Discovery and In Vivo Characterization of Long Noncoding RNAs

Discovery and In Vivo Characterization of Long Noncoding RNAs

A dissertation presented

by

Stephen Liapis

To

The Department of Molecular and Cellular Biology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biochemistry

Harvard University

Cambridge, Massachusetts

5/11/2016

Dissertation Advisor: John L. Rinn Stephen Liapis

Discovery and In Vivo Characterization of Long Noncoding RNAs

Abstract

The noncoding genome, or the portion of the genome that does not encode for proteins, encompasses >95% of the human genome. It has been found that the majority of disease-associated genetic variants identified by genome-wide association studies (GWAS) are located in this noncoding 95%, where they have the potential to affect regions that control transcription (promoters, enhancers) and noncoding RNAs that also can influence gene expression. The discovery of these alterations has already contributed to a better understanding of the etiology of human diseases and has begun to yield insight into the function of these noncoding loci I am interested in studying how the noncoding genome functions and contributes to human development and disease pathology, especially when it is considered that our understanding of human disease is almost entirely contained within the realm of the <5% of the genome that is protein coding. Toward this end, I have focused my studies on one part of the noncoding genome, long noncoding RNAs. In order to identify whether long noncoding RNAs are important for mammalian development and disease, our lab created a set of lincRNA knockout animal models in which a cassette expressing beta-galactosidase (lacZ) replaces the lincRNA DNA sequence. I have used these models for the *in vivo* characterization of several lincRNAs, including *Fendrr* in the lungs, *Brn1b* in the brain, *Tug1* in the testes, and *Cox2* in the innate immune system. Each of these studies reveals perturbations in development induced by loss of function of the respective lincRNA locus, and demonstrates promising potential for further examination of the role these molecules play in human disease.

Abbreviations

RNA	Ribonucleic acid
DNA	Deoxyribonucleic acid
mRNA	Messenger RNA
rRNA	Ribosomal RNA
tRNA	Transfer RNA
ncRNA	Noncoding RNA
miRNA	Micro RNA
H3	Histone 3
K	Lysine
me	Methylation
RNP	Ribonucleoprotein
FISH	Fluorescence *in situ* hybridization
mESC	Mouse embryonic stem cell
MEF	Mouse embryonic fibroblast
CAST	*Mus musculus castaneus*
lincRNA	Long intergenic noncoding RNA
GWAS	Genome-wide association studies
RAP	RNA antisense pulldown
ChIP	Chromatin immunoprecipitation
lacZ	Beta-galactosidase
X-gal	5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside
GSEA	Gene set enrichment analysis
GBA	Guilt by association
WT	Wild type
Het	Heterozygote
KO	Knockout
bp	Base pair

Acknowledgements

I would like to begin this thesis by expressing my deepest gratitude toward my Ph.D. advisor, Dr. John Rinn. John has exemplified the ideals of professorship, and has not only helped guide me on my path through graduate school but has done so with an aura of friendship and passion for science that is exceedingly rare. He has been unbelievably supportive in my career pursuits, has encouraged me in my endeavors both professional and personal, and has inspired me to achieve more than I would have imagined possible before entering his lab just a few short years ago. I cannot thank him enough for his continued support on this journey.

I would also like to thank my thesis committee for their unwavering support. I would like to deeply thank Dr. Hopi Hoekstra, Dr. Alexander Meissner, and Dr. Ya-Chieh Hsu. Dr. Hoekstra has not only been an outstanding collaborator, who contributed to the work on the *lincRNA-Tug1* project described in Chapter 3, but was also incredibly accessible for discussions ranging project details to career choices, and was always happy to offer mentorship at a moment's notice. Dr. Meissner has been a member of my committee from as far back as my qualifying exam, when I was still pitching the ideas that would lead to the *Fendrr* studies described in Chapter 2. He has always provided incredible scientific insight during experimental discussions, and has been unwaveringly supportive of all my choices throughout graduate school. Dr. Ya-Chieh Hsu is the latest member to my committee, but I am incredibly lucky to have had such a talented scientist as a mentor, and her kindness, generosity, and enthusiasm for my work has been inspirational. I am hugely fortunate to have had such prominent figures guiding me throughout my Ph.D. and supporting me in my career choices as well as my research.

This would not be a thesis written by someone from the Liapis family if I did not spend ample time thanking that same family. My mom and dad, Joy and Gus Liapis, are my rocks and my sounding boards, whether voluntary or otherwise. I cannot think of a single aspect of my life that is not made better through them, and I would never have made it to where I am today without their unconditional love. They taught me that the most important thing in life is, when something knocks you down, a strong person picks themselves up and continues trying. I think this lesson is uniquely applicable to the Ph.D. experience, and helped me cope with the multitude of obstacles and challenges I faced along the way. My brothers, Andrew and Vasily Liapis, have been my best friends for my entire lifetime. It's been difficult living so far from them these past few years, but they have continued to support me through love and laughter. My family taught me never to shy away from a challenge, never to be ashamed of who I am, and they gave me the courage to achieve everything I have achieved to date.

In addition to my family, the one person who influences me more than any other is Faith Birnbaum. I count myself the luckiest person in the world to have met such an inspiring, successful, hard-working person, and cannot thank her enough for helping me through my graduate studies. Our thoughtful conversations, incredible adventures, unbounded love, and total support for one another has aided me more times than I can count. In Faith I find strength to overcome obstacles, I find courage to tackle imposing challenges, I find inspiration to be more creative, and I find humility that keeps me levelheaded and always trying my best. With Faith and our dog, Prozac, by my side, I never doubted that I would be able to overcome the challenges of graduate school, and line up the career of my dreams for after my Ph.D. is complete. I will always be thankful to her for all of my successes.

Table of Contents

List of Figures

Chapter 1 – Introduction

1.1 – A Break from the Past: The noncoding genome and noncoding RNAs

The rapid development and advancement of next generation sequencing technologies has allowed biologists to investigate the portion of the genome that does not code for proteins, also known as the "noncoding" genome, with increasing depth and clarity.[1-6] RNA has been in the center of information flow from genomic content to functional output since the Central Dogma of Molecular Biology was conceived;[7] however, the prevailing paradigm of RNA functioning as just a "second messenger" in protein synthesis has long passed.

It has been 55 years since the seminal paper in which Jacob and Monod posited that RNA occupies a central role in the flow of genetic information, wherein they stated that DNA is transcribed into messenger RNAs (mRNA), which in turn serves as the template for protein synthesis.[8] The discovery of extensive transcription in the mammalian genome, far beyond what would be expected given the known number of protein-coding genes,[9] provided an important new perspective on potentially unrecognized roles for RNA. Studies over the past several decades have pointed to the presence of RNA species that do not get translated into proteins. While some of these first sightings could be explained by alternative mRNA splicing, consistent observation and subsequent analysis revealed new classes of functional RNAs. These molecules diverge from the central dogma by ascribing non-protein coding roles to RNA, and include ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), and small noncoding RNA (ncRNAs). These RNA families function in a variety of ways, notably differing from mRNA in that

they are not translated into protein at the ribosome, yet they constitute the majority of RNA mass in the cell. Amongst these noncoding RNA (ncRNA) forms, rRNA is the most abundant.[10] In addition to rRNAs, tRNAs are also necessary components of the protein synthesis machinery. Their specialized structure allows tRNAs to serve as the interface between DNA and the ribosome in the form of mRNA, as well as amino acids in order for translation to properly occur.[11] In addition to rRNA and tRNA that serve as noncoding but core regulatory RNA elements, small ncRNAs were discovered to have various functional roles in gene regulation, chromatin organization, transposon defense, genome stability, nucleotide modification, and splicing.[12-15]

Collectively these studies identified a wide range of RNAs, but still did not account for several gaps in our understanding of genome regulation. Importantly, they illustrated the diverse range of functions that RNA possess, and provided a platform for biologists to examine the still vastly unexplored regions of the noncoding genome. This incredible diversity of RNA species, in conjunction with the known pervasive transcription of the genome, prompted scientists to look for more families of noncoding RNA species. In any given cell type, at any given time, ~1% of transcription events originate from annotated protein coding regions of the genome.[2, 16] RNA sequencing efforts revealed that this phenomenon of "pervasive transcription" in noncoding regions is conserved across a diverse range of eukaryotes, and recent sequencing analysis of the eukaryote transcriptome has catalogued ~180,000 cDNAs, comprising ~20,000 protein coding genes in mice (similar numbers have been detected in other eukaryotic species, including humans).[9, 17] Interestingly, the majority of the transcripts resulted from RNAs that are

alternatively spliced and are generated from alternative promoters are from noncoding regions of the genome.[18, 19]

1.2 – The Discovery of long noncoding RNA *Xist*

The process of X-chromosome inactivation (XCI) in female cells, and the discovery of the RNA responsible, blew the doors off the long noncoding RNA (abbreviated lncRNA, and representing any noncoding RNA molecule that is longer than 200 nucleotides in length) biology field. This lncRNA, named *Xist* for X-inactive specific transcript,[20] remains the gold standard of lncRNA biology to date, and is the benchmark by which novel lncRNAs are tested for functional relevance in mammalian physiology and development.

Insights into X chromosome inactivation came from studying mice and cell lines with structurally rearranged chromosomes.[22] In some of these studies, various sections of an X chromosome were missing. Depending on which parts were missing, the X chromosome in some cases did not inactivate normally. In the interest of clarity, when describing "inactive" vs "active" chromosomes, it is typically in reference to transcription at loci, or the expression of genes, on that chromosome or specific part of that chromosome. In other studies, sections had come off the X chromosome and attached themselves onto an autosome. Again, inactivation of the modified autosome was dependent on which part of the X chromosome had transferred.[21] These experiments showed that there was a region on the X chromosome that was required and sufficient for X inactivation. This region was dubbed the X-Inactivation Center (XIC) and, in 1991, a

group of researchers showed that the XIC contained a gene encoding a stable transcript – *Xist*.[23] This gene was transcribed on only one of two alleles (only coming from one of the two female X chromosomes, and never appearing in male cells).

Attempts were made to identify a protein encoded by the *Xist* transcript but, by 1992, it was clear that the researchers had found something interesting. The *Xist* gene was transcribed from DNA into RNA, as expected. The RNA was then processed just like any other RNA, undergoing intron splicing, 3' polyadenylation (polyA) tailing, and addition of a 5' 7-methylguanosine (m^7G) cap – each of these modifications is also found in mRNA - to improve its stability.[24] Regardless, however, of the similarity in intra-nuclear processing, one aspect of *Xist* remained puzzling. Before RNA molecules can be translated into protein, they have to move out of the nucleus and into the cytoplasm of the cell. This is because ribosomes are only found in the cytoplasmic compartment. The X*ist* RNA never moved out of the nucleus, which meant it could never generate a protein.[25]

Years later, it is now understood that *Xist* is a long noncoding transcript that is expressed from one copy of the X chromosome as a dosage compensation mechanism in female cells.[33, 59] Upon expression, *Xist* spreads in both directions (5' and 3' relative to its transcriptional start site, or TSS).[60] Once transcribed, the *Xist* RNA spreads in *cis*, meaning that the transcript only binds to the allele from which it was expressed (rather than crossing genome space to bind the other allele or even other chromosomes, which is referred to as an 'in *trans*' interaction). Binding of *Xist* leads to the localization of Polycomb Repressive Complex 2 (PRC2) proteins in complex with the lncRNA, resulting in histone methylation-mediated gene repression across the X chromosome.[26, 27, 28] This

process effectively "cakes" the chromosome from which *Xist* is expressed with methylation marks, silencing it in a *cis*-repressive fashion. Studies have shown that this process is not X chromosome-dependent, and that expression of the *Xist* gene outside of its native locus (i.e. on an autosome) results in PRC2-mediated silencing of that chromosome. Yet, while the studies of *Xist* were groundbreaking and opened up the field of lncRNA biology for further study, several questions remained at the forefront of the field: how many lncRNAs are there? Can lncRNAs be grouped into functional "families"? Can lncRNAs affect mammalian development? These questions, among others prompted our group and others to delve further into the mysteries of the noncoding genome.[33]

1.3 – Mass-Identifying Novel lncRNAs

With the discovery of *Xist,* as well as a few other lncRNA genes involved in imprinting and other cellular processes (e.g. *H19* and *AIR*),[29, 30] scientists began to realize that mRNAs were just the tip of the iceberg and that, perhaps, lncRNAs represented an emerging class of regulatory RNAs that control gene expression by regulating the transcription of mRNAs in the nucleus. But, as alluded to above, how much remained below the surface was still a mystery. These early discoveries were largely accomplished using biochemical approaches, which were able to characterize many of the more abundant structural and regulatory RNAs (e.g. those highly expressed), such as those mentioned above. It was not until the advancement of full genome analysis that biologists were able to better appreciate the complexity of RNAs in the cell.

One of the first genome sequencing technologies was the development of automated Sanger sequencing, which allowed for the mapping of expressed sequence tags

(ESTs) to identify genomic regions that were actively transcribed (an early version of transcriptome analysis).[31, 32] These studies were constrained by short sequence reads, low coverage of the genome, and an incomplete reference of an assembled human genome with which to align the ESTs. Even so, many of these reads mapped to previously un-annotated regions of the genome, well outside the loci of known protein coding genes. The coarse methodology, however, prevented scientists from better understanding what these transcripts might be doing. Other technologies such as tiling DNA microarrays refined our view of genomic transcription over time,[34] but it was not until biologists had a better understanding of epigenetic modifications to DNA, and chromatin signatures, that we were able to truly delve into the biology of noncoding RNAs.

With the human and mouse reference genomes available to researchers for the purpose of aligning reads,[35, 36] chromatin signatures could be examined via chromatin immunoprecipitation (ChIP) at a genome-wide scale. Coupled with increasingly powerful sequencing technologies, ChIP allowed scientists to pull down epigenetic markers and see where signatures of active transcription were taking place in a given cell type.[37, 38, 39] Many epigenetic marks have been identified to date, and biologists are still piecing together the roles that some play in regulating gene expression, but two signatures were particularly important in the effort to identify new noncoding genes at a genome scale: Histone H3 lysine 4 trimethylation (H3K4me3), and histone H3 lysine 36 trimethylation (H3K36me3). H3K4me3 signatures tend to arise at the promoters of actively-transcribed genes, and H3K36me3 signatures show up along the "body" of actively transcribed genes. When they appear together, in so-called K4-K36 domains, it is strong evidence of an actively transcribed gene at that locus.[33, 40, 41]

In 2009, Guttman et al. published a genome-wide survey of K4-K36 domains across a variety of human and mouse cell types, in which they overlapped chromatin signature data with previously annotated reference genomes in order to see how many novel "gene signatures" could be identified using this approach.[41] Incredibly, they identified ~5,000 novel loci that possessed K4-K36 signatures and seemingly represented lncRNA genes (the transcribed region was longer than 200bp, and no annotated protein originated from those loci). This fundamentally changed the way scientists looked at long noncoding RNAs: what once was seen as a handful of examples (*Xist, AIR, H19*) suddenly exploded into potentially thousands of genome regulators that had been previously overlooked.

Subsequent studies further characterized these lncRNAs at a genomic scale, using bioinformatics approaches and large data sets. The term long noncoding RNA can now be broken down into four distinct subcategories, based on the genomic positioning of the gene body[33]: (1) long intergenic noncoding RNAs (lincRNAs) reside in empty genome space between known protein coding genes, and do not overlap sequences in either direction; (2) intronic lncRNAs reside in the introns of other spliced transcripts; (3) sense lncRNAs overlap in sequence of protein coding genes on the same strand; and (4) antisense lncRNAs overlap protein coding genes on the opposite strand. In addition, transcripts originating from the lincRNA subcategory might also share a promoter with one of their adjacent protein coding genes (by definition, it has to be the 3' one), in which case the gene is said to be driven by a bidirectional promoter (in that it controls expression of genes in both directions).

This burst of new gene discovery induced a frenzy in the noncoding world, and many people have since contributed to the field by identifying candidates that resemble mRNAs in multiple ways but are clearly lacking in their ability to code for proteins. These criteria were regarded as good benchmarks for transcripts that might be functionally significant, rather than transcriptional noise. Using protein homology queries (BLASTX)[42, 43] and codon substitution frequency (CSF) analyses, lncRNAs were assessed to lack coding potential although later biochemical approaches (such as ribosome profiling)[45] indicates that small peptides can be encoded within some lncRNAs. In subsequent studies, enhanced bioinformatics approaches (such as PhyloCSF)[44] allowed our lab and others to select candidate lincRNA genes with exceedingly low protein coding potential for further functional studies, although the question of how many of those initial 5,000 putative lncRNA genes are truly noncoding remains a point of contention today.

1.4 – Implications in Human Development and Disease

By 2011, a few dozen lncRNA genes had been found to play important regulatory roles in a variety of biological processes.[33, 41, 48] Aside from X-inactivation, which we have already discussed, other lncRNAs like *H19*[46] and *HOTAIR*[47] were found to function in the context of imprinting and development, respectively. One study involved a cell-based screen that identified dozens of lincRNAs required to maintain pluripotency.[41] Cumulatively, these efforts indicate the lincRNAs have significant functional relevance in mammalian development, specifically in mouse – it remained to be seen whether they also played important roles in human development, and potentially human disease, as well.

Previous work in our lab worked to address this problem by comprehensively identifying human lincRNA genes.[48] Unlike in earlier approaches, where novel lincRNA genes were identified based on histone methylation signature, this screen was done using RNA sequencing data from across 24 human tissues and cell lines. Transcriptome reconstruction allowed the researchers to identify all noncoding and unclassified transcripts previously annotated in the human reference genome. Using this approach, the team was able to identify human lincRNA genes by looking for transcripts that were reliably expressed, greater than 200 nucleotides in length, multi-exonic, lacking in protein coding potential, and positioned in the intergenic space between protein coding genes.

The result of this bioinformatics screen was a curated set of 4,662 human lincRNA loci, comprising 14,353 alternatively spliced transcripts and representing a huge trove of loci not present in any other sequencing database (RefSeq, UCSC, etc). As with previous catalogs of mouse lncRNAs,[41, 49] these human transcripts are processed similarly to mRNAs (3' polyA tail, 5' m^7G cap, splicing), are found within K4-K36 domains, and seemingly lack protein coding potential. Interestingly, it appears that lincRNAs are expressed in a more tissue-specific manner than mRNAs. All told, the curated human catalog reveals the lincRNAs are prevalent in humans as well as in mice, and so functional characterization of these genome regulators might have implications in human development.

One surprising facet of this human lincRNA screen was the finding that many lincRNA genes are situated in disease-associated regions of the genome.[48] Previous genome wide association studies (GWAS) identified many disease phenotypes with

associated mutations mapping to intergenic regions.[50] For this reason, the etiology of these diseases remained unexplained, while our knowledge of the noncoding genome was still in a fledgling state. With the latest human screen, however, researchers were able to identify over 400 lincRNA genes that are located within over 1,000 disease-associated regions lacking a protein coding gene. In some cases, the tissue-specific expression of the lincRNA gene correlated perfectly with the observed disease phenotype. For example, there is one human disease that has been characterized as neonatal lethal, known as alveolar capillary dysplasia with misalignment of pulmonary veins (ACD/MPV). ACD/MPV is a rare, congenital, and lethal lung disorder, characterized by decreased organization of the pulmonary vasculature and alveolar sacs, that contributes to pulmonary hypertension leading to death in human neonates.[51] It was previously known that mutations and deletions on human chromosome 16 at the locus that includes the protein-coding gene *Foxf1* are associated with ACD/MPV etiology.[52] Other research has similarly demonstrated that the FOXF1 protein plays an important role in the development of the lung and the gastrointestinal tract.[53] More recently, our lab identified a noncoding transcript originating from a homologous region in mice known as *Fendrr* (FOXF1 adjacent non-coding developmental regulatory RNA) that is encompassed by the mutations associated with the disease. One obvious question is whether mutations in *Fendrr* directly contributes to the development of ACD/MPV. This highlighted the relevance of lincRNAs not only to mouse and human development, but implicated these molecules in disease[54] and warranted further study into (1) how lincRNAs function, and (2) how malfunction can lead to mammalian disease progression.

Examples of lincRNAs being associated with human cancers abound as well. Many have been found to be differentially expressed in cancerous cell lines, while others have been shown to be regulated by cancer-related gene pathways like p53 and NF-kB (such as, as we will discuss in Chapter 4, *linc-Cox2*).[41, 55, 56] We discuss in Chapter 5 a previously described lincRNA, *HOTTIP*, which acts as an activator of gene expression at the HOXA cluster through TRX-group proteins. Knockdown of the *HOTTIP* transcript results in limb deformation and developmental defects in chickens.[57] *Linc-Tug1*, which we discuss in great detail in Chapter 3, has been associated not only with human cancers but is also one of the most downregulated genes in the European Bioinformatics Institute's database for the human disease teratozoospermia (a non-specific characterization of sperm with abnormal morphology).[58] While the functional roles of lincRNAs in these disease remains to be fully elucidated, these examples demonstrate the potential for lincRNAs involvement in diseases other than cancer. Clearly, however, there is need for further studies aimed at identifying if and how lincRNAs contribute to human disease. We have used this uncertainty as the impetus for the characterization of two lincRNAs in this thesis, with the goal of increasing our understanding of how lincRNAs can affect mammalian development and disease.

1.5 – References

1. Amaral, P. P., M. E. Dinger, T. R. Mercer, and J. S. Mattick. 2008. 'The eukaryotic genome as an RNA machine', *Science*, 319: 1787-9.

2. Ponting, C. P., P. L. Oliver, and W. Reik. 2009. 'Evolution and functions of long noncoding RNAs', *Cell*, 136: 629-41.

3. Weinberg, R. A., and S. Penman. 1968. 'Small molecular weight monodisperse nuclear RNA', *J Mol Biol*, 38: 289-304.

4. Paul, J., and J. D. Duerksen. 1975. 'Chromatin-associated RNA content of heterochromatin and euchromatin', *Mol Cell Biochem*, 9: 9-16.

5. Salditt-Georgieff, M., M.M Harpold, M.C. Wilson, J.E. Jr. Darnell. 1981. 'Large heterogeneous nuclear ribonucleic acid has three times as many 5′ caps as polyadenylic acid segments, and most caps do not enter polyribosomes', *Mol. Cell. Biol.* 1:179–87.

6. Salditt-Georgieff, M., J.E. Jr Darnell. 1982. 'Further evidence that the majority of primary nuclear RNA transcripts in mammalian cells do not contribute to mRNA', *Mol. Cell. Biol.* 2:701–7.

7. Crick, F. (1970). CENTRAL DOGMA OF MOLECULAR BIOLOGY. *Nature, 227*(5258), 561-&. doi:10.1038/227561a0

8. Jacob, F., & Monod, J. (1961). GENETIC REGULATORY MECHANISMS IN SYNTHESIS OF PROTEINS. *Journal of Molecular Biology, 3*(3), 318-&. doi:10.1016/s0022-2836(61)80072-7

9. Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., . . . S, R. G. E. R. G. (2005). The transcriptional landscape of the mammalian genome. *Science, 309*(5740), 1559-1563. doi:10.1126/science.1112014

10. Smit, S., Widmann, J., & Knight, R. (2007). Evolutionary rates vary among rRNA structural elements. *Nucleic Acids Research, 35*(10), 3339-3354. doi:10.1093/nar/gkm101

11. Agirrezabala, X., & Frank, J. (2009). Elongation in translation as a dynamic interaction among the ribosome, tRNA, and elongation factors EF-G and EF-Tu. *Quarterly Reviews of Biophysics, 42*(3), 159-200. doi:10.1017/s0033583509990060

12. Ahmad, K., & Henikoff, S. (2002). Epigenetic consequences of nucleosome dynamics. *Cell, 111*(3), 281-284. doi:10.1016/s0092-8674(02)01081-4

13. Ghildiyal, M., & Zamore, P. D. (2009). Small silencing RNAs: an expanding universe. *Nature Reviews Genetics, 10*(2), 94-108. doi:10.1038/nrg2504

14. Kiss, T. (2001). Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. *Embo Journal, 20*(14), 3617-3622. doi:10.1093/emboj/20.14.3617

15. van Wolfswinkel, J. C., & Ketting, R. F. (2010). The role of small non-coding RNAs in genome stability and chromatin organization. *Journal of Cell Science, 123*(11), 1825-1839. doi:10.1242/jcs.061713

16. Core, L. J., Waterfall, J. J., & Lis, J. T. (2008). Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science, 322*(5909), 1845-1848. doi:10.1126/science.1162228

17. Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., . . . Hayashizaki, Y. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics, 38*(6), 626-635. doi:10.1038/ng1789

18. Ravasi, T., Suzuki, H., Pang, K. C., Katayama, S., Furuno, M., Okunishi, R., . . . Mattick, J. S. (2006). Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Research, 16*(1), 11-19. doi:10.1101/gr.4200206

19. Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., . . . Gingeras, T. R. (2004). Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Research, 14*(3), 331-342. doi:10.1011/gr.2094104

20. Clark, M. B., & Mattick, J. S. (2011). Long noncoding RNAs in cell biology. *Seminars in Cell & Developmental Biology, 22*(4), 366-376. doi:10.1016/j.semcdb.2011.01.001

21. Cattanac.Bm, & Isaacson, J. H. (1967). CONTROLLING ELEMENTS IN MOUSE X CHROMOSOME. *Genetics, 57*(2), 331-&.

22. Rastan, S., & Robertson, E. J. (1985). X-CHROMOSOME DELETIONS IN EMBRYO-DERIVED (EK) CELL-LINES ASSOCIATED WITH LACK OF X-CHROMOSOME INACTIVATION. *Journal of Embryology and Experimental Morphology, 90*, 379-388.

23. Brown, C. J., Ballabio, A., Rupert, J. L., Lafreniere, R. G., Grompe, M., Tonlorenzi, R., & Willard, H. F. (1991). A GENE FROM THE REGION OF THE HUMAN X-INACTIVATION CENTER IS EXPRESSED EXCLUSIVELY FROM THE INACTIVE X-CHROMOSOME. *Nature, 349*(6304), 38-44. doi:10.1038/349038a0

24. Brown, C. J., Hendrich, B. D., Rupert, J. L., Lafreniere, R. G., Xing, Y., Lawrence, J., & Willard, H. F. (1992). THE HUMAN XIST GENE - ANALYSIS OF A 17 KB INACTIVE X-SPECIFIC RNA THAT CONTAINS CONSERVED REPEATS AND IS HIGHLY LOCALIZED WITHIN THE NUCLEUS. *Cell, 71*(3), 527-542. doi:10.1016/0092-8674(92)90520-m

25. Brockdorff, N., Ashworth, A., Kay, G. F., McCabe, V. M., Norris, D. P., Cooper, P. J., . . . Rastan, S. (1992). THE PRODUCT OF THE MOUSE XIST GENE IS A 15 KB INACTIVE X-SPECIFIC TRANSCRIPT CONTAINING NO CONSERVED ORF AND LOCATED IN THE NUCLEUS. *Cell, 71*(3), 515-526. doi:10.1016/0092-8674(92)90519-i

26. Cao, R., Wang, L. J., Wang, H. B., Xia, L., Erdjument-Bromage, H., Tempst, P., . . . Zhang, Y. (2002). Role of histone H3 lysine 27 methylation in polycomb-group silencing. *Science, 298*(5595), 1039-1043. doi:10.1126/science.1076997

27. Vire, E., Brenner, C., Deplus, R., Blanchon, L., Fraga, M., Didelot, C., . . . Fuks, F. (2006). The Polycomb group protein EZH2 directly controls DNA methylation. *Nature, 439*(7078), 871-874. doi:10.1038/nature04431

28. Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S., & Brockdorff, N. (1996). Requirement for Xist in X chromosome inactivation. *Nature, 379*(6561), 131-137. doi:10.1038/379131a0

29. Nagano, T., & Fraser, P. (2011). No-Nonsense Functions for Long Noncoding RNAs. *Cell, 145*(2), 178-181. doi:10.1016/j.cell.2011.03.014

30. Bernstein, E., & Allis, C. D. (2005). RNA meets chromatin. *Genes & Development, 19*(14), 1635-1655. doi:10.1101/gad.1324305

31. Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., . . . Grp, R. G. E. R. (2001). Functional annotation of a full-length mouse cDNA collection. *Nature, 409*(6821), 685-690. doi:10.1038/35055500

32. Schuler, G. D., Boguski, M. S., Stewart, E. A., Stein, L. D., Gyapay, G., Rice, K., . . . Hudson, T. J. (1996). A gene map of the human genome. *Science, 274*(5287), 540-546. doi:10.1126/science.274.5287.540

33. Rinn, J. L., & Chang, H. Y. (2012). Genome Regulation by Long Noncoding RNAs. *Annual Review of Biochemistry, Vol 81, 81*, 145-166. doi:10.1146/annurev-biochem-051410-092902

34. Shoemaker, D. D., Schadt, E. E., Armour, C. D., He, Y. D., Garrett-Engele, P., McDonagh, P. D., . . . Boguski, M. S. (2001). Experimental annotation of the human genome using microarray technology. *Nature, 409*(6822), 922-+. doi:10.1038/35057141

35. Lander, E. S., Int Human Genome Sequencing, C., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., . . . Int Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860-921. doi:10.1038/35057062

36. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., . . . Zhu, X. H. (2001). The sequence of the human genome. *Science, 291*(5507), 1304-+. doi:10.1126/science.1058040

37. Barski, A., Cuddapah, S., Cui, K. R., Roh, T. Y., Schones, D. E., Wang, Z. B., . . . Zhao, K. J. (2007). High-resolution profiling of histone methylations in the human genome. *Cell, 129*(4), 823-837. doi:10.1016/j.cell.2007.05.009

38. Bernstein, B. E., Mikkelsen, T. S., Xie, X. H., Kamal, M., Huebert, D. J., Cuff, J., . . . Lander, E. S. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell, 125*(2), 315-326. doi:10.1016/j.cell.2006.02.041

39. Rando, O. J., & Chang, H. Y. (2009). Genome-Wide Views of Chromatin Structure. *Annual Review of Biochemistry, 78*, 245-271. doi:10.1146/annurev.biochem.78.071107.134639

40. Mikkelsen, T. S., Ku, M. C., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., . . . Bernstein, B. E. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature, 448*(7153), 553-U552. doi:10.1038/nature06008

41. Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., . . . Lander, E. S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature, 458*(7235), 223-227. doi:10.1038/nature07672

42. Lin, M. F., Carlson, J. W., Crosby, M. A., Matthews, B. B., Yu, C., Park, S., . . . Kellis, M. (2007). Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. *Genome Research, 17*(12), 1823-1836. doi:10.1101/gr.6679507

43. Lin, M. F., Deoras, A. N., Rasmussen, M. D., & Kellis, M. (2008). Performance and scalability of discriminative metrics for comparative gene identification in 12 Drosophila genomes. *Plos Computational Biology, 4*(4). doi:10.1371/journal.pcbi.1000067

44. Lin, M. F., Jungreis, I., & Kellis, M. (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics, 27*(13), I275-I282. doi:10.1093/bioinformatics/btr209

45. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., & Weissman, J. S. (2009). Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science, 324*(5924), 218-223. doi:10.1126/science.1168978

46. Bartolomei, M. S., Zemel, S., & Tilghman, S. M. (1991). PARENTAL IMPRINTING OF THE MOUSE H19 GENE. *Nature, 351*(6322), 153-155. doi:10.1038/351153a0

47. Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., . . . Chang, H. Y. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by Noncoding RNAs. *Cell, 129*(7), 1311-1323. doi:10.1016/j.cell.2007.05.022

48. Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., & Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development, 25*(18), 1915-1927. doi:10.1101/gad.17446611

49. Ponjavic, J., Oliver, P. L., Lunter, G., & Ponting, C. P. (2009). Genomic and Transcriptional Co-Localization of Protein-Coding and Long Non-Coding RNA Pairs in the Developing Brain. *Plos Genetics, 5*(8). doi:10.1371/journal.pgen.1000617

50. Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.*Proceedings of the National Academy of Sciences of the United States of America, 106*(23), 9362-9367. doi:10.1073/pnas.0903103106

51. Janney, C.G., Askin, F.B., and Kuhn, C. (1981). CONGENITAL ALVEOLAR CAPILLARY DYSPLASIA - AN UNUSUAL CAUSE OF RESPIRATORY-DISTRESS IN THE NEWBORN. American Journal of Clinical Pathology *76*, 722-727.

52. Szafranski, P., Dharmadhikari, A.V., Brosens, E., Gurha, P., Kolodziejska, K.E., Ou, Z.S., Dittwald, P., Majewski, T., Mohan, K.N., Chen, B.*, et al.*(2013). Small

noncoding differentially methylated copy-number variants, including lncRNA genes, cause a lethal lung developmental disorder. Genome Research *23*, 23-33. doi: 10.1101/gr.141887.112.

53. Mahlapuu, M., Enerbäck, S., and Carlsson, P. (2001). Haploinsufficiency of the forkhead gene Foxf1, a target for sonic hedgehog signaling, causes lung and foregut malformations. Development 128, 2397–2406.

54. Wapinski, O., and Chang, H.Y. (2011). Long noncoding RNAs and human disease. Trends in Cell Biology *21*, 354-361.

55. Hung, T., Wang, Y. L., Lin, M. F., Koegel, A. K., Kotake, Y., Grant, G. D., . . . Chang, H. Y. (2011). Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nature Genetics, 43*(7), 621-U196. doi:10.1038/ng.848

56. Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M. J., Kenzelmann-Broz, D., . . . Rinn, J. L. (2010). A Large Intergenic Noncoding RNA Induced by p53 Mediates Global Gene Repression in the p53 Response. *Cell, 142*(3), 409-419. doi:10.1016/j.cell.2010.06.040

57. Wang, K. C., Yang, Y. W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., . . . Chang, H. Y. (2011). Long noncoding RNA programs active chromatin domain to coordinate homeotic gene expression. *Journal of Investigative Dermatology, 131*, S63-S63.

58. Platts, A. E., Dix, D. J., Chemes, H. E., Thompson, K. E., Goodrich, R., Rockett, J. C., . . . Krawetz, S. A. (2007). Success and failure in human spermatogenesis as revealed by teratozoospermic RNAs. *Human Molecular Genetics, 16*(7), 763-773. doi:10.1093/hmg/ddm012

59. Nguyen, D. K., & Disteche, C. M. (2006). Dosage compensation of the active X chromosome in mammals. *Nature Genetics, 38*(1), 47-53. doi:10.1038/ng1705

60. Engreitz, J.M. et al. The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science 341, 1237973 (2013).

Chapter 2 – Development & broad characterization of *in vivo* functional lincRNA models

2.1 – Selection and development of our lincRNA knockout models

In the wake of Cabili et al. (2011),[1] other studies found that a significant proportion of disease-associated genetic variants identified by genome-wide association studies are located in the noncoding genome; in many cases, in places where they have the potential to affect regions that control transcription (via promoters, enhancers) and noncoding RNAs that also can influence gene expression.[2] The discovery of these alterations has already contributed to a better understanding of the etiology of human diseases and has begun to yield insight into the function of these noncoding loci.[3] I was interested in studying how the noncoding genome functions and contributes to human development and disease pathology, especially when it is considered that our understanding of human disease is almost entirely contained within the realm of the <5% of the genome that is protein coding.[2]

At the outset of my graduate studies, several researchers had started to uncover functional roles for lincRNAs in mouse and human cell lines, yet very few had been tested for *in vivo* relevance. None had, at the time, been specifically tied to a mammalian disease. New tools and approaches were needed for us to be capable of assessing the function of a lincRNA in a comprehensive fashion. A long-held standard for determining the function of a previously unknown gene involves engineering mice strains with mutations and/or deletions in endogenous genes of interest.[4] Eliminating or modifying a single gene in the mouse genome provides insight into the role that gene plays in normal physiology and

disease pathogenesis; in the case of a gene with a human ortholog, the animal model approach is often employed as a proxy for understanding the role of that gene in humans.[26] In order to identify ideal candidates for functional studies, we developed a step-wise computational lincRNA selection pipeline to first identify promising lincRNAs from publicly-available datasets, followed by a genetic approach to engineer a cohort of lincRNA knockout mouse strains.[1, 5, 23] We drew from several existing data sets (both computationally-derived and experimental)[6] to create a list of promising loci that could be lincRNA candidates. All transcripts with identifiable protein coding domains or those overlapping known non-lincRNA annotations (including annotated protein coding genes and other families of noncoding RNAs, such as tRNAs or other known ncRNAs) were excluded from the list. After this initial filtering process, we subsequently removed any remaining transcripts with the potential to code for peptides via phylogenetic codon substitution frequency analysis (PhyloCSF).[7] We have previously demonstrated that this approach is capable of identifying open reading frames (ORFs) in the sequence of a transcript that can code for known peptides as small as 11 amino acids in length.[8, 9] Using this technique as well as a set of selective criteria (PhyloCSF scores of less than 200), we eliminated all but those transcripts least likely to be translated at the ribosome – even still, this remaining list was again combed over using existing ribosome profiling datasets.[9] This biochemical approach measures a transcript's capacity to occupy a ribosome; ribosomal occupancy indicates not only that said transcript has left the nucleus (if you recall, this is what alerted people to the uniqueness of *Xist* years ago), but also that it is bound and therefore probably translated in the cell.[27, 28] None of the examined candidates were shown to occupy ribosomes in a significant manner. In a final step, we looked at mass spectrometry data to discard any remaining transcripts with mapped

peptides. We chose to develop a knockout model for 18 lincRNAs selected from the initial pool which, based on extensive analysis, do not appear to exhibit protein coding potential.[5, 23]

These 18 loci possess key characteristics of previously identified lincRNA genes. (1) they each contain the presence of K4-K36 chromatin domains, indicating that they are the sites of actively transcribed genes,[6, 10] (2) RNA polymerase II (polII) binding peaks near the promoter (further evidence of active transcription),[24] and (3) presence of a transcript in humans that could be considered a syntenic ortholog (meaning it remains in the same genomic context) that would indicate conservation among mammals.[11] In order to characterize the function of the 18 lincRNA loci, and to determine whether they are required for mammalian development and/or can be implicated in mammalian disease, we generated knockout mice models for each lincRNA gene.[5] Mutant mice were generated by replacing the gene body with a β-galactosidase (beta-gal, also known as lacZ) expression cassette (see example schematic in Figure 1, below). Linearized targeting constructs, containing the lacZ and neomycin resistance genes, were electroporated into mouse embryonic stem (ES) cells derived from a 129S6Sv/Ev female to a C57BL/6N male mating.[4] Mouse ES cells carrying a heterozygous deletion of the lincRNA gene were identified by neomycin selection and loss-of-function allele screening via qPCR. Simultaneous replacement of the body of each lincRNA gene with the lacZ cassette was confirmed by gain-of-allele qPCR against the lacZ cassette. ES cells that had successfully integrated the construct were subsequently used to generate lincRNA$^{+/-}$ mice via tetraploid complementation.[4]

Figure 1:

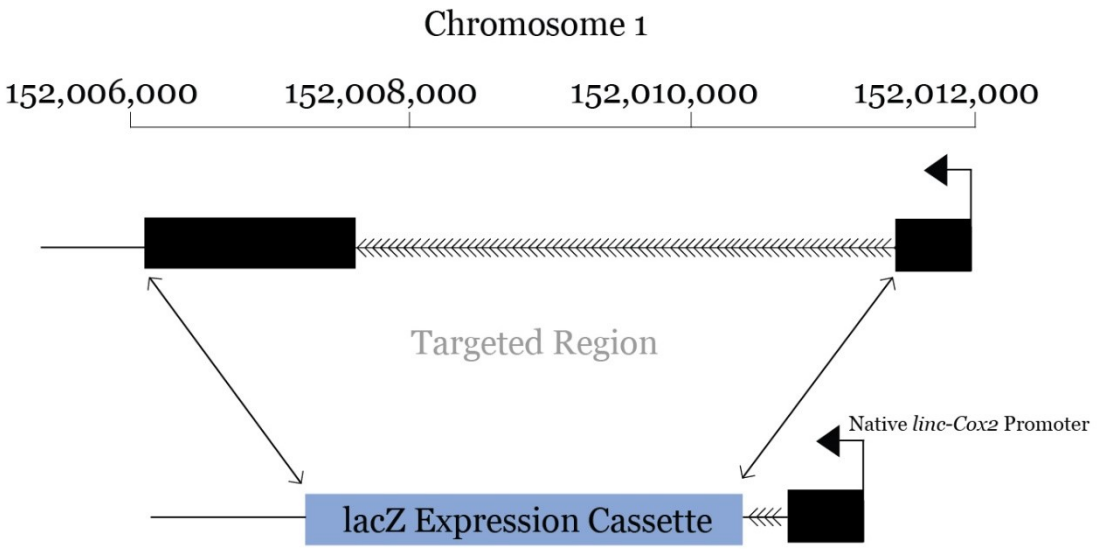


Figure 1: an example of our lincRNA knockout strategy. Represented here is lincRNA-Cox2, found on Chromosome 1 adjacent to the protein coding gene *Ptgs2*. Formerly named *Ptgs2-os2,* since it lies on the opposite strand of *Ptgs2*, linc-Cox2 was knocked out by replacing the gene downstream of exon 1 (from the beginning of intron 1 to the end of the last exon) with a lacZ expression cassette. This allows the lacZ gene to be controlled by the endogenous lincRNA promoter, enabling us to visualize the spatiotemporal expression dynamics of the locus *in vivo*. This strategy was employed with each of the 18 lincRNA knockout models we developed.

2.2 – *In vivo* spatiotemporal expression dynamics of 18 lincRNA loci

We set out to better understand the physiological relevance of lincRNAs *in vivo*, by first identifying where and when (aka spatiotemporally) our candidate pool of lincRNA loci are expressed during murine development. We examined the gene expression patterns of the candidate lincRNAs using RNA sequencing of various adult tissues and cell types (Fig. 2). We used cells and tissues from our own heterozygous mice, as well as publicly available RNA sequencing data from several tissue types. A set of lincRNAs presented highly restricted expression profiles, suggesting strong tissue specificity. Some examples: *lincRNA-Celr* is expressed in neural stem cells (NSCs), *lincRNA-Enc1* in mouse embryonic stem cells (mESCs), and *lincRNA-Cox2*[6, 25] in lungs. One lincRNA, *Tug1*, showed a unique pattern of ubiquitous expression across the panel of tissues we examined. Some lincRNAs, such as *linc-Brn1b*, do not exhibit strong expression or tissue specificity using this approach, although we have subsequently learned that it is seemingly critical for proper development of specialized structures of the brain (as we will discuss later on in the chapter).[5, 23] This could be due to the fact that brain transcriptomes were analyzed by harvesting whole-brain RNA, so if the lincRNA in question was expressed from a specialized cell population within that organ system then its expression could be diluted out. Regardless, this approach provided an interesting first look into these lincRNAs, demonstrating a host of spatiotemporal expression patterns and possible functional roles based on where and when they are expressed in murine development.

Figure 2:

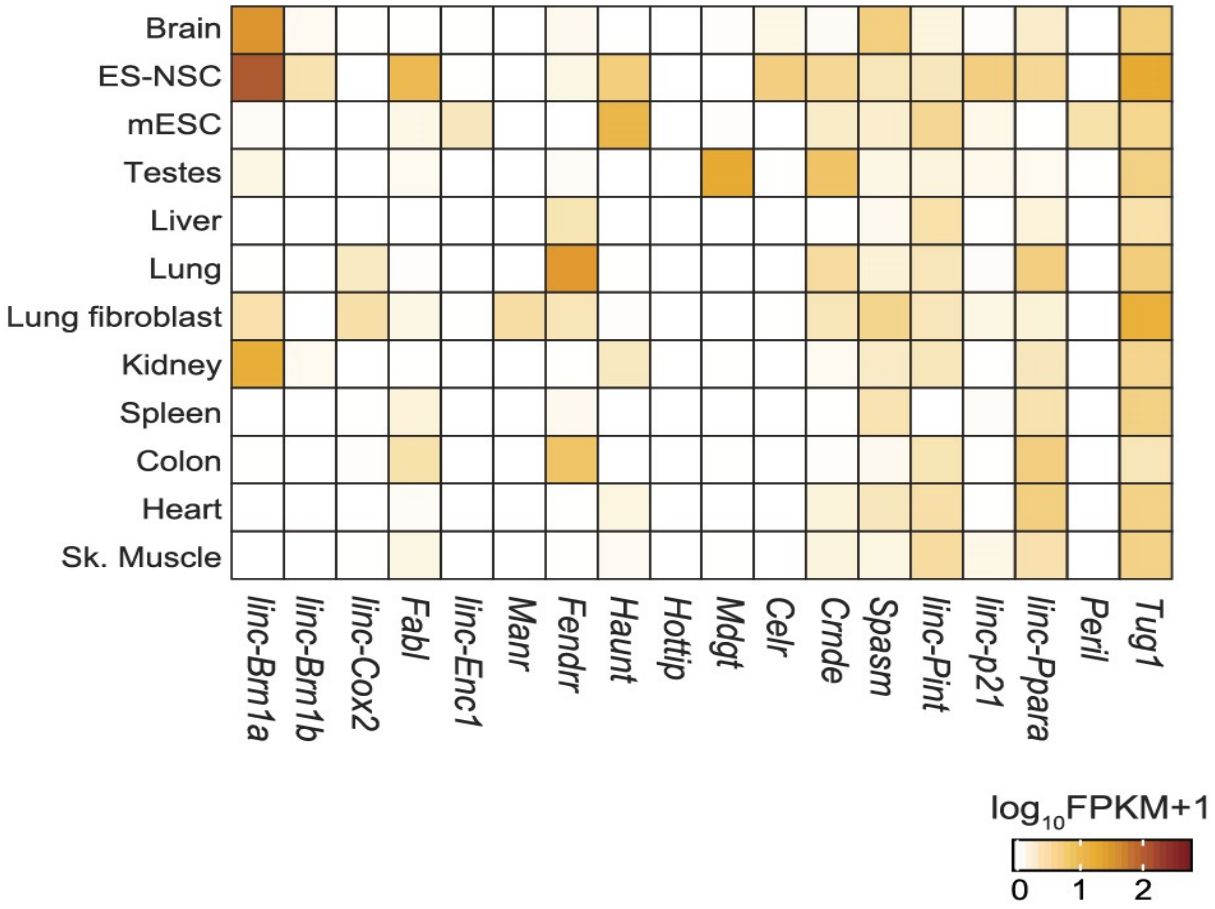


Figure 2: A heat map of expression levels of the 18 lincRNAs across a panel of adult mouse tissues and cell lines as determined by RNA sequencing. Some lincRNAs exhibit potent expression in a specific subset of tissues (such as *linc-Brn1a, linc-Fendrr,* and linc-*Haunt*). Others, including *linc-Cox2*, *linc-Brn1b,* and *linc-Enc1*, were found to be tissue specific but lowly expressed (perhaps due to being expressed only in a small population of cells within the larger organ). *LincRNA-Tug1* was unique among the 18 lincRNAs in that it appears to be highly expressed throughout development, and in a ubiquitous fashion.

In addition to examining lincRNA expression by RNA sequencing, we verified our results via lacZ staining of tissues in embryonic and adult heterozygous mice from each strain. The lacZ imaging studies were conducted by incubating harvested and paraformaldehyde-fixed tissues in 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal). X-gal is a lactose analog and, therefore, substrate for ß-galactosidase (beta-gal, otherwise known as lacZ) which, upon cleavage by the enzyme, dimerizes into 5,5'-dibromo-4,4'-dichloro-indigo.[12, 29] This reaction produces an insoluble blue coloring of the cell(s) or tissue(s) in which lacZ is expressed.[12] Coupled with our strategy for knocking out the 18 lincRNA genes, which, to reiterate, replaced the body of the lincRNA with a lacZ expression cassette, we were able to catalogue expression of each lincRNA locus across a spatiotemporal gradient via X-gal staining. The results mirror those obtained by RNA sequencing (Fig. 1), thus confirming that (1) our knockout strategy and integration of lacZ into the lincRNA locus was successful, and (2) lacZ expression is being driven by the native lincRNA promoter in each of our models. These results demonstrated enormous potential for our new strains, by serving as the first lincRNA knockout models to incorporate a reporter, dramatically increasing the number of lincRNA genetic models available, and comprising an important resource that will be used to better understand the functional contribution of lincRNAs to mammalian developmental biology and disease.

It is important to note that all lacZ staining experiments were performed using heterozygous mutant mice. As such, we have attributed lack of observation of gross morphological and developmental defects to dosage compensation by the wild type

allele.[30] Further analysis of heterozygote crosses, and knockout morphology, would yield greater insights into the functional roles of these lincRNA loci.

2.3 - Knockout mice implicate lincRNAs in mammalian development roles

To assess the requirement for each lincRNA in embryonic development and viability, we examined the progeny derived from heterozygote mating pairs for all 18 strains (Fig. 3). Genotyping of weanlings (21 days old) revealed normal Mendelian segregation of mutant alleles in 15 of the 18 strains.[31] For the three remaining strains *Peril, Mdgt* and *Fendrr,* the progeny of heterozygote intercrosses contained much lower numbers of homozygote mutants than expected. Only 13 *Peril*−/− mice (of an expected 32), and 6 *Mdgt*−/− mice (of an expected 17) were found at weaning age, indicating that deletion of *Peril* and *Mdgt* leads to reduced viability with >50% and 65% penetrance, respectively (Fig. 3, blue rows). Further examination of pups from the *Mdgt* strain revealed that, of the homozygous pups that died, all did so less than 2 weeks after birth. For *Fendrr,* no homozygous mutants were found following postnatal day zero (P0), indicating that the lethal phenotype for this strain is fully penetrant shortly after birth. Thus, 3 out of the 18 (17%) lincRNA knockout strains generated exhibit a lethal phenotype, confirming that ablation of lincRNA genes can affect the viability of affected mutants.

Figure 3:

Strain	+/+	+/-	-/-	Total	p-value
linc-Brn1a	12 (13)	32 (26)	7 (13)	51	0.1168
linc-Brn1b	16 (17)	39 (33)	11 (17)	66	0.1952
linc-Cox2	10 (10)	19 (20)	11 (10)	40	0.9277
Fabl	18 (23)	52 (45)	20 (23)	90	0.3220
linc-Enc1	16 (12)	17 (23)	13 (12)	46	0.2252
Manr	21 (20)	37 (40)	22 (20)	80	0.7886
Fendrr	36 (23)	57 (47)	0 (23)	93	8.9 E-8
Haunt	20 (19)	44 (39)	13 (19)	77	0.2741
Hottip	8 (8)	16 (17)	9 (8)	33	0.9122
Mdgt	25 (17)	37 (34)	6 (17)	68	0.0038
Celr	11 (19)	43 (38)	21 (19)	75	0.1202
Crnde	20 (19)	41 (39)	16 (19)	77	0.7302
Spasm†	13 (22)	29 (22)	47 (45)	89	0.0498
linc-Pint	14 (12)	23 (23)	9 (12)	46	0.5818
linc-p21	19 (20)	40 (39)	19 (20)	78	0.9391
linc-Ppara	13 (14)	35 (28)	8 (14)	56	0.1112
Peril	34 (32)	79 (63)	13 (32)	126	0.0005
Tug1	15 (11)	19 (21)	8 (11)	42	0.2574

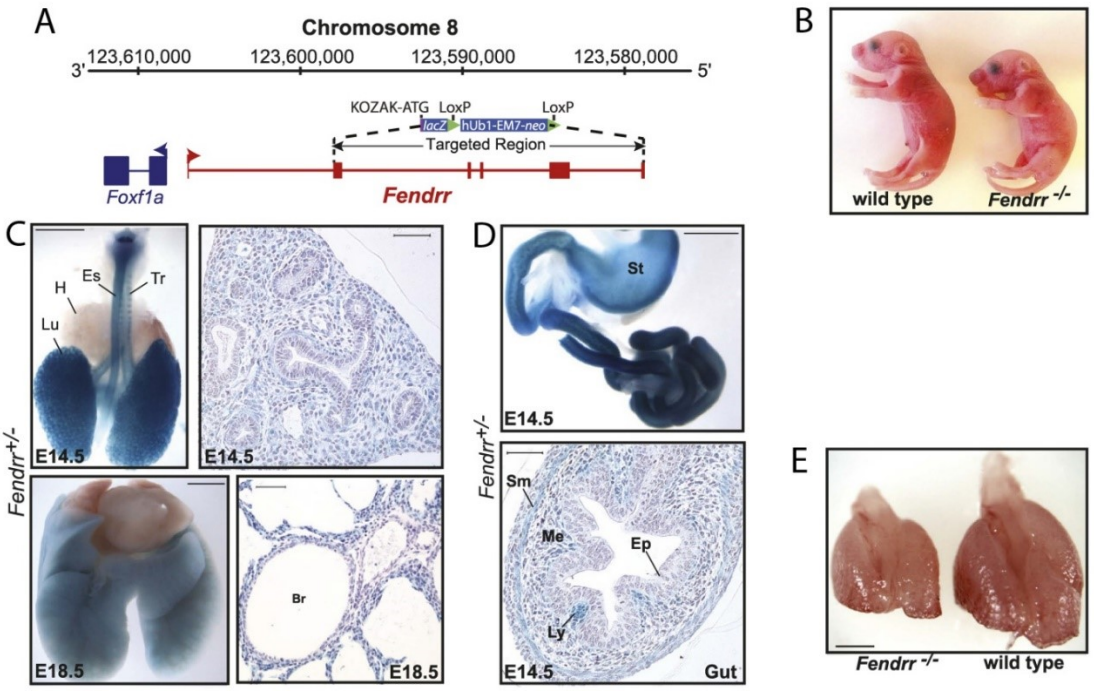
Figure 3: We examined the F1 Mendelian ratios following a series of [+/-] x [+/-] (het x het) crosses for each lincRNA loss of function strain. Of the 18 mutant model strains developed, 3 strains (*linc-Fendrr, linc-Mdgt,* and *linc-Peril)* deviated from expected normal Mendelian ratios (25%/50%/25% ratios of wild type, heterozygous, and knockout mice, respectively).

Since *lincRNA-Fendrr* was one of the three lincRNAs apparently required for postnatal development, and had already been associated via GWAS with the human disease ACD/MPV (see Chapter 1.4), I chose to focus on further characterization of this lincRNA in *in vivo* studies.

2.4 – *In vivo* characterization of defects associated with *lincRNA-Fendrr*

Fendrr is a noncoding locus that has been identified as expressing a 2.4kb transcript consisting of six exons.[5, 6, 13] It is transcribed from a bidirectional promoter shared with the protein-coding gene *Foxf1,* and is located 1.3kb from the *Foxf1* transcriptional start site (TSS, Fig. 4A). A noncoding locus in the human genome that is a positional equivalent of *Fendrr*, expressing a transcript from a syntenic region on human chromosome 16q24 (mouse chromosome 8), was identified from a catalog of human lincRNAs and resides in the region originally defined as underlying ACD/MPV.[1,50] This finding prompted us to question whether mutations in *Fendrr* could contribute to the development of ACD/MPV in humans.

Figure 4:



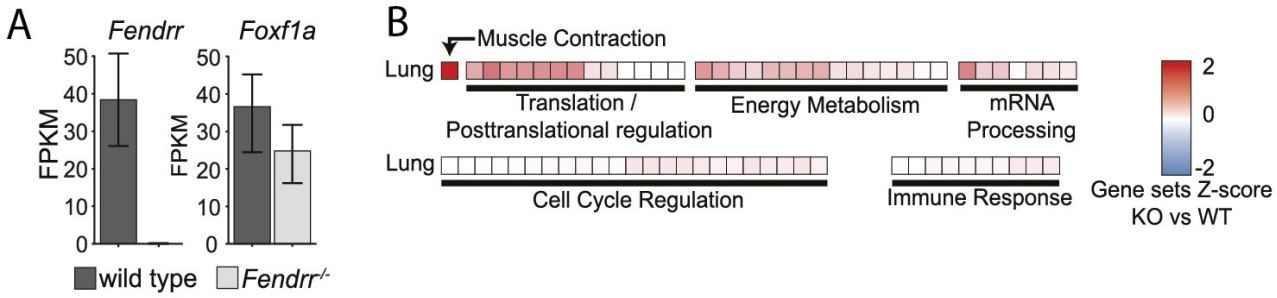
We monitored the Mendelian ratios of F1 pups following the mating of two pairs of *Fendrr*$^{+/-}$ mice. Normal Mendelian ratios were found for wild type and heterozygous pups. Importantly, we observed 8 *Fendrr*$^{-/-}$ mutant newborns (P0), all of which died within 24 hours (Fig. 3). To determine the onset age of *Fendrr* lethality we monitored the survival of embryos at early (E14.5) and late stages (E18.5) of embryonic development.[32] Normal Mendelian ratios were found at both E14.5 and E18.5, with embryos appearing macroscopically normal prior to birth, suggesting that the lethality most likely occurred during or after birth (Fig. 3). That no *Fendrr*$^{-/-}$ pups survived past 24 hours indicated a fully penetrant perinatal lethal phenotype. During the course of our initial studies, another group generated a *Fendrr* loss-of-function mouse model by inserting a three-fold tandem polyadenylation (3xpA) transcriptional

termination sequence immediately downstream of the promoter.[13] This loss of function model differs from our approach in that a 3xpA sequence allows for binding of polII and transcription initiation, but halts transcription at the elongation phase. This effectively stops the act of transcription from taking place, while keeping the DNA sequence of the endogenous locus intact.[33, 34] In contrast to our preliminary results, Grote et al. observed lethality at E13.5 due to heart and body wall (omphalocele) defects.[35] When analyzing E14.5 embryos, we found no resorbed embryos or omphalocele in our *Fendrr* homozygous mutants (Fig. 4B). Although both studies used similar genetic background strains, a possible explanation for this discrepancy may be found in the distinct targeting strategies used to knockout *Fendrr*. Regardless, both studies confirm that loss of *Fendrr* is lethal in mice.

Using RNA sequencing to create a transcriptome profile for *Fendrr* in adult mouse tissues and cell lines, we found that *Fendrr* is expressed at high levels in the adult lung, and that lower levels are detectable in colon, liver, spleen and brain (see Fig. 2). Analysis of lacZ expression using a standard X-gal staining protocol (as described above in Chapter 2.2)[12] in E14.5 and E18.5 embryos confirmed expression of *Fendrr* in these tissues as well as in the respiratory system and along the gastrointestinal tract (Fig. 4C and 4D). Perinatal lethality in mice is often associated with respiratory failure.[14] Since the highest expression levels of *Fendrr* are found in the lungs, we evaluated the ability of E18.5 to initiate breathing following surgical delivery. After cleaning of their airways, all *Fendrr*$^{-/-}$ embryos analyzed either failed to breathe or gasped and stopped breathing within 5 hours (n=7 *Fendrr* KO embryos). In contrast, respiration initiated normally and was maintained for all but one of the heterozygote and wild type embryos (n=15 *Fendrr*

HET, and n=8 WT). *Fendrr*$^{-/-}$ lungs at E14.5 were hypoplastic compared to wild type, and histological evaluation of the lungs revealed a decrease in the number and organization of pulmonary blood vessels (most notably in the arteries), as well as a general failure of vasculogenesis within the lungs of the *Fendrr*$^{-/-}$ mutants compared to wild type (n=3 *Fendrr*$^{-/-}$ and n=3 wild type). At E18.5, *Fendrr*$^{-/-}$ lungs appear to have fewer but larger alveoli (n=3 *Fendrr*$^{-/-}$ and n=3 wild type, Fig. 4C). Together, these results suggest that respiratory failure observed at birth in *Fendrr*$^{-/-}$ mice could be due to a lung maturation and vascularization defect, which recapitulates the disease phenotype of ACD/MPV. Interestingly, *Fendrr* deletion phenocopies the neighboring protein coding gene, *Foxf1*,[15] despite exhibiting a non-significant effect on *Foxf1* gene expression (Fig. 5A).

Figure 5:



We also observed expression of *Fendrr* in the esophagus and gut (Fig. 4D). Importantly, and in contrast to the model put forth by Grote et al., we did not observe *Fendrr* expression in the heart at E14.5, E18.5, or postnatally (lacZ images not shown). Sequencing results, combined with immunostaining from other groups, indicate that *Fendrr* is expressed early in the development of the lateral plate mesoderm. This tissue is

known to serve as the embryonic precursor to mesodermal tissues including vasculature.[36] Our results suggest model in which the *Fendrr* locus functions by regulating the maturation and differentiation of lateral plate mesoderm-derived tissues across several organ systems, and that it may be required for proper development and function of the mammalian lung.

We next investigated if neighboring gene expression is perturbed by the deletion of *linc-Fendrr*. We harvested lungs from E14.5 *Fendrr* KO embryos and WT littermates (n = 2) and performed differential RNA sequencing analysis. A loss of *Fendrr* expression in KO relative to WT lungs confirmed deletion of *Fendrr* (Fig. 5A). No significant change in the expression of the adjacent *Foxf1a* protein coding gene was observed in the *Fendrr* KO mice, indicating that our knockout strategy did not unintentionally affect the DNA sequence of the protein coding gene. Furthermore, genes within 1 Mb of the *Fendrr* locus in either the 5' or 3' directions were not significantly differentially expressed, suggesting that the *Fendrr* gene does not act as a local *cis* enhancer for protein coding gene expression.[37] Gene set enrichment analysis (GSEA)[38] identified gene sets involved in muscle differentiation and contraction as the most significant sets misregulated in *Fendrr* KO lungs compared to wild type (Fig. 5B). This agrees with our identification of defects in the lung vasculature of the *Fendrr* KO mice. Further studies will be needed to understand how specific changes in gene-expression patterns upon deletion of *Fendrr* contribute to the observed defects and perinatal lethality.

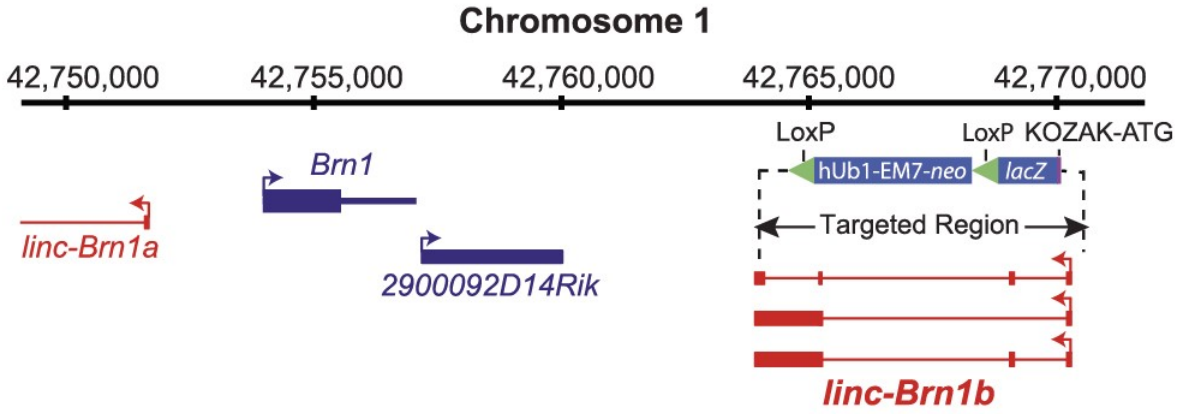
2.5 – lincRNA expression dynamics in the mammalian brain

One recurring theme we observed in our RNA sequencing of mouse tissues (our so-called "bodymap" studies) was the prevalence of lincRNA expression in the mammalian brain. In order to study lincRNA candidates of potential functional relevance in the development of neurons we used syntenic orthology, as described above,[11] and RNA sequencing methods[16, 17] to select those with putative human transcripts, and whose expression was regulated during *in vitro* neural differentiation. In order to identify qualifying candidates, we took transcripts that were expressed during a time course of embryonic stem (hESC) cell-derived human neural stem cells (hNSC) differentiation,[18] assembled and aggregated them with an existing set of RNA-Seq data using our lincRNA discovery pipeline as described above.[1] The resulting catalog contained 24,737 transcripts mapping to 14,259 putative human lincRNA genes. Furthermore, we observed 769 lincRNA genes with significant differential expression ($q<0.01$; Cuffdiff2) between any two time points during hNSC differentiation. 302 of these loci were significantly induced relative to their expression at differentiation day 0. This approach revealed that 7 lincRNAs from our mouse knockout strains have human orthologs that are dynamically induced during in vitro human neuronal differentiation. Interestingly, two of these, *linc–Brn1a* and *linc–Brn1b*, were almost exclusively expressed in NSCs as determined by RNA sequencing. These lincRNAs reside in the genomic region of the protein coding gene *Brn1*, a well-studied transcription factor involved in cortical development.[19, 20, 21]

The hypothesis that lincRNA genes play an important role in brain development is supported by the fact that ablation of certain lincRNA loci perturbs neuronal development. The *linc–Brn1b* gene is a spliced transcript approximately 3 kilobases (Kb)

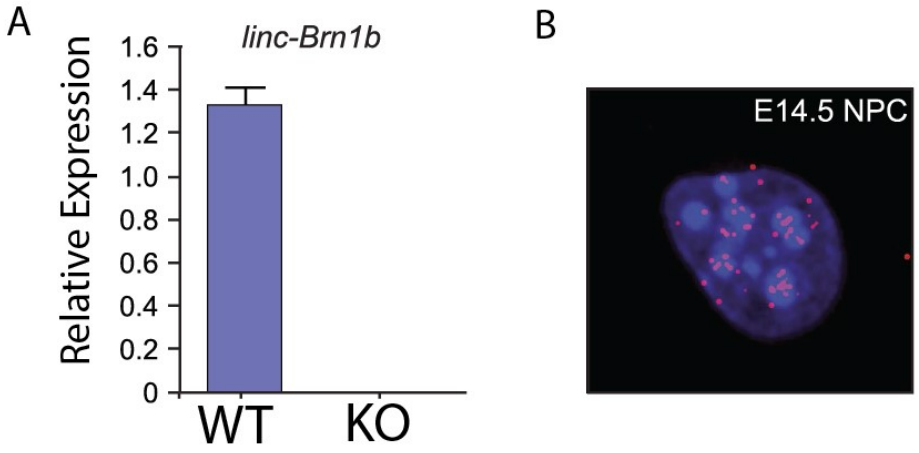
in length after maturing, and originates from a 6.8 Kb genomic locus that resides approximately 10 Kb downstream (in the 5' direction) of the *Brn1* (*Pou3f3*) protein coding gene.[21] We replaced the entire *linc–Brn1b* locus with a lacZ expression cassette (Fig. 6) in order to generate the *linc–Brn1b* knockout mice, as we have done with each of our lincRNA knockout strains.

Figure 6:



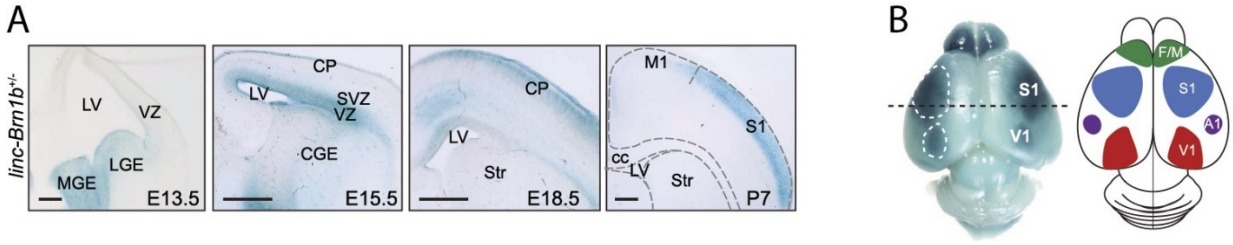
Complete ablation was confirmed by qRT–PCR (Fig. 7A) using adult brain cDNA as template.

Figure 7:



RNA fluorescence *in situ* hybridization (FISH)[39, 40] in mouse E14.5 neural progenitor cells (NPCs) isolated from the cerebral cortex demonstrated that the *linc–Brn1b* transcript is predominantly nuclear with moderate cytoplasmic expression (Fig. 7B).

Figure 8:



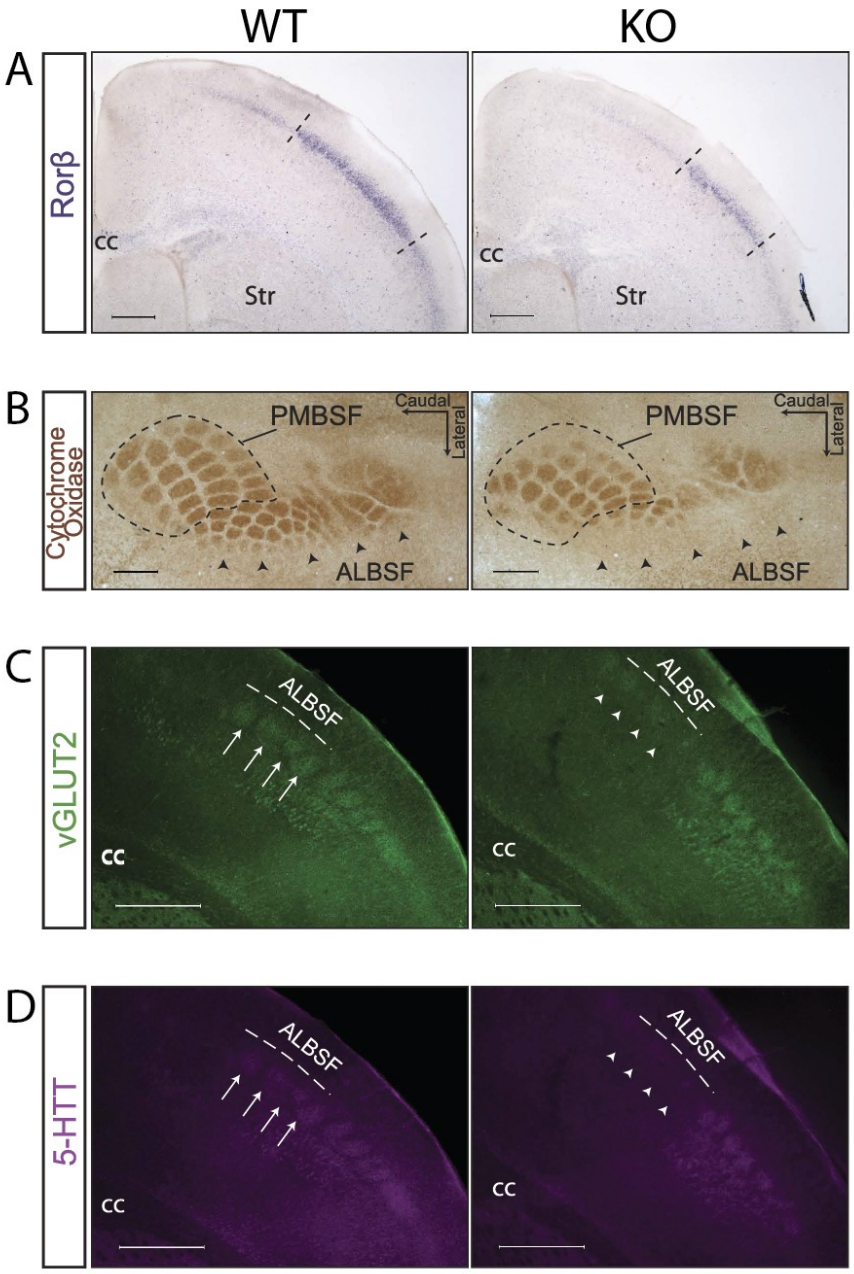
The spatiotemporal distribution of *linc–Brn1b* during brain development was not known at the time we analyzed our knockout mice. Therefore, we used *lacZ* expression from the *Brn1b* locus in heterozygote mutants to define its expression in vivo.[41] In conjunction with neuroscientists from the Arlotta Lab at Harvard, we found that *Brn1b* was expressed within neural progenitors of both the ventral and dorsal telencephalon, from as early in development as E13.5 (Fig. 8A). Characterization of expression in dorsal telencephalon showed that, by E15.5, the lincRNA locus was strongly transcribed in progenitors of the ventricular zone (VZ) and the subventricular zone (SVZ) in the developing cortex. By E18.5, we observed restricted expression in the developing upper cortical layers. In addition, we examined postnatal day 7 (P7) since this time point is regarded as adolescence in mice and provides a good intermediate stage between embryonic (<P0) and adulthood (>P21).[42, 43] Whole mount lacZ staining at P7 showed a specific distribution for *Brn1b* within the primary somatosensory cortex and the primary visual cortex (Fig. 8B). These data cumulatively suggest a potential role for *Brn1b* in the

development of distinct classes of neurons in highly specific spatial contexts within the brain.

Further investigation into the physiology of linc-Brn1b knockout mice indicated that loss of function at this locus results in a decreased number of intermediate progenitors in the developing telencephalon, reduced L2/3 neurons of the cerebral cortex, and disorganization of the barrel cortex.[22] Specifically, investigation of the *Brn1b* lacZ distribution showed high levels of expression in primary somatosensory cortex, an area that receives sensory information from the mustacial vibrissae in rodents.[47] Given the expression of *linc–Brn1b* within this region, and the requirement for proper specification of upper layer neurons, we investigated whether *linc–Brn1b* is required for the proper development of the somatosensory cortex and organization of the barrel structures. Immunostaining against RAR-related orphin beta (*Ror*β), a marker of the barrel cortex,[23] in coronal sections of P7 WT and KO mouse cortices demonstrated a reduction in the size of the barrel cortex in the *Brn1b* KO mice (Fig. 9A), with a more pronounced loss of *Ror*β+ neurons at the medial edge. Histochemical staining of cytochrome-c-oxidase activity[44] in the somatosensory cortex was examined and showed a distinct disruption of the individual barrel structures, particularly within the anteriolateral barrel subfield (ALBSF) (Fig. 9B). A reduction in overall area and number of apparent in *linc–Brn1b* mutants was also observed in the highly organized posteriomedial barrel subfield (PMBSF). These findings, consistent with the *Rorβ in situ* hybridization data, were followed up with P7 coronal sections stained against vGLUT2 (Fig. 9C) and 5-HTT (Fig. 9D), two specific markers of barrel structures.[45, 46] Analysis of both of these markers corroborates impairment of the barrels within the ALBSF, as well as a general disorganization of individual barrel structures in the *Brn1b* KO mice.

These results demonstrate the requirement of *linc–Brn1b* for the proper development of different classes of projection neurons within the cerebral cortex. They also suggest that the loss of *linc–Brn1b* could potentially have broader implications for cortical sensory processing, and warrant further analysis in future studies.

Figure 9:



The effects of *Brn1b* on the barrel structures of the developing cortex prompted us to look more closely at the expression dynamics of lincRNAs in the brain, since preliminary evidence suggested that several of our other lincRNA loss of function models have spatiotemporally-restricted expression similar to *linc-Brn1b*. Of the 18 lincRNA knockout strains that we developed, 13 strains were selected for further neurological studies based on expression in embryonic stem (ES) cell-derived neural stem cells and in brain RNA sequencing datasets.[5, 6] Toward this end, we used the knocked-in lacZ reporter gene to determine the spatiotemporal expression profiles of these lncRNAs in the brain, as described above with *linc-Brn1b*. We then performed RNA sequencing of embryonic and adult whole brains from WT and KO mice to gain insights into the transcriptional profiles of these mice, and to see how they are affected by loss of function of these lincRNA loci *in vivo*. We found that these 13 lincRNAs possess a wide range of spatiotemporal expression profiles in the brain, with several lincRNAs being highly specific to unique brain regions and cell types.[23] Combining lincRNA expression data with the differential transcriptional profiles of KO vs. WT mice allowed us to investigate whether loci annotated as lincRNAs possessed any role in brain development and neuronal physiology.

We collected whole brains from embryonic (E14.5) and adult (4-8 weeks postnatal) time points for the 13 lincRNA mutant strains selected for further neurological examination. LincRNA expression dynamics were studied within these brain samples via coronal sections collected every 80 μm (with E14.5 samples) or 240 μm (with adult samples) and subsequent X-gal staining to detect lacZ activity. Rostro–caudal (the front-to-back vector) images spanning the length of the brains were also collected, as we did previously with the *linc-Brn1b* mice. Of the 13 lncRNA strains analyzed by these methods,

10 (77%) showed clear lacZ signal in the adult brain. Consistent with our previous analysis of the effects of linc-Brn1b ablation in both embryonic and adult mice, little or no β-gal expression was observed for three strains (*linc-Mannr, linc-Halr1, and linc-Trp53cor1*) in the adult brain.[5, 6] Among those lincRNA loci with detectable β-gal signal in the adult brain, β-gal activity was also detected at E14.5 for another five loci (*linc-Enc1, linc-Eldr,* *linc-Pantr1,* and *linc-Pantr2,* and *linc-Peril*). Several of these lincRNAs demonstrated embryonic expression in regions known to give rise to the corresponding β-galactosidase-positive cell populations seen in adult mice from the same strain, indicating that these loci might function in development and maintenance of neuronal populations.

We have endeavored to examine the effects of *linc-Brn1b* LOF on murine brain development, and found a critical role for this noncoding locus in the proper formation of the barrel field in the somatosensory cortex of adolescent and adult mice. Our subsequent lacZ and RNA sequencing screen among 13 of our 18 lincRNA knockout models revealed five candidates whose spatial expression at embryonic time points mirrors that observed in adulthood (*linc-Enc1, linc-Eldr, linc-Pantr1,* and *linc-Pantr2,* and *linc-Peril*). The continuous and spatially restricted expression of these loci might indicate their importance in neurological development, maintenance, and function. Future studies will aim to examine the KO mice from each of these five strains to the extent that we have characterized the *Brn1b* locus and its role in the mammalian brain.

2.6 – Discussion of our spatiotemporal expression screen

Thousands of long noncoding RNAs have been discovered as transcribed units in mammalian genomes. However, the fraction of these new transcripts have general functional significance *in vivo* is debated. While several studies have indicated a role for

lincRNAs in diverse biological processes, it has been suggested that most transcripts could represent nonfunctional transcriptional by-products. Early critical studies of knockout strains (e.g., *Xist* and *Tsix*) did find lincRNAs implicated in X inactivation to be required for life. Yet, of the relatively few lncRNA mouse models derived since, many have displayed subtle defects or no phenotype. Combined with difficulties in finding a phenotype in mouse models such as *Malat1*[48] and *Neat1*[49], these findings have led some to suspect that acute silencing of lincRNAs results in stronger observed phenotypes than constitutive deletions, where compensatory events may obfuscate functional roles. In order to address these questions, we developed a new knockout resource that allows us to elucidate the functional relevance and physiological importance of lincRNAs through genetic ablation of lincRNA loci. The data derived from preliminary and broad-scale screens, as described above, hints at the pivotal role some of these RNA molecules might play in development, and offers up physiological insights that can only be gleaned by constitutive lincRNA knockouts. Deletion of some lincRNAs presents with specific and potent physiological abnormalities, such as *lincRNA-Fendrr* in the developing lung and *lincRNA-Brn1b* in the mammalian brain. In subsequent chapters of this Thesis, we will discuss our efforts to further our understanding of lincRNA biology through probing two of our knockout strains; these strains, which present clear morphological defects, provide unique models for lincRNA function as well as a window through which we can better understand how lincRNA function at the molecular level can affect all levels of development.

2.7 – Resulting Publications and Author Contributions

The work described in this chapter has resulted in two publications in peer-reviewed journals. The creation of our knockout lincRNA mouse models, characterization of *lincRNA-Fendrr,* and initial characterization of *lincRNA-Brn1b* was published in DOI: 10.7554/eLife.01749. Subsequent analysis of lincRNA expression in the brain, and in-depth characterization of the physiological defects of *linc-Brn1b* loss of function in the somatosensory cortex, was published in DOI: 10.1073/pnas.1411263112.

The author of this dissertation, Stephen Liapis (SL) was involved with the conception and design of research plans for the work described on *lincRNA-Fendrr*, alongside Martin Sauvageau (MS) and Loyal Goff (LAG). SL was responsible for acquisition of data in described experiments on *Fendrr* and *Brn1b*, and was specifically involved with drafting and editing corresponding sections of the paper, as well as drafting and revising sections of the entire resulting manuscript. Additionally, SL was responsible for harvesting brain tissue from adult and embryonic mice across all 13 strains analyzed in the second listed publication, and assisted with histology on the barrel cortex. Analysis and interpretation of other data included in these articles (aside from the *Fendrr* and *Brn1b* sections) was performed by Stephen Liapis with the assistance of the other authors on the paper, including Martin Sauvageau, Loyal Goff, Simona Lodato, Boyan Bonev, Abigail Groff, Chiara Gerhardinger, Diana Sanchez, Ezgi Hacisuleyman, Eric Li, Matthew Spence, William Mallard, Michael Morse, Mavis Swerdel, Michael D'Ecclessis, Jennifer Moore, Venus Lai, Guochun Gong, George Yancopoulos, David Frendewey, Manolis Kellis, Ronald Hart, David Valenzuela, Paola Arlotta, and John Rinn. Writing of this thesis chapter was done by SL with suggestions for revision by John Rinn.

2.8 – Materials and Methods

Generation of knockout mice

lincRNA knockout mice were generated by replacing the selected lincRNA gene with a *lacZ* cassette. Briefly, targeting constructs were constructed using VelociGene technology as described previously. Linearized targeting constructs, generated by gap repair cloning containing mouse lincRNA upstream and downstream homology arms flanking a KOZAK-ATG-*lacZ*-pA-*LoxP*-hUb1-EM7-*neo*(superscript R)-pA-*LoxP* cassette, were electroporated into VGF1 hybrid mouse embryonic stem (ES) cells, derived from a 129S6S v/Ev female to a C57BL/6N male mating. Mouse ES cells carrying a heterozygous deletion of the lincRNA gene were identified by loss-of-function allele screening with 2 Taqman qPCR assays. Simultaneous replacement of the lincRNA gene with the *lacZ* cassette was confirmed by gain-of-allele Taqman assays against the *lacZ* and neomycin resistance cassette. Probes were labeled with 6-carboxy-fluorescein (FAM) on their 5′ ends and BHQ-1 on their 3′ ends. Targeted ES clones were introduced into an 8-cell stage mouse embryo using the VelociMouse method. Mice were backcrossed once with C57BL/6J. Mutant mice were identified by genotyping for loss of lincRNA allele and gain of *lacZ* cassette. Toe clips, embryos or yolk sac were digested for 30 min at 95°C in 100 µl of 25 mM Sodium Hydroxide and 0.2 mM EDTA. Tissue digestion was neutralized by adding 100 µl of 40 mM Tris-HCl. PCR reactions using 4 µl of digested tissue with 10 mM *lacZ* specific and lincRNA gene specific primer pairs were then performed and run on a 2% agarose gel. PCR conditions were as follows: 5 min at 95°C followed by 35 cycles of 30 s at 95°C, 45 s at 60°C, 30 s at 72°C and a final step at 72°C for 2 min. Mice were housed under controlled pathogen-free conditions (Harvard University's Biological Research Infrastructure) and experiments were approved by the Harvard University Committee on

the Use of Animals in Research and Teaching. Viability of the 18 lincRNA mutant strains was determined at postnatal day 21 by genotyping the progeny of heterozygous inter-crosses. In the case of lethal strains, the developmental stage at which lethality occurs was determined by genotyping of embryos at E14.5 and E18.5 and newborns. Respiratory function (*Fendrr* mutant strain) was evaluated in surgically delivered E18.5 embryos from heterozygous intercrosses. After cleaning of the airways, pups were placed on a 37°C warm pad and observed for sign of breathing.

RNA isolation & Illumina RNA sequencing libraries preparation

Total RNA from embryonic and postnatal mouse tissues, neural stem cells, and neurospheres was isolated using TRIzol (Life Technology, Carlsbad, CA)/chloroform extraction followed by spin-column purification (RNeasy mini kit, Qiagen, Venlo, Netherlands) according to the manufacturer instructions. RNA concentration and purity were determined using a Nanodrop (Thermo Fisher, Waltham, MA). RNA integrity was assessed on a Bioanalyzer (Agilent, Santa Clara, CA) using the RNA 6000 RNA chip. High-quality RNA samples (RNA Integrity Number ≥8) were used for library preparation. mRNA-seq libraries were constructed using the TruSeq RNA Sample Preparation Kit (Illumina, San Diego, CA) as previously described. 500 ng total RNA was used as input for the TruSeq libraries from mouse tissues, and 200 ng for the libraries from neural stem cells and neurospheres. Prior to sequencing, libraries were run on a Bioanalyzer DNA7500 chip to assess purity, fragment size, and concentration. Libraries free of adapter dimers and with a peak region area (220–500 bp) ≥80% of the total area were sequenced. Individually barcoded samples were pooled and sequenced on the Illumina HiSeq 2000 platform.

RNA sequencing data analysis

Paired-end 101 bp reads were aligned to the mouse (mm9) reference genome assembly and, for the human neuronal differentiation time course also to the human (hg19) assembly, using Tophat2 with default options and assembled into transcripts with Cufflinks. Aligned reads and assembled transcriptome catalog were used as input for Cuffdiff2 to determine expression levels (FPKM, Fragments Per Kilobase per Million mapped reads) and differential expression between conditions using default options. CummeRbund v2.1 (http://compbio.mit.edu/cummeRbund/) was then used to process, index, and visualize the output of the Cuffdiff2 analyses. Guilt-by-Association analysis (GBA) was performed to predict the effect of gene expression changes on biological processes. *Cis*-enhancer activity was tested by determining the number of genes with differential expression in a particular Knockout vs wild type contrast within ±1 Mb window of the targeted lincRNA. 1000 random genomic intervals of the same size were obtained and interrogated in kind to determine how often the same number of differentially expressed (DE) genes could be identified. The ratio of intervals with DE genes >= the number of DE genes in the target-flanking window to the number of iterations, provided a bootstrapped p value and false discovery rate estimate.

Guilt by association (GBA) analysis

Predictive guilt by association analysis for 17/18 tested lincRNAs was conducted as follows: Pearson correlation values of FPKM expression profiles were calculated for each lincRNA to all protein coding genes across a compendium of RNA-Seq samples (combination of in-house samples and samples from. Protein coding genes were then rank-ordered and subjected to the gene set enrichment analysis described above.

Significant gene sets for a given lincRNA represent the most likely pathways/biological processes for which this lincRNA may play a role.

lacZ expression analysis and histology

Expression of the knocked-in *lacZ* reporter gene was assessed in heterozygous mice. Embryos (from E13.5 to E18.5) were fixed in 4% paraformaldehyde (PFA) in phosphate buffered saline (PBS) overnight at 4°C prior to dissection of the brain, lung and respiratory tract, digestive tract, heart, and other organs. P7 brain, from *linc–Brn1b* mutant strain, were dissected from pups transcardially perfused with 4% paraformaldehyde (PFA), and fixed overnight at 4°C. The fixed tissues were rinsed three times at room temperature in PBS, 2 mM MgCl2, 0.01% deoxycholic acid, 0.02% NP-40. X-gal staining was performed by incubating the tissues for up to 16 hr at 37°C in the same buffer supplemented with 5 mM potassium ferricyanide, 5 mM potassium ferrocyanide and 1 mg/ml X-gal. Staining reaction was stopped by washing three times in PBS at room temperature, followed by 2 hr post-fixation in 4% PFA at 4°C. Stained whole organs and sagittal brain sections were imaged using a Leica M216FA stereomicroscope (Leica Microsystems, Buffalo Grove, IL) equipped with a DFC300 FX digital imaging camera. Histology was performed at the Rodent Histopathology Service of the Dana Farber/Harvard Cancer Center Pathology Research Core. Embryos were harvested, fixed in Bouin's solution and embebbed in paraffin. Microtome sections were stained with hematoxilin and/or eosin for histological analysis.

Immunohistochemistry

Embryonic brains, dissected in cold PBS and fixed in 4% PFA/PBS overnight at 4°C, and P7 brains dissected from pups transcardially perfused with 4% PFA and post-fixed as described above, were processed for Nissl staining and immunofluorescence as previously described. Nissl-stained and immunostained sections were imaged using a Nikon 90i fluorescence microscope equipped with a Retiga Exi camera (Q-IMAGING, Surrey, Canada) and acquired with Volocity image analysis software v4.0.1 (Perkin Elmer, Waltham, MA). For quantification of overall cortical thickness, cortical layers and number of CUX1+, CTIP2+ and TLE4+ cells within the primary somatosensory cortical area, anatomically matched sections were processed (n = 3 *linc–Brn1b–/–*; n = 3 wild-type, at P7). Boxes of 300 pixels in width and spanning the thickness of the cortex were super-imposed at matched locations on each section, and the overall cortical thickness was measured as the distance from the *pia* to the white matter in each box, using ImageJ. Specific layer thicknesses were measured at the midpoint of the matched-location 300 pixel images for each of the TLE4+, CTIP2+ and SATB2+ immunofluorescence stainings using ImageJ. Layer VI thickness was measured as the distance between the dorsal edge of the TLE4+ region and the white matter. Layer V thickness was determined by the span of the CTIP2+ region, and layer II–IV thickness were measured as the SATB2+ region between the dorsal edge of the CTIP2+ stain and the pia. In each case results were expressed as mean ± SEM. Cell counts of the specific neuronal subpopulations were obtained using the ITCN plugin for ImageJ and results were expressed as mean ± SEM. A priori criteria were defined for analysis. Statistical analysis was performed using R unpaired Student's *t* test assuming equal variance was used for the pairwise comparisons.

Fluorescence *in situ* hybridization (FISH)

Single molecule FISH was performed as described. Briefly, oligonucleotide probes targeting and tiling *Peril* (48 probes) and *linc–Brn1b* (20 probes) were conjugated to Quasar 570 fluorophores and HPLC purified (Biosearch Technologies, Petaluma, CA). Dissociated E14.5 cortical neurospheres or mouse ES cells were fixed in 2% formaldehyde for 10 min, washed twice with PBS, and permeabilized with 70% ethanol. The cells were then seeded onto previously gelatinized two-chamber cover glasses. Prior to hybridization, the cells were rehydrated in wash buffer containing 10% formamide and 2 × SSC for 5 min. Probes (0.5 ng/µl final concentration) were hybridized in 10% dextran sulfate, 10% formamide, and 2 × SSC at 37°C overnight. After hybridization, cells were washed twice with wash buffer at 37°C for 30 min (with DAPI added to the second wash for nuclear staining), and twice with 2 × SSC. After the SSC wash, the cells were equilibrated in anti-fade buffer (2 × SSC, 0.4% glucose, 10 mM Tris pH 8.0) for 3–5 min. Cells were mounted in 100 µl anti-fade buffer supplemented with 1 µl of glucose oxidase (G2133-10KU; Sigma-Aldrich, St. Louis, MO) and 1 µl of catalase (C3515-10 MG; Sigma-Aldrich) and immediately imaged with a LSM 700 Inverted Confocal microscope (Zeiss, Jena, Germany). 25 Z-stacks were taken per field, using DAPI and laser 639 for excitation.

Neural stem cell (NSC) differentiation and cell culture

H1 human neural stem cells were prepared as described previously and grown at 37°C, 5% CO_2 on 1:4 diluted Matrigel-coated wells in neural proliferation medium (NPM; 50% DMEM/F12 Glutamax, 50% Neurobasal medium, 0.5X N2, 0.5X B27 without vitamin A, 20 ng/ml FGF [Life Technologies]). For differentiation, cells were plated at a density of 106 cells per well in a 6-well plate and allowed to proliferate for one day in the NPM

medium. Neural induction was then initiated by withdrawal of FGF and addition of BDNF by switching the medium to neural differentiation medium (NDM; 100% Neurobasal medium, 1X B27 without vitamin A [Life Technologies], 10 ng/ml BDNF [Peprotech, Rocky Hill, NJ].) Differentiating cultures were maintained by refreshing NDM every other day until collection. Samples of these cultures were collected at days 0, 1, 2, and 4. Remaining cells (those designated for collection at days 5, 11, and 18) were replated at day 4 at a density 106 cells per well of a Poly-D-lysine/laminin-coated 6-well plate. Cells were harvested with Accutase (Stem Cell Technologies, Vancouver, Canada) and RNA collected as described above.

2.9 – References

1. Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., & Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development, 25*(18), 1915-1927. doi:10.1101/gad.17446611

2. Wapinski, O., and Chang, H.Y. (2011). Long noncoding RNAs and human disease. Trends in Cell Biology *21*, 354-361.

3. Kumar, V., Westra, H.-J., Karjalainen, J., Zhernakova, D.V., Esko, T., Hrdlickova, B., Almeida, R., Zhernakova, A., Reinmaa, E., Voesa, U.*, et al.* (2013). Human Disease-Associated Genetic Variation Impacts Large Intergenic Non-Coding RNA Expression. Plos Genetics *9*.

4. Valenzuela, D.M., Murphy, A.J., Frendewey, D., Gale, N.W., Economides, A.N., Auerbach, W., Poueymirou, W.T., Adams, N.C., Rojas, J., Yasenchak, J., et al. (2003). High-throughput engineering of the mouse genome coupled with high-resolution expression analysis. Nat. Biotechnol. 21, 652–659

5. Sauvageau, M., Goff, L.A., Lodato, S., Bonev, B., Groff, A.F., Gerhardinger, C., Sanchez-Gomez, D.B., Hacisuleyman, E., Li, E., Spence, M., Liapis, S.C., et al. (2013). Multiple knockout mouse models reveal lincRNAs are required for life and brain development. Elife 2.

6. Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., . . . Lander, E. S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature, 458*(7235), 223-227. doi:10.1038/nature07672

7. Lin, M. F., Jungreis, I., & Kellis, M. (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics, 27*(13), I275-I282. doi:10.1093/bioinformatics/btr209

8. Guttman, M., & Rinn, J. L. (2012). Modular regulatory principles of large non-coding RNAs. *Nature, 482*(7385), 339-346. doi:10.1038/nature10887

9. Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S., & Lander, E. S. (2013). Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins. *Cell, 154*(1), 240-251. doi:10.1016/j.cell.2013.06.009

10. Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., . . . Regev, A. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology, 28*(5), 503-U166. doi:10.1038/nbt.1633

11. Zhu, J. C., Sanborn, J. Z., Diekhans, M., Lowe, C. B., Pringle, T. H., & Haussler, D. (2007). Comparative Genomics search for losses of long-established genes on the human lineage. *Plos Computational Biology, 3*(12), 2498-2509. doi:10.1371/journal.pcbi.0030247

12. Watson, C.M., Trainor, P.A., Radziewic, T., Pelka, G.J., Zhou, S.X., Parameswaran, M., Quinlan, G.A., Gordon, M., Sturm, K., and Tam, P.P.L. (2008). Application of lacZ transgenic mice to cell lineage studies. Methods Mol. Biol. 461, 149–164. doi: 10.1007/978-1-60327-483-8_10.

13. Grote, P., Wittler, L., Hendrix, D., Koch, F., Währisch, S., Beisaw, A., Macura, K., Bläss, G., Kellis, M., Werber, M., et al. (2013). The Tissue-Specific lncRNA Fendrr Is an Essential Regulator of Heart and Body Wall Development in the Mouse. Dev. Cell 24, 206–214. doi: 10.1016/j.devcel.2012.12.012.

14. Eggan, K.K., Akutsu, H.H., Loring, J.J., Jackson-Grusby, L.L., Klemm, M.M., Rideout, W.M.W., Yanagimachi, R.R., and Jaenisch, R.R. (2001). Hybrid vigor, fetal overgrowth, and viability of mice derived by nuclear cloning and tetraploid embryo complementation. Proc. Natl. Acad. Sci. U.S.a. 98, 6209–6214.

15. Mahlapuu, M., Enerbäck, S., and Carlsson, P. (2001). Haploinsufficiency of the forkhead gene Foxf1, a target for sonic hedgehog signaling, causes lung and foregut malformations. Development 128, 2397–2406.

16. Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology, 31*(1), 46-+. doi:10.1038/nbt.2450

17. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., . . . Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq

experiments with TopHat and Cufflinks. *Nature Protocols, 7*(3), 562-578. doi:10.1038/nprot.2012.016

18. Goff, L. A., Davila, J., Swerdel, M. R., Moore, J. C., Cohen, R. I., Wu, H., . . . Hart, R. P. (2009). Ago2 Immunoprecipitation Identifies Predicted MicroRNAs in Human Embryonic Stem Cells and Neural Precursors.*Plos One, 4*(9). doi:10.1371/journal.pone.0007192

19. McEvilly, R. J., de Diaz, M. O., Schonemann, M. D., Hooshmand, F., & Rosenfeld, M. G. (2002). Transcriptional regulation of cortical neuron migration by POU domain factors. *Science, 295*(5559), 1528-1532. doi:10.1126/science.1067132

20. Sugitani, Y., Nakai, S., Minowa, O., Nishi, M., Jishage, K., Kawano, H., . . . Noda, T. (2002). Brn-1 and Brn-2 share crucial roles in the production and positioning of mouse neocortical neurons. *Genes & Development, 16*(14), 1760-1765. doi:10.1101/gad.978002

21. Dominguez, M. H., Ayoub, A. E., & Rakic, P. (2013). POU-III Transcription Factors (Brn1, Brn2, and Oct6) Influence Neurogenesis, Molecular Identity, and Migratory Destination of Upper-Layer Cells of the Cerebral Cortex.*Cerebral Cortex, 23*(11), 2632-2643. doi:10.1093/cercor/bhs252

22. Petersen, C. C. H. (2007). The functional organization of the barrel cortex. *Neuron, 56*(2), 339-355. doi:10.1016/j.neuron.2007.09.017

23. Goff, L. A., Groff, A. F., Sauvageau, M., Trayes-Gibson, Z., Sanchez-Gomez, D. B., Morse, M., . . . Rinn, J. L. (2015). Spatiotemporal expression and transcriptional perturbations by long noncoding RNAs in the mouse brain.*Proceedings of the National Academy of Sciences, 112*(22), 6855-6862. doi:10.1073/pnas.1411263112

24. Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., . . . Snyder, M. (2010). Variation in Transcription Factor Binding Among Humans. *Science, 328*(5975), 232-235. doi:10.1126/science.1183621

25. Carpenter, S., Aiello, D., Atianand, M. K., Ricci, E. P., Gandhi, P., Hall, L. L., . . . Fitzgerald, K. A. (2013). A Long Noncoding RNA Mediates Both Activation and

Repression of Immune Response Genes. *Science, 341*(6147), 789-792. doi:10.1126/science.1240925

26. Justice, M. J., Siracusa, L. D., & Stewart, A. F. (2011). Technical approaches for mouse models of human disease.*Disease Models & Mechanisms, 4*(3), 305-310. doi:10.1242/dmm.000901

27. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., & Weissman, J. S. (2009). Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science, 324*(5924), 218-223. doi:10.1126/science.1168978

28. Ingolia, N. T., Lareau, L. F., & Weissman, J. S. (2011). Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell, 147*(4), 789-802. doi:10.1016/j.cell.2011.10.002

29. Ozaki, M., Matsumura, K., Kaneko, S., Satoh, M., Watanabe, Y., & Aoyama, T. (1993). A VACCINIA VIRUS VECTOR FOR EFFICIENTLY INTRODUCING INTO HIPPOCAMPAL SLICES. *Biochemical and Biophysical Research Communications, 193*(2), 653-660. doi:10.1006/bbrc.1993.1674

30. Lifschyt.E, & Lindsley, D. L. (1972). ROLE OF X-CHROMOSOME INACTIVATION DURING SPERMATOGENESIS.*Proceedings of the National Academy of Sciences of the United States of America, 69*(1), 182-&. doi:10.1073/pnas.69.1.182

31. Wright, S. (1931). Evolution in Mendelian populations. *Genetics, 16*(2), 0097-0159.

32. Downs, K. M., & Davies, T. (1993). STAGING OF GASTRULATING MOUSE EMBRYOS BY MORPHOLOGICAL LANDMARKS IN THE DISSECTING MICROSCOPE. *Development, 118*(4), 1255-1266.

33. Friedrich, G., & Soriano, P. (1991). PROMOTER TRAPS IN EMBRYONIC STEM-CELLS - A GENETIC SCREEN TO IDENTIFY AND MUTATE DEVELOPMENTAL GENES IN MICE. *Genes & Development, 5*(9), 1513-1523. doi:10.1101/gad.5.9.1513

34. Gnatt, A. L., Cramer, P., Fu, J., Bushnell, D. A., Kornberg, R. D. (2001) "Structural Basis of Transcription: An RNA Polymerase II Elongation Complex at 3.3 Å Resolution" Science 292:1876

35. Doyonnas, R., Kershaw, D. B., Duhme, C., Merkens, H., Chelliah, S., Graf, T., & McNagny, K. M. (2001). Anuria, omphalocele, and perinatal lethality in mice lacking the CD34-related protein podocalyxin. *Journal of Experimental Medicine, 194*(1), 13-27. doi:10.1084/jem.194.1.13

36. Tonegawa, A., Funayama, N., Ueno, N., & Takahashi, Y. (1997). Mesodermal subdivision along the mediolateral axis in chicken controlled by different concentrations of BMP-4. *Development, 124*(10), 1975-1984.

37. Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America, 100*(16), 9440-9445. doi:10.1073/pnas.1530509100

38. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.*Proceedings of the National Academy of Sciences of the United States of America, 102*(43), 15545-15550. doi:10.1073/pnas.0506580102

39. Hacisuleyman, E., Goff, L. A., Trapnell, C., Williams, A., Henao-Mejia, J., Sun, L., . . . Rinn, J. L. (2014). Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nature Structural & Molecular Biology, 21*(2), 198-+. doi:10.1038/nsmb.2764

40. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A., & Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods, 5*(10), 877-879. doi:10.1038/nmeth.1253

41. Lein, E. S., Hawrylycz, M. J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., . . . Jones, A. R. (2007). Genome-wide atlas of gene expression in the adult mouse brain. *Nature, 445*(7124), 168-176. doi:10.1038/nature05453

42. Ashwell, K. W. S., Waite, P. M. E., & Marotte, L. (1996). Ontogeny of the projection tracts and commissural fibres in the forebrain of the tammar wallaby (Macropus eugenii): Timing in comparison with other mammals. *Brain Behavior and Evolution, 47*(1), 8-22. doi:10.1159/000113225

43. Dunlop, S. A., Tee, I. B. G., Lund, R. D., & Beazley, L. D. (1997). Development of primary visual projections occurs entirely postnatally in the fat-tailed dunnart, a marsupial mouse, Sminthopsis crassicaudata. *Journal of Comparative Neurology, 384*(1), 26-40.

44. Capaldi, R. A. (1990). STRUCTURE AND FUNCTION OF CYTOCHROME-C-OXIDASE. *Annual Review of Biochemistry, 59*, 569-596.

45. Takamori, S., Rhee, J. S., Rosenmund, C., & Jahn, R. (2001). Identification of differentiation-associated brain-specific phosphate transporter as a second vesicular glutamate transporter (VGLUT2). *Journal of Neuroscience, 21*(22).

46. Caspi, A., Sugden, K., Moffitt, T. E., Taylor, A., Craig, I. W., Harrington, H., . . . Poulton, R. (2003). Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene. *Science, 301*(5631), 386-389. doi:10.1126/science.1083968

47. Woolsey, T. A., & Vanderlo.H. (1970). STRUCTURAL ORGANIZATION OF LAYER-IV IN SOMATOSENSORY REGION (SI) OF MOUSE CEREBRAL CORTEX . DESCRIPTION OF A CORTICAL FIELD COMPOSED OF DISCRETE CYTOARCHITECTONIC UNITS. *Brain Research, 17*(2), 205-&. doi:10.1016/0006-8993(70)90079-x

48. Tripathi, V., Ellis, J. D., Shen, Z., Song, D. Y., Pan, Q., Watt, A. T., . . . Prasanth, K. V. (2010). The Nuclear-Retained Noncoding RNA MALAT1 Regulates Alternative Splicing by Modulating SR Splicing Factor Phosphorylation.*Molecular Cell, 39*(6), 925-938. doi:10.1016/j.molcel.2010.08.011

49. Clemson, C. M., Hutchinson, J. N., Sara, S. A., Ensminger, A. W., Fox, A. H., Chess, A., & Lawrence, J. B. (2009). An Architectural Role for a Nuclear Noncoding RNA: NEAT1 RNA Is Essential for the Structure of Paraspeckles.*Molecular Cell, 33*(6), 717-726. doi:10.1016/j.molcel.2009.01.026

50. Szafranski, P., Dharmadhikari, A.V., Brosens, E., Gurha, P., Kolodziejska, K.E., Ou, Z.S., Dittwald, P., Majewski, T., Mohan, K.N., Chen, B.*, et al.*(2013). Small noncoding differentially methylated copy-number variants, including lncRNA genes, cause a lethal lung developmental disorder. Genome Research *23*, 23-33. doi: 10.1101/gr.141887.112.

Chapter 3 – *In Vivo* Characterization of *lincRNA-Tug1*

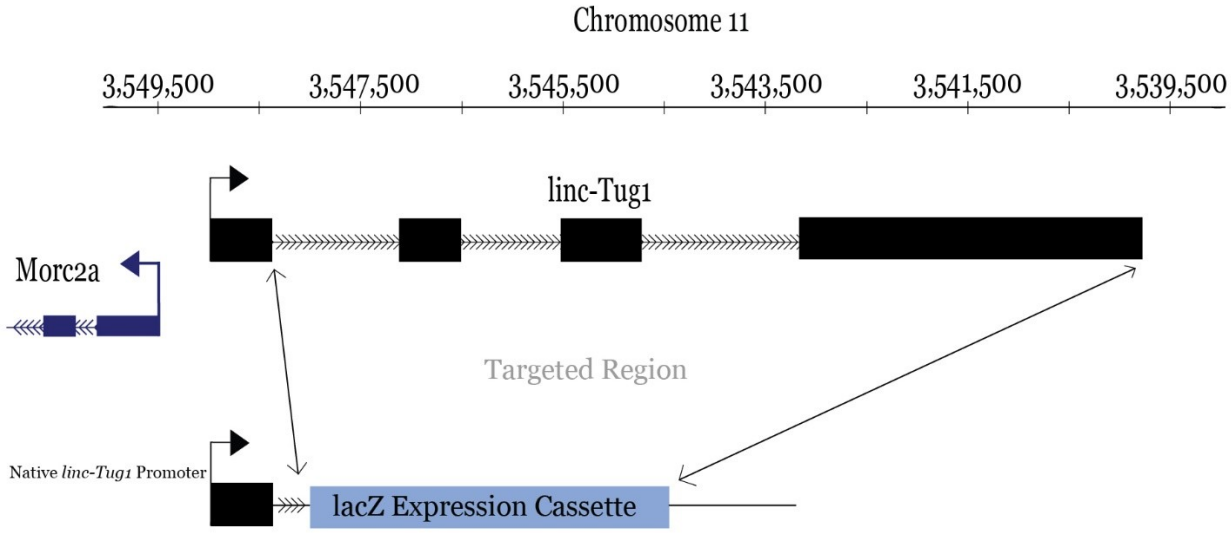
3.1 – Taurine-Upregulated Gene 1 (*Tug1*)

In the post genomic era, thousands of long intergenic noncoding RNAs have been discovered as transcribed units within mammalian genomes. Several studies have indicated a role for some of these lincRNAs in diverse biological processes, and efforts throughout the field (including ours) have cumulatively swayed the previously held belief that most of these transcripts could represent nonfunctional transcriptional "noise".[1-3] Various *in vitro* experimental models are making it increasingly clear that lincRNAs possess a set of diverse features (e.g. local repeats, ability to form ribonucleoprotein complexes, etc.) that point toward functional relevance.[4-6] Furthermore, analysis of our *in vivo* lincRNA knockout mouse strains has confirmed this RNA species to be highly tissue-specific, developmentally regulated, and located in disease-associated loci (see chapters 1 & 2 of this thesis).[7] Ablation of lincRNAs via our loss of function animal models has already yielded insights into the biological functions of these loci, as with the role of *linc-Brn1b* in murine somatosensory development,[7, 8] yet the complete characterization of a lincRNA, from *in vivo* mouse phenotype induced by loss of function, to the molecular mechanism of that RNA molecule, remains an elusive goal. Here we describe the most complete such characterization of a lincRNA locus, to our knowledge, to date. The deletion of one lincRNA, *Tug1* (Taurine Upregulated Gene 1) results in fully penetrant male sterility. Given the presence of a human ortholog (*TUG1*), as well as the recent explosion of interest in lncRNA biology, characterizing *Tug1* from *in vivo* phenotype to molecular mechanism would further demonstrate the importance of lincRNAs in development and human disease.

Tug1 was first identified by Young et al. (2005),[9] following a screen to identify genes that are upregulated during differentiation of the murine retina. Taurine is a derivative of the amino acid cysteine (Cys) that is necessary for a variety of physiological processes, including photoreceptor maturation. Several studies have demonstrated that taurine deficiency in cats, rodents, and primates leads to a failure of photoreceptors to develop properly during maturation and, if adults are deprived of taurine will also induce photoreceptor degeneration in the adult retina.[10, 11] Taurine is present at high levels in the murine retina during development, and exogenous taurine supplementation stimulates the production of rod photoreceptors.[12] This induction has been found to be due to signals mediated by the binding of taurine to glycine-receptor a2 and GABA(A) receptors.[13] Young et al. performed a screen to identify the genes that are regulated after signaling through these receptors, and found *Tug1* to be one of the most consistently and significant upregulated genes.[9] We subsequently identified this gene in collaboration with the Lander lab in the aforementioned lincRNA locus screen,[14] noticing a strong K4-K36 domain at the promoter and later identifying strong patterns of expression in a ubiquitous manner throughout development. Our lacZ knockout model for *linc-Tug1* confirmed this by RNA sequencing and lacZ expression (see chapter 2).[7]

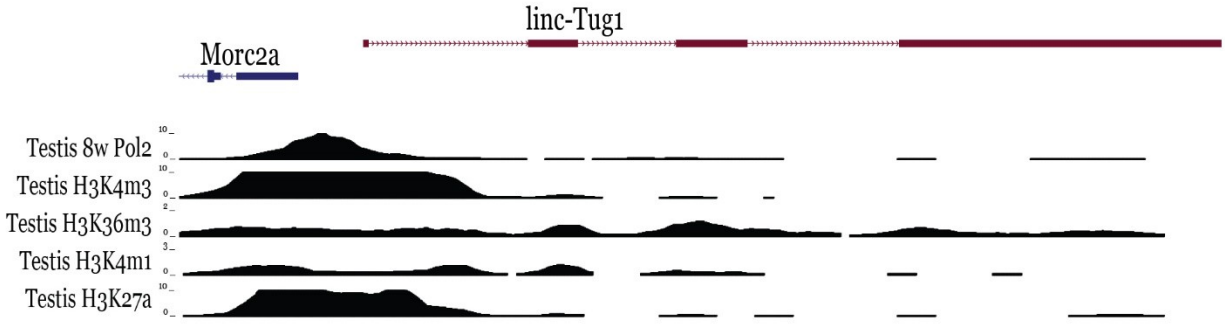
Tug1 is a 9kb transcript originating from a noncoding locus on Chromosome 11 and consists of four exons (Fig. 10). It is transcribed from a bidirectional promoter shared with the protein-coding gene *Morc2a,* and is located 2kb from the *Morc2a* transcriptional start site (TSS).

Figure 10:



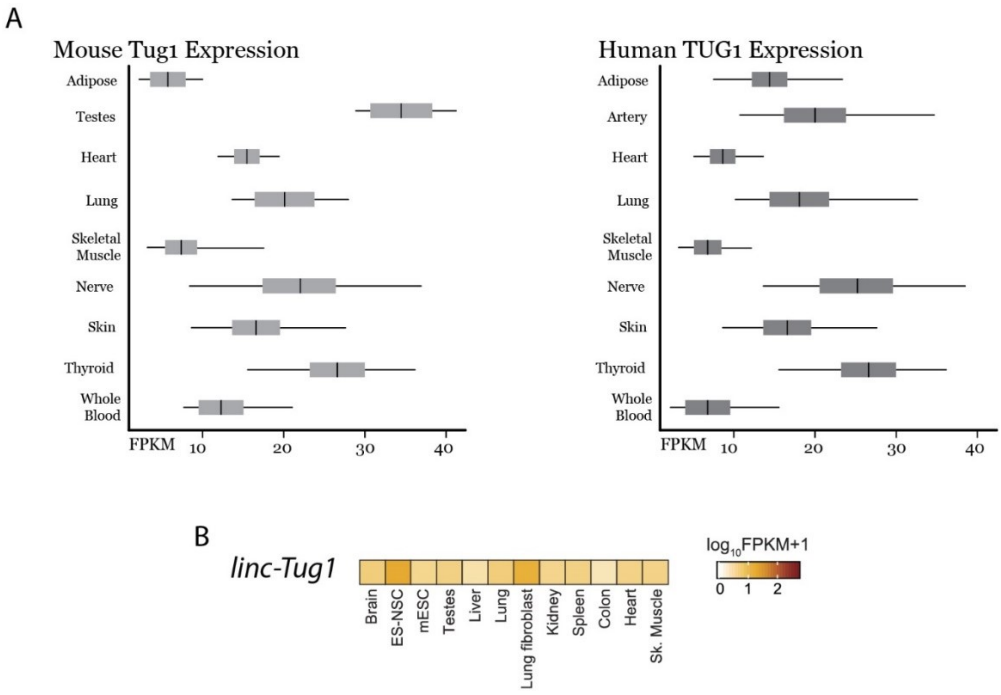
A human ortholog, *TUG1*, was identified within our catalog of human lncRNAs.[15] Our lab previously reported that this lncRNA associates with PRC2 in human cell lines (HeLa, HFF).[16] The *Tug1* locus is, as previously mentioned, demarcated by a strong K4-K36 domain, with histone H3 lysine 4 trimethylation at the promoter and H3 lysine 36 trimethylation peaks across each of the four exons (Fig. 11). RNA polymerase II binding (pol2) is observed at the promoter, further supporting the notion that the locus is actively transcribed. However, we also observe H3k4me1 and H3k27ac signal as well, which are regarded as marks of DNA enhancers.[17, 18] The presence of canonical K4-K36 chromatin marks, as well as RNA sequencing and lacZ expression data, led us to believe that this locus encodes a fully transcribed and processed functional RNA molecule.

Figure 11:



Additionally, *lincRNA-Tug1* is conserved across mammalian species, has high expression across multiple tissues in mice (Fig. 12 A [left panel] and Fig. 12 B)[7, 14] and humans (GTEx, Fig. 12 A [right panel])[15, 19], and exhibits extremely low protein coding potential (see Chapter 1 and Sauvageau et al., 2013).[7]

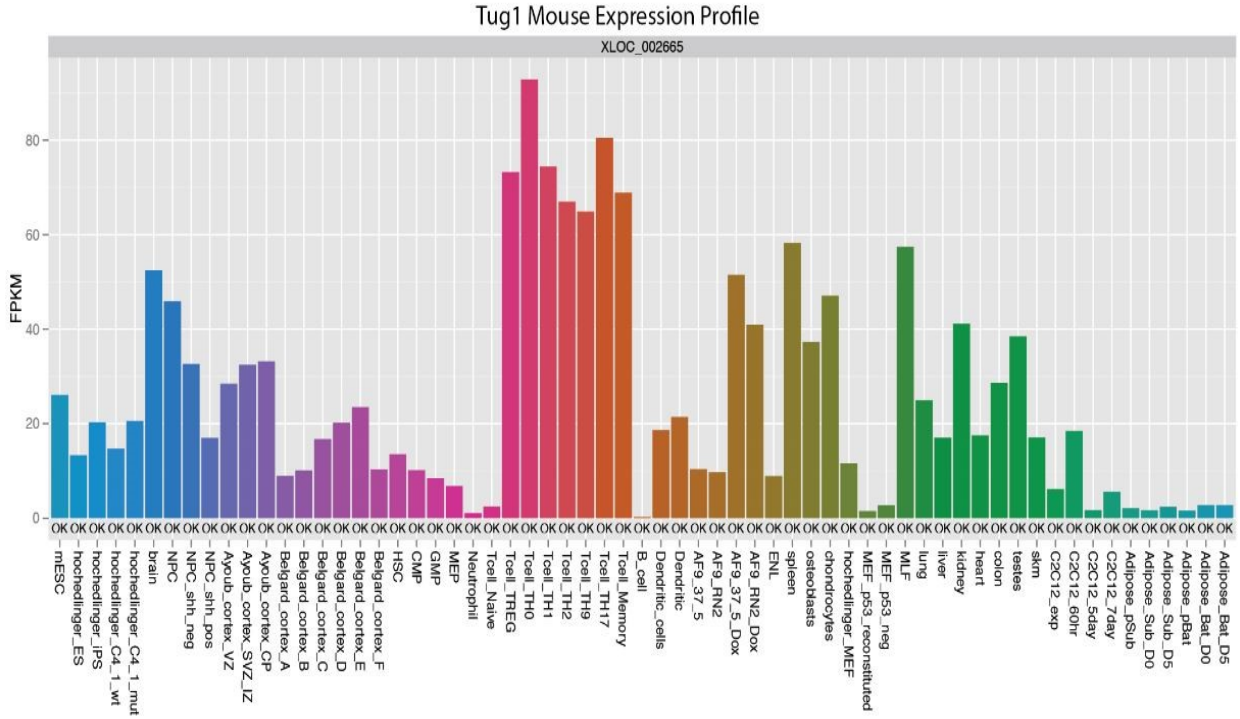
Figure 12:



3.2 – Tug1 deletion results in male-specific infertility in mice

Using RNAseq expression profiling from adult and embryonic (E14.5) *Tug1*$^{+/-}$ mouse tissues and cell lines, we found that *Tug1* is ubiquitously expressed at high levels throughout murine development (E14.5 and adult samples, Fig. 13).

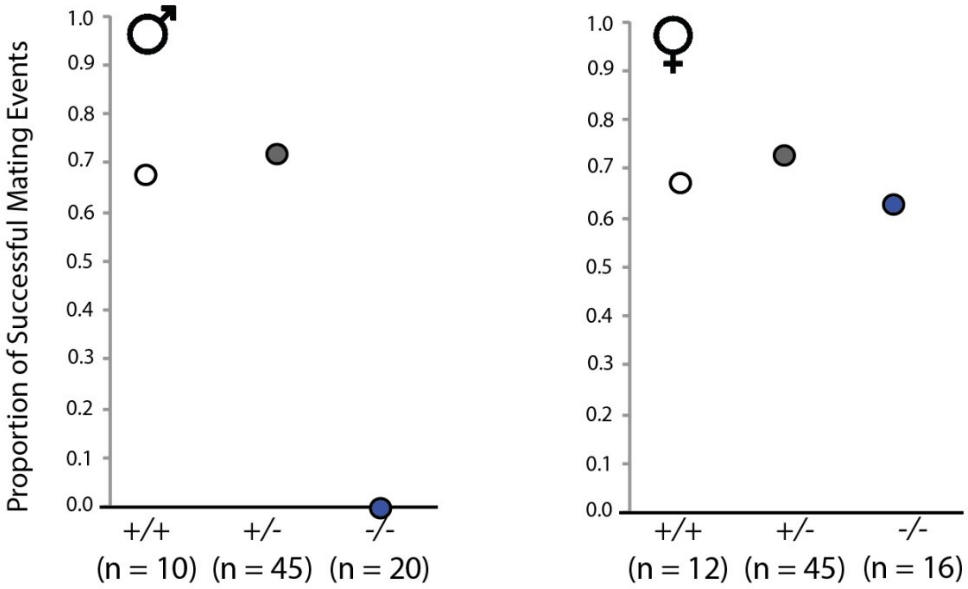
Figure 13:



Analysis of *lacZ* expression using a standard X-gal staining protocol[20] in E14.5 and adult tissues confirmed expression of *Tug1* throughout the body. Our preliminary survey suggested a housekeeping gene-like role for the *Tug1* locus[21] but, even with such ubiquitous expression in mice and humans, it remained unclear what the function of this locus might be. Following this overview of the *Tug1* expression landscape, I transitioned

to examining homozygous knockout mutants (*Tug1*$^{-/-}$) to screen for morphological defects associated with a loss of function at this locus.[22] It quickly became clear that *Tug1*$^{-/-}$ males are infertile**.** To determine the nature of *Tug1*-related infertility, we monitored the plug and litter production rates of both male and female mice. Normal plug rates were observed for *Tug1*$^{-/-}$ mice (data not shown), suggesting that the failure to impregnate most likely occurred after copulation. We observed normal litter production (proportion of plugging events resulting in a viable litter, and litter size) for *Tug1* WT (+/+) and HZ (+/-) males, as well as *Tug1* females of all genotypes (when mated to WT or HZ males). Interestingly, we observed 20 *Tug1*$^{-/-}$ mutant males, all of which failed to produce a viable litter (Fig. 14**)**. It is important to note that a failure to impregnate was noted *after* a plug was identified in each mating pair – this indicates that a lack of litters born to *Tug1*$^{-/-}$ males was not due to decreased sexual activity, but rather a physiological aberration associated with *Tug1* deletion in male mice.

Figure 14:



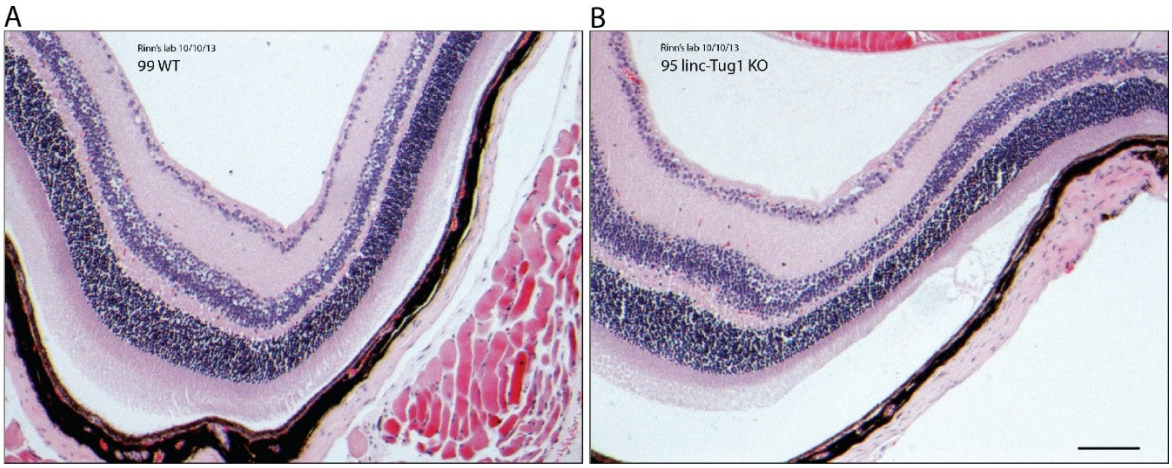
3.3 – Genetic deletion of the *Tug1* locus does not result in retinal defects

In parallel to examining the reproductive fitness of *Tug1* mutant mice, we also briefly investigated the retinas of *Tug1* KO mice, as it was found in a screen for genes involved in photoreceptor development by the aforementioned Young et al. (2005) paper. In this paper, following a screen that identified the Tug1 locus as upregulated in the presence of taurine (see Chapter 3.1), the transcript originating from the *Tug1* genomic locus was knocked down using a small interfering RNA (siRNA) targeting exon 2 of the processed RNA in rats (not mice). These molecules were co-electroporated into developing retinal cells along with a CAG-GFP construct that labels cells into which the electroporation was successful. CAG-GFP, when used in this way, is supposed to primarily label photoreceptor cells since only mitotic cells will incorporate the construct, and over 70% of mitotic cells in the P0 rat retina are rod photoreceptors (the authors do not indicate whether this proportion is comparable in non-rat mammalian retinas). The authors found a significant reduction in green-labeled cells following *Tug1* siRNA knockdown relative to wild type, and cDNA microarrays found a downregulation of genes involved in photoreceptor development in the *Tug1* knockdown samples.[9]

This approach, while appropriate at the time it was published, is technically limited by today's standards and is problematic for several reasons: (1) targeted knockdown of a lincRNA with siRNAs is very difficult due to lack of conservation of primary lincRNA sequence relative to mRNAs,[23] (2) siRNAs are more susceptible to producing off-target effects than genetic deletion,[24] and (3) the readout for photoreceptor development (GFP) can derive up to 30% of its signal from non-photoreceptor mitotic cells in the retina and

lead to false positive conclusions.[9] As such, we felt that our approach of analyzing *Tug1* function by genetic deletion *in vivo* would provide better insight into whether murine photoreceptor development went awry in the absence of this locus. Histological sectioning of *Tug1* WT and KO mice revealed no difference in the morphology of the retina (Fig. 15).

Figure 15:

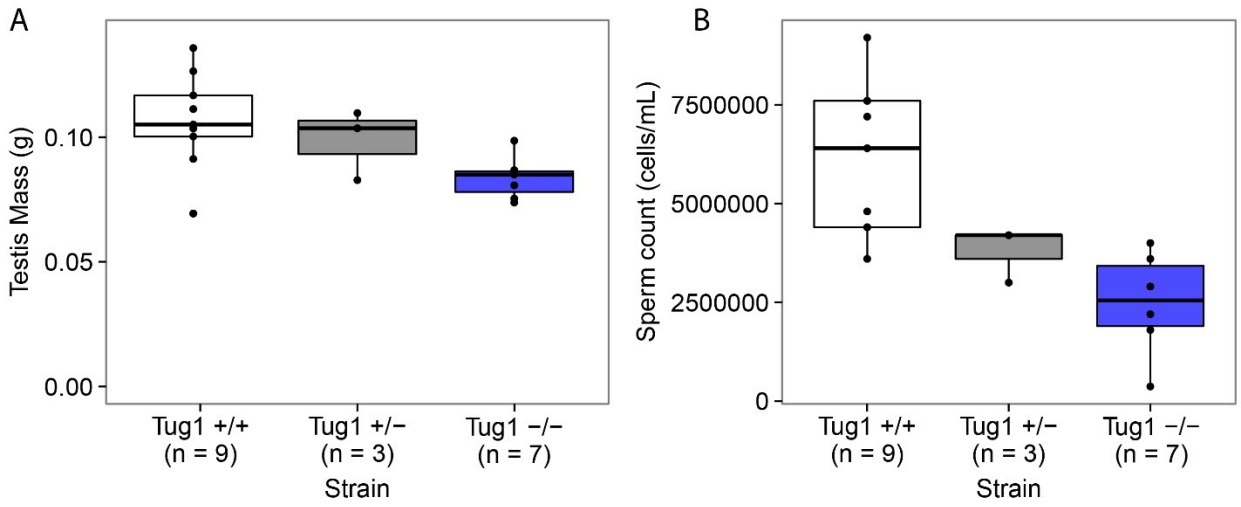


3.4 – Effects of *linc-Tug1* deletion on testes mass and sperm count

At this stage, we had used our lacZ reporter *Tug1* knockout mice to obtain a coarse view of the tissues in which *Tug1* is transcribed throughout development (E14.5 and Adult). To investigate the functional and mechanistic roles of Tug1 in male fertility, I set out to better understand the reasons why *Tug1*$^{-/-}$ males fail to reproduce. I started by isolating testes from Tug1 KO, HZ, and WT males. Collaborating with Emily Jacobs-Palmer from the Hoekstra lab, we measured testes mass, sperm count, and sperm morphology from each of these groups. Relative to WT mice, *Tug1*$^{-/-}$ males exhibit a 33%

reduction in testes mass (Fig. 16A, 0.08g, n=7, p=0.08)**.** Closer examination revealed a 66% reduction in average sperm count (Fig. 16B, 2 million cells/mL, n=7, p=0.008).

Figure 16:

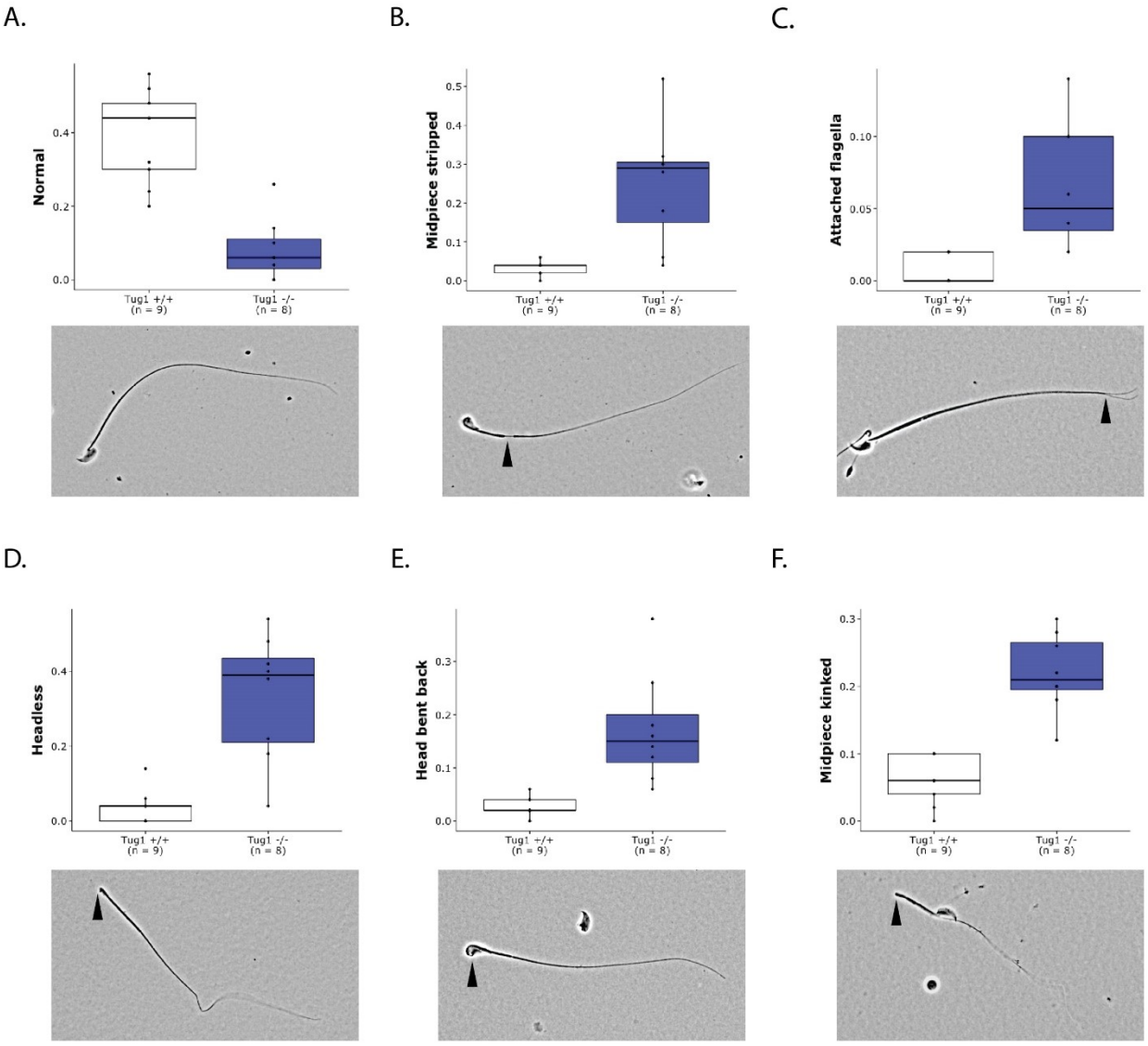


3.5 – *Tug1* knockout sperm exhibit highly penetrant morphological defects

We next collected sperm from one cauda epididymis by surgically bisecting it (opening it lengthwise) and suspending the collected tissue in Biggers-Whitten-Whittingham (BWW) sperm media. After obtaining sperm samples from n=9 *lincRNA-Tug1* male mice between 9-15 weeks of age (reproductive age in mice is 6-8 weeks), we observed that *Tug1*$^{-/-}$ males display normal physiology in around 10% of sperm, versus up to 90% in wild type counterparts (16C, n=8 mice, N=50 sperm/male, p=0.008). Closer examination revealed that the majority of observed sperm aberrations could be

characterized by head and midpiece defects, resulting in high proportions of *Tug1* KO sperm with stripped midpieces (Fig. 17B), attached flagellae (Fig. 17C), headless bodies (Fig. 17D), incorrectly positioned sperm heads (Fig. 17E), and/or broken midpiece sections (Fig. 17F). We also examined motility of collected sperm, monitoring samples after 15 minutes of incubation in BWW at 37°C. There was no observable difference in motility between sperm collected from *Tug1* WT and KO mice. These results are consistent with the observed infertility in *Tug1*$^{-/-}$ males, where the low proportion of normal sperm coupled with the reduction in overall sperm count leaves these males with very few viable gametes.

Figure 17:

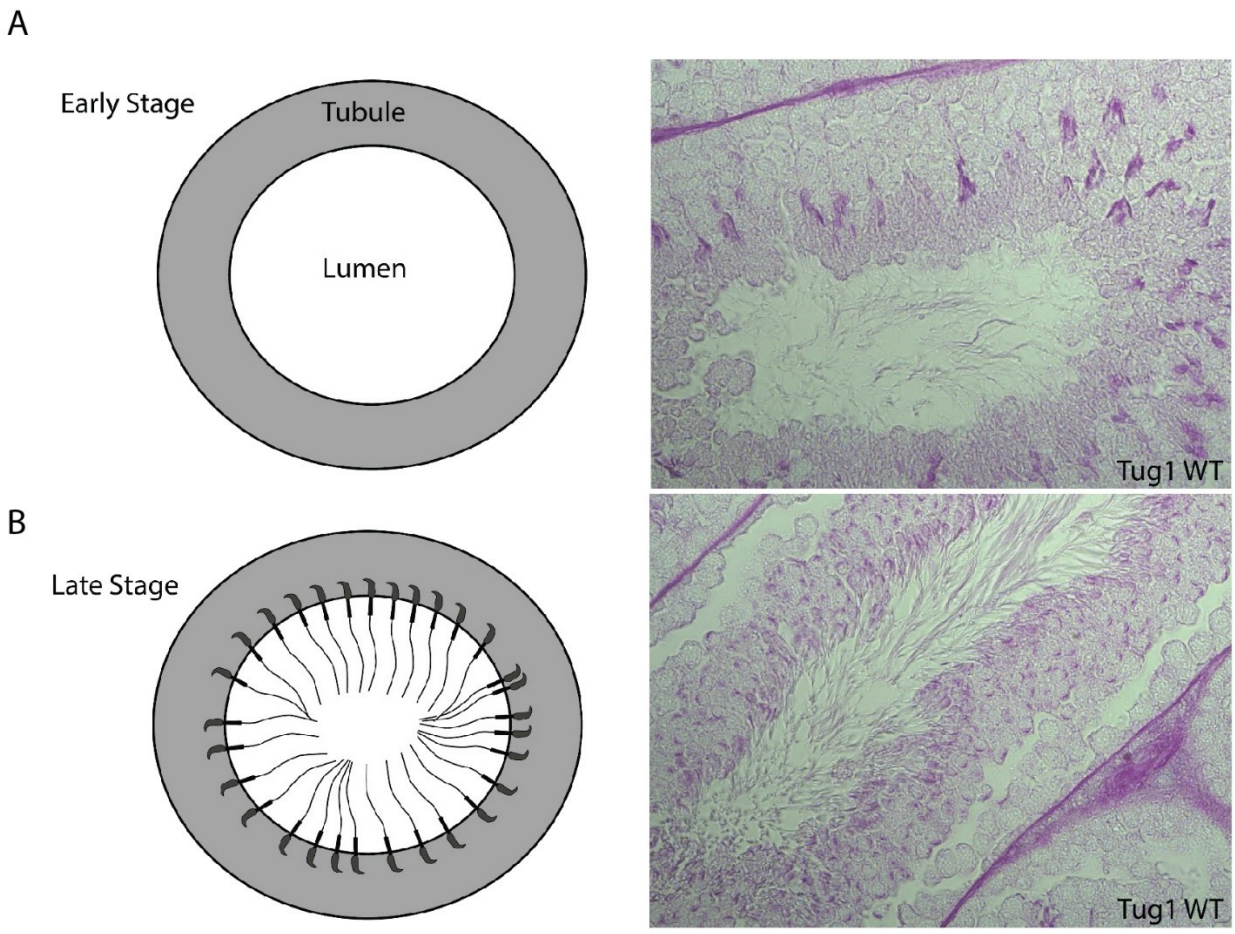


We did, however, observe a low number of physiologically normal sperm during our analysis, which suggests that deletion of *Tug1* might not result in total sterility but rather a strongly reduced fertility. Either way, these findings prompted us to further investigate the role of the *Tug1* locus in this developmental process.

3.6 – Expression of the *Tug1* locus *in vivo*

Tug1 deletion results in an apparent and significant effect on the development of mature and physiologically normal sperm. It therefore became evident that we needed a higher resolution understanding of when and where *Tug1* is expressed in the male gonad. To tackle this we sectioned testes from *Tug1* -/- and +/+ mice, and co-stained with X-gal (described above) and periodic acid-Schiff (PAS) stain[25] in order to identify not only the spatiotemporal dynamics of *Tug1* expression in the testes, but also how transcription at that locus corresponds to defects in male reproductive development. PAS stains the acrosome of developing sperm,[26] allowing for visualization of the discrete stages of spermatogenesis, and so we were able to screen testes sections for where and when *Tug1* turned on. Spermatogenesis is a complex process, in which spermatogonial stem cells form spermatozoa within the seminiferous tubule. Spermatogenesis can be broken into 16 discrete steps, with each subsequent stage moving away from the Sertoli cells at the tubule periphery and toward the interior lumen.[27] The end of spermatogenesis is characterized by dissociation of mature sperm into the lumen, where they subsequently migrate toward the epididymis. Figure 18, below, provides illustrations and histological cross-sections of early- (Fig. 18A) and late-stage (Fig. 18B) seminiferous tubules in a normal (WT) context. In order to create the sections shown below, and to identify discrete stages of spermatogenesis, we stained testes sections with Mayer's hematoxylin, periodic acid, and Schiff's Reagent.

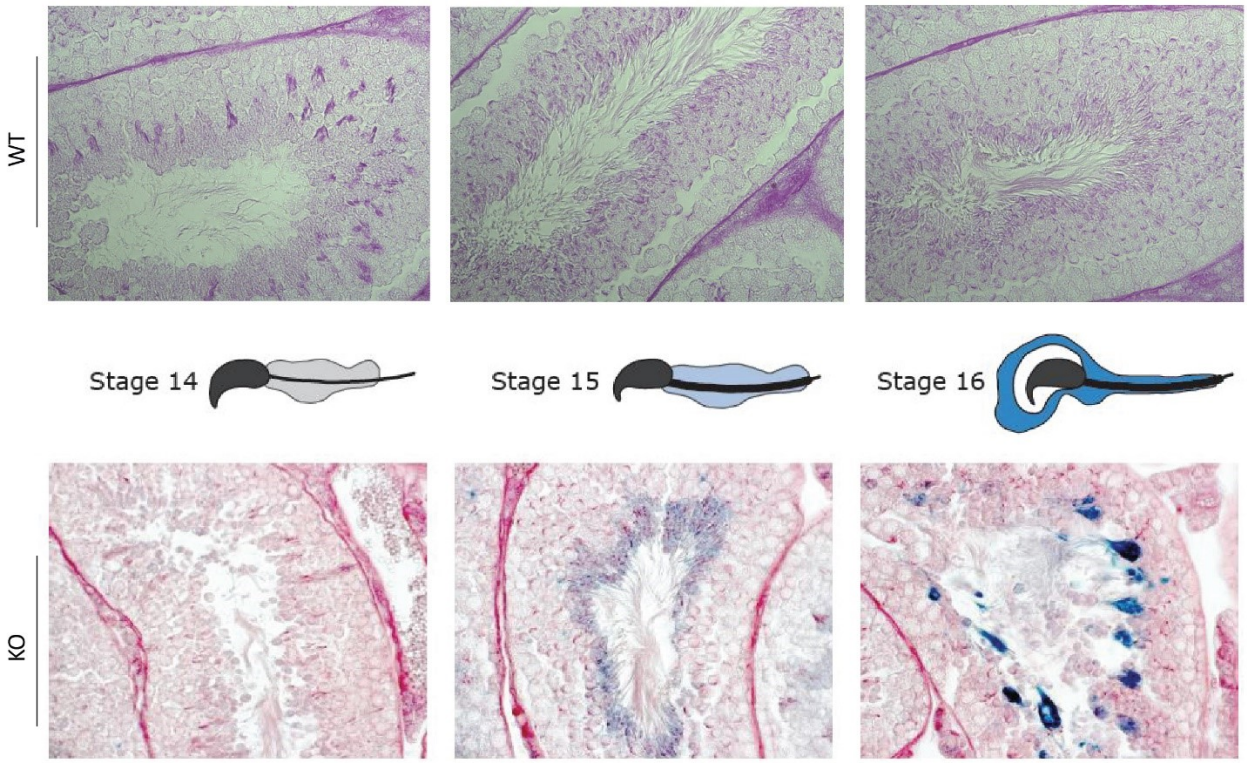
Figure 18:



We harvested testes from 3 WT and 3 KO male mice, and prepared them for sectioning using our previously established protocol (see 2.7.5: lacZ expression analysis and histology). Following fixation and lacZ staining, tissues are co-stained using PAS to aid in staging seminiferous tubules. Of the 16 stages of spermatogenesis, we observe *Tug1* expression via lacZ staining after the protrusion of flagellar tails into the lumen, indicating expression turns on between stages 14-16 of spermatid development (Fig.19). The temporally late expression pattern of the *Tug1* locus in this process indicates that this

lincRNA is required for the final maturation and dissociation of sperm at the end of the gametogenic cycle.

Figure 19:



It is important to note that the observed temporal regulation of *Tug1* expression suggests it turns on very late in spermatogenesis. At this stage of gamete development, transcription has undergone global silencing. Thus, expression of the *Tug1* locus at this stage of development means (1) the lacZ cassette is not translated immediately after transcription, (2) *Tug1* is transcribed in the Sertoli progenitor population of cells at the exterior of the seminiferous tubule, and diffuses through shared cytoplasm with late-stage spermatids exiting the spermatogenic cycle, or (3) the *Tug1* locus escapes transcriptional

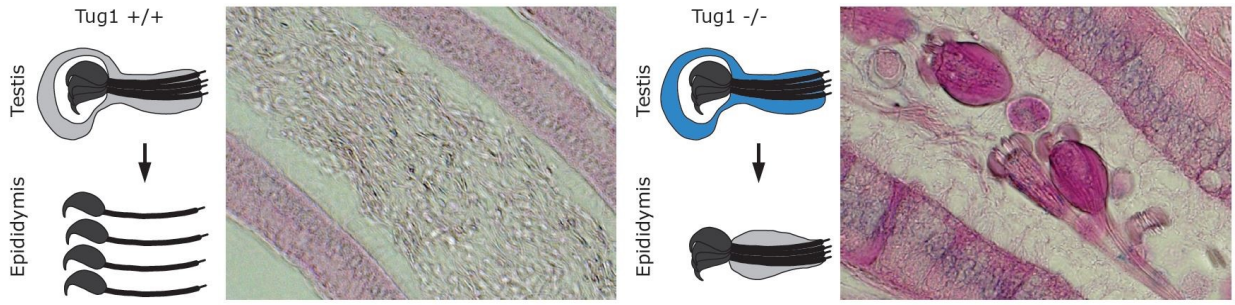
silencing. There is no evidence of delayed translation of the lacZ mRNA in any of our lincRNA knockout models, and if the *Tug1* transcript (in this case, lacZ) is produced in Sertoli cells at the outer edges of the seminiferous tubule then we might expect the entire tubule to stain positive for lacZ activity. Restriction of lacZ expression to the lumen periphery indicates a local site of translation. Thus, signs point to the *Tug1* locus as either being transcribed prior to silencing, or escaping silencing in the male reproductive organ altogether.

3.7 - Tug1 deletion results in a failure of spermatid to individualization

During the course of the histological experiments described above, we observed an interesting phenomenon. Sperm from WT males appeared as physiologically normal, free-swimming gametes in both testes and epididymis sections. By contrast, sperm from *Tug1*$^{-/-}$ males appeared as fused packets of several mature sperms. Figure 20 illustrates this contrast: the left illustration demonstrates how syncytial spermatids separate into individual, mature sperms, and the image on the left shows what this looks like in a WT epididymis. Comparatively, the right illustration indicates what we think is happening with *Tug1* KO spermatids as the develop. Instead of separating from their syncytial cytoplasmic bridge, these gametes remain fused together even as they otherwise mature. The histological section on the right demonstrates what this looks like in a KO epididymis. Interestingly, images such as these lead us to believe that there is potential for these fused packets to cause physical blockages within the epididymis prior to ejaculation, although this hypothesis has not been tested. This cellular phenotype was consistent between testes and epididymis sections when viewed under PAS/X-gal co-stain. Toward the end of

spermatogenesis, premature sperm (known as spermatids) seem to stall at cytokinesis and stay connected by an intercellular bridge. Spermatogenesis concludes with cytokinesis and cleavage of the syncytial bridge at the end of stage 16, in a process known as individualization. The intercellular bridge is considered to be the final visible sign of spermatogonial differentiation.[28, 29] At this point, the molecular mechanisms underlying individualization are largely unknown in mammals – the only published works that we could find (using both Web of Science and Google Scholar article searches) pertain to another animal model, the fruit fly *Drosophila melanogaster*. It seems, however, that *Tug1* plays an integral role in the ability of late-stage spermatids to cleave their syncytial bridges and become free-swimming, functionally mature male gamete.

Figure 20:



3.8 - Tug1 regulates the expression levels of cis genes 3' of its TSS

The observed phenotype is one of the most profound known examples of an *in vivo* phenotype associated with loss of function at a lncRNA locus. We have characterized the morphological defects associated with *Tug1* deletion from the organismal to cellular levels. In order to truly understand how this locus functions in the context of male gametogenesis, specifically in the process of mammalian individualization, we need to study the perturbations associated with *Tug1* loss of function in genetic and molecular contexts.

Tug1 deletion results in malformation of certain tissues during development, as evidenced by histological analysis in our preliminary data (see above). Differential regulation of genes and gene pathways could give rise to these pathological differences between *Tug1* knockout and wild type mice. For this reason, it is critical to identify whether genes are misregulated in these mice, specifically in the male gonad. RNA sequencing of the testes allowed us to identify 67 genes that are significantly different between *Tug1* -/- and +/- mice.[30, 31] When we looked at the distribution of these genes throughout the genome, we noticed that 7 are within a 1 megabase window of the *Tug1* transcription start site (Fig. 21, TSS, $p<0.0005$). Further investigation revealed that these genes all lie immediately downstream of the *Tug1* locus in the 3' direction. Perhaps even more interestingly, each of these genes is upregulated in the *Tug1* -/- samples.

Figure 21:

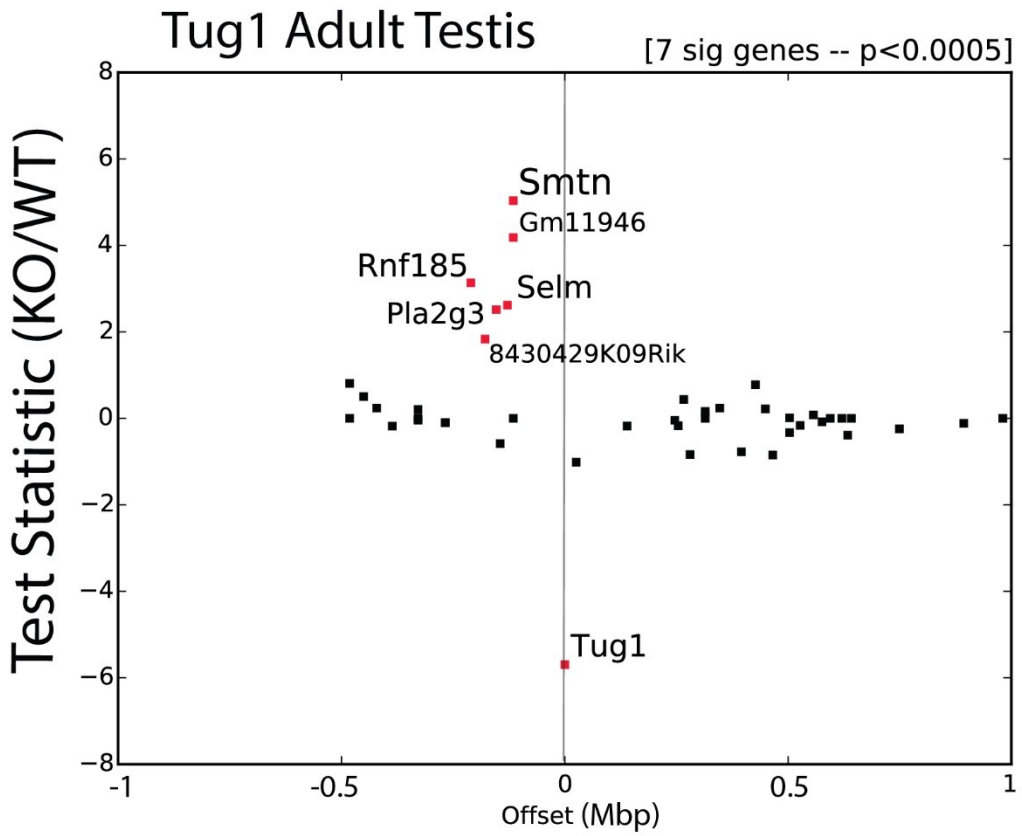
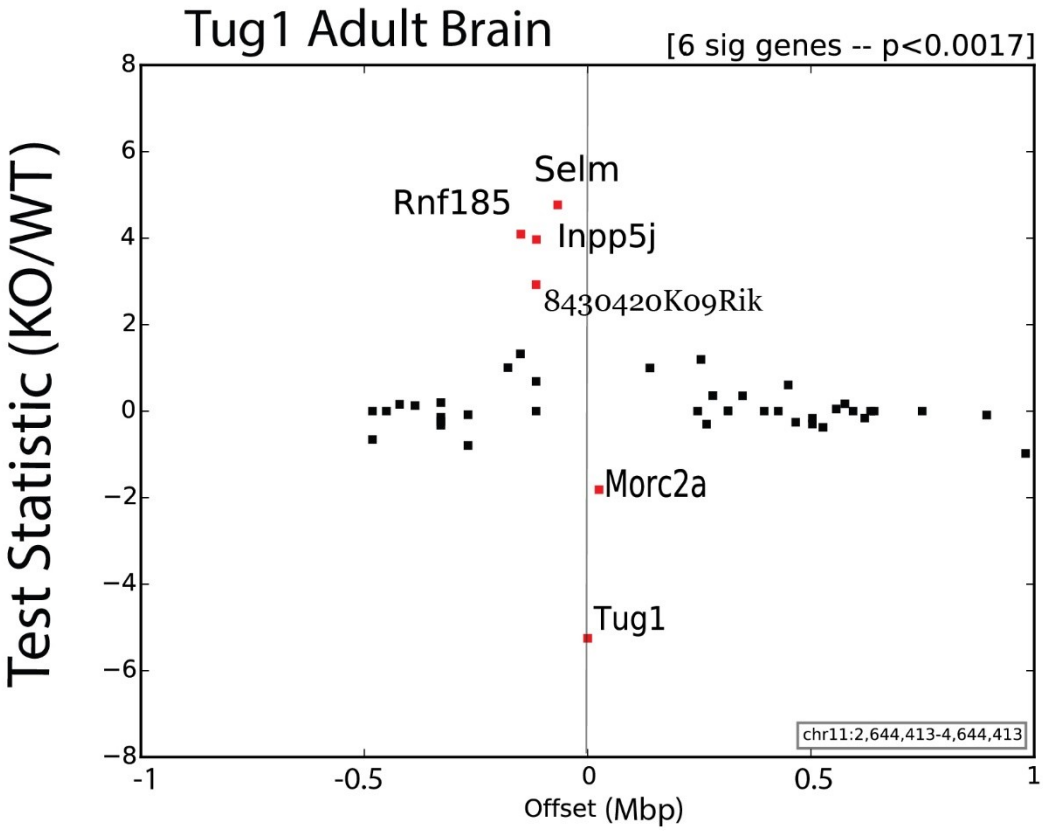


Figure 21: In these graphs, known as "cis plots", we investigate the spatial distribution of differentially expressed genes in primary genome space (linear sequence space). Black dots are those genes labeled as non-significantly differential, whereas red dots indicate those called significant. A dot's placement along the x-axis indicates the position of that gene's TSS relative to *Tug1* so, for example, the genes *Rnf185, Smtn,* and *Selm* all lie downstream (3') of the *linc-Tug1* transcription start site. Positioning along the y-axis correlates with a gene's relative expression level in KO samples relative to its expression in WT samples (with 0 denoting no change).

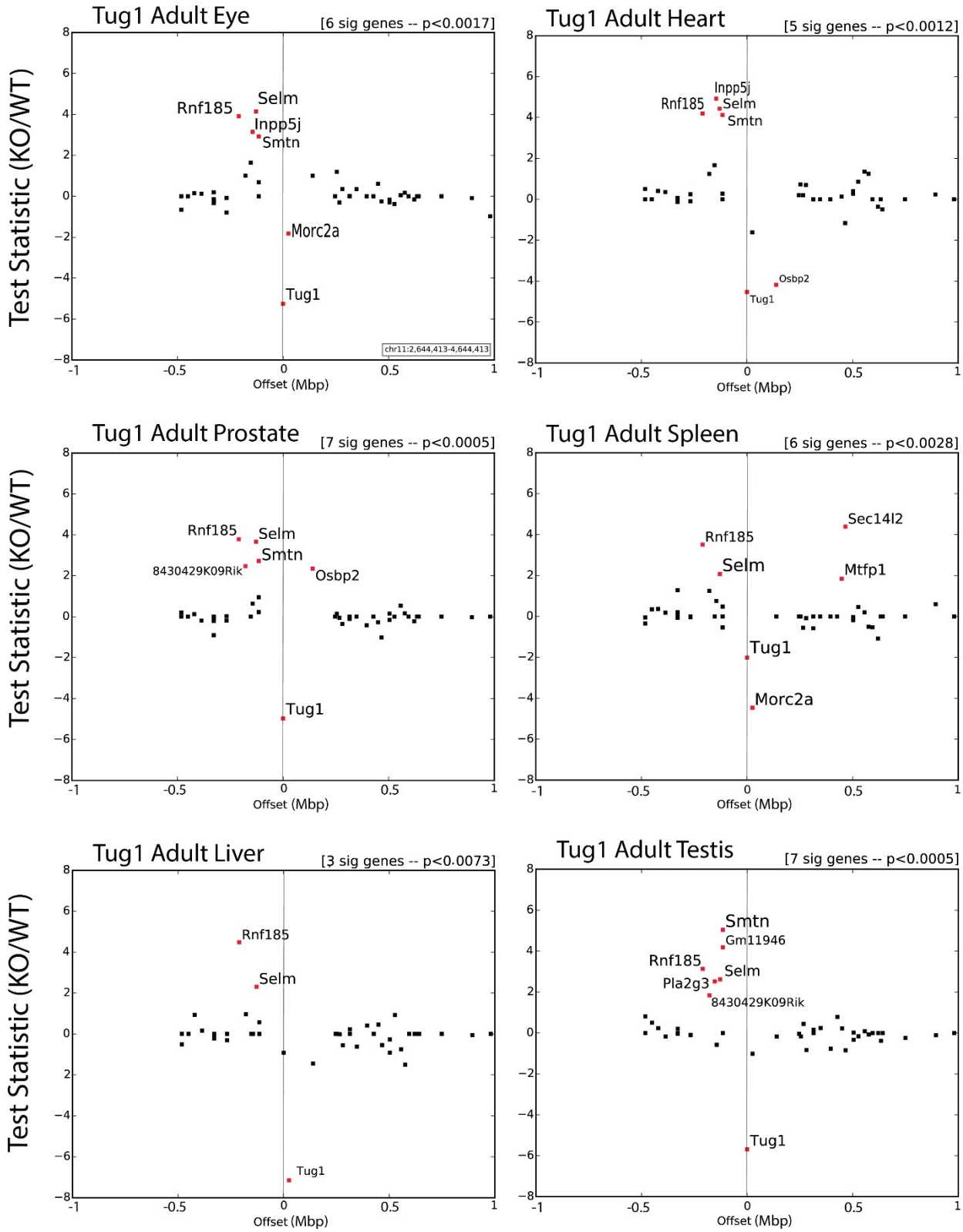
Around that same time, our lab published a study in which we surveyed lncRNA gene profiles in the brain,[8] described in greater detail in Chapter 2. In this study, we noticed the same pattern in the expression profile of *Tug1* -/- brains (Fig. 22).

Figure 22:



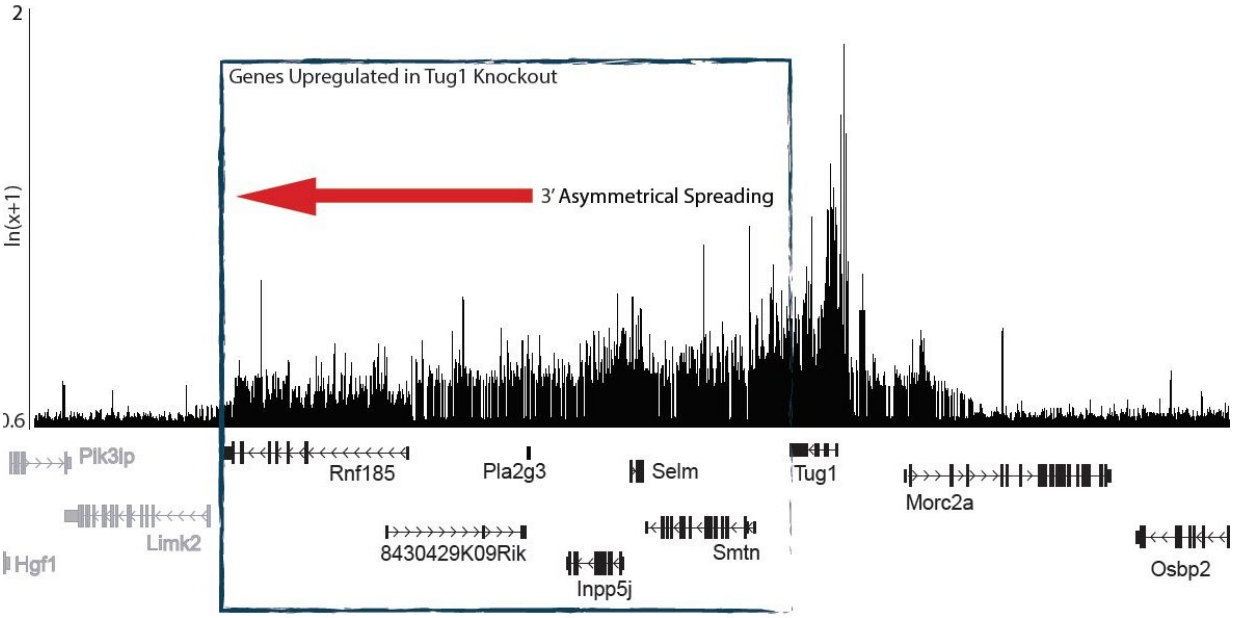
These results hint that *Tug1* might function as a *cis* regulatory locus across the body. *Tug1* is highly expressed in much of the mammalian body, and so we looked into the transcriptomes of a select panel of tissues to see if the observed gene profile was recapitulated in somatic contexts. RNA sequencing of these tissues illustrated similar dynamics in each (Fig. 23).

Figure 23:


Tug1 Adult Eye
[6 sig genes -- p<0.0017]
Rnf185
Selm
Inpp5j
Smtn
Morc2a
Tug1
chr11:2,644,413-4,644,413
Tug1 Adult Heart
[5 sig genes -- p<0.0012]
Inpp5j
Rnf185
Selm
Smtn
Osbp2
Tug1
Tug1 Adult Prostate
[7 sig genes -- p<0.0005]
Rnf185
Selm
Smtn
8430429K09Rik
Osbp2
Tug1
Tug1 Adult Spleen
[6 sig genes -- p<0.0028]
Sec14l2
Rnf185
Selm
Mtfp1
Tug1
Morc2a
Tug1 Adult Liver
[3 sig genes -- p<0.0073]
Rnf185
Selm
Tug1
Tug1 Adult Testis
[7 sig genes -- p<0.0005]
Smtn
Gm11946
Rnf185
Selm
Pla2g3
8430429K09Rik
Tug1
Test Statistic (KO/WT)
Offset (Mbp)

An alternative approach to identify direct genomic targets and protein partners of a lncRNA is to pulldown the RNA (*Tug1)* in a crosslinked cell and see what comes along for the ride. In a parallel set of experiments, we did this using a technique known as RNA antisense purification (RAP), in which biotinylated probes complementary to *Tug1* transcripts are used to pull down the lncRNA and any DNA fragments it associates with.[32] RAP analysis in mouse embryonic stem cells (MESCs) and mouse embryonic fibroblasts (MEFs) shows the same pattern, in which *Tug1* transcript spreads in a unidirectional fashion from its locus and associates with DNA downstream of the TSS (Fig. 24). Reads from biotinylated *Tug1* immunoprecipitation were mapped back to murine reference genome mm19, and were found to overlap the exon-spanning sequences of protein coding genes *Rnf185, Pla2g3, Selm, Smtn, Inpp5j,* as well as one un-annotated gene *8430428K09Rik*. Gene ontology (GO) analysis on these genes did not reveal common pathways that could be implicated in sperm defects. However, from a genomics standpoint, these analyses collectively indicate that the *Tug1* locus functions as a *cis* repressor of gene expression across the body. In conjunction with our previous work, we present a model by which *lincRNA-Tug1* complexes with PRC2 to regulate the expression of these genes through *cis* repression. How the gene profile of *Tug1* KO mice leads to male-specific infertility, however, is still a question that remains to be addressed in future studies.

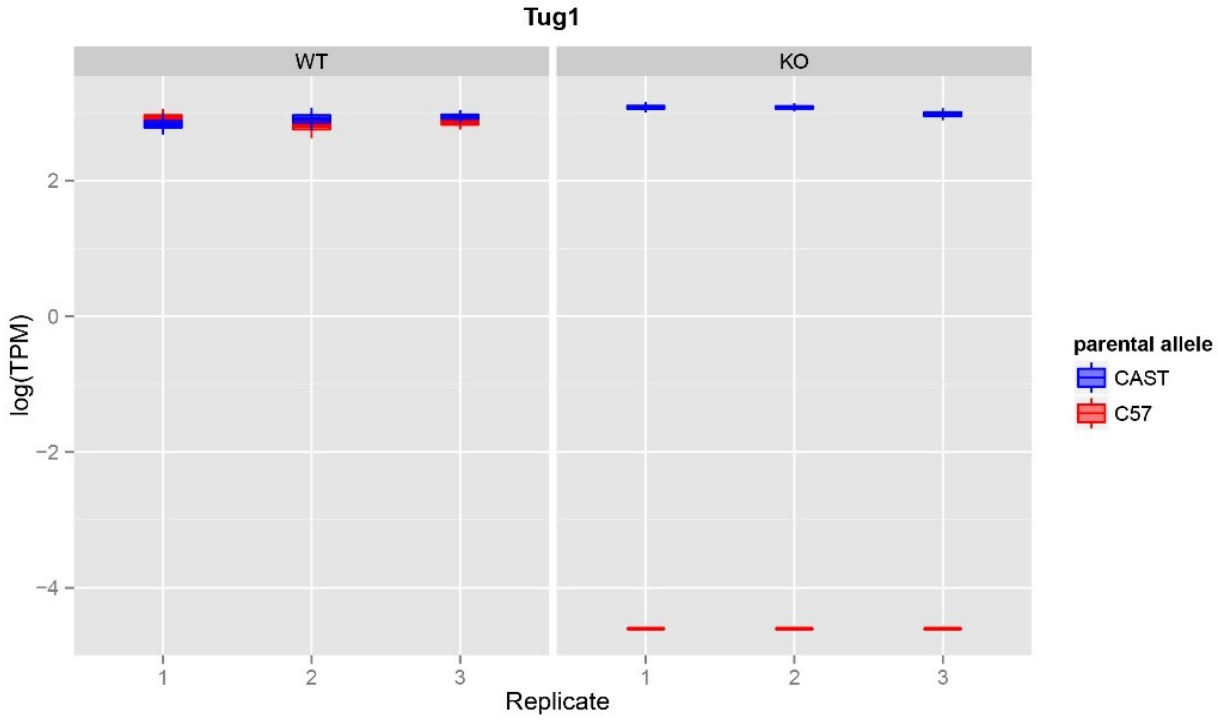
Figure 24:



3.9 - Tug1 regulates genes in an allele-specific manner

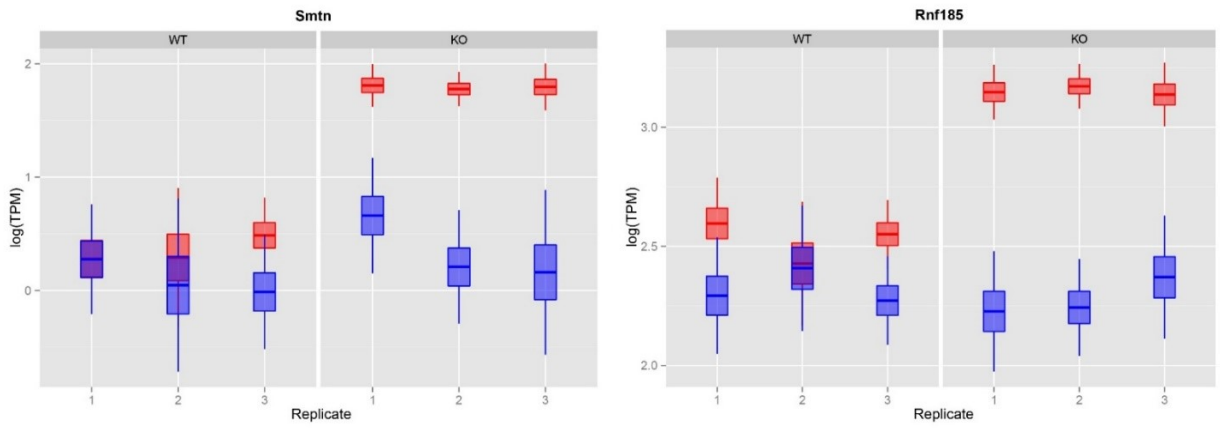
The true definition of *cis*-regulation is the ability to regulate not only in spatial proximity, but also in an allele-specific manner.[33] In order to address this particular detail, we crossed our C57BL/6 *Tug1* +/- mice with another sub-species of *M. musculus*: *M. m. Castaneus* (CAST). These mice breed and form viable, fertile litters with C57 strains, but possess alleles that are readily distinguishable by qPCR or sequencing techniques. We crossed C57BL/6 *Tug1* +/- mice with CAST *Tug1* +/+ mice to obtain heterozygous pups with the loss of function *Tug1* allele originating from the C57 background and the wild type *Tug1* allele contributed by the CAST background (Fig. 25).

Figure 25:



We collected testes from these mice and, in collaboration with Nimrod Rubinstein from the Dulac lab, performed allele-specific transcriptome analysis to determine whether *Tug1* deletion causes perturbations to both the C57 and CAST alleles, or is restricted to the C57 (for a full description of the approach, please see Methods below). We observed differential expression of two genes within the 1MB window 3' of the *Tug1* TSS, *Rnf185* and *Smtn*. Each of these genes was upregulated on just the C57 allele, indicating that *Tug1* is seemingly acting in a truly *cis* fashion (Fig. 26).

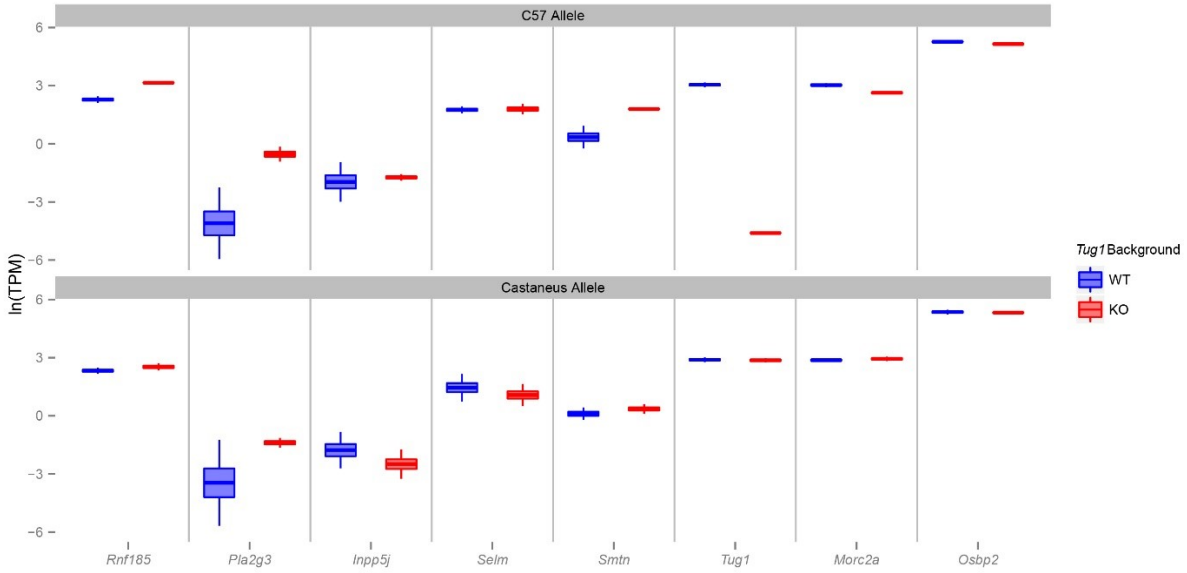
Figure 26:



When you pan out, however, and look at the same genomic window in which the cis plots and RAP analyses were done, the picture becomes a bit muddled, primarily due to the complexity in conveying this experiment in a single graph. Figure 27 attempts to accomplish that, and requires some explanation for the reader to properly assess what is going on. The graph is split horizontally into two rows: the top row, "C57 Allele", represents the allele of each gene that is derived from the C57 parent. Conversely, the bottom row, "Castaneus Allele", represents the allele of each gene that is derived from the CAST parent. Keep in mind that *Tug1* deletion in HET mice would only occur on the C57 allele, and that the CAST allele will always contain an endogenous WT copy of the *Tug1* locus. Therefore, the blue color (*Tug1* background = WT) is essentially labeling the WT copy of *Tug1* while the red color (*Tug1 background* = KO) labels the KO allele carrying lacZ instead of the lincRNA gene. With these parameters in mind, one may visualize the effects of *Tug1* deletion on a neighboring gene by comparing the top row to the bottom row. For example in the bottom row, which represents the Castaneus allele, *Rnf185* shows a similar level of expression between the WT and KO *Tug1* C57 alleles. When you look at

the top row, which represents the C57 allele, you see a different pattern in that the red bar is significantly higher than the blue bar, indicating the *Tug1* KO allele induced an upregulation of the C57-derived *Rnf185* allele relative to the effects that the *Tug1* WT allele had. It is not statistically correct to do a differential on the top row vs. the bottom row, as these values are already differentials themselves. Regardless, the patterns elucidated from this experiment indicate that *Tug1* behaves in a truly cis-regulatory manner as only alleles on the C57 KO background were affected by deletion of this noncoding locus, whereas alleles on the CAST background were never affected.

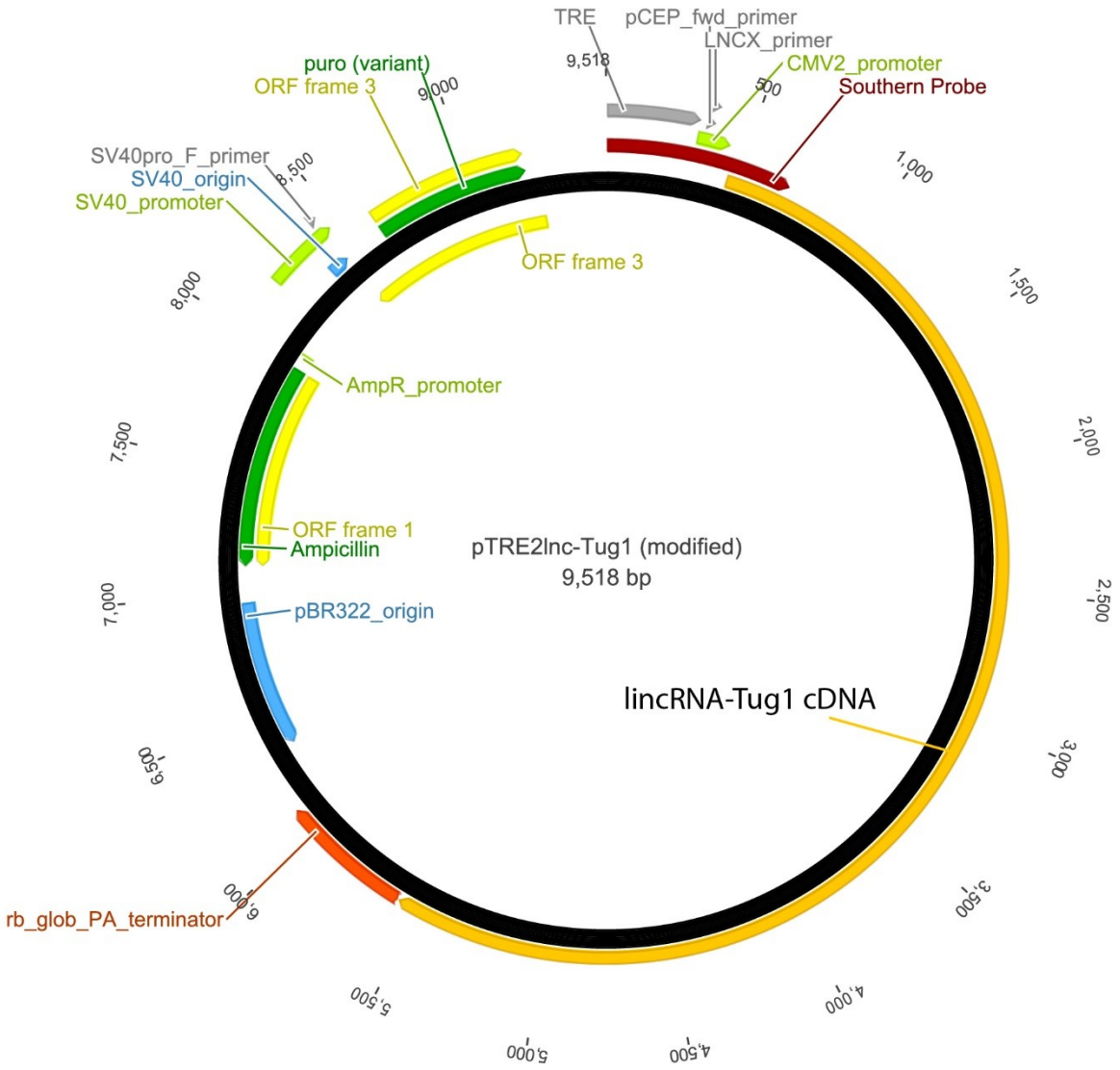
Figure 27:



3.10 –*Tug1* transient overexpression does not rescue genetic phenotype

One question that consistently arises when studying ncRNA loci is: "how do we know it's an RNA?" Our KO strain described above, which removes the DNA sequence of the endogenous *linc-Tug1* gene, maintains the possibility that the DNA element of this locus is responsible for the observed morphological and genetic aberrations, and so parsing the contributions of each remains a challenge. Our model does, however, rule out the possibility that the act of transcription is what's important at this locus, as the inserted lacZ cassette is still transcribed in our mice. Whether the observed gene profile changes and mouse phenotype are tied to the DNA at the *Tug1* locus, or the transcript originating therein (i.e. the *Tug1* lincRNA), remains to be seen. To address this, we recently cloned a transcript originating from the *Tug1* locus and performed a transient overexpression into *Tug1* WT mouse embryonic stem cells (mESCs), as well as both *Tug1* WT and KO mouse embryonic fibroblasts (MEFs). In this inducible system, the *linc-Tug1* cDNA was cloned into a pTRE2 vector containing a tetracycline-controlled transcription activator (Tet-on). In the presence of doxycycline, a commonly used antibiotic, the Tet-on promoter will induce expression of *lincRNA-Tug1*. A complete vector map of our overexpression construct can be found in Fig. 28. Transfection efficiencies were determined through use of a similar construct carrying GFP (data not shown).

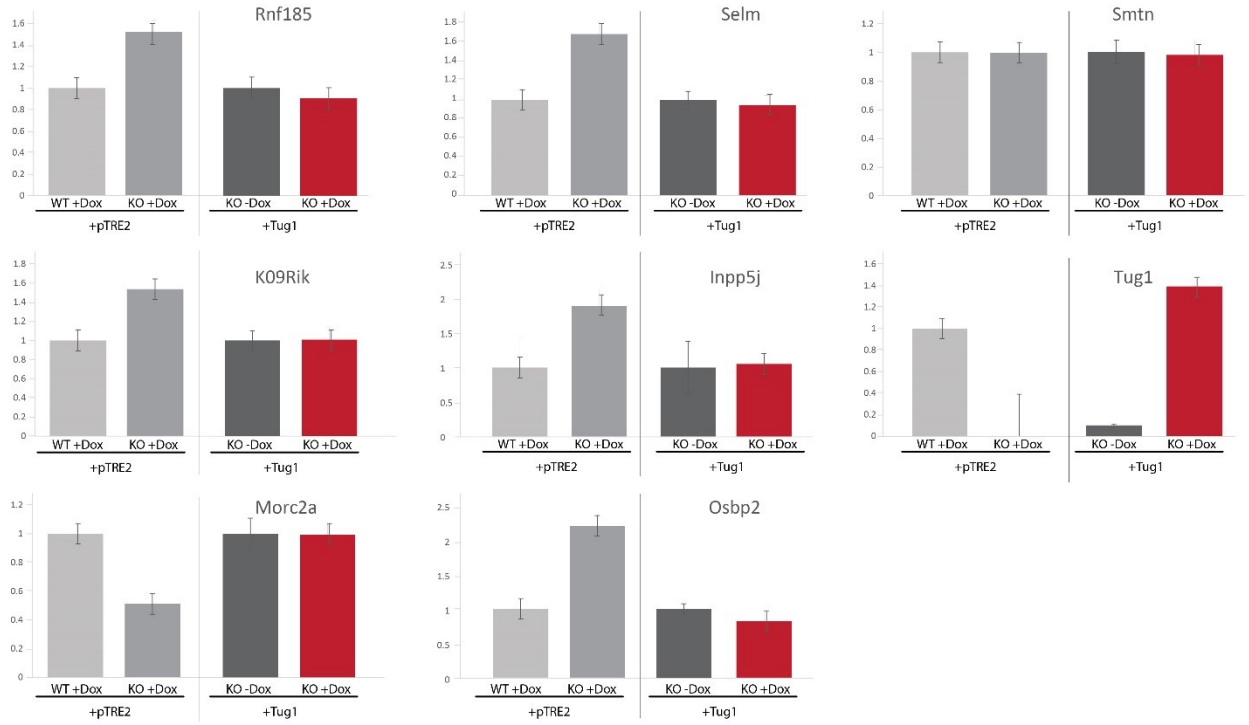
Figure 28:



Utilizing this expression construct, we were able to induce *lincRNA-Tug1* from a non-endogenous location (transient). Quantitative PCR (qPCR) analysis was performed on cells derived from 6 WT and 6 KO mice, with each mouse representing one biological replicate. Each replicate was split into + and – doxycycline groups, and we tested both pTRE2-Tug1 as well as pTRE2-EV (empty vector, just listed as pTRE2 from now on). The

results, as shown in Fig. 30, illustrate that inducing expression of the *Tug1* transcript from a non-endogenous location does not significantly affect expression levels of the genes within its 3' cis window. All other experiments have pointed to the role of *lincRNA-Tug1* as a repressor of gene function, and so we would expect overexpression of the RNA to downregulate genes relative to the empty vector (pTRE2) control. We do not observe such a phenomenon, and some possibilities for why this might be include: (1) the *Tug1* RNA must be spatially adjacent to the genes it regulates, or (2) the *Tug1* DNA is the functional component and must be spatially adjacent to the genes it regulates.

Figure 29:



3.11 – Conclusions and Future Directions

While the number of expressed lincRNA genes might exceed that of protein coding genes, how they function remains a mystery. More than a thousand of these noncoding loci are expressed throughout the stages of spermatogenesis in mice and rats, and some have been identified as differentially expressed in recent studies of human reproductive defects. One lincRNA known as *Neat1* was found to function in female fertility, and we discuss it in greater detail in Chapter 5. All told, there is increasing evidence for the critical roles lincRNAs play in reproductive development, and warrant further studies in this organ system.

We have made great strides in (i) characterizing the morphological phenotype associated with *Tug1* deletion, (ii) identifying the gene profile changes and regulatory role that Tug1 plays, and (iii) uncovering the mechanism by which *Tug1* achieves this regulation in mice. Overexpression of *Tug1* in MESC and MEF cell lines via transient transfection has demonstrated that a *cis* mechanism is required to induce the genetic perturbations observed by RNA sequencing of *Tug1* KO tissues. It remains to be seen whether it is the DNA or the RNA that serves as the functional unit in this circuit. Future experiments include integration and viral overexpression of *lincRNA-Tug1,* as well as introducing the polyadenylation signal to stop transcription of the endogenous *Tug1* gene (described above). Both of these efforts are currently ongoing at the time of writing this thesis. Additional experiments that would prove insightful include *in vitro* fertilization (IVF) of a mouse ovum (egg cell) using sperm derived from either *Tug1* WT or KO sperm. If KO-derived sperm can fertilize at comparable rates to the WT, then this experiment

demonstrates that only the process of individualization has been affected as a result of *Tug1* deletion, and that packaging of genetic material for reproduction has not also gone awry. Disproving this hypothesis (if KO sperm do not create viable embryos following IVF) would open the door for further genetic studies to identify problems arising during gametogenesis with greater resolution than we have currently achieved. Finally, we are in the process of developing an *in vivo* inducible *lincRNA-Tug1* knockout mouse model. This model is the F1 progeny of a tamoxifen-responsive Cre-recombinase (CreER^{TM})[42] mated with a C57 mouse in which the *linc-Tug1* locus has been flanked by loxP sites. The resulting CreER^{TM}:loxP-Tug1-loxP mouse will exhibit "WT" *Tug1* function until administration of tamoxifen, at which point the *Tug1* DNA locus will be excised. Comparison of CreER^{TM}:loxP-Tug1-loxP +TM and – TM will demonstrate *in vivo* the necessity and sufficiency of this genomic locus to effect the spermatogenic phenotype of mice.

3.12 – Author Contributions

Stephen Liapis (SL) and Emily Jacobs-Palmer (EJP) contributed equally to this work. SL and Chiara Gerhardinger (CG) noticed *Tug1* male infertility phenotype, and CG process *Tug1* retinas. SL and EJP designed and implemented research, working together to collect tissue and sperm samples, conduct staining and histology, and quantify defects of reproductive systems. SL performed all lacZ staining. SL and EJP staged seminiferous tubules, and EJP quantified sperm physiological defects. SL created designs for tubule staging and individualization in Adobe Illustrator. EJP managed the colony of *Castaneus* founders, as well as C57/CAST hybrid progeny. Hybrid genome analysis was done in

collaboration with Nimrod Rubinstein. RNA sequencing and cis plot analysis were performed on testes RNA sequencing by Abigail Groff (AG), and subsequent analysis of of other tissue types was done by SL. RAP analysis was performed by AG and SL. Martin Sauvageau (MS) assisted with development of transient inducible *Tug1* construct. Overexpression experiments and qPCR analysis were performed by SL and MS. Overall design of project objectives, experiments, and goals was done by SL, EJP, with the mentorship of principle investigators John Rinn and Hopi Hoekstra. This work has (at the time of writing this thesis) not been published in a peer reviewed journal, but the collective group of authors fully anticipates this work to result in a journal article in the near future. Writing of this chapter was done entirely by SL, with suggestions for revision made by John Rinn and Hopi Hoekstra.

3.13 - Methods and Materials

Castaneus Hybrid Genome Analysis:

Each RNA-seq library was first subjected to quality and adapter trimming using the Trim Galore utility (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore) with stringency level 3. Subsequently, STAR RNA-seq aligner[35] was used to map reads from each RNA-seq library to a C57Bl/6J, Cast/EiJ diploid genome, created by incorporating C57Bl/6J and Cast/EiJ single nucleotide polymorphisms and indels (obtained from obtained from the Mouse Genome Project)[36] into the *Mus musculus* GRCm38 reference genome sequence using the AlleleSeq package.[37] The Gencode M2 mouse gene annotation[37, 40] was lifted over (using the UCSC liftOver utility) according to the coordinates of the C57Bl/6J and Cast/EiJ genomes to create a C57Bl/6J, Cast/EiJ diploid

gene annotation. Subsequently, MMSEQ[41] was used to estimate expression levels of each transcript in the diploid gene annotation. MMSEQ's Fragment Per Kilobase per Million (FPKM) units were then converted to Transcripts per Million (TPM) and any transcript which TPM was lower than 0.01 was set to 0. Next, expression levels of transcripts which cannot be distinguished from each other according to the read data were combined following Perez et al. (2015).[38] Finally, allelic biased expression across all samples was estimated for each transcript according to Perez et al. (2015), where a posterior probability cutoff of 0.95 was used for calling an allelic bias significant.

Examination of *Tug1* testes, epididymes, and sperm

Tug1 knockout, heterozygote, and wild type males were chosen between 60 and 400 days of age for reproductive analysis. The entire male reproductive tract (testes and epididymes) was dissected out and immediately placed into phosphate buffered saline (PBS). One testis per mouse was used for weight measurements, while the other was sent for sectioning at the Harvard Bauer Laboratory Histology Core. Sperm samples were collected via bisection of the cauda epididymis, and subsequent suspension in Biggres-Whitten-Whittingham (BWW) sperm media at an incubation temperature of 37°C. Sperm morphology was examined by fixation in 2% paraformaldehyde, suspension in Fluoromount media (Southern Biotech), and mounting onto microscopy slides. Recordings for cell morphology counted abnormal and normal sperm. Abnormal sperm were categorized by commonality of defect. Statistical analysis of *Tug1* reproductive defects was performed using Kruskal-Wallis one-way analysis of variance (ANOVA).

RAP Analysis

RAP was performed as described. The *Tug1* mature transcript sequence was tiled with 120 bp antisense nucleotides that have been biotinylated. Two distinct pools of antisense probes, one targeting Tug1 and the other containing sense probes as a negative control, were generated using IDT oligonucleotide synthesis. The hybridization was performed in triplicate crosslinked whole testes lysates with 20 ng of oligos. The oligos were subsequently captured by binding to streptavidin beads and elutions for RNA and DNA were collected. Consistent with standard ChIP-Seq assays, duplicate pull-downs were performed and sequenced to control for technical variability.

3.14 - References

1. Guttman, M., and Rinn, J.L. (2012). Modular regulatory principles of large non-coding RNAs. Nature 482, 339–346. doi: 10.1038/nature10887.

2. Mercer, T.R., and Mattick, J.S. (2013). Structure and function of long noncoding RNAs in epigenetic regulation. Nat Struct Mol Biol 20, 300–307. doi: 10.1038/nsmb.2480.

3. Rinn, J.L., and Chang, H.Y. (2012). Genome regulation by long noncoding RNAs. Annu. Rev. Biochem. 81, 145–166. doi: 10.1146/annurev-biochem-051410-092902

4. 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. Nature Structural & Molecular Biology *21*, 198-+.

5. Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., . . . Chang, H. Y. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by Noncoding RNAs. *Cell, 129*(7), 1311-1323. doi:10.1016/j.cell.2007.05.022

6. Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A., et al. (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature 472, 120–124. doi: 10.1038/nature09819.

7. Sauvageau, M., Goff, L.A., Lodato, S., Bonev, B., Groff, A.F., Gerhardinger, C., Sanchez-Gomez, D.B., Hacisuleyman, E., Li, E., Spence, M., Liapis, S.C., et al. (2013). Multiple knockout mouse models reveal lincRNAs are required for life and brain development. Elife 2.

8. L. A. Goff *et al.*, Spatiotemporal expression and transcriptional perturbations by long noncoding RNAs in the mouse brain. *Proceedings of the National Academy of Sciences* **112**, 6855-6862 (2015)

9. Young, T. L., Matsuda, T., & Cepko, C. L. (2005). The noncoding RNA Taurine upregulated gene 1 is required for differentiation of the murine retina. *Current Biology, 15*(6), 501-512. doi:10.1016/j.cub.2005.02.027

10. Imaki, H., Moretz, R., Wisniewski, H., Neuringer, M., & Sturman, J. (1987). RETINAL DEGENERATION IN 3-MONTH-OLD RHESUS-MONKEY INFANTS FED A TAURINE-FREE HUMAN INFANT FORMULA. *Journal of Neuroscience Research, 18*(4), 602-614. doi:10.1002/jnr.490180414

11. Sturman, J. A., Gargano, A. D., Messing, J. M., & Imaki, H. (1986). FELINE MATERNAL TAURINE DEFICIENCY - EFFECT ON MOTHER AND OFFSPRING. *Journal of Nutrition, 116*(4), 655-674.

12. Altshuler, D., Loturco, J. J., Rush, J., & Cepko, C. (1993). TAURINE PROMOTES THE DIFFERENTIATION OF A VERTEBRATE RETINAL CELL-TYPE IN-VITRO. *Development, 119*(4), 1317-1328.
13. Young, T. L., & Cepko, C. L. (2004). A role for ligand-gated ion channels in rod photoreceptor development. *Neuron, 41*(6), 867-879. doi:10.1016/s0896-6273(04)00141-2

14. 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. doi: 10.1038/nature07672.

15. 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 & Development *25*, 1915-1927. 10.1101/gad.17446611.

16. 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. doi: 10.1073/pnas.0904715106.

17. Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes & Development, 16*(1), 6-21. doi:10.1101/gad.947102

18. Jaenisch, R., & Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics, 33*, 245-254. doi:10.1038/ng1089

19. Ardlie, K. G., DeLuca, D. S., Segre, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., . . . Consortium, G. T. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science, 348*(6235), 648-660. doi:10.1126/science.1262110

20. Watson, C.M., Trainor, P.A., Radziewic, T., Pelka, G.J., Zhou, S.X., Parameswaran, M., Quinlan, G.A., Gordon, M., Sturm, K., and Tam, P.P.L. (2008). Application of lacZ transgenic mice to cell lineage studies. Methods Mol. Biol. 461, 149–164. doi: 10.1007/978-1-60327-483-8_10.

21. Eisenberg, E., & Levanon, E. Y. (2013). Human housekeeping genes, revisited. *Trends in Genetics, 29*(10), 569-574. doi:10.1016/j.tig.2013.05.010

22. Lodish H, Berk A, Zipursky SL, et al. Molecular Cell Biology. 4th edition. New York: W. H. Freeman; 2000. Section 8.2, Isolation and Analysis of Mutants.Available from: http://www.ncbi.nlm.nih.gov/books/NBK21735/

23. Ulitsky, I., & Bartel, D. P. (2013). lincRNAs: Genomics, Evolution, and Mechanisms. *Cell, 154*(1), 26-46. doi:10.1016/j.cell.2013.06.020

24. Dykxhoorn, D. M., Novina, C. D., & Sharp, P. A. (2003). Killing the messenger: Short RNAs that silence gene expression. *Nature Reviews Molecular Cell Biology, 4*(6), 457-467. doi:10.1038/nrm1129

25. Mowry, R. W. (1963). SPECIAL VALUE OF METHODS THAT COLOR BOTH ACIDIC AND VICINAL HYDROXL GROUPS IN HISTOCHEMICAL STUDY OF MUCINS - WITH REVISED DIRECTIONS FOR COLLOIDAL IRON STAIN, USE OF ALCIAN BLUE G8X AND THEIR COMBINATIONS WITH PERIODIC ACID-SCHIFF REACTION. *Annals of the New York Academy of Sciences, 106*(2), 402-&.

26. Tanii, I., Yoshinaga, K., & Toshimori, K. (1999). Morphogenesis of the acrosome during the final steps of rat spermiogenesis with special reference to tubulobulbar complexes. *Anatomical Record, 256*(2), 195-201.

27. Phillips, B. T., Gassei, K., & Orwig, K. E. (2010). Spermatogonial stem cell regulation and spermatogenesis.*Philosophical Transactions of the Royal Society B-Biological Sciences, 365*(1546), 1663-1678. doi:10.1098/rstb.2010.0026

28. Fabrizio, J. J., Hime, G., Lemmon, S. K., & Bazinet, C. (1998). Genetic dissection of sperm individualization in Drosophila melanogaster. *Development (Cambridge), 125*(10), 1833-1843.

29. Tokuyasu, K. T., Hardy, R. W., & Peacock, W. J. (1972). DYNAMICS OF SPERMIOGENESIS IN DROSOPHILA-MELANOGASTER .1. INDIVIDUALIZATION PROCESS. *Zeitschrift Fur Zellforschung Und Mikroskopische Anatomie, 124*(4), 479-&. doi:10.1007/bf00335253

30. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., . . . Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols, 7*(3), 562-578. doi:10.1038/nprot.2012.016

31. Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology, 31*(1), 46-+. doi:10.1038/nbt.2450

32. Engreitz, J., Lander, E. S., & Guttman, M. (2015). RNA Antisense Purification (RAP) for Mapping RNA Interactions with Chromatin. *Nuclear Bodies and Noncoding Rnas: Methods and Protocols, 1262*, 183-197. doi:10.1007/978-1-4939-2253-6_11

33. Wittkopp, P. J., & Kalay, G. (2012). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics, 13*(1), 59-69. doi:10.1038/nrg3095

34. Ørom, U.A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., et al. (2010). Long noncoding RNAs with enhancer-like function in human cells. Cell 143, 46–58 doi: 10.1016/j.cell.2010.09.001.

35. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21.

36. Mouse Genome Project: ftp://ftp-mouse.sanger.ac.uk/REL-1303-SNPs_Indels-GRCm38

37. Engström, P.G., Steijger, T., Sipos, B., Grant, G.R., Kahles, A., Rätsch, G., Goldman, N., Hubbard, T.J., Harrow, J., Guigó, R., et al. (2013). Systematic evaluation of spliced alignment programs for RNA-seq data. Nat. Methods 10, 1185–1191.

38. Perez JD, Rubinstein ND, Fernandez DE, Santoro SW, Needleman LA, et al. 2015. Quantitative and functional interrogation of parent-of-origin allelic expression biases in the brain. eLife 4:e07860

39. Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., Leng, J., Bjornson, R., Kong, Y., Kitabayashi, N., et al. (2011). AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol. Syst. Biol. 7, 522.

40. Steijger, T., Abril, J.F., Engström, P.G., Kokocinski, F., Akerman, M., Alioto, T., Ambrosini, G., Antonarakis, S.E., Behr, J., and Bertone, P. (2013). Assessment of transcript reconstruction methods for RNA-seq. Nat. Methods 10, 1177–1184.

41. Turro, E., Su, S.-Y., Gonçalves, Â., Coin, L.J., Richardson, S., and Lewin, A. (2011). Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. Genome Biol. 12, R13.

42. Danielian, P. S., Muccino, D., Rowitch, D. H., Michael, S. K., & McMahon, A. P. (1998). Modification of gene activity in mouse embryos in utero by a tamoxifen-inducible form of Cre recombinase. *Current Biology, 8*(24), 1323-1326. doi:10.1016/s0960-9822(07)00562-3

Chapter 4 – Infection models reveal the role of *lincRNA-Cox2* in responsive immunity

4.1 – Identification of *lincRNA-Cox2*

We have previously discussed our collaborative screen from 2009 in which several thousand novel lincRNA loci were discovered based on the presence of a canonical K4-K36 chromatin domain (see Chapter 1.3).[1] Several cell types were employed in this screen, in order to identify tissues-specific lincRNAs at the broadest scale possible. One of these cell types included CD11C+ bone marrow-derived dendritic cells. Dendritic cells are a central component of the mammalian immune system, and function through antigen presentation to T cells of the adaptive immune system.[2] In order to present an antigen to T cells for further induction of a global immune response, dendritic cells need to first recognize a pathogen through an innate and passive system. This is accomplished through the use of pattern recognition receptors and toll-like receptors (TLRs) on the surface of the dendritic cell. Toll-like receptors recognize, and bind, specific chemical signatures found on pathogens, and specific pathogens will active unique TLRs (for example, bacterial lipopolysaccharide, LPS, is a potent activator of TLR4; we will discuss this example in greater detail below).[3] Once an immature dendritic cell comes into contact with a pathogen it recognizes, it will phagocytose said pathogen in order to "activate". During this activation process, dendritic cells degrade the phagocytosed pathogen into component proteins, while simultaneously migrating toward the nearest lymph node in the body. At the lymph nodes, dendritic cells utilize major histocompatibility complex proteins (MHC-class proteins) to "present" pathogen proteins to T-cells, which in turn

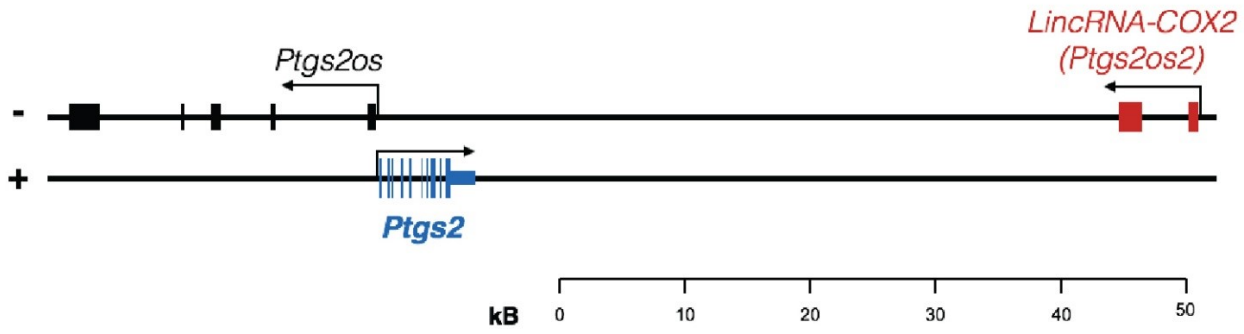
mount an adaptive immune response targeted against the pathogen associated with the presented protein fragments.[4] In this way, dendritic cells act as a key mediator between the innate and adaptive immune systems, and function in a keystone role in the mammalian response to harmful pathogens.[5]

CD11C+ is a gene that encodes a transmembrane protein integral to multiple cell types, including dendritic cells as well as other leukocytes (white blood cells of the immune system).[6] Other CD11C+ cell types include monocytes, macrophages, neutrophils, and B cells.[7] Sorting of cells by this surface protein, therefore, would also capture other non-dendritic immune cells. The screen, therefore, could be considered a non-specific search for lincRNAs involved in the broader immune system. Incubation with LPS, mentioned above, activates a signaling cascade through potent antagonism of the TLR4 receptor, and subsequently induces a strong immune response.[8] This signaling pathway has been previously described, starting with LPS binding to TLR4. Lipopolysaccharides (LPS) are large lipid/polysaccharide molecules that are found embedded in the outer membrane of gram-negative bacteria.[9] LPS acts as the major component in the outer membrane of these bacteria, and contributes not only to the structural integrity of said membrane but has also been associated with adhesion of bacteria to various surfaces, bacteriophage sensing, and chemical defense. LPS elicits a strong immune response upon binding to TLR4, presumably as an evolved defense against gram-negative pathogens invading their mammalian hosts.[1, 9] Binding of LPS to TLR4 induces expression of pathogen resistance genes through NF-kB, a well-studied transcription factor that controls the expression of genes in response to external stimuli such as stress, UV radiation, and pathogen presentation (among other stimuli).[10, 11] In

short, CD11C+ immune cells become “activated”, and turn on immune-specific genes, in the presence of LPS.

Previous work by our collaborators, Susan Carpenter and Kate Fitzgerald at the University of Massachusetts at Amherst, included a lincRNA screen that found stimulation of these CD11C+ with LPS induces the expression of 20 lincRNA loci. These genes demonstrated significant upregulation following LPS-stimulation and, interestingly, most of them resided within genomic proximity to the NF-kB locus on murine chromosome 3, chr3:135,584,655-135,691,547 (human chromosome 4, chr4:102,501,329-102,617,302). One putative lincRNA locus, located outside of the NF-kB locus on murine chromosome 1, showed the single greatest change in gene expression following LPS stimulation. This gene, located 51 kilobases downstream (Fig. 30) of the protein coding gene cyclooxygenase-2 (*Cox2*). For the sake of clarity, we will heretofore refer to this protein coding neighbor as its alternate name, Prostaglandin-endoperoxide synthase 2 *(Ptgs2*).[13] This lincRNA locus, known as *Ptgs2os2* (for *Ptgs2* opposing strand transcript 2) was found to be induced over 1,000x 12 hours after LPS stimulation of TLR4. This locus was renamed *lincRNA-Cox2*, and so use of the name *Cox2* will henceforth refer to the lincRNA, while *Ptgs2* remains reserved for the protein coding gene. Interestingly, stimulation of CD11C+ cells via TLR3 did not induce *linc-Cox2* in any significant way, indicating that this lincRNA responds potently to specific pathogens and could play a functional role in mammalian immune response to TLR4 pathway activation *in vivo*.

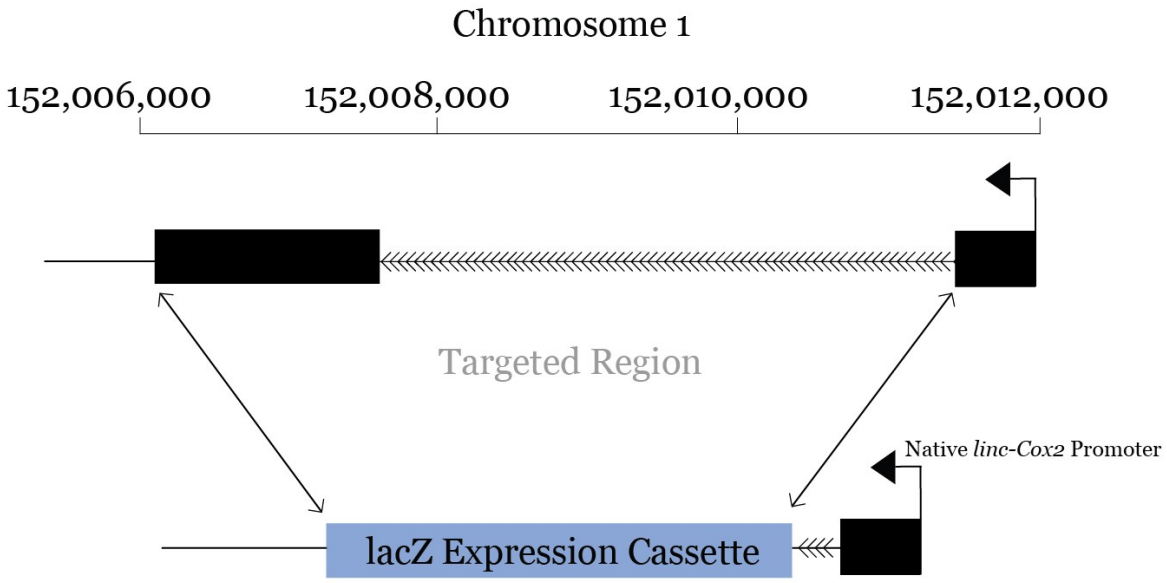
Figure 30:



4.2 – lacZ model of *linc-Cox2* and background results

lincRNA-Cox2 is a 2 exon transcript that originates from a <6kb locus on murine chromosome 1. We selected *Cox2* as one of our 18 lincRNA knockout strains based on the promising expression dynamics found in our early lincRNA screen, described above,[14] as well as other selection criteria describe in greater detail in Chapter 2. We replaced the linc-Cox2 locus with a lacZ expression cassette downstream of exon 1 (Fig. 31), keeping the promoter intact and leaving over 51kb of space between the deleted gene and the neighboring protein coding gene, *Ptgs2*.

Figure 31:



In our initial survey of lincRNA expression by X-gal staining in whole organs from each of our 18 knockout strains, we looked at *linc-Cox2* following CO2 euthanization and cervical dislocation (as per standard IACUC protocol), followed by fixation of tissues and subsequent X-gal staining. We did not observe any lacZ staining in any of the surveyed tissues using this method (data not shown). At the same time, Kate Fitzgerald's lab at the University of Massachusetts at Amherst also published their study of lincRNA-Cox2 in a cell-based system. They were looking for long noncoding RNA genes that were transcribed during the innate immune response, such as the dendritic activation by TLR4 stimulation described above. Aside from dendritic cells, other leukocytes (including some of those mentioned previously in this chapter) are integral to the mammalian innate immune system; included among those are macrophages, large white blood cells that are derived from monocytes produced by hematopoietic differentiation in the bone marrow. In order

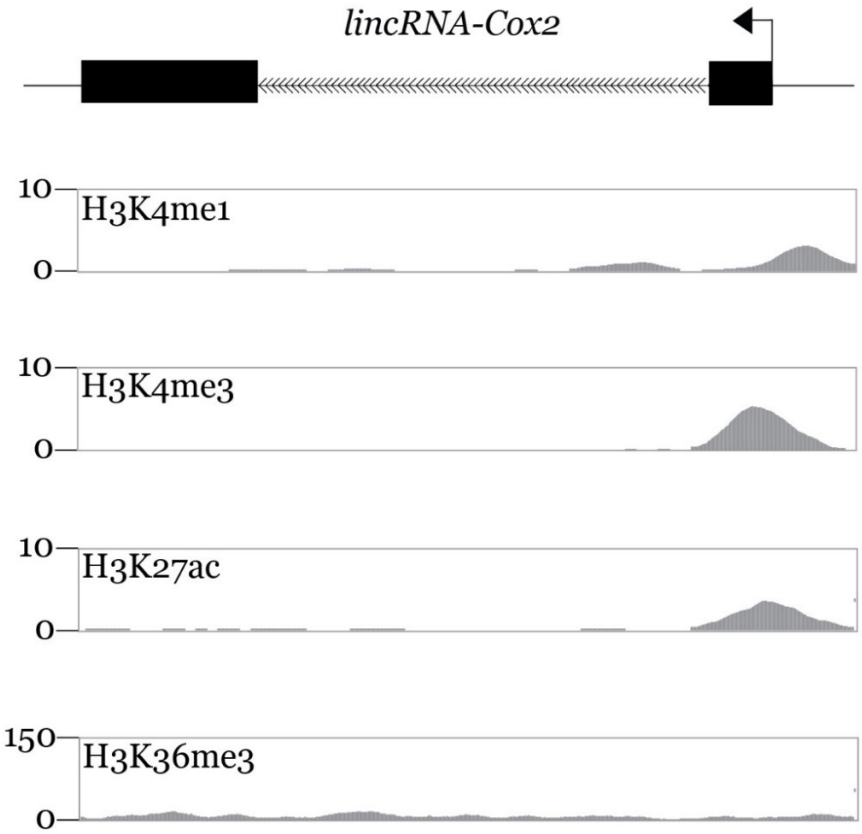
to obtain macrophages from mice, bone marrow cells from wild-type mice were cultured in DMEM with 10% fetal bovine serum (FBS) and 20% L929 supernatants. Bone marrow-derived macrophages (BMDMs) were subsequently immortalized, providing more convenient vessels for long-term cell culture assays, with a J2 virus.[25] These BMDMs were stimulated as per our first experimental ID of linc-Cox2 in dendritic cells (above),[1] by incubating the cells for up to 24 hours in 100ng/mL of LPS (a Tlr4 agonist), 100nM of Pam3CSK4 (a Tlr2/1 agonist),[15] 100ng/mL of Pam2CSK4 (a Tlr2/6 agonist),[16] 25ug/mL of PolyI:C (a Tlr3 agonist),[17] or 5ug/mL of poly(dA-dT) (a synthetic double-stranded RNA that acts as a Tlr3 agonist).[1] These experiments confirmed that transcription of *linc-Cox2* is induced following Tlr4 stimulation, and that Tlr3 activation does not upregulate the lincRNA. Furthermore, activation of Tlr1, 2, 7, and 8 also induce *lincRNA-Cox2* expression. Interestingly, the authors also ran these cell-based infection models with Myd88$^{-/-}$ BMDMs. Myd88 (myeloid differentiation primary response gene 88) acts downstream of all toll-like receptors, except Tlr3, in an immune signaling cascade to activate NF-kB.[18] The authors propose a model whereby TLR signaling induces the expression of *lincRNA-Cox2* through Myd88-mediated NF-kB signaling. As a master regulator of inflammation and immunity, NF-kB points to a model for linc-*Cox2* in which the lincRNA goes on to serve as a regulator of immune response genes through its interactions with various regulatory complexes. This study laid the groundwork for further study of lincRNAs in mammalian immunity, and began to explore the possibility that lncRNAs represent a previously overlooked central component of innate immunity. As such, I teamed up with Susan Carpenter, the study's lead author, and Roland Elling, a MD in the Fitzgerald lab, to utilize our linc-Cox2 knockout strain for *in vivo* infection studies aimed at revealing the role of this noncoding locus in the innate immune response,

potentially paving the way for the identification of both novel drug targets as well as opportunities for the advancement of therapeutics in infectious and inflammatory disease.

4.3 – Characterization of the *lincRNA-Cox2* genomic locus

Chromatin signatures from publicly available ChIP data in macrophage cell lines (UCSC) at the *linc-Cox2* locus reveal strong H3K4 trimethylation, but low signal on H3k36me3. Interestingly, there is strong presence of H3K4me1 and H3K27ac signal, indicating the potential for this locus to possibly function as a DNA enhancer element (Fig. 32).[19,20]

Figure 32:



RNA sequencing of a panel of tissues from unchallenged (non-stimulated) mice, harvest in the same way as described above in Ch. 4.2, revealed (as expected) no significant expression of *linc-Cox2*. It was, however, surprising that even in unstimulated mice we found low levels of *Cox2* expression in lungs and lung fibroblasts (Fig. 33). A phenomenon that remained consistent across replicates, this could indicate either a persistent, low background level of inflammation in the lungs that is inducing low levels of *linc-Cox2* expression, or perhaps the lincRNA locus is performing other functions in these cell types.

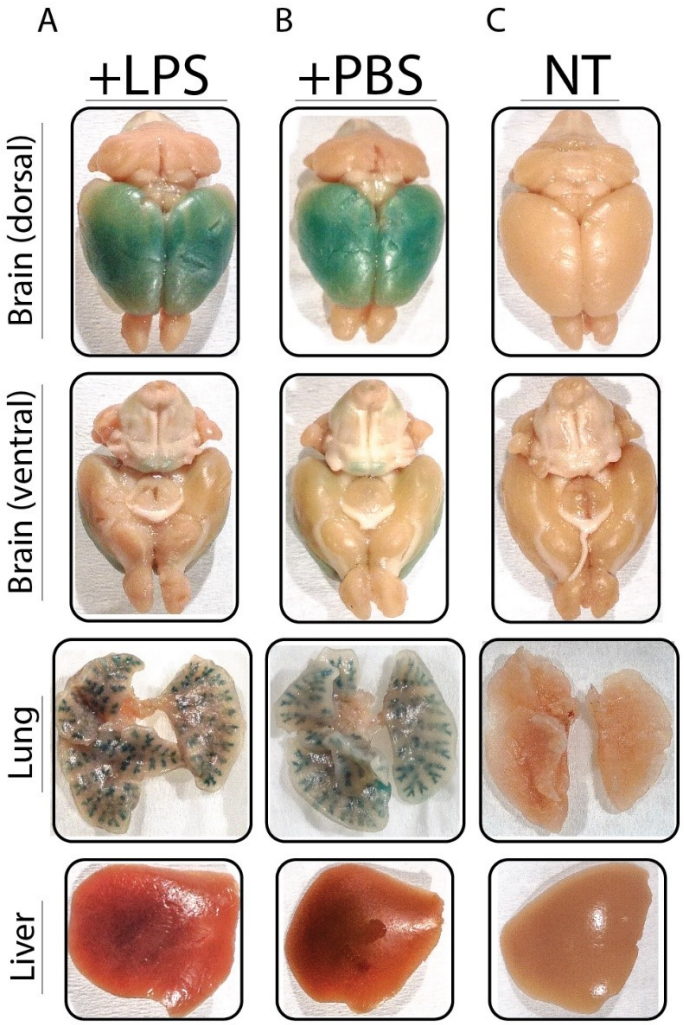
Figure 33:

4.4 – Expression dynamics of *linc-Cox2* in immune-challenged mice

In order to understand the expression dynamics of *linc-Cox2 in vivo*, we performed n=3 intraperitoneal injection (IP, injection into the body cavity, as per IACUC standards) using 100ng/mL LPS. In addition to an LPS-treated group, we also injected n=3 control mice IP with phosphate-buffered saline solution (PBS, GIBCO), and kept n=3 mice as an

un-injected (no treatment, NT) control. All mice were housed for 6 hours of incubation post-treatment in a sterile environment (Harvard mouse procedure room, Sherman Fairchild Biochemistry Building 3rd Floor) at room temperature (25°C). Treatment groups were kept separate from one another, but mice within treatment groups shared an enclosure, food, water, and bedding. At 6 hours, mice were euthanized, and tissues were harvested, fixed, and stained as per our standard protocol. The results were indeed surprising, and Figure 34 summarizes them below:

Figure 34:



From left to right in Fig. 34 (above): column (A) represents HET mice treated with 6 hours of 100ng/mL LPS; column (B) represents HET mice treated with 6 hours of PBS; column (C) represents HET mice in the untreated group. No lacZ staining was observed in the majority of tissues screened, including liver (bottom row). Two tissues stained in all 3 of the +LPS mice: the lungs, and the brain. Surprisingly, those same two tissues stained in all three of the +PBS mice as well. No tissues in any of the NT mice were stained. There are three aspects of these results that warrant further discussion, and I will address them in order: (1) the staining of the +PBS group tissues, (2) the appearance of lacZ staining in the brain, and (3) the highly specific staining pattern observed in the lungs.

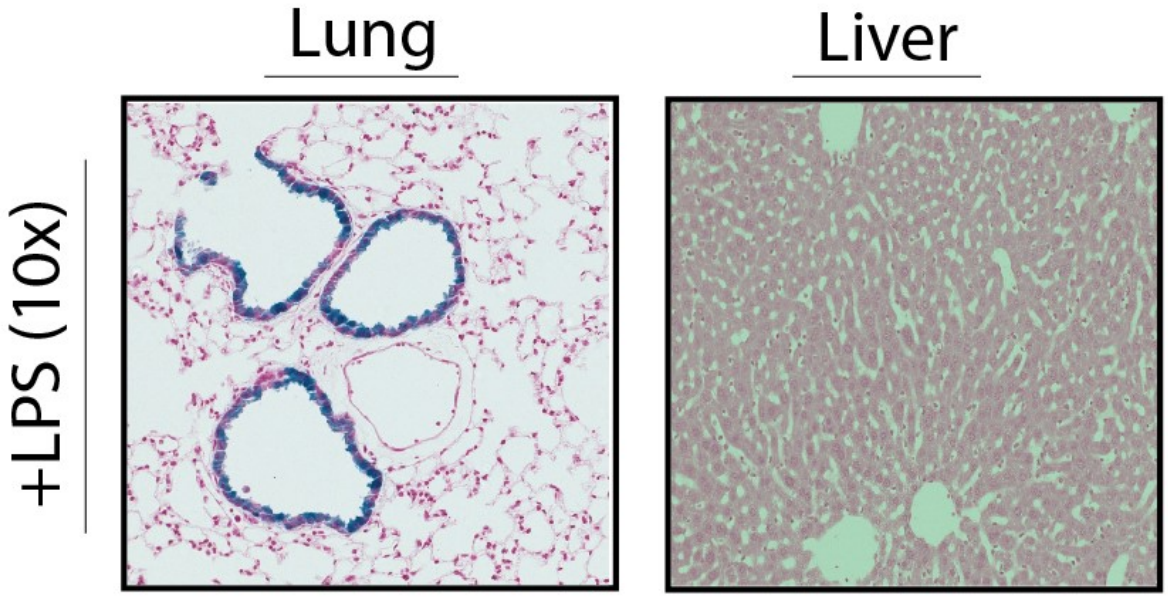
It was initially vexing as to why the PBS-treated mice would exhibit lacZ staining. Not only that but, if we are to believe that LPS treatment reflects a "true signal", and shows the spatial positioning of where *lincRNA-Cox2* should be expressed during immune challenge, why is the +PBS lacZ staining restricted to those tissues as well (and not eliciting a broader, more global response)? In retrospect, this should not surprise the reader, as IP of any substance should theoretically induce a global inflammatory response. Indeed, several papers have come out in immunology journals over recent years that report similar findings, including one study that found interleukin genes being upregulated following PBS-IP (IL4 in particular, which resides downstream of NF-kB in the inflammation signaling pathway). Therefore, we must conclude that IP of PBS induced a global inflammatory response through NF-kB signaling that induced expression of lacZ via the native *lincRNA-Cox2* promoter; whether this was accomplished through direct binding of NF-kB to the promoter, or through a stepwise circuit starting

with NF-kB signaling and ending with transcription of the *Cox2* locus, remains to be unraveled.

The appearance of lacZ staining in the brain remains to this day less clear of a story, as we have yet to follow up on that particular organ system (having favored the lungs and leukocyte cell populations in our study). Through conventional understanding of the blood-brain barrier, it seems unlikely that LPS-IP would be capable of inducing an inflammatory response in the brain capable of inducing *Cox2* expression. However, clear lacZ expression in the upper cortical layers of the brain can be observed not only in +LPS-treated mice, but also in +PBS-treated mice. There is a trickle of newfound evidence coming from various groups that indicates the blood-brain barrier to be more permeable than has been historically considered true, although future studies could endeavor to identify what, if any, roles this locus possesses in the mammalian brain.

Finally, we move to the lungs. The highly specific staining pattern observed in both +LPS and +PBS treated groups, as seen in Figure 34 above, was the most specific staining pattern that I observed among any whole-mount tissues screened from our 18 lincRNA knockout strains. In order to obtain a better understanding of the localization of lacZ expression within this organ system, I sectioned stained and fixed (4% PFA) tissue as previously described and counterstained sections with eosin Y, which colors non-nuclear structures of cells and tissues in order to provide contrast for visualization. Analysis of sections revealed that lacZ expression appeared in the lower bronchi (toward the bronchiolar-alveolar junctions), through which oxygenated air passes on its way to alveolar sacs for gas exchange with the blood. Representative images of eosin-stained bronchiolar and eosin-stained hepatic (control) tissues can be found below in Fig. 36.

Figure 35:



It appears that the cell population labeled in these sections are ciliated pulmonary macrophages (PMs), a leukocyte population that resides in the lungs. There are sub-classes of PMs, including alveolar-specific macrophages (AMs), and bronchiolar-specific macrophages (BMs), but due to the spatial localization of the *linc-Cox2* lacZ staining we were unable to determine which of these subsets was responsible for the observed expression pattern. PMs serve not only as typical macrophages, in that they are a first line of defense in the innate immune battle against invading pathogens (especially airborne, given their distribution in the body), but PMs also function by remediating acute lung damage. As the body's point of oxygen gas exchange with the blood, the lungs are under particular stress from drying out, expansion and contraction of endothelium, pathogen invasion through the dermis, chemical exposure, and other damaging sources. Therefore, one of the roles of the PMs and other mesenchymal cell types residing within the lungs is to alleviate this stress through fibrosis (particularly in the case of acute hemorrhaging),

and pathogen clearance. At this point in our study, it was unclear if *linc-Cox2* was involved in any of these inflammatory response pathways. However, given our understanding of the Tlr-mediated expression dynamics of this locus in cell-based assays, as well as the observed *in vivo* lacZ expression in pulmonary macrophages, it seemed reasonable to hypothesize that lincRNA-Cox2 is involved in regulating inflammation, specifically in lung tissues (and possibly in other tissues, such as the brain).

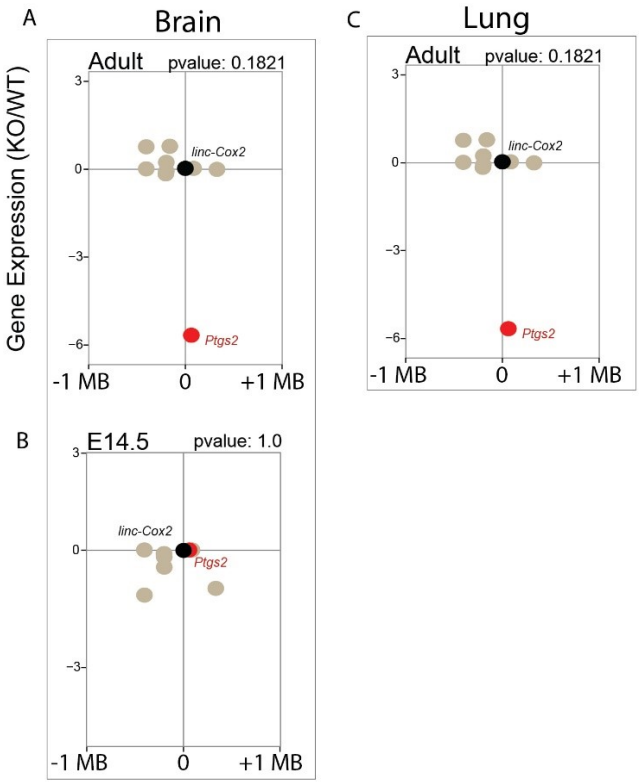
4.5 – RNA sequencing reveals a potential lung-inflammation role for *Cox2*

The observed expression in the lung and brains is highly specific, and so we transitioned to looking at mice with a total loss of function at the *linc-Cox2* locus. As we did with the *linc-Tug1* testes, in order to truly understand how this locus functions in the context of inflammation, specifically in lung inflammation and immune response, we need to study the perturbations associated with *Cox2* loss of function in genetic and molecular contexts.

Since *Cox2* is regulated by NF-kB signaling, we wanted to see whether deletion of this locus results in downstream changes to known inflammatory response genes. Our hypothesis was that linc-Cox2 functions as a regulator of gene expression following immune challenge, though whether it serves as an activator or a repressor of these genes was still unclear. Differential regulation of genes and gene pathways could point to the pathways in which *linc-Cox2* exhibits a functional role, and could indicate pathological differences between *Cox2* knockout and wild type mice as well. For these reasons, it is critical to identify whether genes are misregulated in these mice, specifically in the lungs, brain, and primary macrophages. RNA sequencing of the adult brain (Fig. 37A),

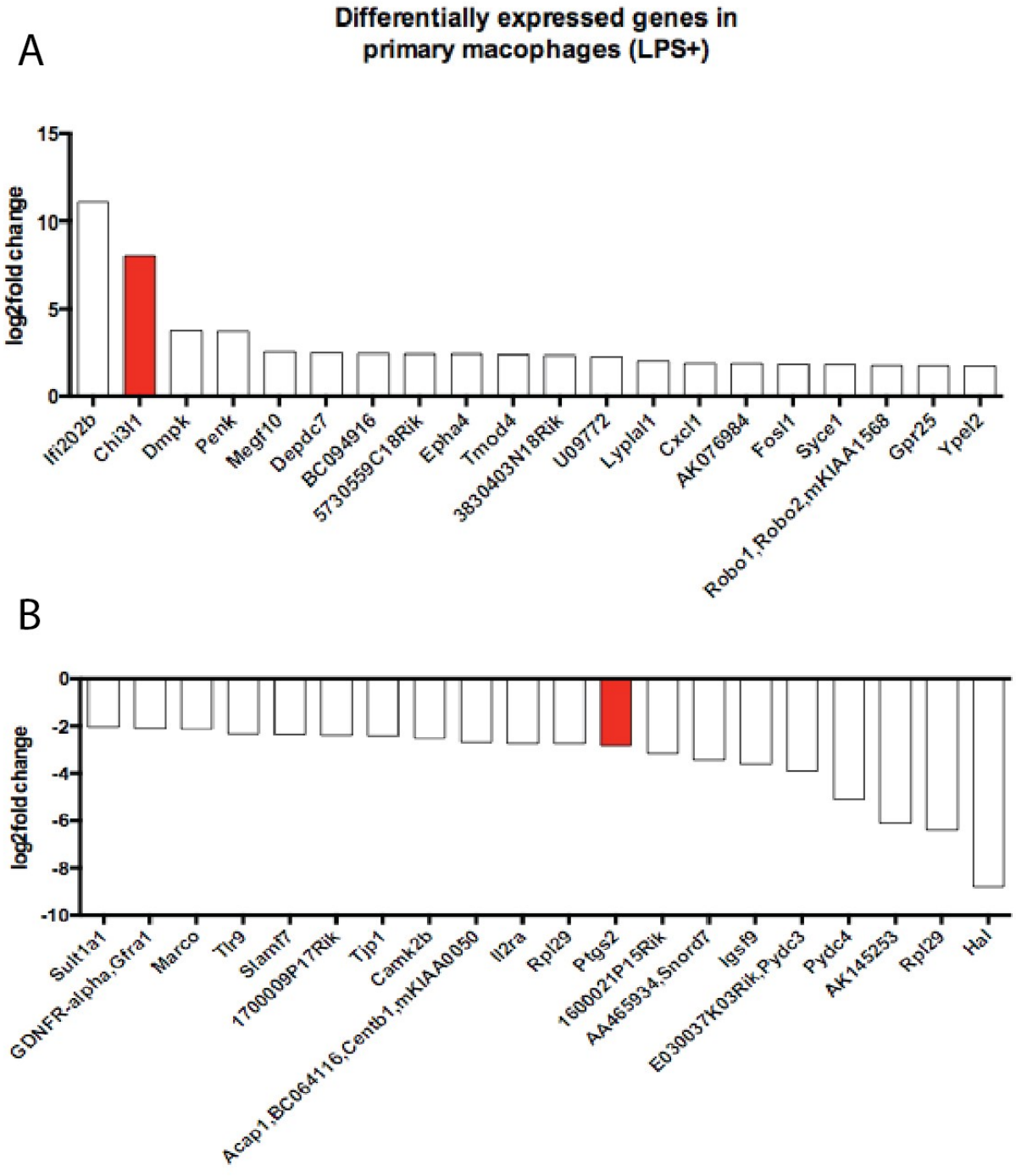
embryonic E14.5 brain (Fig. 37B), and adult lungs (Fig. 37C) showed that, when we looked at the same ±1MB window used for our *Tug1* cis plots, only one gene was downregulated in the *linc-Cox2* KO mice relative to WT: *Ptgs2* (p< 0.1821). This was consistent between adult tissues, but was not observed in embryonic brain sequencing. This is because *Ptgs2* is not expressed in either WT or KO brains at E14.5, so there is no significant change between the genetic backgrounds. Importantly, these mice were not immune challenged (+LPS or +PBS) prior to harvesting tissues and purifying RNA, so we are looking at the so-called "baseline" expression of *linc-Cox2* (hence 0 expression in either WT or KO brain and lung tissues) and *Ptgs2*. Nevertheless, the fact that *Ptgs2* was downregulated in our *linc-Cox2* KO mice even in the absence of immune challenge indicated that the *linc-Cox2* is functioning as a DNA enhancer for this well-known immune response gene.

Figure 36:



From here, we obtained primary macrophages (not immortalized) from both *linc-Cox2* KO and WT mice. We treated these mice with 100ng/mL LPS for 6 hours, purified RNA, and performed RNA sequencing to see what happens in the presence of an acute immune challenge when *linc-Cox2* is absent. We identified 93 genes that were significantly upregulated in the KO relative to WT (log2foldchange>2, p< 0.0008) as well as 216 genes that were significant downregulated (p< 0.0007). Consistent with the observed changes in adult lung and brain tissue (above), *Ptgs2* was the 9th-most downregulated gene in the gene set (Fig. 38B), with an 86% reduction in KO primary macrophages following 6 hours of LPS stimulation (n=3, p<0.0001). Among the genes that were upregulated, however, the two most upregulated stood out in particular: *Ifi202b,* and *Chi3l1*. These inflammatory response genes were massively upregulated in the KO mice, with *Ifi202b* upregulated more than 2,000-fold over WT +LPS (p<0.0001) and *Chi3l1* upregulated almost 300-fold over WT (p<0.0006). Figure 38A illustrates a rank list of the most upregulated genes, and highlights *Chi3l1* in the lineup as we will circle back to discuss this gene further toward the end of this chapter.

Figure 37:


A
Differentially expressed genes in primary macophages (LPS+)
log2fold change
15
10
5
0
Ifi202b
Chi3l1
Dmpk
Penk
Megf10
Depdc7
BC094916
5730559C18Rik
Epha4
Tmod4
3830403N18Rik
U09772
Lyplal1
Cxcl1
AK076984
Fosl1
Syce1
Robo1,Robo2,mKIAA1568
Gpr25
Ypel2
B
log2fold change
0
-2
-4
-6
-8
-10
Sult1a1
GDNFR-alpha,Gfra1
Marco
Tlr9
Slamf7
1700009P17Rik
Tjp1
Camk2b
Acap1,BC064116,Centb1,mKIAA0050
Il2ra
Rpl29
Ptgs2
1600021P15Rik
AA465934,Snord7
Igsf9
E030037K03Rik,Pydc3
Pydc4
AK145253
Rpl29
Hal

4.6 – Cell based infection models of inflammatory gene expression levels

Considering the consistent observation that absence of *linc-Cox2* results in depletion of *Ptgs2* mRNA, we set out to further characterize the interaction between these two components. To do so, we generated primary macrophages from *lincRNA-Cox2* WT, HET, and KO mice and performed a series of infection studies using various Tlr agonists (Fig. 39A). We observe potent activation of *linc-Cox2* following Pam3 (Tlr1, 2) LPS (Tlr4), and R848 (Tlr7/8) stimulation in WT cells, but nothing in KO cells. Additionally, as expected, we observe no induction of *Cox2* in either WT or KO cells following pIC (Tlr3) challenge. Performing similar infection tests with bacterial isolates (Fig. 39B) or intact pathogens (Fig. 39C) demonstrated a significant reduction in *Cox2* expression following these challenges (with the exception of Tlr3 ligands pIC and Sendai Virus).

Figure 38:

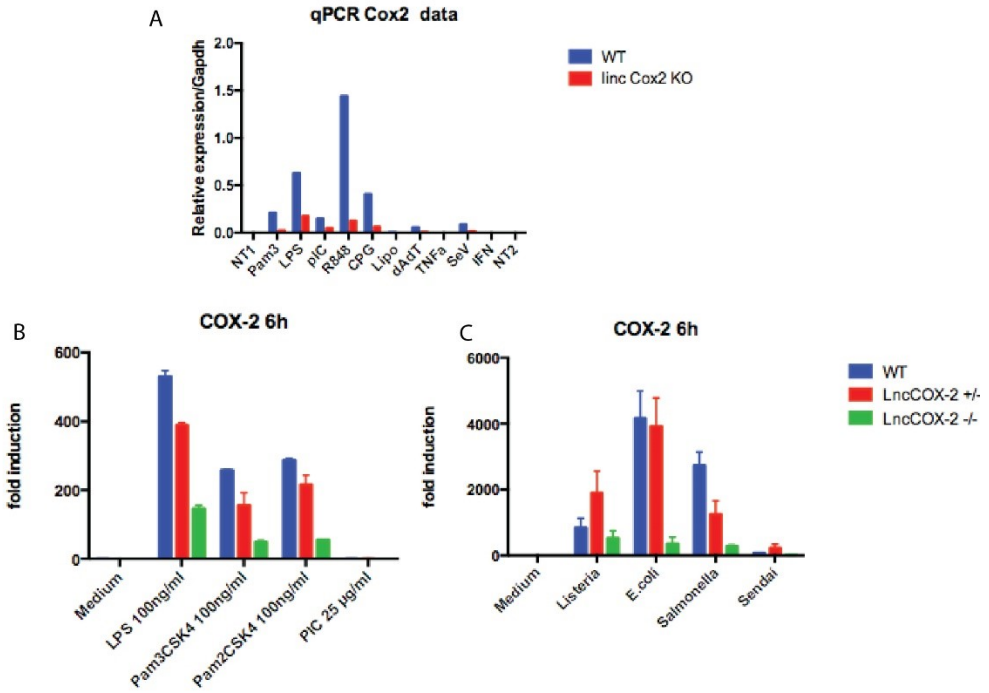
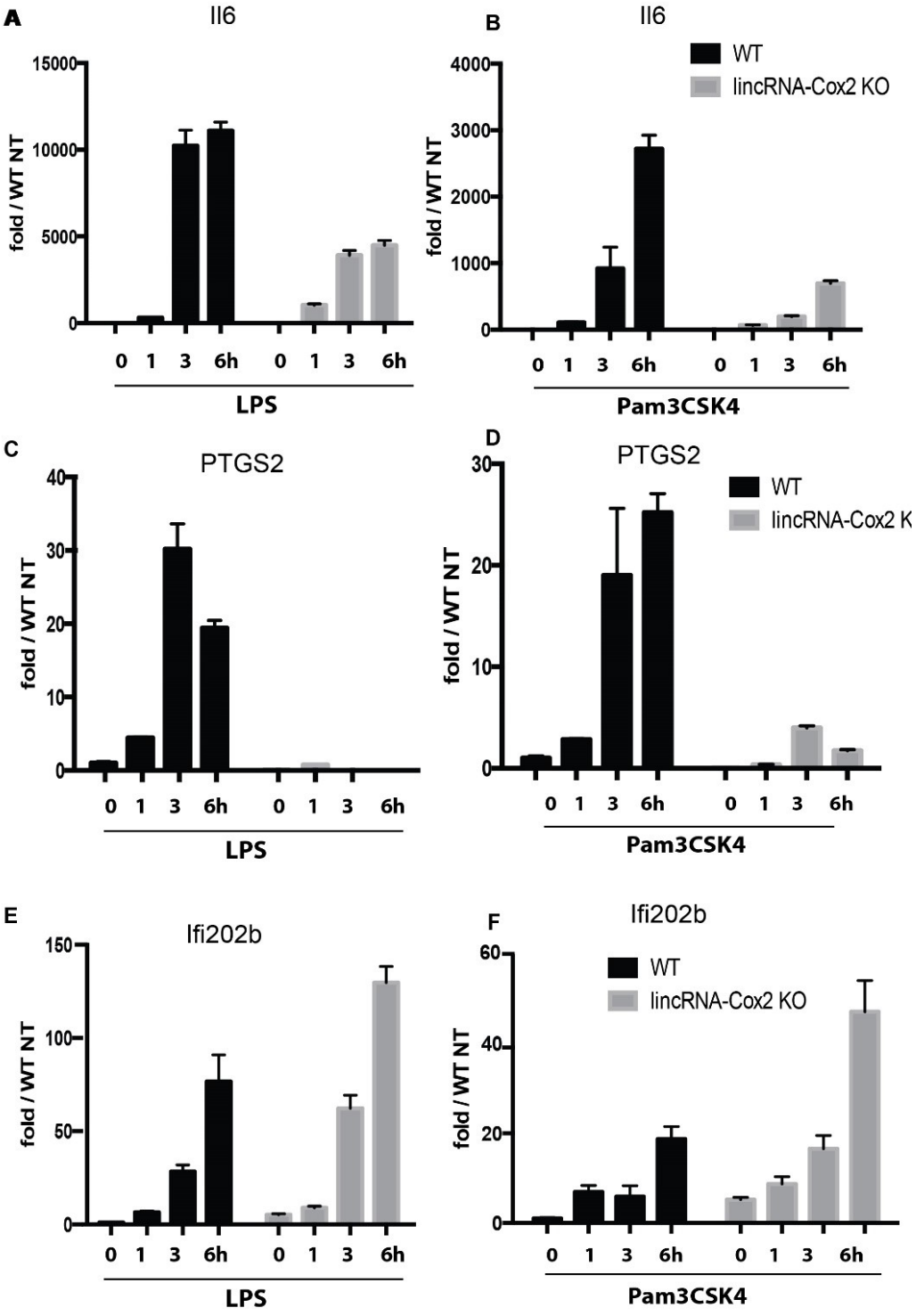


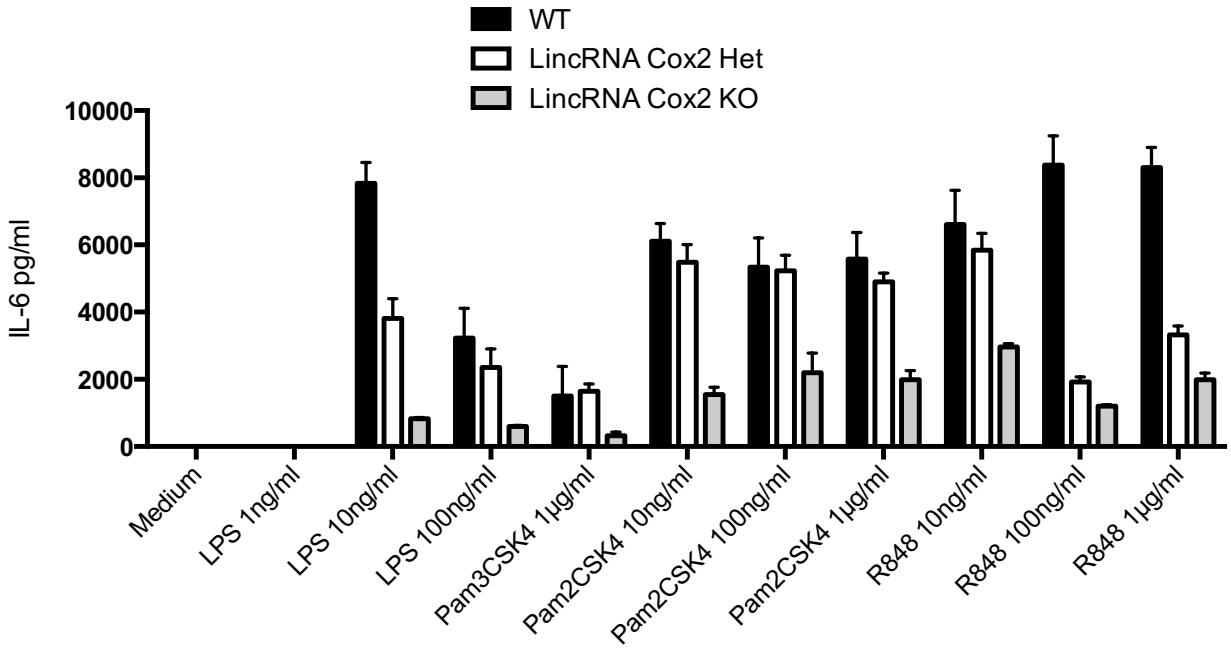
Figure 39 (below) illustrates our experiments with two Tlr ligands: LPS (Fig. 39, left-hand column), a Tlr4 ligand described above; and Pam3CSK4 (Fig. 39, right-hand column), a Tlr2 ligand. We looked at the mRNA expression levels of three genes via qPCR: interleukin 6 (*IL6*), *Ptgs2*, and *Ifi202b*. *IL6* is a very well characterized pro-inflammatory cytokine that is secreted by macrophages, among other leukocytes, in response to inflammation. Once secreted, *IL6* binds to interleukin receptors on target cells in order to stimulate genetic changes for immune response, especially after wounding or tissue damage that leads to inflammation. In Carpenter et al. (2013), IL6 was found to be potently downregulated following *linc-Cox2* knockdown. We also found downregulation of IL6 in our primary macrophage sequencing in KO +LPS relative to WT +LPS (30% reduction, $p<0.0001$). Running a 6-hour treatment time course in WT and KO primary macrophages, we observed a potent (>10,000-fold) increase in IL6 expression following LPS treatment in WT cells (Fig. 40A). This effect was relatively muted in KO cells, but upregulation (<5,000-fold increase) was still observed. This pattern is maintained, albeit at lower levels across the board, for Pam3Csk4 stimulation (Fig. 40B). *Ptgs2* behaved as previously observed, with potent upregulation in WT cells following LPS and Pam3Csk4 stimulation; in KO cells, *Ptgs2* expression is essentially zero in the presence of LPS - in the presence of Pam3Csk4, however, *Ptgs2* induction does happen at low levels relative to WT cells. This indicates that *Ptgs2* induction probably happens downstream of *linc-Cox2*, and that Tlr2 stimulation could possibly **activate transcription at the *Ptgs2* locus by bypassing the NF-kB-mediated Tlr4 stimulation of *linc-Cox2***.

Figure 39:


A
Il6
fold / WT NT
15000
10000
5000
0
0 1 3 6h 0 1 3 6h
LPS
B
Il6
WT
lincRNA-Cox2 KO
fold / WT NT
4000
3000
2000
1000
0
0 1 3 6h 0 1 3 6h
Pam3CSK4
C
PTGS2
fold / WT NT
40
30
20
10
0
0 1 3 6h 0 1 3 6h
LPS
D
PTGS2
WT
lincRNA-Cox2 K
fold / WT NT
30
20
10
0
0 1 3 6h 0 1 3 6h
Pam3CSK4
E
Ifi202b
fold / WT NT
150
100
50
0
0 1 3 6h 0 1 3 6h
LPS
F
Ifi202b
WT
lincRNA-Cox2 KO
fold / WT NT
60
40
20
0
0 1 3 6h 0 1 3 6h
Pam3CSK4

We also looked at the protein levels (via enzyme-linked immunosorbent assay, ELISA)[22] of interleukin 6 (IL6). We confirmed that a *linc-Cox2* KO not only results in downregulation of the IL6 mRNA, but also results in a depletion of the IL6 protein as well (Fig. 40).

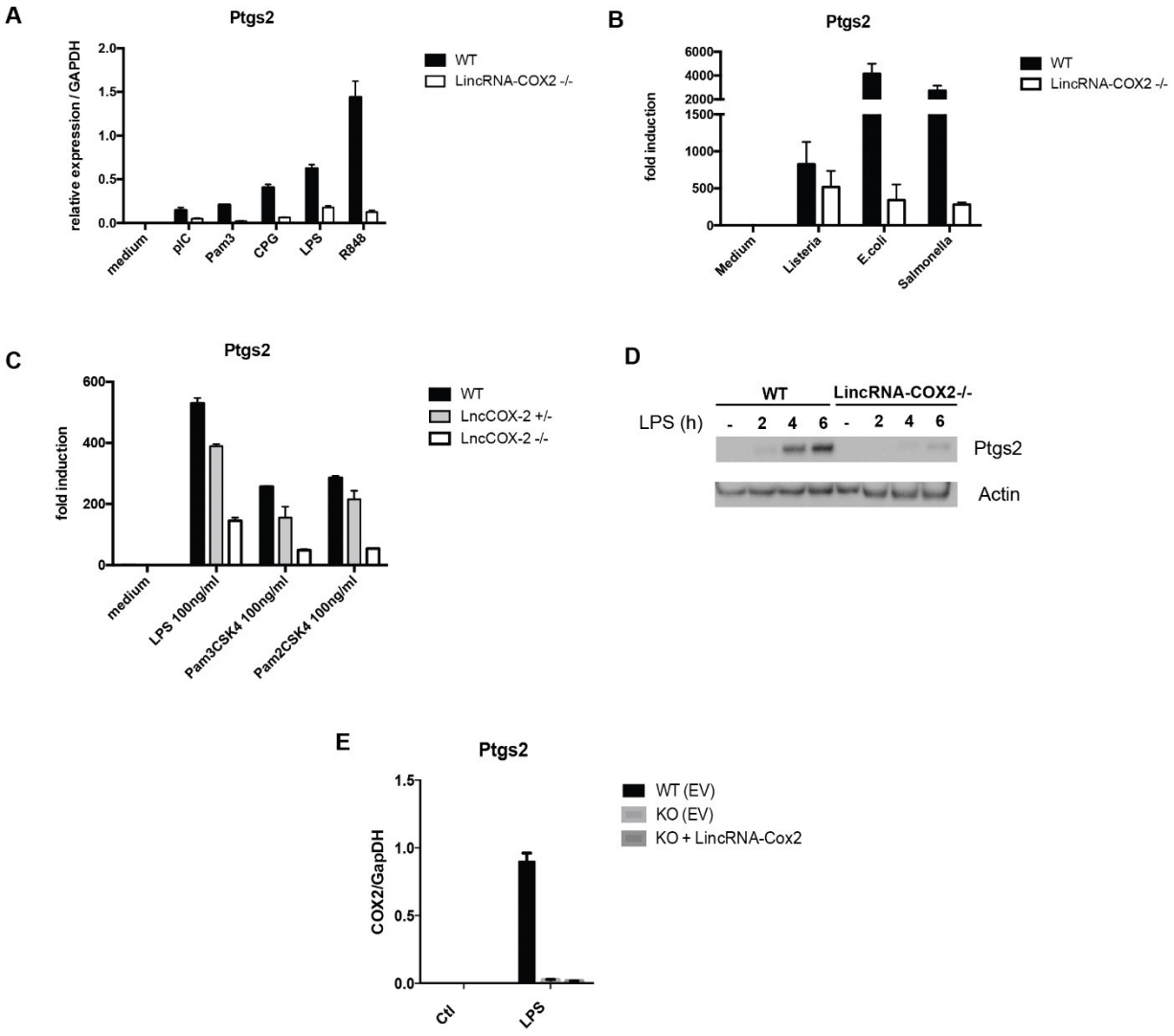
Figure 40:



4.7 – Effects of *linc-Cox2* on *Ptgs2* expression

The recurring observation that *Ptgs2* is downregulated in *linc-Cox2* KO tissues and cell samples (as measured by both RNA sequencing and qPCR) caused us to question whether *linc-Cox2* is functioning upstream of *Ptgs2*, and functions by regulating the expression of this protein-coding gene. We have already demonstrated that *Ptgs2* is induced following LPS stimulation in WT cells, but that it is essentially ablated in the absence of *linc-Cox2*. In order to expand on these findings, we measured *Ptgs2* expression levels following Tlr stimulation using prepared pathogen compounds, like LPS, (Fig. 41A), as well as intact bacterial pathogens (Fig. 41B). The results mirror those observed earlier, wherein Tlr agonists induce *Ptgs2* expression in WT cells, but induction is quelled in KO cells. When we analyzed mice that were heterozygous for the *linc-Cox2* locus, there seems to be an intermediate effect, with *Ptgs2* induction levels falling precisely between those from the WT and KO cell samples (Fig. 41C). Western blot analysis demonstrates that not only does *linc-Cox2* deletion result in downregulation of transcription at the *Ptgs2* locus, but protein levels are also affected by the knockout (Fig. 41D). Finally, neither lentiviral nor transient overexpression of *linc-Cox2* is not sufficient to recover *Ptgs2* expression levels in KO cells following LPS stimulation for 6 hours (lentiviral results Fig. 41E, transient data not shown). It is important to keep in mind that lentiviral integration and overexpression of this construct happens randomly, as with all traditional forms of lentiviral overexpression, and so this experiment simply rules out that the *linc-Cox2* transcript regulates the *Ptgs2* locus in *trans*. It does not offer evidence supporting or disproving the hypothesis that *linc-Cox2* regulates its neighboring gene in *cis*.

Figure 41:

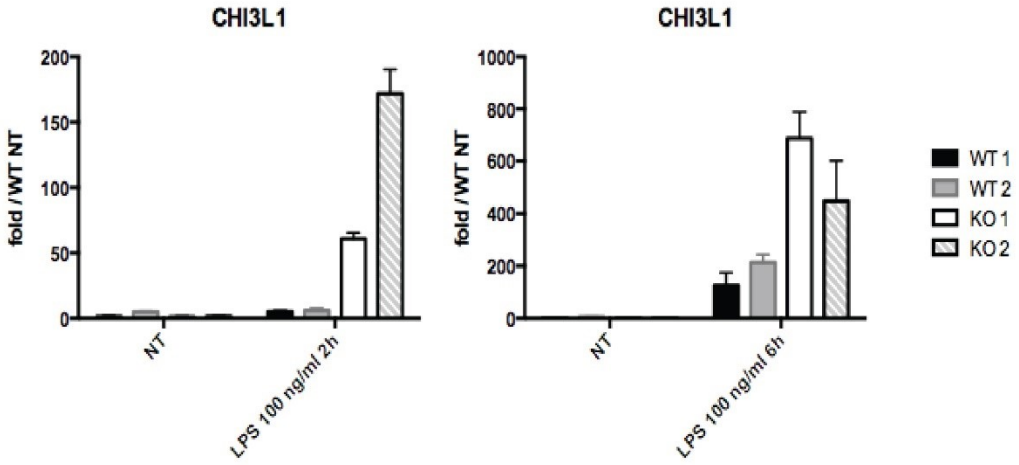


4.8 – Probing the relationship between *Ptgs2* and *linc-Cox2*

Chitinase-3-like 1 (*Chi3l1*) is a protein coding gene located on murine chromosome 1, about 150kb downstream of the *linc-Cox2* locus. Alternatively known as *YKL-40,* this gene codes for a secreted glycoprotein that originates in macrophages and has been associated with remodeling of wound tissue following acute injury – it has been particularly well studied in the mammalian lungs, and in the development of asthma.[21]

We find massive upregulation of Chi3l1 in stimulated cells at both the mRNA and protein levels (Fig. 43). Based on our analysis of *Ptgs2* expression, however, we wondered whether *linc-Cox2* acted directly as a repressor of *Chi3l1*, or if it was the depletion of *Ptgs2* that caused the observed upregulation of the secreted glycoprotein. In order to answer this question, we generated primary macrophages from *Ptgs2* knockout mice (*Ptgs2*$^{-/-}$) and stimulated them with 100ng/mL LPS for 6 hours. Our logic was as follows: if *Ptgs2* is regulating expression of *Chi3l1*, then a *Ptgs2* KO should recapitulate the genetic changes observed in our *linc-Cox2* KO cells. If, however, *linc-Cox2* acts on *Chi3l1* independently of *Ptgs2*, then we would not see those same changes in the *Ptgs2* KO (in which both endogenous copies of *linc-Cox2* are functioning as WT). We monitored the expression of *Ptgs2 (*Fig. 44A), *linc-Cox2* (Fig. 43B), *Chi3l1* (Fig. 44C), and *Ifi202b (*Fig. 44D), in addition to several benchmark inflammatory genes (*Il6, RANTES,* and *TNF-α*). As expected, *linc-Cox2* is upregulated following LPS stimulation in both *Ptgs2*$^{+/+}$ and *Ptgs2*$^{-/-}$ cells, indicating that the lincRNA functions upstream during an immune response. *IL6* and *RANTES* are upregulated in WT and KO, though higher in the KO, which could point toward *Ptgs2* as a negative regulator of these inflammation genes. *Chi3l1*, however, is depleted in both WT and *Ptgs2* KO cells following LPS stimulation. These results are duplicated when LPS treatment dosage is increased 10x to 1μg/mL (Fig. 45). This leaves us with a clear phenotype associated with *linc-Cox2* KO, one that is independent of the neighboring protein coding gene *Ptgs2*. How *linc-Cox2* regulates *Chi3l1* remains unknown, yet we have begun to unravel what the implications are following upregulation of this secreted protein.

Figure 42:



2h LPS 6h LPS

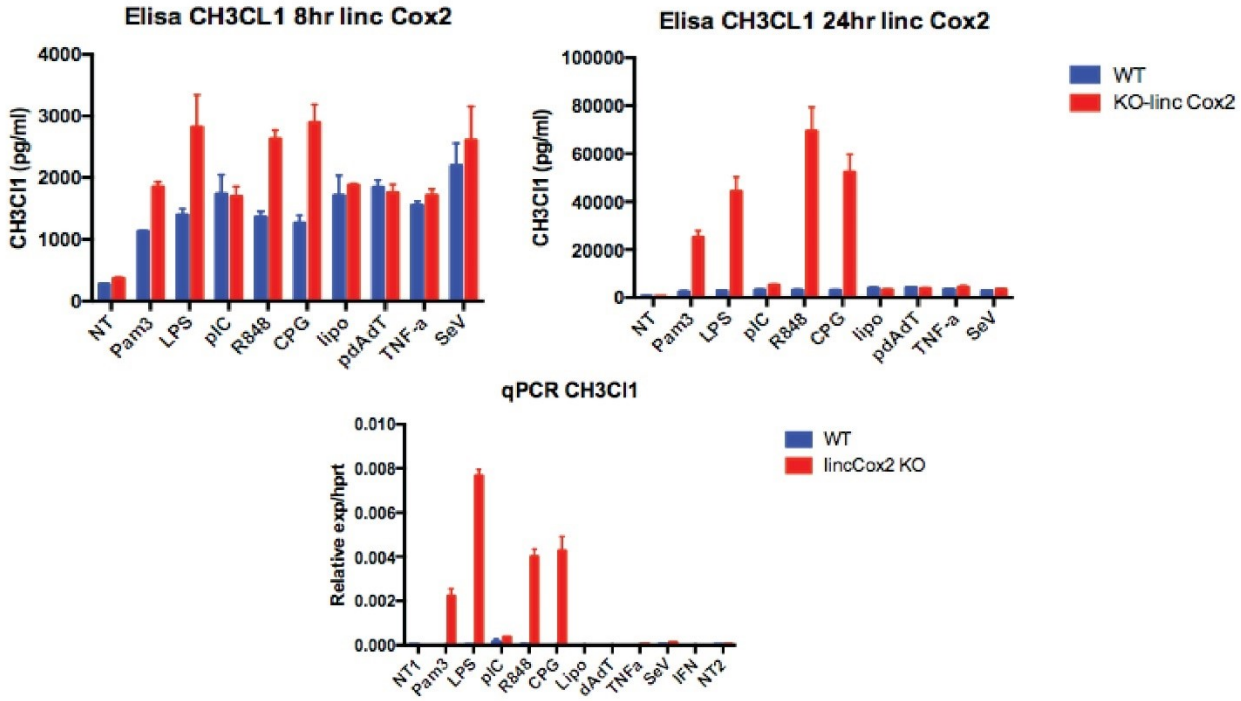


Figure 43:

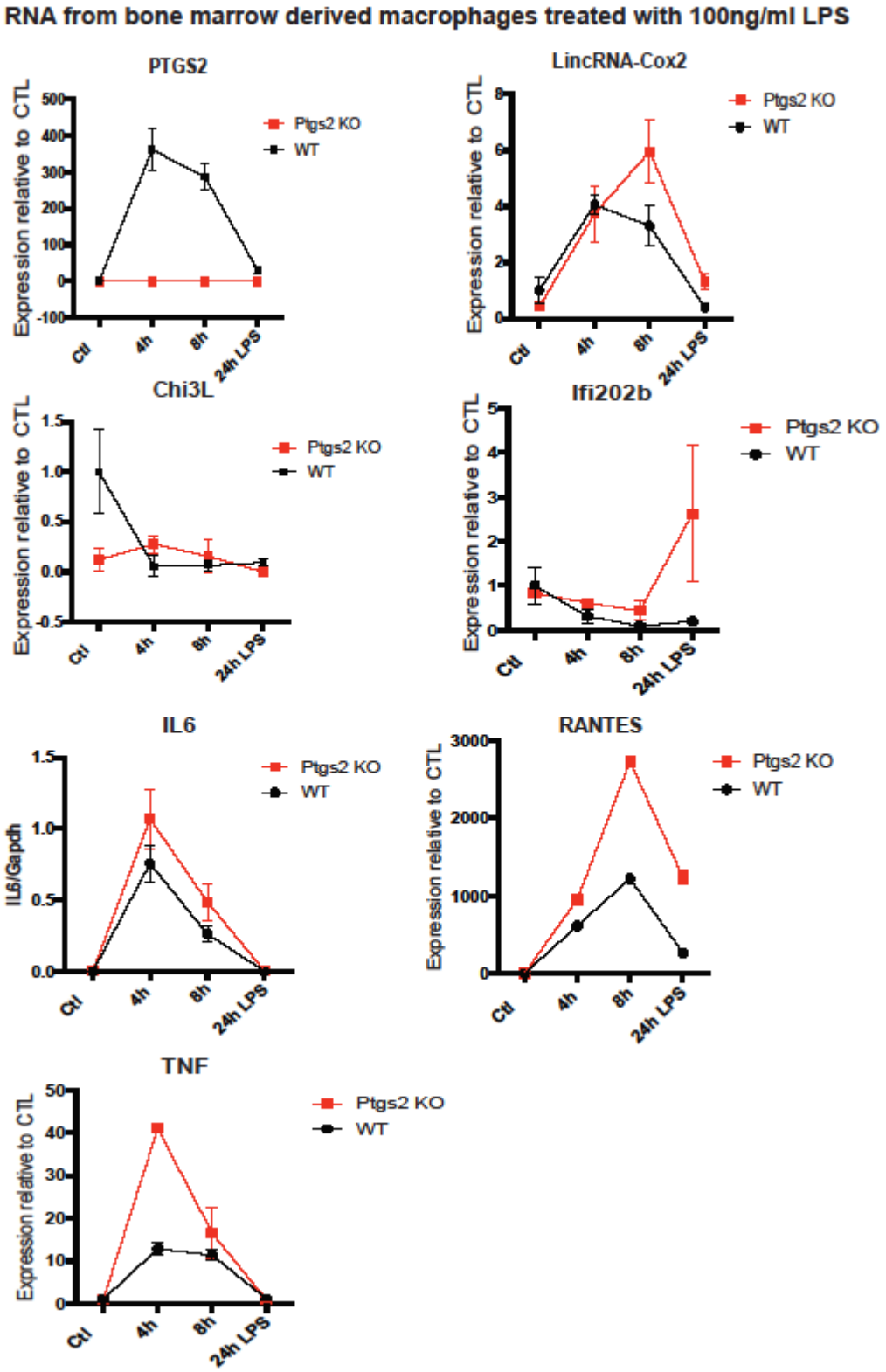
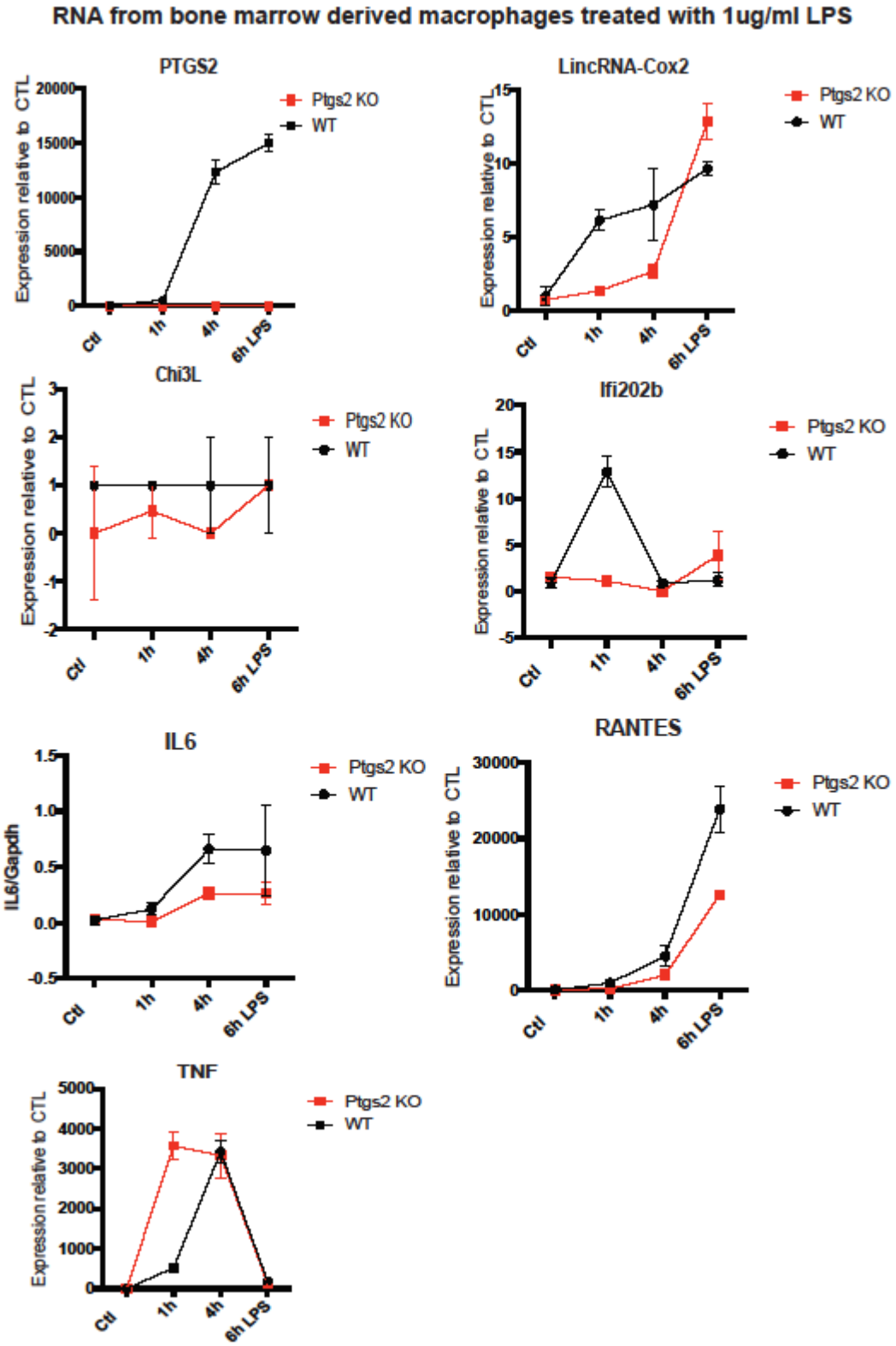

RNA from bone marrow derived macrophages treated with 100ng/ml LPS
PTGS2
LincRNA-Cox2
Chi3L
Ifi202b
IL6
RANTES
TNF
Ptgs2 KO
WT
Expression relative to CTL
IL6/Gapdh
Ctl
4h
8h
24h LPS

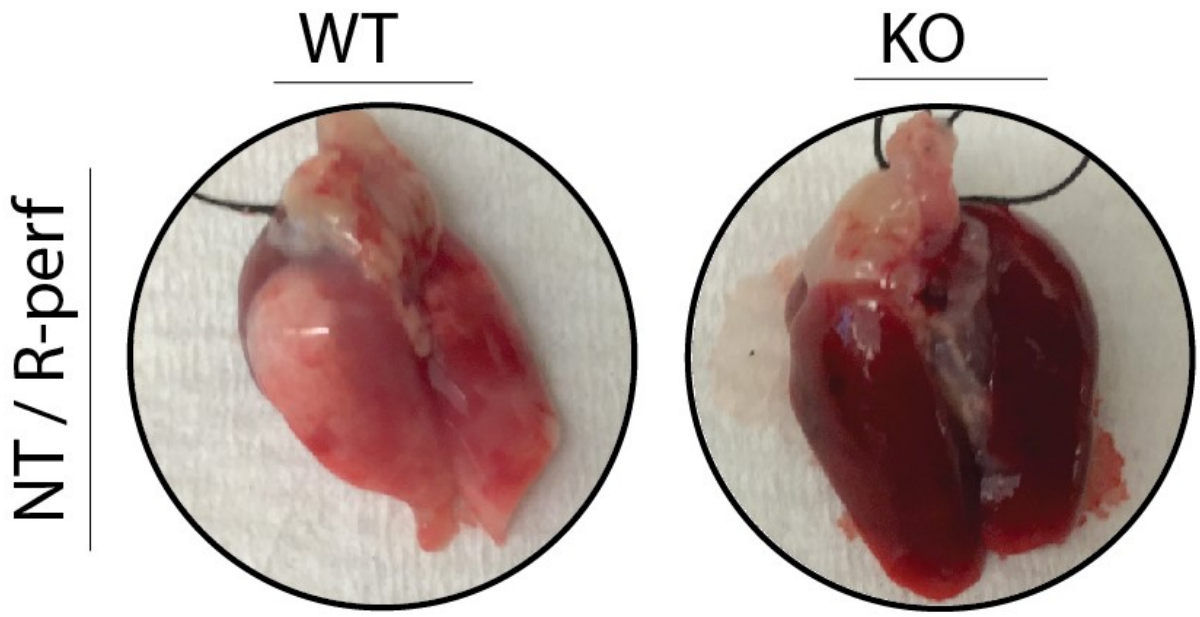
Figure 44:


RNA from bone marrow derived macrophages treated with 1ug/ml LPS
PTGS2
Expression relative to CTL
20000
15000
10000
5000
0
Ctl
1h
4h
6h LPS
Ptgs2 KO
WT
LincRNA-Cox2
Expression relative to CTL
15
10
5
0
Ctl
1h
4h
6h LPS
Ptgs2 KO
WT
Chi3L
Expression relative to CTL
3
2
1
0
-1
-2
Ctl
1h
4h
6h LPS
Ptgs2 KO
WT
Ifi202b
Expression relative to CTL
20
15
10
5
0
-5
Ctl
1h
4h
6h LPS
Ptgs2 KO
WT
IL6
IL6/Gapdh
1.5
1.0
0.5
0.0
-0.5
Ctl
1h
4h
6h LPS
Ptgs2 KO
WT
RANTES
Expression relative to CTL
30000
20000
10000
0
Ctl
1h
4h
6h LPS
Ptgs2 KO
WT
TNF
Expression relative to CTL
5000
4000
3000
2000
1000
0
Ctl
1h
4h
6h LPS
Ptgs2 KO
WT

4.9 – Future directions for *linc-Cox2*

Recent papers have identified *Chi3l1* protein as an important player in the mammalian response to acute injury in the lungs, and that it functions by suppressing injury and promoting fibrosis as a wound-healing mechanism. We have preliminary data to suggest that *linc-Cox2* KO mice develop higher rates of acute lung hemorrhaging (Fig. 46) than their WT counterparts, though these data are still low in number (n=2) and so improved statistics are required before moving forward with any kind of model.

Figure 45:

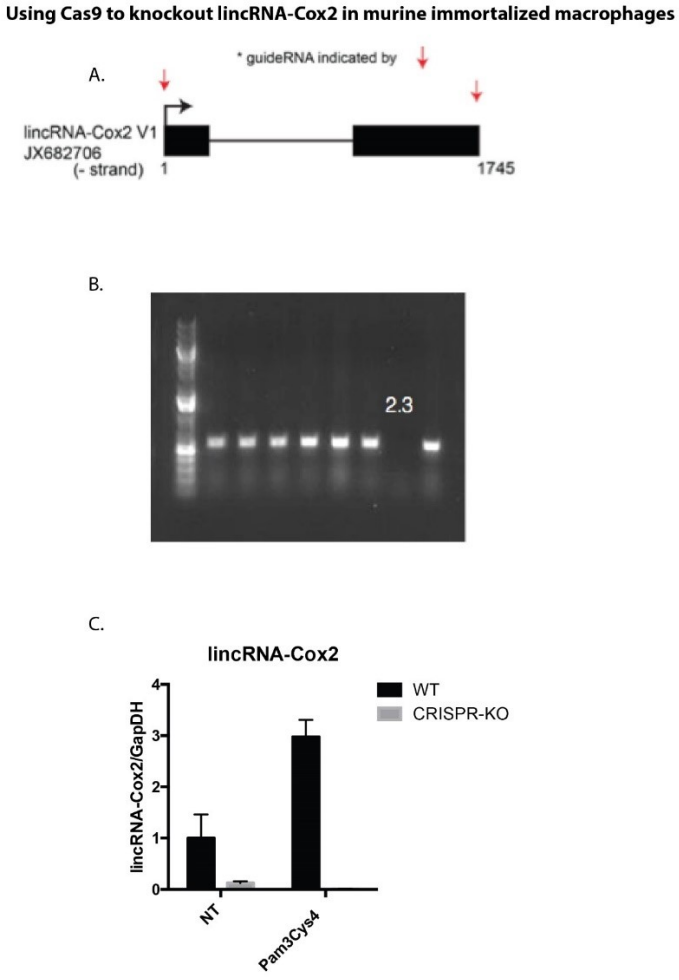


Since YKL-40 has been associated with the proliferation of lung fibrosis, we are also conducting experiments using the antibiotic bleomycin. Intratracheal administration of bleomycin induces acute lung damage and fibrosis in mice, and is the standard for lung fibrosis studies. We are interested in identifying whether *linc-Cox2* KO mice are conferred

resistance to the development of lung fibrosis due to the hyper-expression of *ChI3l1* in the absence of the lincRNA locus.

Furthermore, we are in the process of validating the genetic perturbations observed in our lacZ KO mice by reconstituting novel *linc-Cox2* KO macrophage lines using CRISPR-Cas9 (23, 24) to excise the locus. Preliminary results indicate that we have successfully created these knockout lines, and experiments are carrying forth rapidly (Fig. 46).

Figure 46:



Ultimately, we aim to identify the molecular mechanism by which *linc-Cox2* regulates *Chi3l1* and *Ptgs2*. Based on the observed genetic changes to the expression of these genes following ablation of the *linc-Cox2* locus, it is plausible that the lincRNA serves as a local DNA enhancer for the neighboring *Ptgs2* protein-coding gene, and that the transcribed RNA serves as a repressor of *Chi3l1* transcription. Identifying the protein binding partners of *linc-Cox2* is a sensible first step in this path toward molecular characterization. Complete characterization of this locus could increase our understanding of the role of noncoding genes in inflammatory response, help us discern between enhancer-like and truly functional RNA molecules, and provide insight into the molecular regulation of diseases like pulmonary fibrosis and (potentially) asthma. The *linc-Cox2* noncoding RNA is particularly interesting in that its effects on *Chi3l1 are* independent of *Ptgs2*, a known immune response-related protein, and it will be remarkable to conclude our work in characterizing this gene *in vivo*.

4.10 – Author Contributions

Stephen Liapis (SL), Susan Carpenter (SC), and Roland Elling (RE) contributed equally to this work. SL and SC performed initial characterization of the effects of infection on *linc-Cox2* expression. SL, SC, and RE each contributed to generation and immortalization of BMDM and dendritic cell populations. ELISAs were performed primarily by RE, with assistance by SL. qPCR was performed primarily by SL, with assistance by RE. SL was responsible for harvesting RNA and generation of Illumina sequencing libraries for brain, lung, and primary macrophage tissues. RNA sequencing was performed by SL with the assistance of Abigail Groff (AG). Lung histology was

performed by SL with the oversight of principle investigator Carla Kim. SL performed all lacZ staining. SL managed the colony of *linc-Cox2* mutant mice. Deletion of *linc-Cox2* by CRISPR-Cas9 was designed by the collective author group, and performed in the laboratory of Susan Carpenter. Overall design of project objectives, experiments, and goals was done by SL, SC, and RE, with the mentorship of principle investigators John Rinn and Kate Fitzgerald. This work has (at the time of writing this thesis) not been published in a peer reviewed journal, but the collective group of authors fully anticipates this work to result in a journal article in the near future. Writing of this chapter was done entirely by SL, with suggestions for revision made by John Rinn.

4.11 - Methods and Materials

Generation of bone marrow derived macrophages (BMDMs)

BMDMs were generated from bone marrow cells harvested from tibiae and femora of 6-8 week old mice by flushing the bone marrow space with cell culture media using a syringe with a 27G needle. Cells were cultured in D-MEM with 10% fetal bovine serum supplemented with penicillin/streptomycin or ciprofloxacin. Primary BMDM media was supplemented with 30% L929 supernatants in addition and the cells were used for experiments 6-9 days after differentiation. J2 virus was used on day 3/4 after isolation of bone marrow cells to establish transformed BMDM cell lines.

In vitro macrophage stimulations

Bone marrow cells were stimulated with Toll-like receptor (TLR) ligands for the indicated time points using the following concentrations: Lipopolysaccharide (LPS) 100ng/ml

(TLR4), Pam_3CSK_4 100ng/ml (TLR2/1), Pam_2CSK_4 100ng/ml (TLR2/6), Poly(I:C) 25 µg/ml (TLR3), R848 1µg/ml (TLR7&8). For RNA and protein isolation, 1-2x10^6 cells were seeded in a 12-well format, for cytokine measurement 1-2x10^5 cells were plated in 96-well plates. Bacterial stimulations with Salmonella (MOI XX), E.coli or Listeria have been performed as previously described.Sendai virus was used at a concentration of 200 hemagglutinanting units/ml.

Antibodies:

We used *Ptgs2* polyclonal antibody aa570-598 (Cayman Chemical). https://www.caymanchem.com/product/160106/

ELISAs

Cytokines have been measured by commercial sandwich ELISA kits for IL-6 (ebioscience), CCL5/Rantes (R&D) and Chitinase3like-1 (R&D) according to manufacturer´s instructions.

4.12 – References

1. Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., . . . Lander, E. S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature, 458*(7235), 223-227. doi:10.1038/nature07672

2. Banchereau, J., & Steinman, R. M. (1998). Dendritic cells and the control of immunity. *Nature, 392*(6673), 245-252. doi:10.1038/32588

3. Akira, S., & Takeda, K. (2004). Toll-like receptor signalling. *Nature Reviews Immunology, 4*(7), 499-511. doi:10.1038/nri1391

4. Germain, R. N. (1994). MHC-DEPENDENT ANTIGEN-PROCESSING AND PEPTIDE PRESENTATION - PROVIDING LIGANDS FOR T-LYMPHOCYTE ACTIVATION. *Cell, 76*(2), 287-299. doi:10.1016/0092-8674(94)90336-0

5. Iwasaki, A., & Medzhitov, R. (2010). Regulation of Adaptive Immunity by the Innate Immune System. *Science, 327*(5963), 291-295. doi:10.1126/science.1183021

6. Arnaout, M. A. (1990). STRUCTURE AND FUNCTION OF THE LEUKOCYTE ADHESION MOLECULES CD11 CD18. *Blood, 75*(5), 1037-1050.

7. Gahmberg, C. G. (1997). Leukocyte adhesion: CD11/CD18 integrins and intercellular adhesion molecules. *Current Opinion in Cell Biology, 9*(5), 643-650. doi:10.1016/s0955-0674(97)80117-2

8. Hoshino, K., Takeuchi, O., Kawai, T., Sanjo, H., Ogawa, T., Takeda, Y., . . . Akira, S. (1999). Cutting edge: Toll-like receptor 4 (TLR4)-deficient mice are hyporesponsive to lipopolysaccharide: Evidence for TLR4 as the Lps gene product. *Journal of Immunology, 162*(7), 3749-3752.

9. Raetz, C. R. H., & Whitfield, C. (2002). Lipopolysaccharide endotoxins. *Annual Review of Biochemistry, 71*, 635-700. doi:10.1146/annurev.biochem.71.110601.135414

10. Baldwin, A. S. (1996). The NF-kappa B and I kappa B proteins: New discoveries and insights. *Annual Review of Immunology, 14*, 649-683. doi:10.1146/annurev.immunol.14.1.649

11. Ghosh, S., May, M. J., & Kopp, E. B. (1998). NF-kappa B and rel proteins: Evolutionarily conserved mediators of immune responses. *Annual Review of Immunology, 16*, 225-260. doi:10.1146/annurev.immunol.16.1.225

12. Carpenter, S., Aiello, D., Atianand, M. K., Ricci, E. P., Gandhi, P., Hall, L. L., . . . Fitzgerald, K. A. (2013). A Long Noncoding RNA Mediates Both Activation and Repression of Immune Response Genes. *Science, 341*(6147), 789-792. doi:10.1126/science.1240925

13. Kosaka, T., Miyata, A., Ihara, H., Hara, S., Sugimoto, T., Takeda, O., . . . Tanabe, T. (1994). CHARACTERIZATION OF THE HUMAN GENE (PTGS2) ENCODING PROSTAGLANDIN-ENDOPEROXIDE SYNTHASE-2. *European Journal of Biochemistry, 221*(3), 889-897. doi:10.1111/j.1432-1033.1994.tb18804.x

14. Sauvageau, M., Goff, L.A., Lodato, S., Bonev, B., Groff, A.F., Gerhardinger, C., Sanchez-Gomez, D.B., Hacisuleyman, E., Li, E., Spence, M., Liapis, S.C., et al. (2013). Multiple knockout mouse models reveal lincRNAs are required for life and brain development. Elife 2.

15. Lombardi, V., Van Overtvelt, L., Horiot, S., Moussu, H., Chabre, H., Louise, A., . . . Moingeon, P. (2008). Toll-like receptor 2 agonist Pam3CSK4 enhances the induction of antigen-specific tolerance via the sublingual route.*Clinical and Experimental Allergy, 38*(11), 1819-1829. doi:10.1111/j.1365-2222.2008.03056.x

16. Halliday, A., Turner, J. D., Guimaraes, A., Bates, P. A., & Taylor, M. J. (2016). The TLR2/6 ligand PAM2CSK4 is a Th2 polarizing adjuvant in Leishmania major and Brugia malayi murine vaccine models. *Parasites & Vectors, 9*. doi:10.1186/s13071

17. Ueta, M., Hamuro, J., Kiyono, H., & Kinoshita, S. (2005). Triggering of TLR3 by polyI : C in human corneal epithelial cells to induce inflammatory cytokines. *Biochemical and Biophysical Research Communications, 331*(1), 285-294. doi:10.1016/j.bbrc.2005.02.196

18. Yamamoto, M., Sato, S., Hemmi, H., Hoshino, K., Kaisho, T., Sanjo, H., . . . Akira, S. (2003). Role of adaptor TRIF in the MyD88-independent toll-like receptor signaling pathway. *Science, 301*(5633), 640-643. doi:10.1126/science.1087262

19. Ørom, U.A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., et al. (2010). Long noncoding RNAs with enhancer-like function in human cells. Cell 143, 46–58 doi: 10.1016/j.cell.2010.09.001.

20. Jaenisch, R., & Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics, 33*, 245-254. doi:10.1038/ng1089

21. Ober, C., Tan, Z., Sun, Y., Possick, J. D., Pan, L., Nicolae, R., . . . Chupp, G. L. (2008). Effect of variation in CHI3L1 on serum YKL-40 level, risk of asthma, and lung function. *New England Journal of Medicine, 358*(16), 1682-1691. doi:10.1056/NEJMoa0708801

22. Engvall, E., & Perlmann, P. (1972). ENZYME-LINKED IMMUNOSORBENT ASSAY, ELISA .3. QUANTITATION OF SPECIFIC ANTIBODIES BY ENZYME-LABELED ANTI-IMMUNOGLOBULIN IN ANTIGEN-COATED TUBES. *Journal of Immunology, 109*(1), 129-&.

23. Hsu, P. D., Lander, E. S., & Zhang, F. (2014). Development and Applications of CRISPR-Cas9 for Genome Engineering. *Cell, 157*(6), 1262-1278. doi:10.1016/j.cell.2014.05.010

24. Mali, P., Yang, L. H., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., . . . Church, G. M. (2013). RNA-Guided Human Genome Engineering via Cas9. *Science, 339*(6121), 823-826. doi:10.1126/science.1232033

25. Blasi, E., Radzioch, D., Merletti, L., & Varesio, L. (1989). GENERATION OF MACROPHAGE CELL-LINE FROM FRESH BONE-MARROW CELLS WITH A MYC/RAF RECOMBINANT RETROVIRUS. *Cancer Biochemistry Biophysics, 10*(4), 303-317.

Chapter 5 – Conclusions and Future Perspectives

5.1 – lncRNA screens have identified a curated set of candidates for in-depth functional characterization

The widespread utilization of increasingly powerful sequencing technologies, most notably RNA sequencing and chromatin immunoprecipitation DNA sequencing, has allowed biologists to survey the genomic landscape at an unprecedented scale. In doing so, we have begun to reveal previously unappreciated complexities and nuances of the genome. One area of study that has particularly benefited from this advancement in methodology is that of noncoding RNA. Screening the mammalian genome for hallmarks of active transcription (K4-K36 methylation signatures) resulted in the realization that thousands of long noncoding RNAs (lncRNAs) are transcribed from previously unannotated regions of the genome. This finding, coupled with the discoveries of other noncoding RNA species (tRNAs, rRNAs, miRNAs, etc), demonstrated that the central dogma of molecular biology was indeed an outdated paradigm, and that RNA probably functions in a much more versatile role than anyone expected.

Even with thousands of novel loci being identified by the presence of local K4-K36 domains, however, the functional relevance of these noncoding RNAs remains in question. Several techniques have been developed that leverage the capabilities of bioinformatics and biochemistry in order to curate this list. Weeding out those loci whose transcripts could potentially code for proteins via primary sequence (BLASTX, PhyloCSF) or ribosome-occupancy (ribosome profiling) analyses was a starting point for many, including our lab. In many cases, these results showed that loci annotated as lincRNAs

could still contain open reading frames (ORFs) that translate for peptides, and so this is not an end-all marker of whether a transcript functions as a lincRNA. Utilizing fluorescence *in situ* hybridization (FISH) allowed scientists to query the spatial distribution of transcripts within a cell, and gave credence to those that remained exclusively nuclear as translation happens in the cytoplasmic compartment of the cell. Further investigation into the phylogenetic conservation of lincRNA sequences identified loci with higher degrees of conservation in their exons than in introns and surrounding, non-transcribed, sequence. Conservation of primary sequence is one predictor of gene function, and many lincRNAs exhibited some higher degree of conservation relative to the background. Interestingly, lincRNAs are not conserved to the extent of mRNAs, which some have said marks these transcripts as nonfunctional, while others have suggested lower conservation could mean lincRNAs function through secondary or tertiary structure, rather than primary sequence, through association with protein complexes such as Polycomb (PRC) and Trithorax (TRX) group proteins.

5.2 – Further in-depth functional characterization of lincRNAs is needed to advance the field

While the number of studies meant to identify new noncoding loci have been legion, the effort to mechanistically characterize and categorize of lincRNAs into functional "families" is still in a fledgling state. Perhaps still the best-described example of a long noncoding RNA is the lncRNA *Xist*, described in detail in Chapter 1.2. Many labs have worked to characterize this locus, and more than 25 years after its discovery there is still much to be learned including the mechanism by which *Xist* spreads out along its

originating chromosome *in cis*, and how certain genes can escape from X-chromosome silencing, among other questions. One resounding lesson has been learned from this molecule, however, which is that RNA is a functionally diverse set of molecules that warrants further study beyond its ability to code for proteins.

There has been a surge in comprehensive analysis of lincRNA function in recent years. Very recently our lab published a study describing one lincRNA, known as *Firre* (functional intergenic repeating RNA element). *Firre* is expressed from the X-chromosome, escaping X-inactivation by *Xist*, and spreads in *trans* to specific genomic targets. Localization of *Firre* throughout the nucleus was examined by single-molecule FISH, which showed punctate loci for both the sites of transcription and binding. Deletion of *Firre* results in an aberration of pluripotency gene expression, and GSEA suggests it plays a role in the differentiation of adipose tissue. This is a beautiful example of a thoroughly characterized lincRNA, especially when it is considered that *Firre* was first discovered in a screen for genes involved in adipogenesis.

Other studies, such as those with *lincRNA-Neat1* (see Chapter 2 Reference #49), paint a similar picture of lincRNA function. *Neat1* is of particular interest, due to the physiological defects found in its absence. This lincRNA has been found to be involved in paraspeckle formation in the nucleus, and is exclusively localized to this element of the nuclear body following transcription and processing. Previous cell-based studies from as early as 2007 indicated not only their association, but also that this lincRNA is required for formation of paraspeckle components into a cohesive body. The physiological repercussions for loss of function at the *Neat1* locus, however, remained a mystery until late 2014. The same group that found *Neat1* is required for paraspeckle formation also

identified that *Neat1* (and, therefore, paraspeckles) are not required for the development of viable, fertile offspring. In a twist of events that is surprisingly similar to the story we present for lincRNA-Tug1, they noticed that only half of *Neat1* knockout females produced a viable litter during the course of maintenance breedings. Further investigation revealed that *Neat1* KO females stochastically develop ovarian defects resulting in decreased serum progesterone levels, impaired functionality of the corpus luteum, and ultimately failure to impregnate during normal copulation. How *Neat1* induces these physiological defects remains to be seen, but the authors demonstrated the validity of using an animal model knockout to characterize the functionality of a lincRNA in great detail.

It is also important to remember that these loci were discovered by and large through bioinformatics, and that even though curation efforts tried to retain only those candidates whose protein coding potential was exceedingly low, some putative lincRNA loci might function through protein coding mechanisms. One particular example is the peptide MLN, which was discovered within a long noncoding RNA. Anderson et al. demonstrated that the MLN peptide is the component of the locus that functions in skeletal muscle uptake of calcium. Only through rigorous characterization of existing candidate lincRNAs can the field put to rest questions of whether these molecules possess functional importance.

5.3 – lincRNAs as regulators of gene expression

As increasing number of lincRNAs are studied in the context of cell culture and *in vivo* models, functional roles attributed to these molecules have emerged. Several papers have emerged discussing the role of long noncoding RNAs in regulating gene expression, and the prevailing view is that is how these molecules function. *Xist* is a great example of this: as we have discussed already in the introduction to this thesis, the lncRNA is transcribed, associates with proteins from the Polycomb repressive complex (PRC2), and recruits these proteins in *cis* to the chromatin of the X-chromosome. This is a clear example of what many scientists have found, that lincRNAs function through protein partners to regulate gene expression. Regulating expression can, of course, occur in several ways: ribonucleoprotein (RNP) complexes can increase transcription of a locus through activation or upregulation, and inversely they can decrease transcription of a locus through repression or downregulation. Several studies have shown that lincRNAs can achieve these ends by targeting chromatin modifying RNPs to their target loci (for *Xist* the target locus is the entire X-chromosome while, for other lincRNAs, the target is a specific gene, such as *Rnf185* for *linc-Tug1* or *Chl3l1* for *linc-Cox2)*.

Taking note of this recurring theme, scientists have endeavored to identify which (if any) lincRNAs form complexes with proteins. Not only does this expand our knowledge of how lincRNAs function as a class, but it also provides clues as to the specific functions that might be attributed to individual lincRNAs as well. Examples of other lincRNAs that have been shown to form gene-regulating RNPs include *HOTAIR*, which associates with PRC2 and represses gene expression at hundreds of loci throughout the mouse and human genome. The plant lncRNA *COLDAIR* (cold-inducible intronic lncRNA) is

expressed in response to temperature changes and represses expression of flowering genes by recruitment of PRC2 to those loci in a process known as vernalization. *Tug1*, which we have discussed at great length in Chapter 3, was previously found by early members of our lab to associate with PRC2 in three mouse and human cell lines (hLF, hFF, and HeLa). This study employed RNA immunoprecipitation (RIP) utilizing antibodies against components of the PRC2 component proteins SUZ12 and EZH2, allowing the authors to cast a broad net for lincRNAs that associate with this repressive complex. Many lincRNAs were found to associate with PRC2, and some of them in interesting patterns: *Xist* only associates with PRC2 in female hLFs by nature of its expression, while *Tug1* was found to form associations strongly in all three cell lines. *Fendrr*, interestingly enough, associates most strongly in lung fibroblasts – possibly indicating its underlying role in the physiology we observed in Chapter 2.

While studies have identified many lincRNAs that associate with repressive protein complexes, a few examples have emerged in which the RNA in question upregulates target gene expression instead. Perhaps the best example of this phenomenon to date is the lncRNA *HOTTIP* (HOXA transcript at the distal tip). The gene encoding *HOTTIP* resides at the 5' end of the *HOXA* locus, and has been shown to activate several of the 5' *HOXA* genes. To do so, it associates with WDR5, a protein component of the trithorax-group protein (TRX) complex. Recruitment of TRX proteins to target genes induces trimethylation of H3K4 (H3K4me3), a marker of active gene expression. Consistent with this observation, *HOTTIP* loss of function by shRNA knockdown results in downregulation of *HOXA* genes and results in aberrant wing and leg development in chickens.

The manner in which lincRNAs associate with specific protein complexes remains to be seen, as do the ways in which they subsequently guide their protein partners to specific chromatin targets in the genome. This concept of a "guide lincRNA" can be broken down to address (1) *cis*-acting and (2) *trans*-acting RNAs. *Cis*-acting lincRNAs, such as *Xist*, have been thought to guide associated proteins (like PRC2) to their neighboring genes by simple diffusion, with the lincRNA spreading cotranscriptionally and bringing its proteins along for the ride. This explanation does not make complete sense, however, when you consider the extent to which *Xist* spreads – over 2000 genes encompassed by roughly 153 megabases (MB, 10^6 bases) in humans. This break in logic has been accounted for by the process of chromatin looping, where chromatin folds in upon itself in 3-dimensional space within the nucleus and could therefore (so the argument goes) allow for diffusion of *Xist* to affect the entire chromosome when packed into a small radius. One could imagine that, if this were the case, other chromosomes could be impacted by proximity to the X-chromosome in 3D space within the nucleus and, as such, undergo catalysis of H3K27me3 by *Xist,* resulting in gene silencing at *trans* sites. As this has not been observed, it is unlikely that *Xist* spreads along its originating chromosome by simple diffusion, but rather through some still-to-be-elucidated mechanism of targeted spreading. *Trans*-acting RNAs can best be described through the example of *lincRNA-Firre*, described above. Hacisuleyman et al. (2014) performed RAP experiments, as we did with *Tug1*, in mouse embryonic stem cells (mESCs) and mouse adipocytes by labeling the lincRNA with biotinylated complimentary probes and subjugating the proteins pulled down in this way to mass spectrometry (MS). These pulldowns identified several candidate proteins that associate with *Firre*, with the top-ranked candidate being hnRNPU (heterogeneous nuclear ribonuclear protein U). hnRNPU is a previously known

RNA binding protein that has been shown to bind premature mRNAs and shuttle them around the nucleus during processing. After validating the interaction through hnRNPU pulldowns (which showed that *Firre* came along for the ride), the group knocked down hnRNPU using siRNAs in mESCs and two human cell lines. Silencing hnRNPU expression ablated the punctate foci observed in *Firre* FISH (described above), indicating that this protein is required for the *trans*-localization observed for lincRNA-Firre. As the lincRNA was left intact in these experiments, it seems that hnRNPU is the element that performs a binding role at these loci, leaving us with the question of what it is that *lincRNA-Firre* actually does in this relationship. Perhaps it is still the agent of genomic targeting, rather than binding, or perhaps it serves as a scaffold for still other proteins that are brought to *trans* chromosomal sites in the presence of hnRNPU. Further work, as well as *in vivo* studies in a *lincRNA-Firre* animal model, will aim to address these questions.

5.4 – Molecular Mechanism of lincRNAs Still Unclear

The preceding section should highlight that, although great strides have been made in characterizing a handful of lincRNA loci in recent years, we are still largely unsure of how these RNAs function at the molecular level. Do they recruit proteins to target loci through complimentary binding to single-stranded DNA? Do they rely on protein partners, such as hnRNPU, for targeting and binding to *trans*-sites? Can *cis*-acting lincRNAs, such as *Xist* or *Tug1*, specifically target neighboring genes or do they spread

through simple diffusion? Experiments have begun whittling away at these questions, but answers that resound with surety have yet to come.

Despite these uncertainties, attention to the molecular function of lincRNAs is gaining steam. Loss of function studies revealed a novel class of long noncoding RNAs, known as circular RNAs (circRNAs). circRNAs are single-stranded RNA molecules, which, as their name indicates, form a covalently closed continuous loop with the 5' and 3' ends joining together. These molecules are resistant to exonuclease digestion, as they do not present an exposed end, and are thought to function as regulators of gene expression. Unlike their lincRNA brethren, however, circRNAs do not regulate target genes through recruitment of protein complexes to upregulate or downregulate their expression; rather, they act as complementary sponges to titrate micro RNAs (miRNAs) that would otherwise bind to and catalyze the degradation of mRNAs destined for translation. One example of these "repressor of repressors" is CDR1as. CDR1 is encoded antisense to the human gene *CDR1*. Its target is the miRNA miR-7. miR-7, as with all miRNAs, is a RNA molecule 21 nucleotides in length that represses mRNA translation through complementary binding and cleavage of the mRNA (via the Argonaute protein AGO2, which we won't go further into). CDR1as possesses over 60 miR-7 binding sites, allowing it to soak up miR-7 and prevent it from performing its role on target mRNAs. This example should resonate with the lincRNA field because, through a combination of biochemical and bioinformatic techniques, the molecular mechanism of a vexing class of ncRNAs was uncovered. Similar approaches should be employed to a greater extent in the effort to better understand lincRNAs.

This thesis, and unmentioned experiments in the pipeline from our lab and others, aims to do just that. RAP analysis of the *Tug1* transcript identified unidirectional spreading in a 5'-3' direction that argues against the diffusional spreading model for *cis* targeting and offers evidence in favor of a more targeted localization process. Cis plot analysis on RNA sequencing data allows for the spatial distribution of differentially expressed genes at the primary sequence level, and provides insight into whether the lincRNA locus in question is either (a) an activating lincRNA (e.g.. HOTTIP), or (b) a repressive lincRNA (e.g. HOTAIR, Xist, COLDAIR, etc). Biochemical approaches such as ribosome profiling can be complemented with bioinformatics querying techniques like BLASTX and PhyloCSF to identify protein coding potential of genes, and lend credence to the notion that these molecules function through ways other than enigmatically coding for small peptides.

5.5 – DNA vs. RNA vs. Act of Transcription?

One question that continues to rear its head in lincRNA functional studies is whether the locus in question functions as a DNA element, or an RNA element, or neither of those is of importance and it is the act of transcription that is the functional component of the locus. Indeed, our KO strains described above removes the DNA element of the lincRNA loci as well as the putative lncRNA, and so parsing the contributions of each remains a challenge. Our models do, however, rule out the possibility that the act of transcription is what's important, as the inserted lacZ cassette is still transcribed in our mice. Other groups, such as with Grote et al. (2013)'s study of the *Fendrr* locus, have worked to discern between the DNA element (the primary DNA sequence of the gene) and

the act of transcription through insertion of a strong polyadenylation (pA) sequence downstream of the promoter. “Strong” polyadenylation signals, such as those utilized in the aforementioned paper, utilize triple tandem repeat polyadenylation signals (3xpA) to keep the primary DNA sequence of a locus intact while preventing the extension phase of transcription by RNA polymerase II. While our group and others have employed one or another of these approaches, future characterization of lincRNAs will likely use a combination of genetic deletion, reporter replacement of a gene, and transcriptional stop sequence to comprehensively analyze the functional component of a locus.

Alternative strategies aim to understand whether the RNA product of transcription is functional, regardless of the role of the DNA element in a physiological process. For example, our experiments with transient overexpression of the *Tug1* RNA (Chapter 3) attempt to do just that. To summarize, we recently cloned a transcript originating from the *Tug1* locus, and overexpressed this transcribed cDNA in both *Tug1* +/+ and -/- MESC and MEF cell lines. Expressing the gene from a transient plasmid in KO cells did not recover the expression profile (in terms of FPKM) of differentially expressed genes as expected if the RNA molecule was the functional unit. As we discussed in that Chapter, it is important to remember that both transient and lentiviral expression of a lincRNA locus happens at exogenous loci in the genome due to non-integration or random integration, respectively. As such, it is highly unlikely that introduction of an exogenous lincRNA will result in transcription occurring at the native locus, and so *cis*-acting lincRNAs cannot be functionally validated using this approach. Other, *trans*-acting RNAs (such as *Firre)* can be validated via overexpression experiments. Interestingly, new technologies development might make these studies easier to conduct and analyze. Preliminary

experiments with *cis*-acting *Xist* have demonstrated that targeted integration using CRISPR-Cas9 results in chromosome-wide silencing of autosomes into which the *Xist* gene is introduced. Increased penetration of this technology into overexpression studies could make identification of functional *cis*-acting RNAs much easier and more feasible, and in conjunction with polyA and reporter (lacZ cassette) models will enable scientists to understand the roles of RNA, DNA, and the act of transcription better than ever before.