UNDERSTANDING PROMOTER EVOLUTION AND THE GENESIS OF NON-CODING GENES

ABSTRACT

We take advantage of the recent discovery of hundreds of species-specific long intergenic non-coding RNAs (lincRNAs) to study the evolution of promoters and how non-coding genes are born. Using a high-throughput reporter assay, we will identify active mouse promoter sequences found upstream of transcribed lincRNAs as well as orthologous human fragments that share sequence similarity but do not drive transcription. We will then characterize the essential substitutions that give rise to new transcriptional events and integrate in vivo measurements of DNA methylation and transcription factor binding to understand the interplay between sequence, epigenetics, and transcription. Finally, we will test our understanding of promoter evolution and activity by designing and testing synthetic promoter sequences that we predict to drive transcription. Together, these experiments will be an important step in deciphering the regulatory code and provide insight into evolutionary mechanisms of gene birth.

INTRODUCTION

The promoter is critical to gene regulation. In mammalian species, the core promoter is the major functional region upstream of the transcription start site (TSS) where RNA polymerase II (RNApolII) assembles with general transcription factors to form the preinitiation complex (PIC), which sets the stage for transcription initiation [1-3]. Mammalian promoters are quite diverse in their sequence makeup and do not share a common motif [1-3]. They are also diverse in their ability to drive transcription: some promoters constitutively drive expression across many cell lines while others require enhancer elements to be active [1-7]. Understanding and characterizing the promoter is crucial if we are to understand how transcription is controlled.

A central question in understanding transcriptional control is how does gene regulation evolve? Since the discovery that humans and apes share 99% sequence similarity in protein-coding genes [8], the evolution of regulatory gene has been intensely investigated, as scientists try to resolve the key genomic differences that define each species. Understanding how promoters evolve not only allows us to better map the relationship between sequence and transcriptional activity, but also provides insight into evolutionary mechanisms underlying speciation that are impossible to test in higher species.

Previous comparative studies looking at promoter evolution consist primarily of computational analyses that characterize evolutionary signatures found within promoters [9-11]. Though these studies are important steps towards understanding promoter evolution, they are not without limitations. Genome-wide promoter studies analyze predicted promoter regions (usually, the sequence just upstream of the TSS) that are not experimentally validated to be active, and lump together very different subsets of promoter sequences that have varying levels of transcription activity and tissue specificity. Previous comparative studies on promoter evolution also have not addressed epigenetic changes that may occur alongside nucleotide mutations. Epigenetic changes including DNA methylation and transcription factor binding are thought to play a major role in transcriptional activity and may be a better predictor of promoter activity than sequence [12-17]. Finally, few studies have been conducted on the genesis of new promoters, as this is believed to be a rare event among protein coding genes [18].

Here, we take advantage of the recent discovery in our lab of over 300 long intergenic non-coding RNAs (lincRNAs) present only in the mouse-rat lineage in a highly tissue-specific manner, and not transcribed in humans, chimpanzees, or cows (unpublished data). Despite no evidence of transcriptional activity, over 50% of these lincRNAs can be mapped to orthologous and syntenic regions in the human genome with greater than 30% sequence similarity. These lincRNAs are consistently transcribed across different RNA-sequencing studies and marked by robust chromatin signatures associated with transcription [19, 20]. They do not appear to be a part of any homologous families or have paralogs within the mouse genome, making them unlikely to have arose through duplication events [21, 22]. This "natural" experimental setup allows us to compare functional promoter sequences to their orthologous, non-functional counterpart to understand how new promoters may be born. We can further take advantage of the tissue-specific nature of lincRNAs to integrate relevant epigenetic changes between nonfunctional and active promoter regions. In this proposal, I outline a study that experimentally identifies core promoter sequences that lead to active transcription of lincRNAs but have inactive orthologous counterparts, and characterize the genetic and epigenetic changes that occur in the evolution of an active lincRNA promoter.

SPECIFIC AIMS

***Aim 1**:* To experimentally characterize the activity of lincRNA core promoter sequences from the mouse genome as well as orthologous inactive sequences from the human genome through a high-throughput transient transfection report assay.

***Background**:* Unlike prokaryotic organisms, eukaryotic promoters do not share a consensus sequence motif. Instead, promoter regions are often simply defined as the region directly upstream of a TSS [1, 2, 4, 11]. Promoter regions can also be predicted using chromatin-immunoprecipitation with massively parallel DNA sequencing (ChIP-seq) to find the binding sites of proteins that form the PIC. This has been done with relatively high accuracy on a genome-wide scale in the mouse and human genomes with ~60% of transcribed genes and ~10% of untranscribed genes reported as being occupied by the PIC [5, 6].

Experimentally, putative mammalian promoter sequences can be tested for promoter activity using transient transfection reporter assays. In this assay, a plasmid is made in which a reporter gene is located downstream from the putative promoter sequence and this gene is then transfected into cells. Quantitative measurements of the reporter gene product's activity is used as a proxy measurement of promoter activity. Though this assay is considered an *in vivo* test of promoter function, it is of course limited by the fact that the promoter sequence has been taken out of its genomic contexts and is unable to interact with other, possibly necessary, cis-regulatory elements. Nevertheless, a previous study testing the 500bp upstream and 100bp downstream fragments of 900 TSS in the human genome was able to measure activity in 60% of the fragments tested and found strong correlation between promoter activity and corresponding endogenous RNA levels [7].

Past research into the elements that contribute to the activity of mammalian promoters have focused on promoter sequences from a single species and/or synthetically mutated promoter sequences [4-7]. Genome-wide scans for active promoter sequences have yet to be conducted in a comparative fashion. In our study, we examine active promoters from lincRNAs transcribed in mouse as well as an orthologous sequence that does not appear to lead to transcription. In this way, not only do we gain an understanding for what elements are necessary for promoter activity, but we also learn lessons of how new promoters may be born.

Approach: We will test the activity of promoters of rodent-specific lincRNAs as well as the activity of their orthologous sequence from the human genome. For each rodent-specific lincRNA, we will define the core promoter region as 600bp upstream and 100bp downstream of the lincRNA TSS in the mouse genome. Because lincRNAs are expressed in such a highly tissue-specific manner [19, 20], we can further corroborate our findings with tissue-matched ChIP-seq data for provided by the Mouse ENCODE consortium [23]. Any region that does not contain an RNAPolII peak in tissue in which the lincRNA is expressed will be thrown out.

We will then identify regions orthologous to our set of putative promoters in the human genome using the *LiftOver* tool from the University of California at Santa Cruz Genome Browser [13]. Again, we will integrate RNAPolII ChIP-seq data from the Human ENCODE consortium [25] to ensure no RNAPolII binding in our set of putative non-functional sequence. In this way, we filter out any lincRNAs that are not actually rodent-specific but are simply very lowly transcribed in the human.

Using the methods for a transfection reporter assay outlined in Cooper et al. 2006, we will test promoter activity for our set of putative promoter sequences and corresponding orthologs across the nine cell lines in which the lincRNAs were first identified. Additionally, we will test 6 positive and 102 negative controls previously reported to have no promoter activity [6]. Briefly, each sequence of interest will be PCR amplified and transformed into a firefly luciferase plasmid. The plasmid will then be transfected into cells along with a *Renilla* luciferase control plasmid. The level of activity of each fragment is a ratio of the firefly luciferase to the control luciferase signal, and a fragment is considered to be active if its signal exceeds three standard deviation above signal from the negative control fragments.

Through this experiment, we will identify promoter sequences that can drive transcription without any additional *cis* regulatory elements, as well as their orthologs that share sequence similarity, but are unable to drive transcription.

Aim 2: To characterize the sequence-level and epigenetic-level changes that occur as a neutral genomic region evolves into a core promoter.

Background: Although mammalian promoters do not share a defining sequence motif, they are often associated with elements such as initiator elements (YYANWYY), CpG islands, TFIIB recognition elements (SSRCGCC), or TATA boxes [1-3]. However it is unknown which element or combination of elements are most important for promoter activity. Furthermore, it is unknown how novel promoters evolve into existence. Some studies suggest that promoter regions are evolving faster than neutral regions [10, 11], possibly due to biased gene conversion (BCG), a phenomenon where AT➔GC mutations are more frequently seen because of the GC-biased repair of A:C and G:T mismatches [26]. Another suggested mechanism for promoter genesis that is particularly popular in the lincRNAs community is through insertions of repetitive transposons with promoter potential [22, 27]. A recent study of human lincRNAs revealed that 83% of lincRNAs contain a transposable element (TE) and that these TEs are enriched to be found at the TSS [28].

Beyond promoter sequence, it is understood that the presence of certain epigenetic marks are correlated with promoter regions. Promoters, especially those with CpG islands, are often hypomethylated and become repressed upon methylation [12-15]. In addition, promoters are frequently bound by sequence-specific transcription factors that are thought to stabilize the PIC to further enhance transcription [1-2, 16-17]. However, what drives these epigenetic marks – whether it is sequence, structural, or *trans* effects – remains poorly understood.

Approach: From the results of our reporter assay, we largely expect to see two cases of promoter activity: 1) both the putative promoter and its ortholog show no promoter activity, or 2) the putative promoter shows activity but its ortholog does not. Case 1 suggests that either we are not testing the correct promoter sequence or that additional cis-regulatory elements are needed activate transcription. Because we cannot differentiate between the two possibilities, we focus our downstream analysis on Case 2.

We will first analyze the sequence differences between active promoters and their inactive orthologs. Using the likelihood ratio test (LRT) described in Pollard et al. (used for discovery human accelerated regions) [29], we will test these promoter regions for accelerated evolution. The LRT statistic for the region is the ratio of the likelihood of the model with acceleration on the mouse branch to the model without the acceleration. The significance of the LRT statistics can be assessed by simulation from the genome-wide null model. In addition, we will test if faster evolving promoters have higher proportion of AT➔GC substitutions than neutrally evolving promoters, which may suggest accelerated evolution due to BGC.

To understand what sequence elements may be necessary for promoter genesis, we will test active promoters for enrichment of motif elements described in the background section, as compared to their orthologous counterparts. Finally, we will test for enrichment of repeat elements within these promoters as compared to a set of size- and GC content-matched random regions. Using the UCSC multiple alignment [24], we can further see if these repeat elements are ancestral or if they were inserted in the rodent lineage, providing clues to the mechanism behind the genesis of lincRNA promoters.

Next, we will use bisulfite data and transcription factor ChIP-seq data from the ENCODE consortium [23] to examine the epigenetic differences between inactive regions and active promoters. Because lincRNAs are highly tissue specific, we can narrow our analysis down to looking only at data derived from the appropriate tissue.

We will compare the amount of methylation within active promoters to promoters of highly conserved protein-coding genes expressed at comparable levels in the mouse genome. We will also compare the amount of methylation within inactive orthologous regions with random neutral regions in the human genome. In this way, we can understand whether DNA methylation is "switch-like", where inactive sequences have distinct, lower distribution of methylation compared to active promoters, or if there is a range of DNA methylation as we move on the spectrum from random sequences, to promoter-similar sequences, to newly evolved promoters, to anciently-evolved promoters.

We will repeat this analysis with TF binding data. Again, we will compare TF binding of mouse lincRNA promoters and human orthologous regions to TF binding of mouse protein-coding promoters and human random regions to understand whether TF binding is switch-like or falls on a gradient.

Finally, we will examine whether the epigenetic differences between promoters and orthologous regions are correlated with sequence changes by testing if the regions of epigenetic change contains more substitutions than randomly selected bases within the promoter region. Manual inspection of mutational changes can corroborate whether the mutation actually removed methylation sites or created new transcription factor binding sites. In contrast, it is possible that sequence changes do not drive promoter evolution, but rather, other *trans* effects open up the DNA and allow for transcription factors to bind and initiate transcription.

Our computational analysis on the sequence changes and epigenetic changes that occur between active promoter sequences and their orthologous inactive sequences will lead us to understand what elements are involved in the genesis of a promoter.

Aim 3: To validate discoveries of necessary promoter components by testing novel sequences predicted to have promoter function.

Background: A major limitation of computational analysis is its inability to prove causation. From our first and second aim, we will discover what elements are enriched in active promoters and what sequence changes are correlated with those elements. In our third aim, we plan to experimentally validate our computational findings by designing and testing our own synthetic promoters.

Approach: Using our discoveries of essential promoter components, we will attempt to "turn on" inactive human sequences by incorporating only the mouse substitutions that we believe are necessary to drive promoter activity. As a negative control, we will also randomly select the same number of mouse substitutions to incorporate into the human sequence and test for activity (using the same reporter assay describe in Aim 2). If we truly discovered causative rather than correlative evolutionary changes, our intentionally designed promoters should show more activity as compared to the randomly designed promoters. Being able to predict and design active promoters will provide strong evidence that our computation discoveries are rooted in true biology.

CONCLUSION AND SIGNIFICANCE

The genesis of new genes and promoter regions has hardly been studied, as *de novo* origination of new protein coding genes is a rare phenomenon. The recent discovery of hundreds of non-coding genes, many of which appear to be species specific and tissue specific, provides us with a perfect natural system to study how promoters evolve when new genes are born. This study will characterize what sequence elements transform inactive genomic regions into active promoters and what mechanisms (biased gene conversion, transposon intersection, etc.) drive those transformations. Further, this study will integrate epigenetic data in order to better understand how sequence, methylation, and transcription factor binding interplay to give rise to transcription initiation. This study will discover and validate key evolutionary events that lead to the origination of mouse- and rodent-specific lincRNAs, shedding light on the regulatory genomic code as well as the evolutionary mechanisms that give rise to the diverse phenotypes across species.

References

1. Sandelin A., et al., *Mammalian RNA polymerase II core promoters: insights from genome-wide studies.* Nature Reviews Genetics, 2007. **8**(6): p. 424-436.
2. Butler J.E.F, and J.T. Kadonaga. *The RNA polymerase II core promoter: a key component in the regulation of gene expression.* Genes and Development, 2002. **16**(20): p. 2583-2592.
3. Juven-Gershon, T., et al., *Regulation of gene expression via the core promoter and the basal transcriptional machinery.* Developmental Biology, 2010. **339**(2): p. 225-229.
4. Carninici P., et al. *Genome-wide analysis of mammalian promoter architecture and evolution.* Nature, 2006. **38**(6): p. 626-635.
5. Barrera L.O., et al. *Genome-wide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs.* Genome Research, 2008. **18**(1): p. 46-59.
6. Kim, T.H., et al., *A high-resolution map of active promoters in the human genome.* Nature, 2005. **436**(7052): p. 876-880.
7. Cooper S.J., et al. *Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome.* Genome Research, 2006. **16**(1): p. 1-10.
8. King M.C., and A.C. Wilson. *Evolution at two levels in humans and chimpanzees.* Science, 1975. **188**(4184): p. 107-116.
9. Akan P., and P. Deloukas, *DNA sequence and structural properties as predictors of human and mouse promoters.* Gene, 2008. **410**(1): p. 165-176.
10. Liang H., Lin Y.S., and W.H. Li, *Fast Evolution of Core Promoters in Primate Genomes.* Molecular Biology and Evolution, 2008. **25**(6): p. 1239-1244.
11. Taylor M.S., et al., *Heterotachy in Mammalian Promoter Evolution.* PLoS Genetics, 2006. **2**(4): e30.
12. Weber M., et al. *Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome.* Nature Genetics, 2007. **39**(4): p. 457 - 466.
13. Bird A., *DNA methylation patterns and epigenetic memory.* Genes and Development, 2002. **16**(1): p. 6-21
14. Miranda, T.B., and P.A. Jones. *DNA methylation: the nuts and bolts of repression.* Journal of Cellular Physiology, 2007. **213**(2): 384-390.
15. C.L. Hsieh, *Dependence of transcriptional repression on CpG methylation density.* Molecular and Cellular Biology, 1994. **14**(8): 5487-5494.
16. Tabach, Y., et al., *Wide-scale analysis of human functional transcription factor binding reveals a strong bias towards the transcription start site*. PLoS One, 2007. **2**(8): e807.
17. Whitfield, T.W., et al., *Functional analysis of transcription factor binding sites in human promoters.* Genome Biology, 2012. **13**(9): R50.
18. Long M., et al. *The origin of new genes: glimpses from the young and old*. Nature Reviews Genetics, 2003. **4**(11): p. 856-875.
19. Ponjavic J, et al. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res*. 2007 May; 17(5): 556-65.
20. Guttman M, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*. 2009 Mar 12; 458(7235): 223-7.
21. Wang J, et al. Mouse transcriptome: Neutral evolution of 'non-coding' complementary DNAs. *Nature*. 2004 Oct; 431.
22. Ponting C.P., Oliver P.L., and W. Reik. *Evolution and Functions of Long Noncoding RNAs.* Cell, 2009. **136**(4): p. 629-641.
23. Mouse ENCODE Consortium, et al. *An encyclopedia of mouse DNA elements (Mouse ENCODE).* Genome Biology, 2012. **13**(8): p. 418.
24. Kent W.J., et al. *The human genome browser at UCSC*. Genome Research, 2002. **12**(6): p. 996 - 1006.
25. Encode Project Consortium, et al. *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
26. Galtier N., and L. Duret, *Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution.* Trends in Genetics, 2007. **23**(6): p. 273-277.
27. Thornburg BG, et al. Transposable elements as a significant source of transcription regulating signals. 2006; *Gene* 365:104-110.
28. Kelley D and John Rinn. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biology. 2012; 13: R107.
29. Pollard, K.S., et al., *Forces Shaping the Fastest Evolving Regions in the Human Genome.* PLoS Genet, 2006. **2**(10): e168.