UNIVERSITY OF CALIFORNIA SAN DIEGO

Enhancer screens to study enhancer regulation in development and evolution

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Biology

by

Benjamin Paul Song

Committee in charge:

Professor Emma K. Farley, Chair
Professor Christopher K. Glass
Professor Cornelis Murre
Professor James W. Posakony
Professor Deborah Yelon

2022

The Dissertation of Benjamin Paul Song is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

To my parents, Ying-Hsiu Su and Wei Song, for being my biggest fans and always believing I could do anything I put my mind to.

To my sister, Janet Song, for always letting me follow her lead.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my thesis advisor Professor Emma Farley. Emma's research drew me to UCSD and the fields of development and transcription, but her unwavering support and guidance and all the backup plans when plan A failed helped me get through these five years. I would also like to thank my committee members, Professors Christopher Glass, Cornelis Murre, James Posakony, and Deborah Yelon, for their feedback and guidance through this process.

I would like to thank all the members of the Farley lab for their day-to-day support. My biggest thanks to Michelle Ragsac, who has been my computational partner in all of my work. Without her, I would still be struggling to install packages on UCSC, let alone analyzing the data. I would also like to thank Jessica Grudzien, Fabian Lim, Granton Jindal, Genevieve Ryan, Hannah Finnegan, Sophia Le, Joe Solvason, and Krissie Tellez for their experimental and computational support and valuable feedback with my projects. Special thanks to Meng Zhu from the Tabin lab for her experimental help and discussions with the vertebrate MPRA project.

Finally, I would like to thank the community around me that has supported during this period of my life. First, to my parents, Ying-Hsiu Su and Wei Song, who first tried to convince me to go to medical school, and then to Princeton for graduate school. But once my decision was made, they supported me with whole-heartedly, like how they fostered my love for science from a young age. Thanks to my sister, Janet Song, whose love for science is contagious and helped pave the way for me. She was constantly curious about my projects and always revised my applications and writings, even when she was busier than ever. Thanks to my friends here in San Diego, who made my life full and enriching. I never could have imagined trying so many new activities and experiences before I moved here. And finally, thanks to my best friend and

partner, Erin Schiksnis, for always being by my side. Seeing her smile was always enough to brighten my day, no matter what happened.

Chapter 1 contains material submitted to Cell Reports. Song BP, Ragsac MF, Tellez K, Jindal GA, Grudzien JL, Le SH, Farley EK. “Diverse logics and grammars encode notochord enhancers”. The dissertation author was the primary investigator and author of this paper. Special thanks to the Farley lab and Dennis Schifferl for helpful discussions. Special thanks to Janet H.T. Song for her critical reading of the manuscript. Special thanks to the UCSD IGM Genomics Center for their assistance with sequencing.

Chapter 2 contains unpublished material coauthored with Zhu, Meng and Solvason, Joseph J. The dissertation author was the primary author of this chapter. Special thanks to the Farley lab for helpful discussions. Special thanks to Timothy Sackton for help in designing the library of genomic elements for testing. Special thanks to Sophia Le for help with cloning the library vector. Special thanks to the UCSD IGM Genomics Center for their assistance with sequencing.

VITA

2016 Bachelor of Science in Biology, Massachusetts Institute of Technology

2022 Doctor of Philosophy in Biology, University of California San Diego

ABSTRACT OF THE DISSERTATION

Enhancer screens to study enhancer regulation in development and evolution

by

Benjamin Paul Song

Doctor of Philosophy in Biology

University of California San Diego, 2022

Professor Emma K. Farley, Chair

Enhancers are sequences in the genome that act as switches to turn on gene expression in the right time and place during development. Enhancers are regulated by binding of transcription factors to recruit transcriptional machinery. However, exactly how the sequence of an enhancer encodes this function is poorly understood. Reporter assays test enhancers for activity, and mutational approaches to evaluate important sequences within an enhancer help

us better understand enhancer regulation. However, testing of individual enhancers is slow and tedious process. Massively parallel reporter assays (MPRAs) can test thousands to millions of enhancers within a single experiment. In this dissertation, I discuss the use of two enhancer MPRAs to further our understanding of enhancer regulation in development and evolution. First, I performed an MPRA to study enhancer regulation in the developing notochord of *Ciona robusta* and discovered notochord logics and grammars, the interplay between transcription factor order, orientation, spacing, and binding affinity, important for driving notochord-specific expression. These enhancer logics and grammars show signatures of conservation across chordates. In Chapter 2, I developed enhancer MPRAs in the chicken limb bud, which is the first of its kind in developing vertebrate embryos. Using this method, I identified new enhancers active in the forelimb and hindlimb that could be further studied to understand how sequence changes impact enhancer activity. Overall, the methods I developed to test genomic regions in developing chordate and vertebrate embryos will enable unprecedented insight into how enhancers encode the instructions for development.

INTRODUCTION

Enhancers are genomic elements that act as switches to ensure the precise patterns of gene expression required for development (Levine, 2010). Enhancers regulate the timing, locations and levels of expression by binding of transcription factors (TFs) to sequences within the enhancer known as transcription factor binding sites (TFBSs) (Heinz et al., 2010; Liu and Posakony, 2012; Small et al., 1992; Spitz and Furlong, 2012; Swanson et al., 2010). This binding, along with protein-protein interactions, leads to recruitment of transcriptional machinery and activation of gene expression. While we understand that TFBSs regulate enhancers and mediate tissue-specific expression, we have limited understanding of how the sequence of an enhancer encodes a particular expression pattern and what combinations of binding sites within enhancers are able to mediate enhancer activity. Given that the majority of variants associated with disease and phenotypic diversity lie within enhancers (Maurano et al., 2012; Tak and Farnham, 2015; Visel et al., 2009), it is critical that we understand how the underlying enhancer sequence encodes tissue-specific expression and what types of changes within an enhancer sequence can cause changes in expression, cellular identity and phenotypes.

Reporter assays are a tool to characterize enhancer activity. In this assay, enhancers, with a minimal promoter, drive expression of a reporter gene to measure the activity of an enhancer. Expression of the reporter can be evaluated by measuring the number of transcripts of the reporter (Wong and Medrano, 2005) or by evaluating protein expression of the reporter (GFP: Chalfie et al., 1994; lacZ: Schmidt et al., 1998). Reporter assays can also be used to identify important sequences or binding sites in enhancers by mutating nucleotides in an enhancer and comparing the mutated activity with the wild-type enhancer.

However, in most traditional reporter assays, measuring expression of reporters is often assayed individually, which is a slow and tedious process. Massively parallel reporter assays (MPRAs) combat this problem by using high-throughput sequencing to measure many different enhancers at the same time. This is often done by sequencing expression of a transcribed barcode associated with a unique enhancer (de Boer et al., 2020; Farley et al., 2015; King et al., 2020), or by sequencing transcription of the enhancer itself (Arnold et al., 2013). MPRAs have been used to measure tens to millions of sequences at a time, significantly increasing the number of enhancers tested in a single time.

To functionally test our understanding of the code of enhancer regulation, MPRAs are often performed on putative enhancers in the genome. These are found through a variety of methods. One way to identify enhancers is through the identification of putative TFBSs using DNA binding motifs, like position-weight matrices, as enhancers use TFs to drive gene expression (D'haeseleer, 2006). Most tools that use binding motifs, however, only use optimal binding sites, which has been shown to lead to loss of tissue-specificity (Crocker et al., 2015; Farley et al., 2015). Another method is to identify areas of open chromatin, as enhancers are often in these regions. There are a number of genome-wide, biochemical methods to assay for chromatin structure, like chromatin immunoprecipitation followed by sequencing (ChIP-Seq, Johnson et al., 2007) of histones, DNaseI hypersensitive sites sequencing (DNase-seq, Crawford et al., 2004; Sabo et al., 2004), assay for transposition-accessible chromatin using sequencing (ATAC-seq, Buenrostro et al., 2013), cleavage under targets and release using nuclease (CUT&RUN, Skene et al., 2018), Hi-C (Lieberman-Aiden et al., 2009), and others. All of these methods can be used to generate thousands of putative enhancers; however, the power of these genome-wide approaches to predict functional enhancers can be quite variable (Grossman et al., 2017; Halfon, 2019; King et al., 2020; Ryan and

Farley, 2020). In flies, combining high-affinity binding site clusters with histone marks can lead to strong predictive power (>90% success rate) (Berman et al., 2002; Nègre et al., 2011; Rebeiz et al., 2002). However, this same approach has failed with human genomic regions (28% success rate) (King et al., 2020). Thus, functional testing of putative enhancers is important to improve our understanding of what sequences encode an enhancer.

Testing of putative enhancers should assay for its full expression pattern. However, most MPRAs that are performed are done in cell culture or in single cell types (Friedman et al., 2021; King et al., 2020; Patwardhan et al., 2012). Additionally, enhancers often drive expression in multiple cell types. Cell culture or single cell types limit the full understanding of an enhancer's expression. Thus, functional testing of enhancers should be done within a developing embryo to identify the full picture of an enhancer's activity. To do this, our lab has developed enhancer MPRAs in whole embryos.

In this dissertation, I used MPRAs to test putative enhancers in two different systems to better understand how enhancers are regulated. In Chapter 1, I performed an MPRA on a library of 90 genomic elements from *Ciona robusta* to investigate the role of Zic and ETS in driving notochord expression. Surprisingly, only nine of these elements drove notochord expression. Of these, I found that one of the Zic and ETS enhancers is near an important notochord gene, *laminin alpha* (Veeman et al., 2008). The orientation of binding sites within this *laminin alpha* enhancer is critical for enhancer activity demonstrating the role of enhancer grammar, or the interplay between TFBS affinity, order, orientation and spacing. I found similar clusters of Zic and ETS sites within the introns of *laminin alpha-1* in both mouse and human, and all three of these regions contain the same 12bp spacing between the Zic and ETS hinting at the conservation of grammatical rules across chordates. I also identified novel FoxA and Bra sites in the BraS enhancer and showed

that these sites along with Zic and ETS were necessary and sufficient. Other known *Bra* enhancers within *Ciona* (Corbo et al., 1997) and vertebrates (Schifferl et al., 2021) also harbor this combination of TFs, suggesting that Zic, ETS, FoxA, and Bra is a common feature of *Bra* regulation in chordates. Overall, this study finds that grammar is a key component of functional enhancers with signatures of enhancer logic and grammar seen across chordates.

In Chapter 2, I develop a novel vertebrate MPRA in chicken limb buds. The chicken limb bud is an ideal system to investigate enhancer specificity, as enhancers can be electroporated into the limb buds to identify those that drive expression in the forelimb, hindlimb, or both. Using this new vertebrate MPRA, I performed an initial study to investigate differential activity of enhancers identified as conserved or accelerated in the developing chicken limb bud. This screen identifies many new limb enhancers, including enhancers that have highly conserved sequences across birds but different activities.

References

Arnold, C.D., Gerlach, D., Stelzer, C., Boryń, Ł.M., Rath, M., Stark, A., 2013. Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. Science 339, 1074–1077. https://doi.org/10.1126/science.1232542

Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., Eisen, M.B., 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. Proc. Natl. Acad. Sci. U.S.A. 99, 757–762. https://doi.org/10.1073/pnas.231608898

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., Greenleaf, W.J., 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 10, 1213–1218. https://doi.org/10.1038/nmeth.2688

Chalfie, M., Tu, Y., Euskirchen, G., Ward, W.W., Prasher, D.C., 1994. Green Fluorescent Protein as a Marker for Gene Expression. Science 263, 802–805. https://doi.org/10.1126/science.8303295

Corbo, J.C., Levine, M., Zeller, R.W., 1997. Characterization of a notochord-specific enhancer from the Brachyury promoter region of the ascidian, Ciona intestinalis. Development 124, 589–602. https://doi.org/10.1242/dev.124.3.589

Crawford, G.E., Holt, I.E., Mullikin, J.C., Tai, D., National Institutes of Health Intramural Sequencing Center‡, Green, E.D., Wolfsberg, T.G., Collins, F.S., 2004. Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. Proc. Natl. Acad. Sci. U.S.A. 101, 992–997. https://doi.org/10.1073/pnas.0307540100

Crocker, J., Abe, N., Rinaldi, L., McGregor, A.P., Frankel, N., Wang, S., Alsawadi, A., Valenti, P., Plaza, S., Payre, F., Mann, R.S., Stern, D.L., 2015. Low affinity binding site clusters confer hox specificity and regulatory robustness. Cell 160, 191–203. https://doi.org/10.1016/j.cell.2014.11.041

de Boer, C.G., Vaishnav, E.D., Sadeh, R., Abeyta, E.L., Friedman, N., Regev, A., 2020. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. Nat Biotechnol 38, 56–65. https://doi.org/10.1038/s41587-019-0315-8

D'haeseleer, P., 2006. What are DNA sequence motifs? Nat Biotechnol 24, 423–425. https://doi.org/10.1038/nbt0406-423

Farley, E.K., Olson, K.M., Zhang, W., Brandt, A.J., Rokhsar, D.S., Levine, M.S., 2015. Suboptimization of developmental enhancers. Science 350, 325–328. https://doi.org/10.1126/science.aac6948

Friedman, R.Z., Granas, D.M., Myers, C.A., Corbo, J.C., Cohen, B.A., White, M.A., 2021. Information content differentiates enhancers from silencers in mouse photoreceptors. eLife 10, e67403. https://doi.org/10.7554/eLife.67403

Grossman, S.R., Zhang, X., Wang, L., Engreitz, J., Melnikov, A., Rogov, P., Tewhey, R., Isakova, A., Deplancke, B., Bernstein, B.E., Mikkelsen, T.S., Lander, E.S., 2017. Systematic dissection of genomic features determining transcription factor binding and enhancer function. Proc. Natl. Acad. Sci. U.S.A. 114. https://doi.org/10.1073/pnas.1621150114

Halfon, M.S., 2019. Studying Transcriptional Enhancers: The Founder Fallacy, Validation Creep, and Other Biases. Trends in Genetics 35, 93–103. https://doi.org/10.1016/j.tig.2018.11.004

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., Glass, C.K., 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38, 576–589. https://doi.org/10.1016/j.molcel.2010.05.004

Johnson, D.S., Mortazavi, A., Myers, R.M., Wold, B., 2007. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science 316, 1497–1502. https://doi.org/10.1126/science.1141319

King, D.M., Hong, C.K.Y., Shepherdson, J.L., Granas, D.M., Maricque, B.B., Cohen, B.A., 2020. Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in mouse embryonic stem cells. eLife 9, e41279. https://doi.org/10.7554/eLife.41279

Levine, M., 2010. Transcriptional Enhancers in Animal Development and Evolution. Current Biology 20, R754–R763. https://doi.org/10.1016/j.cub.2010.06.070

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A., Lander, E.S., Dekker, J., 2009. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science 326, 289–293. https://doi.org/10.1126/science.1181369

Liu, F., Posakony, J.W., 2012. Role of Architecture in the Function and Specificity of Two Notch-Regulated Transcriptional Enhancer Modules. PLoS Genet 8, e1002796. https://doi.org/10.1371/journal.pgen.1002796

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A.K., Canfield, T.K., Giste, E., Diegel, M., Bates, D., Hansen, R.S., Neph, S., Sabo, P.J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S.R., Kaul, R., Stamatoyannopoulos, J.A., 2012. Systematic localization of common disease-associated variation in regulatory DNA. Science 337, 1190–1195. https://doi.org/10.1126/science.1222794

Nègre, N., Brown, C.D., Ma, L., Bristow, C.A., Miller, S.W., Wagner, U., Kheradpour, P., Eaton, M.L., Loriaux, P., Sealfon, R., Li, Z., Ishii, H., Spokony, R.F., Chen, J., Hwang, L., Cheng, C., Auburn, R.P., Davis, M.B., Domanus, M., Shah, P.K., Morrison, C.A., Zieba, J., Suchy, S., Senderowicz, L., Victorsen, A., Bild, N.A., Grundstad, A.J., Hanley, D., MacAlpine, D.M., Mannervik, M., Venken, K., Bellen, H., White, R., Gerstein, M., Russell, S., Grossman, R.L., Ren, B., Posakony, J.W., Kellis, M., White, K.P., 2011. A cis-regulatory map of the Drosophila genome. Nature 471, 527–531. https://doi.org/10.1038/nature09990

Patwardhan, R.P., Hiatt, J.B., Witten, D.M., Kim, M.J., Smith, R.P., May, D., Lee, C., Andrie, J.M., Lee, S.-I., Cooper, G.M., Ahituv, N., Pennacchio, L.A., Shendure, J., 2012. Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol 30, 265–270. https://doi.org/10.1038/nbt.2136

Rebeiz, M., Reeves, N.L., Posakony, J.W., 2002. SCORE: A computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. Proc. Natl. Acad. Sci. U.S.A. 99, 9888–9893. https://doi.org/10.1073/pnas.152320899

Ryan, G.E., Farley, E.K., 2020. Functional genomic approaches to elucidate the role of enhancers during development. WIREs Systems Biology and Medicine 12, e1467. https://doi.org/10.1002/wsbm.1467

Sabo, P.J., Humbert, R., Hawrylycz, M., Wallace, J.C., Dorschner, M.O., McArthur, M., Stamatoyannopoulos, J.A., 2004. Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. Proc. Natl. Acad. Sci. U.S.A. 101, 4537–4542. https://doi.org/10.1073/pnas.0400678101

Schifferl, D., Scholze-Wittler, M., Wittler, L., Veenvliet, J.V., Koch, F., Herrmann, B.G., 2021. A 37 kb region upstream of *brachyury* comprising a notochord enhancer is essential for notochord and tail development. Development 148, dev200059. https://doi.org/10.1242/dev.200059

Schmidt, A., Tief, K., Foletti, A., Hunziker, A., Penna, D., Hummler, E., Beermann, F., 1998. lacZ Transgenic mice to monitor gene expression in embryo and adult. Brain Research Protocols 3, 54–60. https://doi.org/10.1016/S1385-299X(98)00021-X

Skene, P.J., Henikoff, J.G., Henikoff, S., 2018. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. Nat Protoc 13, 1006–1019. https://doi.org/10.1038/nprot.2018.015

Small, S., Blair, A., Levine, M., 1992. Regulation of even-skipped stripe 2 in the Drosophila embryo. EMBO J 11, 4047–4057. https://doi.org/10.1002/j.1460-2075.1992.tb05498.x

Spitz, F., Furlong, E.E.M., 2012. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet 13, 613–626. https://doi.org/10.1038/nrg3207

Swanson, C.I., Evans, N.C., Barolo, S., 2010. Structural Rules and Complex Regulatory Circuitry Constrain Expression of a Notch- and EGFR-Regulated Eye Enhancer. Developmental Cell 18, 359–370. https://doi.org/10.1016/j.devcel.2009.12.026

Tak, Y.G., Farnham, P.J., 2015. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. Epigenetics Chromatin 8, 57. https://doi.org/10.1186/s13072-015-0050-4

Veeman, M.T., Nakatani, Y., Hendrickson, C., Ericson, V., Lin, C., Smith, W.C., 2008. Chongmague reveals an essential role for laminin-mediated boundary formation in chordate convergence and extension movements. Development 135, 33–41. https://doi.org/10.1242/dev.010892

Visel, A., Rubin, E.M., Pennacchio, L.A., 2009. Genomic views of distant-acting enhancers. Nature 461, 199–205. https://doi.org/10.1038/nature08451

Wong, M.L., Medrano, J.F., 2005. Real-time PCR for mRNA quantitation. BioTechniques 39, 75–85. https://doi.org/10.2144/05391RV01

CHAPTER 1

ABSTRACT

The notochord is a key structure during chordate development. We have previously identified several enhancers regulated by Zic and ETS that encode notochord activity within the marine chordate *Ciona robusta (Ciona).* To better understand the role of Zic and ETS within notochord enhancers, we tested 90 genomic elements containing Zic and ETS sites for expression in developing *Ciona* embryos using a whole-embryo, massively parallel reporter assay. We discovered that 39/90 of the elements were active in developing embryos; however only 10% (9/90) were active within the notochord, indicating that more than just Zic and ETS sites are required for notochord expression. Further analysis revealed notochord enhancers were regulated by three groups of factors: (1) Zic and ETS, (2) Zic, ETS and Brachyury (Bra), and (3) Zic, ETS, Bra and FoxA. One of these notochord enhancers, regulated by Zic and ETS, is located upstream of *laminin alpha*, a gene critical for notochord development in both *Ciona* and vertebrates. Reversing the ETS sites in this enhancer greatly diminishes expression, indicating that enhancer grammar is critical for enhancer activity. Strikingly, we find clusters of Zic and ETS binding sites within the introns of mouse and human *laminin alpha-1* with conserved enhancer grammar. Our analysis also identified two notochord enhancers regulated by Zic, ETS, FoxA and Bra binding sites: the Bra Shadow (BraS) enhancer located in close proximity to the gene *Bra*, and an enhancer located near the gene *Lrig*. By creating a library of 45 million enhancer variants with the sequence, affinity and position of the Zic, ETS, FoxA and Bra sites fixed while all other nucleotides are randomized, we discover that these sites are necessary and sufficient for notochord expression. Zic, ETS, FoxA and Bra binding sites occur within the *Ciona* Bra434 enhancer and vertebrate notochord *Bra* enhancers, suggesting a conserved regulatory logic. Collectively, this study deepens

our understanding of how enhancers encode notochord expression, illustrates the importance of enhancer grammar, and hints at the conservation of enhancer logic and grammar across chordates.

INTRODUCTION

Enhancers are genomic elements that act as switches to ensure the precise patterns of gene expression required for development (Levine, 2010). Enhancers regulate the timing, locations and levels of expression by binding of transcription factors (TFs) to sequences within the enhancer known as transcription factor binding sites (TFBSs) (Heinz et al., 2010; Liu and Posakony, 2012; Small et al., 1992; Spitz and Furlong, 2012; Swanson et al., 2010). This binding, along with protein-protein interactions, leads to recruitment of transcriptional machinery and activation of gene expression. While we understand that TFBSs regulate enhancers and mediate tissue-specific expression, we have limited understanding of how the sequence of an enhancer encodes a particular expression pattern and what combinations of binding sites within enhancers are able to mediate enhancer activity. Given that the majority of variants associated with disease and phenotypic diversity lie within enhancers (Maurano et al., 2012; Tak and Farnham, 2015; Visel et al., 2009), it is critical that we understand how the underlying enhancer sequence encodes tissue-specific expression and what types of changes within an enhancer sequence can cause changes in expression, cellular identity and phenotypes.

A set of grammatical rules that define how enhancer sequence encodes tissue-specific expression is an attractive idea first suggested almost 30 years ago (Arnone and Davidson, 1997; Barolo, 2016; Levo and Segal, 2014; Thanos and Maniatis, 1995). The hypothesis for grammatical rules is based on the fact that proteins and the enhancer DNA have physical properties. These physical constraints govern the interaction of proteins with DNA and could be read out within the

DNA sequence at the level of TFBSs. Enhancer grammar is composed of constraints on the number, type, and affinity of TFBSs within an enhancer and the relative syntax of these sites (orders, orientations, and spacings) (Jindal and Farley, 2021).

We previously identified grammatical rules governing notochord enhancers regulated by Zic and ETS TFBSs (Farley et al., 2016). We found that there was an interplay between affinity and organization of TFBSs, such that organization could compensate for poor affinity and vice versa. Using these rules, we identified two novel notochord enhancers, Mnx and Bra Shadow (BraS). These enhancers use low-affinity ETS sites in combination with Zic sites to encode notochord expression (Farley et al., 2016). Here, we focus on obtaining a deeper understanding of how enhancers regulated by Zic and ETS encode notochord expression.

Zic and ETS are co-expressed in the developing notochord of the marine chordate *Ciona* (Figure 1) and in vertebrates (Dykes et al., 2018; Matsumoto et al., 2007). The notochord is a key feature of chordates and acts as a signaling center to pattern the neighboring neural tube, paraxial mesoderm, and gut (Herrmann and Kispert, 1994; Stemple, 2005). Specification of the notochord by Brachyury (Bra), also known as T, is highly conserved across chordates (Chesley, 1935; Chiba et al., 2009; Wilkinson et al., 1990; Yasuo and Satoh, 1993). Other conserved TFs important for activation of notochord gene expression include Zic, ETS, a TF downstream of FGF signaling, and FoxA (Dykes et al., 2018; Elms et al., 2004; Imai et al., 2002b; Kumano et al., 2006; Matsumoto et al., 2007; Warr et al., 2008; Yagi et al., 2004) (Imai et al., 2002a; Matsumoto et al., 2007; Miya and Nishida, 2003; Schulte-Merker and Smith, 1995; Yasuo and Hudson, 2007) (Ang and Rossant, 1994; Dal-Pra et al., 2011; José-Edwards et al., 2015; Katikala et al., 2013; Passamaneck et al., 2009; Weinstein et al., 1994).

Our study focuses on the marine chordate, *Ciona intestinalis type A,* also known as *Ciona robusta* (*Ciona*), a member of the urochordates, the sister group to vertebrates (Delsuc et al., 2006). Fertilized *Ciona* eggs can be electroporated with many enhancers in a single experiment which allows for testing of many enhancers in whole, developing embryos (Davidson and Christiaen, 2006; Farley et al., 2015). Furthermore, these embryos are transparent and have defined cell lineages, making it easy to image and determine the location of enhancer activity. These advantages, along with the fast development of *Ciona* and the similarity of notochord development programs between *Ciona* and vertebrates (Davidson and Christiaen, 2006; Di Gregorio, 2020), make it an ideal organism to study the rules governing notochord enhancers during development. Within the *Ciona* genome, we found 1092 elements containing one Zic site and at least two ETS sites within 30bp upstream or downstream of the Zic site. We tested 90 of these for expression in developing *Ciona* embryos. Only 10% of these regions drive notochord expression. These notochord enhancers fall into three categories: enhancers containing Zic and ETS sites, ones with Zic, ETS and Bra sites, and ones with Zic, ETS, FoxA and Bra sites. Within enhancers containing Zic and ETS sites, the organization of sites is important for activity, indicating that grammatical constraints on Zic and ETS encode enhancer activity. We find that one of the Zic and ETS enhancers is near an important notochord gene, *laminin alpha* (Veeman et al., 2008). The orientation of binding sites within this *laminin alpha* enhancer is critical for enhancer activity demonstrating the role of enhancer grammar. We find similar clusters of Zic and ETS sites within the introns of *laminin alpha-1* in both mouse and human. Strikingly, we find the same 12bp spacing between the Zic and ETS conserved across all three species. Additionally, this study identifies two enhancers using a combination of Zic, ETS, FoxA, and Bra to encode notochord expression. One of these is the BraS enhancer. By creating a library of 45 million enhancer variants with the

sequence, affinity and position of the Zic, ETS, FoxA and Bra sites fixed while all other nucleotides are randomized, we discover that these sites are necessary and sufficient for notochord expression. Other known *Bra* enhancers within *Ciona* (Corbo et al., 1997) and vertebrates (Schifferl et al., 2021) also harbor this combination of TFs, suggesting that Zic, ETS, FoxA, and Bra is a common feature of *Bra* regulation in chordates. Collectively, our study finds that grammar is a key component of functional enhancers with signatures of this enhancer logic and grammar seen across chordates.

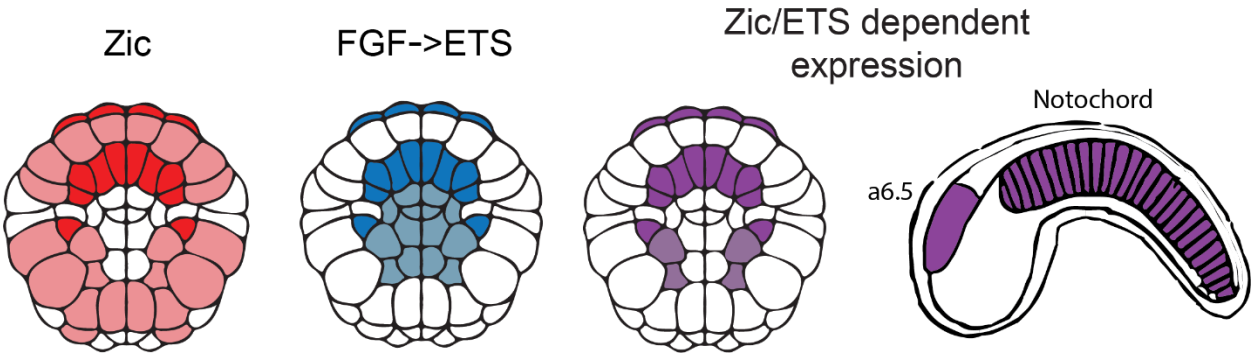


Figure 1. Zic and ETS expression in the 110-cell stage embryo. Co-expression of Zic and ETS is shown in purple and occurs in the notochord, a6.5 lineage, which gives rise to the anterior sensory vesicle and palps, and four mesenchyme cells shown in light purple. A schematic of the tailbud embryo shows the notochord and a6.5 cell types later in development. Dark coloring represents a6.5 and notochord lineages, and light coloring represents other tissues with expression of Zic and/or ETS.

RESULTS:

Searching for clusters of Zic and ETS sites within the *Ciona* genome

To better understand how Zic and ETS sites within enhancers encode notochord expression, we searched the *Ciona* genome (KH2012) for clusters of Zic and ETS sites. To do this, we first identified Zic motifs in the genome. We defined Zic motifs using EMSA and enhancer mutagenesis data from previous studies (see methods for motifs) (Matsumoto et al., 2007; Takahashi et al., 1999; Yagi et al., 2004). Using the Zic site as an anchor, we searched the 30bp

upstream and downstream of the Zic site for ETS sites, using the core motif GGAW (GGAA and GGAT) to consider all ETS sites regardless of affinity (Lamber et al., 2008; Wei et al., 2010), as we have previously found that low-affinity ETS sites are required to encode notochord-specific expression (Farley et al., 2016). This search identified 1092 genomic regions approximately 68bp in length. We define these regions as ZEE elements.

Testing ZEE genomic elements for enhancer activity in developing *Ciona* embryos

We selected 90 ZEE elements (Figure S1 and Table S1) and synthesized these upstream of a minimal promoter (bpFog, Rothbächer et al., 2007; Stolfi et al., 2015) and a transcribable barcode to conduct an enhancer screen (experiment outlined in Figure 2A). Each enhancer was associated with, on average, six unique barcodes. Each different barcode is a distinct measurement of enhancer activity. We electroporated this library into fertilized *Ciona* eggs. We collected embryos at the late gastrula stage (5.5 hours post fertilization, hpf) when notochord cells are developing (Jiang and Smith, 2007) and both Zic and ETS are expressed (Imai et al., 2004; Winkley et al., 2021). At this timepoint, we isolated mRNA and DNA. To determine that all the enhancer plasmids got into the embryos, we isolated the plasmids from the embryos and sequenced the DNA barcodes. We detected barcodes associated with all 90 ZEE elements from the isolated plasmids, indicating that all elements were tested for activity within the developing *Ciona* embryos.

We next wanted to see how many of the 90 ZEE elements act as enhancers to drive transcription. Active enhancers will transcribe the GFP and the barcode into mRNA. To find the functional enhancers, we isolated the mRNA barcodes from our electroporated embryos and sequenced them. We analyzed the sequencing data and measured the reads per million (RPM) for each barcode. To calculate an average RNA RPM for a given enhancer, we averaged the RPM for

each RNA barcode associated with an enhancer. To normalize the enhancer activity to the differences in the amount of plasmid and therefore number of copies of the enhancer electroporated into embryos, we took the log2 of the average enhancer RNA RPM divided by the DNA RPM for the same enhancer to create an enhancer activity score. Enhancer activity scores below zero are non-functional, while elements with scores above zero are considered functional enhancers. The highest activity score is around four. The experiment was repeated in biological triplicate and there was a high correlation between all three biological replicates (Figure S2).

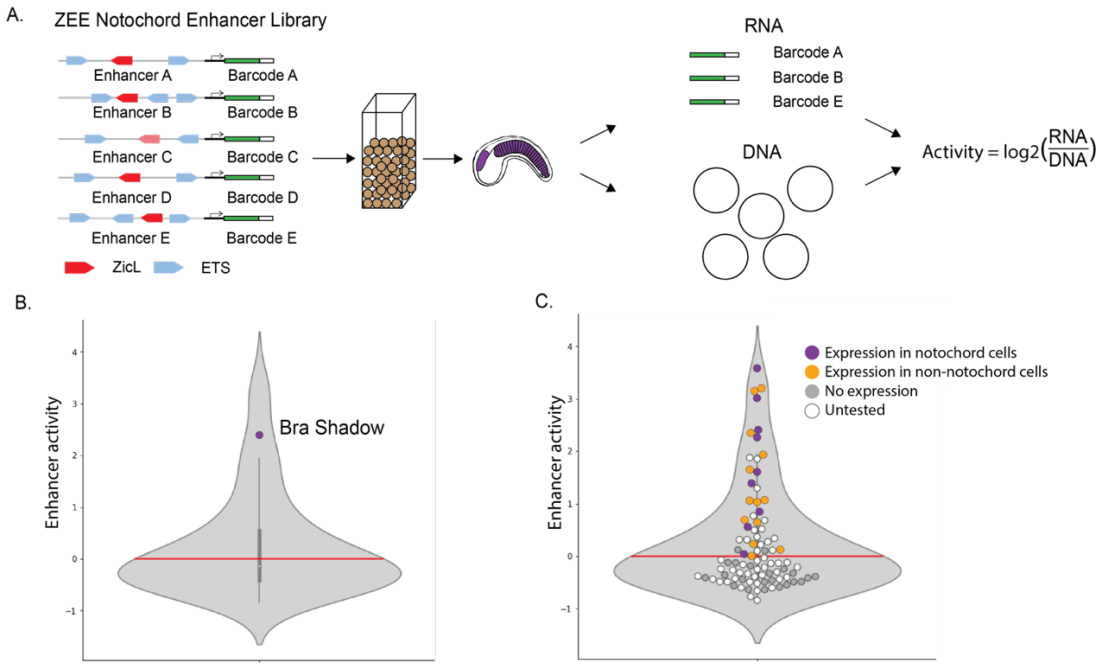


Figure 2. **Screening Zic and ETS genomic elements in *Ciona*. A.** Schematic of enhancer screen. 90 ZEE genomic regions, each associated with on average six unique barcodes were electroporated into fertilized *Ciona* eggs. mRNA and plasmid DNA were extracted from 5.5hpf embryos (tailbud embryo shown to highlight tissues with predicted expression). The mRNA and DNA barcodes were sequenced, and a normalized enhancer activity score was calculated for each enhancer by taking the log2 of the mRNA activity for a given enhancer divided by the number of copies of the plasmid. **B.** Violin plot showing the distribution of enhancer activity. The Bra Shadow enhancer served as a positive control and is labeled. The red line indicates the cut-off for non-functional elements at zero. **C.** Same plot as B, but with all 90 ZEE elements plotted as dots. Dots are colored by the results of an orthogonal screen, where we measured the GFP expression in at least 150 embryos to determine the location of expression (50 embryos per repeat). Enhancers driving notochord expression are shown in purple, enhancers with expression but no notochord expression are shown in orange. ZEE elements that do not drive expression are grey and untested enhancers are shown in white.

Many genomic ZEE elements are not enhancers

As an internal, positive control in our enhancer screen, we included the Bra Shadow (BraS) enhancer. This enhancer drives expression in the notochord and weak expression in the a6.5 lineage, both locations that express Zic and ETS (Farley et al., 2016). The BraS enhancer activity score is 2.4 (Figure 2B), indicating that our library screen is detecting functional enhancers. Thirty-nine of the ZEE elements act as enhancers in our screen, while fifty-one of the ZEE elements drove no expression. This suggests that genomic elements containing a single Zic site and at least two ETS sites are not sufficient to drive expression in the notochord. To further validate our sequencing data and to determine the tissue-specific location of the functional enhancers, we selected 20 non-functional elements and 24 functional enhancers from our screen to test by an orthogonal approach. Each of these ZEE elements were cloned upstream of a minimal bpFog promoter and GFP. We electroporated each enhancer into fertilized eggs and analyzed the GFP expression of these ZEE elements under the microscope at 8hpf in at least 150 embryos across three biological replicates. Collectively, we analyzed expression of these elements in over 6600 embryos with this orthogonal approach.

All 20 ZEE elements defined as non-functional in our library drove no GFP expression, validating our enhancer activity score cut off that we defined for non-functional enhancers (Fig 2C). In the 24 enhancers detected as functional within the enhancer screen, 92% of these enhancers (22/24) showed GFP expression within the embryos when tested individually (Table S2). Nine ZEE elements drove expression in the notochord (Figure S3 and Table S3). Four of these enhancers are active almost exclusively in the notochord (ZEE10, 13, 20, 27). The remaining five are active in the notochord with additional expression in the endoderm and/or nerve cord (b6.5 lineage).

Twelve of the ZEE enhancers drove varying levels of expression in the a6.5 lineage, which gives rise to the neural cell types called the anterior sensory vesicle and the palps, but only one drove expression exclusively in this cell type (ZEE22). Thirteen ZEE elements drove expression in one or more for the following cell types: the nerve cord (b6.5 lineage), mesenchyme, and endoderm. The expression patterns seen for these active enhancers are consistent with the expression patterns of Zic and ETS which are expressed in the muscle, endoderm, ectoderm, mesenchyme, notochord, a6.5 neural lineage and b6.5 neural cell types (Hudson et al., 2016, 2007; Imai et al., 2006; Picco et al., 2007; Wagner and Levine, 2012) (Note S1 discusses the expression patterns of the ZEE elements with notochord expression in more detail). The only cells to co-express both Zic and ETS are the notochord, a6.5, and a small number of mesenchyme cells (Figure 1). Therefore, enhancers under combinatorial control of Zic and ETS are likely to be active in the notochord and the a6.5 neural lineage (Ikeda and Satou, 2016; Matsumoto et al., 2007; Wagner and Levine, 2012). Collectively these results indicate that our enhancer screen accurately detects functional enhancers, and our tissue-specific analysis provides detailed expression patterns for these enhancers.

Elucidating the logic of the enhancers driving notochord expression

Having seen that so few enhancers drive expression in the notochord, we were interested to better understand why these nine functional enhancers were active in the notochord. It is possible that they are functional due to the grammar of the Zic and ETS sites or because other TFBSs are required for notochord expression. To investigate these two hypotheses, we looked at the nine notochord enhancers in more detail. FoxA and Bra are two other TFs important for activation of notochord enhancers in chordates (Ang and Rossant, 1994; Casey et al., 1998; Dal-Pra et al., 2011; José-Edwards et al., 2015; Katikala et al., 2013; Passamaneck et al., 2009; Wilkinson et al., 1990).

We therefore searched all 90 ZEE elements for FoxA and Bra sites. We used EMSA and crystal structure data to define TRTTTAY as the FoxA motif (Katikala et al., 2013; Li et al., 2017; Passamaneck et al., 2009) and TNNCAC as the Bra motif (Casey et al., 1998; Conlon et al., 2001; Di Gregorio and Levine, 1999; Dunn and Di Gregorio, 2009; Müller and Herrmann, 1997).

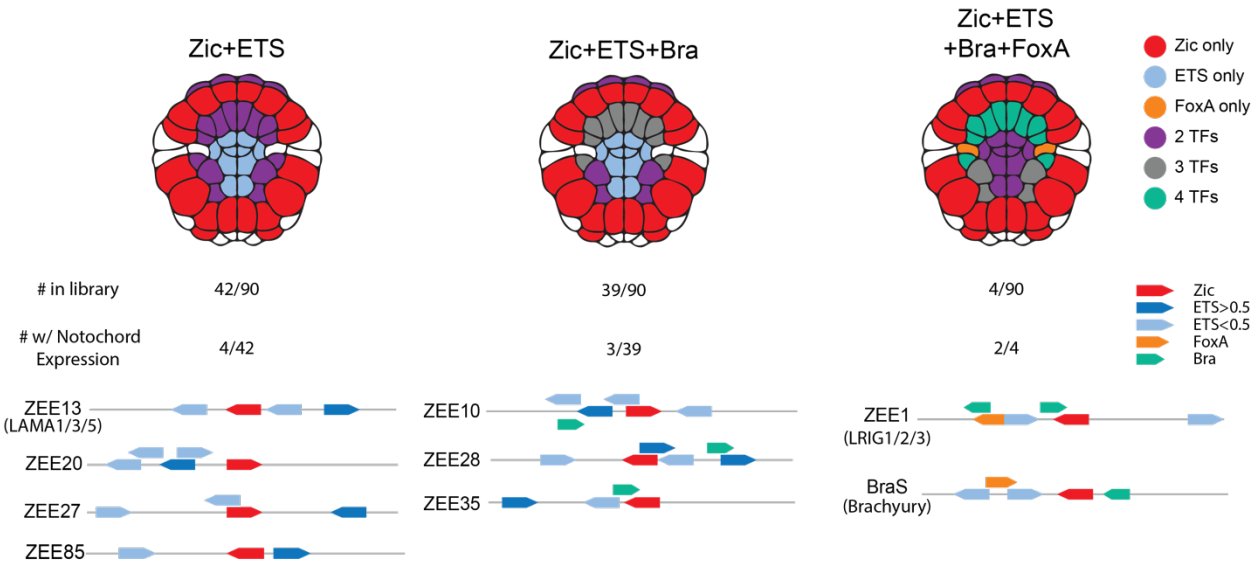


Figure 3. **Combinations of transcription factors in ZEE enhancers that drive notochord expression.** Notochord-expressing ZEE elements were grouped by the combination of transcription factor binding sites present in each element. For each combination, an embryo schematic shows the overlapping region of expression for that given combination. Below the embryo schematic, the number of ZEE elements, the number of ZEE elements with notochord expression and schematics of the ZEE elements with notochord expression within each group. Zic (red), ETS (blue), FoxA (orange), and Bra (green) sites are annotated. Dark blue ETS sites have an affinity of greater than 0.5, light blue sites have an affinity of less than 0.5.

The nine elements that drive notochord expression contain three different combinations of transcription factors

Of the 90 genomic regions we tested, 42 had only Zic and ETS sites, 39 had Zic, ETS and Bra sites, 4 had Zic, ETS, FoxA, and Bra sites and 5 had Zic, ETS and FoxA sites. Ten percent of the enhancers containing only Zic and ETS sites drive notochord expression (4/42). Eight percent (3/39) of the enhancers containing Zic, ETS, and Bra drive notochord expression. None of the

enhancers (0/5) containing Zic, ETS, and FoxA drive notochord expression, while fifty percent (2/4) of the enhancers containing Zic, ETS, FoxA and Bra are active in the notochord (Figure 3 and Figure S4). Thus, there are three groups of notochord enhancers that contain: (1) Zic and ETS sites alone, (2) Zic, ETS and Bra sites, or (3) Zic, ETS, FoxA, and Bra sites. Having found that only a few of the elements containing Zic and ETS sites alone were functional, we wanted to understand if the organization or grammar of sites within these enhancers was important.

Zic and ETS enhancer grammar encodes notochord *laminin alpha* expression

Four enhancers containing Zic and ETS sites only (ZEE13, 20, 27 and 85) drive notochord expression. ZEE13, 20 and 27 drive expression only in the notochord and have similar levels of expression. ZEE85 drives expression predominantly in the nerve cord (b6.5 lineage) with weak notochord expression. ZEE20, 27, and 85 are not in close proximity to known notochord genes, though it is possible that these elements regulate notochord genes further away. The ZEE13 enhancer is located close to *laminin alpha*, which is critical for notochord development (Veeman et al., 2008) (Figure 4A). Given the proximity of this notochord-specific enhancer to *laminin alpha*, we decided to focus further analysis on this enhancer, which we renamed the Lama enhancer. Notably, this enhancer contains three ETS sites. To determine the affinity of these sites, we used Protein Binding Microarray data (PBM) for mouse ETS-1 (Wei et al., 2010), as the binding specificity of ETS is highly conserved across bilaterians (Nitta et al., 2015; Wei et al., 2010). The consensus highest-affinity site has a score of 1.0, and all other 8-mer sequences have a score relative to the consensus. The Lama enhancer contains two ETS sites with exceptionally low affinities of 0.10, or 10% of the maximal binding affinity, while the most distal ETS site is a high-affinity site (0.73).

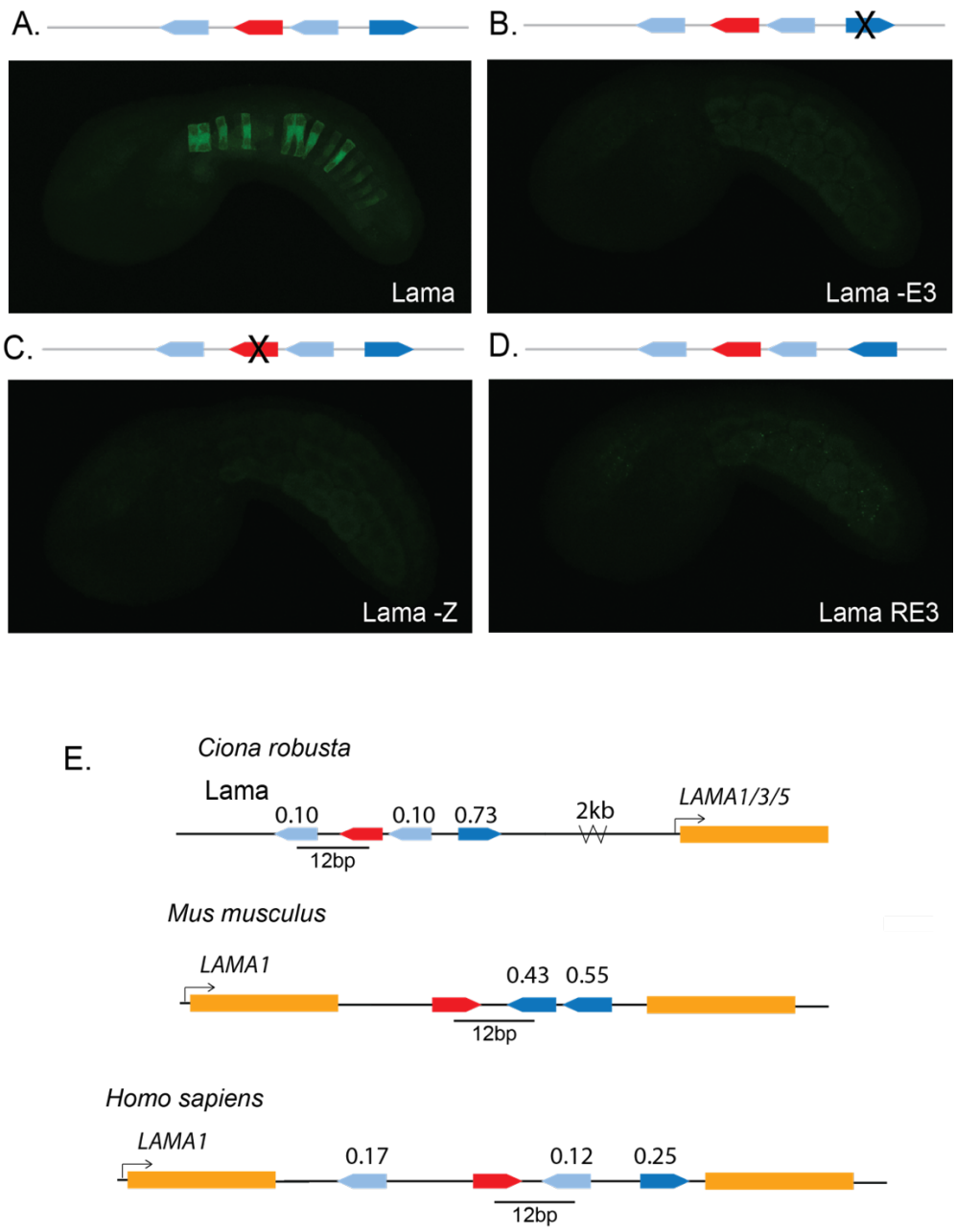


Figure 4. Zic and ETS grammar encodes a notochord *laminin alpha* enhancer. A. Embryo electroporated with the Lama enhancer (ZEE13); GFP expression can be seen in the notochord. **B.** Embryo electroporated with Lama -E3, where ETS3 was mutated to be non-functional; no GFP expression detected. **C.** Embryo electroporated with Lama -Z, where the Zic was mutated to be non-functional; no GFP expression detected. **D.** Embryo electroporated with Lama RE3, where the sequence of ETS3 was reversed; no GFP expression detected. Comparable results were seen when ETS1 was reversed. **E.** Schematics of Zic and ETS clusters near *laminin alpha* in the genome of *Ciona*, mouse, and human. All three *laminin alpha* clusters have a spacing of 12bp between an ETS and Zic site and all contain non-consensus ETS sites. ETS site affinity scores are noted above each site. Dark blue ETS sites have an affinity of greater than 0.5, light blue sites have an affinity of less than 0.5.

To determine if the Zic site and ETS sites are important for enhancer activity, we made a point mutation to ablate the ETS3 site, which we chose because it has the highest affinity (Figure 4B, Figure S5A, and Table S4). This led to a complete loss of notochord activity indicating that this ETS site contributes to enhancer activity. Similarly, ablation of the Zic site results in complete loss of enhancer activity, indicating that both Zic and ETS sites are necessary for activity of this Lama enhancer (Figure 4C, Figure S5A, and Table S4). We did not ablate the low affinity ETS sites of the Lama enhancer. Previously, we saw that the organization of sites within enhancers, a component of enhancer grammar, is critical for enhancer activity in both the Mnx and Bra enhancer. To see if enhancer grammar is important for activity within the Lama enhancer, we altered the orientation of sites within this enhancer and measured the impact on enhancer activity. Reversing the orientation of the first ETS site, which has an affinity of 0.10, led to a dramatic reduction in notochord expression, suggesting the orientation of this ETS site is important for enhancer activity. Similarly, reversing the orientation of the third ETS site (Lama RE3), which has an affinity of 0.73, also causes a loss of notochord expression (Figure 4D, Figure S5A, and Table S4). These two manipulations demonstrate that the orientation of these ETS sites within this enhancer is important for activity, and thus, that there are some grammatical constraints on the *Ciona* Lama enhancer. It is likely that grammar is an important feature of enhancers regulated by Zic and ETS, as we have previously seen similar grammatical constraints on the orientation and spacing of binding sites within the Mnx and BraS enhancer, and because so few genomic elements containing these sites are functional (Farley et al., 2016).

Vertebrate *laminin alpha-1* introns contain clusters of Zic and ETS with conserved spacing.

The expression of laminin in the notochord is highly conserved between urochordates and vertebrates (Reeves et al., 2017; Scott and Stemple, 2005; Veeman et al., 2008). Indeed, laminins play a vital role in both urochordate and vertebrate notochord development, with mutations in laminins or components that interact with laminins causing notochord defects (Machingo et al., 2006; Parsons et al., 2002; Pollard et al., 2006). The *Ciona laminin alpha* is the ortholog of the vertebrate *laminin alpha 1/3/5* family. We therefore sought to determine if we could find a similar combination of Zic and ETS sites in proximity to vertebrate *laminin* genes, as both Zic (Dykes et al., 2018; Warr et al., 2008) and ETS (Barnett et al., 1998; Olivera-Martinez et al., 2012) are important in vertebrate notochord development. Strikingly, we find a cluster of Zic and ETS sites within the intron of both the mouse and human *laminin alpha-1* genes. The affinity of the ETS sites in all three species is also far from the consensus: the human cluster contains three ETS sites of 0.12, 0.17 and 0.25 affinity, while the putative mouse enhancer contains fewer, but higher-affinity, ETS sites (Figure 4E). We have previously seen that the spacing between Zic and adjacent ETS sites affects levels of expression, with spacings of 11 and 13bp seen between ETS and Zic sites in the BraS enhancer and Mnx enhancer, respectively (Farley et al., 2016). In line with this observation, the *laminin alpha-1* clusters in mouse and human and the *Ciona* Lama enhancer have a 12bp spacing between the ETS and adjacent Zic site in all three species, suggesting that such spacings (11-13bp) are a feature of some notochord enhancers regulated by Zic and ETS. The conservation of this combination of sites, the low-affinity ETS sites, and the conserved spacing hints at the conservation of enhancer grammar across chordates.

The Zic, ETS, FoxA and Bra regulatory logic encodes notochord enhancer activity

The group of genomic elements most enriched in notochord expression was the group containing Zic, ETS, FoxA and Bra binding sites, with two of the four driving notochord expression. Both of these enhancers are located near genes expressed in the notochord (Reeves et al., 2017). The first was our positive control BraS, while the second enhancer is in proximity of the *Lrig* gene. Both of these enhancers drive strong notochord expression along with some neural a6.5 expression.

We previously identified the BraS enhancer through a search for rules governing Zic and ETS grammar that included number and type of TFBSs, along with the affinity, spacing, and orientation of TFBSs (Farley et al., 2016). The BraS enhancer contains a Zic and two low-affinity ETS sites (0.14 and 0.25). We previously saw that changing the orientation of the lowest affinity ETS site, located 11bp from the Zic site, leads to loss of expression, indicating that there are grammatical constraints on this enhancer and that the 0.14 affinity ETS site is important for expression (Farley et al., 2016). To further confirm the role of the Zic and two ETS sites within BraS, we ablated these three sites (Zic and both ETS sites) with point mutations; this leads to complete loss of expression, demonstrating that these sites are necessary for notochord expression (Figure 5B, Figure S5B, and Table S4). To test if these sites are sufficient for notochord expression, we created a library of 24.5 million variants in which the Zic and two ETS sites were kept constant in sequence, affinity, and position while all other nucleotides were randomized. We electroporated this library into embryos and counted GFP expression in 8hpf embryos. BraS has notochord expression in 73% of embryos, while the ZEE-randomized BraS enhancer (BraS rZE) has notochord expression in only 28% of embryos. Thus, BraS rZE drives expression within the notochord in significantly fewer embryos than BraS, indicating that there are other sites within the

enhancer that are also important for tissue-specific expression (Figure 5C, Figure S5B, and Table S4). This experiment highlights the importance of understanding sufficiency in addition to necessity of sites.

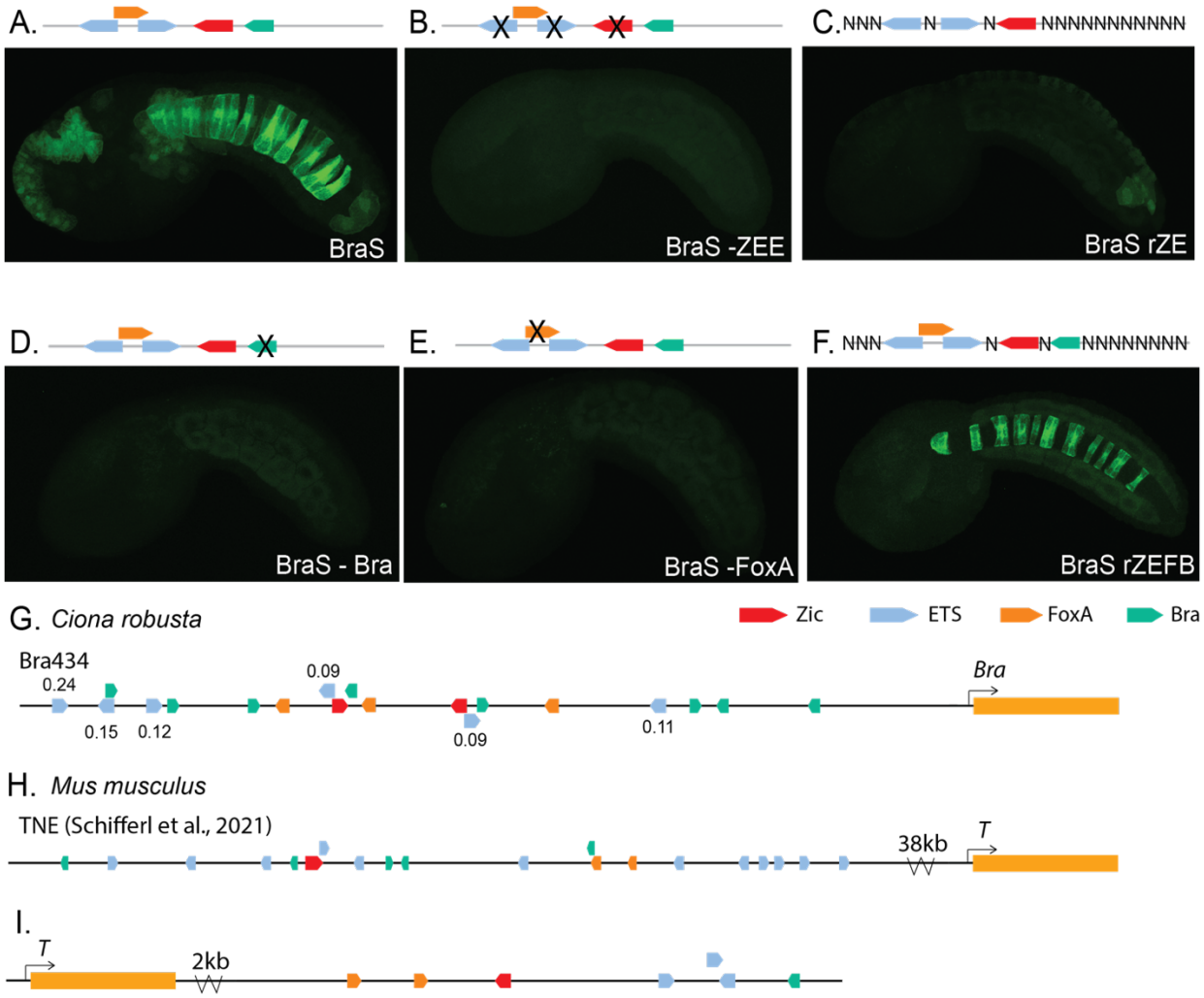


Fig 5. Zic, ETS, FoxA, and Bra may be a common regulatory logic for Brachyury enhancers. **A.** Embryo electroporated with the Bra Shadow (BraS) enhancer; GFP expression can be seen in the notochord. **B.** Embryo electroporated with BraS -ZEE, where the Zic and two ETS sites were mutated to be non-functional; no GFP expression was detected. **C.** Embryo electroporated with BraS rZE, where the Zic and two ETS sites were fixed, and all other nucleotides were randomized; GFP expression was greatly diminished. **D.** Embryo electroporated with BraS -Bra, where the sequence of Bra was mutated to be non-functional; GFP expression was greatly diminished. **E.** Embryo electroporated with BraS -FoxA, where the sequence of FoxA was mutated to be non-functional; GFP expression was greatly diminished. **F.** Embryo electroporated with BraS rZEFB, where the Zic, two ETS, FoxA, and Bra sites were fixed, and all other nucleotides were randomized; GFP expression can be seen in the notochord **G-I**. Schematics of Zic (red), ETS (blue), FoxA (orange), and Bra (green) clusters near *Bra* in the genomes of *Ciona* and mouse.

Two obvious candidates for additional functional sites within BraS are the FoxA and Bra sites, which we detected in this enhancer. Both FoxA and Bra are TFs known to regulate notochord enhancers in urochordates and vertebrates (Ikeda and Satou, 2016; José-Edwards et al., 2015; Kumano et al., 2006; Lolas et al., 2014; Passamaneck et al., 2009; Reeves et al., 2021). To test if the Bra and FoxA sites contribute to expression we ablated these sites. Ablating the Bra site within BraS leads to a significant reduction in expression, as does ablating the FoxA site (Figure 5D and E, Figure S4B, and Table S4). These manipulations suggest that all five sites (Zic, FoxA, Bra, and two ETS sites) are necessary for enhancer activity, and that all four TFs contribute to the activity of BraS.

To test if the Zic, two ETS, FoxA and Bra sites are sufficient for notochord expression, we created another BraS randomization library with 45 million variants in which the Zic, ETS, FoxA, and Bra (ZEFB) sites were fixed in sequence, position and affinity and all other nucleotides within the enhancer were randomized. When we electroporated this library into *Ciona*, the number of embryos showing notochord expression between the BraS ZEFB-randomized library (BraS rZEFB) and BraS WT was not significantly different (73% BraS vs 62% BraS rZEFB) (Figure 5F, Figure S5B, and Table S4), suggesting that these five sites together are sufficient to drive notochord expression in the BraS enhancer. While there is no significant difference in the number of embryos with notochord expression between the BraS rZEFB and BraS enhancers, we noticed that expression in the notochord was slightly weaker for BraS rZEFB ($p=0.03$) (Figure S4C), suggesting that other elements within the randomized region may further augment the levels of notochord expression. We also noted that significantly fewer embryos drive expression in the a6.5 lineage in the BraS rZEFB relative to the BraS enhancer (14% vs 32% of embryos respectively,

$p<0.01$) (Figure S4D) suggesting that sequences within the randomized region are important for the neural a6.5 expression. Studies of enhancers often stop when mutation experiments demonstrate a TF is necessary for enhancer activity. However, this falls short of a full understanding of enhancers. Our results highlight that finding necessary sites is not enough to identify the regulatory logic of an enhancer. These necessity and sufficiency experiments have uncovered a deeper understanding of the BraS enhancer, namely that it is regulated by Zic, ETS, FoxA, and Bra.

Zic, ETS, Bra and FoxA may be a common regulatory logic for *Ciona Brachyury* enhancers

The first and most well-studied *Bra* enhancer is the Bra434 enhancer (Corbo et al., 1997; Fujiwara et al., 1998), which drives strong expression in the notochord (Figure S6A). Bra434 enhancer contains Zic, ETS, FoxA, and Bra sites; ablating these sites within this enhancer lead to reduced expression, suggesting that these sites contribute to enhancer activity (Reeves et al., 2021; Shimai and Veeman, 2021). There are different reports regarding the number and location of Zic, ETS, FoxA, and Bra sites within the Bra434 enhancer depending on the method used to define sites (Corbo et al., 1997; Shimai and Veeman, 2021). Here we annotate the Bra434 enhancer using crystal structure data, enhancer mutagenesis data, EMSA and PBM data (Casey et al., 1998; Conlon et al., 2001; Di Gregorio and Levine, 1999; Dunn and Di Gregorio, 2009; Katikala et al., 2013; Lamber et al., 2008; Li et al., 2017; Matsumoto et al., 2007; Müller and Herrmann, 1997; Passamaneck et al., 2009; Takahashi et al., 1999; Wei et al., 2010; Yagi et al., 2004).

Our approach identifies two Zic sites, six low-affinity ETS sites, three FoxA sites, and eight Bra sites (Figure 5G and Figure S6B). Of these TFs, the least information is available regarding Zic; thus, it is possible that there are other more degenerate Zic sites that may be

identified in future studies. (Corbo et al., 1997; Fujiwara et al., 1998; Reeves et al., 2021; Shimai and Veeman, 2021). Bra434 has stronger expression in the notochord than BraS and this may be due to the longer length of the Bra434 enhancer and the presence of more Zic, ETS, FoxA and Bra sites within Bra434 relative to BraS enhancer. Having seen that clusters of Zic, ETS, FoxA, and Bra are important in the BraS and Bra434 enhancers, we next wanted to see if this logic is found in *Bra* enhancers in vertebrates.

Vertebrate notochord enhancers contain clusters of Zic, ETS, Fox and Bra, suggesting this is a common logic for regulation of *Brachyury* expression in the notochord

In mouse, the most well-defined notochord enhancer to date is within an intron of *T2*, 38kb upstream of *T*, which is the mouse ortholog of *Bra* (Schifferl et al., 2021) (Figure 5H). This mouse *T* enhancer is required for *Bra*/*T* expression, notochord cell specification and differentiation (Schifferl et al., 2021). Homozygous deletion of this *Bra*/*T* enhancer in mouse leads to reduction of *Bra*/*T* expression, a reduction in the number of notochord cells, and halving of tail length. Bra/T and FoxA binding sites have previously been identified within this enhancer (Schifferl et al., 2021). We find that this mouse *Bra*/*T* enhancer also contains Zic and ETS binding sites. Within this enhancer there are 12 ETS sites; 11 of these have affinities ranging from 0.09-0.14, while one site has an affinity of 0.65, indicating that this enhancer contains low-affinity ETS sites.

As we saw with the *Ciona* BraS and Bra434 enhancer, typically there are multiple enhancers that all regulate the same or similar patterns of expression (Frankel et al., 2010; Hong et al., 2008; Perry et al., 2010). This is thought to confer the transcriptional robustness required for successful development (Antosova et al., 2016; Frankel et al., 2010; Osterwalder et al., 2018; Perry et al., 2010). Following this logic, we continued to search the mouse *Bra*/*T* region to see if we

could find other putative notochord enhancers that may regulate *Bra*/*T*. We identified a region located 2kb downstream of *T* that contains a cluster of Zic, low-affinity ETS (0.11-0.12), FoxA and Bra sites (Figure 5I). This putative enhancer occurs within an open chromatin region in mouse E8.25 notochord cells (Pijuan-Sala et al., 2020), suggesting this may be another mouse *T* enhancer. Similarly in zebrafish, a notochord enhancer located 2.1kb upstream of the *Bra* ortholog *ntl* (Harvey et al., 2010) also contains a cluster of Zic, ETS, FoxA, and Bra sites (Table S6). The presence of these four TFs in *Ciona*, zebrafish, and mouse *Bra* enhancers suggests that the use of Zic, ETS, FoxA and Bra could be a common enhancer logic regulating expression of the key notochord-specification gene *Bra* in chordates.

Discussion

In this study we sought to understand the regulatory logic of notochord enhancers by taking advantage of high-throughput studies within the marine chordate *Ciona*. Within the *Ciona* genome, there are 1092 genomic regions containing a Zic site within 30bp of two ETS sites. We tested 90 of these ZEE genomic regions for expression in developing *Ciona* embryos. Surprisingly, only nine of the regions drove notochord expression. Among these nine, we identified a *laminin alpha* enhancer that was highly dependent on grammatical constraints for proper expression. We found a similar cluster of Zic and ETS sites within the intron of the mouse and human *laminin alpha-1* gene; strikingly, these clusters and the *Ciona* laminin enhancer have the same spacing between the Zic and ETS sites. Within the BraS enhancer, although Zic and ETS are necessary for enhancer activity, randomization of the BraS enhancer keeping only the Zic and ETS sites constant in a sea of 24.5 million variants reveals that these sites are not sufficient for notochord activity. FoxA and Bra sites are also necessary for notochord expression. Indeed, creating a library of 45 million BraS

variants in which all five TFBSs are kept constant in position, and affinity while all other nucleotides are randomized leads to notochord expression in a similar proportion of embryos as the WT BraS, which indicates these sites are sufficient for notochord expression. We find that the combination of Zic, ETS, FoxA, Bra occurs within other *Bra* enhancers in *Ciona* and vertebrates suggesting this combination of TFs may be a common logic regulating *Bra* expression. Our study identifies new developmental enhancers, demonstrates the importance of enhancer grammar within developmental enhancers and provides a deeper understanding of the regulatory logic governing *Bra*. Our findings of the same clusters of sites within vertebrates hint at the conserved role of grammar and logic across chordates.

Very few genomic regions containing Zic and two ETS sites are functional enhancers

Our analysis of 90 genomic elements all containing at least one Zic site in combination with two ETS sites strikingly demonstrated that clusters of sites are not sufficient to drive expression. Only 39 of the 90 (43%) elements tested drove any expression, and even more surprisingly, only 15 of these drove expression in lineages that co-express Zic and ETS, namely the a6.5 (anterior sensory vesicle and palps) and/or notochord. These findings indicate that searching for clusters of TFs is only minimally effective in identification of enhancers and suggests that the organization of sites is also important for rendering a cluster of binding sites a functional enhancer. Our findings are in agreement with the work from King et al., that found only 28% of the genomic elements they tested for enhancer function in ES cells drove enhancer activity, despite the fact that these genomic elements contain TF motifs and bound these TFs in ChIP-seq assays (King et al., 2020). Our study and King et al. suggest that having motifs, or even TF binding is not

sufficient to drive expression and suggests that the grammar of these sites is critical for rendering a cluster of TFBSs a functional enhancer (King et al., 2020).

Grammar is a key constraint of the Lama and BraS enhancers

Zic and ETS are necessary for activity of the Lama enhancer. Within the Lama enhancer, the orientation of binding sites relative to each other was critical for expression, providing evidence that enhancer grammar is a critical feature of functional enhancers regulated by Zic and ETS. Flipping the orientation of either the first or last ETS sites relative to the Zic site led to loss of enhancer activity in the *Ciona* Lama enhancer. This mirrors the results of flipping the orientation of the ETS sites within the BraS enhancer (Farley et al., 2016). *Laminin alpha* is a key gene involved in notochord development in both *Ciona* and vertebrates (Pollard et al., 2006; Veeman et al., 2008). Intriguingly, we find that both the human and mouse *laminin alpha-1* have introns that harbor a similar cluster of Zic and ETS sites to those seen within *Ciona*. There is a conservation of 12bp spacing between the Zic and ETS site across all three chordate enhancers, similar to the spacing we have observed between Zic and ETS sites within the notochord enhancers Mnx and BraS (Farley et al., 2016). We note that the vertebrate regions do not drive notochord expression in *Ciona*. It possible that grammar is subtly tweaked between different species. Alternatively, the lack of activity could be due to promoter incompatibility across species, as in our assay we tested the mouse and human Lama enhancers with a Ciona promoter. Reporter assays within mouse embryos could further investigate the functionality of the mouse and human Lama putative enhancers and the role of the 12bp spacing within these elements.

Necessity of sites does not mean sufficiency – a deeper understanding of the BraS enhancer

Our study of the BraS enhancer highlights the importance of testing sufficiency of sites to investigate if we fully understand the regulatory logic of an enhancer. We previously demonstrated that reversing the orientation of an ETS site led to loss of notochord expression in the BraS enhancer. Here, in this study, we show via point mutations that both Zic and ETS sites are required for enhancer activity. However, randomization of the BraS enhancer to create 24.5 million variants in which only the Zic and ETS sites are constant demonstrates that these sites are not sufficient for enhancer activity, as the randomized BraS enhancer (BraS rZE) only drives notochord expression in less than half the number of embryos as the BraS enhancer. Having discovered that Zic and ETS alone were not sufficient, we find that both FoxA and Bra sites also contribute to the enhancer activity. In a library of 45 million variants in which the Zic, ETS, Bra and FoxA sites are kept constant in sequence, affinity and position within a randomized backbone (BraS rZEFB), we see no significant difference in the number of embryos with notochord expression. This indicates that these five sites are necessary and sufficient for enhancer activity. However, the neural expression seen with the BraS enhancer appears to depend on some features within the randomized backbone, as the ZEFB library drives significantly less neural expression. We also note that the BraS rZEFB drives slightly weaker levels of notochord expression. These findings illustrate that enhancers are densely encoded with many features which contribute to expression. This is in line with recent work suggesting that enhancers contain far more regulatory information that previously appreciated (Fuqua et al., 2020). It is possible that degenerate Zic, ETS, FoxA, or Bra sites could be present or novel TFBS are also contributing to this logic. Further analysis conducting MPRAs with these two libraries (BraS rZE and BraS rZEFB) will determine what other features are contributing to notochord and neural expression. Sufficiency experiments are rarely done, and we

are unaware of another study that has tested sufficiency across the entirety of an enhancer in developing embryos. However, our experiments demonstrate the importance of testing sufficiency to determine all the features contributing to enhancer function and illustrate the dense encoding of regulatory information within enhancers.

Partial grammatical rules can provide signatures that identify enhancers, but improved understanding could lead to more accurate predictions

We were able to find the BraS enhancer using grammatical constraints on organization and spacing between Zic and ETS site and affinity of ETS sites (Farley et al., 2016). Interestingly, we did not have all the features required for enhancer activity. As such, this suggests that partial knowledge of grammatical constraints, or partial signatures of grammar could be used to identify functional enhancers. Our previous strategy searched for these grammatical constraints in proximity of known notochord genes, which may be why we were successful in identification of the Mnx and BraS enhancer with only partial grammar rules. Understanding the dependency between all features within an enhancer will likely enable greater success in identification of functional regulatory elements, as current genomic screens have shown limited success of identifying functional enhancers through epigenetic markers and transcription factor binding sites alone (King et al., 2020). Until then, our current knowledge of grammatical constraints may still be useful for pointing us towards putative enhancers.

Zic, ETS, FoxA, and Bra may be a common logic upstream of *Brachyury* in chordates

The Bra434 enhancer also contains the same combination of sites as the BraS enhancer; therefore, it is possible that this is a common logic for regulating *Bra*. Interestingly, we find these

sites within mouse and zebrafish *Bra* enhancers (Harvey et al., 2010; Schifferl et al., 2021). While there are differences in expression dynamics of these factors in vertebrates and ascidians, it is striking to see this combination of sites in validated notochord enhancers across these species. Indeed, our study in both the *laminin* enhancers and *Bra* enhancers provides hints of a conserved regulatory logic across chordates, although future tests of these putative enhancers within mouse are required to see if these are truly conserved enhancers with similar grammar signatures. Our study focuses on conservation of grammatical signatures rather than sequence conservation. A recent study searching for conserved enhancers in syntenic regions suggests that there may be much more conservation of enhancer function than expected based on sequence conservation (Wong et al., 2020). Our approach searching for grammatical signatures rather than sequence conservation may allow for identification of such functionally conserved enhancers.

Approaches to understanding dependency grammar of notochord expression

Searching for grammatical rules governing enhancers requires comparison of functional enhancers with the same features. Although we thought we had the same features in all 90 regions, we actually had at least three distinct types of enhancers within our screen. This illustrates a common problem in mining genomic data for patterns, as the assumption that we are comparing like with like is often an incorrect one. Other screens mining genomic elements have hit similar roadblocks, with only a few functional genomic examples being uncovered and thus limiting the ability to find grammatical rules (King et al., 2020). To uncover the grammatical constraints on enhancers, we need to not only understand the number and types of sites within an enhancer, but also the dependency between these sites, such as affinity, spacing, and orientation (Jindal and Farley, 2021).

Massively or gigantic parallel reporter assays with increased size and complexity and that combine both synthetic enhancers and genomic elements will likely be required to pinpoint the rules governing enhancer activity within genomes. However, integrating synthetic screens with genomic screens is a major challenge as synthetic screens often have limited application within the context of the genome (King et al., 2020). Another approach is to study entirely random sequences for enhancer activity, which has been done in the context of promoters in bacteria and yeast (Yona et al., 2018; de Boer et al., 2020). Indeed, the conclusions of these studies mirror our own findings that grammar and low-affinity sites are critical components of functional regulatory elements. However, as 83% of the random sequences within yeast drove expression, it is unclear how well random sequences mirror the regulatory landscape within the genome that has been shaped by evolutionary constraints over millions of years. Nonetheless, testing random sequences within the context of developing embryos could provide another source of data to understand how enhancers encode tissue-specific expression (Galupa et al., 2022). In the future, integration of genomic regions, synthetic designed, and random sequences will contribute to our understanding of enhancer grammar. Despite the complexity of studying enhancers in developing embryos, our study demonstrates that enhancer grammar is critical for encoding notochord activity and our observation of the same logics and grammar signatures in both *Ciona* and vertebrates hints at conservation of these grammatical constraints across chordates.

Limitations of the study

In this study, we screened 90 ZEE elements for functionality; however, only 10% were active in the notochord. We anticipate that discovering more notochord enhancers regulated by Zic, ETS, or regulated by Zic, ETS, FoxA, and Bra could better inform our understanding of

notochord grammar. Towards this end, testing all 1092 ZEE elements we identified within the *Ciona* genome could strengthen this study. However, this would likely only yield 100 notochord enhancers, which would still not be enough to define grammatical rules. As discussed above, combining assays of genomic regions with synthetic and random enhancer screens could help gain enough data to determine the grammar of notochord enhancers.

Another limitation relates to our identification of conserved enhancer logic and grammar across chordates. While we identified similar signatures with the Lama enhancers in *Ciona*, mouse and humans, we did not test the mouse Lama enhancer for activity in mouse, nor did we functionally interrogate the importance of the 12bp spacing within this enhancer in the context of *Ciona* or mouse. Conducting these studies would deepen our understanding of the conservation of grammar across chordates. We also identified a common logic of Zic, ETS, FoxA and Bra within Bra enhancers. While we know that deletion of the mouse *Bra* TNE enhancer does lead to loss of notochord in mouse, it would strengthen the study to manipulate the Zic, ETS, FoxA, Bra sites within the context of the mouse and zebrafish Bra/T enhancers to determine if the conservation of this logic is important for regulation of *Bra*.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Emma Farley (efarley@ucsd.edu).

Materials availability

Plasmids generated in this study are available upon request.

Data and code availability

- Microscopy and scoring data reported in this paper will be shared by the lead contact upon request.
- All ZEE screen sequencing data will be deposited to GEO and will be made publicly available as of the date of publication.
- All original code from this study is available at https://github.com/farleylab/Diverse-Logics-Notochord-Study
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Tunicates

Adult *C. intestinalis* type A aka *Ciona robusta* (obtained from M-Rep) were maintained under constant illumination in seawater (obtained from Reliant Aquariums) at 18C. *Ciona* are hermaphroditic, therefore there is only one possible sex for individuals. Age or developmental stage of the embryos studied are indicated in the main text.

Method Details

Library Construction

The genomic regions were ordered from Agilent Technologies with adapters containing BseRI sites. This was cloned into the custom-designed SEL-Seq (Synthetic Enhancer Library-Sequencing) vector using type II restriction enzyme BseRI. After cloning, the library was transformed into bacteria (MegaX DHB10 electrocompetent cells), and the culture was grown up until an OD of 1 was reached. DNA was extracted using the Macherey-Nagel Nucleobond Xtra Midi kit. A 30bp barcode with adapters containing Esp3I sites was cloned into this library using type II restriction enzyme Esp3I. The library was transformed into bacteria (MegaX DHB10 electrocompetent cells) and grown up until an OD of 2 was reached. The DNA library was extracted from the bacteria using the Macherey-Nagel Nucleobond Xtra Midi kit.

Electroporation

Dechorionation, *in vitro* fertilization, and electroporation were performed as described previously in Farley et al., 2016.

GFP reporter assays

70 μg DNA was resuspended in 100 μL water and added to 400 μL of 0.96 M D-mannitol. Typically for each electroporation, eggs and sperm were collected from 10 adults. Embryos were

fixed at the appropriate developmental stage for 15 minutes in 3.7% formaldehyde. The tissue was then cleared in a series of washes of 0.3% Triton-X in PBS and then of 0.01% Triton-X in PBS. Samples were mounted in Prolong Gold. GFP images were obtained with an Olympus FV3000, using the 40X objective. All constructs were electroporated in three biological replicates.

ZEE MPRA screen

50 μg of the ZEE library was electroporated into ~5000 fertilized eggs. Embryos developed until 5hrs 30 min at 22^{o}C. Embryos put into TriZol, and RNA was extracted following the manufacturer's instructions (Life Technologies). The RNA was DNase treated using Turbo DNaseI from Ambion following standard instructions. Poly-A selection was used to obtain only mRNA using poly-A biotinylated beads as per instructions (Dyna-beads, Life technologies). The mRNA was used in an RT reaction that was specifically selected for the barcoded mRNA (Transcriptor High Fidelity, Roche). The RT product was PCR amplified and size selected using Agencourt AMPure beads (Beckman Coulter), then checked for quality and size on the 2100 Bioanalyzer (Agilent) and sent for sequencing on the NovaSeq S4 PE100 mode (Illumina). Three biological replicates were sent for sequencing.

The DNA was extracted by mixing the phenol-chloroform and interphase of TriZol extraction with 500uL of Back Extraction Buffer (4M guanidine thiocyanate, 50mM sodium citrate, and 1M Tris-base). DNA was treated with RnaseA (Thermo Fisher). DNA was cleaned up with phenol:chloroform:isoamyl alcohol (25:24:1) (Life Technologies). The DNA was PCR amplified and size selected using Agencourt AMPure beads (Beckman Coulter), then checked for quality and size on the 2100 Bioanalyzer (Agilent) and sent for sequencing on the NovaSeq S4 PE100 mode (Illumina). Three biological replicates were sent for sequencing.

Counting Embryos

For each experiment, once embryos had been mounted on slides, slide labels were covered with thick tape and randomly numbered by a laboratory member not involved in this project. Expression of GFP within embryos on each slide was counted blind. In each experiment, all comparative constructs were present, along with a slide with BraS as a reference. The X-Cite was turned on for 1hr before analysis to ensure the illumination intensity was constant. TO determine levels of expression, high expression was set as visible with less than 25% power on X-Cite illuminator. Fifty embryos were counted for each biological replicate.

Acquisition of Images

For enhancers being compared, images were taken from electroporations performed on the same day using identical settings. For representative images, embryos were chosen that represented the average from counting data. All images are subsequently cropped to an appropriate size. In each figure, the same exposure time for each image is shown to allow direct comparison.

Identification of Putative Notochord Enhancers

We developed a script that allows for the input of any organism's genome in the fasta file format. The script first looks for an exact match of one of seven canonical Zic family binding sites and their reverse complements. We used the following sites in our search: CAGCTGTG (Zic1/2/3), CCGCAGT (Zic7/3/1), CCGCAGTC (Zic6), CCCGCTGTG (Zic1), CCAGCTGTG (Zic3), CCGCTGTG (Zic2/ZicC), and CCCGCAGTC (Zic5) as these have been identified as functional in previous studies (Matsumoto et al., 2007; Yagi et al., 2004). Next, we drew a window of 30 bp from either end of the canonical Zic family binding site and determine if there are at least two Ets binding site cores (i.e., either GGAA or GGAT and their respective reverse complement sequences) present within the window. The location of all regions containing at least a single Zic family binding site and two Ets binding sites are saved as part of the genome search.

Scoring Relative Affinities of Binding Sites

We calculated the relative ETS binding affinity using the median signal intensity of the universal protein binding microarray (PBM) data for mouse Ets-1 proteins from the UniProbe database (http://thebrain.bwh.harvard.edu/uniprobe/index.php) (Hume et al., 2015). Previous studies have shown that the specificity of ETS family members is highly conserved even from flies to humans (Nitta et al., 2015; Wei et al., 2010), and thus ETS-1 is a good proxy for binding affinity in *Ciona* ETS-1 which has a conserved DNA binding domain (Farley et al., 2015). The relative affinity score represents the fractional binding of median signal intensities of the native 8-mer motifs compared to the optimal 8-mer motifs for optimal Ets, which we defined as the CCGGAAGT motif and its corresponding reverse complement.

Enhancer to Barcode Assignment & Dictionary Analysis

We constructed a dictionary of unique barcode tag-enhancer pairs by not allowing for any mismatches in the ~68 bp enhancers in our library and by not allowing barcode tag-enhancer pairs to have a read count of fewer than 150 reads. Additionally, we required all barcode tags to be 29 bp or 30 bp in length. If more than one barcode tag was associated with a single enhancer, we included all associated barcode tags that met the aforementioned barcode length and read count requirements. Within our dictionary, we did not find barcode tags that were matched to multiple enhancers. In total, the dictionary contains 90 enhancers that were uniquely mapped to one or more barcode tags, and a total of 640 barcode tag-enhancer pairs.

SEL-Seq Data Analysis

For the whole embryo library, we sequenced barcode tags from the DNA and RNA libraries on the Illumina HiSeq 4000. Reads that perfectly matched barcode tags in our barcode tag-enhancer dictionary were included in the subsequent analysis.

We extracted all of the read sequences from the sequencing libraries and collapse them based on unique sequences, tabulating the number of times a unique sequence appears in the library. Next, we perform preliminary filtering on the unique sequences, filtering out sequences that (i) have N's present, (ii) are missing the GFP sequence after our expected location of the barcode tag, (iii) contain a barcode that is not an exact match to our enhancer-barcode tag dictionary, (iv) did not meet the minimum read cutoff of 25 reads. For the preliminary filtering step, all DNA and RNA libraries were processed separately.

We normalize our data into RPM. We filter our data to only include the set of barcode tags and enhancers that appear in DNA across all replicates and consolidate the expression for each enhancer by taking the average RPM value across barcode tags. For determining if an enhancer was active, we calculated an "enhancer activity score." This score is calculated by averaging the log2(RNA/DNA) value across a given enhancer's biological replicates.

QUANTIFICATION AND STATISTICAL ANALYSES

To assess statistical differences between enhancer expression, Fischer's exact test was used with the fisher.test function in R. To assess statistical differences between enhancer expression levels, chi-squared test was used with the CHISQ.TEST function in Excel.

Supplementary figures:

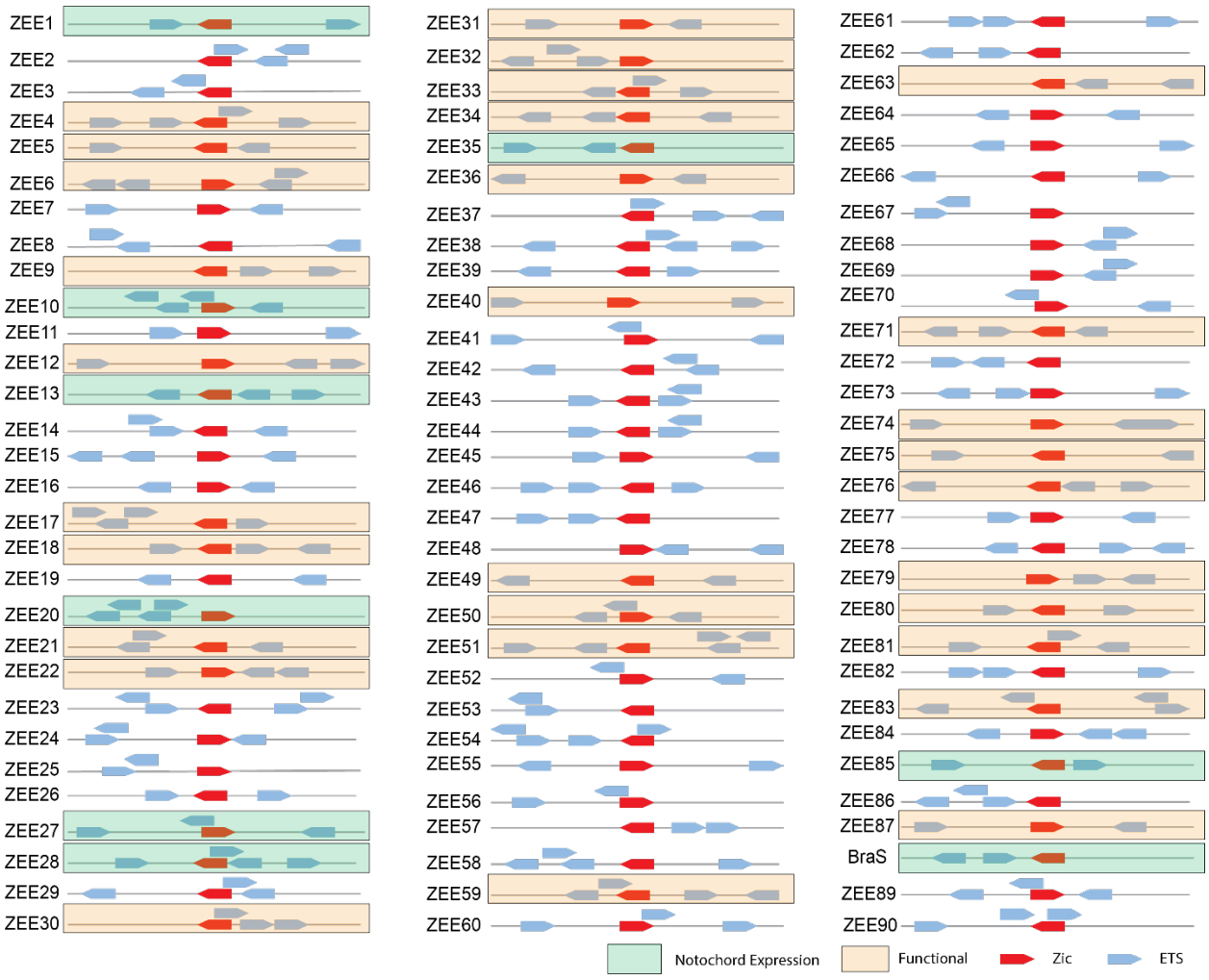


Figure S1. ZEE elements screened. Schematic of each ZEE element tested within our MPRA assay. Zic sites are colored red and ETS sites are colored blue. ZEE elements that were functional are boxed in orange. ZEE elements that drove notochord expression are boxed in green.

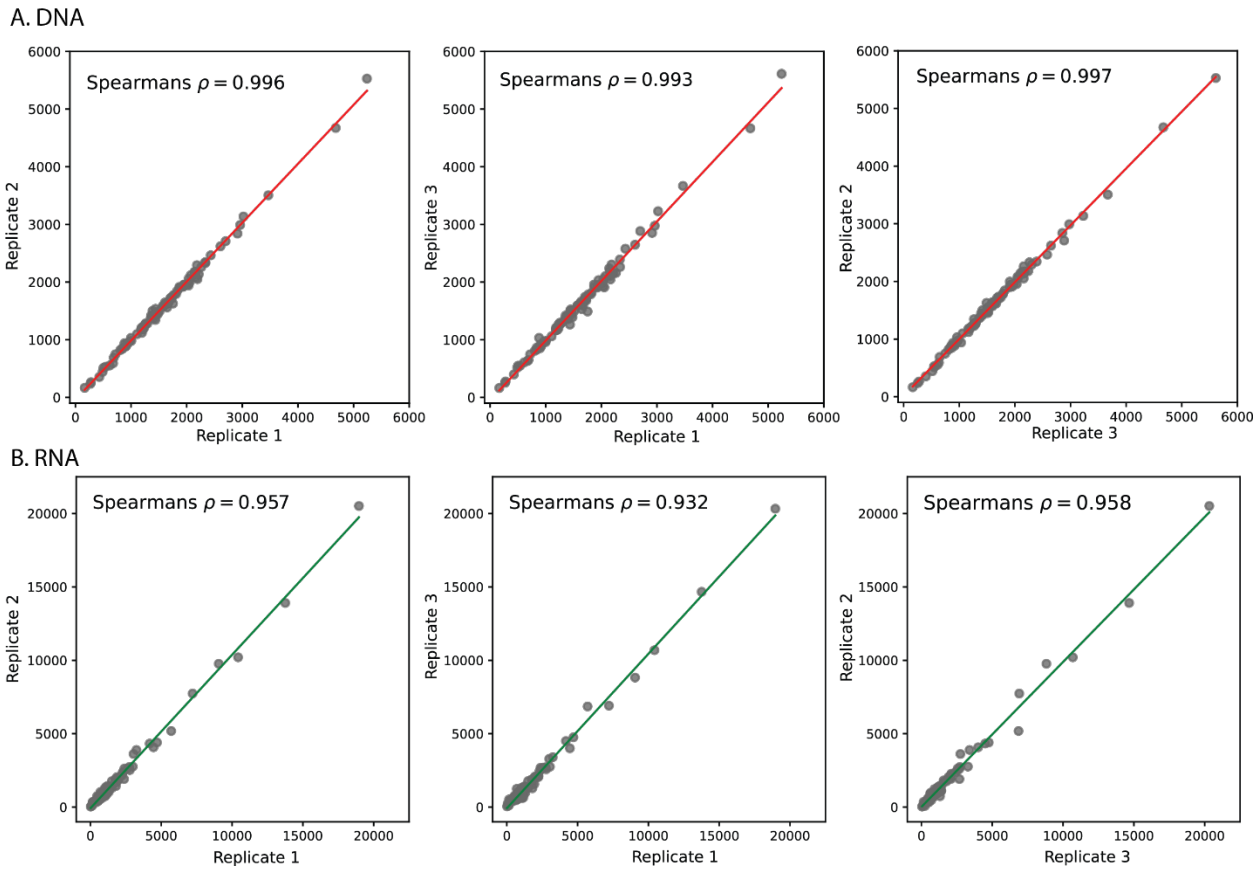


Figure S2. Data quality metrics illustrate high robustness of ZEE genomic screen. A. Correlation of DNA plasmids detected between replicates was plotted. All Spearman correlations between replicates were >0.99. **B.** Correlation of mRNA barcodes detected between replicates was plotted. All Spearman correlations between replicates were >0.9.

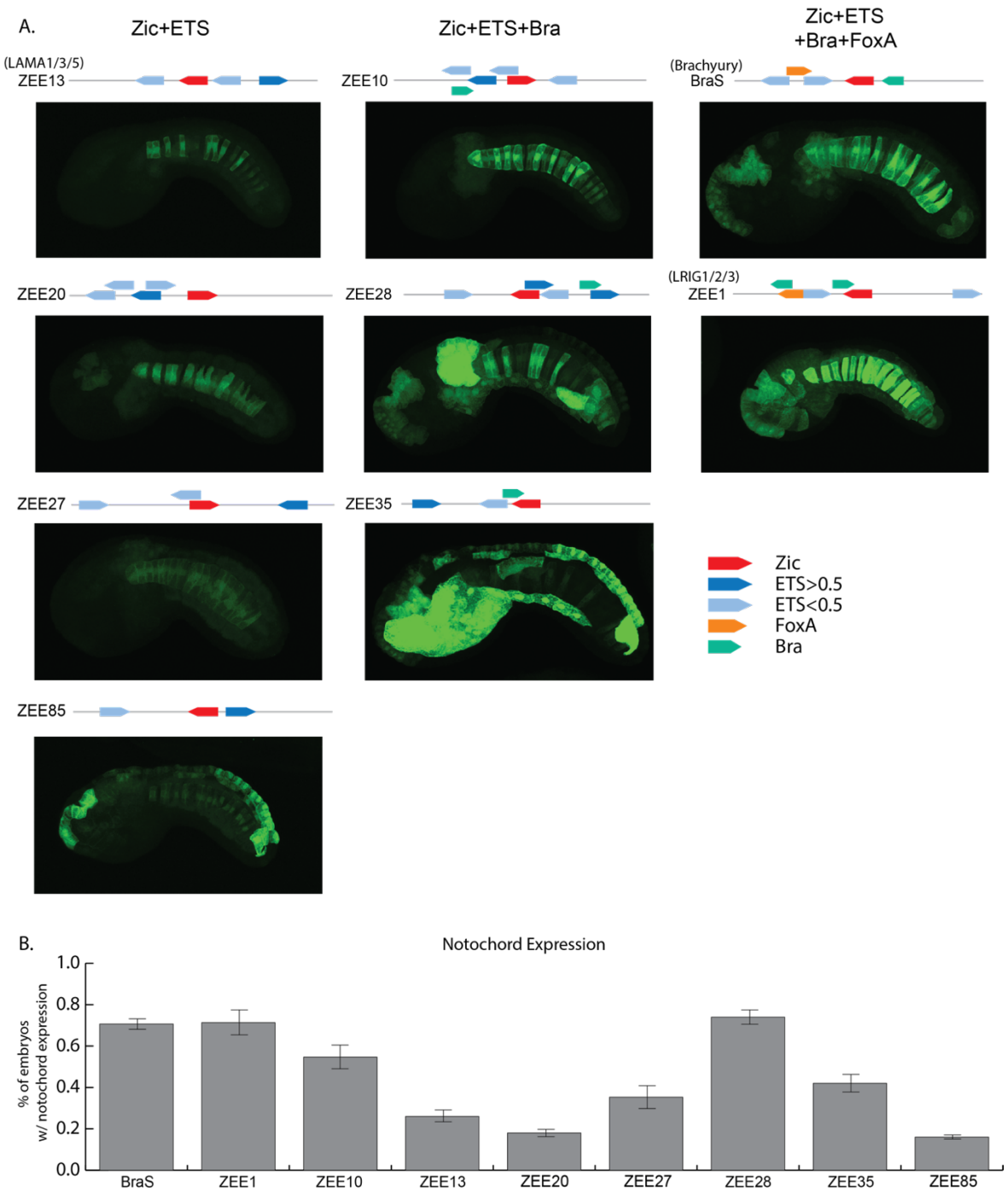


Figure S3. Nine ZEE elements drive notochord expression. A. Images and schematics of the nine notochord enhancers in the ZEE library. Zic (red), ETS (blue), FoxA (orange), and Bra sites (green) are annotated. Dark blue ETS sites have an affinity of greater than 0.5, light blue sites have an affinity of less than 0.5. **B.** Counting data for nine ZEE elements showing the % of embryos with notochord expression. Three biological replicates were performed with 50 embryos per replicate analyzed.

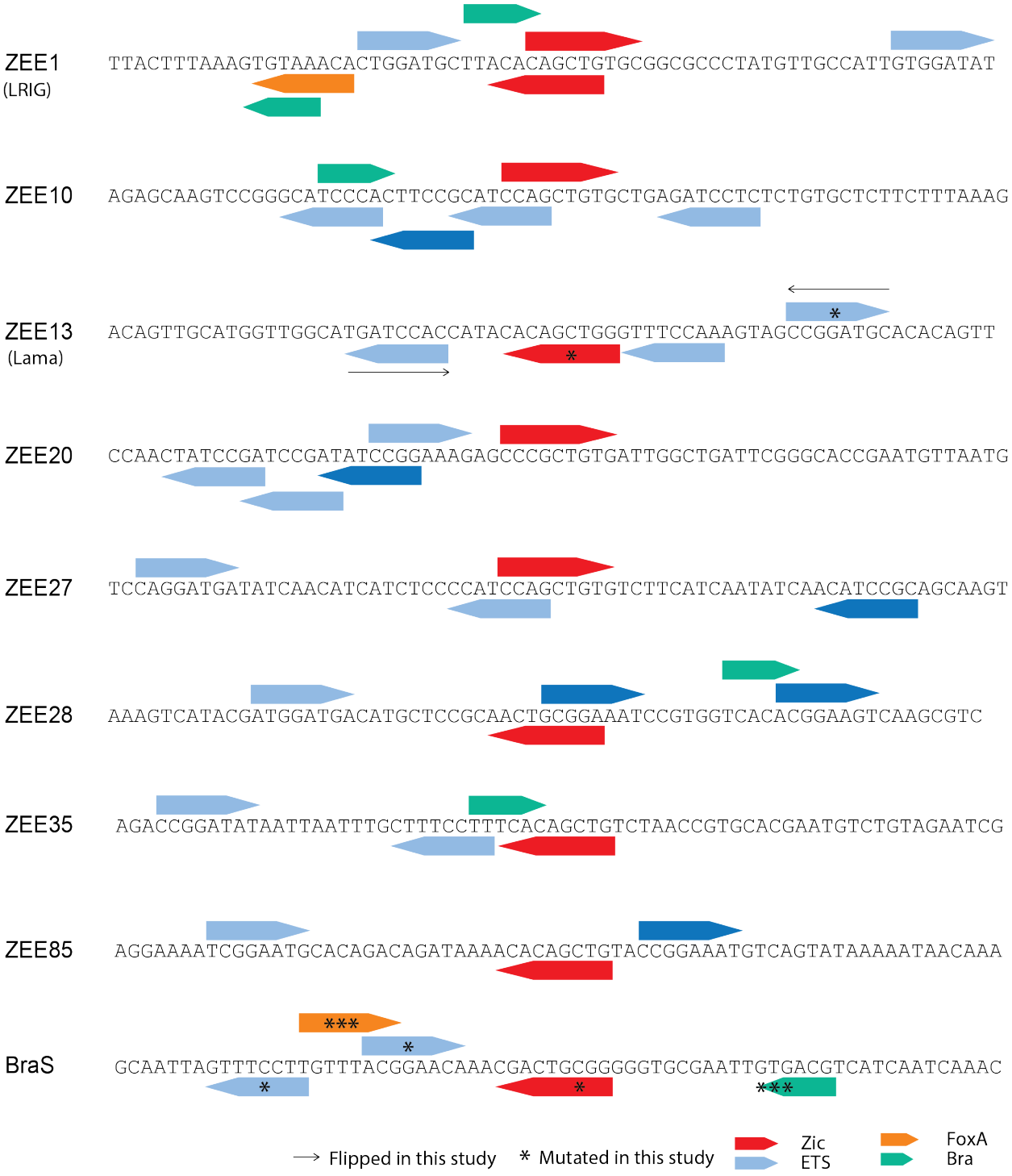
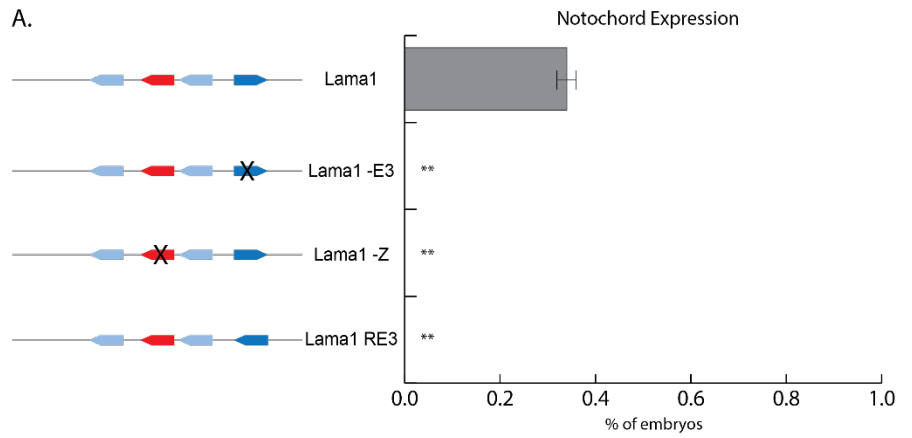
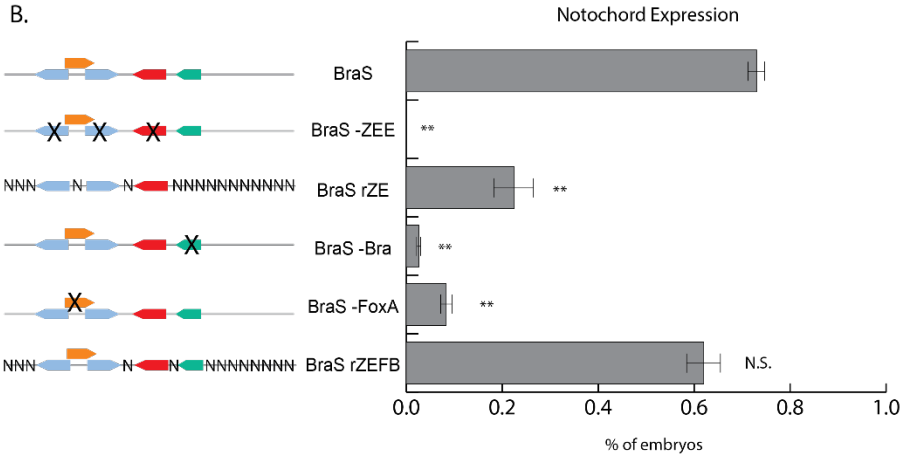
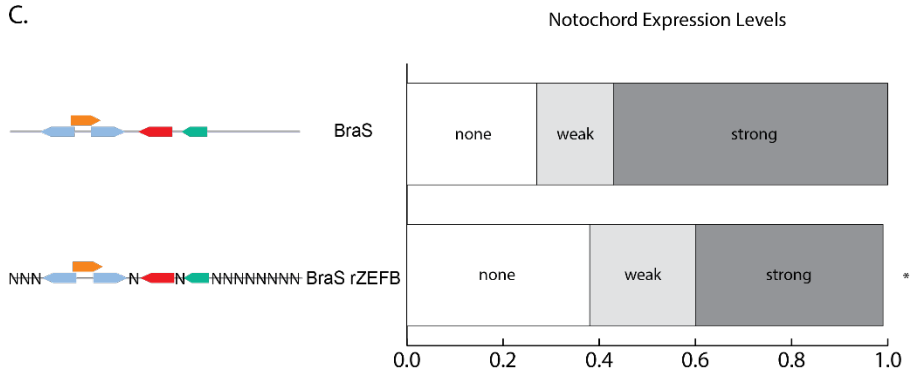


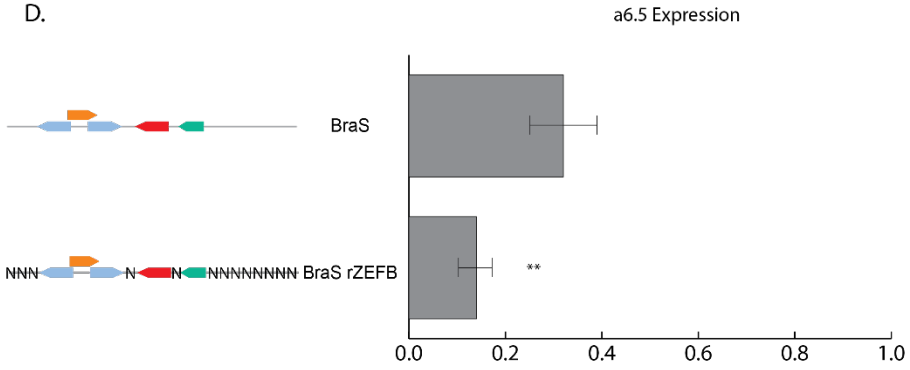
Figure S4. Annotated sequences of the nine ZEE elements that drive notochord expression. Zic (red), ETS (blue), FoxA (orange), and Bra sites (green) are annotated. Asterisk denotes nucleotide that was mutated in this study, arrow denotes a binding site that was flipped. Dark blue ETS sites have an affinity of greater than 0.5, light blue sites have an affinity of less than 0.5.

Figure S5. Scoring of manipulated notochord enhancers. A. Scoring of notochord expression for embryos electroporated with the *laminin alpha* (Lama) enhancer, Lama -E3, Lama -Z, and Lama RE3. Lama -E3, Lama -Z, and Lama RE3 all show no notochord expression. **B.** Scoring of notochord expression for embryos electroporated with Bra Shadow (BraS), BraS -ZEE, BraS rZE, BraS -Bra, BraS – FoxA, and BraS rZEFB. BraS -ZEE, BraS rZE, BraS -Bra, and BraS –FoxA all show statistically significant less notochord expression compared to BraS, while BraS rZEFB is not significantly different. **C.** Scoring of levels of expression in the notochord for embryos electroporated with BraS and BraS rZEFB. BraS rZEFB shows less notochord expression levels compared to BraS **D.** Scoring of a6.5 expression for embryos electroporated with BraS and BraS rZEFB. BraS rZEFB shows statistically significant less a6.5 expression compared to BraS. P values calculated by chi-squared test for expression levels and Fischer's exact test for all other comparisons, *$P<0.05$, ** $P < 0.01$. Dark blue ETS sites have an affinity of greater than 0.5, light blue sites have an affinity of less than 0.5. For counting data in figure A we conducted three biological repeats analyzing 50 embryos per replicate. For counting data shown in B, C and D we conducted two biological repeats analyzing 50 embryos per replicate.


A.
Notochord Expression
Lama1
Lama1 -E3
Lama1 -Z
Lama1 RE3
**
**
**
0.0
0.2
0.4
0.6
0.8
1.0
% of embryos


B.
Notochord Expression
BraS
BraS -ZEE
BraS rZE
BraS -Bra
BraS -FoxA
BraS rZEFB
NNN N N NNNNNNNNNN
NNN N N NNNNNNNN
**
**
**
N.S.
0.0
0.2
0.4
0.6
0.8
1.0
% of embryos


C.
Notochord Expression Levels
BraS
BraS rZEFB
NNN N N NNNNNNNN
none
weak
strong
none
weak
strong
*
0.0
0.2
0.4
0.6
0.8
1.0


D.
a6.5 Expression
BraS
BraS rZEFB
NNN N N NNNNNNNN
**
0.0
0.2
0.4
0.6
0.8
1.0

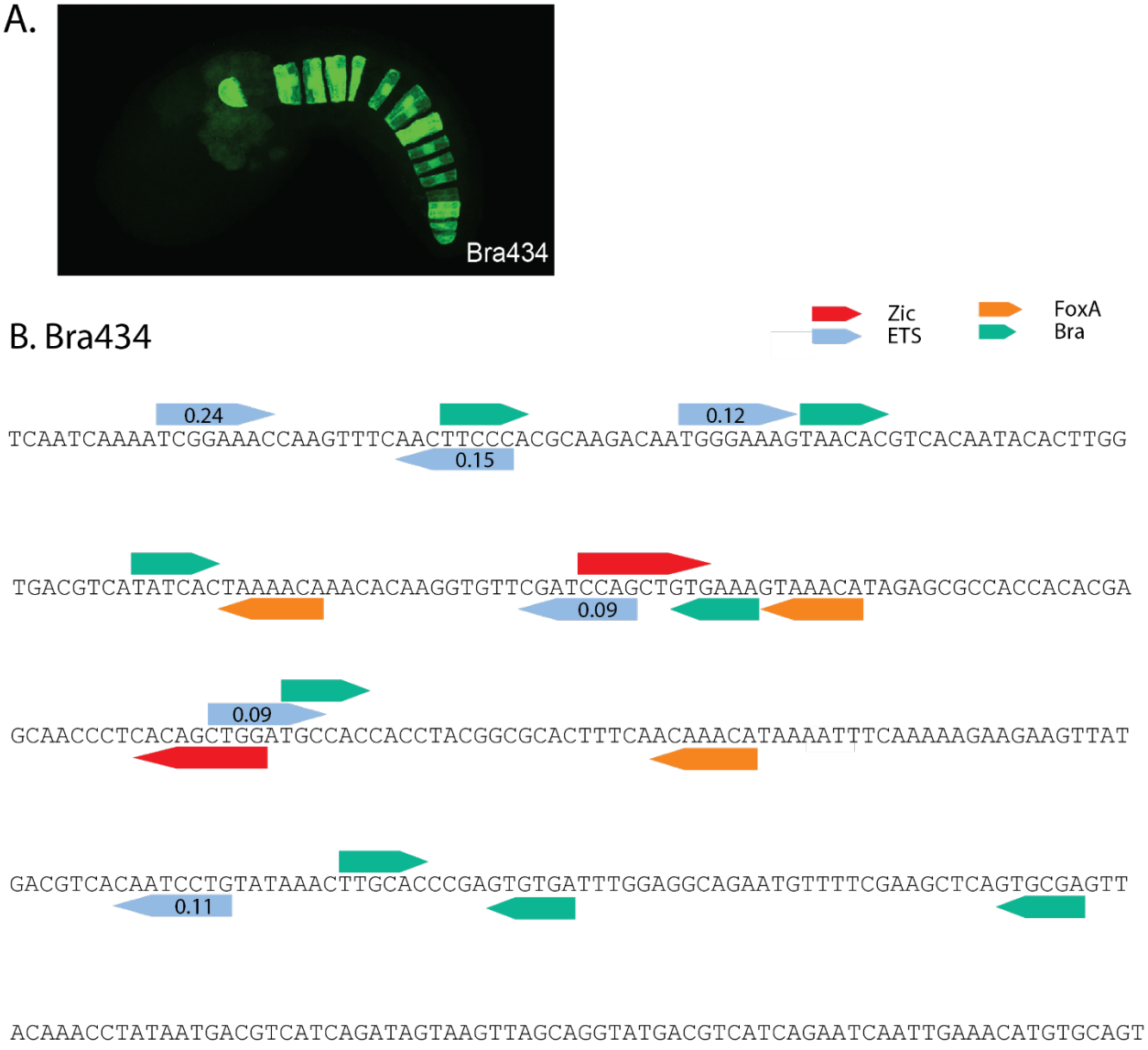


Figure S6. Updated annotation of Bra434. **A.** Image of Bra434 electroporated into Ciona embryo. **B.** Annotation of the Bra434 using PBM, EMSA, and crystal structure data. Zic sites in red, ETS sites in light blue, FoxA sites in orange, and Bra sites in green. Affinities of ETS calculated from PBM data (Wei et al., 2010) are labeled.

Supplementary Note S1: Expression patterns of ZEE elements driving notochord expression.

Levels of expression for notochord-specific enhancers:

There are four notochord-specific enhancers (ZEE10, 13, 20, and 27). The strongest of these is ZEE10, which is the only ZEE element in this group to contain a Bra site in addition to the Zic and ETS sites. We speculate that this additional Bra binding site could maintain and amplify the signal in a positive, feed-forward loop. ZEE13, 20, and 27 are all similar in their levels of expression, and we speculate that these enhancers have an organization of Zic and ETS sites that are permissive to notochord expression.

BraS and ZEE1 a6.5 expression:

BraS and ZEE1 have strong notochord expression, but also a6.5 expression. Zic and ETS are co-expressed in the a6.5 and notochord cell lineages; thus, we think that the a6.5 expression seen in these constructs could be due to an organization of sites permissive to both neural and notochord expression. ZEE1 also has head endoderm expression, which could be due to the expression of FoxA and ETS in the endoderm or other sites that we have yet to identify. The randomization of BraS rZEFB leads to a reduction in the number of embryos with a6.5 expression; this indicates that other sequences beyond the identified sites contribute to the a6.5 expression.

ZEE35 and 85 have weak notochord expression with stronger ectopic expression:

ZEE35 and 85 both have weak notochord and stronger expression in other domains. ZEE85 drives strong expression in the b6.5 nerve cord; this expression could be due to ETS sites working in combination with other sites within the enhancer. ZEE35 drives strong expression in the endoderm, nerve cord, and a6.5 lineage. We speculate that this enhancer may contain an organization of sites that is optimal for binding of ETS in the endoderm and Zic and ETS in the a6.5 lineage. Bra is thought to be able to act as an activator or repressor, so the notochord expression may be dampened

by Bra acting as a repressor in ZEE35. It is also possible that the organization of sites within these enhancers are not optimal for notochord expression, but more optimal for other domains of expression.

References

Ang, S.-L., Rossant, J., 1994. HNF-3β is essential for node and notochord formation in mouse development. Cell 78, 561–574. https://doi.org/10.1016/0092-8674(94)90522-3

Antosova, B., Smolikova, J., Klimova, L., Lachova, J., Bendova, M., Kozmikova, I., Machon, O., Kozmik, Z., 2016. The Gene Regulatory Network of Lens Induction Is Wired through Meis-Dependent Shadow Enhancers of Pax6. PLoS Genet. 12, e1006441. https://doi.org/10.1371/journal.pgen.1006441

Arnone, M.I., Davidson, E.H., 1997. The hardwiring of development: organization and function of genomic regulatory systems. Dev. Camb. Engl. 124, 1851–1864. https://doi.org/10.1242/dev.124.10.1851

Barnett, M.W., Old, R.W., Jones, E.A., 1998. Neural induction and patterning by fibroblast growth factor, notochord and somite tissue in Xenopus. Dev. Growth Differ. 40, 47–57. https://doi.org/10.1046/j.1440-169X.1998.t01-5-00006.x

Barolo, S., 2016. How to tune an enhancer. Proc. Natl. Acad. Sci. 113, 6330–6331. https://doi.org/10.1073/pnas.1606109113

Casey, E.S., O'Reilly, M.A., Conlon, F.L., Smith, J.C., 1998. The T-box transcription factor Brachyury regulates expression of eFGF through binding to a non-palindromic response element. Development 125, 3887–3894. https://doi.org/10.1242/dev.125.19.3887

Chesley, P., 1935. Development of the short-tailed mutant in the house mouse. J. Exp. Zool. 70, 429–459. https://doi.org/10.1002/jez.1400700306

Chiba, S., Jiang, D., Satoh, N., Smith, W.C., 2009. brachyury null mutant-induced defects in juvenile ascidian endodermal organs. Development 136, 35–39. https://doi.org/10.1242/dev.030981

Conlon, F.L., Fairclough, L., Price, B.M.J., Casey, E.S., Smith, J.C., 2001. Determinants of T box protein specificity. Development 128, 3749–3758. https://doi.org/10.1242/dev.128.19.3749

Corbo, J.C., Levine, M., Zeller, R.W., 1997. Characterization of a notochord-specific enhancer from the Brachyury promoter region of the ascidian, Ciona intestinalis. Dev. Camb. Engl. 124, 589–602. https://doi.org/10.1242/dev.124.3.589

Dal-Pra, S., Thisse, C., Thisse, B., 2011. FoxA transcription factors are essential for the development of dorsal axial structures. Dev. Biol. 350, 484–495. https://doi.org/10.1016/j.ydbio.2010.12.018

Davidson, B., Christiaen, L., 2006. Linking Chordate Gene Networks to Cellular Behavior in Ascidians. Cell 124, 247–250. https://doi.org/10.1016/j.cell.2006.01.013

de Boer, C.G., Vaishnav, E.D., Sadeh, R., Abeyta, E.L., Friedman, N., Regev, A., 2020. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. Nat. Biotechnol. 38, 56–65. https://doi.org/10.1038/s41587-019-0315-8

Delsuc, F., Brinkmann, H., Chourrout, D., Philippe, H., 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439, 965–968. https://doi.org/10.1038/nature04336

Di Gregorio, A., 2020. The notochord gene regulatory network in chordate evolution: Conservation and divergence from Ciona to vertebrates. Curr. Top. Dev. Biol. 139, 325–374. https://doi.org/10.1016/bs.ctdb.2020.01.002

Di Gregorio, A., Levine, M., 1999. Regulation of Ci-tropomyosin-like, a Brachyury target gene in the ascidian, Ciona intestinalis. Dev. Camb. Engl. 126, 5599–5609. https://doi.org/10.1242/dev.126.24.5599

Dunn, M.P., Di Gregorio, A., 2009. The evolutionarily conserved leprecan gene: Its regulation by Brachyury and its role in the developing Ciona notochord. Dev. Biol. 328, 561–574. https://doi.org/10.1016/j.ydbio.2009.02.007

Dykes, I.M., Szumska, D., Kuncheria, L., Puliyadi, R., Chen, C., Papanayotou, C., Lockstone, H., Dubourg, C., David, V., Schneider, J.E., Keane, T.M., Adams, D.J., Brown, S.D.M., Mercier, S., Odent, S., Collignon, J., Bhattacharya, S., 2018. A Requirement for Zic2 in the Regulation of Nodal Expression Underlies the Establishment of Left-Sided Identity. Sci. Rep. 8, 10439. https://doi.org/10.1038/s41598-018-28714-1

Elms, P., Scurry, A., Davies, J., Willoughby, C., Hacker, T., Bogani, D., Arkell, R., 2004. Overlapping and distinct expression domains of Zic2 and Zic3 during mouse gastrulation. Gene Expr. Patterns 4, 505–511. https://doi.org/10.1016/j.modgep.2004.03.003

Farley, E.K., Olson, K.M., Zhang, W., Brandt, A.J., Rokhsar, D.S., Levine, M.S., 2015. Suboptimization of developmental enhancers. Science 350, 325–328. https://doi.org/10.1126/science.aac6948

Farley, E.K., Olson, K.M., Zhang, W., Rokhsar, D.S., Levine, M.S., 2016. Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. Proc. Natl. Acad. Sci. 113, 6508–6513. https://doi.org/10.1073/pnas.1605085113

Frankel, N., Davis, G.K., Vargas, D., Wang, S., Payre, F., Stern, D.L., 2010. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. Nature 466, 490–493. https://doi.org/10.1038/nature09158

Fujiwara, S., Corbo, J.C., Levine, M., 1998. The snail repressor establishes a muscle/notochord boundary in the Ciona embryo. Development 125, 2511–2520. https://doi.org/10.1242/dev.125.13.2511

Fuqua, T., Jordan, J., van Breugel, M.E., Halavatyi, A., Tischer, C., Polidoro, P., Abe, N., Tsai, A., Mann, R.S., Stern, D.L., Crocker, J., 2020. Dense and pleiotropic regulatory information in a developmental enhancer. Nature 587, 235–239. https://doi.org/10.1038/s41586-020-2816-5

Galupa, R., Alvarez-Canales, G., Borst, N.O., Fuqua, T., Gandara, L., Misunou, N., Richter, K., Alves, M.R.P., Karumbi, E., Perkins, M.L., Kocijan, T., Rushlow, C.A., Crocker, J., 2022. Enhancer architecture and chromatin accessibility constrain phenotypic space during development. bioRxiv 2022.06.02.494376. https://doi.org/10.1101/2022.06.02.494376

Harvey, S.A., Tümpel, S., Dubrulle, J., Schier, A.F., Smith, J.C., 2010. no tail integrates two modes of mesoderm induction. Development 137, 1127–1135. https://doi.org/10.1242/dev.046318

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., Glass, C.K., 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 38, 576–589. https://doi.org/10.1016/j.molcel.2010.05.004

Herrmann, B.G., Kispert, A., 1994. The T genes in embryogenesis. Trends Genet. TIG 10, 280–286. https://doi.org/10.1016/0168-9525(90)90011-t

Hong, J.-W., Hendrix, D.A., Levine, M.S., 2008. Shadow enhancers as a source of evolutionary novelty. Science 321, 1314. https://doi.org/10.1126/science.1160631

Hudson, C., Lotito, S., Yasuo, H., 2007. Sequential and combinatorial inputs from Nodal, Delta2/Notch and FGF/MEK/ERK signalling pathways establish a grid-like organisation of distinct cell identities in the ascidian neural plate. Dev. Camb. Engl. 134, 3527–3537. https://doi.org/10.1242/dev.002352

Hudson, C., Sirour, C., Yasuo, H., 2016. Co-expression of Foxa.a, Foxd and Fgf9/16/20 defines a transient mesendoderm regulatory state in ascidian embryos. eLife 5, e14692. https://doi.org/10.7554/eLife.14692

Hume, M.A., Barrera, L.A., Gisselbrecht, S.S., Bulyk, M.L., 2015. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein–DNA interactions. Nucleic Acids Res. 43, D117–D122. https://doi.org/10.1093/nar/gku1045

Ikeda, T., Satou, Y., 2016. Differential temporal control of *Foxa.a* and *Zic-r.b* specifies brain versus notochord fate in the ascidian embryo. Development dev.142174. https://doi.org/10.1242/dev.142174

Imai, K.S., Hino, K., Yagi, K., Satoh, N., Satou, Y., 2004. Gene expression profiles of transcription factors and signaling molecules in the ascidian embryo: towards a comprehensive understanding of gene networks. Development 131, 4047–4058. https://doi.org/10.1242/dev.01270

Imai, K.S., Levine, M., Satoh, N., Satou, Y., 2006. Regulatory Blueprint for a Chordate Embryo. Science 312, 1183–1187. https://doi.org/10.1126/science.1123404

Imai, K.S., Satoh, N., Satou, Y., 2002a. Early embryonic expression of *FGF4/6/9* gene and its role in the induction of mesenchyme and notochord in *Ciona savignyi* embryos. Development 129, 1729–1738. https://doi.org/10.1242/dev.129.7.1729

Imai, K.S., Satou, Y., Satoh, N., 2002b. Multiple functions of a Zic-like gene in the differentiation of notochord, central nervous system and muscle in Ciona savignyi embryos. Dev. Camb. Engl. 129, 2723–2732. https://doi.org/10.1242/dev.129.11.2723

Jiang, D., Smith, W.C., 2007. Ascidian notochord morphogenesis. Dev. Dyn. Off. Publ. Am. Assoc. Anat. 236, 1748–1757. https://doi.org/10.1002/dvdy.21184

Jindal, G.A., Farley, E.K., 2021. Enhancer grammar in development, evolution, and disease: dependencies and interplay. Dev. Cell 56, 575–587. https://doi.org/10.1016/j.devcel.2021.02.016

José-Edwards, D.S., Oda-Ishii, I., Kugler, J.E., Passamaneck, Y.J., Katikala, L., Nibu, Y., Di Gregorio, A., 2015. Brachyury, Foxa2 and the cis-Regulatory Origins of the Notochord. PLoS Genet. 11, e1005730. https://doi.org/10.1371/journal.pgen.1005730

Katikala, L., Aihara, H., Passamaneck, Y.J., Gazdoiu, S., José-Edwards, D.S., Kugler, J.E., Oda-Ishii, I., Imai, J.H., Nibu, Y., Di Gregorio, A., 2013. Functional Brachyury binding sites establish a temporal read-out of gene expression in the Ciona notochord. PLoS Biol. 11, e1001697. https://doi.org/10.1371/journal.pbio.1001697

King, D.M., Hong, C.K.Y., Shepherdson, J.L., Granas, D.M., Maricque, B.B., Cohen, B.A., 2020. Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in mouse embryonic stem cells. eLife 9, e41279. https://doi.org/10.7554/eLife.41279

Kumano, G., Yamaguchi, S., Nishida, H., 2006. Overlapping expression of FoxA and Zic confers responsiveness to FGF signaling to specify notochord in ascidian embryos. Dev. Biol. 300, 770–784. https://doi.org/10.1016/j.ydbio.2006.07.033

Lamber, E.P., Vanhille, L., Textor, L.C., Kachalova, G.S., Sieweke, M.H., Wilmanns, M., 2008. Regulation of the transcription factor Ets-1 by DNA-mediated homo-dimerization. EMBO J. 27, 2006–2017. https://doi.org/10.1038/emboj.2008.117

Levine, M., 2010. Transcriptional Enhancers in Animal Development and Evolution. Curr. Biol. 20, R754–R763. https://doi.org/10.1016/j.cub.2010.06.070
Levo, M., Segal, E., 2014. In pursuit of design principles of regulatory sequences. Nat. Rev. Genet. 15, 453–468. https://doi.org/10.1038/nrg3684

Li, J., Dantas Machado, A.C., Guo, M., Sagendorf, J.M., Zhou, Z., Jiang, L., Chen, X., Wu, D., Qu, L., Chen, Z., Chen, L., Rohs, R., Chen, Y., 2017. Structure of the Forkhead Domain of FOXA2 Bound to a Complete DNA Consensus Site. Biochemistry 56, 3745–3753. https://doi.org/10.1021/acs.biochem.7b00211

Liu, F., Posakony, J.W., 2012. Role of Architecture in the Function and Specificity of Two Notch-Regulated Transcriptional Enhancer Modules. PLoS Genet. 8, e1002796. https://doi.org/10.1371/journal.pgen.1002796

Lolas, M., Valenzuela, P.D.T., Tjian, R., Liu, Z., 2014. Charting Brachyury-mediated developmental pathways during early mouse embryogenesis. Proc. Natl. Acad. Sci. 111, 4478–4483. https://doi.org/10.1073/pnas.1402612111

Machingo, Q.J., Fritz, A., Shur, B.D., 2006. A beta1,4-galactosyltransferase is required for convergent extension movements in zebrafish. Dev. Biol. 297, 471–482. https://doi.org/10.1016/j.ydbio.2006.05.024

Matsumoto, J., Kumano, G., Nishida, H., 2007. Direct activation by Ets and Zic is required for initial expression of the Brachyury gene in the ascidian notochord. Dev. Biol. 306, 870–882. https://doi.org/10.1016/j.ydbio.2007.03.034

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A.K., Canfield, T.K., Giste, E., Diegel, M., Bates, D., Hansen, R.S., Neph, S., Sabo, P.J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S.R., Kaul, R., Stamatoyannopoulos, J.A., 2012. Systematic localization of common disease-associated variation in regulatory DNA. Science 337, 1190–1195. https://doi.org/10.1126/science.1222794

Miya, T., Nishida, H., 2003. An Ets transcription factor, HrEts, is target of FGF signaling and involved in induction of notochord, mesenchyme, and brain in ascidian embryos. Dev. Biol. 261, 25–38. https://doi.org/10.1016/s0012-1606(03)00246-x

Müller, C.W., Herrmann, B.G., 1997. Crystallographic structure of the T domain–DNA complex of the Brachyury transcription factor. Nature 389, 884–888. https://doi.org/10.1038/39929

Nitta, K.R., Jolma, A., Yin, Y., Morgunova, E., Kivioja, T., Akhtar, J., Hens, K., Toivonen, J., Deplancke, B., Furlong, E.E.M., Taipale, J., 2015. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. eLife 4, e04837. https://doi.org/10.7554/eLife.04837

Olivera-Martinez, I., Harada, H., Halley, P.A., Storey, K.G., 2012. Loss of FGF-Dependent Mesoderm Identity and Rise of Endogenous Retinoid Signalling Determine Cessation of Body Axis Elongation. PLoS Biol. 10, e1001415. https://doi.org/10.1371/journal.pbio.1001415

Osterwalder, M., Barozzi, I., Tissières, V., Fukuda-Yuzawa, Y., Mannion, B.J., Afzal, S.Y., Lee, E.A., Zhu, Y., Plajzer-Frick, I., Pickle, C.S., Kato, M., Garvin, T.H., Pham, Q.T., Harrington, A.N., Akiyama, J.A., Afzal, V., Lopez-Rios, J., Dickel, D.E., Visel, A., Pennacchio, L.A., 2018. Enhancer redundancy provides phenotypic robustness in mammalian development. Nature 554, 239–243. https://doi.org/10.1038/nature25461

Parsons, M.J., Campos, I., Hirst, E.M.A., Stemple, D.L., 2002. Removal of dystroglycan causes severe muscular dystrophy in zebrafish embryos. Dev. Camb. Engl. 129, 3505–3512. https://doi.org/10.1242/dev.129.14.3505

Passamaneck, Y.J., Katikala, L., Perrone, L., Dunn, M.P., Oda-Ishii, I., Di Gregorio, A., 2009. Direct activation of a notochord cis-regulatory module by Brachyury and FoxA in the ascidian Ciona intestinalis. Dev. Camb. Engl. 136, 3679–3689. https://doi.org/10.1242/dev.038141

Perry, M.W., Boettiger, A.N., Bothma, J.P., Levine, M., 2010. Shadow enhancers foster robustness of Drosophila gastrulation. Curr. Biol. CB 20, 1562–1567. https://doi.org/10.1016/j.cub.2010.07.043

Picco, V., Hudson, C., Yasuo, H., 2007. Ephrin-Eph signalling drives the asymmetric division of notochord/neural precursors in Ciona embryos. Dev. Camb. Engl. 134, 1491–1497. https://doi.org/10.1242/dev.003939

Pijuan-Sala, B., Wilson, N.K., Xia, J., Hou, X., Hannah, R.L., Kinston, S., Calero-Nieto, F.J., Poirion, O., Preissl, S., Liu, F., Göttgens, B., 2020. Single-cell chromatin accessibility maps reveal regulatory programs driving early mouse organogenesis. Nat. Cell Biol. 22, 487–497. https://doi.org/10.1038/s41556-020-0489-9

Pollard, S.M., Parsons, M.J., Kamei, M., Kettleborough, R.N.W., Thomas, K.A., Pham, V.N., Bae, M.-K., Scott, A., Weinstein, B.M., Stemple, D.L., 2006. Essential and overlapping roles for laminin alpha chains in notochord and blood vessel formation. Dev. Biol. 289, 64–76. https://doi.org/10.1016/j.ydbio.2005.10.006

Reeves, W.M., Shimai, K., Winkley, K.M., Veeman, M.T., 2021. Brachyury controls Ciona notochord fate as part of a feed-forward network. Dev. Camb. Engl. 148, dev195230. https://doi.org/10.1242/dev.195230

Reeves, W.M., Wu, Y., Harder, M.J., Veeman, M.T., 2017. Functional and evolutionary insights from the Ciona notochord transcriptome. Dev. Camb. Engl. 144, 3375–3387. https://doi.org/10.1242/dev.156174

Rothbächer, U., Bertrand, V., Lamy, C., Lemaire, P., 2007. A combinatorial code of maternal GATA, Ets and β-catenin-TCF transcription factors specifies and patterns the early ascidian ectoderm. Development 134, 4023–4032. https://doi.org/10.1242/dev.010850

Schifferl, D., Scholze-Wittler, M., Wittler, L., Veenvliet, J.V., Koch, F., Herrmann, B.G., 2021. A 37 kb region upstream of *brachyury* comprising a notochord enhancer is essential for notochord and tail development. Development 148, dev200059. https://doi.org/10.1242/dev.200059

Schulte-Merker, S., Smith, J.C., 1995. Mesoderm formation in response to Brachyury requires FGF signalling. Curr. Biol. 5, 62–67. https://doi.org/10.1016/S0960-9822(95)00017-0

Scott, A., Stemple, D.L., 2005. Zebrafish notochordal basement membrane: signaling and structure. Curr. Top. Dev. Biol. 65, 229–253. https://doi.org/10.1016/S0070-2153(04)65009-5

Shimai, K., Veeman, M., 2021. Quantitative Dissection of the Proximal Ciona brachyury Enhancer. Front. Cell Dev. Biol. 9, 804032. https://doi.org/10.3389/fcell.2021.804032

Small, S., Blair, A., Levine, M., 1992. Regulation of even-skipped stripe 2 in the Drosophila embryo. EMBO J. 11, 4047–4057. https://doi.org/10.1002/j.1460-2075.1992.tb05498.x

Spitz, F., Furlong, E.E.M., 2012. Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. 13, 613–626. https://doi.org/10.1038/nrg3207

Stemple, D.L., 2005. Structure and function of the notochord: an essential organ for chordate development. Development 132, 2503–2512. https://doi.org/10.1242/dev.01812

Stolfi, A., Ryan, K., Meinertzhagen, I.A., Christiaen, L., 2015. Migratory neuronal progenitors arise from the neural plate borders in tunicates. Nature 527, 371–374. https://doi.org/10.1038/nature15758

Swanson, C.I., Evans, N.C., Barolo, S., 2010. Structural Rules and Complex Regulatory Circuitry Constrain Expression of a Notch- and EGFR-Regulated Eye Enhancer. Dev. Cell 18, 359–370. https://doi.org/10.1016/j.devcel.2009.12.026

Tak, Y.G., Farnham, P.J., 2015. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. Epigenetics Chromatin 8, 57. https://doi.org/10.1186/s13072-015-0050-4

Takahashi, H., Mitani, Y., Satoh, G., Satoh, N., 1999. Evolutionary alterations of the minimal promoter for notochord-specific Brachyury expression in ascidian embryos. Dev. Camb. Engl. 126, 3725–3734. https://doi.org/10.1242/dev.126.17.3725

Thanos, D., Maniatis, T., 1995. Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. Cell 83, 1091–1100. https://doi.org/10.1016/0092-8674(95)90136-1

Veeman, M.T., Nakatani, Y., Hendrickson, C., Ericson, V., Lin, C., Smith, W.C., 2008. Chongmague reveals an essential role for laminin-mediated boundary formation in chordate convergence and extension movements. Dev. Camb. Engl. 135, 33–41. https://doi.org/10.1242/dev.010892

Visel, A., Rubin, E.M., Pennacchio, L.A., 2009. Genomic views of distant-acting enhancers. Nature 461, 199–205. https://doi.org/10.1038/nature08451

Wagner, E., Levine, M., 2012. FGF signaling establishes the anterior border of the Ciona neural tube. Dev. Camb. Engl. 139, 2351–2359. https://doi.org/10.1242/dev.078485
Warr, N., Powles-Glover, N., Chappell, A., Robson, J., Norris, D., Arkell, R.M., 2008. Zic2 - associated holoprosencephaly is caused by a transient defect in the organizer region during gastrulation. Hum. Mol. Genet. 17, 2986–2996. https://doi.org/10.1093/hmg/ddn197

Wei, G.-H., Badis, G., Berger, M.F., Kivioja, T., Palin, K., Enge, M., Bonke, M., Jolma, A., Varjosalo, M., Gehrke, A.R., Yan, J., Talukder, S., Turunen, M., Taipale, M., Stunnenberg, H.G., Ukkonen, E., Hughes, T.R., Bulyk, M.L., Taipale, J., 2010. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. EMBO J. 29, 2147–2160. https://doi.org/10.1038/emboj.2010.106

Weinstein, D.C., Ruiz i Altaba, A., Chen, W.S., Hoodless, P., Prezioso, V.R., Jessell, T.M., Darnell, J.E., 1994. The winged-helix transcription factor HNF-3β is required for notochord development in the mouse embryo. Cell 78, 575–588. https://doi.org/10.1016/0092-8674(94)90523-1

Wilkinson, D.G., Bhatt, S., Herrmann, B.G., 1990. Expression pattern of the mouse T gene and its role in mesoderm formation. Nature 343, 657–659. https://doi.org/10.1038/343657a0

Winkley, K.M., Reeves, W.M., Veeman, M.T., 2021. Single-cell analysis of cell fate bifurcation in the chordate Ciona. BMC Biol. 19, 180. https://doi.org/10.1186/s12915-021-01122-0

Wong, E.S., Zheng, D., Tan, S.Z., Bower, N.I., Garside, V., Vanwalleghem, G., Gaiti, F., Scott, E., Hogan, B.M., Kikuchi, K., McGlinn, E., Francois, M., Degnan, B.M., 2020. Deep conservation of the enhancer regulatory code in animals. Science 370, eaax8137. https://doi.org/10.1126/science.aax8137

Yagi, K., Satou, Y., Satoh, N., 2004. A zinc finger transcription factor, ZicL, is a direct activator of *Brachyury* in the notochord specification of *Ciona intestinalis*. Development 131, 1279–1288. https://doi.org/10.1242/dev.01011

Yasuo, H., Hudson, C., 2007. FGF8/17/18 functions together with FGF9/16/20 during formation of the notochord in Ciona embryos. Dev. Biol. 302, 92–103. https://doi.org/10.1016/j.ydbio.2006.08.075

Yasuo, H., Satoh, N., 1993. Function of vertebrate T gene. Nature 364, 582–583. https://doi.org/10.1038/364582b0

Yona, A.H., Alm, E.J., Gore, J., 2018. Random sequences rapidly evolve into de novo promoters. Nat. Commun. 9, 1530. https://doi.org/10.1038/s41467-018-04026-w

Acknowledgements

Chapter 1 contains material submitted to Cell Reports. Song BP, Ragsac MF, Tellez K, Jindal GA, Grudzien JL, Le SH, Farley EK. “Diverse logics and grammars encode notochord enhancers”. The dissertation author was the primary investigator and author of this paper. Special thanks to the Farley lab and Dennis Schifferl for helpful discussions. Special thanks to Janet H.T. Song for her critical reading of the manuscript. Special thanks to the UCSD IGM Genomics Center for their assistance with sequencing.

CHAPTER 2

Abstract

Massively parallel reporter assays (MPRAs) can quantitatively measure the function of thousands to millions of sequences at once. These are used to test putative enhancers, synthetic sequences, and enhancer variants for activity. By testing many sequences at once, MPRAs allow us to better understand how sequences of enhancers encode function. However, most MPRAs are performed in non-endogenous systems, such as cell culture, where the factors required for an enhancer to work may not fully recapitulate the *in vivo* environment. Enhancers drive tissue-specific expression using the transcription factors in those tissues. Thus, an enhancer must be tested in a variety of cell types across development to understand how an enhancer's sequence can encode the precise patterns of gene expression required for development. In this protocol, we developed an embryonic MPRA in the chicken limb bud. This is the first MPRA of putative regulatory elements done in a developing vertebrate embryo, allowing us to assay expression of genomic elements during development towards the goal of understanding how enhancer sequences can encode tissue-specific expression during vertebrate development.

Introduction

Gene transcription is controlled by regulatory elements in the genome. These regulatory elements act as switches to turn on specific gene expression during development and tissue maintenance (Levine, 2010). Nucleotide changes in these regulatory elements can have major phenotypic impacts in both disease and evolution. In disease, a large majority of all human disease genome-wide association studies show associations within noncoding variants (Manolio et al., 2009; Maurano et al., 2012; Tak and Farnham, 2015). In evolution, sequence changes in regulatory

elements have been shown to affect morphology, adaptation, and behavior (Carroll, 2015; Visel et al., 2009). For example, malaria uses the Duffy protein to enter blood cells. Mutation in the *Duffy* gene enhancer leads to loss of expression of Duffy, conferring malarial resistance (Tournamille et al., 1995).

Putative regulatory elements are often identified from genome-wide methods like chromatin immunoprecipitation followed by sequencing (ChIP-Seq, Johnson et al., 2007), DNaseI hypersensitive sites sequencing (DNase-seq, Crawford et al., 2004; Sabo et al., 2004), assay for transposition-accessible chromatin using sequencing (ATAC-seq, Buenrostro et al., 2013), cleavage under targets and release using nuclease (CUT&RUN, Skene et al., 2018), Hi-C (Lieberman-Aiden et al., 2009), and others. However, these methods can only identify putative regulatory elements, not functionally test them. Reporter assays have been historically used to characterize putative regulatory elements (Dinger and Beck-Sickinger, 2002; Hakkila et al., 2002). However, they must be tested on an individual basis, which is a slow and tedious process. More recently, massively parallel reporter assays have been used to simultaneously test thousands to millions of these putative regulatory elements (Arnold et al., 2014; de Boer et al., 2020; Farley et al., 2015; Gordon et al., 2020; King et al., 2020). They measure either RNA expression of barcodes associated to each putative regulatory element, or by using the regulatory element as the barcode itself.

MPRAs have been performed in a wide variety of systems, such as cell culture, dissected tissues, or adult tissues (de Boer et al., 2020; Inoue et al., 2017; King et al., 2020; Shen et al., 2016), but rarely within a developing embryo, where precise enhancer expression is crucial to activate specific gene networks, ultimately leading to development of a healthy organism. *Ciona robusta* is a model system where MPRAs have been performed in whole developing embryos to

great success (Farley et al., 2015; Song et al., 2022). However, *Ciona* are not vertebrates, and thus, the development of a vertebrate MPRA would greatly improve our understanding of how genetic changes in enhancers can drive disease.

The chicken limb bud is an ideal system to develop a vertebrate MPRA for several reasons. Electroporation of reporters into chicken limb buds is an established protocol, allowing for an easy way to assay a library of thousands of sequences in a single embryo (Tomizawa et al., 2022). We can also investigate enhancer specificity within this system, as we can identify enhancers expressed in the forelimb or hindlimb, compared to a related mesenchymal tissue, the flank. Furthermore, a set of putative regulatory elements and control enhancers have been identified previously for rigorous development of this method (Menke et al., 2008; Sackton et al., 2019). Some of these enhancers drive forelimb or hindlimb-specific activity, demonstrating that assaying for activity in both forelimbs and hindlimbs can further our understanding of how enhancers encode expression in a particular limb or in both limbs simultaneously. Anterior and posterior patterning in the developing limb bud has also been studied, and in the future, more detailed MPRA approaches could be used to provide further tissue specificity within the forelimb and hindlimbs. Finally, the development of the limb is well-studied and many of the transcription factors governing morphogenesis and patterning are known (McQueen and Towers, 2020). Thus, identifying binding sites of known limb transcription factors within enhancers would allow us to better understand what sequences are driving tissue-specific expression.

This chapter discusses a newly developed, vertebrate MPRA in chicken limb buds in two parts. First, I provide a detailed protocol that outlines how the MPRA libraries are made and electroporated into the chicken limb, and how the library mRNA and plasmid DNA libraries are subsequently isolated from the developing chicken embryo and prepared for sequencing to identify

active enhancers. I plan to publish this as a protocols paper which will serve as a resource to the research community. Second, I describe an initial study using this new MPRA to investigate differential activity of enhancers identified as conserved or accelerated in the developing bird limb bud. Previous studies comparing the genomic sequences of flying and flightless birds have identified non-coding regions that contain either conserved sequences or accelerated sequences. These regions were overlayed with ATAC-seq to identify putative enhancers with conserved or accelerated sequences between flying and flightless birds. In our preliminary study, we discover many novel limb enhancers, intriguingly, I identify enhancers that are highly conserved in sequence between chicken and emu but drive differential expression.

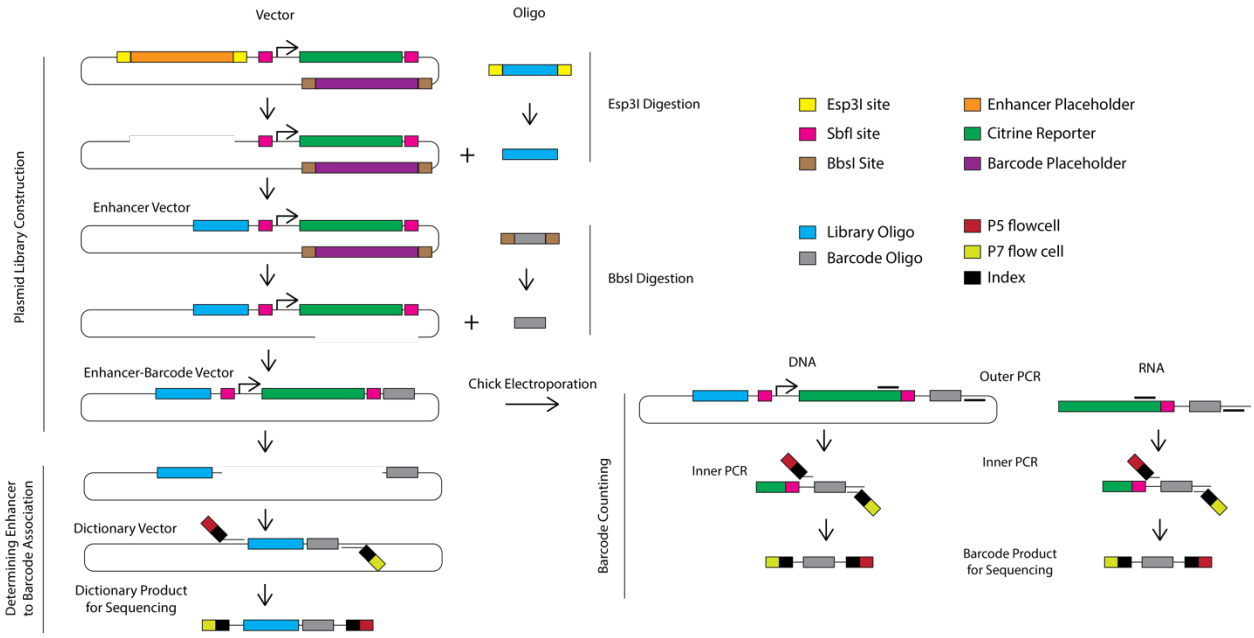


Figure 6. Schematic of chicken limb bud MPRA. A 250bp library of regulatory elements (light blue) and the library vector are digested with Esp3I and ligated together (Esp3I sites in yellow). Barcode oligo (gray) and enhancer vector are digested with BbsI and ligated together (BbsI sites in brown). This creates the enhancer-barcode vector for testing. To determine which barcodes are associated with which enhancers, the promoter and citrine reporter (green) are digested with SbfI (pink) and then intramolecularly ligated together. The dictionary vector is amplified with P5 (maroon)/P7 (mustard) adapters and indices for sequencing. The enhancer-barcode vector is electroporated into chick embryos and then the DNA and RNA barcode counts are extracted. An outer PCR is performed followed by an inner PCR to attach P5/P7 adapters and indices for sequencing.

Materials:

Biological Materials

Chicken embryos

Reagents

Zymo DNA Clean & Concentrator Kit (Zymogen, cat no. D4004)

Zymoclean Gel DNA Recovery Kit (Zymogen, cat no. D4007)

Esp3I (New England Biolabs, cat no. R0734)

BbsI-HF (New England Biolabs, cat no. R3539)

10x CutSmart Buffer (New England Biolabs, cat no. R6004)

MegaX DH10B $T1^R$ Electrocompetent cells (ThermoFisher, cat no. C6400)

NucleoBond Xtra Midi Kit (Macherey-Nagel, cat no. 740410)

TriZOL Reagent (Invitrogen, cat no. 15596026)

Chloroform/Isoamyl Alcohol (24:1, v/v) (Acros Organics, code: 327155000)

2-propanol (Sigma-Aldrich, cat no. 190764)

UltraPure Glycogen (Invitrogen, cat no. 18014010)

UltraPure DNase/RNase-Free Distilled Water (Life Technologies, cat no. 10977015)

Turbo DNA-free Kit (Invitrogen, cat no. AM1907)

Bioanalyzer High Sensitivity DNA kit (Agilent, part no. 5067-4626)

Dynabeads mRNA Purification Kit (Life Technologies Ambion, cat no. 61006)

Primers (custom-made by IDT with HPLC purification)

Transcriptor High Fidelity cDNA Synthesis Kit (Roche, SKU 5081955001)

AMPure XP Reagent (Beckman Coulter cat no. A63881)

Ethanol absolute (KOPTEC, VWR cat no. 89125-186)

Phusion High-Fidelity PCR Master Mix (New England Biolabs, cat no. M0531)

Qubit dsDNA HS Assay Kit (Invitrogen, cat no. Q32851)

Guanidine Thiocyanate (Invitrogen, cat no. AM9422)

Sodium Citrate (Sigma-Aldrich, cat no. S4641)

Tris(hydroxymethyl)aminomethane (Apex, VWR cat no. 33621.20)

UltraPure Phenol:Chloroform:Isoamyl Alcohol (25:24:1, v/v) (Invitrogen, cat no. 15593031)

Guanidine Thiocyanate (Sigma-Aldrich, cat no. G9277)

Sodium Acetate (Sigma-Aldrich, cat no. W302406)

Tris-Base (Sigma-Aldrich, cat no. 10708976001)

RNaseA (Thermo Scientific, cat no. FEREN0531)

Qubit assay tubes (ThermoFisher, cat no. Q32836)

Qubit dsDNA HS Assay Kit (ThermoFisher, cat no. Q32851)

Equipment

Pipettes (2µL, 20µL, 200µL, 1000µL; Gilson, SKU F144801, F123600, F123601, F144802, F123602)

Filter tips (10µL, 20µL, 200µL, 1000µL; Olympus Plastics, 24-403, 24-404, 24-412, 24-430C)

DNA LoBind Tubes, 1.5mL (Eppendorf, cat no. 022431021)

200µL PCR tubes (Olympus Plastics, cat no. 27-125)

DynaMag-2 (Life Technologies, cat no. 12321D)

GenePulser Xcell electroporator (Bio-Rad, cat no. 1652660)

Vortex Mixer

2L flasks

Qubit fluorometer (ThermoFisher, cat no. Q33238)

NanoDrop Spectrophotometer (ThermoFisher, cat no. ND-ONE)

Bioanalyzer (Agilent, cat no. G2939BA) or TapeStation (Agilent, cat no. G2991BA)

Thermal Cycler

Heating Dry bath

Shaking Incubator

Temperature Centrifuge

Methods:

Enhancer Library cloning

In this section, the library of putative enhancers is cloned into the library vector using Esp3I. Amplification of the library is required if the starting material is small, <1µg.

Part I: Library amplification

Primer Extension

1. Dissolve Twist Bioscience library in TE Buffer to 5ng/µL
2. Set up 16 reactions of dsDNA extension

	1x	16x
10µM Forward Primer	2.5µL	40µL
10µM Reverse Primer	2.5µL	40µL
5ng/µL Library oligo	1µL	16µL
Ultra-Pure Water (UPW)	17.5µL	304µL
Phusion-HF 2x Master Mix	25µL	400µL

3. Run extension reaction:

98°C for 30s, (98°C for 15s, 63°C for 40s, 72°C for15s) x 8 cycles, 72°C for 5 min, 4°C hold

4. Pool extension reactions together

dsDNA cleanup

5. Add 5 volumes of Zymo DNA binding buffer to extension reactions
6. Add 750µL of extension reaction to three different Zymo Spin Column
7. Spin columns at 10,000xg for 15s, discard supernatant
8. Repeat steps 6 and 7 until all of the extension reactions have passed through the columns
9. Add 200µL of DNA Wash buffer to each spin column

10. Spin columns at 10,000xg for 15s, discard supernatant
11. Repeat steps 9 and 10 once more
12. Elute each column with 45µL UPW
13. Incubate at room temperature (RT) for 10 minutes
14. Spin columns at 10,000xg for 15s
15. Measure concentration by Nanodrop
16. Run 1µL of sample on the Bioanalyzer

Part 2: Library Digestion

1. Calculate total amount of DNA recovered from dsDNA extension based on Nanodrop (there should be approximately double the input into the dsDNA extension reaction)
2. Determine number of units of Esp3I needed to digest the dsDNA based on the following formula:

$$\frac{Units\ of Esp3I\ per\ \mu g}{\mu g\ of\ library\ dsDNA} = \frac{length\ of\ \lambda\ DNA}{length\ of\ library\ DNA}\ x\ \frac{\#\ of\ Esp3I\ sites\ in\ library\ oligo}{\#\ of\ Esp3I\ sites\ in\ \lambda\ DNA}$$

Size of λ DNA 48,502bp and number of Esp3I sites in λ DNA is 14

Information for any restriction enzyme can be found at: https://nc2.neb.com/NEBcutter2/

3. Set up Esp3I digestion reactions:

	1x
Library dsDNA oligo	1.7µg
Esp3I (10,000 units/mL)*	4µL
10x CutSmart Buffer	5µL
Ultra-Pure Water	Up to 50µL

*Maximum of 4uL per reaction due to Esp3I being stored in glycerol

23 units of Esp3I can digest 1µg of a 300bp library

Calculate number of reactions needed to digest all of the dsDNA library

4. Incubate digestion reactions at 37 °C for one hour
5. Pool all digestion reactions together
6. Add 5 volumes of Zymo DNA binding buffer to extension reactions
7. Add 750µL of digestion reaction to Zymo Spin columns
8. Spin columns at 10,000xg for 15s, discard supernatant
9. Repeat steps 7 and 8 until all of the digestion reactions have passed through the columns
10. Add 200µL of DNA Wash buffer to each spin column
11. Spin columns at 10,000xg for 15s, discard supernatant
12. Repeat steps 9 and 10 once more
13. Elute each column with 10µL UPW
14. Incubate at RT for 10 minutes
15. Spin columns at 10,000xg for 15s
16. Measure concentration of digested library by Nanodrop
17. Run 1µL of sample on the Bioanalyzer

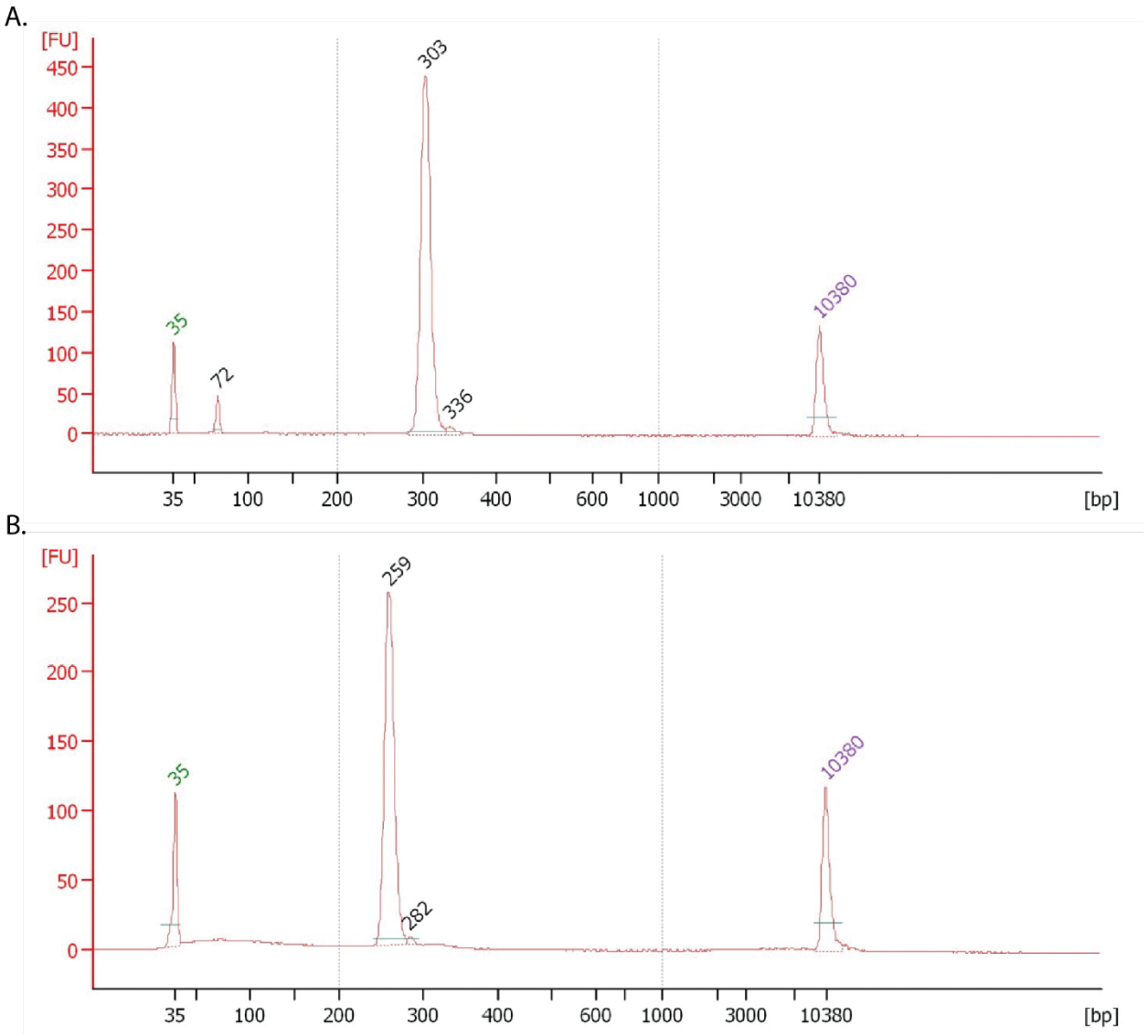


Figure 7. Library oligo bioanalyzers. A. Sharp peak at expected oligo size. Small primer peak sometimes visible (72bp) **B.** Sharp peak at expected digested oligo size. ~50bp size decrease, demonstrating successful digestion. Sometimes, small single cut peak is visible (282bp)

Part 3: Vector Digestion

1. Set up Esp3I digestion reactions:

	1x
Vector DNA	10µg
Esp3I (10,000 units/mL)*	2µL
10x CutSmart Buffer	5µL
Ultra-Pure Water	Up to 50µL

2. Incubate digestion reactions at 37 °C for two hours
3. Run digestion reactions on 1% agarose gel at 80V for 3 hours
4. Excise gel band with enhancer placeholder digested out
5. Weigh gel band and add 3µL of Agarose Dissolving Buffer per mg of gel band
6. Incubate gel band mixture at 55 °C for 10 minutes, vortex mixture at 5 minutes
7. Add 750µL of digestion reaction to Zymo Spin column
8. Spin columns at 10,000xg for 15s, discard supernatant
9. Repeat steps 7 and 8 until all of the digestion reactions have passed through the columns
10. Add 200µL of DNA Wash buffer to each spin column
11. Spin columns at 10,000xg for 15s, discard supernatant
12. Repeat steps 9 and 10 once more
13. Elute each column with 100µL UPW
14. Incubate at RT for 10 minutes
15. Spin columns at 10,000xg for 15s
16. Add 2 volumes of Zymo DNA binding buffer to vector digestion reactions
17. Add 750µL of digestion reaction to Zymo Spin columns

18. Spin columns at 10,000xg for 15s, discard supernatant
19. Add 200µL of DNA Wash buffer to each spin column
20. Spin columns at 10,000xg for 15s, discard supernatant
21. Repeat steps 9 and 10 once more
22. Elute each column with 10µL UPW
23. Incubate at RT for 10 minutes
24. Spin columns at 10,000xg for 15s
25. Measure concentration by Nanodrop
26. Run 1µL of sample on the Bioanalyzer

Part 4: Enhancer Ligation

1. Set up enhancer ligation reactions:

	1x	10x
Digested Vector DNA	50ng	500ng
Digested Library oligo	50ng	500ng
2x Rapid Ligation Buffer	10µL	100µL
T4 DNA Ligase	1µL	10µL
Ultra-Pure Water	Up to 20µL	Up to 200µL

2. Split into 10 Lo-Bind tubes of 20µL each
3. Incubate at RT for 10 minutes
4. Pool all ligation reactions together
5. Add 2 volumes of Zymo DNA binding buffer to ligation reactions
6. Add ligation mixture to Zymo Spin columns

7. Spin columns at 10,000xg for 15s, discard supernatant
8. Add 200μL of DNA Wash buffer to each spin column
9. Spin columns at 10,000xg for 15s, discard supernatant
10. Repeat steps 9 and 10 once more
11. Elute each column with 10μL UPW
12. Incubate at RT for 10 minutes
13. Spin columns at 10,000xg for 15s
14. Measure concentration by Nanodrop
15. Run 1μL of sample on the Bioanalyzer

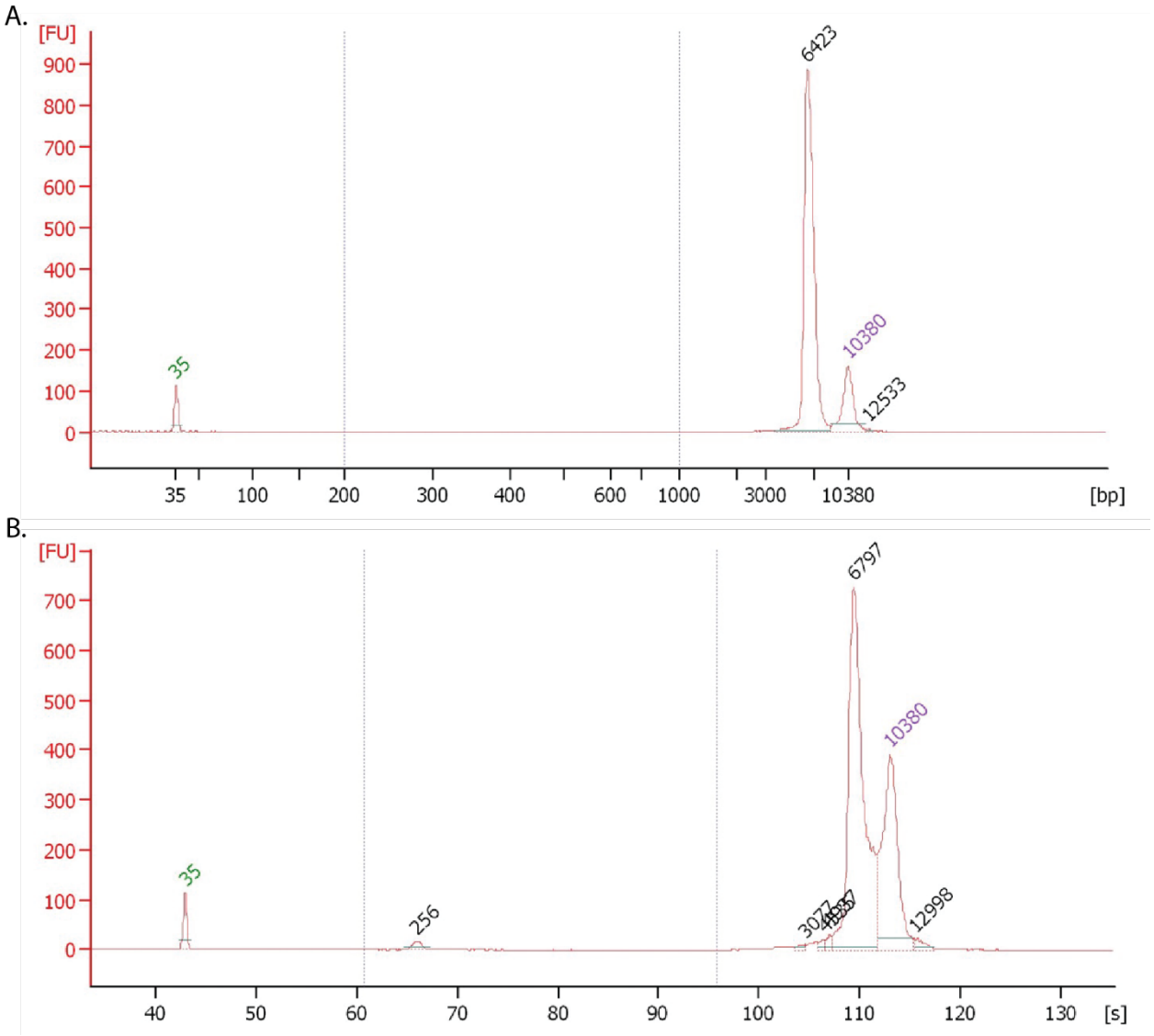


Figure 8. Vector bioanalyzers. A. Sharp peak at expected digested vector size. **B.** Shifted peak with large tail following ligation of library oligo suggests successful circularization of the plasmid. Sometimes, small unligated library oligo peak is visible (282bp)

Part 5: Enhancer ligation transformation

1. Prewarm 5mL of SOC medium from MegaX DH10B $T1^R$ electrocompetent cells at 37°C for 30 minutes
2. Thaw 200µL of MegaX DH10B $T1^R$ electrocompetent cells on ice
3. Aliquot 20µL of cells into 10 Lo-Bind tubes
4. Add 1µL of pUC19 DNA to one aliquot, 1µL of Ultra-Pure Water to another aliquot, and 1µL of enhancer ligation reaction to each of the other eight aliquots

   Note: no more than 100ng of the enhancer ligation per aliquot
5. Transfer competent cells to 1mm-gap cuvette
6. Tap cuvette on counter to move cells to the bottom of the cuvette
7. Place cuvette in GenePulser Xcell electroporator, shock the cells with the following settings: voltage: 1800V; capacitance: 25 µF; resistance: 200ohms; cuvette: 1mm
8. Add 200µL of prewarmed SOC medium at a time to cuvette; pipette up and down, add to 5mL conical tube, repeat until 1mL of SOC has been mixed with shocked cells in conical tube
9. Repeat steps 5-8 for all tubes
10. Incubate cells in shaking incubator (250rpm) at 37°C for 1hr
11. Pool cells from enhancer ligation reaction together and measure total volume of pool
12. Perform two serial 10x dilutions to 1:10,000 of 100µL of enhancer ligation
13. Plate 100µL of 1:1000 and 1:10,000 dilutions on ampicillin plates
14. Perform two serial 10x dilutions to 1:1000 of 100µL of pUC19 bacteria
15. Plate 100µL of 1:100 and 1:1000 dilutions on ampicillin plates
16. Plate 100µL of undiluted Ultra-Pure Water bacteria on ampicillin plates

17. Grow ampicillin plates overnight at 37°C
18. Split undiluted enhancer ligation bacteria into 3 2L flasks of 300mL of 2xYT
19. Incubate flasks at 37C in a shaking incubator (250rpm) overnight

Part 6: Counting colonies and enhancer library purification

1. Count colonies on diluted enhancer ligation plates and calculate enhancer library complexity by calculating number of colonies in 1mL of undiluted pool and multiplying by total volume of the pool
2. Pour flasks into 500mL centrifuge bottles
3. Centrifuge at 4400xg at 4°C for 25 minutes
4. Discard supernatant
5. Purify plasmids using Macherey-Nagel Nucleobond Xtra Midi Kit, following manufacturer's specifications. Use one column per flask.
6. Measure plasmid concentration using Nanodrop

Barcode cloning

In this section, the transcribable barcode is cloned into the library of putative enhancers. The protocol in this section is very similar to the enhancer library cloning section. **Deviations will be noted by asterisks**

Part I: Double stranded extension of barcode oligo

1. Dissolve IDT DNA barcode oligo in Ultra-Pure Water to **50ng/µL**

2. Set up 16 reactions of dsDNA extension

	1x	16x
10μM Reverse Primer	2.5μL	40μL
50ng/μL barcode oligo	1μL	16μL
Ultra-Pure Water (UPW)	17.5μL	344μL
Phusion-HF 2x Master Mix	25μL	400μL

3. Run extension reaction:

98°C for 30s, 98°C for 15s, 63°C for 40s, 72°C for15s, 72°C for 5 min, 4°C hold

4. Pool extension reactions together

dsDNA cleanup

5. Add 5 volumes of Zymo DNA binding buffer to extension reactions
6. Add 750μL of extension reaction to three different Zymo Spin Column
7. Spin columns at 10,000xg for 15s, discard supernatant
8. Repeat steps 6 and 7 until all of the extension reactions have passed through the columns
9. Add 200μL of DNA Wash buffer to each spin column
10. Spin columns at 10,000xg for 15s, discard supernatant
11. Repeat steps 9 and 10 once more
12. Elute each column with 45μL UPW
13. Incubate at RT for 10 minutes
14. Spin columns at 10,000xg for 15s
15. Measure concentration by Nanodrop
16. Run 1μL of sample on the Bioanalyzer

Part 2: Barcode Digestion

1. Calculate total amount of DNA recovered from dsDNA extension based on Nanodrop (there should be approximately double the input into the dsDNA extension reaction)
2. Determine number of units of **BbsI-HF** needed to digest the dsDNA based on the following formula:

$$\frac{Units\ of BbsI\ per\ \mu g}{\mu g\ of\ library\ dsDNA} = \frac{length\ of\ \lambda\ DNA}{length\ of\ library\ DNA}\ x\ \frac{\#\ of\ BbsI\ sites\ in\ library\ oligo}{\#\ of\ BbsI\ sites\ in\ \lambda\ DNA}$$

Size of λ DNA 48,502bp and **number of BbsI sites in λ DNA is 24**

Information for any restriction enzyme can be found at: https://nc2.neb.com/NEBcutter2/

3. Set up BbsI-HF digestion reactions:

	1x
Barcode dsDNA oligo	**3.1µg**
BbsI-HF (20,000 units/mL)	4µL
10x CutSmart Buffer	5µL
Ultra-Pure Water	Up to 50µL

Maximum of 4uL per reaction due to BbsI-HF being stored in glycerol

26 units of BbsI-HF can digest 1µg of a 153bp barcode oligo

Calculate number of reactions needed to digest all of the dsDNA barcode oligo

4. Incubate digestion reactions at 37°C for one hour

5. Pool all digestion reactions together
6. Add 5 volumes of Zymo DNA binding buffer to extension reactions
7. Add 750µL of digestion reaction to Zymo Spin columns
8. Spin columns at 10,000xg for 15s, discard supernatant
9. Repeat steps 7 and 8 until all of the digestion reactions have passed through the columns
10. Add 200µL of DNA Wash buffer to each spin column
11. Spin columns at 10,000xg for 15s, discard supernatant
12. Repeat steps 9 and 10 once more
13. Elute each column with 10µL UPW
14. Incubate at RT for 10 minutes
15. Spin columns at 10,000xg for 15s
16. Measure concentration of digested barcode oligo by Nanodrop
17. Run 1µL of sample on the Bioanalyzer

Part 3: Enhancer Library Digestion

1. Set up BbsI-HF digestion reactions of enhancer library:

	1x
Enhancer Library DNA	10µg
BbsI-HF (20,000 units/mL)	2µL
10x CutSmart Buffer	5µL
Ultra-Pure Water	Up to 50µL

2. Incubate digestion reactions at 37ºC for two hours
3. Run digestion reactions on 1% agarose gel at 80V for 3 hours

4. Excise gel band with barcode placeholder digested out
5. Weigh gel band and add 3μL of Agarose Dissolving Buffer per mg of gel band
6. Incubate gel band mixture at 55C for 10 minutes, vortex mixture at 5 minutes
7. Add 750μL of digestion reaction to Zymo Spin column
8. Spin columns at 10,000xg for 15s, discard supernatant
9. Repeat steps 7 and 8 until all of the digestion reactions have passed through the columns
10. Add 200μL of DNA Wash buffer to each spin column
11. Spin columns at 10,000xg for 15s, discard supernatant
12. Repeat steps 9 and 10 once more
13. Elute each column with 100μL UPW
14. Incubate at RT for 10 minutes
15. Spin columns at 10,000xg for 15s
16. Add 2 volumes of Zymo DNA binding buffer to vector digestion reactions
17. Add 750μL of digestion reaction to Zymo Spin columns
18. Spin columns at 10,000xg for 15s, discard supernatant
19. Add 200μL of DNA Wash buffer to each spin column
20. Spin columns at 10,000xg for 15s, discard supernatant
21. Repeat steps 9 and 10 once more
22. Elute each column with 10μL UPW
23. Incubate at RT for 10 minutes
24. Spin columns at 10,000xg for 15s
25. Measure concentration by Nanodrop
26. Run 1μL of sample on the Bioanalyzer

Part 4: Barcode Ligation

1. Set up Ligation reactions:

	1x	10x
**Digested Enhancer library **	50ng	500ng
Digested barcode oligo	50ng	500ng
2x Rapid Ligation Buffer	10µL	100µL
T4 DNA Ligase	1µL	10µL
Ultra-Pure Water	Up to 20µL	Up to 200µL

2. Split into 10 Lo-Bind tubes of 20µL each
3. Incubate at RT for 10 minutes
4. Pool all ligation reactions together
5. Add 2 volumes of Zymo DNA binding buffer to ligation reactions
6. Add ligation mixture to Zymo Spin columns
7. Spin columns at 10,000xg for 15s, discard supernatant
8. Add 200µL of DNA Wash buffer to each spin column
9. Spin columns at 10,000xg for 15s, discard supernatant
10. Repeat steps 9 and 10 once more
11. Elute each column with 10µL UPW
12. Incubate at RT for 10 minutes
13. Spin columns at 10,000xg for 15s
14. Measure concentration by Nanodrop
15. Run 1µL of sample on the Bioanalyzer

Part 5: Barcode ligation transformation

Note: The transformation step here limits the number of barcodes associated with each enhancer. Ideally, try transforming different amounts of ligation reaction to reach your targeted enhancer to barcode ratio. The following protocol is suggested for 10-50 million barcodes total.

1. Prewarm 5mL of SOC medium from MegaX DH10B $T1^R$ electrocompetent cells at 37°C for 30 minutes
2. Thaw **100µL** of MegaX DH10B $T1^R$ electrocompetent cells on ice
3. Aliquot 20µL of cells into **five** Lo-Bind tubes
4. Add 1uL of pUC19 DNA to one aliquot, 1uL of Ultra-Pure Water to another aliquot, and 1uL of barcode ligation reaction to each of the other **three** aliquots

   Note: no more than 100ng of the barcode ligation reaction per aliquot
5. Transfer competent cells to 1mm-gap cuvette
6. Tap cuvette on counter to move cells to the bottom of the cuvette
7. Place cuvette in GenePulser Xcell electroporator, shock the cells with the following settings: voltage: 1800V; capacitance: 25 µF; resistance: 200ohms; cuvette: 1mm
8. Add 200µL of prewarmed SOC medium at a time to cuvette; pipette up and down, add to 5mL conical tube, repeat until 1mL of SOC has been mixed with shocked cells in conical tube
9. Repeat steps 5-8 for all tubes
10. Incubate cells in shaking incubator (250rpm) at 37°C for 1hr
11. Pool cells from enhancer ligation reaction together and measure total volume of pool
12. Perform two serial 10x dilutions to 1:10,000 of 100µL of enhancer ligation
13. Plate 100µL of 1:1000 and 1:10,000 dilutions on ampicillin plates

14. Perform two serial 10x dilutions to 1:1000 of 100µL of pUC19 bacteria
15. Plate 100µL of 1:100 and 1:1000 dilutions on ampicillin plates
16. Plate 100µL of undiluted Ultra-Pure Water bacteria on ampicillin plates
17. Grow ampicillin plates overnight at 37°C
18. Split undiluted enhancer ligation bacteria into 3 2L flasks of 300mL of 2xYT
19. Incubate flasks at 37°C in a shaking incubator (250rpm) overnight

Part 6: Counting colonies and enhancer-barcode library purification

1. Count colonies on diluted enhancer ligation plates and calculate enhancer library complexity by calculating number of colonies in 1mL of undiluted pool and multiplying by total volume of the pool
2. Pour flasks into 500mL centrifuge bottles
3. Centrifuge at 4400xg at 4°C for 25 minutes
4. Discard supernatant
5. Purify plasmids using Macherey-Nagel Nucleobond Xtra Midi Kit, following manufacturer's specifications. Use one column per flask.
6. Measure plasmid concentration using Nanodrop

Dictionary sequencing preparation

In this section, the enhancer-barcode library is prepared for sequencing, so that the barcodes associated with each putative enhancer can be ascertained.

Part 1: Digestion of enhancer-barcode library for dictionary sequencing

1. Set up SbfI digestion reactions of enhancer-barcode library:

	1x
Enhancer-barcode Library DNA	10µg
SbfI-HF (20,000 units/mL)	2µL
10x CutSmart Buffer	5µL
Ultra-Pure Water	Up to 50µL

2. Incubate digestion reactions at 37°C for two hours
3. Run digestion reactions on 1% agarose gel at 80V for 3 hours
4. Excise gel band with promoter and reporter digested out
5. Weigh gel band and add 3µL of Agarose Dissolving Buffer per mg of gel band
6. Incubate gel band mixture at 55C for 10 minutes, vortex mixture at 5 minutes
7. Add 750µL of digestion reaction to Zymo Spin column
8. Spin columns at 10,000xg for 15s, discard supernatant
9. Repeat steps 7 and 8 until all of the digestion reactions have passed through the columns
10. Add 200µL of DNA Wash buffer to each spin column
11. Spin columns at 10,000xg for 15s, discard supernatant
12. Repeat steps 9 and 10 once more
13. Elute each column with 100µL UPW
14. Incubate at RT for 10 minutes

15. Spin columns at 10,000xg for 15s
16. Add 2 volumes of Zymo DNA binding buffer to vector digestion reactions
17. Add 750µL of digestion reaction to Zymo Spin columns
18. Spin columns at 10,000xg for 15s, discard supernatant
19. Add 200µL of DNA Wash buffer to each spin column
20. Spin columns at 10,000xg for 15s, discard supernatant
21. Repeat steps 9 and 10 once more
22. Elute each column with 10µL UPW
23. Incubate at RT for 10 minutes
24. Spin columns at 10,000xg for 15s
25. Measure concentration by Nanodrop
26. Run 1µL of sample on the Bioanalyzer

Part 2: Ligation of enhancer-barcode library for dictionary sequencing

1. Set up self-ligation reactions:

	1x	10x
Digested Enhancer-barcode library	50ng	500ng
2x Rapid Ligation Buffer	10µL	100µL
T4 DNA Ligase	1µL	10µL
Ultra-Pure Water	Up to 20µL	Up to 200µL

2. Split into 10 Lo-Bind tubes of 20µL each
3. Incubate at RT for 10 minutes
4. Pool all ligation reactions together

5. Add 2 volumes of Zymo DNA binding buffer to ligation reactions
6. Add ligation mixture to Zymo Spin columns
7. Spin columns at 10,000xg for 15s, discard supernatant
8. Add 200µL of DNA Wash buffer to each spin column
9. Spin columns at 10,000xg for 15s, discard supernatant
10. Repeat steps 9 and 10 once more
11. Elute each column with 10µL UPW
12. Incubate at RT for 10 minutes
13. Spin columns at 10,000xg for 15s
14. Measure concentration by Nanodrop
15. Run 1µL of sample on the Bioanalyzer

Part 3: Dictionary ligation transformation

1. Prewarm 5mL of SOC medium from MegaX DH10B $T1^R$ electrocompetent cells at 37°C for 30 minutes
2. Thaw 100µL of MegaX DH10B $T1^R$ electrocompetent cells on ice
3. Aliquot 20µL of cells into 5 Lo-Bind tubes
4. Add 1µL of pUC19 DNA to one aliquot, 1µL of Ultra-Pure Water to another aliquot, and up to 100ng of barcode ligation reaction to each of the other three aliquots

   Note: no more than 2µL of the barcode ligation reaction per aliquot
5. Transfer competent cells to 1mm-gap cuvette
6. Tap cuvette on counter to move cells to the bottom of the cuvette

7. Place cuvette in GenePulser Xcell electroporator, shock the cells with the following settings: voltage: 1800V; capacitance: 25 μF; resistance: 200ohms; cuvette: 1mm
8. Add 200uL of prewarmed SOC medium at a time to cuvette; pipette up and down, add to 5mL conical tube, repeat until 1mL of SOC has been mixed with shocked cells in conical tube
9. Repeat steps 5-8 for all tubes
10. Incubate cells in shaking incubator (250rpm) at 37°C for 1hr
11. Pool cells from enhancer ligation reaction together and measure total volume of pool
12. Perform two serial 10x dilutions to 1:10,000 of 100μL of enhancer ligation
13. Plate 100μL of 1:10,000 dilutions on ampicillin plates
14. Perform two serial 10x dilutions to 1:1000 of 100uL of pUC19 bacteria
15. Plate 100μL of 1:100 and 1:1000 dilutions on ampicillin plates
16. Plate 100μL of undiluted Ultra-Pure Water bacteria on ampicillin plates
17. Grow ampicillin plates overnight at 37°C
18. Grow undiluted enhancer ligation bacteria in a 2L flask with 300mL of 2xYT
19. Incubate flasks at 37°C in a shaking incubator (250rpm) overnight

<u>Part 4: Counting colonies and dictionary plasmid purification</u>

1. Count colonies on diluted enhancer ligation plates and calculate enhancer library complexity by calculating number of colonies in 1mL of undiluted pool and multiplying by total volume of the pool
2. Pour flasks into 500mL centrifuge bottles
3. Centrifuge at 4400xg at 4C for 25 minutes

4. Discard supernatant
5. Purify plasmids using Macherey-Nagel Nucleobond Xtra Midi Kit, following manufacturer's specifications. Use one column per flask.
6. Measure plasmid concentration using Nanodrop

Part 5: PCR to amplify dictionary for sequencing

1. Set up PCR reaction for dictionary amplification

	1x	8x
10μM PCR inner For	2.5μL	20μL
10μM PCR inner Rev	2.5μL	20μL
50ng/μL dictionary plasmid	1μL	8μL
Ultra-Pure Water (UPW)	19μL	152μL
Phusion-HF 2x Master Mix	25μL	200μL

Note: PCR primers have indices and P5/P7 adapter sequences, use different For/Rev primers for each sample to be sequenced

2. Run PCR reaction with the following conditions:

98°C for 30s, [98°C for 15s, 63°C for 40s, 72°C for15s] x 10 cycles, 72°C for 5 min, 4°C hold

dsDNA cleanup

3. Add 5 volumes of Zymo DNA binding buffer to extension reactions
4. Add 750μL of extension reaction to three different Zymo Spin Column
5. Spin columns at 10,000xg for 15s, discard supernatant
6. Repeat steps 6 and 7 until all of the extension reactions have passed through the columns

7. Add 200µL of DNA Wash buffer to each spin column
8. Spin columns at 10,000xg for 15s, discard supernatant
9. Repeat steps 9 and 10 once more
10. Elute each column with 50µL UPW
11. Incubate at RT for 10 minutes
12. Spin columns at 10,000xg for 15s

Bead cleanup

13. Pool reactions of each sample together in Lo-Bind tube
14. Add 0.95x the volume of the pooled sample in AMPure XP beads
15. Pipette to mix
16. Incubate at RT for 5 min
17. Put tubes on magnet for 5 min
18. Discard supernatant
19. Wash with 1mL 85% ethanol on magnet
20. Remove supernatant
21. Repeat steps 7 and 8 one more time
22. Remove all ethanol with vacuum
23. Dry beads at RT until the beads are not shiny anymore (but are not cracked)
24. Add 40µL UPW to beads, pipette to mix
25. Incubate at RT for 5 min
26. Put tubes on magnet for 5 min
27. Transfer supernatant to new Lo-Bind tube.
28. Measure concentration by Qubit

29. Run 1μL of sample on the Bioanalyzer
30. Submit samples for sequencing

    Note: Coverage of 100x the number of barcodes is ideal for sequencing depth

Chick Electroporation

Note: The methods of this section of the protocol will be completed by Dr. Meng Zhu of the Tabin lab at Harvard University.

Briefly, the library is electroporated into HH16 forelimb buds, hindlimb buds, and flank of chicken embryos, along with an electroporation marker. The embryos develop to stage HH21 and then the forelimb bud, hindlimb bud, and flank of each embryo is dissected out, collected and flash frozen in liquid nitrogen. Each replicate consists of tissue collected from ten embryos.

RNA Extraction of barcodes

In this section, the transcribed barcodes driven by enhancer activity are extracted from the tissues and prepared for sequencing.

<u>Part 1: TRIzol extraction</u>

1. Add 1mL of TRIzol reagent to each sample
2. Vortex until all tissue is dissolved completely
3. Add 200μL chloroform
4. Shake vigorously for 1 min
5. Centrifuge at 12,000xg for 5 min at 4^{o}C
6. Transfer the upper, aqueous layer (~650μL) containing RNA into new Lo-Bind tube, place on ice

(Keep the remaining solid and bottom layers for DNA extraction later)

7. Add 600µL of isopropanol to upper aqueous layer
8. Optional: Add 2µL of glycogen for pellet visualization following precipitation
9. Vortex 15s
10. Incubate 5 min at RT
11. Centrifuge 12,000xg for 15 min at 4°C
12. Remove most liquid, leave about 10uL
13. Wash with 1mL of 70% ethanol
14. Centrifuge at 8,000xg for 6 min at 4°C
15. Remove the supernatant
16. Centrifuge at 8,000xg for 30s at 4°C
17. Remove remaining liquid with pipette and/or vacuum pump
18. Let pellet dry until pellet turns nearly transparent
19. Add 90µL of UPW to each sample
20. Let dissolve for 30 min at RT
21. Nanodrop for concentration and send for RNA integrity (RIN) analysis

Part 2: DNase digestion

1. Add 10µL of 10X TURBO buffer and 1.5µL of TURBO DNase enzyme from TURBO DNA-free Kit to each sample
2. Pipette to mix reaction
3. Incubate at 37°C for 30 min
4. Add additional 1.5µL of TURBO DNase enzyme to each reaction

5. Pipette to mix reaction
6. Add 10µL of DNase Inactivation Reagent to each reaction
7. Incubate 5 min at RT, flicking 2-3 during incubation time
8. Centrifuge at 10,000xg for 2 min at RT
9. Transfer supernatant to new Lo-Bind tube
10. Nanodrop for concentration and send for RIN analysis

Part 3: mRNA isolation

1. Heat each total RNA sample at 65ºC for 5 min, then place immediately on ice
2. Aliquot 200µL of Dynabeads Oligo $(dT)_{25}$ to empty Lo-Bind tube for each sample that is being processed
3. Place tubes on magnet for 30s
4. Remove supernatant
5. Resuspend beads with 100µL of Binding buffer, incubate 1 min at RT
6. Place tubes on magnet for 30s
7. Discard supernatant
8. Add equal volume of Binding buffer as total RNA sample to beads, (100µL)
9. Add total RNA to Binding buffer/beads mixture
10. Pipette to mix
11. Put on rotator for 5 min at RT
12. Centrifuge for 5s
13. Place tube on magnet for 30s
14. Remove supernatant

15. Remove tube from magnet and ash with 200uL of washing buffer B, incubate at RT for 1 min
16. Place tube on magnet for 30s
17. Remove supernatant
18. Repeat steps 15-17 one more time
19. Use vacuum pump to remove all supernatant from the beads
20. Elute beads with 20.8µL of Tris-HCl, mix well by pipetting
21. Heat to 75^{o}C for 2min
22. Place tube on magnet for 30s
23. Transfer supernatant to new Lo-Bind tube, place on ice
24. Repeat steps 20-22 once more
25. Transfer supernatant to same tube, place on ice
26. Nanodrop

Part 4: cDNA synthesis

1. Add 4µL of 50uM Rev RT to each mRNA sample
2. Split each sample into two 200µL tubes
3. In thermocycler, incubate at 65ºC for 10 min, then 4ºC hold
4. Create master mix

	For each sample
5x Transcription Buffer	8µL
10mM dNTP	4µL
0.1M DTT	2µL
Reverse Transcriptase	2.2µL
RNase Inhibitor (40U/µL)	1µL

5. Add 17.2µL per 200uL tube
6. In thermocycler, run the following program:

   55ºC 1hr, 85ºC 5 min, 4ºC hold

Bead cleanup of cDNA

7. Pool two reactions of each sample together in Lo-Bind tube
8. Add 0.95x the volume of the pooled sample in AMPure XP beads
9. Pipette to mix
10. Incubate at RT for 5 min
11. Put tubes on magnet for 5 min
12. Discard supernatant
13. Wash with 1mL 85% ethanol on magnet
14. Remove supernatant

15. Repeat steps 7 and 8 one more time
16. Remove all ethanol with vacuum
17. Dry beads at RT until the beads are not shiny anymore (but are not cracked)
18. Add 40µL UPW to beads, pipette to mix
19. Incubate at RT for 5 min
20. Put tubes on magnet for 5 min
21. Transfer supernatant to new Lo-Bind tube.
22. Nanodrop

Part 5: Outer RNA PCR

1. For each sample:

	1x
10µM FVH1 For	5µL
10µM Rev RT	5µL
cDNA	40µL
Phusion-HF 2x Master Mix	50µL

2. Split each mix into two 200µL tubes
3. Run PCR reaction with the following conditions:

   98°C for 30s, [98°C for 10s, 60°C for 10s, 72°C for15s] x 15 cycles, 72°C for 5 min, 4°C hold

Bead cleanup of outer PCR

4. Pool PCR reactions together into Lo-Bind tube
5. Add 0.95x the volume of the pooled sample in AMPure XP beads (95µL)

6. Pipette to mix
7. Incubate at RT for 5 min
8. Put tubes on magnet for 5 min
9. Discard supernatant
10. Wash with 1mL 85% ethanol on magnet
11. Remove supernatant
12. Repeat steps 7 and 8 one more time
13. Remove all ethanol with vacuum
14. Dry beads at RT until the beads are not shiny anymore (but are not cracked)
15. Add 40µL UPW to beads, pipette to mix
16. Incubate at RT for 5 min
17. Put tubes on magnet for 5 min
18. Transfer supernatant to new Lo-Bind tube.
19. Nanodrop

Part 6: Inner RNA PCR

1. For each sample:

	1x
10µM PCR inner For UDI00XX	5µL
10µM PCR inner Rev UDI00XX	5µL
Outer RNA PCR product	40µL
Phusion-HF 2x Master Mix	50µL

Note: PCR primers have indices and P5/P7 adapter sequences, use different For/Rev primers for each sample to be sequenced

2. Run PCR reaction with the following conditions:

   98°C for 30s, [98°C for 10s, 60°C for 10s, 72°C for15s] x 15 cycles, 72°C for 5 min, 4°C hold

Bead cleanup of inner PCR

3. Transfer reaction to Lo-Bind tube
4. Add 0.95x the volume of the PCR reaction in AMPure XP beads
5. Pipette to mix
6. Incubate at RT for 5 min
7. Put tubes on magnet for 5 min
8. Transfer supernatant to new Lo-Bind tube
9. Add 0.85x the volume
10. Wash with 1mL 85% ethanol on magnet
11. Remove supernatant
12. Repeat steps 7 and 8 one more time
13. Remove all ethanol with vacuum
14. Dry beads at RT until the beads are not shiny anymore (but are not cracked)
15. Add 40µL UPW to beads, pipette to mix
16. Incubate at RT for 5 min
17. Put tubes on magnet for 5 min
18. Transfer supernatant to new Lo-Bind tube.
19. Measure concentration by Qubit

20. Run 1µL of sample on the Bioanalyzer
21. Send for Illumina NovaSeq PE100 Sequencing, ideally 10-100x sequencing depth per barcode

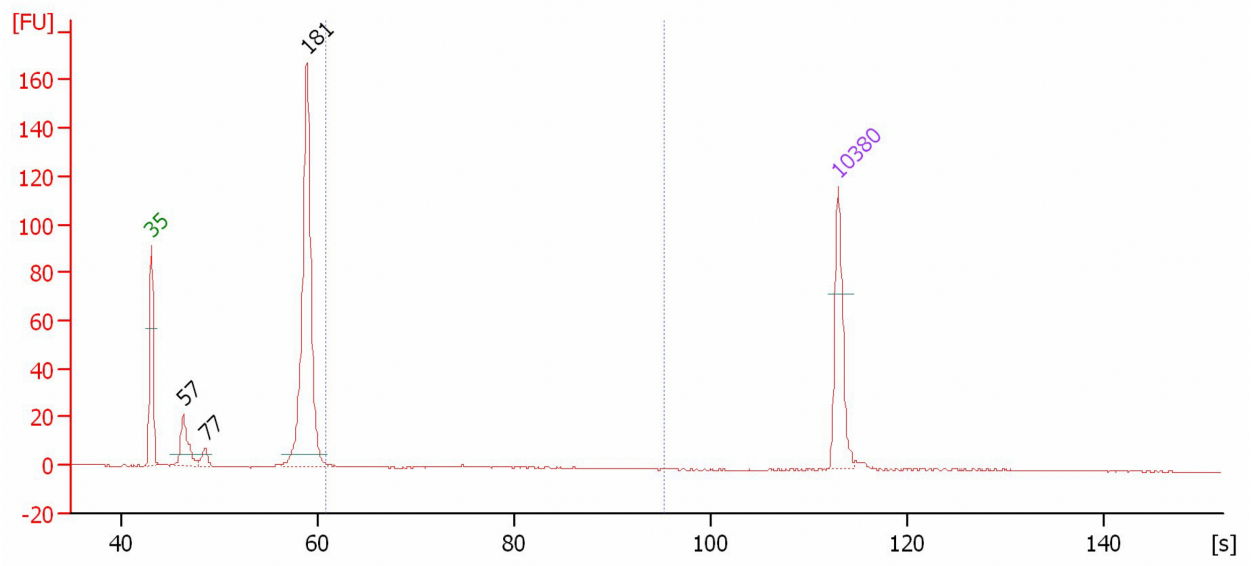


Figure 9. Example Bioanalyzer of inner PCR result. Expected PCR product size of ~180bp. Small primer peaks are sometimes visible.

DNA extraction

In this section, plasmid DNA is extracted and prepared for sequencing to normalize enhancer activity by the number of copies electroporated into the tissue.

Part 1: Total DNA extraction

1. Add 500µL of Back Extraction Buffer (4M Guanidine thiocyanate, 50mM sodium citrate, 1M Tris base)
2. Mix by vigorously inverting for 15s
3. Rock on nutator at RT for 20 min
4. Centrifuge at 14,000xg for 30 min at RT
5. Transfer the upper aqueous phase (~500µL) containing DNA to a new Lo-Bind tube

6. Add 2µL of 20ug/uL glycogen
7. Add 400µL of isopropanol and mix by vortexing
8. Incubate for 5 min at RT
9. Centrifuge at 12,000xg for 15 min at 4°C
10. Remove the supernatant
11. Wash with 1mL 70% ethanol
12. Centrifuge at 8,000xg for 6 min at 4°C
13. Remove the supernatant
14. Centrifuge at 8,000xg for 30s at 4°C
15. Remove remaining supernatant with pipette and vacuum pump
16. Air dry until pellet is nearly transparent
17. Add 100µL of UPW
18. Let it dissolve for 30 min

Part 2: RNaseA digestion

1. Add 5µL of RNaseA to each sample
2. Incubate at 37°C for 1 hour
3. Spin down a phase lock tube per sample at 12,000xg for 30s
4. Add RNaseA reaction to phase lock tube
5. Add equal volume of phenol:chloroform:isoamyl alcohol (25:24:1) to phase lock tube
6. Vigorously shake to mix
7. Centrifuge at 14,000xg for 5 min at RT
8. Remove the upper aqueous phase and transfer to new Lo-bind tube
9. Add 0.1 volume of 3M sodium acetate

10. Add 3 volumes of ice cold 100% ethanol
11. Invert to mix and place in -80°C overnight
12. Centrifuge at 14,000xg for 15 min at 4°C
13. Remove supernatant
14. Wash with 1mL of 70% ethanol
15. Centrifuge at 8,000xg for 5 min at 4°C
16. Repeat steps 13-15 once more
17. Remove supernatant with pipette and vacuum pump
18. Air dry pellet until nearly transparent
19. Resuspend in 40µL UPW
20. Nanodrop

Part 3: Outer DNA PCR

1. For each sample:

	1x
10µM FVH1 For	5µL
10µM Rev RT	5µL
Extracted DNA	40µL
Phusion-HF 2x Master Mix	50µL

2. Split each mix into two 200µL tubes
3. Run PCR reaction with the following conditions:

   98°C for 30s, [98°C for 10s, 60°C for 10s, 72°C for15s] x 15 cycles, 72°C for 5 min, 4°C hold

Bead cleanup of outer PCR

4. Pool PCR reactions together into Lo-Bind tube
5. Add 0.95x the volume of the pooled sample in AMPure XP beads (95uL)
6. Pipette to mix
7. Incubate at RT for 5 min
8. Put tubes on magnet for 5 min
9. Discard supernatant
10. Wash with 1mL 85% ethanol on magnet
11. Remove supernatant
12. Repeat steps 7 and 8 one more time
13. Remove all ethanol with vacuum
14. Dry beads at RT until the beads are not shiny anymore (but are not cracked)
15. Add 40µL UPW to beads, pipette to mix
16. Incubate at RT for 5 min
17. Put tubes on magnet for 5 min
18. Transfer supernatant to new Lo-Bind tube.
19. Nanodrop

Part 4: Inner RNA PCR

1. For each sample:

	1x
10μM PCR inner For UDI00XX	5μL
10μM PCR inner Rev UDI00XX	5μL
Outer DNA PCR product	40μL
Phusion-HF 2x Master Mix	50μL

Note: PCR primers have indices and P5/P7 adapter sequences, use different For/Rev primers for each sample to be sequenced

2. Run PCR reaction with the following conditions:

98°C for 30s, [98°C for 10s, 60°C for 10s, 72°C for15s] x 15 cycles, 72°C for 5 min, 4°C hold

Bead cleanup of inner PCR

3. Transfer reaction to Lo-Bind tube
4. Add 0.95x the volume of the PCR reaction in AMPure XP beads
5. Pipette to mix
6. Incubate at RT for 5 min
7. Put tubes on magnet for 5 min
8. Transfer supernatant to new Lo-Bind tube
9. Add 0.85x the volume
10. Wash with 1mL 85% ethanol on magnet
11. Remove supernatant
12. Repeat steps 7 and 8 one more time

13. Remove all ethanol with vacuum
14. Dry beads at RT until the beads are not shiny anymore (but are not cracked)
15. Add 40μL UPW to beads, pipette to mix
16. Incubate at RT for 5 min
17. Put tubes on magnet for 5 min
18. Transfer supernatant to new Lo-Bind tube.
19. Measure concentration by Qubit
20. Run 1μL of sample on the Bioanalyzer (should look similar to RNA inner PCR)
21. Send for Illumina NovaSeq PE100 Sequencing, ideally 10-100x sequencing depth per barcode

The above provides a detailed protocol for how the library was created, and how we isolate the plasmid DNA and barcode mRNA from the chick limb bud. This work will be incorporated into a protocol paper.

Pilot Study - Testing putative enhancers identified as conserved or accelerated when comparing flying and flightless birds.

In our initial study, we used the MPRA assay in the developing chicken limb bud to study the role of sequences changes identified in a comparative genomic analysis on enhancer activity. A previous study by Sackton et al. studied the genomic changes involved in the evolution of loss of flight. They closely examined the genomes of many flying birds and ratites, a clade of birds that includes many well-known flightless birds, such as the ostrich, rhea, and emu. They hypothesized that genomic regions highly conserved among flying birds, including some flying ratites, but with

highly accelerated mutation rates in flightless ratites, could be contributing to loss of flight. While the genomic data provides an excellent system for comparative genomic analysis, functional studies in ratites is challenging. The chick embryo, which is a flying bird, provides an ideal system to gain functional genomic data, such as epigenetic datasets, and for functional validation of candidate enhancers. These ratite-accelerated regions were therefore compared to embryonic chicken forelimb ATAC and ChIP-Seq peaks to find regions that could be putative enhancers. 54 candidate enhancers were discovered, including one region where the chicken and flying tinamou sequences drove strong expression, compared to weak expression in the flightless rhea version (Sackton et al., 2019). The rhea version of this region was accelerated, suggesting that these mutations could be driving functional divergence of enhancer activity.

Following this initial analysis of one enhancer, we wanted to use these datasets to identify enhancers conserved in limb development across birds and that may be involved in the loss of flight. I developed and performed a chicken limb bud MPRA, as described earlier in Chapter 2. The chicken and emu genomes were used to compare flying and flightless birds, as the chicken is the most commonly experimentally studied flying bird, while the emu is a flightless ratite that has also been used experimentally in limb studies (Young et al., 2019). This initial study tested a library containing 250 conserved regions between chicken and emu within chicken forelimb ATAC peaks, 200 emu-accelerated regions within chicken forelimb ATAC peaks, and 50 ratite-accelerated regions. For each of these regions, both the chicken and emu sequences were included in the library, totaling to 1000 unique sequences (Figure 10A). Following construction of the library, as described earlier in the protocol, I detected 949 out of the 1000 sequences indicating that the library construction was successful, and we have a library with almost all the sequences we wanted. 51 of the sequences that we targeted were not associated with barcodes, due to

limitations on the overall complexity of the library itself. Overall, each putative enhancer sequence within the library was associated with, on average, nine barcodes per genomic region, totaling a library of 8,452 unique enhancer-barcode members.

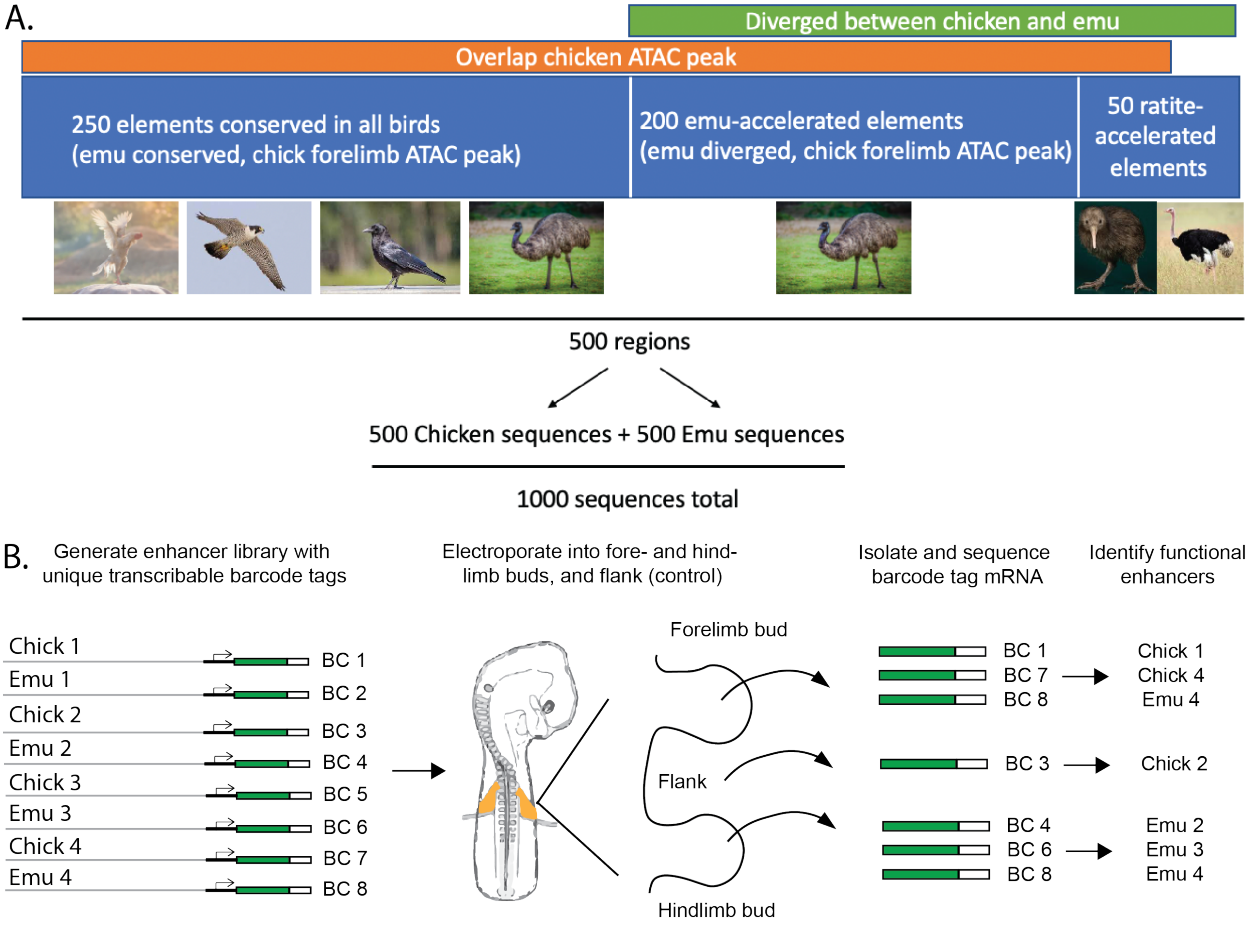


Figure 10. Experimental overview of bird library screen. A. The bird library consists of 250 regions conserved across all birds, 200 emu-accelerated regions, and 50 ratite-accelerated elements, for a total of 500 regions. For each of these regions, the chicken and emu sequences were included in the library, for a total of 1000 sequences. **B.** The bird library containing both chicken and emu sequences of each region was electroporated into forelimb buds, hindlimb buds, and flanks of chicken embryos. These tissues were then collected at stage HH21, and the mRNA was isolated to determine which barcodes were transcribed. These barcodes were then used to identify the functional genomic regions in the library.

The experiment is outlined in Figure 10B. The library was electroporated into the chicken forelimb bud, hindlimb bud, and flank. The flank was used as a control tissue, as both the limb buds and the flank are mesenchymal tissues (Damon et al., 2008). Thus, enhancers driving general

mesenchymal expression should drive expression in all three tissues, compared to limb bud mesenchyme-specific enhancers. The forelimb, hindlimb, and flank were then dissected out, and the mRNA was isolated. Active enhancers are identified by the transcribed barcodes detected in the assay. Three biological replicates of this experiment were performed.

From this initial bird library enhancer MPRA, 93 active enhancers were identified in the forelimbs and 49 active enhancers were identified in the hindlimbs based on their enhancer activity scores (Figure 11A). Because most of the regions we targeted were derived from chicken forelimb ATAC peaks, it was reassuring to see more forelimb enhancers. To further look at these enhancers, activities between chicken and emu sequences of the same region were compared. We noticed that these regions grouped into three categories. Many of the sequences that were conserved between species also had conserved activity levels (Figure 11B). Emu-accelerated sequences sometimes drove differential activity (Figure 11C). Interestingly, there was a cohort of regions where the sequences were conserved between chicken and emu, but the activity levels were different. The sequence conservation in these regions is not 100%, and thus the very small differences in these elements must be associated with the expression changes. In the future, it will be interesting to decipher what few sequence differences in these regions are sufficient to drive such different activities. Overall, from this initial vertebrate MPRA in chicken limb buds, we were able to identify many novel enhancers, both in the forelimb and hindlimb. The most surprising result was the identification of enhancers with largely conserved sequence but highly differential activity. We hope that this initial study will lead to insight into how limb enhancers are regulated and how changes within enhancers can modify enhancer activity. These studies will also provide insight into the efficacy of using sequence conservation and acceleration to predict functional changes in enhancers.

My primary goal in chapter 2 is to create a working protocol for MPRAs in the chick limb bud that can be used for further studies. The results discussed here are preliminary and further replicates and optimizations of this screen are required to have full confidence in the data and results. Validation of the interesting enhancers identified in this screen will greatly bolster our data.

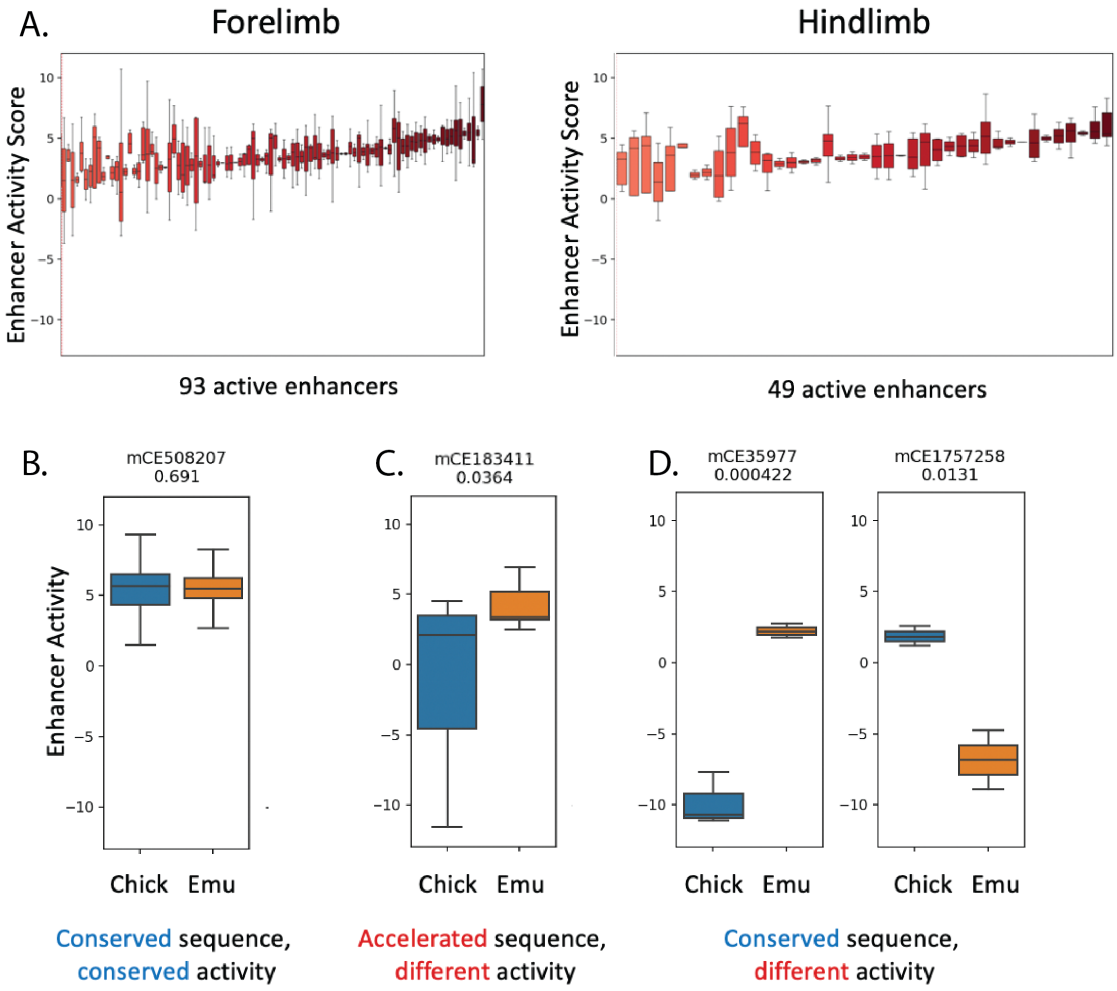


Figure 11. Limb enhancers identified from bird library enhancer screen. A. Enhancer activity scores were calculated for all regions detected. 93 active enhancers were detected in the forelimb and 49 active enhancers were detected in the hindlimbs. **B-D.** Activity of chicken and emu sequences for each region were compared. In B, mCE508207 has conserved sequence between chicken and emu and shows conserved activity. In C, mCE183411 has accelerated mutation rate in emu and shows differential activity. In D, mCE35977 and mCE1757258 have conserved sequences but show differential activity.

References

Arnold, C.D., Gerlach, D., Spies, D., Matts, J.A., Sytnikova, Y.A., Pagani, M., Lau, N.C., Stark, A., 2014. Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. Nat. Genet. 46, 685–692. https://doi.org/10.1038/ng.3009

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., Greenleaf, W.J., 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10, 1213–1218. https://doi.org/10.1038/nmeth.2688

Carroll, S., 2015. Gain of cis-regulatory activities underlies novel domains of wingless gene expression in Drosophila.

Crawford, G.E., Holt, I.E., Mullikin, J.C., Tai, D., National Institutes of Health Intramural Sequencing Center‡, Green, E.D., Wolfsberg, T.G., Collins, F.S., 2004. Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. Proc. Natl. Acad. Sci. 101, 992–997. https://doi.org/10.1073/pnas.0307540100

Damon, B.J., Mezentseva, N.V., Kumaratilake, J.S., Forgacs, G., Newman, S.A., 2008. Limb bud and flank mesoderm have distinct "physical phenotypes" that may contribute to limb budding. Dev. Biol. 321, 319–330. https://doi.org/10.1016/j.ydbio.2008.06.018

de Boer, C.G., Vaishnav, E.D., Sadeh, R., Abeyta, E.L., Friedman, N., Regev, A., 2020. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. Nat. Biotechnol. 38, 56–65. https://doi.org/10.1038/s41587-019-0315-8

Dinger, M.C., Beck-Sickinger, A.G., 2002. The First Reporter Gene Assay on Living Cells Green Fluorescent Protein as Reporter Gene for the Investigation of Gi-Protein Coupled Receptors. Mol. Biotechnol. 21, 009–018. https://doi.org/10.1385/MB:21:1:009

Farley, E.K., Olson, K.M., Zhang, W., Brandt, A.J., Rokhsar, D.S., Levine, M.S., 2015. Suboptimization of developmental enhancers. Science 350, 325–328. https://doi.org/10.1126/science.aac6948

Gordon, M.G., Inoue, F., Martin, B., Schubach, M., Agarwal, V., Whalen, S., Feng, S., Zhao, J., Ashuach, T., Ziffra, R., Kreimer, A., Georgakopoulos-Soares, I., Yosef, N., Ye, C.J., Pollard, K.S., Shendure, J., Kircher, M., Ahituv, N., 2020. lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. Nat. Protoc. 15, 2387–2412. https://doi.org/10.1038/s41596-020-0333-5

Hakkila, K., Maksimow, M., Karp, M., Virta, M., 2002. Reporter Genes lucFF, luxCDABE, gfp, and dsred Have Different Characteristics in Whole-Cell Bacterial Sensors. Anal. Biochem. 301, 235–242. https://doi.org/10.1006/abio.2001.5517

Inoue, F., Kircher, M., Martin, B., Cooper, G.M., Witten, D.M., McManus, M.T., Ahituv, N., Shendure, J., 2017. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. Genome Res. 27, 38–52. https://doi.org/10.1101/gr.212092.116

Johnson, D.S., Mortazavi, A., Myers, R.M., Wold, B., 2007. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science 316, 1497–1502. https://doi.org/10.1126/science.1141319
King, D.M., Hong, C.K.Y., Shepherdson, J.L., Granas, D.M., Maricque, B.B., Cohen, B.A., 2020. Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in mouse embryonic stem cells. eLife 9, e41279. https://doi.org/10.7554/eLife.41279

Levine, M., 2010. Transcriptional enhancers in animal development and evolution. Curr. Biol. CB 20, R754-763. https://doi.org/10.1016/j.cub.2010.06.070

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A., Lander, E.S., Dekker, J., 2009. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science 326, 289–293. https://doi.org/10.1126/science.1181369

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C.N., Slatkin, M., Valle, D., Whittemore, A.S., Boehnke, M., Clark, A.G., Eichler, E.E., Gibson, G., Haines, J.L., Mackay, T.F.C., McCarroll, S.A., Visscher, P.M., 2009. Finding the missing heritability of complex diseases. Nature 461, 747–753. https://doi.org/10.1038/nature08494

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A.K., Canfield, T.K., Giste, E., Diegel, M., Bates, D., Hansen, R.S., Neph, S., Sabo, P.J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S.R., Kaul, R., Stamatoyannopoulos, J.A., 2012. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science 337, 1190–1195. https://doi.org/10.1126/science.1222794

McQueen, C., Towers, M., 2020. Establishing the pattern of the vertebrate limb. Development 147, dev177956. https://doi.org/10.1242/dev.177956

Menke, D.B., Guenther, C., Kingsley, D.M., 2008. Dual hindlimb control elements in the *Tbx4* gene and region-specific control of bone size in vertebrate limbs. Development 135, 2543–2553. https://doi.org/10.1242/dev.017384

Sabo, P.J., Humbert, R., Hawrylycz, M., Wallace, J.C., Dorschner, M.O., McArthur, M., Stamatoyannopoulos, J.A., 2004. Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. Proc. Natl. Acad. Sci. 101, 4537–4542. https://doi.org/10.1073/pnas.0400678101

Sackton, T.B., Grayson, P., Cloutier, A., Hu, Z., Liu, J.S., Wheeler, N.E., Gardner, P.P., Clarke, J.A., Baker, A.J., Clamp, M., Edwards, S.V., 2019. Convergent regulatory evolution and loss of flight in paleognathous birds. Science 364, 74–78. https://doi.org/10.1126/science.aat7244

Shen, S.Q., Myers, C.A., Hughes, A.E.O., Byrne, L.C., Flannery, J.G., Corbo, J.C., 2016. Massively parallel *cis* -regulatory analysis in the mammalian central nervous system. Genome Res. 26, 238–255. https://doi.org/10.1101/gr.193789.115

Skene, P.J., Henikoff, J.G., Henikoff, S., 2018. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. Nat. Protoc. 13, 1006–1019. https://doi.org/10.1038/nprot.2018.015

Song, B.P., Ragsac, M.F., Tellez, K., Jindal, G.A., Grudzien, J.L., Le, S.H., Farley, E.K., 2022. Diverse logics and grammar encode notochord enhancers (preprint). Developmental Biology. https://doi.org/10.1101/2022.07.25.501440

Tak, Y.G., Farnham, P.J., 2015. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. Epigenetics Chromatin 8, 57. https://doi.org/10.1186/s13072-015-0050-4

Tomizawa, R.R., Tabin, C.J., Atsuta, Y., 2022. *In ovo* electroporation of chicken limb bud ectoderm: Electroporation to chick limb ectoderm. Dev. Dyn. 251, 1628–1638. https://doi.org/10.1002/dvdy.352

Tournamille, C., Colin, Y., Cartron, J.P., Le Van Kim, C., 1995. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy–negative individuals. Nat. Genet. 10, 224–228. https://doi.org/10.1038/ng0695-224

Visel, A., Rubin, E.M., Pennacchio, L.A., 2009. Genomic views of distant-acting enhancers. Nature 461, 199–205. https://doi.org/10.1038/nature08451

Young, J.J., Grayson, P., Edwards, S.V., Tabin, C.J., 2019. Attenuated Fgf Signaling Underlies the Forelimb Heterochrony in the Emu Dromaius novaehollandiae. Curr. Biol. 29, 3681-3691.e5. https://doi.org/10.1016/j.cub.2019.09.014

Acknowledgements

Chapter 2 contains unpublished material coauthored with Zhu, Meng and Solvason, Joseph J. The dissertation author was the primary author of this chapter. Special thanks to the Farley lab for helpful discussions. Special thanks to Timothy Sackton for help in designing the library of

genomic elements for testing. Special thanks to Sophia Le for help with cloning the library vector. Special thanks to the UCSD IGM Genomics Center for their assistance with sequencing.

DISCUSSION

In this dissertation, I have described two MPRA enhancer screens in developing embryos. In the first enhancer screen, I sought to understand notochord enhancer regulation by testing 90 genomic regions of *Ciona robusta* with Zic and ETS transcription factor binding sites. Interestingly, only nine of the 90 tested drove notochord enhancers. Among these nine, I identified a *laminin alpha* enhancer that was highly dependent on grammatical constraints for proper expression. I found a similar cluster of Zic and ETS sites within the intron of the mouse and human *laminin alpha-1* gene; strikingly, these clusters and the *Ciona* laminin enhancer have the same spacing between the Zic and ETS sites. Within the BraS enhancer, I demonstrated that newly identified FoxA and Bra sites are necessary for notochord expression and determined that the five TFBSs together in BraS (Zic, 2 ETS, FoxA, and Bra) are sufficient for notochord expression by creating a library of 45 million BraS variants in which all five TFBSs are kept constant in position, and affinity while all other nucleotides are randomized. I find that the combination of Zic, ETS, FoxA, Bra occurs within other *Bra* enhancers in *Ciona* and vertebrates suggesting this combination of TFs may be a common logic regulating *Bra* expression. This study identifies new developmental enhancers, demonstrates the importance of enhancer grammar within developmental enhancers and provides a deeper understanding of the regulatory logic governing *Bra*. These findings of the same clusters of sites within vertebrates hint at the conserved role of grammar and logic across chordates.

In the second enhancer screen, I focused on developing a vertebrate enhancer MPRA in the chicken limb bud. The limb bud is an ideal system for a vertebrate MPRA because tissue-specificity can be easily investigated, as there can be specificity differences between the forelimb and hindlimb, as well as the limb mesenchyme and general mesenchyme, when the flank tissue is

used as a control. Additionally, electroporation of plasmids into the chicken limb is established, allowing for a method to assay many putative enhancers at once. Using this newly developed MPRA, we investigated differential activity of enhancers identified as conserved or accelerated in the developing bird limb bud of flying and flightless birds. Previous studies comparing the genomes of flying and flightless birds identified non-coding regions that contained conserved or accelerated sequences. Overlaying these regions with chicken forelimb ATAC peaks identified putative enhancers, and we tested the activity of 1000 of these conserved or accelerated putative enhancers between chicken and emu implicated in the loss of flight. From this screen, we identified many new forelimb and hindlimb enhancers, including enhancers with highly conserved sequence between chick and emu, but differential activity. While my main goal was to develop this vertebrate MPRA in chicken limb buds, future experiments will further validate these enhancers and provide insight into how enhancers control gene expression during limb development.

Enhancer randomization is a valuable tool to test for sufficiency

Enhancers are regulated by the binding of transcription factors to DNA sequences known as transcription factor binding sites. In Chapter 1, we searched for the transcription factor binding sites Zic, ETS, FoxA, and Bra in 90 genomic regions of the Ciona genome. To demonstrate that these sequences are important, necessity experiments are often performed where binding sites are mutated. Mutations in binding sites that lead to reduced or ablated enhancer activity demonstrate the necessity of that binding site for proper enhancer expression. In addition, many mutagenesis approaches have been used to identify important sequences. For example, in linker-scanning mutagenesis, small blocks of an enhancer are sequentially mutated to identify which nucleotides are contributing to expression (Greene, 1991). Similarly, in saturation mutagenesis, each

nucleotide in a sequence is mutated (Kircher et al., 2019; Patwardhan et al., 2012). These methods inform us of which nucleotides are contributing to expression.

The counterpart to these mutagenesis experiments is sufficiency, which tells us whether the previously identified important sequences in an enhancer can drive gene expression on their own. However, this is much trickier to interrogate. A common technique is to substitute the binding motifs of interest into a new genetic sequence and evaluate the expression levels (Grossman et al., 2017). However, each genetic background is also an opportunity for spurious binding sites that could confound the results. Thus, our lab has pioneered the use of a randomized library to test for sufficiency of transcription factor binding sites, as described in Chapter 1. In this method, we utilize millions of synthetic variants surrounding a fixed set of binding site sequences. This pool is then assayed all at once, such that the individual contribution of a single variant is minimized, while the fixed set of binding sites should be the only significant contributor. In Chapter 1, using this method we found that that Zic and ETS sites alone were not sufficient for notochord expression in BraS. However, inclusion of FoxA and Bra sites, along with the Zic and ETS sites, did demonstrate sufficiency. Thus, we demonstrated the power of this randomization tool to identify sufficient binding sites for an enhancer. This is an unbiased method to test for sufficiency of previously identified binding sites, such as from mutagenesis screens, and we hope that enhancer randomization will be adopted as a gold standard to demonstrate sufficiency of binding sites for an enhancer.

Enhancers are expressed temporally

Enhancers activate gene expression both in specific tissues and with precise timings. In the MPRAs described in this dissertation, each experiment was performed at a single

timepoint. It is possible, then, that different timepoints for collection could uncover previously unnoticed enhancers. For example, many genes specifically expressed in the secondary notochord, or the posterior notochord, only show this expression pattern late in development (Reeves et al., 2014). This could be due to the fact that many transcription factors are part of gene networks that activate in a temporal manner, such that these genes, and the enhancers that drive their expression, are only activated later during development. For example, Xbp1 is a transcription factor that is activated by Brachyury, and many of its target genes are active only after neurulation (Wu et al., 2022). Enhancers that require Xbp1 to drive expression would, therefore, go undetected when assaying for enhancer activity at the gastrula stage. Within our notochord studies, we see that ablation of Zic and ETS sites in the Brachyury Shadow enhancer leads to complete loss of expression. Zic and ETS are early transcription factors activating Brachyury expression (Matsumoto et al., 2007). In comparison, ablation of FoxA or Bra leads to a significant, but not total, loss of expression, which may suggest that these transcription factors my function later to maintain Brachyury Shadow activity. Further experiments to ablate combinations of FoxA, ETS, and Zic would be necessary to discern the role of these factors in activation vs. maintenance of the Brachyury Shadow enhancer.

Overall, the temporal aspect of enhancers is often underappreciated. Thus, it would be fruitful to see whether performing the MPRAs discussed in this at different times during development would identify new enhancers. Comparison of enhancers that show expression at one timepoint, but not another, would allow us to better understand what sequences in these enhancers are regulating this temporal difference.

MPRAs of genomic regions are challenging

Identifying enhancers from genomic regions has yielded mixed results. In flies, the use of high-affinity binding sites clusters and histone markers has been successful in predicting enhancers (Berman et al., 2002; Rebeiz et al., 2002). In the modENCODE project, use of CREB binding protein binding sites combined with histone marks led to identifying enhancers quite successfully (30/33 putative enhancers drive expression) (Nègre et al., 2011). However, the use of chordate genomic regions in MPRAs to try to understand enhancer regulation has proven challenging, as most of the genomic regions tested have not driven gene expression (King et al., 2020; Song et al., 2022). It is likely that evolution has selected against activity of most genomic regions, as spurious transcription of genes could be damaging to cellular integrity. Furthermore, among the few genomic regions that do drive expression, it is difficult to find patterns that are driving similar expression patterns. Each genomic region has a different order, orientation and spacing between their binding sites, and the sequences between binding sites are also completely different. Thus, the sheer variance between each individual genomic region makes identification of rules of enhancer regulation difficult to study. Perhaps more stringent constraints, such as high-affinity binding sites, or more well-defined grammar rules may lead to greater success in identifying active enhancers in chordates.

MPRAs are a powerful tool to test thousands to millions of sequences at once, allowing us to slowly gain knowledge about how DNA sequences encode their function. In the future, I believe we will see a combination of synthetic, random and genomic enhancer screens that will enable patterns that govern tissue-specific expression to be computationally deciphered. Furthermore, analysis of variants with ectopic expression could result in the discovery of novel enhancer logics driving enhancer expression, driving a cascading effect toward unraveling the mystery of the gene

regulatory code. With the rapid rate of advancement in sequencing, oligo synthesis, and computational tools, I believe these futuristic MPRAs may be closer than they seem.

Conserved sequence does not always mean conserved function

It is often assumed that highly conserved enhancers drive similar expression in related species. However, from our chicken limb bud MPRA, we noted that a surprising number of enhancers with conserved sequence between chicken and emu had significantly different activity. This suggests that even in highly conserved sequences, small changes can have dramatic effects. Previous studies have noticed a similar pattern in the ZRS limb enhancer, where point mutations can lead to extra digits in the limbs (Albuisson et al., 2012; Lim et al., 2022). However, there are also many cases where highly conserved enhancers are robust to small levels of mutagenesis (Dickel et al., 2018). Thus, it is important to functionally test enhancers to identify why certain changes, but not others, lead to functional consequences.

Advancements during this dissertation and Future Directions

Many studies have utilized *Ciona* as a model system to study enhancer regulation (Bertrand et al., 2003; Corbo et al., 1997; Imai et al., 2006; José-Edwards et al., 2015; Reeves et al., 2021). However, most of these examine a small set of enhancers. In the largest study I found, 19 unique genomic regions were assayed (Brown et al., 2007). Thus, our MPRA of 90 elements is the largest screen of genomic regions in *Ciona* so far. From this screen, we identified eight novel notochord enhancers that can be further examined to better understand what sequences drive notochord expression.

The 90 elements selected from our screen came from a pool of 1092 genomic regions containing Zic and ETS. As 10% of the regions we tested drove notochord expression, a larger

screen encompassing all 1092 regions would theoretically yield 100 new notochord enhancers. This would allow for stronger identification of patterns driving notochord enhancers, such as organizations of binding sites that are most optimal.

Additionally, we were able to identify signatures of a conserved enhancer logic governing Brachyury regulation. While Zic, ETS, FoxA, and Bra were all known to be independently. important in driving notochord expression in *Ciona*, the combination of these four factors together had not been fully appreciated, especially in the context of the Brachyury Shadow enhancer. Furthermore, no studies in vertebrates have previously linked Zic to notochord regulation, despite studies suggesting Zic is expressed in the early stages of vertebrate notochord development (Dykes et al., 2018; Warr et al., 2008). Thus, our discovery of the combination of Zic, ETS, FoxA, and Bra binding sites in zebrafish and mouse Brachyury enhancers opens an avenue to investigate how these transcription factors interact to drive notochord expression.

The chicken limb bud screen we developed is the first vertebrate enhancer MPRA in a developing embryo. Using this screen, we identified many novel forelimb and hindlimb-specific enhancers, and enhancers with highly conserved sequence but differential activity. Further validation of these enhancers by reporter testing will be needed to confirm these findings; however, these enhancers could be used to better understand enhancer regulation governing limb development. Additionally, transcription factors governing dorsal/ventral patterning of the limb have been studied (Altabef and Tickle, 2002), but enhancers that drive these patterns of expression have remained elusive. Thus, further refinement of this chicken limb bud enhancer screen by dorsal/ventral dissection of the limbs could begin to provide insight into the mechanisms of limb patterning.

Collectively, my studies in the *Ciona* and chicken embryo have provided novel techniques to study enhancers and shed insight into the role of enhancer grammar in encoding tissue-specific enhancers. Future studies implementing these MPRA approaches will bring us closer to deciphering the grammatical constraints on tissue-specific enhancers across chordates.

References

Albuisson, J., Schmitt, S., Baron, S., Bézieau, S., Benito-Sanz, S., Heath, K.E., 2012. Clinical utility gene card for: Leri-Weill dyschondrosteosis (LWD) and Langer mesomelic dysplasia (LMD). Eur. J. Hum. Genet. 20, 3–4. https://doi.org/10.1038/ejhg.2012.64

Altabef, M., Tickle, C., 2002. Initiation of dorso-ventral axis during chick limb development. Mech. Dev. 116, 19–27. https://doi.org/10.1016/S0925-4773(02)00125-9

Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., Eisen, M.B., 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. Proc. Natl. Acad. Sci. 99, 757–762. https://doi.org/10.1073/pnas.231608898

Bertrand, V., Hudson, C., Caillol, D., Popovici, C., Lemaire, P., 2003. Neural Tissue in Ascidian Embryos Is Induced by FGF9/16/20, Acting via a Combination of Maternal GATA and Ets Transcription Factors. Cell 115, 615–627. https://doi.org/10.1016/S0092-8674(03)00928-0

Brown, C.D., Johnson, D.S., Sidow, A., 2007. Functional Architecture and Evolution of Transcriptional Elements That Drive Gene Coexpression. Science 317, 1557–1560. https://doi.org/10.1126/science.1145893

Corbo, J.C., Levine, M., Zeller, R.W., 1997. Characterization of a notochord-specific enhancer from the Brachyury promoter region of the ascidian, Ciona intestinalis. Dev. Camb. Engl. 124, 589–602. https://doi.org/10.1242/dev.124.3.589

Dickel, D.E., Ypsilanti, A.R., Pla, R., Zhu, Y., Barozzi, I., Mannion, B.J., Khin, Y.S., Fukuda-Yuzawa, Y., Plajzer-Frick, I., Pickle, C.S., 2018. Ultraconserved enhancers are required for normal development. Cell 172, 491–499.

Dykes, I.M., Szumska, D., Kuncheria, L., Puliyadi, R., Chen, C., Papanayotou, C., Lockstone, H., Dubourg, C., David, V., Schneider, J.E., Keane, T.M., Adams, D.J., Brown, S.D.M., Mercier, S., Odent, S., Collignon, J., Bhattacharya, S., 2018. A Requirement for Zic2 in the Regulation of Nodal Expression Underlies the Establishment of Left-Sided Identity. Sci. Rep. 8, 10439. https://doi.org/10.1038/s41598-018-28714-1

Greene, J.M., 1991. Linker-Scanning Mutagenesis of DNA. Curr. Protoc. Mol. Biol. 13. https://doi.org/10.1002/0471142727.mb0804s13

Grossman, S.R., Zhang, X., Wang, L., Engreitz, J., Melnikov, A., Rogov, P., Tewhey, R., Isakova, A., Deplancke, B., Bernstein, B.E., Mikkelsen, T.S., Lander, E.S., 2017. Systematic dissection of genomic features determining transcription factor binding and enhancer function. Proc. Natl. Acad. Sci. 114. https://doi.org/10.1073/pnas.1621150114

Imai, K.S., Levine, M., Satoh, N., Satou, Y., 2006. Regulatory Blueprint for a Chordate Embryo. Science 312, 1183–1187. https://doi.org/10.1126/science.1123404

José-Edwards, D.S., Oda-Ishii, I., Kugler, J.E., Passamaneck, Y.J., Katikala, L., Nibu, Y., Di Gregorio, A., 2015. Brachyury, Foxa2 and the cis-Regulatory Origins of the Notochord. PLoS Genet. 11, e1005730. https://doi.org/10.1371/journal.pgen.1005730

King, D.M., Hong, C.K.Y., Shepherdson, J.L., Granas, D.M., Maricque, B.B., Cohen, B.A., 2020. Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in mouse embryonic stem cells. eLife 9, e41279. https://doi.org/10.7554/eLife.41279

Lim, F., Ryan, G.E., Le, S.H., Solvason, J.J., Steffen, P., Farley, E.K., 2022. Affinity-optimizing variants within the ZRS enhancer disrupt limb development (preprint). Genetics. https://doi.org/10.1101/2022.05.27.493789

Matsumoto, J., Kumano, G., Nishida, H., 2007. Direct activation by Ets and Zic is required for initial expression of the Brachyury gene in the ascidian notochord. Dev. Biol. 306, 870–882. https://doi.org/10.1016/j.ydbio.2007.03.034

Nègre, N., Brown, C.D., Ma, L., Bristow, C.A., Miller, S.W., Wagner, U., Kheradpour, P., Eaton, M.L., Loriaux, P., Sealfon, R., Li, Z., Ishii, H., Spokony, R.F., Chen, J., Hwang, L., Cheng, C., Auburn, R.P., Davis, M.B., Domanus, M., Shah, P.K., Morrison, C.A., Zieba, J., Suchy, S., Senderowicz, L., Victorsen, A., Bild, N.A., Grundstad, A.J., Hanley, D., MacAlpine, D.M., Mannervik, M., Venken, K., Bellen, H., White, R., Gerstein, M., Russell, S., Grossman, R.L., Ren, B., Posakony, J.W., Kellis, M., White, K.P., 2011. A cis-regulatory map of the Drosophila genome. Nature 471, 527–531. https://doi.org/10.1038/nature09990

Rebeiz, M., Reeves, N.L., Posakony, J.W., 2002. SCORE: A computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. Proc. Natl. Acad. Sci. 99, 9888–9893. https://doi.org/10.1073/pnas.152320899

Reeves, W., Thayer, R., Veeman, M., 2014. Anterior-posterior regionalized gene expression in the *Ciona* notochord: Differential Expression in the *CIONA* Notochord. Dev. Dyn. 243, 612–620. https://doi.org/10.1002/dvdy.24101

Reeves, W.M., Shimai, K., Winkley, K.M., Veeman, M.T., 2021. Brachyury controls Ciona notochord fate as part of a feed-forward network. Dev. Camb. Engl. 148, dev195230. https://doi.org/10.1242/dev.195230

Song, B.P., Ragsac, M.F., Tellez, K., Jindal, G.A., Grudzien, J.L., Le, S.H., Farley, E.K., 2022. Diverse logics and grammar encode notochord enhancers (preprint). Developmental Biology. https://doi.org/10.1101/2022.07.25.501440

Warr, N., Powles-Glover, N., Chappell, A., Robson, J., Norris, D., Arkell, R.M., 2008. Zic2 - associated holoprosencephaly is caused by a transient defect in the organizer region during gastrulation. Hum. Mol. Genet. 17, 2986–2996. https://doi.org/10.1093/hmg/ddn197

Wu, Y., Devotta, A., José-Edwards, D.S., Kugler, J.E., Negrón-Piñeiro, L.J., Braslavskaya, K., Addy, J., Saint-Jeannet, J.-P., Di Gregorio, A., 2022. Xbp1 and Brachyury establish an evolutionarily conserved subcircuit of the notochord gene regulatory network. eLife 11, e73992. https://doi.org/10.7554/eLife.73992