription or noncoding RNA also contribute to the overall spatial organization of the genome, silencing of transposons, and genome rearrangement.

RNA-mediated genome organization. Noncoding RNAs such as Xist, have long been known to be part of the chromosome, without a function on specific genes. Recently, another repeat-containing noncoding RNA, Firre, has been found to organize a few loci on different chromosomes, without affecting the expression of nearby genes (Hacisuleyman et al., 2014). More recently, a large class of RNA, C_0t-1 RNA, RNA transcribed from the highly repetitive regions of the genome, collectively coats *in cis* all chromosomes except the inactivated X chromosome (Hall et al., 2014). These RNAs are very stably associated with nuclear matrix and help to maintain a decondensed state of the euchromatin. It is possible that many RNAs produced by pervasive transcription also function in a similar way. If that is the case, the loss of individual RNA would have little functional impact to the genome or the cell.

Instigate RNAi silencing of multi-copy repeats. Recently, Cruz et al propose that pervasive transcription could be a mechanism that detects and silences multi-copy repeats in the genome (Cruz and Houseley, 2014). Using a genetically

engineered yeast cells containing Dicer and Argonaute (Drinnenberg et al., 2009), found that endogenous RNAi is driven by copy number, i.e. high-copy loci generate more RNA that could form double-strand RNA, leading to more endogenous siRNA to degrade the RNA or silence the loci, preventing it from further jumping around the genome. For this to work, the genome needs to be pervasively transcribed to count the copy number of repeats.

Divergent transcription facilitates translocation. Two groups perform large-scale capture and sequencing to map translocation sites in the presence of AID, an enzyme that drives class switch recombination and somatic hypermutation in B cells (Chiarle et al., 2011; Klein et al., 2011). Both found a strong correlation between translocation breakpoint and the transcription start site of active genes. This is further supported by ChIP-seq showing AID binds active genes with open chromatin (Yamane et al., 2011), likely through the interaction with Spt5 and paused Pol II (Pavri et al., 2010). Interestingly, RNA exosome, responsible for degrading divergent transcripts, are also implicated in AID targeting to transcribed regions (Basu et al., 2011). Pefanis et al recently showed that translocations near a TSS or within gene bodies preferentially occurs over regions generating exosome substrate ncRNAs. These observations suggest that divergent transcription can recruit the exosome and facilitate AID targeting, which leads to translocation (Pefanis et al., 2014).

Evolutionary impact of the noncoding transcription

It is possible that some if not most noncoding transcripts are evolutionary young and have no biological function in the cell. They serve as raw materials for

evolutionary selection to work on, and may be selected and benefit the organism in the future. Carvunis et al propose the concept of proto-genes, pervasive noncoding transcripts containing short open reading frames that are also translated due to widespread translational activity (Carvunis et al., 2012). Nearly 2,000 proto-genes are detected in budding yeast. Similarly, a set of 24 protein-coding genes in human have been shown to evolve from noncoding transcripts in rhesus or chimpanzee (Xie et al., 2012).

Transcription is a mutagenic process that could accelerate evolution. A variety of transcription-dependent processes could cause mutations and recombination (Kim and Jinks-Robertson, 2012), including deamination, DNA damage, non-B-DNA structure, R-loop formation, and collision with replication complex. Transcription induced mutations, as well as transcription-coupled repair processes both contain sequence bias. Genomics studies have revealed signatures of transcription in transcribed regions, mainly increased G and T content on the coding strand (Green et al., 2003; Mugal et al., 2009; Polak et al., 2010). The functional implication of this mutational bias remains unclear. In Chapter 3 I explore the connections between divergent transcription, the U1-PAS axis, and transcription-induced mutational bias, leading to a model of how transcription drives new gene origination and share genome evolution.

Perspective

Despite significant progress, we are still at the early stages of understanding transcription in the noncoding genome. Most existing models for the regulation of noncoding transcription await further mechanistic details, and more importantly, it

is unclear at this moment the relative contribution of each pathway, or regulation at each stage. Genome-wide unbiased screens, similar to the recent one performed in yeast (Marquardt et al., 2014), have the potential to uncover novel pathways and assess their relative contribution. The recently developed genome-wide CRISPSR-Cas9 gene knockout libraries showed promising specificity and efficiency (Shalem et al., 2014; Wang et al., 2014), and will facilitate the screen for factors involved in the regulation of noncoding transcription.

One of the major challenges in understanding the function of noncoding transcription is the lack of tools for precise manipulation of the lowly abundant, nuclear enriched noncoding RNA, or the transcription activity itself. Again, the CRISPR-Cas9 system, and in particular, the CRISPR transcription interference system (Gilbert et al., 2013), can be used to increase or decrease the transcriptional activity of any noncoding region. We envision that the CRISPR-Cas9 technology will also accelerate the study of the function of noncoding transcription.

References

Adelman, K., and Lis, J.T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nat. Rev. Genet. *13*, 720–731.

Aguilera, A., and García-Muse, T. (2012). R loops: from transcription byproducts to threats to genome stability. Mol. Cell *46*, 115–124.

Almada, A.E., Wu, X., Kriz, A.J., Burge, C.B., and Sharp, P.A. (2013). Promoter directionality is controlled by U1 snRNP and polyadenylation signals. Nature *499*, 360–363.

Andersen, P.K., Lykke-Andersen, S.S., and Jensen, T.H. (2012). Promoter-proximal polyadenylation sites reduce transcription activity. Genes Dev. *26*, 2169–2179.

Andersen, P.R., Domanski, M., Kristiansen, M.S., Storvall, H., Ntini, E., Verheggen, C., Schein, A., Bunkenborg, J., Poser, I., Hallais, M., et al. (2013). The human cap-binding complex is functionally connected to the nuclear RNA exosome. Nat. Struct. Mol. Biol. *20*, 1367–1376.

Anderson, H.E., Wardle, J., Korkut, S.V., Murton, H.E., López-Maury, L., Bähler, J., and Whitehall, S.K. (2009). The fission yeast HIRA histone chaperone is required for promoter silencing and the suppression of cryptic antisense transcripts. Mol. Cell. Biol. *29*, 5158–5167.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.

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

Arigo, J.T., Eyler, D.E., Carroll, K.L., and Corden, J.L. (2006). Termination of cryptic unstable transcripts is directed by yeast RNA-binding proteins Nrd1 and Nab3. Mol. Cell *23*, 841–851.

Van Bakel, H., Nislow, C., Blencowe, B.J., and Hughes, T.R. (2010). Most "Dark Matter" Transcripts Are Associated With Known Genes. Plos Biol. *8*, e1000371.

Basu, U., Meng, F.-L.L., Keim, C., Grinstein, V., Pefanis, E., Eccleston, J., Zhang, T.T., Myers, D., Wasserman, C.R., Wesemann, D.R., et al. (2011). The RNA Exosome Targets the AID Cytidine Deaminase to Both Strands of Transcribed Duplex DNA Substrates. Cell *144*, 353–363.

Batt, D.B., Luo, Y., and Carmichael, G.G. (1994). Polyadenylation and transcription termination in gene constructs containing multiple tandem polyadenylation signals. Nucleic Acids Res. *22*, 2811–2816.

Beaulieu, Y.B., Kleinman, C.L., Landry-Voyer, A.M., Majewski, J., and Bachand, F. (2012). Polyadenylation-Dependent Control of Long Noncoding RNA Expression by the Poly(A)-Binding Protein Nuclear 1. Plos Genet. *8*, 17.

Berg, M.G., Singh, L.N., Younis, I., Liu, Q., Pinto, A.M., Kaida, D., Zhang, Z., Cho, S., Sherrill-Mix, S., Wan, L., et al. (2012). U1 snRNP Determines mRNA Length and Regulates Isoform Expression. Cell *150*, 53–64.

Bieberstein, N.I., Carrillo Oesterreich, F., Straube, K., and Neugebauer, K.M. (2012). First exon length controls active chromatin signatures and transcription. Cell Rep. *2*, 62–68.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799–816.

Bresson, S.M., and Conrad, N.K. (2013). The human nuclear poly(a)-binding protein promotes RNA hyperadenylation and decay. PLoS Genet. *9*, e1003893.

Brinster, R.L., Allen, J.M., Behringer, R.R., Gelinas, R.E., and Palmiter, R.D. (1988). INTRONS INCREASE TRANSCRIPTIONAL EFFICIENCY IN TRANSGENIC MICE. Proc. Natl. Acad. Sci. U. S. A. *85*, 836–840.

Bulger, M., and Groudine, M. (1999). Looping versus linking: toward a model for long-distance gene activation. Genes Dev. *13*, 2465–2477.

Byun, J.S., Fufa, T.D., Wakano, C., Fernandez, A., Haggerty, C.M., Sung, M.-H., and Gardner, K. (2012). ELL facilitates RNA polymerase II pause site entry and release. Nat. Commun. *3*, 633.

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

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

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.

Carvunis, A.R., Rolland, T., Wapinski, I., Calderwood, M.A., Yildirim, M.A., Simonis, N., Charloteaux, B., Hidalgo, C.A., Barbette, J., Santhanam, B., et al. (2012). Proto-genes and de novo gene birth. Nature *487*, 370–374.

Castelo-Branco, G., Amaral, P.P., Engström, P.G., Robson, S.C., Marques, S.C., Bertone, P., and Kouzarides, T. (2013). The non-coding snRNA 7SK controls transcriptional termination, poising, and bidirectionality in embryonic stem cells. Genome Biol. *14*, R98.

Chang, H., Lim, J., Ha, M., and Kim, V.N. (2014). TAIL-seq: Genome-wide Determination of Poly(A) Tail Length and 3' End Modifications. Mol. Cell *53*, 1044–1052.

Cheung, V., Chua, G., Batada, N.N., Landry, C.R., Michnick, S.W., Hughes, T.R., and Winston, F. (2008). Chromatin- and transcription-related factors repress transcription from within coding regions throughout the Saccharomyces cerevisiae genome. PLoS Biol. *6*, e277.

Chiarle, R., Zhang, Y., Frock, R.L., Lewis, S.M., Molinie, B., Ho, Y.-J., Myers, D.R., Choi, V.W., Compagno, M., Malkin, D.J., et al. (2011). Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. Cell *147*, 107–119.

Choi, T., Huang, M., Gorman, C., and Jaenisch, R. (1991). A generic intron increases gene expression in transgenic mice. Mol. Cell. Biol. *11*, 3070–3074.

Churchman, L.S., and Weissman, J.S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature *469*, 368–373.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. Science (80-.). *322*, 1845–1848.

Core, L.J., Waterfall, J.J., Gilchrist, D.A., Fargo, D.C., Kwak, H., Adelman, K., and Lis, J.T. (2012). Defining the status of RNA polymerase at promoters. Cell Rep. *2*, 1025–1035.

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

Cruz, C., and Houseley, J. (2014). Endogenous RNA interference is driven by copy number. Elife *3*, e01581.

Damgaard, C.K., Kahns, S., Lykke-Andersen, S., Nielsen, A.L., Jensen, T.H., and Kjems, J. (2008). A 5 ' splice site enhances the recruitment of basal transcription initiation factors in vivo. Mol. Cell *29*, 271–278.

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

Dawson, M.A., Prinjha, R.K., Dittmann, A., Giotopoulos, G., Bantscheff, M., Chan, W.-I., Robson, S.C., Chung, C., Hopf, C., Savitski, M.M., et al. (2011). Inhibition of BET recruitment to chromatin as an effective treatment for MLL-fusion leukaemia. Nature *478*, 529–533.

DeGennaro, C.M., Alver, B.H., Marguerat, S., Stepanova, E., Davis, C.P., Bähler, J., Park, P.J., and Winston, F. (2013). Spt6 regulates intragenic and antisense transcription, nucleosome positioning, and histone modifications genome-wide in fission yeast. Mol. Cell. Biol. *33*, 4779–4792.

Descostes, N., Heidemann, M., Spinelli, L., Schüller, R., Maqbool, M.A., Fenouil, R., Koch, F., Innocenti, C., Gut, M., Gut, I., et al. (2014). Tyrosine phosphorylation of RNA Polymerase II CTD is associated with antisense promoter transcription and active enhancers in mammalian cells. Elife *3*, e02105.

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.

Djebali, S., Davis, C.A., 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.

Drinnenberg, I.A., Weinberg, D.E., Xie, K.T., Mower, J.P., Wolfe, K.H., Fink, G.R., and Bartel, D.P. (2009). RNAi in budding yeast. Science *326*, 544–550.

Ebert, M.S., and Sharp, P.A. (2010a). Emerging roles for natural microRNA sponges. Curr. Biol. *20*, R858–61.

Ebert, M.S., and Sharp, P.A. (2010b). MicroRNA sponges: progress and possibilities. RNA *16*, 2043–2050.

Ebert, M.S., Neilson, J.R., and Sharp, P.A. (2007). MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. Nat. Methods *4*, 721–726.

Flynn, R.A., Almada, A.E., Zamudio, J.R., and Sharp, P.A. (2011). Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. Proc. Natl. Acad. Sci. U. S. A. *108*, 10460–10465.

Fong, Y.W., and Zhou, Q. (2001). Stimulatory effect of splicing factors on transcriptional elongation. Nature *414*, 929–933.

Frech, B., and Peterhans, E. (1994). RT-PCR: "background priming" during reverse transcription. Nucleic Acids Res. *22*, 4342–4343.

Furger, A., O'Sullivan, J.M., Binnie, A., Lee, B.A., and Proudfoot, N.J. (2002). Promoter proximal splice sites enhance transcription. Genes Dev. *16*, 2792–2799.

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

Glusman, G., Qin, S., El-Gewely, M.R., Siegel, A.F., Roach, J.C., Hood, L., and Smit, A.F.A. (2006). A third approach to gene prediction suggests thousands of additional human transcribed regions. PLoS Comput. Biol. *2*, e18.

Gong, C., and Maquat, L.E. (2011). lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. Nature *470*, 284–288.

Green, P., Ewing, B., Miller, W., Thomas, P.J., Green, E.D., and Progr, N.C.S. (2003). Transcription-associated mutational asymmetry in mammalian evolution. Nat. Genet. *33*, 514–517.

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

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., 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.

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

Hah, N., Murakami, S., Nagari, A., Danko, C.G., and Kraus, W.L. (2013). Enhancer Transcripts Mark Active Estrogen Receptor Binding Sites. Genome Res.

Hainer, S.J., Pruneski, J.A., Mitchell, R.D., Monteverde, R.M., and Martens, J.A. (2011). Intergenic transcription causes repression by directing nucleosome assembly. Genes Dev. *25*, 29–40.

Hall, L.L., Carone, D.M., Gomez, A. V, Kolpa, H.J., Byron, M., Mehta, N., Fackelmayer, F.O., and Lawrence, J.B. (2014). Stable C0T-1 repeat RNA is abundant and is associated with euchromatic interphase chromosomes. Cell *156*, 907–919.

He, Y., Vogelstein, B., Velculescu, V.E., Papadopoulos, N., and Kinzler, K.W. (2008). The antisense transcriptomes of human cells. Science *322*, 1855–1857.

Hennig, B.P., Bendrin, K., Zhou, Y., and Fischer, T. (2012). Chd1 chromatin remodelers maintain nucleosome organization and repress cryptic transcription. EMBO Rep. *13*, 997–1003.

Hicks, M.J., Yang, C.-R., Kotlajich, M. V, and Hertel, K.J. (2006). Linking splicing to Pol II transcription stabilizes pre-mRNAs and influences splicing patterns. PLoS Biol. *4*, e147.

Hobson, D.J., Wei, W., Steinmetz, L.M., and Svejstrup, J.Q. (2012). RNA polymerase II collision interrupts convergent transcription. Mol. Cell *48*, 365–374.

Houalla, R., Devaux, F., Fatica, A., Kufel, J., Barrass, D., Torchet, C., and Tollervey, D. (2006). Microarray detection of novel nuclear RNA substrates for the exosome. Yeast *23*, 439–454.

Houseley, J., LaCava, J., and Tollervey, D. (2006). RNA-quality control by the exosome. Nat. Rev. Mol. Cell Biol. *7*, 529–539.

Hsin, J.-P., Li, W., Hoque, M., Tian, B., and Manley, J.L. (2014). RNAP II CTD tyrosine 1 performs diverse functions in vertebrate cells. Elife *3*, e02112.

Jenal, M., Elkon, R., Loayza-Puch, F., van Haaften, G., Kühn, U., Menzies, F.M., Oude Vrielink, J.A.F., Bos, A.J., Drost, J., Rooijers, K., et al. (2012). The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites. Cell *149*, 538–553.

Jensen, T.H., Patricio, K., McCarthy, T., and Rosbash, M. (2001). A block to mRNA nuclear export in S. cerevisiae leads to hyperadenylation of transcripts that accumulate at the site of transcription. Mol. Cell *7*, 887–898.

Ji, X., Zhou, Y., Pandit, S., Huang, J., Li, H., Lin, C.Y., Xiao, R., Burge, C.B., and Fu, X.-D. (2013). SR proteins collaborate with 7SK and promoter-associated nascent RNA to release paused polymerase. Cell *153*, 855–868.

Johnson, J.M., Edwards, S., Shoemaker, D., and Schadt, E.E. (2005). Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. Trends Genet. *21*, 93–102.

Johnsson, P., Ackley, A., Vidarsdottir, L., Lui, W.-O., Corcoran, M., Grandér, D., and Morris, K. V (2013). A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells. Nat. Struct. Mol. Biol. *20*, 440–446.

Jonkers, I., Kwak, H., and Lis, J.T. (2014). Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. Elife *3*, e02407.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430–435.

Kaida, D., Berg, M.G., Younis, I., Kasim, M., Singh, L.N., Wan, L., and Dreyfuss, G. (2010). U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature *468*, 664–U81.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C.C., Suzuki, M., Kawai, J., et al. (2005). Antisense transcription in the mammalian transcriptome. Science (80-.). *309*, 1564–1566.

Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B.E., 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, N., and Jinks-Robertson, S. (2012). Transcription as a source of genome instability. Nat. Rev. Genet. *13*, 204–214.

Kim, T.-K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182–187.

Klein, I.A., Resch, W., Jankovic, M., Oliveira, T., Yamane, A., Nakahashi, H., Di Virgilio, M., Bothmer, A., Nussenzweig, A., Robbiani, D.F., et al. (2011). Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes. Cell *147*, 95–106.

Kouzine, F., Liu, J.H., Sanford, S., Chung, H.-J.J., and Levens, D. (2004). The dynamic response of upstream DNA to transcription-generated torsional stress. Nat. Struct. Mol. Biol. *11*, 1092–1100.

Kouzine, F., Sanford, S., Elisha-Feil, Z., and Levens, D. (2008). The functional response of upstream DNA to dynamic supercoiling in vivo. Nat. Struct. Mol. Biol. *15*, 146–154.

Kouzine, F., Gupta, A., Baranello, L., Wojtowicz, D., Ben-Aissa, K., Liu, J., Przytycka, T.M., and Levens, D. (2013). Transcription-dependent dynamic supercoiling is a short-range genomic force. Nat. Struct. Mol. Biol. *20*, 396–403.

Kouzine, F., Levens, D., and Baranello, L. (2014). DNA topology and transcription. Nucleus *5*.

Kruesi, W.S., Core, L.J., Waters, C.T., Lis, J.T., and Meyer, B.J. (2013). Condensin controls recruitment of RNA polymerase II to achieve nematode X-chromosome dosage compensation. Elife *2*, e00808–e00808.

Kwek, K.Y., Murphy, S., Furger, A., Thomas, B., O'Gorman, W., Kimura, H., Proudfoot, N.J., and Akoulitchev, A. (2002). U1 snRNA associates with TFIIH and regulates transcriptional initiation. Nat. Struct. Biol. *9*, 800–805.

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

Lam, M.T.Y., Cho, H., Lesch, H.P., Gosselin, D., Heinz, S., Tanaka-Oishi, Y., Benner, C., Kaikkonen, M.U., Kim, A.S., Kosaka, M., et al. (2013). Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. Nature *498*, 511–515.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

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

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

Lemay, J.-F., D'Amours, A., Lemieux, C., Lackner, D.H., St-Sauveur, V.G., Bähler, J., and Bachand, F. (2010). The nuclear poly(A)-binding protein interacts with the exosome to promote synthesis of noncoding small nucleolar RNAs. Mol. Cell *37*, 34–45.

Lemieux, C., Marguerat, S., Lafontaine, J., Barbezier, N., Bähler, J., and Bachand, F. (2011). A Pre-mRNA degradation pathway that selectively targets intron-containing genes requires the nuclear poly(A)-binding protein. Mol. Cell *44*, 108–119.

Lenasi, T., and Barboric, M. Mutual relationships between transcription and pre-mRNA processing in the synthesis of mRNA. Wiley Interdiscip. Rev. RNA *4*, 139–154.

Li, X., and Manley, J.L. (2006). Cotranscriptional processes and their influence on genome stability. Genes Dev. *20*, 1838–1847.

Li, X.L., and Manley, J.L. (2005). Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability. Cell *122*, 365–378.

Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., 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.

Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A.Y., Merkurjev, D., Zhang, J., Ohgi, K., Song, X., et al. (2013). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature *498*, 516–520.

Lim, S.J., Boyle, P.J., Chinen, M., Dale, R.K., and Lei, E.P. (2013). Genome-wide localization of exosome components to active promoters and chromatin insulators in Drosophila. Nucleic Acids Res. *41*, 2963–2980.

Lin, S., Coutinho-Mansfield, G., Wang, D., Pandit, S., and Fu, X.-D. (2008). The splicing factor SC35 has an active role in transcriptional elongation. Nat. Struct. Mol. Biol. *15*, 819–826.

Lubas, M., Christensen, M.S., Kristiansen, M.S., Domanski, M., Falkenby, L.G., Lykke-Andersen, S., Andersen, J.S., Dziembowski, A., and Jensen, T.H. (2011). Interaction profiling identifies the human nuclear exosome targeting complex. Mol. Cell *43*, 624–637.

Luco, R.F., and Misteli, T. (2011). More than a splicing code: integrating the role of RNA, chromatin and non-coding RNA in alternative splicing regulation. Curr. Opin. Genet. Dev. *21*, 366–372.

Luco, R.F., Allo, M., Schor, I.E., Kornblihtt, A.R., and Misteli, T. (2011). Epigenetics in alternative pre-mRNA splicing. Cell *144*, 16–26.

Lunyak, V. V, Prefontaine, G.G., Núñez, E., Cramer, T., Ju, B.-G., Ohgi, K.A., Hutt, K., Roy, R., García-Díaz, A., Zhu, X., et al. (2007). Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. Science *317*, 248–251.

Marinello, J., Chillemi, G., Bueno, S., Manzo, S.G., and Capranico, G. (2013). Antisense transcripts enhanced by camptothecin at divergent CpG-island promoters associated with bursts of topoisomerase I-DNA cleavage complex and R-loop formation. Nucleic Acids Res. *41*, 10110–10123.

Marquardt, S., Escalante-Chong, R., Pho, N., Wang, J., Churchman, L.S., Springer, M., and Buratowski, S. (2014). A Chromatin-Based Mechanism for Limiting Divergent Noncoding Transcription. Cell *157*, 1712–1723.

Martin, G., Gruber, A.R., Keller, W., and Zavolan, M. (2012). Genome-wide Analysis of Pre-mRNA 3 ' End Processing Reveals a Decisive Role of Human Cleavage Factor I in the Regulation of 3 ' UTR Length. Cell Rep. *1*, 753–763.

Masternak, K., Peyraud, N., Krawczyk, M., Barras, E., and Reith, W. (2003). Chromatin remodeling and extragenic transcription at the MHC class II locus control region. Nat. Immunol. *4*, 132–137.

McCloskey, A., Taniguchi, I., Shinmyozu, K., and Ohno, M. (2012). hnRNP C tetramer measures RNA length to classify RNA polymerase II transcripts for export. Science *335*, 1643–1646.

Melgar, M.F., Collins, F.S., and Sethupathy, P. (2011). Discovery of active enhancers through bidirectional expression of short transcripts. Genome Biol. *12*, R113.

Melo, C.A., Drost, J., Wijchers, P.J., van de Werken, H., de Wit, E., Vrielink, J., Elkon, R., Melo, S.A., Leveille, N., Kalluri, R., et al. (2013). eRNAs Are Required for p53-Dependent Enhancer Activity and Gene Transcription. Mol. Cell *49*, 524–535.

Moison, C., Arimondo, P.B., and Guieysse-Peugeot, A.-L. (2011). Commercial reverse transcriptase as source of false-positive strand-specific RNA detection in human cells. Biochimie *93*, 1731–1737.

Mousavi, K., Zare, H., Dell'Orso, S., Grontved, L., Gutierrez-Cruz, G., Derfoul, A., Hager, G.L., and Sartorelli, V. (2013). eRNAs Promote Transcription by Establishing Chromatin Accessibility at Defined Genomic Loci. Mol. Cell.

Mugal, C.F., von Gruenberg, H.-H., and Peifer, M. (2009). Transcription-Induced Mutational Strand Bias and Its Effect on Substitution Rates in Human Genes. Mol. Biol. Evol. *26*, 131–142.

Nam, D.K., Lee, S., Zhou, G., Cao, X., Wang, C., Clark, T., Chen, J., Rowley, J.D., and Wang, S.M. (2002). Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. Proc. Natl. Acad. Sci. U. S. A. *99*, 6152–6156.

Neil, H., Malabat, C., d'Aubenton-Carafa, Y., Xu, Z., Steinmetz, L.M., and Jacquier, A. (2009). Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. Nature *457*, 1038–1042.

Nguyen, A.T., and Zhang, Y. (2011). The diverse functions of Dot1 and H3K79 methylation. Genes Dev. *25*, 1345–1358.

Nicolas, E., Yamada, T., Cam, H.P., Fitzgerald, P.C., Kobayashi, R., and Grewal, S.I.S. (2007). Distinct roles of HDAC complexes in promoter silencing, antisense suppression and DNA damage protection. Nat. Struct. Mol. Biol. *14*, 372–380.

Nott, A., Meislin, S.H., Moore, M.J., and Muslin, S.H. (2003). A quantitative analysis of intron effects on mammalian gene expression. Rna-a Publ. Rna Soc. *9*, 607–617.

Ntini, E., Järvelin, A.I., Bornholdt, J., Chen, Y., Boyd, M., Jørgensen, M., Andersson, R., Hoof, I., Schein, A., Andersen, P.R., et al. (2013). Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. Nat. Struct. Mol. Biol. *20*, 923–928.

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

Palmiter, R.D., Sandgren, E.P., Avarbock, M.R., Allen, D.D., and Brinster, R.L. (1991). Heterologous introns can enhance expression of transgenes in mice. Proc. Natl. Acad. Sci. U. S. A. *88*, 478–482.

Patel, M.C., Debrosse, M., Smith, M., Dey, A., Huynh, W., Sarai, N., Heightman, T.D., Tamura, T., and Ozato, K. (2013). BRD4 coordinates recruitment of pause release factor P-TEFb and the pausing complex NELF/DSIF to regulate transcription elongation of interferon-stimulated genes. Mol. Cell. Biol. *33*, 2497–2507.

Paulsen, R.D., Soni, D. V, Wollman, R., Hahn, A.T., Yee, M.-C., Guan, A., Hesley, J.A., Miller, S.C., Cromwell, E.F., Solow-Cordero, D.E., et al. (2009). A Genome-wide siRNA Screen Reveals Diverse Cellular Processes and Pathways that Mediate Genome Stability. Mol. Cell *35*, 228–239.

Pavri, R., Gazumyan, A., Jankovic, M., Di Virgilio, M., Klein, I., Ansarah-Sobrinho, C., Resch, W., Yamane, A., San-Martin, B.R., Barreto, V., et al. (2010). Activation-Induced Cytidine Deaminase Targets DNA at Sites of RNA Polymerase II Stalling by Interaction with Spt5. Cell *143*, 122–133.

Pefanis, E., Wang, J., Rothschild, G., Lim, J., Chao, J., Rabadan, R., Economides, A.N., and Basu, U. (2014). Noncoding RNA transcription targets AID to divergently transcribed loci in B cells. Nature.

Perales, R., and Bentley, D. (2009). "Cotranscriptionality": the transcription elongation complex as a nexus for nuclear transactions. Mol. Cell *36*, 178–191.

Peterlin, B.M., and Price, D.H. (2006). Controlling the elongation phase of transcription with P-TEFb. Mol. Cell *23*, 297–305.

Pointner, J., Persson, J., Prasad, P., Norman-Axelsson, U., Strålfors, A., Khorosjutina, O., Krietenstein, N., Svensson, J.P., Ekwall, K., and Korber, P. (2012). CHD1 remodelers regulate nucleosome spacing in vitro and align nucleosomal arrays over gene coding regions in S. pombe. EMBO J. *31*, 4388–4403.

Polak, P., Querfurth, R., and Arndt, P.F. (2010). The evolution of transcription-associated biases of mutations across vertebrates. Bmc Evol. Biol. *10*, 187.

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

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., and Jensen, T.H. (2008). RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters. Science (80-.). *322*, 1851–1854.

Preker, P., Almvig, K., Christensen, M.S., Valen, E., Mapendano, C.K., Sandelin, A., and Jensen, T.H. (2011). PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. Nucleic Acids Res. *39*, 7179–7193.

Prescott, E.M., and Proudfoot, N.J. (2002). Transcriptional collision between convergent genes in budding yeast. Proc. Natl. Acad. Sci. U. S. A. *99*, 8796–8801.

Qu, X., Lykke-Andersen, S., Nasser, T., Saguez, C., Bertrand, E., Jensen, T.H., and Moore, C. (2009). Assembly of an export-competent mRNP is needed for efficient release of the 3'-end processing complex after polyadenylation. Mol. Cell. Biol. *29*, 5327–5338.

Rahl, P.B., Lin, C.Y., Seila, A.C., Flynn, R.A., McCuine, S., Burge, C.B., Sharp, P.A., and Young, R.A. (2010). c-Myc regulates transcriptional pause release. Cell *141*, 432–445.

Rigo, F., and Martinson, H.G. (2009). Polyadenylation releases mRNA from RNA polymerase II in a process that is licensed by splicing. RNA *15*, 823–836.

Rinn, J.L., and Chang, H.Y. (2012). Genome regulation by long noncoding RNAs. Annu. Rev. Biochem. *81*, 145–166.

Roy, D., Zhang, Z., Lu, Z., Hsieh, C.-L., and Lieber, M.R. (2010). Competition between the RNA transcript and the nontemplate DNA strand during R-loop formation in vitro: a nick can serve as a strong R-loop initiation site. Mol. Cell. Biol. *30*, 146–159.

Saguez, C., Schmid, M., Olesen, J.R., Ghazy, M.A.E.-H., Qu, X., Poulsen, M.B., Nasser, T., Moore, C., and Jensen, T.H. (2008). Nuclear mRNA surveillance in THO/sub2 mutants is triggered by inefficient polyadenylation. Mol. Cell *31*, 91–103.

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P.P. (2011). A ceRNA Hypothesis: The Rosetta Stone of a Hidden RNA Language? Cell *146*, 353–358.

Sanchez, A., Choubey, S., and Kondev, J. (2013). Regulation of noise in gene expression. Annu. Rev. Biophys. *42*, 469–491.

Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109–113.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*, 467–470.

Schmid, M., and Jensen, T.H. (2008). The exosome: a multipurpose RNA-decay machine. Trends Biochem. Sci. *33*, 501–510.

Schulz, D., Schwalb, B., Kiesel, A., Baejen, C., Torkler, P., Gagneur, J., Soeding, J., and Cramer, P. (2013). Transcriptome surveillance by selective termination of noncoding RNA synthesis. Cell *155*, 1075–1087.

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008). Divergent Transcription from Active Promoters. Science (80-.). *322*, 1849–1851.

Seila, A.C., Core, L.J., Lis, J.T., and Sharp, P.A. (2009). Divergent transcription: a new feature of active promoters. Cell Cycle *8*, 2557–2564.

Shalem, O., Sanjana, N.E., Hartenian, E., Shi, X., Scott, D.A., Mikkelsen, T.S., Heckl, D., Ebert, B.L., Root, D.E., Doench, J.G., et al. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. Science *343*, 84–87.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nat. Biotechnol. *26*, 1135–1145.

Shim, Y.S., Choi, Y., Kang, K., Cho, K., Oh, S., Lee, J., Grewal, S.I.S., and Lee, D. (2012). Hrp3 controls nucleosome positioning to suppress non-coding transcription in eu- and heterochromatin. EMBO J. *31*, 4375–4387.

Sigova, A.A., Mullen, A.C., Molinie, B., Gupta, S., Orlando, D.A., Guenther, M.G., Almada, A.E., Lin, C., Sharp, P.A., Giallourakis, C.C., et al. (2013). Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proc. Natl. Acad. Sci. U. S. A. *110*, 2876–2881.

Singh, B.N., and Hampsey, M. (2007). A transcription-independent role for TFIIB in gene looping. Mol. Cell *27*, 806–816.

Singh, G., Kucukural, A., Cenik, C., Leszyk, J.D., Shaffer, S.A., Weng, Z., and Moore, M.J. (2012). The Cellular EJC Interactome Reveals Higher-Order mRNP Structure and an EJC-SR Protein Nexus. Cell *151*, 750–764.

Subtelny, A.O., Eichhorn, S.W., Chen, G.R., Sive, H., and Bartel, D.P. (2014). Poly(A)-tail profiling reveals an embryonic switch in translational control. Nature *508*, 66–71.

Tan-Wong, S.M., Zaugg, J.B., Camblong, J., Xu, Z., Zhang, D.W., Mischo, H.E., Ansari, A.Z., Luscombe, N.M., Steinmetz, L.M., and Proudfoot, N.J. (2012). Gene loops enhance transcriptional directionality. Science *338*, 671–675.

Thiebaut, M., Kisseleva-Romanova, E., Rougemaille, M., Boulay, J., and Libri, D. (2006). Transcription termination and nuclear degradation of cryptic unstable transcripts: a role for the nrd1-nab3 pathway in genome surveillance. Mol. Cell *23*, 853–864.

Tilgner, H., Knowles, D.G., Johnson, R., Davis, C.A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T.R., and Guigó, 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.

Vasiljeva, L., and Buratowski, S. (2006). Nrd1 interacts with the nuclear exosome for 3' processing of RNA polymerase II transcripts. Mol. Cell *21*, 239–248.

Wang, K.C., and Chang, H.Y. (2011). Molecular mechanisms of long noncoding RNAs. Mol. Cell *43*, 904–914.

Wang, G.-Z., Lercher, M.J., and Hurst, L.D. (2011). Transcriptional Coupling of Neighboring Genes and Gene Expression Noise: Evidence that Gene Orientation and Noncoding Transcripts Are Modulators of Noise. Genome Biol. Evol. *3*, 320–331.

Wang, J., Gong, C., and Maquat, L.E. (2013). Control of myogenesis by rodent SINE-containing lncRNAs. Genes Dev. *27*, 793–804.

Wang, T., Wei, J.J., Sabatini, D.M., and Lander, E.S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. Science *343*, 80–84.

Whitehouse, I., Rando, O.J., Delrow, J., and Tsukiyama, T. (2007). Chromatin remodelling at promoters suppresses antisense transcription. Nature *450*, 1031–1035.

Wilusz, J.E., Sunwoo, H., and Spector, D.L. (2009). Long noncoding RNAs: functional surprises from the RNA world. Genes Dev. *23*, 1494–1504.

Wyers, F., Rougemaille, M., Badis, G., Rousselle, J.-C., Dufour, M.-E., Boulay, J., Régnault, B., Devaux, F., Namane, A., Séraphin, B., et al. (2005). Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. Cell *121*, 725–737.

Xie, C., Zhang, Y.E., Chen, J.-Y., Liu, C.-J., Zhou, W.-Z., Li, Y., Zhang, M., Zhang, R., Wei, L., and Li, C.-Y. (2012). Hominoid-Specific De Novo Protein-Coding Genes Originating from Long Non-Coding RNAs. Plos Genet. *8*, e1002942.

Xie, M.T., He, Y.H., and Gan, S.S. (2001). Bidirectionalization of polar promoters in plants. Nat. Biotechnol. *19*, 677–679.

Xu, Z.Y., Wei, W., Gagneur, J., Perocchi, F., Clauder-Munster, S., Camblong, J., Guffanti, E., Stutz, F., Huber, W., and Steinmetz, L.M. (2009). Bidirectional promoters generate pervasive transcription in yeast. Nature *457*, 1033–U7.

Yadon, A.N., Van de Mark, D., Basom, R., Delrow, J., Whitehouse, I., and Tsukiyama, T. (2010). Chromatin remodeling around nucleosome-free regions leads to repression of noncoding RNA transcription. Mol. Cell. Biol. *30*, 5110–5122.

Yamane, A., Resch, W., Kuo, N., Kuchen, S., Li, Z.Y., Sun, H.W., Robbiani, D.F., McBride, K., Nussenzweig, M.C., and Casellas, R. (2011). Deep-sequencing identification of the genomic targets of the cytidine deaminase AID and its cofactor RPA in B lymphocytes. Nat. Immunol. *12*, 62–U85.

Yoon, J.-H., Abdelmohsen, K., Srikantan, S., Yang, X., Martindale, J.L., De, S., Huarte, M., Zhan, M., Becker, K.G., and Gorospe, M. (2012). LincRNA-p21 suppresses target mRNA translation. Mol. Cell *47*, 648–655.

Zofall, M., Fischer, T., Zhang, K., Zhou, M., Cui, B., Veenstra, T.D., and Grewal, S.I.S. (2009). Histone H2A.Z cooperates with RNAi and heterochromatin factors to suppress antisense RNAs. Nature *461*, 419–422.

Chapter 2: Suppression of noncoding transcription by the U1-PAS axis

In this chapter, we present a mechanism for suppressing noncoding transcription, especially promoter divergent transcription, through activation of two RNA processing signals.

This chapter was published as:

Albert E. Almada*, Xuebing Wu*, Andrea J. Kriz, Christopher B. Burge, Phillip A. Sharp, Promoter directionality is controlled by U1 snRNP and polyadenylation signals, Nature, 2013, 499:360–363 (* equal contribution)

For supplementary tables and data:

http://www.nature.com/nature/journal/v499/n7458/full/nature12349.html#supplementary-information

Author contribution

A.E.A., X.W. and P.A.S. conceived and designed the research. A.E.A. performed experiments. X.W. and A.J.K. performed computational analysis. A.E.A., X.W., C.B.B. and P.A.S. analysed the data and wrote the manuscript.

Abstract

Transcription of the mammalian genome is pervasive but productive transcription outside protein-coding genes is limited by unknown mechanisms (Djebali et al., 2012). In particular, although RNA polymerase II (RNAPII) initiates divergently from most active gene promoters, productive elongation occurs primarily in the sense coding direction (Core et al., 2008; Preker et al., 2008; Seila et al., 2008). Here we show that asymmetric sequence determinants flanking gene transcription start sites (TSS) control promoter directionality by regulating promoter-proximal cleavage and polyadenylation. We find that upstream antisense RNAs (uaRNAs) are cleaved and polyadenylated at poly (A) sites (PAS) shortly after their initiation. De novo motif analysis reveals PAS signals and U1 snRNP (U1) recognition sites as the most depleted and enriched sequences, respectively, in the sense direction relative to the upstream antisense direction. These U1 and PAS sites are progressively gained and lost, respectively, at the 5' end of coding genes during vertebrate evolution. Functional disruption of U1 snRNP activity results in a significant increase in promoter-proximal cleavage events in the sense direction with slight increases in the antisense direction. These data suggests that a U1-PAS axis characterized by low U1 recognition and high density of PAS in the upstream antisense region reinforces promoter directionality by promoting early termination in upstream antisense regions whereas proximal sense PAS signals are suppressed by U1 snRNP. We propose that the U1-PAS axis limits pervasive transcription throughout the genome.

Introduction

Two potential mechanisms for suppressing transcription elongation in the upstream antisense region of gene TSS include inefficient release of paused RNAPII and / or early termination of transcription. RNAPII pauses shortly after initiation downstream of the gene TSS and the paused state is released by the recruitment and activity of p-TEFb (Adelman and Lis, 2012). A detailed characterization of several uaRNAs in mouse embryonic stem cells (mESCs) suggested that p-TEFb is recruited similarly in both sense and antisense directions (Flynn et al., 2011), and in human cells, elongating RNAPII (phosphorylated at serine 2 in the C-terminal domain) occupies the proximal upstream transcribed region (Preker et al., 2011). These data argue that the upstream antisense RNAPII complex undergoes the initial phase of elongation but likely terminates early due to an unknown mechanism.

Results

To globally test whether upstream antisense transcripts undergo early termination (compared to coding mRNA) by a canonical PAS-dependent cleavage mechanism, we mapped by deep sequencing the 3'-ends of polyadenylated RNAs in mESCs (Spies et al., 2013). For most protein-coding genes, transcription termination is triggered by cleavage of the nascent RNA upon recognition of a PAS whose most essential feature is an AAUAAA sequence or a close variant located about 10-30 nucleotides upstream of the cleavage site (Proudfoot, 2011). We sequenced two cDNA libraries and obtained over 230 million reads, of which 114 million mapped uniquely to the genome with at most two mismatches. We developed a

computational pipeline to identify 835,942 unique 3'-ends (cleavage sites) whose poly (A) tails are likely to be added post-transcriptionally and are also associated with the canonical PAS hexamer or its common variants (Supplementary Fig. 1, see Supplementary Methods).

To investigate whether uaRNAs are terminated by PAS-dependent mechanisms, we focused our analysis on cleavage sites proximal to gene TSS and at least 5kb away from known gene termination end sites (TES). Interestingly, in the upstream antisense region we observed a 2-fold higher number of cleavage sites compared to the downstream sense sites flanking protein-coding gene TSS (Fig. 1a). The peak of the upstream antisense cleavage sites is about 700 bps from the coding gene TSS. This observation suggests that upstream antisense transcripts are frequently terminated by PAS-directed cleavage shortly after initiation, a trend we also observe in various tissues of mouse and human (Derti et al., 2012) (Supplementary Fig. 2). Inspection of gene tracks at the Pigt locus reveals upstream antisense cleavage shortly after a PAS (AATAAA) less than 400 bases from the Pigt TSS, whereas in the sense direction cleavage is confined to the TES (Fig. 1b). Similar patterns were observed for subsets of promoters (promoters without nearby genes, GRO-seq defined divergent promoters, and ChIP-seq defined RNAPII-occupied promoters), or for high confidence cleavage sites, cleavage reads, and cleavage clusters (Supplementary Fig. 3). Of all divergent promoters, nearly half (48%) produce PAS-dependent upstream antisense cleavage events within 5kb of coding gene TSS, compared to 33% downstream of the TSS. We validated several of these

promoter proximal sense and antisense cleavage sites using 3'-RACE (Supplementary Fig. 4).

Similar to annotated cleavage sites at TES of genes, these upstream antisense cleavage sites are associated with the PAS located at the expected position, about 22 nucleotides upstream the cleavage site (Supplementary Fig. 5a-b) (Beaudoing et al., 2000; Tian et al., 2005). Moreover, the nucleotide sequence composition flanking the cleavage sites resembles that of TES of genes (Supplementary Fig. 5c-e) including a downstream U-rich region (Gil and Proudfoot, 1987; MacDonald et al., 1994). To determine whether members of the canonical cleavage and polyadenylation machinery bind specifically to uaRNA cleavage sites, we analyzed available cross-linking immunoprecipitation (CLIP) sequencing datasets for 10 canonical 3' end processing factors, including CPSF-160, CPSF-100, CPSF-73, CPSF-30, Fip1, CstF-64, CstF-64τ, CF I_m25, CF I_m59, and CF I_m68 along with poly (A) 3'-end sequencing data generated in HEK293 cells (Martin et al., 2012). We detect specific binding of all 10 factors at uaRNA cleavage sites with positional profiles identical or very similar to that of mRNA cleavage sites (Supplementary Fig. 6). These results indicate the poly (A) tails that we analyzed are products of PAS-dependent cleavage and polyadenylation, rather than either a priming artifact or PAS-independent polyadenylation representing a transient signal for RNA degradation (LaCava et al., 2005; Vanácová et al., 2005; Wyers et al., 2005).

As a first step to understand the molecular mechanism underlying the cleavage bias, we examined the frequency of PAS in a 6 kb region on the four strands flanking the coding gene TSS. We observed an approximately 33% depletion of the

canonical AATAAA PAS hexamer specifically downstream of the TSS on the coding strand of genes as compared to the other regions (Fig. 2a). Since this 33% depletion is unlikely to explain the 2-fold cleavage bias observed (see simulation results in Supplementary Fig. 8a), we searched for additional discriminative 6-mer sequence signals in an unbiased manner. All 4096 hexamers were ranked by enrichment in the first 1 Kb of the sense strand of genes relative to the corresponding upstream antisense region (Fig. 2b). Interestingly, we identified the PAS as the most depleted sequence in sense genes relative to the upstream antisense region of gene TSS. In addition, we identified 5' splice site related sequences (or sequences recognized by U1 referred to as U1 sites) as the most enriched hexamers in sense genes (Fig. 2b) relative to antisense regions. This includes the consensus GGUAAG (first) that is perfectly complementary to the 5' end of U1 snRNA, as well as GGUGAG (third) and GUGAGU (fifth), which represent common 5' splice site sequences (with the first GU in each motif located at the intron start). Consistent with the hexamer enrichment analysis, a metagene plot displaying an unbiased prediction of strong, medium, and weak U1 sites (see Supplementary Methods) revealed strong enrichment of U1 signals in the first 500 bps downstream of the TSS, with essentially only background levels observed in all other regions and a small depletion in the upstream antisense direction (Fig. 2c).

The asymmetric distribution of U1 sites and PAS sites flanking the TSS could potentially explain the biased cleavage pattern shown in Fig. 1a if the U1 snRNP complex suppresses cleavage and polyadenylation near a U1 site, as has been observed in various species including human and mouse (Andersen et al., 2012;

Berg et al., 2012; Kaida et al., 2010). Consistent with this model, we observed a depletion of cleavage sites, especially frequent cleavage sites, downstream of strong U1 sites (Supplementary Fig. 7a). Focusing on the upstream antisense direction, the presence of proximal PAS sites (within 1kb of coding gene TSS) is significantly associated with shorter uaRNAs (p < 1e-15), whereas the presence of proximal U1 sites is significantly associated with longer uaRNAs but only in the presence of proximal PAS sites (p < 0.0006), consistent with a model where U1 promotes RNA lengthening by suppressing proximal PAS (Supplementary Fig. 7b). To test whether the encoded bias in U1 and PAS signal distribution explains the cleavage bias observed from our 3'-end sequencing analysis, we performed a cleavage site simulation using predicted strong U1 sites and canonical PAS (AATAAA) sequences. Specifically, we defined a protection zone of 1 Kb downstream of a strong U1 site and used the first unprotected PAS as the cleavage site. The metagene plot of simulated cleavage events (Fig. 2d) recapitulate the major features of the observed distribution (Fig. 1a), including an antisense peak around 700 bps upstream and a ~2-fold difference between sense and antisense strands. Similar patterns were robustly observed when varying the size of the protection zone (Supplementary Fig. 8). Thus, we identified a U1-PAS axis flanking gene promoters that may explain why uaRNAs undergo early termination.

To validate the U1-PAS axis model, we functionally inhibited U1 snRNP in mESCs. Specifically, we transfected mESCs with either an antisense morpholino oligonucleotide (AMO) complementary to the 5' end of U1 snRNA to block its binding to 5' splice sites (or similar sequences) or a control AMO with scrambled

sequences followed by 3'-end RNA sequencing (Berg et al., 2012; Kaida et al., 2010). Interestingly, we observe in two biological replicates a dramatic increase in promoter-proximal cleavage events in coding genes but only a slight increase in upstream antisense regions, which eliminates the asymmetric bias in promoter-proximal cleavage we observed in either the wild-type cells or cells treated with scrambled control AMOs (Fig. 3). These observations confirm that U1 protects sense RNA in protein-coding genes from premature cleavage and polyadenylation in promoter proximal regions, thus, reinforcing transcriptional directionality of genes. However, in the antisense direction, the activity of U1 is much less and there is little enhancement in cleavage sites upon inhibition of U1 recognition.

The conservation of the asymmetric cleavage pattern across human and mouse (Supplementary Fig. 2) led us to examine if there is evolutionary selection on the U1-PAS axis. Previously, mouse protein-coding genes have been assigned to 12 evolutionary branches and dated by analyzing the presence or absence of orthologs in the vertebrate phylogeny (Zhang et al., 2010). We find strong trends of progressive gain of U1 sites depending on the age of a gene (Fig. 4a) and loss of PAS sites (Fig. 4b) over time at the 5' end (the first 1kb) of protein-coding genes, suggesting that suppression of promoter-proximal transcription termination is important for maintaining gene function. Interestingly, the same trends, although weaker, are observed in upstream antisense regions, suggesting at least a subset of uaRNAs may be functionally important in that over time they gain U1 sites and lose PAS sites to become more extensively transcribed. In addition to the coding strand of genes (downstream sense region), PAS sites were also progressively lost on the

other three strands flanking TSS (Fig. 4b). This observation probably reflects on the increases in CpG-rich sequences within 1 kb of gene TSS and suggests that coding genes acquire CpG islands as they age (Fig. 4c). However, the bias of low PAS site density in the sense direction extends across the total transcription unit (Supplementary Fig. 9) and is distinct from the CpG density near the promoter.

We also propose that some long noncoding RNAs (lncRNAs) generated from bidirectional promoters might represent an evolutionary intermediate between uaRNAs and protein-coding genes. Consistent with this, annotated head-to-head mRNA-lncRNA pairs as a whole showed a bias (in terms of promoter-proximal cleavage site, U1 site, and PAS site distributions flanking coding gene TSS) weaker than head-to-head mRNA-uaRNA pairs but stronger than mRNA-mRNA pairs (Supplementary Fig. 10). This is also consistent with recent results suggesting that *de novo* protein-coding genes originate from lncRNAs at bidirectional promoters (Xie et al., 2012).

The U1-PAS axis likely has a broader role in limiting pervasive transcription throughout the genome. The enrichment of U1 sites and depletion of PAS sites are confined to the sense strand within the gene body, whereas intergenic and antisense regions show relatively high PAS but low U1 density (Supplementary Fig. 9), indicating the U1-PAS axis may serve as a mechanism for terminating transcription in both antisense and intergenic regions.

Discussion

Together, we propose that a U1-PAS axis is important in defining the directionality for transcription elongation at divergent promoters (Supplementary

Fig. 11). Although the U1-PAS axis may explain the observed cleavage bias at promoters surprisingly well, it seems likely that additional cis-elements may influence PAS usage (Hu et al., 2005) and will need to be integrated into this model. There may also be other PAS-independent mechanisms that contribute to termination of transcription in upstream antisense regions and across the genome (Arigo et al., 2006; Connelly and Manley, 1989; Zhang et al., 2013). However, evidence for the U1-PAS axis is found in several different tissues of mouse and human, indicating its wide utilization as a general mechanism to regulate transcription elongation in mammals. Like protein-coding transcripts, lncRNAs must also contend with the U1-PAS axis. These RNAs and short non-coding RNAs from divergent transcription of gene promoters may be considered part of a continuum that varies in the degree of the activity of the U1-PAS axis.

Methods

Total RNA was extracted from V6.5 mESCs that were grown under standard ES cell culture conditions (Seila et al., 2008). Poly (A) RNA was selected, fragmented using a limited RNase T1 digestion, reverse transcribed using an oligo-dT containing primer, and the resulting cDNA was circularized and PCR amplified using Illumina-specific primers. U1 inhibition experiments were performed as previously described (Berg et al., 2012; Kaida et al., 2010).

Cell Culture. V6.5 (C57BL/6-129) mouse embryonic stem cells (mESCs) (Koch Institute Transgenic Facility) were grown under standard ES cell culture conditions (Seila et al., 2008).

Poly(A) 3'-End sequencing. Total RNA was extracted from V6.5 mESCs using Ambion's Ribopure kit (AM1924M). Poly (A) selected RNA was fragmented using RNase T1 (AM2283). Reverse transcription was performed with an RT oligo (Table S1) at 0.25 uM final concentration using Invitrogen's Superscript III Reverse Transcriptase (18080-44) according to the manufacturer's protocol. The resulting cDNA was run on a 6% TBE-Urea polyacrylamide gel (National Diagnostics) and the 100-300 size range of products were gel extracted and eluted overnight. The gel-purified cDNA products were circularized using CircLigase II (CL9025K) according to the manufacturer's protocol. Circularized cDNA was PCR-amplified using the Phusion High-Fidelity DNA Polymerase (MO530L) for 15-18 cycles using the primers described in Table S1. Amplified products were run on a 1.5 % agarose gel and the 200-400 size range was extracted using Qiagen's MinElute Gel Extraction Kit (28604). The 3'-end library was then submitted for Illumina sequencing on the HI-Seq 2000 platform.

U1 snRNP inhibition with antisense morpholino oligonucleotides (AMO). V6.5 mESCs were transfected using the Amaxa Nucleofector II with program A-23 (mESC-specific) according to the manufacturers protocol. Specifically, 2.5 million V6.5 mESCs were transfected with 7.5 uM of U1-targeting or a scrambed AMO for 8 hrs (Berg et al., 2012; Kaida et al., 2010) prior to RNA sequencing analysis.

3'-RACE. Total RNA was extracted using Ambion's Ribopure kit and DNase-treated using Ambion's DNA Free-Turbo. 3'-RACE was performed using Ambion's Gene Racer Kit according to the manufacturer's instructions. 3'-end PCR products

were run on a 1.5% agarose gel, gel extracted using Qiagen's gel extraction kit, and Sanger sequenced. All primers are described in Table S1.

Reads mapping. Raw reads were processed with the program *cutadapt* (Martin, 2011) to trim the adaptor sequence (TGGAATTCTCGGGTGCCAAGGAACTCCAGTCACATCAC) from the 3' end. Reads longer than 15 nts after adaptor trimming are mapped to the mouse genome (mm9) with *bowtie* (Langmead et al., 2009) requiring unique mapping with at most two mismatches (options: -n 2 -m 1 --best --strata). Mapped reads were collapsed by unique 3' end positions.

Internal priming filter. To remove reads whose A-tail is encoded in the genome rather than added post-transcriptionally, we filtered reads that have 1) more than 10 As in the first 20 nt window or 2) more than 6 As in the first 10 nt window downstream the 3' end. The threshold used is based on the bimodal distribution of the number of As downstream of annotated TES.

PAS filter. In addition to a set of 12 hexamers identified previously in mouse and human EST analysis (Beaudoing et al., 2000; Tian et al., 2005), we analyzed the annotated TES in the mouse genome to identify additional potential PAS variants. All hexamers with at most two mismatches to the canonical AATAAA motif were used to search in the sequence up to 100 nts. upsteam of annotated TES. The distribution of the position of each hexamer relative to the TES (a histogram) is compared to that of AATAAA. Hexamers with a position profile similar to AATAAA will have a peak around position 20-24. We quantified the similarity by Pearson correlation coefficient and used a cut-off of 0.5 after manual inspection. In total, 24

new hexamers were identified as potential PAS and a hierarchy was assigned for the 36 hexamers (PAS36): first, the 12 known variants are ranked by their frequency of usage in the mouse genome, and then the newly identified PAS ranked by their correlation with AATAAA in terms of the positional profile defined above. To define a window where most PAS or variants are located, we searched for each of the 36 PAS variants within 100 nts of annotated gene 3' ends and chose the best one according to the designated hierarchy. We summarized the distance of the best PAS to the annotated TES and defined a window of (0-41) around the position 22 peak such that 80% of the annotated TES have their best matched PAS within that window. Using this criteria, we searched for PAS36 variants within the 0-41 window upstream of our experimentally sequenced 3'-ends. If there were multiple PAS hexamers identified within this window for a given 3'-end, we chose the best one defined by the hierarchy described above. Reads without any of the 36 PAS variants within the 0-41 window were discarded.

Remove potential false positive cleavage sites. Due to sequencing error, abundant transcripts such as ribosomal gene mRNAs can produce error-containing 3' end reads that mapped to other locations in the genome, leading to false positive cleavage sites. To remove such potential false positive sites, we defined a set of 71674 (7.5%) abundant cleavage sites that are supported with more than 100 reads from the pooled library. A bowtie reference index was built using sequences within 50 nts upstream of those abundant sites. Non-abundant sites within these 50 nts reference regions were not used to search for false positives. Reads initially mapped to sites outside these reference regions were re-mapped against the new index

allowing up to two mismatches. Reads mapped to any of the reference regions in this analysis were treated as potential false positive reads. Cleavage sites containing only potential false positive reads are defined as potential false positive sites and were removed from subsequent analysis. In total, 7.2% (389185) of initially mapped reads are outside the reference regions. 0.34% of all mapped reads were classified as potential false positive reads and 9.1% (86425) of all cleavage sites were identified as potential false positive sites.

Remove B2 SINE RNA associated cleavage sites. We further removed cleavage sites associated with B2_Mm1a and B2_Mm1t SINE RNAs. These B2 SINE RNAs are transcribed by RNA Pol III but contain AAUAAA sequences near the 3' end. In total, 3.5% (33696) of all cleavage sites passing the internal priming filter and the PAS filter were mapped within B2 regions or within 100 nts downstream of B2 3' end. These sites were removed.

Prediction of U1 sites / putative 5' splice sites. A nucleotide frequency matrix of the 5' splice sites (3 nt in exon and 6 nt in intron) was compiled using all annotated constitutive 5' splice sites in the mouse genome. The motif was then used by FIMO (Grant et al., 2011) to search significant matches ($p<0.05$) on both strands of the genome. Matches were then scored by a Maximum Entropy model (Yeo and Burge, 2004). Maximum entropy scores for all annotated 5' splice sites were also calculated to define thresholds used to classify the predicted sites into strong, medium and weak. Sites with scores larger than the median of annotated 5' splice sites (8.77) were classified as 'strong'. Sites with scores lower than 8.77 but higher than the threshold dividing the first and second quarter of annotated 5' splice sites

(7.39) were classified as 'medium', and the rest of the predicted sites with scores higher than 4 were classified as 'weak'. Sites with scores lower than 4 were discarded.

Define a set of divergent promoters. GRO-seq data from mouse embryonic stem cells (Min et al., 2011) were used to define a set of active promoters and divergent promoters. Active promoters were defined as promoters with GRO-seq signal detected within the first 1kb downstream sense strand. Divergent promoters were defined as active promoters that further have detected GRO-seq signal within the first 2kb upsteam antisense strand. A minimum number of two reads within the defined window (downstream 1kb or upstream 2kb) were used as a cut-off for background signals.

Define Ser5p RNA Pol II bound TSS. ChIP-seq data for ser5p RNA Pol II and corresponding input was downloaded from GEO database (accession number GSE20530 (Rahl et al., 2010)) and peaks called using MACS (Zhang et al., 2008) with default settings. TSS less than 500bps away from a peak summit are defined as bound.

Discriminative hexamer analysis. An unbiased exhaustive enumeration of all 4096 hexamers was performed to find hexamers that are discriminative of downstream sense and upstream antisense strands of protein-coding gene promoters. Specifically, the first 1000 nucleotides downstream sense and upstream antisense of all protein-coding gene TSS were extracted from repeat masked genome (from UCSC genome browser, non-masked genome sequence gave similar results). For each hexamer, the total number of occurrences on each side was

counted and then the log2 ratio of the occurrences on sense versus antisense strand was calculated as a measure of enrichment on the sense but depletion on the antisense strand.

Cleavage site simulation. Protein-coding genes and 10kb upstream antisense regions were scanned for strong U1 sites and PAS sites (AATAAA). Starting from protein coding gene TSS, the first unprotected PAS was predicted to be the cleavage site. A PAS is protected only if it is within a designated protection window (in nucleotides) downstream (+) of a strong U1 site.

Binding of 3' end processing factors in uaRNA regions. RNA 3' end cleavage and polyadenylation sites and CLIP-seq read density of ten 3' end processing factors in wild type HEK293 cells were downloaded from Gene Expression Omnibus (GEO) dataset GSE37401. A cleavage site is defined as a uaRNA cleavage site if it is outside any protein-coding gene but locates within 5kb upstream antisense of a protein-coding gene. mRNA cleavage sites are defined as cleavage sites within 100 bases of annotated protein-coding gene ends. For each 3' end-processing factor, CLIP read density within 200 bases of all cleavage sites are added up every 5bp bin and then normalized such that the max value is 1.

Evolutionary analysis of U1 sites, PAS sites, and CpG islands. Mouse protein-coding gene branch/age assignment was obtained from a previous analysis (Zhang et al., 2010). The number of strong U1 sites, PAS (AATAAA) sites, and CpG islands (UCSC mm9 annotations) in the first 1kb region flanking TSS on each strand were calculated, and the average number of sites in each branch/age group was plotted against gene ages. Pearson correlation coefficient and linear regression

fitting were done using R. Significance of the correlation was assessed by comparing to a null distribution of correlation coefficients calculated by shuffling gene branch/age assignments 1000 times.

Bidirectional promoter analysis. For each annotated TSS the closest upstream antisense TSS was identified and those TSS pairs within 1kb were defined as head-to-head pairs. LncRNAs were defined as noncoding RNAs longer than 200 bps. UCSC mm9 gene annotations were used in this analysis.

Figures

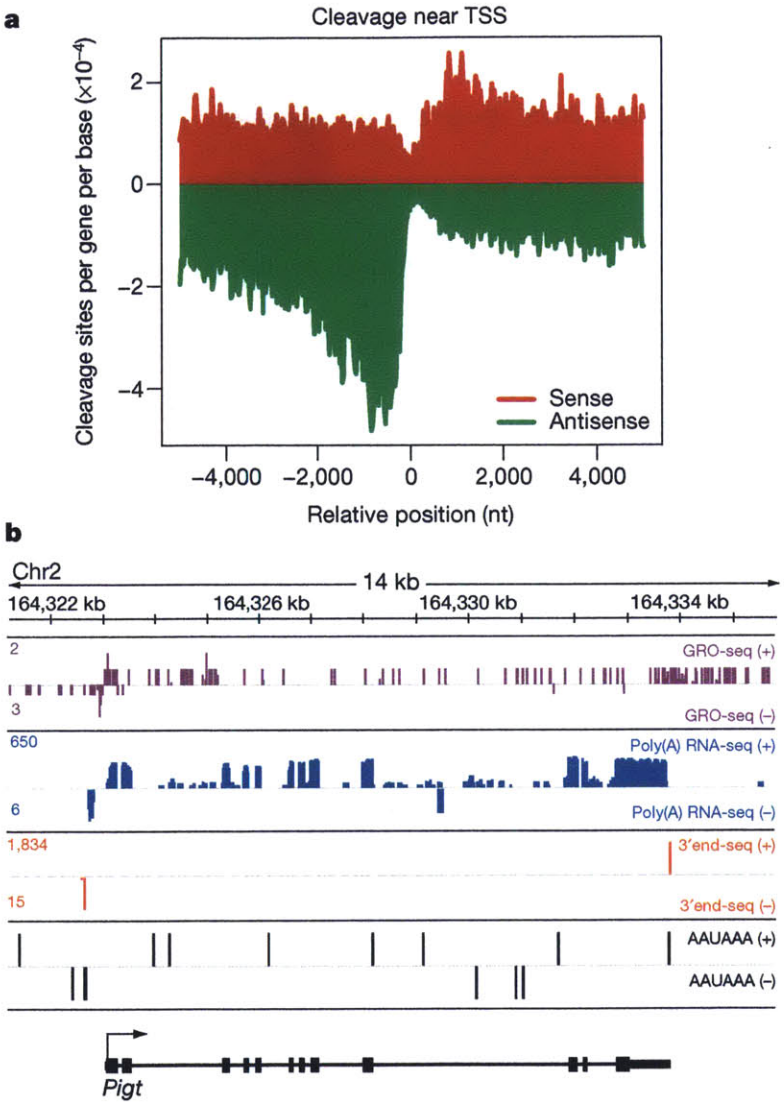


Figure 1. Promoter-proximal PAS-dependent termination of uaRNA. a, Metagene plot of sense (red) or antisense (green) unique cleavage sites flanking coding gene TSS. The number of unique cleavage sites per gene per base in each 25 bp bin across 5 Kb upstream and downstream of the TSS is plotted. Mean cleavage density of first 2 Kb: sense/antisense = 1.45/3.10. **b**, Genome browser view from the Pigt locus (shown in black on the + strand) displaying the following tracks with + strand (top) and – strand (bottom) represented: GRO-Seq (purple) (Min et al., 2011), Poly (A)+ RNA-Seq (blue) (Sigova et al., 2013), 3'end RNA-Seq (orange), and PAS (AAUAAA, black). For each gene track, the x-axis represents the linear sequence of genomic DNA. The numbers on the top left corner represent the maximum read density on each track.

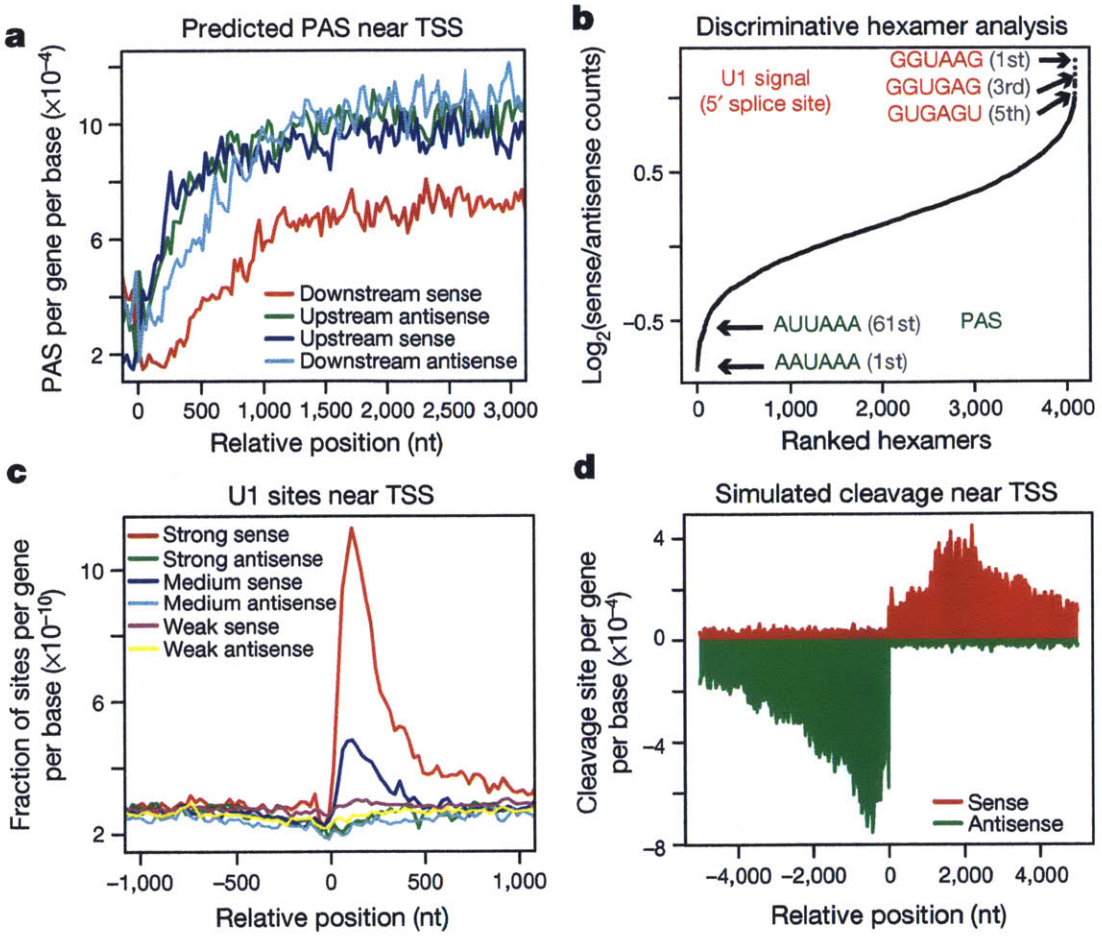


Figure 2. Asymmetric distribution of PAS and U1 signals flanking coding gene TSS. a. Number of AATAAA sites per gene per base in each 25 bp bin within a 3 Kb region flanking gene TSS on the downstream sense (red), downstream antisense (light blue), upstream antisense (green), and upstream sense (dark blue) strands. **b**, Rank of all 4096 hexamers by enrichment (log2 ratio) in the first 1 Kb of all coding genes in the sense direction relative to 1 Kb in the upstream antisense direction of the TSS. **c**, Density of predicted 5' splice sites within a 1 Kb region flanking gene TSS. Strong, medium, and weak 5' splice sites are defined in Methods. **d**, Metagene plot of simulated cleavage sites around gene TSS. The first unprotected PAS (AAUAAA) that is not within 1 Kb downstream of a strong U1 site for all coding genes is plotted. Mean cleavage density of first 2kb: sense/antisense = 2.08/4.99.

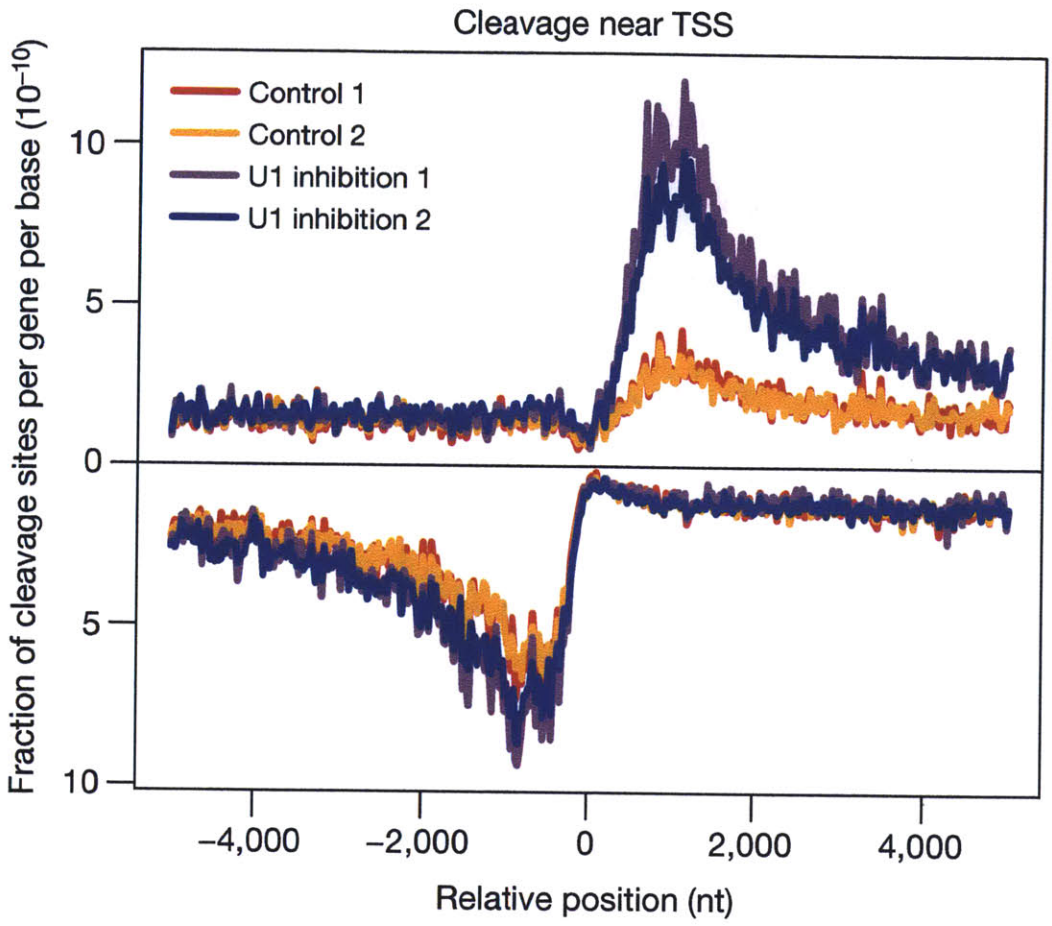


Figure 3. Promoter-proximal cleavage sites are altered upon functional U1 snRNP inhibition. Y-axis is the number of cleavage sites per gene per base divided by the total number of cleavage sites identified in each 3' end-sequencing library in a 5 Kb region flanking coding gene TSS. Signal for the antisense strand is set as negative. U1 inhibition 1 (purple) and U1 inhibition 2 (blue) represent 3'-end sequencing libraries generated from mESCs treated with a U1-targeting AMO. Control 1 (red) and Control 2 (orange) represent 3'-end sequencing libraries generated from mESCs treated with a scrambled control AMO. Mean cleavage density of first 2kb: sense/antisense = 2.5/4.4 (Control 1), 2.4/4.3 (Control 2), 7.0/5.8 (U1 inhibition 1), 5.9/5.5 (U1 inhibition 2).

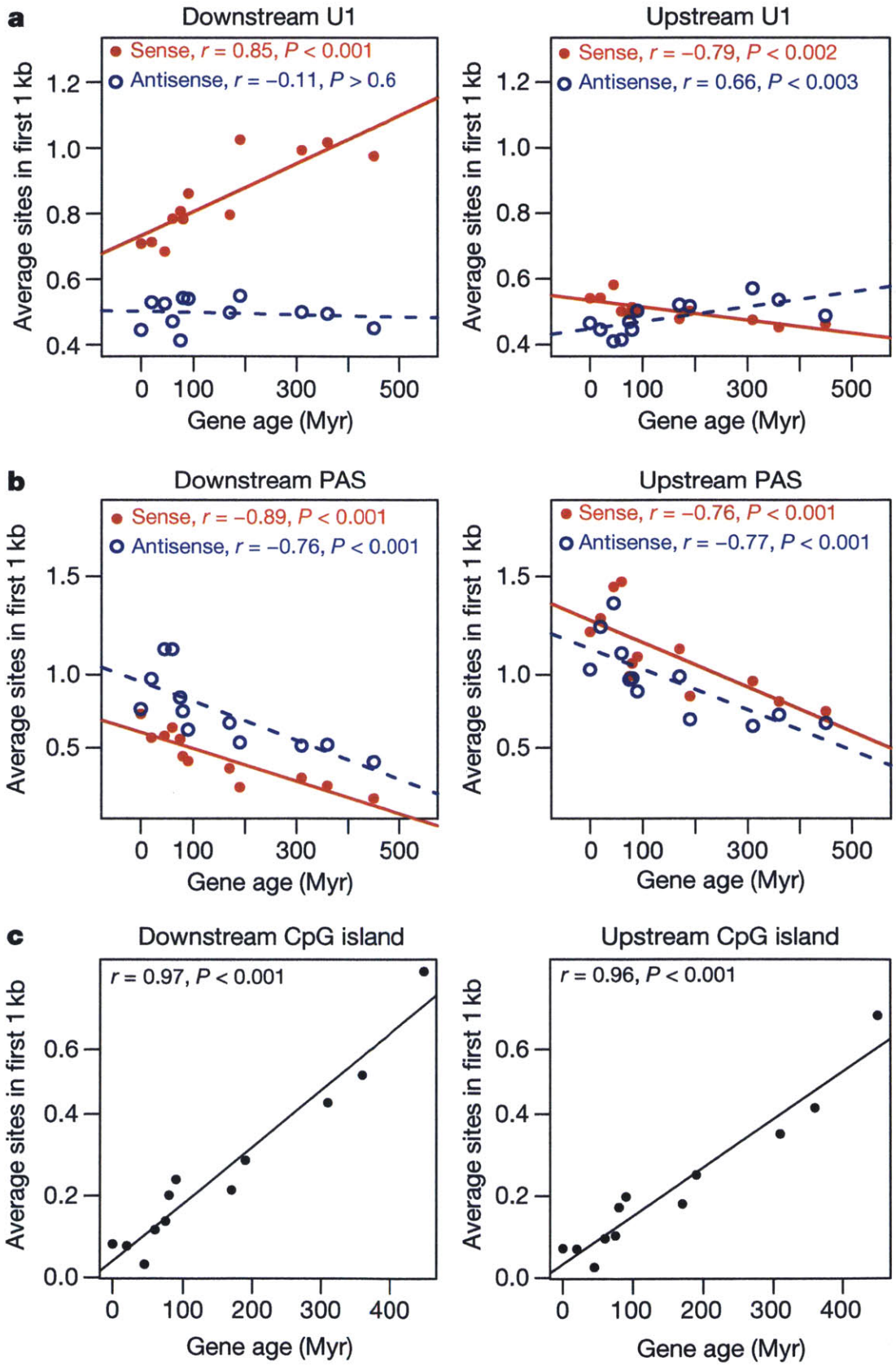
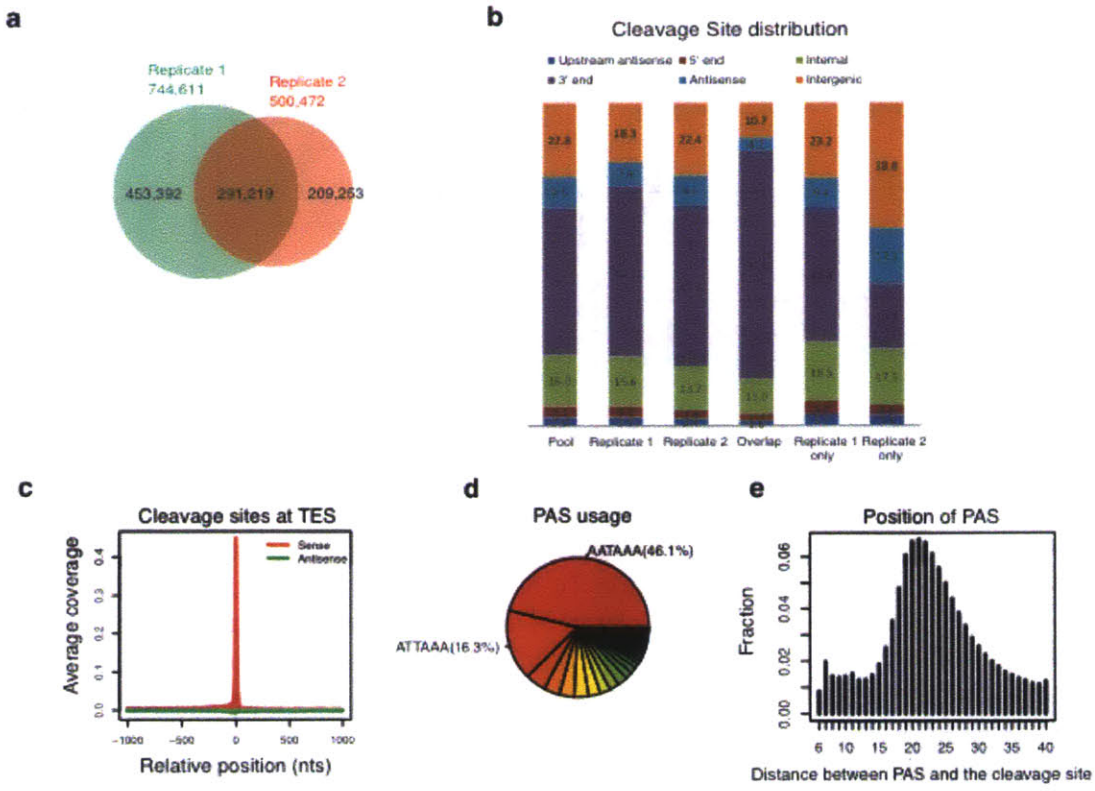
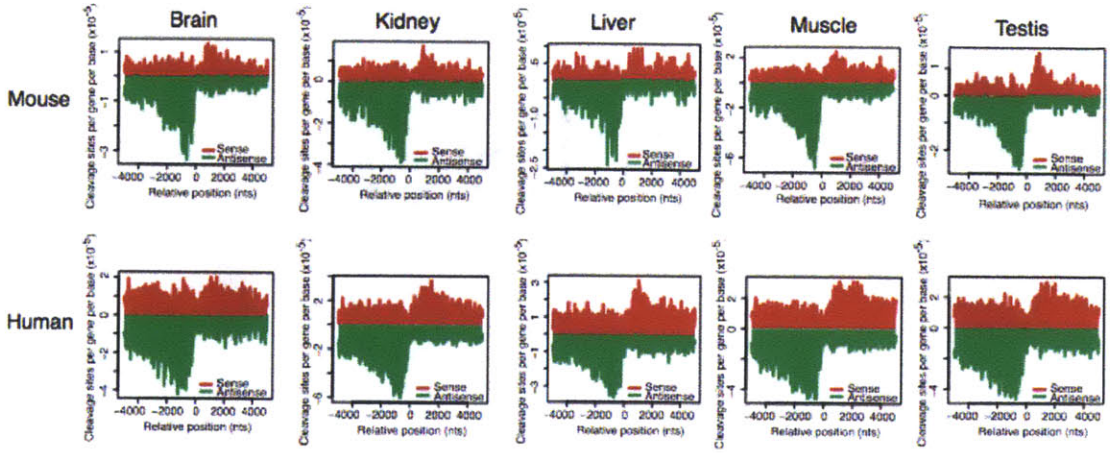

a
Downstream U1
Sense, r = 0.85, P < 0.001
Antisense, r = −0.11, P > 0.6
Average sites in first 1 kb
Gene age (Myr)
Upstream U1
Sense, r = −0.79, P < 0.002
Antisense, r = 0.66, P < 0.003
b
Downstream PAS
Sense, r = −0.89, P < 0.001
Antisense, r = −0.76, P < 0.001
Upstream PAS
Sense, r = −0.76, P < 0.001
Antisense, r = −0.77, P < 0.001
c
Downstream CpG island
r = 0.97, P < 0.001
Upstream CpG island
r = 0.96, P < 0.001

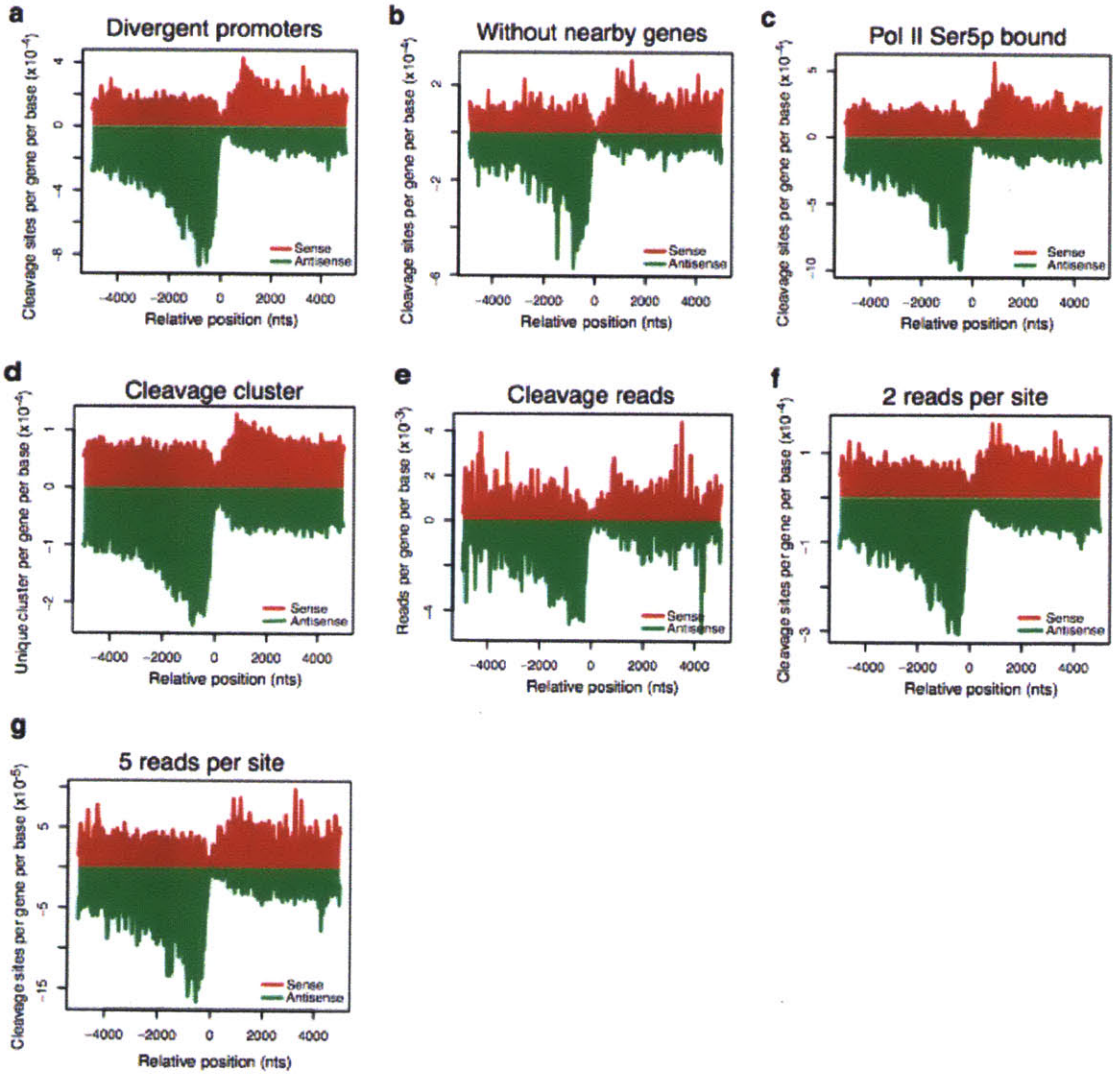
Figure 4. Evolutionary gain and loss of U1 and PAS sites. a, Average number of strong U1 sites in the first 1 Kb of protein-coding genes and upstream regions. **b**, Average number of PAS sites in the first 1 Kb downstream and upstream of coding gene TSS, respectively. **c**, Average number of CpG islands overlapping the first 1 Kb of protein-coding genes and upstream regions. Genes are divided into 12 ordered groups by gene age. X-axis indicates the age (myr, million years) of gene groups. The number of genes in each group (from old to young): 11934, 1239, 914, 597, 876, 1195, 279, 175, 198, 315, 926, and 1143. Solid red dots and blue circles indicate sites on the sense and antisense strands, respectively.



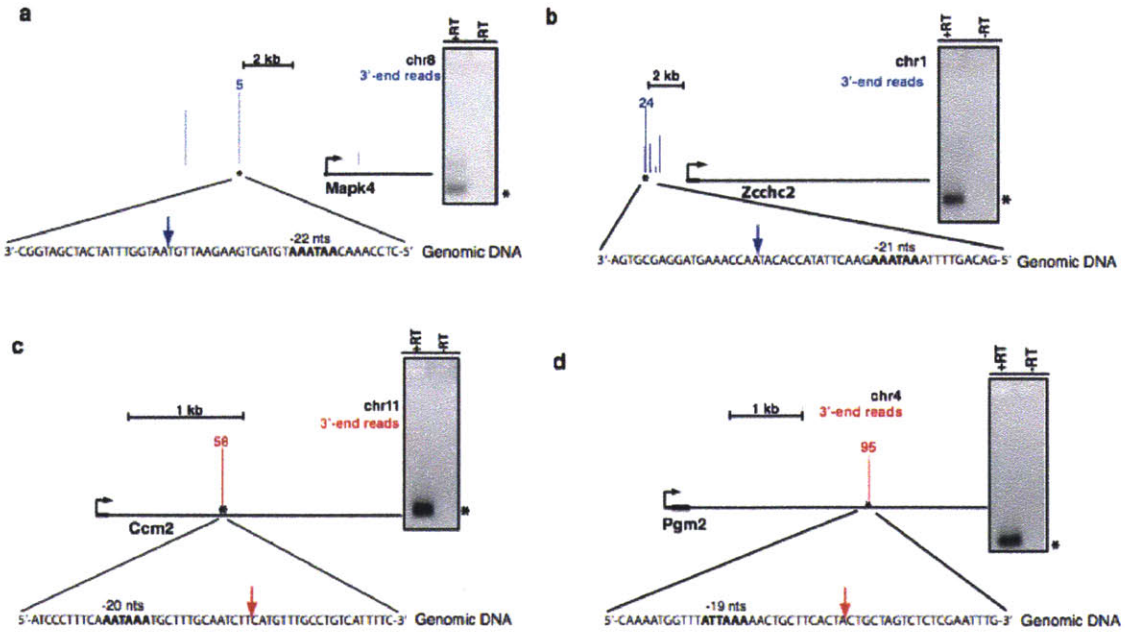
Supplementary Figure 1. Mapping the 3' ends of polyadenylated RNAs by deep sequencing in mESCs. (a) Venn diagram depicts the overlap of unique cleavage sites between two 3'-end libraries that were constructed and denoted as library replicate 1 and replicate 2. (b) The fraction of cleavage sites in six non-overlapping categories including: 2 kb flanking 3' end of the gene (3' end), 5 kb downstream the TSS in the gene (5' end), internal of the gene (Internal, not 5' end or 3' end), upstream antisense of the TSS within 5 kb (Upstream antisense), antisense to the gene (Antisense), and other intergenic regions (Intergenic) in pool (combining replicate 1 and 2), replicate 1, replicate 2, overlap (only common to replicate 1 and 2), and sites unique to replicate 1 or replicate 2. (c) Density of unique cleavage sites at annotated 3' ends of genes with sense and antisense sites shown in red and green, respectively. Position zero denotes the annotated TES. Average coverage equals the number of unique cleavage sites per nucleotide per gene. (d) Pie chart displaying the usage of each PAS (all percentages shown in Supplementary Table 4a) among all unique cleavage sites. (e) Histogram showing the distance of the PAS (all 36 hexamers) 5' end relative to the cleavage site (indicated as position zero on the x-axis) and fraction of all cleavage sites that have a PAS at each position is shown on the y-axis.



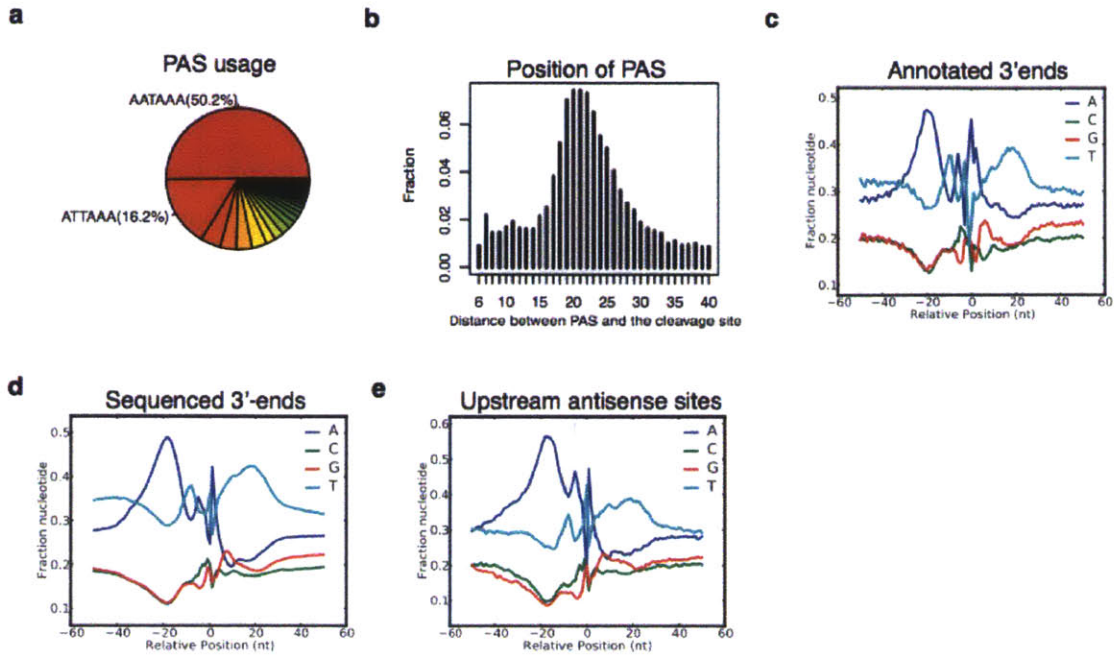
Supplementary Figure 2. The cleavage bias near gene TSS is conserved in various tissues in mouse and human. To determine if the bias found in mouse ES cells can be observed in other mouse tissues or another mammalian species, we examined published 3'-end sequencing data. Panels display metagene plots of sense (red) or antisense (green) unique cleavage sites flanking coding gene TSS. The number of unique cleavage sites in each 25 bp bin across 5 kb upstream and downstream of the TSS is plotted and unique cleavage sites within 5 kb of annotated 3'-ends were removed. In all tissues of human and mouse, we observed more upstream antisense cleavage and a promoter proximal antisense peak. Despite different sets of genes being expressed across various tissues and analyzing 3'-end sequencing data generated from another mammalian species, the pattern is consistent with the biased distribution of PAS and U1 sites that is generally encoded in gene sequences.



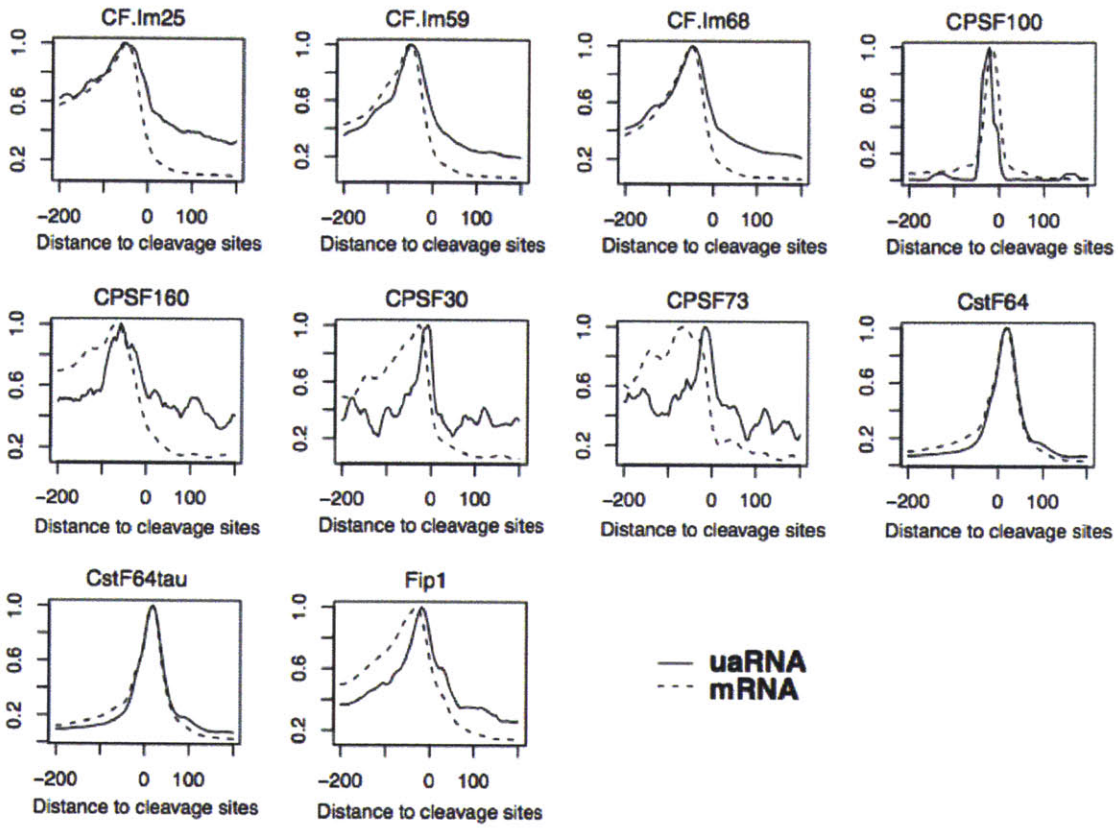
Supplementary Figure 3. Metagene analysis of cleavage sites near gene TSS. Displayed metagene plots (a-g) were generated in the same way as Figure 1a with the specified modifications. (a) Plot focusing on divergent promoters (details in methods), (b) or a subset of promoters where the gene is at least 6 kb in size and there are no other TSS or TES within the 10 kb window. Unlike Figure 1a, sites within 5 kb of TES were not removed. (c) A plot displaying a subset of promoters that showed significant Ser5 phosphorylated Pol II peaks in mESCs. For metagene plots a-c, only unique cleavage sites are being plotted. (d) Plotting the density of unique cleavage clusters (cleavage sites within 24 bps were clustered together and the most 5' sites are used as a reference site of the cluster). (e) Plotting read density instead of unique cleav- age sites. Sites with more than 500 supporting reads were removed from the plot since they could be unannotated gene ends. Metagene plots (f-g) were generated in the same way as Figure 1a except taking a subset of unique cleavage sites with at least two (f) or five (g) supporting reads.



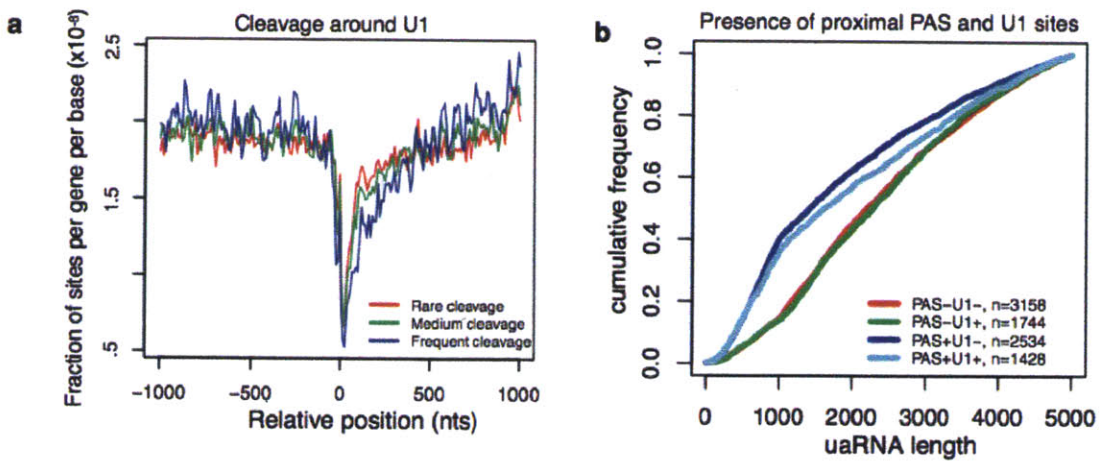
Supplementary Figure 4. Validation of promoter proximal antisense (a-b) and sense (c-d) cleavage sites using 3'-RACE. Each panel displays a genome browser view of the promoter proximal region at four coding genes: Mapk4 (a), Zcchc2 (b), Ccm2 (c), Pgm2 (d) with the gene TSS denoted with a black arrow pointing towards the right. Promoter proximal 3'-end cleavage reads for uaRNA (blue) and mRNA (red) are displayed above each gene schematic shown in black. The assayed cleavage site is denoted with an asterisk and the number of reads supporting each site is displayed above each site. We validated the most prominent cleavage site (supported by the most number of reads) for each uaRNA loci. Agarose gels of 3'-RACE PCR products are displayed to the right and each assayed cleavage site (asterisk) was cloned and sequenced using Sanger sequencing methods. Scale bars are represented in black above genes. The encoded genome sequence is displayed including the sequence of the PAS (bold) and the distance between the cleavage site (blue and red arrow for uaRNA and mRNA, respectively) and the 5'-end nucleotide of the PAS is noted above.



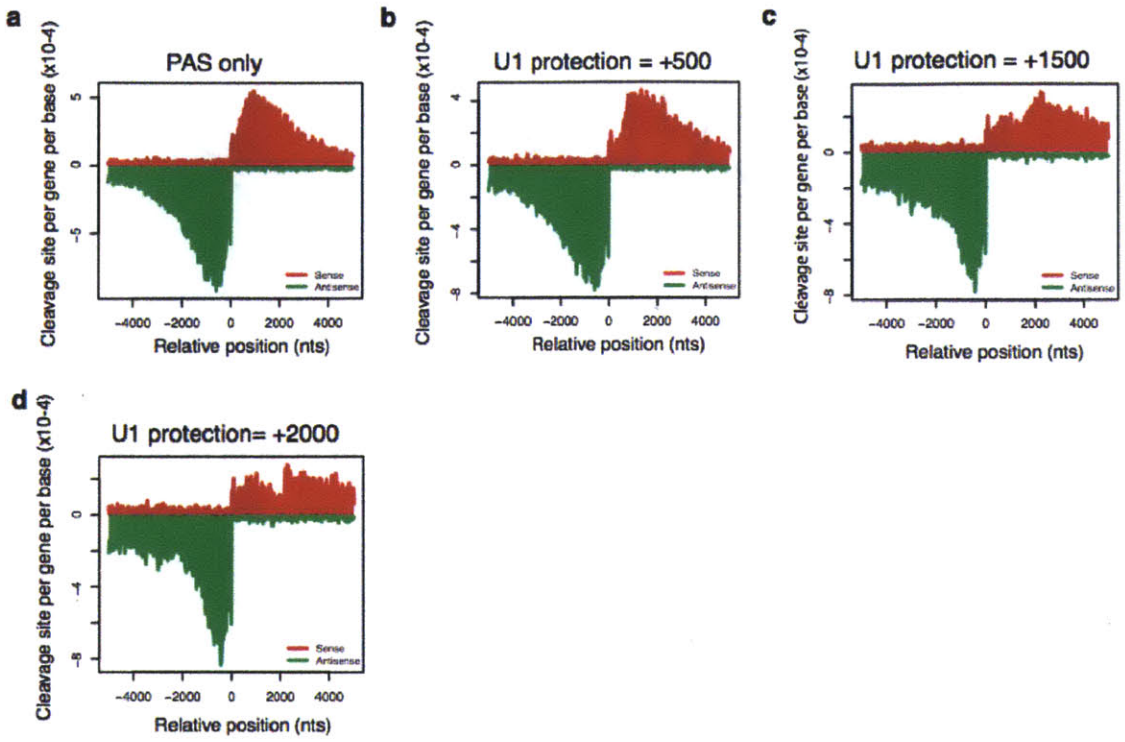
Supplementary Figure 5. Upstream antisense cleavage sites resemble annotated gene TES. (a) Pie chart displaying the usage of each PAS among unique cleavage sites in the upstream antisense region. (b) Histogram showing the distance of the PAS 5' end relative to the cleavage site indicated as position zero. For (a-b), figures include all 36 PAS hexamers with the percentage of all PAS hexamers in (a) described in Supplementary Table 4b. (c-e) The nucleotide frequency flanking cleavage sites (position 0): annotated end of genes (c), cleavage sites detected from our 3' end sequencing -- sites within 2 kb of annotated gene ends (d), and upstream antisense sites (e).



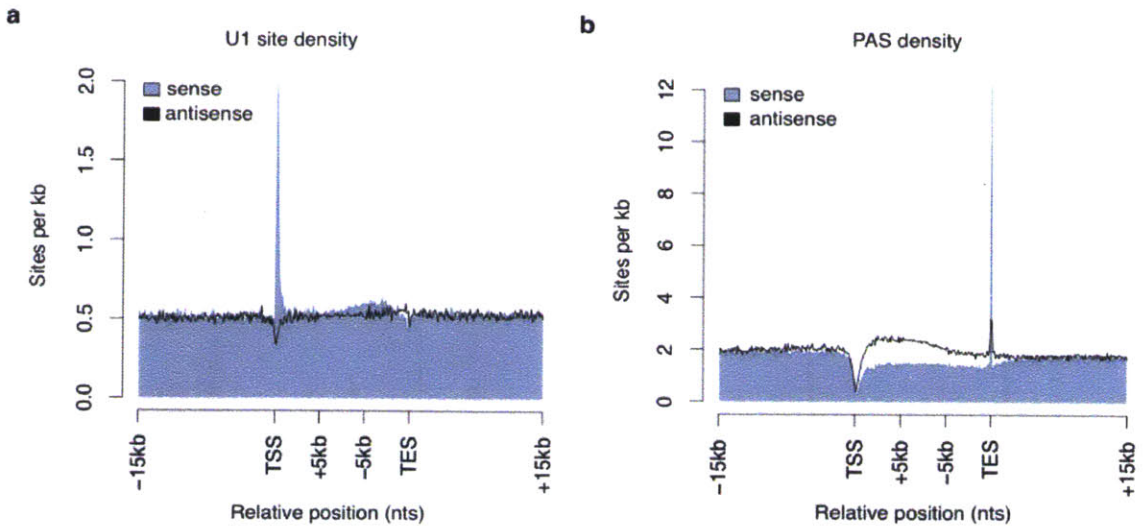
Supplementary Figure 6. Binding profiles of ten 3' end processing factors around cleavage sites in uaRNA regions and mRNA ends. A cleavage site is defined as a uaRNA cleavage site if it is outside any protein coding gene but locates within 5 kb upstream antisense of a protein-coding gene TSS. mRNA cleavage sites are defined as cleavage sites within 100 bases of annotated protein-coding gene ends. For each 3' end processing factor, CLIP read density within 200 bases of all cleavage sites are summed up in every 5 bp bin and subsequently normalized such that the max value is 1.



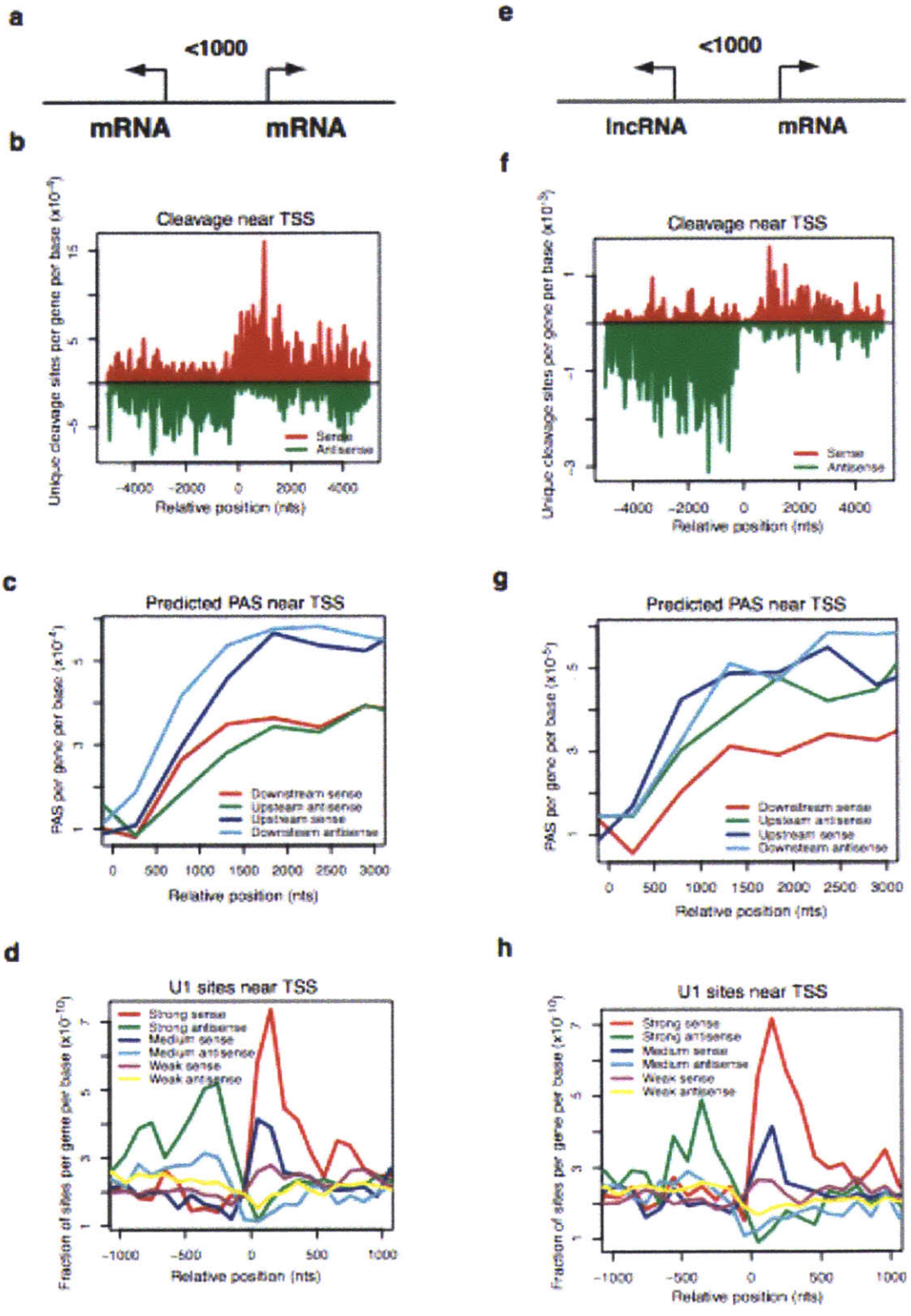
Supplementary Figure 7. Proximal U1 sites are associated with uaRNA length. (a) Distribu- tion of cleavage sites flanking strong U1 sites (position 0). Cleavage sites are classified as rare, medium, and frequent sites based on the number of reads supporting each cleavage site (rare: 1 read, medium: 2-9 reads, frequent: >9 reads). Y-axis is shown as the fraction of sites per gene per base. (b) CDF plot comparing the length of uaRNAs grouped by the presence/absence of promoter proximal PAS and U1 sites. PAS+/- (U1+/-) indicates the presence/absence of PAS or U1 sites in the first 1 kb of uaRNA region. The length of uaRNAs is estimated using the distance from cleav- age sites to coding gene TSS.



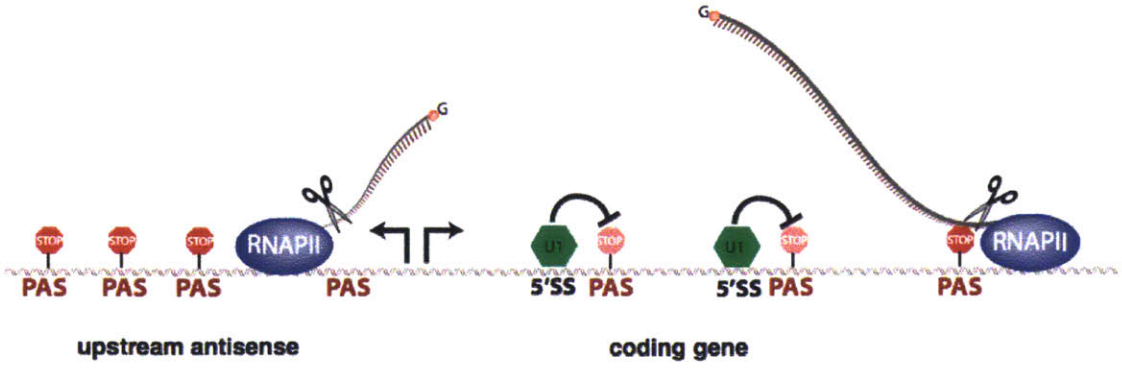
Supplementary Figure 8. Cleavage site simulation near coding gene TSS. Plots were generated in the same way as Figure 2d with unique simulated cleavage sites being plotted. Above each simulation plot, U1 protection refers to the zone of protection in nucleotides downstream (+) conferred by a strong U1 site. Metagene plot of simulated cleavage events considering the PAS (AATAAA) alone (a), or parameters where a PAS is protected if it contains a strong U1 site at least 500 (b), 1500 (c), or 2000 (d) nts upstream. These data demonstrate that the cleavage bias from the simulation is robust when considering protection zones of various sizes.



Supplementary Figure 9. Density of U1 and PAS signals at coding genes and intergenic regions. The density of strong U1 sites (a) and AAUAAA polyadenylation signals (b) in sites per kb for protein-coding genes longer than 15 kb and flanking 15 kb of intergenic sequences. U1 or PAS signals located on sense or antisense regions are depicted in purple and black, respectively. In addition to the strong U1 enrichment in the proximal sense direction of the gene, we observe a modest increase in the frequency of strong U1 signals internal to genes. We also observed a strong strand bias of PAS in coding transcription units, both exon and intron sequences, as compared to intergenic regions. Specifically, PAS are depleted on the sense strand when compared to the antisense strand throughout coding genes prior to the TES. In absolute terms, the genome back- ground has a relatively high density of PAS (~ 2 sites per kb on average) but lower density of strong U1 sites (~0.5 sites per kb on average). Together, the observed distributional patterns support a general model of a U1-PAS axis favoring elongation to produce long transcripts such as precursors to mRNA but limiting transcription from antisense and intergenic regions.


a
<1000
mRNA
mRNA
b
Cleavage near TSS
Unique cleavage sites per gene per base (x10^-3)
Sense
Antisense
Relative position (nts)
c
Predicted PAS near TSS
PAS per gene per base (x10^-4)
Downstream sense
Upstream antisense
Upstream sense
Downstream antisense
Relative position (nts)
d
U1 sites near TSS
Fraction of sites per gene per base (x10^-10)
Strong sense
Strong antisense
Medium sense
Medium antisense
Weak sense
Weak antisense
Relative position (nts)
e
<1000
lncRNA
mRNA
f
Cleavage near TSS
Unique cleavage sites per gene per base (x10^-3)
Sense
Antisense
Relative position (nts)
g
Predicted PAS near TSS
PAS per gene per base (x10^-5)
Downstream sense
Upstream antisense
Upstream sense
Downstream antisense
Relative position (nts)
h
U1 sites near TSS
Fraction of sites per gene per base (x10^-10)
Strong sense
Strong antisense
Medium sense
Medium antisense
Weak sense
Weak antisense
Relative position (nts)

Supplementary Figure 10. U1-PAS axis at mRNA:mRNA and lncRNA:mRNA gene pairs. 1047 and 629 mRNA:mRNA and lncRNA:mRNA gene pairs, respectively, were analyzed similarly as in Fig 1a, Fig. 2a, and Fig. 2c, except that larger bins were used (500 bps bin for PAS and 100 bps bin for U1) to smooth the curve due to the low number of genes used to make the plot. For mRNA:mRNA gene pairs position zero represents the TSS of all genes on the + strand. For lncRNA:mRNA gene pairs position zero represents the TSS of the coding gene.



Supplementary Figure 11. Illustration of the U1-PAS axis for divergent non-coding RNA control. At divergent promoters, RNAPII (depicted as a purple oval) transcribes in both downstream sense and upstream antisense directions, yet upstream antisense RNAs are frequently terminated shortly after initiation due to the high density of PAS (red stop sign) and a lack of strong U1 signals to suppress these sites. In contrast, PAS signals are low in the downstream sense direction and are generally protected by the binding of U1 snRNP (green hexagon) to a nearby 5' splice site denoted as 5'SS in black. A pink stop sign denotes a protected PAS. The U1-PAS axis may function to promote continued elongation throughout the gene and to ensure transcription is suppressed outside protein-coding genes.

References

Adelman, K., and Lis, J.T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nat. Rev. Genet. *13*, 720–731.

Andersen, P.K., Lykke-Andersen, S.S., and Jensen, T.H. (2012). Promoter-proximal polyadenylation sites reduce transcription activity. Genes Dev. *26*, 2169–2179.

Arigo, J.T., Eyler, D.E., Carroll, K.L., and Corden, J.L. (2006). Termination of cryptic unstable transcripts is directed by yeast RNA-binding proteins Nrd1 and Nab3. Mol. Cell *23*, 841–851.

Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J.M., and Gautheret, D. (2000). Patterns of variant polyadenylation signal usage in human genes. Genome Res. *10*, 1001–1010.

Berg, M.G., Singh, L.N., Younis, I., Liu, Q., Pinto, A.M., Kaida, D., Zhang, Z., Cho, S., Sherrill-Mix, S., Wan, L., et al. (2012). U1 snRNP Determines mRNA Length and Regulates Isoform Expression. Cell *150*, 53–64.

Connelly, S., and Manley, J.L. (1989). A CCAAT box sequence in the adenovirus major late promoter functions as part of an RNA polymerase II termination signal. Cell *57*, 561–571.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. Science (80-.). *322*, 1845–1848.

Derti, A., Garrett-Engele, P., Macisaac, K.D., Stevens, R.C., Sriram, S., Chen, R., Rohl, C.A., Johnson, J.M., and Babak, T. (2012). A quantitative atlas of polyadenylation in five mammals. Genome Res. *22*, 1173–1183.

Djebali, S., Davis, C.A., 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.

Flynn, R.A., Almada, A.E., Zamudio, J.R., and Sharp, P.A. (2011). Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. Proc. Natl. Acad. Sci. U. S. A. *108*, 10460–10465.

Gil, A., and Proudfoot, N.J. (1987). Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit beta-globin mRNA 3' end formation. Cell *49*, 399–406.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018.

Hu, J., Lutz, C.S., Wilusz, J., and Tian, B. (2005). Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. RNA *11*, 1485–1493.

Kaida, D., Berg, M.G., Younis, I., Kasim, M., Singh, L.N., Wan, L., and Dreyfuss, G. (2010). U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature *468*, 664–U81.

LaCava, J., Houseley, J., Saveanu, C., Petfalski, E., Thompson, E., Jacquier, A., and Tollervey, D. (2005). RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. Cell *121*, 713–724.

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.

MacDonald, C.C., Wilusz, J., and Shenk, T. (1994). The 64-kilodalton subunit of the CstF polyadenylation factor binds to pre-mRNAs downstream of the cleavage site and influences cleavage site location. Mol. Cell. Biol. *14*, 6647–6654.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 10–12.

Martin, G., Gruber, A.R., Keller, W., and Zavolan, M. (2012). Genome-wide Analysis of Pre-mRNA 3 ' End Processing Reveals a Decisive Role of Human Cleavage Factor I in the Regulation of 3 ' UTR Length. Cell Rep. *1*, 753–763.

Min, I.M., Waterfall, J.J., Core, L.J., Munroe, R.J., Schimenti, J., and Lis, J.T. (2011). Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. Genes Dev. *25*, 742–754.

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., and Jensen, T.H. (2008). RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters. Science (80-.). *322*, 1851–1854.

Preker, P., Almvig, K., Christensen, M.S., Valen, E., Mapendano, C.K., Sandelin, A., and Jensen, T.H. (2011). PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. Nucleic Acids Res. *39*, 7179–7193.

Proudfoot, N.J. (2011). Ending the message: poly(A) signals then and now. Genes Dev. *25*, 1770–1782.

Rahl, P.B., Lin, C.Y., Seila, A.C., Flynn, R.A., McCuine, S., Burge, C.B., Sharp, P.A., and Young, R.A. (2010). c-Myc regulates transcriptional pause release. Cell *141*, 432–445.

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008). Divergent Transcription from Active Promoters. Science (80-.). *322*, 1849–1851.

Sigova, A.A., Mullen, A.C., Molinie, B., Gupta, S., Orlando, D.A., Guenther, M.G., Almada, A.E., Lin, C., Sharp, P.A., Giallourakis, C.C., et al. (2013). Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proc. Natl. Acad. Sci. U. S. A. *110*, 2876–2881.

Spies, N., Burge, C.B., and Bartel, D.P. (2013). 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. Genome Res. *23*, 2078–2090.

Tian, B., Hu, J., Zhang, H., and Lutz, C.S. (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res. *33*, 201–212.

Vanácová, S., Wolf, J., Martin, G., Blank, D., Dettwiler, S., Friedlein, A., Langen, H., Keith, G., and Keller, W. (2005). A new yeast poly(A) polymerase complex involved in RNA quality control. PLoS Biol. *3*, e189.

Wyers, F., Rougemaille, M., Badis, G., Rousselle, J.-C., Dufour, M.-E., Boulay, J., Régnault, B., Devaux, F., Namane, A., Séraphin, B., et al. (2005). Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. Cell *121*, 725–737.

Xie, C., Zhang, Y.E., Chen, J.-Y., Liu, C.-J., Zhou, W.-Z., Li, Y., Zhang, M., Zhang, R., Wei, L., and Li, C.-Y. (2012). Hominoid-Specific De Novo Protein-Coding Genes Originating from Long Non-Coding RNAs. Plos Genet. *8*, e1002942.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. *11*, 377–394.

Zhang, L., Ding, Q., Wang, P., and Wang, Z. (2013). An upstream promoter element blocks the reverse transcription of the mouse insulin-degrading enzyme gene. Biochem. Biophys. Res. Commun. *430*, 26–31.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zhang, Y.E., Vibranovski, M.D., Landback, P., Marais, G.A.B., and Long, M.Y. (2010). Chromosomal Redistribution of Male-Biased Genes in Mammalian Evolution with Two Bursts of Gene Gain on the X Chromosome. Plos Biol. *8*.

Chapter 3: Shaping genome's evolution through noncoding transcription

In this chapter, I present a model suggesting noncoding transcription and especially promoter divergent transcription could drive new gene origination and rearrange the genome.

This work was published as:

Xuebing Wu, Phillip A. Sharp, Divergent transcription: a driving force for new gene origination? ***Cell***, 2013, 155:990-996

The mammalian genome is extensively transcribed, a large fraction of which is divergent transcription from promoters and enhancers that is tightly coupled with active gene transcription. Here we propose that divergent transcription may shape the evolution of the genome by new gene origination.

Widespread divergent transcription

The vast majority of the human genome, including half of the region outside known genes, is transcribed (Djebali et al., 2012). However, most intergenic transcription activity produces short and unstable noncoding transcripts whose abundances are usually an order of magnitude lower than those from typical protein-coding genes. Except for a few well-studied cases (see review in (Guttman and Rinn, 2012; Lee, 2012; Mercer et al., 2009; Ponting et al., 2009; Rinn and Chang, 2012; Ulitsky and Bartel, 2013; Wang and Chang, 2011; Wei et al., 2011; Wilusz et al., 2009), it's unclear whether most intergenic transcription is regulated or has cellular function.

Recent evidence has shown that most intergenic transcription occurs near or is associated with gene transcription, such as transcription from promoter and enhancer regions (Sigova et al., 2013). The majority of mammalian promoters direct transcription initiation on both sides with opposite orientations, a phenomenon known as divergent transcription (Core et al., 2008; Preker et al., 2008; Seila et al., 2008). Divergent transcription generates upstream antisense RNAs (uaRNAs, or PROMPTs, promoter upstream transcripts) near the 5' end of genes that are typically short (50-2,000 nucleotides) and relatively unstable (Flynn et al., 2011; Ntini et al., 2013; Preker et al., 2008, 2011). Similar divergent transcription also occurs at distal

enhancer regions, giving rise to RNAs termed enhancer RNAs (eRNAs) (Kim et al., 2010; De Santa et al., 2010). In mouse and human embryonic stem (ES) cells most long noncoding RNAs (lncRNAs, longer than 100 nucleotides) are associated with protein-coding genes, including ~50% as uaRNAs and ~20% as eRNAs (Sigova et al., 2013). These observations suggest that divergent transcription from promoters and enhancers of protein-coding genes is the major source of intergenic transcription in ES cells.

In the textbook model of a eukaryotic promoter, the directionality is set by the arrangement of an upstream cis-element region followed by a core promoter (Fig 1A). The cis-elements are bound by sequence-specific transcription factors whereas the core promoter is bound by TATA-binding protein (TBP) and other factors that recruit the core transcription machinery. Most mammalian promoters lack a TATA element (TATA-less) and are CpG rich (Sandelin et al., 2007). For these promoters, TBP is recruited through sequence specific transcription factors such as Sp1 that bind CpG rich sequences and components of the TFIID complex that have little sequence specificity. Thus, in the absence of strong TATA elements such as for CpG island promoters, TBP-complexes are recruited on both sides of the transcription factors to form pre-initiation complexes in both orientations (Fig 1B). This model is supported by the observation that divergent transcription occurs at most promoters that are associated with CpG islands in mammals, whereas promoters with TATA elements in mammals and worm are associated with unidirectional transcription (Core et al., 2008; Kruesi et al., 2013). In addition, divergent transcription is less common in Drosophila where CpG islands are rare (Core et al., 2012). Since tran-

scription factors with chromatin remodeling potential and transcription activation domains also bind at enhancer sites, it is not surprising that these are also sites of divergent transcription. In fact, promoters and enhancers have many properties in common, and it has been shown recently that many intragenic enhancers can act as alternative promoters producing tissue-specific lncRNAs (Kowalczyk et al., 2012).

The U1-PAS axis and gene maturation

Promoter-proximal noncoding transcription in both yeast and mammals has been shown to be suppressed at the chromatin level, including nucleosome remodeling (Whitehouse et al., 2007), histone deacetylation (Churchman and Weissman, 2011), and gene loop formation (Tan-Wong et al., 2012). We and others recently found that in mammals promoter upstream antisense transcription is frequently terminated due to cleavage of the nascent RNA by the same process responsible for the generation of the poly A tract at the 3' ends of genes (Almada et al., 2013; Ntini et al., 2013). In both cases, the primary signal directing this process is the poly (A) signal (PAS) motif, AAUAAA or similar (Proudfoot, 2011). Pol II terminates transcription within several kb after such cleavage (Anamika et al., 2012; Richard and Manley, 2009). Computational analysis showed that relative to the 5' end of the sense regions, PAS motifs are enriched whereas potential U1 snRNP binding sites, or 5' splice site-like sequences, are depleted in the upstream antisense regions. The binding of U1 snRNP is known to suppress PAS directed cleavage over regions of thousand nucleotides downstream (Berg et al., 2012; Kaida et al., 2010). Thus, the bias in the distribution of U1 snRNP binding sites and PAS promotes expression of full-length mRNAs by

suppressing premature cleavage and polyadenylation but favors early termination of uaRNAs. This conclusion is strongly supported by the finding that inhibition of U1 snRNP dramatically increased termination and polyadenylation of sense-oriented transcripts in the gene region (Almada et al., 2013).

If the U1-PAS axis defines the length of a transcribed region, then it might be expected that for a typical protein-coding gene (~20 kb) to evolve from intergenic noncoding DNA would involve strengthening of the U1-PAS axis by gaining U1 sites and losing PAS in the sense orientation. Examining the distributions of U1 and PAS sites in bidirectional promoters involving UCSC-annotated mRNA-mRNA, mRNA-lncRNA, and mRNA-uaRNA pairs, we found that lncRNAs showed properties resembling intermediates between mRNA genes and uaRNA regions in terms of the density of U1 sites and PAS sites (Almada et al., 2013). That is, the density of PAS decreases from regions producing uaRNA to lncRNA to mRNA, whereas U1 sites show the opposite trend, consistent with the differences in the length and abundance of these transcripts. We also studied the evolution of the U1-PAS axis in vertebrates, and found that older genes exhibit progressive gain of U1 sites and loss of PAS sites at their 5' ends. Together these observations suggest that strengthening of the U1-PAS axis may be associated with the origination and maturation of genes.

De novo gene origination from divergent transcription

Below we propose a model (Fig 2) arguing that the act of transcription in germ cells strengthens the U1-PAS axis in the upstream antisense region of an active gene, or the associated enhancer regions, creating a feedback loop amplifying transcription

activity, which eventually may drive origination of a new antisense-oriented gene (Fig 3).

One consequence of transcription is that it can cause mutations, especially on the coding (non-transcribed) strand. During transcription, transient R-loops can be formed behind the transcribing RNA polymerase II, exposing the coding strand as single-stranded DNA whereas the non-coding strand is base-paired with and thus protected by the nascent RNA (Aguilera and García-Muse, 2012). The lack of splicing signals in the divergent transcript also makes it more vulnerable to R-loop formation, as splicing factors have been implicated in suppressing R-loop formation (Li and Manley, 2006, 2005; Paulsen et al., 2009). In addition, divergent transcription generates negative supercoiling at promoters which facilitates DNA unwinding and promotes R-loop formation (Aguilera and García-Muse, 2012; Seila et al., 2009). As a consequence of R-loop formation, the single-stranded coding strand is vulnerable to mutagenic processes, such as cleavage, deamination, and depurination. Genomics studies have shown that during mammalian evolution, transcribed regions accumulate G and T bases on the coding strand, relative to the non-coding strand or non-transcribed regions (Green et al., 2003; Mugal et al., 2009; Park et al., 2012; Polak et al., 2010). Evidence suggests that such strand bias may result from passive effects of deamination, transcription-coupled repair, and somatic hypermutation pathways in germ cell-transcribed genes, in the absence of selection (Green et al., 2003; McVicker and Green, 2010; Polak and Arndt, 2008).

Accumulation of G and T content on the coding strand will strengthen the U1-PAS axis (Fig. 2). A-rich sequences such PAS (AATAAA) is likely to be lost when the

genomic DNA accumulates G and T. In contrast, G+T rich sequences, such as U1 snRNP binding sites (e.g., resembling 5' splice sites, G|GTAAGT and G|GTGAGT), are likely to emerge in these regions. Since promoter-proximal PAS reduces transcriptional activity (Andersen et al., 2012), the loss of PAS and gain of U1 sites should contribute to lengthening of the transcribed region as well as its more robust transcription. The gain of U1 sites could also enhance transcription by recruiting basal transcription initiation factors (Damgaard et al., 2008; Furger et al., 2002; Kwek et al., 2002) or elongation factors (Fong and Zhou, 2001). Therefore a positive feedback loop is formed: active transcription causes the coding strand to accumulate sequence changes favoring higher transcription activity.

As noted above, strengthening of the U1-PAS axis also favors extension of the transcribed region. Being longer gives the transcript several advantages: by chance longer RNAs are more likely to contain additional splicing signals such as a 3' splice site to become spliced, or binding sites for splicing-independent nuclear export factors, thus escaping nuclear exosome degradation by packaging and exporting to cytoplasm (Nott et al., 2003; Singh et al., 2012). Longer RNAs are also more likely to carry an open reading frame, either generated de novo or by incorporation of gene remnants.

Once in the cytoplasm, the RNA should at some frequency be translated into short polypeptides due to widespread translational activity (Carvunis et al., 2012). Some of the polypeptides may provide advantage to the organism and become fixed in the population, thereby forming a new gene.

Accelerating other new gene origination processes

In addition to *de novo* gene origination, the model described above also facilitates new gene origination via other mechanisms in regions of divergent transcription. Tandem duplication, retroposition, and recombination of existing genes or gene fragments are the major mechanisms for new gene origination (Chen et al., 2013; Long et al., 2013). Most duplicated genes or gene fragments are silenced due to the lack of required elements such as a promoter. In contrast, genes or gene fragments inserted into regions of divergent transcription, such as upstream of a promoter or flanking an enhancer, will be transcribed, likely under different regulation than prior to their insertion, and thus could evolve to carry out functions different than the original gene. In support of this, a recent survey of human and mouse genes evolved from "domesticated" transposons (Kalitsis and Saffery, 2009) showed that a significant proportion of them are located in bidirectional promoters. Promoter upstream regions also preferentially accumulate transposable elements, which can carry 5' splice site sequences that may accelerate the process of new gene origination (Gotea et al., 2013).

New gene origination from enhancers

Similar to promoters, enhancers are also divergently transcribed, and as a result, new genes might originate at enhancer regions through the same mechanism described above. The possibility of enhancer derived new genes has not been previously discussed. Manual inspection of a list of 24 hominoid-specific *de novo* protein-coding genes (Xie et al., 2012) revealed that *MYEOV* (myeloma overexpressed), a

gene implicated in various types of cancer (Janssen et al., 2000, 2002; Leyden et al., 2006; Moss et al., 2006), is likely derived from an intergenic enhancer in mouse. The mouse syntenic region of *MYEOV* is within a 5 kb region about 100 kb away from any gene, but covered by intensive H3K4me1 marks, diagnostic of an enhancer, and positive for Mediator binding in mouse ES cells, as well as nascent transcription signals (GRO-seq) indicating divergent transcription, all indicating this region is an active enhancer in mouse ES cells. Further analysis is needed to firmly establish the role of enhancer transcription in the origination of the *MYEOV* gene. For example, it will be interesting to examine the evolutionary dynamics of the spatial and functional relationship between the enhancer/*MYEOV* locus and the corresponding target gene.

Predictions and supporting evidence

A recent comparative analysis of human-mouse gene annotations detected over a thousand lncRNAs annotated in the upstream antisense region of human genes whereas lncRNAs divergent from the corresponding mouse protein-coding genes could not be detected (Gotea et al., 2013). This observation suggests that promoter divergent transcription could be capable of generating large number of primate-specific transcripts. Another study (Xie et al., 2012) identified 24 hominoid-specific *de novo* protein-coding genes in human, five of which derive from bidirectional promoters ($P < 0.01$, compared to shuffled gene positions), confirming promoter divergent transcription as an important source of *de novo* gene origination, and enhancer transcription may drive other new genes, as noted above.

An important feature of genes originated in the proposed model is that both the new gene and the ancestral gene are likely to be expressed in germ cells. This is because for the transcription-induced G and T bias to accumulate and spread in a population, these mutations should occur in germ cells. A prediction of the model is that new genes are preferentially expressed in germ cells, or tissues with high fraction of germ cells. Consistent with this, previous reports showed that lineage-specific genes in human, fly, and zebrafish genomes are preferentially expressed in reproductive organs or tissues, such as testis (Clark et al., 2007; Levine et al., 2006; Tay et al., 2009; Yang et al., 2013). Moreover, divergent gene pairs in the human genome are enriched for housekeeping genes, such as DNA repair and DNA replication genes (Adachi and Lieber, 2002) that are actively transcribed in germ cells. In addition, the strand bias of G and T content correlates with germ cell but not somatic tissue gene expression levels (Majewski, 2003).

The model could explain the origin of divergent protein-coding gene pairs separated by less than 1 kb (usually less), which account for 10% of human protein-coding genes (Adachi and Lieber, 2002; Li et al., 2006; Piontkivska et al., 2009; Trinklein et al., 2004; Wakano et al., 2012; Xu et al., 2012), far higher proportion than would be expected if genes were randomly distributed in the genome. The model proposed here provides a natural explanation for the evolutionary origin of these gene pairs. It is likely that many more genes originated from divergent transcription, with the bidirectional organization having been disrupted by transposon insertion, recombination, or other genome rearrangement events. The model also predicts that divergent gene pairs commonly have unrelated functions, although

they frequently might share co-expression. Except for a few cases, such as histone gene pairs and collagen gene pairs that are likely results of tandem duplication, the majority of divergent gene pairs in the human genome do not share higher functional similarity compared to random gene pairs (Li et al., 2006; Xu et al., 2012). For example, 35 of the 105 annotated DNA repair genes have bidirectional promoters, making DNA repair the most over-represented pathway for genes involved in bidirectional promoters, yet all 35 DNA repair genes are paired with non-DNA repair genes (Xu et al., 2012). Similarly, genes coding subunits of protein complexes are enriched in bidirectional pairs in human, yet none of these pairs code for two subunits of the same complex (Li et al., 2006). A similar observation has been reported for yeast and is consistent with the argument that the bidirectional conformation reduces expression noise and is not strongly selected for share functionality (Wang et al., 2011). The lack of functional relatedness is also illustrated by the parallel evolution of bidirectional promoters of *RecQ* helicases (Piontkivska et al., 2009). The five *RecQ* paralogs were duplicated early during metazoan evolution, yet all evolve to have divergent partners in human. However, these partner genes showed no functional or sequence similarity with each other (Piontkivska et al., 2009), suggesting parallel and independent origination of new genes from all five promoters.

Impact on genome organization and evolution

Divergent transcription likely facilitates the rearrangement events that reshape the genome, and also introduces unique features into genome organization, including

the sharing of promoters, physical linkage in three-dimensional space, and co-expression of distal genes.

Although vertebrates share most of their genes, the genomic position and orientation of specific genes differ significantly due to genome rearrangement events, such as translocation, recombination, and duplication followed by the loss of the original copy. The survival of the gene or gene fragments at the new position can be facilitated by divergent transcription as discussed above. The role of divergent transcription in preserving the function of the new gene copy is likely significant, given that translocation preferentially occurs near active promoters (Chiarle et al., 2011; Klein et al., 2011). The correlation between transcription and translocation could potentially increase the chance that the translocated gene is still expressed and thus functional, therefore reducing the cost of translocation. For example, although ~40% of human protein coding genes can be traced back to fish, fewer than 7% (83/1262) of human bidirectional gene pairs are also bidirectional in the fish genome (Li et al., 2006), suggesting that most human bidirectional gene pairs formed with young genes, or by bringing together old genes through translocation facilitated by divergent transcription.

In addition to bidirectional organization, spatial and functional coupling between distal gene pairs would be introduced through new gene origination from enhancer transcription. Due to the tight coupling between gene transcription and enhancer transcription, an enhancer derived new gene will share a significant co-expression pattern with the old gene, despite the distance in the linear genome. Such coupled transcription of distal gene pairs brought together by chromatin inter-

actions could contribute to the formation of transcription factories, nuclear foci where multiple genes are transcribed together without the requirement of shared function (Edelman and Fraser, 2012; Sutherland and Bickmore, 2009). The existence of transcription factories has been supported by increasing evidence, including *in vivo* live imaging (Ghamari et al., 2013) and chromatin interaction mapping (Li et al., 2012). These are probably related to super-enhancers where many genes that are coordinately expressed are associated with a common enhancer region (Lovén et al., 2013; Whyte et al., 2013). Overlaying comparative genomics analysis onto high-throughput chromatin interaction mapping data across multiple species (Dixon et al., 2012; Li et al., 2012) may help to reveal the evolutionary origin of transcription factories.

Conclusions

In conclusion, we propose that divergent transcription at promoters and enhancers results in changes of the transcribed DNA sequences that over evolutionary time drive new gene origination in the transcribed regions. Although the models proposed here are consistent with significant available data, systematic tests of these models await further advances such as in-depth characterization of additional genomes and experiments designed to test specific hypothesis. Over evolutionary times, genes formed through divergent transcription can be shuffled to other locations losing their evolutionary context. We envision future studies will uncover more functional surprises from divergent transcription, and illuminate how intergenic transcription is integrated into the cellular transcriptome.

Figures

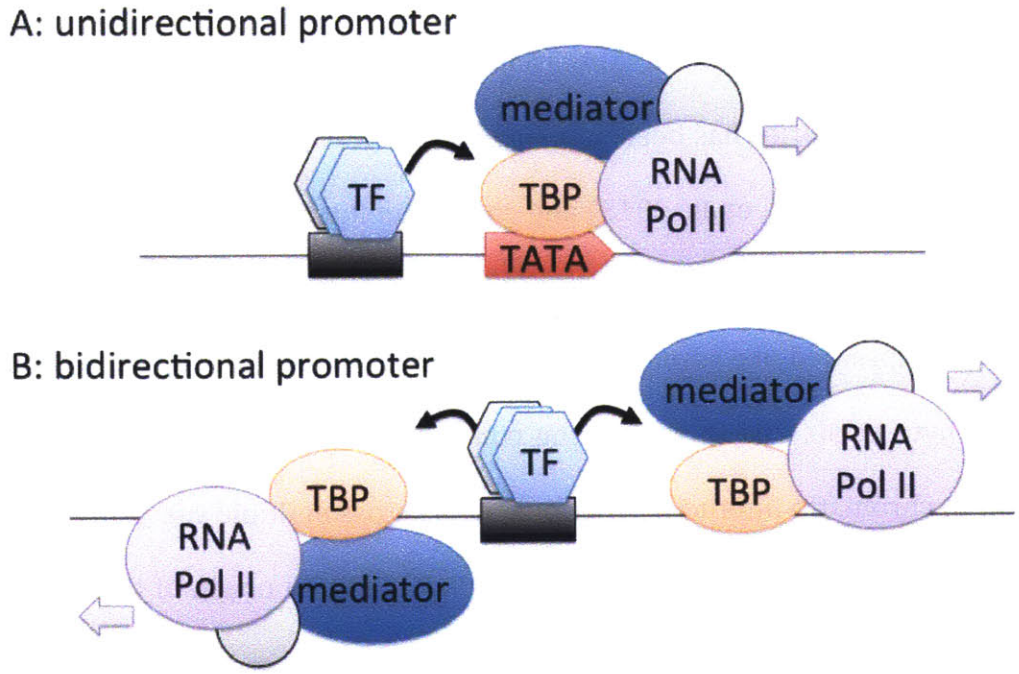


Figure 1: Transcription factors drive divergent transcription. A) Transcription factor (TF) binding helps to recruit TATA-binding protein (TBP) and associated factors, which binds the directional TATA element in the DNA and orientates RNA Pol II to transcribe downstream DNA. B) In the absence of strong TATA elements common of CpG island promoters, TF-recruited TBP and associated factors binds to low specificity sequences and forms initiation complexes at similar frequencies in both directions.

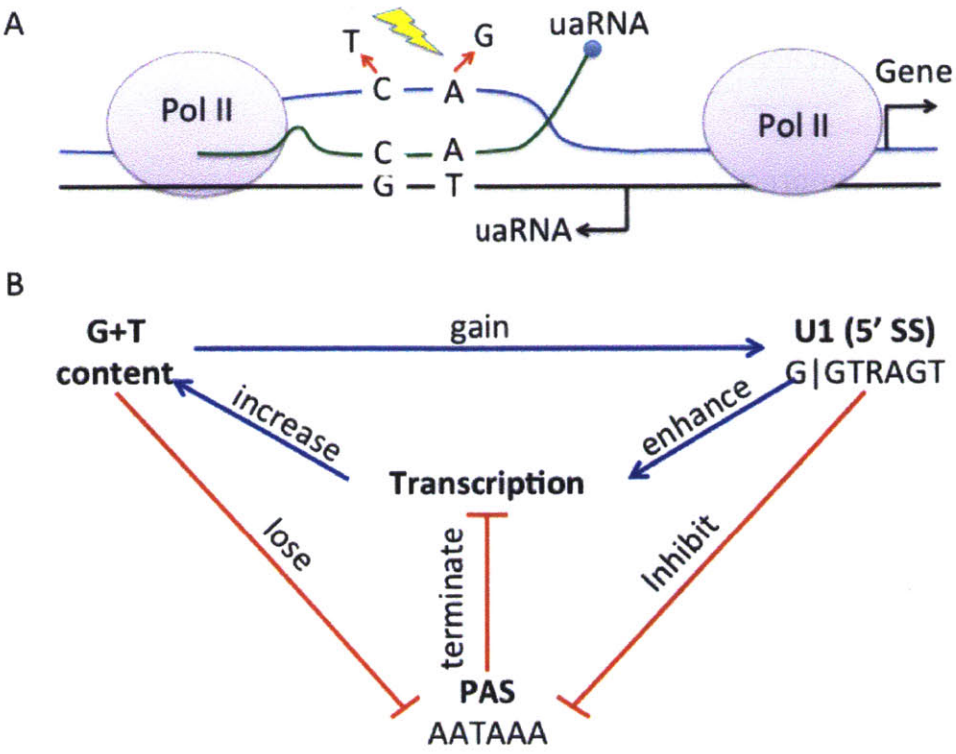


Figure 2: Feedback loops between transcription, U1, and PAS signals. (A) Germ cell transcription exposes the coding strand (non-template, which has the same sequence as the RNA) single-stranded and vulnerable to mutations towards G and T bases, (B) which increases the chance of gaining GT-rich sequences such as U1 binding site (5' splice site (5' SS)) and also increases the chance of losing A-rich sequences such as PAS, which terminates transcription. U1 binding can enhance transcription through promoting transcription initiation and reinitiation, and also inhibiting the usage of nearby PAS.

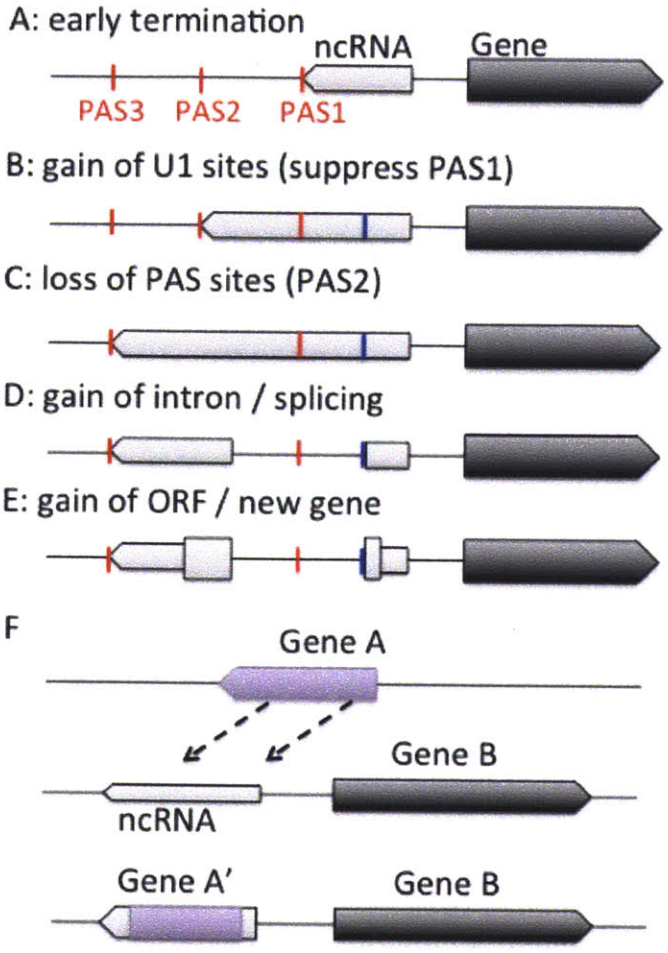


Figure 3: Divergent transcription drives new gene origination. A-E) *De novo* protein-coding gene origination, and F) gene duplication or translocation. A) Divergent transcription of a gene (right dark block) generates divergent noncoding RNA (ncRNA) in the upstream antisense direction, which is terminated by PAS-dependent mechanism (PAS: red bars). B) Transcription increases G and T frequency on the coding strand, thus increases the chance of encoding a U1 site (blue bar) which suppress a downstream PAS (PAS1), favoring the usage of a downstream PAS (PAS2). C) Increase in G+T content also increases the chance of losing PAS sites (PAS2) which activates a further downstream site (PAS3) and extends the transcribed region. D) The longer transcript acquires splicing signals, which makes it more stable and exported to the cytoplasm. E) The longer transcript encodes a short ORF and the resulting short peptide is selected and fixed in the population and becomes a new protein-coding gene. F) Gene A is translocated or duplicated into the promoter upstream antisense region of gene B, and evolves into a new gene A'. Thin and thick blocks represent transcribed noncoding and coding regions, respectively.

References

Adachi, N., and Lieber, M.R. (2002). Bidirectional gene organization: A common architectural feature of the human genome. Cell *109*, 807–809.

Aguilera, A., and García-Muse, T. (2012). R loops: from transcription byproducts to threats to genome stability. Molecular Cell *46*, 115–124.

Almada, A.E., Wu, X., Kriz, A.J., Burge, C.B., and Sharp, P.A. (2013). Promoter directionality is controlled by U1 snRNP and polyadenylation signals. Nature *499*, 360–363.

Anamika, K., Gyenis, À., Poidevin, L., Poch, O., and Tora, L. (2012). RNA polymerase II pausing downstream of core histone genes is different from genes producing polyadenylated transcripts. PloS One *7*, e38769.

Andersen, P.K., Lykke-Andersen, S., and Jensen, T.H. (2012). Promoter-proximal polyadenylation sites reduce transcription activity. Genes & Development *26*, 2169–2179.

Berg, M.G., Singh, L.N., Younis, I., Liu, Q., Pinto, A.M., Kaida, D., Zhang, Z., Cho, S., Sherrill-Mix, S., Wan, L., et al. (2012). U1 snRNP Determines mRNA Length and Regulates Isoform Expression. Cell *150*, 53–64.

Carvunis, A.R., Rolland, T., Wapinski, I., Calderwood, M.A., Yildirim, M.A., Simonis, N., Charloteaux, B., Hidalgo, C.A., Barbette, J., Santhanam, B., et al. (2012). Proto-genes and de novo gene birth. Nature *487*, 370–374.

Chen, S., Krinsky, B.H., and Long, M. (2013). New genes as drivers of phenotypic evolution. Nature Reviews. Genetics *14*, 645–660.

Chiarle, R., Zhang, Y., Frock, R.L., Lewis, S.M., Molinie, B., Ho, Y.-J., Myers, D.R., Choi, V.W., Compagno, M., Malkin, D.J., et al. (2011). Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. Cell *147*, 107–119.

Churchman, L.S., and Weissman, J.S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature *469*, 368–373.

Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., Kaufman, T.C., Kellis, M., Gelbart, W., Iyer, V.N., et al. (2007). Evolution of genes and genomes on the Drosophila phylogeny. Nature *450*, 203–218.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. Science *322*, 1845–1848.

Core, L.J., Waterfall, J.J., Gilchrist, D.A., Fargo, D.C., Kwak, H., Adelman, K., and Lis, J.T. (2012). Defining the status of RNA polymerase at promoters. Cell Reports *2*, 1025–1035.

Damgaard, C.K., Kahns, S., Lykke-Andersen, S., Nielsen, A.L., Jensen, T.H., and Kjems, J. (2008). A 5 ' splice site enhances the recruitment of basal transcription initiation factors in vivo. Molecular Cell *29*, 271–278.

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

Djebali, S., Davis, C.A., 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.

Edelman, L.B., and Fraser, P. (2012). Transcription factories: genetic programming in three dimensions. Current Opinion in Genetics & Development *22*, 110–114.

Flynn, R.A., Almada, A.E., Zamudio, J.R., and Sharp, P.A. (2011). Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. Proceedings of the National Academy of Sciences of the United States of America *108*, 10460–10465.

Fong, Y.W., and Zhou, Q. (2001). Stimulatory effect of splicing factors on transcriptional elongation. Nature *414*, 929–933.

Furger, A., O'Sullivan, J.M., Binnie, A., Lee, B.A., and Proudfoot, N.J. (2002). Promoter proximal splice sites enhance transcription. Genes & Development *16*, 2792–2799.

Ghamari, A., van de Corput, M.P.C., Thongjuea, S., van Cappellen, W.A., van Ijcken, W., van Haren, J., Soler, E., Eick, D., Lenhard, B., and Grosveld, F.G. (2013). In vivo live imaging of RNA polymerase II transcription factories in primary cells. Genes & Development *27*, 767–777.

Gotea, V., Petrykowska, H.M., and Elnitski, L. (2013). Bidirectional Promoters as Important Drivers for the Emergence of Species-Specific Transcripts. Plos One *8*, e57323.

Green, P., Ewing, B., Miller, W., Thomas, P.J., Green, E.D., and Progr, N.C.S. (2003). Transcription-associated mutational asymmetry in mammalian evolution. Nature Genetics *33*, 514–517.

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

Janssen, J.W.G., Vaandrager, J.W., Heuser, T., Jauch, A., Kluin, P.M., Geelen, E., Bergsagel, P.L., Kuehl, W.M., Drexler, H.G., Otsuki, T., et al. (2000). Concurrent activation of a novel putative transforming gene, myeov, and cyclin D1 in a subset of multiple myeloma cell lines with t(11;14)(q13;q32). Blood *95*, 2691–2698.

Janssen, J.W.G., Cuny, M., Orsetti, B., Rodriguez, C., Valles, H., Bartram, C.R., Schuuring, E., and Theillet, C. (2002). MYEOV: A candidate gene for DNA amplification events occurring centromeric to CCNDI in breast cancer. International Journal of Cancer *102*, 608–614.

Kaida, D., Berg, M.G., Younis, I., Kasim, M., Singh, L.N., Wan, L., and Dreyfuss, G. (2010). U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature *468*, 664–U81.

Kalitsis, P., and Saffery, R. (2009). Inherent promoter bidirectionality facilitates maintenance of sequence integrity and transcription of parasitic DNA in mammalian genomes. Bmc Genomics *10*, 498.

Kim, T.-K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182–187.

Klein, I.A., Resch, W., Jankovic, M., Oliveira, T., Yamane, A., Nakahashi, H., Di Virgilio, M., Bothmer, A., Nussenzweig, A., Robbiani, D.F., et al. (2011). Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes. Cell *147*, 95–106.

Kowalczyk, M.S., Hughes, J.R., Garrick, D., Lynch, M.D., Sharpe, J.A., Sloane-Stanley, J.A., McGowan, S.J., De Gobbi, M., Hosseini, M., Vernimmen, D., et al. (2012). Intragenic enhancers act as alternative promoters. Molecular Cell *45*, 447–458.

Kruesi, W.S., Core, L.J., Waters, C.T., Lis, J.T., and Meyer, B.J. (2013). Condensin controls recruitment of RNA polymerase II to achieve nematode X-chromosome dosage compensation. eLife *2*, e00808–e00808.

Kwek, K.Y., Murphy, S., Furger, A., Thomas, B., O'Gorman, W., Kimura, H., Proudfoot, N.J., and Akoulitchev, A. (2002). U1 snRNA associates with TFIIH and regulates transcriptional initiation. Nature Structural Biology *9*, 800–805.

Lee, J.T. (2012). Epigenetic regulation by long noncoding RNAs. Science (New York, N.Y.) *338*, 1435–1439.

Levine, M.T., Jones, C.D., Kern, A.D., Lindfors, H.A., and Begun, D.J. (2006). Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. Proceedings of the National Academy of Sciences of the United States of America *103*, 9935–9939.

Leyden, J., Murray, D., Moss, A., Arumuguma, M., Doyle, E., McEntee, G., O'Keane, C., Doran, P., and MacMathuna, P. (2006). Net1 and Myeov: computationally identified mediators of gastric cancer. British Journal of Cancer *94*, 1204–1212.

Li, X., and Manley, J.L. (2006). Cotranscriptional processes and their influence on genome stability. Genes & Development *20*, 1838–1847.

Li, X.L., and Manley, J.L. (2005). Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability. Cell *122*, 365–378.

Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., 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.

Li, Y.-Y., Yu, H., Guo, Z.-M., Guo, T.-Q., Tu, K., and Li, Y.-X. (2006). Systematic analysis of head-to-head gene organization: Evolutionary conservation and potential biological relevance. Plos Computational Biology *2*, 687–697.

Long, M., Vankuren, N.W., Chen, S., and Vibranovski, M.D. (2013). New Gene Evolution: Little Did We Know. Annual Review of Genetics.

Lovén, J., Hoke, H.A., Lin, C.Y., Lau, A., Orlando, D.A., Vakoc, C.R., Bradner, J.E., Lee, T.I., and Young, R.A. (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell *153*, 320–334.

Majewski, J. (2003). Dependence of mutational asymmetry on gene-expression levels in the human genome. American Journal of Human Genetics *73*, 688–692.

McVicker, G., and Green, P. (2010). Genomic signatures of germline gene expression. Genome Research *20*, 1503–1511.

Mercer, T.R., Dinger, M.E., and Mattick, J.S. (2009). Long non-coding RNAs: insights into functions. Nature Reviews. Genetics *10*, 155–159.

Moss, A.C., Lawlor, G., Murray, D., Tighe, D., Madden, S.F., Mulligan, A.M., Keane, C.O., Brady, H.R., Doran, P.P., and MacMathuna, P. (2006). ETV4 and Myeov knockdown impairs colon cancer cell line proliferation and invasion. Biochemical and Biophysical Research Communications *345*, 216–221.

Mugal, C.F., von Gruenberg, H.-H., and Peifer, M. (2009). Transcription-Induced Mutational Strand Bias and Its Effect on Substitution Rates in Human Genes. Molecular Biology and Evolution *26*, 131–142.

Nott, A., Muslin, S.H., and Moore, M.J. (2003). A quantitative analysis of intron effects on mammalian gene expression. Rna-a Publication of the Rna Society *9*, 607–617.

Ntini, E., Järvelin, A.I., Bornholdt, J., Chen, Y., Boyd, M., Jørgensen, M., Andersson, R., Hoof, I., Schein, A., Andersen, P.R., et al. (2013). Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. Nature Structural & Molecular Biology *20*, 923–928.

Park, C., Qian, W.F., and Zhang, J.Z. (2012). Genomic evidence for elevated mutation rates in highly expressed genes. Embo Reports *13*, 1123–1129.

Paulsen, R.D., Soni, D. V, Wollman, R., Hahn, A.T., Yee, M.-C., Guan, A., Hesley, J.A., Miller, S.C., Cromwell, E.F., Solow-Cordero, D.E., et al. (2009). A Genome-wide siRNA Screen Reveals Diverse Cellular Processes and Pathways that Mediate Genome Stability. Molecular Cell *35*, 228–239.

Piontkivska, H., Yang, M.Q., Larkin, D.M., Lewin, H.A., Reecy, J., and Elnitski, L. (2009). Cross-species mapping of bidirectional promoters enables prediction of unannotated 5 ' UTRs and identification of species-specific transcripts. Bmc Genomics *10*, 189.

Polak, P., and Arndt, P.F. (2008). Transcription induces strand-specific mutations at the 5 ' end of human genes. Genome Research *18*, 1216–1223.

Polak, P., Querfurth, R., and Arndt, P.F. (2010). The evolution of transcription-associated biases of mutations across vertebrates. Bmc Evolutionary Biology *10*, 187.

Ponting, C.P., Oliver, P.L., and Reik, W. (2009). Evolution and functions of long noncoding RNAs. Cell *136*, 629–641.

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., and Jensen, T.H. (2008). RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters. Science *322*, 1851–1854.

Preker, P., Almvig, K., Christensen, M.S., Valen, E., Mapendano, C.K., Sandelin, A., and Jensen, T.H. (2011). PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. Nucleic Acids Research *39*, 7179–7193.

Proudfoot, N.J. (2011). Ending the message: poly(A) signals then and now. Genes & Development *25*, 1770–1782.

Richard, P., and Manley, J.L. (2009). Transcription termination by nuclear RNA polymerases. Genes & Development *23*, 1247–1269.

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

Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y., and Hume, D.A. (2007). Mammalian RNA polymerase II core promoters: insights from genome-wide studies. Nature Reviews. Genetics *8*, 424–436.

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H., Ragoussis, J., Wei, C.-L., and Natoli, G. (2010). A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers. Plos Biology *8*, e1000384.

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008). Divergent Transcription from Active Promoters. Science *322*, 1849–1851.

Seila, A.C., Core, L.J., Lis, J.T., and Sharp, P.A. (2009). Divergent transcription: a new feature of active promoters. Cell Cycle (Georgetown, Tex.) *8*, 2557–2564.

Sigova, A.A., Mullen, A.C., Molinie, B., Gupta, S., Orlando, D.A., Guenther, M.G., Almada, A.E., Lin, C., Sharp, P.A., Giallourakis, C.C., et al. (2013). Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proceedings of the National Academy of Sciences of the United States of America *110*, 2876–2881.

Singh, G., Kucukural, A., Cenik, C., Leszyk, J.D., Shaffer, S.A., Weng, Z., and Moore, M.J. (2012). The Cellular EJC Interactome Reveals Higher-Order mRNP Structure and an EJC-SR Protein Nexus. Cell *151*, 750–764.

Sutherland, H., and Bickmore, W.A. (2009). Transcription factories: gene expression in unions? Nature Reviews Genetics *10*, 457–466.

Tan-Wong, S.M., Zaugg, J.B., Camblong, J., Xu, Z., Zhang, D.W., Mischo, H.E., Ansari, A.Z., Luscombe, N.M., Steinmetz, L.M., and Proudfoot, N.J. (2012). Gene loops enhance transcriptional directionality. Science (New York, N.Y.) *338*, 671–675.

Tay, S.-K., Blythe, J., and Lipovich, L. (2009). Global discovery of primate-specific genes in the human genome. Proceedings of the National Academy of Sciences of the United States of America *106*, 12019–12024.

Trinklein, N.D., Aldred, S.F., Hartman, S.J., Schroeder, D.I., Otillar, R.P., and Myers, R.M. (2004). An abundance of bidirectional promoters in the human genome. Genome Research *14*, 62–66.

Ulitsky, I., and Bartel, D.P. (2013). lincRNAs: Genomics, Evolution, and Mechanisms. Cell *154*, 26–46.

Wakano, C., Byun, J.S., Di, L.-J., and Gardner, K. (2012). The dual lives of bidirectional promoters. Biochimica Et Biophysica Acta-Gene Regulatory Mechanisms *1819*, 688–693.

Wang, K.C., and Chang, H.Y. (2011). Molecular mechanisms of long noncoding RNAs. Molecular Cell *43*, 904–914.

Wang, G.-Z., Lercher, M.J., and Hurst, L.D. (2011). Transcriptional Coupling of Neighboring Genes and Gene Expression Noise: Evidence that Gene Orientation and Noncoding Transcripts Are Modulators of Noise. Genome Biology and Evolution *3*, 320–331.

Wei, W., Pelechano, V., Jarvelin, A.I., and Steinmetz, L.M. (2011). Functional consequences of bidirectional promoters. Trends in Genetics *27*, 267–276.

Whitehouse, I., Rando, O.J., Delrow, J., and Tsukiyama, T. (2007). Chromatin remodelling at promoters suppresses antisense transcription. Nature *450*, 1031–1035.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307–319.

Wilusz, J.E., Sunwoo, H., and Spector, D.L. (2009). Long noncoding RNAs: functional surprises from the RNA world. Genes & Development *23*, 1494–1504.

Xie, C., Zhang, Y.E., Chen, J.-Y., Liu, C.-J., Zhou, W.-Z., Li, Y., Zhang, M., Zhang, R., Wei, L., and Li, C.-Y. (2012). Hominoid-Specific De Novo Protein-Coding Genes Originating from Long Non-Coding RNAs. Plos Genetics *8*, e1002942.

Xu, C., Chen, J., and Shen, B. (2012). The preservation of bidirectional promoter architecture in eukaryotes: what is the driving force? BMC Systems Biology *6 Suppl 1*, S21.

Yang, L., Zou, M., Fu, B., and He, S. (2013). Genome-wide identification, characterization, and expression analysis of lineage-specific genes within zebrafish. Bmc Genomics *14*, 65.

Chapter 4: The RNA-guided CRISPR-Cas9 system

In this chapter, I review the CRISPR-Cas9 system with a focus on target specificity.

This chapter was published as:

Xuebing Wu, Andrea J. Kriz, Phillip A. Sharp, Target specificity of the CRISPR-Cas9 system, Quantitative Biology, 2014, in press

Contribution

X. W. wrote this chapter with the help from A. J. K., especially for the survey of existing tools. P. A. S. supervised this work. M. Lindstrom assisted figure preparation.

ABSTRACT

The CRISPR-Cas9 system, naturally a defense mechanism in prokaryotes, has been repurposed as an RNA-guided DNA targeting platform. It has been widely used for genome editing and transcriptome modulation, and has shown great promise in correcting mutations in human genetic diseases. Off-target effects are a critical issue for all of these applications. Here we review the current status on the target specificity of the CRISPR-Cas9 system.

THE CRISPR-CAS9 SYSTEM

The CRISPR-Cas system is widely found in bacterial and archaeal genomes as a defense mechanism against invading viruses and plasmids (Barrangou and Marraffini, 2014; Deveau et al., 2010; Horvath and Barrangou, 2010; Marraffini and Sontheimer, 2010; van der Oost et al., 2009; Terns and Terns, 2011). The type II CRISPR-Cas system from *Streptococcus pyogenes* relies on only one protein, the nuclease Cas9, and two noncoding RNAs, crRNA and tracrRNA, to target DNA (Jinek et al., 2012). These two noncoding RNAs can further be fused into one single guide RNA (sgRNA). The Cas9/sgRNA complex binds double-stranded DNA sequences that contain a sequence match to the first 17-20 nucleotides of the sgRNA if the target sequence is followed by a protospacer adjacent motif (PAM) (Fig. 1). Once bound, two independent nuclease domains in Cas9 will each cleave one of the DNA strands 3 bases upstream of the PAM, leaving a blunt end DNA double stranded break (DSB). DSBs can be repaired mainly through either the nonhomologous end joining (NHEJ) pathway or homology-directed repair (HDR). NHEJ typically leads to short

insertion/deletion (indels) near the cutting site, whereas HDR can be used to introduce specific sequences into the cutting site if exogenous template DNA is provided. This discovery paved the way for use of Cas9 as a genome-engineering tool in other species. In this review, we focus on target specificity of the CRISPR-Cas9 system. We refer readers to other excellent reviews for further discussion of the CRISPR-Cas9 technology (Hsu et al., 2014; Mali et al., 2013a; Sander and Joung, 2014; Zhang et al., 2014a).

APPLICATIONS OF CRISPR-CAS9

Genome editing

The use of the CRISPR-Cas9 system as a tool to manipulate the genome was first demonstrated in 2013 in mammalian cells (Cong et al., 2013; Mali et al., 2013b). Both studies showed that expressing a codon-optimized Cas9 protein and a guide RNA leads to efficient cleavage and short indels of target loci, which could inactive protein-coding genes by inducing frameshifts. Up to five genes have been mutated simultaneously in mouse and fish cells by delivering five guide RNAs (Jao et al., 2013; Yang et al., 2013). Targeting two sites on the same chromosome can be used to create deletions and inversions of regions range from 100 bps to 1,000,000 bps (Canver et al., 2014; Xiao et al., 2013). Defined interchromosomal translocation such as those found in specific cancers can be created by targeting Cas9 to different chromosomes (Torres et al., 2014). With exogenous template oligos, specific sequences such as HA-tag or GFP could be inserted into genes to label proteins (Auer et al., 2014; Hruscha et al., 2013), or to correct mutations in disease genes in human and mouse (Schwank et al., 2013; Wu et al., 2013; Yin et al., 2014). The system has

also been adapted to many other species as well, including monkey, pig, rat, zebrafish, worm, yeast, and several plants (see review (Sander and Joung, 2014)).

Transcriptome modulation

Mutating the two nuclease domains of Cas9 generates the catalytically inactive Cas9 (dCas9), or nuclease-null Cas9, which can bind DNA without introducing cleavage or mutation (Jinek et al., 2012). When targeted to promoters, dCas9 binding alone can interfere with transcription initiation, likely by blocking binding of transcription factors or RNA polymerases. When targeted to the non-template strand within the gene body, dCas9 complex blocks RNA polymerase II transcription elongation (Gilbert et al., 2013; Larson et al., 2013; Qi et al., 2013). Fusing dCas9 with transcription repressor domains such as the Krueppel-associated box (KRAB) leads to stronger silencing of mammalian genes, a technology termed CRISPRi (Larson et al., 2013). Activation of transcription is also possible by fusing dCas9 with activator domains such as VP64. However, several studies showed that multiple sgRNAs targeting the same promoter need to be used simultaneously to change target gene expression substantially (Cheng et al., 2013; Kearns et al., 2013; Mali et al., 2013c). The position of target sites with respective to transcription start site (TSS) affects the efficiency of silencing or activation, a subject that needs to be further investigated for optimal target design (Farzadfard et al., 2013).

Genomic loci imaging and other applications

To enable site-specific labeling and imaging of endogenous loci in living cells, GFP has also been fused to dCas9 (Chen et al., 2013). In this case, tens of sgRNAs are required to target the same locus such that individual loci show up as punctate dots, unless the target locus

contains targetable tandem repeats. The fusion of dCas9 with other heterologous effector domains could enable many other applications. For example, one could fuse dCas9 with chromatin modifiers to change the epigenetic state of a locus. Other potential applications of the system have been previously reviewed extensively (Mali et al., 2013a; Sander and Joung, 2014).

ASSESSING CAS9 TARGET SPECIFICITY

The original characterization of the Cas9/sgRNA system showed that not every position in the guide RNA needs to match the target DNA, suggesting the existence of off-target sites (Jinek et al., 2012). Concerns about off-target effects depend on the purpose of the targeting. As discussed above and below, Cas9/sgRNA binding at a site does not necessarily lead to DNA cutting or mutation, and binding or cutting may not have any functional consequence either, especially when the off-target sites are outside of genes or regulatory elements. The off-target effects of Cas9 cutting/mutation have been studied extensively but sensitive and unbiased genome-wide characterization is still missing. Below we review existing approaches that have been or can be used to study Cas9 target specificity.

Assay of predicted off-targets

Typically a list of potential off-target sites are predicted based on sequence homology to the on-target, or using more sophisticated tools that incorporate various rules previously described in literature (see section "Tools for target design and off-target prediction"). Two types of assays are commonly used to detect and quantify indels formed at those selected sites: mismatch-detection nuclease assay and next generation sequencing (NGS). In the

mismatch-detection nuclease assay, genomic DNA from cells treated with Cas9 and sgRNA is PCR amplified, denatured and rehybridized to form heteroduplex DNA, containing one wildtype strand and one strand with indels. Mismatches can be recognized and cleaved by mismatch detection nucleases, such as Surveyor nuclease (Qiu et al., 2004) or T7 endonuclease I (Mashal et al., 1995), enabling quantitation of the products by electrophoresis. It is challenging to use this assay to detect loci with less than 1% indels and this assay is difficult to scale-up. Alternatively, the PCR product can also be sequenced directly using NGS platform. The fraction of reads with indels is quantified after mapping to the genome or directly to the amplicon. When combined with proper controls and statistical models, NGS based approaches are more accurate and sensitive than nuclease based assays.

Systematic mutagenesis

To characterize Cas9/sgRNA specificity, several groups performed systematic mutagenic analysis of the sgRNA or target DNA to evaluate the importance of the position, identity, and number of mismatches in the RNA/DNA duplex (Cong et al., 2013; Fu et al., 2013; Hsu et al., 2013). These studies revealed a very complicated picture of Cas9 specificity (Carroll, 2013). However, it is unclear whether the observed variation truly reflects specificity requirement, or is confounded by unintended changes caused by the mutations introduced in the sgRNA or target DNA. For example, mutations in the sgRNA could change the sgRNA abundance dramatically, which would alter the targeting efficiency (Wu et al., 2014, see below). Also mutations in DNA might create or disrupt binding sites for endogenous proteins that interfere with Cas9 binding. The number of variants evaluated is also limited

in these studies. Finally, each study typically examines less than four target sites, leaving questions whether the observations can be generalized.

In vitro cleavage site selection

A more comprehensive way to study Cas9 cutting specificity is *in vitro* selection. In this assay a large pool of partially randomized targets are synthesized and cleaved by Cas9 or other nucleases *in vitro* (Guilinger et al., 2014a; Pattanayak et al., 2011, 2013). The cleavage leaves a 5' phosphate group in the DNA, which can then be ligated to an adaptor and selectively amplified using PCR. The advantages of this approach are that the sequence space explored by the target library can be very large (10^{12} molecules, even larger than all possible sites in any genome), and that target specificity can be evaluated independently of genome or species used and is not affected by chromatin structure that is usually cell-type specific. However these advantages also impose potential limitations of this assay. Although the sequence space of the library can be huge, most substrates contain on average only 4-5 mismatches to the on-target (Pattanayak et al., 2013). Given that efficient cleavage with 7 mismatches has been observed (Jinek et al., 2012), such an assay could still miss a significant fraction of genomic off-targets. For example, when the *in vitro* cleavage site selection approach was applied to another type of nuclease, the Zinc Finger Nuclease (ZFN), only one of the four off-targets identified by an *in vivo* assay was detected (see below) (Gabriel et al., 2011; Pattanayak et al., 2011; Sander et al., 2013). Alternatively, instead of using partially randomized synthetic DNA library, one could perform the same assay with genomic DNA to detect possible genomic off-targets.

It has also been reported that compared to *in vivo* conditions, Cas9 cutting is more promiscuous *in vitro* (Cho et al., 2014), i.e. off-targets are cleaved at much higher frequency

in vitro than *in vivo*. This can be potentially explained by chromatin blockage of accessibility of the off-target sites *in vivo* (Kuscu et al., 2014; O'Geen et al., 2014; Wu et al., 2014). Therefore a potential solution is to perform *in vitro* selection assay using native or fixed chromatin prepared from cells. However, the higher rate of off-target cutting could also be due to higher effective concentrations of Cas9/sgRNA used *in vitro*. A titration series of Cas9/sgRNA concentration is needed to assess the *in vivo* relevance of off-target sites identified by *in vitro* approaches.

DSB capture and sequencing

Cas9 and other DNA endonucleases typically induce DSBs, and several assays have been developed to capture DSBs induced in cells (Chailleux et al., 2014; Crosetto et al., 2013; Gabriel et al., 2011), although none of them have been applied to Cas9 system. Gabriel et al transformed human cells with integrase-defective lentiviral vectors (IDLVs), which are incorporated into DSBs via NHEJ pathway, thus tagging those transient cutting events (Gabriel et al., 2011). This approach uncovered four *in vivo* off-target cleavage sites for a ZFN targeting the CCR5 locus. In another *in situ* assay called BLESS (Crosetto et al., 2013), cells are fixed first and then chromatin are purified and ligated with biotinylated DNA linkers. Both approaches could in principle be applied to Cas9 treated cells to uncover genome-wide cutting sites. Compared to *in vitro* cleavage site selection approach, DSB capture approaches are physiologically more relevant, but can be less efficient since most DSBs exist very transiently, and the capture can be biased since both *in vivo* IDLV labeling and *in situ* linker ligation can be affected by local chromatin and sequence composition near the cutting site. Thus certain DSBs induced by the nuclease will not be tagged. For instance, of the 36 ZFN off-target sites identified by *in vitro* selection approach, only one is

identified by the IDLV-based DSB capture (Sander et al., 2013). In addition, DSB capture approaches may identify large number of false positive sites, since DSBs can be generated by endogenous cellular process independent of Cas9 cutting, or during the library preparation process. Proper controls, such as cells treated with no Cas9 or no sgRNA can be used to filter false positives.

Whole genome sequencing

Compared to assays described above, whole genome sequencing (WGS) would be a less biased assessment of off-target mutations caused by Cas9, although it will miss off-target sites that are bound without cutting, or are cut but then always perfectly repaired. In addition to small indels, WGS can also detect Cas9 induced structural changes, such as inversions (Canver et al., 2014). So far relatively high coverage (30-60X) of WGS has been performed in single clones of Cas9 treated cells in a variety of species, including worm (Chiu et al., 2013), Arabidopsis (Feng et al., 2014), rice (Zhang et al., 2014b), and human pluripotent stem cells (Smith et al., 2014; Veres et al., 2014). Interestingly, although a number of mutations were identified in Cas9 treated clones, none were found to be near sites with sequences similar to the target, indicating Cas9 induced off-target mutations are rare and it is possible to obtain clones without off-target mutations. However, due to the high cost, only a few clones have been sequenced for each target, which would miss most low-frequency off-target events. For example, if there was a single possible off-target site per genome mutated at a 40% frequency relative to the on-target site, this could have escaped detection in these experiments. However, if there were 10 possible off-target sites per genome mutated at a 40% frequency, then at least one of these sites should have been detected. Therefore, WGS is ideal for screening individual clones for off-targets, but at the

moment, it is not practical for systematic study of a large number of guide RNAs to determine the rules governing Cas9 specificity.

Whole genome binding

Chromatin immunoprecipitation (ChIP) is a widely used technique for assaying genome-wide binding of proteins on DNA *in vivo* (Park, 2009). Briefly, live cells are crosslinked, lysed and chromatin fragmented and then immunoprecipitated to pull down DNA bound by a specific protein. The DNA is then purified and assayed by microarray or NGS. Compared to other readouts, such as indels that are downstream of the repair pathway, or gene expression changes, which are also affected by relative position of binding to the transcription start site, ChIP provides direct evidence for Cas9 binding on the genome. We and other groups recently generated the first maps of dCas9 binding on mammalian genomes (Kuscu et al., 2014; O'Geen et al., 2014; Wu et al., 2014); all three studies revealed a large number of binding sites, for example up to six thousand in mouse embryonic cells, as well as substantial variation (200 fold) in the number of off-target sites between sgRNAs (Wu et al., 2014). Specificity was not altered by fusion to an effector domain, as dCas9-KRAB had a similar binding profile to dCas9 alone (O'Geen et al., 2014). Surprisingly, two of these studies observed little cutting/mutation at most off-targets tested, while one study observed significant cleavage at 30 out of 57 selected off-target sites, albeit at a substantially lower rate than on-target cleavage (Kuscu et al., 2014). We further observed little to none of the off-target gene expression change which would presumably result from strong dCas9 binding at many off-target sites (Wu et al, unpublished data). It is possible that most of the off-targets detected by ChIP are weak and transient interactions stabilized by crosslinking. Native ChIP without crosslinking may help to clarify this question. The

other limitation of ChIP approach is that it is inherently biased towards open chromatin and highly transcribed genes (Teytelman et al., 2013). There could be other biases that remain to be discovered. For example, we failed to detect binding at previously validated off-target sites using an NAG PAM (Wu et al., 2014). It is also unclear whether the two mutations introduced in dCas9 alter the target binding specificity as compared to wild type active Cas9.

Transcriptome profiling

For application in transcription modulation, transcriptome profiling by either microarray or RNA-seq is the ultimate read out for assessing off-target effects. In all published cases (Cheng et al., 2013; Gilbert et al., 2013; Perez-Pinera et al., 2013), no significant off-target gene expression changes were observed, which again is unexpected given the large number of off-target binding sites reported in ChIP-based studies, and that off-target binding is enriched in accessible active regulatory elements (Wu et al., 2014). It also remains to be seen whether marginally affected genes are enriched for off-target binding sites.

DETERMINANTS OF CAS9/sgRNA SPECIFICITY

Despite potential bias, the assays and studies described above revealed many factors that could affect Cas9/sgRNA targeting specificity (Fig. 2), and these can be broadly classified into two categories. First, the intrinsic specificity encoded in the Cas9 protein, which likely determines the relative importance of each position in the sgRNA for target recognition, which may vary for different sgRNA sequences. Secondly, the specificity also depends on the relative abundance of effective Cas9/sgRNA complex with respect to effective target concentration. Below we discuss factors that could affect target specificity.

PAM

The protospacer-adjacent motif (PAM) is strictly required to be immediately next to the 3' end of the target sequence. PAM is recognized by an individual domain in the Cas9 protein (Nishimasu et al., 2014), and the PAM sequence varies across bacteria species (Garneau et al., 2010; Zhang et al., 2013). Presumably species with longer PAM, having less targetable sites in the genome, will have correspondingly fewer off-targets, although this has not been directly tested. For the widely used Cas9 from *Streptococcus pyogenes*, the PAM is typically NGG, where the first position shows no nucleotide bias. Recent data suggested that PAM binding is required for both opening the DNA and target cleavage (Nishimasu et al., 2014; Sternberg et al., 2014). Both *in vitro* (Pattanayak et al., 2013) and *in vivo* (Hsu et al., 2013; Mali et al., 2013c; Ran et al., 2013) cleavage data suggested that NAG is also tolerated to some extent, especially when Cas9/sgRNA is in excess to target DNA. In addition, other variants that contain at least one of the two G's at position 2 and 3, i.e. NNG or NGN, could lead to some cleavage activity *in vitro* under Cas9 excess conditions (Pattanayak et al., 2013). Interestingly recent genome-wide ChIP-seq data revealed no significant Cas9 binding at NAG targets (Kuscu et al., 2014; O'Geen et al., 2014; Wu et al., 2014), including previously validated off-target NAG cleavage sites, suggesting ChIP may not be able to detect off-target sites with certain PAMs.

Seed

In the original characterization of CRISPR-Cas9 (Jinek et al., 2012), mismatches in the first 7 positions (PAM-distal) of the guide RNA are well tolerated in terms of cleavage of a plasmid *in vitro*. Further studies in bacteria and mammalian cells showed that mismatches in the 10-12 base pairs in the PAM-proximal region usually lead to decrease or even

complete abolishment of target cleavage activity. Another study reported that Cas9 can even cleave DNA sequences that contain insertions or deletions relative to the guide RNA; however many of these sites could be alternatively aligned to contain only mismatches to the guide (Lin et al., 2014). Thus, the PAM-proximal 10-12 bases have been defined as the seed region for Cas9 cutting activity (Cong et al., 2013; Jiang et al., 2013). However, a relatively comprehensive *in vitro* cleavage and selection approach revealed no clearly defined seed region for four guide RNAs, although the results confirmed that mismatches near the PAM region are less tolerated (Pattanayak et al., 2013). In contrast, in two genome-wide binding datasets, one out of two (O'Geen et al., 2014) and three of the four (Wu et al., 2014) sgRNAs tested showed a clearly defined seed region, only the first 5 nucleotides next to PAM. A third genome-wide binding dataset detected no obvious seed for twelve sgRNAs tested, although PAM proximal bases tended to be more preserved than PAM distal bases in binding sites (Kuscu et al., 2014). However, the same data, when analyzed with our pipeline, revealed the 5-nucleotide seed region for three out of twelve sgRNAs (Wu et al., unpublished data); this is likely due to differences in selecting the best match to the guide region near binding sites, e.g., accepting matches with alternative PAMs. Hundreds of binding sites detected by ChIP *in vivo* contain only seed match with mismatches at all the other 15 positions in the guide RNA (Wu et al., 2014). We also showed that seed-only sites could be bound by Cas9/sgRNA complex *in vitro* using a gel shift assay. The variation in the length of the seed detected by different assays likely stems from different concentrations of factors and lengths of dwell times required for Cas9 binding and cleavage.

Cas9/sgRNA abundance

Cas9 cutting becomes less specific at higher effective concentrations of Cas9/sgRNA complexes. For example, *in vitro*, when excessive amounts of Cas9/sgRNA complex are present, mismatches in the guide matching region are more tolerated, and Cas9 can even cut at sites with mismatches in the PAM region (Pattanayak et al., 2013). Hsu et al also showed that *in vivo* the specificity (ratio of indel frequency at target vs off-target) increases when decreasing amounts of Cas9 and sgRNA plasmids are transfected into cells (Hsu et al., 2013). Genome-wide, we have found that increasing Cas9 protein levels by 2.6 fold leads to a 2.6 fold increase in the number of off-target binding peaks in the genome. On the other hand, at a constant level of Cas9 protein, titrating the amount of sgRNA expression plasmid transfected, and thus the abundance of sgRNA, largely determines the number of off-target binding sites in mouse genome (Wu et al., 2014).

Target or guide sequence

In addition to targeting Cas9 to a certain region in the genome, the sequence of the sgRNA alone appears to affect specificity (Fu et al., 2013; Hsu et al., 2013; Mali et al., 2013b; Pattanayak et al., 2013). For example, the tolerance of mismatches at each position varies dramatically between different sgRNAs, an observation that remains to be understood. Possible mechanisms whereby a change in sgRNA sequence could affect Cas9 specificity include: 1) Changes that alter the effective concentration of sgRNA (by modulating transcription of the sgRNA, the stability of the sgRNA, or sgRNA loading into Cas9). For example, we found that two mutations in the seed region can increase U6 promoter transcribed sgRNA's abundance by at least 7 fold (Wu et al., 2014). 2) Changes that alter the number of seed-matching sites in the genome, which can vary by 100-fold (see below).

3) Changes that depend on the local chromatin environment of the target DNA sequences (ie. chromatin accessibility). 4) Changes that might cause off-target effects by blocking the binding of trans-acting factors that may potentially affect Cas9 binding or reporter gene transcription. 5) Changes that alter the thermodynamic stability of the guide RNA-DNA duplex. It is likely that the observed effects of sgRNA sequence on specificity are the result of multiple mechanisms described above. Below we will discuss some of these effects in detail.

Accessibility of seed match genomic sites

In cells DNA is packed in chromatin and may have limited accessibility for Cas9 PAM recognition and target binding. DNase I hypersensitivity (DHS) is typically considered to be an indicator of chromatin accessibility. We have shown that DHS is a strong predictor of whether a 5-nucleotide seed followed by NGG (seed+NGG) site will be bound *in vivo* (Wu et al., 2014), and others have also observed a strong correlation between Cas9-bound sites and open chromatin (Kuscu et al., 2014; O'Geen et al., 2014). In fact, the number of seed+NGG sites in DHS peaks (accessible seed+NGG sites) accurately predicts the number of ChIP peaks detected *in vivo* (R^2=0.92) (Wu et al., 2014). Interestingly, designed target sites not in DHS peaks show significant ChIP enrichment over background, in our case comparable to that of target sites in open chromatin, suggesting that chromatin accessibility is not a requirement for binding to the on-target site (Kuscu et al., 2014; Wu et al., 2014). This is consistent with previous studies showing that dCas9-VP64 fusion protein could be targeted to non-open chromatin regions to activate target gene transcription (Perez-Pinera et al., 2013). In sum, chromatin accessibility seems to be preferentially facilitating off-target binding.

The preferential enrichment of off-targets in accessible chromatin has implications for dCas9-based transcriptome modulation. In fact, we found that regulatory elements of active genes, such as promoter and enhancers, are significantly enriched for off-target binding since those elements are accessible when active. To what extent these off-target binding events lead to gene expression change remains to be addressed.

Abundance of seed match genomic sites

Given that the binding seed length is relatively short (5 to 12), each guide RNA potentially has thousands to hundreds of thousands of seed match sites in the mammalian genome that are followed by NGG (Wu et al., 2014). However, due to mutational bias and other sequence bias in the genome, the occurrence of specific seed sequences could vary dramatically. For example, there are about 1 million AAGGA+NGG sites in the mouse genome, compared to less than 10,000 CGTCG+NGG sites. Therefore it is important to consider abundance of seed sites when designing sgRNA targets for dCas9 based applications. We have shown that the number of accessible seed+NGG sites in the genome can very accurately predict the number of peaks detected by ChIP ($R^2=0.92$), although we only tested four guide RNAs (Wu et al., 2014).

Epigenetics

In addition to chromatin accessibility, we have also shown that for target sites with CpG dinucleotides, methylation status strongly correlates with ChIP signal (Wu et al., 2014). Specifically, more methylation is associated with less binding, a correlation even stronger than DHS for the same set of sites. Consistent with the observation that CpG methylation is typically associated with chromatin silencing, we observed a strong negative correlation

between DHS and CpG methylation. However, the correlation between CpG methylation and Cas9 binding remained strong even after subtracting the effect of DHS. Previously Hsu et al. showed that *in vitro* CpG methylation has no effects on Cas9 cutting of substrates with no mismatches to the guide RNA, and *in vivo*, Cas9 could mutate a promoter that is highly methylated, albeit with low indel frequency (Hsu et al., 2013). Taking this information together, we speculate that CpG methylation may represent chromatin accessibility not detected by DHS and like DHS, CpG methylation only affects binding at off-target sites. Similarly histone modifications may affect target site accessibility, although so far this has not been investigated.

Target sequence length

One might expect that if the guide region is longer than 20 nucleotides, a longer RNA-DNA duplex may be formed and thus the Cas9/sgRNA complex might have higher specificity. Ran et al increased the length of the guide region to 30 nucleotides by extending the 5' end of the sgRNA. Interestingly Northern blots detected that the extended 5' end was trimmed *in vivo* (Ran et al., 2013), suggesting that Cas9 only protects about 20 nucleotides of the guide RNA and free sgRNA is largely unstable. On the other hand, it has been recently reported that when sgRNA is truncated to 17 or 18 nucleotides, the specificity increases dramatically (Fu et al., 2014). The mechanism underlying this increased specificity is unclear. It was assumed the increased specificity is because the first 2-3 nucleotides are not necessary for on-target binding but instead stabilize off-target binding (Fu et al., 2014). The other possibility is that shortened sgRNA may simply be less abundant or less efficiently loaded into Cas9.

sgRNA scaffold

In addition to the 5' end, various modifications have been introduced to the scaffold region of the guide RNA, although their impact on target specificity is not well studied. Extension or truncation at the 3' end can drastically change sgRNA expression levels (Hsu et al., 2013), likely due to change in transcription or RNA stability, which in principle could affect specificity by tuning the effective concentration of Cas9/sgRNA complexes. Modifications have also been introduced to stabilize the sgRNA by flipping an A-U base pair at the beginning of the scaffold (Chen et al., 2013). Increasing the length of a hairpin that is supposed to be bound by Cas9 also helps to increase the efficiencies for both imaging and transcription regulation, likely due to more efficient loading of sgRNA into Cas9. The effect of these modifications on the specificity of binding or cutting remains unclear, although it is reported that these modifications lead to higher signal to background ratio for imaging (Chen et al., 2013).

STRATEGIES TO INCREASE SPECIFICITY

Controlling Cas9/sgRNA abundance and duration

Typically Cas9 and sgRNA are expressed in cells by transient transfection of expressing plasmids. Titrating down the amount of plasmid DNA used in transfection increases specificity, although there is a trade-off for decreased efficiency at the on-target site. This is particularly an issue when the promoter is very strong, i.e. successfully transfected cells express a large amount of Cas9 and sgRNA leading to off-target effects. More recently, sgRNA has been expressed by RNA Pol II transcription and processed from introns,

microRNAs, ribozymes, and RNA-triplex-helix structures, providing more flexible control of the sgRNA abundance (Kiani et al., 2014; Nissim et al., 2014).

Alternative delivery methods have also been developed to increase specificity. Compared to plasmid transfection based delivery, direct delivery of recombinant Cas9 protein and *in vitro* transcribed sgRNA, either individually or as purified complex, reduces off-targets in cells (Kim et al., 2014; Ramakrishna et al., 2014). This is likely due to the rapid degradation of the protein and RNA in cells, which would lower the effective concentration of the Cas9-sgRNA effector complex and its duration in cells.

Paired nickase

The Cas9 "nickase" generated by mutating only one nuclease domain can only cleave one strand of the target DNA, which is thought to be repaired efficiently in cells. When the nickase is targeted to two neighboring regions on opposite strands, the offset double nicking leads to a double stranded break with tails that are degraded and subsequently indels in the target region. The requirement of dual Cas9 targeting to a nearby region dramatically increases the specificity, since it is generally unlikely that two guide RNAs will also have nearby off-targets. The limitation of this strategy is that nicks induced by Cas9 could still lead to mutations in off target sites via unknown mechanisms (Fu et al., 2014; Mali et al., 2013b, 2013c; Ran et al., 2013).

dCas9-FokI dimerization

FokI nuclease only cuts DNA when dimerized (Bitinaite et al., 1998). Fusion of dCas9 to FokI monomers creates an RNA-guided nuclease that only cuts the DNA when two guide RNAs bind nearby regions with defined spacing and orientation, thus substantially

reducing off-target cleavage (Guilinger et al., 2014b; Tsai et al., 2014). It has been reported that RNA-guided FokI nuclease is at least four fold more specific than paired Cas9 nickase (Guilinger et al., 2014b), likely due to FokI nuclease only functioning when dimerized whereas Cas9 nickase can cleave as a monomer (Tsai et al., 2014). Similar to paired nickases, the requirement of two nearby PAM sites with defined spacing and orientation reduces the frequency of target sites in the genome.

TOOLS FOR TARGET DESIGN AND OFF-TARGET PREDICTION

Several tools have been developed for designing sgRNA targets, with the primary consideration to avoid off-targets in the genome (Aach et al., 2014; Bae et al., 2014; Gratz et al., 2014; Heigwer et al., 2014; Hsu et al., 2013; Ma et al., 2013; Montague et al., 2014; Sander et al., 2007, 2010; Xiao et al., 2014; Xie et al., 2014). These tools typically consider an input sequence, a genomic region, or a gene and output potential target/guide sequences with predicted minimized off-target effects. Many of the tools also provide predicted off-target sites for a given sgRNA. These tools vary in their scheme for scoring potential guides and off-targets. Some tools incorporate data from previous systematic mutagenic studies (Hsu et al., 2013) or user-input penalties (Aach et al., 2014; Heigwer et al., 2014) to individually score off-targets based on location and number of mismatches to the guide. Other tools have binary criteria for off-targets, such as sites with less than a certain number of mismatches to the entire guide region (Bae et al., 2014; Sander et al., 2007, 2010), or to some defined PAM proximal or distal region (Gratz et al., 2014; Ma et al., 2013; Montague et al., 2014; Xiao et al., 2014; Xie et al., 2014). Potential guides are generally ranked by a weighted sum of off-target scores, or by number of off-targets.

Several tools consider factors beyond position and number of mismatches. Some tools (Aach et al., 2014) include the option to score off-targets with alternate PAMs based on the finding that Cas9 cleaves these sites with lower efficiency (Hsu et al., 2013; Mali et al., 2013c; Pattanayak et al., 2013; Ran et al., 2013). In terms of the on-target site, various tools consider presence of SNPs and secondary structure (Ma et al., 2013) in the potential guide, which could impact targeting and loading of the sgRNA (Makarova et al., 2011), genomic context of the guide (e.g. exons, transcripts, CpG islands), which could impact the intended purpose of the sgRNA (Heigwer et al., 2014; Montague et al., 2014), and GC content, which could impact effectiveness of the sgRNA (Heigwer et al., 2014; Montague et al., 2014; Wang et al., 2014; Xie et al., 2014).

Information from these tools is usually downloadable and sometimes viewable in an interactive format (Hsu et al., 2013; Montague et al., 2014). In addition, some tools provide support beyond finding potential guides, such as sequences of oligonucleotides for sgRNA construction (Sander et al., 2007, 2010; Xie et al., 2014) or primers for validation of cleavage at the target site (Montague et al., 2014; Xie et al., 2014). Some tools also provide specialized modes for design of sgRNA with paired Cas9 nickases (Heigwer et al., 2014; Hsu et al., 2013; Sander et al., 2007, 2010; Xiao et al., 2014; Xie et al., 2014) or RNA-guided FokI nucleases (Sander et al., 2007, 2010; Xie et al., 2014).

Each of these tools has its advantages and disadvantages. Researchers seeking to design CRISPR-Cas9 targets in less well-studied organisms or alternative species of Cas9 will need to use tools that accept user-input genomes (Aach et al., 2014; Bae et al., 2014; Ma et al., 2013; Xiao et al., 2014; Xie et al., 2014), are tailored for their organism (Gratz et al., 2014), accept alternate PAM (Bae et al., 2014; Montague et al., 2014; Xiao et al., 2014) or

user-input PAM (Aach et al., 2014). The desired purpose of the CRISPR-Cas9 guide is also an important factor to consider. For example, some tools focus on designing sgRNAs to target genes with high efficacy (Montague et al., 2014). If off-target effects are more of a concern, it may be helpful to use a tool that scores predicted off-targets quantitatively (Aach et al., 2014; Heigwer et al., 2014; Hsu et al., 2013). The type of off-targets detected by each tool also varies; most tools only search for off-targets with few (typically three or less) PAM-proximal or total mismatches to the guide (Gratz et al., 2014; Ma et al., 2013; Montague et al., 2014; Sander et al., 2007, 2010). Considering what we have discussed in this review, especially for applications of dCas9, these may fail to detect many potential off-targets compared to tools that consider off-targets with more mismatches to the guide (Aach et al., 2014; Bae et al., 2014; Heigwer et al., 2014; Hsu et al., 2013; Xiao et al., 2014; Xie et al., 2014). Since almost every tool has unique features, it may be useful to incorporate multiple tools during the design process. We refer readers to Supplementary Table 1 for a more detailed comparison.

Overall, these tools could aid in designing sgRNA targets that have minimal sequence homology to other sites in the genome. However, many features that are important to sgRNA specificity, as we have discussed, remain to be implemented, such as impact of seed sequence on sgRNA abundance, seed abundance in the genome, and epigenetic features. These factors, as we have discussed, are currently thought to primarily affect binding, or dCas9 based, applications.

PERSPECTIVE

Despite intense study, the rules governing the specificity of Cas9/sgRNA targeting, especially target cutting and mutation remain elusive. At this stage, it is still challenging to predict genome-wide off-targets of Cas9 with any significant confidence. Although our genome-wide binding data set shows that the number of off-target peaks can be accurately predicted from the number of accessible seed+NGG sites, predicting binding at individual sites remains challenging (Wu et al., 2014). This suggests that there could be other factors, such as higher-level chromatin structure, that further limit binding of Cas9.

In addition, the relationship between Cas9 binding and functional consequences such as cleavage, mutation and transcription perturbation remains elusive. Several lines of evidences suggest that most Cas9 off-target binding events may be transient and have little functional impact. First, in two separate studies, only one of the 295 off-target binding sites (Wu et al., 2014) or one out of 473 off-target binding sites (O'Geen et al., 2014) tested showed evidence of mutations in cells expressing active Cas9 and corresponding sgRNAs. Secondly, transcriptome profiling revealed negligible off-target gene expression change (Cheng et al., 2013; Gilbert et al., 2013; Perez-Pinera et al., 2013). Furthermore, theoretical calculation implies an exponential decay in activity from Cas9 binding to downstream effects such as gene expression change (Mali et al., 2013c). However, a direct comparison between genome-wide binding, cutting, and transcriptome change will be needed to support this claim.

The current rules of Cas9/sgRNA specificity are likely incomplete and biased. Most assays described here are biased, and may only detect a fraction of the off-target sites in cells and predict many false positives. Integration of multiple assays will likely lead to more

comprehensive and more accurate identification of off-targets. For example, intersecting ChIP-detected Cas9 binding sites with whole-genome sequencing data will likely lead to authentic Cas9 target sites while removing Cas9-independent false positives, such as sequencing error or ChIP bias.

In addition to biased assays, the rules learned from each study are also likely biased by the small number of sgRNAs studied, given that the target specificity highly depends on the target sequence. Most of the assays described here are difficult to scale-up, such as ChIP, *in vitro* selection, and whole-genome sequencing. Further development of multiplexable unbiased assays, such as DSB capture with barcoded linkers, could facilitate the study of large number of sgRNAs at the same time.

The issue of off-targets is most critical in use of the Cas9 system to mutate specific genes. Here off-targets could generate spurious phenotypes and mistaken interpretations. This is particularly a concern when a large library of Cas9 vectors is screened with selective conditions for specific phenotypes. In this case a rare off-target mutation could be selected and the phenotype accredited to the on-target gene. The only really valid assay under these conditions is the deep sequencing of the total genome of the cloned mutated cell. However, this is much too expensive for most experiments and will only be done in particular cases. The principles summarized here about specificity of the Cas9 system hopefully will lead to experimental designs that optimize the probability of obtaining desired on-target mutants in the absence of unknown off-target changes.

Lastly, alternative Cas9 protein and guide RNA architecture may improve specificity. Several alternative Cas9 proteins from various bacteria have been studied and display very different PAM sequences (Esvelt et al., 2013). Comprehensive characterization of the

specificity, such as genome-wide binding and cutting, may identify novel Cas9 proteins with dramatically improved specificity. With available crystal structure (Nishimasu et al., 2014), it is also possible to design a new Cas9 protein with increased specificity via protein engineering and *in vitro* evolution.

REFERENCES

Aach, J., Mali, P., and Church, G.M. (2014). CasFinder: Flexible algorithm for identifying specific Cas9 targets in genomes. bioRxiv.

Auer, T.O., Duroure, K., De Cian, A., Concordet, J.-P., and Del Bene, F. (2014). Highly efficient CRISPR/Cas9-mediated knock-in in zebrafish by homology-independent DNA repair. Genome Res. *24*, 142–153.

Bae, S., Park, J., and Kim, J.-S. (2014). Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. Bioinformatics *30*, 1473–1475.

Barrangou, R., and Marraffini, L.A. (2014). CRISPR-Cas Systems: Prokaryotes Upgrade to Adaptive Immunity. Mol. Cell *54*, 234–244.

Bitinaite, J., Wah, D.A., Aggarwal, A.K., and Schildkraut, I. (1998). FokI dimerization is required for DNA cleavage. Proc. Natl. Acad. Sci. U. S. A. *95*, 10570–10575.

Canver, M.C., Bauer, D.E., Dass, A., Yien, Y.Y., Chung, J., Masuda, T., Maeda, T., Paw, B.H., and Orkin, S.H. (2014). Characterization of Genomic Deletion Efficiency Mediated by CRISPR/Cas9 in Mammalian Cells. J. Biol. Chem. advanced online publication.

Carroll, D. (2013). Staying on target with CRISPR-Cas. Nat. Biotechnol. *31*, 807–809.

Chailleux, C., Aymard, F., Caron, P., Daburon, V., Courilleau, C., Canitrot, Y., Legube, G., and Trouche, D. (2014). Quantifying DNA double-strand breaks induced by site-specific endonucleases in living cells by ligation-mediated purification. Nat. Protoc. *9*, 517–528.

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

Cheng, A.W., Wang, H., Yang, H., Shi, L., Katz, Y., Theunissen, T.W., Rangarajan, S., Shivalila, C.S., Dadon, D.B., and Jaenisch, R. (2013). Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. Cell Res. *23*, 1163–1171.

Chiu, H., Schwartz, H.T., Antoshechkin, I., and Sternberg, P.W. (2013). Transgene-Free Genome Editing in Caenorhabditis elegans Using CRISPR-Cas. Genetics *195*, 1167–1171.

Cho, S.W., Kim, S., Kim, Y., Kweon, J., Kim, H.S., Bae, S., and Kim, J.-S. (2014). Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. Genome Res. *24*, 132–141.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819–823.

Crosetto, N., Mitra, A., Silva, M.J., Bienko, M., Dojer, N., Wang, Q., Karaca, E., Chiarle, R., Skrzypczak, M., Ginalski, K., et al. (2013). Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. Nat. Methods *10*, 361–365.

Deveau, H., Garneau, J.E., and Moineau, S. (2010). CRISPR/Cas system and its role in phage-bacteria interactions. Annu. Rev. Microbiol. *64*, 475–493.

Esvelt, K.M., Mali, P., Braff, J.L., Moosburner, M., Yaung, S.J., and Church, G.M. (2013). Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. Nat. Methods *10*, 1116–1121.

Farzadfard, F., Perli, S.D., and Lu, T.K. (2013). Tunable and Multifunctional Eukaryotic Transcription Factors Based on CRISPR/Cas. ACS Synth. Biol. *2*, 604–613.

Feng, Z., Mao, Y., Xu, N., Zhang, B., Wei, P., Yang, D.-L., Wang, Z., Zhang, Z., Zheng, R., Yang, L., et al. (2014). Multigeneration analysis reveals the inheritance, specificity, and patterns of CRISPR/Cas-induced gene modifications in Arabidopsis. Proc. Natl. Acad. Sci. U. S. A. *111*, 4632–4637.

Fu, Y., Foden, J.A., Khayter, C., Maeder, M.L., Reyon, D., Joung, J.K., and Sander, J.D. (2013). High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat. Biotechnol. *31*, 822–826.

Fu, Y., Sander, J.D., Reyon, D., Cascio, V.M., and Joung, J.K. (2014). Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. Nat. Biotechnol. *32*, 279–284.

Gabriel, R., Lombardo, A., Arens, A., Miller, J.C., Genovese, P., Kaeppel, C., Nowrouzi, A., Bartholomae, C.C., Wang, J., Friedman, G., et al. (2011). An unbiased genome-wide analysis of zinc-finger nuclease specificity. Nat. Biotechnol. *29*, 816–823.

Garneau, J.E., Dupuis, M.-È., Villion, M., Romero, D.A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadán, A.H., and Moineau, S. (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature *468*, 67–71.

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

Gratz, S.J., Ukken, F.P., Rubinstein, C.D., Thiede, G., Donohue, L.K., Cummings, A.M., and O'Connor-Giles, K.M. (2014). Highly specific and efficient CRISPR/Cas9-catalyzed homology-directed repair in Drosophila. Genetics *196*, 961–971.

Guilinger, J.P., Pattanayak, V., Reyon, D., Tsai, S.Q., Sander, J.D., Joung, J.K., and Liu, D.R. (2014a). Broad specificity profiling of TALENs results in engineered nucleases with improved DNA-cleavage specificity. Nat. Methods *11*, 429–435.

Guilinger, J.P., Thompson, D.B., and Liu, D.R. (2014b). Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. Nat. Biotechnol. *32*, 577–582.

Heigwer, F., Kerr, G., and Boutros, M. (2014). E-CRISP: fast CRISPR target site identification. Nat. Methods *11*, 122–123.

Horvath, P., and Barrangou, R. (2010). CRISPR/Cas, the immune system of bacteria and archaea. Science *327*, 167–170.

Hruscha, A., Krawitz, P., Rechenberg, A., Heinrich, V., Hecht, J., Haass, C., and Schmid, B. (2013). Efficient CRISPR/Cas9 genome editing with low off-target effects in zebrafish. Development *140*, 4982–4987.

Hsu, P.D., Scott, D.A., Weinstein, J.A., Ran, F.A., Konermann, S., Agarwala, V., Li, Y., Fine, E.J., Wu, X., Shalem, O., et al. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. Nat. Biotechnol. *31*, 827–832.

Hsu, P.D., Lander, E.S., and Zhang, F. (2014). Development and Applications of CRISPR-Cas9 for Genome Engineering. Cell *157*, 1262–1278.

Jao, L.-E., Wente, S.R., and Chen, W. (2013). Efficient multiplex biallelic zebrafish genome editing using a CRISPR nuclease system. Proc. Natl. Acad. Sci. U. S. A. *110*, 13904–13909.

Jiang, W., Bikard, D., Cox, D., Zhang, F., and Marraffini, L.A. (2013). RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat. Biotechnol. *31*, 233–239.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science *337*, 816–821.

Kearns, N.A., Genga, R.M.J., Enuameh, M.S., Garber, M., Wolfe, S.A., and Maehr, R. (2013). Cas9 effector-mediated regulation of transcription and differentiation in human pluripotent stem cells. Development *141*, 219–223.

Kiani, S., Beal, J., Ebrahimkhani, M.R., Huh, J., Hall, R.N., Xie, Z., Li, Y., and Weiss, R. (2014). CRISPR transcriptional repression devices and layered circuits in mammalian cells. Nat. Methods Epub ahead of print.

Kim, S., Kim, D., Cho, S.W., Kim, J., and Kim, J.-S. (2014). Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. Genome Res. *24*, 1012–1019.

Kuscu, C., Arslan, S., Singh, R., Thorpe, J., and Adli, M. (2014). Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. Nat. Biotechnol. Epub ahead of print.

Larson, M.H., Gilbert, L.A., Wang, X., Lim, W.A., Weissman, J.S., and Qi, L.S. (2013). CRISPR interference (CRISPRi) for sequence-specific control of gene expression. Nat. Protoc. *8*, 2180–2196.

Lin, Y., Cradick, T.J., Brown, M.T., Deshmukh, H., Ranjan, P., Sarode, N., Wile, B.M., Vertino, P.M., Stewart, F.J., and Bao, G. (2014). CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. Nucleic Acids Res. Epub ahead of print.

Ma, M., Ye, A.Y., Zheng, W., and Kong, L. (2013). A guide RNA sequence design platform for the CRISPR/Cas9 system for model organism genomes. Biomed Res. Int. *2013*, 270805.

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J.M., Wolf, Y.I., Yakunin, A.F., et al. (2011). Evolution and classification of the CRISPR-Cas systems. Nat. Rev. Microbiol. *9*, 467–477.

Mali, P., Esvelt, K.M., and Church, G.M. (2013a). Cas9 as a versatile tool for engineering biology. Nat. Methods *10*, 957–963.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013b). RNA-guided human genome engineering via Cas9. Science *339*, 823–826.

Mali, P., Aach, J., Stranges, P.B., Esvelt, K.M., Moosburner, M., Kosuri, S., Yang, L., and Church, G.M. (2013c). CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. Nat. Biotechnol. *31*, 833–838.

Marraffini, L.A., and Sontheimer, E.J. (2010). CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat. Rev. Genet. *11*, 181–190.

Mashal, R.D., Koontz, J., and Sklar, J. (1995). Detection of mutations by cleavage of DNA heteroduplexes with bacteriophage resolvases. Nat. Genet. *9*, 177–183.

Montague, T.G., Cruz, J.M., Gagnon, J.A., Church, G.M., and Valen, E. (2014). CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. Nucleic Acids Res. Epub ahead of print.

Nishimasu, H., Ran, F.A.A., Hsu, P.D.D., Konermann, S., Shehata, S.I.I., Dohmae, N., Ishitani, R., Zhang, F., and Nureki, O. (2014). Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA. Cell *156*, 935–949.

Nissim, L., Perli, S.D., Fridkin, A., Perez-Pinera, P., and Lu, T.K. (2014). Multiplexed and Programmable Regulation of Gene Networks with an Integrated RNA and CRISPR/Cas Toolkit in Human Cells. Mol. Cell *54*, 698–710.

O'Geen, H., Henry, I.M., Bhakta, M.S., Meckler, J.F., and Segal, D.J. (2014). A genome-wide analysis of Cas9 binding specificity using ChIP-seq and targeted sequence capture. bioRxiv.

Van der Oost, J., Jore, M.M., Westra, E.R., Lundgren, M., and Brouns, S.J.J. (2009). CRISPR-based adaptive and heritable immunity in prokaryotes. Trends Biochem. Sci. *34*, 401–407.

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

Pattanayak, V., Ramirez, C.L., Joung, J.K., and Liu, D.R. (2011). Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. Nat. Methods *8*, 765–770.

Pattanayak, V., Lin, S., Guilinger, J.P., Ma, E., Doudna, J.A., and Liu, D.R. (2013). High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. Nat. Biotechnol. *31*, 839–843.

Perez-Pinera, P., Kocak, D.D., Vockley, C.M., Adler, A.F., Kabadi, A.M., Polstein, L.R., Thakore, P.I., Glass, K.A., Ousterout, D.G., Leong, K.W., et al. (2013). RNA-guided gene activation by CRISPR-Cas9-based transcription factors. Nat. Methods *10*, 973–976.

Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P., and Lim, W.A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell *152*, 1173–1183.

Qiu, P., Shandilya, H., D'Alessio, J.M., O'Connor, K., Durocher, J., and Gerard, G.F. (2004). Mutation detection using Surveyor nuclease. Biotechniques *36*, 702–707.

Ramakrishna, S., Kwaku Dad, A.-B., Beloor, J., Gopalappa, R., Lee, S.-K., and Kim, H. (2014). Gene disruption by cell-penetrating peptide-mediated delivery of Cas9 protein and guide RNA. Genome Res. *24*, 1020–1027.

Ran, F.A., Hsu, P.D., Lin, C.-Y., Gootenberg, J.S., Konermann, S., Trevino, A.E., Scott, D.A., Inoue, A., Matoba, S., Zhang, Y., et al. (2013). Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity. Cell *154*, 1380–1389.

Sander, J.D., and Joung, J.K. (2014). CRISPR-Cas systems for editing, regulating and targeting genomes. Nat. Biotechnol. *32*, 347–355.

Sander, J.D., Zaback, P., Joung, J.K., Voytas, D.F., and Dobbs, D. (2007). Zinc Finger Targeter (ZiFiT): an engineered zinc finger/target site design tool. Nucleic Acids Res. *35*, W599–605.

Sander, J.D., Maeder, M.L., Reyon, D., Voytas, D.F., Joung, J.K., and Dobbs, D. (2010). ZiFiT (Zinc Finger Targeter): an updated zinc finger engineering tool. Nucleic Acids Res. *38*, W462–8.

Sander, J.D., Ramirez, C.L., Linder, S.J., Pattanayak, V., Shoresh, N., Ku, M., Foden, J.A., Reyon, D., Bernstein, B.E., Liu, D.R., et al. (2013). In silico abstraction of zinc finger nuclease cleavage profiles reveals an expanded landscape of off-target sites. Nucleic Acids Res. *41*, e181.

Schwank, G., Koo, B.-K., Sasselli, V., Dekkers, J.F., Heo, I., Demircan, T., Sasaki, N., Boymans, S., Cuppen, E., van der Ent, C.K., et al. (2013). Functional Repair of CFTR by CRISPR/Cas9 in Intestinal Stem Cell Organoids of Cystic Fibrosis Patients. Cell Stem Cell *13*, 653–658.

Smith, C., Gore, A., Yan, W., Abalde-Atristain, L., Li, Z., He, C., Wang, Y., Brodsky, R.A., Zhang, K., Cheng, L., et al. (2014). Whole-Genome Sequencing Analysis Reveals High Specificity of CRISPR/Cas9 and TALEN-Based Genome Editing in Human iPSCs. Cell Stem Cell *15*, 12–13.

Sternberg, S.H., Redding, S., Jinek, M., Greene, E.C., and Doudna, J.A. (2014). DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature *507*, 62–67.

Terns, M.P., and Terns, R.M. (2011). CRISPR-based adaptive immune systems. Curr. Opin. Microbiol. *14*, 321–327.

Teytelman, L., Thurtle, D.M., Rine, J., and van Oudenaarden, A. (2013). Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. Proc. Natl. Acad. Sci. U. S. A. *110*, 18602–18607.

Torres, R., Martin, M.C., Garcia, A., Cigudosa, J.C., Ramirez, J.C., and Rodriguez-Perales, S. (2014). Engineering human tumour-associated chromosomal translocations with the RNA-guided CRISPR–Cas9 system. Nat. Commun. *5*, 3964.

Tsai, S.Q., Wyvekens, N., Khayter, C., Foden, J.A., Thapar, V., Reyon, D., Goodwin, M.J., Aryee, M.J., and Joung, J.K. (2014). Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. Nat. Biotechnol. *32*, 569–576.

Veres, A., Gosis, B.S., Ding, Q., Collins, R., Ragavendran, A., Brand, H., Erdin, S., Talkowski, M.E., and Musunuru, K. (2014). Low Incidence of Off-Target Mutations in Individual CRISPR-Cas9 and TALEN Targeted Human Stem Cell Clones Detected by Whole-Genome Sequencing. Cell Stem Cell *15*, 27–30.

Wang, T., Wei, J.J., Sabatini, D.M., and Lander, E.S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. Science *343*, 80–84.

Wu, X., Scott, D.A., Kriz, A.J., Chiu, A.C., Hsu, P.D., Dadon, D.B., Cheng, A.W., Trevino, A.E., Konermann, S., Chen, S., et al. (2014). Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. Nat. Biotechnol. Epub ahead of print.

Wu, Y., Liang, D., Wang, Y., Bai, M., Tang, W., Bao, S., Yan, Z., Li, D., and Li, J. (2013). Correction of a Genetic Disease in Mouse via Use of CRISPR-Cas9. Cell Stem Cell *13*, 659–662.

Xiao, A., Wang, Z., Hu, Y., Wu, Y., Luo, Z., Yang, Z., Zu, Y., Li, W., Huang, P., Tong, X., et al. (2013). Chromosomal deletions and inversions mediated by TALENs and CRISPR/Cas in zebrafish. Nucleic Acids Res. *41*, e141.

Xiao, A., Cheng, Z., Kong, L., Zhu, Z., Lin, S., Gao, G., and Zhang, B. (2014). CasOT: a genome-wide Cas9/gRNA off-target searching tool. Bioinformatics Epub ahead of print.

Xie, S., Shen, B., Zhang, C., Huang, X., and Zhang, Y. (2014). sgRNAcas9: A Software Package for Designing CRISPR sgRNA and Evaluating Potential Off-Target Cleavage Sites. PLoS One *9*, e100448.

Yang, H., Wang, H., Shivalila, C.S., Cheng, A.W., Shi, L., and Jaenisch, R. (2013). One-step generation of mice carrying reporter and conditional alleles by CRISPR/Cas-mediated genome engineering. Cell *154*, 1370–1379.

Yin, H., Xue, W., Chen, S., Bogorad, R.L., Benedetti, E., Grompe, M., Koteliansky, V., Sharp, P.A., Jacks, T., and Anderson, D.G. (2014). Genome editing with Cas9 in adult mice corrects a disease mutation and phenotype. Nat. Biotechnol. *32*, 551–553.

Zhang, F., Wen, Y., and Guo, X. (2014a). CRISPR/Cas9 for genome editing: progress, implications and challenges. Hum. Mol. Genet. Epub ahead of print.

Zhang, H., Zhang, J., Wei, P., Zhang, B., Gou, F., Feng, Z., Mao, Y., Yang, L., Zhang, H., Xu, N., et al. (2014b). The CRISPR/Cas9 system produces specific and homozygous targeted gene editing in rice in one generation. Plant Biotechnol. J.

Zhang, Y., Heidrich, N., Ampattu, B.J., Gunderson, C.W., Seifert, H.S., Schoen, C., Vogel, J., and Sontheimer, E.J. (2013). Processing-independent CRISPR RNAs limit natural transformation in Neisseria meningitidis. Mol. Cell *50*, 488–503.

FIGURES

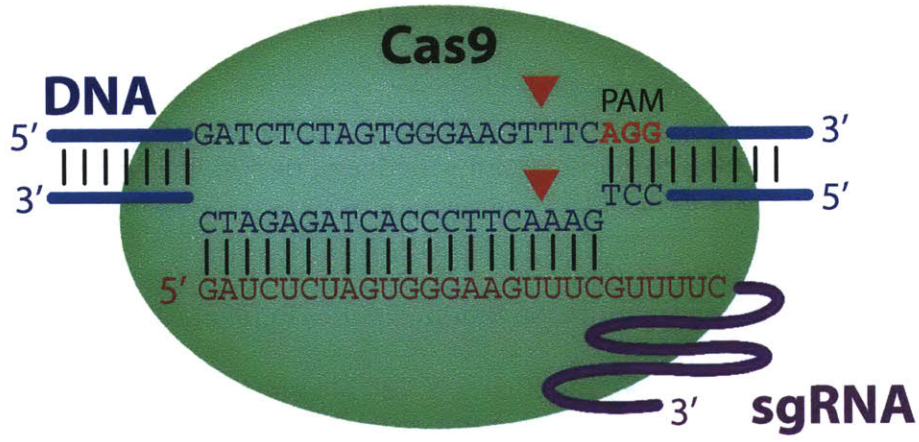


Figure 1: The CRISPR-Cas9 system. The sgRNA (blue) targets the Cas9 protein to genomic sites containing sequences complementary to the 5' end of the sgRNA. The target DNA sequence needs to be followed by a proto-spacer adjacent motif (PAM), typically NGG. Cas9 is a DNA endonuclease with two active domains (red triangles) cleaving each of the two DNA strands three nucleotides upstream of the PAM. The five nucleotides upstream of the PAM is defined as the seed region for target recognition.

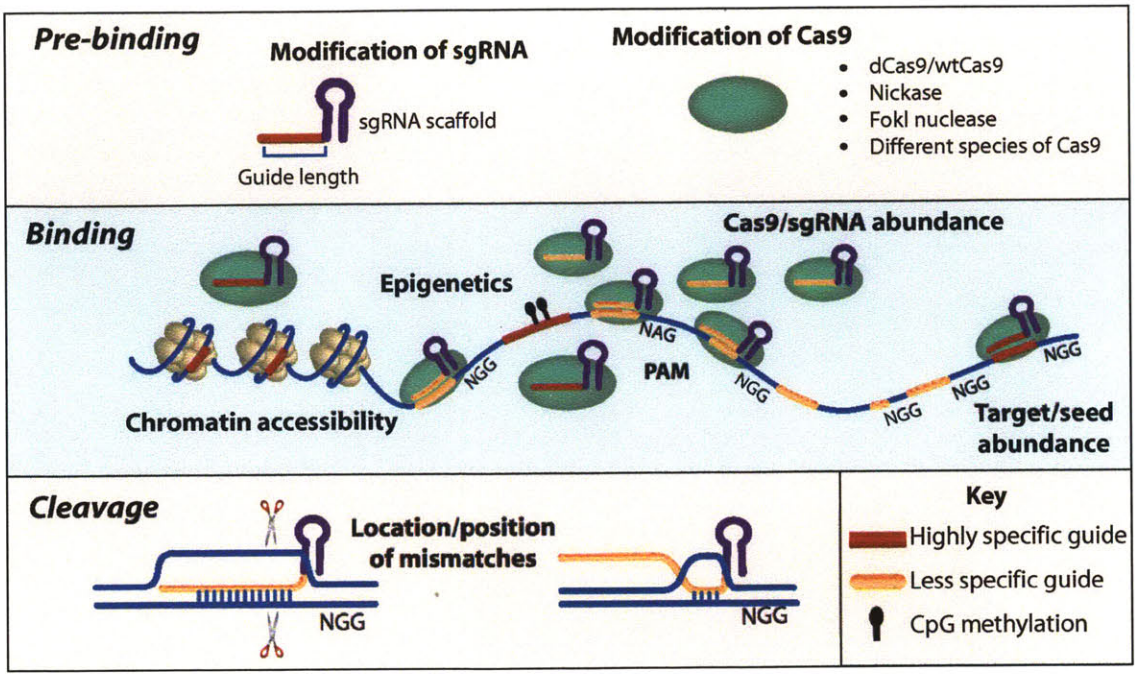


Figure 2: Factors that impact Cas9 specificity. (Top) Before Cas9 is introduced to the system, specificity can be modified by altering the architecture of the single guide RNA (sgRNA) or the Cas9 protein itself. (Middle) At the DNA level, beyond the PAM requirement for binding, closed chromatin and methylated DNA negatively impact Cas9 binding, while increased abundance of Cas9/sgRNA complexes and guide sequences in the genome positively impact Cas9 binding. (Bottom) Although Cas9 can transiently bind DNA that is complementary to only a small seed sequence in the sgRNA, only sequences with extensive complementarity to the guide will be cleaved or direct activation or silencing of targeted genes.

Chapter 5: Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells

In this Chapter, I describe a genome-wide characterization of CRISPR-Cas9 target specificity.

This chapter was published as:

Xuebing Wu, David A. Scott, Andrea J. Kriz, Anthony C. Chiu, Patrick D. Hsu, Daniel B. Dadon, Albert W. Cheng, Alexandro E. Trevino, Silvana Konermann, Sidi Chen, Rudolf Jaenisch, Feng Zhang, Phillip A. Sharp, Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. ***Nature Biotechnology***, 2014, advanced online publication

Supplementary table and data:
http://www.nature.com/nbt/journal/vaop/ncurrent/full/nbt.2889.html#supplementary-information

Contributions
X.W., F.Z. and P.A.S. designed experiments; X.W. and A.J.K. performed most experiments; D.A.S. performed targeted indel sequencing; A.W.C. and D.B.D. cloned the piggyBac dCas9 and sgRNA expressing vectors; A.C.C. generated the dCas9 stable cell line; P.D.H., A.E.T. and S.K. purified Cas9; P.D.H. contributed to in vitro binding assay; S.C. contributed to ChIP experiments with transient transfection. X.W., F.Z. and P.A.S. wrote the manuscript with help from all other authors. R.J., F.Z. and P.A.S. supervised the research.

Abstract

Bacterial type II CRISPR-Cas9 systems have been widely adapted for RNA-guided genome editing and transcription regulation in eukaryotic cells, yet their *in vivo* target specificity is poorly understood. Here we mapped genome-wide binding sites of a catalytically inactive Cas9 (dCas9) from *Streptococcus pyogenes* loaded with single guide RNAs (sgRNAs) in mouse embryonic stem cells (mESCs). Each of the four sgRNAs tested targets dCas9 to tens to thousands of genomic sites, characterized by a 5-nucleotide seed region in the sgRNA, in addition to an NGG protospacer adjacent motif (PAM). Chromatin inaccessibility prevents dCas9 binding to other sites with matching seed sequences, and consequently 70% of off-target sites are associated with genes. Targeted sequencing of 295 dCas9 binding sites in mESCs transfected with catalytically active Cas9 identified only one site mutated above background. We propose a two-state model for Cas9 binding and cleavage, in which a seed match triggers binding but extensive pairing with target DNA is required for cleavage.

Introduction

Many bacterial and archaeal genomes encode clustered regularly interspaced short palindromic repeats (CRISPR), which are transcribed and processed into short RNAs that guide CRISPR-associated (Cas) proteins to cleave foreign nucleic acids(Deveau et al., 2010; Horvath and Barrangou, 2010; Marraffini and Sontheimer, 2010; van der Oost et al., 2009; Terns and Terns, 2011). To target particular genomic loci in eukaryotic cells, the type II CRISPR-Cas system from *Streptococcus pyogenes* has been adapted so that it requires the nuclease Cas9 and one sgRNA(Cong et al., 2013; Hsu et al., 2013; Jinek et al., 2012; Mali et al., 2013a). The first ~20 nucleotides of the sgRNA (the guide region) are complementary to the target DNA site, which also needs to contain a sequence called the protospacer adjacent motif (PAM), typically NGG(Mojica et al., 2009).

The simplicity of targeting any locus with a single protein and a programmable sgRNA has quickly led to widespread use of Cas9(Gasiunas and Siksnys, 2013; Mali et al., 2013b) in applications such as genome editing(Cong et al., 2013; Jiang et al., 2013; Jinek et al., 2013; Mali et al., 2013a; Shalem et al., 2014; Wang et al., 2014), disease gene repair(Schwank et al., 2013; Wu et al., 2013) and knock-in of specific tags(Cong et al., 2013; Dickinson et al., 2013). The catalytically inactive dCas9 (D10A and H840A mutations) alone or when fused to activators or repressors has been used to modulate transcription(Cheng et al., 2013; Gilbert et al., 2013; Maeder et al., 2013; Mali et al., 2013c; Perez-Pinera et al., 2013; Qi et al., 2013) and dCas9 has also been fused to GFP to allow imaging of genomic loci in living cells(Chen et al., 2013).

However, the mechanism of target recognition and target specificity of the Cas9 protein remains poorly understood(Carroll, 2013; Chiu et al., 2013; Cho et al., 2014; Cong et al., 2013; Cradick et al., 2013; Fu et al., 2013; Hsu et al., 2013; Mali et al., 2013c; Pattanayak et al., 2013). Most previous studies have analyzed a set of candidate off-target sites with up to five mismatches to the designed on-target. These studies have examined *in vitro* cleavage, cleavage induced indels or reporter gene expression change as the read-out rather than direct binding(Fu et al., 2013; Hsu et al., 2013; Mali et al., 2013c; Pattanayak et al., 2013). Base pairing in the first 10-12 nucleotides adjacent to PAM (defined as the "seed") was found to be generally more important than pairing in the rest of the guide region(Cong et al., 2013; Jiang et al., 2013; Jinek et al., 2012; Sternberg et al., 2014). However, large variations were observed across target sites, cell types and species regarding the importance of base pairing at each position(Carroll, 2013). Some studies have shown that Cas9 is highly specific(Chiu et al., 2013; Cho et al., 2014; Gilbert et al., 2013), whereas other studies have demonstrated substantial Cas9 off-target activity(Cradick et al., 2013; Fu et al., 2013; Hsu et al., 2013; Mali et al., 2013c; Pattanayak et al., 2013). Epigenetic features such as CpG methylation and chromatin accessibility have been reported to have little effect on targeting(Hsu et al., 2013; Perez-Pinera et al., 2013).

To our knowledge, there has been no previous report of genome-wide binding maps of dCas9. Our data reveal a well-defined seed region for target binding and a very large number of off-target binding sites, most of which do not seem to undergo substantial cleavage by Cas9. Our observations explain some of the

previously observed heterogeneity, provide insights into target recognition and the cleavage process and could guide future target design.

Results

Genome-wide binding of dCas9-sgRNA

To map dCas9 *in vivo* binding sites, we generated mESCs with a stably integrated vector encoding HA-tagged dCas9 (Fig. 1a), and performed chromatin immunoprecipitation followed by sequencing (ChIP-seq) with cells transfected with either no sgRNA or one of each of 4 sgRNAs (Phc1-sg1, Phc1-sg2, Nanog-sg2 and Nanog-sg3) targeting the promoters of *Phc1* or *Nanog*, respectively. For each sgRNA, we observed ~100 fold enrichment for dCas9 at the on-target site compared to flanking regions, and the spatial resolution is sufficient to distinguish between two binding sites separated by 22 base pairs (bps) (Nanog-sg2 and Nanog-sg3) (Fig. 1b).

Using the standard ChIP-seq peak-calling procedure MACS(Zhang et al., 2008) – comparing immunoprecipitated material and input (whole cell extract) DNA – we identified between 2,000 and 20,000 peaks in each sequencing library (Supplementary Fig. 1a). Cells expressing dCas9 but not transfected with sgRNAs (dCas9-only ChIP) exhibited 2,115 peaks. Most (77%) of the peaks detected in the dCas9-only ChIP were also detected in libraries prepared from dCas9-sgRNA immunoprecipitations (Supplementary Fig. 1b). The peaks in dCas9-only ChIP were enriched in open chromatin regions (Supplementary Fig. 2a) and 41% contained GG/CC-rich motifs that closely resemble CTCF binding motifs (Supplementary Fig. 2b-d). Such peaks could either represent 'sampling' by dCas9 of accessible sites

containing NGG(Sternberg et al., 2014), or transcription-dependent artifacts as previously reported for GFP ChIP in yeast(Teytelman et al., 2013).

To identify sgRNA-dependent dCas9 binding sites, we matched sequencing depth by randomly sampling an equal number of reads from all six libraries and then performed pair-wise peak calling with MACS using each of the other five libraries as the control; we only retained peaks that were enriched over all five controls (Fig. 1c). Only 3 background peaks were called using this approach for dCas9-only ChIP. The number of sgRNA-specific peaks varied substantially; for example there were nearly 6,000 peaks for Nanog-sg3 but only 26 peaks for Nanog-sg2 (Fig. 1b). Many of the off-target peaks showed high binding levels, as defined by the peak height relative to on-target peaks after subtracting dCas9-only reads. For example, there were 91 off-target peaks with more than 50% of the binding level of the on-target site for Nanog-sg3 (Supplementary Table 1). These results suggest that there are substantial numbers of off-target binding sites and the majority of the dCas9-sgRNA complex binds outside the designed target site.

A 5-nucleotide seed for dCas9 binding

Sequence motifs enriched within 50 bps of peak summits were identified using MEME-ChIP(Machanick and Bailey, 2011). The top motif found for each ChIP library matched the PAM-proximal region of the transfected sgRNA plus the PAM NGG (Fig. 2a and Supplementary Fig. 3). For 3 of the 4 sgRNAs, only PAM-proximal positions 1 to 5 in the target DNA showed preference of base match to the guide (Fig. 2a). We therefore define position 1-5 as the 'seed' region of the sgRNA. For Nanog-

sg2, the guide match extends to about 10-12 bases to the 5' end, possibly due to the presence of multiple Us in the seed that lowers the thermodynamic stability of the sgRNA-DNA interaction. For Nanog-sg3 and Phc1-sg2, exact match to the 5-nucleotide seed followed by NGG (seed+NGG) within 50 bps of peak summits explained 96% and 97% of the peaks, respectively. When the 50 nucleotides flanking peak summits were shuffled preserving dinucleotide frequency, less than 5.7% of the shuffled sequences contained seed+NGG (Fig. 2a) for all four sgRNAs. Moreover, the seed+NGG sites were highly enriched at the center of the peak (Fig. 2a, right), suggesting the target sites identified are directly bound by sgRNA-guided dCas9.

We found that seed+NGG alone is sufficient for Cas9 binding *in vivo* and *in vitro*. For example, there were 92 peaks in the Nanog-sg3 sample containing only seed+NGG matches, i.e. mismatches at all the other 15 positions. The strongest peak containing only seed+NGG showed 52% binding activity relative to the on-target (Fig. 2b). *In vitro* gel shift assays confirmed specific binding to seed+NGG only substrates but with lower affinity than the on-target site (Fig. 2c).

The peak motif analysis (Supplementary Fig. 3) revealed no enrichment of binding at seed sites followed by NAG, an alternative PAM previously reported to function in Cas9-mediated cleavage(Hsu et al., 2013; Jiang et al., 2013; Pattanayak et al., 2013). For example, of all 996 (33%) Phc1-sg1 ChIP peaks without seed+NGG sites, only 18 had seed+NAG within 50-bp of the peak summit, even less than expected by chance (Supplementary Fig. 4). ChIP-seq in human HEK293FT cells transfected with dCas9 and the same sgRNAs used in a previous study(Hsu et al.,

2013), where NAG cleavage was reported, also failed to detect binding at those NAG off-target sites (Supplementary Fig. 5). *In vitro* we observed >10 fold decrease in affinity when NGG was mutated to NAG in the on-target substrate (Supplementary Fig. 6). Our *in vivo* and *in vitro* binding data are consistent with previous *in vitro* cleavage data showing that NAG or other variants rarely function as PAMs under enzyme-limiting conditions(Pattanayak et al., 2013).

Chromatin accessibility is the major determinant of in vivo binding

There are hundreds of thousands of seed+NGG sites in the genome for each sgRNA, for example, 621,651 for Nanog-sg3. To understand why only a small fraction of sites (<1%) were bound, we first looked for a correlation between the number of base match to the 20-nucleotide guide region and binding levels of ChIP peaks. Overall the correlation was very weak (Pearson correlation coefficient r=0.03, 0.12, 0.15 and 0.55 for Nanog-sg3, Phc1-sg2, Phc1-sg1 and Nanog-sg2, respectively (Fig. 3a and Supplementary Fig. 7)).

To identify determinants influencing *in vivo* binding, we applied a linear regression model of a set of sequence (mono- and di-nucleotide frequency), structural (melting temperature, DNA energy and flexibility(Packer et al., 2000)) and epigenetic (chromatin accessibility as assayed by DNase I hypersensitivity (DHS)(Stamatoyannopoulos et al., 2012) and DNA CpG methylation(Stadler et al., 2011)) features around the seed+NGG sites for each sgRNA (Online Methods). We found that chromatin accessibility (DHS) is the strongest indicator of binding *in vivo*, explaining up to 19% of the variation in binding when considering all individual

seed+NGG sites in the genome (Fig. 3b). The difference in the number of seed+NGG sites in DHS peaks (i.e. accessible seed+NGG sites) explained 92% of the variation in the number of dCas9 peaks among the four sgRNAs (Fig. 3c, n = 4, $p<0.05$, F-test). Although this is based on a limited set of sgRNAs, it suggests that it might be possible to predict the approximate number of off-target peaks based on the seed sequence in cell types where chromatin accessibility data are available.

Previous data suggested that Cas9 cleavage activity is not affected by DNA CpG methylation(Hsu et al., 2013). However, for the 17% of seed+NGG sites in the genome that contain CpG dinucleotides within the 20mer guide match and NGG, CpG methylation became the strongest predictor of dCas9 binding and negatively correlated with binding (Fig. 3d, Supplementary Fig. 8a-b). In a regression model, adding CpG methylation to DHS for sites containing CpGs almost doubled the amount of variation explained (Supplementary Fig. 8c). Our data suggests that CpG methylation likely reflects an aspect of chromatin accessibility not fully captured by DHS or that when combined with extensive mismatches, CpG methylation may impede binding.

The correlation with chromatin accessibility suggested that dCas9 off-target binding might preferentially occur at active genes. For Nanog-sg3, 70% of the off-target sites were associated with genes, including 18% in promoter region (< 2 kb upstream of gene TSS), 6% near enhancer regions and 46% within genes (Fig. 3e). For example, an off-target peak that co-localized with the *Dusp19* gene TSS and a DHS peak showed 74% binding relative to the on-target with only 7 base matches to Nanog-sg3 (Fig. 3f).

Seed sequences influence sgRNA abundance and specificity

The Nanog-sg2 sgRNA had substantially fewer off-target binding sites than predicted by accessible seed+NGG sites (Fig. 3c). Although the same amount of sgRNA plasmids were transfected, the abundance of Nanog-sg2 was more than seven fold lower than the other three sgRNAs as determined by Northern blot (Fig. 4a). The same pattern of sgRNA abundance was observed when cells were transfected with sgRNA expression plasmids without co-transfecting dCas9, although all four sgRNAs showed substantially decreased levels of abundance, consistent with previous reports that Cas9 stabilizes sgRNA in cells(Jinek et al., 2013).

To test if sgRNA abundance influences the number of off-target sites bound, we repeated the ChIP experiments after transfection with various amounts of sgRNA plasmids. Northern blots confirmed the decrease in sgRNA when less plasmid was transfected (Fig. 4a) and we identified decreased numbers of peaks with decreased amounts of plasmid (Fig. 4b). When the level of Nanog-sg3 was reduced to a similar level as Nanog-sg2 (Fig 4a, comparing lane 13 to lanes 16 and 17), the number of peaks for Nanog-sg3 was still much higher than for Nanog-sg2, presumably due to the presence of more accessible Nanog-sg3 seed+NGG sites in the genome (Fig. 3c). When 0.02 μg plasmid was transfected, Nanog-sg3 RNA was barely detected (lane 14); the 122 peaks identified in this library showed low overlap (9%) with our previous Nanog-sg3 ChIP, suggesting these were mostly non-specific signals.

A comparison of the seed regions of the four sgRNAs suggested that UUU in the seed of Nanog-sg2 might be responsible for decreased sgRNA abundance and increased specificity, consistent with recent observation that U in PAM-proximal position 1-4 leads to low gene knockout efficacy(Wang et al., 2014). Indeed, two mutations (U to G and U to A) in the Nanog-sg2 seed region that converted the seed (GUUUC) to the same sequence as the Phc1-sg2 seed (GGUAC), led to higher levels of sgRNA (sgRNA N2b in Fig. 4c). Considering the presence of GUUUUA adjacent to the seed and because sgRNAs are transcribed by RNA polymerase III which is terminated by U-rich sequences (Nielsen et al., 2013; Orioli et al., 2011), we speculate that together with the downstream U-rich region, multiple Us in the seed might induce termination of sgRNA transcription. Consistent with this, three sgRNAs with seeds UUAUU, ACUUU and UUUUU also showed very low abundance (Fig. 4c, sgRNA P3, N5 and N6). When GUUUC was placed upstream of the seed thus away from GUUUUA in the sgRNA, the sgRNA was well expressed (sgRNA C4 in Fig. 4c).

One of 295 off-target sites is mutated above background

To test if binding correlates with Cas9 nuclease-induced mutation, we examined the indel frequencies of the four on-target sites and 295 selected off-target sites by targeted PCR and sequencing (Hsu et al., 2013). These sites were selected to cover a broad range of binding levels and numbers of mismatches to the sgRNA. We ranked all peaks by binding (background subtracted read counts) and for each binding level selected a peak with the fewest mismatches and another peak with most mismatches to the guide.

We determined the indel frequency of the 299 selected binding sites in wild type mESCs transfected with active Cas9 and each of the four sgRNAs, for three independent biological replicates (Supplementary Table 3). The level of Cas9 protein transiently expressed in the cells was 2.6 fold higher than in cells with stably integrated dCas9 used for ChIP (Supplementary Fig. 9a, comparing lane 1 to lane 8). The same ChIP and peak calling procedures in cells transiently transfected with dCas9 identified 2.7 times more Nanog-sg3 peaks (16,119 versus 5,957 in dCas9 stable cell lines), including 96% (85) of the 89 peaks selected for indel analysis. The amount of Cas9/dCas9 plasmids used for transfection is similar to levels used for genome editing applications by the field (Supplementary Fig. 9b).

Using our previously validated model (Hsu et al., 2013), the background indel frequencies due to sequencing noise were determined for each individual target using two biological replicates transfected with only Cas9 but no sgRNA ("control"). Importantly the control samples showed no evidence of targeted mutations by Cas9 (note that background indels in the absence of Cas9 might also occur). We manually reviewed sequencing alignments of all loci with indel frequencies above 0.03%. We found 12% to 37% sequencing reads from the on-target sites contained indels, yet only one off-target, which was from Nanog-sg2, was mutated at a frequency of 0.7% (Fig. 5). There was no detectable correlation between binding and indel frequency (sites in Fig. 5 are ranked by decreasing binding from left to right for each sgRNA). The selected sites include 7 of the top 10 (including all the top 6) and 36 of the top 50 Nanog-sg3 binding sites with the strongest ChIP signals, and 4 of the 8 Nanog-sg3

off-target binding sites that have fewer than four mismatches to the sgRNA; none of these off-target sites showed cleavage significantly above the background level.

Discussion

We have shown that dCas9 binding is more promiscuous than previously thought. The low binding specificity is explained by the limited requirement for an accessible match to a 5-nucleotide seed followed by an NGG PAM. The position of the seed region next to PAM is consistent with previous observations that base pairing near PAM is critical for targeting (Cong et al., 2013; Jiang et al., 2013; Jinek et al., 2012; Sternberg et al., 2014), but the seeds we identified for 3 of the 4 sgRNAs are shorter than those previously reported; seed lengths of 8-13 nucleotides have been described as required for cleavage by Cas9 (Cong et al., 2013; Jiang et al., 2013; Jinek et al., 2012; Sternberg et al., 2014).

The seed sequence influences the specificity of Cas9-sgRNA binding in several ways. Firstly, seed composition determines the frequency of a seed+NGG site in the genome. Secondly, seed composition determines how likely a seed+NGG site will be in open chromatin. Thirdly, seed composition affects sgRNA abundance, probably at the level of transcription, and thus the effective concentration of Cas9-sgRNA complex. Lastly, seed composition may also affect loading into Cas9 and again tune the level of functional Cas9 (Wang et al., 2014). Through all four mechanisms U-rich seeds are likely to increase target specificity.

Our results suggest that applications based on dCas9 or dCas9-effector fusions, such as transcription modulation, imaging, and epigenome editing, could be

complicated by substantial off-target binding. Previous studies suggest that several sgRNAs targeting the same gene are frequently necessary for gene activation (Cheng et al., 2013; Mali et al., 2013c; Perez-Pinera et al., 2013); this could potentially reduce off-target effects due to the requirement of co-targeting. However, the use of multiple sgRNAs increases the number of potential off-target binding sites, which might complicate interpretation. Although we only detected indels at a low frequency (0.7%) above background for one off-target binding site among 295 selected sites, 295 is a small fraction of all possible binding sites and may not be representative of the complete off-target mutation profile of each sgRNA. This is an important question as low frequencies of indels could complicate certain CRISPR-Cas9 applications, such as genome-wide screening that involves selective growth (Shalem et al., 2014; Wang et al., 2014). Therefore, to minimize the likelihood of false positive screening hits resulting from off-targeting, we recommend using multiple guide RNAs to target each gene and using the concordance among multiple guides to interpret screening results. We further note that although binding sites with NAG PAMs are not enriched in the ChIP data, a previous study has shown that NAG-flanked genomic loci can contribute to off-target indel-mutations. Therefore, unbiased and more sensitive detection of genome-wide mutations will be needed to determine Cas9 cutting specificity.

The observation that most of the sites bound by Cas9 do not seem to have substantial cleavage is reminiscent of the eukaryotic Argonaute-microRNA system, in which most target mRNAs bearing partial microRNA match are bound without cleavage and only a few targets with extensive pairing are cleaved (Bartel, 2009).

We propose a two-state model (Fig. 6) similar to the Argonaute-microRNA system, in which pairing of a short seed region triggers binding after PAM recognition and subsequent DNA unwinding. Targets with only seed complementary remain bound by Cas9 without cleavage; only those with extensive pairing undergo efficient cleavage. This suggests a conformation change between binding and cleavage as observed for Argonaute-microRNA complexes (Bartel, 2009; Jinek and Doudna, 2009). While this paper was under review, a pair of Cas9 structural studies were published (Jinek et al., 2014; Nishimasu et al., 2014), including a crystal structure of dCas9 in complex with sgRNA and target DNA, which not only supports our observation of a PAM-proximal 5-nucleotide seed but also suggests a large conformation change during the inactive-active state transition (Nishimasu et al., 2014).

Methods

Oligonucleotides

All oligonucleotides used in this study were purchased from Integrated DNA Technologies. Sequences are listed in Supplementary Table 2.

Cloning

A two-step fusion PCR was used to amplify Cas9 nickase ORF from pX335 vector (Addgene: 42335) and incorporate H840A mutation to create a nuclease deficient Cas9 (dCas9). This PCR product was inserted into the Gateway donor backbone pCR8/GW/TOPO to create pAC84 (Addgene: 48218). The dCas9 ORF in

pAC84 was then transferred to a piggyBac-based destination vector pAC150 (PB-Lox-HygroR-Lox-4xHSInsulators-EF1a-DEST) by LR Clonase reaction (Invitrogen) to create pAC159 (PB-LHL-4xHS-EF1a-dCas9). The sgRNA expression cassette was amplified by PCR from pX335 vector and cloned into a piggyBac vector pAC158 (PB-neo-4xHSInsulators) to create pAC103 (PBneo-sgExpression). sgRNA was then cloned into BbsI-digested pAC103 by oligo cloning method as described previously (Cong et al., 2013). Cas9 transient transfection constructs consisted CBh-driven WT-Cas9 or Cas9-D10AH840A (dCas9) containing a C-terminal HA-tag.

Cell culture

V6.5 mouse embryonic stem cells (mESCs) were cultured in DMEM supplemented with 10% FBS, pen/strep, L-glutamine, nonessential amino acids and LIF. For generation of cells stably integrating dCas9, cells were transfected in a 6-well and selected using Hygromycin B at 100 ug/mL 24 hours post transfection, then raised to 150 ug/mL 48 hours post transfection. Cells were split onto 10 cm plates, and single clones were isolated, expanded, and used for all experiments described. HEK293FT cells were cultured as previously described (Hsu et al., 2013). All transfection were done with Lipfectamine 2000 (Invitrogen).

ChIP

Three million cells were seeded on to 10cm plates on day 1, transfected with sgRNAs plasmids (or together with HA-dCas9 plasmids) on day 2, transferred to 15cm plates on day 3, and crosslinked on day 4 with roughly 50 million cells.

Crosslinking is done by adding 2 mL (0.1 volume) 37% formaldehyde to the plate, incubating at room temperature for 15min, and quenched by adding 1 mL 2.5M glycine. Cells were rinsed twice with cold PBS and scraped to collect in cold PBS. Cells were centrifuged at 1,350g for 5min at 4°C and washed again in cold PBS. Cells were flash frozen in a dry ice/ethanol mix and stored at -80°C. Cell pellet was resuspended in 5mL cold Lysis Buffer 1 (50mM HEPES-KOH pH 7.5, 140mM NaCl, 1mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Trition X-100, 1x Roche complete protease inhibitors), rotated at 4°C for 10min followed by centrifugation at 1,350g for 5min at 4°C. Pellet was resuspended in 5mL Lysis Buffer 2 (10mM Tris-Cl pH 8, 200mM NaCl, 1mM EDTA, 0.5mM EGTA, 1x Roche complete protease inhibitors), rotated at 4°C for 10min followed by centrifugation at 1,350g for 5min at 4°C. Nuclear pellet was resuspended in 2mL Sonication Buffer (20mM Tris-Cl pH 8, 150mM NaCl, 2mM EDTA, 0.1% SDS, 1% Triton X-100, 1x Roche complete protease inhibitors) and sonicated (60min total time, 30sec on, 30sec off, in 6 rounds of 10min) in a Bioruptor (Diagenode). Lysate was centrifuged in eppendorf tubes in a microfuge at 4°C a max speed for 20min. Supernatant was collected and 50mL of this was saved as input. Protein G Dynabeads were conjugated to 5 ug rabbit anti-rat antibody (Thermo) in 0.1M Na-Phosphate pH 8 buffer at 4°C with rotation followed by conjugation to 5 ug HA antibody (Roche 3F10, #11867431001). Beads were resuspended in 50 ul Sonication Buffer and added to samples to immunoprecipitate overnight. The next day, beads were washed twice in sonication buffer, once in sonication supplemented with 500mM NaCl, once in LiCl Buffer (10mM Tris-Cl pH 8, 250mM LiCl, 1mM EDTA, 1% NP-40) and once in TE + 50mM NaCl. Each wash was

accomplished with rotation at 4°C for 5min. Chromatin was eluted at 65°C for 15min in Elution Buffer (50mM Tris-Cl pH 8, 10mM EDTA, 1% SDS). Input was combined with elution buffer and both input and IP crosslinks were reversed at 65°C overnight. RNA was digested with RNAse A at 0.2mg/ml final concentration (Sigma) at 37 °C for 2hrs and protein was digested with proteinase K at 0.2mg/mL final concentration (Life Technologies) at 55 °C for 45min. DNA was phenol:chloroform:isoamyl alcohol (Life Technologies) extracted and ethanol precipitated. Barcoded libraries were prepared and sequenced on Illumina HiSeq2000.

ChIP-seq data analysis

Reads were de-multiplexed and mapped to mouse genome mm9 using bowtie (Langmead et al., 2009), requiring unique mapping with at most two mismatches (-n 2 -m 1 --best --strata). Mapped reads were collapsed and the same number of reads (about 9 million) was randomly sampled from each library to match sequencing depth. Peaks were called using MACS (Zhang et al., 2008) with default settings. For each sample, the other samples are each used as a control and only peaks called over all five controls are defined as target sites. To quantify relative binding strength, reads were first extended at the 3' end to the average fragment length (d) estimated by MACS, and then the number of fragments (extended reads) overlapping with the seed+NGG region is counted and normalized by subtracting counts from dCas9-only control. If multiple seed+NGG match sites were found, the one with the highest relative binding was assigned to the peak.

Analysis on determinants of binding

Mouse ES cell DNase Hypersensitivity data (bigwig file and narrow peak file) were downloaded from UCSC genome browser hosting the mouse ENCODE project (Stamatoyannopoulos et al., 2012). DNA CpG methylation data was downloaded from GEO dataset GSE30202. Melting temperature (Tm) was calculated using the *oligotm* program in *primer3* version 2.3.6. DNA stability and flexibility were calculated using a table of tetranucleotide scores derived from X-ray crystal structures in a previous study (Packer et al., 2000). The linear regression is performed by using the *lm* function in *R*, one feature a time to calculate the R^2 value for each feature.

Northern blot

Total RNA was isolated using TRIzol (Life Technologies) and 5 ug of total RNA was loaded on 8% denaturing PAGE. Northern blot was done as previously described (Hsu et al., 2013), using a probe targeting the scaffold shared by all sgRNAs.

Protein Purification

Human codon-optimized Cas9 (Addgene plasmid 42230) was subcloned into a custom pET-based expression vector with an N-terminal hexahistidine (6xHis) tag followed by a SUMO protease cleavage site. The fusion construct was transformed into E. coli Rosetta 2 (DE3) competent cells (Millipore), grown in LB media to

OD600 0.6, and induced with 0.2 mM IPTG for 16h at room temperature. Cells were pelleted, resuspended and washed with Milli-Q H2O supplemented with 0.2 mM PMSF, and lysed with lysis buffer (20 mM Trizma base, 500 mM NaCl, 0.1% NP-40, 2 mM DTT, 10 mM imidazole). The lysis buffer was supplemented with protease inhibitor cocktail (Roche) immediately prior to use. Whole lysate was sonicated at 40% amplitude (Biologics Inc., 2s on, 4s off) prior to ultracentrifugation (30,000 rpm for 45m). The clarified lysate was applied to cOmplete His-tag purification columns (Roche), washed with wash buffer 1 (20 mM Trizma base, 500 mM NaCl, 0.1% NP-40, 2 mM DTT, 10% glycerol, 10 mM imidazole) and wash buffer 2 (20 mM Trizma base, 250 mM NaCl, 0.1% NP-40, 2 mM DTT, 10% glycerol, 50 mM imidazole). The 6xHis affinity tag was released via SUMO protease cleavage and bound protein was eluted with a linear gradient of 150 mM – 500 mM imidazole. Eluted protein was concentrated with Amicon centrifugal filter units with Ultracel membrane (Millipore) and stored at -80°C.

In vitro transcription

A T7 promoter forward oligo was annealed to an sgRNA template oligo by heating to 95°C for 3 min in 1x T4 DNA ligase buffer and then cooled at room temperature for 30 min. The annealed product was used as template and transcribed with MEGAshortscript T7 Kit (Life Technologies). RNAs were purified by MEGAclear Kit (Life Technologies) and frozen at -80 °C.

Gel shift assay

Single stranded DNA oligos of 50 nucleotides were purchased from IDT and PAGE purified. Double stranded substrate were generated by mixing 100 pmol each strand in water (10 ul total), heating to 95°C for 3 min and cool to room temperature. The substrates were then 5' end labeled with [γ-32P]-ATP using T4 PNK (New England Biolabs) for 30 min at 37°C, and free ATP removed by G-25 column (GE Healthcare). For each reaction, 100 nM Cas9 was mixed with a 1:4 dilution series of sgRNA (from 0 to 100 nM) in 1x NEBuffer 3 at 37°C for 10 min, and then about 0.5 nM labeled substrate oligos were added and incubated for 5 min at 37°C in a 10 ul reaction. Reactions were stopped on ice and added 1/2 volume of 50% glycerol. Samples were loaded on to 12% native PAGE and run at 300V for 2 hours at room temperature. Gels were visualized by phosphorimaging. Gel quantification is done with ImageJ. The fraction bound shown in Fig. 2c was calculated as the ratio of intensity from the specific binding band to the total intensity of the entire lane.

Targeted sequencing and indel detection

For replicate 1, cells were seeded in 6-well plates (300,000 cells per well), transfected with 2 ug sgRNA plasmid, 2 ug Cas9 plasmid, using 10 ul Lipofectmine 2000 reagent per sample for 3 hours. For replicate 2 and 3, 50% more plasmids were used. DNA was extracted and selected target sites were PCR amplified, normalized, and pooled in equimolar proportions. Pooled libraries were denatured, diluted to a 14pM concentration and sequenced using the Illumina MiSeq Personal

Sequencer (Illumina). Sequencing data was demultiplexed using paired barcodes, mapped to reference amplicons, and analyzed for indels as described previously (Hsu et al., 2013).

Figures

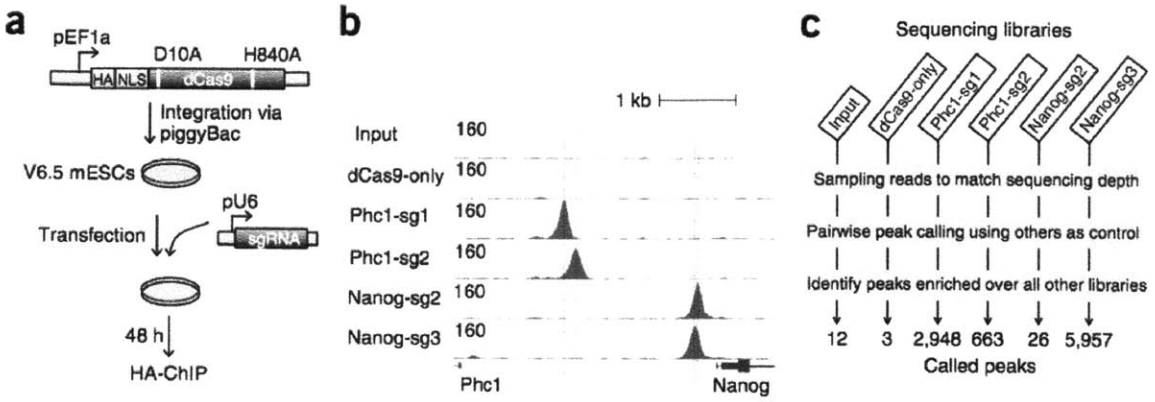


Figure 1: Genome-wide *in vivo* binding of dCas9-sgRNA. (**a**) Schematic of dCas9 ChIP. EF1a promoter-driven HA-tagged dCas9 with nuclear localization signal (NLS) is integrated into the genome of mESCs via the piggyBac system. Plasmids containing U6 promoter-driven sgRNAs were transfected and ChIP was carried out two days later with HA antibody. (**b**) ChIP signals (normalized read counts) around on-target sites. Vertical dashed lines indicate designed target sites (the region complementary to the sgRNA). (**c**) Peak calling. Reads were sampled from each library, and peaks were called using each other library as a control (Online Methods). Only peaks called over all other five controls were retained. The numbers at the bottom indicate the numbers of peaks called for each library using these criteria.

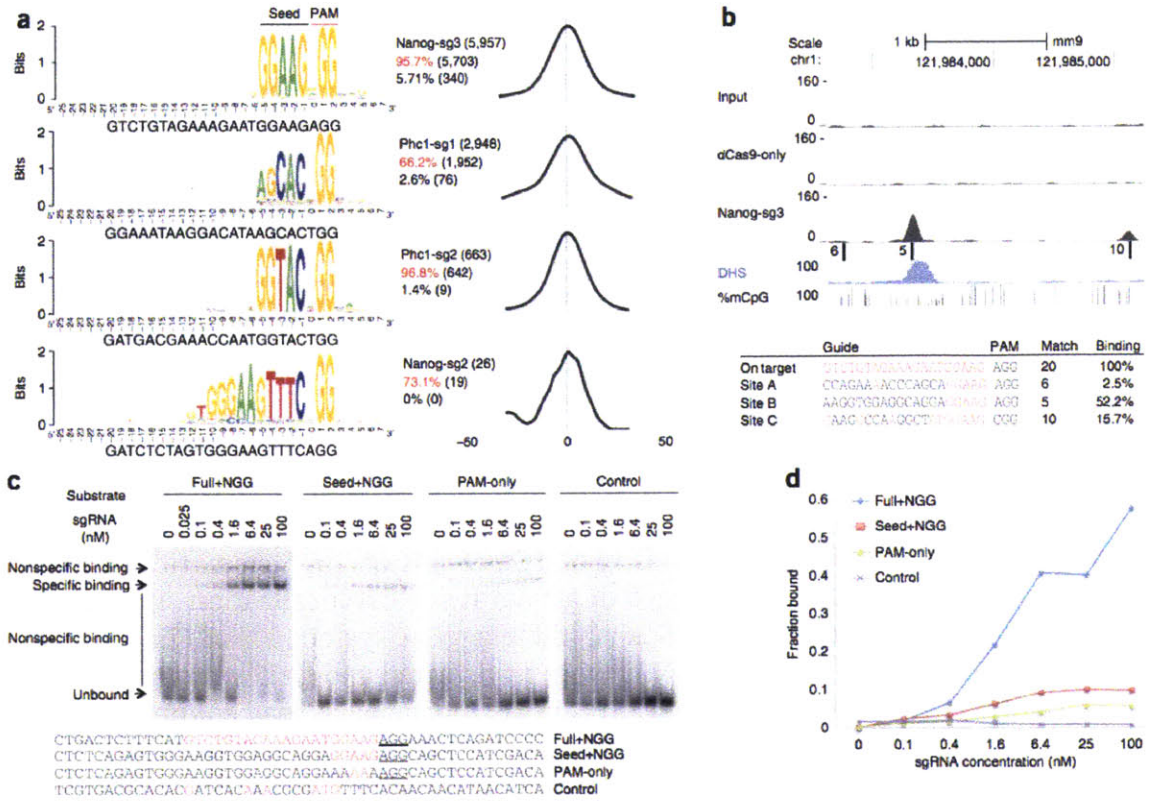


Figure 2: A 5-nucleotide seed for dCas9 binding. (**a**) Most peaks are associated with seed+NGG matches. The best match to the sgRNA followed by NGG within 50bp flanking peak summits were aligned to generate the sequence logo using WebLogo (Crooks et al., 2004). The text to the right of the logos indicates the total number of peaks (top line), percentage and number of peaks with exact 5-nucleotide seed+NGG match within 50 bps of peak summits (middle line, in red) or when the 100 nucleotides sequence were shuffled while maintaining dinucleotide frequency (bottom line). The distribution of the exact seed+NGG match relative to the peak summit was shown on the right (the numbers indicate nucleotide positions) (**b**) Example of binding at seed+NGG only sites. On the top are six tracks: Input, dCas9-only IP, and Nanog-sg3 IP read density, seed+NGG sites (position indicated by bars, named as A/B/C, and the numbers to the left indicates the number of matches to the guide), DHS read density and fraction of methylated alleles at CpG sites. Below are the target sequences, PAM, number of matches to the sgRNA and relative binding at each site. Guide-matched bases are in red. (**c**) Gel shift assay for 50 bp double-stranded DNA substrates with sequences matching the Nanog-sg3 on-target site ("Full+NGG") and a seed+NGG only off-target site ("Seed+NGG", site B in Fig 2b). "PAM only" is the "Seed+NGG" substrate with a mutated seed. The negative control substrate ("Control") was designed to contain no NGG or NAG. Complete substrate sequences are shown at the bottom, with PAM underlined and guide-matched bases in red. (**d**) The quantification of the gels in (**c**). Shown is the percentage of the specific binding band relative to the entire lane at each sgRNA concentration.

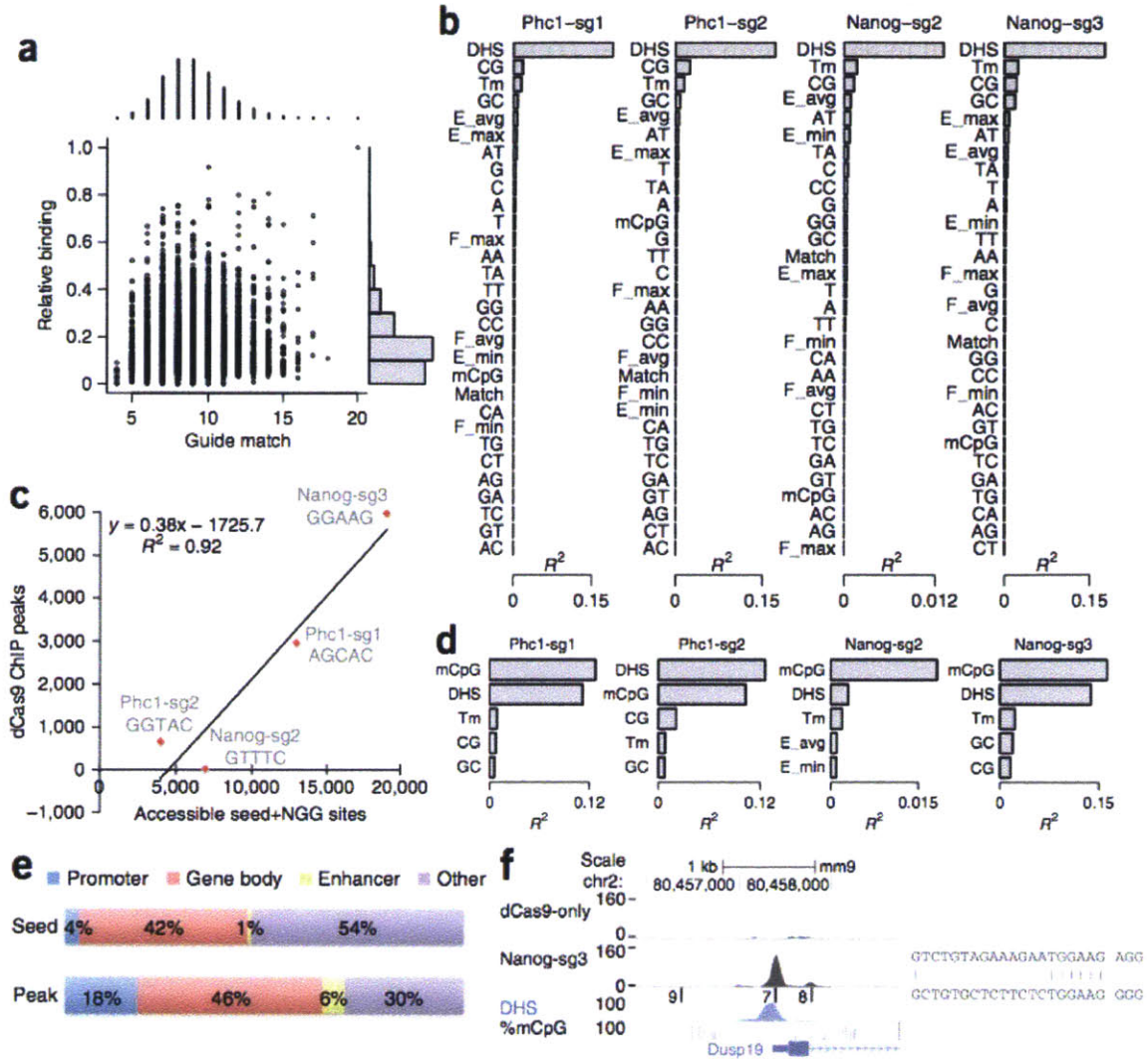


Figure 3: Chromatin accessibility is the major determinant of binding *in vivo*. (**a**) Scatter (center) and histogram (top and right) plots of the number of matches to the sgRNA guide region (x-axis) and binding relative to the on-target site (y-axis) for all Nanog-sg3 peaks. Relative binding levels (0 to 1) are divided into 10 equal bins and the number of peaks in each bin is shown on the right of the scatter plot. (**b**) Ranking of features based on R^2, the percent of variation in binding explained by each feature in a linear regression model (using R, one feature a time). DHS: DNase I hypersensitivity read density; Tm: melting temperature; Match: number of bases that match the sgRNA; E (F)_min/max/avg: minimum, maximum, and average tetranucleotide energy (flexibility) score within the guide+NGG region; A/C/G/T or their combination indicates mono- and di-nucleotide frequency in the guide+NGG region; %mCpG: average fraction of methylated CpG in the guide+NGG region. (**c**) Scatter plot and linear regression between the number of dCas9 ChIP peaks and the number of accessible seed+NGG sites (i.e. sites overlapping with DHS peaks). (**d**) As for (**b**) but only plotting the top five features after regression was done using sites

containing CpG dinucleotides. (**e**) Off-target peaks are preferentially associated with genes. Shown is the percentage of Nanog-sg3 seed+NGG sites (top) or ChIP peaks (bottom) that fall in each region category. (**f**) Example of off-target binding at the *Dusp19* promoter. Tracks are the same as Fig. 2b. On the right is the alignment of the off-target site with 7 matches (bottom) to the guide sequence (top).

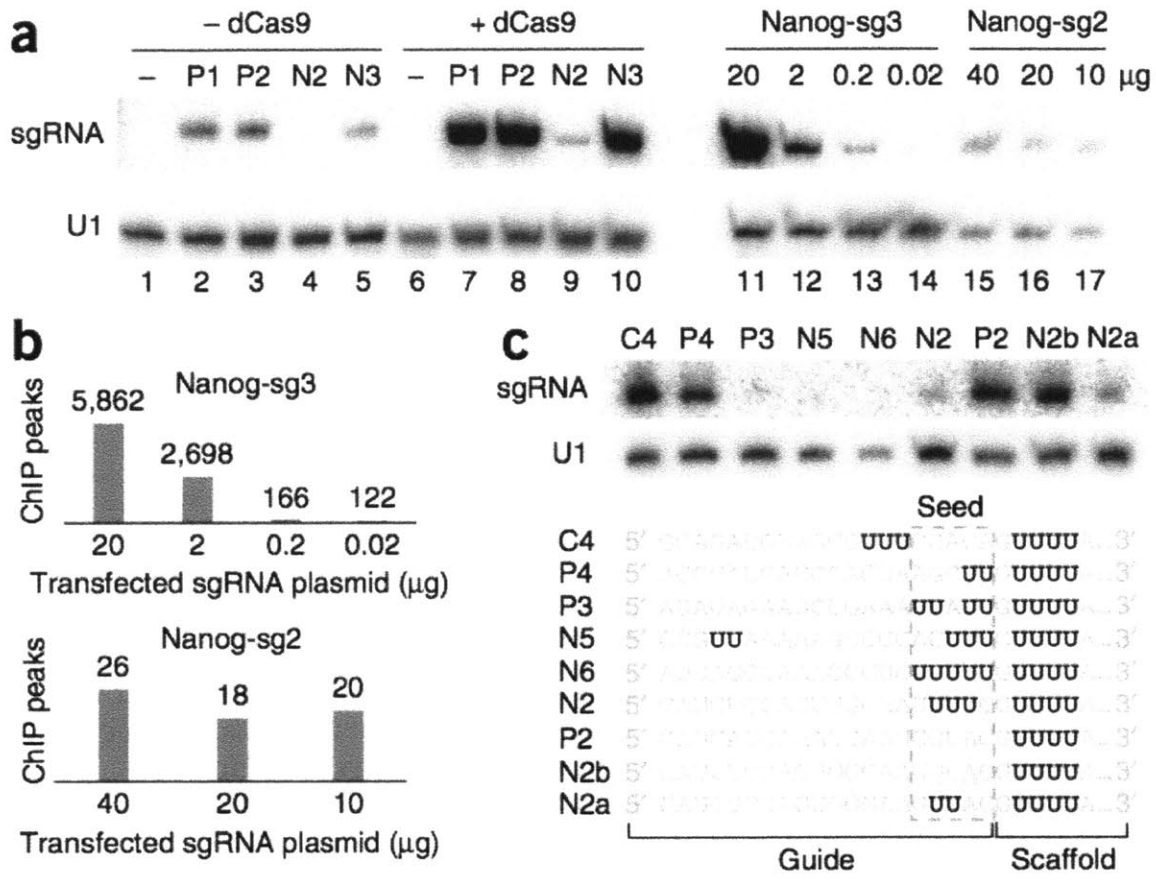


Figure 4: Seed sequences influence sgRNA abundance and specificity. (**a**) Northern blot showing the abundance of sgRNAs. Lanes 1-10: from cells transfected with dCas9 (lanes 6-10) or without dCas9 (lanes 1-5), and with either no sgRNA (lanes 1 and 6) or one of the four sgRNAs (P1: Phc1-sg1; P2: Phc1-sg2; N2: Nanog-sg2; N3: Nanog-sg3). Lanes 11-14: Nanog-sg3 abundance from dCas9-mESCs transfected with 20, 2, 0.2 or 0.02 µg Nanog-sg3 plasmid. Lanes 15-17: Nanog-sg2 abundance from dCas9-mESCs transfected with 40, 20 or 10 µg Nanog-sg2 plasmid. (**b**) The number of ChIP peaks detected from cells transfected with decreasing amount of sgRNA plasmids. (**c**) U-rich seed limits sgRNA abundance. Northern blot from dCas9 cells transfected with the sgRNAs listed below. Consecutive Us are highlighted in bold black.

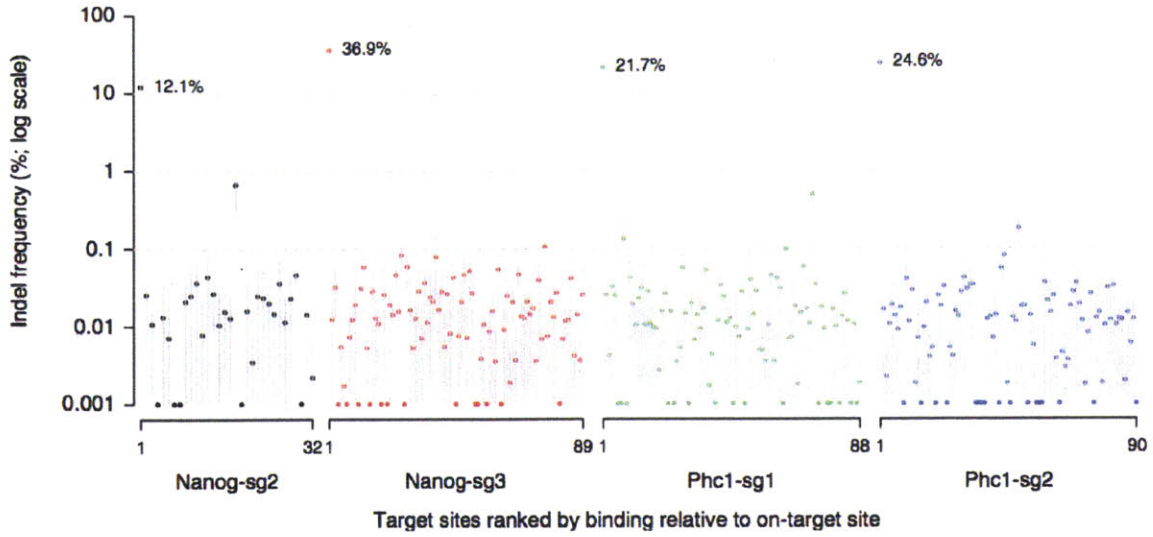


Figure 5: Indel frequencies at on-target sites and 295 off-target sites. For each sgRNA, selected target sites (Supplementary Table 3) were ranked by decreasing ChIP binding relative to on-target. Dots and gray bars indicate the mean and standard deviation of indel frequency from three biological replicates, respectively. The Y-axis was truncated at 0.001% for visualization at log scale. The indel frequencies for the four on-target sites are labeled with percentages.

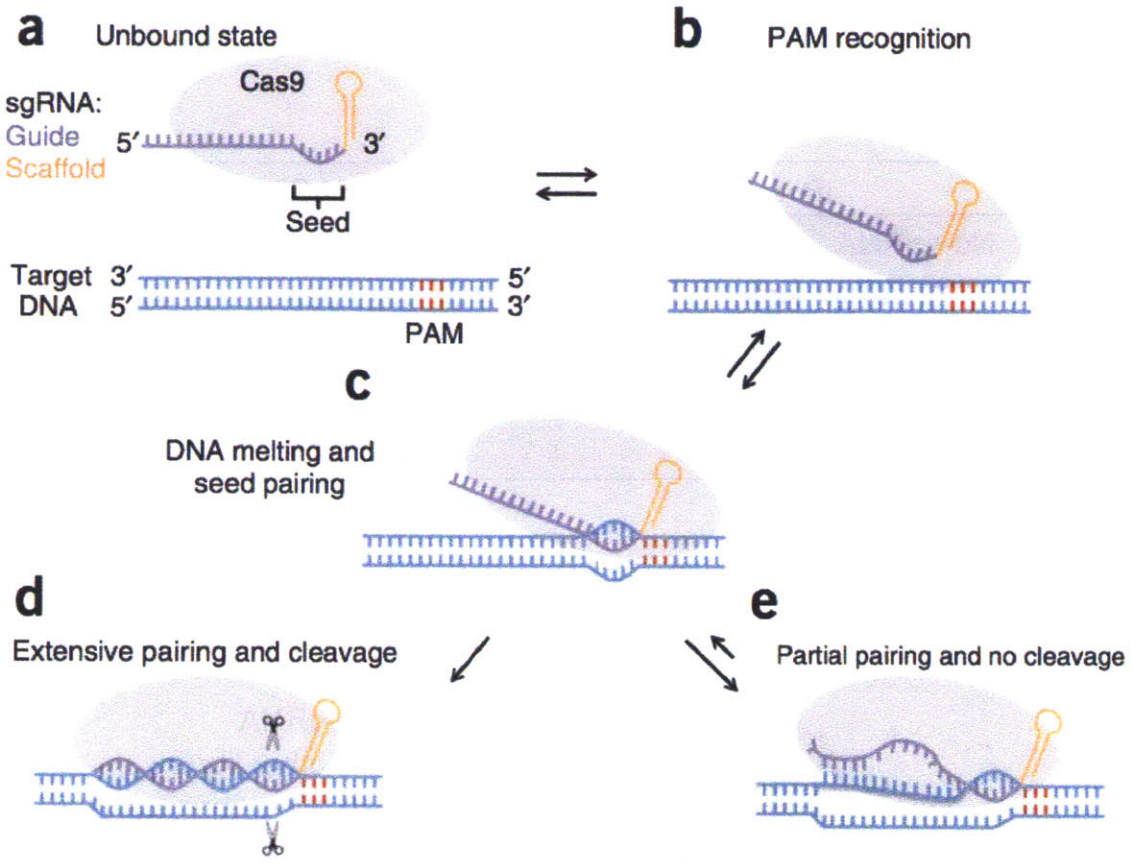
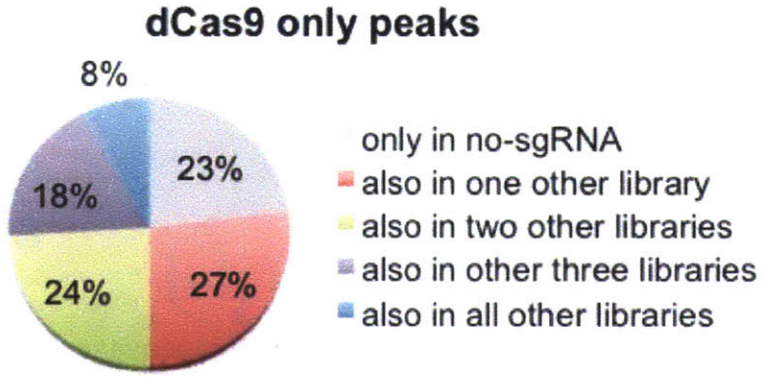


Figure 6: A model for Cas9 target binding and cleavage. (**a**) In the unbound state, Cas9 is loaded with sgRNA but not bound to DNA. The PAM region in the DNA is colored in red. (**b**) Recognition of the PAM by Cas9. (**c**) Cas9 melts the DNA target near the PAM to allow seed pairing. (**d**) If base pairing can be propagated to PAM-distal regions, the two Cas9 nuclease domains may be able to 'clamp' the target DNA and cleave it. (**e**) If only partial pairing occurs, there is no cleavage and Cas9 remains bound to the target.

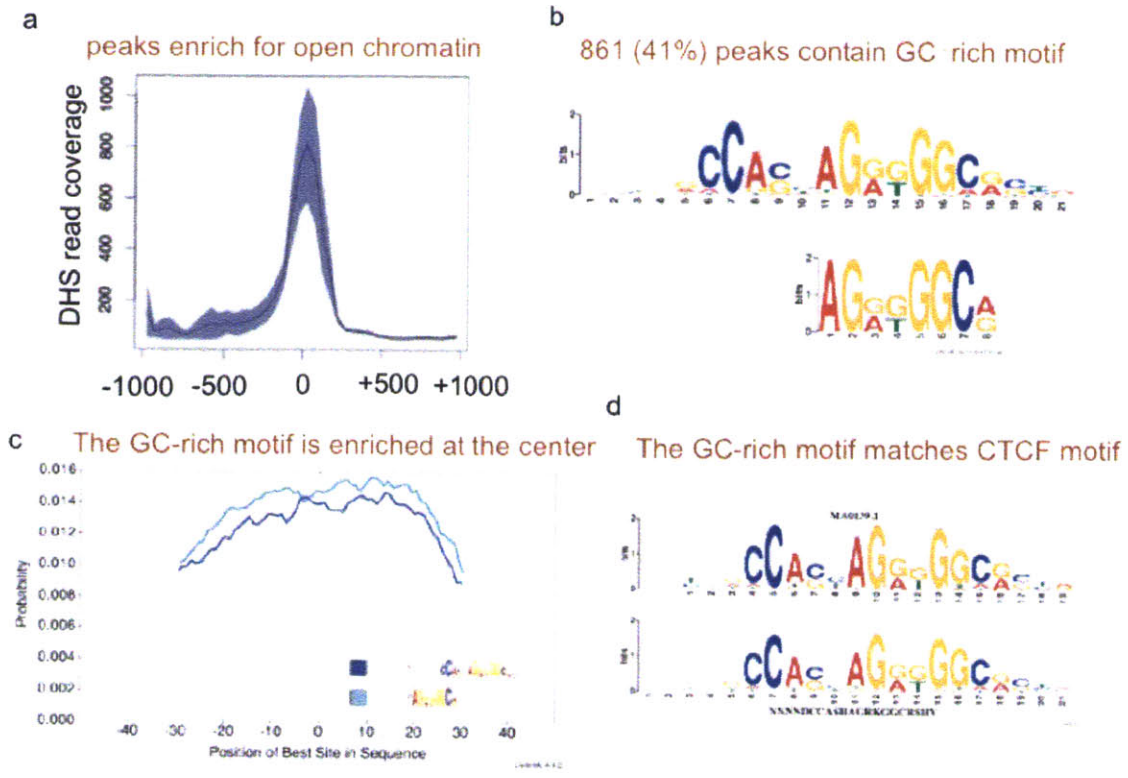
a

IP	Peaks called (MACS) over input
No-sgRNA / dCas9 only	2,115
Phc1-sg1	21,328
Phc1-sg2	4,568
Nanog-sg2	4,424
Nanog-sg3	19,857

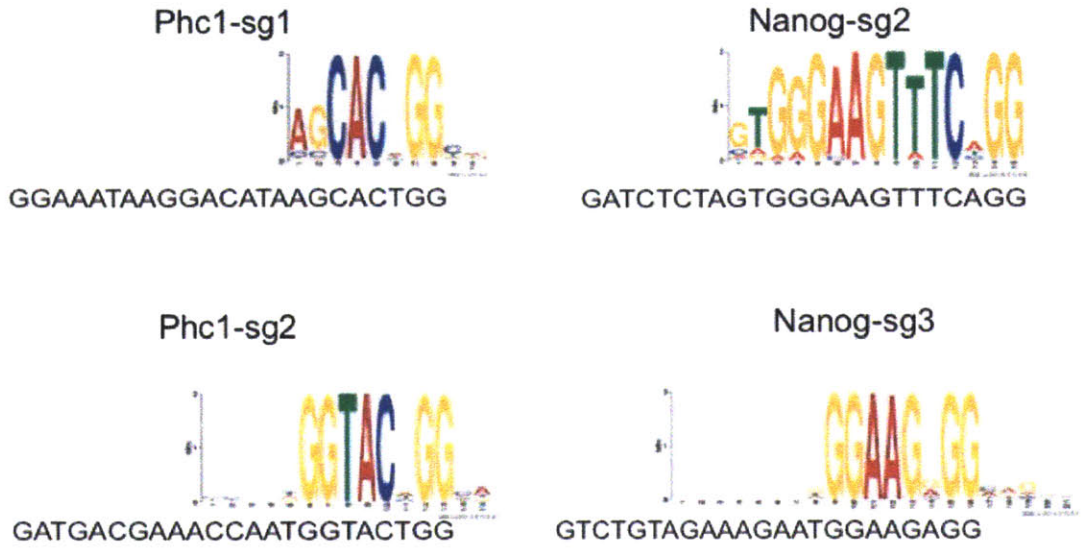
b



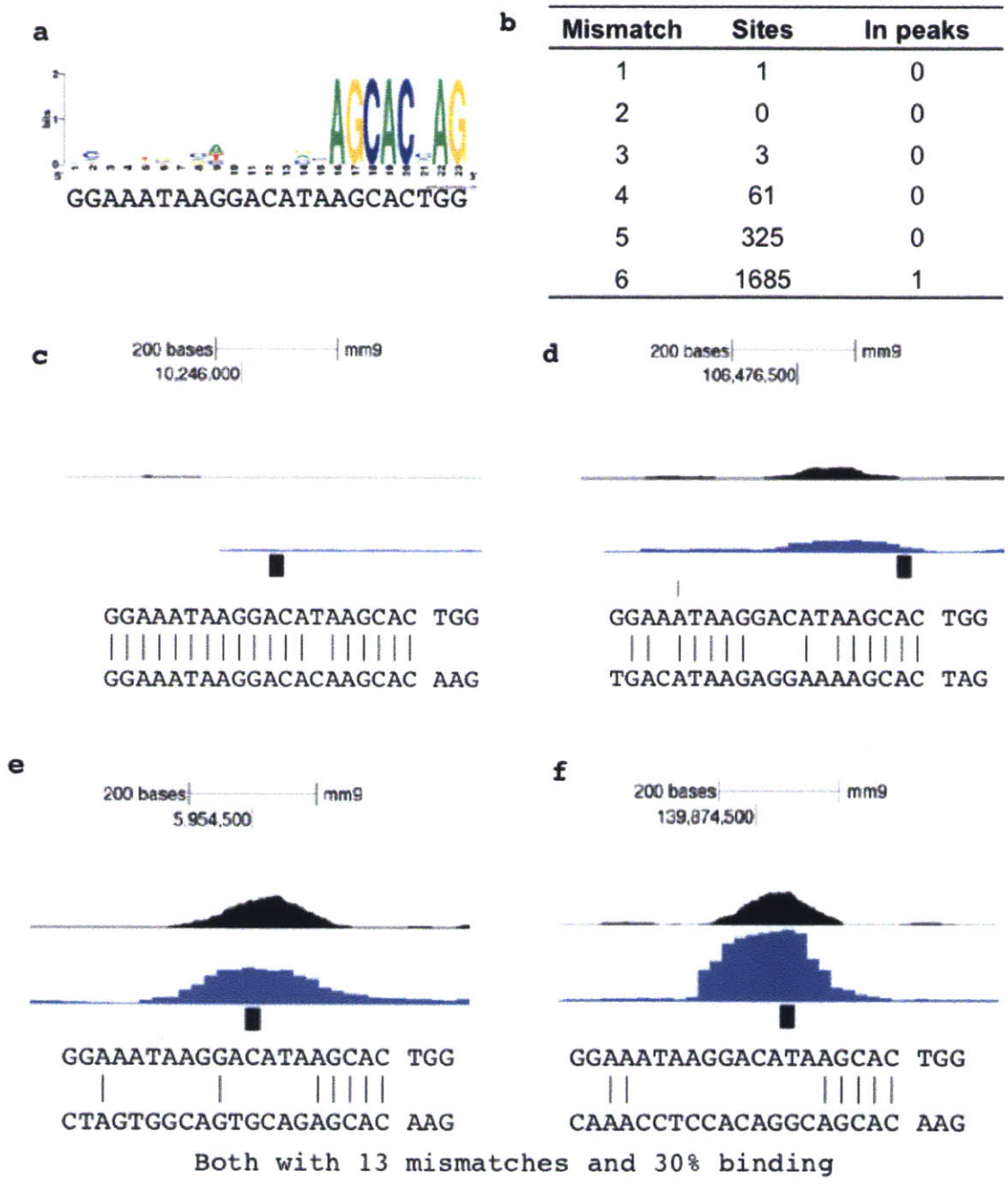
Supplementary Figure 1 | Conventional peak calling comparing IP to input. (**a**) The number of peaks called by MACS using default settings. (**b**) The fraction of dCas9-only peaks that are also detected in one, two, three, or all other four IP samples.



Supplementary Figure 2 | Characteristics of dCas9-only peaks. (**a**) Peaks are enriched for open chromatin regions. Shown is the average density of Dnase Hypersensitivity reads per 50bp bin in a 2kb window centered on peak summits. Blue area indicates standard error. (**b**) De novo motif finding within 50bp of peak summits by MEME-ChIP uncovered two related GG/CC-rich motifs. (**c**) Relative position of the motif within the peak, 0 indicates peak summits. (**d**) The longer motif (bottom) closely resembles CTCF binding motif (top, $p < 1e\text{-}23$)

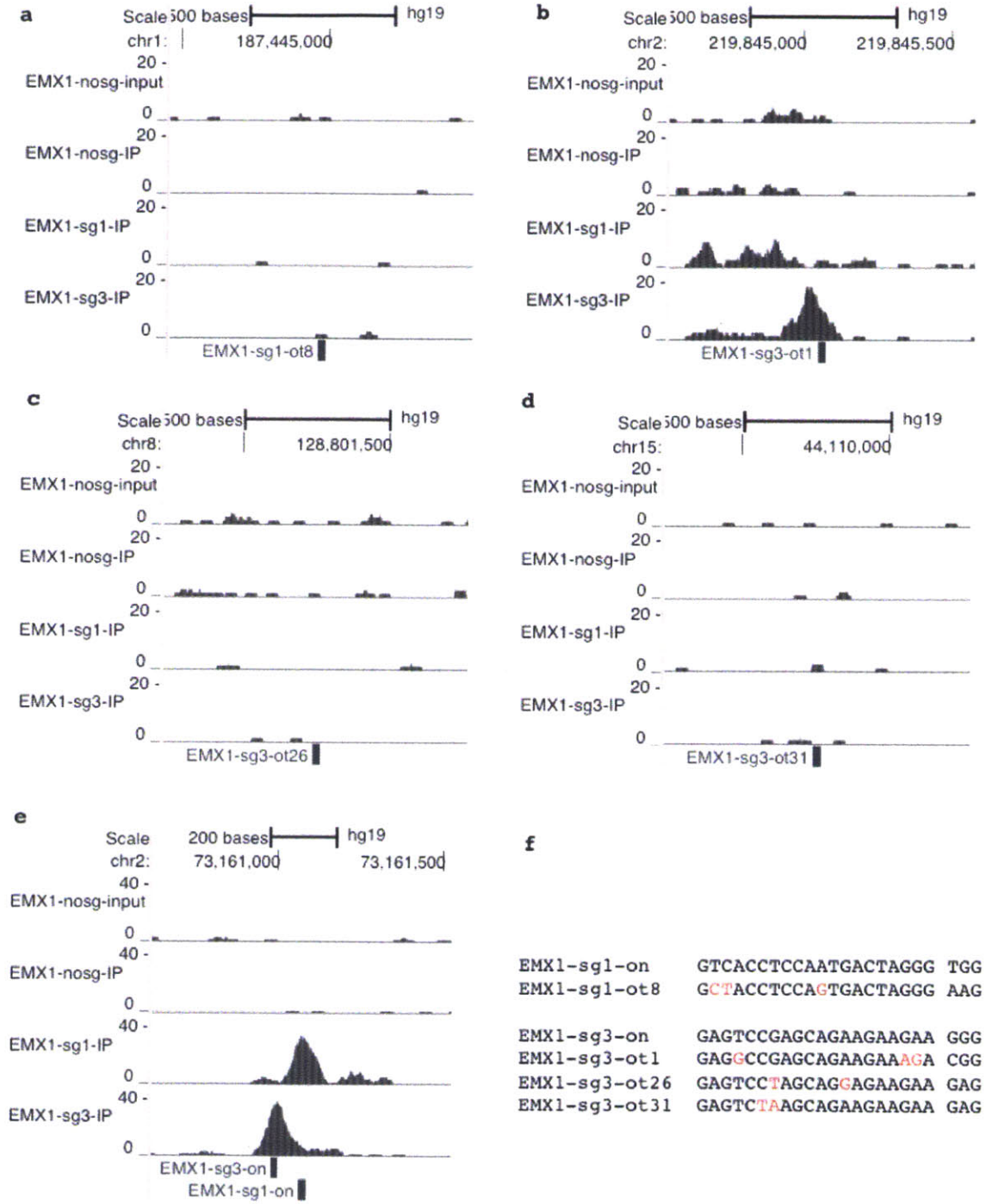


Supplementary Figure 3 | De novo motif discovery in ChIP peaks. Motifs detected by MEME-ChIP using default settings and sequences within 50 bps of peak summits. The guide RNA sequences were shown below the motif.

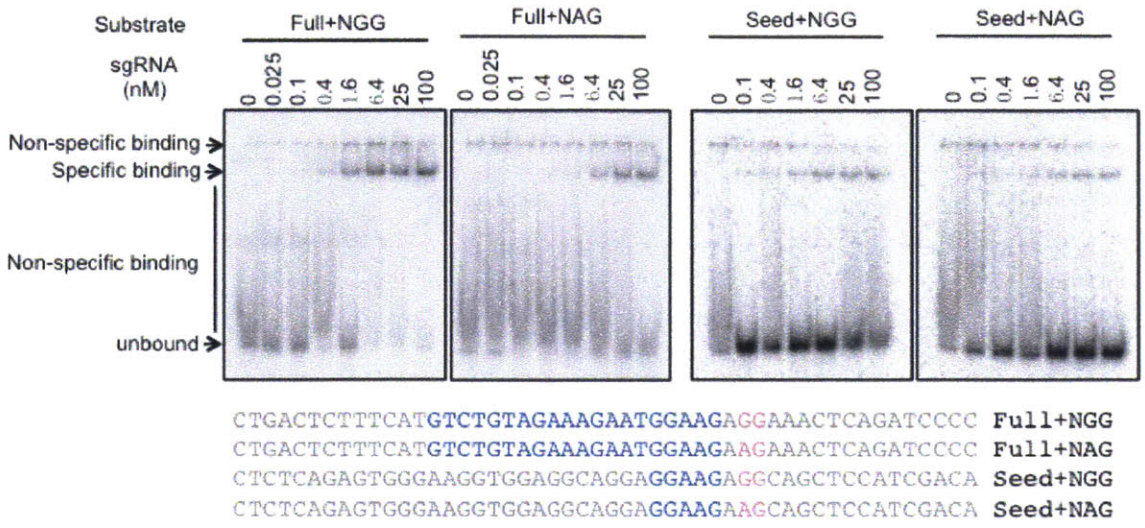


Supplementary Figure 4 | Lack of ChIP enrichment at seed+NAG sites. (**a**) of the 996 (33%) Phc1-sg1 ChIP peaks without seed+NGG sites, only 18 (1.8%) contain seed+NAG (AGCACNGG) within 50bp of peak summits, which is not higher than random (2.7%). Motif logo of the 18 NAG peaks showing the lack of base pairing outside the seed region. (**b**) The number of seed+NAG sites in the genome (column 2) and within ChIP peaks (column 3) that contain specific number of mismatches in the guide region (column 1). None of the seed+NAG sites with less than 6

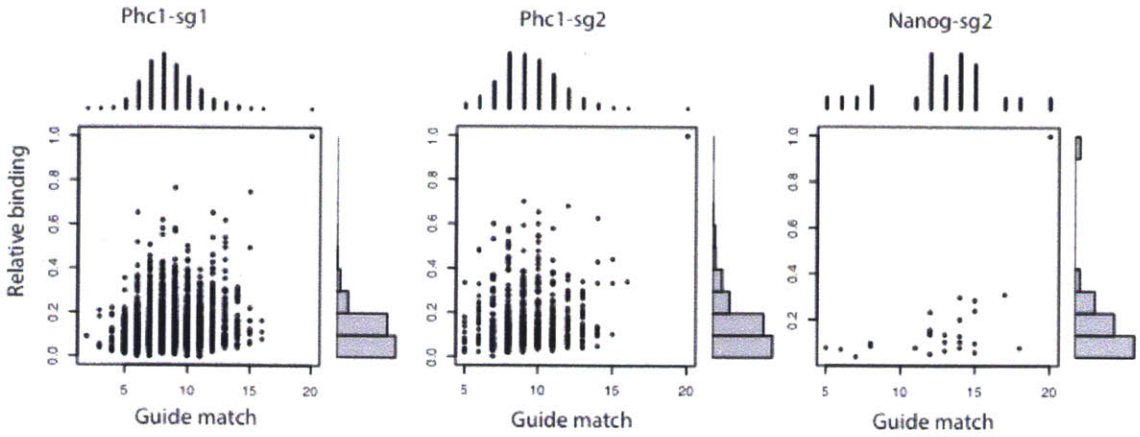
mismatches showed ChIP signals strong enough to be defined as peaks. (**c**) The best match contains only 1 mismatch and showed no ChIP signals. (**d**) Only one site with 6 mismatches is within a peak, yet the seed+NAG site is not in the center of the peak. (**e-f**) Similar to (**c-d**) but showing the two strongest peaks that are associated with seed+NAG sites. For (**c-f**), the top track is ChIP signal, and the bottom track is open chromatin. The scale is the same as Fig. 2b.



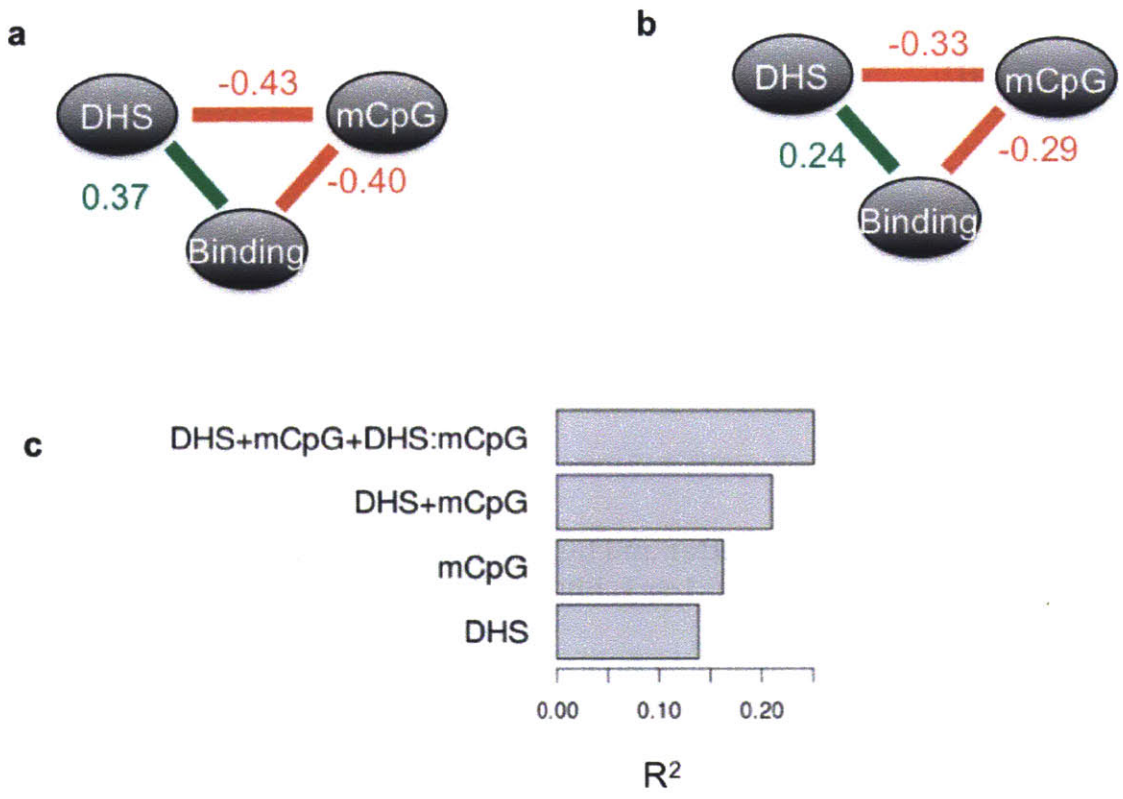
Supplementary Figure 5 | ChIP signals in HEK293FT cells. ChIP read density at four off-tar- gets (**a-d**) and on-targets (**e**). Sequences of the guide match and PAM are shown in (**f**), with mismatches highlighted in red. The four tracks are: input DNA, dCas9 transfected with no sgRNA, dCas9 transfected with EMX1-sg1, and dCas9 transfected with EMX1-sg3.



Supplementary Figure 6 | Gel shift assay for NAG substrates. The assay was done under the same condition as Fig. 2c. Sequences are shown at the bottom, with AG in pink and guide-matched bases in blue. Gels for NGG substrates were taken from Fig. 2c for comparison.

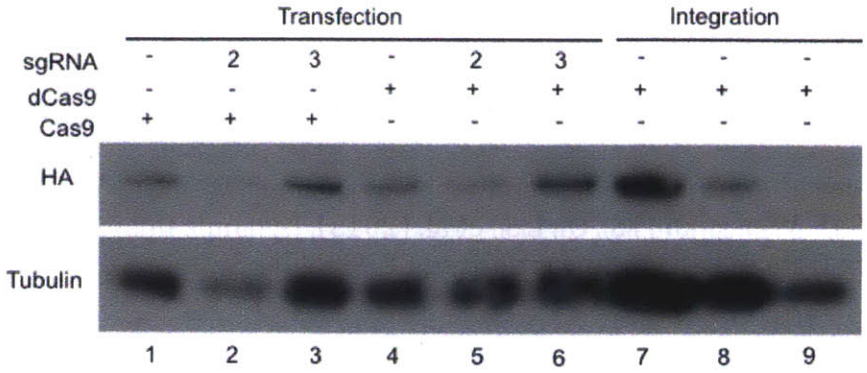


Supplementary Figure 7 | Scatter and histogram plots of guide match and relative binding for sgRNA Phc1-sg1, Phc1-sg2, and Nanog-sg2. The legends were the same as Fig 3a.



Supplementary Figure 8 | CpG methylation is negatively correlated with DHS and ChIP signals. (**a**) Pearson correlation coefficients between DHS, CpG methylation, and binding. (**b**) Partial Pearson correlation coefficients between DHS, CpG methylation, and binding. (**c**) The fraction of variation in binding explained by DHS, CpG methylation, DHS and CpG methylation without interaction, or DHS and CpG methylation with interaction.

a



b

Study	Plates	Cas9 plasmid used per well (ug)	10 cm equivalent (ug)
Schwank et al, 2013, Cell Stem Cell	48-well	0.7	38.5
Cong et al, 2013, Science	24-well	0.8	22
This study	10-cm	20	20
Hsu et al, 2013, Nature Biotechnology	24-well	0.5	13.7
Pattanayak et al, 2013, Nature Biotechnology	6-well	1	6.1

Supplementary Figure 9 | Cas9/dCas9 expression. (**a**) Western blot using lysates from cells with either transiently transfected Cas9 (lanes 1-3), or dCas9 (lanes 4-6), or cells stably integrat- ed with dCas9 (lanes 7-9). All Cas9/dCas9 proteins contain an HA tag. Tubulin was used as loading control. Cells for lanes 2 and 5 were also transfected with Nanog-sg2, and lanes 3 and 6 were from cells transfected with Nanog-sg3. Lanes 7-9 were the same lysate with 1:1, 1:2, and 1:4 dilution. After normalizing to the loading control, the HA band in lane 1 is about 2.6 times of the HA band in lane 8, suggesting that the expression of dCas9 in our stable cells is much lower than the cells with transiently transfected dCas9. (**b**) Comparison of Cas9 plasmid used in various studies, including the references, type of plates used, and the amount of Cas9/dCas9 plasmids transfected per well, and the equivalent amount on a 10 cm plate based on the area on each plate.

References

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell *136*, 215–233.

Carroll, D. (2013). Staying on target with CRISPR-Cas. Nat. Biotechnol. *31*, 807–809.

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

Cheng, A.W., Wang, H., Yang, H., Shi, L., Katz, Y., Theunissen, T.W., Rangarajan, S., Shivalila, C.S., Dadon, D.B., and Jaenisch, R. (2013). Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. Cell Res. *23*, 1163–1171.

Chiu, H., Schwartz, H.T., Antoshechkin, I., and Sternberg, P.W. (2013). Transgene-Free Genome Editing in Caenorhabditis elegans Using CRISPR-Cas. Genetics *195*, 1167–1171.

Cho, S.W., Kim, S., Kim, Y., Kweon, J., Kim, H.S., Bae, S., and Kim, J.-S. (2014). Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. Genome Res. *24*, 132–141.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819–823.

Cradick, T.J., Fine, E.J., Antico, C.J., and Bao, G. (2013). CRISPR/Cas9 systems targeting β-globin and CCR5 genes have substantial off-target activity. Nucleic Acids Res. *41*, 9584–9592.

Crooks, G.E., Hon, G., Chandonia, J.-M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res. *14*, 1188–1190.

Deveau, H., Garneau, J.E., and Moineau, S. (2010). CRISPR/Cas system and its role in phage-bacteria interactions. Annu. Rev. Microbiol. *64*, 475–493.

Dickinson, D.J., Ward, J.D., Reiner, D.J., and Goldstein, B. (2013). Engineering the Caenorhabditis elegans genome using Cas9-triggered homologous recombination. Nat. Methods *10*, 1028–1034.

Fu, Y., Foden, J.A., Khayter, C., Maeder, M.L., Reyon, D., Joung, J.K., and Sander, J.D. (2013). High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat. Biotechnol. *31*, 822–826.

Gasiunas, G., and Siksnys, V. (2013). RNA-dependent DNA endonuclease Cas9 of the CRISPR system: Holy Grail of genome editing? Trends Microbiol. *21*, 562–567.

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

Horvath, P., and Barrangou, R. (2010). CRISPR/Cas, the immune system of bacteria and archaea. Science *327*, 167–170.

Hsu, P.D., Scott, D.A., Weinstein, J.A., Ran, F.A., Konermann, S., Agarwala, V., Li, Y., Fine, E.J., Wu, X., Shalem, O., et al. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. Nat. Biotechnol. *31*, 827–832.

Jiang, W., Bikard, D., Cox, D., Zhang, F., and Marraffini, L.A. (2013). RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat. Biotechnol. *31*, 233–239.

Jinek, M., and Doudna, J.A. (2009). A three-dimensional view of the molecular machinery of RNA interference. Nature *457*, 405–412.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science *337*, 816–821.

Jinek, M., East, A., Cheng, A., Lin, S., Ma, E., and Doudna, J. (2013). RNA-programmed genome editing in human cells. Elife *2*, e00471.

Jinek, M., Jiang, F., Taylor, D.W., Sternberg, S.H., Kaya, E., Ma, E., Anders, C., Hauer, M., Zhou, K., Lin, S., et al. (2014). Structures of Cas9 Endonucleases Reveal RNA-Mediated Conformational Activation. Science *343*, 1247997.

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.

Machanick, P., and Bailey, T.L. (2011). MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics *27*, 1696–1697.

Maeder, M.L., Linder, S.J., Cascio, V.M., Fu, Y., Ho, Q.H., and Joung, J.K. (2013). CRISPR RNA-guided activation of endogenous human genes. Nat. Methods *10*, 977–979.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013a). RNA-guided human genome engineering via Cas9. Science *339*, 823–826.

Mali, P., Esvelt, K.M., and Church, G.M. (2013b). Cas9 as a versatile tool for engineering biology. Nat. Methods *10*, 957–963.

Mali, P., Aach, J., Stranges, P.B., Esvelt, K.M., Moosburner, M., Kosuri, S., Yang, L., and Church, G.M. (2013c). CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. Nat. Biotechnol. *31*, 833–838.

Marraffini, L.A., and Sontheimer, E.J. (2010). CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat. Rev. Genet. *11*, 181–190.

Mojica, F.J.M., Díez-Villaseñor, C., García-Martínez, J., and Almendros, C. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology *155*, 733–740.

Nielsen, S., Yuzenkova, Y., and Zenkin, N. (2013). Mechanism of eukaryotic RNA polymerase III transcription termination. Science *340*, 1577–1580.

Nishimasu, H., Ran, F.A.A., Hsu, P.D.D., Konermann, S., Shehata, S.I.I., Dohmae, N., Ishitani, R., Zhang, F., and Nureki, O. (2014). Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA. Cell *156*, 935–949.

Van der Oost, J., Jore, M.M., Westra, E.R., Lundgren, M., and Brouns, S.J.J. (2009). CRISPR-based adaptive and heritable immunity in prokaryotes. Trends Biochem. Sci. *34*, 401–407.

Orioli, A., Pascali, C., Quartararo, J., Diebel, K.W., Praz, V., Romascano, D., Percudani, R., van Dyk, L.F., Hernandez, N., Teichmann, M., et al. (2011). Widespread occurrence of non-canonical transcription termination by human RNA polymerase III. Nucleic Acids Res. *39*, 5499–5512.

Packer, M.J., Dauncey, M.P., and Hunter, C.A. (2000). Sequence-dependent DNA structure: tetranucleotide conformational maps. J. Mol. Biol. *295*, 85–103.

Pattanayak, V., Lin, S., Guilinger, J.P., Ma, E., Doudna, J.A., and Liu, D.R. (2013). High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. Nat. Biotechnol. *31*, 839–843.

Perez-Pinera, P., Kocak, D.D., Vockley, C.M., Adler, A.F., Kabadi, A.M., Polstein, L.R., Thakore, P.I., Glass, K.A., Ousterout, D.G., Leong, K.W., et al. (2013). RNA-guided gene activation by CRISPR-Cas9-based transcription factors. Nat. Methods *10*, 973–976.

Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P., and Lim, W.A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell *152*, 1173–1183.

Schwank, G., Koo, B.-K., Sasselli, V., Dekkers, J.F., Heo, I., Demircan, T., Sasaki, N., Boymans, S., Cuppen, E., van der Ent, C.K., et al. (2013). Functional Repair of CFTR by CRISPR/Cas9 in Intestinal Stem Cell Organoids of Cystic Fibrosis Patients. Cell Stem Cell *13*, 653–658.

Shalem, O., Sanjana, N.E., Hartenian, E., Shi, X., Scott, D.A., Mikkelsen, T.S., Heckl, D., Ebert, B.L., Root, D.E., Doench, J.G., et al. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. Science *343*, 84–87.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490–495.

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

Sternberg, S.H., Redding, S., Jinek, M., Greene, E.C., and Doudna, J.A. (2014). DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature *507*, 62–67.

Terns, M.P., and Terns, R.M. (2011). CRISPR-based adaptive immune systems. Curr. Opin. Microbiol. *14*, 321–327.

Teytelman, L., Thurtle, D.M., Rine, J., and van Oudenaarden, A. (2013). Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. Proc. Natl. Acad. Sci. U. S. A. *110*, 18602–18607.

Wang, T., Wei, J.J., Sabatini, D.M., and Lander, E.S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. Science *343*, 80–84.

Wu, Y., Liang, D., Wang, Y., Bai, M., Tang, W., Bao, S., Yan, Z., Li, D., and Li, J. (2013). Correction of a Genetic Disease in Mouse via Use of CRISPR-Cas9. Cell Stem Cell *13*, 659–662.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Chapter 6: Future Directions

In this chapter I present a hypothesis regarding the regulation the U1-PAS axis during mouse embryonic development.

Introduction

As an essential component of the spliceosome, U1 snRNA is indispensable for the proper splicing of most genes, including most house-keeping genes. Therefore the abundance of U1 snRNA in cells is robustly maintained at relatively constant level. At least four mechanisms ensure the level of U1 is insensitive to most variations in cells. First, there are multiple copies of the U1 gene encoding the snRNA (Lund and Dahlberg, 1984). Second, U1 snRNA is present in cells at extremely high level, on the order of one million copy per cell (Baserga and Steitz, 1993). Third, the half-life of U1 snRNA is exceptionally long, i.e. 4 to 5 days (Sauterer et al., 1988). Lastly, U1 snRNA auto-regulates its own abundance through an unknown mechanism (Cáceres et al., 1992; Mangin et al., 1985). Nonetheless, there are endogenous mechanisms that could modulate the functional activity of U1 snRNA.

Here I propose that some organisms hijack the auto-regulation mechanism to down-regulate U1 snRNA levels by increased expression of a variant U1 snRNA defective in suppressing polyadenylation.

Developmental regulation of U1 snRNA variants

In a few species there are variants of U1 snRNA accumulating to high levels at specific developmental stages, especially during early embryonic development. The first example was found in *Xenopus laevis*, where a class of U1 variants are specifically expressed in oocytes and embryos (Forbes et al., 1984). These U1 variants are called embryonic U1, or U1b, as compared to the normal adult/somatic

U1, or U1a. Soon similar embryonic variants were identified in other species including mouse (Lund et al., 1985), sea urchin (Nash et al., 1989), and fly (Lo and Mount, 1990).

The most well studied variant is mouse embryonic mU1b, which includes three variants mU1b1, mU1b3, and mU1b6 that differ in sequence but show the same developmental expression patterns (Lund et al., 1985). In fetal tissues such as brain and liver, mU1b accumulates to roughly the same level as mU1a, and remains detectable by northern blot in adult tissues that contain stem cells, including testis, spleen, thymus, and to some extent, ovary. Lund et al examined the time course of U1 variants during mouse development (Lund et al., 1985), in testis, brain, and liver, from embryonic day 13 to 14 weeks after birth. In brain and liver, the fraction of mU1b monotonically decreases from 40% at embryonic day 13, to undetectable around 6 weeks after birth. Interestingly, in testis, mU1b first decreases but goes back after birth and remain around 40%. Moreover, the embryonic mU1b can be detected in transformed cells but not in non-transformed cells.

Whether embryonic U1 has unique functions or not remains unclear. The developmental regulation of these variants in multiple species suggests they are likely functional, although so far none of these variants have been characterized in terms of molecular function or cellular phenotype. Two sequence features differ between the two classes of mU1, mU1a (mU1a1 and mU1a2) and mU1b (mU1b1, mU1b3, and mU1b6). There are seven nucleotides distinguishing all mU1b from mU1a, and they cluster around position 60 and 70. In addition to the sequence difference, the ribose-methylation at A_{70} is absent from all mU1b RNAs. Interestingly,

the A_{70} methylation is also lost in frog U1b. Both these two differences occur in stem loop B, where the U1 specific protein U1-A binds (Bach et al., 1990). In deed, although mU1b is assembled properly into snRNP, U1-A protein has less affinity to mU1b (Bach et al., 1990). Although it is speculated that mU1b might have altered specificity for 5' splice sites and regulate splicing, neither stem loop B nor U1-A protein is required for splicing (Heinrichs et al., 1990; Will et al., 1996).

Embryonic U1 silencing coincides with global lengthening of 3' UTR

The developmental and cell type expression pattern of mU1b strikingly coincides with previously reported global change in mRNA 3' UTR (Ji et al., 2009; Mayr and Bartel, 2009; Sandberg et al., 2008). Sandberg et al first reported global shortening of mRNA 3' UTR via activation of promoter proximal alternative polyadenylation sites in proliferating cells (Sandberg et al., 2008). Ji et al further showed that the average length of 3' UTR of mRNAs progressively lengthens during mouse embryonic development (Ji et al., 2009), coinciding with the loss of mU1b. Moreover, like the exception of mU1b in testis, the trend of mRNA 3' UTR length is reversed to shortening after birth in testis. Furthermore, Mayr et al also showed that independent of proliferating rate, transformed cells have shorter mRNA 3' UTR length, again coincide with the presence of mU1b in transformed cells (Mayr and Bartel, 2009).

The correlation between embryonic U1 silencing and global 3' UTR lengthening suggests potential role of embryonic U1 in 3' UTR shortening. Recent studies have demonstrated that U1 suppresses proximal polyadenylation events in

mRNA (Kaida et al., 2010), and down-regulation of U1 leads to shorter 3' UTR (Berg et al., 2012). Below I propose a model that explains the global change in 3' UTR length by developmental regulation of U1 variants.

Embryonic variants as dominate negatives of U1 snRNA

I hypothesize that mouse embryonic U1, mU1b, is defective in suppressing cleavage and polyadenylation, and thus doesn't regulate 3' UTR length. If this is the case, as the fraction of mU1b decreases during embryonic development, the fraction of somatic mU1a increases, protecting proximal polyadenylation sites from being used, leading to longer 3' UTR.

As discussed above, it seems the only difference in mU1b snRNP and mU1a snRNP is the lack of U1-A protein, which seems to be dispensable for splicing. The question is then whether U1-A is required for suppressing cleavage and polyadenylation. The inhibition of mRNA cleavage and polyadenylation by U1 snRNP was discovered recently (Kaida et al., 2010), and the molecular details remain to be addressed. However, there are extensive interactions between U1-A and the 3' end processing machinery, including the poly-A polymerase (PAP), and the CPSF-160 complex. U1-A protein homodimer directly binds and inhibits PAP when two copies of U1-A protein are recruited to its own mRNA by specific sequence motifs, leading to degradation of the message (Boelens et al., 1993; van Gelder et al., 1993; Gunderson et al., 1994; Klein Gunnewiek et al., 2000; Varani et al., 2000). This auto-regulatory mechanism does not inhibit cleavage, and seems to be independent of U1 snRNP. In addition, U1-A also directly binds CPSF-160 complex,

but in this case, increases polyadenylation efficiency *in vitro* (Lutz et al., 1996). Again, no inhibition on cleavage was observed. In two cases, U1-A was reported to inhibit cleavage, and in both cases, U1-A binds to two motifs that might also function as CstF-64 binding sites, thus inhibiting cleavage and polyadenylation, and again, this is independent of U1 snRNP (Ma et al., 2006; Phillips et al., 2004; Workman et al., 2014).

In summary, it is possible that U1-A in U1 snRNP is required for suppressing cleavage and polyadenylation, and thus regulating 3' UTR length, although currently no direct evidence is available.

On the other hand, it is worth noting that there doesn't seem to be such embryonic U1 variants in human (Lund, 1988; O'Reilly et al., 2013), yet the global regulation of 3' UTR length is conserved between human and mouse (Sandberg et al., 2008). It doesn't rule out that embryonic mU1b is regulating this process, but suggest other mechanisms exist to regulate global 3' UTR length.

References

Bach, M., Krol, A., and Lührmann, R. (1990). Structure-probing of U1 snRNPs gradually depleted of the U1-specific proteins A, C and 70k. Evidence that A interacts differentially with developmentally regulated mouse U1 snRNA variants. Nucleic Acids Res. *18*, 449–457.

Baserga, S.J., and Steitz, J.A. (1993). The RNA World. R.F. Gesteland, and J.F. Atkins, eds. (Cold Spring Harbor Laboratory Press), pp. 359–381.

Berg, M.G., Singh, L.N., Younis, I., Liu, Q., Pinto, A.M., Kaida, D., Zhang, Z., Cho, S., Sherrill-Mix, S., Wan, L., et al. (2012). U1 snRNP Determines mRNA Length and Regulates Isoform Expression. Cell *150*, 53–64.

Boelens, W.C., Jansen, E.J., van Venrooij, W.J., Stripecke, R., Mattaj, I.W., and Gunderson, S.I. (1993). The human U1 snRNP-specific U1A protein inhibits polyadenylation of its own pre-mRNA. Cell *72*, 881–892.

Cáceres, J.F., McKenzie, D., Thimmapaya, R., Lund, E., and Dahlberg, J.E. (1992). Control of mouse U1a and U1b snRNA gene expression by differential transcription. Nucleic Acids Res. *20*, 4247–4254.

Forbes, D.J., Kirschner, M.W., Caput, D., Dahlberg, J.E., and Lund, E. (1984). Differential expression of multiple U1 small nuclear RNAs in oocytes and embryos of Xenopus laevis. Cell *38*, 681–689.

Van Gelder, C.W., Gunderson, S.I., Jansen, E.J., Boelens, W.C., Polycarpou-Schwarz, M., Mattaj, I.W., and van Venrooij, W.J. (1993). A complex secondary structure in U1A pre-mRNA that binds two molecules of U1A protein is required for regulation of polyadenylation. EMBO J. *12*, 5191–5200.

Gunderson, S.I., Beyer, K., Martin, G., Keller, W., Boelens, W.C., and Mattaj, L.W. (1994). The human U1A snRNP protein regulates polyadenylation via a direct interaction with poly(A) polymerase. Cell *76*, 531–541.

Heinrichs, V., Bach, M., Winkelmann, G., and Lührmann, R. (1990). U1-specific protein C needed for efficient complex formation of U1 snRNP with a 5' splice site. Science *247*, 69–72.

Ji, Z., Lee, J.Y., Pan, Z., Jiang, B., and Tian, B. (2009). Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. Proc. Natl. Acad. Sci. U. S. A. *106*, 7028–7033.

Kaida, D., Berg, M.G., Younis, I., Kasim, M., Singh, L.N., Wan, L., and Dreyfuss, G. (2010). U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature *468*, 664–U81.

Klein Gunnewiek, J.M., Hussein, R.I., van Aarssen, Y., Palacios, D., de Jong, R., van Venrooij, W.J., and Gunderson, S.I. (2000). Fourteen residues of the U1 snRNP-specific U1A protein are required for homodimerization, cooperative RNA binding, and inhibition of polyadenylation. Mol. Cell. Biol. *20*, 2209–2217.

Lo, P.C., and Mount, S.M. (1990). Drosophila melanogaster genes for U1 snRNA variants and their expression during development. Nucleic Acids Res. *18*, 6971–6979.

Lund, E. (1988). Heterogeneity of human U1 snRNAs. Nucleic Acids Res. *16*, 5813–5826.

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

Lund, E., Kahan, B., and Dahlberg, J.E. (1985). Differential control of U1 small nuclear RNA expression during mouse development. Science *229*, 1271–1274.

Lutz, C.S., Murthy, K.G., Schek, N., O'Connor, J.P., Manley, J.L., and Alwine, J.C. (1996). Interaction between the U1 snRNP-A protein and the 160-kD subunit of cleavage-polyadenylation specificity factor increases polyadenylation efficiency in vitro. Genes Dev. *10*, 325–337.

Ma, J., Gunderson, S.I., and Phillips, C. (2006). Non-snRNP U1A levels decrease during mammalian B-cell differentiation and release the IgM secretory poly(A) site from repression. RNA *12*, 122–132.

Mangin, M., Ares, M., and Weiner, A.M. (1985). U1 small nuclear RNA genes are subject to dosage compensation in mouse cells. Science *229*, 272–275.

Mayr, C., and Bartel, D.P. (2009). Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell *138*, 673–684.

Nash, M.A., Sakallah, S., Santiago, C., Yu, J.C., and Marzluff, W.F. (1989). A developmental switch in sea urchin U1 RNA. Dev. Biol. *134*, 289–296.

O'Reilly, D., Dienstbier, M., Cowley, S.A., Vazquez, P., Drozdz, M., Taylor, S., James, W.S., and Murphy, S. (2013). Differentially expressed, variant U1 snRNAs regulate gene expression in human cells. Genome Res. *23*, 281–291.

Phillips, C., Pachikara, N., and Gunderson, S.I. (2004). U1A inhibits cleavage at the immunoglobulin M heavy-chain secretory poly(A) site by binding between the two downstream GU-rich regions. Mol. Cell. Biol. *24*, 6162–6171.

Sandberg, R., Neilson, J.R., Sarma, A., Sharp, P.A., and Burge, C.B. (2008). Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. Science *320*, 1643–1647.

Sauterer, R.A., Feeney, R.J., and Zieve, G.W. (1988). Cytoplasmic assembly of snRNP particles from stored proteins and newly transcribed snRNA's in L929 mouse fibroblasts. Exp. Cell Res. *176*, 344–359.

Varani, L., Gunderson, S.I., Mattaj, I.W., Kay, L.E., Neuhaus, D., and Varani, G. (2000). The NMR structure of the 38 kDa U1A protein - PIE RNA complex reveals the basis of cooperativity in regulation of polyadenylation by human U1A protein. Nat. Struct. Biol. *7*, 329–335.

Will, C.L., Rümpler, S., Klein Gunnewiek, J., van Venrooij, W.J., and Lührmann, R. (1996). In vitro reconstitution of mammalian U1 snRNPs active in splicing: the U1-C protein enhances the formation of early (E) spliceosomal complexes. Nucleic Acids Res. *24*, 4614–4623.

Workman, E., Veith, A., and Battle, D.J. (2014). U1A regulates 3' processing of the survival motor neuron mRNA. J. Biol. Chem. *289*, 3703–3712.